Pacific Symposium on Biocomputing 5:563-572 (2000)

SEARCHING FOR MOLECULES WITH SIMILAR BIOLOGICAL ACTIVITY: ANALYSIS BY FINGERPRINT PROFILING

JEFFREY W. GODDEN[1], LING XUE[1], FLORENCE L. STAHURA[1], and JÜRGEN BAJORATH[1,2]

[1]*New Chemical Entities, Inc., 18804 North Creek Parkway South, Bothell, WA 98011, and*
[2]*Department of Biological Structure, University of Washington, Seattle, WA 98195, USA*

We have recently developed a mini-fingerprint (MFP) representation for small molecules that performs well in database searches for compounds with similar biological activity. The MFP consists of only 54 bit positions that account for numerical ranges of three two-dimensional (2D) descriptors or the presence or absence of defined structural fragments. Here we present an analysis method, termed fingerprint profiling, to systematically compare bit patterns of compounds belonging to different biological activity classes. Some but not all bit positions were variably occupied in seven different activity classes and responsible for the detection of structure-activity differences. The analysis has made it possible to rank bit positions and encoded molecular descriptors according to their importance for our similarity search calculations. Fingerprint profiling can be applied to any keyed bit string representation and should be helpful, for example, to analyze descriptor distributions in large compound databases.

1 Introduction

Binary bit string representations of molecular structure and properties, often called fingerprints, have become popular tools to analyze chemical similarity[1,2]. Widely used fingerprints account for intramolecular atomic distance or connectivity patterns and molecular descriptors[3-5]. Such fingerprints are often highly complex (hashed or folded) and consist of many bit positions (~1,000 or more). In hashed fingerprints, molecular properties and patterns are mapped to overlapping bit segments and, in consequence, single bit positions can not be associated with a specific descriptor or property. Such complex fingerprints are designed to be sensitive to subtle differences in molecular structure and properties.

The assessment of molecular similarity typically relies on pairwise comparison of molecular bit strings using metrics such as the Tanimoto coefficient (Tc)[1], defined as Tc = B/(B1+B2-B), where B is the number of bits set on (i.e., 1) in common in fingerprints of molecules 1 and 2, B1 is the number of bits set on in molecule 1 and B2 the number of bits set on in molecule 2.

We were interested in the design of bit string representations to identify compounds with similar biological activity. Such fingerprints must be able to capture essential chemical features responsible for a specific activity[6]. At the same

time, they should not discriminate between minor structural variations of compounds that are well tolerated within a given activity class[7].

The design of such fingerprints was investigated in a two step analysis. First, we assembled a test database consisting of a total of ~400 compounds belonging to seven distinct biological activity classes[8] and partitioned this database using a classification method based on principal component analysis[9]. In this study, all possible combinations of 57 structural key-type[10,11] fragments (SSKeys) and 17 other 2D molecular descriptors[8] were analyzed. We found that a combination of 32 SSKeys and only three additional 2D descriptors (accounting for hydrogen bonding acceptors, aromatic character, and molecular flexibility) effectively classified compounds with similar biological activity[8]. On the basis of these findings, we designed several small fingerprints (mini-fingerprints, MFPs) and evaluated their performance in exhaustive one-against-all similarity searches in our test database (with systematic variation of Tc cut-off values for detection of similarity)[12].

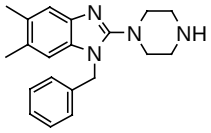
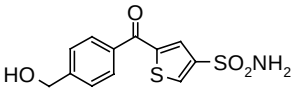
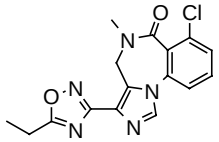
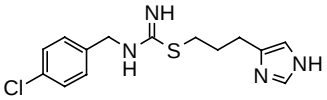
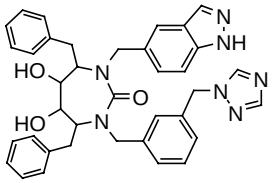
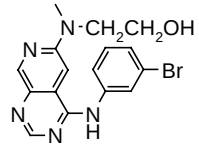
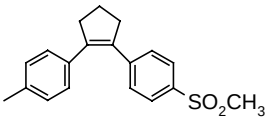
Overall best performance was obtained for an MFP consisting of 54 bit positions (termed SSKey-3DS) that correctly recognized similar biological activity of 54% of the test compounds and showed only ~2 % false positives (Tc = 0.7). In comparison, a complex 2D pharmacophore fingerprint with 1,024 bit positions[5] recognized a maximum of 35% of compounds (Tc = 0.6) with similar biological activity with 0.5% false positives[12]. At a Tc cut-off value of 0.85, this complex fingerprint totally eliminated false positive recognition but found only 9% of compounds with similar activity, while the MFP still correctly recognized 24% of compounds with similar activity with very few (0.03%) false positives. These findings indicated that the "medium resolution" of our MFP design was suitable for similarity searching focused on biological activity, perhaps more so than highly complex fingerprints that may often be too sensitive to structural variations of compounds with comparable activity.

Here we extend the study of MFP performance and introduce fingerprint profiling as an analysis tool. Fingerprint profiles were calculated for seven activity classes by averaging each bit position within the fingerprint for the class. Comparison of fingerprint profiles and calculation of standard deviations of average bit occupancy at each bit position made it possible to identify those MFP positions (and descriptor settings) that were important to distinguish biological activities of test compounds.

2 Methods

The current version of our test database for similarity searching focused on biological activity includes a total of 364 compounds in seven different biological activity classes, collected from the literature as described[8]. The compound composition of the database is summarized in Table 1.

Table 1. Bioactivity classes in the test database.

Biological activity	Number of compounds	Example structure
Serotonin receptor ligands (5-HT)	71	
Carbonic anhydrase II (CA) inhibitors	68	
Benzodiazepine receptor ligands (BEN)	59	
H3 antagonists (H3)	52	
HIV protease (HIV) inhibitors	48	
Tyrosine Kinase (TK) inhibitors	35	
Cyclooxygenase-2 (Cox-2) inhibitors	31	

The MFP analyzed in this study, termed SSKey-3DS, consists of 32 bit positions that detect the presence or absence of a particular structural key and, in addition, 22 bits that account for numerical ranges of three 2D molecular descriptors, identified by systematic compound classification analysis[8] (as described above). These are the number of aromatic bonds in a molecule (ARB), the number of hydrogen bonding acceptors (HBA), and the fraction of rotatable bonds (FRB). The design of SSKey-3D is illustrated in Figure 1. Encoded numerical ranges for these descriptors were determined by a survey of descriptor distributions in large compound databases[12] and the relative importance of SSKeys, ARB, HBA, and FRB (and thus the lengths of their bit segments) was estimated from principal component analysis in activity-based compound classification[8,12].

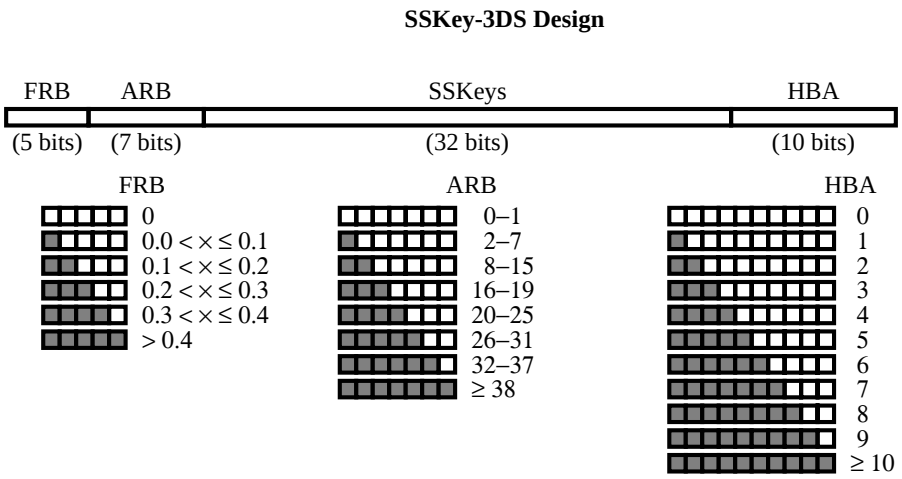


Figure 1: Mini-fingerprint design. A schematic representation of SSKey-3DS is shown. Each bit position reports the presence (i.e., "1") or absence (i.e., "0) of a structural fragment or, alternatively, sets a numerical range for ARB (aromatic bonds), HBA (hydrogen bonding acceptors), or FRB (fraction of rotatable bonds), as indicated by gray shading of bit segments (gray means "1").

The MFP and all routines required for systematic database searching were generated using SVL code[13] and implemented in MOE[14] that was used for all calculations. For each compound in our test database, SSKey-3DS was generated. Then, the sum of bit positions was calculated for each activity class and divided by the number of compounds belonging to this class, thus providing the average bit occupancy at each position. The profiles were compared and standard deviations of bit occupancy were calculated for each bit position.

3 Results

Fingerprint profiles of the seven activity classes and a consensus profile for all activity classes are shown in Figure 2. The results show that activity classes display significant differences in their patterns of bit occupancy.

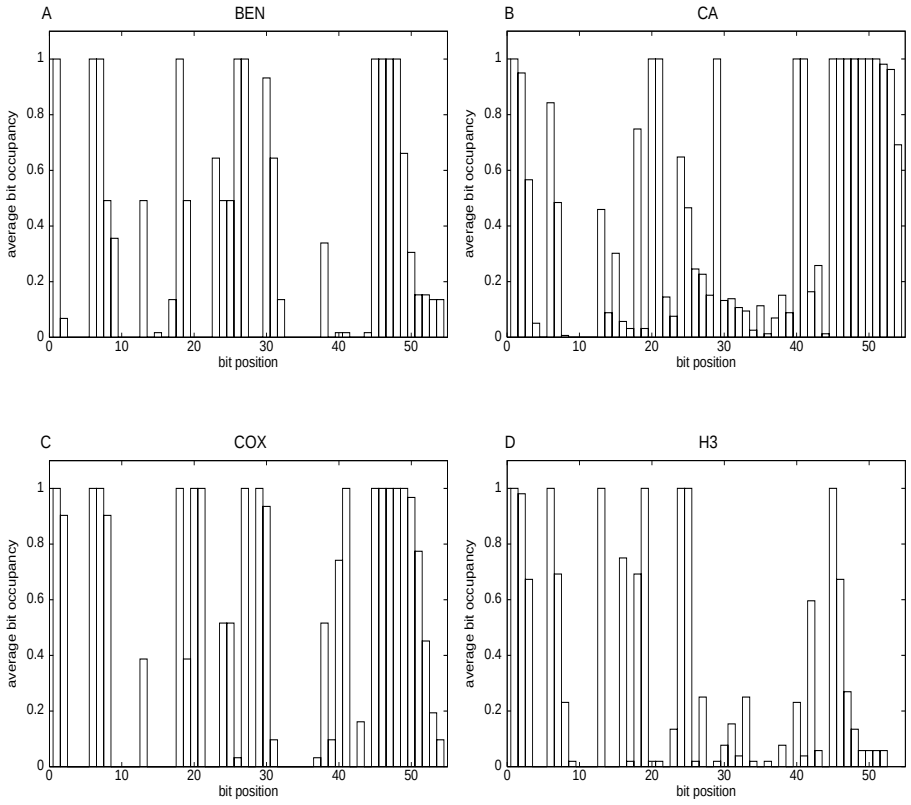
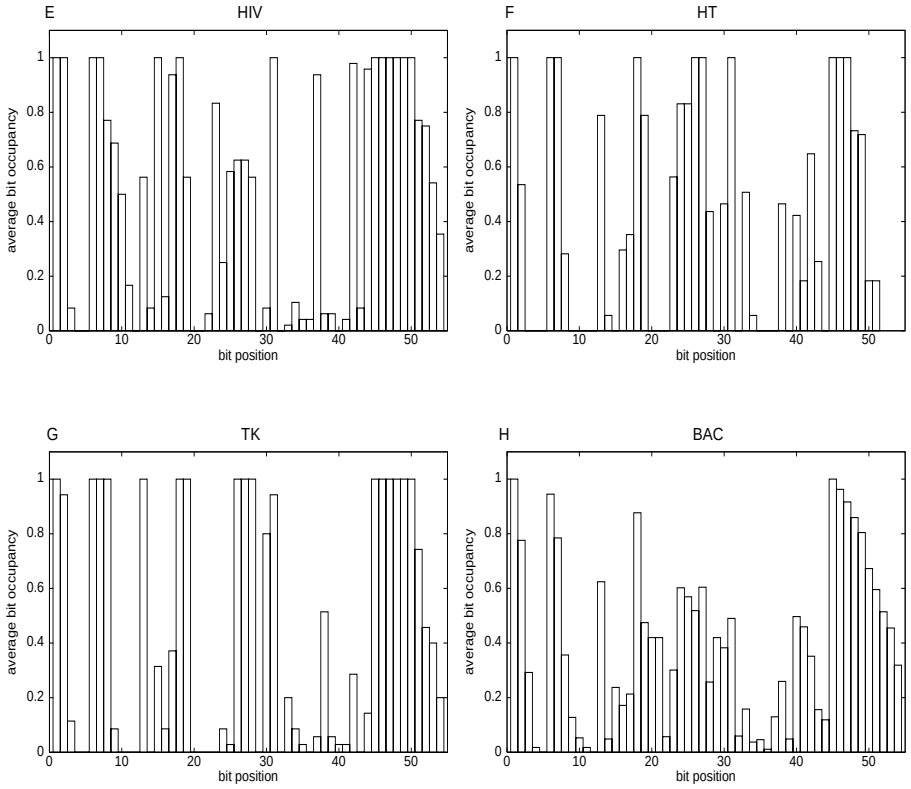


Figure 2 (continued on next page): Fingerprint profiles for compounds belonging to different activity classes. The average of each of the 54 bit positions of SSKey-3DS is shown for each activity class and the sum of all classes. A. Benzodiazepine receptor ligands (BEN) B. Carbonic anhydrase-II inhibitors (CA) C. Cyclooxygenase-2 inhibitors (COX) D. H3 antagonists (H3) E. HIV protease inhibitors (HIV) F. Serotonin receptor ligands (HT) G. Tyrosine kinase inhibitors (TK) H. All activity classes (BAC).



The MFP profiles display high variability in some regions and are similar in others. This is illustrated in the consensus profile (panel H in Figure 2). Regardless of the activity class, a few bit positions were rarely or (in two instances) never set on. By contrast, several other bit positions were often or, in two cases, always set on. Bit positions rarely or mostly set on in all activity classes (approximately 10 in total) did not contribute to the detection of activity differences. Thus, in our search calculations, MFP performance was dependent on a subset of approximately 40 bit positions and their combinations. Most important were bit positions with high variability of occupancy. These positions were identified by calculation of standard deviations of bit occupancy over all compound classes, as shown in Table 2.

Table 2. Variability of SSKey-3DS Fingerprint within Biological Activity Classes

Standard Deviation	Bit Occupancy	Bit Position	Descriptor
0.5000	0.4967	40	Non-H atom linked to 3 heteroatoms
0.4999	0.4901	31	Nitrogen attached to α-carbon of aromatic system
0.4998	0.5143	52	HBA=8
0.4997	0.5187	26	9-membered or larger (fused) ring
0.4994	0.4747	19	Nitrogen-containing aromatic ring
0.4983	0.4593	41	Quaternary atom
0.4980	0.4549	53	HBA=9
0.4952	0.5692	25	5-membered aromatic ring
0.4935	0.4198	20	$-SO_2$
0.4935	0.4198	21	-SO
0.4935	0.4198	29	-OSO
0.4908	0.5956	51	HBA=7
0.4894	0.6022	24	5-membered non-aromatic ring
0.4890	0.6044	27	Fused ring system
0.4860	0.3824	30	Halogen atom
0.4843	0.6242	13	Heterocycle
0.4788	0.3560	8	ARB 16 to 19
0.4775	0.3516	42	2 methylenes separated by 2 atoms
0.4693	0.6725	50	HBA=6
0.4660	0.3187	54	HBA >10
0.4587	0.3011	23	Amide
0.4548	0.2923	3	FRB 0.2 to 0.3
0.4383	0.2593	38	Methyl attached to hetero atom
0.4371	0.2571	28	Fused aromatic ring system
0.4255	0.2374	15	Aliphatic OH
0.4170	0.7758	2	FRB 0.1 to 0.2
0.4111	0.7846	7	ARB 8 to 15
0.4096	0.2132	17	Aliphatic tertiary amine
0.3967	0.8044	49	HBA=5
0.3769	0.1714	16	Aliphatic secondary amine
0.3650	0.1582	33	Rings separated by 2-3 non-ring atoms
0.3629	0.1560	43	Non-ring oxygen attached to aromatic system
0.3477	0.8593	48	HBA=4
0.3359	0.1297	37	Oxygens separated by 2 atoms
0.3335	0.1275	9	ARB 20 to 25
0.3285	0.8769	18	Phenyl ring
0.3234	0.1187	44	2 non-C,H atoms separated by 2 atoms
0.2767	0.9165	47	HBA=3
0.2363	0.0593	32	$-NO_2$
0.2321	0.0571	22	Ester
0.2279	0.9451	6	ARB 2 to 7
0.2235	0.0527	10	ARB 26 to 31
0.2145	0.0484	14	Aromatic OH
0.2145	0.0484	39	Double bond
0.2098	0.0462	35	NN
0.1896	0.0374	34	Rings separated by 4-5 non-ring atoms
0.1896	0.9626	46	HBA=2
0.1314	0.0176	4	FRB 0.3 to 0.4
0.1314	0.0176	11	ARB 32 to 37
0.1043	0.0110	36	C attached to 3 carbons and a hetero atom
0.0000	1.0000	1	FRB 0 to 0.1
0.0000	0.0000	5	FRB >0.4
0.0000	0.0000	12	ARB >38
0.0000	1.0000	45	HBA=1

Non-ARB, -FRB, and -HBA bits detect the presence or absence of SSKey-type fragments.

The results in Table 2 helped to understand the relative importance of molecular descriptors encoded in SSKey-3DS (standard deviations calculated here range from 0.0-0.5 and their differences are subtle due to the binary setting of bit values). A few low and high ranges of numerically encoded descriptors (e.g., HBA = 1, ARB > 38; bottom of Table 2) were, as to be expected, detected in all or none of the test compounds and could therefore be excluded from further MFP designs. The lower part of the table characterizes features that were common to compounds in our database, regardless of their activity (e.g., aromatic character, certain functional groups etc.; see also Table 1). The upper part shows descriptors with variable bit occupancy that were more important to distinguish compounds with different activity. Among the top scoring descriptors were a variety of structural keys (known to be powerful 2D descriptors[8,15]). An interesting observation is that higher ranges of hydrogen bonding acceptors (i.e., HBA 7-9) were very important, more so than other numerically encoded descriptors.

4 Discussion

Approaches to database searching for compounds with similar activity are conceptually based on the idea that significant similarities in molecular structure and properties are responsible for similar biological activity[6]. However, structure and activity can relate in many different ways and it is difficult to generate molecular representations that capture structure-activity relationships for diverse sets of molecules. On the basis of extensive molecular descriptor analysis, we investigated the design of short binary bit string representations of molecules that are conceptually much more simple than other commonly used fingerprints[8,12].

Results obtained so far include complete similarity searches for ~400 compounds belonging to seven different biological activity classes. Our findings indicate that the use of relatively few and simple descriptors is sufficient to describe molecular features at a level of resolution suitable to distinguish biological activities. However, our current studies have at least two significant limitations. First, the number of activity classes analyzed to date is still small, and the results may vary to some extent with the composition of the compound database used for benchmarking. Second, we currently do not differentiate between activity levels of compounds belonging to the same class (e.g., weakly versus highly active), akin to the scenario of binary QSAR calculations[16,17].

In order to analyze MFP performance in more detail, we have calculated fingerprint profiles. As an analysis tool, fingerprint profiling can be immediately applied to any keyed fingerprint-type representation (i.e., where each bit is associated with a particular pattern or descriptor). It can also be used to study descriptor distributions in large compound database and identify structural keys and descriptors that characterize, for example, molecules with drug-like properties[18,19].

It is important to note that fingerprint profiling is a diagnostic tool to visualize and study patterns and variability of bit occupancy and not in itself a method to identify or classify biologically active compounds[8,20,21].

When analyzing our MFP and benchmark compound database, we found that different compound classes showed characteristic patterns of bit occupancy. Analysis of these patterns, aided by calculation of standard deviations, has provided a differentiated view of our MFP design and made it possible to rank molecular descriptors according to their importance. Only a subset of bit positions, with highly variable occupancy, was responsible for the performance of the MFP, consistent with the idea that combinations of relatively few encoded descriptors suffice to capture essential features of tested compounds.

References

1. D. R. Flower, "On the properties of bit string-based measures of chemical similarity" *J. Chem. Inf. Comput. Sci*. **38**, 379 (1998)
2. P. Willett, "Chemical similarity searching", *J. Chem. Inf. Comput. Sci*. **38**, 983 (1998)
3. C. A. James, D. Weininger, D, Daylight theory manual. Daylight Chemical Information Systems, Inc. (URL: www.daylight.com), Irvine, CA (1995)
4. UNITY. Tripos, Inc. (URL: www.tripos.com), St. Louis, MO (1995)
5. R. P. Sheridan, B. L. Bush, "Patty: a programmable atom typer and language for automatic classification of atoms in molecular databases" *J. Chem. Inf. Comput. Sci*. **33**, 756 (1993)
6. "Concepts and applications of molecular similarity" Eds. M. Johnson, G. M. Maggiora (Wiley, New York, 1990)
7. R. E. Babine, S. L. Bender "Molecular recognition of protein-ligand complexes: applications to drug design" *Chem. Rev*. **97**, 1359 (1997)
8. L. Xue, J. Godden, H. Gao, J. Bajorath "Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis" *J. Chem. Inf. Comput. Sci*. **39**, 699 (1999)
9. P. Labute "QuaSAR-Cluster: A different view of molecular clustering" (URL: www.chemcomp.com/article/cluster.htm), Chemical Computing Group, Inc., Montreal (1998)
10. M. J. McGregor, P. V. Pallai "Clustering of large databases of compounds: using MDL "Keys" as structural descriptors" *J. Chem. Inf. Comput. Sci*. **37**, 443 (1997)
11. MDL Informations Systems, Inc., San Leandro, CA (1997)
12. L. Xue, J. W. Godden, J. Bajorath "Database searching for compounds with similar biologcial activity using short binary bit string representations of molecules" *J. Chem. Inf. Comput. Sci*. **39**, in press (1999)

13. M. Santavy, P. Labute, "SVL: The scientific vector language" (URL: www.chemcomp.com/feature/svl.htm), Chemical Computing Group, Inc., Montreal (1998)
14. MOE (Molecular Operating Environment) (URL: www.chemcomp.com), Chemical Computing Group, Inc., Montreal (1998)
15. R. D. Brown, Y. C. Martin "The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding" *J. Chem. Inf. Comput. Sci*. **37**, 1 (1997)
16. P. Labute "Binary QSAR: A new method for the determination of quantitative structure activity relationships" *Pac. Symp. Biocomput.* **7**, 444 (1999)
17. H. Gao, C. Williams, P. Labute, J. Bajorath "Binary quantitative structure-activity relationship (QSAR) analysis of estrogen receptor ligands" *J. Chem. Inf. Comput. Sci*. **39**, 164 (1999)
18. V. J. Gillet, P. Willett, J. Bradshaw "Identification of biological activity profiles using substructural analysis and genetic algorithms" *J. Chem. Inf. Comput. Sci*. **38**, 165 (1998)
19. A. Ajay, W. P. Walters, M. A. Murcko "Can we learn to distinguish between "drug-like" and "nondrug-like" molecules?" *J. Med. Chem.* **41**, 3314 (1998)
20. G. Klopman. "The MultiCASE program II. Baseline activity identification algorithm (BAIA)" *J. Chem. Inf. Comput. Sci*. **38**, 78 (1998)
21. X. Chen, A. Rusinko, S. S. Young. "Recursive partitioning analysis of a large structure-activity data set using three-dimensional descriptors. *J. Chem. Inf. Comput. Sci*. **38**, 1054 (1998)