Identification of Active Molecular Sites Using Quantum-Self-Similarity Measures

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

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

Robert Ponec

Institute of Chemical Process Fundamentals, Czech Academy of Sciences, Prague 6, Suchdol 2, 165 02, Czech Republic

Received December 7, 2000

A novel approach to construct theoretical QSAR models is proposed. This technique, based on the systematic use of quantum similarity measures as theoretical molecular descriptors, opens the possibility to localize and to identify the position of the bioactive part of drug molecules in situations, where the nature of the pharmacophore is not known. To test the reliability of this new approach, the method has been applied to the study of steroids binding to corticosteroid-binding human globulin. The studied molecules involved the set of 31 Cramer's steroids, often used as a benchmark set in QSAR studies. It has been shown that theoretical QSAR models based on the present procedure are superior to those derived from alternative existing approaches. In addition, a new method to measure the statistical significance of multiparameter QSAR models is also proposed.

INTRODUCTION

In the last two decades quantum similarity theory[1−22] has increasingly been applied as a new means of the construction of theoretical QSAR models.[23−44] Some molecular similarity techniques have been focused on characterizing biological mechanisms by means of identifying molecular fragments, which are important for recognition by the enzyme.[31−34,38,39,42,43] Our own approach to quantum similarity measures (QSM) have clearly shown that the appropriate theoretical measures correlate remarkably well with empirical molecular descriptors such as log P[30] or Hammett σ constants[31,32] and, consequently, can replace them in QSAR equations. Thus, for example, the Hammett σ constants were found to correlate with the quantum self-similarity measures (QS-SM) of the fragment COOH in a series of substituted aromatic carboxylic acids,[31,32] so that the theoretical LFER equations equivalent to empirical Hammett pK vs σ plots could be straightforwardly formulated. In a similar way it was also possible to obtain alternative theoretical QSAR models for the correlation of biological activities.[33,34] Although this approach proved to be useful for many systems, it is nevertheless true that it can be straightforwardly applied only to systems in which the bioactive part of the molecule (pharmacophore) is known beforehand, so that the identification of the appropriate theoretical descriptor is self-evident. This, however, is not often the case, and in this situation the design of the appropriate QSAR model, whether empirical or theoretical, is extremely difficult. Because of the importance of these models for rational drug design, several approaches were proposed in recent years in which the problem of identification of the unknown pharmacophore was addressed. Two most commonly used approaches, GRID[45] and CoMFA,[46] are based on the calculations of interaction energies at grid points in the space surrounding the target structure. Two interaction energy contributions are computed: the steric by measures of Lennard-Jones potentials and the electrostatic based on the Coulomb potential. From these approaches other methods were derived, such as CoMSIA,[47] CoMMA,[48] or more recently SOMFA.[49]

Our aim in this study is to complement the above-mentioned techniques by a simple procedure based on the systematic use of the so-called fragment QS-SM as a source of new theoretical molecular descriptors. To demonstrate the basic idea of this approach, the method will be first applied to the dissociation of substituted benzoic acids, where the chemical process is clearly localized into the COOH group, and as it will be shown, this molecular fragment is also correctly identified as the active, reaction center by the present procedure. This molecular set was previously studied using semiempirical methods and only evaluating the correlation between the QS-SM of the fragment COOH and the σ constant.[31] Now, the study is extended to ab initio calculations, and several molecular fragments are analyzed in order to find the best ones which better correlates with the experimental values.

Based on this result, the approach is then applied to the study of the affinity of a series of 31 steroids interacting with the corticosteroid binding globuline (CBG). This steroid series, known also as the Cramer's set, is widely used as a benchmark for testing various theoretical QSAR models,[23,25,36,46−63] and it has also been used in two previous studies,[23,25] based on QSM. In the first work,[23] the theoretical QSAR models were derived from Topological Quantum Similarity Indices (TQSI) and Molecular Quantum Similarity Measures (MQSM), while in the second one,[25] the formalism was based on the so-called tuned MQSM derived from the incorporation of MQSM into the general convex set theory.[19] Within this approximation, two or more quantum similarity

* Corresponding author fax: 34 972 418356; e-mail: director@iqc.udg.es.

10.1021/ci000160u CCC: $20.00
Published on Web 06/16/2001

matrices are combined so as to optimize the precision and the predicting power of the corresponding QSAR model. However, this approach is computationally demanding since the calculation of the individual elements of quantum similarity matrices Z_{AB} requires the optimization of the relative position of the corresponding molecular pairs *A* and *B*.[17] In the present approach, these computational demands are considerably reduced by using fragment intramolecular QS-SM as molecular descriptors which allows to avoid the problems of the molecular alignment completely.

THEORETICAL

Although theoretical considerations underlying the introduction of the concept of QSM were sufficiently described in several previous studies, it is worthwhile to be reminded of briefly the basic ideas, which provide an appropriate theoretical background for the theory of empirical structure−activity relationships.[18−22] The crucial role in introducing the concept of QSM has played an effort in finding an appropriate theoretical tool for the quantitative exploitation of the old, intuitive, but for chemistry extremely fruitful idea that similar molecules also have similar properties. As electron distribution, characterized by the quantum chemically generated density function $\rho(\mathbf{r})$, is in fact the ultimate molecular descriptor, it seems quite natural to base the definition of the QSM on this simplest observable quantum chemical quantity. Within this approach, the QSM of two molecules *A* and *B*, described by the corresponding density functions $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$, can be quantitatively characterized by the value of the integral

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

where $\Omega(\mathbf{r}_1,\mathbf{r}_2)$ is an arbitrary positive definite two-electron operator which acts in the formula as a general weighting factor. Several kinds of QSM can be introduced depending on the actual choice of the operator Ω. Thus, for example, the identification of Ω with the Dirac delta function reduces the general expression to the formula

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

which represents the so-called overlap-like QSM.[1]

Another possible option is to identify Ω with the Coulomb repulsion term $\mathbf{r}_{12}^{-1}$, and in this case the general formula (1) transforms into the so-called Coulomb-like QSM:

$$Z_{AB} = \int\int \rho_A(\mathbf{r}_1)|\mathbf{r}_1 - \mathbf{r}_2|^{-1}\, \rho_B(\mathbf{r}_2)d\mathbf{r}_1\, d\mathbf{r}_2 \qquad (3)$$

Although it is in general possible to introduce also other types of QSM, only overlap-like and Coulomb-like similarity measures will be considered in this study.

Connection between MQSM and QSAR. One of the most interesting and the most important applications for the above introduced QSM consists of the possibility of their use in the design of theoretical QSAR models. This use is based on the possibility to transform the formula for the expectation value of the nondifferential operator ω associated with the measured molecular property y_I

$$y_I = \int \omega(\mathbf{r})\, \rho_I(\mathbf{r})\, d\mathbf{r} \qquad (4)$$

into a discrete form, analogous to traditional multilinear QSAR equations.[18−22] This transformation relies on the concept of molecular similarity. The basic idea of this procedure will be briefly reminded below. For this purpose, imagine that we have a series of molecules $\{A, B, \ldots N\}$ and the pairwise QSM Z_{IJ} for all possible pairs are calculated from the set. The quantities Z_{IJ} form the square $n \times n$ matrix **Z**, the so-called similarity matrix, in terms of which the formula (4) can be rewritten as

$$y_I = \sum_K c_K Z_{KI} \qquad (5)$$

where c_K are the components of the vector representing the operator ω in the discrete basis of density functions $\{\rho_I\}$. This equation, which represents in fact the discrete counterpart of the continuous formulation (4), can be regarded as the most general form of QSAR equations. Although the above formalism represents the most direct approach to the design of theoretical QSAR models, the straightforward application of this approach has one unpleasant side-effect—it is computationally very demanding. This is due to the fact that the values of the pairwise similarity measures Z_{IJ}, forming the molecular similarity matrix, depend on the distance and the mutual orientation of the corresponding molecules. This implies that, in order to get meaningful results, the position of the molecules has to be optimized so as to give the maximum similarity for each pair *I* and *J*.[17]

In this study we propose a new approach, which to a considerable extent reduces the computational requirements of the original formalism, while still retaining the sufficient accuracy of the corresponding to theoretical QSAR models. This approach avoids the lengthy process of molecular alignment by considering only the diagonal elements Z_{II} of the similarity matrix **Z**. These elements, the so-called quantum self-similarity measures (QS-SM), play, within this simplified approach, the role of theoretical QSAR descriptors, and as it has been shown in several previous published papers,[30−33] these QS-SM can be successfully used as an alternative to traditional QSAR descriptors. Thus, for example, the QS-SM Z_{II} was found to correlate with the molecular hydrophobicity empirical descriptor, the log P.[30] Similarly, it was also possible to describe the substituent effect, traditionally characterized by the Hammett σ constant, by the so-called fragment QS-SM.[31,32]

Fragment Self-Similarity Measures. The definition of these QSM is analogous to the original formula (1) from which it differs only in that the electron densities of an appropriate fragment *X* are compared instead of the total densities of the whole molecule

$$Z_{II}^X(\Omega) = \int\int \rho_I^X(\mathbf{r}_1)\Omega(\mathbf{r}_1,\mathbf{r}_2)\rho_I^X(\mathbf{r}_2)d\mathbf{r}_1\, d\mathbf{r}_2 \qquad (6)$$

The choice of the appropriate fragment is always a matter of certain arbitrariness, but intuitively one feels that the best chance to describe properly the effect of the systematic structural variation on a certain molecular property is when the corresponding fragment QS-SM is associated with the functional group responsible in any given case for the observed property. The correctness of this intuition was clearly confirmed in several previous studies, where the original empirical QSAR models were substituted by the

corresponding to theoretical counterparts.[30−33] However, such a straightforward approach is applicable only for systems and processes, in which the fragment responsible for the observed activity is known. An example in this respect can be constituted by the study of the biological activity of substituted phenylisothiocyanates[31] and benzensulfonamides,[33] whose activity is most probably due to the presence of NCS or SO_2NH_2 groups, respectively. Unfortunately such a situation is not often the case and our aim here is to propose a simple general method, allowing for the identification and the localization of the position of the active fragment in the molecule in any given case. The method is based on the generation of various molecular fragments and on the subsequent evaluation of the quality of all possible QSAR models, based on the use of the corresponding fragment QS-SM as theoretical molecular descriptors. In this initial development, the proposed technique only applies to molecules which present common structural features, but it can be adapted to structurally diverse molecules following some schemes given in the literature.[42]

The QS-SM QSAR models are generated using the following systematic procedure:

1. The set of considered molecular fragments is defined. In this study, the fragments were defined as a group of several (1, 2, or 3) neighboring atoms contributing to the molecular skeleton. The corresponding fragment electron densities, *X*, can then be obtained from the total electron density of the whole molecule *I*

$$\rho_I(\mathbf{r}) = \sum_{\mu} \sum_{\nu} D_{\mu\nu}\, \chi^*_{\mu}(\mathbf{r})\chi_{\nu}(\mathbf{r}) \tag{7}$$

by appropriately restricting the summations over the basis functions. Thus for example, if the molecular fragment *X* consists of atoms a_1, a_2, ... a_n, then the corresponding density is given by

$$\rho^X_I(\mathbf{r}) = \sum_{i}^{n} \sum_{\mu \in a_i} \sum_{\nu \in I} D_{\mu\nu}\, \chi^*_{\mu}(\mathbf{r})\chi_{\nu}(\mathbf{r}) \tag{8}$$

Based on these densities, the associated fragment QS-SM are computed according to eq 6.

2. Having defined the set of *m* molecular fragments and the corresponding QS-SM for each fragment, the next step consists of the systematic evaluation of the quality of all possible QSAR models based on the fragment QS-SM as descriptors. Here it is necessary to stress that the generated QSAR models do not need to be only one-dimensional but can be constructed and analyzed using any of the available multilinear correlation equations, with the number of independent descriptors ranging from 1 to any selected value *k* ($k < m$). The total number of *k*-parameter correlation equations emerging from this scheme is $\binom{m}{k}$. All these alternatives were systematically generated using a nested summation algorithm.[64,65] The quality of any individual correlation was characterized by the value of the regression coefficient *r*. In addition, and in order to estimate the predictive power of the model, cross-validation (CV) following the leave-one-out (LOO) scheme was performed. The representation of experimental values against the cross-validated values gives the cross-validation regression coefficient r_{cv}. Recently, in our laboratory, a new basic approach to directly obtain r_{cv} has been employed. See the Appendix for more details.

Evaluation of the Statistical Significance of Generated Theoretical QSAR Models. As it was stressed above, the proposed method of detection and localization of active molecular fragments is based on the evaluation of the quality of all systematically generated QSAR models and on their subsequent selection. This would be a trivial problem if the compared QSAR models were based on the same number of parameters (descriptors), since in this case the quality of the correlation can unequivocally be evaluated by the value of the correlation coefficient *r*. However, such a simple evaluation is not applicable in the present case, since the above-reported systematic procedure always generates not only linear, $k = 1$, but also all possible *k*-parameter multilinear correlation equations for any selected value of *k*. It is apparent, that when comparing QSAR models which differ in the number of parameters, the comparison of correlation coefficients is useless. This is due to the fact that the inclusion of any new additional parameter into a QSAR model always increases the value of the correlation coefficient, but obviously such an increase does not necessary guarantee the increase of the statistical importance of the correlation. The solution of this important problem of QSAR analysis was recently addressed by one of us,[66] and the basic idea of this approach will be briefly summarized below.

Suppose a general multilinear correlation equation

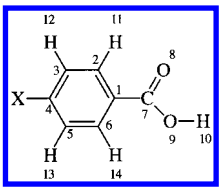
$$y_i = \sum_j a_j x_{ij} + b \mid j \in 1,2...k \wedge i \in 1,2...n \tag{9}$$

is based on the set of variables $\{y_1,y_2,...y_n\}$ $\{x_{i1},x_{i2},...x_{ik}|i=1,2,...n\}$ and assume that the actually observed correlation coefficient is *r*. Then, instead of using the actual values of variables y_i and x_{ij}, one can analyze the same correlation using another set of randomly generated variables (λ_i,ν_{ij}). It is clear, that the correlation coefficient *R* of this randomly generated correlation will be, with high probability, quite low. But, repeating the same random experiment many times, there is a certain (nonzero) probability *P* that the correlation coefficient of such an accidental randomly generated correlation will be equal to or greater than *r*. Such probability depends on the number of points *n*, on the value of the correlation coefficient *r*, and on the number of parameters *k*. It is clear that the lower the value of the probability *P* is, the more difficult it will be to obtain the correlation with $R > r$ accidentally or, in other words, the higher is the statistical significance of the original correlation. This implies that the values of the probability *P* (or its negative logarithm p*P*) can be used as a simple universal measure of the statistical significance of the correlations, irrespective of how many parameters they involve. This probability is closely related to the so-called confidence level of the correlation. The relation of these quantities is given by the formula

$$CL(\text{in }\%) = (1 - P)100 \tag{10}$$

The main goal of the study[66] in which the above criterion was introduced is that it was possible to propose a simple geometrical model allowing the analytical calculation of the corresponding probability for any values of *n*, *k*, and *r*. This

Scheme 1. Numbering of Atoms for Benzoic Acids



probability is given by the formula

$$P = \frac{\int_0^{\arccos(r)} \cos^{k-1}\theta \sin^{n-k-2}\theta d\theta}{\int_0^{\pi/2} \cos^{k-1}\theta \sin^{n-k-2}\theta d\theta} \qquad (11)$$

The numerical calculation of this probability is not difficult and can be done using any existing mathematical programs. Here, an original Fortran code has been used and can be obtained upon request. The quality of the empirical correlations can also be characterized by the value of the cross-validated correlation coefficient r_{cv}. In terms of this approach, acceptable correlations are characterized by $r_{cv} > 0.75$, and this simple, but widely used, criterion has been also used in this study.

Computations. Several kinds of calculation steps were followed in this study. The first of them consists of the quantum chemical generation of wave functions and electron densities for the studied series of molecules. These calculations were performed by the Hartree−Fock method using a 3-21G* basis set for fully optimized molecular geometries. These calculations were carried out using Gaussian 98 software package.[67] In the next step, fragment densities for a series of systematically generated fragments were calculated using eq 8. Finally, the above generated data were employed for the calculation of the corresponding fragment QS-SM, which served afterward as descriptors for the construction of theoretical QSAR models. These models were systematically generated using a nested summation algorithm,[64,65] and the statistical parameters of these correlations were analyzed using the above-reported criteria.

RESULTS AND DISCUSSION

(A) Para-Substituted Benzoic Acids. To convincingly demonstrate the ability of the present approach to detect and localize a molecular fragment responsible, in a given case, for the observed activity, the results of the application of the above formalism to the dissociation of substituted para-substituted benzoic acids are reported first. It is apparent that the active molecular fragment in this example is the COOH group, and, as it will be shown, the systematic search of the best theoretical descriptor also confirms this intuitive expectation. The application of the above-reported approach will be described in detail next.

1. Molecular Set. Twelve para-substituted benzoic acids, with substituents, are as follows: NO_2, CN, CF_3, CCl_3, Br, Cl, F, H, CH_3, CH_2CH_3, OCH_3, and $N(CH_3)_2$.

2. Fragments for the Calculation of QS-SM. The classification of fragments used for the construction of theoretical QSAR descriptors is based on the numbering of individual atoms as shown in Scheme 1.

Based on this numbering, the set of the molecular fragments was arbitrarily selected according to the following strategy: (a) all monatomic fragments corresponding to individual C and O atoms—nine fragments in total: C1−C7, O8, and O9; (b) all biatomic fragments between pairs of directly bonded atoms—14 fragments in total; (c) all triatomic fragments involving directly bonded triads of atoms—20 fragments in total; and (d) the fragment COOH. The total number of such generated fragments was $m = 44$.

3. Selection of the Best Theoretical QSAR Model. As the dissociation of substituted benzoic acids is empirically described by the Hammett equation, involving only one parameter, the σ constant, a restriction of only one-parameter QSAR models, $k = 1$ was adopted. The theoretical QSAR models are based on the computation of the correlation between the σ constant and the QS-SM of each fragment. As the total number of generated theoretical descriptors, m, is 44, the total number of all generated and analyzed linear equations is also 44. The results of these calculations for QSAR models derived from overlap and Coulomb-like fragment QS-SM are summarized in Tables 1 and 2, respectively.

The first, and most important, conclusion resulting from Tables 1 and 2 is that the best QSAR models are indeed generated from QS-SM associated with COOH fragment or any of its subfragments. This fact is preferably observed for Coulomb QS-SM, presented in Table 2. In this respect, the obtained results confirm the correctness of the original expectation. However, the situation is slightly more complex. This is due to the fact that in addition to "expected" correlations with the QS-SM related to the COOH fragment, there is also a considerable number of QSAR models whose theoretical descriptors seem to be in no relation to this group, but whose precision is comparable or only slightly lower. This result seems to be unexpected since in traditional QSAR models there is usually just one empirical descriptor appropriate for the evaluation of the given property, and it can be hardly imagined that the same property could be described by so many different descriptors. To explain the above findings, it has to be realized that a strict specificity characteristic of the empirical parameters does not exist for theoretical descriptors, especially if they are based, as in this case, on quantities derived from electron density. The reason for this greater flexibility of quantum chemical QSAR descriptors, compared to classical ones, may be related to the recently formulated "holographic electron density theorem".[68] This theorem states that any finite fragment of the electron density, considered to be the ultimate molecular descriptor, contains the same amount of structural information as the total electron density of the whole molecule. This implies that all possible fragments of electron density are equivalent in their information content, so that any of them could, in principle, be used for the generation of the appropriate theoretical descriptor. However, the actual situation is slightly less favorable, and some differences between individual fragments and their associated theoretical descriptors can be observed. This is due to the fact that although the holographic electron density theorem guarantees the same information content within any molecular fragment, it says nothing about how this information could be extracted. The present approach, based on the application of QS-SM, represents one of these possibilities. It is just here, in the selection of the particular method of extracting the structural information, where the differences between individual fragments enter into play.

Table 1. Molecular Fragments and Statistical Parameters of QSAR Models for the Dissociation of Para-Substituted Benzoic Acids Based on Overlap QS-SM

fragment[a]	r^b	$r_{cv}{}^c$	CL[d]	pP^e
O9H10	0.980	0.971	100.000	7.609
O9	0.974	0.961	100.000	7.079
C1	0.957	0.945	100.000	5.950
C7O8	0.948	0.905	100.000	5.555
C7O8O9	0.937	0.892	99.999	5.138
C2C1C6	0.911	0.885	99.996	4.411
C1C7O8	0.894	0.857	99.991	4.048
C7O8O9H10	0.893	0.835	99.991	4.029
C1C6	0.875	0.830	99.981	3.712
C2C1	0.873	0.806	99.979	3.687
C7	0.820	0.748	99.890	2.957
C2C1H11	0.808	0.712	99.854	2.836
O8	0.803	0.719	99.835	2.782
C1C6H14	0.776	0.698	99.701	2.525
C7O9	0.751	0.669	99.515	2.314
C2C1C7	0.665	0.514	98.180	1.740
C3C2H12	0.652	0.525	97.841	1.666
C3H12	0.563	0.365	94.316	1.245
C3C2H11	0.523	0.326	91.929	1.093
C1C7	0.475	0.058	88.092	0.924
C5H13	0.467	0.210	87.381	0.899
C1C6C7	0.459	−0.146	86.701	0.876
C1C6C5	0.451	−0.042	85.910	0.851
C3C2	0.422	0.121	82.860	0.766
C6C5H13	0.399	0.112	80.144	0.702
C3C2C4	0.382	0.024	77.968	0.657
C3C4	0.380	−0.017	77.665	0.651
C7O9H10	0.376	−0.379	77.125	0.641
C6C5C4	0.347	−0.053	73.027	0.569
C3	0.335	−0.061	71.342	0.543
C4	0.333	0.011	71.008	0.538
C6	0.318	−0.401	68.573	0.503
C3C4H12	0.311	−0.199	67.405	0.487
C5C4	0.304	−0.213	66.330	0.473
C1C7O9	0.303	−0.291	66.119	0.470
C2	0.279	−0.502	61.937	0.420
C3C2C1	0.267	−0.654	59.831	0.396
C3C5C4	0.253	−0.491	57.241	0.369
C5C4H13	0.207	−0.464	48.099	0.285
C5	0.164	−0.578	38.861	0.214
C6C5H14	0.140	−0.644	33.463	0.177
C6H14	0.128	−0.785	30.783	0.160
C2H11	0.123	−0.819	29.623	0.153
C6C5	0.016	−0.906	3.826	0.017

[a] For numbering see Scheme 1. [b] Standard correlation coefficient. [c] Cross-validation regression coefficient. Negative r_{cv} values close to −1 indicate a reverse relationship between observed and predicted properties. [d] Confidence level defined in eq 10. [e] pP = −log P.

Table 2. Molecular Fragments and Statistical Parameters of QSAR Models for the Dissociation of Para-Substituted Benzoic Acids Based on Coulomb QS-SM

fragment[a]	r^b	$r_{cv}{}^c$	CL[d]	pP^e
O8	0.992	0.987	100.000	9.636
O9H10	0.992	0.988	100.000	9.576
C7O8O9H10	0.983	0.977	100.000	8.006
C7O8	0.983	0.976	100.000	7.928
C7O8O9	0.982	0.975	100.000	7.847
C2C1C7	0.980	0.973	100.000	7.665
C1C7O8	0.976	0.962	100.000	7.225
C1C7	0.971	0.955	100.000	6.790
C1C6C7	0.968	0.950	100.000	6.597
C1C7O9	0.954	0.933	100.000	5.818
C2C1H11	0.946	0.930	100.000	5.477
C1C6H14	0.946	0.910	100.000	5.474
C1	0.938	0.914	99.999	5.172
C6C5H13	0.920	0.875	99.998	4.653
O9	0.919	0.888	99.998	4.617
C7O9H10	0.917	0.878	99.997	4.570
C3C2H12	0.909	0.852	99.996	4.384
C5H13	0.899	0.840	99.993	4.152
C3H12	0.895	0.815	99.992	4.073
C1C6	0.875	0.806	99.981	3.715
C2C1	0.868	0.821	99.975	3.599
C1C6C5	0.865	0.779	99.972	3.550
C3C2C1	0.864	0.746	99.971	3.532
C7	0.853	0.784	99.957	3.370
C6C5H14	0.845	0.737	99.946	3.264
C3C2H11	0.834	0.702	99.925	3.123
C6C5	0.764	0.592	99.617	2.417
C5	0.752	0.593	99.525	2.324
C3C2	0.749	0.537	99.491	2.294
C6	0.742	0.633	99.432	2.245
C2	0.732	0.637	99.324	2.170
C3	0.722	0.443	99.203	2.099
C2C1C6	0.674	0.562	98.370	1.788
C4	0.488	0.177	89.271	0.969
C3C2C4	0.442	0.034	84.931	0.822
C3C4	0.434	0.015	84.108	0.799
C6C5C4	0.429	−0.007	83.558	0.784
C5C4	0.412	−0.057	81.620	0.736
C3C4H12	0.359	−0.189	74.835	0.599
C5C4H13	0.335	−0.273	71.217	0.541
C3C5C4	0.334	−0.311	71.075	0.539
C7O9	0.291	−0.261	64.179	0.446
C6H14	0.158	−0.567	37.544	0.204
C2H11	0.110	−0.755	26.560	0.134

[a] For numbering see Scheme 1. [b] Standard correlation coefficient. [c] Cross-validation regression coefficient. Negative r_{cv} values close to −1 indicate a reverse relationship between observed and predicted properties. [d] Confidence level defined in eq 10. [e] pP = −log P.

Nevertheless, the situation where a given molecular property can be correlated with so many theoretical descriptors raises necessarily the question of the interpretation of these individual correlations, especially with respect to the possibility to extract from them the information about the nature of the active molecular fragment. In the following part the basic idea of such an interpretation will be sketched.

For this purpose, first it is possible to consider the set of all generated QSAR models, and one can evaluate how many times each of the atoms contribute to the examined molecular fragments. Then, in the next step one can do the same but only for the set of fragments, which generated the QSAR models of satisfactory quality. The quality or the statistical significance of the correlations was evaluated using the confidence level (*CL*) and the cross-validated correlation coefficient r_{cv} criterion as commented previously. The correlations with *CL* > 99.9%, which, in this case, roughly coincides with the criterion $r_{cv} > 0.75$, were considered as having acceptable statistical significance. The results of this analysis are summarized in Table 3.

Looking into Table 3 it is possible to see that some atoms contribute to fragments generating statistically significant correlations more often that others. A typical example in this respect is the carbon atom C1, which contributes to 13 fragments, from the whole considered set of 44. For Coulomb QS-SM, 12 times these fragments are associated with the descriptors yielding the statistically significant QSAR models. Similarly high frequency of participating in fragments, generating the statistically significant QSAR models, can be also observed for the carbon atom C7, oxygen atoms O8 and O9, and hydrogen atom H10. On the other hand, there are other atoms, like C3, C4, and C5, for which the corresponding frequency is much lower.

Table 3. Appearance of Individual Atoms in Selected QSAR Models of Dissociation of Substituted Benzoic Acids for Overlap and Coulomb QS-SM

atoms[a]	total (N_t)[b]	$r_{cv} > 0.75$[c]	N_g/N_t	CL > 99.9%[d]	N_g/N_t
		Overlap QS-SM			
C1	13	5	0.385	5	0.385
C2	11	2	0.182	2	0.182
H11	3	0		0	
C3	10	0		0	
H12	3	0		0	
C4	8	0		0	
C5	10	0		0	
H13	3	0		0	
C6	11	2	0.182	2	0.182
H14	3	0		0	
C7	11	4	0.364	4	0.364
O8	5	4	0.800	4	0.800
O9	7	4	0.571	4	0.571
H10	3	2	0.667	2	0.667
		Coulomb QS-SM			
C1	13	11	0.846	12	0.923
C2	11	4	0.364	6	0.545
H11	3	1	0.333	2	0.667
C3	10	2	0.200	4	0.400
H12	3	2	0.667	2	0.667
C4	8	0		0	
C5	10	3	0.300	4	0.400
H13	3	2	0.667	2	0.667
C6	11	5	0.455	6	0.545
H14	3	1	0.333	2	0.667
C7	11	10	0.909	10	0.909
O8	5	5	1.000	5	1.000
O9	7	6	0.857	6	0.857
H10	3	3	1.000	3	1.000

[a] For numbering see Scheme 1. [b] Appearance of each atom in all generated molecular fragments (N_t). [c] Appearance of a given atom in statistically significant QSAR models satisfying the criterion $r_{cv} > 0.75$ (N_g). [d] Appearance of a given atom in statistically significant QSAR models satisfying the criterion $CL > 99.9\%$ (N_g).

This result is very interesting since the atoms identified by the above frequency analysis exactly coincide with the atoms of COOH group, or its close neighbors, expected to be the active molecular fragment on the basis of classical chemical considerations. Based on this result, we report in the following part the application of the presented methodology to the study of the biological activity of Cramer's steroids.

(B) Cramer Steroid Data Set. This molecular set involves a series of 31 steroids whose biological activity is due to their affinity to corticosteroid binding globuline (CBG).[23,25,36,46−63] The structures of these molecules are depicted in Figure 1, and the corresponding biological activities are summarized in Table 4.

In keeping with the above introduced general methodology, the biological activity of this set of steroids was correlated with a series of systematically generated QSAR models based on fragment QS-SM as the corresponding to theoretical descriptors. The fragments considered for the generation of these descriptors were selected using the same protocol as in the previous case of benzoic acids. This involves, in the first step, the generation of the wave function and the density matrix at HF/3-21G* level for the whole series of molecules. As the geometry optimization of such large molecules is time-consuming, the molecular geometries were optimized at semiempirical AM1 level[69] using AMPAC program.[70] The above optimized geometries were used in the next step for the quantum chemical generation of HF/3-21G* electron densities using Gaussian 98 software package.[67] Based on these data, the fragments were selected according to the following systematic procedure:

(a) All H atoms were excluded so that only the fragments involving heavy atoms (C, O) of the common molecular skeleton were considered. This skeleton is composed of 18 atoms: 17 are carbons forming the rings A, B, C and D, and the remaining one is the oxygen bonded to the carbon atom C1 of the ring A. The classification of the fragments is based on the numbering shown in Scheme 2.

(b) The following fragments were considered: (b1) all monoatomic fragments corresponding to individual C and O atoms—C1−C17 and O: 18 fragments in total; (b2) all biatomic fragments between the pairs of directly bonded atoms—21 fragments in total; (b3) all triatomic fragments involving directly bonded triads of atoms—31 fragments in total; and (b4) the whole basic skeleton. In this way, the total number of generated active fragments is $m = 71$.

The theoretical QSAR fragment models were generated for both overlap-like and Coulomb-like QS-SM as descriptors. These models were constructed in the form of multilinear regression equations with the number of parameters k ranging from 2 to 6. The results of this systematic search are summarized in Table 5. As it is evident from the results of Table 5, the number of QSAR models to be considered rapidly increases with the number of parameters k. Moreover, the number of statistically significant QSAR models satisfy-

Table 4. Cramer Steroids CBG Binding Affinity

Figure 1 no.	CBG (pK_a)	Figure 1 no.	CBG (pK_a)
1	−6.279	**17**	−5.225
2	−5.000	**18**	−5.000
3	−5.000	**19**	−7.380
4	−5.763	**20**	−7.740
5	−5.613	**21**	−6.724
6	−7.881	**22**	−7.512
7	−7.881	**23**	−7.553
8	−6.892	**24**	−6.779
9	−5.000	**25**	−7.200
10	−7.653	**26**	−6.144
11	−7.881	**27**	−6.247
12	−5.919	**28**	−7.120
13	−5.000	**29**	−6.817
14	−5.000	**30**	−7.688
15	−5.000	**31**	−5.797
16	−5.225		

Scheme 2. Numbering of Atoms for Cramer Steroids

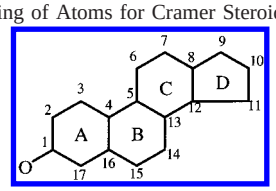


Table 5. Number of Statistically Significant QSAR Equations Satisfying the r_{cv} and Confidence Level Criterions for the Cramer Steroids Set

		overlap		Coulomb	
k	$\binom{m}{k}$	$r_{cv} > 0.75$	$CL > 99.9\%$	$r_{cv} > 0.75$	$CL > 99.9\%$
2	2485	502	1184	264	1354
3	57155	17820	37080	10587	40307
4	971635	386670	749690	243679	788624
5	13019909	5976443	11161540	3751968	11487055
6	143218999	71355131	130973686	42842979	132716856

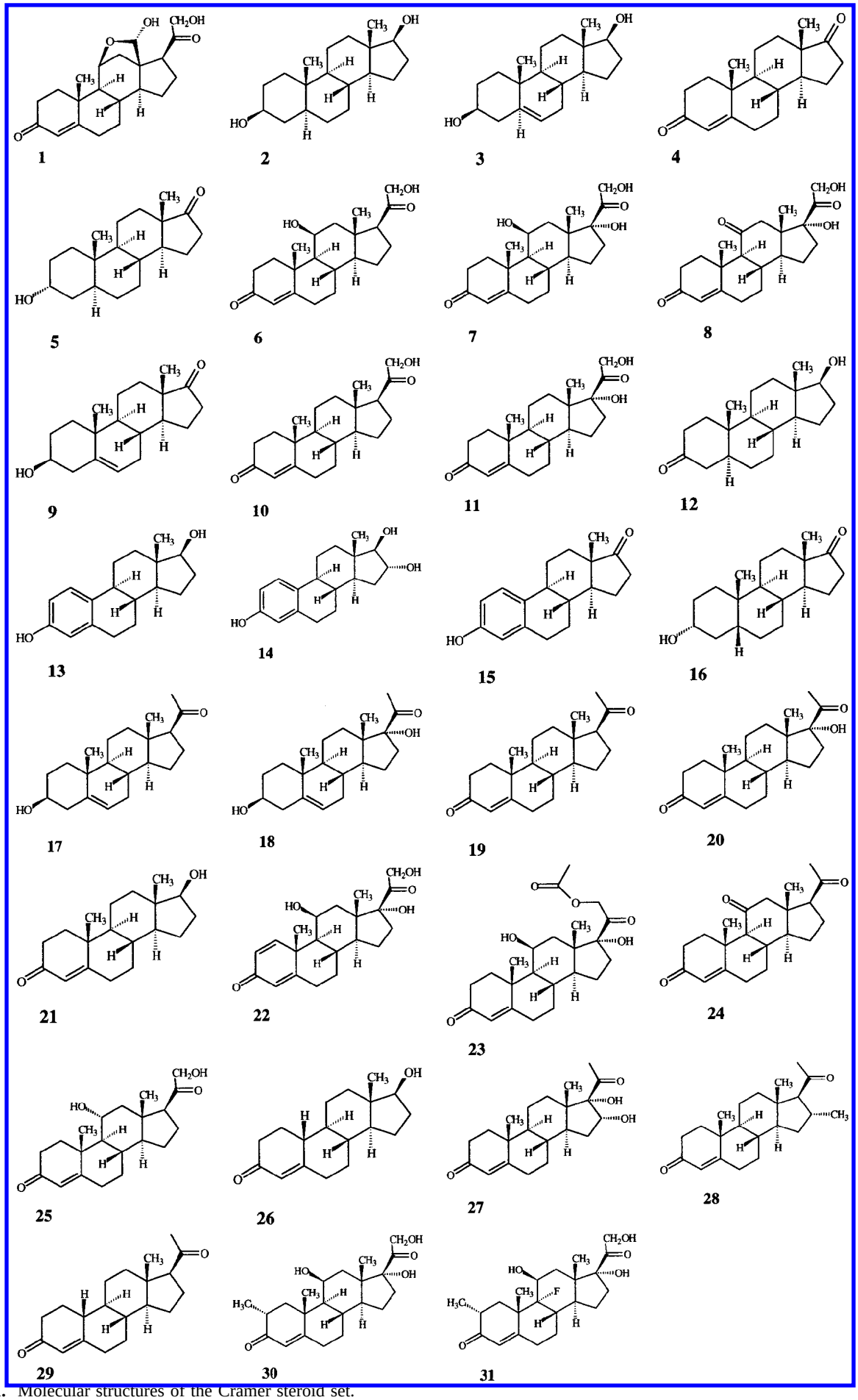


Figure 1. Molecular structures of the Cramer steroid set.

Table 6. Statistical Significance of Cramer Steroids QSAR Models for Different Number of Parameters *k*

	overlap			Coulomb		
k^a	r_{cv}	r	pP	r_{cv}	r	pP
2	0.873	0.889	9.475	0.864	0.890	9.554
3	0.924	0.941	12.139	0.897	0.919	10.285
4	**0.941**	**0.958**	**12.935**	0.916	0.941	11.161
5	0.946	0.960	12.339	0.929	0.955	11.613
6	0.945	0.962	11.684	0.909	0.960	11.313
6^b	0.947	0.961	11.444	0.938	0.953	10.562

[a] The values in individual rows correspond to the QSAR model with highest r value for a given number of parameters k. For $k = 2-5$, they also are the QSAR models with highest r_{cv}. [b] QSAR model with highest r_{cv} using $k = 6$ descriptors.

ing the r_{cv} ($r_{cv} > 0.75$) or *CL* (*CL* > 99.9%) criterion promptly increases with *k*. The fact that the number of statistically significant models is bigger for *CL* than for r_{cv} criterion means that the selected confidence level 99.9% is, in this example, a less severe criterion that $r_{cv} > 0.75$. To reduce the number of statistically significant QSAR models so as to be comparable with the number resulting from r_{cv} criterion, it would be necessary to increase the *CL* to roughly 99.9999%.

In contrast to the simple form of linear one-parameter correlation equations, describing the dissociation of substituted benzoic acids, the set of statistically significant QSAR models, generated for the studied series of steroids, is not homogeneous and involves equations with the number of parameters *k* ranging from 2 to 6. This raises the question of the evaluation of the statistical significance of the observed correlations and, consequently, of the selection of the optimally significant ones. To overcome this problem, two independent methodologies were used. First of them is represented by the recently proposed analytical model,[66] based on the comparison of calculated probabilities of the random generation of the QSAR model with the same number of points *N*, parameters *k*, and the correlation coefficient *r* as the actually observed one. The results of such comparison are summarized in Table 6 from which it is evident, seeing the values of the negative logarithm of probability, p*P*, that the statistically most important is the four-parameter correlation equation based on overlap-like QS-SM. This theoretical model corresponds to the expression

$$y = -6.592 - 0.762 \times \mathrm{Z(C17)} + 0.511 \times \mathrm{Z(C12C13)} + 0.982 \times \mathrm{Z(C13C14C15)} + 0.438 \times \mathrm{Z(C8C12C11)} \quad (12)$$

A graphical representation of the actual CBG affinities versus the predicted ones for the four-parameter QSAR model (12) is shown in Figure 2, based on a LOO CV analysis. The main characteristic of this representation is that compound **31** becomes not an outlier, a normal feature stated in many QSAR studies.[23,25,36,46-60,62,63]

Randomization Test Analysis. To verify the conclusions of the analytical model, the evaluation of the statistical significance of the correlations was independently performed by the numerical randomization test. This test consists of randomly reorienting the set of observed CBG activities and in the subsequent determination of the optimal QSAR model based on fragment QS-SM which gives the maximal value

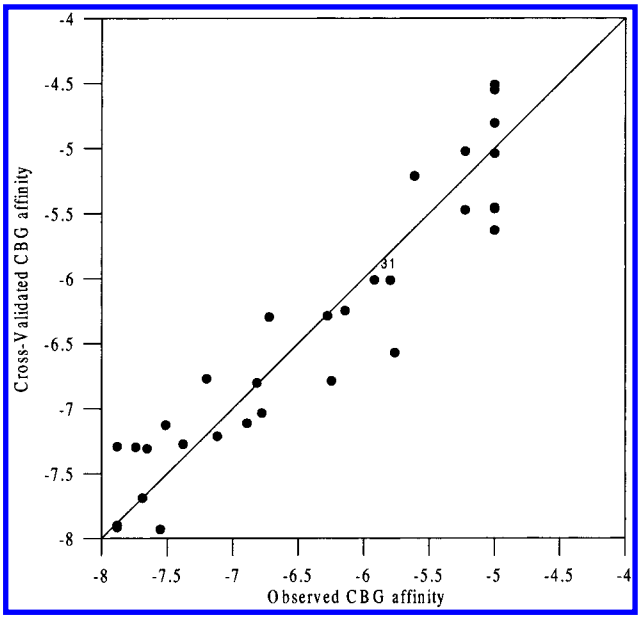


Figure 2. Cross-validated versus experimental CBG affinities for the Cramer steroids QSAR model (12).

of cross-validated correlation coefficient r_{cv} in a LOO analysis. Repeating this process many times, in our case we used 100 random runs, one obtains a set of r_{cv} values which are plotted against the ordinary correlation coefficient of the optimal QSAR model. The results of this randomization test are summarized in Figure 3 from which it is possible to see that the first possibility of random generation of statistically significant QSAR model ($r_{cv} > 0.75$) is observed for $k = 5$.

Comparison with Previous QSM Results on Steroid Data Set. In connection with eq 12 it is perhaps worth mentioning that the same series of steroids was also studied in two previous QSM studies.[23,25] Table 7 shows the main results of the corresponding QSAR models. As it is possible to see from this table, the four-parameter correlation eq 12 obtained in this study is clearly superior to any related previously described QSAR model based on QSM. The best previous results were obtained for a tuned QSAR analysis, based on the combination of three quantum similarity matrices and a six-parameter model.[25] For this study, the $q^{(2)}$ value[71,72] is 0.842, whereas the four-parameter eq 12 yields a value of $r_{cv}^2 = 0.886$.

Identification of the Bioactive Molecular Fragment. Although the basic philosophy of the localization of the bioactive molecular fragment is exactly the same as in the above analyzed case of substituted benzoic acids, the situation with the steroid data set is obviously much more complex. This is due to the fact that the total number of statistically significant theoretical QSAR models in this case is huge (see Table 5), and the statistical parameters of individual models are often quite close. In this situation, it is very difficult to base the selection of bioactive molecular fragment on only a few of the very best correlation equations, like eq 12, and in order to get reliable predictions, the whole set of statistically significant QSAR models has to be considered. For this purpose a simple universal procedure is proposed. It is based on the construction of histograms, depicting the frequency where each atom of the basic skeleton contributes to statistically significant QSAR models.

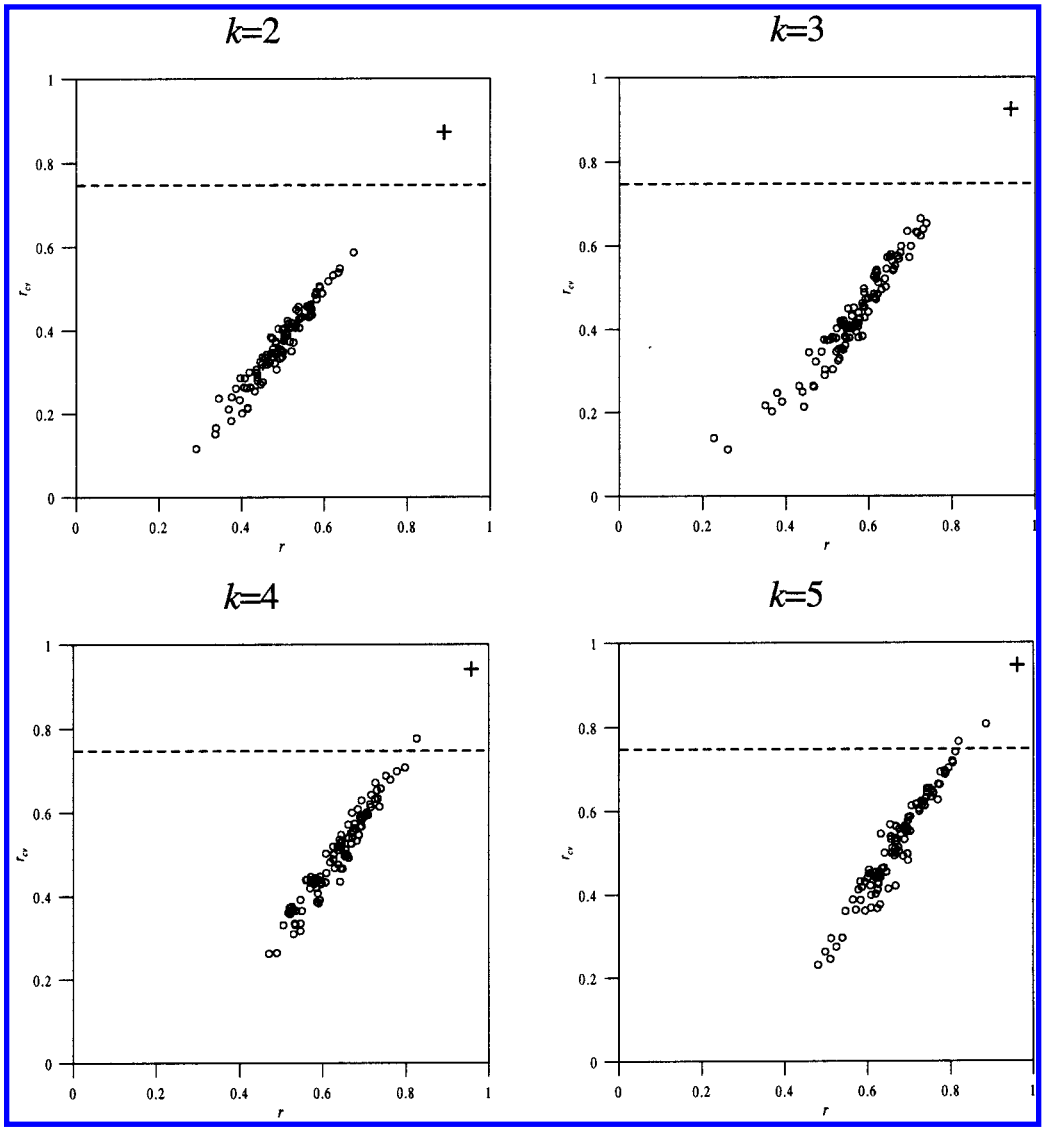


Figure 3. Numerical randomization test analysis for the Cramer steroids set. Multilinear regressions using overlap QS-SM and two up to five descriptors.

Table 7. Comparison of Statistical Significance of the Cramer Steroids QSAR Model (12) with Other Related QSM Models from Previous Studies

	TQSI[a]	MQSM[b]	MQSM[c]	tuned MQSM[d]	fragment QS-SM
$r_{cv}^2 \mid q^{(2)}$ [e]	0.775	0.705	0.759	0.842	0.886
r^2	0.837	0.781	0.833	0.903	0.917
pP	9.17	8.29	8.20	10.25	12.93
no. of PCs (k)	4	3	5	6	4

[a] QSAR study using topological quantum similarity indices. From ref 23. [b] QSAR study using simple quantum similarity matrices, PCA technique for variable reduction and no PCs selected. From ref 23. [c] Simple similarity matrices using classical scaling for variable reduction and selection of PCs. From ref 25. [d] Tuned QSAR model using a mixture of three similarity matrices. From ref 25. [e] In previous works[23,25] $q^{(2)}$ has been used to determine the predictive power of the models.

This frequency is for each individual atom defined as the ratio N_g/N_t, where N_g is the number of "favorable" cases in which a given atom contributes to statistically significant QSAR models and N_t is the total number of correlations involving a given atom.

The corresponding histograms based on the set of overlap-like and Coulomb-like fragment QS-SM are depicted in Figure 4. This figure has been only constructed for QSAR models obtained using four-parameter multilinear equations. The selection of statistically significant QSAR models for the calculation of the corresponding frequencies was based on two criteria: (a) $r_{cv} > 0.75$ and (b) $CL > 99.9999\%$. Then, the selected number of QSAR models which satisfy the above criterions are very similar. For example, 386670 and 307797, respectively, using overlap QS-SM.

As can be deduced from Figure 4, there is no significant difference between the histograms obtained from both criteria. A close parallelism is also observed for the histograms based on overlap-like and Coulomb-like similarity measures. Based on these histograms it is possible to identify the (bio)active part of the molecule with the fragment involving the atoms: C1, C2, C16, C17, O. This suggests that the biological activity of the Cramer set of steroids is very probably due to the presence of a carbonyl group bonded to the ring A of the basic skeleton. It is interesting that the set of most active steroids, the molecules **6**, **7**, **10**, **11**, **19**, **20**, **22**, **23**, **25**, **28**, and **30**, all possess this common structural feature. The main difference between overlap and Coulomb histograms is that for the last one, important contributions

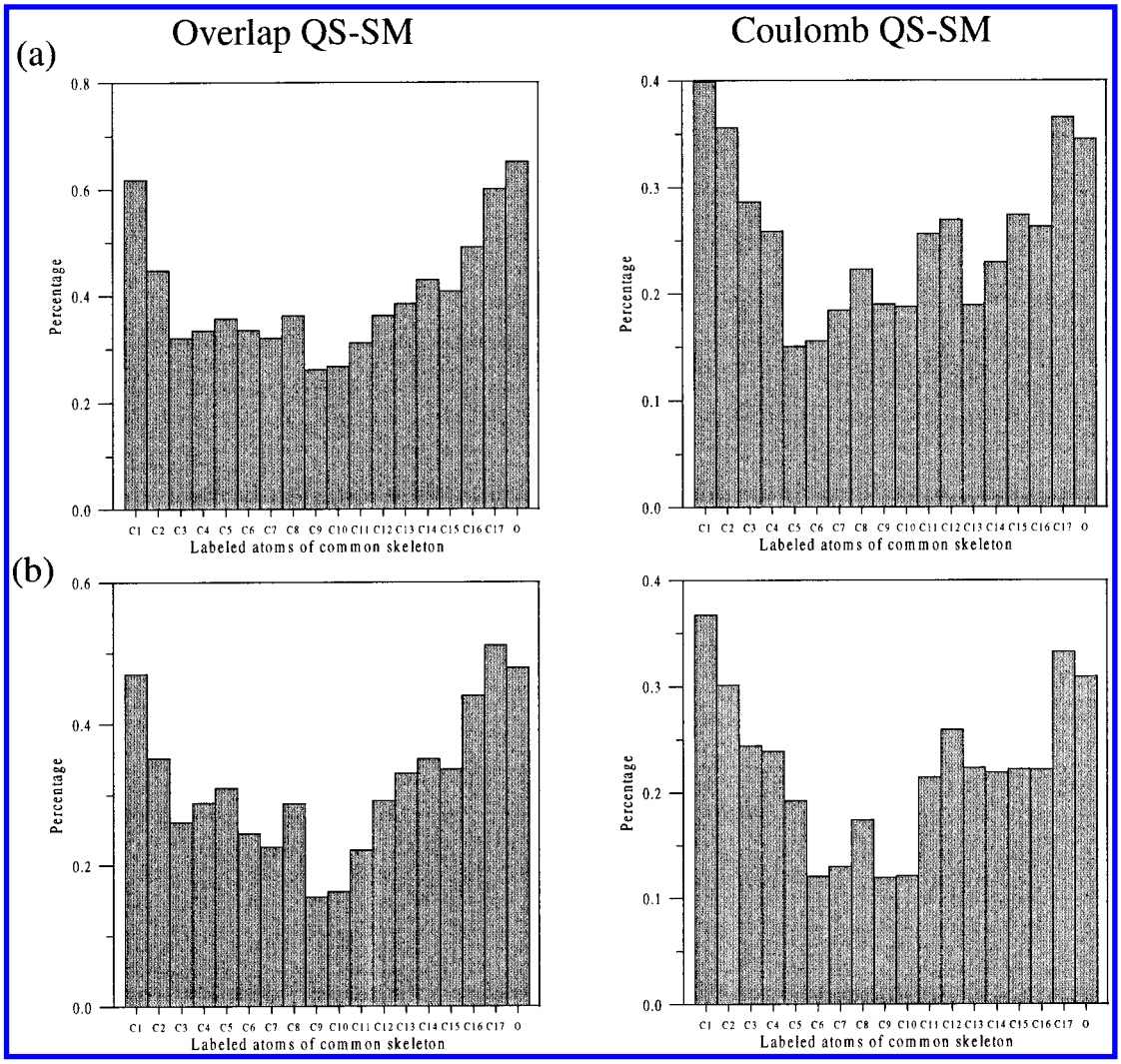


Figure 4. Histograms for overlap and Coulomb QS-SM for Cramer steroids. QSAR models using four-parameter regression equations. Selected statistically significant models using criterions: (a) $r_{cv} > 0.6$ and (b) $CL > 99.9999\%$.

can also be expected from the atoms C11 and especially C12 belonging to the five-membered ring D. Very similar conclusions result also from the use of SOMFA methodology,[49] which identifies two important areas in the steroids skeleton: a large area of negative potential around atom C1 and a large area of positive potential around the five-membered ring. To emphasize the proposed spatial regions determining the activity for the Cramer steroids series, two QSAR models have been computed, combining overlap QS-SM of fragment CCO in ring A and Coulomb QS-SM of atom C12:

$$y = -6.384 - 0.668 \times Z^{ove}(C2C1O) + 0.416 \times Z^{cou}(C12)$$

$$n = 31,\ r = 0.887,\ r_{cv} = 0.858,\ pP = 9.375 \quad (13)$$

$$y = -6.384 - 0.668 \times Z^{ove}(C17C1O) + 0.416 \times Z^{cou}(C12)$$

$$n = 31,\ r = 0.884,\ r_{cv} = 0.856,\ pP = 9.372 \quad (14)$$

Sound correlations have been obtained, comparable with the optimally significant model presented in Table 6 for $k = 2$ parameters. These elucidated fragments could be considered to be representative of the regions where the differences in the electron density preferentially influence the binding affinity of the steroids.

Predictive Ability of the QSAR Models Associated to the Fragments Favoring Activity. The information extracted from the above-mentioned histograms can be used for designing new inhibitors with unknown activity. But in this case, it could be more convenient to compare the predictive capacity of proposed models (13) and (14) with the conclusions of other related studies reported for the same series of steroids in the literature.[25,36,46−63] However, most of these studies do not use the whole set of 31 steroids, but, instead, the set of the first 21 molecules (**1**−**21**) is considered as the training set and the quality of the models is then assessed by comparison of their predictions for the test set of last molecules (**22**−**31**). The standard deviation of errors of prediction (SDEP) is employed as a coefficient to estimate the quality of the model, which is a root-mean-square error of the predictions: $[\sum(y_{pred} - y_{obs})^2/n]^{1/2}$. To make possible direct comparison with Cramer's set earlier studies, QSAR models (13) and (14) have been recalculated for the training

Table 8. Predicted Values of Cramer Steroids Test Set (**22**−**31**) for Different Approaches

steroid	actual activity	CoMFA (FFD)[a]	compass[a]	MS-WHIM[a]	SOMFA[a]	TQSAR[b]	fragment QS-SM
22	−7.512	−7.883	−7.062	−7.300	−7.279	−7.237	−7.036
23	−7.553	−7.430	−7.729	−8.332	−7.034	−7.879	−7.221
24	−6.779	−6.642	−6.462	−6.821	−6.925	−6.648	−7.023
25	−7.200	−7.705	−7.466	−7.445	−7.232	−7.809	−7.307
26	−6.144	−6.495	−5.994	−6.121	−5.744	−6.832	−6.345
27	−6.247	−6.962	−6.383	−6.901	−6.800	−7.318	−7.322
28	−7.120	−6.848	−6.625	−6.532	−6.603	−7.363	−7.536
29	−6.817	−6.816	−7.403	−6.838	−6.692	−7.540	−7.296
30	−7.688	−7.767	−7.741	−7.860	−7.345	−7.628	−7.264
31	−5.797	−7.793	−7.779	−7.491	−7.283	−7.537	−6.675
	SDEP	0.716	0.705	0.662	0.584	0.762	0.544
	SDEP[c]	0.356	0.339	0.411	0.367	0.555	0.493

[a] See ref 49, Table 4. [b] Reference 25. [c] Excludes steroid **31**.

set of 21 steroids. The results of the corresponding analysis are

$$y = -6.454 - 0.726 \times Z^{ove}(C1C2O) + 0.401 \times Z^{cou}(C12)$$

$$n = 21,\ r = 0.909,\ r_{cv} = 0.865,\ pP = 6.844,\ SDEP = 0.544 \quad (15)$$

$$y = -6.449 - 0.719 \times Z^{ove}(C1C17O) + 0.407 \times Z^{cou}(C12)$$

$$n = 21,\ r = 0.908,\ r_{cv} = 0.865,\ pP = 6.817,\ SDEP = 0.550 \quad (16)$$

As can be seen, the coefficients r and r_{cv} increase with respect those of eqs 13 and 14, but the negative logarithm of the probability decreases because the number of molecules is has been reduced. A comparison of the predictive powers of fragment QS-SM and other QSAR approaches is given in Table 8.

It would be interesting to note that some involuntary mistake was performed in a previous study[25] on the same steroid Cramer set. Erroneous SDEP values were given for tuned MQSM, which now have been corrected in Table 8.

CONCLUSIONS

In this study we report a new systematical procedure for the detection and localization of molecular fragments, responsible for the observed activity in a series of structurally related molecules. The approach, based on the use of QS-SM as new set of theoretical molecular descriptors, was applied to description of the CBG activity in the series of 31 Cramer's steroids. The results of our approach are equivalent, even better, in comparison to those obtained using other alternative approaches. In addition, a new methodology to measure the number of statistically significant multilinear regression parameters is proposed.

APPENDIX

Direct Computation of r^2_{cv} Coefficient in Multiple Linear LOO Procedures. Given a $n \times k$ descriptors matrix, $\mathbf{X} = \{x_{ij}\}$, and the vector containing the dependent parameters, $\mathbf{y} = (y_1, y_2, ..., y_n)^T$, the vector collecting the coefficients of the attached multilinear regression is

$$\mathbf{c} = [\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{y} \quad (17)$$

Defining the prediction matrix

$$\mathbf{H} = \{h_{ij}\} = \mathbf{X}[\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T \quad (18)$$

the dependent values fitted by the model are obtained by means of the product

$$\mathbf{y}' = (y'_1, y'_2, ..., y'_n)^T = \mathbf{Xc} = \mathbf{Hy}$$

In this context it is defined as the coefficient of multiple determination:

$$r^2 = 1 - \frac{\sum_{p=1}^{n}(y_p - y'_p)^2}{\sum_{p=1}^{n}(y_p - \bar{y})^2} \quad (19)$$

Here, $\bar{y}$ is the mean value of the observed variables. This term coincides with the correlation coefficient between the **y** variables and the ones fitted by the model (**y′**).

Considering a standard process of cross-validation and collecting the cross-validated property values in the vector $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)^T$. The hat over the variables indicates that they must be obtained for each molecule, say p, from a linear model fit constructed without considering the pth observation, that is, in a LOO procedure. It is customary to represent, in a bidimensional plot, the dependent cross-validated values ($\hat{\mathbf{y}}$) against the experimental ones (**y**). By analogy with expression 19, an *estimation* of the correlation coefficient for the cross-validation procedure is

$$q^{(2)} = 1 - \frac{\sum_{p=1}^{n}(y_p - \hat{y}_p)^2}{\sum_{p=1}^{n}(y_p - \bar{y})^2} = 1 - \frac{PRESS}{S_{yy}}$$

This constitutes the usual definition for the $q^{(2)}$ parameter. The two summations are identified with the prediction error of the sum of squares (*PRESS*) statistic[73] and the sum of quadratic errors from the mean value, S_{yy}.

The $q^{(2)}$ parameter can be *negative*, as it is discussed in ref 74. That is the reason in this paper is preferred to use the $q^{(2)}$ notation instead of the standard Q^2 or q^2 ones.

Nevertheless, it is straightforward to compute the correlation coefficient attached to the linear LOO procedure. In standard textbooks,[73] a demonstration is carried out allowing the computation of *PRESS* variable once the matrix of predictions has been obtained. The demonstration in ref 73 is exact, but it is focused into obtaining the $\mathbf{y}_p - \hat{\mathbf{y}}_p$ differences. Here, a very similar algebraic procedure will be followed, but this time only the $\hat{\mathbf{y}}_p$ values will be computed.

Defining the column vector $\mathbf{x}_p = (x_{p1},x_{p2},...,x_{pm})^T$ as the one collecting the original independent descriptors for the molecule number p coming from the pth row of the $\mathbf{X}$ matrix. Then, if the data of molecule p are eliminated, that is, the value y_p and the vector $\mathbf{x}_p$ are set to zero, a new properties vector $\mathbf{y}_{(p)}$ and descriptors matrix $\mathbf{X}_{(p)}$ are being defined. Following a similar notation as in (17), the coefficients of the linear model are

$$\mathbf{c}_{(p)} = [\mathbf{X}_{(p)}^T\mathbf{X}_{(p)}]^{-1}\,\mathbf{X}_{(p)}^T\mathbf{y}_{(p)}$$

This vector allows the computation of the cross-validated property value for the molecule number p:

$$\hat{y}_p = \mathbf{x}_p^T\mathbf{c}_{(p)} = \mathbf{x}_p^T[\mathbf{X}_{(p)}^T\mathbf{X}_{(p)}]^{-1}\,\mathbf{X}_{(p)}^T\mathbf{y}_{(p)}$$

On the other hand, from (18), the prediction matrix elements are defined as

$$h_{ij} = \mathbf{x}_i^T[\mathbf{X}^T\mathbf{X}]^{-1}\,\mathbf{x}_j \qquad (20)$$

and it is straightforward to demonstrate that the following relationship

$$[\mathbf{X}_{(p)}^T\mathbf{X}_{(p)}]^{-1} = [\mathbf{X}^T\mathbf{X}]^{-1} + \frac{[\mathbf{X}^T\mathbf{X}]^{-1}\,\mathbf{x}_p\mathbf{x}_p^T[\mathbf{X}^T\mathbf{X}]^{-1}}{1-h_{pp}}$$

holds.[73] In this way, one is able to write

$$\hat{y}_p = \mathbf{x}_p^T\left\{[\mathbf{X}^T\mathbf{X}]^{-1} + \frac{[\mathbf{X}^T\mathbf{X}]^{-1}\,\mathbf{x}_p\mathbf{x}_p^T[\mathbf{X}^T\mathbf{X}]^{-1}}{1-h_{pp}}\right\}\mathbf{X}_{(p)}^T\mathbf{y}_{(p)} = \frac{\mathbf{x}_p^T[\mathbf{X}^T\mathbf{X}]^{-1}\,\mathbf{X}_{(p)}^T\mathbf{y}_{(p)}}{1-h_{pp}}$$

Since $\mathbf{X}^T\mathbf{y} = \mathbf{X}_{(p)}^T\mathbf{y}_{(p)} + \mathbf{x}_p\mathbf{y}_p$, then

$$\hat{y}_p = \frac{\mathbf{x}_p^T[\mathbf{X}^T\mathbf{X}]^{-1}[\mathbf{X}^T\mathbf{y} - \mathbf{x}_p\mathbf{y}_p]}{1-h_{pp}} = \frac{\mathbf{x}_p^T[\mathbf{X}^T\mathbf{X}]^{-1}\,\mathbf{X}^T\mathbf{y} - \mathbf{x}_p^T[\mathbf{X}^T\mathbf{X}]^{-1}\,\mathbf{x}_p\mathbf{y}_p}{1-h_{pp}}$$

and from (17) and (20) it is easily obtained

$$\hat{y}_p = \frac{1}{1-h_{pp}}\sum_{i\neq p}^{n} h_{pi}y_i$$

The correlation coefficient between the elements contained in vectors $\mathbf{y}$ and $\hat{\mathbf{y}} = (\hat{y}_1,\hat{y}_2,...,\hat{y}_n)^T$ gives the value of r_{cv}.

The methodology described here can be also applied in leave-two- or leave-many-out procedures.[74]

ACKNOWLEDGMENT

The present work was supported in part by the *Fundació Maria Francisca de Roviralta* as well as the European Commission contract #ENV4-CT97-0508, a UdG grant #3/00, and the CICYT project #SAF2000-223. This research has been carried out using the CESCA and CEPBA resources, coordinated by C[4]. One of us (R. Ponec) acknowledges a CEPBA grant and also the support from the Czech Ministry of Education grant No. D09.20. The authors also thank the referees for their constructive criticism, which improved several aspects of this work.

REFERENCES AND NOTES

(1) 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.
(2) Bowen-Jenkins, P. E.; Richards, W. G. Ab intio computations of molecular similarity. *J. Phys. Chem.* **1985**, *89*, 2195−2197.
(3) Carbó, R.; Domingo, Ll. LCAO-MO Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *23*, 517−545.
(4) Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem. Biol. Symp.* **1987**, *14*, 105−110.
(5) Ponec, R. Topological aspects of chemical reactivity. On the similarity of molecular structures. *Collect. Czech. Chem. Commun.* **1987**, *52*, 555−561.
(6) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G., Eds.; John Wiley & Sons: New York, 1990.
(7) Cooper, D. L.; Allan, N. L. A novel approach to molecular similarity. *J. Comput.-Aided Mol. Design* **1989**, *3*, 253−259.
(8) Cioslowski, J.; Fleischmann, E. D. Assessing molecular similarity from results of ab initio electronic structure calculations. *J. Am. Chem. Soc.* **1991**, *113*, 64−67.
(9) Allan, N. L.; Cooper, D. L. A momentum space approach to molecular similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 587−590.
(10) *Shape in chemistry: and introduction to molecular shape and topology*; Mezey, P. G., Eds.; VCH: New York, 1993.
(11) Carbó, R.; Calabuig, B.; Vera, L.; Besalú, E. Molecular quantum similarity: theoretical framework, ordering principles, and visualization techniques. *Adv. Quantum Chem.* **1994**, *25*, 253−313.
(12) Solà, M.; Mestres, J.; Carbó, R.; Duran, M. Use of ab initio quantum molecular similarities as an interpretative tool for the study of chemical reactions. *J. Am. Chem. Soc.* **1994**, *116*, 5909−5915.
(13) *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Carbó, R., Ed.; Kluwer Academic: Amsterdam, 1995.
(14) Molecular Similarity I. In *Topics in Current Chemistry*; Sean, K. D., Ed.; Springer-Verlag: Berlin, 1995; Vol. 173.
(15) Molecular Similarity II. In: *Topics in Current Chemistry*; Sean, K. D., Ed.; Springer-Verlag: Berlin, 1995; Vol. 174.
(16) *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1. 1998; Vol. 2.
(17) 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.
(18) 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.
(19) Carbó-Dorca, R. Tagged sets, convex sets and quantum similarity measures. *J. Math. Chem.* **1998**, *23*, 353−364.
(20) Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *J. Mol. Struct. (THEOCHEM)* **1998**, *451*, 11−23.
(21) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. Quantum mechanical origin of QSAR: theory and applications. *J. Mol. Struct. (THEOCHEM)* **2000**, *504*, 181−228.
(22) Carbó-Dorca, R. Stochastic transformation of quantum similarity matrices an their fuse in quantum QSAR (QQSAR) models. *Int. J. Quantum Chem.* **2000**, *79*, 163−177.
(23) 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.

(24) Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular quantum similarity measures tuned 3D QSAR: an antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624−631.
(25) Robert, D.; Amat, L.; Carbó-Dorca, R. Three-dimensional quantitative structure−activity relationships from tuned molecular quantum similarity measures: prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333−344.
(26) Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput. Aided Mol. Design* **1999**, *13*, 597−610.
(27) Robert, D.; Carbó-Dorca, R. Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401−422.
(28) Robert, D.; Amat, L.; Carbó-Dorca, R. Quantum similarity QSAR: Study of inhibitors binding to thrombin, trypsin, and factor Xa, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem.* **2000**, *80*, 265−282.
(29) Robert, D.; Gironés, X.; Carbó-Dorca, R. Quantification of the influence of single-point mutations on Haloalkane Dehalogenase activity: a molecular quantum study. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 839−846.
(30) Amat, L.; Carbó-Dorca, R.; Ponec, R. Molecular quantum similarity measures as an alternative to log *P* values in QSAR studies. *J. Comput. Chem.* **1998**, *19*, 1575−1583.
(31) Ponec, R.; Amat, L.; Carbó-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): a quantum similarity approach. *J. Comput. Aided Mol. Design* **1999**, *13*, 259−270.
(32) Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum similarity approach to LFER: substituent and solvent effects on the acidities of carboxylic acids. *J. Phys. Org. Chem.* **1999**, *12*, 447−454.
(33) Amat, L.; Carbó-Dorca, R.; Ponec, R. Simple linear QSAR models based on quantum similarity measures. *J. Med. Chem.* **1999**, *42*, 5169−5180.
(34) Carbó-Dorca, R.; Robert, D.; Amat, L.; Gironés, X.; Besalú, E. Molecular quantum similarity in QSAR and drug design. In *Lecture Notes in Chemistry*; Springer: Berlin, 2000; Vol. 73.
(35) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Similarity screening of molecular data sets. *J. Comput.-Aided Mol. Design* **1992**, *6*, 513−520.
(36) Good, A. C.; So, S.-S.; Richards, W. G. Structure−activity relationships from molecular similarity matrices. *J. Med. Chem.* **1993**, *36*, 433−438.
(37) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem.* **1993**, *36*, 2929−2937.
(38) Lee, C.; Smithline, S. An approach to molecular similarity using density functional theory. *J. Phys. Chem.* **1994**, *98*, 1135−1138.
(39) Measures, P. T.; Mort, K. A.; Allan, N. L.; Cooper, D. L. Applications of momentum-space similarity. *J. Comput.-Aided Mol. Design* **1995**, *9*, 331−340.
(40) Benigni, R.; Cotta-Ramusino, M.; Giorgi, F.; Gallo, G. Molecular similarity matrices and quantitative structure−activity relationships: a case study with methodological implications. *J. Med. Chem.* **1995**, *38*, 629−635.
(41) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular field-based similarity approach to pharmacophoric pattern recognition. *J. Mol. Graphics Modelling* **1997**, *15*, 114−121.
(42) Measures, P. T.; Mort, K. A.; Cooper, D. L.; Allan, N. L. A quantum molecular similarity approach to anti-HIV activity. *J. Mol. Struct. (THEOCHEM)* **1998**, *423*, 113−123.
(43) Popelier, P. L. A. Quantum molecular similarity. 1. BCP space. *J. Phys. Chem. A* **1999**, *103*, 2883−2890.
(44) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular-field-based similarity study of nonnucleoside HIV-1 reverse transcriptase inhibitors. *J. Comput. Aided-Mol. Des.* **1999**, *13*, 79−93.
(45) Goodford, P. J. A Computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(46) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(47) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.
(48) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129−2140.
(49) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field analysis: A Tool for Structure−Activity Studies. *J. Med. Chem.* **1999**, *42*, 573−583.
(50) Oprea, T. I.; Ciubotariu, D.; Sulea, T. L.; Simon, Z. Comparison of the minimal Steric Difference (MTD) and Comparative Molecular Field Analysis (CoMFA) Methods for Analysis of Binding of Steroids to Carrier Proteins. *Quant. Struct.-Act. Relat.* **1993**, *12*, 21−26.
(51) Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.
(52) Hahn, M.; Rogers, D. Receptor Surface Models. 2. Application to Quantitative Structure−Activity Relationships Studies. *J. Med. Chem.* **1995**, *38*, 2091−2102.
(53) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic *Ah* Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.
(54) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state Fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513−520.
(55) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteroid-binding globulin activity of steroids. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 521−534.
(56) Norinder, U. 3D-QSAR Investigation of the Tripos Benchmark Steroids and some Protein-Tyrosine Kinase Inhibitors of Styrene Type using the TDQ Approach. *J. Chemom.* **1996**, *10*, 533−545.
(57) Schnitker, J.; Gopalaswamy, R.; Crippen, G. M. Objective models for steroid binding sites of human globulins. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 93−110.
(58) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 79−92.
(59) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409−422.
(60) Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. Alignment of Molecules by the Monte Carlo Optimization of Molecular Similarity Indices. *J. Comput. Chem.* **1997**, *18*, 1344−1353.
(61) So, S.-S.; Karplus, M. Three-Dimensional Quantitative Structure−Activity Relationships from Molecular Similarity Matrixes an Genetic Neural Networks. 1. Method and Validations. *J. Med. Chem.* **1997**, *40*, 4347−4359.
(62) Tominaga, Y.; Fujiwara, I. Prediction-Weighted Partial Least-Squares Regression Method (PWPLS) 2: application to CoMFA. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1152−1157.
(63) Chen, H.; Zhou, J.; Xie, G. PARM: A Genetic Evolved Algorithm To Predict Bioactivity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 243−250.
(64) Carbó, R.; Besalú, E. Definition, mathematical examples an quantum chemical applications of nested summation symbols and logical Kronecker deltas. *Computers Chem.* **1994**, *18*, 117−126.
(65) Carbó, R.; Besalú, E. Definition and quantum chemical applications of nested summations symbols and logical functions: Pedagogical artificial intelligence devices for formulae writing, sequential programming and automatic parallel implementation. *J. Math. Chem.* **1995**, *18*, 37−72.
(66) Pecka, J.; Ponec, R. Simple analytical method for evaluation of statistical importance of correlations in QSAR studies. *J. Math. Chem.* **2000**, *23*, 13−22.
(67) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *GAUSSIAN 98*, *Revision A.6*; Gaussian, Inc.: Pittsburgh, PA, 1998.
(68) Mezey, P. G. The Holographic Electron Density Theorem and Quantum Similarity Measures. *Mol. Phys.* **1999**, *96*, 169−178.
(69) 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.
(70) AMPAC 6.01; Semichem, Inc.: 7128 Summit, Shawnee, KS 66216.

D.A.

(71) Wold, S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **1978**, *20*, 397−405.
(72) Wold, S. Validation of QSARs. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191−193.
(73) Montgomery, D. C.; Peck, E. A. *Introduction to linear regression analysis*; Wiley: New York, 1992.
(74) Besalú, E. Fast computation of cross-validated properties in full linear leave-many-out procedures. IT-IQC-00-36; Institute of Computational Chemistry: *J. Math. Chem.* (in press).

CI000160U