Accurate Partitioning of Compounds Belonging to Diverse Activity Classes

Ling Xue[†] and Jürgen Bajorath*[,#,†]

Albany Molecular Research, Inc. (AMRI), Bothell Research Center, 18804 North Creek Pkwy, Bothell, Washington 98011, and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received December 1, 2001

Diverse sets of compounds were classified according to biological activity by use of a partitioning approach based on principal component analysis in conjunction with a genetic algorithm for molecular descriptor evaluation. Combinations of 236 molecular property and structural key descriptors were explored for their performance in classifying 317 molecules belonging to 21 distinct biological activity classes from various sources. Preferred descriptor combinations were further explored by complete factorial analysis. In these calculations, compounds having similar specific activity were predicted with greater than 80% accuracy.

INTRODUCTION

Compound classification techniques are typically based on clustering or partitioning algorithms.[1−3] Cluster analysis or partitioning is frequently used to group molecules according to chemical properties or biological criteria, identify active compounds in databases, distinguish between different activity classes, or identify representative compound subsets.[2,3] In these calculations, multidimensional chemical space is initially defined by selection of a number of molecular descriptors. Clustering of compounds in chemical space involves the calculation of intermolecular distances so that compounds that are close to each other may be combined into clusters. By contrast, in partitioning, chemical space is divided into sections based on ranges of calculated descriptor values, and compounds that fall into the same sections are combined. This process is critically influenced by binning of descriptor value ranges, which produces cells in chemical space.[4] Cell-based partitioning methods in low-dimensional chemistry spaces, in particular the BCUT metric,[5] and other partitioning concepts[6] have become popular compound classification tools in chemoinformatics and drug discovery.[2,3,7]

Cluster algorithms, on the other hand, are generally divided into two major classes, hierarchical and nonhierarchical methods. In hierarchical clustering, a first generation of clusters is combined or divided until a final result is obtained. By contrast, nonhierarchical clustering proceeds in a single step that does not depend on relationships or hierarchy between clusters. For clustering of chemical structures, Ward's hierarchical method[8] and Jarvis-Patrick clustering,[9] a nonhierarchical approach, have long been among the most widely used methods.[10,11] In several studies, hierarchical methods, in particular Ward's clustering, gave superior results relative to nonhierarchical techniques.[11,12] However, for chemical applications, the performance of clustering or partitioning methods is highly influenced by the choice of molecular descriptors, which represents an important variable in comparative studies. In many cases, descriptors are still selected intuitively, rather than based on systematic evaluation.

Previously, we have reported a compound classification approach[13,14] that combines principal component analysis (PCA)[15] for cell-based partitioning with a genetic algorithm (GA)[16] for systematic descriptor selection and evaluation of PCA calculation parameters. PCA removes the correlation between selected descriptors and derives an orthogonal lower-dimensional space defined by principal components (PCs) that are linear combinations of the original descriptors. Binning of PC axes produces cells for partitioning of compounds according to their PC values. We have applied this methodology to classify compounds in a compound database consisting of seven biological activity classes (including different enzyme inhibitors and receptor ligands). The number of single molecular descriptors for evaluation was increased in subsequent studies to more than 200,[13,14] and, ultimately, greater than 95% classification or prediction accuracy was achieved.[14]

Our major concern with these findings has been the potential database-dependence of prediction accuracy, since only seven biological activity classes were initially used and descriptor combinations were specifically "trained" to distinguish between these sets. We have now addressed these issues by substantially expanding our test database from seven to 21 classes of molecules with drug-like properties and by carrying out a de novo analysis on these classes. Using the combined GA-PCA approach, further complemented by factorial analysis of preferred descriptor combinations, we have achieved greater than 80% classification accuracy in these calculations. Thus, our findings suggest that accurate activity assignments can be obtained for different and increasingly challenging data sets and support the predictive value of the partitioning method investigated here. In addition, the obtained results make it possible to draw some conclusions regarding selection of descriptor combinations for compound classification.

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jurgen.bajorath@albmolecular.com.
[†] Albany Molecular Research, Inc. (AMRI), Bothell Research Center.
[#] University of Washington.

10.1021/ci010248n CCC: $22.00
Published on Web 03/30/2002

Table 1. Compound Classes with Specific Biological Activity[a]

biological activity	code	no. of compds
1. cyclooxygenase-2 (Cox-2) inhibitors	COX	17
2. tyrosine kinase (TK) inhibitors	TKE	20
3. HIV protease inhibitors	HIV	18
4. H3 antagonists	H3E	21
5. benzodiazepine receptor ligands	BEN	22
6. serotonin receptor ligands (5-HT)	5HT	21
7. carbonic anhydrase II inhibitors	CAE	22
8. β-lactamase inhibitors	BLC	14
9. protein kinase C inhibitors	PKC	15
10. estrogen antagonists	ESTR	11
11. antihypertensive (ACE inhibitor)	ACE	17
12. antiadrenergic (β-receptor)	ADR	16
13. glucocorticoid analogues	GLU	14
14. angiotensin AT1 antagonists	ANG	10
15. aromatase inhibitors	ARO	10
16. DNA topoisomerase I inhibitors	TOP	10
17. dihydrofolate reductase inhibitors	DIH	11
18. factor Xa inhibitors	FAC	14
19. farnesyl transferase inhibitors	FAR	10
20. matrix metalloproteinase inhibitors	MAT	12
21. vitamin D analogues	VIT	12

[a] Compound sets 1−7 were assembled from the literature as described,[17] and sets 8 and 9 were collected from the Chapman and Hall compendium[20] (thus, they exclusively consist of natural products). Compound sets 10−13 and 14−21 were taken from the Comprehensive Medicinal Chemistry database[18] and Synthline,[19] respectively (and thus consist of drug-like molecules or known drugs).

MATERIALS AND METHODS

Compound Database. Our database consists of 317 molecules belonging to 21 distinct biological activity classes. Only compounds with previously reported activity were selected. The composition of the benchmark database is reported in Table 1 and Figure 1 shows an example structure for each of the 21 activity classes, illustrating the diverse nature of compound classes investigated herein as well as the complexity of some of these molecules. Seven of the 21 activity classes were taken from our initial studies and assembled from the literature as described.[17] The other classes were either assembled from the Comprehensive Medicinal Chemistry database,[18] the drug subset of Synthline,[19] or the Chapman & Hall dictionary of natural products.[20] Thus, our test database includes compounds with specific activity from both synthetic and natural sources. For some classes (e.g., carbonic anhydrase inhibitors), a large number of active molecules could be obtained, whereas for others, only a fairly limited number was available. However, to balance the composition of our test database (and thus avoid prevalence of only a few classes), we selected between 10 and 22 molecules in each case. Since we wanted to avoid that activity classes predominantly consisted of very similar analogues, we did not include many activity classes available in commercial databases consisting of drugs or drug-like molecules.

Molecular Descriptors. The descriptors investigated in our analysis can roughly be divided into three different categories, diverse 1D and 2D property descriptors,[21,22] structural keys (molecular fragment-type descriptors),[23,24] and a new class of implicit 3D descriptors[25] that map various properties (e.g., partial charge) on van der Waals surfaces approximated from 2D representations of molecules. Of the publicly available set of 166 MACCS keys,[23] 147 were included in our evaluation, excluding those that were present in either all or none of our database compounds. Furthermore, a set of 63 diverse 1D/2D molecular property descriptors[14] and 26 complex surface descriptors[25] were selected, yielding a total of 236 descriptors that were treated as independent descriptors in all calculations. Table 2 defines the molecular property and surface descriptors, and Table 3 lists those structural keys that occurred in preferred descriptor combinations identified in this study, as discussed below. Descriptor values were calculated using the Molecular Operating Environment (MOE).[26]

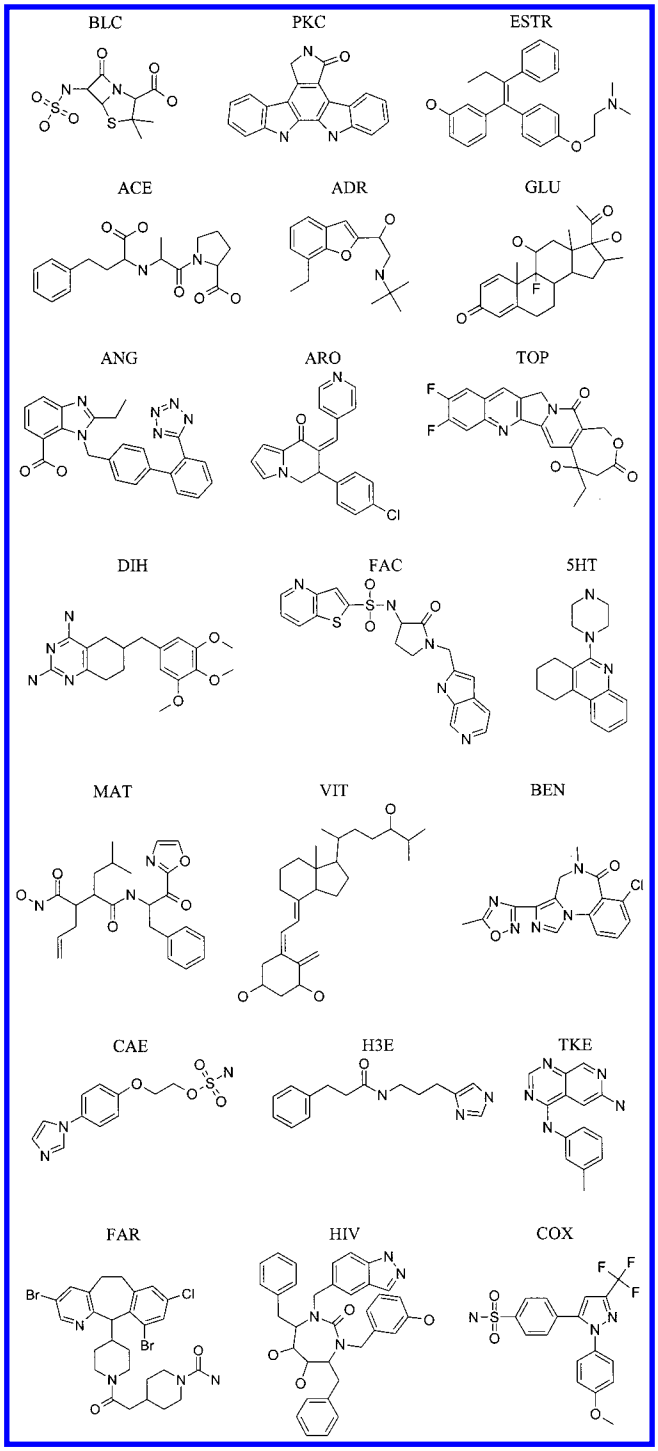

Figure 1. Representative structures. For each of the 21 compound classes in our benchmark database, an example structure is shown. Abbreviations for biological activity classes are given according to Table 1.

Table 2. Property Descriptors Evaluated in This Study and Complex van der Waals Surface Area Descriptors

descriptor	definition
	a. Property Descriptors[a]
logP(o/w)	log of the octanol/water partition coefficient
SlogP	log of the octanol/water partition coefficient[28]
density	molecular weight divided by the van der Waals volume
SMR	molecular refractivity[28]
apol	sum of the atomic polarizabilities of all atoms
Fcharge	total molecular charge (sum of formal charges)
PC+	sum of the positive q_i
PC-	sum of the negative q_i
RPC+	the largest positive q_i divided by the sum of the positive q_i
RPC−	the smallest negative q_i divided by the sum of the negative q_i
a_aro	number of aromatic atoms
HB-a	rule-based definition of the number of hydrogen bond acceptors[17]
HB-don	rule-based definition of the number of hydrogen bond donors[17]
a_nC	number of carbon atoms
a_nN	number of nitrogen atoms
a_nO	number of oxygen atoms
a_nF	number of fluorine atoms
a_nP	number of phosphorus atoms
a_nS	number of sulfur atoms
a_nCl	number of chlorine atoms
a_nBr	number of bromine atoms
a_nI	number of iodine atoms
a_acc	number of H-bond acceptors[26]
a_don	number of H-bond donors[26]
b_heavy	number of bonds between heavy atoms
b_rotN	number of nonring bonds
b_rotR	fraction of nonring bonds
b_1rotN	number of single nonring bonds
b_1rotR	fraction of single nonring bonds
b_ar	number of aromatic bonds
b_double	number of double nonaromatic bonds
b_triple	number of triple bonds
f_c=o	number of C═O group
f_conh2	number of CONH2 group
f_nh2	number of primary NH2 group
f_so2n	number of SO2N group
f_so2nh2	number of SO2NH2 group
i_so2nh	indicator variable for SO2NH group
i_so2nh2	indicator variable for SO2NH2 group
chi0_C	sum of the inverse square roots of the d_i of the carbon atoms
chi1	sum of the inverse square roots of $d_i d_j$ for all bonded heavy atoms i and j
chi1_C	sum of the inverse square roots of $d_i d_j$ for all bonded carbon atoms i and j
Kier2	$(n-1)(n-2)^2/p_2 2$ (Kier shape index[29])
KierA2	$(s-1)(s-2)^2/p_2 2$, where $s = n + a$
KierA3	$(s-1)(s-3)^2/p_3 2$ for odd n, and $(s-3)(s-2)^2 / p_3 2$ for even n, where $s = n + a$
radius	if r_i is the largest distance matrix entry in row i of A, then the radius is defined as the smallest of the r_I
petitjean	shape descriptor[30]
VDistMa	if m is the sum of the distance matrix entries, then VDistMa is defined as the sum of $a_{ij} * \log_2 a_{ij}/m - \log m$ over all i and j
VAdjMa	vertex adjacency information (magnitude): $1 + \log_2 m$ where m is the number of heavy-heavy bonds
VAdjEq	vertex adjacency information (equality): $-(1-f)\log_2(1-f) - f \log_2 f$ where $f = (n^2 - m)/n^2$, n is the number of heavy atoms and m is the number of bonds between heavy atoms
vsa_acc	approximation to the sum of VDW surface area of hydrogen-bond acceptors
vsa_pol	approximation to the sum of VDW surface area of polar atoms
vsa_acid	approximation to the sum of VDW surface area of acidic atoms
vsa_base	approximation to the sum of VDW surface area of basic atoms
	b. Complex van der Waals Surface Area Descriptors[b]
PEOE_VSA+6	sum of v_i where p_i is greater than 0.3
PEOP_VSA+5	sum of v_i where p_i is in the range [0.25,0.30)
PEOE_VSA+4	sum of v_i where p_i is in the range [0.20,0.25)
PEOE_VSA+3	sum of v_i whre p_i is in the range [0.15,0.20)
PEOE_VSA+2	sum of v_i where p_i is in the range [0.10,0.15)
PEOE_VSA+1	sum of v_i where p_i is in the range [0.05,0.10)
PEOE_VSA+0	sum of v_i where p_i is in the range [0.00,0.05)
PEOE_VSA-0	sum of v_i where p_i is in the range [−0.05,−0.00)
PEOE_VSA-1	sum of v_i where p_i is in the range [−0.10,−0.05)
PEOE_VSA-2	sum of v_i where p_i is in the range [−0.15,−0.10)
PEOE_VSA-3	sum of v_i where p_i is in the range [−0.20,−0.15)
PEOE_VSA-4	sum of v_i where p_i is in the range [−0.25,−0.20)
PEOE_VSA-5	sum of v_i where p_i is in the range [−0.30,−0.25)
PEOE_VSA-6	sum of v_i where p_i is less than −0.30
PEOE_VSA_PPOS	sum of v_i where p_i is greater than 0.2
PEOE_VSA_PNEG	sum of v_i where p_i is less than −0.2
PEOE_VSA_POL	sum of v_i where $\|p_i\|$ is greater than 0.2
SlogP_VSA0	sum of v_i such that $L_i <= -0.4$
SlogP_VSA1	sum of v_i such that L_i is in (−0.4,−0.2]
SlogP_VSA2	sum of v_i such that L_i is in (−0.2,0]
SlogP_VSA3	sum of v_i such that L_i is in (0,0.1]

Table 2 (Continued)

descriptor	definition
SlogP_VSA4	sum of υ_i such that L_i is in (0.1,0.15]
SlogP_VSA5	sum of υ_i such that L_i is in (0.15,0.20]
SlogP_VSA6	sum of υ_i such that L_i is in (0.20,0.25]
SlogP_VSA7	sum of υ_i such that L_i is in (0.25,0.30]
SlogP_VSA8	sum of υ_i such that L_i is in (0.30,0.40]
SlogP_VSA9	sum of υ_i such that $L_i > 0.40$
SMR_VSA0	sum of υ_i such that R_i is in [0,0.11]
SMR_VSA1	sum of υ_i such that R_i is in (0.11,0.26]
SMR_VSA2	sum of υ_i such that R_i is in (0.26,0.35]
SMR_VSA3	sum of υ_i such that R_i is in (0.35,0.39]
SMR_VSA4	sum of υ_i such that R_i is in (0.39,0.44]
SMR_VSA5	sum of υ_i such that R_i is in (0.44,0.485]
SMR_VSA6	sum of υ_i such that R_i is in (0.485,0.56]
SMR_VSA7	sum of υ_i such that $R_i > 0.56$

[a] q_i is the partial charge of atom i in a molecule, p_i represents the partial charge of atom i calculated using the PEOE method,[31] and υ_i is the van der Waals (VDW) surface area of atom i. d_i is the number of heavy atoms bonded to atom i. $\upsilon_i = (p_i - h_i)/(z_i - p_i - 1)$, where p_i is the number of s and p valence electrons of atom i, h_i is the number of hydrogen bonded to atom i, and z_i is the atomic number of atom i. n is the number of atoms in the non-hydrogen graph of the molecule, m is the number of bonds, and a is the sum of $(r_i/r_c - 1)$. r_i is the covalent radius of atom i, and r_c is the covalent radius of a carbon atom. p_2 is the number of paths of length 2, and p_3 is the number of paths of length 3. The graph distance matrix of a molecule with n (non-hydrogen) atoms is defined as the $n \cdot n$ matrix A, where a_{ij} is the length of the shortest path in graph between atoms i and j. [b] The surface are descriptors[25] are based on accessible van der Waals surface area calculated for each atom, υ_i, along with another atomic property, p_i. The υ_i values were calculated using a 2D connection table approximation. Each descriptor in a series is defined as the sum of υ_i values over all atoms i such that p_i is in a specified range $(a,b]$. Partial charge values were calculated using the PEOE method.[31] L_i denotes the contribution to logP(o/w) for atom i calculated using SlogP.[28] R_i captures the contribution to molar refractivity for atom i as calculated using the SMR formalism.[28]

Table 3. Definition of Selected Structural Keys

MACCS no.	definition
19	seven member ring
25	carbon atom bonded to three nitrogen atoms
26	fused ring
33	nitrogen atom bonded to sulfur atom
51	sulfur atom bonded to both carbon and oxygen atom
56	three- or four-coordinated nitrogen bonded to two oxygen and one carbon atom
58	sulfur atom bonded to two heteroatoms
62	ring attached to another ring through a nonring bond
71	nitrogen atom bonded to oxygen atom
72	two oxygen atoms positioned two atoms away from each other
75	nitrogen atom bonded to a ring and a nonring atom
76	three or four-coordinated carbon atom doubly bonded to one carbon atom
77	two nitrogen atoms bonded to the same atom
79	two nitrogen atoms separated by two other atoms
80	two nitrogen atoms separated by three other atoms
81	sulfur atom bonded to two heavy atoms
83	five member heterocyclic ring
85	three or four-coordinated nitrogen bonded to three carbon atoms
86	two methylene groups attached to one heteroatom
88	sulfur present
91	a heteroatom with hydrogen three bonds away from a methylene group
92	three- or four-coordinated carbon atom bonded to one C, one N, and one O
96	five member ring
98	six member heterocyclic ring
105	three atoms all in the rings and bonded to a central atom also in a ring
106	three- or four-coordinated heavy atom bonded to three heteroatoms
110	carbon atom bonded to a nitrogen and an oxygen atom
111	an atom bonded to a methylene group and nitrogen atom
116	a methyl and a methylene group three bonds away
117	nitrogen and oxygen bonded to the same atom
120	more than one heteroatom
122	three- or four-coordinated nitrogen
124	two bonded heteroatoms
125	more than one aromatic ring
128	two methylene groups four bonds away
130	more than one pair of heteroatoms that are connected
131	more than one H attached to a heteroatom
136	double bonded oxygen
137	heteroatoms in ring
140	more than three oxygen atoms
151	NHx group
158	C−N single bond
163	six member ring

Compound Classification. The GA-PCA algorithm, a flowchart of which is shown in Figure 2, was implemented in MOE using SVL[27] code and makes use of MOE's built-in PCA function. This figure also shows a schematic view of the chromosome designed for GA calculations. Each of 236 bits in segment I was assigned to one of the descriptors

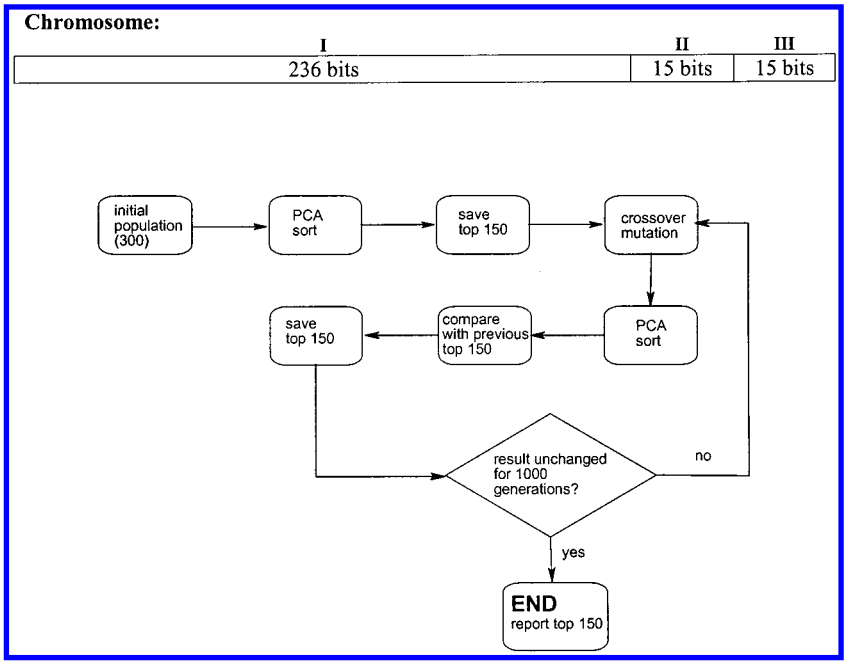

Figure 2. GA-PCA method. At the top, the outline of a chromosome designed to encode molecular descriptors and PCA calculation parameters is shown. It consists of three segments, which account for the presence (1) or absence (0) of single descriptors (I; 236 bits), the number of principal components (II; 15 bits), and the number of axis intervals or bins (III; 15 bits). If set on (1), each bit position in these segments adds a principal component or bin to the calculation. At the bottom, a flowchart of a genetic algorithm is shown that illustrates how principal component analysis and scoring function were embedded in these calculations.

evaluated here, encoding its inclusion or omission from the calculation, and 30 bits (segments II and III) of control PCA parameters, encoding the number of PCs and PC axis bins, respectively. The combination of these parameters determines the number of cells for compound partitioning. In our calculations, both the number of PCs and bins was allowed to vary from 1 to 15. An initial population of 300 chromosomes was randomly generated with initial bit occupancy of 16% (determined on the basis of various test calculations[14]). Mutation and crossover rates were set to 5% and 25%, respectively. If a new chromosome had more than 20% of its bits set on after crossover and mutation, each bit was assigned a 50% probability to reach the next generation. As indicated in Figure 2, assessment of the calculations depends on implementation of a "sort" or scoring function to quantify the predictions and identify best performing descriptor combinations. For biological activity-oriented compound classification, a desired result is to have as many compounds as possible in "pure" cells or classes (i.e., cells populated exclusively with compounds sharing similar activity), as opposed to "mixed" cells (i.e., consisting of compounds with different activity) or singletons ("unassigned" compounds). Therefore, the following scoring function was implemented:

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

N_p is the total number of the compounds in pure classes, N_m the number of compounds in mixed classes, and N_{total} the total number of compounds in the database. C is the total number of classes obtained by PCA analysis, and C_a is the number of different activity classes in the database. C_s here is the number of singletons. A factor of 100 was applied to scale the resulting values and thus obtain top scores greater than 1. It follows that S values are high if many compounds occur in a small number of pure classes. In addition, "overall prediction accuracy" was defined as $A = N_p/N_{total}$. Following PCA, scores were calculated for the descriptor combination encoded in each chromosome, the top scoring 150 chromosomes were saved and again subjected to crossover and mutation operations, and the resulting chromosomes represented the next generation. We considered the calculations converged, if results remained unchanged for at least 1000 generations. In this study, convergence was reached after a total of 59 250 generations. Preferred descriptor combinations were further explored by complete factorial analysis using the same PCA and scoring scheme.

RESULTS AND DISCUSSION

GA-PCA Analysis. The results of our extensive GA-PCA calculations are summarized in Table 4, sorted according to obtained scores. As can be seen, the top 20 descriptor combinations, consisting of six to 17 descriptors, achieve prediction accuracy between 70% and 74.4%. For random predictions on this data set, less than five percent prediction accuracy would be expected. A few clear trends can be observed. The number of mixed cells and compounds populating these cells is generally low for preferred descriptor combinations. However, the number of singletons, i.e., molecules not considered similar to any other database compound, is higher, for example, ~60 molecules for the top four descriptor combinations. This can in part be attributed to the fact that compounds within activity classes were intentionally selected to be quite diverse, rather than being analogues of single chemotypes (i.e., having common core structure or scaffold). For example, average Tc values for pairwise comparison of all compounds in each class using MACCS keys ranged from 0.58 to 0.79. Thus, the challenge here was not only to distinguish between activity classes but also to recognize molecular similarity within each class.

Table 4. Top 20 Descriptor Combinations Identified by GA-PCA Analysis[a]

	descriptors	nDS	PC	bins	score	PA	P	nP	S	M	nM
1	a_aro, 19, 47, 58, 65, 77, 96, 105, 116, 128, 131, 136, 163	13	8	8	0.85	74.4	70	236	65	3	16
2	a_aro, f_c=o, vsa_acid, 60, 75, 83, 88, 96, 110, 131, 137	11	9	8	0.83	73.8	68	234	59	4	24
3	a_aro, f_so2nh2, RPC-, PEOE_VSA+3, 75, 92, 105, 122	8	7	8	0.81	73.5	71	233	66	5	18
4	b_ar, f_so2n, 25, 29, 71, 76, 80, 86, 98, 131, 136, 158	12	8	8	0.80	73.2	61	232	62	6	23
5	a_aro, a_nP, 19, 80, 92, 106, 120, 130	8	10	11	0.79	72.6	60	230	32	10	55
6	a_aro, b_ar, 11, 19, 26, 43, 56, 62, 70, 73, 79, 104, 111, 122, 124, 140	16	10	10	0.77	72.6	68	230	79	3	8
7	a_aro, f_conh2, 26, 91, 96, SlogP_VSA1	6	10	13	0.75	71.9	73	228	75	3	14
8	a_aro, a_nS, b_1rotR, b_rotR, f_so2n, vsa_acid, 15, 33, 62, 81, 85, 91, 117, 125, 136, 151, 162	17	7	9	0.74	71.6	68	227	66	5	24
9	19, 33, 48, 62, 72, 80, 84, 85, 86, 110, 151, 158, 162, SlogP_VSA7	14	6	6	0.74	71.6	66	227	68	6	22
10	a_nI, b_ar, RPC+, PEOE_VSA-5, 15, 51, 75, 77, 80, 83, 131, 151	12	8	10	0.74	71.6	70	227	71	6	19
11	b_ar, density, f_so2n, RPC+, 8, 19, 40, 48, 55, 72, 78, 98, 126, 136, 152	15	5	6	0.74	71.3	62	226	47	11	44
12	19, 27, 45, 53, 56, 79, 81, 94, 104, 132, 136, 144, 158, 163, SMR_VSA4	15	5	9	0.74	71.6	66	227	81	3	9
13	a_nS, b_ar, PEOE_VSA-5, 38, 49, 58, 62, 67, 84, 105, 155	11	10	10	0.73	71.3	67	226	70	5	21
14	a_nP, RPC-, 23, 26, 30, 32, 52, 80, 96, 106, 131, 144, 146, 152, 154, 157, 159	17	8	8	0.72	70.7	60	224	48	10	45
15	I_so2nh, a_nP, f_c=o, RPC-, 19, 29, 53, 69, 81, 98, 153, 156, SMR_VSA4	13	8	8	0.72	71.0	68	225	77	5	15
16	b_rotR, 13, 30, 45, 62, 68, 83, 84, 106, 117, 126, 127, 128, 136, 138, 146, 151	17	8	7	0.72	71.0	70	225	78	5	14
17	a_nP, b_ar, f_c=o, 19, 20, 40, 47, 63, 67, 80, 94, 102, 106, 115, 123, 125, 157	17	7	7	0.71	71.0	68	225	83	3	9
18	17, 26, 37, 62, 73, 74, 84, 98, 99, 106, 108, 110, 120, 123, 127, 145, 156	17	8	8	0.71	70.7	62	224	76	5	17
19	a_aro, 26, 47, 52, 85, 94, 106, SlogP_VSA1	8	7	5	0.71	70.3	61	223	47	10	47
20	b_ar, PEOE_VSA+3, 29, 72, 75, 80, 92, 99, 136	9	6	7	0.71	70.7	67	224	78	4	15

[a] "nDs" is the number of descriptors in each combination, "PC" is the number of principal components, "Bins" is the number of axis intervals for the calculation identifying the preferred descriptor combination, "PA" stands for prediction accuracy, "P" is the number of pure and "M" is the number of mixed classes, "S" is the number of singletons, and "nP" and "nM" give the total number of compounds in pure and mixed classes, respectively. Descriptors are abbreviated according to Table 2, and numbers indicate structural keys as defined in Table 3.

Failure to do so is expected to result in an increase in the number of singletons. This also means that effective descriptor combinations must achieve a balance between the level of chemical detail they capture and the detection of key features that determine specific activities and distinguish activity classes from each other. In general, we find that preferred descriptor combinations consist of both structural keys and property descriptors, with keys representing the majority of descriptors within the top scoring combinations. Structural keys that occur within the top 30 combinations are described in Table 3, representing a subset of 43 preferred MACCS keys. Selected property descriptors are also recurrent among the top scoring combinations, in particular a simple descriptor counting the number of aromatic atoms in a molecule and, in addition, some of the complex van der Waals surface descriptors.[25] Several preferred descriptor combinations are similar, whereas others are distinct from each other, indicating that reasonable predictive performance in these partitioning calculations can be achieved in a variety of ways.

Calculation Parameters. The data in Table 4 also reveal that preferred descriptor combinations were obtained using a relatively large number of PCs and bins for compound partitioning, on average eight in each case. We have analyzed the reasons for these findings by systematic variation of both the number of principal components and bins for PCA of top scoring descriptor combinations. Representative results are reported in Table 5. As can be seen, inclusion of the first five principal components, which together account for greater than 80% of the data variance in our benchmark database, with respect to the selected descriptor combination, yielded greater than 70% prediction accuracy, if eight bins were used per axis. However, to achieve the best performance (~74% prediction accuracy), it was necessary to account for more than 99% of the data variance, corresponding to nine principal components. We also found that the partitioning results were not greatly influenced by the number of bins, as long as a necessary minimum of cells was generated. For the example reported in Table 5, four bins per principal component axis were sufficient to produce 70% prediction accuracy, but inclusion of up to eight bins, the identified optimum value, increased performance by only ~3%. In turn, further increase in the number of bins to 12 reduced prediction accuracy by only ~1.5%. Thus, the majority of newly created cells for partitioning remained "empty". These control calculations rationalize the automatic parameter selection of the genetic algorithm calculations. Inclusion of, on average, eight principal components accounted for all, or almost all, of the variance within the data sets and produced slightly better scores and ~3% improved prediction accuracy than calculations based on approximately 80% coverage of data variance. In practical terms, these differ-

Table 5. Principal Components, Data Set Variance, and Predictive Performance[a]

PCs	var (%)	PA (%), four bins	PA (%), eight bins
1	27.0	0	0
2	47.4	3.2	16.4
3	64.2	19.6	47.3
4	75.3	48.9	63.4
5	83.7	67.8	71.0
6	89.5	68.4	72.2
7	94.0	68.7	72.2
8	97.2	68.7	72.3
9	99.4	71.9	73.8
10	100	70.7	73.8

[a] "PCs" gives the number of principal components included in the analysis, "Var" means variance within the benchmark database with respect to the original descriptor combination, and "PA" means prediction accuracy. For PCA, both the number of principal components and bins were systematically varied between 1 and 15. For each number of principal components, only two PA values are shown for clarity, one for four axis intervals and one for eight. In this case, optimum performance was observed for a combination of 9 PCs and 8 bins. As a representative example, data are reported for descriptor combination number two in Table 4 (i.e., a_aro, f_c=o, vsa_acid, 60, 75, 83, 88, 96, 110, 131, 137).

Table 6. Top Scores Obtained by Complete Factorial Analysis of Preferred Descriptor Combinations[a]

DS	descriptors by CFA	nDS	PC	bins	score	PA	P	nP	S	M	nM
6	a_aro, 19, 26, 56, 62, 79, 111, 122, 124, 140	10	7	4	1.22	80.6	69	256	38	7	23
10	b_ar, RPC+, PEOE_VSA-5, 51, 75, 77, 80, 83, 131	9	9	4	1.09	79.2	78	251	71	2	5
1	a_aro, 19, 58, 77, 105, 116, 128, 131, 136, 163	10	10	6	0.99	77.3	67	245	52	4	20
4	b_ar, f_so2n, 25, 71, 76, 80, 86, 98, 131, 136, 158	11	4	13	0.97	77.0	64	244	58	5	15
8	a_aro, b_1rotR, b_rotR, f_so2n, vsa_acid, 33, 62, 81, 91, 117, 125, 136, 151	13	7	14	0.95	76.3	67	242	49	5	26
9	19, 33, 62, 72, 80, 85, 86, 110, 158, SlogP_VSA7	10	6	5	0.91	75.7	70	240	52	6	25
2	a_aro, f_c=o, 75, 83, 88, 131, 137	6	4	9	0.89	75.1	67	238	43	9	36
3	a_aro, f_so2nh2, RPC-, PEOE_VSA+3, 75, 92, 105, 122	8	8	4	0.84	74.1	67	235	67	5	15
5	a_aro, a_nP, 19, 80, 92, 106, 120, 130	8	7	11	0.79	72.6	60	230	32	10	55
7	a_aro, f_conh2, 26, 91, 96, SlogP_VSA1	6	4	13	0.77	71.9	73	228	72	4	17

[a] Each of the top 10 descriptor combinations in Table 4 was subjected to complete factorial analysis, and, in each case, the best descriptor combination is reported. The resulting "secondary" combinations are sorted according to their scores. "DS" indicates the rank of the "primary" top 10 descriptor combinations within the original top 30 list in Table 4, and "CFA" stands for complete factorial analysis. All other abbreviations are used according to Table 4.

ences are significant, as they avoid prediction errors for more than 10 compounds.

Exploration by Complete Factorial Analysis. Since genetic algorithm calculations generally produce reasonable but not necessarily optimal solutions to problems under investigation,[16] we tried to further improve prediction accuracy by complete factorial analysis of descriptor combinations in Table 4 followed by PCA. In Table 6, the best combination obtained by complete factorial analysis of each of the original top 10 combinations is reported. In eight of 10 cases, prediction accuracy could be further improved by omission of several descriptors from the initial combinations. The top scoring combination, with 80.6% prediction accuracy, now consisted of a descriptor accounting for aromatic atoms and nine structural keys, and the second best combination, yielding 79.2% accuracy, consisted of four property descriptors and five structural keys. For comparison, the average prediction accuracy over 60 control calculations, using 10 randomly selected descriptors in each case, was 35%. Complete factorial analysis of the top combination in Table 4 led to the omission of three structural keys, which improved prediction accuracy from 74.4% to 77.3%, producing the overall third best score. For the top scoring combination identified by factorial analysis, omission of six descriptors led to an 8% increase in prediction accuracy. The data in Table 6 reveal another trend. For five of the seven best descriptor combinations, including the top scoring combination, improvements in scoring and prediction accuracy were achieved by decreasing the number of singletons more significantly than increasing the number of compounds in mixed classes. This is consistent with the idea that fewer descriptors produce a less complex and discriminatory "chemical grid" for compound classification. Thus, reducing the number of descriptors leads to improved recognition of compound similarity within activity classes but, on the other hand, slightly increases the probability of false-positive activity assignments. Therefore, finding a favorable balance between these effects is a critical factor for successful compound partitioning.

Conclusions. Much effort has been spent to assemble a benchmark database consisting of diverse molecules and activity classes from different sources. Partitioning of compounds belonging to 21 compound classes with distinct specific activities was thought to represent a rather challenging test case for molecular classification analysis. In our calculations, we were able to partition this database with up to ~80% prediction accuracy. Many preferred descriptor combinations were initially identified by GA-PCA analysis, and the best scoring combinations were further explored and improved by complete factorial analysis of descriptor combinations. The influence of PCA parameters on the accuracy of cell-based partitioning could be rationalized. Of 236 molecular descriptors evaluated in this study, only 10 descriptors were required to achieve the best predictions and several combinations yielded comparably high prediction accuracy. Consistent with previous findings, combinations of structural keys and a few property descriptors performed best. We deliberately excluded descriptors from this study that could only be calculated from 3D conformations of molecules, due to the uncertainty of predicting the bioactive conformations of our test compounds. However, the obtained results suggest that 3D descriptors may not be necessary to achieve reasonably accurate compound classifications. Some of the descriptors evaluated herein, including a set of 43 preferred structural keys, were recurrent within top scoring combinations and should thus be of more general interest in the study of similarities and differences between drug-like molecules.

ACKNOWLEDGMENT

The authors wish to thank F. Stahura and J. Godden for help in assembling the benchmark database and critical review of the manuscript.

REFERENCES AND NOTES

(1) Barnard, J. M.; Downs, G. M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.

(2) Mason, J. S.; Pickett, S. D. Partition-based selection. *Perspect. Drug Discov. Design.* **1997**, *7/8*, 85−114.

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

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

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

(6) 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.

(7) Schnur, D. Design and diversity analysis of large compound libraries using cell-based methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.

(8) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.

(9) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(10) Willett, P.; Wintermann, V.; Bawden, D. Implementation of nonhierarchic cluster 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.
(11) 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.
(12) Wild, D. J.; Blankley, C. J. Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155−162.
(13) 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.
(14) Xue, L.; Godden, J.; Bajorath, J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227−1234.
(15) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal component analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.
(16) Forrest, S. Genetic algorithms − Principles of natural selection applied to computation. *Science* **1993**, *261*, 872−878.
(17) 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.
(18) CMC-3D (Comprehensive Medicinal Chemistry Database), version 99.1; MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA.
(19) Synthline Drug Database on STN International, taken from Drugs of the Future (comprehensive drug monographs, Prous Science), 1984-present; Prous Science: Provenza 388, Barcelona, Spain.
(20) Chapman & Hall, Dictionary of Natural Products, CD-ROM version 1999; CRC Press LLC: 2000 NW Corporate Blvd., Boca Raton, FL, U.S.A.
(21) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.
(22) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combin. Chem. High Throughput Screen.* **2000**, *3*, 363−372.
(23) MACCS keys; MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA 94577.
(24) 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.
(25) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464−477.
(26) Molecular Operating Environment, version 2001.01; Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada.
(27) Sanatvy, M.; Labute, P. SVL: The Scientific Vector Language. *J. Chem. Computing Group*. http://www.chemcomp.com/feature/svl.htm.
(28) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(29) Kier, L. B. Indexes of molecular shape from chemical graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.
(30) 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.
(31) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

CI010248N