Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Component Analysis Identified by a Genetic Algorithm

Ling Xue[†] 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 January 6, 2000

We have evaluated combinations of 111 descriptors that were calculated from two-dimensional representations of molecules to classify 455 compounds belonging to seven biological activity classes using a method based on principal component analysis. The analysis was facilitated by application of a genetic algorithm. Using scoring functions that related the number of compounds in pure classes (i.e., compounds with the same biological activity), singletons, and mixed classes, effective descriptor sets were identified. A combination of only four molecular descriptors accounting for aromatic character, hydrogen bond acceptors, estimated polar van der Waals surface area, and a single structural key gave overall best results. At this performance level, ∼91% of the compounds occurred in pure classes and mixed classes were absent. The results indicate that combinations of only a few critical descriptors are preferred to partition compounds according to their biological activity, at least in the test cases studied here.

INTRODUCTION

Clustering or partitioning methods divide compound collections into different classes on the basis of specific structural features, molecular properties, and/or distinct biological activities.[1−9] Such methods are widely used to, for example, identify biologically active molecules in databases[1,8,9] or to aid in diversity analysis[4,10,11] and compound subset selection.[12,13] Successful application of these methods depends on the algorithms used[2,3] and also on the choice of descriptors that capture characteristic molecular features[12−15] and associate molecular similarity with biological activity.[16]

In a previous study,[8] we have investigated an algorithm to partition compounds based on principal component analysis (PCA)[17] and explored all possible combinations of 17 one-/two-dimensional (1/2D) molecular descriptors, including a set of 57 structural keys[6] (or fragments), to classify compounds belonging to seven biological activity classes. We found that structural keys in conjunction with two or three additional 2D descriptors effectively partitioned compounds according to biological activity.[8] We also compared PCA-based compound classification to other clustering approaches. Using preferred descriptor combinations, the performance of the PCA-based algorithm was at least comparable to Jarvis−Patrick clustering,[18] one of the most widely used nonhierarchical clustering methods.[1] Preferred descriptor combinations identified in our analysis were subsequently used to design small fingerprint representations of molecules[l9] for application in similarity searching,[20] aiming at the detection of compounds with specific activity similar to a query molecule.

We have now extended our evaluation of molecular descriptors for efficient PCA-based compound classification to the combinatorial analysis of a total of 111 descriptors, including 110 descriptors that could be calculated from 2D molecular representations and a set of 166 structural keys, treated as a single complex descriptor. Due to the significantly increased magnitude of the combinatorial problem, complete factorial analysis of descriptor combinations to partition compounds belonging to seven biological activity classes became a computationally nonfeasible task. Therefore, we implemented a genetic algorithm to identify desirable descriptor combinations. Our analysis is presented herein. Consistent with our initial study, we have found that combinations of only a few descriptors effectively partition compounds according to their specific biological activity. More than 30 novel combinations of descriptors with improved performance were identified.

METHODS

Computations. All programs required for the calculations reported herein, including scoring functions and a genetic algorithm, as described below, were generated using SVL code[21] and implemented in MOE.[22]

Compound Classification. For PCA-based[17] partitioning of compounds, the QuaSAR-Cluster function[23] of MOE was used as described.[8] The principle of the PCA-based method is illustrated in Figure 1. Briefly, molecules are expressed using descriptors as a vector in *n*-dimensional descriptor space. PCA performs a linear transformation of the *n*-vectors, reduces the dimensionality of the descriptor space, removes the correlation between descriptors, and estimates a normalized set of *p* principal components. Each of the principal

* To whom any correspondence should be addressed at New Chemical Entities, Inc. Telephone: (425) 424-7297. Fax: (425) 424-7299. E-mail: jbajorath@nce-mail.com.
[†] New Chemical Entities, Inc.
[‡] University of Michigan.

10.1021/ci000322m CCC: $19.00
Published on Web 04/04/2000

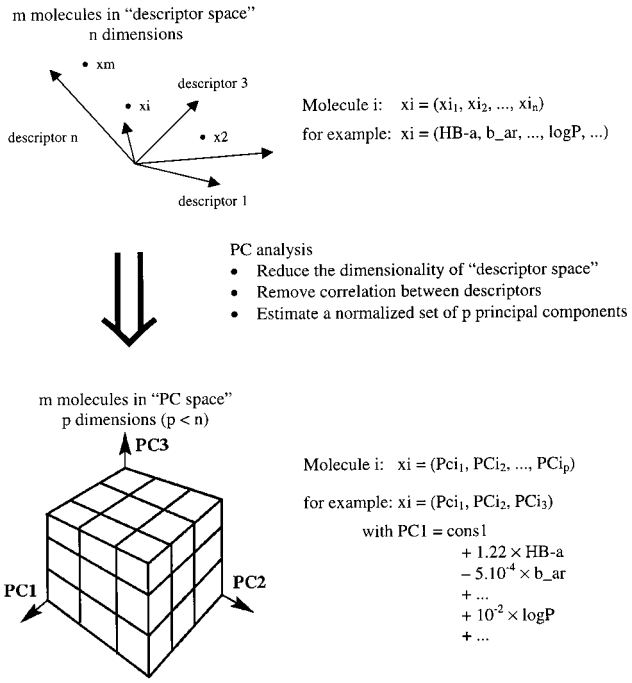

Figure 1. Principal component analysis. The schematic representation illustrates how compounds are classified (partitioned) on the basis of principal component analysis of molecular descriptor combinations.

Table 1. Composition of Test Database

biological activity	no. of compds	% of database
cyclooxygenase-2 (Cox-2) inhibitors	31	6.8
tyrosine kinase (TK) inhibitors	35	7.7
HIV protease inhibitors	48	10.5
H3 antagonists	52	11.4
benzodiazepine receptor ligands	59	13.0
serotonin receptor ligands (5-HT)	71	15.6
carbonic anhydrase II inhibitors	159	34.9

component axes is divided into intervals or bins, and coordinates of each molecule are assigned to an axis interval. Molecules within the same cell in principal component space form a class. A detailed description of the algorithm is available in an online publication.[23] Different from our initial study (where the number of principal components and axis bins were kept constant),[8] both principal components and bins were variable and could adopt values between 2 and 15.

Test Database. The same compound database as in our initial study was used, assembled from the literature as described,[8] so that results obtained for different descriptor combinations could be directly compared. This database consists of 455 compounds with "activity" against seven different targets (including diverse sets of enzyme inhibitors, receptor agonists, and antagonists). Table 1 reports the composition of the database.

Scoring Functions. The results of PCA-based compound classification were ranked using the following two functions we implemented:

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

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 singletons. A scaling factor of 10 was used to penalize the occurrence of mixed clusters. If only pure classes and no mixed classes or singletons occur, S_1 would not be defined. Thus, in these instances, the score would be set to

$$S_1 = 1 + 7/C_p$$

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

where N_p is the number of the compounds in pure classes, N_m is the number of compounds in mixed classes, N_{total} is the total number of compounds in the database, C is the total number of the classes obtained after PCA analysis, and C_a is the number of distinct activity classes in the database. The factor of 100 was introduced to make values of S_1 and S_2 comparable in their magnitude (i.e., approximately 1).

The scoring functions have different focal points. S_1 gives high values if the number of mixed classes is small, and S_2 is high if the number of compounds in the pure classes is large and the total number of classes is small. However, if compound classification is "successful" (i.e., few mixed classes and singletons), both scores are high.

Molecular Descriptors. We explored a total of 111 descriptors that could be calculated from 2D representations of molecules. Of these, 110 descriptors represent physicochemical parameters, 2D descriptors, and some implicit 3D descriptors. The evaluated descriptors are listed and classified in Table 2. Many but not all of these descriptors could be calculated with MOE, version 1999.05. As an aggregate descriptor, the complete set of 166 MACCS/MDL structural keys was added.[24] To separately test a limited number of single structural key-type descriptors, specific functional groups were also added, as shown in Table 2d, for example, the carbonyl or sulfonamide group. If all 166 MDL keys were treated as separate descriptors, increasing the total number of descriptors to more than 260, convergence could not be reached in our genetic algorithm calculations.

Genetic Algorithm. A genetic algorithm (GA)[25,26] was implemented to identify preferred descriptor combinations for PCA-based compound classification. A GA-based analysis was recently reported to correlate substructure analysis and biological activity profiles of compounds in chemical databases.[27] Our GA procedure is summarized in Figure 2. We generated a (binary bit string) chromosome consisting of a total of 141 bits: 110 descriptors (110 bits), 166 MDL keys (1 bit), principal component number (15 bits), and number of bins for each principal component (15 bits). Test calculations were carried out with different initial bit occupancy of chromosomes, followed by PCA, to determine the approximate scoring range for initially set bit densities. On the basis of these calculations, highest intermediate scores (i.e., scores monitored over the first several hundred generations) were obtained for bit occupancies at or below 10%. Above this level, intermediate scores generally decreased with increasing initial bit densities. Thus, to generate the initial chromosome population, each bit was allowed an 8% chance to be set on (i.e., 1). An initial population of 200 chromosomes was randomly generated. Each chromosome was ranked based on its S_2 score following PCA. The top

Table 2.

descriptor	definition
	(a) Physicochemical Properties and Charge Descriptors[a]
weight	molecular weight
mr	molar refractivity
log *P*(o/w)	log of the octanol/water partition coefficient
apol	sum of the atomic polarizabilities of all atoms
bpol	sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule
vdw_vol	van der Waals volume
density	molecular weight divided by the van der Waals volume
vdw_area	van der Waals surface area
Fcharge	total charge of the molecule (sum of formal charges)
PC+	sum of the positive q_I
PC−	sum of the negative q_I
RPC+	largest positive q_i divided by the sum of the positive q_i
RPC−	smallest negative q_i divided by the sum of the negative q_i
PEOE_PC+	sum of the positive p_i
PEOE_PC−	sum of the negative p_i
PEOE_RPC+	largest positive p_i divided by the sum of the positive p_i
PEOE RPC−	smallest negative p_i divided by the sum of the negative p_i
PEOE VSA+6	sum of ν_i where p_i is greater than 0.3
PEOE_VSA+5	sum of ν_i where p_i is in the range [0.25,0.30)
PEOE_VSA+4	sum of ν_i where p_i is in the range [0.20,0.25)
PEOE_VSA+3	sum of ν_i where p_i is in the range [0.15,0.20)
PEOE_VSA+2	sum of ν_i where p_i is in the range [0.10,0.15)
PEOE_VSA+l	sum of ν_i where p_i is in the range [0.05,0.10)
PEOE_VSA+O	sum of ν_i where p_i is in the range [0.00,0.05)
PEOE_VSA-0	sum of ν_i where p_i is in the range [−0.05,0.00)
PEOE_VSA-1	sum of ν_i where p_i is in the range [−0.10,−0.05)
PEOE_VSA-2	sum of ν_i where p_i is in the range [−0.15,−0. 10)
PEOE_VSA-3	sum of ν_i where p_i is in the range [−0.20,−0.15)
PEOE_VSA-4	sum of ν_i where p_i is in the range [−0.25,−0.20)
PEOE_VSA-5	sum of ν_i where p_i is in the range [−0.30,−0.25)
PEOE VSA-6	sum of ν_i where p_i is less than −0.30
PEOE_VSA_POS	sum of ν_i where p_i is greater than or equal to 0
PEOE_VSA_NEG	sum of ν_i where p_i is less than 0
PEOE_VSAPPOS	sum of ν_i where p_i is greater than 0.2
PEOE_VSA_PNEG	sum of ν_i where p_i is less than −0.2
PEOE_VSA_HYD	sum of ν_i where $\|p_i\|$ is less than or equal to 0.2
PEOE_VSA POL	sum of ν_i where $\|p_i\|$ is greater than 0.2
	(b) Atom, Bond, and Electron Pair Count Descriptors[b]
a_count	no. of atoms
a_heavy	no. of heavy atoms
a_aro	no. of aromatic atoms
HB-a	rule-based definition of the number of hydrogen bond acceptors[8]
HB-don	rule-based definition of the number of hydrogen bond donors[8]
LP	no. of lone pair electrons on oxygen and nitrogen atoms
a_nH	no. of hydrogen atoms
a_nC	no. of carbon atoms
a_nN	no. of nitrogen atoms
a_nO	no. of oxygen atoms
a_nF	no. of fluorine atoms
a_nP	no. of phosphorus atoms
a_nS	no. of sulfur atoms
a_nCl	no. of chlorine atoms
a_nBr	no. of bromine atoms
a_nI	no. of iodine atoms
a_ICM	entropy of the distribution of elements in the molecule
a_IC	no. of atoms in the molecule times a_ICM
b count	no. of bonds
b heavy	no. of bonds between heavy atoms
b_rotN	no. of nonring bonds
b_rotR	fraction of nonring bonds
b_lrotN	no. of single nonring bonds
b_lrotR	fraction of single nonring bonds
b_ar	no. of aromatic bonds
b_single	no. of single nonaromatic bonds
b double	no. of double nonaromatic bonds
b_triple	no. of triple bonds
VadjEq	entropy of the distribution of values in the adjacency matrix
VadjMa	log (2 times the number of bonds)

Table 2 (Continued)

descriptor	definition
(c) Connectivity Indices, Kier κ Shape Indices,[29] Graph Distance Matrix Descriptors, and Surface Area Descriptors[c]	
Zagreb	sum of the squares of the d_i of the heavy atoms
ChiO	sum of the inverse square roots of the d_i[29]
ChiOC	sum of the inverse square roots of the d_i of the cabon atoms
chi 1	sum of the inverse square roots of $d_i d_j$ for all bonded heavy atoms i and j
chi2	bond connectivity index (order 2)
chil_C	sum of the inverse square roots of $d_i d_j$ for all bonded carbon atoms i and j
chi0v	sum of the inverse square roots of the ν_1
chi0v_C	sum of the inverse square roots of the ν_i of the carbon atoms
chilv	sum of the inverse square roots of $\upsilon_i \upsilon_j$ for all bonded heavy atoms i and j
chi2	bond valence connectivity index (order 2)
chilv_C	sum of the inverse square roots of $\upsilon i \upsilon j$ for all bonded carbon atoms i and j
Kierl	$n(n-1)^2/m^2$
Kier2	$(n-1)(n-2)^2/p_2^2$
Kier3	$(n-1)(n-3)^2/p_3^2$ for odd n, and $(n-3)(n-2)^2/p_3^2$ for even n
KierA1	$s(s-1)^2/m^2$, where $s = n + a$
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 $s = n + a$
KierFlex	KeirAl $\times$ KeirA2/n
WienerPath	Wiener path no., which is half the sum of all the distance matrix entries[30]
WienerPol	Wiener parity no., which is half the number of entries in the distance matrix with a value of 3 [30]
diameter	largest distance matrix entry[31]
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 [31]
petitjean	value of (diameter − radius)/diameter[31]
VdistMa	if m is the sum of the distance matrix entries, then VDistMa is defined to the sum of a_{ij} (log a_{ij}/m − log m over all i and j (logarithms are taken in base 2)
VdistEq	entropy of the distribution of values in the distance matrix
BalabanJ	Balaban's connectivity topological index[32]
Vsa_don	approx to the sum of VDW surface area of hydrogen bond donors
Vsa_acc	approx to the sum of VDW surface area of hydrogen bond acceptors
Vsa_pol	approx to the sum of VDW surface area of polar atoms
Vsa_hyd	approx to the sum of VDW surface area of hydrophobic atoms
Vsa_acid	approx to the sum of VDW surface area of acidic atoms
Vsa_base	approx to the sum of VDW surface area of basic atoms
Vsa_other	approx to the sum of VDW surface area of other atoms
(d) Single Structural Key-like Functional Groups[d]	
f_c=o	no. of C═O groups
f_conh2	no. of $CONH_2$ groups
f_nh2	no. of primary NH_2 groups
f_so2n	no. of SO_2N groups
f_so2nh2	no. of SO_2NH_2 groups
i_so2nh	indicator variable for SO_2NH group
i_so2nh2	indicator variable for SO_2NH_2 group

[a] q_i is the partial charge of atom i in a molecule and p_i represents the partial charge of atom i calculated according to the PEOE method,[28] and ν_i is the van der Waals surface area of atom i. [b] "HB-a" and "LP" provide similar, yet distinct, descriptions for hydrogen bond acceptors ("LP" generally represents a greater number of electron lone pairs). [c] d_i is the number of heavy atoms bonded to atom i. $\nu_i = (p_i - h_i)/(z_i - p_i - 1)$, 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 atoms is defined as the n by n matrix, **A**, where a_{ij} is the length of the shortest path in graph between atoms i and j. The descriptors represent values derived from the graph distance matrix of the non-hydrogen molecular graph of a molecule. [d] These seven structural key descriptors were implemented in MOE and used in addition to 166 MACCS keys that were applied together as a complex descriptor.

scoring 100 chromosomes were retained, and the pairwise crossover operation was performed for 25% of the population. Then, mutation was carried out by randomly inverting 5% of the bits in each chromosome, providing the next generation. If a new chromosome was generated with more than 20% of bits set on after crossover and mutation, each bit was assigned a survival chance of 50% to reach the next generation. Descriptor combinations in this population were again subjected to PCA, crossover, and mutation operations, and the top scoring 100 combinations were saved. If the results did not change for 200 generations, the calculation was considered converged. For the calculation with 8% initial probability of bit occupancy, this was the case after 4313 generations.

RESULTS AND DISCUSSION

Descriptors and Search Space. A major goal of this study was to evaluate a large number of molecular descriptors under conditions of PCA-based compound classification to further increase the performance of the approach. Table 3 compares the parameters of our initial[8] and current study. In our first analysis, combinations of 17 descriptors (16 single descriptors plus one descriptor for structural keys) were tested, using constant principal component and bin values, to partition compounds in our test database. Thus, the total number of possible descriptor combinations was $2^{17} = 131\,072$, and it was feasible, albeit computationally expensive, to explore all descriptor combinations by complete factorial analysis. By extending the number of molecular

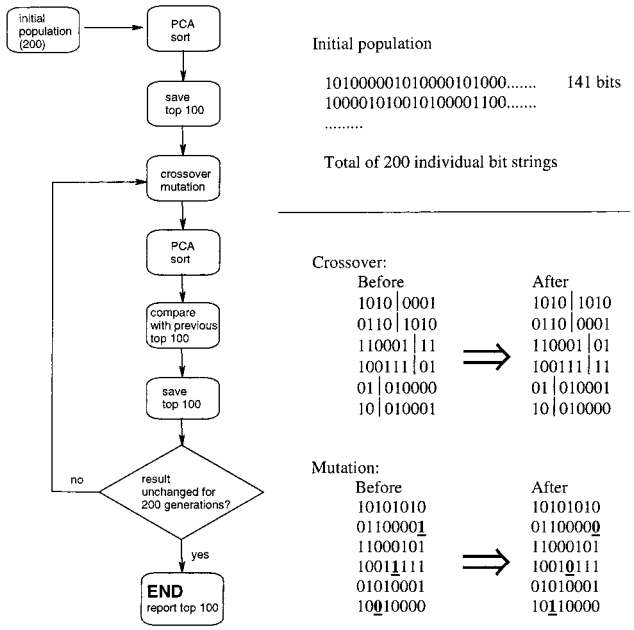

Figure 2. Genetic algorithm. The diagram summarizes the genetic algorithm approach applied in the study. Crossover and mutation operations are illustrated for a population of six model chromosomes, each of which consists of only eight bits. Following sorting by score, each pair of chromosomes is subjected to crossover at a random positions to generate a new pair of chromosomes, and a specified number of bits in all chromosomes are inverted during the mutation stage.

Table 3. Parameters for PCA-Based Compound Classification[a]

	Xue et al.[8]	current study
compounds/activity classes	455/7	455/7
no. of descriptors	16	110
structural keys	57 (SS)	166 (MACCS)
PCA limit	3	2−15
no. of bins	8	2−15
scoring function	S_1	S_2, S_1
calculation method	factorial analysis	genetic algorithm

[a] The table compares our initial and current calculations to identify preferred combinations of molecular descriptors for PCA-based classification of test compounds according to their biological activity. In our current study, GA-based optimization was carried out on the basis of S_2 scores, and S_1 scores were subsequently calculated for the top 100 descriptor combinations. The initial set of 57 structural keys (abbreviated SS) was selected on the basis of their high frequency of occurrence in large compound databases[8] and included a subset of 42 MACCS keys. In the current study, the complete set of 166 MACCS keys[24] was used.

descriptors to 111, the number of possible descriptor combinations increases to $2^{111} = 2.6 \times 10^{33}$ and, by permitting flexible principal component and bin limits (from 2 to 15), the search space further increases to approximately 10^{42} possible combinations. This by far exceeds the limits of complete factorial analysis and, therefore, a chromosome representation of descriptors was designed and a genetic algorithm implemented to facilitate the analysis.

Chromosome Design. In order to explore descriptor combinations, we encoded a binary chromosome consisting of 141 bit positions that can be divided into four segments, as illustrated in Figure 3. Each of the first 110 bit positions accounts for one of the numerical descriptors reported in Tables 2a−c or one of seven structural key-like functional groups shown in Table 2d. If a bit was set on (i.e., set to 1)

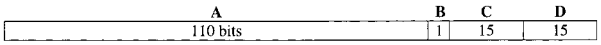

Figure 3. Design of binary chromosomes. The schematic illustrates how molecular descriptor combinations and calculation parameters were encoded in a chromosome consisting of a total of 141 bit positions. The chromosome can be divided into four parts: **A**, each of 110 bit positions accounts for a single descriptor (see text for further details); **B**, a single bit accounts for the presence of absence of 166 MACCS keys (in an "all or nothing" fashion); **C**, each of 15 bits, if set on, adds one principal component to the calculation; **D**, each of 15 bits, if set on, adds one bin segment for division of principal components following PCA.

for a numerical descriptor, values of this descriptor were calculated for all compounds in our test database and subjected to PCA (together with other "active" descriptors). If a bit was set on for one of the functional groups, the presence or absence of this group was monitored in each compound and used in PCA. Following these 110 bit positions, a single bit accounts for the presence or absence of 166 MACCS keys (which were thus applied as a complex descriptor). If this bit was set on, the presence or absence of all MACCS keys in database compounds was determined and subjected to PCA. The following two bit segments, each consisting of 15 bit positions, determine the number of principal components and axis bins for each calculation, respectively. In these cases, each bit position that was set on accounted for one PC or bin. For example, if 3 of 15 bits were set on in the third segment and 8 of 15 bits set on in the fourth, three principal components were calculated and, following PCA, each principal component was divided into eight bins.

Genetic Algorithm. GA-based approaches typically provide reasonable but not necessarily optimal solutions to a problem under investigation.[25−27] Thus, better combinations of descriptors may exist than the best performing combinations identified in this study. In our calculations, we achieved best performance of descriptor combinations, as assessed with our scoring functions, after testing approximately 863 000 combinations. This is a large number but still only a small fraction of the possible search space. Convergence was reached within a reasonable period of calculation time, which was likely due to the fact that we tested a number of initial bit densities for chromosomes (see Methods), thus enabling us to narrow down the range of bit settings leading to effective compound classification. Bit densities are proportional to the number of active descriptors, and initial bit settings at or below 10% generally resulted in the best intermediate scores. By contrast, a reference calculation with 50% probability of initial bit occupancy produced lower scores and did not reach our convergence criterion.

Test Compounds and Scoring Functions. Our analysis depends on at least two case-specific parameters, the source database (thus, the chemical nature and diversity of selected compounds and their biological activity) and our scoring or fitness functions. Our analysis was designed to identify descriptor combinations that effectively classify compounds according to biological activity, rather than chemical similarity. Thus, our scoring functions penalize false positive assignments of similar activity and favor correct assignments. The initial scoring function S_1 attempted to minimize the number of mixed classes. We implemented a more complex scoring function S_2 to not only maximize the number of compounds in pure classes but also minimize the total

Table 4. Top 15 Combinations of Descriptors for PCA-Based Compound Classification Using Genetic Algorithm[a]

descriptors	PC/bin	S_1	S_2	N_p	N_s	N_m
a_aro, b_ar, b_double, b_triple, a_nP, fso2nh2, LP	4/5	1.28	1.73	415 (60)	27	13 (2)
a_aro, f_c=o, LP, vsa_pol	9/8	1.78	1.59	414 (73)	41	0 (0)
a_aro, a_nO, LP	7/10	0.901	1.36	404 (64)	31	20 (4)
b_ar, b_double, a_nN, f_so2nh2	6/9	1.27	1.34	401 (56)	24	30 (2)
VdistMa, a_aro, b_triple, VadjEq, PEOE_VSA+2, PEOE_VSA+5, I_so2nh	3/9	0.667	1.30	402 (64)	36	17 (6)
b_double, PEOE_VSA-6, HB-a, LP	10/5	1	1.30	401 (67)	27	27 (4)
Petitjean, radius, b_lrotR, b_double, b_rotN, KierA2, MACCS	5/3	0.75	1.20	397 (72)	26	32 (7)
a_aro, PEOE_VSA+5, PEOE_VSA-4, PEOE VSA-5	6/8	0.973	1.19	397 (71)	33	25 (4)
b_ar, a_nN, PEOE_VSA_POL, vsa_po	7/5	0.987	1.18	397 (74)	35	23 (4)
a_aro, a_nS, fc=o, LP, vsa_pol	5/6	1.25	1.17	399 (79)	53	3 (1)
a_aro, PEOE_VSA+5, LP, HB-don	4/6	0.909	1.15	397 (80)	38	20 (5)
a_aro, PEOE_VSA+2, PEOE_VSA-6	11/9	0.688	1.14	394 (64)	33	28 (6)
b_double, PEOE_VSA+6, PEOE_VSA-4, f_so2nh2, LP	8/8	0.897	1.14	395 (70)	38	22 (4)
a_aro, chi2, PEOE_VSA+2, PEOE_VSA-5, PEOE_VSA_POL, f_so2n, f_so2nh2, logP(o/w)	4/6	1.42	1.14	399 (91)	54	2 (1)
Petitjean, radius, b_lrotN, chiO_C, b_heavy, PEOE_VSA+2, RPC+, MACCS	4/4	0.709	1.14	395 (73)	33	27 (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 numbers 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. Descriptors are defined in Table 2.

number of classes. Therefore, in this analysis, we used S_2 rather than S_1 scores as our primary performance criterion.

Selected Molecular Descriptors. Combinations of descriptors with highest S_2 scores identified in our calculations are listed in Table 4. The three top scoring combinations consist of only seven, four, and three (in part similar) descriptors, respectively. Recurrent among top scoring combinations are descriptors that account for aromatic character, hydrogen bond acceptors, partial charges, bond counts, and single structural key-type fragments. Principal component and bin settings vary greatly among the 15 top scoring combinations, but at least three principal components are required in each case. This is in accord with our previous observations that the first three principal components accounted for most of the variability of tested descriptor combinations.[8] In our calculations, hydrogen bond acceptors are clearly more important than donors, and LP is a higher performance descriptor for hydrogen bond acceptors than HB-a. The complex MACCS key descriptor does not occur in the top six scoring combinations. The top and fourth ranked combinations include a single structural key detecting the sulfonamide group. The importance of this descriptor can be rationalized by the fact that approximately one-third of the compounds in the test database are carbonic anhydrase inhibitors that usually depend on the presence of this functional group. However, the descriptor combination with the second highest score does not include this descriptor but a key that accounts for carbonyl groups, and the third ranked combination lacks any structural key.

Performance of PCA-Based Compound Classification. The results in Table 4 show that the performance of the PCA-based compound classification method is generally high. The top 15 descriptor combinations partition between 87 and 91% of all compounds into pure classes and a maximum of 7% of compound into mixed classes. The second ranked descriptor combination partitioned 91% of the compounds into pure classes and did not produce any mixed classes. This combination consisted of only four descriptors, the number of aromatic atoms, hydrogen bond acceptors, a structural key accounting for carbonyl groups, and polar van der Waals surface area. Overall, we favor this descriptor combination because the top scoring combination, consisting of seven descriptors, partitioned 13 compounds into mixed classes. However, this combination resulted in 14 fewer singletons and 13 fewer classes than the second ranked combination and thus scored slightly higher.

Comparison with Jarvis−Patrick Clustering. Although descriptors were exclusively optimized in this study for PCA-based classification, the top scoring descriptor combinations were also applied in Jarvis−Patrick clustering[18] using a previous implementation of the algorithm.[8] The results are reported in Table 5. Under these conditions, the performance of Jarvis−Patrick clustering was significantly lower than PCA-based compound classification. Although Jarvis−Patrick clustering produced only very few singletons, many compounds occurred in mixed clusters. Thus, many false positive assignments were made. This comparison shows that the performance of preferred descriptor combinations depends on the specifics of the algorithm used.

Database Dependence of Classification Performance. As mentioned above, optimization of descriptor combinations depends, at least to some extent, on the composition of test database(s) under investigation. In the current database, carbonic anhydrase inhibitors represent the largest biological activity class (Table 1). In two of the top four descriptor combinations shown in Table 4, we find a single descriptor accounting for the presence of sulfonamide fragments. Many carbonic anhydrase inhibitors contain sulfonamide groups that complex the zinc ion in the active site of the enzyme. Thus, one would anticipate that descriptors accounting for the presence or absence of sulfonamide groups are important for distinguishing carbonic anhydrase inhibitors from other compounds in our benchmark database. Since Cox-2 inhibitors in our database also contain sulfonamide groups, these functional groups are not a single class-specific feature. However, since carbonic anhydrase inhibitors represent the largest activity class in our test database, we investigated the question of whether the performance of our top scoring descriptor combinations was dependent on the presence of carbonic anhydrase inhibitors. Thus, we omitted carbonic

Table 5. Comparison of PCA-Based Compound Classification with Jarvis−Patrick Clustering[a]

descriptor	S_2	JP-S_2	JP-N_P	JP-N_s	JP-N_m
a_aro, b_ar, b_double, b_triple, a_nP, f_so2nh2, LP	1.73	0.48	314 (20)	1	140 (8)
a_aro, f_c=o, LP, vsa_pol	1.59	0.28	258 (18)	3	194 (9)
a_aro, a_nO, LP	1.36	0.28	258 (17)	3	194 (9)
b_ar, b_double, a_nN, f_so2nh2	1.34	0.23	236 (20)	1	218 (10)
VDistMa, a_aro, b_triple, VAdjEq, PEOE_VSA+2, PEOE_VSA+5, I_so2nh	1.30	0.13	172 (16)	4	279 (15)
b_double, PEOE_VSA-6, HB-a, LP	1.30	0.28	256 (15)	4	195 (8)
petitjean, radius, b_lrotR, b_double, b_rotN, KierA2, MACCS	1.20	0.79	359 (17)	9	87 (2)
a_aro, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-5	1.19	0.09	135 (11)	1	319 (16)
b_ar, a_nN, PEOE_VSA_POL, vsa_po	1.18	0.16	194 (13)	2	259 (9)
a_aro, a_nS, f_c=o, LP, vsa_pol	1.17	0.32	271 (19)	1	183 (9)
a_aro, PEOE_VSA+5, LP, HB_don	1.15	0.28	259 (17)	1	195 (10)
a_aro, PEOE_VSA+2, PEOE_VSA-6	1.14	0.24	241 (15)	3	211 (9)
b_double, PEOE_VSA+6, PEOE_VSA-4, f_so2nh2, HB-a	1.14	0.16	193 (18)	2	260 (14)
a_aro, chi2, PEOE_VSA+2, PEOE_VSA-5, PEOE_VSA_POL, f_so2n, f_so2nh2, logP(o/w)	1.14	0.16	191 (11)	5	259 (10)
petitjean, radius, b_lrotN, chiO_C, b_heavy, PEOE_VSA+2, RPC+, MACCS	1.14	0.30	263 (11)	5	187 (5)

[a] Top scoring descriptor combinations for PCA-based compound classification and S_2 scores are reported as in Table 4. "JP-S_2" reports corresponding scores obtained by Jarvis−Patrick clustering. "JP-N_p" is the total number of the compounds in pure clusters (the number of pure clusters is reported in the parentheses) obtained by the Jarvis−Patrick approach. "JP-N_s" is the number of singletons. "JP-N_m" is the total number of compounds in mixed clusters (the number of the mixed clusters is given in parentheses).

Table 6. Performance of Top Scoring Descriptor Combinations for PCA-Based Compound Classification in the Absence of Carbonic Anhydrase Inhibitors[a]

descriptor	S_2	$*S_2$	$*N_p$	$*N_s$	$*N_m$
a_aro, b_ar, b_double, b_triple, a_nP, f_so2nh2, LP	1.73	3.4	277 (37)	13	6 (1)
a_aro, f_c=o, LP, vsa_pol	1.59	1.59	253 (45)	16	27 (4)
a_aro, a_nO, LP	1.36	1.7	255 (39)	15	26 (4)
b_ar, b_double, a_nN, f_so2nh2	1.34	1.55	250 (33)	16	30 (2)
VDistMa, a_aro, b_triple, VAdjEq, PEOE_VSA+2, PEOE_VSA+5, I_so2nh	1.30	3	275 (40)	19	2 (1)
b_double, PEOE_VSA-6, HB-a, LP	1.30	2.79	274 (45)	22	0 (0)
petitjean, radius, b_lrotR, b_double, b_rotN, KierA2, MACCS	1.20	1.29	245 (45)	30	21 (5)
a_aro, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-5	1.19	0.95	226 (38)	20	50 (3)
b_ar, a_nN, PEOE_VSA_POL, vsa_po	1.18	1.46	251 (50)	26	19 (3)
a_aro, a_nS, f_c=o, LP, vsa_pol	1.17	2.49	271 (50)	20	5 (1)
a_aro, PEOE_VSA+5, LP, HB_don	1.15	1.46	252 (51)	31	13 (3)
a_aro, PEOE_VSA+2, PEOE_VSA-6	1.14	1.86	262 (51)	29	5 (2)
b_double, PEOE_VSA+6, PEOE_VSA-4, f_so2nh2, HB-a	1.14	1.72	257 (46)	20	19 (2)
a_aro, chi2, PEOE_VSA+2, PEOE_VSA-5, PEOE_VSA_POL, f_so2n, f_so2nh2, logP(o/w)	1.14	1.45	252 (52)	34	10 (3)
petitjean, radius, b_lrotN, chiO_C, b_heavy, PEOE_VSA+2, RPC+, MACCS	1.14	1.24	245 (61)	32	19 (3)

[a] Top scoring descriptor combinations and S_2 scores for PCA-based classification of compounds in the test database consisting of seven activity classes (Table 1) are reported as in Table 4. Asterisks indicate the corresponding performance values obtained by partitioning of the database after omission of carbonic anhydrase inhibitors (i.e., six biological activity classes).

anhydrase inhibitors from the database, partitioned the remaining compounds belonging to six activity classes using our top 15 descriptor combinations (Table 4), and calculated corresponding S_2 scores. The results in Table 6 show that classification performance is not dependent on the presence of carbonic anhydrase inhibitors. In fact, scores for the best descriptor combinations generally increase, which is probably due to two factors. The number of classified compounds decreases by approximately 35% (thus reducing the statistical error rate), and, in addition, the carbonic anhydrase inhibitors used in this study consisted of a diverse array of structures (and were thus a "difficult" class).

Biological Activity and Structural Diversity. The results in Table 4 also reveal that the minimum number of calculated pure classes among the top 15 combinations is 56 (having on average seven correctly assigned compounds per pure class), whereas the test database consists of seven biological activity classes (with, on average, 65 compounds per class). Thus, the calculations generate subclasses of compounds sharing equivalent biological activity. These subclasses are distinguished by minor differences in structure and/or properties that are permitted within a series of compounds having similar activity. In addition, our database contains structurally distinct types of compounds with similar activity.[19] Taken together, these findings demonstrate the sensitivity of the approach and preferred descriptor combinations. On the other hand, sensitivity to minor differences in structure and properties of compounds having similar activity also reflects a drawback of the approach, since it causes the occurrence of singletons (i.e., compounds for which activity relationships were missed). Thus, compound classification on the basis of biological activity requires finding a balance, or compromise, between assessment of molecular similarity and biological activity.

Table 7. Comparison of Previously Identified and Current Top Five Descriptor Combinations for PCA-Based Compound Classification According to S_1 Scores[a]

molecular descriptors	PC/bin	C_p	C_m	C_s	S_1
Current Study					
a_aro, f_c=o, LP, vsa_pol	9/8	73	0	41	1.78
a_nC, RPC-, KierA3, MACCS	5/4	100	7	58	1.47
chiO_C, chil, a_nCl, PC+, PEOE_VSA+2, apol, MACCS	4/5	91	6	53	1.44
a_aro, PEOE_VSA+4, PEOE_VSA_PNEG, PEOE_VSA_POL, vsa_pol	7/6	91	0	63	1.44
a_aro, b_double, PEOE_VSA-6, PEOE_VSA_PNEG, I_so2nh2, vsa_acc	7/5	85	0	59	1.44
Xue et al.[8]					
b_ar, SS, HB-a	3/8	75	4	34	1.01
b_lrotR, $^1\chi$, PEOE_PC+, SS, $^1\kappa$, $^2\kappa$, $^3\kappa$	3/8	88	4	52	0.96
b_lrotR, b_ar, SS, HB-a	3/8	74	4	38	0.95
$^0\chi$, PEOE_PC+, SS, HB-a, CMR	3/8	96	4	71	0.86
b_ar, PEOE_PC+, SS, $^2\kappa$, $^3\kappa$	3/8	77	5	42	0.84

[a] "C_p", "C_m", and "C_s" are the number of pure classes, mixed classes, and singletons, respectively. "PC" gives the number of principal components and "bin" the number of intervals per component. S_1 scores were calculated as defined in Methods.

Application of the Method. Successful partitioning of test compounds into pure classes, as achieved in our study, is an important criterion for application of the approach, for example, to search databases for compounds with activity similar to query molecules. In this type of analysis, one would first identify preferred descriptor combinations for biological activity classes of interest, then add selected query compounds to databases, partition these compound collections, and closely inspect classes into which query molecules fall (as these classes may contain functional analogues). In this regard, it should be noted that the number of classes obtained in a partitioning experiment is not only dependent on the compound collections under investigation but also on the parameters applied in the calculations. For example, an increase in the number of bins set on principal components (see Methods) will lead to a larger number of cells in principal component space (representing potential classes) and may thus further divide classes of compounds.

Comparison with Previous Results. To directly compare the results obtained in this study with our initial analysis, we have also ranked our top 100 descriptor combinations according to S_1 scores. Table 7 compares previously identified and current top scoring combinations. As to be expected (see Methods), effective descriptor combinations score high (i.e. >1) using both scoring functions but the relative rankings of the results change. The top S_2 scoring combination is only number 11 on the S_1 list and our preferred descriptor combination, ranked second by S_2 scores, becomes the top S_1 scoring combination. In total, we found 36 new descriptor combinations with S_1 scores better than 1.01, the highest score achieved previously (Table 7). These findings support our efforts to extend the analysis to a much larger descriptor set. On the new S_1 list, two combinations that include the MACCS key descriptor are in the top five combinations, in addition to combinations with single structural keys. All of the previously identified top scoring combinations include the set of 57 structural keys (SS). These findings emphasize the importance of structural key-type descriptors.[14] However, we also find that the presence of very few and, in some case, only a single structural key yields performance comparable to or better than the complete set. In addition, our overall top scoring descriptor combination, which does not include a structural key, illustrates that the contribution of these keys is not necessarily required. Descriptors accounting for aromatic character and hydrogen bond acceptors are consistently found in newly identified and previous top scoring combinations, confirming that these descriptors are particularly important for classification of our test compounds according to their biological activity.

CONCLUSIONS

We have evaluated combinations of a large number of molecular descriptors to partition 455 compounds according to their specific activity by application of principal component analysis of molecular descriptor space. The magnitude of the combinatorial problem suggested the implementation of a genetic algorithm. A number of well-performing descriptor combinations were identified. The majority of compounds were correctly classified, and only few, if any, false positive structure−activity relationships were detected. However, the performance of the approach is still somewhat limited by the occurrence of (approximately 9%) singletons. This suggests that the calculations put too much weight on the detection of differences in molecular structure and properties of test compounds, rather than capturing those features that are most critical for a specific activity. Thus, in general, increasingly detailed descriptions of molecules are likely to overestimate the role of fine structural details in similarity searching focused on biological activity. This view is consistent with the finding that combinations of only a few molecular descriptors performed well, albeit not yet optimally, in our analysis.

ACKNOWLEDGMENT

We thank J. Godden and F. Stahura for many helpful discussions and critical review of the manuscript.

REFERENCES AND NOTES

(1) 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.

(2) Hodes, L. Clustering a large number of compounds. 1. Establishing the method on an initial sample. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66−71.

(3) 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.

(4) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.
(5) 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.
(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) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead discovery using stochastic cluster analysis (SCA): A new method for clustering structurally similar compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305−312.
(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) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of large structure/biological activity data sets using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(10) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.
(11) Mason, J. S.; Hermsmeier, M. A. Diversity assessment. *Curr. Opin. Chem. Biol.* **1999**, *3*, 342−349.
(12) Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(13) Matter, H.; Pötter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.
(14) 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.
(15) Brown, R. D. Descriptors for diversity analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.
(16) Johnson, M.; Maggiora, G. M. *Concepts and applications of molecular similarity*; Wiley: New York, 1990.
(17) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal component analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.
(18) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(19) 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.
(20) Willett, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(21) Sanatvy, M.; Labute, P. SVL: The Scientific Vector Language. *J. Chem. Comput. Group* (http://www.chemcomp.com/feature/svl.htm).
(22) *MOE (Molecular Operating Environment)*; MOE 1999.05; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 1995.
(23) Labute, P. QuaSAR-Cluster: A different view of molecular clustering. *J. Chem. Comput.* (http://www.chemcomp.com/article/cluster.htm). This function is part of the Molecular Operating Environment.
(24) MDL Information Systems Inc., San Leandro, CA.
(25) Forrest, S. Genetic algorithms−Principles of natural selection applied to computation. *Science* **1993**, *261*, 872−878.
(26) Clark, D. E.; Westhead, D. R. Evolutionary algorithms in computer-aided molecular design. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 337−358.
(27) Gillett, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf Comput. Sci.* **1998**, *38*, 165−179.
(28) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity−A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(29) Kier, L. B. Indexes of molecular shape from chemical graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.
(30) Balaban, A. T. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta* **1979**, *53*, 355−375.
(31) 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.
(32) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

CI000322M