SHED: Shannon Entropy Descriptors from Topological Feature Distributions

Elisabet Gregori-Puigjané and Jordi Mestres*

Chemogenomics Laboratory, Research Unit on Biomedical Informatics,
Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra,
Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain

Received February 14, 2006

A novel set of molecular descriptors called SHED (SHannon Entropy Descriptors) is presented. They are derived from distributions of atom-centered feature pairs extracted directly from the topology of molecules. The value of a SHED is then obtained by applying the information-theoretical concept of Shannon entropy to quantify the variability in a feature-pair distribution. The collection of SHED values reflecting the overall distribution of pharmacophoric features in a molecule constitutes its SHED profile. Similarity between pairs of molecules is then assessed by calculating the Euclidean distance of their SHED profiles. Under the assumption that molecules having similar pharmacological profiles should contain similar features distributed in a similar manner, examples are given to show the ability of SHED for scaffold hopping in virtual chemical screening and pharmacological profiling compared to that of substructural BCI fingerprints and three-dimensional GRIND descriptors.

INTRODUCTION

The generation of mathematical representations for molecules has long been an active line of research in computational drug discovery. As a result, a large number and variety of molecular descriptors reflecting the one-dimensional, two-dimensional, and three-dimensional features of chemical structures have been devised.[1] Once formulated, the relevance of these descriptors is often established according to their ability to reflect the pharmacological properties of molecules.[2] Their potential impact in drug discovery is ultimately assessed when used, for example, in deriving quantitative structure−activity relationships from sets of molecules for which biological data is experimentally available[3] or in performing similarity searches of large chemical libraries against a panel of reference active compounds.[4]

Within this context, atom-centered feature pairs constitute an attractive family of molecular descriptors.[5] In their various formulations, they have been proven to show a decent performance on a diverse range of computational aspects in drug discovery covering quantitative structure−activity relationships,[5] compound selection,[6] virtual chemical screening,[7−10] and virtual pharmacological profiling.[11] In all of these studies, the actual computational encoding of the atom-centered feature-pair descriptors attempts to capture their overall distribution within a molecule by storing the occurrence of feature pairs at different distance ranges, either at the topological[5−8] or geometrical level,[9−11] to form a so-called binned fingerprint representation of each molecule. These molecular fingerprints are then used to assess the degree of resemblance between molecules according to different similarity metrics.[12−14]

Two aspects are worth emphasizing to understand the scope and limitations of this family of descriptors. On one hand, formulations based on geometric atom pair descriptors require three-dimensional coordinates and, thus, provide representations which are dependent on the conformation of the molecules.[9−11] Consequently, this involves a 2D to 3D conversion of atomic coordinates and potentially the generation of multiple conformers for each molecule. On the other hand, formulations based on topological atom pair descriptors do not need three-dimensional coordinates and do not have this conformational dependency.[5−8] However, they result in crisp representations of molecules that may not capture some of the essential information present only when using three-dimensional coordinates. A formulation based on a fuzzy description of topological features could be a good balance between the two approaches.

It is along these lines that a novel set of molecular descriptors is introduced. This new formulation takes advantage of the information-theoretical concept of Shannon entropy,[15] an approach that is increasingly being applied to process chemical information.[16−19] Accordingly, these new descriptors will be referred to as SHED, for SHannon Entropy Descriptors, and represent a means to quantify the variability displayed by topological distributions of atom-centered feature pairs in molecules. The following sections describe, first, the methodological details for obtaining SHED and, second, several application examples to assess their potential ability for scaffold hopping in virtual chemical screening and pharmacological profiling.

METHODS

The process of obtaining SHED from a chemical structure is illustrated in Figure 1 for dimetindene, a histamine H1 antagonist. The original input structure should be in MDL's SD file format.[20] From an SD file, each atom in a molecule is first mapped to a Sybyl atom type.[21] Subsequently, each atom type is assigned currently to one or more of four atom-centered features, namely, hydrophobic (H), aromatic (R),

* Corresponding author e-mail: jmestres@imim.es.

10.1021/ci0600509 CCC: $33.50
Published on Web 05/11/2006

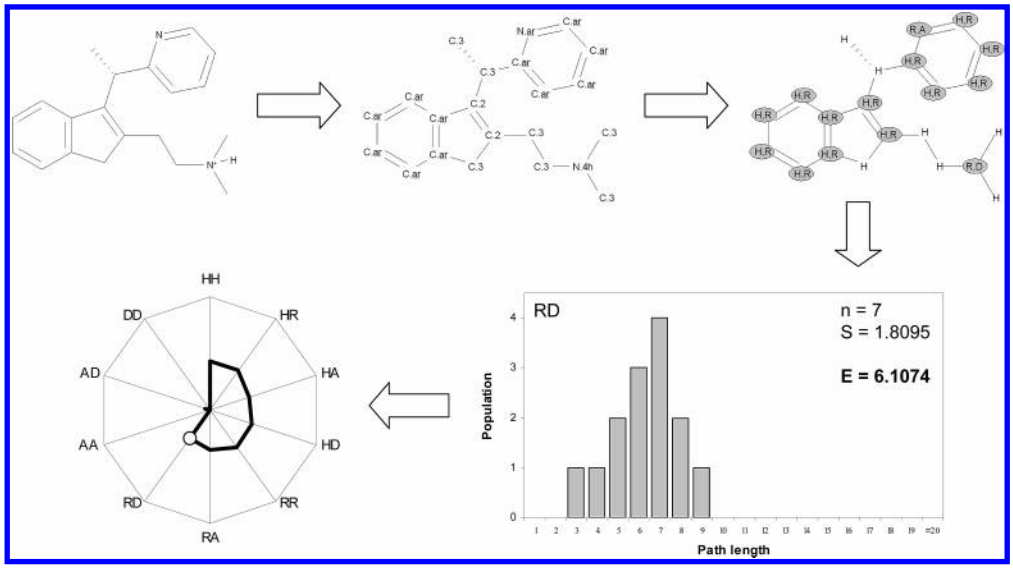

Figure 1. Generation of a SHED profile from chemical structure.

acceptor (A), and donor (D). For example, an aliphatic C.3 carbon will be assigned to a hydrophobic feature (H), whereas a protonated N.4h nitrogen will be assigned to both aromatic and donor features (RD). Then, the shortest path length between atom-centered feature pairs is derived and its occurrence at different path lengths stored to create a feature-pair distribution. A maximum path length of 20 bonds was used. Feature pairs at distances over 20 bonds are accumulated in the last bin. As an example, the distribution of RD feature pairs within dimetindene is displayed. An equivalent distribution is derived for each of the 10 possible feature pairs resulting from all pair combinations of the four features used.

At this stage, the concept of Shannon entropy[15] is applied to determine the variability of feature-pair distributions. Within this approach, the entropy, *S*, of a population, *P*, distributed in a certain number of bins (representing in this case the different path lengths), $N = 20$, is given by

$$S = -\sum_{i=1}^{N} \rho_i \ln \rho_i \qquad \rho_i = p_i/P \tag{1}$$

where ρ_i and p_i are, respectively, the probability and the population at each bin *i* of the distribution. The values of *S* range between 0, reflecting the situation of the entire population being concentrated in a single bin, and a maximum number, $S_{max} = \ln N$, reflecting the situation of a uniformly distributed population among all bins. In the case of dimetindene (Figure 1), RD pairs can be found at path lengths occupying seven bins and the variability in their population gives rise to a distribution with an entropy value of 1.8095. To have a more intuitive measure that can be linearly related to the situation of full uniform occupancy, entropy values are transformed into projected entropy values, $E = e^S$. Correspondingly, *E* values provide a measure of the expected maximum uniform occupancy from the corresponding *S* value. Now, for any given population $P > 0$, the values of *E* can vary from 1, reflecting the situation of zero entropy in which the population is totally concentrated in a single bin, to *N*, reflecting the situation of maximum entropy in which the population is uniformly distributed among all bins. In the limit case of $P = 0$, then *E* will be assigned to $E = 0$. For the RD feature pair in dimetindene (Figure 1), the maximum achievable *E* value for a population uniformly occupying seven bins would be $E = 7$. The obtained *E* value of 6.1074 reflects a slight deviation from the situation of full uniform occupancy on seven bins. This *E* value will ultimately be the SHED for the RD feature pair. The set of SHED values obtained for the 10 possible feature pairs constitute the SHED profile of a molecule. As illustrated in Figure 1, SHED profiles are represented using a wheel chart, the circle in the chart indicating the *E* value (SHED) for the RD feature pair in dimetindene.

RESULTS AND DISCUSSION

The basic assumption is that molecules having similar features arranged in a similar way should display similar SHED profiles. The underlying question is to what extent SHED profiles derived from topology-based atom-centered feature-pair distributions of molecules are well-suited to recognize the presence of similar features arranged in a similar way around significantly different molecular scaffolds, an ability usually referred to as "scaffold hopping".[7] To investigate this issue, an analysis of SHED profiles for molecules directed to different targets is presented next, followed by application examples on the use of SHED profiles for virtual chemical screening and pharmacological profiling.

Scaffold Hopping. Three diverse sets of molecules containing comparable features arranged similarly around essentially different scaffolds were selected. The structures of the molecules and their corresponding SHED profiles are collected in Figure 2. The first set includes a list of five known cyclooxygenase-1 inhibitors (COX-1: EC 1.14.99.1), namely, indomethacin (**1**), sulindac (**2**), fenbufen (**3**), ketoprofen (**4**), and indoprofen (**5**); the second set is a selection of five known thrombin inhibitors (Factor IIa: EC 3.4.21.5), namely, BM14.1224 (**6**), BM51.1047 (**7**), DAPA (**8**), 4-TAPAP (**9**), and 3-TAPAP (**10**); and the third set contains five estrogen receptor subtype α antagonists (ERα: NR 3.A.1), namely, raloxifene (**11**), bazedoxifene (**12**), a tetrahydroisoquinoline ligand (**13**), LY326315 (**14**), and EM343 (**15**). As can be observed, despite the significant scaffold

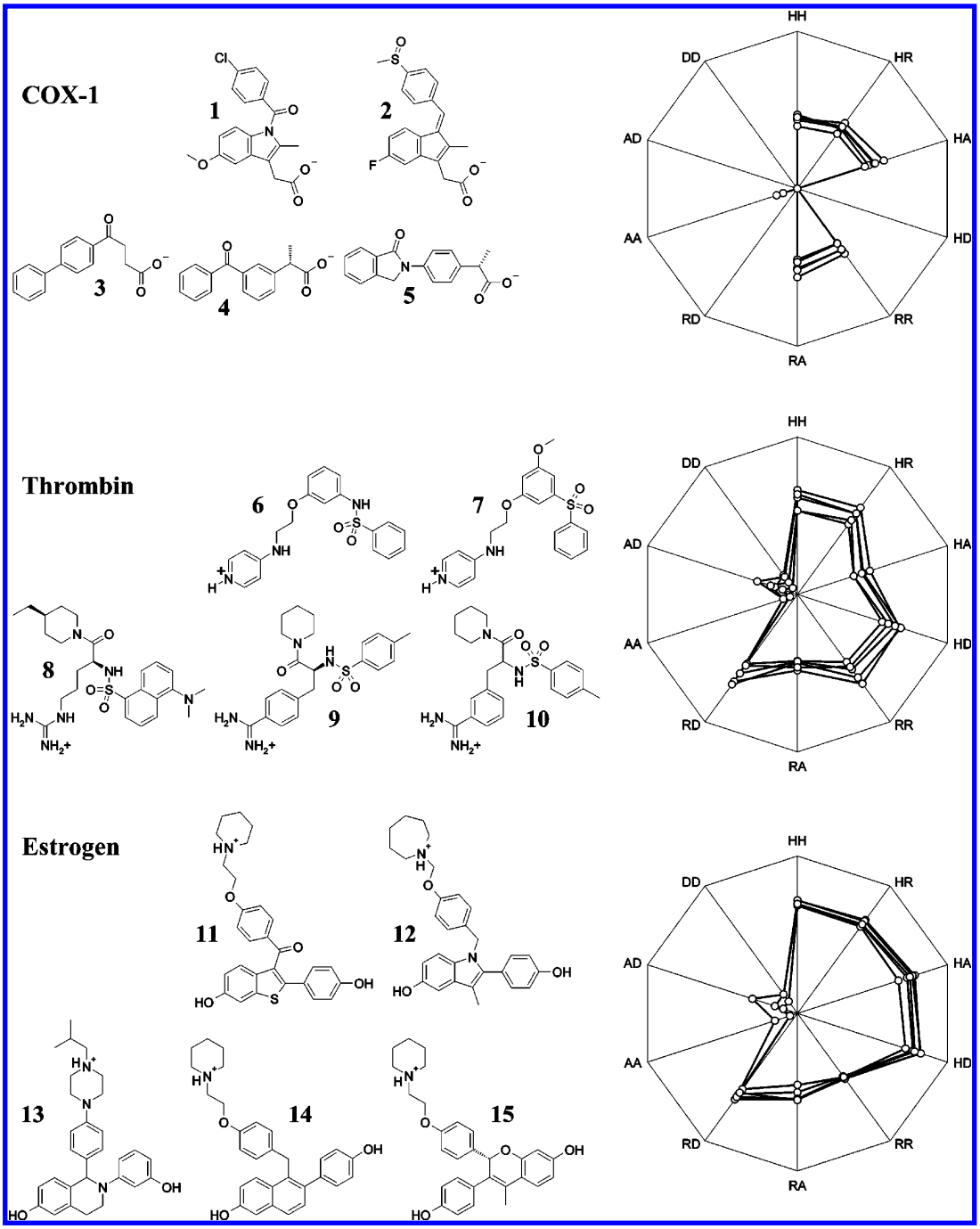

Figure 2. SHED profiles for three diverse sets of target-directed molecules, namely, cyclooxygenase inhibitors (top), thrombin inhibitors (middle), and estrogen α antagonists (bottom).

diversity present in the three families of chemical structures, reasonably equivalent SHED profiles are obtained within each set. At the same time, the various target-directed SHED profiles are found to be essentially different from one another. Altogether, these results provide evidence of the potential applicability of SHED for identifying molecules containing similar features distributed similarly around diverse molecular scaffolds.

A characteristic worth emphasizing is the fact that size appears to be implicitly accounted for in the SHED profile of molecules, particularly in the SHED values corresponding to feature pairs involving hydrophobic and aromatic centers. For example, average values and standard deviations of SHED for the HH pairs found in COX-1 inhibitors, thrombin inhibitors, and ERα antagonists are, respectively, 4.40 ± 0.27, 5.95 ± 0.61, and 7.01 ± 0.17, which are indicative of the increasing size of compounds associated with those targets. In general, as compounds become bigger and more complex (in terms of combinations of features), the area filled by their SHED profiles would tend to be larger as well.

To assess quantitatively the degree of discrimination obtained when comparing different types of molecules, average Euclidean distances and standard deviations derived from the SHED profiles of the three sets of ligands are presented in Figure 3. The clear separation obtained between comparisons of pharmacophorically similar molecules and comparisons of molecules having essentially different feature distributions is remarkable. For example, average intraset

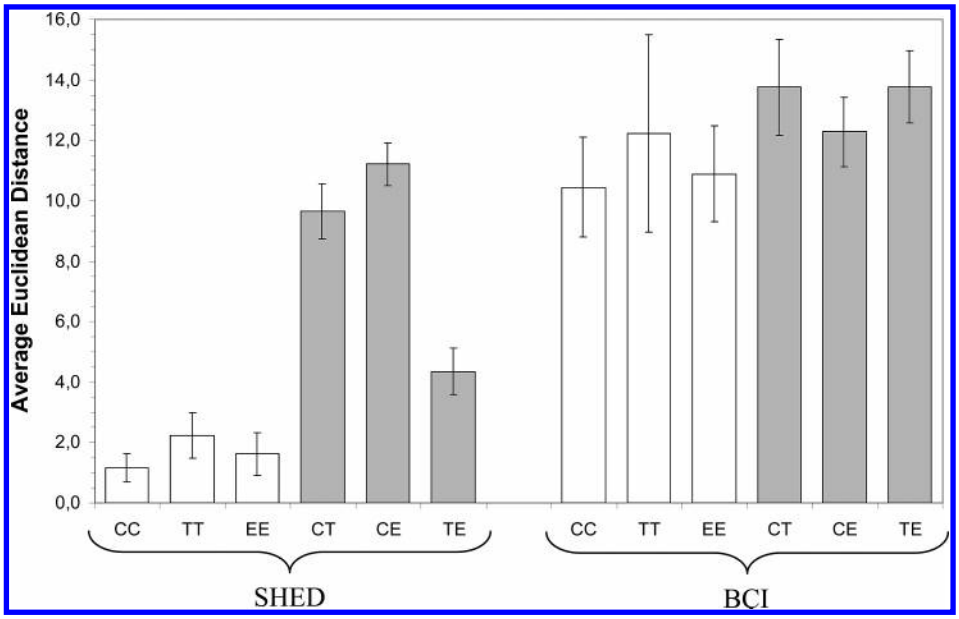

Figure 3. Average Euclidean distances and standard deviations obtained when using SHED profiles and BCI fingerprints for the three sets of ligands active to COX-1 (C), thrombin (T), and estrogen (E) shown in Figure 2. White bars correspond to intraset distances (e.g., CC refers to the 10 nonzero Euclidean distances between all pairwise combinations of the five COX-1 inhibitors), and gray bars correspond to interset distances (e.g., CT refers to the 25 Euclidean distances between all pairwise combinations of the five COX-1 and five thrombin inhibitors).

distances and standard deviations between COX-1, thrombin, and estrogen ligands are 1.17 ± 0.46, 2.23 ± 0.75, and 1.62 ± 0.71, respectively. In contrast, average interset distances and standard deviations when comparing COX-1/thrombin, COX-1/estrogen, and thrombin/estrogen ligands are 9.65 ± 0.92, 11.21 ± 0.70, and 4.35 ± 0.78, respectively. To put these results into perspective, the same exercise was done using BCI fingerprints,[22] a representative 2D substructural fingerprint-based method widely used in compound clustering and similarity searching.[23] In this case, the discrimination between intraset and interset comparisons is not as clear as that with SHED profiles, and a more fuzzy (less separation between intraset and interset distances) and less compact (larger values for standard deviations) picture appears. This outcome emphasizes the potential use of SHED in virtual screening applications.

Despite the attractive resemblance observed in the target-directed SHED profiles, both qualitatively in Figure 2 and quantitatively in Figure 3, those compounds represent only a focused subset extracted from the ample diversity of active compounds that could be identified and generated for a given target. In fact, having similar SHED profiles may well be a reflection of making analogous interactions within similar pockets in their respective targets. In reality, depending largely on the size of the protein binding cavity, its flexibility, and its degree of exposure to the solvent, ligands will bind to different pockets, exploit different interactions, and have a variety of solvent-exposed functional groups. In this scenario, a less compact picture will certainly emerge when visualizing the corresponding SHED profiles, but each SHED profile will be representative of a particular distribution of features allowed in an active compound for a given target.

Virtual Chemical Screening. Computational drug discovery of novel chemical modulators for members of the therapeutically relevant family of G protein-coupled receptors (GPCRs) is still dominated by ligand-based approaches

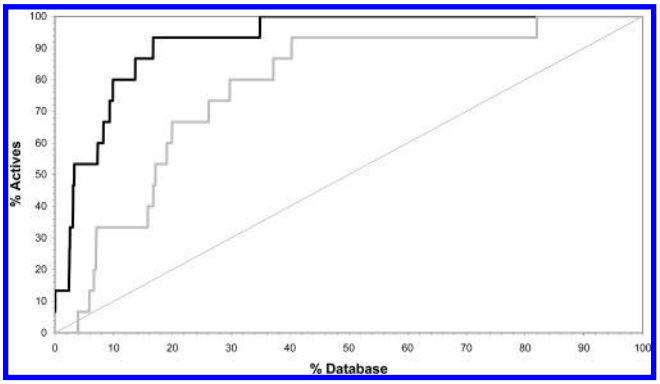

Figure 4. Enrichment curve for the retrieval of 15 α_{1A}-adrenoceptor antagonists from a database of 3033 drugs, using a reference set of 24 α_{1A}-adrenoceptor antagonists using SHED (black bold line) and GRIND (gray bold line) descriptors.

mainly because of, on one hand, the technical difficulties encountered in crystallizing these receptors and, on the other hand, the large amount of biological data available for molecules acting on these receptors. Accordingly, to illustrate the applicability and performance of SHED for virtual chemical screening on GPCRs, the structures of a set of 24 highly diverse α_{1A}-adrenoceptor antagonists with known binding affinities (Ki < 300 nM) were extracted from a recent publication.[24] The SHED profiles derived from those 24 reference compounds were then used to score a database composed of 3033 drugs and a test set of 15 α_{1A}-adrenoceptor antagonists.[24] The scoring of each compound in the database was simply assigned to the minimum value of all Euclidean distances calculated between the SHED profile of the compound and each one of the 24 reference SHED profiles.

The percentage of actives found with SHED within each percentage of the database is plotted in Figure 4 (black bold line). After rank ordering, selection of the top-ranked 5% and 10% of the compounds in the database would have included 53.3% (8) and 80.0% (12), respectively, of the 15

α_{1A}-adrenoceptor antagonists in the test set. In terms of overall enrichment, in the ideal situation that all 15 actives were found in the 15 top-ranked compounds, the value for the normalized area under the curve (AUC) would be 0.9975 (AUC_T), whereas a random identification of actives (symbolized by the thin diagonal line in Figure 4) would result in a normalized AUC of 0.5000 (AUC_R). Correspondingly, the normalized AUC of the resultant active identification line (black bold line) is 0.9220. From these AUC values, an enrichment factor can be defined as $E = (AUC - AUC_R)/(AUC_T - AUC_R)$. This enrichment factor can have values in the range of [−1.0, 1.0]. A value of $E = -1.0$ would reflect the worst scenario of finding all actives in the database in the last bottom-ranked compounds, whereas a value of $E = 1.0$ would reflect the ideal situation of finding all actives in the database in the first top-ranked compounds. A random identification of actives would result in $E = 0.0$. On the basis of this definition, the current virtual screening returns an enrichment of $E = 0.8482$.

For the sake of comparison, alignment-independent GRIND descriptors derived from three-dimensional molecular interaction fields were also calculated for all compounds using the program ALMOND[25] with three-dimensional structures derived by CORINA.[26] In this case, the scoring of each compound in the database was assigned to the minimum value of all Euclidean distances calculated between the first five scaled principal component analysis scores of the compound and those corresponding to each one of the 24 reference compounds. The percentage of actives found with GRIND within each percentage of the database is plotted in Figure 4 (gray bold line). After rank ordering, selection of the top-ranked 5% and 10% of the compounds in the database would have included 6.7% (1) and 33.3% (5), respectively, of the 15 α_{1A}-adrenoceptor antagonists in the test set. In terms of overall enrichment, the normalized AUC of the resultant active identification line (gray bold line) is 0.7770, which corresponds to an enrichment of $E = 0.5568$.

The performance of SHED for identifying the test set of 15 α_{1A}-adrenoceptor antagonists within the top positions of the rank-ordered database may have been masked to some extent by the presence of drugs that might as well have some affinity for the α_{1A}-adrenoceptor. For instance, droperidol and trazodone were found at ranks 9 and 29, respectively, and were neither in the reference nor in the test set of α_{1A}-adrenoceptor antagonists. Some examples of drugs that were found within the top ca. 1% of the rank-ordered database, together with the reference α_{1A}-adrenoceptor antagonist with the closest SHED profile to that of each drug, are collected in Figure 5. At rank 4, we found WAY 100635, an antagonist of the serotonin 5-HT_{1A} receptor. The closest α_{1A}-adrenoceptor antagonist found in the reference set is ARC 239 ($pK_i = 9.0$).[24] Despite having essentially different scaffolds, the similarities between the structural features of both compounds are remarkable. This result would thus alert one to the possibility of WAY 100635 hitting the α_{1A}-adrenoceptor as an off target. In fact, evidence can be found in the literature that WAY 100635 induces hypotension in anesthetized rats and that this effect could be partially explained by the antagonism of vascular α_1-adrenoceptors.[27] Ketanserin, a serotonin 5-HT_2 antagonist, was found at rank 14. The closest α_{1A}-adrenoceptor antagonist found in the reference set is spiperone ($pK_i = 8.1$), with which ketanserin shares a similar

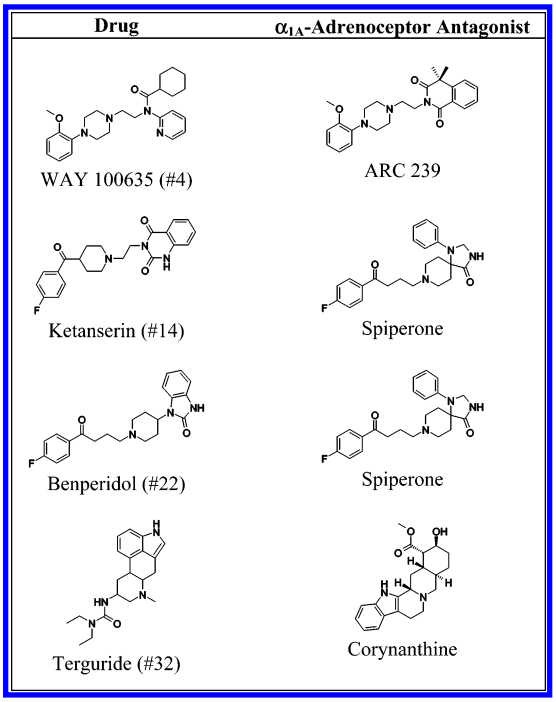

Figure 5. Selection of top-ranked drugs identified in the virtual chemical screening for α_{1A}-adrenoceptor antagonists. The reference α_{1A}-adrenoceptor antagonist having the closest SHED profile to each drug is also included. The rank position of the drug is given in parentheses.

distribution of pharmacophoric features.[24] Most interestingly, ketanserin was recently reported to be a potent antagonist for the α_{1A}-adrenoceptor ($pK_i = 8.0$).[28] Also benperidol, a dopamine D_2 receptor antagonist, is found at rank 22. As for ketanserin, spiperone was the closest reference α_{1A}-adrenoceptor antagonist to benperidol. Benperidol is structurally related to droperidol and shows a striking resemblance to spiperone, suggesting that the α_{1A}-adrenoceptor could be an off target for this drug as well. A final fourth example was extracted from rank 32, where terguride, a dopamine D2 partial agonist, was located. The closest α_{1A}-adrenoceptor antagonist found in the reference set is corynanthine ($pK_i = 7.5$).[24] The two compounds present no obvious structural similarities, but the relative distribution of the different features appears to be remarkably equivalent. Of mention is the fact that a recent study confirmed experimentally that terguride indeed displays potent antagonist properties at the α_{1A}-adrenoceptor ($pK_b = 8.0$).[29] On the basis of these results, the potential use of SHED profiles beyond virtual chemical screening will be investigated next.

Virtual Pharmacological Profiling. The basic assumption sustaining virtual chemical screening activities is that similar compounds are expected to have similar affinities for a given target. However, further than having affinity for a particular target, the ultimate biological effect of compounds is established by their pharmacological profile against a set of biologically relevant targets.[30] Therefore, for virtual pharmacological profiling, the original statement can be extended to assuming that similar compounds should display similar pharmacological profiles. To explore this aspect further, we took a list of 47 compounds for which the experimental pharmacological profile on a panel of 75 targets was known.[31] The majority of compounds had some affinity for the μ opiate receptor, but against a panel of diverse targets,

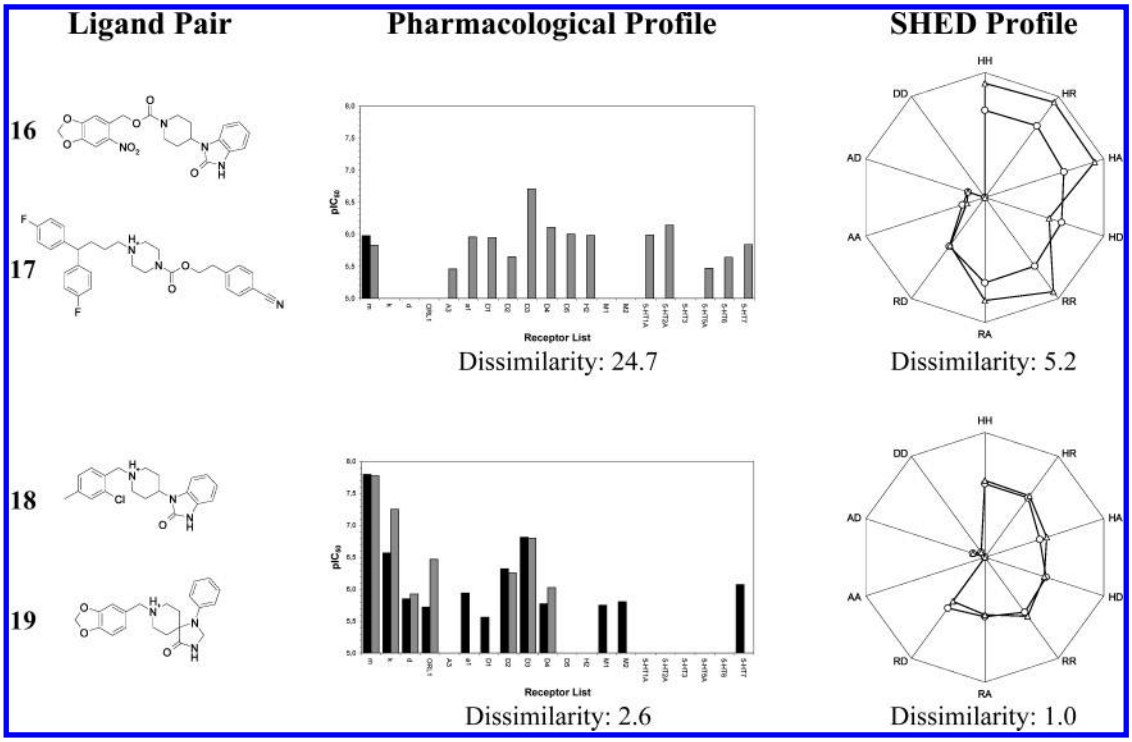

Figure 6. Structures, pharmacological profiles (extracted from ref 22), and SHED profiles for two selected pairs of molecules.

they displayed essentially different pharmacological profiles. As illustrative examples, two selected pairs of compounds are shown in Figure 6. Despite the apparent structural differences, compounds **16** and **17** had a similar affinity for the μ opiate receptor, with pIC_{50} values of 5.98 and 5.83, respectively. However, while the pharmacological profile of compound **16** shows a high selectivity toward the μ opiate receptor, compound **17** has a poor specificity, with a low micromolar affinity for 14 out of 20 GPCR targets. In contrast, compounds **18** and **19** not only had similar high affinities for the μ opiate receptor, with pIC_{50} values of 7.80 and 7.78, respectively, but showed also comparable overall pharmacological profiles against the panel of 20 GPCRs.

Having the ability to estimate in advance potential deviations in the pharmacological profiles of a list of hits identified in the early stages of a drug discovery project would be of great value to prioritize further optimization activities on those hits. Recent studies have explored the use of similarity metrics to analyze the relationship between the degree of pharmacophore similarity in a pair of compounds and the similarity of their respective pharmacological profiles.[31] Accordingly, the extent to which the SHED profiles introduced in this work reflect the relative pharmacological profiles of compounds is an aspect worth investigating at this stage. To this aim, the SHED profiles for the two pairs of compounds described above have been included also in Figure 6. A comparison of the SHED profiles obtained for compounds **16** and **17** evidences dissimilar feature distributions, the profile for compound **17** covering clearly a wider SHED area consistent with the larger size of this molecule relative to that of **16**. In contrast, compounds **18** and **19** showed visibly similar SHED profiles, consistent with a pair of compounds of approximately the same size containing similar features arranged similarly.

At a more quantitative level, as proposed in a previous study,[31] an activity dissimilarity score of the pharmacological profiles of two compounds can be defined as

$$D(\mathrm{A,B}) = \Sigma\, \Delta[|\%\mathrm{inh}_i(\mathrm{A}) - \%\mathrm{inh}_i(\mathrm{B})|]$$

where $\%\mathrm{inh}_i$ stand for the percentages of inhibition (at 10 μM) of A and B in the test i among 75 tests of the profile. The term

$$\Delta[|\%\mathrm{inh}_i(\mathrm{A}) - \%\mathrm{inh}_i(\mathrm{B})|]$$

was originally defined as an empirical measure of how different the two compounds behave with respect to that test:

$$\Delta(x) = \begin{cases} 0 \text{ if } x \le 30 \\ (x - 30)/40 \text{ if } 30 < x \le 70 \\ 1 \text{ if } x > 70 \end{cases}$$

As an example, Figure 6 contains also the values obtained for the dissimilarity of pharmacological profiles and SHED profiles between the two illustrative pairs of compounds. The visual observation that the pharmacological profiles of compounds **16** and **17** differ much more than those of compounds **18** and **19** is reflected in dissimilarity values of 24.7 and 2.6, respectively. Correspondingly, the differences observed in the SHED profiles of compounds **16** and **17** compared to those of compounds **18** and **19** result in dissimilarity values of 5.2 and 1.0, respectively. These results are illustrative of the potential use of SHED profiles as a means for alerting one to likely differences or similarities in the pharmacological profiles of series of molecules for which experimental affinities are available only for a particular target.

To investigate this issue further, SHED and pharmacological profile dissimilarities were evaluated for each of the pairwise combinations of the 47 compounds, in an analogous way to that reported in the previous study from which the

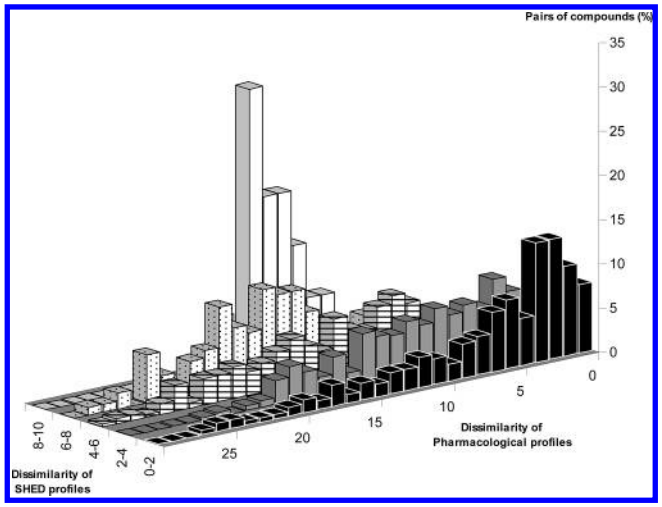

Figure 7. Dissimilarity of pharmacological profiles versus dissimilarity of SHED profiles for all compound pairs derived from a list of 47 compounds for which experimental affinity data on 75 targets are available (extracted from ref 31).

structures and experimental binding affinities of compounds were extracted.[31] Then, all pairs of compounds were sorted into dissimilarity categories. The resultant distribution of the pairs in each category versus the dissimilarity scores obtained for the calculated SHED profiles and the observed pharmacological profiles is given in Figure 7. For each pair of compounds, the dissimilarity of SHED profiles refers to the Euclidean distance between their SHED profiles, whereas the dissimilarity of pharmacological profiles accounts for the differences in their respective percentages of inhibition accumulated over the entire set of biologically relevant targets to which the compounds were tested against, as detailed above.[31] Remarkably, the overall distribution presented in Figure 7 reproduces qualitatively the results obtained in the earlier parent study (see Figure 2 in ref 31). In the class of compounds having the most similar feature distributions (dissimilarity of SHED profiles $\leq$ 2), the occurrence of pharmacologically similar compound pairs (dissimilarity of pharmacological profiles $\leq$ 5) is significantly high. With an increasing dissimilarity of SHED profiles, the relative probability of finding pairs of compounds with similar pharmacological profiles decreases considerably. Consequently, pairs of compounds having similar SHED profiles are more likely to have similar pharmacological profiles than any random pair of dissimilar molecules.

CONCLUSIONS

We have introduced SHED as a novel set of molecular descriptors based on the information-theoretical concept of Shannon entropy applied to quantifying the variability displayed by topological distributions of atom-centered feature pairs in chemical structures. Under this new representation, molecules containing comparable features arranged similarly around essentially different scaffolds give rise to similar SHED profiles, illustrated in this work for the cases of cyclooxygenase-1 inhibitors, thrombin inhibitors, and estrogen α antagonists. This property was then further assessed in a virtual chemical screening exercise to retrieve 15 α_{1A}-adrenoceptor antagonists from a database of 3033 drugs. Using a reference set of 24 diverse α_{1A}-adrenoceptor antagonists, selecting the top-ranked 10% of the compounds in the database would have included 80.0% of the α_{1A}-adrenoceptor antagonists in the test set, with an overall enrichment factor of 0.8482. In addition, the use of SHED helped to identify the α_{1A}-adrenoceptor as a potential off target for several top-ranked drugs. Finally, SHED profiles demonstrated a decent performance for estimating differences in the virtual pharmacological profiling of molecules, with pairs of compounds having similar SHED profiles showing a trend toward also having similar pharmacological profiles.

Given the large amount of experimental data available currently on the affinity of molecules to targets, ligand-based approaches to drug discovery remain still competitive against more sophisticated structure-based methods. In view of the results presented here, the use of SHED appears as a simple, yet attractive, low-dimensional representation of molecules with promising applicability for the virtual identification and profiling of novel hits at the early stages of drug discovery projects.

ACKNOWLEDGMENT

This research is supported by a grant from the Instituto de Salud Carlos III (Ministerio de Sanidad y Consumo), research project reference number 02/3051. We are indebted to Manuel Pastor (Universitat Pompeu Fabra, Spain) for deriving the GRIND descriptors, Xavier Fradera (Organon, Scotland) for obtaining the BCI fingerprints, and Rafael Gozalbes (Cerep, France) for making available the structures of the compounds used for the pharmacological profiling test case and generating Figure 7.

Supporting Information Available: The mapping of Sybyl atom types to pharmacophoric features; feature-pair distributions and the SHED profile for dimetindene; SHED profiles for COX-1, thrombin, and estrogen ligands; and the structures of the 39 α_{1A}-adrenoceptor antagonists 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) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000.
(2) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(3) Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. D. Chem-bioinformatics: Comparative QSAR at the interface between chemistry and biology. *Chem. Rev.* **2002**, *102*, 783−812.
(4) Maggiora, G. M.; Johnson, M. A. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(5) Carhart, R. E.; Smith, E. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(6) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(7) 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*, 2894−2896.
(8) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.
(9) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.
(10) Makara, G. M. Measuring molecular similarity and diversity: Total pharmacophore diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.
(11) Horvath, D.; Jeandenans, C. Neighborhood behaviour of in silico structural spaces with respect to in vitro activity spaces. A novel understanding of the molecular similarity principle in the context of

multiple receptor binding profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680−690.
(12) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(13) Maggiora, G. M.; Petke, J. D.; Mestres, J. A general analysis of field-based molecular similarity indices. *J. Math. Chem.* **2002**, *31*, 251−270.
(14) Mestres, J.; Maggiora, G. M. Putting molecular similarity into context: asymmetric indices for field-based similarity measures. *J. Math. Chem.* **2006**, *39*, 107−118.
(15) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1949.
(16) Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 796−800.
(17) Staura, F. L.; Godden, J. W.; Bajorath, J. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.
(18) Miller, J. L.; Bradley, E. K.; Teig, S. L. Luddite: An information-theoretic library design tool. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 47−54.
(19) Graham, D. J. Information content in organic molecules: Aggregation states and solvent effects. *J. Chem. Inf. Model.* **2005**, *45*, 1223−1236.
(20) Elsevier MDL, San Leandro, CA. http://www.mdli.com/ (accessed May 2006).
(21) Tripos, Inc., St. Louis, MO. http://www.tripos.com/mol2/atom_types.html (accessed May 2006).
(22) Barnard Chemical Information Ltd., Leeds, U. K. http://www.bci.gb.com/ (accessed May 2006).
(23) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.
(24) Li, M.-Y.; Tsai, K.-C.; Xia, L. Pharmacophore identification of α_{1A}-adrenoceptor antagonists. *Bioorg. Med. Chem. Lett*. **2005**, *15*, 657−664.
(25) 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.
(26) Molecular Networks, GmbH, Erlangen, Germany. http://www.mol-net.de/ (accessed May 2006).
(27) Villalobos-Molina, R.; Lopez-Guerrero, J. J.; Gallardo-Ortiz, I. A.; Ibarra, A. Evidence that the hypotensive effect of WAY 100635, a 5-HT1A receptor antagonist, is related to vascular alpha 1-adrenoceptor blockade in the adult rat. *Auton. Autacoid. Pharmacol*. **2002**, *22*, 171−176.
(28) Israilova, M.; Suzuki, F.; Tanaka, T.; Nagatomo, T.; Taniguchi, T.; Muramatsu, I. Binding and functional affinity of sarpogrelate, its metabolite m-1 and ketanserin for human recombinant alpha-1-adrenoceptor subtypes. *Pharmacology* **2002**, *65*, 69−73.
(29) Newman-Tandredi, A.; Cussac, D.; Audinot, V.; Nicolas, J.-P.; De Ceuninck, F.; Boutin, J.-A.; Millan, M. J. Differential actions of antiparkinson agents at multiple classes of monoaminergic receptor. II. Agonist and antagonist properties at subtypes of dopamine D2-like receptor and a1/a2-adrenoceptor. *J. Pharmacol. Exp. Ther.* **2002**, *303*, 805−814.
(30) Giaever, G.; Flaherty, P.; Kumm, J.; Proctor, M.; Nislow, C.; Jaramillo, D. F.; Chu, A. M.; Jordan, M. I.; Arkin, A. P.; Davis, R. W. Chemogenomic profiling: Identifying the functional interactions of small molecules in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 793−798.
(31) Poulain, R.; Horvath, D.; Bonnet, B.; Eckhoff, C.; Chapelain, B.; Bodinier, M.-C.; Déprez, B. From hit to lead. Analyzing structure−profile relationships. *J. Med. Chem.* **2001**, *44*, 3391−3401.

CI0600509