Database Searching for Compounds with Similar Biological Activity Using Short Binary Bit String Representations of Molecules

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

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

Received April 24, 1999

In an effort to identify biologically active molecules in compound databases, we have investigated similarity searching using short binary bit strings with a maximum of 54 bit positions. These "minifingerprints" (MFPs) were designed to account for the presence or absence of structural fragments and/or aromatic character, flexibility, and hydrogen-bonding capacity of molecules. MFP design was based on an analysis of distributions of molecular descriptors and structural fragments in two large compound collections. The performance of different MFPs and a reference fingerprint was tested by systematic "one-against-all" similarity searches of molecules in a database containing 364 compounds with different biological activities. For each fingerprint, the most effective similarity cutoff value was determined. An MFP accounting for only 32 structural fragments showed less than 2% false positive similarity matches and correctly assigned on average ∼40% of the compounds with the same biological activity to a query molecule. Inclusion of three numerical two-dimensional (2D) molecular descriptors increased the performance by 15%. This MFP performed better than a complex 2D fingerprint. At a similarity cutoff value of 0.85, the 2D fingerprint totally eliminated false positives but recognized less than 10% of the compounds within the same activity class.

INTRODUCTION

Efforts to explore and exploit molecular similarity[1,2] play a significant role in contemporary chemical and drug discovery research.[3,4] The increasing availability of large compound collections[5,6] and virtual libraries[7,8] has made computational methods an essential component of database analysis, and attempts have been made to identify molecules with specific activities.[8,9] A variety of concepts to mine compound databases and classify molecules according to predefined similarity criteria have been introduced. Recent activities in this area include, for example, the development and assessment of methods to cluster molecules[9] and to select representative subsets from large compound collections,[10] the use of binary QSARs to obtain predictive models for selection of bioactive molecules,[11] and efforts to identify druglike molecules.[12,13]

Searching for compounds that are similar to a query molecule,[1] in two[14] or three[15] dimensions, is a widely applied concept in database mining. Structural or chemical similarity can be defined and measured in many ways.[2,3] The approach has its roots in substructure analysis[14,16] (i.e., searching for molecules that share a structural fragment with the query compound), and increasingly abstract and complex representations of molecules have been constructed. These include binary bit string descriptions of molecules,[17] often called fingerprints. These fingerprints capture, for example, atomic distances or pharmacophore patterns,[18] unique structural paths,[19] or other molecular descriptors.[20,21] These types of fingerprints, as implemented in major software packages,[19,22,23] are highly complex. They are usually hashed (i.e., different patterns or properties are matched to the same or overlapping bit segments) and consist of many bit positions, often more than 1000.

In this study, we have explored similarity searching using much simpler bit string representations of molecules. Other investigations have shown that limited sets of 2D molecular descriptors, including structural key-type fragments (SS-Keys),[24,25] are sufficient to effectively cluster compounds and study their diversity. We have therefore encoded three bit string representations using a few preferred 2D descriptors to systematically carry out all possible similarity searches in a database of biologically active molecules. Fingerprint-based searching for compounds with similar activity is consistent with the idea that structurally similar molecules often exhibit similar biological effects.[3] In early-phase drug discovery, the identification of molecules with defined activity by virtual database screening is a primary goal. However, activity-oriented similarity searching is more complicated than searching for structural similarity. This is due to the fact that different compound classes can show activity against the same or similar biological targets.[28] The results of our study suggest that relatively simple fingerprint-like descriptions of molecules are effective tools for biological similarity searching. These minifingerprints (MFPs) limit the number of false positives and correctly identify more than 50% of the compounds with similar biological activity. They are less sensitive than complex fingerprints to minor structural differences of compounds that do not significantly alter biological activity.

* Corresponding author. Phone: (425) 424-7297. Fax: (425) 424-7299. E-mail: jbajorath@nce-mail.com.
[†] New Chemical Entities, Inc.
[‡] University of Washington.

10.1021/ci990308d CCC: $18.00
Published on Web 07/31/1999

Table 1. Composition of Compound Database for Similarity Search

biological activity	no. of compds
H3 antagonists (H3)	52
carbonic anhydrase II (CA) inhibitors	68
HIV protease (HIV) inhibitors	48
serotonin receptor ligands (5-HT)	71
benzodiazepine receptor ligands (BEN)	59
cyclooxygenase-2 (COX) inhibitors	31
tyrosine kinase (TK) inhibitors	35

MATERIALS AND METHODS

Programs required for our analysis were generated using SVL code[29] and implemented in MOE.[23] These include routines to (1) analyze the distributions of structural keys and numerical values for selected 2D molecular descriptors in compound databases, (2) encode molecular descriptors as bit patterns, (3) calculate different MFPs for each molecule in a compound database, (4) systematically compare the MFP of each compound in the test database to all other compounds, (5) systematically vary the cutoff value for similarity in MFP comparisons, and (6) calculate and list percentage values and scores for MFP performance.

Two conceptually different databases were used to survey the distribution of structural fragments and molecular descriptors, Optiverse (OV)[6] and Maybridge (MB).[30] OV is a combinatorial database based on diversity design and contains 117 976 compounds, while MB consists of 58 239 compounds and intermediates frequently used in medicinal chemistry. In this study, we did not aim to specifically survey descriptor distributions among known drug molecules. As a test database for similarity searching, seven sets of compounds (with each set having a different biological activity) were collected from the literature as described previously.[26] The database consists of a total of 364 compounds, and its composition is summarized in Table 1.

The design of the MFPs analyzed here is described in the Results and Discussion section. As a reference, the ph4_ph2D_Fingerprint[18] (PH2D) was used, as implemented in MOE. This complex 2D fingerprint (1024 bit positions) calculates all pairwise distances for atoms in a molecule on the basis of graphs and represents these values as a pharmacophore or signature key, i.e., a bit string with each position capturing the presence or absence of a unique pattern.

As a similarity cutoff value for pairwise comparison of fingerprints, the Tanimoto coefficient[2,17] (TC) was calculated as $TC = BC/(B1 + B2 - BC)$. BC is the number of common bits set on (i.e., 1), and B1 and B2 are bits set on in the fingerprints of molecules 1 and 2, respectively. Molecules with a TC greater than the specified cutoff value were considered "similar". If these molecules had the same biological activity, their similarity was classified as "correct". If molecules belonged to different activity classes, the assignment was classified as "incorrect". For each molecule in our test database, correct and incorrect matches and their averages were calculated. A simple score was calculated to measure overall MFP performance: $S = (C - I)/N$. C is the number correctly identified and I the number of incorrectly identified compounds. N is the total number of compounds with the same biological activity. The score was set to zero if C was smaller than I. Scores were calculated for each activity class and averaged over all seven classes. For each fingerprint, similarity cutoff values were determined to achieve the best overall performance. Scores were also calculated for a cutoff value of 0.85, which is often considered to correspond to similar biological activity of compared compounds.[21,31]

RESULTS AND DISCUSSION

Molecular Descriptors and Fingerprints. A number of studies have suggested that limited sets of molecular descriptors, in particular SSKey-type structural fragments,[24,25] are sufficient to capture important molecular features. Limited descriptor sets have successfully been used to cluster compound collections, study their diversity, and predict their biological properties.[8,9,20,27] However, highly complex binary bit string representations of molecules (i.e., fingerprints) are mostly used to quantify molecular diversity or similarity.[18,19,21,22] In a previous study,[26] we systematically explored a number of 2D molecular descriptors for compound classification based on principal component analysis.[32,33] We found that a set of 57 structural fragments or keys, similar to MDL SSKeys,[24,25] and 2 or 3 additional 2D descriptors accounting for aromatic character, hydrogen-bonding capacity, and molecular flexibility were sufficient to effectively partition compounds according to biological activity. Here we wished to determine whether short binary bit string representations of molecules based on such descriptors (which are much simpler than the widely used complex fingerprints[18,19,22]) would suffice to identify compounds with similar biological activity via similarity searching, a conceptually different approach.[1,2] To do so, we have constructed MFPs using only the number of aromatic bonds (ARB) and hydrogen-bonding acceptors (HBA),[34−36] the fraction of rotatable bonds (FRB) per molecule, and structural key-type fragments as descriptors. The design of these fingerprints is described below as well as their assessment using a test database of 364 compounds belonging to 7 different biological activity classes (Table 1).

Distribution of the Molecular Descriptors. To identify the appropriate ranges of values for encoding numerical descriptors, the distributions of HBA, ARB, and FRB in two large compound databases, OV and MB, were analyzed (a total of ~176 000 compounds; see Materials and Methods section). The results are shown in Figure 1. The descriptor distributions in these databases vary but display similar trends. HBA values from one to nine are well populated in OV and MB and peak between four and five hydrogen-bond acceptors per molecule. In both databases, ARB values show discrete peaks at six and 12, corresponding to the presence of one or two 6-membered aromatic rings per molecule. Five-membered aromatic rings are less frequently observed. Some compounds in both databases have as many as 40 aromatic bonds in total. FRB values show broad distributions that are well populated between 0.1 and 0.3. We also analyzed the occurrence of 57 previously used SSKey-type fragments[26] in OV and MB to determine if some of these fragments are over- or underrepresented. The results are summarized in Table 2. Two fragments occurred in more than 90% of all compounds, and 23 were found in less than 10% of the compounds. These fragments were omitted from further analysis, and only 32 structural keys were considered that were present in greater than 10% and less than 90% of the

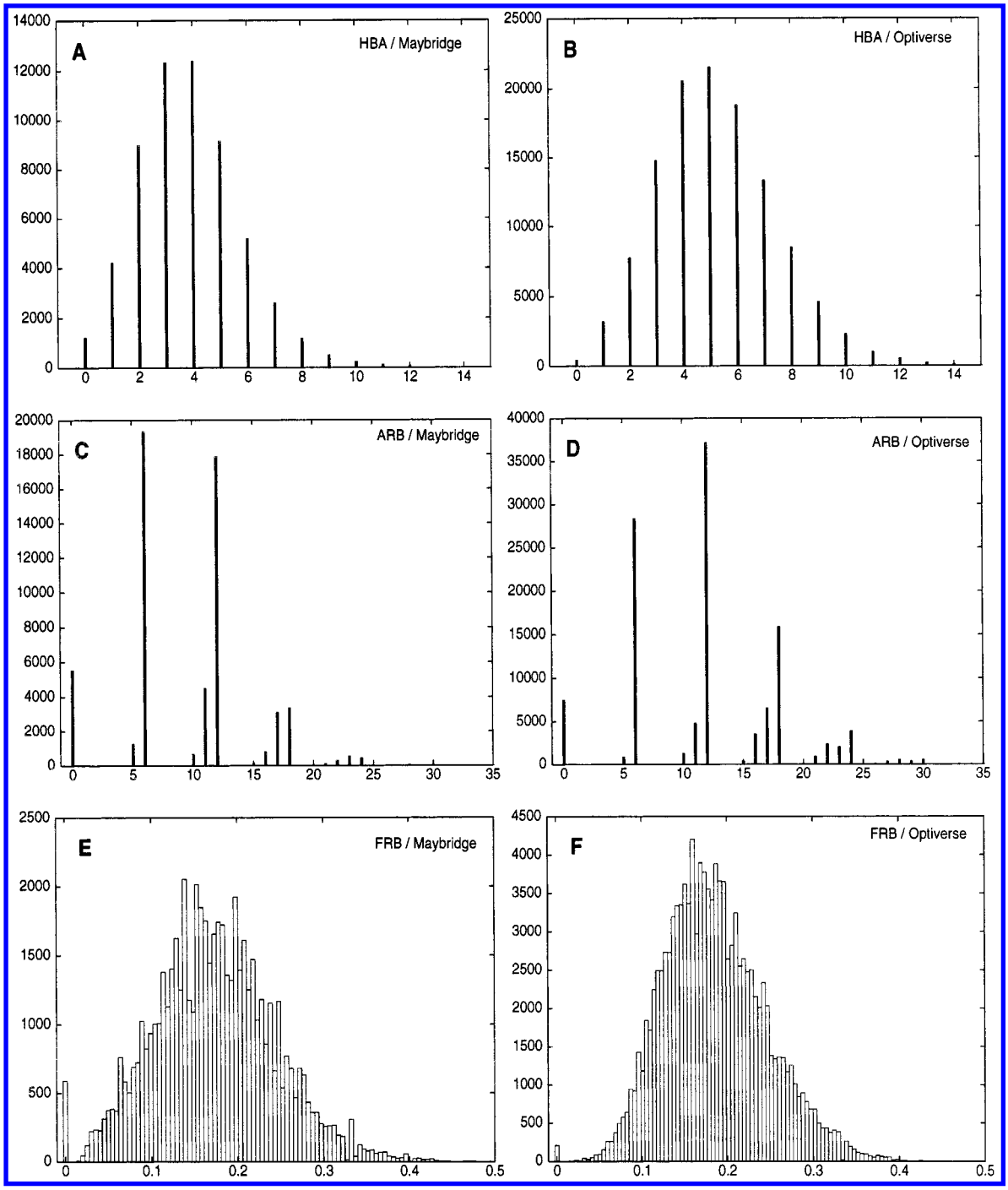

Figure 1. Descriptor distributions. Histograms are shown to summarize the distribution of three numerical molecular descriptors in the MB (panels A, C, and E) and OV (B, D, and F) compound databases (HBA, hydrogen-bond acceptors; ARB, number of aromatic bonds; FRB, fraction of rotatable bonds). In A−D, the number of database compounds (*y* axis) with a specific number of HBAs (A, B) or ARBs (C, D) (*x* axis) is shown. In E and F, the observed FRB range (from 0 to 0.5) is divided into 100 equally spaced bins (*x* axis), and the number of compounds per bin is reported.

database compounds. These values were arbitrarily chosen to exclude fragments that were present in almost all or nearly none of the compounds. However, more stringent values (e.g., 20% and 80%) could have also been selected. A similar descriptor analysis would be very difficult, if not impossible, for highly complex or hashed fingerprints.

Fingerprint Design. Three MFPs were encoded on the basis of the results presented above. Figure 2 shows a schematic representation of the MFP design. "SSKey" is the simplest MFP where each bit position detects the presence or absence of one of the 32 fragments. In "SSKey-3DS", the three numerical descriptors were combined with the SSKey representation. The number of bits assigned to a particular descriptor weights this descriptor relative to others in the similarity calculations. Therefore, the importance of structural keys (32 bits) relative to the 3 numerical descriptors was estimated on the basis of principal component analysis as described[26] with MOE.[33] From the top three principal components that accounted for greater than 80% of the observations, we estimated the relative importance of the descriptors as SSKey 1.0, HBA 0.7, ARB 0.5, and FRB 0.5. Based on the relative importance and the observed numerical

Table 2. Distribution of Structural Key-Type Fragments in Maybridge and Optiverse Databases[a]

SSKey	MB, %	OV, %	SSKey	MB, %	OV, %
[a#X][r]	34	27	C[N][r]	39	79
C[OH]	10	17	NO2	10	12
[a#Q][r5]	33	25	[#Qr][Aq0]$_{2-3}$[#Qr]	22	45
a[OH]	5	10	[#Qr][Aq0]$_{4-5}$[#Qr]	8	19
C[NH]C	15	42	NN	13	2
CN(C)C	10	46	CC(C)(C)[#Q]	8	15
c1ccccc1	82	90	O[#Q][#Q]O	7	13
An[r]a	34	27	[#X]CH3	26	33
SO2	11	13	C=C	35	30
S(=O)	11	13	[#Q]([#X])([#X])[#X]	28	22
C(=O)OC	14	21	[#Q][#Q]([#Q])([#Q])[#Q]	34	36
C(=O)N	33	75	[#Q]CH2[#Q][#Q]CH2[#Q]	15	61
[A#Q][r5]	47	43	aa[Oq0]	23	36
[#Q][r≥9]	22	33	[#XH][#Q][#Q][#XH]	5	18
[#Qr][#Qr]([#Qr])[#Qr]	46	49	OSO	11	13
Aa(a)a	8	15	[#G7]	55	33
[#Q][r6]	92	97	[A#X][r]	91	95
C#C	1	2	[#Qr][Aq0]$_{10-11}$[#Qr]	<1	1
OC(=O)O	<1	<1	[#Qr][Aq0]$_{12-13}$[#Qr]	<1	<1
AC(=O)OH	2	<1	[#Qr][Aq0]$_{\geq 14}$[#Qr]	<1	<1
CC(=O)OH	4	3	[#XH][#XH]	4	<1
a[NH2]	4	<1	[#Qr][Aq0]$_{6-7}$[#Qr]	2	7
C[NH2]	6	1	[#Qr][Aq0]$_{8-9}$[#Qr]	<1	3
SO2NH2	0	0	[#X][#G7]	<1	0
C(=O)H	<1	<1	NC(C)N	<1	<1
[#X](=O)OH	<1	<1	OS=O)N	0	0
[#Q][r7]	2	3	S[#Q]N	7	7
[#Q][r8]	2	2	[#Q]CH2OH	1	4
[A#Q]=S	7	5			

[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], number of bonds in a ring; A, nonaromatic atom; a, aromatic atom; #, triple bond; =, double bond. For example, [#X] CH3 describes a methyl group bound to any non-carbon, non-hydrogen atom, and [#Q][r6] describes any non-hydrogen atom attached to a six-membered ring (SSKey = structural keys, OV = Optiverse, and MB = Maybridge). For each SSKey, the percentage of compounds in the databases containing the fragment is given. SSKeys below the dashed line occur in less than 10% or more than 90% of the database compounds and were not used to design MFPs.

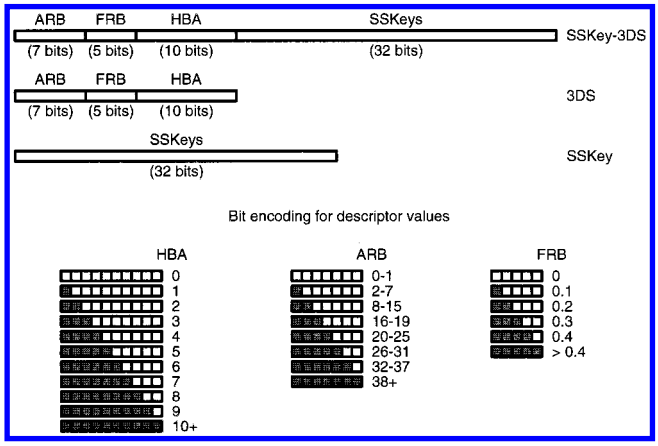

Figure 2. Fingerprint design. A schematic representation of three MFPs is shown. "SSKey" uses a 32 bit representation to account for the presence (i.e., "1") or absence ("0") of 32 structural fragments. "3DS" uses 22 bits to encode numerical values for HBA, ARB, and FRB, as indicated by gray shading of bit segments (i.e., gray means "1"). For example, a molecule with 4 hydrogen-bonding acceptors, 18 aromatic bonds, and a value of 0.2 for FRB would be encoded as the following bit string: [1111000000 1110000 11000]. "SSKey-3D" combines numerical descriptors and structural keys in a 54 bit representation.

ranges of the descriptors, bits to encode these descriptors were set as follows: HBA, 10 bits; ARB, 7 bits; and FRB, 5 bits. Thus, SSKey-3DS consists of 54 bits, and 3DS, representing only 3 numerical descriptors, consists of 22 bits, as illustrated in Figure 2. In contrast to folded or hashed fingerprints, the MFPs are "keyed"; i.e., each bit position is associated with one particular fragment or value. These simple molecular representations were tested and compared to PH2D as described in the following.

MFP Performance. Similarity searches were carried out for each of the 364 compounds in our test database against all other compounds using 3 MFPs and PH2D. In each case, similarity cutoff values for fingerprint overlap were systematically varied in 0.01 increments between 0 and 1. Thus, ~147 000 similarity searches were carried out in total. The overall performance and cutoff values are reported in Table 3. The best overall performance was observed for cutoff values between 0.6 and 0.7 rather than 0.85, which is widely accepted as an indication of a close correlation of compound structure and biological activity. As further discussed below, activity classes in our database contain different structural subclasses. Therefore, in this case, lower cutoff values may capture activity relationships in a more effective way. It is important to note that all cutoff values between 0 and 1 were systematically tested in our calculations. In our analysis, the best performing MFP was SSKey-3DS, which correctly recognized 54% of the similar compounds at a cutoff value of 0.7 and 2.4% false positives. 3DS, consisting of only three numerical descriptors, was least effective, but addition of 3DS to SSKey improved the performance of the structural keys by 15%. These findings are in accord with previous results obtained for compound classification.[26] PH2D recognized a maximum of 35% of the compounds with similar activity at a cutoff value of 0.6 and only 0.5% false positives. At the more stringent similarity cutoff of 0.85, PH2D totally

Table 3. Overall Performance of Fingerprints and Similarity Cutoff Values[a]

	cutoff	correct, %	incorrect, %	score	cutoff	correct, %	incorrect, %	score
3DS	0.85	47	7	0.22	0.93	24	1	0.24
SSKey	0.85	22	0.03	0.21	0.69	39	1.7	0.31
SSKey-3DS	0.85	24	0.03	0.24	0.70	54	2.4	0.40
PH2D	0.85	9	0	0.09	0.60	35	0.5	0.32

[a] On the left, scores are reported for a similarity cutoff value for a fingerprint overlap of 0.85. On the right, the similarity cutoff value that yields the optimum score for each fingerprint is shown. Scores are $S = (C - I)/N$. C is the number of correctly identified compounds in the similarity search (i.e., from the same biological activity group). I is the number of compounds that are incorrectly identified (i.e., from a different biological activity group). N is the total number of compounds in the same biological activity group. The score was set to zero if C was smaller than I.

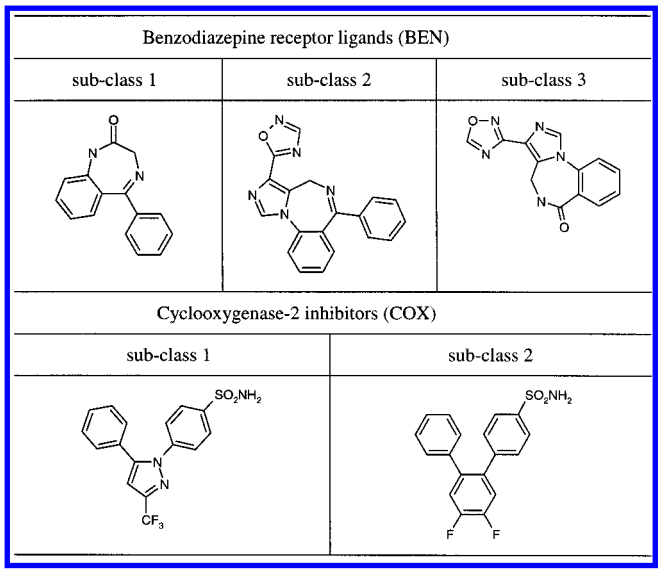

Figure 3. Representative subclasses of bioactive compounds. For two activity classes, different prototypic structures are shown. Each of these structures represents the conserved core of structurally distinct groups of compounds having similar biological activity.

eliminated the false positives but recognized only 9% of the compounds with the same activity. These observations suggest that this complex fingerprint is too sensitive to fine structural details in similarity searching focused on biological activity. In many cases, compounds with equivalent biological activity display some significant structural variations. This is illustrated in Figure 3, which shows subclasses of compounds with the same biological activity that were not recognized as "similar" using PH2D. On the basis of our calculations, the more limited "resolution" of SSKey-3DS is better suited to identify compounds with biological activities similar to a query molecule.

How did the fingerprints perform for different activity classes in our database? Table 4 summarizes the results obtained for each of the seven classes and reports the number of correctly and incorrectly identified molecules on a per compound basis. This allows a more detailed assessment than the overall MFP performance. The relative performance of the tested fingerprints varies somewhat from class to class. None of the seven activity classes appears to be an "easy" case, suggesting that our database represented a meaningful test system. The results show that 3DS found more incorrect than correct molecules per compound and confirm that this rudimentary molecular description is insufficient for similarity searching. However, detection of only 32 structural fragments, as encoded in SSKey, showed increased discriminatory power. At the best cutoff value, SSKey performed almost as good as SSKey-3DS, when assessed on a per compound basis. On average, SSKey correctly recognized 19 similar molecules per compound and 4 false positives, while SSKey-3DS correctly recognized 28 similar molecules and 8 false positives. At this level, PH2D detected no false positives but only 11 similar molecules per compound. At a similarity cutoff of 0.85, few, if any, molecules were incorrectly recognized by SSKey-3DS, SSKey, and PH2D,

Table 4. Class-Specific Performance of Different Fingerprints in Similarity Searching[a]

	BEN	CA	H3	TK	5-HT	HIV	COX
3DS							
Cutoff Value: 0.85							
correct	16.59	61.00	14.04	17.63	23.28	20.79	18.42
incorrect	19.78	19.46	4.33	49.43	16.62	25.13	48.23
score	0.10	0.61	0.26	0	0.14	0.21	0
Cutoff Value: 0.70							
correct	24.25	68.00	27.08	33.46	44.77	31.79	29.32
incorrect	83.32	79.12	22.60	155.54	85.75	117.29	159.52
score	0	0.15	0.45	0	0.01	0	0
SSKey-3DS							
Cutoff Value: 0.85							
correct	16.56	15.68	11.31	7.69	10.83	15.96	10.10
incorrect	0	0.19	0.07	0.06	0.03	0	0.54
score	0.28	0.23	0.22	0.22	0.15	0.33	0.31
Cutoff Value: 0.70							
correct	23.00	52.88	24.35	28.20	23.82	21.16	23.06
incorrect	10.10	6.12	1.44	12.49	8.48	3.50	13.68
score	0.26	0.68	0.45	0.45	0.22	0.37	0.39
SSKey							
Cutoff Value: 0.85							
correct	18.93	6.24	15.08	4.66	10.44	15.13	7.90
incorrect	0	0	0	0.46	0.23	0	0
score	0.32	0.09	0.29	0.12	0.14	0.31	0.25
Cutoff Value: 0.70							
correct	22.08	19.24	25.85	15.11	19.45	18.88	12.74
incorrect	6.37	2.68	0.44	5.34	7.27	0.25	5.87
score	0.29	0.25	0.50	0.30	0.18	0.39	0.24
PH2D							
Cutoff Value: 0.85							
correct	9.78	1.53	4.38	3.29	6.61	3.46	3.00
incorrect	0	0	0	0	0	0	0
score	0.17	0.02	0.08	0.09	0.09	0.07	0.10
Cutoff Value: 0.70							
correct	20.42	4.82	12.92	5.40	11.51	14.50	8.29
incorrect	0	0	0.02	0.09	0.03	0	0
score	0.34	0.07	0.25	0.15	0.16	0.30	0.28

[a] For each fingerprint and activity class, the average number of recognized similar compounds per query compound with the same biological activity is reported ("correct") as well as the average number of similar compounds with different activity ("incorrect"). Activity classes are abbreviated as in Table 1. Results are reported for two similarity cutoff values (0.70 and 0.85) for fingerprint overlap between two molecules. Scores were calculated as in Table 3.

and per compound, SSKey-3DS correctly identified 13 molecules, SSKey found 11, and PH2D identified 5. In the calculations reported in Table 4, SSKey-3DS performed better than PH2D in all but one case (BEN). Thus, in summary, we consider SSKey-3DS as the preferred molecular representation for the activity-based similarity analysis presented herein.

CONCLUSIONS

In this study, we have investigated whether simple bit string representations of molecules are appropriate tools for similarity searching in compound databases. To augment fingerprint design, a statistical analysis of descriptor distributions in large compound databases was carried out. We have systematically analyzed all possible combinations of 364 active test compounds, 4 fingerprint representations, and 101 similarity cutoff values and determined the optimum performance levels. The results suggest that conceptually simple and short bit string representations (32−54 bit positions), only accounting for a small number of structural keys and three 2D descriptors, function well in searching for compounds with similar biological activity. Highly complex fingerprints may be too sensitive to structural variations in compounds with similar biological activity. The results of this analysis further emphasize the value of structural key-type molecular descriptors.

REFERENCES AND NOTES

(1) Downs, G. M.; Willett, P. Similarity searching in databases of chemical structures. *Rev. Comput. Chem.* **1997**, *7*, 1−66.
(2) Willett, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.
(3) Johnson, M., Maggiora, G. M., Eds. *Concepts and applications of molecular similarity*; Wiley: New York, 1990.
(4) Dean, P. M., Ed. *Molecular similarity in drug design*; Chapman and Hall: Glasgow, 1994.
(5) Thompson, L. A.; Ellman, J. A. Synthesis and application of small molecule libraries. *Chem. Rev.* **1996**, *96*, 555−600.
(6) 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.
(7) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual compound libraries: A new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010−1023.
(8) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.
(9) 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.
(10) Higgs, R. E.; Bemis, K. G.; Watson, I. A.; Wikel, J. H. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 861−870.
(11) 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.
(12) Ajay, A.; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem*. **1998**, *41*, 3314−3324.
(13) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(14) Hagadone, T. R. Molecular substructure similarity searching: Effective retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 515−521.
(15) Good, A. C.; Mason, J. M. Three-dimensional structure database searches. *Rev. Comput. Chem.* **1995**, *7*, 67−117.
(16) Cramer, R. D.; Redl, G.; Berkhoff, C. E. Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem*. **1973**, *17*, 533−535.
(17) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 379−386.
(18) Sheridan, R. P.; Bush, B. L. "Patty": A programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762.
(19) James, C. A.; Weininger, D. Daylight theory manual. Daylight Chemical Information Systems, Inc. (URL: www.daylight.com), Irvine, CA, 1995.
(20) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1−9.
(21) Matter H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(22) UNITY. Tripos, Inc. (URL: www.tripos.com), St. Louis, MO, 1995.
(23) MOE (Molecular Operating Environment) Chemical Computing Group, Inc. (URL: www.chemcomp.com), 1255 University St, Montreal, Quebec, Canada H3B 3X3, 1998.
(24) MDL Information Systems, Inc., 14600 Catalina St, San Leandro, CA 94577.
(25) 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.
(26) 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*., in press.
(27) Gillet, 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) Babine, R. E.; Bender, S. L. Molecular recognition of protein−ligand complexes: applications to drug design. *Chem. Rev.* **1997**, *97*, 1359−1472.
(29) Santavy, M.; Labute, P. SVL: The scientific vector language. Journal of the Chemical Computing Group (URL: www.chemcomp.com/feature/svl.htm), 1998.
(30) Maybridge Chemical Co. Ltd., Trevillett, Tintagel, Cornwall PL34 OHW, U.K.
(31) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: A useful concept for validation of molecular descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(32) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal component analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.
(33) Labute, P. QuaSAR-Cluster: A different view of molecular clustering. Journal of the Chemical Computing Group (URL: www.chemcomp.com/article/cluster.htm), 1998.
(34) Fujita, T.; Nishioka, T.; Nakajima, M. Hydrogen-bonding parameter and its significance in quantitative structure−activity study. *J. Med. Chem.* **1977**, *20*, 1071−1081.
(35) Charton, M.; Charton, B. I. The structural dependence of amino acid hydrophobicity parameters. *J. Theor. Biol.* **1982**, *99*, 629−644.
(36) Yang, G.; Lien, E. J.; Guo, Z. Physical factors contributing to hydrophobic constant π. *Quant. Struct.-Act. Relat*. **1986**, *5*, 12−18.

CI990308D