Relevance of Feature Combinations for Similarity Searching Using General or Activity Class-Directed Molecular Fingerprints

Eugen Lounkine, Ye Hu, José Batista, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany

Received October 13, 2008

We introduce a methodology for the systematic identification of feature combinations derived from fingerprints of bioactive compounds. Structural features were organized in co-occurrence networks from which reference set-based feature cliques were extracted. A similarity search strategy is presented that is based on frequency ranking of cliques. Three types of fingerprints have been compared that are either general in their design or incorporate, to a different degree, compound class information. Taking control calculations into account, search performance was overall best for a molecule-specific extended connectivity fingerprint. For compound class-directed fingerprints, reference set-derived feature cliques occurring with low frequency in a screening database were found to consistently enrich active compounds in database selection sets, even if the features are not highly conserved among reference compounds. Thus, in contrast to general fingerprints, feature combinations play a crucial role in similarity searching using activity-class-directed fingerprints.

INTRODUCTION

Molecular fingerprints are widely used representations of compounds in chemoinformatics that report the presence of chemical features in molecules.[1] Different types of fragment-based fingerprints exist. Keyed fingerprints such as the MACCS keys[2] or the BCI fingerprint[3] encode predefined feature dictionaries and report the presence or absence of individual features in test compounds. By contrast, circular fingerprints such the ECFP fingerprints that are incorporated in Scitegic's Pipeline Pilot[4] derive features in a molecule-oriented manner by sampling atom environment properties in circular layers around individual atoms. Furthermore, ACCS-FPs were introduced as particularly short and activity-class-directed fingerprints that utilize activity-class characteristic substructures extracted from random fragment populations of active compounds.[5] Distribution analysis of activity-class characteristic substructures (ACCSs) among database compounds revealed that ACCS combinations that occur in compounds with defined biological activities are often only sparsely distributed in large databases.[6] Moreover, a recent study has revealed that pairs and triplets of hierarchical fragments extracted from bioactive compounds have significant potential to discriminate between ligands of closely related targets.[7]

In keyed fingerprints, different bit positions account for individual features and fingerprint similarity is typically quantified by calculating fingerprint overlap through the use of similarity coefficients. For this purpose, features are counted independently and there is no need to determine whether or not features occur in combinations. Consequently, the potential role of feature combinations for fingerprint search performance has thus far only been little explored. Fingerprint bits that are strongly conserved in active compounds have been identified[8] as well as characteristic bit patterns[8,9] and consensus fingerprints have been derived for compound classes.[10] However, whether or not feature combinations might influence fingerprint search performance has not yet been investigated.

Therefore, we have evaluated the potential role of defined feature combinations for similarity searching using structural fingerprints. To investigate different fingerprint designs, we have analyzed three fingerprints that do or do not take activity class-related information into account: MACCS keys, which have been selected as a prototype of a general structural fragment-based fingerprint; ECFP_4, which is a more abstract, molecule-directed fingerprint; and ACCS-FP, which is a fingerprint type for which fragments are selected in a compound class-dependent manner. To determine whether fingerprint features are set in combination and assess the potential impact of such combinations on similarity searching, we have developed a generally applicable methodology for the automatic identification of feature combinations of variable size that are preferentially found in active reference compounds. For this purpose, we introduce feature co-occurrence networks (FCoNs) that are calculated based on conditional probabilities of feature pair occurrence in active compounds. A clique detection algorithm is then applied to these networks to extract feature combinations that are prevalent in different activity classes. The frequency of these feature combinations in a large screening database (containing 500 000 ZINC[11] compounds) has been determined, thus enabling compound selection on the basis of selective feature cliques.

Similarity search performance utilizing FCoN feature cliques was compared to standard search calculations using the Tanimoto coefficient (Tc),[1] in combination with centroid and nearest-neighbor search strategies.[9] Feature cliques that rarely occurred in the screening database were capable of substantially enriching active compounds in selection sets

* Author to whom correspondence should be addressed. Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

10.1021/ci800377n CCC: $40.75
Published on Web 02/12/2009

Table 1. Description of Activity Classes

activity class	biological activity	number of compounds	number of ACCSs	avg. similarity	SD
AA2	adrenergic alpha 2 antagonists	35	25.3	0.39	0.15
BK2	bradykinin BK2 antagonists	22	31.0	0.55	0.11
CAL	calpain inhibitors	28	23.7	0.48	0.14
DD1	dopamine D1 agonists	30	55.9	0.56	0.14
F7I	factor VIIa inhibitors	23	16.1	0.46	0.11
GLG	glucagon receptor antagonists	33	34.1	0.44	0.13
GLY	glycoprotein IIb−IIIa antagonist	34	34.0	0.57	0.12
KRA	kainate receptor antagonists	22	15.7	0.55	0.17
LAC	lactamase (beta) inhibitors	29	33.2	0.44	0.18
SQE	squalene epoxidase inhibitors	25	7.7	0.40	0.16
SQS	squalene synthetase inhibitors	42	65.4	0.50	0.17
THI	thiol protease inhibitors	34	42.8	0.49	0.09
ULD	LDL upregulator	21	16.6	0.43	0.16
XAN	xanthine oxidase inhibitors	35	33.8	0.56	0.18

The number of compounds in each activity class is provided, together with the average number of ACCSs generated per reference set. The "avg. similarity" data column in Table 1 reports the average intraclass pair-wise similarity of active compounds, based on MACCS Tc calculations, and the "SD" column gives the corresponding standard deviations.

and defined feature combinations were shown to play a crucial role for the search performance of activity class-directed fingerprints.

METHODS

Datasets. Fourteen activity classes were assembled from the MDDR.[12] Table 1 summarizes the biological activities and composition of these classes. The structural homogeneity of activity classes was assessed by pairwise Tc calculations, using the MACCS fingerprint. These 14 classes consisted of 6 relatively homogeneous ones (average MACC Tc $\geq$ 0.5) and 8 more heterogeneous ones (average MACCS Tc 0.4−0.5). A randomly selected ZINC[11] subset containing 500 000 compounds was used as a database for virtual screening trials. For these calculations, each activity class was randomly divided into 10 reference and test sets of equal size.

Molecular Fingerprints. The popular and widely used MACCS structural keys represent a general fingerprint that accounts for 166 predefined structural features. No reference set-specific information is taken into account in search calculations using MACCS. As a control, we have also investigated a fingerprint of equivalent keyed design, but much larger key numbers. Therefore, a fingerprint was assembled from 733 Pubchem structural keys[13] with non-ambiguous SMARTS[14] representations (Pubchem-FP). These keys are similar in their design to MACCS keys. Different from keyed fingerprints, the ECFP_4 fingerprint is a circular fingerprint that derives features independently from individual test compounds. For each atom of a molecule, neighbor atoms are sampled that are separated from the central atom by up to four bonds. An integer representation is calculated from their atom and bond types using a hash function. This type of fingerprint does not rely on a predefined library of substructures, but generates features in a compound-oriented manner. If an ECFP_4 feature is present in a molecule, the corresponding substructure consisting of a central atom and the additional atoms separated by up to four bonds can be extracted. ACCS-FP have recently been introduced as short, compound class-directed fingerprints encoding activity class characteristic substructures (ACCS) extracted from random fragment populations of compound reference sets.[5] In ACCS-FP, structural information from the entire reference set is utilized in the selection of features. For the current study, ACCS-FP have been generated for each reference set independently, as described in detail in the Supporting Information. Hashed fingerprints such as the Daylight fingerprint[14] were not used in this study, because individual structural features and combinations cannot be directly extracted from them.

Feature Co-occurrence Networks and Clique Detection. To identify structural features that predominantly occur in combination in a compound reference set, we calculate FCoNs based on molecular fingerprints. Individual features encoded by a fingerprint are connected in these networks if they preferably occur in combination in active molecules. To build a FCoN, the frequency of occurrence is calculated first for each feature, by dividing the number of compounds exhibiting the feature by the total number of compounds in the reference set. Pairs of features are assigned a score based on the conditional probabilities of feature pair occurrence:

$$\text{score}(A, B) = \min\left(\frac{f(AB)}{f(A)}, \frac{f(AB)}{f(B)}\right)$$

Here, $f(A)$ denotes the frequency of feature A, $f(B)$ the frequency of feature B, and $f(AB)$ the frequency of the pair AB. Feature pairs that exclusively occur in combination achieve the maximal score of 1, whereas features that never occur in combination in any compound are assigned the minimal score of 0. Scores serve as weights for edges that connect individual features in the FCoN. To identify combinations of frequently co-occurring features, a co-occurrence threshold (ν) is applied and only edges are retained that have a weight equal to or greater than ν. The largest cliques are completely connected subgraphs that are not contained in any other completely connected subgraph. These cliques are identified in the pruned network and corresponding feature combinations reported. Figure 1 shows parts of an exemplary network including detected cliques. For clique detection, the Bron−Kerbosch algorithm[15] was implemented in the Scientific Vector Language (SVL).[16] The clique detection procedure is illustrated in Figure 2. The co-occurrence threshold was systematically varied from ν = 0.5 to 1, in increments of 0.1. Increasing threshold values yield cliques that are highly conserved among reference compounds. In addition to cliques identified for each co-occurrence threshold, we have also pooled all cliques that were unique to an activity class, irrespective of the threshold value. Clique numbers and sizes were determined for each co-occurrence threshold and the pooled sets. In addition, the distribution of cliques among screening database compounds was analyzed. Statistical analyses were performed using in-house generated Perl scripts.

Feature Clique Search Strategy. For search calculations using cliques (rather than fingerprints), a strategy has been designed that takes the degree of compound class-specificity of individual cliques into account. For each clique, the number of database molecules that contain all features is determined. Cliques are ranked in ascending order of their database frequency (i.e., cliques occurring in only small numbers of database compounds are prioritized). Database compounds are selected sequentially, according to the ranked

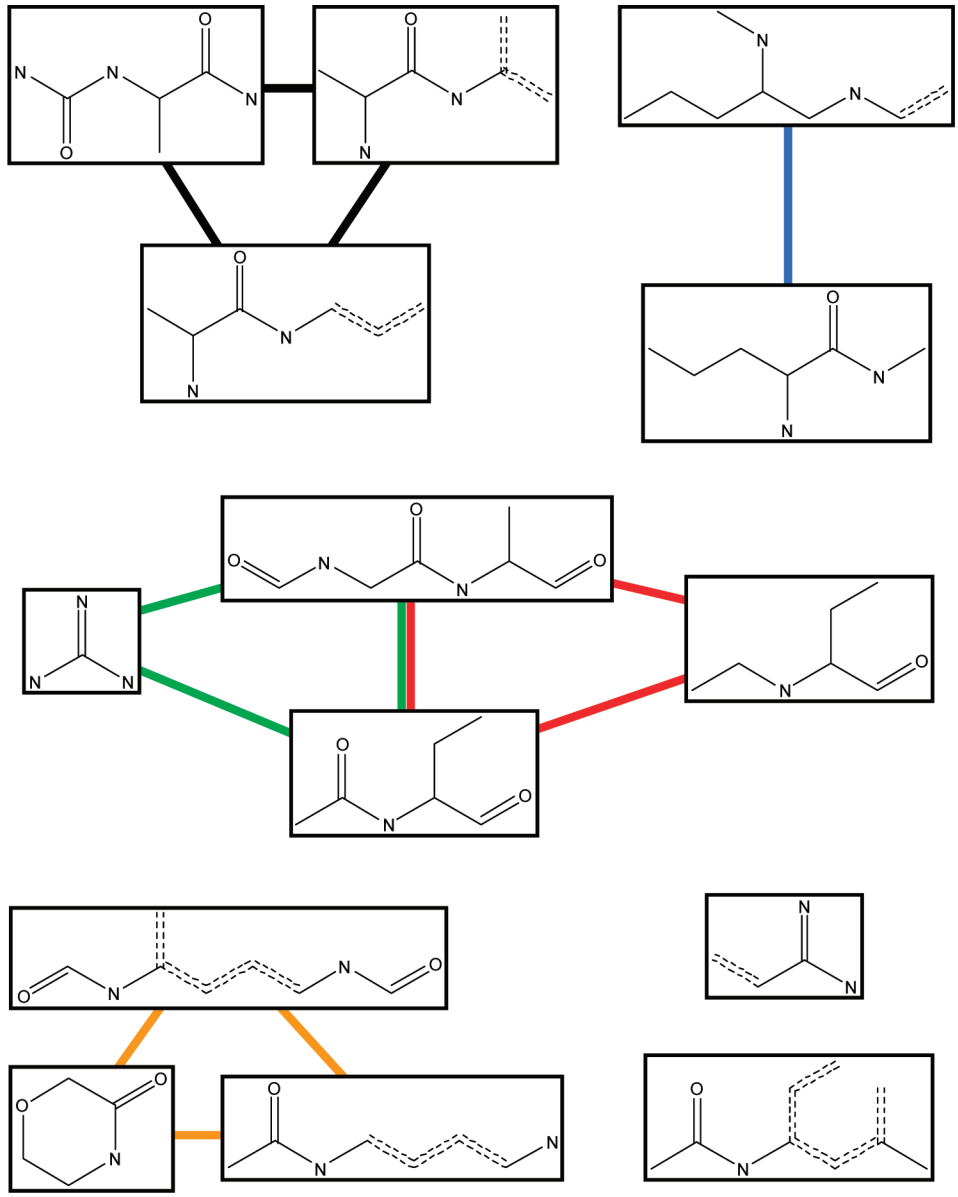

Figure 1. Feature co-occurrence networks. Part of an exemplary feature co-occurrence network (FCoN) is shown for ACCS-FP and activity class F7I (with a co-occurrence threshold of $\nu = 0.6$). Colored lines represent five cliques extracted from the FCoN. Dashed lines in the fragment structures indicate aromatic bonds.

clique list. Thus, each feature clique adds compounds to the selection set that have not been retrieved in previous steps by cliques with lower overall database frequency, thereby producing a compound ranking. This selection procedure is illustrated in Figure 3.

The clique detection strategy extends the concept of conserved bit positions utilized in the centroid approach, fingerprint scaling, and modal fingerprints. Calculating relative frequencies of individual features does not provide information about individual feature combinations of varying size that are present in reference set compounds. By contrast, the FCoN strategy explicitly utilizes conditional probabilities of feature co-occurrence, rather than relative frequencies. Thus, it provides an ensemble of feature cliques that are independent of each other and of varying size. Moreover, individual features may occur in different cliques. However, a clique is only matched to a database compound if all of its features match. Hence cliques behave as features that are defined in an activity class-directed manner.

Feature clique searching is computationally more complex than standard fingerprint calculations, because cliques must be identified and mapped. Thus, clique searching has the complexity $O(n^2)$, whereas the centroid approach is realized in linear time $O(n)$, with n being the number of reference compounds. In addition, clique searching has $O(m^2)$ complexity, with regard to the total number of unique features m. However, feature combinations are only extracted from small numbers of active reference molecules, the number of unique features in a small compound set is limited, and mapping to database compounds utilizes their fingerprint representations. Thus, additional computational costs are low and practical clique search requirements are comparable to standard fingerprinting.

Virtual Screening Trials. FCoN selection sets were transformed to ranked compound lists with equal scores assigned to compounds belonging to identical FCoN selection sets. Such noncontiguous score distributions also occur for ranking methods when compounds are assigned the same

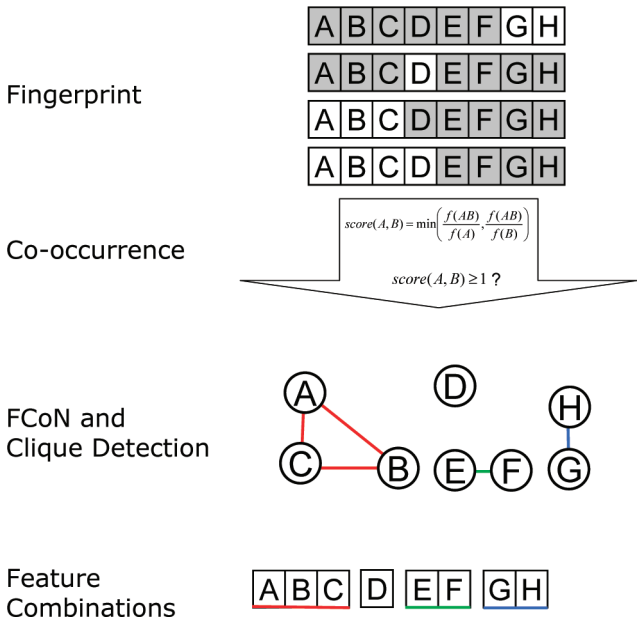

Figure 2. Feature clique detection. FCoN generation is illustrated for a co-occurrence threshold of $\nu = 1$. For a fingerprint calculated for molecules of an activity class, the relative frequency of each feature (bit position) is determined. Shaded boxes correspond to fingerprint bit positions that are set on. The corresponding co-occurrence network (FCoN) is calculated by connecting features that are found in combination in a predefined fraction of compounds containing each individual feature. Cliques are detected and feature combinations are identified. Although features A, B, and C are only present in two molecules, they form a clique based on a threshold value of one, because neither feature occurs without the other two in any compound. Feature D is not part of a clique and therefore is not considered in the search calculations.

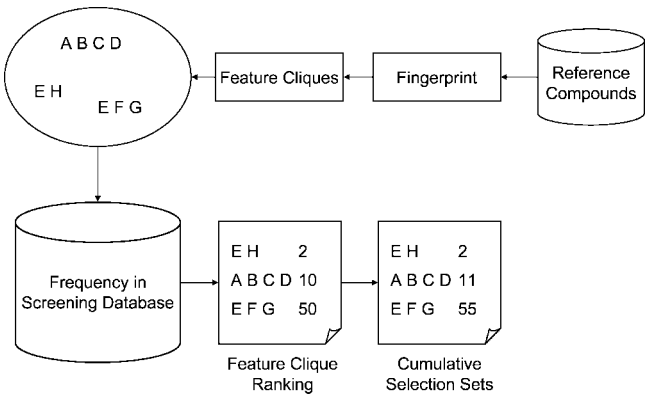

Figure 3. Feature clique search strategy. From fingerprints of active compounds, feature cliques are identified. For each clique, the number of database compounds that contain all of its features is determined. Cliques are sorted by ascending database frequencies (i.e., rarely occurring cliques are preferred). From this ranking, cumulative compound selection sets are generated. In this example, cliques EH and ABCD share one compound, resulting in a selection set size of 11 (2 + 10 − 1) compounds for clique ABCD.

similarity value. This binning effect is taken into account by calculating the expected number of retrieved compounds for each selection set. For example, given two selection sets with 90 and 110 compounds and 5 and 7 active compounds, respectively, suppose one wants to calculate the number of retrieved compounds for the selection set size of 100. We retrieve 5 active compounds contained in the set of 90 compounds. The remaining 10 (= 100 − 90) compounds are randomly selected from 20 (= 110 − 90) compounds that are additionally present in the set of 110 compounds. The expected number of additional actives provided by the set of 110 compounds in this randomly selected set is given by (7 − 5) × 10/20 = 1. Thus, for a selection set of 100 compounds, the method is expected to retrieve 6 compounds.

As a measure of search performance, recovery rates for selection sets of 100 top-ranked compounds were calculated for both Tc- and clique frequency-based compound rankings and averaged over 10 independent trials. For fingerprint similarity searching using multiple reference compounds, two standard *k*-nearest-neighbor approaches (1NN and 3NN),[17] the centroid (fingerprint averaging) technique,[17] and modal fingerprints[10] were applied. In 1NN calculations, the similarity value of the most similar reference compound is assigned to a database molecule as the final similarity value and in 3NN calculations, similarity values are averaged over the three most similar reference compounds to yield the final score. In centroid calculations, fingerprint settings of all reference compounds are averaged to generate a single fingerprint for similarity searching that takes contributions from all reference compounds into account. In modal fingerprints, a threshold is applied to the centroid fingerprint and bits are set on that meet or exceed the threshold. We have applied three thresholds: 50%, 70%, and 100%.

RESULTS AND DISCUSSION

ACCS-FP Generation. Activity-class characteristic substructures (ACCSs) have been generated for all activity classes. Following a previously reported approach, ACCSs were defined as randomly generated fragments that occurred in at least two compounds of a given activity class but in none of 500 randomly selected ZINC compounds.[5,6] A small decorrelated subset of ACCSs is selected based on the evaluation of dependency relationships of fragment co-occurrence. As reported in Table 1, the average number of ACCSs for different classes ranged from ∼8 (SQE) to 65 (SQS). There was no obvious correlation between the number of ACCSs and activity class size or structural heterogeneity. For each class and compound reference set, individual ACCS-FPs were generated by combining ACCS in a keyed bit string format.

Feature Co-occurrence Networks and Clique Distribution. To identify feature combinations that are representative of each reference set, FCoNs were generated for features of three different fingerprints (MACCS, ECFP_4, and ACCS-FP) and utilized to systematically identify maximal feature cliques. Table 2 reports the medians of feature clique numbers that were first calculated for individual compound reference sets for all activity classes and then pooled, and Table 3 provides the clique distribution data for individual co-occurrence thresholds. Figure S1 in the Supporting Information illustrates the distribution of cliques. ACCS-FP produced the smallest number of feature cliques (4−83) and the median number of cliques correlated with the ACCS-FP length, yielding a Pearson correlation coefficient of $R^2 = 0.89$. MACCS and ECFP_4 produced, on average, ∼200 and 300 cliques, respectively; these are, hence, significantly more than ACCS-FP. As also shown in Table 2, the median number of features per clique was comparable for MACCS and ECFP_4 (5−13 features), whereas ACCS-FP cliques were smaller (2−6 features). However, for all fingerprints, large individual cliques were also

Table 2. Clique Statistics for Three Different Fingerprints (MACCS, ECFP_4, and ACCS-FP)[a]

	Number of Cliques			Clique Size			Database Distribution		
activity class	MACCS	ECFP_4	ACCS-FP	MACCS	ECFP_4	ACCS-FP	MACCS	ECFP_4	ACCS-FP
AA2	156	136	21	5	7	3	23193	62	101
BK2	184	715	20	11	9	5	3273	20	96
CAL	133	201	15	5	7	4	22725	636	282
DD1	109	229	46	5	8	6	17246	12	46
F7I	248	346	13	8	6	3	4886	775	77
GLG	312	325	28	8	7	3	9131	1410	81
GLY	138	350	24	6	6	3	16744	100	51
KRA	141	154	8	8	11	5	2272	7	15
LAC	184	275	32	8	12	4	2713	16	51
SQE	224	296	4	7	6	2	16490	7	29
SQS	239	499	83	10	13	6	3285	4	13
THI	243	238	47	5	5	3	17646	3173	151
ULD	245	287	13	8	8	4	10153	747	80
XAN	123	182	23	5	10	3	21485	7	6

[a] Reported are the median values for the number of cliques from pooled sets, their size, and database distribution (i.e., the number of database molecules containing each clique). For the calculation of medians, ten random reference sets were combined.

Table 3. Clique Numbers for Three Different Fingerprints (MACCS, ECFP_4, and ACCS-FP)[a]

activity	Clique Number Distribution (Median Values)						
class	$\nu = 0.5$	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$	$\nu = 1$	pooled
				MACCS			
AA2	62	46	36	29	23	20	156
BK2	70	52	45	33	19	18	184
CAL	52	44	33	25	19	20	133
DD1	40	29	25	23	18	16	109
F7I	103	66	50	40	25	22	248
GLG	129	91	58	38	24	19	312
GLY	49	36	30	22	17	13	138
KRA	52	36	28	25	18	19	141
LAC	66	46	32	30	22	18	184
SQE	89	55	44	34	22	20	224
SQS	64	58	58	46	26	12	239
THI	120	75	43	27	15	12	243
ULD	93	65	46	34	20	19	245
XAN	38	33	35	24	21	18	123
				ECFP_4			
AA2	57	46	40	36	35	34	136
BK2	182	199	158	88	62	56	715
CAL	58	64	63	55	55	47	201
DD1	81	75	59	46	41	40	229
F7I	178	95	61	49	41	40	346
GLG	144	108	70	65	55	50	325
GLY	158	112	80	75	63	58	350
KRA	57	54	44	36	29	29	154
LAC	105	81	60	46	38	36	275
SQE	143	88	72	77	59	59	296
SQS	132	165	132	88	54	52	499
THI	94	83	66	55	45	43	238
ULD	136	89	60	50	38	38	287
XAN	59	64	60	47	43	41	182
				ACCS-FP			
AA2	9	8	6	7	5	5	21
BK2	7	8	5	6	7	7	20
CAL	6	6	5	6	6	6	15
DD1	12	15	12	10	9	10	46
F7I	7	5	3	3	3	3	13
GLG	14	12	10	10	10	10	28
GLY	14	9	7	7	6	6	24
KRA	3	4	3	4	4	4	8
LAC	11	11	10	10	8	8	32
SQE	3	1	2	2	1	1	4
SQS	15	19	20	20	17	17	83
THI	20	20	13	11	9	9	47
ULD	5	4	4	4	4	4	13
XAN	8	8	7	8	9	9	23

[a] Medians of clique numbers are reported for different co-occurrence threshold (ν) values and pooled clique sets.

Table 4. Clique Size for Three Different Fingerprints (MACCS, ECFP_4, and ACCS-FP)[a]

activity	Clique Size Distribution (Median Values)						
class	$\nu = 0.5$	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$	$\nu = 1$	pooled
				MACCS			
AA2	7	4	3	3	2	2	5
BK2	20	10	8	4	2	2	11
CAL	8	5	3	3	3	3	5
DD1	7	5	3	3	3	3	5
F7I	15	8	6	3	2	2	8
GLG	13	9	5	3	2	2	8
GLY	7	8	5	3	3	2	6
KRA	12	7	6	4	3	2	8
LAC	14	9	6	4	3	3	8
SQE	11	6	4	3	3	2	7
SQS	20	11	8	5	3	3	10
THI	9	5	3	2	2	2	5
ULD	15	8	5	3	3	3	8
XAN	6	5	4	3	3	2	5
				ECFP_4			
AA2	9	7	6	6	6	6	7
BK2	14	9	7	5	3	3	9
CAL	13	8	7	4	4	4	7
DD1	9	8	7	5	6	6	8
F7I	7	5	3	3	3	3	6
GLG	9	6	5	4	4	5	7
GLY	9	5	5	5	4	3	6
KRA	14	10	9	9	9	9	11
LAC	15	16	6	5	4	4	12
SQE	9	5	4	4	3	3	6
SQS	12	7	12	17	5	6	13
THI	7	5	4	4	4	5	5
ULD	11	7	5	4	6	6	8
XAN	13	10	7	6	5	5	10
				ACCS-FP			
AA2	3	3	2	2	2	2	3
BK2	4	4	4	4	3	3	5
CAL	4	4	2	2	2	2	4
DD1	9	7	5	4	3	3	6
F7I	3	3	2	2	2	2	3
GLG	5	3	2	2	2	2	3
GLY	3	3	3	2	2	2	3
KRA	5	4	3	3	3	3	5
LAC	6	4	3	3	2	2	4
SQE	2	3	2	2	2	2	2
SQS	8	9	7	5	2	2	6
THI	5	3	3	2	2	2	3
ULD	5	5	4	3	3	3	4
XAN	4	3	2	2	2	2	3

[a] Medians of clique sizes (numbers of features in cliques) are reported for different co-occurrence threshold (ν) values and the pooled clique sets.

detected (in some cases, containing >40 features), as reported in Table 4 and illustrated in Figure S2 in the Supporting Information.

Furthermore, the distribution of identified cliques in 500 000 ZINC database compounds was analyzed. As shown in Table 2 and Figures 4A−C, the three fingerprints

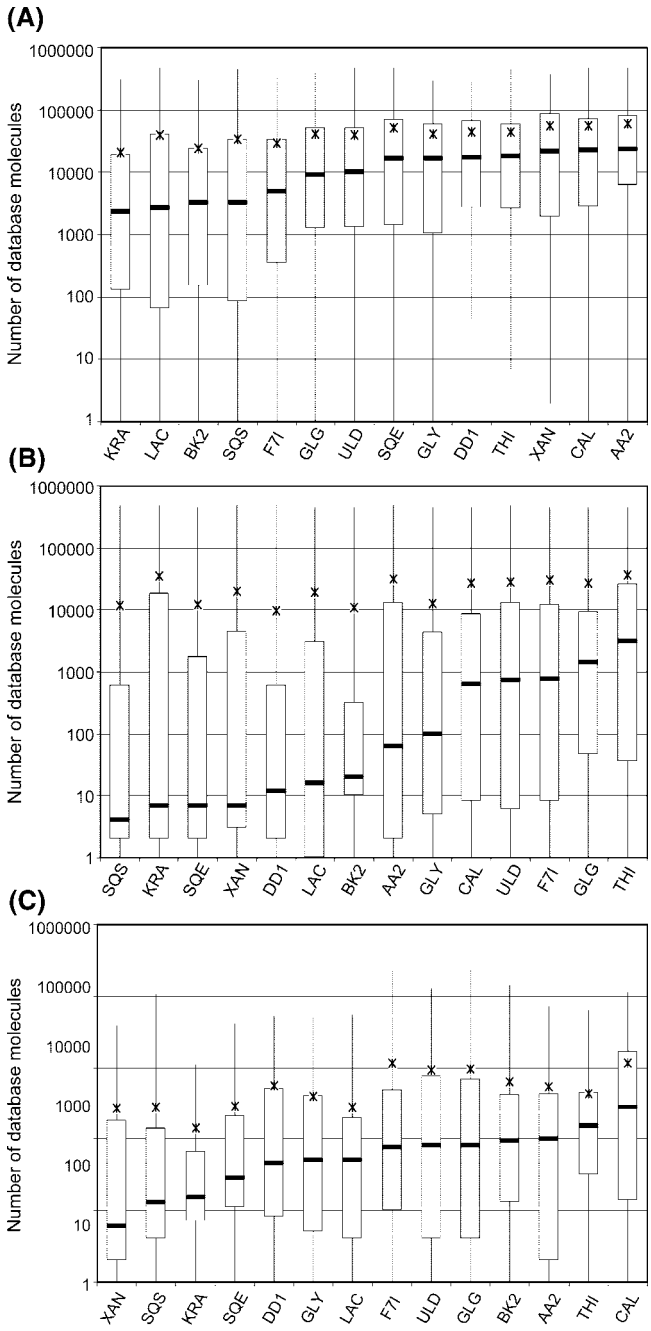

Figure 4. Feature clique distribution. The distribution of cliques from activity classes ((A) MACCS, (B) ECFP_4, and (C) ACCS-FP) among screening database compounds is shown in a box plot representation calculated from 10 independent reference sets. Thick black lines mark the median and asterisks the mean of the distribution.

significantly differed in the average number of database compounds that contained their feature cliques. However, for all three fingerprints, feature combinations were also identified that matched only a few molecules or a single database molecule. MACCS keys produced cliques that were typically found in large numbers (~2000−23000) of database compounds. Thus, the most general of the three fingerprints that we compared produced feature combinations that were least specific for active compounds, as one would expect. Cliques extracted from ECFP_4 occurred, on average, in considerably smaller numbers of database compounds (~5−3000). However, for both MACCS and ECFP_4, individual cliques were also detected that occurred in almost all database molecules. By contrast, ACCS-FP cliques matched only ~5−300 database molecules and the most generic ones were found in ~25 000 compounds (i.e., 5% of the database). Thus, compound class-directed ACCS-FP produced the most specific feature combinations. Note that specific cliques could also be incorporated as additional bits in keyed fingerprints. However, this would increase the redundancy of structural information encoded in fingerprints.

MACCS structural keys are frequently correlated.[8] For example, a number of individual keys account for combinations of others and, consequently, their bits are typically set in concert. Furthermore, many ECFP_4 features are overlapping and therefore describe similar or identical substructures. By contrast, ACCS utilized for fingerprint generation are selected to be maximally independent of each other.[5] Thus, correlation effects are expected to play a different role for these three fingerprints. In particular, ACCS-FP feature combinations are not necessarily a consequence of fragment correlation effects. Cliques can also be formed by fragments that overlap and, hence, describe larger substructures. Such

Table 5. Clique-Based Retrieval of Database Compounds[a]

activity class	Clique Database Distributions						
	$\nu = 0.5$	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$	$\nu = 1$	pooled
				MACCS			
AA2	7730	20885	53094	65435	91088	88006	23193
BK2	158	3541	6425	38528	75803	102747	3273
CAL	7078	24961	54214	59199	54214	64120	22725
DD1	7532	13740	37286	55423	69663	88004	17246
F7I	348	6333	18515	48227	74604	79421	4886
GLG	1136	7842	40932	85941	144425	113971	9131
GLY	3255	11182	31568	50112	75455	92245	16744
KRA	192	2207	7201	19914	43101	44758	2272
LAC	212	1449	5212	26080	67093	71593	2713
SQE	3803	17649	52751	81292	102744	102745	16490
SQS	129	1311	3256	13539	88005	102750	3285
THI	3102	26918	64501	92377	92552	92376	17646
ULD	914	9454	32052	64088	91528	101083	10153
XAN	8175	15494	22415	43582	57061	92247	21485
				ECFP_4			
AA2	60	108	103	61	58	13	62
BK2	7	19	58	1354	2336	414	20
CAL	8	38	959	1741	1632	960	636
DD1	3	17	32	46	7	6	12
F7I	89	1354	6971	9026	8065	3751	775
GLG	156	1959	3610	3962	3610	2655	1410
GLY	9	276	573	279	818	384	100
KRA	5	7	77	27	24	24	7
LAC	9	2	223	225	51	25	16
SQE	3	10	32	11	10	10	7
SQS	5	47	5	2	7	6	4
THI	305	3964	8743	11894	2013	759	3173
ULD	127	421	4649	5054	2671	2671	747
XAN	5	7	19	188	188	188	7
				ACCS-FP			
AA2	37	63	219	287	249	249	101
BK2	15	58	99	188	309	309	96
CAL	22	227	344	476	1117	1118	282
DD1	8	20	94	377	511	655	46
F7I	14	95	491	1062	1062	1062	77
GLG	5	175	668	668	702	702	81
GLY	14	82	232	411	434	434	51
KRA	11	12	34	32	32	32	15
LAC	4	34	74	169	432	432	51
SQE	25	26	26	80	218	218	29
SQS	5	6	15	25	289	513	13
THI	75	158	367	517	694	694	151
ULD	8	25	241	662	760	760	80
XAN	2	5	9	8	11	11	6

[a] The medians of database compound numbers that were retrieved using feature cliques detected at different co-occurrence threshold (ν) values are reported.

Table 6. Recovery Rates at Different Co-occurrence Thresholds

activity class	Recovery Rate						
	$\nu = 0.5$	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$	$\nu = 1$	pooled
				MACCS			
AA2	**4.8**	0.1	0.5	0.1	0.1	0.1	4.2
BK2	10.8	**18.4**	13.6	8.0	0.2	0.0	18.2
CAL	**14.3**	13.1	4.1	2.3	1.8	1.8	**14.3**
DD1	9.4	**15.1**	9.9	6.4	1.1	0.2	13.8
F7I	5.5	5.7	2.6	1.6	1.6	1.6	**6.4**
GLG	**4.9**	2.3	0.3	0.0	0.0	0.0	4.2
GLY	15.7	13.8	12.1	9.2	3.5	1.0	**19.6**
KRA	61.3	54.5	32.0	26.2	6.9	6.2	**61.9**
LAC	34.2	**36.1**	34.8	34.9	31.2	31.2	35.7
SQE	31.6	33.6	32.0	25.4	10.3	10.3	**35.0**
SQS	33.4	29.8	31.2	10.1	1.0	0.3	**36.5**
THI	**1.0**	0.1	0.1	0.1	0.1	0.1	**1.0**
ULD	3.4	**5.7**	3.9	3.8	3.8	3.8	3.4
XAN	39.1	44.9	21.6	6.1	0.5	0.1	**45.5**
				ECFP_4			
AA2	29	**34.8**	32.8	34	34.3	34.3	32.1
BK2	61.8	66.4	63.6	63.6	62.7	62.7	**69.1**
CAL	50.7	48.7	47.4	52.4	53.6	54.8	**60.0**
DD1	63.6	**76.3**	72.4	67.7	60.8	60.8	75.7
F7I	63.7	58.8	56.8	50.7	53.6	53.6	**67.3**
GLG	36.1	**39.1**	38.2	32.8	34.0	34.0	36.9
GLY	76.2	74.1	74.7	76.1	76.7	76.7	**78.8**
KRA	62.1	64.1	70.0	73.6	**75.5**	**75.5**	**75.5**
LAC	57.9	69.3	69.4	70.0	68.7	68.7	**72.3**
SQE	59.8	62.2	59.4	56.8	57.3	57.3	**66.4**
SQS	39.7	41.2	45.9	47.7	46.0	46.0	**48.5**
THI	40.8	40.7	46.4	**46.5**	46.4	46.4	45.9
ULD	36.3	36.0	35.9	36.9	36.9	36.9	**40.3**
XAN	56.2	67.1	65.3	69.2	69.4	70.6	**72.9**
				ACCS-FP			
AA2	14.6	15.9	15.5	15.0	15.0	15.0	**16.8**
BK2	32.3	43.5	44.5	**44.6**	41.5	41.5	**44.6**
CAL	28.3	**30.5**	16.9	12.6	11.7	11.7	28.8
DD1	56.4	64.1	64.4	59.2	48.4	44.4	**65.6**
F7I	11.7	**13.9**	6.9	5.9	5.9	5.9	13.6
GLG	25.9	27.7	22.0	18.9	11.5	11.5	**35.3**
GLY	37.0	27.8	21.6	17.4	15.2	15.2	**37.9**
KRA	36.0	47.7	47.7	51.3	50.9	50.9	**57.0**
LAC	47.7	49.3	47.2	34.6	22.7	22.7	**55.0**
SQE	28.0	29.4	26.9	19.7	17.9	17.9	**32.4**
SQS	34.0	41.7	50.3	43.4	44.0	43.5	**52.9**
THI	**12.3**	8.3	6.1	5.7	5.9	5.9	11.0
ULD	22.6	27.1	23.4	18.7	18.7	18.7	**31.7**
XAN	36.9	46.3	46.7	51.3	57.1	57.1	**61.2**

[a] Average recovery rates are reported for virtual screening trials using the FCoN clique search strategy with cliques detected at different co-occurrence threshold (ν) values. The top recovery rates for each activity class and fingerprint are shown in bold. In addition, the performance of pooled clique sets is reported.

Table 7. Virtual Screening Performance

activity class	Virtual Screening Performance (Recovery Rates, %)				
	FCoN	centroid	modal	1NN	3NN
			MACCS		
AA2	4.23	13.33	10.78	24.47	**28.33**
BK2	18.18	21.82	6.36	**26.36**	17.27
CAL	14.29	22.86	24.25	**32.86**	32.14
DD1	13.81	47.33	56.00	**65.07**	64.67
F7I	6.36	1.67	1.67	**9.17**	7.50
GLG	4.17	5.29	2.35	0.00	**7.65**
GLY	19.60	19.41	15.88	**41.67**	32.94
KRA	61.94	20.91	11.98	**82.73**	63.64
LAC	35.71	24.67	5.87	**60.67**	52.00
SQE	**35.00**	20.77	16.15	3.85	27.69
SQS	36.53	18.57	16.19	46.67	**46.90**
THI	0.97	0.59	1.76	**10.59**	7.06
ULD	3.38	**5.45**	4.74	0.00	3.64
XAN	45.48	43.33	38.33	**58.26**	56.11
			ECFP_4		
AA2	32.13	28.61	0.00	46.11	**51.11**
BK2	69.09	46.36	19.09	69.09	**73.64**
CAL	**60.04**	40.71	42.86	55.71	54.29
DD1	75.69	63.33	17.68	84.67	**86.67**
F7I	**67.26**	3.33	1.67	25.00	24.17
GLG	**36.87**	0.59	0.00	6.47	33.53
GLY	**78.82**	34.71	26.47	48.82	67.65
KRA	75.45	50.00	10.00	**89.09**	86.36
LAC	72.26	40.00	2.67	**76.00**	68.00
SQE	**66.36**	31.54	3.08	40.00	63.85
SQS	48.47	25.71	0.00	57.14	**57.62**
THI	**45.86**	4.71	0.00	29.41	33.53
ULD	40.30	13.64	0.00	29.09	**42.73**
XAN	72.94	53.33	15.00	**73.33**	67.22
			ACCS-FP		
AA2	**16.76**	1.24	NA	2.50	9.33
BK2	44.63	46.36	26.28	21.15	**48.18**
CAL	**28.76**	14.84	10.44	7.63	24.53
DD1	65.62	51.01	38.00	54.93	**73.88**
F7I	13.65	5.14	**15.14**	0.62	5.42
GLG	**35.26**	6.47	NA	1.70	12.07
GLY	**37.86**	17.07	2.54	27.98	26.08
KRA	**57.02**	33.94	7.51	26.56	55.12
LAC	**54.99**	23.93	3.57	52.87	42.39
SQE	**32.39**	11.83	13.19	7.53	16.09
SQS	**52.89**	36.19	7.29	12.91	42.86
THI	**10.99**	1.76	6.67	8.18	3.33
ULD	**31.67**	1.82	6.82	0.36	13.78
XAN	**61.21**	42.78	22.78	28.19	55.62

[a] Reported are average recovery rates of 10 independent virtual screening trials using the FCoN clique, centroid, modal, 1NN, and 3NN search strategies in combination with each fingerprint. The top recovery rates for each activity class and fingerprint are shown in bold. For modal fingerprints, the recovery rates for a threshold of 50% are given.

cliques are expected to frequently occur in database molecules, because the substructures that they represent tend to be more generic than cliques formed by nonoverlapping features. However, the finding that all three fingerprints produce feature combinations that rarely occur in ZINC compounds, as discussed previously, also shows that MACCS and ECFP_4 produce compound class-specific feature cliques that cannot be attributed to general correlation effects and can be identified using FCoN clique ranking. Thus, in addition to feature combinations that result from general correlation effects, activity-class-specific combinations are formed that can be utilized in clique searching, as described below.

In Table 5, clique database distributions at different co-occurrence threshold (ν) levels are reported. We observed the general trend that feature cliques identified at higher thresholds (i.e., highly conserved combinations) were smaller than cliques at lower thresholds but occurred in more database compounds. This finding indicated that feature combinations that are highly conserved in sets of active compounds do not always effectively discriminate between active and inactive database compounds, whereas some (but not all) larger cliques that occur in subsets of active compounds are often only found in small numbers of database compounds. Thus, fingerprint features that are highly conserved in reference sets are not necessarily a compound class-specific signature, because they might also be generic and be present in many different compounds.

Taken together, the results of FCoN clique detection show that fingerprint features frequently occur in combination; for all activity classes and reference sets, multiple cliques were

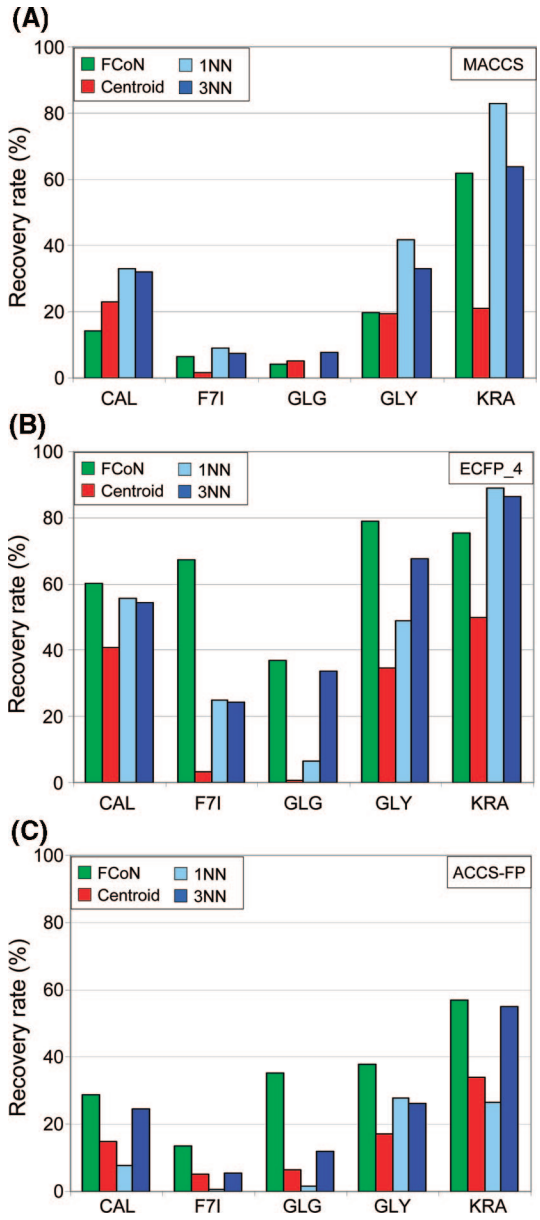

Figure 5. Virtual screening trials. Average recovery rates over 10 independent trials are shown for selection sets of 100 top-scoring compounds for three structurally heterogeneous classes (CAL, F7I, GLG) and two homogeneous classes (GLY, KRA), using different search strategies ((A) MACCS, (B) ECFP_4, and (C) ACCS-FP).

detected. These cliques are often small in size (2−10 features per clique) and the total clique numbers roughly scale with the length of the original fingerprints. Moreover, the distribution of reference set cliques in a large compound database is an indicator of fingerprint information content. Feature cliques of active molecules extracted from the general MACCS fingerprint are found in many database compounds, the molecule-centric ECFP_4 feature combinations occur in fewer database compounds, and many feature combinations produced by the compound class-directed ACCS-FP are only matched by fingerprints of small numbers of database molecules.

Clique-Based Similarity Searching. The distribution of ACCS-FP cliques in active and database molecules suggested the design of a clique-based similarity search strategy that takes activity-class specificity of individual cliques into account. Hence, feature cliques are initially identified from fingerprint representations of active compounds. For each clique, the number of database compounds that contain all of its features then is determined. Cliques are sorted by ascending database frequencies (i.e., rarely occurring cliques are preferred). From this ranking, cumulative compound selection sets are generated. Thus, database compounds are selected based on cliques that occur with increasing frequency. For FCoN-based searching, we have calculated recovery rates for 100 top-ranked database compounds for clique sets derived using different co-occurrence thresholds (ν) as well as for the pooled sets. Table 6 reports the recovery rates of active compounds at different ν values. Generally, pooled clique sets produced results that were better or comparable to those obtained at individual threshold levels. Therefore, pooled cliques were used for further analysis.

Significance of Feature Combinations for Search Performance. To assess the contribution of feature combinations to search performance, clique-based compound selection was compared to established centroid, 1NN, and 3NN fingerprint search strategies on the basis of Tanimoto similarity and also to less-popular modal fingerprints. Table

Table 8. Pubchem Fingerprint Search Results[a]

activity class	Virtual Screening Results (Recovery Rates, %)					Clique Statistics		
	FCoN	centroid	modal	1NN	3NN	size	number	distribution
AA2	6.2	0	0.00	**15**	12.78	5	101	23780
BK2	**30.4**	1.82	0.00	9.2	8.18	7	179	1419
CAL	53.2	34.29	37.99	46.43	**57.86**	7	64	1982
DD1	2.2	24.67	24.83	37.33	**44**	5	57	16740
F7I	**16.7**	0.83	3.33	6.25	2.5	9	145	5211
GLG	**17.3**	1.18	0.00	5.29	12.94	6	142	6281
GLY	17.7	20	20.35	38.82	**40**	5	141	3288
KRA	57.5	20	8.18	**70**	56.36	9	88	2320
LAC	41.7	18	8.67	**65.33**	52	12	138	376
SQE	**44.3**	7.69	4.02	29.01	40	7	99	13324
SQS	48.4	19.05	9.52	**50**	40	16	436	35
THI	9.5	1.76	1.18	**30**	22.94	5	126	5149
ULD	17.4	3.64	6.36	4.56	**20**	8	107	5549
XAN	10.8	34.44	28.89	47.37	**55.56**	5	69	20544

[a] Recovery rates for selection sets of 100 compounds are shown for each activity class. Top recovery rates are marked in bold. Median clique numbers, clique sizes, and the median number of database molecules are provided for pooled clique sets.

Table 9. Fingerprint Comparison[a]

parameter	Value				
	FCoN	Centroid	Modal	1NN	3NN
			MACCS		
RR	21.40%	19.00%	15.17%	33.03%	31.97%
# Best	1	1	0	9	3
			Pubchem-FP		
RR	26.66%	13.38%	10.95%	32.47%	33.22%
# Best	4	0	0	4	6
			ECFP_4		
RR	60.11%	31.18%	9.89%	52.14%	57.88%
# Best	6	0	0	3	5
			ACCS-FP		
RR	38.84%	21.03%	11.45%	18.08%	30.62%
# Best	11	0	1	0	2

[a] Average recovery rates (abbreviated as "RR") over 14 activity classes are reported, as well as the number of classes with the highest recovery rate for each search strategy (denoted as "# Best").

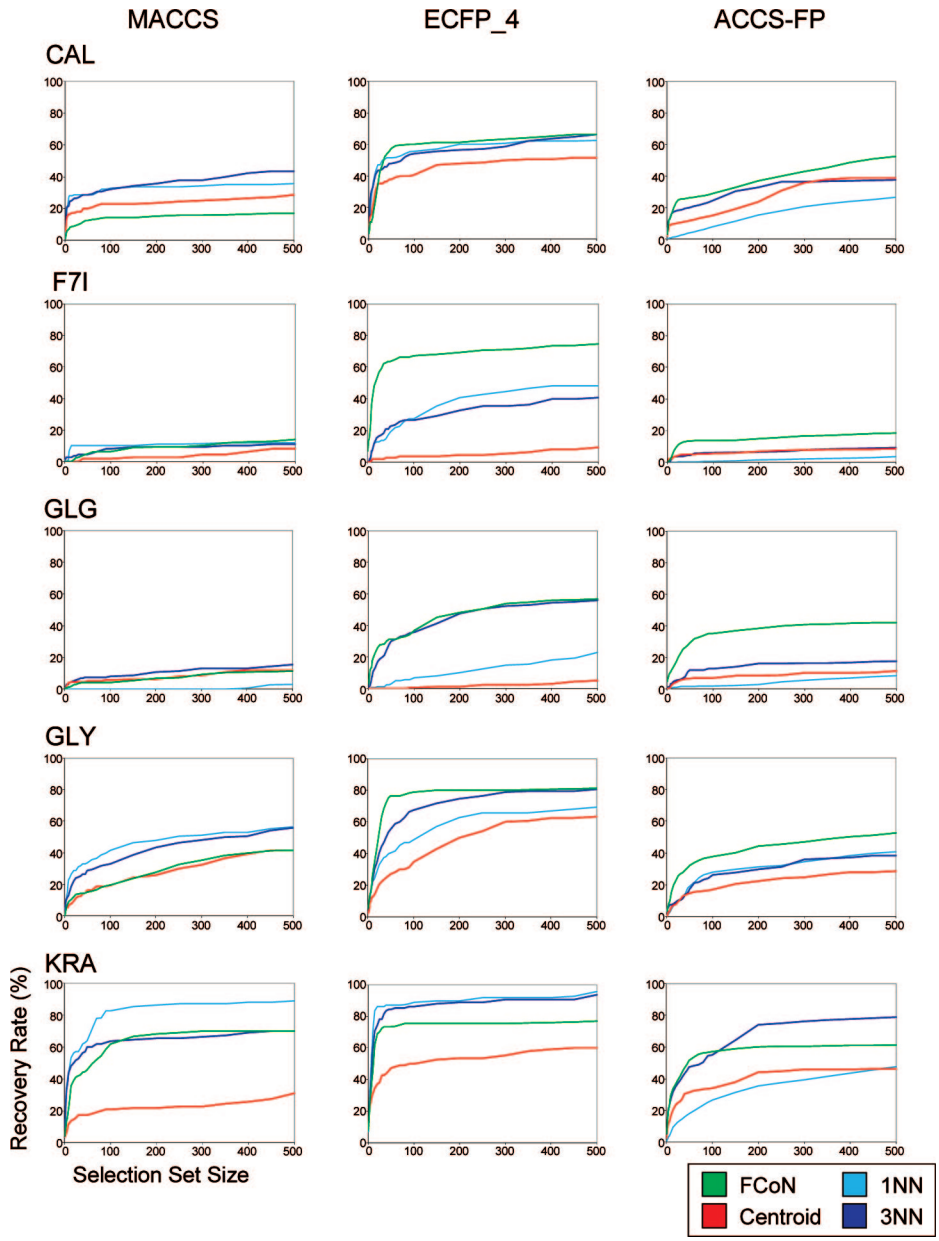

Figure 6. Compound recall curves. For the screening trials reported in Figure 5, cumulative recall curves for active compounds are provided.

7 reports the recovery rates for all activity classes and database selection sets of 100 compounds. For modal fingerprints, the 50% threshold performed consistently better for all fingerprints than the other thresholds; thus, we report only the best results for modal fingerprints. Figures 5A−C show recovery rate bar charts for three structurally heterogeneous (CAL, F7I, GLG) and two homogeneous (GLY, KRA) activity classes, and Figure 6 shows the corresponding recall curves. Depending on the fingerprint, clique-based search performance substantially varied. For MACCS, feature combinations produced lower recovery rates than standard search strategies for all but one activity class, which is consistent with the high database frequency of many MACCS-derived cliques and mirrors the fact that MACCS structural keys have been designed for a broad spectrum of small molecules.[18,19] For the ECFP_4 fingerprint that includes features derived from individual molecules, the overall results of clique searching were comparable to those of the other search strategies (see Figure 5B). For 6 out of 14 classes, cliques produced the highest recall rates. In particular, this was the case for structurally heterogeneous classes (F7I, GLG, THI) where standard search strategies had difficulties in retrieving active compounds. However, for ACCS-FPs, clique searching produced the highest recall for 11 of 14 compound classes (see Table 7 and Figure 5C). Similar to ECFP_4, a notable improvement in recovery rates was observed for heterogeneous classes that represented difficult test cases for nearest-neighbor or centroid searching (see Table 7).

As a control calculation for general fingerprints, we have also evaluated the Pubchem-FP, which is MACCS-like in its design but includes many more structural keys. The Pubchem-FP clique statistics and search results are reported in Table 8, and Table 9 provides a summary of the results obtained for all fingerprints. Despite the large difference in key numbers, Pubchem-FP behaved in a manner that was very similar to the 166 MACCS keys and did not increase its search performance in standard calculations. The clique

distribution among database compounds was also very similar for these two fingerprints. However, because of the larger number of keys, less-generic cliques were obtained for the Pubchem fingerprint, which has improved performance over MACCS keys in clique searching in a few cases. Overall, the highest recovery rates were obtained with ECFP_4, which is the most complex and molecule-centric fingerprint design that has been evaluated here.

The consistently high performance of ACCS-FP clique searching suggests that these cliques preferentially consist of activity-class characteristic features that have a high potential to retrieve active compounds, even if the feature combinations are only partially conserved in reference sets. These findings also indicate that specific feature combinations, rather than individual features, contain most class-specific information in ACCS-FP, which is consistent with the observation that ACCS often form coherent cores in active compounds.[20] By contrast, for structural fingerprints of general design such as MACCS or Pubchem-FP, the feature combinations do not determine search performance. In this case, counts of individual features and fingerprint overlap are a more-reliable measure, which is exploited in the calculation of Tanimoto similarity. For ECFP_4, emphasizing individual features or feature combinations produces comparable search results. For ACCS-FP, feature combinations effectively discriminate between active and database compounds, and, hence, clique searching is much superior to established search strategies for multiple reference compounds such as nearest-neighbor or centroid calculations.

CONCLUSIONS

In this study, we have introduced a methodology for the extraction of feature cliques from molecular fingerprints and their frequency-based prioritization, and we have investigated the role of feature combinations for similarity searching. The design of co-occurrence networks has made it possible to identify cliques systematically. Our analysis has revealed three major findings:

(1) Fingerprint features frequently occur in well-defined combinations.

(2) Feature combinations are highly relevant for the performance of compound class-directed fingerprints, in contrast to MACCS and Pubchem-FP. Class-directed cliques rarely occur in the screening database and are capable of significantly enriching selection sets with active compounds.

(3) Clique-based similarity searching represents a generally preferred strategy for ACCS-FPs, which are activity-class-directed fingerprints that utilize activity-class characteristic substructures extracted from random fragment populations of active compounds. Moreover, clique detection in co-occurrence networks can generally be applied to identify feature combinations that are conserved in active compounds.

Supporting Information Available: Supporting Information provides details of the ACCS-FP generation protocol. Figures S1 and S2 show box plots of clique and clique size distributions, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Willett, P.; Barnard, J.; Downs, G. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(2) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
(3) *BCI*; Digital Chemistry, Ltd.: Leeds, U.K., 2006.
(4) *Scitegic Pipeline Pilot*; Accelrys, Inc.: San Diego, CA, 2008.
(5) Batista, J.; Bajorath, J. Similarity Searching using Compound Class-Specific Combinations of Substructures Found in Randomly Generated Molecular Fragment Populations. *ChemMedChem* **2008**, *3*, 67–73.
(6) Batista, J.; Bajorath, J. Distribution of Randomly Generated Activity Class Characteristic Substructures in Diverse Active and Database Compounds. *Mol. Divers.* **2008**, *12*, 77–83.
(7) Lounkine, E.; Auer, J.; Bajorath, J. Formal Concept Analysis for the Identification of Molecular Fragment Combinations Specific for Active and Highly Potent Compounds. *J. Med. Chem.* **2008**, *51*, 5342–5348.
(8) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218–1225.
(9) Williams, C. Reverse Fingerprinting, Similarity Searching by Group Fusion and Fingerprint Bit Importance. *Mol. Divers.* **2006**, *10*, 311–332.
(10) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: An Algorithm to Determine Structural Commonalities in Diverse Datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.
(11) Irwin, J. J.; Shoichet, B. K. ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(12) *MDL Drug Data Report (MDDR)*; Symyx Software: San Ramon, CA, 2005.
(13) Pubchem Substructure Fingerprint. Available via the Internet at http://pubchem.ncbi.nlm.nih.gov.
(14) *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 2008.
(15) Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575–577.
(16) *Molecular Operating Environment (MOE)*; Chemical Computing Group, Inc.: Montreal, Canada, 2007.
(17) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
(18) McGregor, M.; Pallai, P. Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.
(19) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
(20) Lounkine, E.; Batista, J.; Bajorath, J. Mapping of Activity-Specific Fragment Pathways Isolated from Random Fragment Populations Reveals the Formation of Coherent Molecular Cores. *J. Chem. Inf. Model.* **2007**, *47*, 2133–2139.

CI800377N