DOI: 10.1002/cmdc.200700199

Similarity Searching using Compound Class-Specific Combinations of Substructures Found in Randomly Generated Molecular Fragment Populations

José Batista and Jürgen Bajorath*[a]

Substructure- or fragment-type descriptors have long been widely used and effective tools for chemical similarity searching[1,2] and other applications in chemoinformatics and computer-aided drug discovery.[2–4] Currently available substructure-type descriptors are generally well designed and based on chemical knowledge, predefined molecular organization schemes,[5,6] or retrosynthetic criteria.[7,8] Popular sets of fragment descriptors include MACCS structural keys (166 publicly available fragments)[9,10] or the BCI standard dictionary (1052 fragments).[11,12] Early substructure generation methods systematically derived molecular fragments[13] or grew fragments along evolutionary trees.[14] Such procedures generated series of fragments whose presence depended on each other.[15] Given the large numbers of systematically derived fragments, statistical analyses were employed to identify frequencies of fragment occurrence and weight fragments accordingly in database searching[16] or generate sets of equifrequently occurring fragments.[17] For similarity searching, fragment-type descriptors are typically encoded as molecular fingerprints where each bit position accounts for the presence or absence of a particular fragment.[1,2]

Herein, we depart from systematic or knowledge-based substructure design and, by contrast, mine randomly generated fragment populations for substructures that are associated with different compound classes. In this study, we demonstrate that activity class-specific combinations of random substructures can be systematically identified and used as fingerprints for similarity searching. These findings open up new avenues for the generation of structural descriptors and compound class-directed fingerprints.

The conceptual basis for our substructure analysis is provided by previous studies where we have shown that random fragment populations generated with MolBlaster[18] can be used to detect molecular similarity relationships. MolBlaster randomly deletes rows in connectivity tables of test molecules and samples the resulting fragments. Compounds having similar activity have been identified by comparing their fragment populations using information-theoretic metrics.[18,19] A major conclusion from these studies has been that random fragment populations must contain specific molecular information. What exactly is this information? This question has been addressed by organizing fragment populations as tree structures that capture conditional probabilities of fragment occurrence.[20] The approach is described in Figure 1. For the purpose of our analysis, Activity Class-Characteristic Fragments (ACCS) are defined as fragments that are produced by at least two active molecules within a reference set but no compounds of a background database. Fragment trees are found to contain pathways with ACCS combinations that are specific for different compound activity classes.[20]

We now ask two fundamental questions: First, can combinations of random fragments be used as substructure descriptors for different activity classes? Second, are such molecular representations capable of detecting diverse structure–activity relationships? It is intuitive that active compounds should contain structural patterns that distinguish them from inactive ones. However, key issues of our analysis are whether random fragment populations contain this information and, in addition, whether predictive patterns can be isolated from them.

To address these questions, we have analyzed five high-throughput screening data sets available in PubChem.[21] These data sets include three screens for cathepsin B, L, and S, cysteine protease inhibitors, a screen for JNK3 tyrosine kinase inhibitors, and another one for protein kinase A inhibitors. A summary is provided in Table 1. We have chosen experimental screening data sets because they consist not only of confirmed active but also confirmed inactive compounds and contain the type of hits one searches for in practical in silico screening applications. Furthermore, hits in screening data sets are often structurally diverse and thus provide challenging test cases for the analysis of structure–activity relationships. The structural diversity of active compounds in all five screening sets is reflected by low average pairwise Tanimoto similarity reported in Table 1 and can be further appreciated in Supporting Information Figure 1 that shows representative examples of hits.

Considering the total number of hits available in each screening set, ten subsets of 11–16 active molecules were randomly taken from each set as reference molecules (Table 1). Each reference set was fragmented together with 500 randomly selected ZINC compounds[22] using 3000 MolBlaster iterations with randomized numbers of deletions per step, as described previously.[19] For each reference set, the resulting fragment populations were used to determine cumulative numbers of ACCS for the top three levels (0, 1, 2) of their fragment trees. In Table 2, we report average numbers of ACCS for all reference sets. Independent of their biological activity, active reference molecules from each screening set consistently produced ACCS. At tree level 0, the average number of character-

[a] *J. Batista, Prof. Dr. J. Bajorath*
Department of Life Science Informatics, Bonn–Aachen International Center for Information Technology, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, 53113 Bonn (Germany)
Fax: (+49) 228-2699-341
E-mail: bajorath@bit.uni-bonn.de

Supporting information for this article is available on the WWW under http://www.chemmedchem.org or from the author.

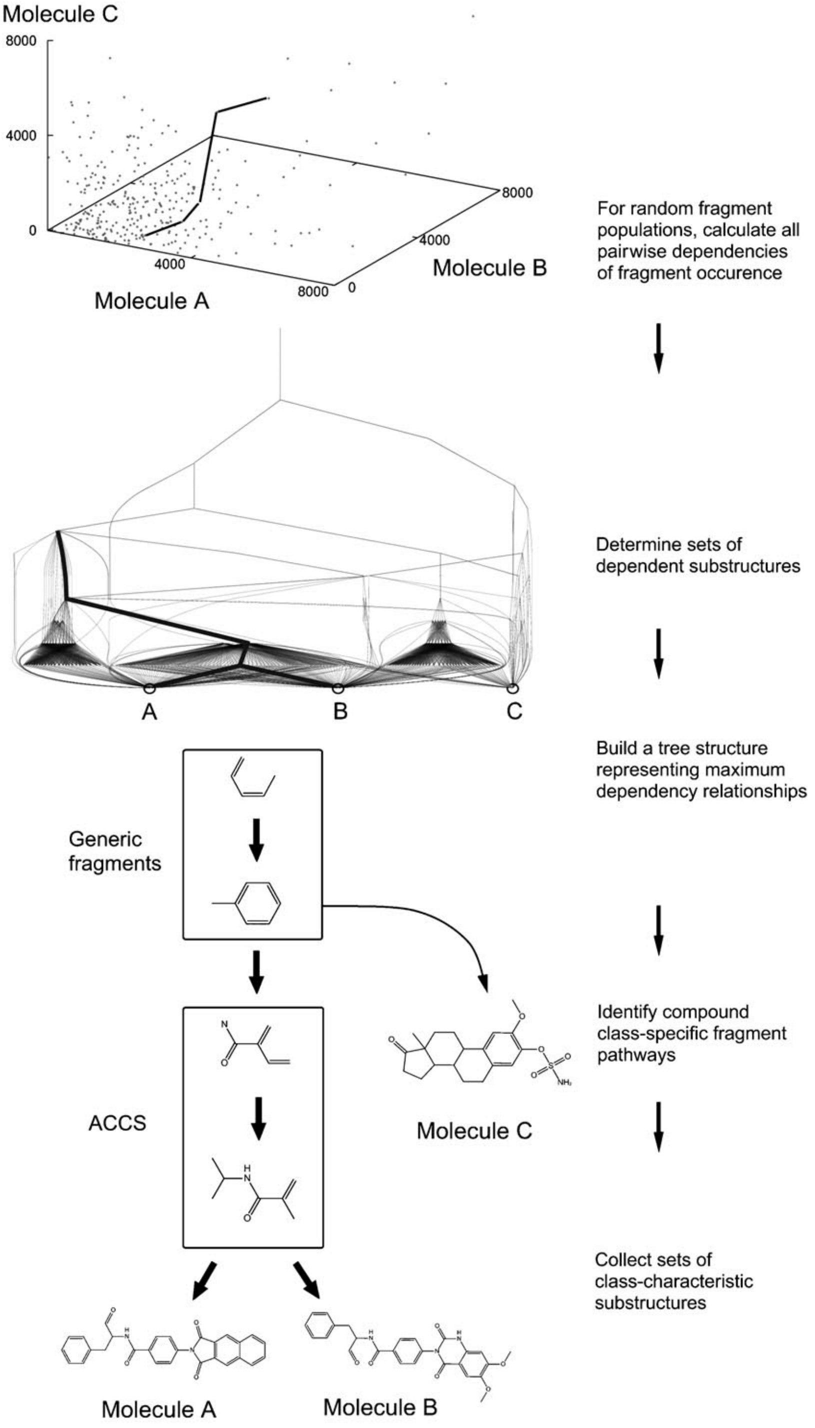
Molecule C
8000
4000
0
4000
8000
Molecule A
0
4000
8000
Molecule B
For random fragment populations, calculate all pairwise dependencies of fragment occurence
Determine sets of dependent substructures
A
B
C
Build a tree structure representing maximum dependency relationships
Generic fragments
Identify compound class-specific fragment pathways
Molecule C
ACCS
Collect sets of class-characteristic substructures
Molecule A
Molecule B

Figure 1. Identification of activity-class specific fragment pathways. The general procedure is outlined in the flow chart on the right. On the left, an example is shown; three molecules (A, B, and C), two of which (A and B) share the same activity. At the top, randomly generated fragment populations of these molecules are displayed within a reference system accounting for relative frequencies of fragment occurrence. From this representation, a tree structure is calculated that captures conditional probabilities of fragment co-occurrence. In tree structures, fragments are organized at different levels. Level 0 represents the begin of pathways, level 1 defines direct dependence on root fragments, level 2 second order dependence, and so on. Tree levels reflect different degrees of fragment generality (that is, fragments become increasingly characteristic for individual molecules). An exemplary fragment pathway leading to molecules A and B is highlighted in the fragment population graph and the corresponding tree structure. The first two fragments within this path (shown below the tree) also occur in the random fragment population of the inactive molecule C. By contrast, the other two fragments are only found in the fragment populations of molecules A and B and thus meet our criteria for activity class-characteristic fragments (ACCS). Thus, the section of this pathway that consists of ACCS is activity-class specific. Random fragments only found in active molecules are called class-characteristic because they often do not occur in all compounds within a class and might also be produced by a database compound as background databases grow in size. By contrast, hierarchical combinations of ACCS in fragment trees are unique features of activity classes. Therefore, fragments taken from activity class-specific pathways are considered class signatures.

Table 1. Screening data sets.[a]

Code	Inhibitors	N_{actives}	$N_{\text{scaffolds}}$	Activity range	Avg. Tc	N_{ref}	$N_{\text{inactives}}$
CAB	Cathepsin B	36	26	46 nm–44 μm	0.45	12	63 287
CAL	Cathepsin L	49	39	3 nm–36 μm	0.43	16	57 764
CAS	Cathepsin S	34	28	4 nm–33 μm	0.54	12	61 723
JNK	JNK3	33	22	1 nm–15 μm	0.37	11	8420
PKA	PKA	94	62	682 nm–357 μm	0.45	16	64 797

[a] A summary of the screening data used in our analysis is provided. N_{actives} is the number of hits per data set and $N_{\text{scaffolds}}$ the number of unique scaffolds that represent these hits. Activity range reports the IC_{50} value range for the hits. $N_{\text{inactives}}$ gives the number of inactive screening set compounds and N_{ref} the number of active reference molecules used in similarity search calculations. The average Tanimoto coefficient (Avg. Tc) for pairwise comparison of hits is calculated using MACCS keys and reflects the structural heterogeneity of active compounds. Tc is defined as Nab/(Na+Nb −Nab) where Na is the number of bits set on in the fingerprint of molecule a, Nb the number of bits set on in b, and Nab the number of bits common to both molecules. All screening sets are publicly available in PubChem-Bioassays under the following AIDs: CAB, 453; CAL, 460; CAS, 501; JNK, 530; PKA, 524.

Table 2. ACCS and average hit rates.[a]

Activity class	Method	Tree level	ACCS	Top 5	Top 10	Top 50	Top 100
CAB	1-NN	0	9.7	0.74	0.51	0.10	0.05
		≤1	33.9	0.89	0.60	0.12	0.06
		≤2	49.1	0.89	0.60	0.12	0.06
	3-NN	0		0.58	0.36	0.08	0.04
		≤1		0.69	0.40	0.10	0.05
		≤2		0.69	0.40	0.09	0.05
CAL	1-NN	0	25.7	0.94	0.60	0.13	0.07
		≤1	69.3	0.96	0.68	0.15	0.07
		≤2	93.9	0.96	0.68	0.14	0.07
	3-NN	0		0.24	0.17	0.06	0.03
		≤1		0.22	0.16	0.06	0.04
		≤2		0.26	0.16	0.07	0.04
CAS	1-NN	0	21.2	0.94	0.86	0.19	0.10
		≤1	50.1	0.94	0.79	0.17	0.09
		≤2	82.5	0.94	0.73	0.16	0.08
	3-NN	0		0.84	0.72	0.19	0.11
		≤1		0.80	0.69	0.19	0.11
		≤2		0.78	0.67	0.19	0.10
JNK	1-NN	0	25.6	0.84	0.68	0.15	0.08
		≤1	47.9	0.78	0.55	0.12	0.06
		≤2	59.6	0.71	0.49	0.11	0.06
	3-NN	0		0.38	0.22	0.09	0.05
		≤1		0.36	0.23	0.09	0.05
		≤2		0.31	0.22	0.08	0.05
PKA	1-NN	0	36.5	1.00	1.00	0.52	0.26
		≤1	99.4	1.00	1.00	0.42	0.21
		≤2	156.5	1.00	0.96	0.40	0.20
	3-NN	0		0.52	0.41	0.11	0.08
		≤1		0.40	0.29	0.10	0.10
		≤2		0.34	0.23	0.11	0.11

[a] For each screening data set, average numbers of ACCS in reference sets are reported at different tree levels. Also reported are average hit rates for similarity searching using ACCS fingerprints. For each of ten reference sets, independent search calculations were carried out. Hit rates were calculated for the top-ranked 5, 10, 50, and 100 screening set molecules on the basis of Tanimoto similarity.

istic substructures ranged from 9.7 (CAB) to 36.5 (PKA). With the exception of PKA, cumulative numbers of ACCS at tree level 2 were always smaller than 100. Figure 2 shows representative ACCS examples for CAB. As can be seen, these substructures are diverse and relatively small. Depending on the tree level, larger substructures are also found.

To investigate whether these substructures could be directly used to detect structure–activity relationships, we encoded ACCS for each reference set and tree level as keyed fingerprints, where each bit position detects the presence or absence of a specific fragment. These small compound class-directed ACCS fingerprints (ACCS-FPs) were then used to search each screening data set for the remaining active molecules (that is, total number of hits minus reference compounds). In these calculations, we applied nearest neighbor methods as a similarity search strategy for multiple reference compounds.[23]. These methods separately calculate the similarity of a database compound to each individual reference molecule.[23] Then either the largest similarity value is used, which is called the 1-NN strategy, or the similarity scores of *k* nearest neighbors are averaged (*k*-NN).[23] For all reference sets, 1-NN and 3-NN calculations were carried out on the basis of Tanimoto similarity[1] with ACCS-FPs, MACCS keys (166 bits), and three other fingerprints; TGD[24] (420 bits), TGT[25] (1704 bits), and Molprint2D.[26] TGD is an atom pair-type fingerprint recording pairs of seven different atom types over a maximum path length of 15 bonds. TGT is a three-point pharmacophore-type 2D fingerprint that captures triangles of four atomic features using graph distances divided into six distance ranges.

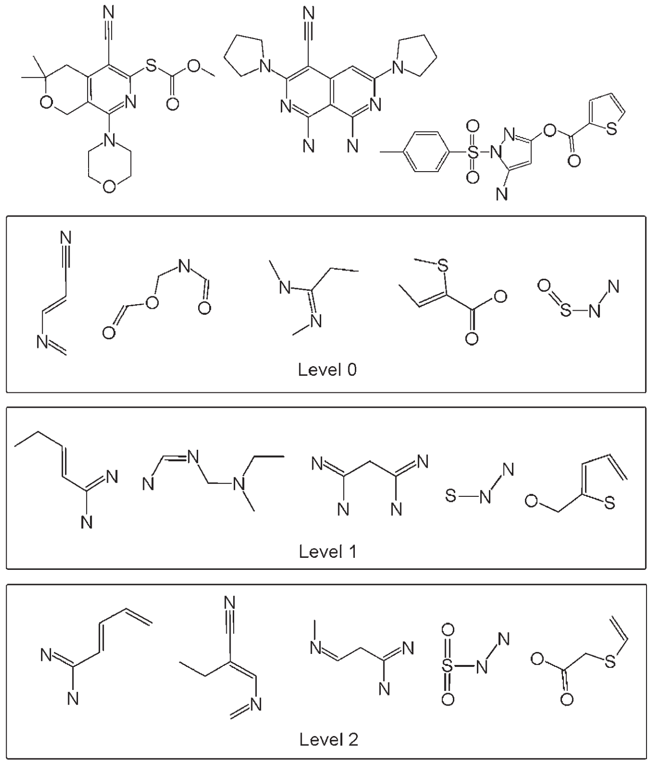

Figure 2. Representative ACCS. For activity class CAB, an exemplary ACCS subset is shown. These substructures occur in fragment populations of a CAB reference set that includes the three compounds shown at the top. ACCS are classified according to their levels in fragment trees.

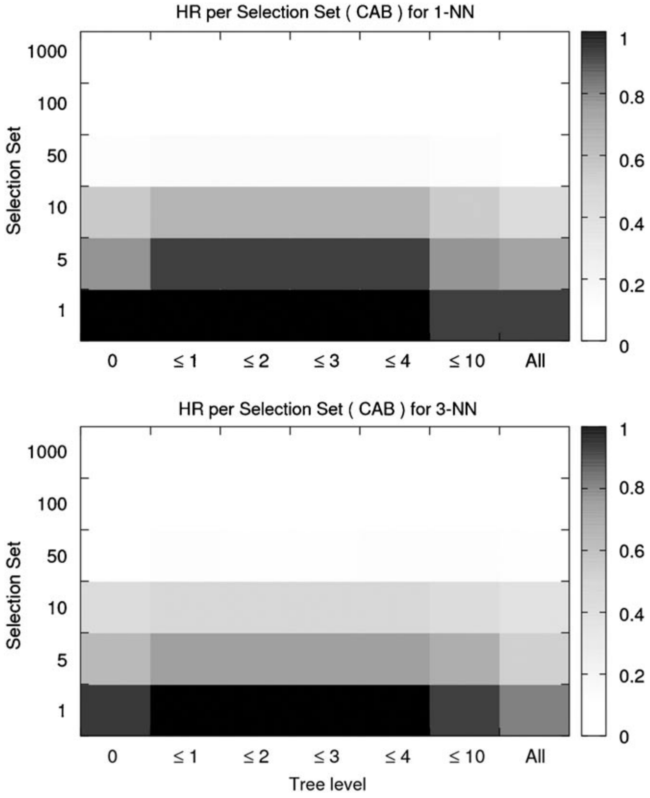

Figure 3. Hit rate map for activity class CAB. Average hit rates are reported for the 1-NN (top) and 3-NN (bottom) search strategies for substructure combinations at different tree levels. Hit rates are color-coded using a continuous spectrum from black (hit rate 1.0, that is, 100%) to white (hit rate 0.0).

Molprint2D generates layered atom environments and varying numbers of strings per molecule.

Table 2 reports average hit rates for ACCS-FPs for the first three tree levels and the different screening data sets. ACCS-FPs consistently retrieved active molecules and displayed a strong tendency to enrich hits in small selection sets of 5 or 10 compounds. The majority of selection sets of 50 compounds also contained 10%–20% active molecules. A graphical representation of search performance is provided in Figure 3, which shows hit rate maps for CAB as an example. These graphs monitor hit rates over all tree levels and reveal that ACCS-FPs containing fragments of the first few tree levels already displayed top search performance. Corresponding hit rate maps for the remaining screening data sets are shown in Supporting Information Figure 2. 1-NN calculations performed overall better than 3-NN, although differences were subtle in a number of cases. Even the smallest ACCS-FPs consisting only of substructures identified at tree level 0 produced hit rates comparable to those of larger versions of ACCS-FPs. Their search performance was not dominated by large fragments. At tree level 0, removal of fragments larger than 50% of the average size (number of atoms) of reference molecules typically reduced ACCS sets by less than 10% and did not notably change search performance. These observations emphasize the importance of ACCS combinations, rather than individual fragments.

ACCS at tree level 0 represent starting points of class-specific fragment pathways and thus occur independently of each other. The observation that the addition of dependent fragments at tree levels 1 and 2 did not increase search performance, indicates that combinations of ACCS at origins of fragment pathways capture much class-specific information, although they are typically small. Moreover, the addition of dependent fragments at increasing tree levels can produce substructure combinations that are not represented by individual reference molecules. The use of such combinations is likely to increase the probability of detecting other database compounds. For example, Figure 3 shows that hit rates for CAB decreased when substructures up to tree level 10 were added. Therefore, it is not required to consider all ACCS produced by a reference set. Rather, combinations of small subsets of ACCS from different class-specific pathways encode sufficient information.

Table 3 reports the results of corresponding similarity search calculations using four different fingerprints. For selection sets of up to 50 compounds, ACCS-FPs recovered on average more hits than MACCS and for sets of 100 compounds, search performance was overall comparable. Thus, variably composed small ACCS-FPs consisting of class-directed random fragments met or exceeded hit rates produced by MACCS that is based on a generally applicable and well-defined fragment dictionary. For small selection sets, ACCS-FPs also produced consistently higher hit rates than the TGD and TGT fingerprints. Compared to Molprint2D, ACCS-FPs performed notably better using the 1-NN search strategy. For 3-NN calculations, hit rates obtained with Molprint2D were a few percent higher in four of five cases. Our primary objective has been to investigate whether random fragment sets could be successfully used for similarity searching, which we have been able to demonstrate. However, the results in Table 2 and Table 3 show that ACCS-FP search performance compared favorably to other 2D fingerprints. At tree level 0, ACCS-FPs contain on average about 24 substructures and are thus even smaller in size than so-called mini-fingerprints (MFPs) that were introduced several years ago as hybrid fingerprints consisting of selected MACCS keys and property descriptors.[27] MFPs have a minimum number of about 60 bit positions and have thus far been the smallest 2D fingerprints.

We also determined unique scaffolds[5] for active compounds in screening data sets (Table 1) and hits identified by similarity searching. Table 4 reports the number of unique scaffolds identified in each similarity search trial. Table 1 shows that the ratio of hits and unique scaffolds ranges from 1.2 to 1.5 for the five screening sets, which further illustrates the diversity of active compounds studied here. Table 4 reveals that ACCS-FP calculations displayed a clear tendency to detect diverse scaffolds. Furthermore, comparison with results of reference calculations also reported in Table 4 shows that ACCS-FPs recognized in most cases at least as many distinct scaffolds as the other fingerprints, and often more. ACCS-FPs detected more scaffolds than Molprint2D in 1-NN

Table 3. Average hit rates for reference calculations.[a]

Fingerprint	Activity class	Method	Top 5	Top 10	Top 50	Top 100
MACCS	CAB	1-NN	0.42	0.36	0.13	0.07
		3-NN	0.52	0.39	0.11	0.05
	CAL	1-NN	0.22	0.19	0.08	0.04
		3-NN	0.08	0.05	0.02	0.01
	CAS	1-NN	0.36	0.37	0.17	0.10
		3-NN	0.54	0.52	0.19	0.10
	JNK	1-NN	0.48	0.41	0.12	0.07
		3-NN	0.50	0.36	0.11	0.07
	PKA	1-NN	0.08	0.13	0.08	0.05
		3-NN	0.12	0.12	0.05	0.04
TGD	CAB	1-NN	0.44	0.33	0.11	0.06
		3-NN	0.52	0.32	0.09	0.05
	CAL	1-NN	0.20	0.16	0.07	0.04
		3-NN	0.18	0.13	0.06	0.04
	CAS	1-NN	0.14	0.13	0.07	0.04
		3-NN	0.34	0.22	0.07	0.05
	JNK	1-NN	0.36	0.21	0.07	0.04
		3-NN	0.36	0.19	0.05	0.04
	PKA	1-NN	0.24	0.21	0.09	0.05
		3-NN	0.28	0.19	0.08	0.05
TGT	CAB	1-NN	0.68	0.38	0.08	0.04
		3-NN	0.66	0.37	0.08	0.04
	CAL	1-NN	0.30	0.18	0.06	0.03
		3-NN	0.30	0.17	0.06	0.03
	CAS	1-NN	0.32	0.24	0.08	0.05
		3-NN	0.38	0.28	0.08	0.04
	JNK	1-NN	0.56	0.35	0.08	0.04
		3-NN	0.56	0.36	0.08	0.04
	PKA	1-NN	0.08	0.16	0.05	0.03
		3-NN	0.10	0.11	0.04	0.02
Molprint2D	CAB	1-NN	0.50	0.47	0.14	0.08
		3-NN	0.76	0.46	0.13	0.07
	CAL	1-NN	0.34	0.30	0.09	0.06
		3-NN	0.30	0.20	0.09	0.05
	CAS	1-NN	0.44	0.49	0.21	0.11
		3-NN	0.74	0.57	0.20	0.11
	JNK	1-NN	0.30	0.28	0.17	0.10
		3-NN	0.34	0.24	0.16	0.10
	PKA	1-NN	0.38	0.29	0.18	0.15
		3-NN	0.48	0.40	0.25	0.18

[a] Reported are average hit rates for reference calculations using MACCS keys, TGD, TGT, and Molprint2D fingerprints, presented according to Table 2.

Table 4. Distinct scaffolds.[a]

Fingerprint	Activity class	Method	Tree level	Top 5	Top 10	Top 50	Top 100
ACCS-FPs	CAB	1-NN	0	2.0 (3.7)	2.8 (5.1)	3.0 (5.2)	3.0 (5.2)
			≤1	1.8 (4.4)	3.0 (6.0)	3.2 (6.1)	3.2 (6.1)
			≤2	1.8 (4.4)	3.0 (6.0)	3.2 (6.1)	3.2 (6.1)
		3-NN	0	1.5 (2.9)	1.6 (3.6)	1.9 (4.1)	2.2 (4.4)
			≤1	1.4 (3.4)	1.5 (4.0)	2.2 (5.0)	2.4 (5.2)
			≤2	1.4 (3.4)	1.6 (4.0)	1.9 (4.3)	2.1 (4.7)
	CAL	1-NN	0	3.7 (4.7)	5.1 (6.0)	5.6 (6.6)	5.6 (6.6)
			≤1	3.4 (4.8)	5.6 (6.8)	6.1 (7.3)	6.2 (7.4)
			≤2	3.7 (4.8)	5.5 (6.8)	5.9 (7.0)	5.9 (7.0)
		3-NN	0	1.0 (1.2)	1.2 (1.7)	2.4 (3.0)	2.4 (3.0)
			≤1	0.9 (1.1)	1.1 (1.6)	2.5 (3.1)	3.1 (3.8)
			≤2	1.0 (1.3)	1.1 (1.6)	2.8 (3.4)	3.3 (4.2)
	CAS	1-NN	0	2.4 (4.7)	6.3 (8.6)	7.7 (9.5)	7.7 (9.6)
			≤1	2.4 (4.7)	6.0 (7.9)	7.1 (8.7)	7.2 (8.8)
			≤2	2.4 (4.7)	5.7 (7.3)	6.5 (8.0)	6.8 (8.3)
		3-NN	0	2.3 (4.2)	5.1 (7.2)	7.4 (9.7)	8.3 (10.6)
			≤1	2.1 (4.0)	4.9 (6.7)	7.2 (9.5)	8.3 (10.6)
			≤ 2	2.2 (3.9)	4.6 (6.7)	7.3 (9.5)	8.0 (10.4)
	JNK	1-NN	0	2.6 (4.2)	4.9 (6.8)	5.4 (7.4)	5.5 (7.8)
			≤1	2.3 (3.9)	4.1 (5.5)	4.6 (6.0)	4.6 (6.0)
			≤2	3.0 (3.6)	3.0 (4.9)	3.0 (5.4)	3.0 (5.6)
		3-NN	0	1.1 (1.9)	1.1 (2.2)	2.7 (4.5)	3.3 (5.4)
			≤1	0.9 (1.8)	1.1 (2.3)	2.9 (4.5)	3.0 (4.7)
			≤2	1.0 (1.6)	1.0 (2.2)	2.0 (4.0)	3.0 (4.7)
	PKA	1-NN	0	3.9 (5.0)	8.1 (10.0)	19.0 (26.1)	19.0 (26.1)
			≤1	3.6 (5.0)	7.8 (10.0)	15.3 (20.9)	15.3 (20.9)
			≤2	3.6 (5.0)	7.7 (9.6)	14.7 (19.9)	14.7 (19.9)
		3-NN	0	2.1 (5.0)	3.3 (9.6)	4.4 (19.9)	6.7 (19.9)
			≤1	1.5 (2.6)	2.4 (4.1)	4.2 (5.5)	7.5 (8.3)
			≤2	1.2 (1.7)	2.4 (2.3)	4.4 (5.7)	8.0 (10.6)
MACCS	CAB	1-NN	N.A.	1.3 (2.1)	1.8 (3.6)	3.3 (6.5)	3.3 (6.8)
		3-NN		1.4 (2.6)	1.5 (3.9)	1.9 (5.3)	2.0 (5.4)
	CAL	1-NN		0.8 (1.1)	1.4 (1.9)	3.4 (3.9)	4.0 (4.5)
		3-NN		0.3 (0.4)	0.4 (0.5)	0.9 (1.0)	1.2 (1.3)
	CAS	1-NN		1.7 (1.8)	3.5 (3.7)	6.6 (8.7)	7.2 (9.7)
		3-NN		2.5 (2.7)	4.5 (5.2)	7.1 (9.5)	7.6 (10.1)
	JNK	1-NN		1.6 (2.4)	1.9 (4.1)	3.2 (5.8)	3.5 (7.0)
		3-NN		1.1 (2.5)	1.2 (3.6)	2.2 (5.5)	3.3 (7.2)
	PKA	1-NN		0.4 (1.3)	1.3 (2.4)	3.8 (3.9)	5.0 (5.2)
		3-NN		0.6 (0.6)	1.1 (1.2)	2.4 (2.6)	3.7 (3.9)

calculations but this trend was reversed for 3-NN searching, which parallels differences in hit rates, as discussed above. Taken together, the results show that class-specific combinations of random fragments encoded in ACCS-FPs have the potential to recognize structurally diverse compounds.

In summary, we have been able to demonstrate that specific combinations of substructures can be extracted from random fragment populations and successfully used for similarity searching. These findings extend currently available approaches to the design of structure-based descriptors and similarity search tools. ACCS-FPs are introduced as prototypic fingerprint representations of substructure combinations derived from compound class-specific fragment pathways. Combinations of only approximately 20 ACCS successfully detect different structure–activity relationships. In addition to their small size, characteristic features of ACCS-FPs include that they are compound class-directed and highly variable in composition. Thus, we conclude that random fragment populations are a valuable source for the identification of substructure combinations that are signatures of different compound classes. Such substructure combinations provide a basis for the development of class-specific 2D similarity search tools.

Keywords: chemoinformatics · fingerprints · random fragment populations · similarity searching · structural descriptors

[1] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983 – 996.

[2] J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233 – 245.

[3] C. Merlot, D. Domine, C. Cleva, D. J. Church, *Drug Discovery Today* **2003**, *8*, 594 – 602.

Table 4. (Continued)

Fingerprint	Activity class	Method	Tree level	Top 5	Top 10	Top 50	Top 100
TGD	CAB	1-NN	N.A.	2.0 (2.2)	2.1 (3.3)	3.1 (5.3)	3.1 (5.5)
		3-NN		2.1 (2.6)	2.1 (3.2)	2.7 (4.6)	3.0 (5.1)
	CAL	1-NN		0.7 (1.0)	1.0 (1.6)	2.7 (3.3)	3.5 (4.1)
		3-NN		0.7 (0.9)	0.8 (1.3)	2.2 (2.8)	3.6 (4.2)
	CAS	1-NN		0.6 (0.7)	0.9 (1.3)	2.7 (3.7)	3.0 (4.2)
		3-NN		1.1 (1.7)	1.6 (2.2)	2.4 (3.6)	3.4 (4.7)
	JNK	1-NN		1.8 (1.8)	2.1 (2.1)	3.1 (3.7)	3.4 (4.2)
		3-NN		1.7 (1.8)	1.8 (1.9)	2.4 (2.5)	3.0 (3.5)
	PKA	1-NN		1.2 (1.2)	2.1 (2.1)	3.8 (4.3)	4.5 (5.2)
		3-NN		1.4 (1.4)	1.9 (1.9)	3.3 (3.8)	4.1 (4.7)
TGT	CAB	1-NN	N.A.	2.2 (3.4)	2.2 (3.8)	2.2 (3.8)	2.3 (3.9)
		3-NN		2.2 (3.3)	2.2 (3.7)	2.2 (3.8)	2.2 (3.8)
	CAL	1-NN		0.9 (1.5)	1.2 (1.8)	2.5 (3.1)	2.9 (3.5)
		3-NN		0.9 (1.5)	1.1 (1.7)	2.3 (2.9)	2.5 (3.1)
	CAS	1-NN		1.1 (1.6)	1.4 (2.4)	2.9 (4.2)	3.5 (4.9)
		3-NN		1.2 (1.9)	1.5 (2.8)	2.5 (3.9)	3.0 (4.4)
	JNK	1-NN		2.6 (2.8)	2.7 (3.5)	2.8 (4.1)	2.8 (4.2)
		3-NN		2.6 (2.8)	2.7 (3.6)	2.8 (4.0)	2.8 (4.0)
	PKA	1-NN		0.4 (0.4)	1.6 (1.6)	2.4 (2.4)	3.0 (3.0)
		3-NN		0.5 (0.5)	1.1 (1.1)	2.1 (2.1)	2.2 (2.2)
Molprint2D	CAB	1-NN	N.A.	1.9 (2.5)	2.3 (4.7)	4.0 (7.1)	4.7 (8.2)
		3-NN		1.5 (3.8)	1.7 (4.6)	3.1 (6.4)	3.3 (6.6)
	CAL	1-NN		1.3 (1.7)	2.4 (3.0)	3.8 (4.4)	5.5 (6.1)
		3-NN		1.3 (1.5)	1.8 (2.0)	4.1 (4.7)	4.5 (5.3)
	CAS	1-NN		2.1 (2.2)	3.8 (4.9)	8.3 (10.3)	8.8 (11.1)
		3-NN		2.8 (3.7)	4.5 (5.7)	7.8 (9.9)	8.4 (10.7)
	JNK	1-NN		1.5 (1.5)	2.7 (2.8)	5.0 (8.4)	5.6 (9.6)
		3-NN		1.4 (1.7)	1.9 (2.4)	4.6 (8.0)	5.8 (9.7)
	PKA	1-NN		1.9 (1.9)	2.9 (2.9)	7.1 (9.1)	10.6 (15.5)
		3-NN		2.4 (2.4)	3.6 (4.0)	9.1 (12.5)	12.5 (18.3)

[a] For all similarity search trials according to Tables 2 and 3, the average number of distinct scaffolds representing the hits is reported. For each selection set, the average number of correctly identified hits is given in parentheses. Tree levels only apply to ACCS-FP calculations.

[4] J. Bajorath, *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.

[5] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.

[6] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch, H. Waldmann, *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

[7] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

[8] G. Schneider, M. L. Lee, M. Stahl, P. Schneider, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487–494.

[9] M. J. McGregor, P. V. Pallai, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.

[10] MACCS structural keys, MDL Elsevier, San Leandro, CA, USA, **2005** (http://www.mdl.com.

[11] J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.

[12] BCI, Digital Chemistry Ltd, Leeds, United Kingdom, **2006** (http://www.digitalchemistry.co.uk).

[13] G. W. Adamson, S. E. Creasey, M. F. Lynch, *J. Chem. Doc.* **1973**, *13*, 158–162.

[14] A. Feldman, L. Hodes, *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 147–152.

[15] G. W. Adamson, D. R. Lambourne, M. F. Lynch, *J. Chem. Soc.* **1972**, *C*, 2428–2433.

[16] G. W. Adamson, J. Cowell, M. F. Lynch, A. W. H. McLure, W. G. Town, A. M. Yapp, *J. Chem. Doc.* **1973**, *13*, 153–157.

[17] P. Willett, *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 147–152.

[18] J. Batista, J. W. Godden, J. Bajorath, *J. Chem. Inf. Model.* **2006**, *46*, 1937–1944.

[19] J. Batista, J. Bajorath, *J. Chem. Inf. Model.* **2007**, *47*, 59–68.

[20] J. Batista, J. Bajorath, *J. Chem. Inf. Model.* **2007**, *47*, 1405–1413.

[21] PubChem. (http://pubchem.ncbi.nlm.nih.gov).

[22] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

[23] J. Hert, P. Willet, D. J. Wilton, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

[24] TGD, implemented in the Molecular Operating Environment (MOE), Chemical Computing Group Inc., Montreal, Quebec, Canada, **2005** (http://www.chemcomp.com).

[25] TGT, implemented in the Molecular Operating Environment (MOE), Chemical Computing Group Inc., Montreal, Quebec, Canada, **2005** (http://www.chemcomp.com).

[26] A. Bender, Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.

[27] L. Xue, J. W. Godden, J. Bajorath, *SAR QSAR Environ. Res.* **2003**, *14*, 27–40.

Received: August 3, 2007
Revised: September 5, 2007
Published online on October 19, 2007