



Facet diagrams for quantum similarity data

David Robert, Xavier Gironés & Ramon Carbó-Dorca*

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

Received 24 October 1998; Accepted 2 February 1999

Key words: Atomic Shell Approximation (ASA) density functions, Carbó index, classical scaling, Cramer steroids, facet theory, quantum similarity, quantitative structure-activity relationships (QSAR), toxicity

Summary

The objective of this work is to demonstrate that an appropriate treatment of quantum similarity matrices can reveal hidden data grouping related to relevant structural features and even to biological properties of interest. Classical scaling is used here to extract the information contained in the similarity relationships between the elements of a molecular set. Facet theory is invoked to relate, in a qualitative way, the spatial regions to structural characteristics as well as to properties of interest. Two application examples are discussed: the Cramer steroid set and a benzene, toluene and xylene derivatives set.

Introduction

The study of structure-activity relationships (QSAR) is a promising subject in the theoretical chemistry field. Many different chemical descriptors [1] and statistical tools [2] have been used to establish valuable quantitative relations between the molecular structure and some relevant pharmacological or biological properties. On the other hand, the study of molecular similarity has provided a new point of view when building molecular descriptors, and it has been proved that it can also be employed as an efficient QSARs source [3]. The key concept that makes possible the construction of structure-activity equations is the fact that *similar* molecules ought to possess *similar* properties. Quantum Similarity provides a suitable quantification of the resemblance between two molecules, based on steric aspects.

Sometimes this is not entirely satisfactory. For instance, in the substrate-receptor interactions, the geometry of the molecule is not the only factor to be taken into account, but also certain properties of the molecular surfaces such as hydrophobicity, electrostatic or hydrogen-bonding potentials must be pointed out. In most cases, however, the structure-activity re-

lationships are clearly determined by the type of the substituents, especially for highly homogeneous series. In these cases, Quantum Similarity can provide a useful approach to analyse the data, and the similarity measures can be used as descriptors to discriminate between high and low active molecules. In fact, the use of MQSM to find out quantitative structure-activity equations has been widely discussed and proved [4].

In relation to the evaluated properties, additional difficulties can appear. One problem is that, in most cases, the property cannot be accurately known. For instance, the error in the measurement of some toxicity levels is high enough to avoid the use of quantitative equations, and then it seems better to transform the numerical values into discrete classes and handle them with other structure-activity methods. Another problem is that the property can present a categorical form, that is, it can originally take on discrete values. In both cases it is more interesting to perform a *qualitative* study of the data, instead of finding quantitative relationships of suspect reliability. Thus, the usual linear regressions (such as principal component regression or partial least squares regression) or neural network algorithms must be disavowed in order to define a relational technique where the regional grouping is associated to a given level of a property value.

*To whom correspondence should be addressed. E-mail: director@iqc.udg.es.

In the present study, Quantum Similarity theory is applied to examine the structure-property relationships of two different molecular sets in a qualitative way. The first one is a widely studied set, which has become a benchmark data set: the 31 Cramer steroids and their binding affinity to the corticosteroid-binding globulin (CBG) [5]. The second application example is a set of 36 benzene, toluene and xylene derivatives and their toxic action to aquatic species [6]. The previous results on these data sets will be referred to later.

Methods

Molecular quantum similarity measures (MQSM)

Molecular Quantum Similarity Theory [7] was developed in order to obtain a formal comparison between the elements of a molecular set in terms of a quantum mechanical descriptor: their electronic density functions. The Quantum Mechanics postulates assume that all the information of a system is contained in the wavefunction, and due to the one-to-one existing relationship between it and the density function, the latter can also be used to describe completely a microscopic system. In this sense, Molecular Quantum Similarity can be considered a theoretical body constructed within the quantum mechanical formalism, particularly based on Löwdin and McWeeny's density function framework [8, 9].

Given the first-order density functions of two molecules: $\rho_A(\mathbf{r}_1)$ and $\rho_B(\mathbf{r}_2)$, where \mathbf{r}_1 and \mathbf{r}_2 are the molecular coordinates, the general form of a *molecular quantum similarity measure* (MQSM) [7a] can be defined as the integral:

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

where $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ is a positive definite operator. This operator can adopt several forms, and the one used in the current study is the Coulomb-like, yielding the so-called *Coulomb MQSM*:

$$Z_{AB}(\Omega) = \iint \rho_A(\mathbf{r}_1) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2. \quad (2)$$

This type of MQSM gives, except for a numeric factor, the Coulomb repulsion energy between two charge distributions, the electronic density functions. In this sense it can be considered as an electrostatic descriptor. The complete set of MQSM can be expressed in matrix form, yielding the quantum similarity matrix

$\mathbf{Z} = \{Z_{AB}\}$. The MQSM can be scaled or normalised by means of quantum similarity indices. The transformation used here is known in the literature as *Carbó index* [7a], defined as:

$$C_{AB} = Z_{AB}(Z_{AA}Z_{BB})^{-1/2}. \quad (3)$$

The Carbó index gives the cosine of the angle between the functions $\rho_A(\mathbf{r}_1)$ and $\rho_B(\mathbf{r}_2)$, and ranges from zero to one: the closer to one, the more similar will be both molecules. In this sense, Carbó index is called to be a C-Class index: a similarity index, in contrast to the D-Class or dissimilarity indices [10]. These similarity matrices can be used to extract information from the molecular set, and suitable transformations of them can be employed as QSAR parameters.

Visualisation of quantum similarity data: Classical scaling

Classical scaling [11] is a multidimensional scaling (MDS) technique for the analysis of proximity data on a set of objects, based upon the theorems of Eckart and Young [12] and Young and Householder [13]. This method considers the similarities as distances and then finds coordinates to explain them; that is, classical scaling performs a mapping of the proximities into distances in an Euclidean space of lower dimension. This is made as follows: first, the original similarity matrix, in our case the Carbó index matrix \mathbf{C} , is *doubly centred* using the centring matrix \mathbf{J} , defined as $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$. This yields:

$$\mathbf{B} = \mathbf{J}\mathbf{C}\mathbf{J}. \quad (4)$$

This transformation sets the centre of coordinates at the centroid of the data. The matrix \mathbf{B} is considered as a matrix of scalar products $\mathbf{B} = \mathbf{X}\mathbf{X}^T$, and the configuration \mathbf{X} is recovered by the eigendecomposition of \mathbf{B} :

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T. \quad (5)$$

Therefore, the scores \mathbf{X} are

$$\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^{1/2}. \quad (6)$$

The k -dimensional Euclidean space used to represent the data is defined by the first k eigencomponents of \mathbf{B} , which are called the principal coordinates of the system:

$$\mathbf{X}_{(k)} = \mathbf{V}_{(k)}\mathbf{\Lambda}_{(k)}^{1/2}. \quad (7)$$

The eigenvalues of **B** give an idea of the explained variance accounted for each PC. From a variational point of view, classical scaling minimises a loss function called *Strain* [14], defined as:

$$L(\mathbf{X}) = \|\mathbf{XX}^T - \mathbf{B}\|^2. \quad (8)$$

The facet theory

The interpretation of the pictorial representations of the molecular relationships and their relation to relevant features (structure and even properties, if possible) is the main objective of this work. The location of the point-molecules in the different MDS subspaces should be attached to some important characteristics of the compounds.

By the MQSM construction, two molecules will be more similar to each other if they possess common shape (structural) features. In this sense, it is reasonable to expect finding regions in the MDS space related to the presence or absence of determined substituents or chemical groups.

Regional interpretation of the MDS configurations is aided by *Facet theory* [15], which provides a systematic framework for the interpretation of regions in MDS solutions. The main idea of this approach is that the MDS space can be partitioned into regions, with regular or irregular contours, which can be associated to features of the data. Most of these features are usually categorical divisions of a domain of interest. The scheme used to classify the elements of these domains into types is called a *facet*. Several facets can be used at the same time, yielding the *multifaceted configurations*.

The two- or three-dimensional MDS subspaces analysed are called *facet diagrams*, which can be partitioned into regions in such a way that each region corresponds to a facet of the data. Partitions in the MDS space can be performed in several ways. If the partitioning lines of a facet cut the space into subspaces that look like parallel stripes of the plane, the facet is said to be partitioned in an axial way, yielding the *axial facets*. If the space is partitioned in concentric bands, the facet is called *modular*. Finally, a *polar facet* divides the space by rays emanating from a common origin.

Combinations of these types of facets in two- or three-dimensional spaces yield to different multifacets. A multifacet of particular interest in this work is the *multiplex*, made up of two, *duplex*, or more, *triplex*... axial partitions.

Finally, it must be emphasised that facet theory is general enough to encompass clusters, viewed as a particular case of the aforementioned regions. A cluster is a region whose points are closer to each other than to any other point in the diagram. Clusters look like 'islands' densely packed and isolated from the rest.

Computational details

The MQSM have been calculated using MOLSIMIL97 program, developed in our laboratory [16]. In order to avoid expensive ab initio calculations, the molecular density functions have been constructed using the Atomic Shell Approximation (ASA) [17,18], which considers molecular densities as a sum of discrete atomic densities centred at atomic positions. The atomic density functions are built as a linear combination of spherical gaussian functions, with coefficients and exponents determined by minimising the error to the respective atomic ab initio density functions [19]. The following rule has been used: one gaussian function for hydrogen, three for carbon, nitrogen, oxygen, fluorine and four for chlorine atoms.

The molecular geometries were obtained from different sources, and the compounds were aligned for a maximum Coulomb-like similarity. A set of selected values of the electronic density functions were rendered and plotted later with the GiD program [20]. The overall ASA isodensity contours exhibited in the current work can be seen and downloaded from our website [21].

The discretization of the molecular properties was based uniquely upon property value criterion, allowing for heterogeneous size classes. The mathematical expression to find the class boundaries is:

$$\frac{n}{k}(y_{\max} - y_{\min}), \quad n = 0, \dots, k \quad (9)$$

where y_{\max} and y_{\min} are the maximum and minimum property value, respectively, and k is the desired number of classes. In this study, 3 classes have been used ($k = 3$), corresponding to low (L), intermediate (M) and high (H) value. This yields:

$$\begin{aligned} y_{\min} &\leq L < y_{\min} + \frac{1}{3}(y_{\max} - y_{\min}) \\ y_{\min} + \frac{1}{3}(y_{\max} - y_{\min}) &\leq M < y_{\min} + \frac{2}{3}(y_{\max} - y_{\min}) \\ y_{\min} + \frac{2}{3}(y_{\max} - y_{\min}) &\leq H \leq y_{\max} \end{aligned} \quad (10)$$

The explicit boundaries for each example set are given in the corresponding section.

Results and discussion

Two molecular sets were examined in order to test the applicability of MDS to Quantum Similarity data and to recognise the existing regions in the configuration and their relation to some molecular facets. The first set is made up of 31 steroids, a widely studied set by novel QSAR methods [5]. The property analysed was the binding affinity to the corticosteroid-binding globulin (CBG). The second set is made up of 36 benzene, toluene and xylene derivatives, which possess an acute toxicity to aquatic species [6].

First application example: the Cramer steroid set

The Cramer steroid set is a widely studied data set by different QSAR models. The correlation to the CBG binding affinity has been traditionally employed to test novel 3D-QSAR methods. The great amount of existing reports on this subject recommends to not expose here a summary of the obtained results. A brief survey of the results of the different QSAR approaches dealing with this data set can be seen in Reference 5. Figure 1 shows the elementary skeleton of the steroid family, where the rings and the substituent positions have been marked. Figure 2 shows the structures of the 31 steroids studied, together with their identification number. The CBG binding affinity and its division into classes, discussed later, are given in Table 1. The steroid geometries were obtained from Reference 22, and the construction of the density functions and MQSM was made as pointed out in the last subsection.

Analysis of the ASA isodensity plots

It has been commented that the molecular electronic density functions used in this study were not the ab initio ones, but an approximation obtained using the so-called ASA approach. This section tries to prove that the major structural characteristics of the molecules, essential for the structure-activity relationships, are in fact contained in the approximated molecular densities, and its use is therefore justified. To do this, the ASA density functions were analysed by means of the observation of their plots. The graphical representation procedure of the ASA electron densities is

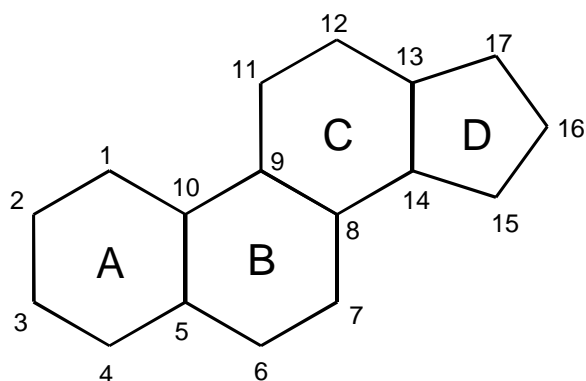


Figure 1. Schematic representation of the steroid structure. A letter is assigned to each ring and the substituent positions are numbered. Almost all the atoms lie nearby the XY plane.

Table 1. Cramer steroid set, identification number, experimental CBG binding affinities (pK_a) and discrete levels^a

Compound	ID	Aff. (pK_a)	Aff. class
Aldosterone	1	6.279	M
Androstenediol	2	5.000	L
Androstenediol	3	5.000	L
Androstenedion	4	5.763	L
Androsterone	5	5.613	L
Corticosterone	6	7.881	H
Cortisol	7	7.881	H
Cortisone	8	6.892	M
Dehydroepiandrosterone	9	5.000	L
Deoxycorticosterone	10	7.653	H
Deoxycortisol	11	7.881	H
Dihydrotestosterone	12	5.919	L
Estradiol	13	5.000	L
Estriol	14	5.000	L
Estrone	15	5.000	L
Etiocolanolone	16	5.225	L
Pregnenolone	17	5.225	L
Hydroxypregnenolone	18	5.000	L
Progesterone	19	7.380	H
Hydroxyprogesterone	20	7.740	H
Testosterone	21	6.724	M
Prednisolone	22	7.512	H
Cortisolacetat	23	7.553	H
4-Pregnene-3,11,20-trione	24	6.779	M
Epicorticosterone	25	7.200	H
19-Nortestosterone	26	6.144	M
16a-17a-Dihydroxyprogesterone	27	6.247	M
17a-Methylprogesterone	28	7.120	H
19-Norprogesterone	29	6.817	M
2a-Methylcortisol	30	7.688	H
2a-Methyl-9a-fluorocortisol	31	5.797	L



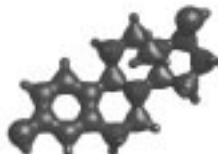

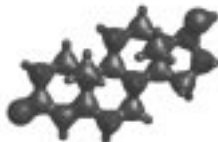

^aH: high; M: intermediate; L: low.

described elsewhere [23], so only the major features of the method are given here. In order to construct isodensity surfaces, a grid surrounding the molecules was defined in such a way that its size and spacing were adequate to obtain a fine representation. The ASA approximated electronic density value was computed at each point of the grid, and the surface reconstruction was done by means of the *marching cubes* algorithm [24]. A value of the density was then fixed, and an isodensity plot was obtained, where only the regions with a density equal to the chosen value were plotted. Throughout this section values of 0.15 and 0.22 au^{-3} were used, two arbitrary values which are sufficient to provide a good visualisation of the molecular features. In addition, the structures have been oriented in the space in such a way that the plots show as best as possible the relevant molecular regions.

The analysis was performed as follows. The Cramer steroids present a common skeleton made up of three hexagonal and one pentagonal rings, and the major part of these molecules combine a few different substitutions at a few number of positions. As it will be discussed later, the most relevant features are focused over two zones: the configuration of ring A and the substituents at position 17. Thus, the substituent at position 17 was first frozen and the effect of the different configurations for ring A was later analysed. The second step consisted of fixing the configuration of ring A and then examining the changes in the density when different substituents at position 17 are used. Finally, the effect of the fluorine atom in the density function of molecule **31** is briefly discussed.

The substituent at position 17 is fixed to be a hydroxyl group, and the differences in the configuration of ring A are studied by means of the analysis of the isodensity plots. Table 2 shows the electron density for compounds **3**, **13** and **21** at values of 0.15 and 0.22 units. These molecules are representatives of the different configurations of ring A. It can be clearly observed how the shape of carbon atoms is spherical when the ring is delocalized (molecule **13**), whereas they possess a tetragonal geometry in the other cases (compounds **3** and **21**). In addition, the orientation of the hydrogen atoms bonded to these carbons is different in both situations: they can be located in the same plane as the ring (molecule **13**) or can adopt the typical tetragonal distribution. In the 0.15 au^{-3} isodensity plot, the double bond of position 3 substituent and in the 4–5 edge in molecule **21** can be differentiated from the single bonds present in the other two compounds by a slightly wider region between the car-

Table 2. ASA isodensity contour plots for molecules **3**, **13**, **21**

Molecule	ASA isodensity XY contour plots	
	0.15 au^{-3}	0.22 au^{-3}
5-Androstenendiol (3)		
Estradiol (13)		
Testosterone (21)		

bon and oxygen atoms, which is even clearer in the 0.22 au^{-3} isodensity plot. At the latter density level, double bonds are the only ones linked.

Now the ring A is fixed at its most common form: a carbonyl group at position 3 and a double conjugated bond between positions 4 and 5, and the most usual substituents at position 17 are analysed in a similar manner. Table 3 shows the electron density for compounds **4**, **21**, **10**, **19** at values of 0.15 and 0.22 units. Differences between molecules with light substituents (**4** and **21**) can be appreciated: the double bond between the carbon and the oxygen atom in molecule **4** is again recognised by a slightly wider density region

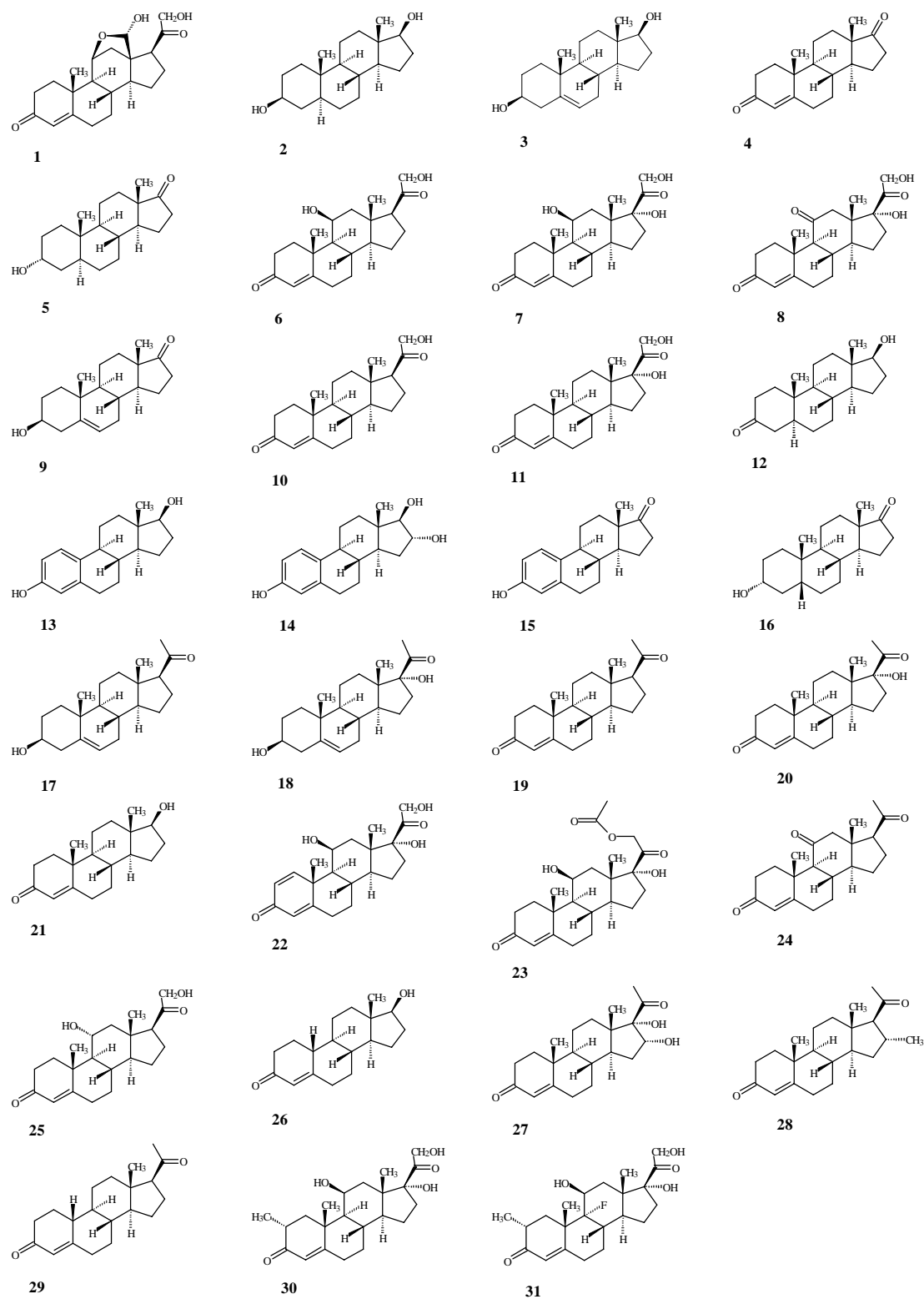


Figure 2. Structures of the 31 steroids used in the present study, with the corresponding identification number.

Table 3. ASA isodensity contour plots for molecules **4**, **21**, **10**, **19**

Molecule	ASA isodensity XY contour plots	
	0.15 au ⁻³	0.22 au ⁻³
4-Androstenedione (4)		
Testosterone (21)		
11-Deoxycorticosterone (10)		
Progesterone (19)		

Table 4. ASA isodensity contour plots for molecule **31**

Molecule	Plane	Perspective of ASA isodensity contour plots	
		0.15 au ⁻³	0.22 au ⁻³
2a-Methyl-9a-fluorocortisol (31)	XY		
	XZ		

entations in order to visualise the fluorine atom in a clearer way.

Summarising, all the relevant topological aspects and the subtle differences in the molecular structure are contained in the approximated density functions, and subsequently, in the MQSM. Such accuracy ensures a good description of the actual differences between the objects when they are compared between them. This is a general capacity of the method, and

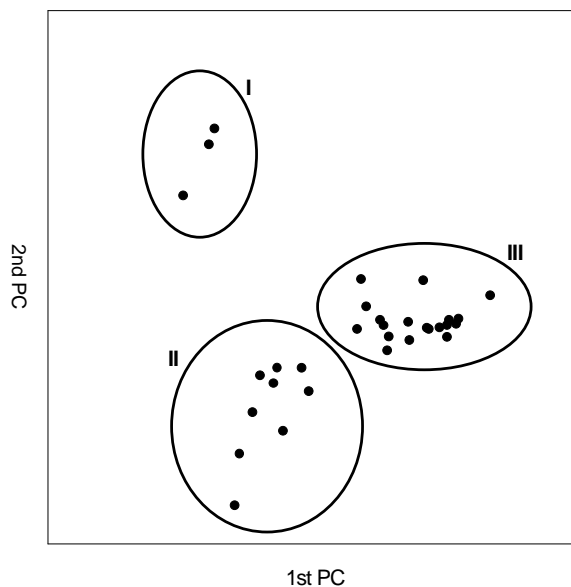


Figure 3. Two-dimensional MDS solution from the Carbó index matrix for the Cramer steroid set. *Natural* clusters have been marked and numbered. Cluster I is made up of elements {**13**, **14**, **15**}, cluster II of {**2**, **3**, **4**, **5**, **9**, **12**, **16**, **21**, **26**}, and cluster III of {**1**, **6**, **7**, **8**, **10**, **11**, **17**, **18**, **19**, **20**, **22**, **23**, **24**, **25**, **27**, **28**, **29**, **30**, **31**}.

this analysis was not repeated for the compounds of the second example case.

Discussion of the MDS solution

The analysis of the quantum similarity data by means of facet theory was performed using the first two principal coordinates (PCs) of the system, containing the 47.7% of the variance of the data. As will be seen a posteriori, the low variance explained by these two PCs is enough to represent satisfactorily the existing relations in the data, and this choice allows one to graphically represent the objects in a visualizable way. In spite of the homogeneity of the set, the distribution of the explained variance of the PCs is rather smooth, and several axes are needed to explain a high percentage of the variance. Fortunately, in most cases the information for a discrimination between the property classes is contained in some degree in the first PC.

Figure 3 shows a two-dimensional plot where the steroid set is represented by points, numbered using their identification number (cf. Figure 2). In this plot, three clusters can be apparently distinguished. To assess that the clustering really exists, the first two principal coordinates were introduced as a data table in a clustering analysis algorithm. The clustering technique employed was the so-called *partitioning around*

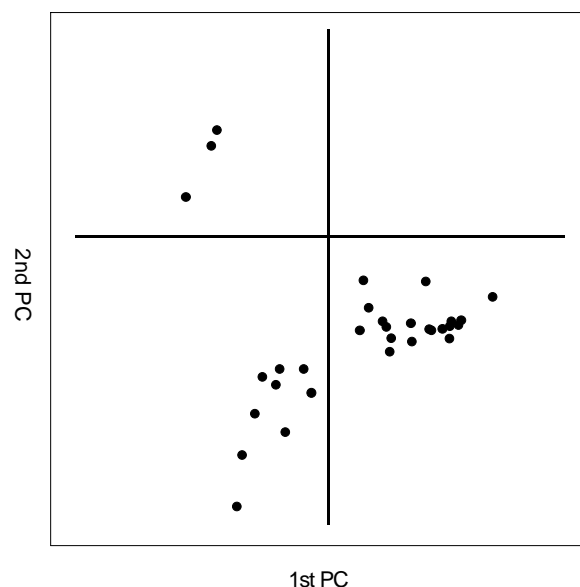


Figure 4. Two-dimensional MDS solution from the Carbó index matrix for the Cramer steroid set. Solid lines are axial facets. See text for formal details.

medoids (PAM) [25], and when a proximity matrix or a data table is given, the algorithm finds the optimal division of the objects, and quantifies the degree of hardness of the partition. In this case, the best classification was found to have three clusters: a first one made up by the steroids {13, 14, 15}, a second one by {2, 3, 4, 5, 9, 12, 16, 21, 26}, and finally a third cluster containing compounds {1, 6, 7, 8, 10, 11, 17, 18, 19, 20, 22, 23, 24, 25, 27, 28, 29, 30, 31}. These 'natural' or 'intuitive' groups have been marked and numbered in the figure. The cluster boundaries have been arbitrarily drawn.

Nevertheless, the facets of the data are the most interesting methodological features. If the previous plot is divided according to different criteria, another diagram is obtained (Figure 4), where two axial facets are distinguished: a vertical axial facet (VAF) and a horizontal axial facet (HAF). This multifaceted diagram contains relevant information on the structure of the steroids. Thus, the VAF differentiates between the steroids possessing light or heavy substituents at position 17. The left side of the VAF includes all the compounds with light substituents at this position, oxygen or hydroxyl groups: {2, 3, 4, 5, 9, 12, 13, 14, 15, 16, 21, 26}, whereas the right side contains the molecules with more complex substituents: {1, 6, 7, 8, 10, 11, 17, 18, 19, 20, 22, 23, 24, 25, 27, 28, 29, 30, 31}. Note that the left side of the VAF

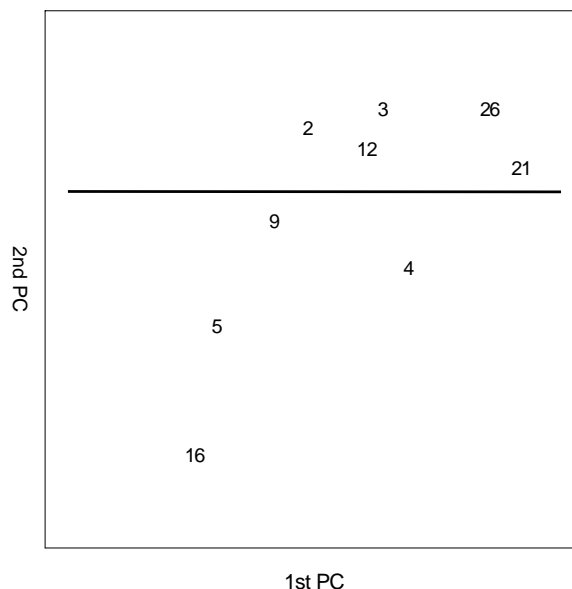


Figure 5. Amplification of the second cluster of the two-dimensional MDS solution for the Cramer steroid set. The solid line is an axial facet. See text for formal details.

coincides with the union of cluster I and II, and the right side with cluster III. On the other hand, the HAF separates those compounds with a delocalized ring at substructure A from those that have not. Thus, the molecules with this delocalized ring, {13, 14, 15}, are located above the HAF, and below those which do not possess this structure. This diagram allows a classification of the molecules in terms of their structural features, as it was announced. This fact is not surprising if we think of the MQSM construction: they are a weighted overlapping of the molecular structures, and when the molecular set is so homogeneous, the slight differences in the MQSM values must reflect the slight differences between the substituents of the common skeleton.

This information is not the only one extractable from this diagram. Going further on in this scheme, each initial natural cluster can be analysed separately to obtain a more accurate description of the system. Thus, Figure 5 shows an amplification of the second cluster, and a new set of differences arise. A new HAF appears which separates steroids with an oxygen atom at position 17 from molecules with OH group at the same position. The quantum similarity measures are able to differentiate between two similar light substituents, and this information can be extracted from the similarity matrix.

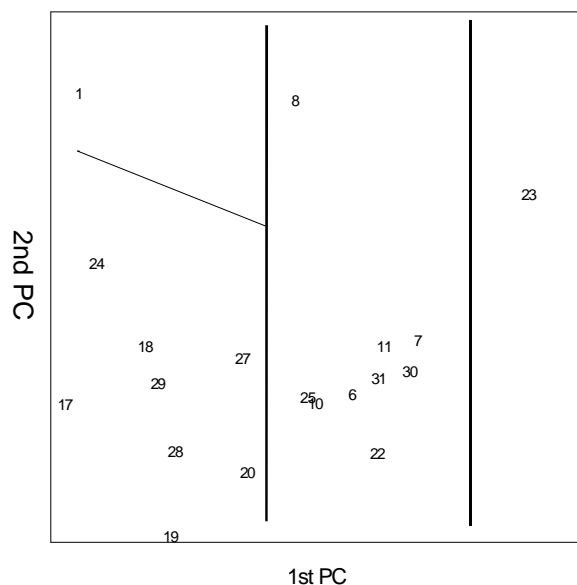


Figure 6. Amplification of the third cluster of the two-dimensional MDS solution for the Cramer steroid set. Solid lines are axial facets, and the thin line is a modification to include point **1** in the second facet. See text for formal details.

A parallel treatment can be carried out for the third cluster, as Figure 6 shows. The multifaceted diagram, a vertical duplex, shows that the differences between the compounds with bulky substituents are also contained in the data. Thus, the rightmost region contains the molecules with the COCH_3 substituent at position 17, whereas the centre region is made up by the steroids with the COCH_2OH chain at the same location. Compound **1** is the only exception, and can be considered a misclassification in the model if we take the facets as purely axial, or otherwise the facets can be distorted as shown by the thin line in the figure to include this point in the correct group. Finally, the leftmost region is formed uniquely by steroid **23**, which possesses the $\text{COCH}_2\text{OCOCH}_3$ substituent, the largest one.

Qualitative structure-activity relationships

Much more interesting than to identify common structural features of the compounds, is to relate these groups to discrete levels of a determined property, given a priori. The most usual property evaluated in the Cramer steroid set is the binding affinity to the corticosteroid-binding globulin (CBG). The quantitative values of this property were reported by Dunn et al. [26]. The discretization of these values has been performed according to Equation 10: low activity (*L*),

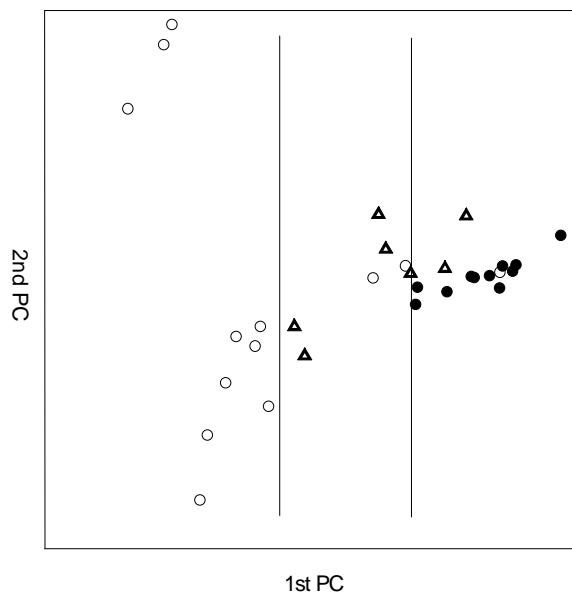


Figure 7. Two-dimensional MDS solution for the Cramer steroid set. Compounds are labelled using the discrete activity levels. Solid lines are axial facets. See text for formal details. ○: High activity level. △: Intermediate activity level. ●: Low activity level.

ranging from 5.000 to 5.960; intermediate activity (*M*), from 5.960 to 6.921; and high activity (*H*), from 6.921 to 7.881. This classification is slightly different from that proposed by Anzali et al. [27]. Actually, compounds **12**, **28** and **31** are classified in a different way.

Taking this into account, the facet diagram for the first two PCs is constructed in relation to the facet 'CBG binding affinity', with low, intermediate and high levels. This is shown in Figure 7, where geometrical symbols have been used to get a better visualisation. The facetization is made in a clear duplex way, and the major part of the compounds are correctly classified in terms of their CBG affinity, in spite of only half of the information is explained by the first two principal axes. The misclassifications present uniquely a 'one-level jump', that is, the compounds misclassified are located in a region with only one activity class of difference (for instance, molecules with low affinity could be set in the intermediate region, but not in the high region), except for molecule **31**, an outlier in most of the QSAR studies on this steroid set, probably due to the existence of a fluorine atom at position 9. Some of these misclassifications can be attached to the choice of the activity level contours and could be reduced by a different discretization of the affinities, such as Anzali's one.

If the configuration of ring A is analysed, it is observed that when the steroid possesses a carbonyl at position 3 and a conjugated double bond between 4–5 positions, their CBG affinity is intermediate or high, except for compounds **4** and **31**. This fact almost coincides with the natural grouping represented by cluster III, which also includes molecules **17** and **18** and does not contain steroids **4**, **21** and **26**. These two last compounds are located at the top right corner of cluster II, and they are the closest molecules to the third cluster. Subsequently, this configuration of ring A seems to be the most appropriate for a good binding to CBG globulin, and the molecular quantum similarity measures can discriminate, in a very good approximation, the elements of the set which possess it from those which do not.

Furthermore, it is reasonable to think that the extension of the study to the following PCs, in the sense of more variance could be explained, should improve the description of the property. In fact, the first PC is responsible for the discrimination into the different activity levels, so a combination of it with other axes of lower variance could allow a better visual class separation. In any case, these results show some interesting trends in the steroid structure-activity relationships that can be used to improve the molecular design models for this family. The homogeneous structure of this family, and the different substituents limited to a few positions, permit the association of the ability of CBG binding to the presence or absence of some substructures at those positions, essentially the configuration of ring A or the substitutions at position 17, and in less degree at positions 2, 3, 9 or 11. Thus, the combination of Figure 4 and Figure 7 shows how the existence of ring A with a carbonyl group at position 3 and a conjugated double bond between positions 4 and 5 lead to an increase of the compounds activity. Further, molecules with light substituents (oxygen or hydroxyl groups) at position 17 present low activity values, whereas the presence of bulky substituents, specially the chain COCH_2OH usually increases the CBG affinity of the steroids. These two effects are complementary: thus, molecules **17** and **18** possess COCH_3 chains but ring A in a non-optimal form, and their affinity is low. On the other hand, compound **4** possesses the optimal form for ring A, but a light substituent at position 17, and it also has a low CBG affinity. Molecules **21** and **26** present the same features as this last steroid, and their activity is intermediate. There are no molecules with the two decisive features that do not possess a high CBG affinity, except for the outlier **31**, where the flu-

orine atom becomes an unfavourable structure for the protein interaction. All these factors indicate the clear structure-activity relationships existing in this system.

Second application example: Benzene, toluene and xylene derivatives

The second example of application of MDS to deal with Quantum Similarity data consists of the study of a set of benzene, toluene and xylene derivatives. These compounds are observed to have an acute toxicity to aquatic species. Toxicity is a good property to be studied with qualitative techniques, since their measurement is usually made by observing the effects on animal species; for instance, by measuring the dose levels necessary to reduce a population to 50% of the initial value. These experiments are hard to reproduce, and there is no sense in assigning a high accuracy on the derived results. So, it seems more convenient to perform QSARs to relate, in a qualitative way, molecular features to discrete property ranges, namely high, medium or low toxic action. A previous study reported the toxicity levels of a wide range of polar and nonpolar narcotic pollutants to three different aquatic organisms: the pond snail *Lymnaea stagnalis*, the water flea *Daphnia magna* and the gummy *Poecilia reticulata* [6]. From all the molecules analysed, a homogeneous subset made up of 36 analogues of benzene, toluene and xylene has been selected. As in the steroid case, Coulomb MQSM were computed for each molecular pair, and the similarity matrix, conveniently transformed into Carbo indices, was dealt with MDS techniques. The most representative two-dimensional subspace, made up of the first and second principal axes, was analysed by means of the Facet theory in order to relate structural features to location in the space, as well as to find qualitative associations to their properties.

It is commonly accepted that the toxic action of narcotic pollutants is produced by the disruption of the functioning of cell membranes, leading to the death by narcosis [28]. Hence the toxic capacity of a compound is related to its propensity to accumulate in them. Phospholipid cell membranes are often modelled by apolar mediums, namely octanol, so it is not surprising that satisfactory correlations are obtained when the octanol/water partition coefficient, $\log P$, is used as a QSAR descriptor [29]. This relation is known as 'baseline toxicity'. Recently, it has been proved that a close connection of quantum self-similarity measures and $\log P$ exists for highly homogeneous series [30].

Hence even better correlations are expected to be obtained when the extra-diagonal terms of the similarity matrix are included.

The property examined is the lethal concentration necessary to reduce the initial population of the fish *Poecilia reticulata* to 50% in 96 h, denoted by LC_{50} . Experimental data was reported by Urrestarazu et al. [31]. Again the toxicity data have been partitioned into three classes, following the distribution given in Equation (10). The boundaries obtained were: low toxicity (*L*), ranging from 2.97 to 4.03; intermediate toxicity (*M*), from 4.03 to 5.09; and high toxicity (*H*), from 5.09 to 6.15. This classification did not lead to groups of homogeneous size, as can be observed in the table.

Table 5 shows the studied compounds, together with an identification number, their numerical toxicity and the corresponding discrete level. The pollutant geometries were calculated at a semiempirical AM1 level using the AMPAC 6.0 software [32]. The simplicity of the structures, namely a hexagonal ring substituted with light atoms or simple chemical groups, made unnecessary higher computational level calculations. Molecular density functions and MQSM were computed as indicated in the Methods section.

Discussion of the MDS solution

As in the steroid case, the study of the quantum similarity relationships between the 36 pollutants was performed by means of the analysis of the two-dimensional MDS representation, using the first two PCs as axes. These two axes contain 43.9% of the explained variance of the system. As in the previous example, the distribution of the variance in relation to the number of PCs is quite smooth, and 8 PCs are necessary to reach 90% of the explained variance. Figure 8 shows the two-dimensional MDS solution, where the identification number was used as a label for each compound.

In this figure there are no clear clusters, and the points are spread all over the plot. This is not a problem for the Facet theory, which is more general than cluster analysis. As it has been stated, clusters are a particular case of data spatial distribution, but not the only one that can be dealt within the facet framework.

This representation can be analysed in terms of the facet scheme, associated to molecular structural features. Thus, a multifaceted diagram arises in relation to the facet 'size/complexity of the substituents'. A vertical triplex can be drawn, which divides the data into 4 regions, as shown in Figure 9. Each one of these

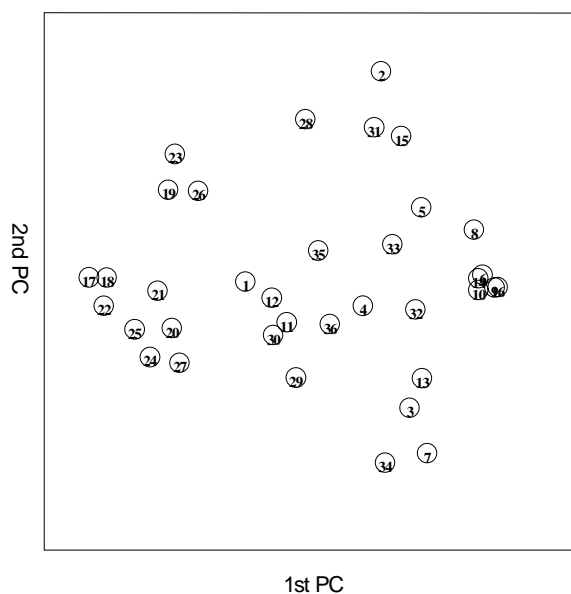


Figure 8. Two-dimensional MDS solution from the Carbó index matrix for the benzene, toluene and xylene derivatives. Identification numbers have been used as labels for the compounds.

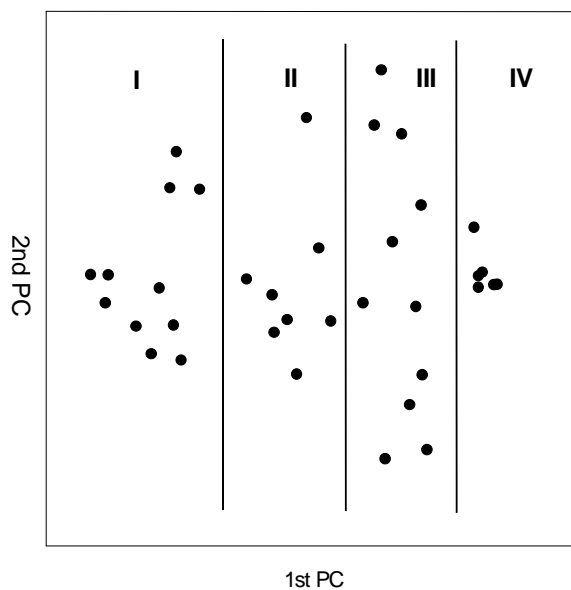


Figure 9. Two-dimensional MDS solution from the Carbó index matrix for the benzene, toluene and xylene derivatives. Solid lines are axial facets. Each region has been marked with a roman number. See text for formal details.

Table 5. Benzene, toluene and xylene derivatives set, identification number, experimental toxicity (LC₅₀) and discrete levels^a

Compound	ID	Tox. (LC ₅₀)	Tox. Class
Chlorobenzene	1	3.77	L
1,2-Dichlorobenzene	2	4.40	M
1,3-Dichlorobenzene	3	4.28	M
1,4-Dichlorobenzene	4	4.56	M
1,2,3-Trichlorobenzene	5	4.89	M
1,2,4-Trichlorobenzene	6	4.83	M
1,3,5-Trichlorobenzene	7	4.74	M
1,2,3,4-Tetrachlorobenzene	8	5.35	H
1,2,3,5-Tetrachlorobenzene	9	5.43	H
1,2,4,5-Tetrachlorobenzene	10	5.85	H
3-Chlorotoluene	11	3.84	L
4-Chlorotoluene	12	4.33	M
2,4-Dichlorotoluene	13	4.54	M
2,4,5-Trichlorotoluene	14	5.06	M
3,4-Dichlorotoluene	15	4.60	M
Pentachlorobenzene	16	6.15	H
Benzene	17	3.09	L
Toluene	18	3.13	L
2-Xylene	19	3.48	L
3-Xylene	20	3.45	L
4-Xylene	21	3.48	L
Nitrobenzene	22	2.97	L
2-Nitrotoluene	23	3.59	L
3-Nitrotoluene	24	3.65	L
4-Nitrotoluene	25	3.67	L
2,3-Dimethylnitrobenzene	26	4.39	M
3,4-Dimethylnitrobenzene	27	4.21	M
2-Chloronitrobenzene	28	3.72	L
3-Chloronitrobenzene	29	4.01	L
4-Chloronitrobenzene	30	4.42	M
2,3-Dichloronitrobenzene	31	4.66	M
2,4-Dichloronitrobenzene	32	4.46	M
2,5-Dichloronitrobenzene	33	4.59	M
3,5-Dichloronitrobenzene	34	4.58	M
2-Chloro-6-nitrotoluene	35	4.52	M
4-Chloro-2-nitrotoluene	36	4.44	M

^a H: high; M: intermediate; L: low.

regions, marked with a roman number, is connected to a structural characteristic of the molecules. Thus, the first region is made up of the compounds with light substituents: nitrogen and methyl, whereas regions II, III and IV contain molecules with bulky substituents, namely chlorine. Each one of these last three regions has its own particularities. Thus, region II contains all the pollutants with one chlorine atom, region III is

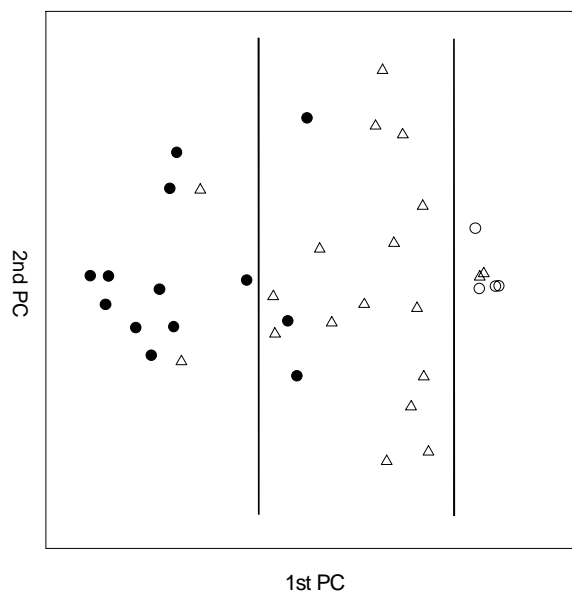


Figure 10. Two-dimensional MDS solution from the Carbó index for the benzene, toluene and xylene derivatives. Solid lines are axial facets. Compounds are labelled using the discrete toxicity levels. ○: High toxicity level. △: Intermediate toxicity level. ●: Low toxicity level.

made up of those molecules with two chlorine atoms, and region IV contains those compounds with three or more chlorines. MQSM are strongly dependent on heavy atoms, since they produce high density peaks. The only misclassifications in this scheme are the molecules 5 and 7, which possess three chlorine substitutions but they are located in the third region. This representation confirms that Quantum Similarity tends to locate close each other those molecules with similar characteristics, as expected by the definition of the MQSM.

Qualitative structure-activity relationships

Can the previous point distribution be associated to the discrete toxicity levels? To answer this question, a new facet diagram of the MDS solution was plotted, now marking the molecules with geometrical symbols according to its toxicity level (cf. Table 4). Thus, the analysis of the data was based on the facet 'toxic action', with low, intermediate and high levels. This configuration is shown in Figure 10.

Again a satisfactory molecular classification was attained, in spite of the fact that the first two PCs only accounted for half the explained variance. As in the steroid case, this was not a drawback to find valuable structure-property relationships, due to the fact that

the relevant molecular features for toxicity are mainly contained in the first PC. Two vertical axial facets have been drawn, which divide the configuration into three regions. As can be observed, almost all the compounds were well classified. As in the steroid case, the few misclassifications presented a 'one-level jump'. Moreover, most of the misclassified compounds are close to the boundaries between classes, and they are then difficult to discriminate. For instance, pollutants **6** and **14**, which possess LC_{50} equal to 4.83 and 5.06, respectively, were incorrectly assigned to the high toxicity class. These are two of the highest medium toxicities, close to the intermediate class upper bound: 5.09. Further, three low toxic agents were classified as medium ones, namely **11**, **32** and **33**. These three compounds also lie near the interclass boundary: they possess 3.84, 3.72 and 4.01 LC_{50} values, respectively; three of the highest ones of their class. Finally, molecules **30** and **31**, the only dimethylnitrobenzenes of the set, were located in the low toxicity region, whereas they have an intermediate action: 4.42 and 4.66. These compounds were also poorly predicted by a previous QSAR study, reported by Urrestarazu et al. [6]. In that study, molecules **30** and **31** were predicted to have a low toxicity, 3.89 and 3.86, respectively.

Comparison with the structure of the set leads to some valuable conclusions. The previous results indicate that chlorine substitutions in the hexagonal ring are decisive for aquatic toxicity to *Poecilia reticulata*. The more chlorine atoms the molecule possesses, the more toxic it is. The addition of a nitrogen atom or a methyl group to the chlorine substituted molecules does not significantly increase the toxicity, as can be deduced by comparing different series of molecules. For instance, for single chlorine substitution, compounds **11**, **12** or **28**, **29**, **30** are not substantially more toxic than molecule **1**. For doubly chlorinated agents, the nitrochlorobenzenes **31**, **32**, **33**, **34** or the chlorotoluenes **13**, **14**, **15** do not present a higher acute toxicity than molecules **2**, **3**, **4**. This information can be used to improve the molecular design of nonpollutant agents, or to discriminate between possible toxic or non-toxic compounds.

Conclusions

In the present work it has been demonstrated how Quantum Similarity Theory is a valuable tool to evidence the information contained in the molecular density functions. Furthermore, it has been proved

that MDS is a suitable geometrical technique to extract the information from the MQSM, capable to carry out interesting pictorial representations of the molecular relationships using only two or three dimensions, especially when congeneric compounds are studied. Moreover, the identification of regions in the MDS subspaces and their easy connection to molecular structural features has been also presented. Each region can be assigned to a facet, in the sense of Facet theory, and further, divisions into discrete activity levels can provide visual qualitative structure-activity relationships, an interesting procedure when the property is not known with enough accuracy, or when one deals with categorical properties.

In the first example case analysed, the Cramer steroids, clear axial facets divide the data into different groups, in relation to the size of the substituents at position 17 and the ring A configuration. These regions have been more deeply studied, and subregions arose giving more precise information on the substituents. The region connection to the CBG affinity have permitted the extraction of valuable conclusions on the steroid structure-activity relationships, and possible ways to improve the molecular design models for this family are outlined. In the second example, a set of benzene, toluene and xylene analogues, the MDS solution separates the data in an axial way, according to the size/weight of the substituents. Further, a clear association between the type of substituents of the hexagonal ring and the corresponding aquatic toxicity was found.

Acknowledgements

The authors gratefully acknowledge Prof. J. Gasteiger and coworkers for providing us with the steroid geometries used in the present study. We also want to thank the reviewers for their advice, which doubtless helped us to substantially improve the manuscript. This research was partially supported by a CICYT grant SAF 96-0158, the EU project ENV4-CT97-0508 and the *Fundació Maria Francisca de Roviralta*. Thanks are also due to the *Centre de Supercomputació de Catalunya (CESCA)* and the *Centre Europeu de Parallelisme de Barcelona (CEPBA)* for a generous amount of computation time. Finally, thanks to E. Oñate from the *Centre Internacional per als Mètodes Numèrics a l'Enginyeria (CIMNE)* for allowing us the free use of the rendering program GiD, developed in this laboratory.

