Determination and Mapping of Activity-Specific Descriptor Value Ranges for the Identification of Active Compounds

Hanna Eckert and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received November 3, 2005

MAD (Mapping to Activity class-specific Descriptor value ranges) is a novel molecular similarity method that relies on the identification of activity-specific descriptors. Applying a categorical descriptor scoring function, value ranges of molecular descriptors in screening databases are compared with those in classes of active compounds and descriptors displaying significant deviations are selected. In order to identify new actives, database molecules are mapped to class-specific value ranges and ranked using a similarity function. As a mapping algorithm, MAD is distinct from many other molecular similarity and virtual screening methods. In systematic virtual screening trials, for small selection sets of only 30 database compounds, average hit and recovery rates over six activity classes ranged from about 10% to 25% and about 25% to 75%, respectively. Moreover, when mining a database of bioactive molecules many similar compounds were selected (with hit rates between about 15% and 79%). Our findings suggest that it is possible to generate compound class-directed descriptor reference spaces for molecular similarity analysis.

1. Introduction

Computational approaches used to assess molecular similarity, evaluate structure−activity relationships, and predict active compounds generally rely on combinations of molecular descriptors[1] for the design of appropriate chemical reference spaces.[2] In the context of QSAR,[3] molecular similarity analysis,[4] or ligand-based virtual screening,[5] there is a long-standing interest in the identification of molecular descriptors that selectively respond to specific compound activities.[6] The availability of "selective" descriptors provides an important basis for distinguishing between active and inactive molecules and for the identification of novel hits and leads, using methods such as clustering,[7] partitioning,[8] or similarity searching.[9] Several approaches operate on the basis of compound class-directed descriptor selection. For example, recursive partitioning[10] makes it possible to trace descriptor pathways that enrich active compounds in terminal nodes, and partitioning based on principal component analysis[11] permits the detection of descriptors that contribute the most to desired classification results.[11] In addition, techniques such as nonlinear mapping[12] or multidimensional scaling[13] have been applied to focus on most important descriptor contributions for diversity analysis or library design. For similar purposes, self-organizing maps[14] have been integrated into QSAR analysis for descriptor selection. Furthermore, the BCUT metric[15] in conjunction with the concept of receptor-relevant subspaces[16] can be used to concentrate compounds around certain descriptor axes in orthogonal chemical reference spaces.

Although the above approaches allow prioritization and selection of descriptor sets for specific purposes, little progress has been made in identifying descriptors that have a strong tendency to respond to unique features of active compounds. This is in part due to the fact that compound classes often respond very differently to alternative classification and virtual screening methods.[17] Only very few studies have attempted to analyze descriptor value distributions on a large scale in order to aid in descriptor selection. For example, we systematically studied and compared the information content of molecular descriptors in various databases by application of the Shannon entropy (SE) concept.[18] Through the introduction of differential Shannon entropy (DSE),[19] we were able to introduce a value range dependence of information content calculations, leading to the development of the SE-DSE metric[20] for database profiling. Application of this concept combining SE and DSE calculations made it possible to classify descriptors and assign them to various information content categories.[19,20] Furthermore, it was possible to identify descriptors that responded differently to systematic chemical differences between, for example, natural and synthetic molecules.[20] In principle, the SE−DSE metric could also be applied to explore differences between active and inactive molecules and find descriptors that are sensitive to chemical characteristics of an activity class. However, what makes it difficult to apply this approach to the analysis of activity classes is the fact that sets of active compounds are usually orders of magnitude smaller than screening databases. If only 10 or 50 compounds are available as a reference set, which is too small to be a statistical sample, SE−DSE calculations are no longer a reliable measure of differences in information content.

This situation has led us to explore the possibility of identifying activity class-specific descriptor value ranges in a more direct manner. The presence of "signature" value ranges that distinguish a set of active compounds from other database molecules is likely to provide a basis for the identification of novel actives. We found that many descriptors indeed adopt narrow value distributions for sets of active compounds that differ from their value distribution in screening databases. On the basis of these findings, we investigated whether it was possible to develop a virtual screening method by mapping molecules to multiple activity-specific descriptor value ranges and identifying those that closely match these ranges. Here we report the results of our studies designed to determine activity class-specific descriptors and utilize them for virtual screening calculations. These investigations have led to the development of a new method termed "mapping to activity class-specific descriptor value ranges" (MAD) to effectively screen large

* To whom correspondence should be addressed. Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

10.1021/jm051110p CCC: $33.50
Published on Web 03/09/2006

compound databases. When MAD was applied to six compound activity classes, significant hit and recovery rates were achieved in simulated virtual screening situations that compared very well to those obtained with established reference methods. Moreover, many compounds with biological activity similar to that of template molecules could be identified with MAD when a collection of active molecules and drugs was mined.

2. Methodology

2.1. Concept. As a first step, we investigated whether it was possible to determine value ranges of molecular property descriptors that systematically differed between sets of active compounds and random database molecules. Then we explored if they could also be used to identify other active compounds. This was attempted by application of a mapping algorithm. Given their calculated descriptor values, database compounds falling into multiple activity-class specific value ranges were assigned a high score and probability to be active. The MAD approach introduced here is suitable for large-scale virtual screening because compound databases can be efficiently scanned to select molecules that closely map descriptor settings of active template compounds.

2.2. Algorithm.

2.2.1. Descriptor Statistics. For the compound database used as the source for screening, a descriptor statistic is generated as follows: for each descriptor, we calculate the minimum value in the database (dbMin), the maximum value (dbMax), and its median, defined as the value that divides a population of values into two equal halves (above and below the median). Furthermore, we calculate the 25%-quantile ($q^{0.25}$), the 75%-quantile ($q^{0.75}$), and the standard deviation (stdDev) of the value distribution of each descriptor. For compounds belonging to an activity class, the minimum (classMin) and maximum (classMax) values of each descriptor are determined. Then, the descriptor value range (exactRange) of the activity class is defined as exactRange = [classMin, classMax], and the size of the value range is sizeRange = classMax − classMin.

2.2.2. Descriptor Scoring. The scoring function is designed to compare descriptor value ranges of activity classes and database molecules. If a descriptor with a broad value distribution in the source database produces only a narrow value range for a given class of active compounds, it might correspond to feature(s) important for specific activity. For the database distribution of each descriptor, we distinguish the central 50% of the values (termed centralRange), from the highest 25% and lowest 25% of values for each descriptor. This makes it possible to avoid a biasing influence of extreme values on descriptor scoring. Similar correction procedures are implemented for the definition of exactRange by omitting a defined number of active compounds with highest and lowest descriptor values. However, for the activity classes tested here, introducing these corrections did not measurably change the results, as reported below. Based on the definition of these three value subranges, we can distinguish between three different scoring categories for each descriptor:

(a) The value range of an activity class falls within the centralRange or overlaps with it: classMax ≥ $q^{0.25}$ or classMin ≤ $q^{0.75}$. Then the descriptor score is calculated as

$$\text{score} = \frac{q^{0.75} - q^{0.25}}{\text{sizeRange}}$$

(b) The value range of an activity class completely falls below the "low" 25%-quantile of the database: classMax < $q^{0.25}$. Then we calculate the descriptor score as follows:

$$\text{score} = 2\frac{q^{0.25} - \text{dbMin}}{\text{sizeRange}}$$

The factor 2 is applied in order to account for the fact that the value range of only 25% of the database molecules is used here, whereas in (a) the value range of 50% of the molecules is used. Thus, the score in (b) is adjusted relative to that in (a).

(c) The value range of an activity class completely falls above the "high" 25%-quantile of the database: classMin > $q^{0.75}$. Then the score is calculated as

$$\text{score} = 2\frac{\text{dbMax} - q^{0.75}}{\text{sizeRange}}$$

The same factor adjustment as in (b) applies. A modification is introduced when sizeRange becomes zero, i.e., when all active compounds have the same descriptor value. In this case, we correct sizeRange by use of delta, defined as a fraction of the standard deviation of the descriptor database distribution:

$$\text{delta} = \frac{\text{stdDev}}{200}$$

By use of delta, the magnitude of the correction is made dependent on the individual standard deviation of each affected descriptor in the compound database. Applying delta, we adjust classMin and classMax as follows:

$$\text{classMin\_new} = \text{classMin} - \text{delta}$$

$$\text{classMax\_new} = \text{classMax} + \text{delta}$$

These modifications produce an extended value range [classMin_new, classMax_new]:

$$\text{sizeRange} = 2\frac{\text{stdDev}}{200} = \frac{\text{stdDev}}{100}$$

The final division by 100 was empirically chosen in order to ensure that descriptors with delta correction achieve one of the top scores within the typically observed scoring range (see below). This is justified because, for these descriptors, all active template compounds adopt exactly the same value (likely to be a class-specific setting).

2.2.3. Score Distribution and Descriptor Selection. Applying the above scoring scheme, only positive scores of 0 or greater are obtained. Descriptors in category (a) produce a score smaller than 1, if the database value range between the 25%- and 75%-quantiles is smaller than the descriptor value range. This generally means that more than 50% of the database molecules match the value range of the descriptor and implies that the descriptor does not display sensitivity to the activity class. By contrast, a descriptor score greater than 1 is usually obtained when less than half of the database molecules match the value range of an activity class, and we thus consider a score greater than 1 a minimum threshold for descriptor selection. Descriptors belonging to categories (b) and (c) produce scores greater than 2 reflecting the situation that only 25% or fewer of the database compounds match the value range of an activity class. Accordingly, these descriptors are selective with respect to the value ranges of activity classes and highly preferred in our studies. Although descriptor scores do not provide the information of how many database compounds actually match the value range of an activity class, they clearly reflect selectivity tendencies given the quantile boundaries applied here.

A typically observed score range for our set of 124 descriptors (see below) would consist of approximately 50% of the descriptors having a score smaller than 1, 30%−40% of the descriptors having a score between 1 and 2, and 10%−20% a score greater than 2. Thus, on average, about 10% of the descriptors specifically responded to the compound classes investigated here, but responsive descriptors significantly varied across these classes. Also, the score distribution among the top 10%−20% of the descriptors displayed strong activity class-dependence. For best descriptors (excluding those with highly discrete settings), we typically observed scores of 5 or greater.

For descriptor selection, we have applied relatively liberal criteria in our calculations. For the reasons discussed above, we consider a score of 1 to be a minimum threshold value for selection. For each activity class, we have selected top scoring descriptors in descending order to a score of 1, but the maximum number of allowed descriptors was set to 60 (nearly 50% of the basis set). These selection criteria cannot be generalized but proved to be very appropriate for our virtual screening trials.

2.2.4. Descriptor Value Ranges for Compound Mapping. Activity class-selective value ranges of descriptors provide the basis for mapping of database molecules to multiple ranges and compound selection. In order to further increase the ability to recognize compounds with similar activity but diverse structures, we investigated the possibility to moderately expand the exact descriptor value range (i.e., exactRange) of each descriptor, thereby effectively increasing the likelihood for database compounds to match it. In preliminary calculations, we observed a tendency to retrieve more active compounds when expanding exactRange. Therefore, exactRange was extended by adding an averaged form of variance:

$$\mathrm{dExp} = \frac{\mathrm{sizeRange}}{|\mathrm{Baits}| - 1}$$

“Baits” refers here to the set of active compounds used to determine exactRange and as templates for virtual screening trials. For compound mapping, the effective value range was set to

$$\mathrm{expandedRange} = [\mathrm{classMin} - \mathrm{dExp}, \mathrm{classMax} + \mathrm{dExp}]$$

For dExp, it is important to note that range expansion becomes smaller with increasing numbers of active template molecules. Increasingly large bait sets are expected to produce larger value ranges due to a higher probability of intraset property variations, whereas smaller bait sets are thought to benefit more from expansion, especially for the purpose of compound recovery.

2.2.5. Mapping and Scoring of Database Compounds. Mapping of database compounds to multiple activity class-selective descriptor value ranges requires a similarity metric to quantify the overlap between value ranges of bait and database compounds. Potential hits would be expected to closely match multiple descriptor settings of the bait set. In this context, M is the number of descriptors where the value calculated for a database compound falls within the value range of an activity class and D is the total number of descriptors selected for this activity class. Thus, for every database molecule, a similarity score s is calculated by dividing the number of matching descriptors by the total number of selected descriptors:

$$s = \frac{M}{D}$$

Accordingly, similarity scores between 0 and 1 can be obtained.

2.3. Calculations.

2.3.1. Activity Classes, Descriptors, and Database Compounds. Two source databases were used for our studies, a compound collection containing ~1.34 million molecules that were collected from various medicinal chemistry vendors,[21] termed background database (BGDB), and the Molecular Drug Data Report (MDDR),[22] containing approximately 160 000 entries. In our analysis, all BGDB molecules were considered inactive (and thus potential false positives), although it is conceivable that BGDB contains novel hits for the activity classes studied here. By contrast, every MDDR molecule is annotated with a certain activity. As descriptors, a previously published basis set of 124 1D, 2D, and implicit 3D descriptors was used in our calculations.[21] These descriptors represent a subset of those implemented in the Molecular Operating Environment (MOE),[23] and their values for activity classes and database compounds were calculated with MOE. The MAD approach was tested on six different activity classes that were originally assembled from the literature for partitioning analyses.[24,25] These classes consisted of 22 benzodiazepines (BEN), 17 cyclooxygenase-2 inhibitors (COX), 22 carbonic anhydrase II inhibitors (CAE), 21 serotonin receptor ligands (5HT), 21 H3 antagonists (H3E), and 20 tyrosine kinase inhibitors (TKE).[25]

2.3.2. Virtual Screening Trials and Performance Measures. For each of the activity classes, 100 sets of 10 compounds each were randomly selected for the determination of specific value ranges and as baits for virtual screening. The remaining compounds were added to BGDB as potential hits. Thus, for each activity class, 100 different search calculations were carried out, thereby limiting bias due to chance effects in the selection of baits and potential hits. In BGDB virtual screening trials, expandedRange was applied for compound mapping and selection. As performance measures, both hit rates (number of active compounds relative to selected database molecules) and recovery rates (number of selected active compounds relative to the total number of potential database hits) were calculated. The numbers of hits and false positives were determined in compound selection sets of increasing size and averaged over 100 trials for each activity class, permitting the final calculation of average hit and recovery rates for each complete experiment. Further analysis was carried out using MDDR as source database. However, in this case, all active compounds were used as baits and value range expansion was not applied to these larger compound sets. For each activity class, a single run was carried out and the distribution of activities among the top scoring MDDR compounds was analyzed.

2.3.3. Reference Calculations. In order to compare the MAD results obtained in our virtual screening calculations with other methods, we carried out 2D similarity search calculations on the six activity classes using a fingerprint consisting of 166 publicly available MACCS structural keys.[26] In these calculations, one active compound at a time was taken as the search template and the remaining active molecules were added to the source database as potential hits. Similarity search calculations were carried out for each active molecule, and database compounds were ranked according to values of the Tanimoto coefficient (Tc). In each case, hit and recovery rates were calculated for the 50 top scoring compounds and the results were averaged for each activity class. MAD and similarity search results were also compared to literature data for five of our six activity classes.

Furthermore, we assessed the degree of structural diversity within each activity class by systematic pairwise comparison of all compounds using the MACCS fingerprint. For each class,

Table 1. Score Calculation Examples[a]

descr	class value range	min of DB	$q^{0.25}$ of DB	$q^{0.75}$ of DB	max of DB	std dev of DB	category	formula	resulting score
fcharge	[0.00, 0.00]	−51.00	0.00	0.00	12.00	0.51	a		0.0
density	[0.68, 0.90]	0.43	0.73	0.83	2.53	0.09	a	(0.83 − 0.73)/(0.90 − 0.68)	0.5
weight	[254.40, 349.26]	35.82	332.49	495.94	994.98	117.15	a		1.7
radius	[6.00, 7.00]	0.00	6.00	9.00	21.00	1.84	a	(9.00 − 6.00)/(7.00 − 6.00)	3.0
b_heavy	[15.00, 24.00]	0.00	25.00	37.00	100.00	9.39	b		5.6
logP(o/w)	[0.51, 2.67]	−4.00	2.69	5.17	8.00	1.81	b	2(2.69 + 4.00)/(2.67 − 0.51)	6.2
kier1	[13.07, 17.81]	0.00	18.34	27.59	61.56	6.84	b		7.7
vdw_vol	[311.26, 401.80]	40.80	418.87	632.44	1288.51	156.10	b	2(418.87 − 40.80)/(401.80 − 311.26)	8.3
a_don	[3.00, 5.00]	0.00	0.00	2.00	12.00	1.09	c		10.0
vsa_pol	[11.37, 29.12]	0.00	0.00	0.00	162.80	7.81	c	2(162.80 − 0.00)/(29.12 − 11.37)	18.3

[a] Examples are given for the different categories (a, b, c) of descriptor score calculations for activity class H3E. "DB" stands for compound database. For clarity, only half of the actual calulations are shown. Descriptors are abbreviated according to MOE[23] implementations: fcharge, sum of formal atom charges; density, mass density; weight, molecular weight; radius, smallest vertex eccentricity in graph; b_heavy, number of heavy-heavy bonds; logP(o/w), log of octanol/water partition coefficient; kier1, first kappa shape index; vdw_vol, van der Waals volume; a_don, number of hydrogen bond donor atoms; vsa_pol, polar van der Waals surface area.

average Tc values were calculated that ranged from 0.62 for TKE to 0.68 for COX (with minimum observed Tc values of ∼0.4). These findings confirmed that the activity classes studied here did not consist of analogue series but contained structurally diverse molecules.

3. Results

3.1. Sensitivity of Descriptors. The MAD approach is based on the identification of activity class-specific descriptor value ranges. A key finding has been that descriptors with class-specific value ranges could be determined for all six compound sets studied here. In each case, significant differences in descriptor scores were observed. Table 1 reports some examples that illustrate how these scores are calculated and provide a good impression of the spread of obtained scores. Figure 1 shows the relationship between value distributions and range sizes for different descriptors to further illustrate the approach. Between three and 20 high-scoring descriptors (score ≥ 5) were identified for each class. High-scoring descriptors often include, among others, relatively simple and chemically intuitive formulations such as logP(o/w), volume descriptors, or donor atom counts (shown in Table 1 for class H3E). The activity classes produced different descriptor score distributions. This is illustrated in Figure 2 showing a comparison of activity classes 5HT and H3E (scores were calculated and averaged over 100 randomly selected bait sets). Class 5HT presents an example of a descriptor score distribution with low sensitivity (Figure 2a). Only a small subset of the descriptors (∼12%) scores greater than 2. Nevertheless, an average of three descriptors produced scores greater than 5. Overall 5HT produced the least favorable descriptor score distribution of the activity classes studied here. By contrast, the distribution of H3E reflects a high degree of descriptor sensitivity (Figure 2b). About 30% of the descriptors have scores greater than 2, high scores are spread out over the scoring range, and on average about 20 descriptors achieve scores greater than 5. Table 2 summarizes the distribution of scores over intervals and confirms the general trend to identify high-scoring descriptors that respond to compound class-specific features. It also shows that the majority of descriptors have little, if any, sensitivity. In our initial studies, we selected on average between 54 and 60 (permitted maximum) descriptors for mapping (Table 2). Mapping sets likely included some descriptors with borderline sensitivity. Including a number of descriptors falling within score interval [1,2) might be expected to increase the probability of matching database compounds and increase false-positive rates in virtual screening calculations. Therefore, the results of our virtual screening trials reported

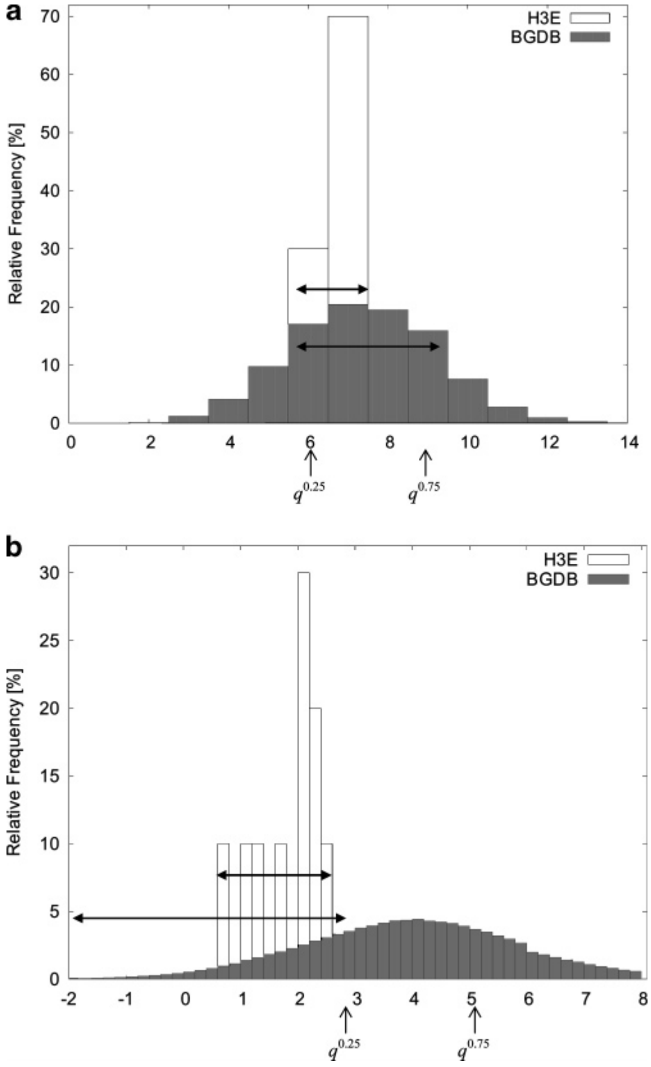

Figure 1. Relationship between value distributions and range sizes of descriptor values. In (a), the value distributions of BGDB and activity class H3E for descriptor radius (smallest vertex eccentricity) are shown, and in (b), the corresponding distributions for descriptor logP(o/w) (log of octanol/water partition coefficient). On the horizontal axis, $q^{0.25}$ and $q^{0.75}$ mark the positions of the 25%- and 75%-quantile of BGDB. Double headed arrows indicate which value ranges are used for the descriptor score calculations. Descriptor radius belongs to category "a", as defined in the text, that is, the class value range falls within the central range of BGDB. By contrast, descriptor logP(o/w) is an example for category "b" because the class value range falls below the 25%-quantile of BGDB.

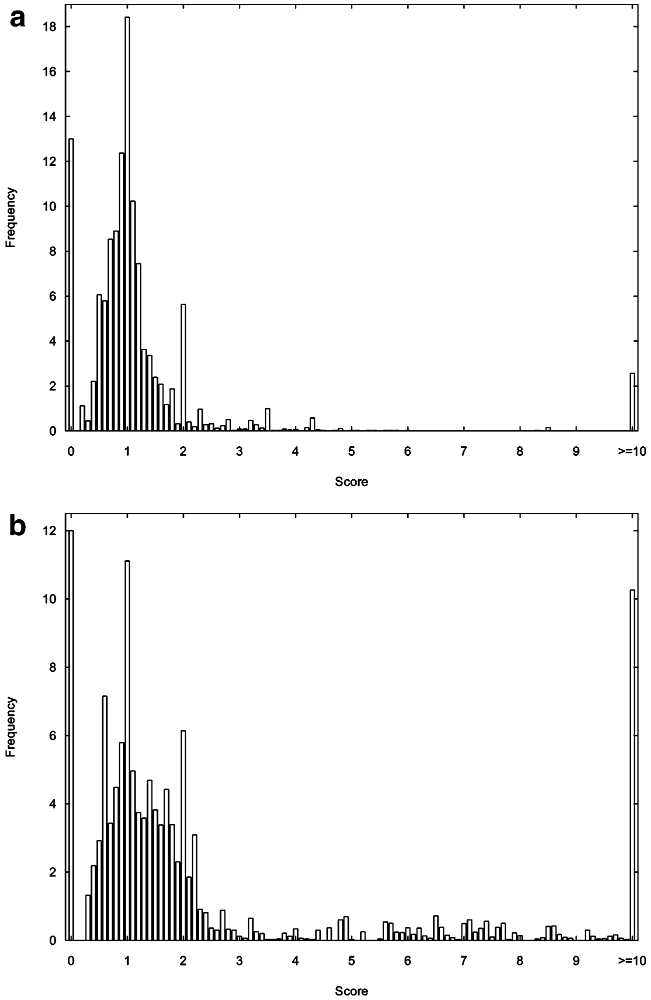

Figure 2. Descriptor score distributions. In (a), the distribution for activity class 5HT is shown, and in (b), the corresponding distribution for class H3E. Reported are averages for 100 randomly selected bait sets of 10 molecules each. The distribution for 5HT represents an example of a nonfavorable case because only relatively few descriptors achieve scores greater than 2. Thus, according to our scoring scheme, not many descriptors measurable respond to 5HT class-specific features. By contrast, the distribution for H3E represents a favorable case. Here many descriptors have scores greater than 2 that are spread out over a wide score range and the top 20 descriptor scores are greater than 5. Thus, approximately 15% of the descriptors tested here specifically respond to molecular features of class H3E.

Table 2. Distribution of Descriptor Scores for Different Activity Classes[a]

score interval	5HT	BEN	CAE	COX	H3E	TKE
[5,inf)	2.9	7.9	14.9	8.5	20.2	15.6
[4,5)	0.9	0.9	4.0	1.1	2.4	0.6
[3,4)	2.2	3.1	0.6	4.8	1.6	0.7
[2,3)	8.4	6.3	1.7	25.8	13.5	7.3
[1,2)	46.1	51.4	41.4	39.8	43.7	40.5
[0,1)	63.5	54.5	61.3	44.0	42.5	59.4
AvSelected	54	60	59	60	60	58

[a] Average numbers of descriptors falling into distinct score intervals are listed; [5,inf) means scores equal to or greater than 5. "AvSelected" reports the average number of descriptors that were used for compound mapping.

below are thought to represent a lower end performance level of MAD calculations, which is appropriate for proof-of-principle investigations.

3.2. Quality of Descriptor Score Distributions Determines Search Performance. We investigated whether the sensitivity of score distributions was directly responsible for the outcome of search calculations. Therefore, we compared results of virtual screening trials for activity classes 5HT, which produced a low-sensitivity descriptor score distribution (Figure 2a), and H3E, for which a high-sensitivity distribution was obtained (Figure 2b). A summary of these search calculations is presented in Figure 3. The results clearly show that for H3E much better hit and recovery rates were achieved than for 5HT and confirm a direct relationship between the quality of descriptor score distributions and database search results. This represents one of the key aspects of the MAD approach. Despite significant differences in descriptor sensitivity between these classes, the results obtained for 5HT were still acceptable. For example, for a selection set consisting of 50 compounds, hit and recovery rates of ~7% and 33%, respectively, were produced for 5HT (Figure 3). This suggests that inclusion of only a few class-selective descriptors is sufficient for correct recognition of at least some active compounds and that the availability of more sensitive descriptors leads to an increase in hit and recovery rates.

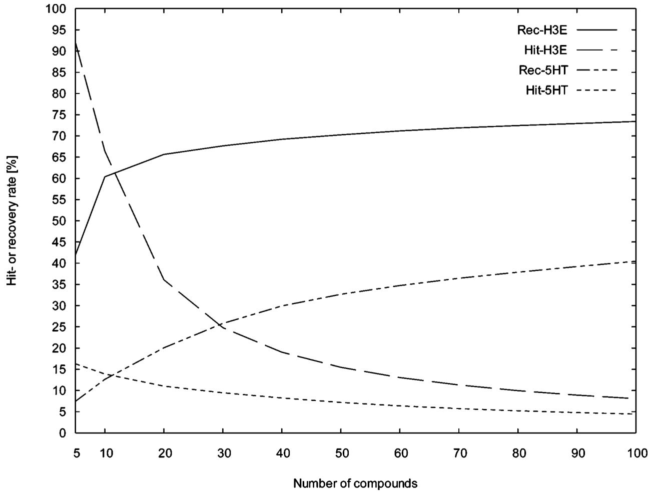

Figure 3. Representative hit and recovery rates. Shown is a comparison of hit and recovery rates for activity classes 5HT and H3E and compound selection sets of varying size. As expected, recovery rates increase with the size of selection sets whereas hit rates decrease. Differences in performance between 5HT and H3E are consistent with the descriptor score distributions shown in Figure 2: H3E has a more favorable score distribution than 5HT, and, accordingly, MAD calculations consistently produce higher hit and recovery rates for H3E.

3.3. Expansion of Descriptor Value Ranges for Mapping. We introduced expanded descriptor value ranges for compound mapping in order to increase the probability of identifying structurally diverse active compounds. This approach was tested by comparing virtual screening trials using exact and expanded value ranges. Representative results are shown in Figure 4. For example, for activity class CAE, a notable increase in recovery rates of about 5−10% was observed over compound selection sets of varying size and a slight yet consistent increase in hit rates (Figure 4a). By contrast, under equivalent calculation conditions, such effects were less obvious for class TKE where only a minor increase in average recovery rates occurred (Figure 4b). The magnitude of these effects displayed compound class dependence, but we generally observed an increase in recovery rates as a consequence of mapping to expanded descriptor value ranges, while hit rates remained largely unaffected. However, mapping to expanded value ranges did not lower hit or recovery rates, indicating that moderate expansion of value ranges did not increase net false-positive rates. Thus, we used expanded value ranges for virtual screening calculations.

3.4. Class-Specific Search Performance. Results of systematic virtual screening trials over all activity classes are

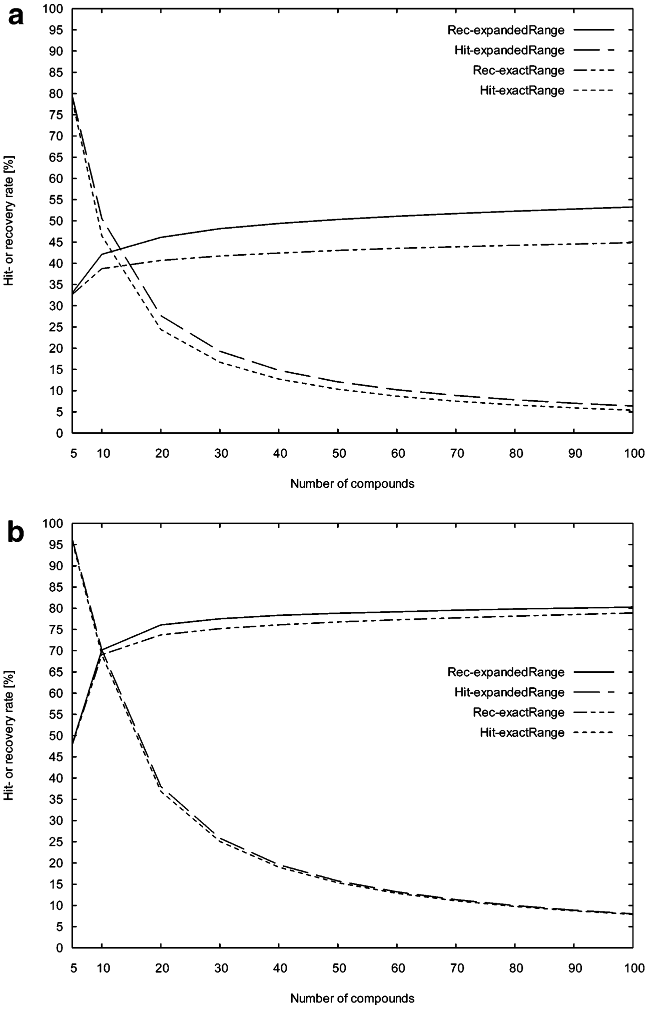

Figure 4. Performance of alternative mapping functions. Results of MAD calculations are shown for two activity classes, (a) CAE and (b) TKE, when different descriptor value ranges are used for mapping of database compounds. For each class, a comparison of hit and recovery rates is shown for compound selection sets of increasing size and mapping to either exactRange or expandedRange, as detailed in the Methodology section. For example, "Rec-expandedRange" reports recovery rates for mapping of database compounds to expandedRange and "Hit-exactRange" hit rates for mapping to exactRange. For CAE, both hit and recovery rates further improve under expandedRange conditions, whereas no significant differences between exactRange and expandedRange are seen for TKE.

reported in Table 3. The trials provided a challenging benchmark scenario for MAD calculations, since only between seven and 12 known active compounds were available as potential hits within more than 1.3 million BGDB molecules that were considered potential false positives. For calculation of hit and recovery rates, compound selection sets were varied in size between five and 100 compounds. Taking the mid-size set of 50 compounds as a reference point, classes 5HT (as discussed) and BEN showed lowest hit and recovery rates of ∼7% and 30−33%, respectively. For class CAE, a hit rate of 12% was achieved and a recovery rate of 50%, while COX gave hit and recovery rates of about 10% and 68%, respectively. Best results were obtained for H3E and TKE that produced hit rates of 15−16% and recovery rates of 70% (H3E) and 79% (TKE). A general trend was the good performance of MAD when small compound sets were selected. For example, reducing selection set size from 50 to 30 compounds only slightly reduced recovery rates but hit rates increased by up to 10%. Table 4 summarizes an "extreme" case of selecting only 10 database compounds, which roughly corresponds to the number of potential hits available in BGDB. On average, between 1.4 and 7 hits were selected within the top scoring 10 compounds. For four of our six classes (except 5HT and BEN), recovery rates were not much affected compared to larger selection sets, but in all cases, hit rates further improved. In three of six cases, hit rates in part significantly exceeded the 50% level.

3.5. Comparison with Other Methods. The rather promising results achieved in these test calculations were compared to those with other methods to put MAD performance into perspective. Five of our six activity classes have previously been studied using Recursive Median Partitioning (RMP)[27] and also another mapping algorithm, Dynamic Mapping of Consensus Positions (DMC),[28] which is discussed in more detail below. The availability of these studies permitted comparison of MAD with literature data. In addition, we have carried out for all six compound sets 2D similarity search calculations using a structural fragment-type fingerprint (2D-FP), which is a widely accepted and intuitive similarity search approach.[2,6] Table 5 reports the comparison of these different approaches. For MAD calculations, the smallest number of potential database hits was available (and the smallest compound sets were selected). Nevertheless, MAD produced overall best results. For example, it gave consistently higher recovery rates than 2D-FP and significantly higher hit rates in four of six cases. Using relatively small selection sets comparable to those of the other methods, the limiting factor of 2D-FP search calculations were low recovery rates. Hit rates of MAD and DMC were comparable, but MAD produced significantly higher recovery rates for three classes. Both methods performed better than RMP.

3.6. Screening Bioactive Compounds. As an additional test, MAD was applied to screen the MDDR database that exclusively contains active compounds. Table 6 summarizes the results. Consistently high hit rates ranging from about 16% to 79% were also achieved in these calculations. For four of six classes, close to or more than half of the compounds of the selection set belong to the same activity class. Hit rates obtained here cannot be directly compared to those of virtual screening trials due to the significant differences in calculation conditions and database composition. For example, MDDR contains many serotonin receptor ligands and, consequently, we noted a significant relative increase in hit rate for 5HT. In addition, Table 6 lists a number of compounds from the selection sets that are structurally similar to baits and have either similar or different activity annotations. Figure 5 shows some examples for activity classes BEN and H3E. These compounds represent interesting molecular similarity relationships that merit further analyses. Figure 6 shows another molecular similarity relationship detected with MAD. For 5HT, compounds belonging to two structural series were identified with either distinct or overlapping dopamine or serotonin antagonist activities.

4. Discussion

4.1. Behavior of the Descriptor Scoring and Selection Scheme. The determination of descriptors whose value ranges for compounds with similar activity significantly depart from those of many database molecules makes it possible to exploit these descriptor settings as a "signature" of activity. We found that, on average, about 10% of the descriptors analyzed here displayed a detectable tendency, albeit at substantially varying levels, to respond to compound class-specific features. Thus, combinations of descriptors with some potential for compound class selectivity were chosen as a basis for compound mapping and proved to be effective virtual screening tools for the activity

Table 3. Hit and Recovery Rates for Six Activity Classes[a]

no. of compounds	recovered ADCs	rec rate, %	hit rate, %	similarity score	no. of compounds	recovered ADCs	rec rate, %	hit rate, %	similarity score
(a) 5HT[b]									
5	0.8	7.4	16.3	0.999	60	3.8	34.7	6.4	0.980
10	1.4	12.6	13.9	0.995	70	4.0	36.5	5.7	0.979
20	2.2	20.1	11.0	0.993	80	4.2	37.9	5.2	0.977
30	2.8	25.8	9.5	0.988	90	4.3	39.2	4.8	0.976
40	3.3	29.9	8.2	0.985	100	4.5	40.5	4.5	0.974
50	3.6	32.7	7.2	0.981					
(b) BEN[c]									
5	1.4	11.4	27.3	0.996	60	3.8	31.5	6.3	0.978
10	2.0	16.6	19.9	0.991	70	4.0	33.1	5.7	0.977
20	2.7	22.4	13.5	0.986	80	4.1	34.5	5.2	0.976
30	3.1	25.7	10.3	0.983	90	4.3	35.9	4.8	0.975
40	3.4	28.0	8.4	0.981	100	4.5	37.3	4.5	0.974
50	3.6	29.8	7.2	0.979					
(c) CAE[d]									
5	4.0	33.1	79.4	0.973	60	6.1	51.1	10.2	0.937
10	5.1	42.1	50.5	0.957	70	6.2	51.7	8.9	0.935
20	5.5	46.1	27.7	0.948	80	6.3	52.3	7.8	0.934
30	5.8	48.2	19.3	0.943	90	6.3	52.8	7.0	0.932
40	5.9	49.4	14.8	0.941	100	6.4	53.3	6.4	0.932
50	6.0	50.3	12.1	0.939					
(d) COX[e]									
5	3.9	55.6	77.8	0.955	60	4.8	69.1	8.1	0.895
10	4.3	61.4	43.0	0.932	70	4.9	69.8	7.0	0.893
20	4.5	63.7	22.3	0.920	80	4.9	70.2	6.1	0.889
30	4.6	65.4	15.3	0.911	90	4.9	70.6	5.5	0.886
40	4.7	66.8	11.7	0.905	100	5.0	71.1	5.0	0.883
50	4.8	68.2	9.6	0.899					
(e) H3E[f]									
5	4.6	41.9	92.1	0.980	60	7.8	71.2	13.1	0.910
10	6.6	60.4	66.4	0.948	70	7.9	71.9	11.3	0.907
20	7.2	65.6	36.1	0.929	80	8.0	72.5	10.0	0.905
30	7.4	67.7	24.8	0.923	90	8.0	72.9	8.9	0.903
40	7.6	69.2	19.0	0.917	100	8.1	73.4	8.1	0.901
50	7.7	70.3	15.5	0.914					
(f) TKE[g]									
5	4.8	48.1	96.2	0.964	60	7.9	79.2	13.2	0.861
10	7.0	70.2	70.2	0.902	70	8.0	79.5	11.4	0.858
20	7.6	76.1	38.0	0.881	80	8.0	79.8	10.0	0.855
30	7.8	77.5	25.8	0.874	90	8.0	80.1	8.9	0.852
40	7.8	78.4	19.6	0.867	100	8.0	80.3	8.0	0.851
50	7.9	78.8	15.8	0.865					

[a] "ADC" stands for active database compounds and reports the total number of potential hits for each activity class ("hidden" in BGDB). "No. of compounds" refers to the size of selection sets. For example, in the case of TKE, 7 hits were identified on average (over 100 trials) when the 10 compounds most similar to the baits were selected from BGDB. "Rec rate" stands for recovery rate, and "similarity score" reports the similarity cutoff value for each selection set. [b] Number of active compounds: 21. Number of baits: 10. Number of ADCs: 11. [c] Number of active compounds: 22. Number of baits: 10. Number of ADCs: 12. [d] Number of active compounds: 22. Number of baits: 10. Number of ADCs: 12. [e] Number of active compounds: 17. Number of baits: 10. Number of ADCs: 7. [f] Number of active compounds: 21. Number of baits: 10. Number of ADCs: 11. [g] Number of active compounds: 20. Number of baits: 10. Number of ADCs: 10.

Table 4. MAD Recovery and Hit Rates for the Top 10 Compounds[a]

activity class	no. of compounds	recovered ADCs	rec rate, %	hit rate, %
5HT	10	1.4	12.6	13.9
BEN	10	2.0	16.6	19.9
CAE	10	5.1	42.1	50.5
COX	10	4.3	61.4	43.0
H3E	10	6.6	60.4	66.4
TKE	10	7.0	70.2	70.2

[a] Summary of recovery and hit rates when selecting only the 10 top scoring compounds from BGDB (averages over 100 trials). Abbreviations are used as reported in footnote *a* of Table 3.

classes we analyzed. In fact, the virtual screening results reported herein were obtained under liberal descriptor selection criteria and should represent a low-end performance range of MAD analysis, since false-positive rates could likely even be further reduced by eliminating some of the descriptors. However, MAD already has low false-positive rates, as it performs well when small compound sets are selected. The high hit and recovery rates achieved here also suggest that it is not required to identify large numbers of highly activity-selective descriptors for given compound classes, as long as it is possible to use combinations of descriptors displaying at least some sensitivity. These assumptions are well in accord with previous observations that combinations of simple or binary-transformed descriptors can be highly discriminatory in database partitioning[27] and the selection of active compounds.[28] Regardless of whether soft or hard descriptor scoring schemes are ultimately applied, our results show that descriptors with compound class-specific preferences can be identified and that combinations of such descriptors have significant predictive ability in compound mapping and virtual screening. The observation that expansion of activity-sensitive descriptor value ranges for mapping of database compounds revealed a general tendency to improve recovery rates further supports the view that combinations of activity-sensitive descriptors are highly discriminatory in the recognition of active compounds.

Table 5. Reference Calculations[a]

activity class	method	ADC	no. of compounds	rec rate, %	hit rate, %
5HT	MAD	11	10	13	14
	2D-FP	20	50	15	6
	DMC	61	155	23	9
	RMP	61	75	22	21
BEN	MAD	12	10	17	20
	2D-FP	21	50	22	9
	DMC	49	17	9	26
	RMP	49	83	18	12
CAE	MAD	12	10	42	51
	2D-FP	21	50	22	9
	DMC				
	RMP				
COX	MAD	7	10	61	43
	2D-FP	16	50	15	5
	DMC	21	18	18	50
	RMP	21	72	11	3
H3E	MAD	21	10	60	66
	2D-FP	20	50	33	13
	DMC	42	24	29	74
	RMP	42	61	5	3
TKE	MAD	20	10	70	70
	2D-FP	19	50	16	6
	DMC	25	25	9	26
	RMP	25	74	40	13

[a] All search calculations were carried out in BGDB. RMP and DMC results were taken from refs 27 and 28. For MAD, RMP, and DMC the size of the bait set was always 10 compounds. "No. of compounds" reports the sizes of selection sets. Depending on the method, selection set sizes vary.

Table 6. Recognition of Bioactive Compounds[a]

activity class	no. of baits	no. of selected compounds	no. of hits	hit rate, %	structurally related compounds with diverse activities
5HT	21	74	32	43.2	11 [7 DDA, 4 ATP]
BEN	22	77	12	15.6	15 [9 P4I, 2 5HT, 2 RTI, 2 COX]
CAE	22	139	51	36.7	14 [8 COX, 6 ESI]
COX	17	90	71	78.9	2 [2 ENA]
H3E	21	100	24	24.0	17 [9 5HT, 3 DAU, 3 MEL, 2 PAF]
TKE	20	72	40	55.6	5 [3 PKC, 2 DRI]

[a] The table summarizes MAD screening of the MDDR database. Hits are MDDR compounds having the same activity as the baits. Also reported for each class are numbers of selected compounds with structural similarity to bait molecules and related or different activities. On the basis of the observed similarity score distributions and score differences among compounds in the top 100 list, fewer than 100 compounds were selected for four classes. In the case of CAE, more than 100 MDDR molecules were selected because compounds at positions 100−139 had the same score. MDDR activity class abbreviations: 5HT, serotonin receptor ligands; ATP, H+/K+-ATPase inhibitors; COX, cyclooxygenase inhibitors; DAU, dopamine autoreceptor agonists; DDA, dopamine (D1, D2, D4) antagonists; DRI, dihydrofolate reductase inhibitors; ENA, endothelin antagonists; ESI, estrone sulfatase inhibitors; MEL, melatonin agonists; P4I, phosphodiesterase IV inhibitors; PAF, platelet-activating factor receptor antagonists; PKC, protein kinase C inhibitors; RTI, reverse transcriptase inhibitors.

4.2. Activity-Related Property Discrepancies. The quantitative treatment of activity-selective descriptor values provides the basis of MAD and goes beyond our information-theoretic studies of systematic descriptor and property differences between compound databases[18−20] and also other investigations that have analyzed such differences in screening data sets.[29] Results of our descriptor value analysis are consistent with previous observations that different descriptors contribute very differently to the accurate classification of active compounds[24,25] or that some molecular properties of different activity classes have statistically different value distributions.[30] Thus, our studies represent an extension of such concepts for direct application in molecular similarity analysis and virtual screening.

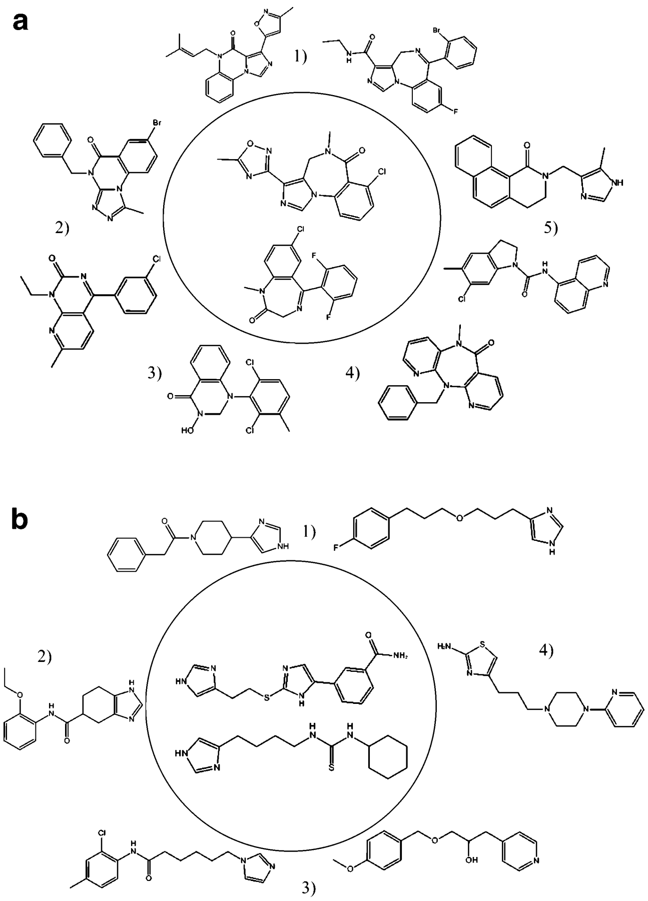

Figure 5. Bioactive compounds detected with MAD. The figure shows a spectrum of bioactive molecules identified in a virtual screen of the MDDR for (a) BEN and (b) H3E (see also Table 6). Two representative bait compounds are shown in the center (circled), and some example compounds from the selection sets are arranged around them and numbered in counterclockwise direction. For BEN, molecules occurring within the top 77 compound selection set include (1) benzodiazepines (same class), (2) phosphodiesterase IV inhibitors (one of which shares the benzodiazepine scaffold), (3) an inhibitor of cyclo-oxygenase and lipo-oxygenase, (4) reverse transcriptase inhibitor, and (5) serotonin receptor antagonists. For H3E, molecules within the 100 top scoring compounds include, for example, (1) histamine H3 antagonists (same class), (2) 5-HT3 antagonist, (3) PAF antagonists, and (4) dopamine receptor agonist.

4.3. Compound Scoring. The results of our virtual screening trials also demonstrate that application of a simple compound scoring function was sufficient for hit identification in all six test cases. To further refine score distributions and increase their resolution, it is possible to apply more complex scoring schemes, for example, by introducing descriptor score-dependent weighting factors in the calculation of compound scores. However, such refinements were not crucial for achieving consistently high hit and recovery rates in our studies.

4.4. Characteristic Features of MAD. As a molecular similarity and ligand-based virtual screening method, MAD is designed to recognize both close and distant molecular similarity relationships. This was taken into account when differentiating between exactRange and expandedRange for descriptor scoring and compound mapping. Dependent on the search problem under investigation, expandedRange can easily be adjusted. Furthermore, the ability of MAD to detect diverse similarity relationships was tested by mining a large number of biologically active compounds from the MDDR. For each class, a spectrum of structures was identified, including many closely but also more distantly related ones; in terms of structure and/or activity.

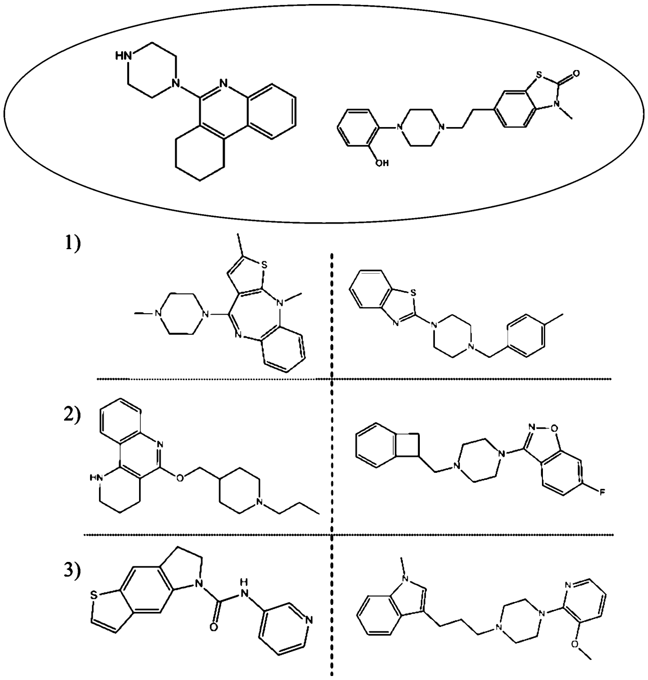

Figure 6. Compounds belonging to two distinct structural series (left and right of the vertical dashed line) with overlapping biological activities were recognized in MAD calculations using a set of 21 serotonin receptor ligands (5HT) as template molecules: (1) dopamine receptor antagonists only, (2) both 5HT and dopamine antagonists, (3) 5HT only. For each activity type, two compounds are shown, each representing one of the two structural series.

These findings also very well illustrate some of the fundamental ideas and challenges of molecular similarity analysis.[4,6]

MAD operates in chemical reference spaces constituted by descriptors with continuous value ranges and not in spaces that utilize binary descriptor formulations.[10,27] The ability to easily adjust continuous descriptor value ranges for mapping makes MAD in principle less sensitive to boundary effects than, for example, cell-based partitioning algorithms.[8,11] As a mapping technique, MAD is computationally very efficient and readily applicable to the analysis of very large compound databases. The only time-limiting factor is the calculation of descriptor values for source compounds, which needs to be done only once, given a descriptor pool and compound database.

4.5. Related Methods. As a mapping algorithm, MAD is conceptually more similar to cell-based partitioning approaches than decision tree methods or clustering techniques.[6] Decision tree methods rely on the generation of sequential and predominantly binary descriptor pathways and on compound learning sets, while clustering approaches, despite algorithmic diversity, ultimately depend on pairwise compound distance or similarity comparisons.[8] The conceptual resemblance of mapping and cell-based partitioning is limited to the fact that both approaches utilize independent descriptor “coordinates” to position compounds in reference spaces. Among cell-based partitioning techniques, the “receptor-relevant subspace concept”[16] is somewhat related to the basic idea behind MAD. The former approach attempts to identify descriptor axes in BCUT spaces[15] around which compounds with a specific activity concentrate.[16] However, apart from this analogy, from an algorithmic point of view, cell-based partitioning and MAD are completely distinct. Moreover, cell-based partitioning aims at generating low-dimensional and orthogonal descriptor spaces, whereas MAD lacks this requirement and, in fact, operates in high-dimensional spaces.

The approach perhaps most similar to MAD is DMC that was previously developed in our laboratory.[28] DMC is also a mapping technique but relies on finding consensus positions for sets of active compounds in binary descriptor spaces of stepwise increasing dimensionality.[28] Binary transformation of molecular property descriptors for DMC is achieved based on calculation of statistical medians for descriptor distributions in compound source databases, underlying the median partitioning approach.[21] Thus, in addition to differences in the way compounds are mapped, DMC does not utilize activity-dependent descriptor value ranges, which is a key feature of MAD. In addition, DMC is much more restricted than MAD in selecting compound sets of a predefined size. This is due to the fact that in DMC the number of compounds mapping to consensus positions of baits is critically determined by dimension extension levels and is often greatly reduced when proceeding to the next level.[28] Similarity selection in DMC is ultimately a binary (yes/no) decision. By contrast, MAD uses a continuous similarity score and creates compound rankings that easily allow for selection of variably sized sets.

4.6. MAD Performance. Without exception, MAD calculations produced meaningful results in the virtual screening trials that were carried out to benchmark the approach. These calculations provided a rather challenging test situation due to the presence of only a few potential hits in a large background database. Although we need to consider the generally strong compound class dependence of the performance of many virtual screening methods,[17] which makes comparisons often difficult, hit and recovery rates obtained with MAD are similar to or better than those reported for application of other ligand-based methods,[6,17] including DMC.[28] This was confirmed by direct comparison of MAD with 2D similarity search calculations for all six activity classes studied here and RMP and DMC results for five of these classes.

4.7. Implications for Chemical Space Design. The design of descriptor spaces for compound classification, diversity and similarity analysis, or virtual screening continues to be a major topic in chemoinformatics research.[2,31] Universally applicable chemical space representations that are suitable for many different applications are yet to be developed, if they exist at all.[2] Accordingly, much emphasis is put on the development of application-dependent reference spaces. This means one typically attempts to identify the “best” descriptors for each application. While it is generally thought that low-dimensional space representations are preferred for many applications[2] including partitioning[15] or diversity design,[31] mapping algorithms such as DMC or MAD depart from this theme because they operate in high-dimensional descriptor spaces. This is not a unique feature. For example, support vector machines, a machine learning approach, project compound data sets into high-dimensional space representations through the use of kernel functions to facilitate compound classification.[32,33] By contrast, MAD descriptor spaces are conceptually simple. They are high-dimensional but composed of “short” descriptor axes (small value ranges) that are specifically derived for an activity class. The results of our analysis suggest that it is readily possible to generate such activity-oriented descriptor reference spaces for molecular similarity analysis without application of machine learning techniques for descriptor selection.

4.8. Conclusions. We have reported the development and evaluation of a new method that is based on the identification of multiple activity-dependent descriptor value ranges and mapping of test compounds to these ranges. Systematic analysis is facilitated by introduction of specifically designed descriptor scoring and molecular similarity functions. As a mapping algorithm that utilizes property descriptor value ranges, MAD adds to the spectrum of previously developed molecular similarity and virtual screening methods. Further improvements of the MAD approach are conceivable by modifying descriptor and similarity scoring functions. In its current implementation, MAD produces promising results on the test cases studied here, under challenging virtual screening conditions. It is worth noting that only relatively few (∼10) active compounds were used for descriptor scoring and virtual screening. In our calculations, the MAD approach displayed the tendency to produce generally high compound recovery rates. For selection sets of small size (e.g., 50 or fewer compounds), MAD calculations produced increasingly high hit rates while largely retaining recovery rates, indicating high sensitivity of the approach. We have observed the same tendency in MAD calculations on diverse compound sets beyond the classes reported here. Taken together, our initial findings suggest that MAD and related compound mapping techniques should merit further investigations and large-scale evaluations.

Acknowledgment. The authors thank Jeffrey W. Godden for helpful discussions and critical review of the manuscript.

References

(1) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(2) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.

(3) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure−activity relationship paradigm. *Methods Mol. Biol.* **2004**, *275*, 131−214.

(4) Johnson, M., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.

(5) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189−1202.

(6) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.

(7) 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.

(8) Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* **2003**, *8*, 707−715.

(9) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(10) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(11) Xue, L.; Bajorath, J. Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.

(12) Agrafiotis, D. K.; Lobanov, V. S. Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1356−1362.

(13) Agrafiotis, D. K.; Lobanov, V. S. Multi-dimensional scaling of combinatorial libraries without explicit enumeration. *J. Comput. Chem.* **2001**, *22*, 1712−1722.

(14) Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.* **1994**, *37*, 3758−3767.

(15) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.

(16) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(17) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(18) 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.

(19) Godden, J. W.; Bajorath, J. Differential Shannon entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060−1066.

(20) Godden, J. W.; Bajorath, J. Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 87−93.

(21) Godden J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Schermerhorn, E. J.; Bajorath, J. Median partitioning: a novel method for the selection of representative subsets from large compound pools. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 885−893.

(22) *Molecular Drug Data Report (MDDR)*; MDL Information Systems Inc.: San Leandro, CA, 2005.

(23) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada, H3B 3X3, 2005.

(24) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 699−704.

(25) Xue, L.; Bajorath, J. Accurate partitioning of compounds belonging to diverse activity classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757−764.

(26) *MACCS structural keys*; MDL Information Systems Inc.: San Leandro, CA, 2002.

(27) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive median partitioning for virtual screening of large databases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182−188.

(28) Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. Molecular similarity analysis and virtual screening in binary-transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 21−29.

(29) Diller, D. J.; Hobbs, D. W. Deriving knowledge through data mining high-throughput screening data. *J. Med. Chem.* **2004**, *47*, 6373−6383.

(30) Gribbon, P.; Sewing, A. High-throughput drug discovery: what can we expect from hits? *Drug Discovery Today* **2005**, *10*, 17−24.

(31) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, R. F. Combinatorial informatics in the post-genomics era. *Nat. Drug Discovery Rev.* **2002**, *1*, 337−346.

(32) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667−673.

(33) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *44*, 549−561.

JM051110P