Recursive Median Partitioning for Virtual Screening of Large Databases

Jeffrey W. Godden,[†,‡] John R. Furr,[†] and Jürgen Bajorath*[,†,‡,§]

Department of Computer-Aided Drug Discovery, Albany Molecular Research, Inc. (AMRI), 21 Corporate Circle, Albany, New York 12212-5098, AMRI Bothell Research Center (AMRI-BRC), 18804 North Creek Parkway, Bothell, Washington 98011, and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received September 27, 2002

Recently, we have introduced the median partitioning (MP) method for diversity selection and compound classification. The MP approach utilizes property descriptors with continuous value ranges, transforms these descriptors into a binary classification scheme by determining their medians in source databases, and divides database molecules in subsequent steps into populations above or below these medians. Having previously demonstrated the usefulness of MP for the classification of molecules according to biological activity, we have now gone a step further and extended the methodology for application in virtual screening. In these calculations, a series of bait molecules having desired activity is added to large compound databases, and subsequent iterations or recursions are carried out to reduce the number of candidate molecules until a small number of compounds are found in partitions enriched with bait molecules. For each recursion step, descriptor combinations are identified that copartition as many active molecules as possible. Descriptor selection is facilitated by application of a genetic algorithm (GA). The recursive MP approach (RMP) has been applied to five diverse biological activity classes in virtual screening of a database consisting of approximately 1.34 million molecules to which different types of active compounds were added. RMP analysis produced hit rates of up to 21%, dependent on the biological activity class, and led to an average ∼3600-fold improvement over random selection for the activity classes that were used as test cases.

INTRODUCTION

Partitioning methods have become important components of the virtual screening repertoire, with recursive partitioning (RP)[1−4] and cell-based methods[5−8] being especially popular at present. Similar to other molecular diversity or compound classification methods based on single value decomposition,[9,10] multidimensional scaling,[11] or nonlinear mapping,[12] cell-based partitioning relies on dimension reduction of predefined[6] or randomly generated[7] chemical descriptor spaces. By contrast, RP divides molecular data sets along decision trees and identifies combinations of binary molecular descriptors that ultimately copartition compounds in learning sets that share the same activity.[1,2] So derived descriptor trees and rules can then be used to search databases for similar compounds. A conceptually related iterative classification approach, although distinct in its algorithmic details, is the recently introduced multidomain clustering method[13] that combines a cluster algorithm with elements of neural network and genetic algorithm simulations and is also capable of extracting structure−activity relationships and rules from molecular learning sets.

Similar to RP, the MP method is an iterative statistically based approach for compound classification. However, different from RP, it does not utilize a decision tree structure to subdivide molecular data sets (according to the presence or absence of specific descriptors or chemical features) and also does not depend on processing of learning sets for the derivation of rules and predictive models. In its original implementation, MP uses one property descriptor per partitioning step to iteratively divide the complete population of molecules into subpopulations above and below the median value of each selected descriptor.[14] The median is defined as the value within a distribution that divides the population into two equal subsets above and below this value.[15] Thus, the choice of n descriptors results in a total of 2^n different partitions, each of which is characterized by a unique binary partition code. Through the use of median values for compound classification, MP effectively transforms value distributions of molecular property descriptors in compound databases into a binary classification scheme ("above" and "below"). However, binary descriptors such as structural fragments (that are either "present" or "absent") cannot be used for MP because median values cannot be calculated for these descriptors.

Originally, the MP approach was designed for the efficient selection of diverse or representative subsets from large compound data sets.[14] In this context, the key feature of MP is that it does not depend on pairwise molecular comparisons, different from other dissimilarity-based methods,[16,17] and can therefore be applied to very large compound collections. For automatic descriptor selection, MP was coupled to a GA (GA-MP). Subsequently, we have investigated the performance of MP in the classification of compounds according to biological activity, a prerequisite for application in virtual screening. We considered this an interesting test case because

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jurgen.bajorath@albmolecular.com. Address correspondence at AMRI-BRC.

† Albany Molecular Research, Inc. (AMRI).

‡ AMRI Bothell Research Center (AMRI-BRC).

§ University of Washington.

10.1021/ci0203848 CCC: $25.00
Published on Web 01/08/2003

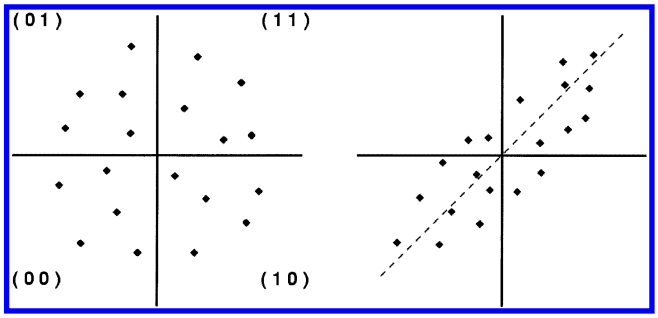

Figure 1. Median partitioning and descriptor correlation. The schematic representations illustrate the concept of MP. As an example, a two-dimensional descriptor space is shown. In the left graph, the axes represent the median values of two descriptors dividing the compound set. Each partition has a unique binary signature. In the right graph, the dashed axis represents a diagonal of correlation for medians of two strongly correlated descriptors, which results in a skewed compound distribution along the diagonal. As a consequence of descriptor correlation, the resulting partitions do not contain equal subpopulations but are either under- or overpopulated.

different from cell-based methods that partition compounds in low-dimensional chemical spaces, MP is by design a multidimensional and direct method. In our initial calculations, a total of 317 molecules belonging to 21 different activity classes were classified in the presence or absence of 2000 randomly selected ACD[18] molecules. In this study, a prediction accuracy of, on average, 73% was achieved[19] and MP performed overall slightly better than cell-based partitioning based on principal component analysis.[8,19] Thus, we concluded that MP was suitable for both diversity selection and activity-based compound classification.

These findings have encouraged us to investigate the MP approach for virtual screening. Our focal point has been how to best adapt MP for efficient screening of very large compound collections, if series of known hits with desired activity are available as a starting point. Herein we present an extension of the MP methodology that processes databases in a recursive manner, thereby iteratively reducing the number of compounds under consideration and increasing their relative similarity to active template molecules. The RMP approach has been successfully applied to search for active compounds belonging to five diverse biological activity classes, yielding a very significant improvement over random compound selection. The results suggest that RMP is a promising and computationally efficient method for virtual screening.

MATERIALS AND METHODS

The MP Concept and Suitable Molecular Descriptors. The basic idea behind median partitioning is illustrated in Figure 1. If the median is calculated for the value distribution of a given property descriptor in a database of molecules, the population of molecules is divided, by definition, into subpopulations with values above or below the median. By designating values above the median as 1 and below the median as 0, we achieve a binary classification scheme. In independent and subsequent steps, this process is repeated for the entire molecular population, whereby n steps (utilizing n distinct descriptors) produce a total of 2^n unique partitions. Molecules that fall into the same partition are considered to be similar at the level of these calculations. For diversity analysis, representative compounds can be selected from each partition. By contrast, in compound classification, one would like to copartition compounds with similar activity and differentiate them from others. Figure 1 also illustrates the influence of descriptor correlation effects on MP. Almost all molecular descriptors studied by us to date display at least some and frequently very significant pairwise correlation,[14] which can be attributed to the fact that there are very few, if any, molecular properties that are completely independent of each other. Correlation between descriptors used for MP leads to the presence of over- and underpopulated or even empty partitions. This effect is not desired for diversity analysis and, consequently, descriptors with minimal database correlation are selected for this purpose.[14] On the other hand, for classification of active compounds, we have found that descriptor correlation effects are irrelevant or, in some cases, even favorable for achieving high prediction accuracy.[19] This is probably due to the fact that recognizing similar compounds and distinguishing them from others can be achieved by consecutive evaluation of related yet distinct molecular features (as captured by correlated descriptors). Thus, descriptor correlation effects should not be deliberately minimized when searching for molecules with similar activity.[19] However, regardless of descriptor correlation effects, the MP concept implies that molecular descriptors are generally preferred that have high information content[20] in the database under investigation because broad distributions of diverse values favor the calculation of meaningful medians.[19]

For our current analysis, a previously described set consisting of 127 diverse 1D and 2D molecular descriptors was used.[8,14] Descriptor values were calculated with MOE.[21] All other calculations were carried out with Perl programs written by the authors. Initially, descriptor value distributions in our test database (see below) were surveyed, and a small number of descriptors with essentially no information content were eliminated.

The Recursive MP Extension and Its GA-RMP Implementation. Conceptually similar to RP,[1,2] the RMP approach relies on two principles, copartitioning of active compounds and enrichment of partitions with actives. In both approaches, copartitioning is achieved by identifying descriptor settings that group compounds with similar activity together. By contrast, enrichment of partitions with actives in RMP is achieved by subset selection and repartitioning and not by consecutive division of a data set, like in RP. The RMP approach is summarized in Figure 2. Following the initial partitioning calculation, all database molecules in partitions containing at least two active (successfully copartitioned) or "bait" compounds are combined as a new pool. Database molecules in partitions containing only a single bait compound are not included in the pool because, in this case, a bait was not recognized as similar to any other active compound (thus violating the copartitioning rule). This process is iteratively reinitialized by calculation of new median values for this smaller population followed by partitioning calculation until sufficiently small partitions are obtained (see Results). Since descriptor selection during each recursion is initialized de novo, all bait molecules are again added to the new compound pool, even if they were found to be singletons in the previous round. Thus, for each recursion, novel partition bit strings are assigned to each

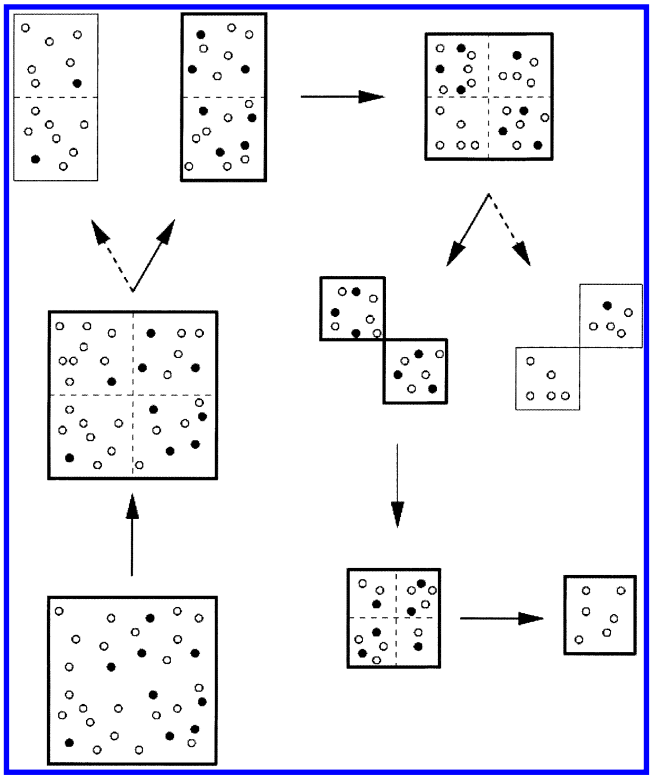

Figure 2. Recursive median partitioning. Shown is a schematic outline of the RMP approach, as discussed in the text, involving only two descriptors and a number of compounds (dots) dispersed in two-dimensional chemical space. Filled dots indicate bait molecules and open dots database compounds. Calculated median values are depicted as thin dashed lines. Dashed arrows point at subsets of database compounds (but not baits) that are eliminated during RMP analysis, based on the absence of copartitioning events. The initial partitioning step (lower left) is followed by two recursions, and the final result (lower right) is a small set of database compounds that are most similar to copartitioned baits. In practice, these molecules would represent candidates for testing.

compound in the pool, including all known actives, independent of the previous subset selection. Thus, the molecules are redispersed in a more narrowly defined chemical space whereby novel descriptor combinations are not biased by previous steps. In addition, the calculations become more focused as the total number of compounds decreases.

For GA-based descriptor selection, each of ~100 bits of the chromosome determines whether a particular descriptor is included (if set on to "1") or not ("0") in the calculation of the associated fitness function.[19] The GA calculation starts with 200 randomly generated chromosomes and the top scoring 40 (20%) are subjected to crossover and mutation operations (at a 5% mutation rate). The calculations are repeated until convergence is reached, in this case, 1000 GA cycles without improving the score S. The GA-RMP fitness function is defined as

$$S = \mathrm{Act(cp)} \times \mathrm{Pa(pop)}$$

with Act(cp) being the total number of copartitioned known active compounds and Pa(pop) the total number of populated partitions. This simple fitness function directs the GA to select descriptor sets that favor copartitioning of known active compounds and, at the same time, maximally disperse the database molecules over unique partitions. This situation is thought to be optimal for obtaining a subset of database molecule most similar to the bait compounds.

Table 1. Biological Activity Classes for RMP Analysis

activity class	baits compds	active database molecules
benzodiazepine receptor ligands	10	49
serotonin receptor ligands	10	61
tyrosine kinase inhibitors	10	25
histamine H3 antagonists	10	42
cyclooxygenase-2 inhibitors	10	21

Molecular Data Sets. Partitioning calculations were carried out using an in-house source database containing ~1.34 million molecules collected from various external sources and vendor catalogs. As test cases, different classes of molecules with specific biological activity were randomly selected from our structurally diverse biological activity database, as reported previously.[8] In each case, the set of active compounds was randomly split into 10 baits and a residual set of active database (or potential "hit") molecules. These activity classes and the number of bait and hit compounds per class are reported in Table 1.

RESULTS AND DISCUSSION

Virtual Screening and RMP. Virtual screening is an attractive approach if it is possible to reduce, in a computationally efficient way, the number of candidate compounds for experimental testing to a small number, while substantially increasing the hit rate relative to random selection. The ultimate goal is to consistently identify novel hits by testing only a limited number of database compounds (for example, ~100). The size of our source database, containing more than 1 000 000 molecules, is quite representative of databases currently used for virtual or even high-throughput screening. Similar to RP, median partitioning is very fast,[14] given its operation in original descriptor space, its linear nature, and the absence of pairwise compound comparisons. Its only significant time-limiting step is the initial calculation of molecular descriptors for the source database. This is a one-time calculation since the statistical characteristics of this large database do not measurably change when adding a small number of bait molecules. For us an important question has been as to whether RMP calculations are indeed capable of identifying hits for diverse biological activities, leading to the design of our benchmark study.

Benchmark Calculations. To evaluate RMP performance, virtual screening "experiments" were simulated as follows. For each of the five activity classes investigated here, active compounds were added to our source database that was then subjected to RMP using automatically selected descriptor combinations. During each recursion, copartitioning of bait molecules was monitored, while the other active compounds (potential hits) were hidden in the database. For each class, five recursion steps were carried out, and after the final recursion the number of hits obtained in each step was determined. For each activity class, three independent RMP calculations were carried out, and the results were averaged. For each calculation, a different set of 10 bait molecules was randomly selected from the pool of active compounds. This was done in order to avoid potential bias toward specific structural subsets (or chemotypes) within an activity class that were by chance selected as (or excluded from) bait molecules.

Table 2. RMP Virtual Screening Results[a]

recursion level	database compds	bait compds	active database compds	hit rate	improvement factor
Benzodiazepine Receptor Ligands					
0	1340848	10	49	3.6e-05	
1	164423.7	8	35.7	0.00022	6.1
2	20596	7.7	24	0.0012	33.3
3	3268.7	7.3	15.7	0.0048	133.3
4	468.4	6.3	11.7	0.025	694.4
5	**73.7**	**6.3**	**8.7**	**12%**	**3333.3**
Serotonin Receptor Ligands					
0	1340860	10	61	4.6e-05	
1	172409.6	6	46.3	0.00027	5.9
2	19229	6.3	38	0.002	43.5
3	3366.7	5.7	28.7	0.0085	184.8
4	399.6	4	19.3	0.048	1043.5
5	**62**	**4.3**	**13.3**	**21%**	**4565.2**
Tyrosine Kinase Inhibitors					
0	1340824	10	25	1.9e-05	
1	205276	10	19	9.3e-05	4.9
2	24359.7	9.3	16	0.00066	34.7
3	3980.4	9.3	13.7	0.0034	178.9
4	480.3	8	12.3	0.026	1368.4
5	**74.3**	**8**	**10**	**13%**	**6842.1**
Histamine H3 Antagonists					
0	1340841	10	42	3.1e-05	
1	274605.3	6.7	19	6.9e-05	2.2
2	29417.3	3	9.3	0.00032	10.3
3	3718.3	2.7	4.3	0.0012	38.7
4	536.6	2.3	3.3	0.0062	0.19
5	**59.3**	**2**	**2**	**3.4%**	**1096.8**
Cyclooxygenase-2 Inhibitors					
0	1340820	10	21	1.6e-05	
1	191183.7	7.7	15.7	8.2e-05	5.1
2	21927	7	10	0.00046	28.8
3	2866.3	7.3	8	0.0028	175.0
4	467.6	5.3	4.3	0.0092	575.0
5	**70**	**4**	**2.3**	**3.3%**	**2062.5**

[a] For each test case, three independent analyses with five recursions were carried out and the results were averaged. The final results are shown in boldface. Recursion level 0 shows the initial database composition. For each recursion, the total number of bait compounds that copartition is reported. Also shown is the total number of active compounds found among the database compounds that fall into partitions containing at least two bait molecules. Hit rate is calculated by dividing the number of active molecules (excluding baits) by the total number of compounds in these partitions. For recursion level 0, hit rate reports the fraction of active molecules (excluding baits) in the database. Improvement factor over random compound selection is calculated by dividing the hit rate by the fraction of active molecules (recursion level 0).

RMP Performance. The results obtained in our test calculations are summarized in Table 2. As can be seen, the virtual screening exercise was successful in each case. The number of residual database compounds was reduced by about an order of magnitude during each recursion and, after five recursions, less than 100 database compounds needed to be selected in order to identify several hits. As one would expect, the obtained results are activity class-dependent, with the H3 antagonists displaying the overall lowest yield (two hits among 59 database compounds) and the serotonin receptor ligands (13 hits among 62 database compounds) the top performance (corresponding to a hit rate of 21%). On average, approximately 68 database compounds had to be selected to identify seven hits (corresponding to an average hit rate of more than 10%). The percentage of active database or potential hit molecules that were identified by RMP calculations from the 1.34 million compound source ranged from 5% (two of 42) for H3 antagonists to 40% (10 of 25) for tyrosine kinase inhibitors, with an average of 18%. Relative to random compound selection, improvement factors in hit identification were very significant. In all calculations, improvement factors increased steadily (a desired effect) and, ultimately, at least a 1000-fold improvement was observed in each case, with a maximum factor of 6842 for tyrosine kinase inhibitors. The calculations revealed a ~3600-fold average improvement over random selection.

Descriptor Distribution. We have also analyzed and compared the GA-based descriptor selections that gave the final RMP results, as reported in Table 3. Between 20 and 39 property descriptors were required to achieve the observed level of performance, and, on average, about half of these descriptors were conserved in the three RMP runs for each activity class. The distribution of descriptor categories is relatively similar for all compound classes. Prevalent is a descriptor type that we call surface property descriptors. These descriptors were designed to map various physical properties (e.g., partial atomic charges) to molecular surface segments approximated from 2D representations of molecules[22] and have very high information content.[20] Only five molecular descriptors occurred in all 15 RMP calculations. Four of these were surface property descriptors and the other one a simple bond count descriptor accounting for the fraction of rotatable bonds in a molecule (and, therefore, molecular flexibility). As one might expect, the descriptor statistics indicate that a core set of descriptors was important for recognizing molecular similarity within each activity class but that quite different descriptor combinations were required to copartition diverse sets of active molecules.

Scoring Scheme. A noteworthy finding of this study has been the effectiveness of the relatively simple RMP scoring function we implemented. Clearly, as shown in Table 2, the scoring function accomplished its major goal, the copartitioning of bait molecules during subsequent recursion steps. However, it did so by optimizing only one additional parameter, the diversity spread of the database molecules (by maximizing the number of populated partitions). These rather surprising observations suggested the investigation of the performance of this scoring scheme in more detail. Therefore, we have carried out a number of additional benchmark calculations and compared the function used here with an alternative and more complex scoring function that was specifically developed for the MP classification of different classes of active compounds.[19]

$$S = \frac{100}{N_{\text{total}}} \times \frac{1}{(N_{\text{total}} - N_{\text{p}}) + C/C_{\text{act}}}$$

Here N_{total} is the total number of active compounds in a databases, and N_{p} is the number of compounds that occur in "pure" partitions (i.e., those that only contain compounds having similar activity). The number of compounds in mixed partitions and also singletons are regarded as classification failures. C is the total number of partitions that contain active compounds (pure, mixed, or singletons), and C_{act} is the total number of different activity classes in the compound database. A scaling factor of 100 is applied to obtain top scores greater than 1. Compared to our simple RMP scoring

Table 3. Descriptor Statistics for the Final Recursions[a]

av no. of descriptors	no. of common descriptors	comm. descr. (%)	common descriptors (categorized)					
			surface property	surface area	connectivity indices	topology indices	physical property	atom/bond counts
Benzodiazepine Receptor Ligands								
29.7	19	63.9	12	2	2	2		1
Serotonin Receptor Ligands								
32.7	16	48.9	7	1	2	2	2	2
Tyrosine Kinase Inhibitors								
19.7	15	76.1	5	2	2	1	3	2
Histamine H3 Antagonists								
38.7	13	33.6	6		1	3	1	2
Cyclooxygenase-2 Inhibitors								
31.3	13	41.5	6	1	2	1		3

[a] Common descriptors consistently occurred in all three simulations per activity class.

Table 4. Comparison of Alternative Scoring Functions[a]

		$S = \text{Act(cp)} \times \text{Pa(pop)}$				$S = \frac{100}{N_{total}} \times \frac{1}{(N_{total} - N_p) + C/C_{total}}$			
activity class	N	Nc	Pac	Fp	Pocc	Nc	Pac	Fp	Pocc
angiotensin antagonists	10	10	4	0	2083	8	3	8	753
aromatase inhibitors	10	6	2	0	2153	5	2	8	753
carbonic anhydrase inhibitors	22	22	6	0	1941	18	6	0	753
ACE inhibitors	17	17	6	1	2155	15	2	9	753
estrogen antagonists	11	8	2	1	2132	5	2	0	753
glucocorticoid analogues	14	11	5	0	2199	10	1	7	753
matrix metalloproteinase inhibitors	12	12	3	0	2166	11	3	3	753
β-lactamase inhibitors	14	13	5	0	2115	8	4	0	753
protein kinase C inhibitors	15	11	3	4	2195	9	2	6	753
vitamin D analogues	12	12	2	0	2239	12	3	7	753

[a] Test calculations were carried out in a database consisting of 2317 molecules including 317 molecules belonging to 21 diverse activity classes.[8] The alternative scoring functions are described in the text. "N" is the number of molecules per activity class. "Nc" reports the number of active molecules that were successfully copartitioned and "Pac" the number of "active" partitions containing these molecules. "Fp" gives the number of compounds with different activity that occurred in "active" partitions. These compounds represent false positives (calculation errors). Similarly, active compounds that were not copartitioned ($N-$Nc) also represent calculation errors. "Pocc" reports the total number of populated partitions that were generated.

scheme, there are two major differences: the more complex scoring function was designed for simultaneous partitioning of several activity classes and to minimize the total number of "active" partitions that are generated (which means that each active partition should contain as many molecules as possible having similar activity). Consequently, high scores are obtained if many active compounds occur in a small number of pure partitions. By contrast, the RMP scoring function was designed for partitioning analysis of one activity class at a time (a typical virtual screening situation) to achieve copartitioning of bait compounds, while maximizing the overall number of partitions populated with database compounds. To analyze the performance of these scoring schemes, we have carried out single recursion partitioning calculations on 10 activity classes (different from those subjected to RMP) in a smaller test set. This test database consisted of 317 compounds belonging to 21 diverse activity classes and, in addition, 2000 randomly selected synthetic compounds, as described previously.[8] To evaluate the RMP scoring function, all compounds belonging to a specific activity class were considered bait molecules and a single recursion step was performed. For the more complex scoring function, the complete database was partitioned (using the best previously identified MP descriptor combination[19]), and resulting compound distributions were calculated individually for each of the 10 activity classes. The obtained results are summarized in Table 4. Median partitioning of this test database did not only permit a direct comparison of two alternative scoring schemes but also provided internal standards to assess the accuracy of copartitioning calculations. For example, in each case, the test database contained approximately 300 molecules with known specific activity different from the chosen activity class, and copartitioning of molecules with different activity was an obvious classification error. The data shown in Table 4 support the choice of our simple scoring function for RMP and reveal some trends that help to better understand its strong performance. First of all, both scoring functions achieved the desired copartitioning of molecules in each activity class. For the RMP scoring function, the copartitioning rate was overall slightly better, since only one activity class was partitioned per calculation, not all of them. Importantly, the simple RMP scoring function led to few, if any, false positive recognitions, whereas the more complex scoring function produced on average close to seven false positives for seven of 10 activity classes (which nevertheless is still a meaningful result). This discrepancy most likely resulted from the significantly different number of partitions that were populated. In the single partitioning analysis guided by the alternative scoring function, a total of 753 partitions were populated, whereas

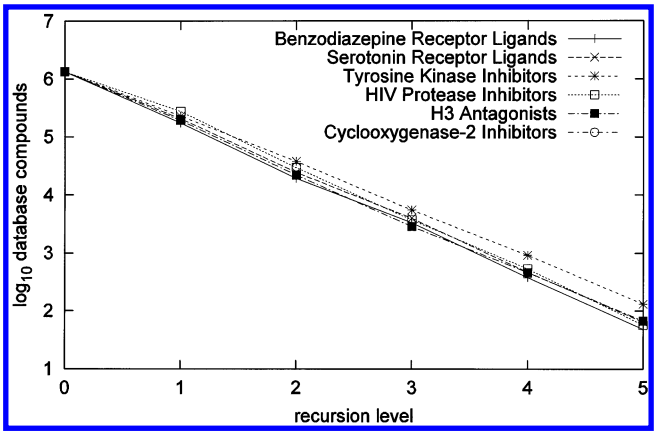

Figure 3. Reduction in the number of database compounds during RMP. For six biological activity classes subjected to RMP, the similar reduction rates of database compounds during subsequent recursion steps are displayed on a logarithmic scale. Efficient reduction in the number of database compounds that are not similar to bait molecules is an important factor for virtual screening. In each of the six cases studied here, five recursion steps were sufficient to reduce the number of potential test compounds to less than 100, while achieving hit rates between 3% and 21%.

the RMP scoring function produced on average approximately 2000 populated partitions in each of the 10 calculations. Thus, the simple scoring function selected for descriptor combinations that created more partitions with lower average population. Simply put, it created a finer "diversity grid" and placed the majority of database compounds in "nonactive" partitions, thereby also reducing false positive recognitions. This increase in diversity produced by our simple RMP scoring function was further illustrated by the larger number of "active" partitions it produced. These effects also explained the more or less monotonic reduction in the number of database compounds that was observed during each recursion in the virtual screening analysis of our large database containing 1.34 million molecules, as reported in Table 2 and illustrated Figure 3. For virtual screening of large databases by RMP, an efficient reduction in the number of background compounds per recursion step is of course highly desired. At the same time, however, compounds most similar to bait molecules must be retained, as illustrated by the significant hit rates documented in Table 2.

Comparison with Related Approaches. The results of our RMP benchmark calculations compare very favorably with other clustering or partitioning techniques with relevance for virtual screening. Although it is difficult to directly compare many of these studies, due to their different design, calculation parameters, and evaluation criteria, overall performance levels can de discussed and put into perspective. For example, using RP, a 15-fold increase in hit rate over random selection was achieved when analyzing a monoamine oxidase inhibitor set.[2] Furthermore, a retrospective analysis of a number of two-step sequential screening investigations suggested that about an 8-fold improvement over random selection could be achieved by application of a clustering or partitioning technique.[23] For comparison, the average improvement factor after the second recursion in our calculations was ~30. In addition, in different iterative computational screening studies, ~100-fold hit enrichment over random selection has been achieved,[23] which can be regarded as a current standard in the field. On the other hand, simulated sequential screening by two-stage cluster analysis on data sets from various cancer cell line assays produced hit rates of up to 40% when compounds were cherry-picked from clusters.[23] However, in these cases, the data sets contained approximately 1% actives, which significantly increases the statistical probability of identifying hits, when compared to our calculations. In our test cases, the presence of 1% active database compounds would correspond to about 10 000 potential hits per activity class and not between 20 and 60. Thus, the present study is much more similar to a "needles in haystacks" scenario, and, from this point of view, the performance level achieved in these calculations is considered encouraging. Additional approaches with potential relevance for virtual screening have recently been introduced. For example, multidomain clustering, another iterative and tree structure-based classification method reminiscent of RP, has been shown to accurately identify different structural families in an anti-HIV data set[13] (but is yet to be evaluated for virtual screening). Other state-of-the-art compound classification methods such as the binary kernel discrimination approach can be expected to yield up to a 10-fold enrichment in the identification of active compounds when applied to different data sets.[24] One of the most significant improvements in hit rate versus random screening was recently reported not for clustering techniques but structure-based virtual screening. In a comparison of virtual and high-throughput screening for inhibitors of tyrosine phosphatase-1B, docking calculations achieved a 1700-fold hit enrichment,[25] comparable in magnitude to the results of our (simulated) virtual screening study.

Conclusions. In this study, the MP approach, originally developed for diversity analysis and subsequently adopted for classification of active compounds, has been extended for virtual screening of large databases. Among the spectrum of virtual screening methods, RMP is conceptually similar to, yet distinct from, recursive partitioning. By design, RMP is very suitable for virtual screening applications, as it can be directly applied and does not require learning sets for predictive model building. Facilitated by its GA implementation, an important and characteristic aspect of RMP is that its recursions proceed in independently initialized descriptor spaces, which is expected to increase the resolution and performance of the approach. In essence, as the pool of database compounds becomes successively smaller, multiple unbiased descriptor settings are employed to improve the recognition of compounds having similar activity. The focal point and guiding principle of RMP is recursive copartitioning of known active compounds that are added as baits to source databases. The underlying idea is that descriptor combinations capable of copartitioning bait molecules also recognize other compounds having similar activity. The results of our benchmark calculations on diverse activity classes support the predictive value of this copartitioning principle. In each case, a number of hits could be identified after five recursions by selecting less than 100 database molecules, yielding significant hit rates and improvements over random selection. On the basis of our findings, we conclude that the RMP approach merits further evaluation and application in virtual screening.

ACKNOWLEDGMENT

The authors thank Florence Stahura, Ling Xue, and Douglas B. Kitchen for helpful discussions.

REFERENCES AND NOTES

(1) Chen, X.; Rusinko, A., III; Young, S. S. Recursive partitioning analysis of a large structure−activity data set using three-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.

(2) 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.

(3) Miller, D. A. Results of a new classification algorithm combining K nearest neighbors and recursive partitioning. *J. Chem. Inf. Comput. Sci.* **2001**, *42*, 168−175.

(4) Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 393−404.

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

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

(7) 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.

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

(9) Xie, D.; Tropsha, A.; Schlick, T. An efficient projection protocol for chemical databases: single value decomposition combined with truncated Newton minimization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 167−177.

(10) Hull, R. D.; Singh, S. B.; Nachbar, R. B.; Sheridan, R. P.; Kearsley, S. K.; Fluder, E. M. Latent semantic structure indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.* **2001**, *44*, 1177−1184.

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

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

(13) Tamura, S. Y.; Bacha, P. A.; Gruver, H. S.; Nutt, R. F. Data analysis of high-throughput screening results: application of multidomain clustering to the NCI anti-HIV. *J. Med. Chem.* **2000**, *45*, 3082−3093.

(14) 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.

(15) Meier, P. C.; Zünd, R. E. *Statistical methods in analytical chemistry*; John Wiley & Sons: New York, 2000.

(16) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.* **1997**, *15*, 372−285.

(17) Willett, P. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *J. Comput. Biol.* **1999**, *6*, 447−457.

(18) ACD (Available Chemicals Directory), MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

(19) Godden, J. W.; Xue, L.; Bajorath, J. Classification of biologically active compounds by median partitioning. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1263−1269.

(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) MOE (Molecular Operating Environment), version 2001.01, Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.

(22) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464−477.

(23) Engels, M. F. M.; Venkatarangan, P. Smart screening: approaches to efficient HTS. *Curr. Opin. Drug. Discov. Dev.* **2001**, *4*, 275−283.

(24) Harper, G.; Bradshaw, J.; Gittin, J. C.; Green, D. V. S.; Leach, A. R. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.

(25) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P., Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213−2221.

CI0203848