Identification of a Preferred Set of Molecular Descriptors for Compound Classification Based on Principal Component Analysis

Ling Xue, Jeff Godden, Hua Gao, and Jürgen Bajorath*,[†]

Computational Chemistry and Informatics, MDS Panlabs, 11804 North Creek Parkway,
Bothell, Washington 98011, and MDS Panlabs and Department of Biological Structure,
University of Washington, Seattle, Washington 98195

Received December 23, 1998

An algorithm based on principal component analysis was investigated to classify molecules in a database consisting of 455 compounds with activities against seven different biological targets. Diversity profiles of these compound sets were calculated and compared. To effectively classify compounds with similar biological activity, all possible combinations of 17 molecular descriptors were tested by complete factorial analysis, and preferred descriptor combinations were identified. High efficiency was achieved for a combination of a limited set of structural keys and two or three additional 2D descriptors. The performance of the approach was compared to Jarvis−Patrick clustering.

INTRODUCTION

Cluster analysis of large compound databases has become a popular tool to aid in the generation of specialized compound libraries and the selection of compounds for experimental testing.[1−10] In cluster analysis, compounds are divided into different classes according to predefined similarity criteria. The measure of similarity depends on the use of molecular descriptors. Clustering or partitioning of compound sets is conceptually based on the similar property principle,[11] which states that structurally similar molecules exhibit similar chemical and biological properties.

Effective classification of molecules is critically dependent on at least two factors, the algorithm and molecular descriptors used. Cluster algorithms can be divided into two major categories, hierarchical and nonhierarchical methods.[12] In nonhierarchical clustering, a set of compounds is divided into a number of generally nonoverlapping clusters. Jarvis−Patrick (JP) clustering is one of the most widely used nonhierarchical methods.[4,9]

Many different molecular descriptors have been reported in the literature, and the choice of such descriptors is often intuitive. Molecular descriptors can be classified as bulk properties (e.g., logP), two-dimensional (2D) (e.g., molecular connectivity), or 3D descriptors (e.g., conformational states). The 2D molecular descriptors including Kier shape indices[14] and structural keys (SSKey-type descriptors)[4,9] are among the most widely used descriptors. Structural key-type descriptors were originally developed for substructure searching and capture the presence or absence of defined structural fragments in a molecule.[15] Their usefulness in both cluster and diversity analysis has previously been established.[4,9]

In this study, we have investigated a novel nonhierarchical algorithm to classify compounds represented by molecular descriptors. This algorithm is based on principal component analysis (PCA) and was developed by P. Labute.[13] Although the algorithm was termed a molecular clustering method, the approach is conceptually more similar to partitioning methods (see URL: http://www.chemcomp.com/article/cluster.htm). To avoid ambiguous terminology, we use the term compound classification rather than clustering or partitioning in this study.

The primary goal of this analysis has been to identify a set of molecular descriptors that perform efficiently in PCA-based compound classification. Thus, we have systematically explored, by complete factorial analysis, all possible combinations of 17 widely used molecular descriptors. An important variable in this study is the choice of the database for compound classification. We have assembled seven distinct sets of compounds from the literature.[16−31] Each set is active against a different biological target. The combined set consists of 455 bioactive compounds. Thus, following the similar property principle[11] we are using similar biological activity as a criterion for classification of compounds. An advantage of doing so is that we are able to apply a simple scoring scheme to assess the efficiency of the calculations.

MATERIALS AND METHODS

Test Compounds. A set of 455 active compounds was collected from the literature[16−31] including inhibitors of carbonic anhydrase II (CAII), cyclooxygenase-2 (Cox-2), HIV-1 protease, and tyrosine kinases, ligands for the benzodiazepine and 5-HT receptors, and histamine receptor (H3) antagonists.

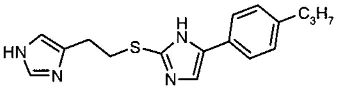
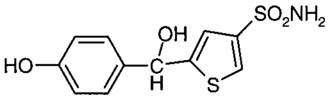
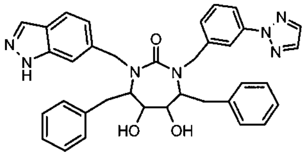
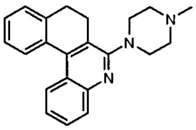
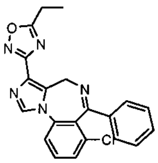
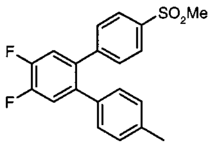
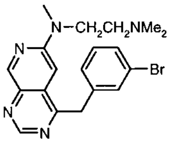
Molecular Descriptors. The SSKey-type descriptors used in this study are a subset containing 41 of 166 MDL structural keys, which represent small molecular fragments.[15] The choice of this subset of MDL structural keys was essentially subjective and based on our observation that these fragments frequently occur in compounds of the Maybridge and Optiverse databases.[32,33] The 41 selected MDL fragments were complemented by a set of 16 structural fragments that

* To whom any correspondence should be addressed. Phone: (425) 487-8297. Fax: (425) 487−8262. E-mail: jbajorath@panlabs.com.

[†] University of Washington.

10.1021/ci980231d CCC: $18.00 © 1999 American Chemical Society
Published on Web 05/18/1999

Table 1. Representative Compounds

Biological activity	Representative structure	Number of compounds	Percentage of database
H3 antagonists		52	11.4
Carbonic anhydrase II inhibitors		159	34.9
HIV protease inhibitors		48	10.5
Serotonin receptor ligands (5-HT)		71	15.6
Benzodiazepine receptor ligands		59	13.0
Cyclooxygenase-2 (Cox-2) inhibitors		31	6.8
Tyrosine Kinase (TK) inhibitors		35	7.7

we designed. In addition to the set of 57 SSKey-type descriptors, other commonly used molecular descriptors, as implemented in MOE, version 1998.03,[13] were investigated (see Results). Complete factorial analysis, i.e., evaluation of all possible combinations of 17 molecular descriptors, was performed for PCA-based compound classification.

PCA-Based Compound Classification. The QuaSAR-Cluster function of MOE[13] is based on PCA.[34] The approach has been discussed in detail (see URL: http://www.chemcomp.com/article/cluster.htm) and is briefly described here. Compounds are expressed as an *n*-vector of molecular descriptors in *n*-dimensional space. PCA performs an *n* by *p* linear transform of the *n*-vectors and calculated *p*-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 *p* axes is divided into *k* intervals and the coordinates of a molecule are assigned a letter code designating their respective axis interval. Thus, each molecule is characterized by a signature code consisting of *k* letters. Molecules that share a signature code form a cluster or class. In our MOE QuaSAR-Cluster calculations, we found that three principal components and eight intervals for signature coding were sufficient for effective compound classification. Additional principal components did not significantly improve classification efficiency.

Comparison with Jarvis−Patrick Clustering. JP cluster analysis was used to benchmark the performance of PCA-based compound classification. The JP algorithm[35] was implemented in MOE by the authors using SVL code (see URL: http://www.chemcomp.com/feature/svl.htm) to test multiple descriptor combinations. To identify nearest neighbors of a given molecule the geometric distance between descriptor vectors was calculated. Compounds were ranked by distance, and a nearest neighbor list was calculated. Compounds are considered to form a cluster if they (i) appear in each other's list of 14 nearest neighbors and (ii) share eight of their 14 nearest neighbors.[36]

Scoring Function. The results of PCA-based compound classification and JP clustering were scored using the following function:

$$\text{Efficiency} = C_p/(10C_m + C_s)$$

C_p is the number of pure classes/clusters (i.e., containing

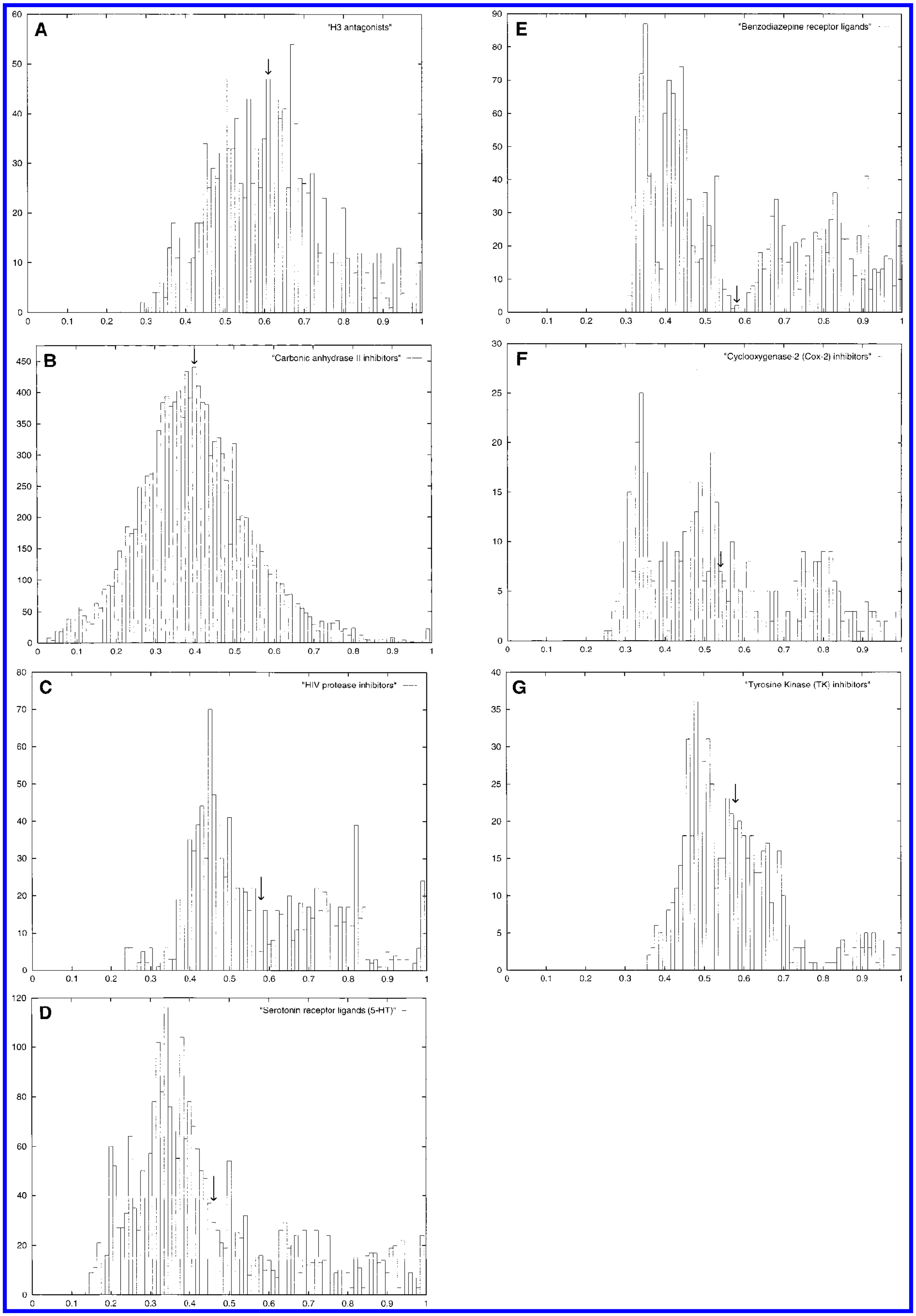


Figure 1. Diversity profiles of compound sets. In panels A−G, histograms are shown for the seven sets of compounds with different biological activities. The number of T_c values for pairwise comparison of compounds within each 0.01 interval is reported. Arrows indicate the average T_c value for each distribution.

only compounds with the same biological activity). C_s is the number of singletons (i.e., containing only one compound), and C_m represents the number of mixed classes/clusters (i.e., containing compounds with different biological activities). An arbitrary scaling factor of 10 was used to penalize the occurrence of mixed clusters.

Molecular Diversity. Tanimoto coefficients (T_c) and average T_c values were calculated using the ph4_ph2D_Fingerprint[37] implemented in MOE. This fingerprint calculates 2D pairwise-distance pharmacophore keys for atoms in a molecule. The pharmacophore key (or signature) is a bit string with each bit position referring to the presence or absence of a unique pattern. The Tanimoto coefficient for comparison of two molecules was calculated as follows:

$$T_c = B_c/(B_1 + B_2 - B_c)$$

B_c is the number of common bits set on (i.e., 1) and B_1 and B_2 are the bits set on in the fingerprints of molecule 1 and 2, respectively.

We have analyzed distributions of T_c values using a histogram method. In these "diversity profiles", the T_c range, from 0 to 1, was separated into 100 intervals of 0.01, and T_c values that fell in each interval were shown as bars.

RESULTS AND DISCUSSION

Compound Database. Representative structures for each of the seven compound sets with distinct biological activity are shown in Table 1. For each set diversity profiles were calculated (Figure 1). The analysis shows highly variable distributions of T_c values in compound sets of different biological activity. In some cases, the T_c distribution displays a single peak (e.g., Figure 1B), whereas other sets show multimodal distributions (e.g., Figure 1C). Such multimodal distributions may be due to the presence of distinct subsets of compounds within an activity class. The observed differences in the diversity profiles suggest that our compound collection provides a meaningful test set for classification. Figure 1E illustrates that an average T_c value is not representative of the diversity distribution within a compound set. We therefore use average T_c calculations only to assess the enrichment of similar compounds within calculated pure classes/cluster of compounds belonging to the same set (see below).

Molecular Descriptors. The 57 SSKey-type descriptors are listed in Table 2 and the other 16 molecular descriptors are shown in Table 3. These descriptors are either bulk properties (e.g., CMR) or 2D descriptors. The Kier shape descriptors[14] implicitly capture some 3D structural information.

Compound Classification. The major objective of our study was to identify a set of molecular descriptors for efficient PCA-based compound classification. Thus, all possible combinations of the 17 molecular descriptors (SSKeys plus 16 single descriptors) were tested. Effective compound classification should predominantly yield pure classes and minimize the number of singletons and, most importantly, mixed classes. Therefore, mixed clusters are weighted by a factor of 10 in the scoring function.

Table 2. Structural Key-Type (SS) Descriptors[a]

SS no	representation	SS no	representation
1	[a#X][r]	30	C[N][r]
2	a[OH]	31	[A#X][r]
3	C[OH]	32	NO2
4	aC(=O)OH	33	[#Qr][Aq0]$_{2-3}$[#Qr]
5	CC(=O)OH	34	[#Qr][Aq0]$_{4-5}$[#Qr]
6	a[NH2]	35	[#Qr][Aq0]$_{6-7}$[#Qr]
7	C[NH2]	36	[#Qr][Aq0]$_{8-9}$[#Qr]
8	C[NH]C	37	[#Qr][Aq0]$_{10-11}$[#Qr]
9	CN(C)C	38	[#Qr][Aq0]$_{12-13}$[#Qr]
10	c1ccccc1	39	[#Qr][Aq0]$_{\geq 14}$[#Qr]
11	an[r]a	40	[#XH][#XH]
12	SO2	41	C#C
13	S(=O)	42	OC(=O)O
14	SO2NH2	43	[#X][#G7]
15	C(=O)H	44	NC(C)N
16	C(=O)OC	45	OS(=O)N
17	C(=O)N	46	S[#Q]N
18	[A#Q][r5]	47	NN
19	[#X](=O)OH	48	CC(C)(C)[#Q]
20	[a#Q][r5]	49	[#Q]CH2OH
21	[#Q][r6]	50	O[#Q][#Q]O
22	[#Q][r7]	51	[#X]CH3
23	[#Q][r8]	52	C=C
24	[#Q][r≥9]	53	[#Q]([#X])([#X])[#X]
25	[#Qr][#Qr]([#Qr])[#Qr]	54	[#Q][#Q]([#Q])([#Q])[#Q]
26	aa(a)a	55	[#Q]CH2[#Q][#Q]CH2[#Q]
27	OSO	56	aa[Oq0]
28	[#G7]	57	[#XH][#Q][#Q][#XH]
29	[A#Q]=S		

[a] [#X]: non-carbon, non-hydrogen atom; [#Q]: non-hydrogen atoms; [#Gn]: element belonging to group n of the periodic table; [r]: ring system; [q]: ring degree; A: non-aromatic atom; a: aromatic atom; #: triple bond; =: double bond.

Table 3. Molecular Descriptors

symbol	description
b_ar	number of aromatic bonds in a molecule
b_1rotR	fraction of rotatable single bonds
$^0\chi$	zero-order atomic connectivity index
$^1\chi$	first-order atomic connectivity index
$^2\chi$	second-order atomic connectivity index
$^1\kappa$	Kier first-shape index
$^2\kappa$	Kier second-shape index
$^3\kappa$	Kier third-shape index
Φ	Kier molecular flexibility index
logP(o/w)	partition coefficient octanol/water
HBd	number of hydrogen bond donors
HBa	number of electron lone pairs that can accept hydrogen bonds
CMR	molecular refractivity
ASA_H	total accessible hydrophobic surface area
PEOE-PC+	total positive charge using the Gasteiger−Marsili charge model
apol	atomic polarizibility

Descriptors are listed that were used in addition to the set of 57 SSKey-type fragments.

Preferred Descriptor Sets. The results of our systematic analysis of descriptor combinations for PCA-based compound classification are reported in Table 4. Overall, classification efficiency was high. For the top 20 combinations of descriptors, the majority of calculated clusters were pure, and many combinations (all of them including SSKeys) produced reasonable results. SSKeys were the most important single descriptor and produced by themselves better results than the majority of descriptor combinations. This parallels the findings of Brown and Martin who compared clustering methods, including JP clustering.[4] However, SSKey-based classification efficiency could be improved by adding a few other descriptors. For PCA-based

Table 4. Top 20 Combinations of Descriptors for PCA-Based Compound Classification[a]

molecular descriptors	C_p	C_m	C_s	efficiency
b_ar, SS, HBa	75 (21)	4 (7)	34 (1)	1.01 (0.30)
b_1rotR,1χ, PEOE_PC+, SS,1κ, $^2\kappa$, $^3\kappa$	88 (14)	4 (5)	52 (7)	0.96 (0.25)
b_1rotR, b_ar, SS, HBa	74 (20)	4 (6)	38 (3)	0.95 (0.32)
$^0\chi$, PEOE_PC+, SS, HBa, CMR	96 (10)	4 (7)	71 (5)	0.86 (0.13)
b_ar, PEOE_PC+, SS, $^2\kappa$, $^3\kappa$	77 (16)	5 (6)	42 (0)	0.84 (0.27)
b_1rotR, b_ar, PEOE_PC+, SS, $^2\kappa$, $^3\kappa$	77 (16)	5 (6)	42 (0)	0.84 (0.27)
b_1rotR, $^0\chi$, $^1\chi$, SS, HBa, $^1\kappa$, $^2\kappa$, $^3\kappa$, CMR	93 (11)	5 (7)	62 (11)	0.83 (0.14)
$^0\chi$, $^1\chi$, SS, HBa, $^1\kappa$, $^2\kappa$, $^3\kappa$, CMR	92 (11)	5 (7)	63 (10)	0.81 (0.14)
b_ar, PEOE_PC+, SS, HBa	70 (22)	5 (6)	37 (2)	0.80 (0.35)
b_1rotR, $^1\chi$, SS, HBa, $^1\kappa$, $^2\kappa$, $^3\kappa$, CMR	83 (16)	4 (5)	65 (8)	0.79 (0.28)
SS	67 (22)	4 (6)	47 (4)	0.77 (0.34)
b_1rotR, $^1\chi$, SS, $^3\kappa$, logP(o/w)	96 (22)	7 (4)	55 (4)	0.77 (0.50)
$^2\chi$, SS, $^3\kappa$, apol, logP(o/w)	103 (8)	6 (8)	75 (9)	0.76 (0.09)
b_1rotR, $^0\chi$, $^1\chi$, $^2\chi$, SS, $^1\kappa$, $^3\kappa$	90 (9)	7 (9)	48 (10)	0.76 (0.09)
b_1rotR, $^1\chi$, PEOE_PC+, SS, $^1\kappa$, $^3\kappa$, CMR	86 (11)	5 (7)	63 (8)	0.76 (0.14)
$^2\chi$, PEOE_PC+, SS, $^3\kappa$, apol, logP(o/w)	105 (10)	7 (8)	68 (7)	0.76 (0.11)
b_1rotR, $^2\chi$, PEOE_PC+, SS, $^3\kappa$, apol, logP(o/w)	105 (10)	7 (8)	68 (7)	0.76 (0.11)
b_ar, PEOE_PC+, SS	64 (24)	6 (4)	25 (1)	0.75 (0.59)
$^0\chi$, $^1\chi$, $^2\chi$, SS, $^2\kappa$, $^3\kappa$, CMR	94 (13)	6 (7)	66 (7)	0.74 (0.17)
b_1rotR, b_ar, PEOE_PC+, SS, HBa, HBd, $^3\kappa$, logP(o/w)	78 (18)	6 (4)	45 (2)	0.74 (0.43)

[a] For comparison with the PCA-based classification method, cluster distributions and efficiency were calculated for Jarvis−Patrick (JP) clustering. JP values are given in parentheses.

classification, the combination of the 57 SSKey-type fragments with b_ar (number of aromatic bonds), and HBa (number of hydrogen bond accepting electron lone pairs) was most effective. Seventy-five of 115 classes were pure, 34 singletons, and only four mixed. This was closely followed by a combination of seven descriptors including Kier shape indices and, with almost indistinguishable efficiency, a combination of SSKeys, b_ar, HBa, and b_1rotR (fraction of flexible bonds). Thus, SSKey-type descriptors and two or three additional 2D descriptors accounting for hydrogen bonding, aromaticity, and molecular flexibility gave the overall best results. Why were hydrogen-bonding acceptors more important than donors in our analysis? For the molecules tested here, acceptors were more abundant than donors. In addition, the two prevalent donors, −OH and $-NH_2$, are also acceptors (and implicitly encoded in some SSKeys).

Comparison with JP Clustering. The performance of the PCA-based approach was compared with JP clustering for the top 20 descriptor combinations in Table 4. The efficiency of JP clustering was overall lower than observed for PCA classification when our scoring function was applied. Since descriptor combinations were not systematically selected for JP clustering, we do not expect optimum performance of that method in our comparison. However, SSKey-type descriptors, for example, are known to perform well in cluster analysis.[4] As observed for PCA-based classification, the efficiency of JP clustering can be improved by adding a few 2D descriptors to SSKeys. Table 4 shows interesting trends when comparing class/cluster distributions. PCA classification consistently produces more pure classes, whereas JP clustering minimizes the occurrence of singletons. The number of mixed classes and clusters is comparable. These distributions lead to better efficiency values for PCA classification due to the presence of more pure classes than clusters. This also suggests that PCA-based compound classification is more sensitive than JP clustering. However, on the basis of our study, we cannot conclude that the PCA method is superior or generally more efficient.

Table 5. Compound Sets and Classes Obtained by Most Efficient PCA Classification[a]

compd set	avg T_c of compound set	no. of pure classes	% compds in pure classes	avg T_c of pure classes
benzodiazepine receptor ligands	0.59	3	62.7	0.87
CAII inhibitors	0.40	30	86.2	0.63
histamine H3 antagonists	0.60	9	94.2	0.73
tyrosine kinase inhibitors	0.58	8	85.7	0.80
5-HT ligands	0.45	12	83.1	0.87
HIV protease inhibitors	0.58	7	68.8	0.78
Cox-2 inhibitors	0.54	5	80.6	0.86

[a] Only nonsingletons were considered. Values are reported for the most efficient descriptor combination: b_ar, SS, HBa (see Table 4).

Compound Sets and Classes. For the PCA-based method, we also analyzed the distribution of classified compounds belonging to each of the seven activity sets. Table 5 shows that the majority of compounds (greater than 80% for five of the seven sets) occur in pure classes. In each case, the calculated average T_c values are significantly larger for pure classes than for the original set. This indicates that PCA-based classification correctly identifies similar subsets of compounds sharing the same biological activity and supports the view that the method is sensitive.

CONCLUSIONS

We have systematically explored combinations of different molecular descriptors for compound classification based on PCA. Using a set of 57 structural keys and two or three 2D descriptors, the PCA-based method efficiently classified compounds on the basis of biological activity. We find that PCA classification performs at least as well as JP clustering. Analysis of class distributions suggests that the PCA-based approach effectively partitions subsets of compounds in a biological activity class.

ACKNOWLEDGMENT

We thank Paul Labute, Chris Williams, and Ken Kelly of Chemical Computing Group, Inc. for helpful discussions.

Furthermore, we thank one of our referees for a number of suggestions that helped to improve the manuscript.

REFERENCES AND NOTES

(1) Hansch, C.; Unger, S. H.; Forsythe, A. B. Strategy in drug design. Cluster analysis as an aid in the selection of substituents. *J. Med. Chem.* **1973**, *16*, 1212−1217.

(2) 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.

(3) Ajay; Walters, P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(4) 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.

(5) 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.

(6) 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.

(7) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(8) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.

(9) 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.

(10) 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.

(11) Johnson, M.; Maggiora, G. M. *Concepts and applications of molecular similarity*; Wiley: New York, 1990.

(12) 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.

(13) Chemical Computing Group Inc. MOE 1998.03. 1255 University Street, Montreal, Quebec, Canada, H3B 3 × 3.

(14) Kier, L. B. Indexes of molecular shape from chemical graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.

(15) MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577.

(16) Mor, M.; Bordi, F.; Silva, C.; Rivara, S.; Crivori, P.; Plazzi, P. V.; Ballabeni, V.; Caretta, A.; Barocelli, E.; Impicciatore, M.; Carrupt, P.-A.; Testa, B. H3-receptor antagonists: synthesis and structure−activity relationships of para- and meta-substituted 4(5)-phenyl-2-[[2-[4(5)-imidazolyl]ethyl]thio]imidazoles. *J. Med. Chem.* **1997**, *40*, 2571−2578.

(17) Hadjipavlou-Litina, D.; Hansch, C. Quantitative structure−activity relationships of the benzodiazepines. A review and reevaluation. *Chem. Rev.* **1994**, *94*, 1483−1505.

(18) Gao, H.; Williams, C. I.; Labute, P.; Bajorath, J. Binary-QSAR analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.

(19) Leurs, R.; Vollinga, R. C.; Timmerman, H. The medicinal chemistry and therapeutic potentials of ligands of the histamine H3 receptor. *Prog. Drug. Res.* **1995**, *45*, 107−165.

(20) Van der Goot, H.; Schepers, M. J. P.; Sterk, G. J.; Timmerman, H. Isothiourea analogues of histamine as potent agonists or antagonists of the histamine H3-receptor. *Eur. J. Med. Chem.* **1992**, *27*, 511−517.

(21) Vollinga, R. C.; Menge, W. M. P. B.; Leurs, R.; Timmerman, H. New analogues of burimamide as potent and selective histamine H3 receptor antagonists: The effect of chain length variation of the alkyl spacer and modifications of the N-thiourea substituent. *J. Med. Chem.* **1995**, *38*, 2244−2250.

(22) Bordi, F.; Mor, M.; Morini, G.; Plazzi, P. V.; Silva, C.; Vitali, T. QSAR study on H3-receptor affinity of benzothiazole derivatives of thioperamide. *IL. Farmaco*. **1994**, *49*, 153−166.

(23) Rewcastle, G. W.; Murray, D. K.; Elliott, W. L.; Fry, D. W.; Howard, C. T.; Nelson, J. M.; Roberts, B. J.; Vincet, P. W.; Showalter, H. H. D.; Winters, T. R.; Deriny, W. A. Tyrosine kinase inhibitors. 14. Structure−activity relationships of methyl-amino-substituted derivatives of 4-[(3-bromophenyl)amino]-6-(methylamino)-pyrido[3,4-d]-pyrimidine (PD158780), a potent and specific inhibitors of the tyrosine kinase activity of receptors for the EGF family of growth factors. *J. Med. Chem.* **1998**, *41*, 742−751.

(24) Morreale, A.; Galvez-Ruano, E.; Iriepa-Canaha, I.; Boyd, D. B. Arylpiperazines with serotonin-3- antagonist activity: A comparative molecular field analysis. *J. Med. Chem.* **1998**, *41*, 2029−2039.

(25) Bromidge, S. M.; Dabbs, S.; Davies, D. T.; Duckworth, D. M.; Forbes, I. T.; Ham, P.; Jones, G. E.; King, F. D.; Saunders: D. V.; Starr, S.; Thewlis, K. M.; Wyman, P. A.; Blaney, F. E.; Naylor, C. B.; Bailey, F.; Blackburn, T. P.; Holland, V.; Kennett, G. A.; Riley, G. J.; Wood, M. D. Novel and selective 5-HT$_{2c}$/$_{2b}$ receptor antagonists as potential anxiolytic agents: Synthesis, quantitative structure−activity relationships, and molecular modeling of substituted 1-(3-pyridylcarbamolyl)-indolines. *J. Med. Chem.* **1998**, *41*, 1598−1612.

(26) Taverne, T.; Diouf, O.; Depreux, P.; Poupaert, J. H.; Lesieur, D.; Guardiola-lemaitre, B.; Renard, P.; Rettori, M.-C.; Caignard, D.-H.; Pfeiffer, B. Novel benzothiazolin-2-one and benzoxazin-3-one arylpiperazine derivatives with mixed 5HT$_{1A}$/D2 affinity as potential atypical antipsychotics. *J. Med. Chem.* **1998**, *41*, 2010−2018.

(27) Jadhav, P. K.; Woerner, F. J.; Lam, P. Y. S.; Hodge, C. N.; Eyermann, C. J.; Man, H.-W.; Daneker, W. F.; Bacheler, L. T.; Rayner, M. M.; Meek, J. L.; Erickson-Viitanen, S.; Jackson, D. A.; Calabrese, J. C.; Schadt, M.; Chang, C.-H. Nonpeptide cycliccyanoguanidines as HIV-1 protease inhibitors: Synthesis, structure−activity relationships, and X-ray crystal structure studies. *J. Med. Chem.* **1998**, *41*, 1446−1455.

(28) De Lucca, G. V.; Kim, U. T.; Liang, J.; Cordora, B.; Klabe, R. M.; Garber, S.; Bacheler, L. T.; Lam, G. N.; Wright, M. R.; Logue, K. A.; Erickson-Viitanen, S.; Ko, S. S.; Trainor, G. L. Nonsymmetric P2/P2′ cyclic urea HIV protease inhibitors. Structure−activity relationship, bioavailability, and resistance profile of monoindazole-substituted P2 analogues. *J. Med. Chem.* **1998**, *1*, 2411−2423.

(29) Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seiberst, K.; Veenhrizen, A. W.; Zhang, Y. Y.; Isakson, P. C. Synthesis and biological evaluation of the 1,5-diarylpyrazole class of cyclooxygenase-2 inhibitors: Identification of 4-[5-(4-methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]benzene sulfonamide (SC-58635, Celecoxib*). J. Med. Chem.* **1997**, *40*, 1347−1365.

(30) Li, J. J.; Norton, M. B.; Renhard, E. J.; Anderson, G. D.; Gregory, S. A.; Isakson, P. C.; Koboldt, C. M.; Masferrer, J. L.; Perkins, W. E.; Seibert, K.; Zhang, Y.; Zweifel, B. S.; Retz, D. B. Novel terphenyls as selective cyclooxygenase-2 inhibitors and orally active antiinflammatory agents. *J. Med. Chem.* **1996**, *39*, 1846−1856.

(31) Reitz, D. B.; Li, J. J.; Norton, M. B.; Reinhard, E. J.; Collins, J. T.; Anderson, G. D.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K. Selective cyclooxygenase inhibitors: Novel 1,2-diarylcyclopetenes are potent and orally active Cox-2 inhibitors. *J. Med. Chem.* **1994**, *37*, 3878−3881.

(32) Maybridge Chemical Co. LTD, Trevillett, Tintagel, Cornwall PL34 OHW, U.K.

(33) Garr, C. D.; Peterson, J. R.; Schultz, L.; Oliver, A. R.; Underiner, T. L.; Cramer, R. D.; Ferguson, A. M.; Lawless, A. S.; Patterson, D. E. J. Solution phase synthesis of chemical libraries for lead discovery. *J. Biomol. Screen*. **1996**, *1*, 179−186.

(34) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal Component analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.

(35) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.

(36) Weininger, D. Clustering package, Daylight Chemical Information Systems, Irvine, CA 1993.

(37) Sheridan, R. P.; Bush, B. L. "Patty": A programmable atom typer and language for automatic classification of atoms in molecular database. *J. Chem. Info. Comput. Sci.* **1993**, *33*, 756−762.

CI980231D