Classification of Biologically Active Compounds by Median Partitioning

Jeffrey W. Godden,[†] Ling Xue,[†] and Jürgen Bajorath*[,†,‡]

Department of Computer-Aided Drug Discovery, Albany Molecular Research, Inc. (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 May 6, 2002

The median partitioning (MP) method was originally developed for the selection of diverse subsets from compound databases. Following this approach, property descriptors are used in subsequent steps to divide compounds into defined partitions from which representative molecules are selected. For descriptor analysis, MP was coupled to a genetic algorithm. MP subset selection does not depend on pairwise comparison of molecules and is therefore applicable to very large compound pools. Here the MP approach was evaluated for the classification of molecules according to biological activity. A total of 317 molecules belonging to 21 different activity classes were studied. MP compound classification calculations were carried out both in the presence and absence of 2000 randomly selected "background" molecules. The performance of MP was compared to cell-based partitioning and found to be at least comparable, with up to approximately 82% of active molecules occurring in "pure" partitions consisting only of molecules sharing the same activity. Different from cell-based methods, MP classification is based on "direct" and "sequential" contributions of molecular property descriptors. Our results suggest that MP in not only an effective method for the selection of diverse subsets but also for the classification of active compounds and searching for molecules with desired activity.

INTRODUCTION

Recently, we have introduced the MP approach as a novel statistical method for the selection of diverse or representative molecular subsets.[1] The median is defined as the value within a distribution that divides the population into two equal subsets (above and below the median value).[2] MP uses n property descriptors and, in each of n subsequent steps, divides a population of molecules into subpopulations above and below the median value of each descriptor until a desired number of unique 2^n partitions are obtained. Figure 1 illustrates the underlying concept of the MP technique. Its key feature is the sequential transformation of molecular descriptor contributions into a binary partitioning scheme, which is not order-dependent. Compound analysis is carried out directly in descriptor space and each of the resulting 2^n partitions is characterized by a unique signature code of n bits. Unlike many other dissimilarity-based methods,[3,4] MP does not depend on pairwise comparison of source compounds and can therefore be effectively applied to large compound databases, which has been a major reason for its design. Essentially, the only time-limiting step is the calculation of descriptor values for large numbers of compounds. Although MP is only based on a binary classification scheme relative to each descriptor, without the ability to scale or combine descriptor contributions, we found that the method was capable of creating a similar degree of diversity as cell-based partitioning approaches, as assessed using various diversity metrics.[1]

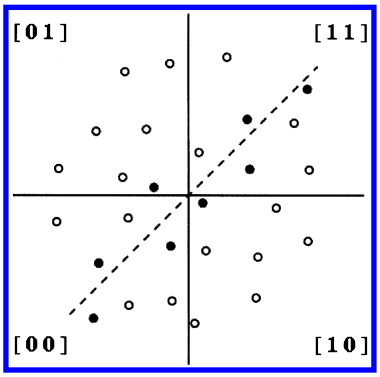

Figure 1. Median partitioning. The schematic representation illustrates the principle of MP by displaying a two-dimensional chemical space. The orthogonal axes represent the medians of two completely uncorrelated descriptors that divide a compound set (nonfilled dots) into equal subpopulations, and n such descriptors can be used in subsequent steps to create 2^n partitions for subset selection of compound classification. As shown, each partition is characterized by a unique binary code. The dashed axis represents a diagonal of correlation for medians of two strongly correlated descriptors. In this case, the compound distribution (filled dots) is skewed along this diagonal, and, in consequence, the partitions do not contain equal subpopulations but either fewer or more molecules (under- or overpopulated partitions).

Compound partitioning or clustering[5] is often also applied when attempting to identify active compounds in databases, for example, by selecting candidates that closely map to known hits or leads in however defined chemical spaces. Among partitioning approaches developed or adapted for computational chemistry applications, MP is in some ways reminiscent of, yet methodological distinct from, recursive partitioning.[6,7] This technique also divides compound databases in subsequent steps according to descriptor values.

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jurgen.bajorath@albmolecular.com.
[†] AMRI-BRC.
[‡] University of Washington.

10.1021/ci020372m CCC: $22.00
Published on Web 08/02/2002

However, recursive partitioning utilizes learning sets of active and nonactive molecules to identify suitable descriptor combinations by dividing these molecules into statistically distinct subsets along decision trees. These procedures are continued until a minimal subset enriched in compounds with desired activity is obtained, and this protocol is then applied to relevant test sets (for example, HTS data).

Among partitioning methods, cell-based approaches based on, for example, the well-known BCUT metric[8,9] or principal component analysis (PCA)[10,11] have become particularly popular. These methods and also other similarity search methods based on singular values decomposition[12] depend on combining descriptor combinations and reducing the dimensionality of descriptor space. Partitioning in low-dimensional chemistry spaces is conceptually elegant and generally considered to be a powerful approach for compound classification. Relative to the original n-dimensional descriptor spaces, often cited advantages of low-dimensional spaces include that compound distributions are usually faster to compute and easier to visualize and that principal axes forming these spaces are orthogonal (and thus easier to bin) and combine weighted contributions from the original descriptors.

Although MP was originally developed for diverse subset selection, rather than prediction of compound similarity, we have now evaluated its potential for the classification of molecules according to biological activity. We felt this was an interesting test case because, as mentioned above, MP is a "direct" descriptor-based approach focused on subsequent pseudobinary classification steps and does not share essential features with dimension reduction methods. Therefore, we have tested MP in this context and compared its performance with PCA-based partitioning of active compounds that was previously established in our laboratory.[10,11] To facilitate the analysis, MP was coupled to a genetic algorithm for descriptor selection. We found that MP was capable of classifying biologically active compounds with reasonable to high prediction accuracy, at least comparable to PCA-based partitioning. The study also revealed some additional characteristics of the MP approach. For example, in contrast to MP-based subset selection, descriptor correlation effects did not significantly influence compound classification.

MATERIALS AND METHODS

Compound Databases. For compound classification, we used a previously assembled database consisting of 317 compounds belonging to 21 different biological activity classes,[11] including diverse sets of enzyme inhibitors, receptor agonists and antagonists, and both synthetic and naturally occurring molecules. The composition of this database is summarized in Table 1. It was originally assembled as a benchmark compound set for PCA-based partitioning and using it in this study made it possible to directly compare previous results with MP classification. In addition, we added 2000 randomly collected background molecules from the ACD[13] to this database to further increase the degree of difficulty for compound classification.

Relevant Descriptors. All calculations reported herein, including both MP and PCA-based partitioning, were based on a previously reported set of a total of 147 1D, 2D, and implicit 3D descriptors[11] and a publicly available set of 166 structural keys.[14] Implicit 3D descriptors refer to a class of composite descriptors that map diverse properties to molecular surfaces approximated from 2D representations of molecules.[15] Values for all property descriptors were calculated with MOE.[16] A detailed description of our basic descriptor set has been reported.[11] Nevertheless, those descriptors that occurred in the best scoring combinations, as identified in the course of our calculations, are also defined in Table 2. Since MP relies on the calculation of medians of descriptor value distributions, binary or two-state descriptors such as structural fragments cannot be applied here. Our only requirement for the preselection of property descriptors for MP was that they had nonzero descriptor entropy[17] in the compound databases studied and thus value distributions[17,18] for which meaningful median values could be calculated. This effectively reduced the number of suitable property descriptors from 147 to 130, which were used as the initial descriptor pool for MP. By contrast, for PCA-based partitioning, a set of 236 descriptors containing 147 MACCS keys was used. The 89 property descriptors chosen for PCA were also a subset of the original 147 descriptors, which was selected based on best performance in preliminary calculations.[11]

Table 1. Biological Activity Classes

biological activity	no. of compds
cyclooxygenase-2 (Cox-2) inhibitors	17
tyrosine kinase (TK) inhibitors	20
HIV protease inhibitors	18
H3 antagonists	21
benzodiazepine receptor ligands	22
serotonin receptor ligands (5-HT)	21
carbonic anhydrase II inhibitors	22
β-lactamase inhibitors	14
protein kinase C inhibitors	15
estrogen antagonists	11
antihypertensive (ACE inhibitor)	17
antiadrenergic (β-receptor)	16
glucocorticoid analogues	14
angiotensin AT1 antagonists	10
aromatase inhibitors	10
DNA topoisomerase I inhibitors	10
dihydrofolate reductase inhibitors	11
factor Xa inhibitors	14
farnesyl transferase inhibitors	10
matrix metalloproteinase inhibitors	12
vitamin D analogues	12

Partitioning Calculations. To facilitate descriptor selection and optimization, both PCA and MP were coupled to a genetic algorithm (GA),[19] for PCA-based partitioning as described previously (GA-PCA)[10] and for MP as summarized in Figure 2 (GA-MP). In both cases, randomly chosen descriptor combinations are encoded in chromosomes, and the partitioning calculations are carried out and evaluated via a scoring function, which is then optimized during GA cycles by altering descriptor combinations using mutation (inversion of single bit positions) and crossover (bit segment swapping) operations until a predefined convergence criterion is reached. In contrast to GA-MP, GA-PCA also requires the application of a binning scheme[20] to the PC axes selected to form low-dimensional chemistry space, which effectively determines the number of cells for partitioning. The number of PCs and bins per PC axis represent additional calculation parameters, which are also encoded in chromosomes for the GA-PCA algorithm.[10] Since such additional calculation

Table 2. Definition of Selected Descriptors[a]

descriptor	definition	median (317)	median (2317)
apol	sum of the atomic polarizabilities of all atoms	55.26	44.49
a_aro	number of aromatic atoms	12	10
a_don	number of H-bond donors	2	1
a_heavy	number of heavy atoms	26	21
a_hyd	number of hydrophobic atoms	17	14
a_nN	number of nitrogen atoms	3	1
a_nF	number of fluorine atoms	0	0
a_nS	number of sulfur atoms	0	0
a_nI	number of iodine atoms	0	0
b_heavy	number of bonds between heavy atoms	29	22
b_ar	number of aromatic bonds	12	11
b_double	number of double nonaromatic bonds	1	1
chi0	atomic connectivity index (order 0)[23]	19.07	15.28
chi1v_C	carbon valence connectivity index (order 1)	5.93	4.55
chi1_C	carbon connectivity index (order 1)	7.83	6.02
diameter	largest value in the distance matrix[24]	13	11
KierA3	third kappa shape index[23]	3.87	3.59
PEOE_RPC-	relative negative partial charge[25]	0.17	0.21
PEOE_VSA+3	sum of v_i where p_i is in the range [0.15,0.20)	10.68	0.00
PEOE_VSA-1	sum of v_i where p_i is in the range [−0.10,−0.05)	55.88	56.24
PEOE_VSA-3	sum of v_i where p_i is in the range [−0.20,−0.15)	0.00	0.00
PEOE_VSA-4	sum of v_i where p_i is in the range [−0.25,−0.20)	5.51	0.00
PEOE_VSA-5	sum of v_i where p_i is in the range [−0.30,−0.25)	13.57	13.57
PEOE_VSA_POS	total positive van der Waals surface area	195.83	146.89
PEOE_VSA_FPNEG	fractional negative polar van der Waals surface area	0.09	0.08
PEOE_VSA_FHYD	fractional hydrophobic van der Waals surface area	0.84	0.86
SlogP_VSA2	sum of v_i such that L_i is in (−0.2,0]	23.86	19.41
SlogP_VSA7	sum of v_i such that L_i is in (0.25,0.30]	124.85	88.22
SMR_VSA0	sum of v_i such that R_i is in [0,0.11]	32.16	23.86
SMR_VSA1	sum of v_i such that R_i is in (0.11,0.26]	36.39	22.00
SMR_VSA4	sum of v_i such that R_i is in (0.39,0.44]	6.37	2.76
SMR_VSA5	sum of v_i such that R_i is in (0.44,0.485]	158.79	126.75
VAdjMa	vertex adjacency information (magnitude)	5.86	5.46
VDistEq	vertex distance equality index	3.44	3.24
VDistMa	vertex distance magnitude index	9.13	8.47
vsa_acc	VDW surface area of hydrogen-bond acceptors	27.93	19.25
vsa_don	VDW surface area of hydrogen-bond donors	0.00	0.00
vsa_other	VDW surface area of nondonor/-acceptor atoms	35.78	27.10
vsa_pol	VDW surface area of polar atoms	19.25	0.00
vdw_vol	VDW volume calculated using a connection table	480.21	389.72
zagreb	Zagreb index[26]	142	106

[a] In the above, v_i is the van der Waals (VDW) surface area of atom i. p_i represents the partial charge of atom i calculated using the PEOE method.[25] L_i denotes the contribution to logP(o/w) for atom i as calculated in the SlogP descriptor.[27] R_i denotes the contribution to molar refractivity for atom i as calculated in the SMR descriptor.[27] The design of "VSA" descriptors has also been reported.[15] For each listed descriptor, calculated median values are shown for both compound databases analyzed here (consisting of 317 and 2317 molecules, respectively).

parameters are not required for MP, the design of chromosomes for MP-GA is simpler, as illustrated in Figure 2. Here initially assembled chromosomes only represent the total number of available descriptors, 130 in this case, and each bit, if set on, adds a specific descriptor to the calculations. Equivalent conditions were established for both GA-MP and GA-PCA calculations. First 200 chromosomes were randomly generated with an initial occupancy rate of less than 10%, and the top scoring 25% of the chromosomes were subjected to pairwise crossover operations, followed by random mutation of all remaining chromosomes at a rate of 5%. GA cycles were continued until no change in score for 1000 cycles was observed. Three independent GA optimizations were carried out, GA-MP for our database consisting only of active compounds (317 molecules) and both GA-MP and GA-PCA for the expanded database also containing background compounds (2317 molecules). For GA-MP on 317 molecules, convergence was reached after 3502 GA cycles and for GA-MP on 2317 molecules after 6713 cycles. For GA-PCA on 2317 molecules, 13 657 iterations were required to reach convergence. For comparison, the GA-PCA data for 317 molecules were taken from ref 11, our previous GA-PCA study on 21 activity classes.

Classification and Scoring Scheme. The general goal of compound classification calculations, as reported herein, is to obtain as many compounds as possible in "pure" partitions or cells (that exclusively consist of molecules sharing the same activities), while minimizing the number of compounds in mixed partitions (i.e., consisting of molecules having different activity) or singletons (active molecules not predicted to be similar to others). Furthermore, we aim to identify those descriptor combinations that yield best predictive performance. To meet the former goal, an appropriate scoring function is required, to meet the latter, an algorithm to facilitate descriptor selection (as described above). Therefore, for both types of calculations (PCA or MP, with or without background molecules), the following scoring function was implemented and optimized during GA cycles:

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

Table 3. Top 10 Scoring Descriptor Sets from GA-MP on 21 Biological Activity Classes[a]

descriptors	nDS	score	%P	P	nP	S	M	nM	cc_{av}
PEOE_VSA+3, PEOE_VSA-3, PEOE_VSA-5, RPC-, SMR_VSA0, SMR_VSA4, a_aro, a_nO, a_nS, b_ar, chi1v_C, vdw_vol, vsa_don	13	1.27	81.7	79	259	46	5	12	0.18
PEOE_VSA+3, PEOE_VSA-3, PEOE_VSA-5, RPC-, SMR_VSA0, SMR_VSA4, a_aro, a_nO, a_nS, chi1v_C, vdw_vol, vsa_don	12	1.27	81.7	79	259	46	5	12	0.17
PEOE_VSA+3, PEOE_VSA-3, PEOE_VSA-5, RPC+, SMR_VSA0, SMR_VSA4, a_aro, a_nO, a_nS, chi1v_C, vdw_vol, vsa_don	12	1.27	81.7	79	259	46	5	12	0.17
PEOE_VSA+3, PEOE_VSA-3, PEOE_VSA-5, RPC-, SMR_VSA0, SMR_VSA4, a_aro, a_nO, a_nS, b_ar, chi1v_C, vdw_vol, vsa_don	13	1.27	81.7	79	259	46	5	12	0.18
PEOE_VSA+3, PEOE_VSA-3, PEOE_VSA-5, RPC-, SMR_VSA0, SMR_VSA4, a_aro, a_nO, a_nS, chi1v_C, vdw_vol, vsa_don	12	1.27	81.7	79	259	46	5	12	0.17
PEOE_VSA+3, PEOE_VSA-3, PEOE_VSA-5, RPC-, SMR_VSA0, SMR_VSA4, a_aro, a_nO, a_nS, chi1, chi1v_C, vsa_don	12	1.27	81.7	82	259	48	4	10	0.17
PEOE_VSA-5, RPC-, SMR_VSA0, SMR_VSA4, SlogP_VSA1, VAdjMa, a_aro, a_nO, a_nS, b_1rotN, b_ar, vsa_don	12	1.26	81.4	73	258	42	7	17	0.23
PEOE_VSA-5, RPC-, SMR_VSA0, SMR_VSA4, SlogP_VSA1, VAdjMa, a_aro, a_nO, a_nS, b_1rotN, vsa_don	11	1.26	81.4	73	258	42	7	17	0.23
PEOE_VSA-5, RPC+, SMR_VSA0, SMR_VSA4, SlogP_VSA1, VAdjMa, a_aro, a_nO, a_nS, b_1rotN, b_ar, vsa_don	12	1.26	81.4	73	258	42	7	17	0.23
PEOE_VSA-5, RPC-, SMR_VSA0, SMR_VSA4, SlogP_VSA1, VAdjMa, a_aro, a_nO, a_nS, b_1rotN, b_ar, vsa_don	12	1.26	81.4	73	258	42	7	17	0.23
a_aro, a_nO, a_nS, PEOE_VSA-5, SMR_VSA0, SMR_VSA4, vsa_don	7				consensus				

[a] The selected descriptors are defined in Table 2. The "consensus" combination includes those descriptors that are shared among the top scoring combinations. The following abbreviations are used: "nDS", number of descriptors; "%P", percentage of active compounds in pure partitions; "P", number of pure partitions; "nP", total number of compounds in pure partitions; "S", number of singletons; "M", number of mixed partitions; "nM", total number of compounds in mixed partitions; cc_{av}, average pairwise descriptor correlation coefficient.

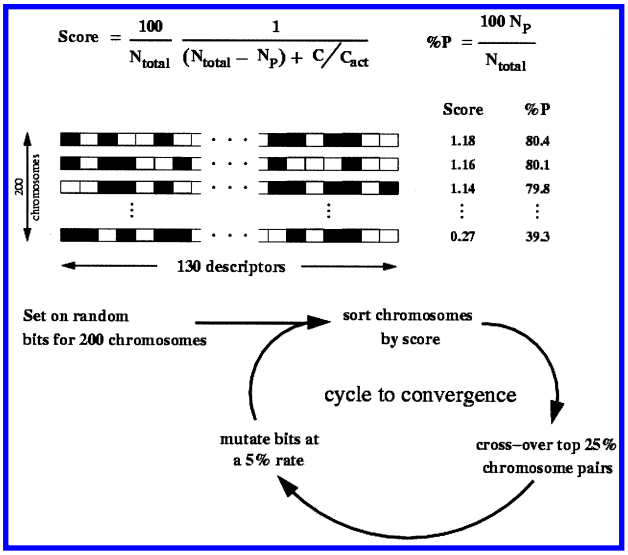

Figure 2. GA-MP method. The figure illustrates how MP calculations are coupled to a genetic algorithm to optimize a scoring function (see Methods) and facilitate automated selection of preferred descriptor combinations.

In this formulation, N_{total} is the total number of active compounds (here 317), and N_p is the number of compounds occurring in pure partitions. Both the number of compounds in mixed classes and singletons are regarded as classification failures. In addition, C is the total number of partitions that contain active compounds (pure, mixed, or singletons) and C_{act} is the number of different activity classes in the database (21 in this case). Thus, the scoring function also attempts to minimize the total number of "active" partitions or cells that are created. Consequently, high scores are obtained if many compounds occur in a small number of pure partitions. A scaling factor of 100 is applied to obtain top scores greater than 1. The addition of background compounds increases the degree of difficulty for the classification calculations because the statistical probability of producing mixed partitions or cells becomes significantly higher. In addition, as an intuitive measure of overall classification accuracy for each calculation, we also define the fraction of compounds in pure partitions as $\%P = 100 \cdot N_p/N_{total}$. This additional metric was not applied to guide descriptor selection during GA calculations but constantly monitored.

RESULTS AND DISCUSSION

MP for Subset Selection and Compound Classification. MP was originally developed as an efficient method to facilitate the selection of diverse and representative subsets from compound collections that did not depend on pairwise comparisons of molecules.[1] In that case, a large compound pool is divided into as many partitions as desired or possible from which representative compounds are selected. Subsequently, we asked the question of whether the MP approach would also be suitable to classify compounds according to biological activity. This is of course a principally different task, since here the similarity of certain molecules with

Table 4. Top 10 Scores from GA-MP on 21 Biological Activity Classes in the Presence of 2000 Background Compounds[a]

descriptors	nDS	score	%P	P	nP	S	M	nM	cc_{av}
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SMR_VSA4, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, VAdjMa, a_hyd, a_nN, a_nS, vsa_acc, vsa_pol, zagreb	19	0.50	63.1	69	200	86	22	31	0.23
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, a_hyd, a_nN, a_nS, b_heavy, vsa_acc, vsa_pol, zagreb	18	0.50	62.8	67	199	74	28	44	0.24
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, VAdjMa, a_hyd, a_nN, a_nS, vsa_acc, vsa_pol, zagreb	18	0.50	62.8	67	199	74	28	44	0.24
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, VAdjMa, a_hyd, a_nN, a_nS, b_heavy, vsa_acc, vsa_pol, zagreb	19	0.50	62.8	67	199	74	28	44	0.25
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, a_hyd, a_nN, a___nS, b_heavy, vsa_acc, vsa_pol, weinerPol, zagreb	19	0.49	62.5	67	198	78	25	41	0.25
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, VAdjMa, a_hyd, a_nN, a_nS, vsa_acc, vsa_pol, weinerPol, zagreb	19	0.49	62.5	67	198	78	25	41	0.25
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, VAdjMa, a_hyd, a_nN, a_nS, vsa_acc, vsa_other, vsa_pol, zagreb	19	0.49	62.5	70	198	77	26	42	0.25
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, VAdjMa, a_hyd, a_nN, a_nS, b_heavy, vsa_acc, vsa_other, vsa_pol, Zagreb	20	0.49	62.5	70	198	77	26	42	0.27
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, a_hyd, a_nN, a_nS, b_heavy, vsa_acc, vsa_other, vsa_pol, zagreb	19	0.49	62.5	70	198	77	26	42	0.25
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SMR_VSA4, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, VAdjMa, a_hyd, a_nN, a_nS, b_heavy, vsa_acc, vsa_other, vsa_pol, weinerPol, Zagreb	22	0.49	62.5	72	198	91	19	28	0.27
a_hyd, a_nN, a_nS, Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, vsa_acc, vsa_pol, zagreb	17				consensus				

[a] For definitions and abbreviations, see legend of Table 3.

respect to their specific activity must be recognized and distinguished from others. Whether or not MP would be capable of doing so was difficult to predict, and therefore the analysis presented herein was carried out. A potentially attractive feature of the MP approach for compound classification is that it does not depend on learning sets to derive predictive models of activity. Furthermore, in contrast to popular cell-based partitioning approaches, which create low-dimensional chemistry space for compound classification, MP operates in n-dimensional descriptor space and does not involve dimension reduction or secondary manipulations, other than transforming each descriptor contribution into a binary classification scheme. Thus, an additional question was as to how MP would compare with such approaches when applied to equivalent classification problems.

Median Values of Property Descriptors. As illustrated in Figure 1, median partitioning requires the calculation of values of each descriptor for all database compounds, followed by calculation of their median values, and separation of the molecular population into subpopulations above and below the medians. As mentioned above, an important point is that these requirements do not permit the use of two-state descriptors for MP, in particular, structural fragment-type descriptors or keys.[14] Table 2 reports the median values for selected descriptors in the two compound databases that we analyzed here. Although medians for some descriptors were identical in both data sets, the median values were generally database-dependent, as one would expect. Despite the fact that the 21 activity classes were a subset of the database containing 2317 compounds, medians for some of the descriptors were found to be significantly different, which reflects the intrinsic variability of these descriptors.[17]

MP Classification Performance. Coupling MP to a genetic algorithm to enable automated descriptor selection (Figure 2) proved to be an effective approach for classification calculations, similar to what was observed previously for PCA-based partitioning.[10] Table 3 summarizes the results obtained for MP classification of the active 317 compounds. We found that overall classification accuracy was high with up to 81.7% of our compounds occurring in pure partitions. As a control, we carried out 5000 GA cycles with random descriptor settings and no score optimization. For these random predictions, an average score of 0.04 was obtained (as opposed to 1.27, the best score in Table 3), and only 11.2% of the compounds were found in pure partitions. Between 11 and 13 descriptors were sufficient to achieve

Table 5. Ten Most Correlated Descriptor Pairs within the Top Scoring Combination for GA-MP on 21 Activity Classes[a]

descriptor pair		cc	first DS omitted		second DS omitted	
			score	%P	score	%P
a_aro	b_ar	0.994	1.27	81.7	1.27	81.7
chi1v_C	vdw_vol	0.913	1.25	81.4	0.99	77.3
SMR_VSA0	a_nO	0.802	1.25	81.4	0.98	77.0
PEOE_VSA-5	SMR_VSA0	0.753	0.83	73.8	1.25	81.4
PEOE_VSA+3	vsa_don	0.671	0.74	71.3	0.76	71.9
PEOE_VSA-5	a_nO	0.638	0.83	73.8	0.98	77.0
RPC-	chi1v_C	−0.549	0.72	70.7	1.25	81.4
RPC-	vdw_vol	−0.497	0.72	70.7	0.99	77.3
SMR_VSA0	a_nS	0.470	1.25	81.4	0.90	75.4
a_nS	chi1v_C	−0.458	0.90	75.4	1.25	81.4

[a] The pairwise correlation coefficients (cc) were analyzed for the top scoring descriptor set reported in Table 3. To investigate the influence of descriptor correlation on the predictive performance of MP classification, each of these descriptors was omitted, one at a time, the MP calculation was repeated, and the scores and percentages of compounds in pure classes were recalculated. "First DS omitted" means that the first descriptor of each pair was omitted prior to partitioning and "second DS omitted" means that the other descriptor of each pair was omitted.

this level of accuracy, and the top scoring descriptor combinations were quite similar, having seven descriptors in common. Shared descriptors range from rather simple ones (e.g., counting the number of aromatic or oxygen atoms in a molecule) to fairly complex descriptors. Among classification errors, singletons (i.e., unassigned active compounds) were three to four times more frequent than molecules in mixed partitions (i.e., false positive recognitions).

Table 4 shows the results for corresponding calculations on the database containing 2000 background compounds (thought to be "inactive"), which increased the degree of difficulty for the classification of active molecules. As to be expected, the scores and overall classification accuracy decreased, but approximately two-thirds of the active compounds were still correctly classified, with up to 63.1% of active molecules occurring in pure partitions. In this case, for random predictions, an average score of 0.03 was obtained and a classification accuracy of 9.2%. Thus, the achieved enrichment of compounds with similar activity in unique partitions was still significant. For the expanded database, both the number of singletons and compounds in mixed partitions increased relative to the results obtained for the 21 activity classes only. However, among classification errors, the trend seen in Table 3 reversed, and approximately twice as many compounds were found in mixed partitions than singletons. This can be rationalized by the significantly increased probability of obtaining mixed partitions in the presence of background compounds. As evident in Table 4, the number of descriptors among the top scoring combinations also increased with the number of database compounds, and 18 or 19 descriptors were required to achieve best performance. However, as seen before, the best descriptor combinations revealed in our calculations were also very similar in this case.

Descriptor Correlation Effects. We further investigated the potential influence of descriptor correlation on MP classification results. This was done because for subset selection, descriptor correlation proved to be an important factor. As illustrated in Figure 1, the use of correlated descriptors in MP produces both under- and overpopulated partitions. This is problematic if one attempts to evenly populate as many partitions as possible for diversity selections, where it is thus advisable to reduce descriptor correlation as much as possible.[1] This can be accomplished, for example, by incorporating correlation coefficients into the GA fitness function for descriptor selection and by minimizing them during GA calculations.[1] However, the influence of descriptor correlation effects on MP in compound classification was unclear. Tables 3 and 4 report the average pairwise correlation coefficients for the descriptors in top scoring combinations. They were low in the first case (Table 3), although no attempt was made here to reduce correlation effects during GA calculations, and increased somewhat with increasing number of descriptors (Table 4). However, despite relatively low average values, top scoring MP descriptor combinations included some highly correlated descriptor pairs, as shown in Table 5. To investigate whether strongly correlated descriptors contributed to the accuracy of the calculations, we systematically omitted descriptors from correlated pairs, one at a time, and recalculated the

Table 6. Comparison of Top Scoring Descriptor Sets from GA-MP and GA-PCA[a]

descriptors	nDS	PC	bins	score	%P	P	nP	S	M	nM
1.1. GA-PCA, 317 Active Compounds										
a_aro, 19, 26, 56, 62, 79, 111, 122, 124, 140	10	7	4	1.22	80.6	69	256	38	7	23
1.2. GA-MP, 317 Active Compounds										
PEOE_VSA+3, PEOE_VSA-3, PEOE_VSA-5, RPC-, SMR_VSA0, SMR_VSA4, a_aro, a_nO, a_nS, b_ar, chi1v_C, vdw_vol, vsa_don	13	n/a	n/a	1.27	81.7	79	259	46	5	12
2.1. GA-PCA, 317 Active Plus 2000 Background Compounds										
PC+, PC-, a_nI, f_c=o, vsa_acc, PEOE_VSA-5, 25, 40, 57, 71, 72, 96, 98, 101, 131, 158, 159, SMR_VSA0	18	6	7	0.37	55.2	57	175	82	19	60
2.2. GA-MP, 317 Active Plus 2000 Background Compounds										
Kier3, PEOE_RPC-, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, RPC-, SMR_VSA4, SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, TPSA, VAdjMa, a_hyd, a_nN, a_nS, vsa_acc, vsa_pol, zagreb	19	n/a	n/a	0.50	63.1	69	200	86	22	31

[a] "PC" reports here the number of principal components and "bins" the number of axis intervals used for PCA-based partitioning. These parameters were selected by GA calculations and determine the dimensionality of the descriptor space and total number of cells for the partitioning experiments. Numbers (e.g., 19, 26, 56 ...) indicate structural keys or fragment-type descriptors.[14] All other abbreviations are used according to the legend of Table 3. Results for partitioning study 1.1. (GA-PCA, 317 active compounds) were taken from a previous report.[11]

partitions. The results in Table 5 show that most of these descriptors indeed contributed to classification accuracy, as omitting them generally reduced the scores. In addition, to further investigate if some of the descriptors were perhaps simply "carried along" during the GA calculations, which could not be ruled out, we also subjected top scoring combinations to complete factorial analysis and determined to what extent the scores changed. Again, we found that omission of descriptors from our preferred combinations, be they correlated or not, reduced the scores, and no better scoring combination could be identified by factorial analysis. Thus, in contrast to MP subset selection, descriptor correlation effects do not need to be reduced if MP is applied to compound classification. In fact, our data suggest that contributions from correlated descriptors contributed to classification accuracy.

Comparison with GA-PCA. How do the results of GA-MP compare to cell-based partitioning? Table 6 summarizes the results of our comparisons. For classification of 317 active compounds, the results for GA-MP were similar, or slightly better, than the best results obtained previously for PCA-based partitioning. In each case, greater than 80% of the compounds were correctly classified. For predictions in the presence of 2000 background compounds, GA-MP achieved about 8% greater accuracy than GA-PCA. The number of descriptors selected by GA-MP and GA-PCA for classification of the two databases was similar in each case. GA-MP had the tendency to produce more singletons than GA-PCA, whereas GA-PCA produced more compounds to mixed cells (i.e., has a higher rate of false positive assignments). Comparison of the results also revealed different descriptor preferences. Relatively complex descriptors were prevalent among those selected by GA-MP, in particular composite surface descriptors,[15] which represent an "information-rich" class,[17] whereas structural keys dominated the GA-PCA selections. Following early observations by Brown and Martin,[21] a number of independent studies have demonstrated the value of structural key-type descriptors for various applications[22] including compound clustering and partitioning. However, as shown here, dependent on the method used, various combinations of other descriptors yield at least comparable accuracy in partitioning calculations.

Conclusions. We have investigated the application of median partitioning to compound classification. Although MP was originally developed for diverse subset selection, we found that the method was capable of classifying compounds according to biological activity with reasonable to high classification accuracy, which was at least comparable to PCA-based partitioning. When coupled to a genetic algorithm, selection of well-performing descriptor combinations for MP was straightforward. MP is conceptually simpler than cell-based partitioning methods and computationally very effective, as its only significant time limiting step is the calculation of descriptors values for database compounds. Moreover, in contrast to MP subset selection, the presence of descriptor correlation effects was not a limiting factor for MP-based compound classification. In fact, correlated descriptors detectably contributed to MP classification performance. Thus, taken together, our findings indicate that MP is a promising and easy to apply method to search for compounds having similar activity. In addition, the results obtained in this study suggest that low-dimensional chemistry spaces, although elegant in their design, may not always be required for accurate compound partitioning.

REFERENCES AND NOTES

(1) 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.

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

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

(4) 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.

(5) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.

(6) Friedman, J. A. Recursive partitioning decision rules for nonparametric classification. *IEEE Trans. Comput.* **1977**, *26*, 404−408.

(7) 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.

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

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

(10) 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.

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

(12) 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.

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

(14) MACCS keys; MDL Information Systems, Inc.: 14600 Catalina Street, San Leandro, CA 94577.

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

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

(17) 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.

(18) 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.

(19) Forrest, S. Genetic algorithms − Principles of natural selection applied to computation. *Science* **1993**, *261*, 872−878.

(20) Bayley, M. J.; Willett, P. Binning schemes for partition-based compound selection. *J. Mol. Graph. Model.* **1999**, *17*, 10−18.

(21) 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.

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

(23) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling. *Rev. Comput. Chem.* **1991**, *2*, 367−422.

(24) Petitjean, M. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.

(25) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(26) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological indices: their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891−898.

(27) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

CI020372M