Evaluation of Descriptors and Mini-Fingerprints for the Identification of Molecules with Similar Activity

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

New Chemical Entities, Inc., 18804 North Creek Parkway, Suite 100, Bothell, Washington 98011, and
Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received April 18, 2000

Combinations of 65 preferred 1D/2D molecular descriptors and 143 single structural keys were evaluated for their performance in compound classification focused on biological activity. The analysis was based on principal component analysis of descriptor combinations and facilitated by use of a genetic algorithm and different scoring functions. In these calculations, several descriptor combinations with greater than 95% prediction accuracy were identified. A set of 40 preferred structural keys was incorporated into a small binary fingerprint designed to search databases for compounds with biological activity similar to query molecules. The performance of mini-fingerprints was tested by systematic similarity search calculations in a database consisting of compounds belonging to seven biological activity classes, which had not been used to select effective descriptors. In these blind test calculations, mini-fingerprints correctly identified approximately 54% of compounds sharing similar biological activity and with 1% false positives. Thus, although the design of mini-fingerprints is conceptually simple, they perform well in activity-oriented similarity searching.

INTRODUCTION

Computational methods that correlate structural features and properties of small molecules with specific biological activity are focal points of our research efforts. As part of these studies, we explore combinations of molecular descriptors for effective classification of compounds according to biological activity. Furthermore, we investigate the design of small binary bit string representations of molecules, called mini-fingerprints (MFPs),[1] to search databases for compounds with biological activity similar to query molecules. MFPs typically consist of only 50−60 bit positions[1] and are thus much smaller than other more widely used binary fingerprints.[2,3]

The design of MFPs was originally based on our observations that combinations of relatively few 1D/2D molecular descriptors[4,5] and structural fragments or keys[4−7] were sufficient to classify compounds according to different specific activities (e.g., different enzyme inhibitors, receptor agonists, or antagonists).[8] These initial findings encouraged us to explore combinations of a larger number of molecular descriptors for their performance in compound partitioning, one of the goals being to encode preferred descriptors as MFPs for similarity searching.

To classify compounds, we combine principal component analysis (PCA)[9,10] of molecular descriptor combinations, division of principal component space into cells, and assignment of test molecules to these cells (based on their positions in principal component space).[10] Using this approach in conjunction with a genetic algorithm[11,12] has enabled us previously to evaluate combinations of 111 molecular descriptors that could be calculated from 2D representations of molecules, including a set of structural keys that were treated as a single complex descriptor.[12] In these calculations, a number of descriptor combinations were identified that effectively partitioned sets of test compounds.[12]

In the current study, we first continued the evaluation of descriptor combinations for compound classification by testing 143 structural keys as separate descriptors together with 65 other 1D/2D descriptors. These calculations were performed using a benchmark database consisting of seven classes of compounds with different activities. Several descriptor combinations were identified that yielded greater than 95% prediction accuracy. Second, on the basis of these results, we designed a new MFP and tested MFPs in blind test calculations. To do so, we assembled a new database consisting of compounds belonging to seven biological activity classes different from those used to select effective descriptor combinations. The best MFP correctly identified ∼57% of all compounds with similar activity and detected only ∼1% false positives. Taken together, our results provide evidence that small fingerprints are of considerable value to search databases for compounds with similar biological activities.

MATERIALS AND METHODS

PCA calculations were carried out using the QuaSAR−Cluster function[10] of MOE[13] using a previously described protocol.[8] A detailed description of the implementation of the PCA function is available in an electronic publication.[10] Figure 1a illustrates the underlying idea. PCA performs a linear transformation of the vectors defining n-dimensional descriptor space and calculates a normalized set of principal components capturing critical descriptor combinations.[9] It

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jbajorath@nce-mail.com. Correspondence should be addressed at NCE, Inc.

[†] New Chemical Entities, Inc.

[#] University of Washington.

10.1021/ci000327j CCC: $19.00
Published on Web 08/24/2000

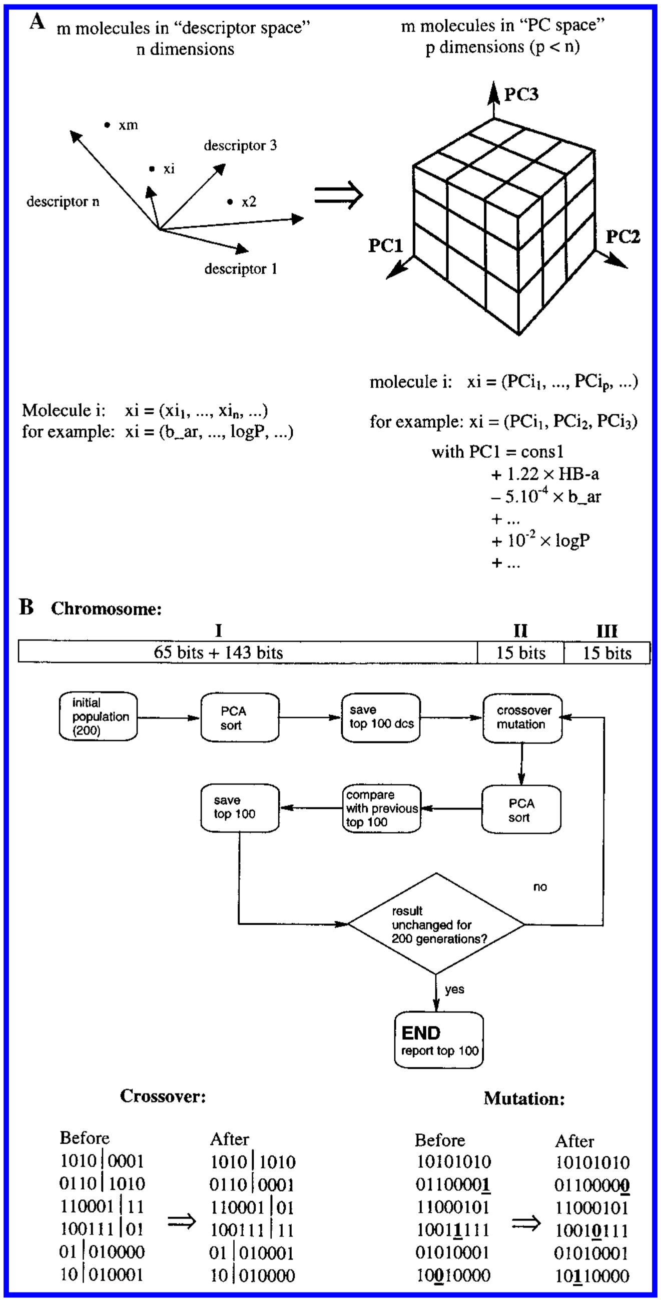


Figure 1. Identification of preferred molecular descriptors. The schematic summarizes key components of our approach to identify combinations of descriptors that effectively partition compound collections according to biological activity. Part (**A**) describes the start and end points of principal component analysis. Part (**B**) illustrates the design of a chromosome and implementation of a genetic algorithm to facilitate exploration of a large number of descriptor combinations. In the flowchart, "dcs" means descriptor combinations and "PCA" principal component analysis. The chromosome consists of three parts: (I) 65 bits each of which accounts for the presence (i.e., 1) or absence (i.e., 0) of one 1D/2D descriptor plus 143 bits each of which accounts for a single structural key; (II) each of 15 bits, if set on (i.e., 1), adds one principal component to the calculation; and (III) each of 15 bits, if set on, adds one bin segment for division of principal components.

reduces the dimensionality of the descriptor space and removes correlation between descriptors. In our calculations, the number of principal components and bins on each component axis were variable (between two and 15 in each case). We used the same benchmark database as in our initial studies[8,12] for compound partitioning to ensure that obtained results can be directly compared. This database consists of 455 compounds belonging to seven activity classes (inhibitors of cyclooxygenase-2, tyrosine kinases, carbonic anhydrase II, and HIV protease, ligands for benzodiazepine and serotonin receptors, and H3 antagonists).[8]

To make the analysis of a large number of descriptors possible, a genetic algorithm[11] (GA) was implemented.[12] Figure 1B shows the organization of the "chromosome" designed for our current analysis. It also illustrates how GA calculations are carried out. Calculation parameters were used as established previously.[12] The chromosome encodes descriptors and calculation parameters. To generate an initial chromosome population, each bit positions is assigned an 8% chance to be set on (i.e., 1) and 200 chromosomes are randomly generated. Following PCA, S_2 scores (as defined below) are calculated for the descriptor combination encoded in each chromosome, and the top scoring 100 chromosomes are retained and subjected to pairwise crossover operation for 25% of the population. Then mutation is carried out by randomly inverting 5% of the bits in each chromosome, and the resulting chromosomes represent the next generation. If a new chromosome has more than 20% of its bits set on after crossover and mutation, each bit is assigned a 50% probability to reach the next generation. Descriptor combinations in this population are again subjected to PCA, crossover, and mutation operations and the top scoring 100 combinations are saved. Convergence is reached if results do not change for 200 generations.

To determine what level of compound classification accuracy could be expected by chance, 300 randomly picked descriptor combinations (resulting in an average of 17 descriptors per calculation) were used to partition the test database. In these calculations, the number of principal components and bins per axis were also randomly selected within the ranges defined in the chromosome (resulting in an average of eight principal components and seven bins per calculation). To provide a reference value, the results were averaged over 300 calculations.

For evaluation of compound classification calculations, S_1 and S_2 scores[8,12] were defined as follows

$$S_1 = \frac{C_p}{10 \times C_m + C_s}$$

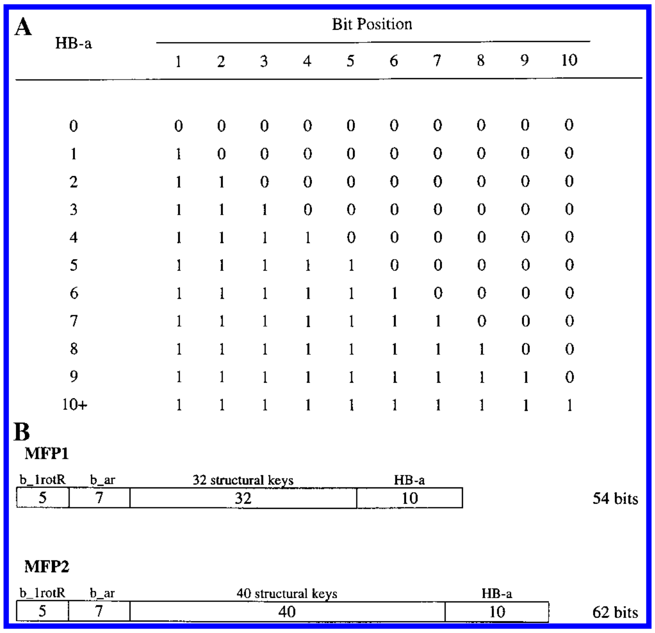
where C_p is the number of the pure classes, C_m is the number of the mixed classes, and C_s is the number of the singletons. A scaling factor of 10 was used to penalize the occurrence of mixed clusters. If only pure classes occur, S_1 is not defined and thus set to

$$S_1 = 1 + (7/C_p)$$

Pure classes contain only compounds having similar activity (a desired result) and mixed classes consist of compounds with different activity (not desired).

$$S_2 = 100 \times \frac{N_p}{N_m + C_s + (C/C_a)} \frac{1}{N_{total}}$$

Here N_p is the total number of the compounds in pure classes, N_m is the number of compounds in mixed classes, and N_{total} is the total number of compounds in the database. C is the total number of the classes obtained after PCA analysis, and



A

HB-a	Bit Position 1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0	0
5	1	1	1	1	1	0	0	0	0	0
6	1	1	1	1	1	1	0	0	0	0
7	1	1	1	1	1	1	1	0	0	0
8	1	1	1	1	1	1	1	1	0	0
9	1	1	1	1	1	1	1	1	1	0
10+	1	1	1	1	1	1	1	1	1	1

Figure 2. Design of mini-fingerprints. (**A**) illustrates how a numerical descriptor is encoded as a binary bit string using HB-a, the number of hydrogen bond acceptors in a molecule, as an example. "10+" means 10 or more acceptors per molecule. (**B**) shows organization of two mini-fingerprints, MFP1 and MFP2, that combine three numerically encoded descriptors (HB-a, b_1rotR, and b_ar) with 32 (MFP1) or 40 (MFP2) structural keys. "b_1rotR" accounts for the fraction of rotatable bonds in a molecule and "b_ar" for the number of aromatic bonds. MFP1 and MFP2 are composed of a total of 54 and 62 bit positions, respectively.

C_a is the number of different activity classes in the database. The factor of 100 was arbitrarily introduced to produce S_2 values comparable in magnitude to S_1. S_1 values are high if the number of mixed classes is small, and S_2 values are high if many compounds occur in a small number of pure classes. During genetic algorithm calculations, descriptor combinations were optimized on the basis of S_2 scores, and, for comparison, S_1 values were also calculated. "Overall prediction accuracy" was defined as $A = N_p/N_{total}$.

To test whether any of the seven activity classes in our compound database significantly influenced the results of compound partitioning, we carried out seven sets of reference calculations. In these calculations, each of the seven activity classes in our benchmark database was omitted once from the database, and the 15 top scoring descriptor combinations were used to partition the remainder of the database. In addition, top scoring descriptor combinations were also used for Jarvis-Patrick clustering[14] (a popular nonhierarchical clustering method)[15] of our test database.

Mini-fingerprints were designed as binary bit strings. Bit segments represented either a range of numerical descriptors or accounted for the presence or absence of preferred structural keys (identified by compound partitioning). Numerical descriptors were encoded as described.[1] For example, 10 bits were assigned to capture the number of hydrogen bond acceptors in a molecule as illustrated in Figure 2A. Bit positions in other segments detect the presence (1) or absence (0) of structural keys. Figure 2B schematically shows the design of two MFPs (described in more detail in the Results section).

The performance of fingerprints was assessed by systematic one-against-all similarity searching in a test database (see Results) consisting of compounds belonging to seven biological activity classes assembled from the Comprehensive Medicinal Chemistry (CMC)[16] or the Chapman and Hall (CH)[17] database. CMC contains many known drug molecules, while CH consists of molecules isolated from natural products. As a reference fingerprint, PH2D was used, as implemented in MOE.[13] This fingerprint (consisting of 1024 bit positions) captures all pairwise distances of atoms in a molecule on the basis of graph representations and generates a pharmacophore or signature key. As similarity criterion for pairwise comparison of fingerprints, different values of the Tanimoto coefficient (Tc) were used, defined as Tc = Bc/(B1+B2−Bc). Bc is the number of common bits set on (1) in molecules 1 and 2 and B1 and B2 are the number of bits set on in molecule 1 and 2, respectively.

For each compound, a search experiment identifies a varying number of similar compounds that reach a specified Tc threshold value (see Results). These compounds either belong to the same activity class as the query molecule (correct identification) or, alternatively, have a different activity (incorrect). Statistically there is a much greater chance to identify compounds incorrectly, since compounds with activity different from the query molecule represent the majority of the test database. As a simple scoring function, we calculate "% of correctly identified compounds" by dividing the number of correct identifications by the total number of molecules belonging to the same activity class. Accordingly, the "% of incorrectly identified compounds" is calculated by dividing the number of incorrect identifications by the total number of compounds having an activity different from the query molecule.

MFPs were generated using SVL[18] code and implemented in MOE. Likewise, all programs required for GA calculations and systematic similarity searching were written in SVL and run via the MOE platform.

RESULTS AND DISCUSSION

The current analysis is a continuation of our studies to identify combinations of molecular descriptors that effectively classify compounds according to specific biological activity. Table 1 summarizes stepwise analyses that systematically tested an increasing number of separate molecular descriptors (from 17 to 111 to 208). To ensure that results obtained in each study could be directly compared, we used the same benchmark database throughout. On the basis of the obtained results, we have constructed short binary bit string representations of molecular structure and properties, which we call mini-fingerprints, and tested the performance of these MFPs in similarity search calculations.

Compound Classification Approach. The method applied here combines principal component analysis of descriptor space and a genetic algorithm for combinatorial exploration (Figure 1). PCA reduces the dimensionality of the descriptor space and removes correlation between molecular descriptors. Thus, selected combinations should consist of largely noncorrelated descriptors. The end result is an orthogonal space defined by p principal components that are linear combinations of the original descriptors. Binning of these axes defines cells in principal component space to which molecules can be assigned based on their "coordinates" calculated from original descriptor values. In our approach,

Table 1. Calculations to Identify Preferred Descriptor Combinations for Compound Classification Based on Principal Component Analysis[a]

	study 1[8]	study 2[12]	current study
compounds/activity classes	455/7	455/7	455/7
number of 1D/2D descriptors	16	110	65
number of structural keys	57	166	143
structural keys treated as	complex descriptor	complex descriptor	single descriptors
principal component limit	3	2−15	2−15
number of bins per PC axis	8	2−15	2−15
scoring function	S_1	S_2, S_1	S_2, S_1
calculation method	factorial analysis	genetic algorithm	genetic algorithm

[a] Parameters of calculations for PCA-based analysis of molecular descriptor combinations are reported. When a genetic algorithm was used as the calculation method, descriptors were selected on the basis of S_2 scores, and S_1 scores were subsequently calculated for the top 100 descriptor combinations to make a direct comparison with study 1 possible. The initial set of 57 structural keys was selected based on their high frequency of occurrence in large compound databases.[8] In study 2, the complete set of 166 MACCS keys[7] was used as a complex descriptor. In the current study, 143 of these keys were evaluated as single descriptors. Twenty-three keys could be omitted from the calculations because they did not occur in any of the compounds in our test database.

Table 2. Top 15 Combinations of Descriptors for PCA-Based Compound Classification Including Single Structural Keys[a]

no.	descriptors	PC/bin	S_1	S_2	N_p	N_s	N_m
1	a_nI, b_triple, f_conh2, 25, 47, 53, 61, 62, 65, 81, 87, 122, 124, 158, 159	7/8	3.24	3.53	438 (55)	17	0 (0)
2	38, 59, 62, 69, 80, 96, 112, 127, 139, 145, 156	6/6	1.38	3.51	436 (44)	12	7 (2)
3	a_aro, 47, 62, 71, 112, 134	9/6	1.81	3.23	435 (49)	17	3 (1)
4	I_so2nh2, vsa_pol, 38, 62, 67, 71, 79, 106, 107, 110, 134, 164	5/7	1.93	3.14	435 (54)	18	2 (1)
5	I_so2nh2, b_ar, 47, 62, 80, 135, 159	8/8	2.24	3.11	434 (47)	21	0 (0)
6	vsa_pol, 15, 19, 22, 45, 62, 75, 80, 106, 114, 124, 166	6/9	1.52	2.89	432 (47)	21	2 (1)
7	a_nN, 15, 42, 62, 88, 92, 129, 130	7/6	1.61	2.85	432 (50)	21	2 (1)
8	a_nN, f_so2n, 22, 62, 81, 119, 129, 144	6/6	1.32	2.8	430 (45)	14	11 (2)
9	PEOE_VSA+3, b_double, 62, 75, 130, 154, 156	5/8	1.28	2.78	432 (59)	16	7 (3)
10	VAdjEq, a_don, 28, 34, 62, 64, 71, 81, 107, 110, 122, 159, 161	6/4	2.91	2.78	433 (64)	22	0 (0)
11	a_aro, 50, 61, 62, 83, 105, 124, 130, 152, 164	8/8	1.47	2.77	431 (53)	16	8 (2)
12	b_ar, 59, 61, 62, 79, 102, 146, 150	6/10	1.39	2.75	429 (39)	18	8 (1)
13	a_nO, 45, 58, 62, 81, 84, 91, 94, 110, 120, 148, 151	7/6	1.9	2.75	432 (59)	21	2 (1)
14	PEOE_VSA_PPOS, vsa_pol, 62, 80	10/10	1.5	2.71	431 (57)	18	6 (2)
15	FCharge, a_aro, b_1rotR, 16, 28, 52, 62, 69, 80, 95, 105, 131, 137, 154	6/10	1.84	2.68	431 (57)	21	3 (1)
	av of 300 randomly picked descriptor combinations and PC/bin values		0.43	0.29	262 (80)	168	63 (7)

[a] “PC” reports the number of principal components and “bin” the number of intervals per component. “N_p” is the total number of compounds in pure classes (“N_m”: in mixed classes, “N_s”: singletons). The number of obtained pure and mixed classes are given in parentheses. Descriptor combinations are sorted according to S_2 scores and S_1 scores are also calculated. Structural key descriptors are reported using their MACCS key number.[7]

each populated cell defines a “class” of compounds. The analysis aims to identify descriptor combinations that yield a maximum number of “pure classes”, defined as cells that contain only molecules sharing similar activity or equivalent specificity. Cells containing only one compound are called “singletons”, and their occurrence is, similar to “mixed classes”, penalized by our scoring functions (see Methods).

Preselection of Descriptors. To limit the combinatorial space for the current calculations, we have built on the results of our previous analysis.[12] Of 110 1D/2D descriptors studied, any descriptor was selected that occurred at least once in a combination of descriptors that yielded an S_2 score of greater than one. This scoring level reflects an overall prediction accuracy of greater than 80%. Prediction accuracy is discussed below in more detail. Using this criterion, 65 1D/2D descriptors were selected. These descriptors are reported and defined in Table 1 of the Supporting Information. Twenty-three of the complete set of 166 MDL/MACCS keys[7] did not occur in any of the compounds in our benchmark database and were therefore omitted. Thus, combinations of a total of 208 single descriptors were tested in this study, defining a theoretical search space of $2^{208} = 4.1 \times 10^{62}$ possible combinations, which of course cannot be fully explored. Therefore, a genetic algorithm was applied. Since these calculations typically provide reasonable, albeit not optimal solutions to a combinatorial problem under investigation,[11] we cannot expect to find the “best” descriptor combination. However, as demonstrated below, a variety of descriptor combinations were found to yield very high, in part close to optimum, classification performance.

Preferred Descriptor Combinations. Genetic algorithm-based evaluation of 208 single descriptors reached convergence (i.e., no changes in the results occurred for 200 iteration) after 8243 generations (which required more than a month calculation time on an SGI Octane workstation). The top 15 combinations of descriptors with highest S_2 scores identified in these calculations are listed in Table 2. The top scoring combinations consist of between four and 15 descriptors and are dominated by structural keys. However, only one combination, the second best, consists entirely of MACCS keys. The compositions of best performing combinations vary significantly. For example, the top two combinations have only one structural key in common. In addition to keys, various descriptors accounting for aromatic and polar character (or charge) are recurrent in the top 15 list. The results in Table 2 demonstrate that the performance of PCA-based compound partitioning using preferred descriptors is high. In the top 15 list, false positives (compounds in mixed classes) are nearly or, in three cases, completely eliminated, and, on average, only 18 singletons (i.e., less than

Table 3. Comparison of PCA-Based Compound Classification with Jarvis-Patrick Clustering[a]

descriptors	S_2	JP−S_2	JP−N_p	JP−N_s	JP−N_m
a_nI, b_triple, f_conh2, 25, 47, 53, 61, 62, 65, 81, 87, 122, 124, 158, 159	3.53	0.31	270 (19)	3	182 (6)
38, 59, 62, 69, 80, 96, 112, 127, 139, 145, 156	3.51	0.57	331 (18)	0	124 (8)
a_aro, 47, 62, 71, 112, 134	3.23	0.26	249 (19)	0	206 (8)
I_so2nh2, vsa_pol, 38, 62, 67, 71, 79, 106, 107, 110, 134, 164	3.14	0.63	340 (20)	2	113 (6)
I_so2nh2, b_ar, 47, 62, 80, 135, 159	3.11	0.67	345 (21)	1	109 (6)
vsa_pol, 15, 19, 22, 45, 62, 75, 80, 106, 114, 124, 166	2.89	0.48	315 (20)	0	140 (7)
a_nN, 15, 42, 62, 88, 92, 129, 130	2.85	0.26	250 (21)	4	201 (9)
a_nN, f_so2n, 22, 62, 81, 119, 129, 144	2.8	0.43	305 (22)	1	149 (10)
PEOE_VSA+3, b_double, 62, 75, 130, 154, 156	2.78	0.10	142 (14)	2	311 (18)
VAdjEq, a_don, 28, 34, 62, 64, 71, 81, 107, 110, 122, 159, 161	2.78	1.48	401 (27)	9	45 (3)
a_aro, 50, 61, 62, 83, 105, 124, 130, 152, 164	2.77	0.44	307 (22)	1	147 (7)
b_ar, 59, 61, 62, 79, 102, 146, 150	2.75	0.34	279 (17)	1	175 (9)
a_nO, 45, 58, 62, 81, 84, 91, 94, 110, 120, 148, 151	2.75	0.31	268 (20)	2	185 (8)
PEOE_VSA_PPOS, vsa_pol, 62, 80	2.71	0.15	187 (14)	1	267 (13)
FCharge, a_aro, b_1rotR, 16, 28, 52, 62, 69, 80, 95, 105, 131, 137, 154	2.68	0.97	376 (32)	6	73 (4)

[a] Top scoring descriptor combinations for PCA-based compound classification and S_2 scores are reported as in Table 2. For comparison, "JP−S_2" reports corresponding scores obtained by Jarvis-Patrick clustering. "JP−Np" is the corresponding total number of the compounds in pure clusters (the number of the pure clusters is reported in the parentheses). "JP−Ns" is the number of singletons. "JP−Nm" is the total number of compounds in mixed clusters (the number of the mixed clusters is given in parentheses).

4%) occur. Table 2 also shows that average prediction accuracy is low for many randomly picked descriptor combinations. Thus, the achieved level of accuracy is unlikely to result from chance events.

To what extent is the performance of descriptor combinations compound class- and/or algorithm-specific? Repartitioning of the test database using top scoring descriptor combinations after omitting one activity class at a time confirms that single classes of active compounds in our database do not dominate the performance level of the approach. As to be expected, obtained scores differ somewhat from those reported in Table 2. For a number of descriptor combinations, scores are even higher than for the complete compound collection, probably because partitioning six instead of seven classes reduces the statistical chance to make errors. Nevertheless, the overall performance is very similar for partitioning of all seven subsets. The results of all calculations are reported in Table 2 of the Supporting Information. By contrast, comparison with Jarvis-Patrick clustering[14] shows that the performance of descriptor combinations depends on the algorithm applied. Calculations reported in Table 3 reveal that the top 15 descriptor combinations, selected for PCA-based compound partitioning, do not perform well when used in Jarvis-Patrick clustering of our database. With one exception, no scores greater than 1 were obtained. In these calculations, Jarvis-Patrick clustering produced few singletons but many false positives (i.e., compounds in mixed clusters), which results in overall low performance.

Improving Prediction Accuracy. The comparison reported in Table 4 reveals that the accuracy of compound partitioning was significantly improved in subsequent analyses by increasing the number of molecular descriptors analyzed and identifying more powerful combinations. It should be noted that we deliberately limited the analysis to descriptors that could be calculated from 2D representations of molecules to avoid bias due to predicted database conformations of biologically active compounds. In the course of the studies, maximum prediction accuracy was improved from approximately 75% to 91% to 96%. This is reflected in the improvement of best scores that increased from approximately 1 to 1.8 to 3.2. As to be expected, the relative ranking of descriptor combinations varies with the calculation of S_1 (Table 4, part A) or S_2 (Table 4, part B) scores, since they weigh classification results in a different manner (see Methods). The comparison shows that overall improvement was mostly achieved by gradually eliminating false positives. In fact, the best performing combinations identified in this study totally eliminate false positive assignments. However, the classification is not perfect because few singletons remain.

Analysis of Descriptor Combinations. What are preferred descriptor combinations? Several conclusions can be drawn from the data presented in Tables 2 and 4. First, combinations of relatively few descriptors, from four to 15, give overall best results. Second, different combinations of descriptors can achieve high prediction accuracy. For example, combinations ranking at number two and three in Table 2 show high prediction accuracy. However, the second ranked combination consists of 11 structural keys and the third ranked combination of five structural keys and an additional descriptor accounting for aromatic atoms, and only two structural keys are common to both combinations. Since descriptors in a combination selected based on PCA should not be strongly correlated, these findings suggest the possibility that compounds with different activity studied here are distinguished by a number of features some of which are sensitive to different descriptors. However, different descriptor combinations with comparable performance may also capture similar molecular characteristics.

Our current analysis suggests that structural keys are powerful 2D descriptors, as suggested by other studies,[4,8] probably because they implicitly capture a variety of molecular properties. However, their presence is not critical for high performance. For example, as shown in Table 4, study 2 produced a variety of combinations of 2D descriptors and no MACCS keys that achieved between 88% and 91% prediction accuracy. Two of these combinations consisted of only three to four descriptors (consistent with results of study 1). Moreover, on the basis of S_1 scores, the complex MACCS descriptor does not occur in any of the top five combinations. Compared to study 2, further improvement of prediction accuracy by overall 5% could only be achieved by treating structural keys as single descriptors. Thus,

Table 4. Comparison of Previously Identified and Current Top Five Descriptor Combinations for PCA-Based Compound Classification According to (A) S_1 and (B) S_2 Scores[a]

(A) S_1 Scores

molecular descriptors	C_p	C_m	C_s	S_1	A (%)
Current Study					
a_nI, b_triple, f_conh2, 25, 47, 53, 61, 62, 65, 81, 87, 122, 124, 158, 159	55	0	17	3.24	96.3
VAdjEq, a_don, 28, 34, 62, 64, 71, 81, 107, 110, 122, 159, 161	64	0	22	2.91	95.2
PEOE_VSA+5, a_aro, b_ar, f_conh2, vsa_acid, 11, 30, 33, 62, 63, 64, 94, 99, 134, 165	73	0	26	2.81	94.3
PEOE_VSA_PPOS, a_aro, a_don, 36, 63, 83, 105	70	0	26	2.69	94.3
a_aro, 36, 40, 62, 76, 88, 96, 134, 142, 154	62	0	24	2.58	94.7
Study 2					
a_aro, f_c=o, LP, vsa_pol	73	0	41	1.78	91.0
a_nC, RPC-, KierA3, MACCS	100	7	58	1.47	85.7
chi0_C, chi1, a_nCl, PC+, PEOE_VSA+2, apol, MACCS	91	6	53	1.44	87.0
a_aro, PEOE_VSA+4, PEOE_VSA_PNEG, PEOE_VSA_POL, vsa_pol	91	0	63	1.44	86.2
a_aro, b_double, PEOE_VSA-6, PEOE_VSA_PNEG, I_so2nh2, vsa_acc	85	0	59	1.44	87.0
Study 1					
b_ar, SS, HB-a	75	4	34	1.01	74.9
b_1rotR,$^1\chi$, PEOE_PC+, SS,1κ, $^2\kappa$, $^3\kappa$	88	4	52	0.96	71.9
b_1rotR, b_ar, SS, HB-a	74	4	38	0.95	74.9
$^0\chi$, PEOE_PC+, SS, HB-a, CMR	96	4	71	0.86	68.6
b_ar, PEOE_PC+, SS, $^2\kappa$, $^3\kappa$	77	5	42	0.84	77.6

(B) S_2 Scores

molecular descriptors	N_p	C_p	N_s	N_m	C_m	S_2	A (%)
Current Study							
a_nI, b_triple, f_conh2, 25, 47, 53, 61, 62, 65, 81, 87, 122, 124, 158, 159	438	55	17	0	0	3.24	96.3
38, 59, 62, 69, 80, 96, 112, 127, 139, 145, 156	436	44	12	7	2	3.51	95.8
a_aro, 47, 62, 71, 112, 134	435	49	17	3	1	3.23	95.6
I_so2nh2, vsa_pol, 38, 62, 67, 71, 79, 106, 107, 110, 134, 164	435	54	18	2	1	3.14	95.6
I_so2nh2, b_ar, 47, 62, 80, 135, 159	434	47	21	0	0	3.11	95.4
Study 2							
a_aro, b_ar, b_double, b_triple, a_nP, f_so2nh2, LP	415	60	27	13	2	1.73	91.2
a_aro, f_c=o, LP, vsa_pol	414	73	41	0	0	1.59	91.0
a_aro, a_nO, LP	404	64	31	20	4	1.36	88.8
b_ar, b_double, a_nN, f_so2nh2	401	56	24	30	2	1.34	88.1
VdistMa, a_aro, b_triple, VadjEq, PEOE_VSA+2, PEOE_VSA+5, I_so2nh	402	64	36	17	6	1.30	88.4

[a] "C_p", "C_m", and "C_s" describe the number of pure classes, mixed classes, and singletons, respectively. Scores and "A" (overall prediction accuracy) were calculated as defined in the Methods section. "SS" represents the set of 57 structural keys shown in Table 1.

combinations of a few selected structural keys perform better than all keys together, and, in addition, combinations of a limited number of structural keys and 2D descriptors perform better than any combination of descriptors and the entire set of MACCS keys.

Application to the Design of Mini-Fingerprints. If combinations of relatively few 2D descriptors and structural keys can effectively partition different sets of bioactive compounds, as discussed above, then such combinations, if encoded as search tools, should also be capable of recognizing molecules with similar activity by database searching. We investigated this hypothesis by encoding preferred descriptor combinations as binary MFPs, as illustrated in Figure 2B. The initial design of mini-fingerprints[1] that produced MFP1 was based on our earlier studies[8] (study 1 in Tables 1 and 4). MFP1 consists of three 2D descriptors accounting for the fraction of rotatable bonds in a molecule and the number of aromatic bonds and hydrogen bond acceptors and, in addition, a set of 32 structural keys. These keys were selected based on their occurrence in at least 10% but not more than 90% of compounds in large databases.[8] When searching a subset of our benchmark database, MFP1 performed better than the structural keys or other MFPs even though it consisting of only a few 2D descriptors.[1]

Based on the results of this study, we have constructed a new mini-fingerprint by focusing on the structural key component. Therefore, structural keys within the top six descriptor combinations in Table 2 were assembled, resulting in a set of 40 preferred structural keys, and these keys were used to replace the set of 32 keys in MFP1. The selected structural keys are described in Table 3 of the Supporting Information. Only four keys were common to both sets. The new mini-fingerprint, termed MFP2, consists of 62 bit positions, and its design is illustrated in Figure 2B.

Fingerprint Performance. As a first step to test and compare the performance of these MFPs, we carried out exhaustive similarity searching in our benchmark database consisting of 455 compounds. In these search calculations, each compound was removed once from the database and searched four times against the remaining compounds using different Tc cutoff values (0.85, 0.80, 0.75, 0.70) for fingerprint overlap, thereby covering a Tc range suitable for similarity searching.[1] The results are shown in Table 5. MFP performance increases from about 24% to 59% correctly identified compounds (with decreasing Tc cutoff values). On average, only ~0.5% false positives are detected. Both MFPs outperform the reference fingerprint (Table 5). MFP2 performs consistently better than MFP1 in our benchmark database, by 3% to 8%, and reaches a maximum performance of 58.9% correctly and 0.9% incorrectly identified compounds at Tc value of 0.70. The improved performance of MFP2 was expected because its design was based on a set

Table 5. MFP Performance and Different Similarity Tc Cutoff Values for Compounds in Seven Biological Activity Classes (Benchmark Database) Used To Identify Preferred Descriptor Combinations[a]

fingerprint	Tc	correct (%)	incorrect (%)
MFP1	0.85	24.2	0.0
MFP2	0.85	27.2	0.0
MFP1	0.80	31.8	0.2
MFP2	0.80	38.4	0.1
MFP1	0.75	41.9	1.1
MFP2	0.75	49.9	0.2
MFP1	0.70	54.0	2.4
MFP2	0.70	58.9	0.9

[a] In this range, the best performance of reference fingerprint PH2D was 21.4% correctly identified compounds and one false positive (Tc of 0.70).

Table 6. Biological Activity of Compound Classes for Similarity Search Calculations[a]

source	biological activity	number of compounds
CH	β-lactamase inhibitors	14
CH	protein kinase C inhibitors	15
CMC	estrogen antagonists	11
CMC	histamine H2-antagonists	12
CMC	antihypertensive (ACE inhibitor)	17
CMC	antiadrenergic (β-receptor)	16
CMC	glucocorticoid analogues	14

[a] The test database consists of compounds belonging to seven different biological activity classes. "CMC" indicates that compounds were collected from the Comprehensive Medicinal Chemistry[16] database, and "CH" means that the source of compounds was the Chapman and Hall[17] collection. Thus, these two classes consist entirely of naturally occurring molecules.

Table 7. MFP Performance at Different Similarity Cut-Off Values in a Test Database Consisting of Compounds with Different Activity[a]

fingerprint	Tc	correct (%)	incorrect (%)
MFP1	0.85	31.8	0.0
MFP2	0.85	35.0	0.0
MFP1	0.80	41.3	0.1
MFP2	0.80	39.2	0.1
MFP1	0.75	48.6	0.5
MFP2	0.75	44.5	0.3
MFP1	0.70	56.8	1.3
MFP2	0.70	52.2	1.2

[a] Blind test calculations were carried out in the compound database shown in Table 6. The best performance of PH2D in this ranges was 14.4% correctly identified compounds with no false positives (Tc of 0.70).

of structural keys that were, as described above, specifically selected on the basis of our partitioning experiments.

A more challenging test case was to test MFP performance using classes of compounds for which descriptors were not specifically selected. Such calculations were expected to provide some answers to the question whether such MFPs have more general value as similarity search tools or whether their performance is largely database-dependent. Therefore, we assembled a test database consisting of 99 compounds belonging to seven different activity classes, as summarized in Table 6. None of these compounds was previously tested or used to select descriptor combinations. Since this database consists of drug-like molecules and natural products, we consider it a relatively difficult test case. Results of systematic search calculations in this database are reported in Table 7. In this case, the performance of both MFPs is quite similar (differences are 2% to 4% over the scoring interval) and ranges from ~32% to 57% correctly identified compounds with a maximum of about 1% false positives. Both MFPs perform significantly better than the reference fingerprint. Overall, MFP performance in these blind test calculations is comparable with the one achieved when searching our benchmark database (Table 5). In both cases, at optimum performance level, our prototypic MFPs have a greater than 50% chance to correctly identify compound with similar activity and detect only about 1% false positives, regardless of the compound database. Thus, these findings suggest that such MFPs should be useful tools to search databases for compound with activity similar to query molecules.

CONCLUSIONS

In this study, we have evaluated a total of 208 (1D/2D and structural-key type) descriptors for their performance in activity-based compound classification and have been able to improve the prediction accuracy of the PCA-based partitioning approach to better than 95%. In conjunction with our earlier study, we conclude that a considerable number of combinations of relatively few descriptors are capable of detecting and discriminating molecular features that are responsible for a specific biological activity. We have applied these findings and developed mini-fingerprints to search databases for molecules with similar activity. MFPs are much smaller and conceptually simpler than other more widely used fingerprint representations. In test calculations, these MFPs were found to have a greater than 50% chance to correctly identify such molecules. Equally important, they detected only a few false positives. We think that MFPs perform well in searches for molecules with similar activity because they evaluate compounds at a medium level of "chemical resolution". For carefully selected descriptor combinations, this level is sufficient to capture molecular features critical for a specific activity but not too sensitive to, for example, minor structural differences which are tolerated within a series of compounds having similar specificity. Taken together, our findings also suggest that application of relatively simple molecular descriptors and their combinations would be sufficient to successfully analyze a wide range of structure−activity relationships.

Supporting Information Available: Tables of descriptors in PCA-based compound classification and tables reporting the performance of top scoring descriptor combinations in the absence of single compound classes. References 19−21 occur only in tables of the Supporting Information. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Xue, L.; Godden, J.; Bajorath, J. Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.

(2) James, C. A.; Weininger, D. *Daylight theory manual*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1995.

(3) UNITY; Tripos, Inc.: St. Louis, MO, 1995.

(4) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D molecular descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.

(5) Brown, R. D. Descriptors for diversity analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.
(6) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using MDL "Keys" as structural descriptors. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 443−448.
(7) *MACCS keys*; MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA 94577.
(8) 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.
(9) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal component analysis and partial least squares regression. *Tetrahedron Comput. Methodol*. **1989**, *2*, 349−376.
(10) Labute, P. QuaSAR−Cluster: A different view of molecular clustering. *J. Chem. Comput. Group.* http://www.chemcomp.com/article/cluster.htm.
(11) Forrest, S. Genetic algorithms − Principles of natural selection applied to computation. *Science* **1993**, *261*, 872−878.
(12) 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.
(13) MOE (Molecular Operating Environment); version 1999.05; Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.
(14) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(15) Willett, P.; Winterman, V.; Bawden, D. Implementation of nonhierarchical clustering analysis methods in chemical information systems: Selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.
(16) *CMC-3D (Comprehensive Medicinal Chemistry Database*); version 99.1; MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA 94577.
(17) *Chapman & Hall, Dictionary of Natural Products*; CD-ROM version 1999; CRC Press LLC: 2000 NW Corporate Blvd, Boca Raton, FL 33431.
(18) Sanatvy, M.; Labute, P. SVL: The Scientific Vector Language. *J. Chem. Comput. Group.* http://www.chemcomp.com/feature/svl.htm.
(19) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(20) Kier, L. B. Indexes of molecular shape from chemical graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.
(21) 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.

CI000327J