

Enrichment of Ligands for the Serotonin Receptor Using the *Shape Signatures* Approach

Karthigeyan Nagarajan,[†] Randy Zauhar,[‡] and William J. Welsh^{*,†}

Department of Pharmacology, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School (UMDNJ-RWJMS), 675 Hoes Lane, Piscataway, New Jersey 08854, and Department of Chemistry & Biochemistry, The University of the Sciences in Philadelphia, 600 S. 43rd Street, Philadelphia, Pennsylvania 19104

Received August 16, 2004

Shape Signatures, a new 3-dimensional molecular comparison method, has been adapted to rank ligands of the serotonin receptors. A set of 825 agonists and 400 antagonists together with approximately 10 000 randomly chosen compounds from the NCI database were used in this study. Both 1D and 2D Shape Signature databases were created, and enrichment studies were carried out. Results from these studies reveal that the 1D Shape Signature approach is highly efficient in separating agonists from a mixture of molecules which includes compounds randomly selected from the NCI database taken as inactive. It is also equally effective at separating agonists and antagonists from a pool of active ligands for the serotonin receptor. Parallel enrichment studies using 2D shape signatures showed high selectivity with more restricted coverage due to the high specificity of 2D signatures. The influence of conformational variation of the shape signature on enrichment was explored by docking a subset of ligands into the crystal structure of serotonin N-acetyltransferase. Enrichment studies on the resulting “docked” conformations produced only slightly improved results compared with the CORINA-generated conformations.

INTRODUCTION

With the advent of massively large chemical databases, investigators are bound to screen literally millions of compounds in an effort to select those that might show biological activity against a selected target, such as an enzyme active site or membrane-bound receptor. One approach to this prodigious task is to employ virtual screening, where in-silico methods reduce huge databases to a perceptibly enriched subset of promising molecules for further experimental testing.

Excellent review articles on virtual screening¹ in general and virtual screening methods² in particular have recently appeared in the literature, thus a brief overview will suffice here. Virtual screening techniques can be broadly classified into two major groups namely, receptor based and ligand based virtual screening. When experimentally determined structures of the protein are available, receptor-based virtual screening by molecular docking and ranking (“scoring”) is used for enhancing efficiency in lead optimization. Comprehensive discussions on receptor-based enrichment techniques are available elsewhere for the interested reader.^{3–5} In the absence of suitable structural data on the receptor, ligand-based virtual screening methods can achieve similar enrichments to those obtained via molecular docking. Several ranking methods such as binary kernel discrimination, similarity searching, and (sub-)structure searching for ligand-based virtual screening have been reported and compared.⁶ Once a set of active inhibitors has been identified, virtual

screening using fragment searching techniques represents a viable avenue for lead discovery.⁷

Ligand-based virtual screening via similarity searching is the method where known ligands (agonists, antagonists) for the receptor in question are utilized as queries to retrieve molecules from a large database of compounds and ranked in terms of their similarity to query or queries with respect to polarity, logP, molecular weight, presence of particular functional groups, etc., that are deemed good indicators of similar biological activity (e.g., binding affinity). This task of enrichment is a challenging problem for large databases, since it is very likely that many compounds of dissimilar activity will nonetheless share many properties in common. As a consequence, many enrichment methods evince a high rate of false-positive selections.

Similarity searching has been an active area of research for several years.^{8,9} There are numerous ways to describe the similarity between two molecules for chemical similarity searching, including substructure and fragment-based methods.¹⁰ The crux of the similarity searching problem rests in identifying key properties of molecules, which effectively distinguish between actives and inactive in a mixture of compounds. Among the properties derived from 2D and 3D representations of molecules, and the empirical bioactivities, a descriptor representing molecular shape is essential in carrying out such a similarity search. The shape of a molecule is often represented by topological descriptors such as Wiener indices, connectivity indices, valence connectivity indices, Kier’s shape indices, Kier’s alpha-modified shape indices, Balaban indices, etc. These indices are computed from molecular graphs, which encode two-dimensional connectivity data. Descriptors derived from the 3-D spatial arrangement of atoms of the molecule include molecular surface

* Corresponding author phone: (732)235-3234; fax: (732)235-3475; e-mail: welshwj@umdnj.edu.

[†] University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School.

[‡] The University of the Sciences in Philadelphia.

area, radius of gyration, principal moment of inertia, molecular volume, and distances between pairs or triplets of substructures.¹¹

Having obtained a numerically tractable form of shape information in terms of such descriptors, effective comparison of these descriptors is crucial. Certain distance comparison metrics and similarity coefficients commonly used in chemical information systems are Hamming distance, Euclidean distance, Soergel distance, Tanimoto coefficient (also called Jaccard coefficient), and the dice coefficient. Most similarity descriptors are derived directly from two-dimensional chemical connectivity and, hence, are inefficient in encoding the three-dimensional shape of a molecule.

In contrast, the present Shape Signatures approach¹² compactly encodes molecular shape information, optionally in conjunction with properties mapped onto the molecular surface. The procedure employs a form of ray tracing, in which the volume of a molecule (defined by its solvent-accessible surface) is explored by a single ray, which is propagated in the interior of the molecule by the rules of optical reflection. "Shape Signatures" are probability distributions derived from the ray-trace. The simplest of these is the distribution of segment lengths observed for the ray-trace, where a segment is any portion of the ray lying between two consecutive reflections. We call these "1D" signatures to emphasize the fact that the domain of the distribution is one-dimensional. Signatures with domains of higher dimension may be defined as joint probability distributions for observing a particular value for the sum of the segment lengths on either side of a reflection point, together with the value of certain surface property defined at the reflection. Such "2D" signatures thus involve a two-dimensional domain for the probability distribution. In applications of the Shape Signatures method conducted thus far, the surface property chosen is the molecular electrostatic potential (MEP) leading to the designation "2D-MEP" signature.

The Shape Signatures approach represents an efficient and useful means of encoding molecular shape and surface-property information, and, furthermore, it provides the means for very fast comparisons of molecules on the basis of shape and polarity.¹² The shape signatures of any two molecules can be compared using simple metrics, and the distances computed between signatures can be immediately used in virtual screening, clustering, and classification schemes.

The primary goal of the present work is to demonstrate the ability of the Shape Signatures approach in retrieving agonists of the serotonin 5HTA receptor in various realistic scenarios. We have taken 1225 ligands of the serotonin receptor from the MDDR database, which had been previously classified into two groups of 825 agonists and 400 antagonists. By splitting a random set of agonists into training set A' and test set A, we have attempted to retrieve A from a mixture D, where D contains A together with inactive molecules. We have conducted two case studies: (1) the mixture D contained A and antagonist molecules and (2) the mixture D contained A together with molecules randomly chosen from the NCI database all of which abide by Lipinski's rule of five. Key "drug like" properties of the NCI molecules, such as the molecular weight (<500) and log P (<5), lie in the same range as the active molecules. In the first case, we explored the efficiency of Shape Signatures in

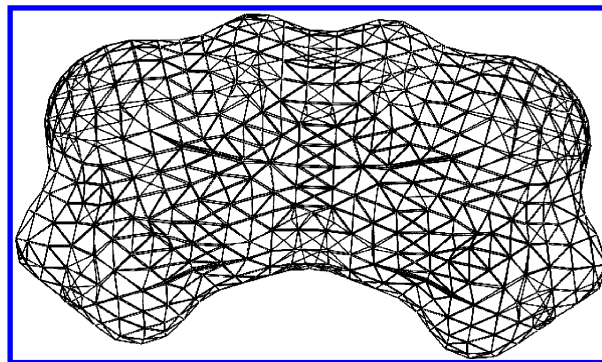


Figure 1. Triangulated solvent accessible surface.

separating agonists within the actives; the second case study was directed at retrieving agonists from a mixture of inactives. After establishing a suitable criterion for selection, we obtained acceptable results for the first case and excellent results for the second case.

To address the possible influence of conformational flexibility on shape signatures and thereupon on enrichment, a subset of compounds was subjected to docking studies. The top scoring conformations obtained from GOLD¹³ docking were used in a separate set of enrichment calculations. The results, although improved somewhat, were not dramatically better than that found for the CORINA-generated conformation.

It is observed that the Shape Signatures approach completely and compactly encodes the 3D shape of the molecules in the form of a histogram. It works reasonably well in distinguishing between agonist and antagonist molecules where the structural differences between them are narrow; this is precisely the situation where many of the general descriptor-based classification systems fail. Furthermore, it performs very well in the more general case of distinguishing between active molecules (serotonin receptor ligands) and inactives (background) represented by the NCI compound database. When compared against MACCS fingerprints combined with Tanimoto Index (TI) to gauge similarity, the Shape Signatures approach gave equivalent or superior enrichments in both of these case studies.

MATERIALS AND METHODS

Shape Signatures. A shape signature is a histogram representation of the ray segment lengths obtained from the ray tracing within the triangulated solvent accessible volume (SAV) of the molecule. The triangulated SAV, shown in Figure 1, was generated by the SMART algorithm.¹⁴ The histogram is obtained by binning the ray lengths from a ray-tracing algorithm, which traces the reflection of several random rays inside a triangulated molecular surface. By using a large number (10 000–50 000) of ray segments, we obtain a histogram that is unique and reproducible for each molecule regardless of the selected point of initiation of the rays. It is worth noting that shape signatures are inherently invariant to rotation of the molecule. Unlike other shape-based comparison approaches, no molecular alignment procedure is required. A shape signature which contains only the length information is termed a 1D shape signature. In addition, as described above, we can create 2D-MEP shape signatures

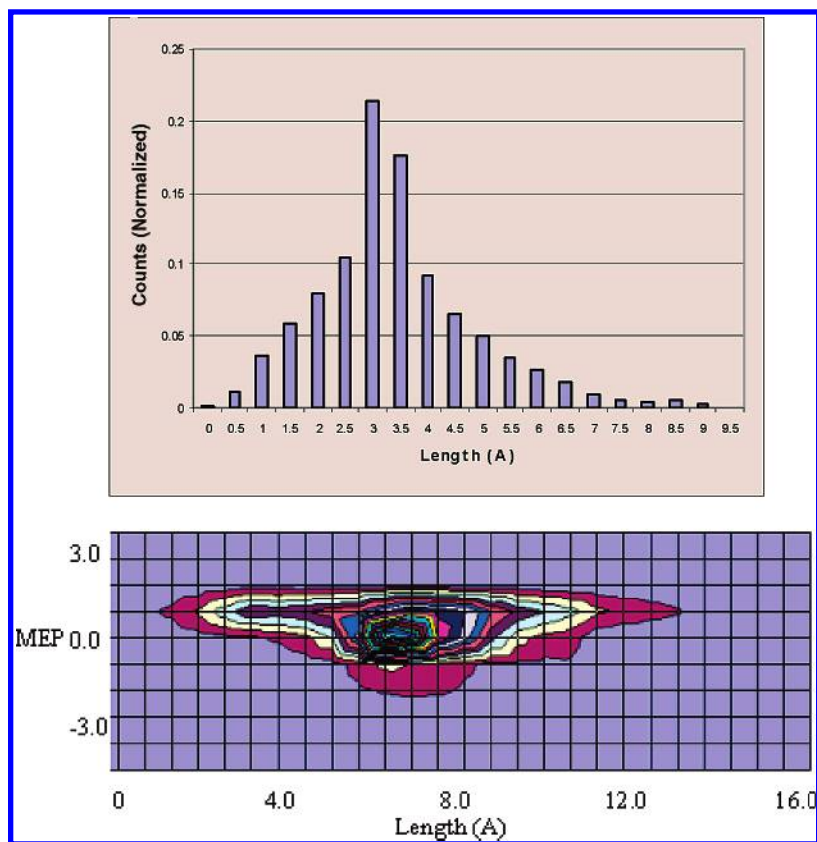


Figure 2. Typical 1D and 2D histograms.

that encode both segment length and electrostatic potential information associated with the point of incidence inside the SAV. Representative 1D and 2D-MEP shape signatures are shown in Figure 2.

Comparing the Shape Signatures of Molecules. Once the shape signature for a molecule is generated, we employ two common metrics to compare it with other shape signatures.¹² The first and most elementary metric is similar to the Manhattan distance, or L_1

$$L_1 = \sum_i |H_i^1 - H_i^2| \quad (1)$$

where H_i^1 and H_i^2 are bin heights of the i th bin for the two histograms compared. The second metric, which emphasizes the longer segment lengths associated with molecular shape, is called the ramp metric, or R_1

$$R_1 = \sum_i d_i |H_i^1 - H_i^2| \quad (2)$$

where d_i is the length corresponding to the i th bin. For further details on the algorithmic features of the Shape Signatures method, the reader is directed to Zauhar et al.¹²

MATERIALS

Figure 3 shows the scheme for the analysis of results and the terminologies used in this work. Our goal is to determine the effectiveness of the Shape Signatures algorithm to separate agonist and antagonist from a general pool. In the first experiments the pool consists of only antagonist and agonist. Later we also included a randomly selected set of

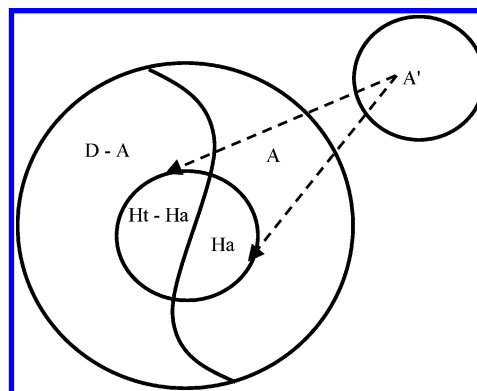


Figure 3. Scheme of typical database and hitlist. A' is the agonist molecules in the training set; A is the agonist test-set molecules in the database. D is molecules in the entire database. H_t is the hitlist and H_a is the agonist present in the hitlist.

molecules from the NCI database. In both these cases, the analysis procedure remained the same.

The schema for construction of the test and training sets of compounds are shown in Figure 4. In the first case, we selected 1225 serotonin ligands which were separated into two groups of 825 agonists and 400 antagonists. A subset of molecules was randomly extracted from the set of agonists and considered as the training set A' . The size of the training set was varied between 20% and 60% of the entire set of agonists for this study. The remaining agonists were mixed with the antagonists to comprise the agonist-antagonist test set. In the second case, 10 033 compounds were randomly selected from the NCI database and mixed with the agonist set to constitute what is called the "combined" (i.e., actives plus inactives) test set. For the entire study, a single conformation for each molecule was adopted. The 3D

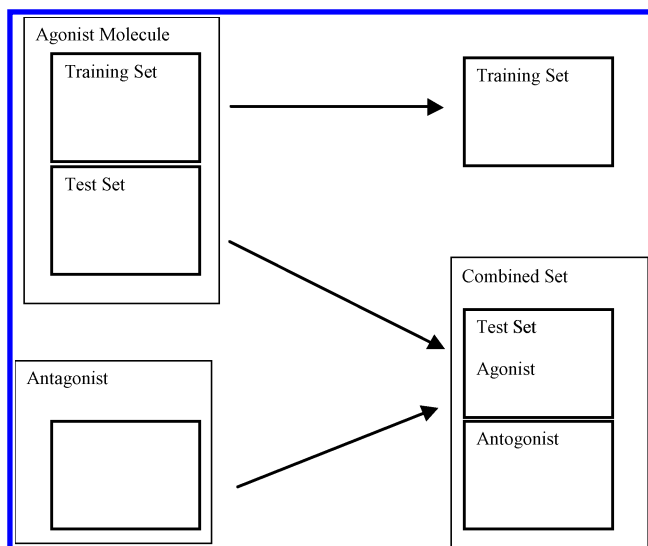


Figure 4. Scheme for constructing test and training sets.

coordinates of this conformation were generated from the corresponding SMILES string using CORINA.

In a separate study, the sensitivity of the shape signatures and subsequent classification analysis to conformation was evaluated using a smaller set (~800) of the serotonin ligands. Specifically, each ligand was docked 25 independent times inside the serotonin N-acetyltransferase receptor site¹⁵ using the GOLD program.¹³ The pose with the highest GOLD score for each compound was retained for classification analysis.

Procedure. The 1D and 2D shape signature histograms were first generated for all the molecules under study. Each molecule from the training set is compared with each molecule in the test set, and the shape signature “1D score” is calculated. The score ranges from 0 to 2 when the L_1 metric is used, where $L_1 = 0$ denotes “identity” and $L_1 = 2$ denotes “maximum dissimilarity”. The score list is sorted in ascending order, and hits are selected from the top of the list (i.e., minimal score corresponds to maximal similarity). Figure 5a represents the comparison results of 300 molecules and the hits they fetched. The agonist query is shown along the y axis where the top 300 hits are shown along the x-axis. The agonist and antagonist hits are shown as black dots and gray dots, respectively. Among various possible ways for

defining a molecule as a hit, we considered two separate criteria as variables and assessed their significance. The first criterion is based on the score, i.e., hits with scores less than a threshold value are selected; the second criterion collects a specific number of hits from the top of the list, irrespective of their scores.

Criteria for Defining Compounds as Hits. Figure 5 shows three representations of the same hitlist from a query subset of 100 agonist molecules. The red dots are true positives (agonists) and black dots are false positives (antagonists). The white space denotes no hit. Figure 5a contains the top 100 hits from the list, whereas parts b and c show only those hits with scores that fall within cutoff values set at $L_1 = 0.1$ and 0.05, respectively. It is seen that stipulation of a cut off enhances the quality of the results; the $L_1 \leq 0.05$ criterion results in considerably improved enrichment. If we define the optimal cutoff value as ‘s’, one of the objectives of this study is to determine the best cutoff value in the present application.

Due to the random selection of training-set compounds, we encounter two extreme types of query molecules (Figures 6 and 7). Molecule **M1** is linear and nearly planar, and its 1D-shape signature histogram closely matches with most of the molecules in the combined set. If we use the threshold criterion, even a reasonably stringent value, this molecule will retrieve several molecules from the mixture of agonists and antagonists resulting in poor yield. In the present example (Figure 6), molecule **M1** retrieved 205 molecules using a threshold value of 0.06.

At the other extreme, molecule **M2** is structurally distinct from the other molecules in the combined test database. Its 1D-shape signature histogram shows a bimodal distribution, differing considerably from the majority of molecules. If we use the top ‘n’ criterion, where the n top-ranked molecules are retrieved as hits irrespective of their scores, then the top of the hitlist will have molecules with high scores and will again be a mixture of agonists and antagonists. In the example shown the score ranged from 0.05 for the top hit to 0.28 for the 50th ranked molecule. By using both criteria, and adjusting the values used for n and s we can attain an acceptable balance of yield and coverage. The SAVs of the molecules are shown in Figure 8.

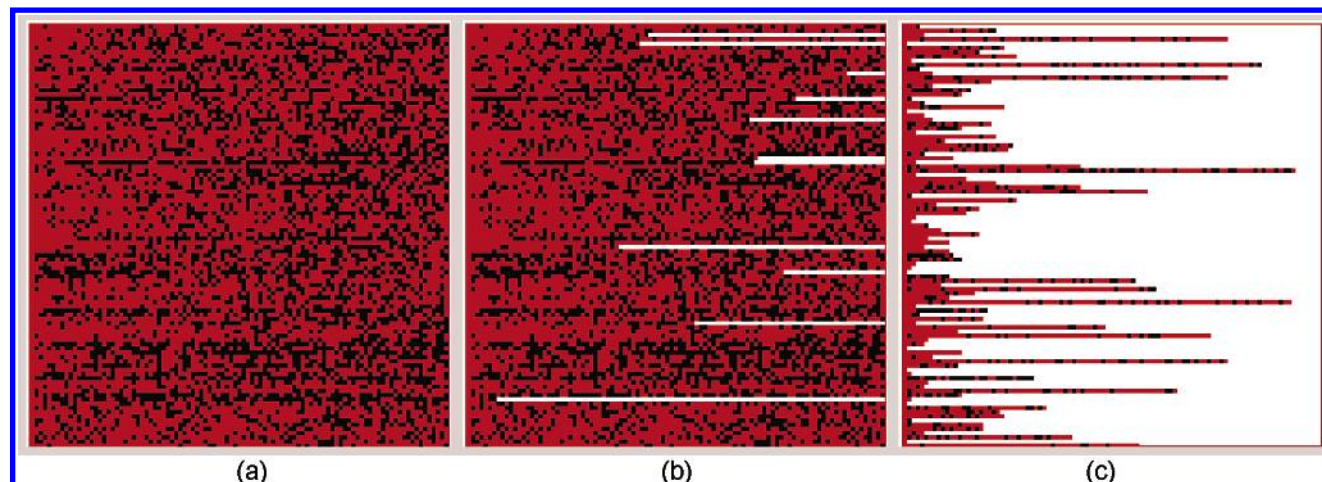


Figure 5. Representation of the hitlist, where the Y-axis is a generalized reference to the array of 100 query agonists. The red and black dots indicate agonists and antagonists, respectively, in the hitlist. A maximum of 100 hits was considered in this example. Part a shows all 100 top hits, while parts b and c show hits whose score is respectively less than 0.1 and 0.05.

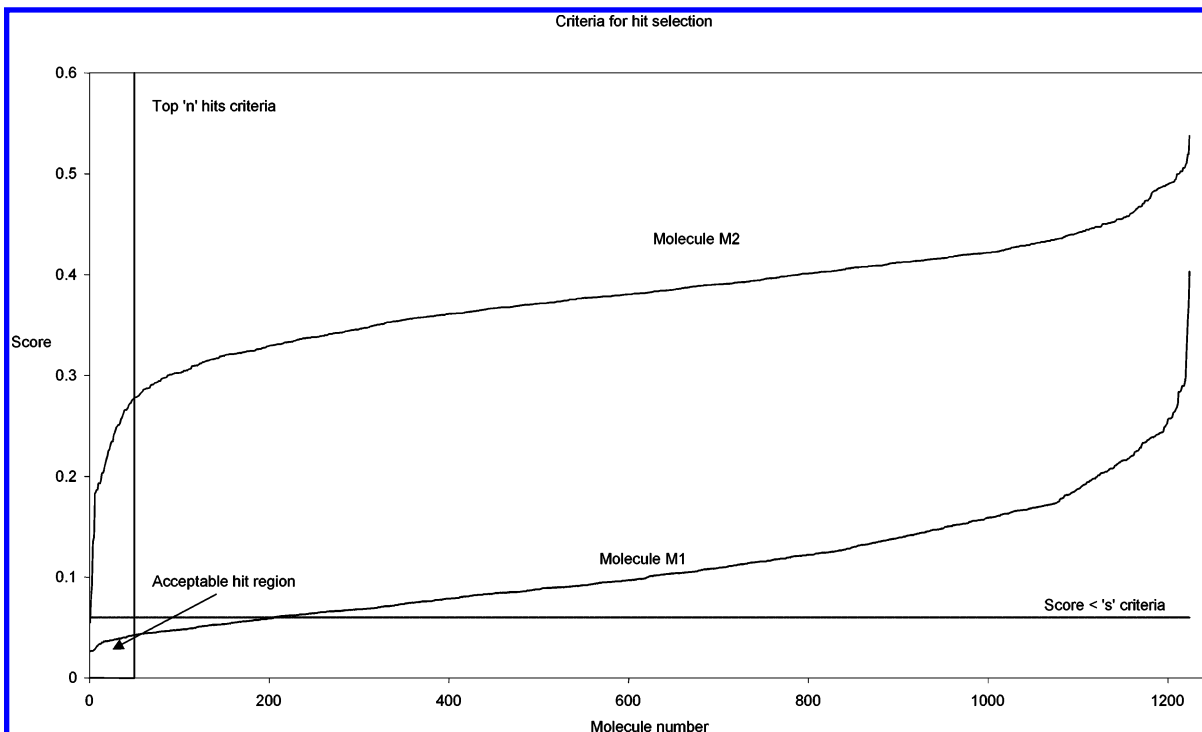


Figure 6. Criteria of hitlist generation from comparison of results.

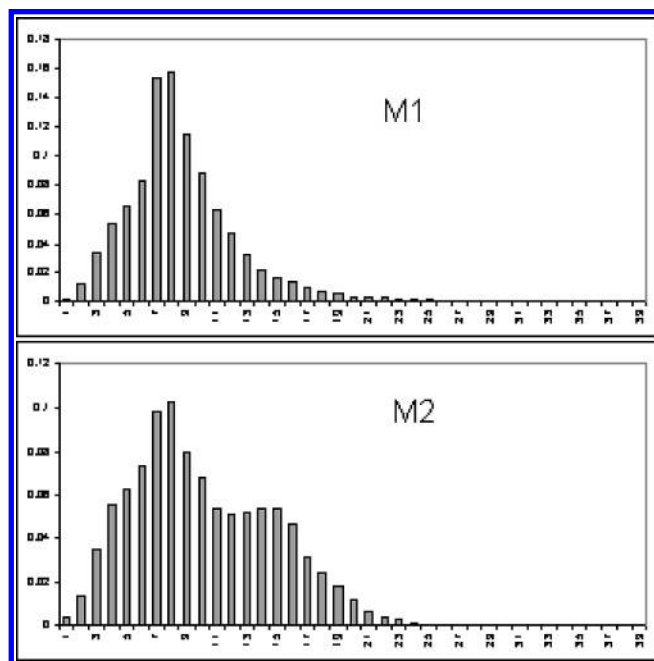


Figure 7. Histograms of molecules M1 and M2.

Comparison Metrics. The metrics used in this paper to analyze the hits are yield, enrichment, and coverage as proposed by Güner.¹⁶ These metrics are defined by the formulas below.

Yield: Yield is defined as the percentage of known actives in the hitlist. Yield is a measure of the purity of the hitlist. If the hitlist contains only actives then the yield will be 100%.

$$Y = \frac{H_a}{H_t} * 100 \quad (3)$$

Enrichment: Enrichment is defined as the ratio of yield of actives in the hitlist (H_a/H_t) relative to the yield of actives

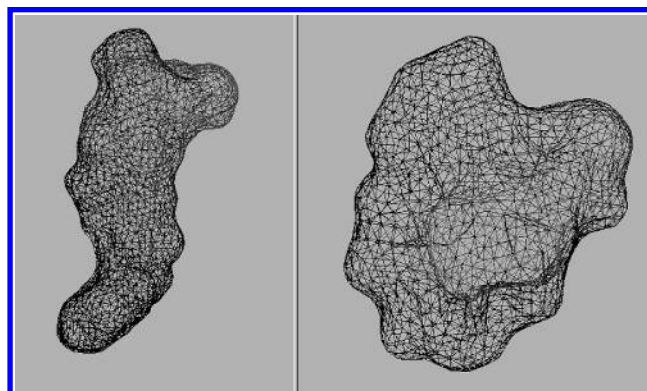


Figure 8. Triangulated SAV of molecules M1 and M2.

in the database (A/D).

$$E = \frac{\frac{H_a}{H_t}}{\frac{A}{D}} \quad (4)$$

Coverage: Coverage is defined as the percentage of known active compounds retrieved from the database. Coverage is the ability of the procedure to retrieve the actives from the database. If all the actives are retrieved, coverage will be 100%.

$$C = \frac{H_a}{A} * 100 \quad (5)$$

Under normal circumstances, yield and coverage tend to move in opposite directions. That is, if the yield is high the coverage will be low and vice versa. An ideal virtual screening method should retrieve all the actives and result in only actives in the hitlist, where both yield and coverage will be 100% since $H_t = H_a = A$. The enrichment will equal

D/A according to eq 4. This ideal situation is a rarity, hence most applications establish a preference for high yield or high coverage depending on the situation.

Docking Studies. In the absence of structural data for the serotonin receptor, the X-ray crystal structure of the 14–3–3 ζ : serotonin N-acetyltransferase complex was retrieved from the Protein Data Bank (PDB 1IB1) for use in this study. About 50% of the agonist and antagonist molecules were chosen randomly and docked inside the receptor pocket using the GOLD software. The conformation with high fitness score was taken, and shape signatures were generated followed by enrichment studies. The shape signatures of conformations of the same molecules generated by CORINA and GOLD were compared.

Comparison with MACCS Fingerprints. To evaluate the present results obtained from Shape Signatures, we made comparisons with a popular method. MACCS fingerprints were generated for 836 agonist and 411 antagonist molecules using MOE.¹⁷ Each molecule was compared with other molecules in the database using the Tanimoto Index (TI) set to $TI = 0.85$ as a similarity criterion, and the similarity matrix was calculated using MOE. This matrix was further employed to calculate the enrichment, yield, and coverage.

RESULTS AND DISCUSSION

To assess the behavior of this ranking procedure, comparison metrics were calculated after adjusting several variables. The values of n (number of top ranked molecules) and s (cutoff similarity threshold) were varied to verify the influence of these parameters on the metrics defined in the previous section. The additional parameter of interest is the ratio of the size of training set to that of the test set. Initially we carried out all the calculations with this ratio set to 50%. The database contained 412 agonists and 400 antagonists, therefore $D = 812$. The training set also contained 412 agonists ($A = 412$), hence the maximum possible value of E will be $812/412 = 1.97$. We obtained $D = 1.44$ and 1.55 from the 1D and 2D Shape Signature methods, respectively.

Table 1 provides the metrics calculated for a range of n and s values within the domain of interest using 1D and 2D methods, respectively. It is observed from the table that, as the values of s and n increase, the coverage approaches 100% and yield decreases to 50%. A yield of 50% corresponds to that expected for random selection from a database containing 50% actives, such as the present case. To achieve a reasonable yield, the s and n values must be confined within tight bounds. In this example the optimum values of s and n were 0.04 and 1, respectively, based on the enrichment value of 1.44. That is the maximum enrichment when only the nearest neighbor ($n=1$) is considered with a tight threshold of $s = 0.04$. The yield was 73% and the coverage was 44%. To achieve 100% coverage—corresponding to all the agonists being retrieved as hits from the training set of molecules—the yield was similar to that of a random selection.

Table 2 summarizes the results of the enrichment studies using 1D and 2D histograms and both the L_1 (linear) and R_1 (ramp) metrics. The 2D method gave better enrichment for both metrics, although at a higher cost in compute time (i.e., 3 s vs 43 s on a Pentium IV 2.4 GHz PC with Suse Linux operating system). In both 1D and 2D comparisons, the ramp

Table 1. Table of Enrichment Results for Varying s and n Values in Case 1 (Agonist and Antagonist Mixture)

enrichment		n						
s	1	2	3	4	5	10	50	100
0.03	1.29	1.29	1.3	1.31	1.3	1.29	1.29	1.29
0.04	1.44	1.34	1.28	1.27	1.25	1.22	1.22	1.22
0.06	1.36	1.16	1.08	1.06	1.05	1.05	1.04	1.04
0.09	1.35	1.07	1.02	1.01	1.01	1.01	1.01	1.01
0.2	1.35	1	1	1	1	1	1	1

yield		n						
s	1	2	3	4	5	10	50	100
0.03	65.57	65.33	66.01	66.23	65.81	65.61	65.61	65.61
0.04	73.2	67.91	64.84	64.21	63.5	61.93	61.93	61.93
0.06	69.21	58.91	55.02	53.97	53.43	53.03	52.8	52.8
0.09	68.61	54.05	51.72	51.19	51	51	51	51
0.2	68.52	50.8	50.74	50.74	50.74	50.74	50.74	50.74

coverage		n						
s	1	2	3	4	5	10	50	100
0.03	19.42	23.79	24.51	24.76	24.76	25	25	25
0.04	44.42	57.52	60.44	61.41	61.65	62.38	62.38	62.38
0.06	63.83	89.08	91.75	92.48	92.72	93.45	93.69	93.69
0.09	68.45	97.09	98.54	99.03	99.27	99.51	99.51	99.51
0.2	68.69	99.76	100	100	100	100	100	100

Table 2. Comparison of 4 Different Scoring Methods of Molecular Similarity

	1D	1D ramp	2D	2D ramp
max enrichment	1.44	1.49	1.55	1.54
yield	73.20	75.65	78.42	78.02
coverage	44.42	42.23	36.17	43.93
s	0.04	0.20	0.10	1.00
n	1	1	2	1

metric is found to provide better results compared with the linear metric.

With $s = 0.04$ and $n = 1$, the yield and coverage were calculated using 1D shape signatures for 6 different agonist training set/database ratios. It is observed that coverage improves (Figure 9) while yield diminishes (Figure 10) as the ratio of molecules in the test set and the combined database increases from 20 to 70. The reduction in yield can be explained by the fact that, as this ratio increases, the proportion of agonists left in the combined database is reduced and hence the number of hits is reduced. Table 3 summarizes the metrics calculated for the various training sets together with the theoretical maximum enrichment.

When comparative similar studies were conducted with MACCS fingerprints using a similarity threshold of $TI = 0.85$, we observed that the yield and enrichments were noticeably lower although the coverage was high. Table 4 lists the key results from MACCS fingerprint-based screening with various training-set ratios. At low values of training set to test set ratio, both approaches show slightly better enrichment. At high ratios, however, the enrichment achieved by the Shape Signatures approach is much better than the fingerprint-based method.

We observed that the agonist and antagonist molecules in our database exhibit strong similarity and, hence, by achieving high coverage and yield simultaneously was practically impossible. When the agonists were mixed with compounds taken from the NCI database and the same procedures repeated, we could achieve a high enrichment ratio and 100%

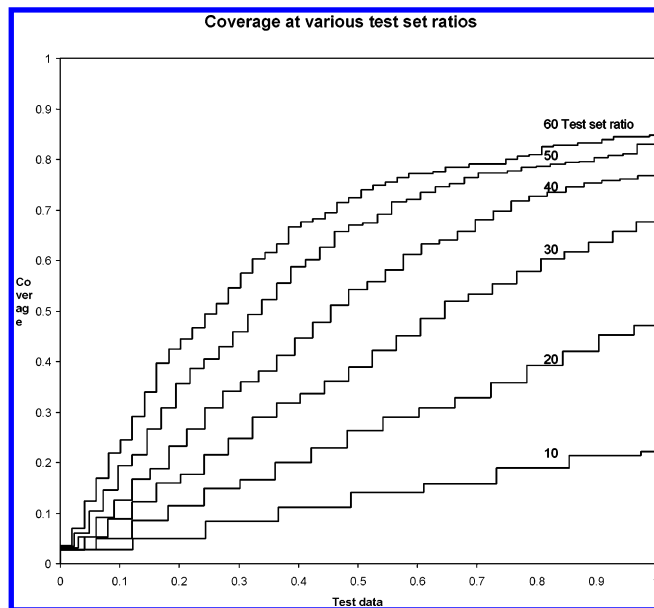


Figure 9. Coverage calculated for various ratios of agonists in the training set versus the remaining data set ranging from 10% to 70%, with $s = 0.04$ and $n = 1$.

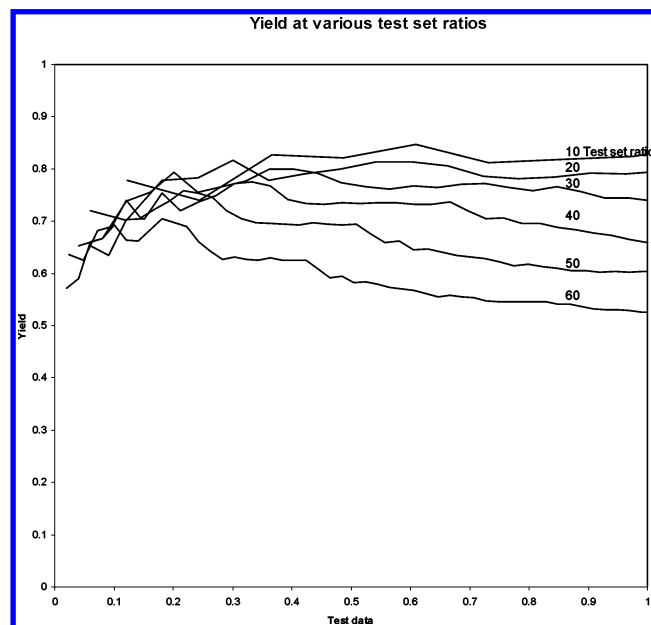


Figure 10. Yield calculated for various ratios of agonists in the training set to those in the remaining data set ranging from 10% to 70%, with $s = 0.04$ and $n = 1$.

Table 3. Enrichment (E), Yield (Y), and Coverage (C) Calculated at Various Ratios as Defined in Figure 10

training set, %	no.	E	Y	C	theoretical max E
20	165	1.28	80.00	33.33	6.42
30	248	1.32	78.14	53.90	3.94
40	330	1.28	70.87	69.29	2.71
50	413	1.25	63.39	79.85	1.97
60	495	1.20	54.37	83.03	1.47
70	578	1.20	53.40	83.18	1.22

purity with reasonable coverage of 39%, i.e., only the agonists are found as hits satisfying the criteria. The theoretical maximum enrichment $E = 25.35$ ($D = 10445$, $A = 412$) can also be realized. We could also observe a purity of 100% with a high coverage of 99.5%. Unlike the first case, where the high enrichment was seen only for $n = 1$

Table 4. Enrichment, Yield, and Coverage Calculated Using MACCS Fingerprints and $TI = 0.85$

training set, %	no.	E	Y	C	theoretical max E
20	167	1.17	68.12	52.52	6.46
30	250	1.13	65.57	48.32	3.9
40	334	1.10	64.65	44.89	2.73
50	418	0.96	54.22	42.71	1.98
60	501	0.90	50.56	33.54	1.48
70	585	0.76	41.03	25.43	1.13

(nearest neighbor consideration), here the enrichment is high even at higher values of n and s . These results suggest that this method is well suited for realistic situations. When the same molecules were subjected to MACCS fingerprint-based studies with $TI = 0.85$, we observed that the coverage was not much improved. For 50% of the molecules in training set the yield achieved was significantly lower than the Shape Signature method (91% versus 100%) which is reflected in the enrichment $E = 23$ (theoretical maximum 25.35).

Comparison of Docked and CORINA-Generated Conformations. The shape signatures of docked and CORINA-generated conformations of the same molecules were compared against each other. It was observed that, of the 800 molecules compared, the difference score for 80% of the molecules was found to be less than 0.085 and 0.225, respectively, for 1D and 2D comparison. For certain molecules, however, the scores were as high as 0.26 and 0.6, respectively, for the 1D and 2D shape signatures. The 3D images of the SAVS of corresponding molecules from CORINA and GOLD are given in Figure 11. The presence of rotatable (i.e., single) bonds in the molecules does not play an obvious role in these shape signature differences. This can be observed in Figure 12, where both 1D and 2D differences between CORINA-generated and GOLD-docked conformations are plotted. (For the sake of clarity, the symbols are shifted in the plot).

In studies using a subset comprising 50% of the molecules in our original database, it was observed that the enrichments obtained for conformations generated by docking with GOLD and CORINA-generated conformations were very similar. The theoretical maximum of enrichment is 2.91 ($D = 614$, $A = 207$), whereas a maximum of 2.51 was achieved here.

CONCLUSION

We have demonstrated the utility of the Shape Signatures method to the problem of classifying molecules as agonists or antagonists for a common receptor, working from a variety of test databases that include a mixture of these compound types together and, alternatively, in combination with compounds selected at random from the NCI database. The key parameters for searching and matching the molecules were determined. The cutoff parameters n and s can be adjusted such that any extreme cases in the training set can be balanced, and the results can be made consistent and independent of the selection of test-set molecules. We have found that, with optimal search parameters, it is possible to correctly identify a majority of the agonists in a mixture of agonists and antagonists with good yield and reasonable coverage. Excellent results are achieved with databases that also include randomly selected compounds from NCI. The present approach produces much better enrichments com-

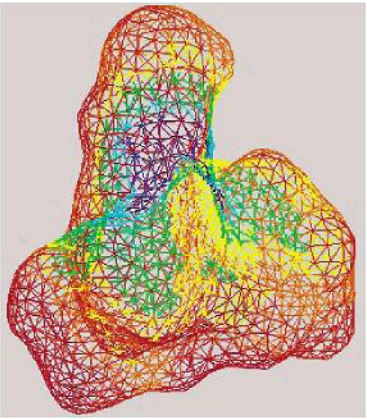
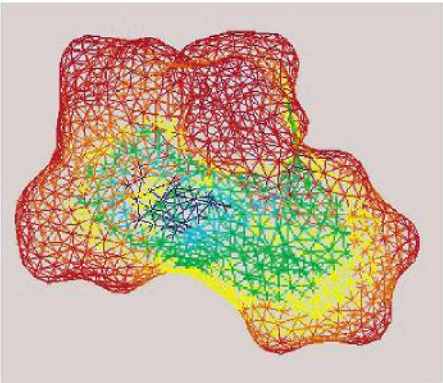
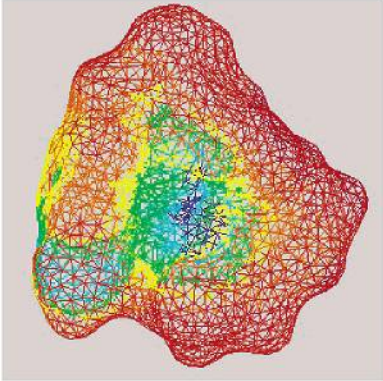
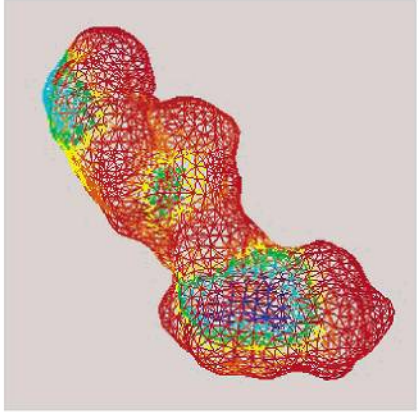
Max Difference	Corina	Gold	Score
1D			0.26
2D			0.57

Figure 11. Comparing molecular structure from CORINA and GOLD-docked conformations showing high differences.

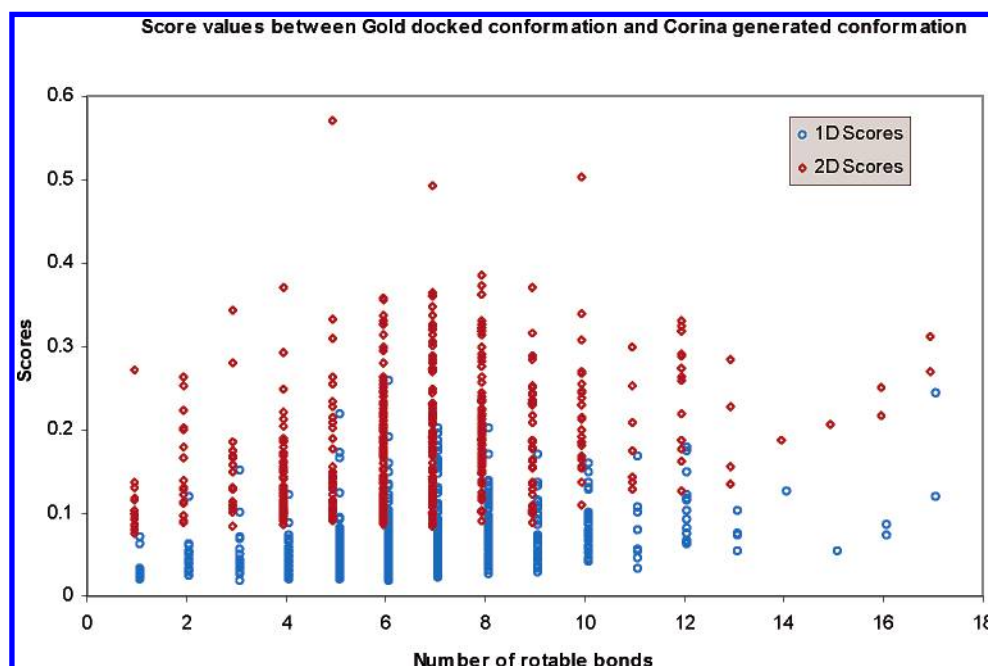


Figure 12. Comparison between CORINA generated and docked conformations.

pared to the widely adopted MACCS fingerprints using the standard similarity threshold $TI = 0.85$.

We also found by separate docking studies, in which the receptor-docked conformations of the ligands were used for

the classification analysis, that conformational variations posed little complication for Shape Signature classification at least in the present application. Studies in our laboratory are underway to explore the generality of this observation. The Shape Signatures method can be readily applied in situations where a receptor structure is unavailable and where ligand conformations are generated automatically (e.g., using CORINA). We find that Shape Signatures is a promising method for molecular classification and expect it to find applications in those cases where molecules cannot be easily classified on the basis of chemical structure alone but where shape and electrostatic properties must be taken into account.

ACKNOWLEDGMENT

We are grateful to Drs. Andrew Maynard and James Damewood of AstraZeneca (Wilmington, DE) for providing the set of serotonin receptor agonists and antagonists.

Supporting Information Available: Complete tables of results from the enrichment studies described in the present paper. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Jain, A. N. Virtual Screening in lead discovery and optimization. *Curr. Opin. Drug Discovery Dev.* **2004**, *4*, 396–403.
- (2) Xu, H.; Agrafiotis, D. K. Retrospect and Prospect of Virtual Screening in Drug Discovery. *Curr. Topics Med. Chem.* **2002**, *2*, 1305–1320.
- (3) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002** Oct 15; *7*(20): 1047–55.
- (4) Waszkowycz, B. Structure-based approaches to drug design and virtual screening. *Curr. Opin. Drug Discovery Dev.* **2002** May; *5*(3): 407–13.
- (5) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002** Jun 1; *47*(4): 409–43.
- (6) David, W.; Willett, P. Comparison of Ranking Methods for Virtual Screening in Lead Discovery Programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469–474.
- (7) Verdonk, M. L.; Hartshorn, M. J. Structure-guided fragment screening for lead discovery. *Curr. Opin. Drug Discovery Dev.* **2004**, *4*, 404–10.
- (8) Barnard, J.; Downs, G. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (9) Downs, G. M.; Willett, P. Similarity Searching in Database of Chemical Structures. *Rev. Comput. Chem.* **1995**, *7*, 1–66.
- (10) Villar, H. O.; Koehler, R. T. Comments on the design of chemical libraries for screening. *Mol. Diversity* **2000**, *5*, 13–24.
- (11) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D Structural Descriptors Relevant to Ligand–Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (12) Zauhar, R.; Moyna, G.; Tian, L.; Li, Z.; Welsh, W. J. Shape Signatures: A New Approach to Computer-Aided Ligand- and Receptor-Based Drug Design. *J. Med. Chem.* **2003**, *46*, 5674–5690.
- (13) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
- (14) Zauhar, R. SMART: A solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J. Comput.-Aided Mol. Design* **1995**, *9*, 149–159.
- (15) Obsil, T.; Ghirlando, R.; Klein, D. C.; Ganguly, S.; Dyda, F. Crystal Structure of the 14–3–3 ζ Serotonin N–Acetyltransferase Complex: A Role for Scaffolding in Enzyme Regulation. *Cell (Cambridge, Mass.)* **2001**, *105*, 257.
- (16) Waldman, M.; Güner, O. F. Effective Analysis of Data Mining Results Molecular Simulations Inc. (San Diego CA) Web Publications http://www.lib.uchicago.edu/cinf/221_nm/talks/221_nm067.pdf.
- (17) MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Quebec, Canada.

CI049746X