Graph Theoretical Similarity Approach To Compare Molecular Electrostatic Potentials

Ray M. Marín, Nestor F. Aguirre, and Edgar E. Daza*

Grupo de Química Teórica, Universidad Nacional de Colombia, Bogotá D. C., Colombia

Received May 31, 2007

In this work we introduce a graph theoretical method to compare MEPs, which is independent of molecular alignment. It is based on the edit distance of weighted rooted trees, which encode the geometrical and topological information of Negative Molecular Isopotential Surfaces. A meaningful chemical classification of a set of 46 molecules with different functional groups was achieved. Structure−activity relationships for the corticosteroid binding affinity (CBG) of 31 steroids by means of hierarchical clustering resulted in a clear partitioning in high, intermediate, and low activity groups, whereas the results from quantitative structure−activity relationships, obtained from a partial least-squares analysis, showed comparable or better cross-validated correlation coefficients than the ones reported for previous methods based solely in the MEP.

INTRODUCTION

There are many areas in chemistry where molecules need to be compared. The comparison may be performed among a set of molecules in order to classify them and monitor changes, e.g., in their shape, in their properties, in their reactivity, etc. The need of a quantitative molecular comparison has led to several molecular similarity measures, even thought the concept of molecular similarity itself is not easy to define, and hence many efforts have been done to measure how similar molecules are. Molecular similarity may be considered as a fuzzy concept, and to a great extent its fuzziness arises because of the lack of a rigorous definition of molecular structure, which is also another fuzzy concept.[1−7] Therefore, there are almost as many definitions of molecular similarity as representations of molecular structure. Molecular similarity may be regarded as shape similarity from a classical[8−10] or quantum perspective,[3,4,11] graph similarity,[12−14] or as a relation among quantum mechanical descriptors,[15−17] just to mention some examples.

In the context of the Pearson's Hard and Soft Acids and Bases (HSAB) principle, chemical reactions are mainly directed by soft−soft and hard−hard interactions. The former ones may be correctly described making use of the Fukui's Frontier Molecular Orbital Theory;[18−20] however, in counter of the Li-Evans proposition,[20] the Fukui function is a poor descriptor for the hard−hard interactions, which are much better described by atomic charges or even better by the Molecular Electrostatic Potential (MEP).[18,21,22] Because of its importance in a number of molecular recognition processes, involving hard−hard interactions, the MEP has been widely used to represent and compare molecules, enhancing our knowledge of chemical reactivity.[23] In recent years, advances in software and hardware have been allowed to consider large sets of molecules and even proteins and nucleic acids.[24,25] Visual analysis of MEP calculated on particular surfaces has been taken as a quick and very intuitive approach to get insights into chemical interaction.[23,26−29] However, the visual comparison could be tedious, subjective, and not reliable when large sets of molecules are considered.[30]

Several methods to compare MEPs and correlate them with some properties or activities, that avoid the problems related to visual comparisons, have been developed.[4,30−39] When examining the wide variety of them, two classes can be clearly identified: those methods requiring the superposition of the corresponding molecular skeletons, i.e., alignment dependent (AD) approaches and those that are alignment independent (AI). While AD methods compare directly the MEP values on a set of points in the 3D ordinary space, AI approaches usually represent molecules by *n*-tuples whose components correspond to some descriptors that characterize the MEP field. The maximum MEP value on a molecular surface, the minimum value, the standard deviation of MEP values on the whole surface, the area of a particular surface, its volume, etc. are some of the descriptors introduced by Murray et al.[31] and subsequently complemented by Chalk et al.[33] Other interesting approaches are the autocorrelation vectors proposed by Wagener et al.[37] and the slightly different autocorrelation vector proposed by Pastor et al.[38] When representing molecules as *n*-tuples of their properties, we assume the existence of an underlying *molecular space* of unknown dimension.[40] In the particular case of the MEP, the minimum number of descriptors required to achieve an appropriated representation is also an open question.

Prior to use an AD method it is necessary to ensure an optimal superposition of molecular frames, but the alignment procedure is meaningful among molecules sharing part of their skeleton or at least an important substructure, otherwise the method will require additional chemical information. A wide variety of methods to superimpose molecules following different criteria are now available. According to Lemmen and Langauer[41] molecular alignment methods may be classified as (i) matching,[42−45] (ii) optimization,[46−48] (iii) grid,[49−51] and (iv) graph based techniques.[52,53] However one must be aware that depending on the particular molecular comparison

* Corresponding author phone: 57-1-3165000 ext 18324; fax: 57-1-3165220; e-mail: eedazac@unal.edu.co.

10.1021/ci7001878 CCC: $40.75 © 2008 American Chemical Society
Published on Web 01/01/2008

method, different alignment techniques may be used and considerably different results may be achieved from one technique to another. For example, when using the Carbó similarity index, molecules may be aligned according to the TGSA[42] or QSSA[46] methods. When comparing both methods[46,54] it is found that TGSA alignments may produce more "chemically correct" similarity values, but the more robust QSSA alignments may produce considerably higher values, in some cases being in counter of the "common chemical sense". Despite the disadvantages mentioned above, AD approaches have been extensively used; two of the most successful ones are the Comparative Molecular Field Analysis method (CoMFA)[34] and the Molecular Quantum Similarity Measures (MQSM) introduced by Carbó,[4] which have shown remarkable results in different sets of molecules.

In this paper we introduce a similarity measure based on negative molecular isopotential surfaces. We propose to represent the MEP through a graph, more precisely, by a weighted rooted tree that encodes some geometrical information and topological relations of successive isopotential surfaces. This alternative approach overcomes the difficulties of molecular alignment and avoids the definition of some particular descriptors to represent the MEP or the molecule itself. For a first example we took 46 small molecules, which represent eight different functional groups. A classification by similarity gives rise to a clear partition of molecules according to their *chemical function*. In a second example, we evaluated the proposed similarity measure, regarding the most popular methods, by means of Quantitative Structure−Activity Relationship (QSAR) and Structure−Activity Relationship (SAR), over the well-known 31 steroid set;[34] similar or better results were obtained when compared with the most widely used methods.

METHODS

Molecular Isopotential Surfaces. The molecular electrostatic potential, MEP, is an scalar field that can be calculated from the expression

$$V(r) = \sum_A \frac{Z_A}{|R_A - r|} - \int \frac{\rho(r')dr'}{|r' - r|} \qquad (1)$$

where $\rho(r')$ is the electronic density function obtained from the standard electronic wave function.

In order to grasp the behavior of the electrostatic potential in the surroundings of a molecule one may consider Molecular Isopotential Surfaces. As they may be computed for positives or negatives cutoffs, we may distinguish two big families of isopotential surfaces for neutral molecules: Positive Molecular Isopotential Surfaces (PMISs) or Negative Molecular Isopotential Surfaces (NMISs). The positive ones resemble the isodensity surfaces, since they wrap the nuclear skeleton.[3,55] Therefore the shape of these surfaces is relatively easy to visualize since their values are strongly determined by the positive charges at nuclear positions (see Figure 1). On the other hand, the Negative Molecular Isopotential Surfaces are fundamentally determined by accumulation of the electronic charge, i.e., they are strongly influenced by the quantum nature of electrons.[55] In Figure 1 we show the NMIS for three different values for the aldosterone steroid (−0.05, −0.08, and −0.11 au), calculated at the HF/6-31G level. One can see that each NMIS may be composed by one or more connected components. For example, the surface corresponding to a potential value of −0.11 au (blue) has two components, while the one at −0.08 au (red) has three components. Furthermore, we can see that all the components at −0.05 au *contain* or *enclose* all the components at −0.08 au and that these also contain all the components obtained at −0.11 au. One may imagine the components as expanding balloons merging each other as the potential approaches to zero.

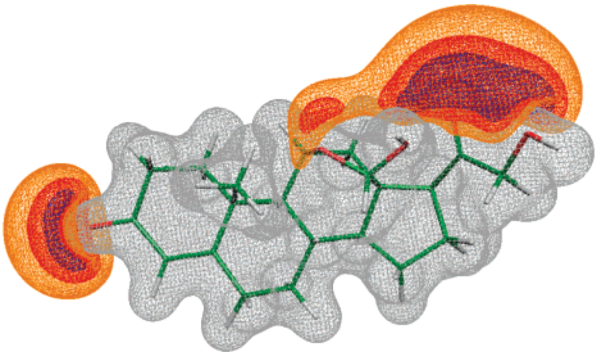

Figure 1. Some isopotential surfaces for aldosterone. In gray, the Positive Molecular Isopotential Surface (PMIS) calculated at $V = 0.11$ au, which encloses the molecular skeleton. In blue, red, and yellow, the Negative Molecular Isopotential Surfaces (NMISs) at $V = -0.11$, $V = -0.08$, and $V = -0.05$ au, respectively.

A detailed examination of the NMISs obtained in a MEP scan, for several neutral molecules, led us to conclude that for two successive NMISs A and B such that $V_A < V_B$, where V_i is the cutoff for NMIS i, all the components of the NMIS A are encompassed by one or more components of the NMIS B. These topological properties exhibited by the NMISs that reflect the concept of connectedness and interior subset were used to map the MEP into a rooted tree. These trees will be the mathematical objects upon which the similarity measure will be defined.

Mapping Negative Molecular Isopotential Surfaces into Rooted Trees. Rooted trees may encode not only the number of components but the hierarchical ordering derived from the parent-child relations of NMISs as it is illustrated in the following example. Six successive NMISs for aldosterone and their mappings into a rooted tree (drawn simultaneously at the right side) are shown in Figure 2a−f. As we will always use rooted trees, in the following we will refer to them just as trees. In this example the potential scan goes from −0.15 au to −0.05 au, with a step-size of 0.02 au. For the first cutoff ($V = -0.15$ au) the NMIS has two connected components, each one of them represented by a node in the tree (nodes 1 and 2). For $V = -0.13$ au there is just one component, to which corresponds one node in the tree (node 3). Besides, it must be noted that the component represented by node 3 encompasses the two former components represented by nodes 1 and 2, that merge by passing from $V = -0.15$ to $V = -0.13$. These parental relations are mapped into edges joining the new node (the parent) with each of the previous ones (the children).

When increasing the potential up to $V = -0.11$ the corresponding NMIS also has two connected components, one containing the previous one (node 4), with only one child represented by the edge 3−4, and another that is represented by the node 10. The scan procedure is followed until the

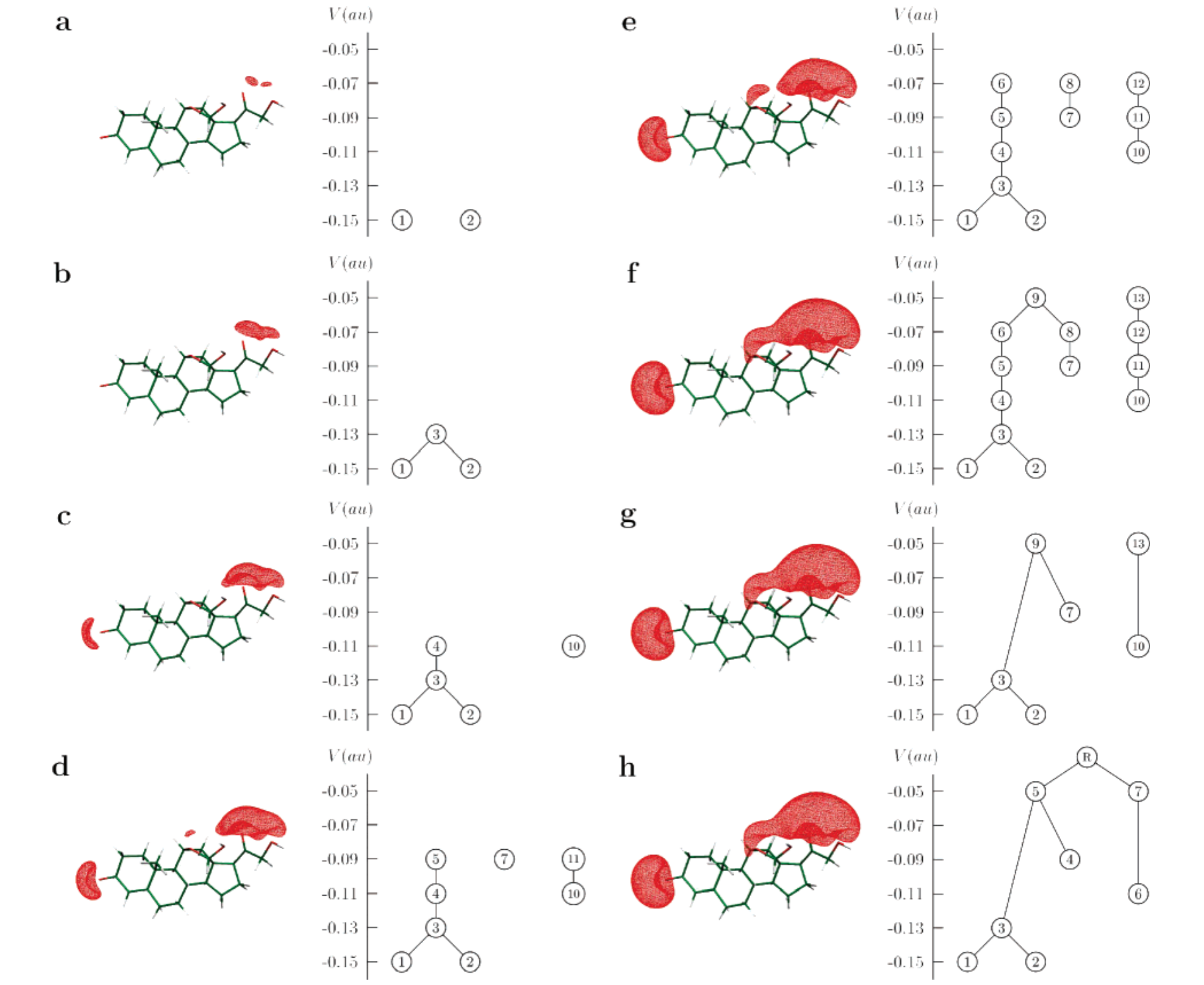

Figure 2. Mapping NMISs into a rooted tree. The potential scan goes from −0.15 to −0.05, with a step-size of 0.02 au: (a) $V = -0.15$ au, (b) $V = -0.13$ au, (c) $V = -0.11$ au, (d) $V = -0.09$ au, (e) $V = -0.07$ au, (f) $V = -0.05$ au, and (g) nodes with only one child, except nodes in the higher value are deleted. (h) In the final simplified rooted tree, the not-connected branches are connected by means of an artificial root node and nodes are numbered again.

maximum cutoff is reached and the tree is almost complete (Figure 2f). Thus, the fusion of components is observed at −0.13 au and −0.05 au, and the new not-connected components appear at −0.15 au, −0.11 au, and −0.09 au. The parent-child relations are given by the respective edges. The number of nodes in the tree depends on the step-size of the potential scan, thus for a step-size two times smaller than the one used to obtain Figure 2, the resulting tree would have one additional node between any pair of connected nodes. In order to reduce this dependence and avoid an excessive number of nodes we have decided to delete all nodes having exactly one child except those corresponding to the maximum cutoff (nodes 9 and 13 in Figure 2f). Notice that we are just interested in connectedness and inclusion relations. The resulting simplified tree is shown in Figure 2g. The final step in the mapping process is to connect all the branches into a single root node R to obtain the tree (Figure 2h). Since there is not any theoretical reason or empirical evidence that demonstrates that this root node will always arise in a natural way in neutral molecules, as a NMIS of only one component that encloses all the previous NMISs, it must be clear that this node is artificial, and we have proposed it in order to obtain only one connected graph per molecule. The existence of a single-component NMIS is demonstrated for negative charged molecules by Pathak et al.,[56] and several examples are shown by Gadre et al.[57] Once we represent each molecule by one simplified tree, we have gone from a 3D shape comparison problem to a tree comparison problem, the latter with many available solutions.[58−62] In other words, the similarity between MEPs will be regarded as the similarity between their corresponding trees.

The Tree Edit Distance and the Similarity Measure. In order to measure how close a couple of trees is, we used a tree edit distance, which is based on the optimum sequence of edit operations needed to convert one tree into another.[58] Let $S = s_1,...,s_n$ be a sequence of edit operations transforming the tree T_1 into the tree T_2. The cost of S is the sum of the costs of each edit operation in S given by $\gamma(S) = \sum_{i=1}^{n} \gamma(s_i)$. The edit distance between trees T_1 and T_2, denoted by $\delta(T_1, T_2)$, is the cost of the sequence S such that $\gamma(S)$ is minimum. The formal definition of the edit distance between two trees T_1 y T_2 is

$$\delta(T_1,T_2) = \min\{\gamma(S)\} \tag{2}$$

The possible edit operations considered in this metric are *inserting*, *deleting*, and *changing* nodes, and their respective costs are given by some weighting factor associated with each node. Since large NMIS are associated with zones around the molecule with large negative values of the MEP, i.e., high reactive sites, we have chosen the superficial area (A) of each connected component as weighting factor. It is worth noting that the superficial area has been used before to characterize MEP surfaces.[30−33] Accordingly, the edit operations have the following associated costs: (i) $\gamma(i \rightarrow \O) = A_i$ for the deletion of the node i, (ii) $\gamma(\O \rightarrow j) = A_j$ for the insertion of the node j, and (iii) $\gamma(i \rightarrow j) = |A_i - A_j|$ for the change of the node i into j. Since the superficial areas are always positive real numbers, the cost function is defined in the one-dimensional metric space $\mathbb{R}^+$, where

$$(1)\ \gamma(i \rightarrow j) \geq 0 \text{ and } \gamma(i \rightarrow i) = 0$$

$$(2)\ \gamma(i \rightarrow j) = \gamma(j \rightarrow i)$$

$$(3)\ \gamma(i \rightarrow k) \leq \gamma(i \rightarrow j) + \gamma(j \rightarrow k)$$

then the γ cost function is a metric, and by definition δ also fulfills the requirements to be a metric distance.[62−64]

Once the costs are defined the distance between trees may be computed, and a corresponding molecular similarity measure between any pair of molecules A and B is defined as

$$S_{AB} = \left(1 - \frac{\delta(A, B)}{\delta_{\max}}\right) \times 100 \tag{3}$$

where $\delta(A, B)$ is the edit distance between the trees of molecules A and B, and $\delta_{\max}$ is the maximum distance value for any pair of molecules in the set.

We have called our method to compare MEPs and measure molecular similarity: Tree Analysis and Representation of Isopotential Surfaces, TARIS. We have also developed a program to use the method with diverse sets of molecules, and it is available at http://taris.sourceforge.net. The software needs as input information a file containing the MEP values in a 3D grid enclosing the molecule (for example a Gaussian *.cube* file[65]) and the minimal and maximal cutoffs with the corresponding step size for the MEP scan. Once the program has loaded the cube files for all the molecules considered in the similarity analysis, it builds the tree for each molecule, computes the edit distance for all the possible pair of trees, and calculates the corresponding similarity values.

Molecular Set for a Classification by Functional Groups. As a first test of the discriminating capabilities of our methodology, we carried out a similarity classification of 46 small molecules that represent eight functional groups: (i) alcohols, (ii) ethers, (iii) aldehydes, (iv) ketones, (v) carboxylic acids, (vi) esters, (vii) amides, and (viii) amines (see Table 1). To classify these molecules the 46 × 46 similarity matrix provided by TARIS was analyzed by means of hierarchical clustering using the average linkage method.[66] The geometry of these molecules was optimized by B3LYP/6-31G** calculations, and the 3D grids with the MEP information were obtained with a resolution of 1 000 000 points/box. The potential scan was made from −0.2 to −0.04 au each 0.001 au.

Molecular Set for SAR and QSAR. We also tested our methodology studying the binding affinity to the human corticosteroid-binding globulins (CBG) for the well-known set of 31 steroids introduced by Cramer et al.[34] (see Table 2); these molecules have also been used as a benchmark set for several methods.[34−37,67,68] Single point quantum chemical calculations over this set were carried out at a HF/6-31G level using the nuclear geometries, previously corrected by Wagner et al.[37] as they are available from the Institute of Computational Chemistry of the University of Girona.[69] The MEP 3D grids were obtained with a resolution of 64 000, 216 000, 512 000, and 1 000 000 points/box, and the scan of the MEP was made from −0.2 to −0.05 au.

The trees obtained from those molecules were compared, and the corresponding similarity matrix was used in a hierarchical clustering analysis —using the average linkage method— to find out structure−activity relationships. A QSAR study using Partial Least-Squares (PLS) was performed following the scheme introduced by Good et al.,[36] which considers that the row vector containing the similarity indices among the molecule i and the remaining ones can be considered a good representation of the molecule in the set (eq 4). For other examples of QSAR models based on similarity matrices see refs 67, 68, and 71−73.

$$\mathbf{S}_i = [S_{i1}\ S_{i2} \cdots S_{ii} \cdots S_{in}] \tag{4}$$

All the quantum chemical calculations (for both molecular sets) were carried out with the Gaussian98 Rev. A11 program suite.[70] All the similarity calculations were carried out on a personal computer with 2.0 GB RAM and a 3.0 GHz Xeon processor.

RESULTS AND DISCUSSION

Classification by Functional Groups. In Figure 3 we show the dendrogram obtained from the similarity matrix of the 46 organic molecules specified in Table 1. It may be seen that molecules were classified according to their functional group, which is in accordance with the more basic chemical classification of organic molecules. Besides, when examining the clustering in more detail it is also possible to observe additional agreement with the classical chemical knowledge: aldehydes appear very similar to ketones (carbonyl compounds) and carboxylic acids appear very similar to esters (carboxylic compounds). In contrast to the next example, where relatively similar molecules are treated, in this molecular set we have covered a broad variety of molecules with different molecular formulas (not easy to superimpose) presenting different chemical behaviors, e.g., from acids to bases, from polar to nonpolar, from electrophilic to nucleophilic compounds. Thus we may say that the MEP analysis that we are introducing is able to recognize very similar molecules (from the chemical point of view) at the same time that molecules with different chemical behavior may be distinguished.

Some open questions remain when chemical similarity is measured as shape similarity within a molecular set of so-called *non-congener* molecules. For example if we compare *n*-hexylamine, methylamine, and *n*-hexanol according to their shapes, we should expect that *n*-hexylamine and *n*-hexanol

Table 1. Formula and Id for the 46 Small Organic Molecules Used for the Functional Group Classification

functional group	molecule Id	R[a]	R_1	functional group	molecule Id	R	R_1
alcohol	**A1**	methyl		carboxylic acid	**E1**	methyl	
	A2	ethyl			**E2**	ethyl	
	A3	*n*-propyl			**E3**	*n*-propyl	
	A4	*n*-butyl			**E4**	*n*-butyl	
	A5	*n*-pentyl			**E5**	*n*-pentyl	
	A6	*n*-hexyl			**E6**	*n*-hexyl	
ether	**B1**	methyl	*n*-pentyl	ester	**F1**	methyl	*n*-pentyl
	B2	methyl	*n*-butyl		**F2**	ethyl	*n*-butyl
	B3	methyl	*n*-propyl		**F3**	*n*-propyl	*n*-propyl
	B4	ethyl	*n*-butyl		**F4**	*n*-butyl	ethyl
	B5	ethyl	*n*-propyl		**F5**	*n*-pentyl	methyl
	B6	ethyl	ethyl	amide	**G1**	methyl	*n*-pentyl
	B7	*n*-propyl	*n*-propyl		**G2**	ethyl	*n*-butyl
aldehyde	**C1**	methyl			**G3**	*n*-propyl	*n*-propyl
	C2	ethyl			**G4**	*n*-butyl	ethyl
	C3	*n*-propyl			**G5**	*n*-pentyl	methyl
	C4	*n*-butyl		amine	**H1**	methyl	
	C5	*n*-pentyl			**H2**	ethyl	
	C6	*n*-hexyl			**H3**	*n*-propyl	
ketone	**D1**	methyl	*n*-butyl		**H4**	*n*-butyl	
	D2	methyl	*n*-propyl		**H5**	*n*-pentyl	
	D3	methyl	ethyl		**H6**	*n*-hexyl	
	D4	ethyl	*n*-propyl				
	D5	ethyl	ethyl				

[a] These are the carbon chains in the following general nomenclature: for alcohols R-OH, ethers R-O-R_1, aldehydes R-CO, ketones R-CO-R_1, carboxylic acids R-COOH, esters R-COO-R_1, amides R-CONH-R_1, and amines R-NH_2

Table 2. Steroids and Their Corticosteroid-Binding Globulin (CBG) Binding Affinities[37]

no.	name	CBG (log*K*)	no.	name	CBG (log*K*)
1	aldosterone	6.279	**17**	pregnenolone	5.225
2	androstanediol	5.000	**18**	hydroxypregnenolone	5.000
3	5-androstanediol	5.000	**19**	progesterone	7.380
4	4-androstenedione	5.763	**20**	hydroxyprogesterone	7.740
5	androsterone	5.613	**21**	testosterone	6.724
6	corticosterone	7.881	**22**	prednisolone	7.512
7	cortisol	7.881	**23**	cortisolacetate	7.553
8	cortisone	6.892	**24**	4-pregnene-3,11,20-trione	6.779
9	dehydroepiandrosterone	5.000	**25**	epicorticosterone	7.200
10	11-deoxycorticosterone	7.653	**26**	19-nortestosterone	6.144
11	11-deoxycortisol	7.881	**27**	16a,17a-dihydroxyprogesterone	6.247
12	dihydrotestosterone	5.919	**28**	17a-methylprogesterone	7.120
13	estradiol	5.000	**29**	19-norprogesterone	6.817
14	estriol	5.000	**30**	2a-methylcortisol	7.688
15	estrone	5.000	**31**	2a-methyl-9a-fluorocortisol	5.797
16	ethiochonalonone	5.225			

appear more similar than *n*-hexylamine and methylamine, in counter of some aspects of their chemical behavior. Since our approach does not take into account explicitly all the molecule but only some *relevant sites*, when *non-congener* molecules are studied, chemical similarity will not be determined by shape or size similarity but for those particular sites that could give to the molecules some of their chemical identity in the context of MEP.

The relative small size and simplicity of these molecules have allowed us to describe some features of the geometrical and topological information encoded in the trees. In Figure 3 it is possible to observe that the similarity values among amines, ethers, and alcohols are considerably high in comparison with the other groups. This occurs because the NMIS for these molecules begins as only one connected component that remains as the only one during the whole potential scan. Therefore these molecules are represented by very simple trees with just three nodes, one for the minimum MEP value, another for the maximum cutoff and the root node. Since there are not topological differences then the distances among these trees are based only on the weighting factors (i.e., the area values). For the other molecules, the NMISs have more than one connected component giving rise to topologically different trees. When topological differences appear, larger distance values are obtained, making similarity values not too high.

From this example it is clear that the proposed methodology shows a high discrimination capability, despite some exceptions (**C1**, **E1**, and **F1**).

SAR and QSAR for the Steroids Set. First, we will show the PLS results obtained for different conditions: we studied the effect of changing the step-size, the resolution of the MEP grid and the maximum cutoff value. We built several QSAR models to evaluate the behavior of the method on these three parameters. For each PLS analysis we have also considered different numbers of molecules ($n = 1, \ldots, 5$) in the leave-*n*-out cross-validation scheme (*lno*). In all cases the first two Principal Components (PCs) were employed.

Step-Size Effect. In Table 3 we show the q^2 and r^2 coefficients for six different step-sizes. After the examination

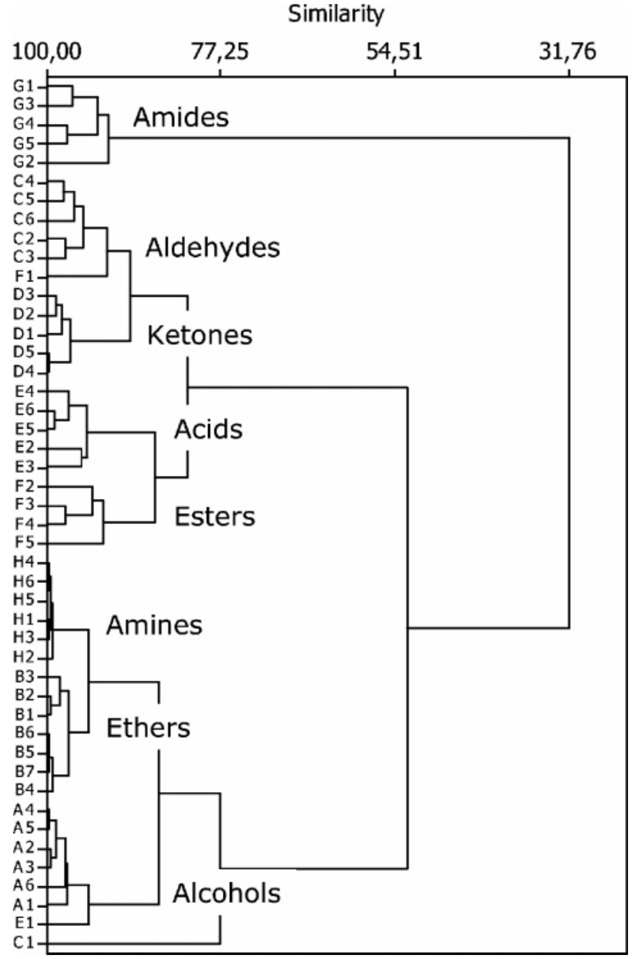

Figure 3. Dendrogram obtained from the similarity matrix for the 46 organic molecules with a 0.001 au stepsize and a 10^6 points/box resolution, using average linkage.

Table 3. Cross-Validated and Fitted Correlation Coefficients for the 31 Steroid Set

	q^2						
step-size[a]	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	r^2	time[b] (s)
0.05	0.64	0.65	0.65	0.66	0.63	0.73	134
0.02	0.67	0.68	0.66	0.62	0.65	0.78	140
0.01	0.69	0.69	0.68	0.63	0.67	0.75	290
0.005	**0.71**	**0.71**	**0.70**	**0.66**	**0.69**	**0.76**	490
0.002	0.66	0.67	0.66	0.64	0.67	0.74	1100
0.001	0.64	0.63	0.62	0.58	0.64	0.72	2160
average	0.67	0.67	0.66	0.63	0.66	0.75	
deviation	0.04	0.04	0.04	0.04	0.03	0.03	

[a] The resolution of the grid in all cases is 1 million points per box. [b] Time to generate the similarity matrix.

of q^2, it is clear that for each *lno* scheme there is a maximum for a step-size of 0.005, and it is also clear that the extreme step-sizes (0.05 and 0.001) exhibit the lowest cross-validated coefficients. Only one exception to this tendency is observed for $n = 4$. In this particular case q^2 is the same for two step-sizes: 0.05 and 0.005. The appearance of a common optimal step-size is explained by the loss of information in the trees or noise introduced with very small steps. Notice that low q^2 values are obtained for large step-sizes because there is a deficient encoding of the topology of the connected components of the NMISs. A big step-size reduces the computational time, but it also reduces the level of detail in the MEP analysis. Under these conditions the parent-child relations and the connected NMISs may be too crude to be a good representation of MEP. For small step-sizes the mapping into a tree becomes quite slow, and significant quantities of "noise" are included. A small step-size generates several tiny NMISs components closely located, so trees with many leaves are obtained, and they may look very different according to the distance we use. However, it must be noticed that in both cases, all deviations from the average q^2 are below 0.04. This small deviation let us to state that there is not a significant dependence on the step-size and that 0.005 is the step-size that offers the highest q^2 values and reasonable low computational time.

Table 4. Cross-Validated and Fitted Correlation Coefficients for Several Resolutions of the MEP Grid

points/box	$q^{2\,a}$	r^2	time (s)
40^3	0.66	0.74	23
60^3	0.66	0.72	88
80^3	0.68	0.74	245
100^3	0.71	0.76	490

[a] Results obtained with a 0.005 step-size and the leave-one-out scheme using the first two PCs

It is worth mentioning that the average values of the cross-validated coefficients through the five *lno* schemes show almost no dependence on the size of the groups used. Only when $n = 4$ slightly low coefficients are obtained, but for the other four *lno* schemes q^2 is between 0.69 and 0.71, for the 0.005 step-size. We believe that this is a relevant result since the predictive power of the model is preserved even when the number of individuals utilized in the construction of the model decreases. This is also important because in larger sets of molecules the leave-one-out scheme could be considerably time-consuming, and we may expect that leaving more molecules out will not produce a considerable decrease in q^2.

MEP Grid Resolution Effect. The second parameter we tested was the resolution of the grid. A higher resolution of the grid implies a better approximation to the surfaces, since they are smoother, and, as a consequence, the detection of the connected components is more precise. In Table 4 we show the correlation coefficients for four different resolutions and the time required for each calculation. As it is expected, for higher resolutions larger times are required; however, changes in q^2 are small. Decrease of the resolution, e.g., from 10^6 to 80^3 (in time from 490 to 245 s) only reduces the coefficient from 0.71 to 0.68; therefore, if the size of the molecules in the set is large enough to make the calculation of high-resolution grids not practicable, reducing the resolution will not have a strong effect on the quality of the results.

Maximum Cutoff Effect. Since the main idea underlaying the TARIS procedure is to characterize the negative MEP field, if wider potential ranges are scanned, then more information about molecules will be encoded, and better QSAR results might be obtained. In order to corroborate this a priori idea, several PLS models were constructed varying the maximum cutoff from −0.110 to −0.050 au. The cross-validated correlation coefficients for each case are in Table 5. For cutoffs from −0.110 to −0.095 au there is a random behavior of the q^2 values, but from −0.090 to −0.050 au there is a clear tendency that shows that for wider scans better correlation values are obtained as we expected to happen.

Table 5. Cross-Validated Correlation Coeffcients and Average Similarity Values for Different Maximum Cutoff Values

cutoff (au)	$q^{2\,a}$	$\delta_{\max}$	average S_{AB}
−0.110	0.68	47.030	68.5
−0.105	0.75	48.842	67.5
−0.100	0.71	52.140	66.8
−0.095	0.66	56.064	66.0
−0.090	0.64	59.138	64.0
−0.085	0.65	62.496	62.4
−0.080	0.65	66.076	61.3
−0.075	0.66	69.869	59.4
−0.070	0.67	75.093	58.9
−0.065	0.69	89.508	62.2
−0.060	0.70	158.933	74.4
−0.055	0.70	166.810	73.7
−0.050	0.71	175.680	72.7

[a] Results obtained with a 0.005 step-size and the leave-one-out scheme using the first two PCs.

This led us to state that the ideal upper limit for the maximum cutoff value is 0 au. However due to fact that the size of the NMISs grows as the MEP approximates to 0 au, the main limiting factor is the space covered by the 3D grid. Since some molecules present NMISs that fell out of the MEP box (i.e., these NMISs are truncated by the box walls), this makes it impossible to determine some NMIS connected components and their superficial areas, thus affecting the construction of the tree. In the particular case of the 31 steroids this happens when cutoffs greater than −0.05 au are considered.

In Table 5 there are also reported the $\delta_{\max}$ values and the average similarity values of the corresponding similarity matrices for every cutoff. Yet the maximum distances always grow when the cutoff is increased (because more and bigger NMIS components are involved); the similarity values do not follow a clear tendency, they rather show a random behavior. Since $\delta_{\max}$ values grow and the same happens with $\delta(A, B)$, the effect on the average of $(\delta(A,B))/(\delta_{\max})$ is small and not well defined, as it is shown in the fourth column in Table 5.

The use of a normalized similarity measure like eq 3 has a bonus since it measures the similarity of a pair of molecules having a third one as reference and not considering similarity as a fixed measure between a couple of molecules, i.e., it is more meaningful to say that molecule A is similar to molecule B regarding molecule C (the context) than just saying that molecules A and B are similar without taking into account the context of the comparison. In order to illustrate this let us examine the next example: based on the chemical behavior, one may say that acetic acid and phenol are not very similar molecules if the reference molecule is chloroacetic acid. However if the reference molecule is not chloroacetic acid but *n*-octane they appear now to be more similar. Thus even when the distance values do not depend on the context, the similarity values do.

The results above show that the best parameters for the method are as follows: step-size = 0.005 au, resolution = 10^6 points/box, maximum cutoff = −0.05 au for which q^2 = 0.71 and r^2 = 0.76. These conditions are used in further analyzes and comparisons.

Internal Consistency of Distance Matrices. Although the similarity measure is based on a distance measure that fulfills the requirements of a metric distance,[63] and therefore it must satisfy the triangular inequality, we have performed an internal consistence analysis of the distance matrices to ensure that the numerical procedure has not led to erroneous results.[54]

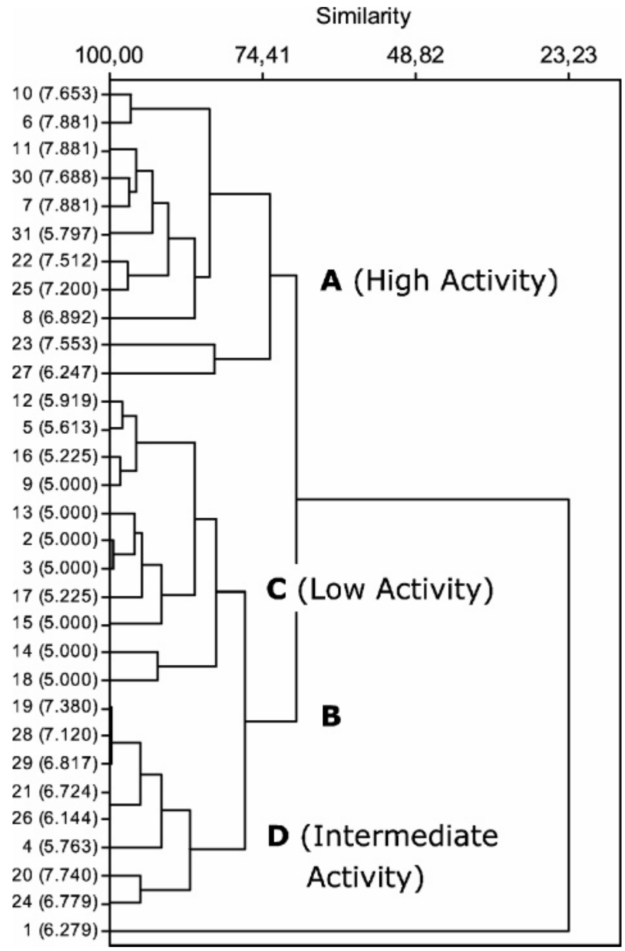

Figure 4. Dendrogram obtained from the 31 steroids similarity matrix with a 0.005 au stepsize and a 10^6 points/box resolution, using average linkage. The log*K* values are in parentheses.

In this work, all the distance matrices used were checked for internal consistency following the methodology described by Bultinck et al.,[54] and in all cases no violations were found. In this way we confirm the internal consistency of our similarity study and the metric behavior of the tree edit distance under the implementation of TARIS.

Clustering Analysis. The similarity values among the trees representing MEPs of the steroids were directly used in a hierarchical clustering analysis to find out if the groups deduced from the similarity measure are in agreement with their experimental activity. In Figure 4 we show the dendrogram obtained for the 31 steroids, using average linkage, for a 0.005 au step-size and a 10^6 points/box resolution.

Three main groups may be clearly identified, each one corresponding to high, intermediate, and low activities. For a 70% of similarity, there are two branches (A and B). The A branch corresponds to the high activity steroids ($7.881 \geqslant \log K \geqslant 6.892$). The branch B splits up in other two branches (C and D), which correspond to intermediate ($6.892 > \log K \geqslant 6.144$) and low activities ($6.144 > \log K \geqslant 5.000$). Only a few steroids are misclassified: **19**, **20**, and **28** should belong to the high activity group but appear in the intermediate group; **27** is classified as high but it should be intermediate;

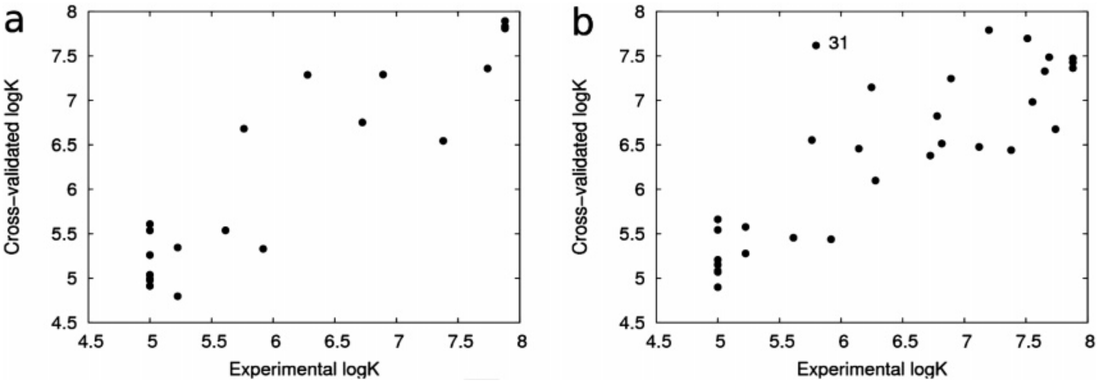

Figure 5. Cross-validated versus experimental log*K* values of CBG affinities with the best TARIS parameters: (a) for the set of the first 21 steroids and (b) for the set of 31 steroids.

Table 6. Cross-Validated Correlation Coefficients for Some Methods Based on the MEP

method	steroid set[a]	q^2	PCs[b]	alignment-dependent
Richards SM[36]	21	0.501	1	yes
CoMSIA[35]	21	0.526	4	yes
CoMFA[36]	21	0.644	1	yes
CoMFA[c 34]	21	0.662	2	yes
TQSAR[d 67]	21	0.832	6	yes
TARIS	21	0.84	5	no
Wagener[37]	31	0.63		no
TARIS	31	0.71	2	no
MQSM[67]	31	0.759	5	yes
Hodking SM[68]	31	0.903	6	yes

[a] The set of 21 steroids correspond to the first 21 in Table 2. [b] When it is applicable, specifies the number of components. [c] For this particular method besides MEP information, steric effects are considered too.[d] This procedure is based on a combination of the overlap, coulomb, and triple-density similarity matrices.

and **4** should be low but is classified as intermediate. Finally, **31** should belong to the group of low activities, but it is classified in the high activity group; this is the only molecule that is totally misclassified. One must notice that this molecule is considered as an outlier in most of the preceding studies.[34−37,67,68] We found that the steroid **1** is out of the groups defined by the clustering, and a similar result was obtained by Bultinck et al., using the Carbó similarity index and hierarchical clustering. However, in that work it was not possible to classify the set of molecules properly, i.e., in agreement with their biological activity.[71] Our results are also as good as those obtained with other methods such as Neural Networks.[36,37]

To compare with previous studies, based solely on the comparison of molecular electrostatic potentials, we consider two molecular sets: one composed by the whole set of 31 steroids and another only by the first 21 of them. In Figure 5 we show the cross-validated versus the experimental CBG affinities for both sets. For the whole set, it is evident that the **31** steroid is an outlier, as previously reported.[34−37,67,68]

In Table 6 we show the results obtained by such methods for either the set of 31 or the subset of the first 21 steroids. For a quick reference of these methods the reader may see ref 67. For a detailed description of each method, the original reference is cited in the table.

In order to identify some of the advantages of our methodology, we focus on the obtained q^2 and on the dependence on the molecular alignment of each method. First, it should be noticed that for the set composed by 21 steroids, the cross-validated coefficient corresponding to the present calculation is the highest one (0.84). Similar results are only obtained by the TQSAR method (0.832); however, this procedure requires not only electrostatic but also additional information. Here the analysis is carried out over a combination of three similarity matrices (overlap, coulomb, and triple-density matrices) which demands more resources to be calculated. The next highest q^2 is that of the CoMFA method with 2 PCs (0.662), which also contains additional information. The methods that only contain MEP information have q^2 below 0.65, which is considerably less than the value reported in this work. For the entire 31 steroids set the best results are obtained by the Hodking similarity measure employing a Genetic Algorithm/Neural Networks method (0.903), but molecular alignment must be achieved previously. Most of the methods considered in Table 6 are alignment-dependent, only that one proposed by Wagener avoids this previous step.

The molecular alignment step is undesirable because it constitutes a bottleneck in similarity studies, not only the procedure itself but also because of the dependence of the similarity measures on the chosen method.[4,5,42,46,54,53] A remarkable issue of the approach that we are introducing is that the molecular superposition is avoided, thus the effort can be directed to the proper characterization of the molecules for example through quantum mechanical properties.

In this work, the compared trees are derived from several NMISs, which are 3D objects defined from quantum mechanical electronic features of the molecule, encoding topological properties of the MEP. Besides, the scan of the MEP assures that a representative range of potential values on a wide region of the surroundings of the molecule is considered. Thus, it is not only one particular surface but the scalar field which is being compared in a nonvisual way. In this way we avoid choosing a limited 3D surface as to where to calculate the MEP, e.g., a given van der Waals or isodensity surface.

This indirect representation of molecular structure by means of trees of negative molecular electrostatic potential encodes information that led us to establish net relationships with the reactivity of a set of molecules, as it was shown with the example above.

CONCLUSIONS

A methodology for nonvisual comparison of Molecular Electrostatic Potentials from a graph theoretical perspective was developed. It has been shown that geometrical and topological information of successive Negative Isopotential Surfaces can be encoded in a weighted rooted tree (a graph) and that the comparison of such trees through a metric measure, an edit distance in this case, may be used to obtain a molecular similarity measure. The proposed method does not need the molecular alignment, which is required in most of the MEP molecular similarity methods; this characteristic can be consider as one of the most remarkable advantages of this approach.

The similarity study of a representative set of molecules encompassing different functional groups gave a classification in close agreement with the classical ideas of organic chemistry, showing that the method can be applied to compare non-congener molecules without the issues associated with the optimal alignment of these type molecules.

In this work the similarity matrix for the well-known 31 steroids set was analyzed by means of hierarchical clustering and Partial Least-Squares. A proper partition in high, intermediate, and low activities was obtained. The cross-validated correlation coefficients of the QSAR are comparable to those from other MEP based similarity methods and showed a low dependence on the step-size and on the 3D grid resolution.

The method we are proposing can be applied to the analysis of MEP obtained not only from quantum mechanical calculations but from other classical or semiclassical sources. This Tree Analysis and Representation of Isopotential Surfaces method is also applicable to the analysis of Positive Molecular Isopotential surfaces and could be extended to electronic isodensity surfaces or to other molecular scalar fields. Further applications are being developed in our group.

ACKNOWLEDGMENT

Particular acknowledge is to our colleges A. Reyes and A. Bernal, for their useful discussions during the preparation of the manuscript. We want to thank one of the referees for his helpful comments to improve this manuscript. We also acknowledge COLCIENCIAS for the financial support.

Supporting Information Available: Distance matrices for the 46 organic molecules, the 21 and the 31 steroids sets, together with the corresponding $\delta_{\max}$ values. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Rouvray, D. H. Are the concepts of chemistry all fussy? In *Concepts in Chemistry: A Contemporary Challenge*; Rouvray, D. H., Ed.; John Wiley and Sons Inc.: New York, 1997; pp 1−15.

(2) Rouvray D. H. Definition and Role of Similarity Concepts in the Chemical and Physical Sciences. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 580−586.

(3) Mezey, P. G. *Shape in Chemistry. An Introduction to Molecular Shape and Topology*; VCH Publishers: New York, 1993; pp 83−88.

(4) Bultinck, P.; Gironés, X.; Carbó-Dorca, R. Molecular Quantum Similarity: Theory and Applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Larter, R., Cundari, T. R., Eds.; John Wiley and Sons, Inc. Publishers: New York, 2005; Vol. 21, pp 127−207.

(5) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. Quantum Molecular Similarity: Theory and Applications to the Evaluation of Molecular Properties, Biological Activities and Toxicity. In *Fundamentals of Molecular Similarity*; Carbó-Dorca, R., Gironés, X., Mezey, P. G., Eds.; Kluwer Academic/Plenum Publishers: New York, 2001; pp 187−320.

(6) Villaveces, J. L.; Daza, E. E. The Concept of Chemical Structure. In *Concepts in Chemistry: A Contemporary Challenge*; Rouvray, D. H., Ed.; John Wiley and Sons Inc.: New York, 1997; pp 101−132.

(7) Villaveces, J. L.; Daza, E. E. On the Topological Approach to the Concept of Chemical Structure. *Int. J. Quantum Chem. Quantum Chem. Symp.* **1990**, *24*, 97−106.

(8) Good, A. C.; Richards, W. G. Rapid Evaluation of Shape Similarity Using Gaussian Functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112−116.

(9) Goldman, B. B.; Wipke, W. T. Quadratic Shape Descriptors. 1. Rapid Superposition of Dissimilar Molecules Using Geometrically Invariant Surface Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 644−658.

(10) Duca, J. S.; Hopfinger, A. J. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367−1387.

(11) Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem. Quantum Biol. Symp.* **1987**, *14*, 105−110.

(12) Randić, M. On Characterization of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672−687.

(13) Randić, M. The Connectivity Index 25 Years After. *J. Mol. Graphics Modell.* **2001**, *20*, 19−35.

(14) Galindo, J. F.; Bermúdez, C. I.; Daza, E. E. A Classification of Central Nucleotides Induced by The Influence of Neighboring Nucleotides in Triplets. *J. Mol. Struct.−THEOCHEM* **2006**, *769*, 103−109.

(15) Popelier, P. L. A. Quantum Molecular Similarity. 1. BCP Space. *J. Phys. Chem. A* **1999**, *103*, 2883−2890.

(16) McCoy, E. F.; Sykes, M. J. Quantum-Mechanical QSAR/QSPR Descriptors from Momentum-Space Wave Functions. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 545−553.

(17) Niño, M.; Daza, E. E.; Tello, M. A Criteria To Classify Biological Activity of Benzimidazoles from a Model of Structural Similarity. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 495−504.

(18) Klopman, G. Chemical Reactivity and the Concept of Charge- and Frontier-Controlled Reactions. *J. Am. Chem. Soc.* **1968**, *90*, 223−234.

(19) Parr, R. G.; Yang, W. Density Functional Approach to the Frontier-Electron Theory of Chemical Reactvity. *J. Am. Chem. Soc.* **1984**, *106*, 4049−4050.

(20) Li, Y.; Evans, J. N. S. The Fukui Function: A Key Concept Linking Frontier Molecular Orbital Theory and the Hard-Soft-Acid-Base Principle. *J. Am. Chem. Soc.* **1995**, *117*, 7756−7759.

(21) Chattaraj, P. K. Chemical Reactivity and Selectivity: Local HSAB Principle versus Frontier Orbital Theory. *J. Phys. Chem. A* **2001**, *105*, 511−513.

(22) Melin, J.; Aparicio, F.; Subramanian, V.; Galván, M.; Chattaraj, P. K. Is the Fukui Function a Right Descriptor of Hard−Hard Interactions? *J. Phys. Chem. A* **2004**, *108*, 2487−2491.

(23) Politzer, P.; Murray, J. S. Molecular Electrostatic Potentials and Chemical Reactivity. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Ed.; VCH Publishers: New York, 1991; Vol. 2, pp 273−312.

(24) Honig, B.; Nicholls, A. Classical Electrostatics in Biology and Chemistry. *Science* **1995**, *268*, 1144−1149.

(25) Chin, K.; Sharp, K. A.; Honig, B.; Pyle, A. M. Calculating the Electrostatic Properties of RNA Provides New Insights into Molecular Interactions and Function. *Nat. Struct. Biol.* **1999**, *6*, 1055−1061.

(26) Tworowski, D.; Safro, M. The Long-Range Electrostatic Interactions Control tRNA−Aminoacyl-tRNA Synthetase Complex Formation. *Protein Sci.* **2003**, *12*, 1247−1251.

(27) Tworowski, D.; Feldman, A. V.; Safro, M. Electrostatic Potential of Aminoacyl-tRNA Synthetase Navigates tRNA on its Pathway to the Binding Site. *J. Mol. Biol.* **2005**, *350*, 866−882.

(28) Cárdenas, C.; Villaveces, J. L.; Bohórquez, H.; Llanos, E.; Suárez, C.; Obregón, M.; Patarroyo M. E. Quantum Chemical Analysis Explains Hemagglutinin Peptide-MHC Class II Molecule HLA-DRβ1*0101 Interactions. *Biochem. Biophys. Res. Commun.* **2004**, *323*, 1265−1277.

(29) Cárdenas, C.; Ortiz, M.; Balbín, A.; Villaveces, J. L.; Patarroyo, M. E. Allele Effects in MHC-Peptide Interactions: A Theoretical Analysis of HLA-DRβ1*0101-HA and HLA-DRβ1*0401-HA Complexes. *Biochem. Biophys. Res. Commun.* **2005**, *330*, 1162−1167.

(30) Arteca, G. A.; Hernández-Laguna, A.; Rández, J. J.; Smeyers, Y. G.; Mezey, P. A Topological Analysis of Molecular Electrostatic Potential on van der Waals Surfaces for Histamine and 4-Substituted Derivatives as H_2-Receptor Agonists. *J. Comput. Chem.* **1991**, *12*, 705−716.

(31) Politzer, P.; Lane, P.; Murray, S. J.; Brinck, T. Investigation of Relationships between Solute Molecule Surface Electrostatic Potentials and Solubilities in Supercritical Fluids. *J. Phys. Chem.* **1992**, *96*, 7938−7943.

(32) Murray, S. J.; Politzer, P. Statistical Analysis of the Molecular Surface Electrostatic Potential: an Approach to Describing Noncovalent

Interactions in Condensed Phases. *J. Mol. Struct.−THEOCHEM* **1998**, *425*, 107−114.

(33) Chalk, A. J.; Beck, B.; Clark, T. A Quantum Mechanical/Neural Net Model for Boiling Points with Error Estimation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 457−462.

(34) 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.

(35) 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.

(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) 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.

(38) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(39) Rodrigo, J.; Barbany, M.; Gutiérrez-de-Terán, H.; Centeno, N. B.; de-Càceres, M.; Dezi, C.; Fontaine, F.; Lozano, J. J.; Pastor, M.; Villà, J.; Sanz, F. Comparison of Biomolecules on the Basis of Molecular Interaction Potentials. *J. Braz. Chem. Soc.* **2002**, *13*, 795−799.

(40) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity - a Review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.

(41) Lemmen, C.; Lengauer, T. Computational Methods for the Structural Alignment of Molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.

(42) Gironés, X.; Robert, D.; Carbó-Dorca, R. TGSA: A Molecular Superposition Program Based on Topo-Geometrical Considerations. *J. Comput. Chem.* **2001**, *22*, 255−263.

(43) Mills, J. E. J.; Perkins, T. D. J.; Dean, P. M. An Automated Method for Predicting the Positions of Hydrogen-bonding Atoms in Binding Sites. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 229−242.

(44) Krämer, A.; Horn, H. W.; Rice, J. E. Fast 3D Superposition and Similarity Search in Databases of Flexible Molecules. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 13−38.

(45) Meurice, N.; Maggiora, G. M.; Vercauteren D. P. Evaluating Molecular Similarity Using Reduced Representations of the Electron Density. *J. Mol. Model.* **2005**, *11*, 237−247.

(46) Bultinck, P.; Kuppens, T.; Gironés, X.; Carbó-Dorca, R. Quantum Similarity Superposition Algorithm (QSSA): A Consistent Scheme for Molecular Alignment and Molecular Similarity Based on Quantum Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1143−1150.

(47) Rönkkö, T.; Tervo, A. J.; Parkkinen, J.; Poso, A. BRUTUS: Optimization of a Grid-Based Similarity Function for Rigid-Body Molecular Superposition. II. Description and Characterization. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 227−236.

(48) Jewell, N. E.; Turner, D. B.; Willett, P.; Sexton, G. J. Automatic Generation of Alignments for 3D QSAR Analyzes. *J. Mol. Graphics Modell.* **2001**, *20*, 111−121.

(49) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A Molecular-Field-Based Similarity Study of Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 79−93.

(50) Klebe, G.; Mietzner, T.; Weber, F. Methodological Developments and Strategies for a Fast Flexible Superposition of Drug-Size Molecules. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 35−49.

(51) Cosgrove, D. A.; Bayada, D. M.; Johnson, A. P. A Novel Method of Aligning Molecules by Local Surface Shape Similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573−591.

(52) Thorner, D. A.; Wild, D. J.; Willett, P.; Wright, P. M. Similarity Searching in Files of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900−908.

(53) Marialke, J.; Körner, R.; Tietze, S.; Apostolakis, J. Graph-Based Molecular Alignment. *J. Chem. Inf. Model.* **2007**, *47*, 591−601.

(54) Bultinck, P.; Carbó-Dorca, R.; Van Alsenoy, C. Quality of Approximate Electron Densities and Internal Consistency of Molecular Alignment Algorithms in Molecular Quantum Similarity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1208−1217.

(55) Politzer, P.; Murray, J. S. The Fundamental Nature and Role of the Electrostatic Potential in Atoms and Molecules. *Theor. Chem. Acc.* **2002**, *108*, 134−142.

(56) Pathak, R. K.; Gadre, S. R. Maximal and Minimal Characteristics of Molecular Electrostatic Potentials. *J. Chem. Phys.* **1990**, *93*, 1770−1773.

(57) Gadre, S. R.; Shrivastava, I. H. Shapes and Sizes of Molecular Anions via Topographical Analysis of Electrostatic Potential. *J. Chem. Phys.* **1991**, *94*, 4384−4390.

(58) Zhang, K.; Shasha, D. Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM J. Comput.* **1989**, *18*, 1245−1262.

(59) Bunke, H.; Shearer, K. A Graph Distance Metric Based on the Maximal Common Subgraph. *Pattern Recognit. Lett.* **1998**, *19*, 255−259.

(60) Torsello, A.; Hancock, E. R. A Skeletal Measure of 2D Shape Similarity. *Comput. Vis. Image Und.* **2004**, *95*, 1−29.

(61) Zhong, Y.; Meacham, C. A.; Paramanik, S. A General Method for Tree-Comparison Based on Subtree Similarity and Its Use in a Taxonomic Database. *Biosystems* **1997**, *42*, 1−8.

(62) Neuhaus, M.; Bunke, H. Automatic Learning of Cost Functions for Graph Edit Distance. *Inf. Sci.* **2007**, *177*, 239−247.

(63) Bille, P. A survey on Tree Edit Distance and Related Problems. *Theor. Comput.* Sci. **2005**, *337*, 217−239.

(64) Conte, D.; Foggia, P.; Sansone, C; Vento, M. Thirty Years of Graph Matching in Pattern Recognition. *Int. J. Pattern Recognit.* **2004**, *18*, 265−298.

(65) Frisch, E.; Frisch, M. J. Gaussian 98 Users's Reference. Gaussian Inc.: Pittsburg, PA 15106 U.S.A, 1999; pp 66−68.

(66) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*, 5th ed.; Prentice Hall: Upper Saddle River, NJ 07458, 2002; pp 668−692.

(67) Robert, D.; Amat, L.; Carbó-Dorca, R. Three-Dimensional Quantitative-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.

(68) So, S.-S.; Karplus, M. Three-Dimensional Quantitative-Activity Relationships from Molecular Similarity Matrices and Genetic Neural Networks. 1. Method and Validations. *J. Med. Chem.* **1997**, *40*, 4347−4359.

(69) http://iqc.udg.es/cat/similarity/QSAR/steroids/ (accessed Mar 16, 2007).

(70) 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.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; 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.; 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.11*; Gaussian, Inc.: Pittsburgh, PA, 2001.

(71) Carbó-Dorca, R.; Bultinck, P. Molecular Quantum Similarity Matrix Based Clustering of Molecules Using Dendrograms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 170−177.

(72) 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.

(73) 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.

CI7001878