RED: A Set of Molecular Descriptors Based on Rényi Entropy

Laura Delgado-Soler,[†] Raul Toral,[‡] M. Santos Tomás,[§] and Jaime Rubio-Martinez*[,†]

Departament de Química Física, Universitat de Barcelona (UB) and the Institut de Recerca en Química Teòrica i Computacional (IQTCUB), Martí i Franqués 1, E-08028 Barcelona, Spain, Instituto de Física Interdisciplinar y Sistemas Complejos (IFISC), CSIC-UIB, Campus UIB, 07122-Palma de Mallorca, Spain, and Secció Matemàtiques ETSAB, Universitat Politècnica de Catalunya (UPC), ETS d'Arquitectura, Avinguda Diagonal 649, E-08028 Barcelona, Spain

Received July 28, 2009

New molecular descriptors, RED (Rényi entropy descriptors), based on the generalized entropies introduced by Rényi are presented. Topological descriptors based on molecular features have proven to be useful for describing molecular profiles. Rényi entropy is used as a variability measure to contract a feature-pair distribution composing the descriptor vector. The performance of RED descriptors was tested for the analysis of different sets of molecular distances, virtual screening, and pharmacological profiling. A free parameter of the Rényi entropy has been optimized for all the considered applications.

INTRODUCTION

Screening of small molecules is often an involved task in early stages of the drug discovery process. As new technologies have enabled the rapid synthesis and high-throughput screening of large libraries of compounds, the number of available structures to screen for novel potential drugs against new targets or pathways has largely increased. However, despite technical advances, the cost of synthesizing and testing the biological properties of every possible molecule of potential interest makes this huge experimental work infeasible nowadays. In consequence, much effort has been directed toward the development of complementary tools for the management of this vast amount of molecules, especially to solve the problem of how to select a small group of them with a high ratio of active compounds. Among these tools, the development of algorithms that generate numerical information reflecting important molecular properties has been an active research field in computational drug discovery. The main goal is to obtain mathematical representations relating molecular structure to its biological or physicochemical properties.[1,3]

Chemoinformatics is an emerging branch of theoretical chemistry that involves the application of informatics methods to solve chemistry problems.[4] At variance with quantum chemistry, where molecules are considered as an ensemble of nuclei and electrons, or molecular mechanics, based on classical molecular models (atoms and bonds), under the chemoinformatic point of view molecules are represented as objects in a chemical space defined by a convenient set of molecular descriptors.[1] After defining a suitable measure in this chemical space, standard optimization and search strategies can be used to find an optimal solution to a given problem. Successful measures based on ligand similarity, such as Euclidean distance, Manhattan distance, or Tanimoto similarity, are widely used to quantify the similarity degree between molecules. These measures can be used to search for compounds similar to a given representative structure by comparing the molecular descriptors that should contain important molecular attributes. Many applications, such as diversity analysis, have been found to obtain enriched subsets in virtual screening or for pharmacological profiling,[2,3,5] among others.

Approaches based on atom-pair distributions are considered to be an attractive family of descriptors.[6] They have been broadly used in different fields of the drug discovery process, covering QSAR, compound selection, virtual screening, and pharmacological profiling.[7] A frequent approach is to store the occurrence of each feature pair for different distances (either topological or geometrical distances), with the aim to capture the global feature distribution.[3]

Despite recent advances in the use of three-dimensional (3D) descriptors, most current similarity searches are based on topological features[8,9] because of the great efficiency shown in most of the reported cases.[10,11] Ligand flexibility is implicitly included when using topological features. Then, descriptors have no conformational dependency and are well-suited for scaffold hopping studies when seeking for a diverse structure set. In this line, descriptors based on topological feature distribution were successfully developed by Gregori-Puigjane et al. leading to promising results for virtual screening or pharmacological profiling of ligands and targets.[3,5]

In this work we propose new molecular descriptors based on the generalized entropy of Rényi. Good results reported using the concept of Shannon entropy to describe a feature-pair distribution[3] reveal this approach as a suitable molecular description and suggest that a more general representation of this concept could generate new descriptors with improved efficiency. For biochemical applications, Rényi entropy has been used as a metric for DNA profiling and several works have been published using the Rényi entropy for analyzing

* Corresponding author phone: 34-93-4039263; fax: 34-93-4021231; e-mail: jaime.rubio@ub.edu.

[†] UB and IQTCUB.

[‡] CSIC-UIB.

[§] UPC.

10.1021/ci900275w CCC: $40.75
Published on Web 10/21/2009

the DNA sequence or DNA binding site.[12-15] These new descriptors, referred to as RED (Rényi entropy descriptors), are introduced in the first methodological section. Results and applications are presented in the following sections.

METHODS

Structure Generation. To analyze the usefulness of RED descriptors to discern between different groups of molecules and the influence of the free parameter involved in the definition of the Rényi entropy, three groups of diverse molecules with reported activity against different targets were analyzed: five cyclooxygenase-1 (COX-1) inhibitors (C), five thrombin (Factor IIa) inhibitors (T), and five estrogen receptor α (ERα) antagonists (E). The molecular structures of these three sets are depicted in Figure 1. Molecules were drawn using the Gaussian Graphical Interface[16] and then optimized using the AM1 Hamiltonian[17] as implemented in the Gaussian 03 program.[18] The resulting structures were converted to mol2 format using Babel 1.6[19] for a proper assignment of Sybyl atom types due to discrepancies with other molecule convertors.

Three small molecule databases were prepared for the virtual screening assays. The first database screened is the initial collection of the Prestwick Chemical Library composed by 880 small molecules (90% being marketed drugs and 10% bioactive alkaloids or related substances) selected for their high chemical and pharmacological diversity as well as for their known bioavailability and safety in humans.[20] The second analyzed library is the MyriaScreen Diversity Collection that is publicly available on the Sigma-Aldrich Web site.[21] It results from the evaluation, filtering, and refinement of selections from each of the Sigma-Aldrich screening compound collections. The MyriaScreen Diversity Collection is comprised of 10 000 high-purity screening compounds handpicked to maximize chemical diversity while maintaining drug-likeness. The last collection was previously used by Sutherland, O'Brien, and Weaver (to be referred to as SOW) to validate a wide variety of QSAR methods.[22] It consists of a collection of eight sets of reported inhibitors for different targets: 114 angiotensin converting enzyme (ACE) inhibitors, 111 acetylcholinesterase (AchE) inhibitors, 147 benzodiazepine receptor (BZR) ligands, 282 cyclooxygenase-2 (COX2) inhibitors, 361 dihydrofolate reductase (DHFR) inhibitors, 66 glycogen phosphorylase B (GPB) inhibitors, 76 thermolysin (THERM) inhibitors, and 88 thrombin (THR) inhibitors. Structures and references to the original works were published by Melville and Hirst.[23]

As the Prestwick and MyriaScreen databases were initially SDF files, CORINA[24] software was used to convert two-dimensional databases into 3D structures. Then, databases were cleaned by removing the small fragments and counterions and the most probable protonation state was determined for each molecule. The SOW database was obtained from the supporting information of ref 23. For all three databases, net charges were assigned to the resulting structures by our in-house program Gen_Conf and then optimized using the AM1 Hamiltonian[17] as implemented in the Gaussian 03 program.[18] The optimized structures were converted to mol2 format using Babel 1.6.[19]

In order to perform a virtual screening assay, 24 of the 39 α_{1A}-adrenoreceptors antagonists reported in the supporting information of ref 3 were taken as the reference set and the remaining 15 molecules were used as a test set. The 3D structure in mol2 format for these compounds was obtained as described before for the three groups of diverse molecules. Test molecules were finally added to all databases to validate the new methodology proposed.

Descriptor Calculations. The process for obtaining RED descriptors is equivalent to that reported by Gregori-Puigjane et al.[3] for the SHED (Shannon entropy descriptors). A structure described in mol2 format is used as the starting point where each atom has been classified into a Sybyl atom type. To avoid possible mistakes in the Sybyl atom type assignment, a suitable molecular 3D structure is needed. Then, each atom type is assigned to one or more of the four atom-centered features: hydrophobic (H), aromatic (R), acceptor (A), and donor (D), according to the supporting information of ref 3. It is important to note that, at this stage, hydrogen atoms are taken into account in order to distinguish atoms with different H-bond properties, for example N.4 and N.4h atoms. We measure all possible path lengths, defined as the number of bonds of the shortest path connecting two non-hydrogen atoms with a given feature pair (a maximum path length of $N = 20$ bonds was used and distances over this limit were accumulated in the last bin). If P_i, $i = 1, ..., N$ is the population of path length i and $P = \sum_{i=1}^{N} P_i$, the total number of path lengths, a probability distribution $\rho_i = P_i/P$ can be derived directly from the molecular topology for each one of the 10 possible atom-centered feature pairs.

Then, Rényi entropy can be obtained as[12]

$$H_\alpha = \frac{1}{1-\alpha} \ln\left(\sum_{i=1}^{N} \rho_i^{\alpha}\right)$$

where $\alpha \neq 1$ is a real positive number, defined as the order of the Rényi entropy. In the limiting case $\alpha \to 1$, Rényi entropy converges to Shannon entropy given by

$$H_{\alpha=1} = -\sum_{i=1}^{N} \rho_i \ln(\rho_i)$$

For a uniform distribution of equally probable values (maximum disorder), we have $\rho_i = 1/N$. Then, Rényi entropy reaches its maximum value at $H_\alpha = \ln(N)$. For a single value distribution (minimum disorder) with $\rho_i = 1$ and $\rho_j = 0$ for all $j \neq i$, Rényi entropy takes its minimum value at $H_\alpha = 0$. To have a more intuitive measure that scales linearly with the maximum entropy value, projected entropies are calculated as $E_\alpha = e^{H_\alpha}$. Thus, for a given population, projected entropy values vary from $E_\alpha = 1$ (single value distribution) to $E_\alpha = N$ (uniform distribution). The set of E_α values calculated for all feature-pair distributions constitutes the 10-dimensional RED descriptor vector.

RESULTS AND DISCUSSION

The identification of homogeneous subgroups from a collection of heterogeneous objects is one of the most common tasks in molecular modeling. This goal is easier to achieve when the attributes or descriptors used for profiling generate pair distances that clearly group the compared objects. As SHED descriptors have proven efficient in accomplishing this task, we first investigated the role of the

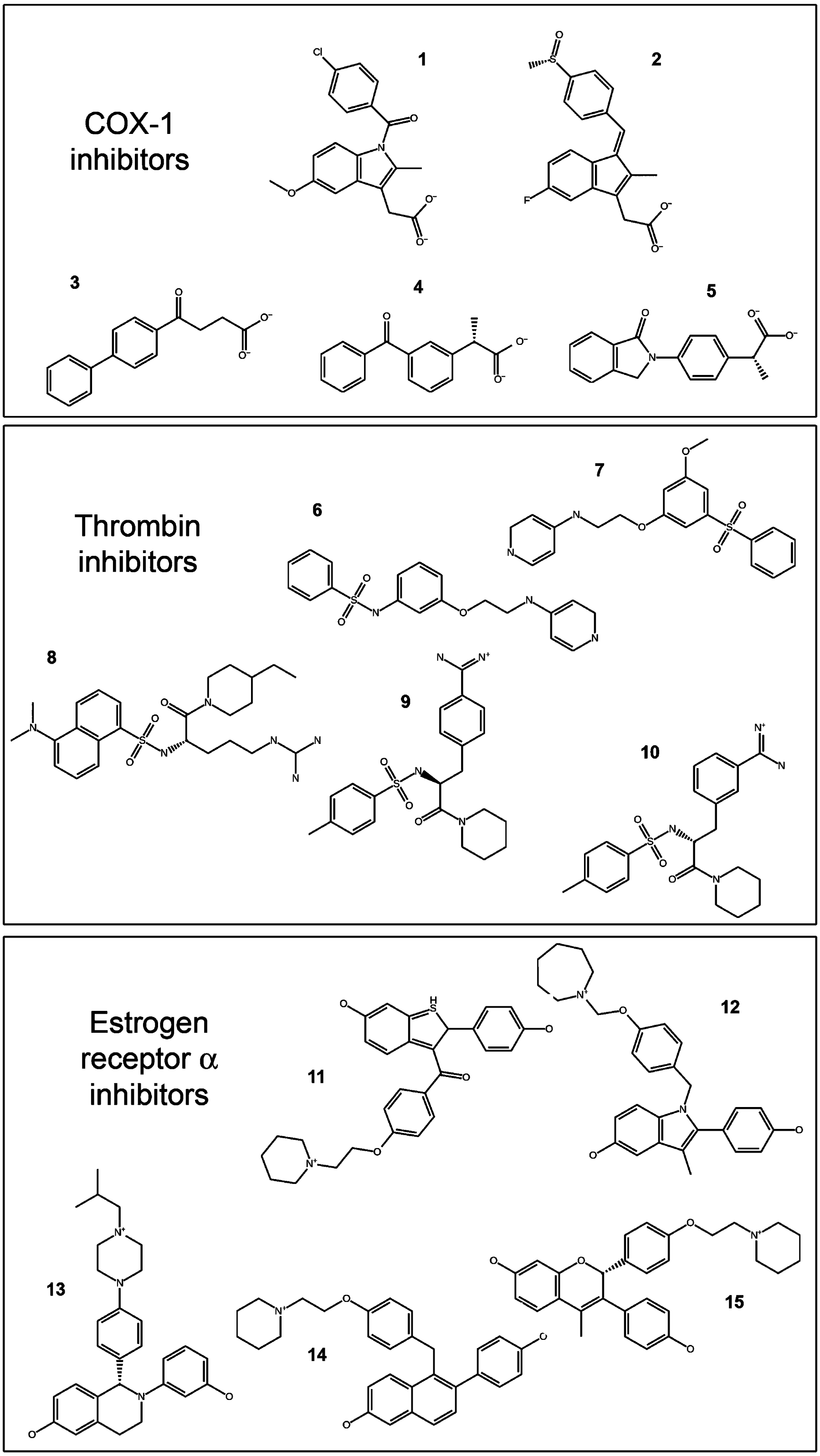


Figure 1. Set of three groups of diverse molecules with reported activity against different targets: five cyclooxygenase-1 (COX-1) inhibitors (C), five thrombin (Factor IIa) inhibitors (T), and five estrogen receptor α (ERα) antagonists (E).

Rényi parameter α included in the RED descriptors for discerning between three different molecular sets. Once we determined the optimum α value, virtual screening and pharmacological profiling assays were analyzed to validate the chosen optimum parameter.

Distance Analysis vs α Rényi Parameter. When a suitable description is used, molecular-pair distances must reflect the different molecular groups composing the analyzed collection. To validate RED descriptors, a test set that includes three groups of diverse molecules (see Figure 1) with reported activity against different targets was used: five cyclooxygenase-1 (COX-1) inhibitors (C), five thrombin (Factor IIa) inhibitors (T), and five estrogen receptor α (ERα) antagonists (E).

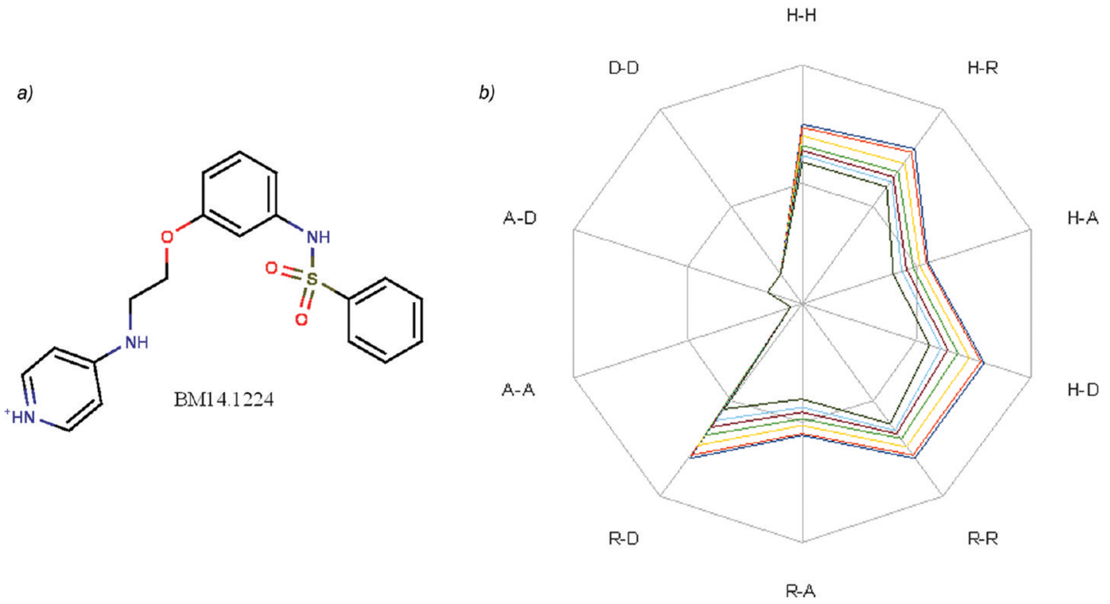


Figure 2. (a) Structure of BM14.1224 compound, a thrombin inhibitor. (b) RED profile obtained for BM14.1224 compound using different values for α: 0.01 (dark blue), 0.1 (red), 0.5 (yellow), 1.0 (green), 1.5 (purple), 2 (light blue), and 3 (dark green).

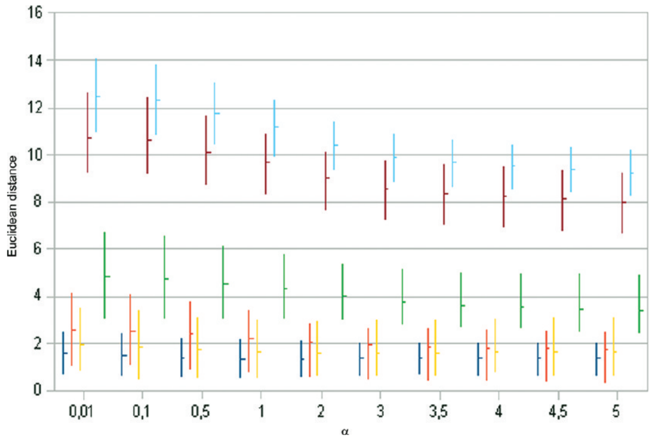


Figure 3. RED intragroup distances (C−C distance in dark blue, T−T distance in red, E−E distance in yellow) and intergroup distances (C−T distances in purple, C−E distances in light blue, T−E distance in green) for different values of α. Calculated mean values and root-mean-square deviation (rmsd) within molecules belonging to each group are represented.

RED descriptors associated with the test set molecules were calculated for different values of the α parameter of the Rényi entropy. As shown in Figure 2, for a given molecule belonging to the group of thrombin inhibitors, the α value does not qualitatively change the descriptor profile. However, from a quantitative point of view there is a contraction or expansion of the descriptor values with respect to $\alpha = 1$. This is simply explained by the monotonicity property of the Rényi entropy: if $\alpha_1 \leq \alpha_2$, then $H_{\alpha_1} \geq H_{\alpha_2}$. Moreover, despite the scaffold diversity present in the different groups of molecules, an equivalent descriptor profile was found for molecules within the same family.

To quantitatively compare the descriptor profiles obtained for the different families, Euclidean distances were calculated between the descriptor vectors corresponding to each molecule. Mean values and standard deviations were calculated for each inhibitor set as well as for intergroup distances. Distance values are presented in Figure 3 for different values of α.

Table 1. Maximum (dmax), Minimum (dmin), Average (davg), and Root-Mean-Square Deviation (rmsd) of the Intergroup and Intragroup Euclidean Distances Associated with SHED (α = 1) and RED (α = 2) Descriptors

	dmax	dmin	davg	rmsd
		SHED (α = 1)		
C−C	2.14	0.57	1.33	0.55
T−T	*3.41*	0.77	2.23	0.75
E−E	2.98	0.57	1.63	0.71
C−T	10.89	8.31	9.66	0.91
C−E	12.35	9.95	11.17	0.71
T−E	5.79	*3.08*	4.54	0.87
		RED (α = 2)		
C−C	2.09	0.61	1.34	0.58
T−T	2.81	0.60	2.03	0.66
E−E	*2.93*	0.65	1.57	0.70
C−T	10.08	7.65	9.00	0.90
C−E	11.46	9.30	10.41	0.64
T−E	5.37	*3.02*	4.02	0.69

As shown, by using mean values RED descriptors can discern between intragroup and intergroup distances for all the α values analyzed. However, an atypically low intergroup mean distance was observed for thrombin (Factor IIa) inhibitors and estrogen receptor α (ERα) antagonists due to similar structural features shared for the two types of inhibitors.

As the value of α increases, the Rényi entropy decreases, tending to $H_\infty = -\ln(\sup_{i=1,\ldots,N} \rho_i)$ for $\alpha \to \infty$ and causing a global reduction of distances. This value is denoted as min-entropy. From Figure 3, the limit for achieving distance convergence can be considered reached for $\alpha = 2$. For larger α, distance variations were negligible. It is observed that, for this α value, intragroup distances have converged to values that allow discerning between different molecular clusters. However, for larger α values a small overlap appears between T−E intergroup distances and E−E intragroup distances.

Despite the fact that different distance ranges are observed for intergroup and intragroup distances, values of $\alpha < 2$, including SHED descriptors ($\alpha = 1$), lead to an overlap

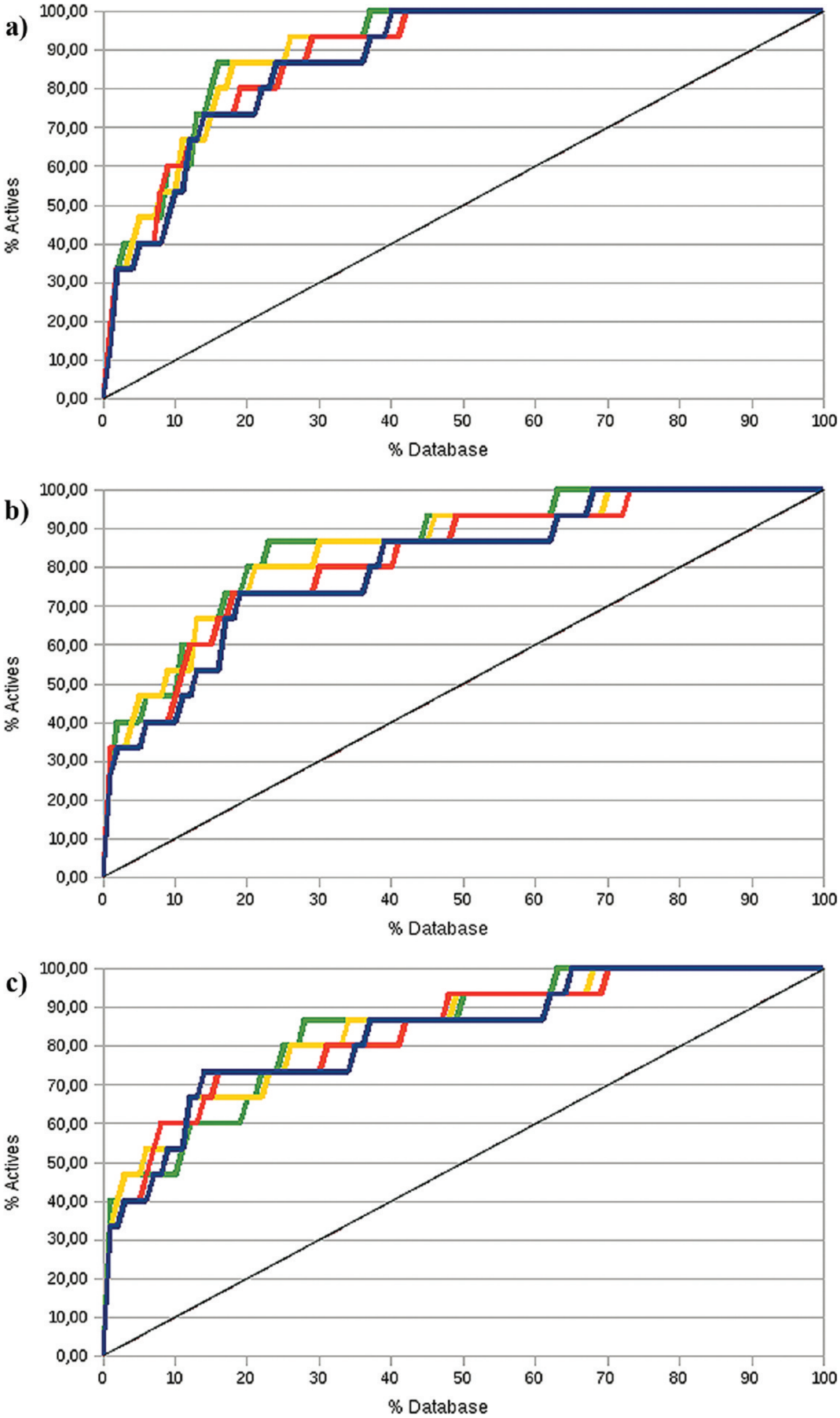


Figure 4. ROC curves for (a) Prestick database, (b) MyriaScreen Diversity Collection, and (c) SOW database. Different descriptors are represented for each database: RED (0.5) in blue, SHED in orange, RED (2.0) in yellow, and RED (3.0) in green.

between both types of distances. The minimum intergroup distance (T−E distance) is larger than the maximum intragroup distance (T−T distance for SHED descriptors and E−E distance for RED descriptors), hindering the establishment of a limit value that separates between intergroup and intragroup distances (italicized values shown in Table 1).

Taking into account the discussed tendencies, $\alpha = 2$ has been selected as the optimal value of the parameter of the Rényi entropy for discerning between different sets of molecules. It has a low dispersion on distance values having nearly reached the convergence. Moreover, it makes possible the determination of a limit distance between intergroup and intragroup distances.

It is noticeable that the Rényi entropy for the selected optimal value $\alpha = 2$, referred to as Rényi's quadratic entropy,[25] allows an important computational simplification that makes its calculation easier.[14,26,27]

Virtual Chemical Screening. In order to perform virtual screening assays, the performance of RED descriptors was studied, including the influence of the α parameter. A set of 24 α_{1A}-adrenoreceptor antagonists with known activity (K_i < 300 nM) was considered for calculating the reference

Table 2. For Different Values of α, Percentage of Actives Recovered for 5% of the Different Databases Screened[a]

α	Prestwick	MyriaScreen	SOW
0.01	40 (6)	33 (5)	40 (6)
0.10	40 (6)	33 (5)	40 (6)
0.50	33 (5)	33 (5)	40 (6)
1.00	33 (5)	33 (5)	40 (6)
1.50	33 (5)	33 (5)	47 (7)
2.00	47 (7)	47 (7)	47 (7)
2.50	40 (6)	47 (7)	47 (7)
3.00	40 (6)	40 (6)	47 (7)
4.00	47 (7)	40 (6)	47 (7)

[a] The number of test compounds retrieved are in parentheses.

descriptors. Three databases (Prestwick, MyriaScreen, and SOW) were screened for a test set constituted by 15 α_{1A}-adrenoreceptor antagonists that were previously added to each database.

The descriptors were calculated for the reference set and the three databases using different α values. Euclidean distances between the databases and the reference structures were used as a dissimilarity metric. The minimum distance between a given molecule and some of the reference structures was considered as the score function for the virtual screening. For each database, all molecules were rearranged according to the score function.

The recovery of known actives was depicted in the ROC (receive operating characteristic) curves using different α values (see Figure 4). Generally, for larger α values a better actives recovery is obtained, reaching a convergence again for $\alpha > 2$. Using $\alpha = 2$, it can be observed (see Table 2) that for just 5% of a database screened, RED descriptors lead to 47% actives recovery independently of the database size. However, by using the SHED profile 33% or 40% actives recovery is obtained, a result which is now database dependent.

In practice, 5% of a database still represents a huge number of compounds (around 100 in the case of the Prestwick and SOW databases, and around 500 for the MyriaScreen database) which would not be practical to acquire for biological evaluation. Hence, only the 10 first compounds found in each database, according to the score function, were analyzed using the RED profile for $\alpha = 2$.

Table 3. Ratio between Active and Total Molecules under a Cutoff Distance Corresponding to the Seventh Test Molecule Minimum Distance to the Reference Set for Different α Values

α	Prestwick	MyriaScreen	SOW	cutoff distance
0.5	7/62	7/1226	7/169	1.84
1.0	7/59	7/1212	7/164	1.79
2.0	7/33	7/986	7/108	1.54
3.0	7/32	7/1201	7/117	1.50

The highest hit rate corresponds to the Prestwick database with six actives retrieved within the top 10 molecules, four of them with 0 distance to some reference molecules. When these molecules were analyzed, topologically identical structures were found with only stereochemical differences from the nearest reference molecule. Thus, in order to obtain comparable results within all databases, these molecules were removed for the remaining analysis.

For smaller databases, Prestwick and SOW, four and five actives were respectively found within the top 10 positions. However, for MyriaScreen only one active molecule was present since the next test molecule appeared at the 47th position. We looked for activity reports related to the potential hits found on the top 10 positions. For referenced compounds, activities described were all related to analgesic or sedative properties, typically associated with α_{1A}-adrenoreceptor antagonism.

Independently of the descriptors used, the last test molecule retrieved was observed to appear around 50% of database screened (see ROC curves in Figure 4). This means the existence of a large amount of molecules with lower distances to the reference set that should be considered as potential actives. In order to determine the distance needed for recovering the whole test set, which can be considered as the cutoff distance for potential actives, distances between the test and reference set were analyzed for different values of α. Results (shown in Figure 5) reflect that some test molecules are really far from the reference set making difficult its retrieval, independently of the value of α used.

It was commented before that a global distance reduction can be observed when increasing α, especially for large distances. This fact can be translated into a lower distance to retrieve all molecules in the test set. As

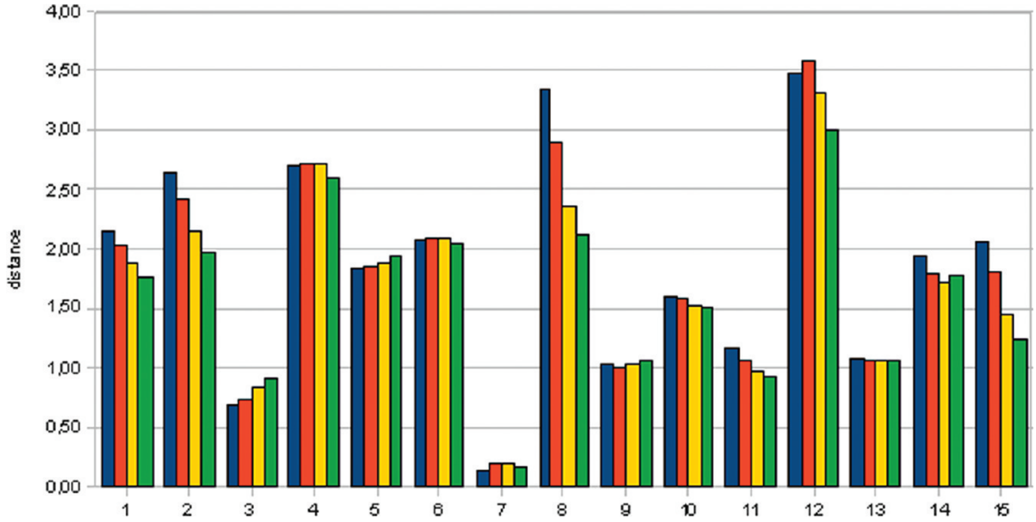
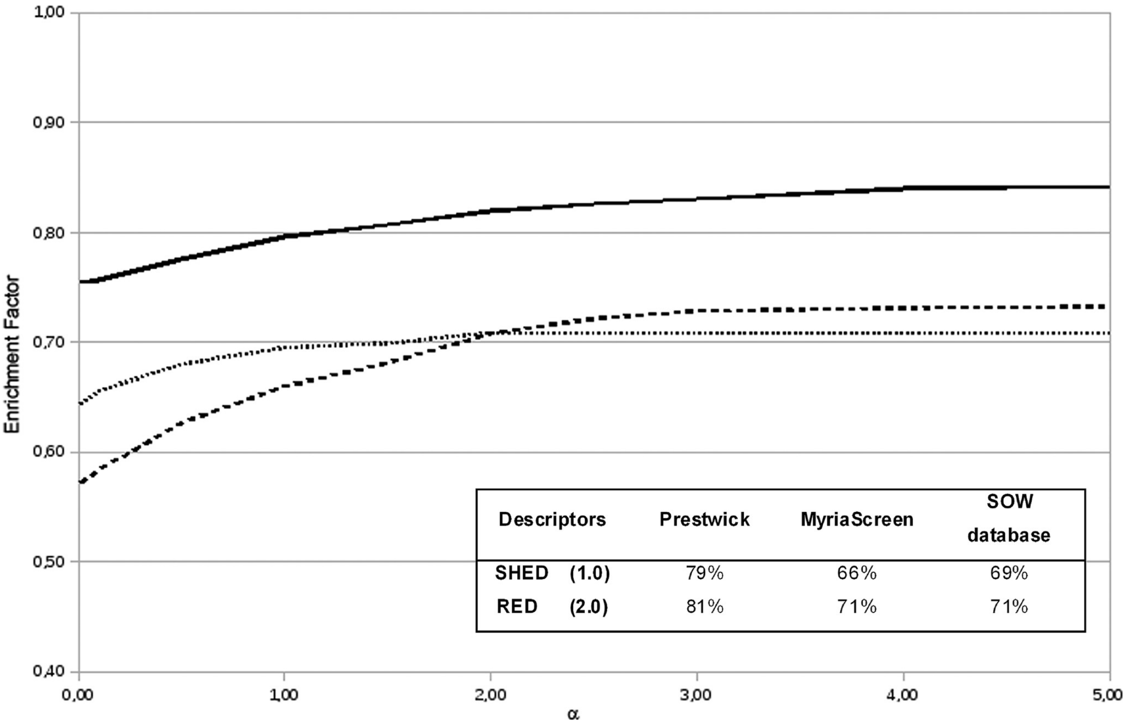


Figure 5. Smallest RED distances between molecules of the test and reference sets. Dark blue values correspond to $\alpha = 0.5$, orange values correspond to SHED results ($\alpha = 1$), yellow values correspond to $\alpha = 2$, and green values correspond to $\alpha = 3$.



Descriptors	Prestwick	MyriaScreen	SOW database
SHED (1.0)	79%	66%	69%
RED (2.0)	81%	71%	71%

Figure 6. Enrichment factor achieved (Ef) with RED descriptors and different α values for the three analyzed databases: Prestwick, solid line; MyriaScreen, dotted line; SOW, dashed line. The enrichment factor percentages corresponding to SHED and RED (α = 2) descriptors are shown in the inset table.

previously discussed, around 50% of database molecules should be considered as potential actives when the cutoff is established as the distance to the further test molecule. Then, a lower limit distance would be useful to reduce the number of candidate molecules. Using the distance needed to retrieve 50% of the test set as cutoff, the ratio of active versus total molecules was calculated for different α values (results shown in Table 3). The optimum value of α considerably increases the ratio of active compounds, mainly by the use of a lower cutoff distance, and fewer molecules should be acquired when looking for a given number of actives.

In order to analyze the overall enrichment, we use the enrichment factor (Ef), defined as

$$\mathrm{Ef} = \frac{\mathrm{AUC} - \mathrm{AUC_R}}{\mathrm{AUC_I} - \mathrm{AUC_R}}$$

where AUC, $\mathrm{AUC_R}$, and $\mathrm{AUC_I}$ stand for the normalized area under the curve, $\mathrm{AUC_R}$ corresponds to the case of random identification of actives (area under the diagonal line, $\mathrm{AUC_R} = 0.5$) and $\mathrm{AUC_I}$ corresponds to the ideal situation where n actives are located at the n first positions (for $n = 15$, $\mathrm{AUC_I} = 0.9975$), respectively. The enrichment factor, Ef, was calculated for different values of α as depicted in Figure 6. A value of Ef = 1.0 indicates the ideal situation where AUC = $\mathrm{AUC_I}$. For all databases the enrichment factor increases with α, reaching a stable value for α > 2, which confirms its utility as the optimal parameter.

When comparing enrichment factors obtained with RED (α = 2) and SHED (α = 1) descriptors, all databases show an improvement for larger α values. However, this improvement is database size dependent since it is more remarkable for the MyriaScreen database than for the remaining two databases. Thus, for larger databases better performance can be expected.

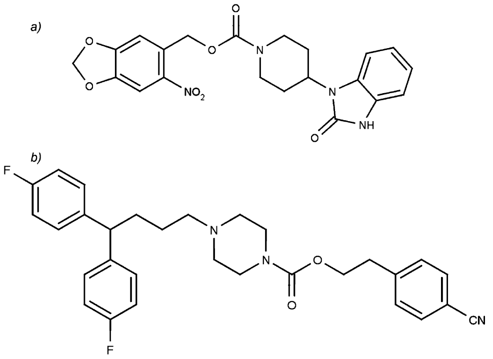


Figure 7. Molecular structures of the two reference molecules: (a) molecule 1; (b) molecule 2.

Virtual Pharmacological Profiling. Similar compounds are likely to present similar activity against a set of targets. Hence, it is possible to assign a pharmacological profile using Euclidean distance between RED descriptors as a similarity measure to compare structural features between different molecules.

To analyze theoretical and experimental descriptions of pharmacological profiles, a previously published set of 47 compounds with known activity against 38 targets was selected.[28] Two compounds, having similar activities over the μ opiate receptor but completely different activities over the other opiate receptors considered (molecule 1 and molecule 2, whose structures are shown in Figure 7), were used as a starting point to generate a set of compounds having different pharmacological profiles. Reported activity data were expressed as $\mathrm{IC_{50}}$ values ranging from 9380 to 0.9 nM. First, activity data were transformed by using $\Delta G = -RT \ln(\mathrm{IC_{50}})$ to express the activity in kcal/mol units. Thus, for T = 298 K one unit expressed in kcal/

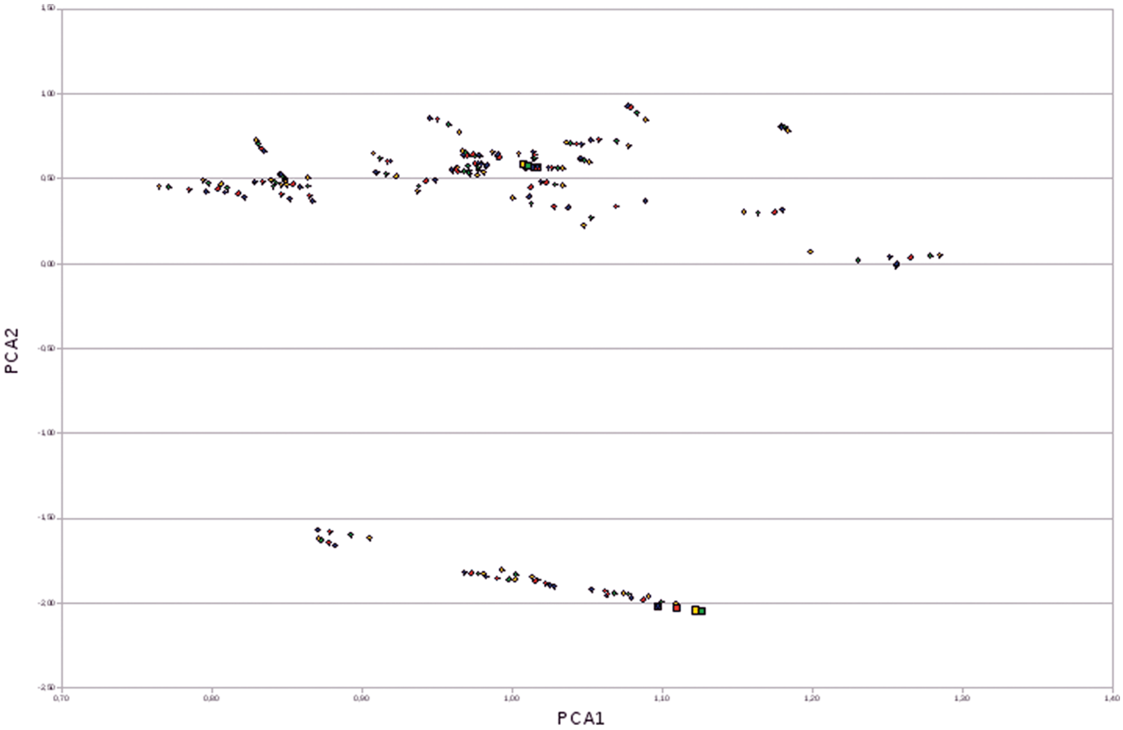


Figure 8. Representation of the analyzed molecules using the two first principal components from the PCA. As expected, two well-separated clusters of molecules appear corresponding to molecule 1 and molecule 2 derivatives. Results for different α values are shown: α = 0.5 in blue; α = 1.0 in red; α = 2.0 in green; and α = 3.0 in yellow. Big squares represent the coordinates for the two reference molecules: molecule 1 (PCA2 > 0) and molecule 2 (PCA2 < 0).

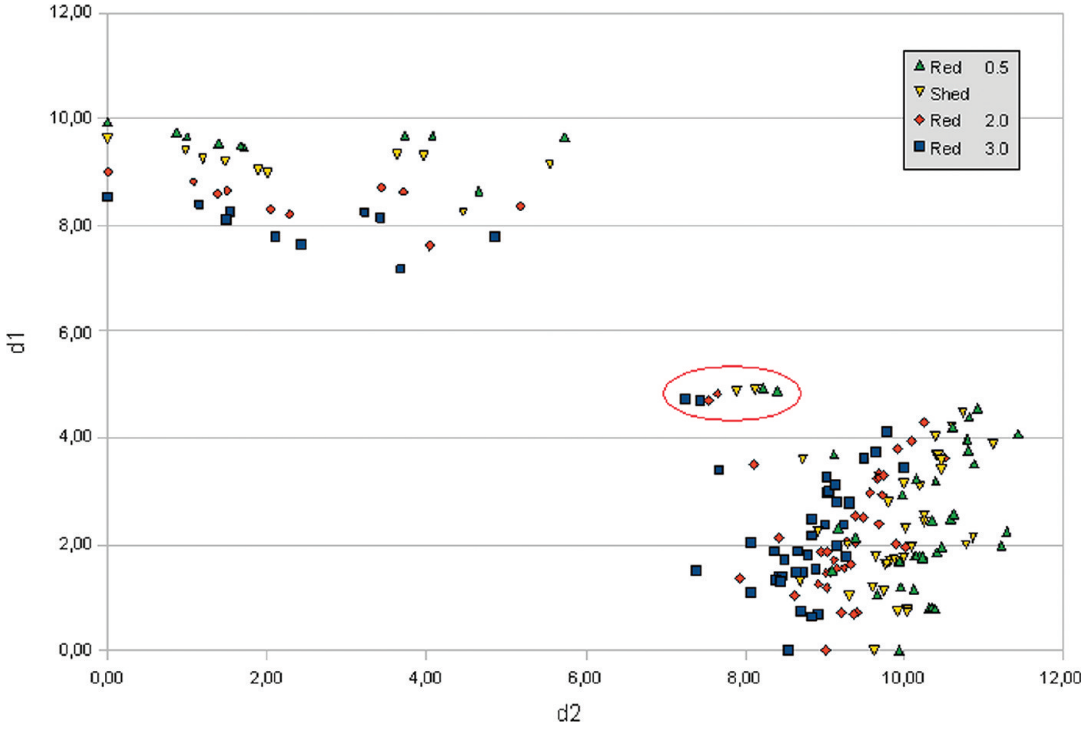


Figure 9. Molecule 1 derivative (lower right) and molecule 2 derivative (upper left) clusters obtained for different descriptor profiles (α values shown in the legend). Each molecule is represented by a spatial point (d2,d1), where d2 and d1 are the distances to the two reference molecules (molecule 2 and molecule 1, respectively). The points inside the circle correspond to molecules 44 and 46, which appear on the wrong cluster.

mol corresponds approximately to 1 order of magnitude in the IC_{50} value. The newly transformed data range from 12.33 to 6.85 kcal/mol.

A modification of the previously proposed activity dissimilarity[28] was used to score the pharmacological profiles of two molecules. Experimental dissimilarity between compounds A and B was estimated as $D(\mathrm{A,B}) = \sum_{i=1}^{\mathrm{targets}} \Delta_i[|\Delta G_\mathrm{A} - \Delta G_\mathrm{B}|]$, with $\Delta_i[|\Delta G_\mathrm{A} - \Delta G_\mathrm{B}|] = \Delta_i[\Delta_\mathrm{AB}]$ and $\Delta_\mathrm{AB} = |\Delta G_\mathrm{A} - \Delta G_\mathrm{B}|$. Here $\Delta_i[\Delta_\mathrm{AB}]$ represents an empirical measure of how different two compounds are with respect to the target i:

$$\Delta_i[\Delta_{AB}] = \begin{cases} 1 & \text{if A active and B inactive} \\ 1 & \text{if B active and A inactive} \\ 0 & \text{if A and B inactives} \\ 1 & \text{if } \Delta_{AB} > \Delta_{AB,\max} \\ \dfrac{\Delta_{AB} - \Delta_{AB,\min}}{\Delta_{AB,\max} - \Delta_{AB,\min}} & \text{if } \Delta_{AB} > \Delta_{AB,\min} \end{cases}$$

The value $\Delta_{AB,max}$ represents the maximum difference for which two compounds are considered to behave differently with respect to a given target and $\Delta_{AB,min}$ corresponds to the minimum value for which compounds A and B are considered to have similar activities against one target. Values of $\Delta_{AB,max} = 3$ and $\Delta_{AB,min} = 1$ (in kcal/mol) were used in this work to compute experimental dissimilarities. However, small changes of these values were also tested (results not shown) and do not affect the global shape of the generated pharmacological profile.

The correlation between experimental and theoretical distances was analyzed for different values of α, by depicting theoretical distance versus experimental distance for all molecules (data not shown). A global correlation was not clearly present, pointing out the difficulty of describing a complete pharmacological profile for a very diverse set of molecules. However, values of theoretical distances follow a bimodal distribution, one for molecule 1 and a second for molecule 2. Also, the clusters appeared more compact when the α value was increased.

To show the capability of RED descriptors to discern between molecules belonging to each group, a principal components analysis (PCA) was performed. Also the influence of the α parameter was analyzed. Figure 8 shows the distribution of molecules using the two major components as coordinates. As expected, two clusters of molecules were observed for all α values considered.

In order to be able to directly compare molecular distributions around molecules 1 and 2 from theoretical and experimental points of view with a more intuitive representation, relative distances to the two reference molecules were calculated. Then, instead of the principal component coordinates, each molecule was represented by a spatial point composed by the coordinates (d2,d1), where d2 (respectively d1) is the distance to the initial molecule 2 (respectively molecule 1).

When observing the chart (see Figure 9), we can now clearly recognize two clusters of molecules: molecule 1 derivatives (lower-right cluster) which appear at low d1 distances and molecule 2 derivatives (upper-left cluster) appearing at low d2 distances. Nevertheless, molecules 44 and 46 (structures shown in Figure 10) follow a different pattern. Although they are obtained as molecule 2 derivatives and share almost equivalent structural features, they appear proximal to the molecule 1 derivative cluster. This behavior can be explained by differences in the hydrogen donor profile of these compounds: at variance from molecule 2, compounds 1, 44, and 46 have one hydrogen donor atom on their structures. Donor-feature distribution is included in four of the 10 descriptor values and when comparing to molecule 2, large distances are found between the descriptor vectors.

The effect of α was also analyzed, observing that the global shape was conserved within all the descriptors used. However, as noted before, an increase of α produces a clusterization of the different molecular groups.

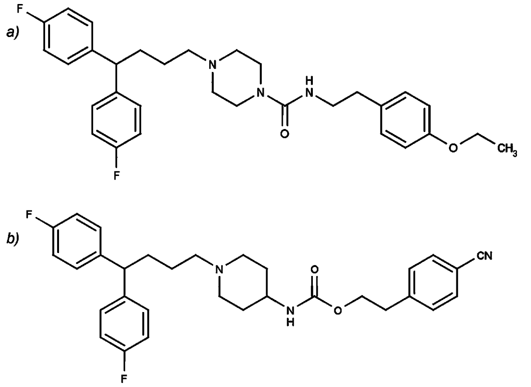


Figure 10. Molecular structures of (a) molecule 44 and (b) molecule 46.

The same analysis was performed for calculated experimental distances. As shown in Figure 11, the experimental metrics used also allow discerning the two groups of molecules. Now, molecules 41, 42, 43, 47, and 48 (obtained as molecule 2 derivatives) appear in the molecule 1 derivative cluster. However, according to their pharmacological activity, these molecules belong to the right cluster because they should be considered as molecule 1 like compounds due to their selectivity against the μ opiate receptor. Thus, the experimental description used properly reflects the pharmacological profile for the whole analyzed set.

It seems clear that RED descriptors allow grouping structurally similar molecules. Therefore, in order to assign a pharmacological profile to a given molecule, it is necessary to correlate this structural similarity with the biological activity. For this purpose theoretical profiles obtained with RED descriptors (with α = 2) and the activity displayed against a set of targets were compared for molecule 28 (structure shown in Figure 12) with molecule 1 and molecule 2 as reference molecules.

The considered structure, molecule 28, is a molecule 1 derivative. As expected, the pharmacological profiles for molecule 28 and molecule 1 are very similar and with remarkable differences from the one obtained for molecule 2 (shown in Figure 13). Therefore, it is expected that the descriptor profiles obtained for those molecules will be able to highlight the mentioned differences.

Figure 14 depicts the RED profile (α = 2) for molecule 28 compared with the two reference molecules. Large differences can be observed when comparing the profiles for the analyzed molecule and molecule 2, while comparison with molecule 1 leads to a completely overlapped profile. At right in the figure, the table shows theoretical distances calculated for different values of α. For all cases considered, those distances reflect the similarities and differences between molecule 28 and the two reference molecules.

In order to compare theoretical and experimental profiles, as well as to characterize the influence of α on the correlation between the experimental activities and the descriptors obtained, the distribution of experimental distances was depicted for different α values for all molecule pairs with a theoretical distance between 0 and 2 (data shown in Figure 15). As should be expected, the experimental distance distribution shows a maximum at short distances, which means that molecules with low theoretical distances also have similar activity profiles. When α is increased, all distances concentrate around this maximum, converging for α = 2.

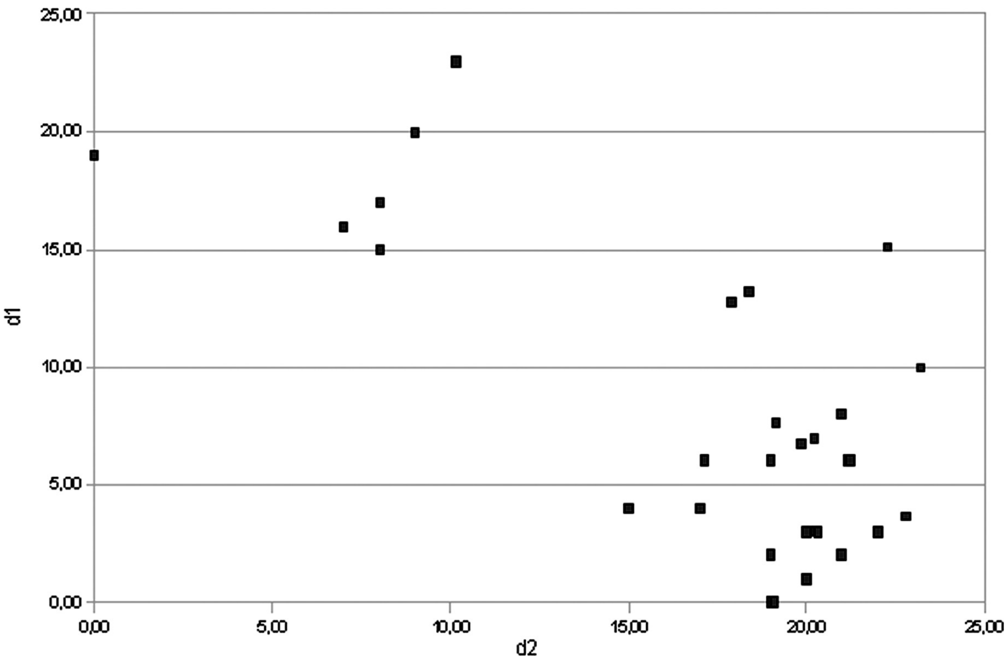


Figure 11. Clusters corresponding to molecule 1 derivatives (lower right) and molecule 2 derivatives (upper left) obtained by representing each compound as an spatial point (d2,d1), where d2 and d1 are experimental calculated distances to the reference molecules.

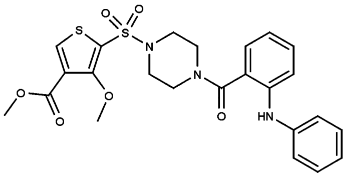


Figure 12. Structure of molecule 28, a molecule 1 derivative.

However, the convergence of distances observed for large α values induces a distortion on distance distribution, with the appearance of a new maximum for large experimental distances which is minimized in the $\alpha = 2$ curve.

Finally, it is worth noting that RED descriptors are very sensitive to the presence or not of one pharmacophoric feature. There are 10 possible distributions in the descriptor vector, and one particular feature is involved in four of them.

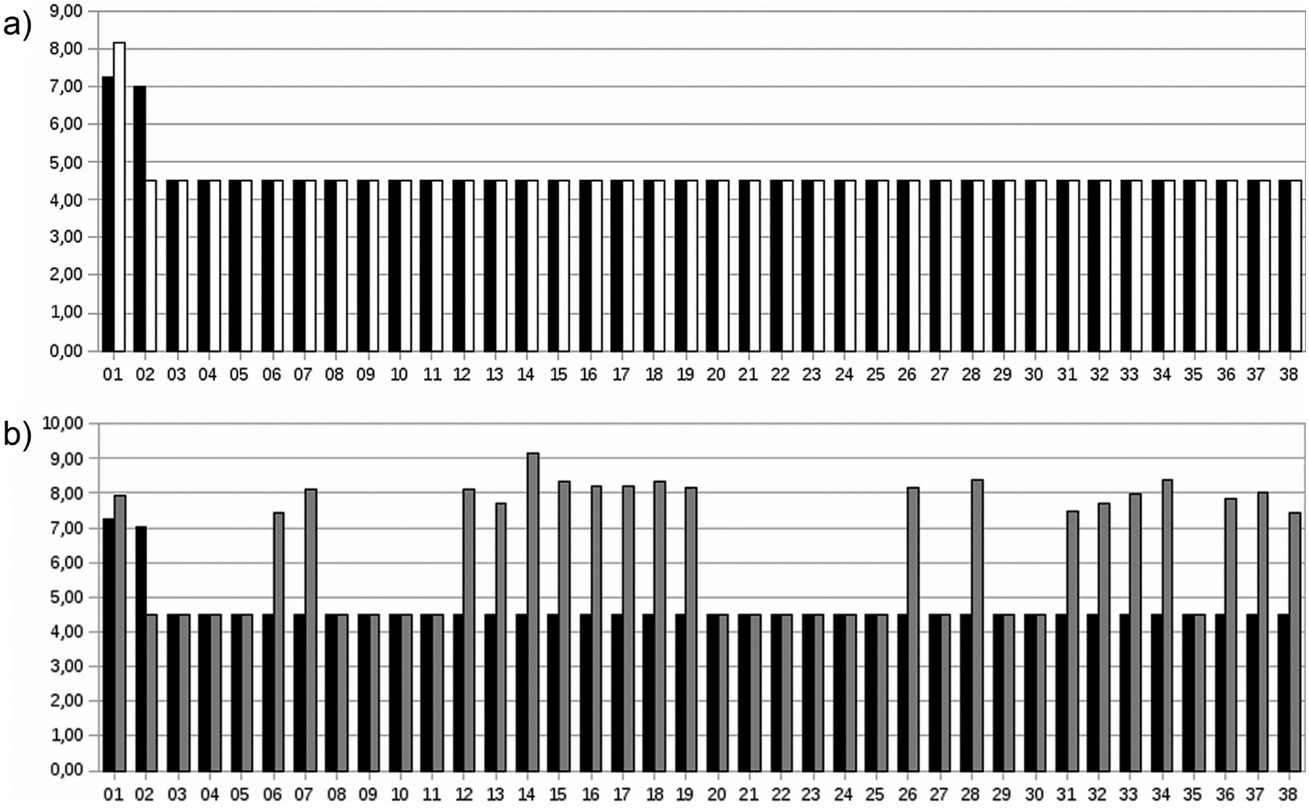
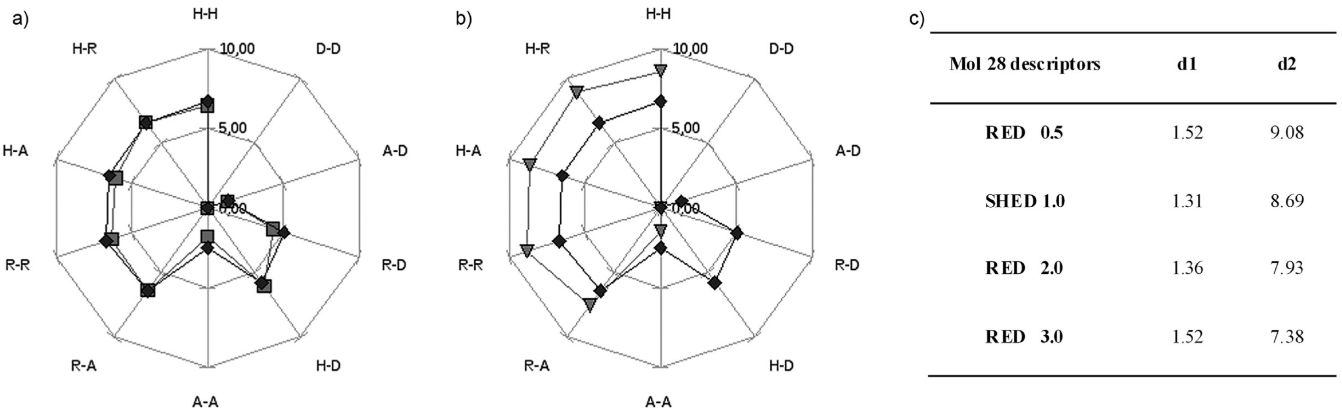


Figure 13. Experimental calculated values of ΔG (on kcal/mol) for molecule 28 (in black) and two reference molecules: (a) molecule 1 (in white); (b) molecule2 (in gray).



Mol 28 descriptors	d1	d2
RED 0.5	1.52	9.08
SHED 1.0	1.31	8.69
RED 2.0	1.36	7.93
RED 3.0	1.52	7.38

Figure 14. RED profiles for molecule 28 compared to reference molecules. (a) RED profile for molecule 28 (triangles) and molecule 1 (squares). (b) Profile for molecule 28 (triangles) and molecule 2 (diamonds). (c) Calculated theoretical distances between the descriptor vectors of molecule 28 and the reference molecules (d1 represents the distance to molecule 1 and d2 the distance to molecule 2) for different α values.

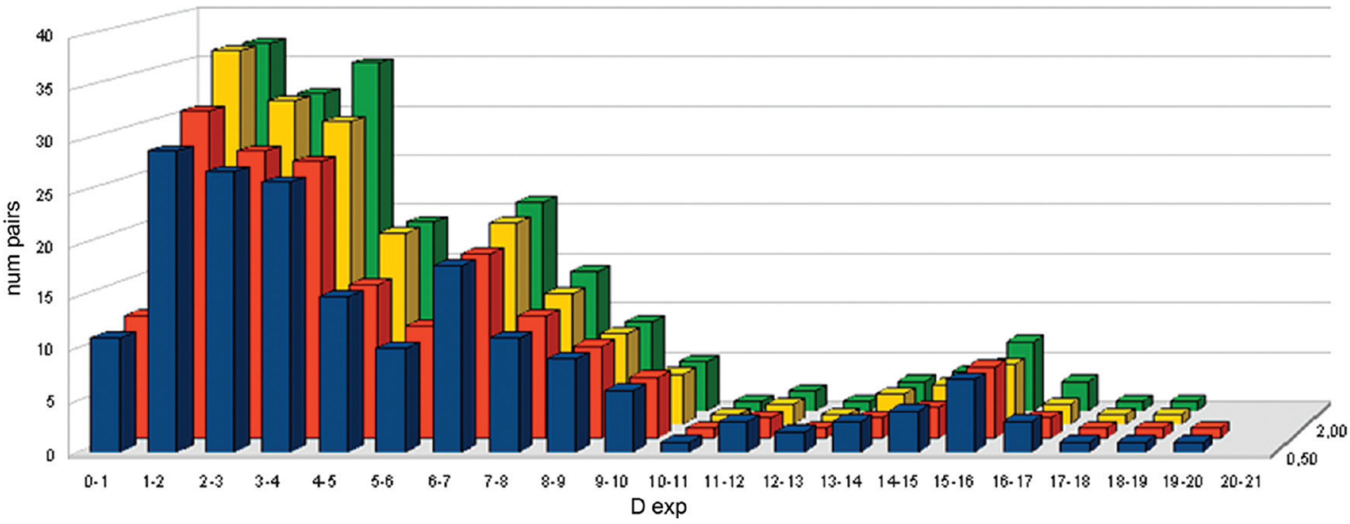


Figure 15. Distribution of experimental distances for molecule pairs with a theoretical distance between 0 and 2. Different values of α are tested: 0.5 (dark blue), 1 (orange), 2 (yellow), and 3 (green).

Then, if one molecule does not have a particular feature, its RED descriptors are very different from those obtained from molecules having this feature. Thus, RED descriptors could be used to study situations where a specific pharmacophoric feature is determinant for biological activity.

CONCLUSIONS

In this work, we have generalized the SHED descriptors reported by Gregori-Puigjane et al.[3] to develop new molecular descriptors based on the Rényi entropy as a variability measure for a feature-pair distribution. As observed for SHED descriptors, molecules with equivalent features rearranged similarly around different scaffolds also lead to equivalent RED profiles independent of the value of α used.

The order of α of the Rényi entropy for RED descriptors was optimized for discerning between different molecule groups. A value of $\alpha = 2$ was selected as the optimum value, allowing the separation of different molecular clusters without exceeding the limitations of the methodology used associated with large α values.

The performance of RED descriptors for *in silico* screening was also analyzed. Compared with SHED descriptors, the optimization of α improves the hit rate for the analyzed databases. The effect of this improvement is especially remarkable for larger databases, reducing the amount of acquired compounds needed to reach a given active retrieval. Moreover, referenced molecules appearing in the top 10 positions were found as potential actives against the considered target.

A set of 47 compounds with known activity against 38 targets was selected to test RED descriptors as a similarity measure for the virtual pharmacological profiling. When comparing a given molecule with two reference molecules with equivalent or totally different activities, RED profiles obtained reflect the pharmacological differences. Correlation between experimental distances and RED distances was also analyzed. Experimental distance distribution was depicted for molecules whose theoretical distance ranged between 0 and 2. A maximum was found for low experimental distances in good agreement with low theoretical distances considered. The optimum value of α concentrates the population around the maximum observed.

In view of the presented results, RED descriptors appear as a suitable representation of molecular structures to be used in similarity-based drug discovery projects. Feature-pair variability represents not only the structural features contained in a given molecule but also its biological activity or binding pattern. Thus, good results were found for discerning different molecule groups as well as for virtual screening or pharmacological profiling. However, as discussed recently,[29,30] the performance of different similarity methods strongly depends on the studied systems, often producing different sets of actives when working with different sets of descriptors. Then only experience can determine the real efficiency of any particular method.

ACKNOWLEDGMENT

We acknowledge financial support from the Spanish Ministry of Science and Technology and FEDER (EU), projects CTQ2006-06588/BQU and FIS2007-60327; the Generalitat de Catalunya, project 2009SGR1308; and the EU NoE BioSim, LSHB-CT-2004-005137. We are also grateful to the "Departament d'Universitats, Recerca i Societat de la informació de la Generalitat de Catalunya i del Fons Social Europeu", and to Dr. Rafael Gozalbes and Prof. Enrique Pérez-Payá (Centro de Investigación Príncipe Felipe, Spain) for making available to us the structures of the compounds used for the pharmacological profile study and the structures of the Prestwick Chemical Library and the MyriaScreen Diversity Collection for the virtual screening process.

Supporting Information Available: AM1 optimized 3D structures for the five COX-1 inhibitors, five Factor IIa inhibitors, five ERα antagonists, and the reference and test molecules used in the virtual chemical screening. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Varnek, A.; Tropsha, A. *Chemoinformatics Approaches to Virtual Screening*; Royal Society of Chemistry: Cambridge, 2008.

(2) Fechner, U.; Schneider, G. Optimization of a Pharmacophore-based Correlation Vector Descriptor for Similarity Searching. *Comb. Sci.* **2004**, *23*, 19–22.

(3) Gregori-Puigjané, E.; Mestres, J. SHED: Shannon Entropy Descriptors from Topological Feature Distribution. *J. Chem. Inf. Model.* **2006**, *46*, 1615–1622.

(4) Gasteiger, J.; Funatsu, K. Chemoinformatics—An Important Scientific Discipline. *J. Comput. Chem.* **2006**, *5* (2), 53–58.

(5) Mestres, J.; Martín-Couce, L.; Gregori-Puigjané, E.; Cases, M.; Boyer, S. Ligand-Based Approach to In Silico Pharmacology: Nuclear Receptor Profiling. *J. Chem. Inf. Model.* **2006**, *46*, 2725–2736.

(6) Carhart, R. E.; Smith, R. E. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

(7) Pérez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixidó, J. APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49* (5), 1245–1260.

(8) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38* (19), 2894–1896.

(9) Morales-Helguera, A.; Combes, R. D.; Perez-Gonzalez, M.; Cordeiro, M. N. Applications of 2D Descriptors in Drug Design: A DRAGON Tale. *Curr. Top. Med. Chem.* **2008**, *8* (18), 1628–1655.

(10) Matter, H.; Potter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211–1225.

(11) Schuffenhauer, A. Similarity searching in files of three dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295–307.

(12) Perera, A.; Vallverdú, M.; Clarià, F.; Soria, J. M.; Caminal, P. DNA Binding Sites Characterization by Means of Rényi Entropy Measures on Nucleotide Transitions. *Engineering in Medicine and Biology Society*; Proceedings of the 28th Annual International Conference of the IEEE, New York, August 30−Sept 3, 2006; IEEE: Piscataway, NJ, 2006; pp 5783−5786.

(13) Vinga, S.; Almeida, J. S. Local Renyi entropic profiles of DNA sequences. *Bioinformatics* **2007**, *393* (8), 1471–2105.

(14) Vinga, S.; Almeida, J. S. Rényi continuous entropy of DNA sequences. *J. Theor. Biol.* **2004**, *231*, 377–388.

(15) Krishnamachari, A.; Mandal, V. m.; Karmeshu. Study of DNA binding sites using the Rényi parametric entropy measure. *J. Theor. Biol.* **2004**, *227*, 429–436.

(16) , Dennington, R., II; Keith, T.; Millam, J.; Eppinnett, K.; Hovell, W. L.; Gilliland, R. *GaussView*, Version 3.09; Semichem, Inc.: Shawnee Mission, KS, 2003.

(17) Dewar, M. J.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. The development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(18) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; 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.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.01; Gaussian, Inc.: Wallingford, CT, 2004.

(19) Babel 1.6. [Online] http://www.ccl.net/cca/software/UNIX/babel/babel-1.6 (accessed May 2009).

(20) Prestwick Chemical. [Online] http://www.prestwickchemical.fr (accessed May 2009).

(21) Sigma Aldrich. [Online] http://www.sigmaaldrich.com (accessed May 2009).

(22) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.

(23) Melville, J. L.; Hirst, J. D. TMACC: Interpretable Correlation Descriptors for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2007**, *47*, 626–634.

(24) CORINA. [Online] http://www.molecular-networks.com/software/corina/index.html#addinfo (accessed April 2009).

(25) Principe, J.; Xu, D.; Fisher, J. *Information Theoretic Learning*; John Wiley: New York, 2000; Vol. 1, Chapter 7.

(26) Jenssen, R.; Hild, K. E.; Erdogmus, D.; Principe, J. C.; Eltoft, T. Clustering using Renyi's entropy. *Proceedings of the International Joint Conference on Neural Networks*; Portland, OR, July 20−24, 2003; IEEE: Piscataway, NJ, 2003; Vol. 1, pp 523−528.

(27) Gokcay, E.; Principe, J. C. A new clustering evaluation function using Renyi's information potential. *Acoustics, Speech, and Signal Processing*; Proceedings of the 2000 IEEE International Conference of Acoustics, Speech, and Signal Processing, Istanbul, Turkey, June 5−9, 2000; IEEE: Piscataway, NJ, 2000; Vol 6, pp 3490−3493.

(28) Poulain, R.; Horvath, D.; Bonnet, B.; Eckhoff, C.; Chapelain, B. From Hit to Lead. Analyzing Structure-Profile Relationships. *J. Med. Chem.* **2001**, *44*, 3391–3401.

(29) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2000**, *7*, 903–911.

(30) Ecker, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2006**, *12*, 225–233.

CI900275W