CHEMMEDCHEM

DOI: 10.1002/cmdc.200800408

Improving the Search Performance of Extended Connectivity Fingerprints through Activity-Oriented Feature Filtering and Application of a Bit-Density-Dependent Similarity Function

Ye Hu, Eugen Lounkine, and Jürgen Bajorath*[a]

The Pipeline Pilot extended connectivity fingerprints (ECFPs) are currently among the most popular similarity search tools in drug discovery settings. ECFPs do not have a fixed bit string format but generate variable numbers of structural features for individual test molecules. This variable string design makes ECFP representations amenable to compound-class-directed modification. We have devised an intuitive feature-filtering technique that focuses ECFP search calculations on feature string ensembles of given compound activity classes. In combination with a simple bit-density-dependent similarity function, feature filtering consistently improved the search performance of ECFP calculations based on Tanimoto similarity and state-of-the-art data fusion techniques on a diverse array of activity classes. Feature filtering and the bit density similarity metric are easily implemented in the Pipeline Pilot environment. The approach provides a viable alternative to conventional similarity searching and should be of general interest to further improve the success rate of practical ECFP applications.

Introduction

Molecular substructures and fingerprints have a long history in chemical database mining[1–4] and are currently in wide use for similarity searching.[5–8] In addition to fingerprints that represent hashed connectivity pathways[9] and two-dimensional (2D) pharmacophore patterns,[10] fingerprints that record structural features[3,11–17] are among the most popular 2D fingerprints.[5,8] Various types of structural fingerprints are distinguished, particularly those based on substructure dictionaries such as BCI,[3,11] MACCS keys,[12,13] and ACCS-FPs,[17] as well as circular feature fingerprints such as layered atom environment[14] and extended connectivity fingerprints (ECFPs).[15,16] Dictionary-based fingerprints are of fixed length. For example, the standard version of the BCI fingerprint and the publicly available version of MACCS consist of 1052 and 166 structural descriptors, respectively, and each bit position accounts for the presence or absence of a specific structural feature. Modal versions of these types of fingerprints are also available that report fragment counts in test molecules, not only the presence or absence of catalogued fragments. BCI and MACCS are general in their design and do not incorporate compound-class-specific structural information. In contrast, with circular fingerprints[14–16] atom environments are systematically recorded in multiple layers around individual atoms. These features are then mapped to unique integer codes using a hash function.[16] However, different from hashed fingerprints,[9] substructures can be extracted from individual features. Furthermore, different from keyed fingerprint designs, circular fingerprints generate features in a molecule-directed manner, and the number of feature strings typically varies depending on the test molecules. Systematic comparisons of fingerprint search performance have shown that circular fingerprints are often the top-performing search tools,[18,19] in addition to compound-class-directed and trainable 2D molecular property fingerprints,[20] which perform particularly well on compound classes of increasing structural diversity.[21] Moreover, in addition to MACCS keys, Scitegic's ECFPs[15] are probably among the most widely applied fingerprints in current use, and this is due, in part, to their implementation in the popular Pipeline Pilot software environment.[16]

The performance of fingerprint similarity search calculations generally benefits from the availability of multiple reference compounds,[6,8] due to the increased content in chemical and structure–activity relationship information.[8] Accordingly, various strategies have been introduced for fingerprint searching using multiple reference compounds; these include scoring[10] or scaling[22,23] of compound set-characteristic bit patterns, fingerprint averaging techniques,[24,25] and data fusion approaches such as *k* nearest-neighbor (*k*NN) calculations.[26] In comparisons of alternative search strategies on different compound classes, nearest-neighbor methods often produce the highest recall of active compounds.[6,7]

The performance of fingerprint search calculations is generally influenced by three factors: the characteristics of the fingerprint, the similarity metric applied to quantitatively compare fingerprint bit settings, and the search strategy. Similar to

[a] *Y. Hu, E. Lounkine, Prof. Dr. J. Bajorath*
Department of Life Science Informatics, B-IT, LIMES
Program Unit Chemical Biology and Medicinal Chemistry
Rheinische Friedrich-Wilhelms-Universität Bonn
Dahlmannstr. 2, 53113 Bonn (Germany)
Fax: (+49) 228-2699-341
E-mail: bajorath@bit.uni-bonn.de

exploring different fingerprint designs and search strategies, attempts have also been carried out to identify similarity metrics that improve compound recall achieved with the Tanimoto coefficient (Tc)[4] that is most often applied to quantify molecular similarity based on fingerprint overlap. However, systematic explorations of many alternative similarity coefficients and their combinations have essentially failed to identify more promising alternatives.[27,28] Hence, Tc calculations continue to dominate the evaluation of fingerprint similarity searching.

We recently introduced a similarity function for a specific fingerprint design, the so-called ACCS-FPs,[17] which consist of a limited number of substructures that are characteristic of a given compound class. In their currently most advanced version, these class-directed structural fingerprints consist of only 20–30 bit positions on average,[29] and are thus much smaller than conventional fingerprints. Due to this very small size, Tc calculations are no longer a reliable similarity measure for ACCS-FPs, and we therefore designed a simple bit-density-dependent similarity metric that has produced promising search results.[29]

On the basis of these findings, we asked whether such an unconventional similarity measure for small bit string representations might also be applicable to general state-of-the-art fingerprints. In addition, we devised a simple filtering method that is applicable to molecule-oriented fingerprints and focuses the search on features that are prevalent in active compounds. We therefore applied the bit density metric to ECFP fingerprints in combination with feature filtering and compared the approach to Tc calculations and nearest-neighbor search strategies. The results of our analysis showed that bit density metric calculations with feature filtering consistently produced higher compound recall than standard calculations, thus improving the search performance of ECFP further. The simplicity of the approach introduced herein enables its routine application to circular fingerprints and provides an attractive alternative to currently preferred search strategies.

Table 1. Activity classes.

Code	Activity Class	Number of Molecules	Number of Unique Scaffolds	Avg_Tc[a]	STD_Tc[b]
5HT	5HT Reuptake Inhibitor	146	80	0.42	0.11
ACAT	ACAT Inhibitor	114	61	0.39	0.12
ARI	Aldose Reductase Inhibitor	134	63	0.39	0.12
COX	Cyclooxygenase Inhibitor	190	72	0.34	0.12
COX2	Cyclooxygenase-2 Inhibitor	96	42	0.35	0.13
FPT	Farnesyl Protein Transferase Inhibitor	106	66	0.38	0.14
FXa	Factor Xa Inhibitor	96	72	0.46	0.11
IL1	IL-1 Inhibitor	116	67	0.34	0.13
LKA	Leukotriene Antagonist	130	74	0.38	0.12
LPI	Lipid Peroxidation Inhibitor	176	93	0.36	0.12
LSI	Leukotriene Synthesis Inhibitor	128	63	0.36	0.11
MRI	Mediator Release Inhibitor	118	66	0.41	0.12
NOS	Nitric Oxide Synthase Inhibitor	106	62	0.41	0.13
PA2	Phospholipase A2 Inhibitor	106	52	0.33	0.13
PAF	PAF Antagonist	162	97	0.40	0.12
PDE	Phosphodiesterase III Inhibitor	94	48	0.42	0.12
PKC	Protein Kinase C Inhibitor	98	54	0.38	0.13
RTI	Reverse Transcriptase Inhibitor	214	85	0.37	0.11
TNF	TNF Inhibitor	194	94	0.38	0.13
TPK	Tyrosine-Specific Protein Kinase Inhibitor	218	102	0.36	0.12
TSI	Thromboxane Synthetase Inhibitor	96	58	0.38	0.12

[a] Average pair-wise Tanimoto coefficient values. [b] Standard deviations for Avg_Tc values were calculated using MACCS structural keys as an approximate measure of intra-class structural heterogeneity.

Methods and Materials

Compound data sets

A total of 21 activity classes were assembled from the MDL Drug Data Report (MDDR),[30] each of which contains between 94 and 218 compounds, as reported in Table 1. The number of unique core structures present in each activity was calculated by using a hierarchical scaffold analysis algorithm[31] and is also reported in Table 1. As can be seen, the activity classes were structurally diverse, as indicated by low average similarity values and the presence of many unique scaffolds. The composition of the MDDR compound data sets is freely available via http://www.lifescienceinformatics.uni-bonn.de (see Downloads). Each activity class was randomly divided into 10 different reference and test sets that were used in independent search trials. Each reference set contained 20 randomly selected compounds of an activity class, and the remaining active molecules were added to the background database as potential hits (i.e., between 74 and 198 active database compounds were available). The background database for similarity search trials consisted of 120000 randomly selected ZINC[32] compounds, all of which were considered inactive.

Fingerprints

Two versions of extended connectivity fingerprints, ECFP_4 and ECFP_6,[15] were used in their Pipeline Pilot[16] implementation. These fingerprints consist of molecule-specific sets of layered atom environments. A code is assigned to each non-hydrogen atom in a molecule that combines its mass, charge, element type, and the number of bonds to other atoms. A number of iterations are then performed by fusing the initial atom code with codes of neighboring atoms until a predefined bond diameter is reached, that is, a four-bond layer for ECFP_4 and a six-bond layer for ECFP_6. The resulting features are sampled, transformed through a hashing procedure, and recorded as integers. The choice of the ECFP_4 and ECFP_6 fingerprints over alternative ECFPs with larger bond diameters

was supported by similarity search test calculations reported in the Results section.

Feature filtering

Activity class features (ACF) were defined as all unique ECFP features produced by a reference set of active compounds. Reference compounds might generate overlapping yet distinct feature sets, and the union of all features represents the ACF set. As illustrated in Figure 1, ACF were calculated and then used to filter ECFP features generated by database compounds. Non-ACF, that is, features that only occured in database compounds but not in active reference molecules, were determined and removed from fingerprint representations of the database compounds. Thus, for the evaluation of the similarity of reference and database compounds, only ACF were considered.

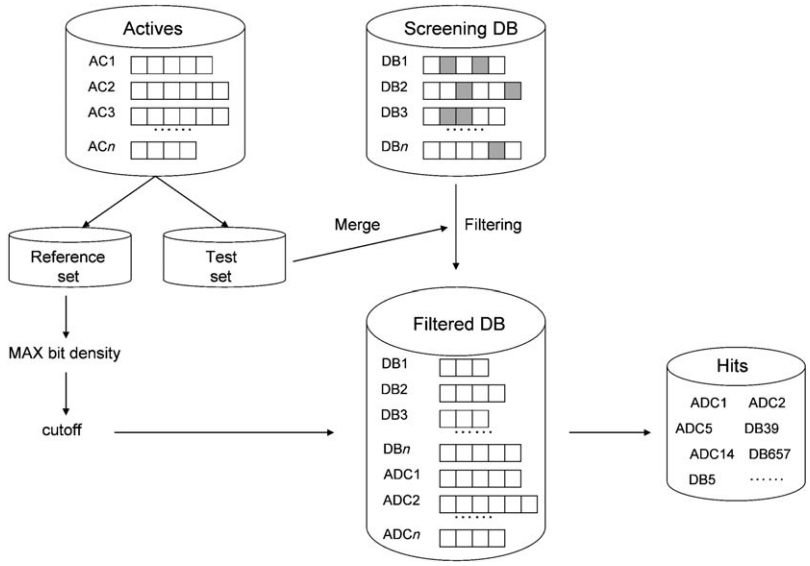

Figure 1. ACF BDM fingerprint searching: An outline of the ACF BDM approach to similarity searching using extended connectivity fingerprints is shown. Active (AC) and database (DB) compounds are encoded using ECFP_4 and ECFP_6. Features that do not occur in active compounds are deleted from database compound representations (gray shaded squares). From a reference set of active compounds, a bit or feature density cutoff is determined based on the maximal feature density observed within the set and the chosen cutoff coefficient. Using this similarity cutoff value, the feature-filtered database is searched, and a compound selection set (putative "Hits") is determined. ADC stands for active database compound.

Bit-density-dependent similarity metric

For similarity evaluation, we applied our previously introduced bit-density-dependent similarity metric (BDM),[29] which was originally designed for the comparison of short fragment fingerprints of fixed formats as an alternative to Tc calculations. For short fingerprints, BDM is calculated by dividing the number of bit positions that are set "on" (i.e. set to 1) in a given fingerprint by its total number of bit positions, hence producing a bit density. A preset bit density is then used as a similarity threshold or cutoff. It is defined as the maximal bit density *MAX* found in active molecules of a reference set multiplied by a coefficient within the range of [0.5, 1] to yield the final similarity cutoff:

$$cutoff = MAX * coefficient$$

Database molecules with a bit density equal to or greater than the final cutoff value are then considered active. Thus, different from Tc calculations, BDM in its original implementation does not produce a compound ranking; rather, the application of the cutoff value determines the size of the database selection set. For keyed fingerprints with a constant length, the application of the cutoff coefficient has the advantage that compound similarity can be assessed by using a sliding bit density window that captures the bit density within the reference set, whereas compound ranking does not use this information and only depends on the fingerprints of the database compounds.

The bit density metric as defined above can be easily adopted for the comparison of ECFP features. Following our ACF approach, BDM is calculated by dividing the number of ACF present in a test molecule by the total number of ACF present in the active reference set. The major difference between BDM applied to keyed fingerprints of fixed format, in its original application, and BDM applied to ACF is that ECFPs represent molecule-specific feature ensembles with essentially no limit on the number of features. Thus, the total number of ACF present in a reference set is variable and provides the basis for feature density calculations of individual reference and test compounds. For application of the cutoff coefficient, as described above, the highest density calculated for a reference compound becomes *MAX*, and database molecules with an ACF density equal to or greater than the cutoff are then considered to be active.

ECFPs are molecule-specific feature strings, and the union of features in active reference compounds represents all ACF, which provides an activity-class-dependent reference point for feature density calculations. Therefore, in this case, BDM can also be easily converted into a continuous ranking function by omitting the cutoff calculations. Database compounds are then directly ranked by their calculated ACF density, that is, the ratio of ACF they contain over the total number of ACF per reference set. In our analysis, we also explore this alternative feature density ranking scheme, termed BDMRank.

Similarity searching with feature filtering, BDM, and BDMRank

For ACF BDM calculations, six different coefficients were applied, ranging from 0.5 to 1 in increments of 0.1. For each reference set, a final coefficient was selected that produced a database selection set of a maximal size of close to 100 molecules, and for this selection set the recovery rate (RR) of active database compounds was determined. For each activity class, the average selection set size and recovery rate were then calculated over 10 individual search trials. ACF BDMRank calculations were carried out for all activity classes and recorded in cumulative recall curves for up to 1000 database compounds.

Reference calculations

ACF BDM calculations with ECFP_4 and ECFP_6 were compared with standard ECFP Tc calculations without feature filtering (unfiltered) and filtered ECFP Tc. For all calculations, the same reference sets were used. The Tanimoto coefficient is defined as:

$$Tc = c/(a + b - c)$$

in which c is the number of bit positions common to fingerprints A and B, a is the number of bit positions set on in A, and b is the number of bit positions set on in B. For Tc calculations, series of search calculations that apply four kNN search strategies[25] were carried out: 1NN, 5NN, 10NN, and 20NN. In 1NN calculations, a database compound is assigned the highest similarity value calculated against any of the 20 reference molecules. In the alternative calculations, similarity values are averaged over the five, 10, and 20 most similar reference molecules, respectively, to produce the final similarity value. Hence, for reference sets of 20 molecules, the 20NN calculations take into account contributions from all reference molecules equally. ACF BDMRank calculations were compared with 1NN similarity searching, which was found to be the best performing nearest-neighbor search strategy (see below).

Results and Discussion

Combining feature filtering and feature density calculations

The ACF BDM approach is outlined in Figure 1. It initially depends on generating and comparing ECFP feature ensembles for available active reference molecules and screening database compounds. Features not present in active compounds are removed from the ECFP representations of all database compounds. The maximum number of features per reference set compound is then determined and used as an upper-limit similarity threshold. A preferred similarity cutoff value is calculated by varying the cutoff coefficient in order to obtain a database selection set of suitable size. Thus, a difference from standard Tc calculations is that BDM does not provide continuous database compound ranking but database selection sets of varying size. For ensembles of molecule-specific feature strings (instead of constant-length fingerprints), BDM calculations can also be easily converted into a continuous ranking scheme, BDMRank, by omitting cutoff calculations. Feature filtering essentially focuses the search on fingerprint patterns of active reference compounds. In combination with BDM, compounds are retrieved that closely match the feature distribution of the reference set.

Fingerprint selection

To compare the basic similarity search performance of ECFPs with different bond diameters, 1NN Tc calculations were carried out over all activity classes with ECFP_4, _6, _8, _10, and _12, and average recovery rates were calculated for database selection sets of 100 compounds. For these fingerprints of increasing bond diameter, average recovery rates of 18.9, 18.6, 18.4, 18.3, and 18.2% were obtained, respectively. Thus, the search performance was overall very similar for the alternative ECFPs, but recovery rates decreased slightly with increasing bond diameters. Therefore, we focused our analysis on ECFP_4 and ECFP_6.

Feature density, filtering, and cutoff coefficients

The number of features generated by ECFP_4 and ECFP_6 was compared for different activity classes. Table 2 reports the median values of features in activity classes as well as the median values of features in database compounds after filtering. For ECFP_4 and ECFP_6, active compounds contained 41–53 and 55–80 features, respectively. The median numbers of

Table 2. Median feature density.[a]

AC	ECFP_4		ECFP_6	
	ACF[b]	Filtered DB	ACF[b]	Filtered DB
5HT	42.0	25.0	59.0	27.0
ACAT	50.0	26.0	69.0	27.0
ARI	46.5	27.0	62.0	28.0
COX	41.0	28.0	55.0	29.0
COX2	45.0	25.0	61.0	26.0
FPT	57.5	26.0	80.5	28.0
FXa	55.0	25.0	76.0	26.0
IL1	44.0	27.0	61.0	28.0
LKA	52.0	26.0	73.0	27.0
LPI	45.5	28.0	61.0	29.0
LSI	46.5	25.0	63.0	26.0
MRI	45.0	27.0	62.0	28.0
NOS	43.0	25.0	57.5	26.0
PA2	44.0	25.0	61.0	26.0
PAF	53.0	27.0	72.0	29.0
PDE	45.0	24.0	61.0	25.0
PKC	48.0	26.0	66.5	27.0
RTI	45.0	28.0	60.5	29.0
TNF	45.0	28.0	62.0	29.0
TPK	47.0	29.0	64.5	30.0
TSI	50.0	25.0	69.0	27.0

[a] For each activity class the median number of ECFP_4 or ECFP_6 features is reported. Additionally, the median numbers of activity class features in database compounds are given. [b] Activity class features.

unfiltered ECFP features present in the database compounds were 44 and 61 for ECFP_4 and ECFP_6, respectively. ECFP_6 generally produces more (and larger) features than ECFP_4 because of its larger bond diameter. Overall, active compounds and database molecules contained similar numbers of ECFP features and were therefore indistinguishable on the basis of feature frequency. However, after removal of non-ACF from database compounds, their feature median values were consistently decreased to $\leq$30 features. Thus, as a consequence of filtering, many database molecules could be separated from active compounds on the basis of ACF density. Consequently, we applied BDM that scores compounds according to the number of ACF they exhibit. Database compounds are selected if they reach an adjusted threshold feature density that is dependent on the maximal feature density observed in a reference set. In our ACF BDM calculations, cutoff coefficients were varied in order to obtain selection sets of close to 100 or fewer compounds (i.e. reasonably sized selection sets for practical applications). Table 3 reports average selection set sizes for different coefficients. Selection set size generally decreased with increasing coefficients and similarity thresholds. For all 21 activity classes and both fingerprints, coefficients were identified that yielded selection sets containing ~100 or fewer database compounds. Table 4 reports the average selection set size for each activity class. These sets contained between 42.1 (COX) and 111.3 (LSI) compounds. For these sets, recovery rates of active compounds were calculated.

Table 3. Database selection set sizes for different cutoff coefficients (ECFP_4 and ECFP_6).[a]

AC	0.5	0.6	0.7	0.8	0.9	1
ECFP_4:						
5HT	14462.8	3154.2	553.3	103.5	22.8	7.4
ACAT	12326.9	2589.9	480.1	94.5	21.6	7.3
ARI	19402.1	3846.7	634.4	93.0	15.1	4.1
COX	41386.4	11636.4	2086.3	252.3	32.8	7.9
COX2	17286.1	3454.0	580.7	62.4	11.4	4.5
FPT	3273.7	327.3	56.4	30.0	11.0	3.3
FXa	4898.1	548.6	81.8	33.2	10.8	3.1
IL1	22273.6	4213.6	577.4	64.1	14.7	5.3
LKA	8927.1	1491.0	238.9	58.2	17.7	4.2
LPI	28130.6	8420.0	1967.4	323.8	38.5	8.8
LSI	9037.6	1605.6	312.0	88.0	20.9	3.3
MRI	25347.0	7259.3	1577.3	270.5	47.2	12.4
NOS	20704.9	4440.7	629.1	82.6	16.3	5.0
PA2	24779.1	5438.5	1042.5	170.3	33.1	7.5
PAF	14774.2	2548.4	549.9	119.1	32.7	11.1
PDE	8076.4	1300.3	184.7	43.3	14.8	4.3
PKC	10430.1	1907.0	309.9	52.4	14.8	4.4
RTI	42863.8	15069.4	3589.3	772.6	112.5	25.8
TNF	20210.1	4789.4	657.2	96.7	19.9	6.0
TPK	31398.5	8719.7	1565.5	282.8	53.5	13.0
TSI	10368.7	1578.8	239.0	68.4	21.3	7.2
ECFP_6:						
5HT	1635.7	312.9	90.5	38.7	14.4	5.2
ACAT	2082.9	422.8	133.7	47.3	17.3	9.1
ARI	2499.1	401.4	104.8	36.1	12.9	5.1
COX	4902.3	633.1	124.8	47.8	20.9	7.4
COX2	1591.1	168.3	46.3	17.7	8.8	4.2
FPT	301.3	91.0	49.7	29.0	12.0	3.3
FXa	417.7	105.7	60.6	32.3	11.7	2.3
IL1	1726.1	204.5	60.5	29.5	12.7	5.2
LKA	819.4	276.3	105.1	40.7	13.0	3.3
LPI	4446.8	711.3	146.4	43.9	15.9	6.4
LSI	1254.0	441.0	176.8	55.8	17.8	3.7
MRI	3212.6	628.8	171.6	57.3	19.3	6.2
NOS	2396.3	251.0	61.4	27.7	11.8	4.0
PA2	2827.5	469.6	135.7	64.3	22.6	5.8
PAF	1604.8	483.1	188.1	66.4	25.9	9.7
PDE	550.9	139.6	55.9	22.7	7.9	3.4
PKC	867.6	190.0	72.1	28.8	10.8	4.5
RTI	9437.3	1823.1	315.3	102.2	37.2	18.5
TNF	2743.7	436.2	112.6	40.9	16.0	5.8
TPK	4848.1	750.4	230.0	92.0	28.9	10.6
TSI	591.0	206.7	92.9	35.0	10.6	3.2

[a] Average selection set sizes calculated from ten independent trials are reported for each activity class and cutoff coefficient for ECFP_4 or ECFP_6 fingerprints.

Similarity searching

Systematic ACF BDM calculations were carried out and compared with standard ECFP *k*NN Tc and control calculations. For each activity class, compound selection sets for ACF BDM, ACF Tc, and standard Tc calculations were of equal size. The results are summarized in Table 4, and Figure 2 shows representative recovery rate comparisons. For Tc calculations, search performance generally increased from 20NN to 1NN, consistent with earlier observations.[26] For ECFP_4, *k*NN Tc calculations produced maximal recovery rates per activity class between 5% (COX) and 26% (FXa), and ACF BDM between 10% (LPI) and 56% (FXa). For ECFP_6, maximal *k*NN Tc recovery rates between 10% (LPI) and 27% (FXa) were observed, whereas ACF BDM achieved maximum rates between 25% (MRI) and 85% (FXa). For both fingerprints and all 21 activity classes, the recall of active compounds was consistently higher for ACF BDM than for any other search calculation. Control calculations revealed that combining ACF filtering and the bit density metric was crucial for achieving top ECFP search performance. When

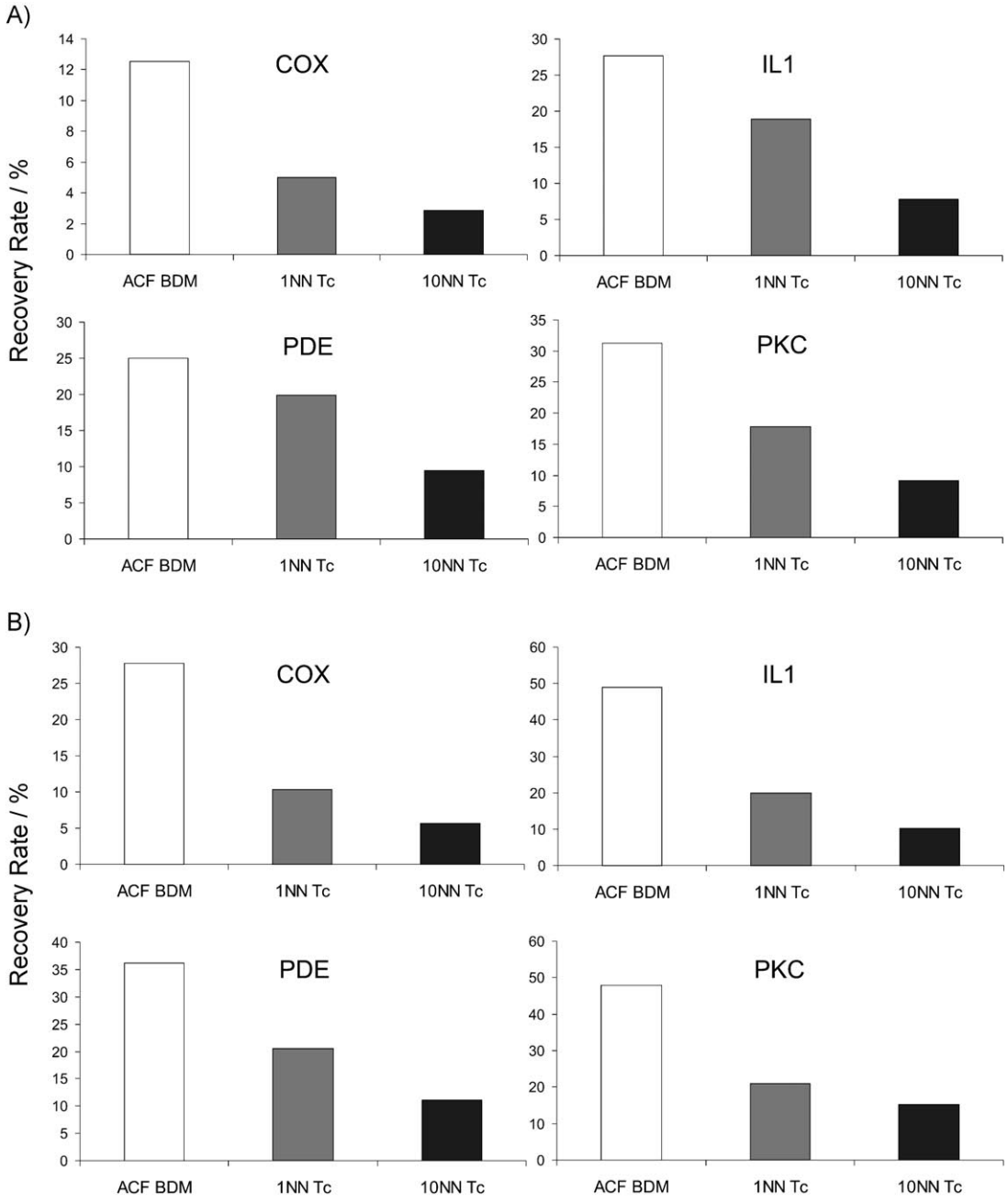

Figure 2. Exemplary recovery rate profiles: For A) ECFP_4 and B) ECFP_6, ACF BDM recovery rates are shown for four representative activity classes (COX, IL1, PDE, and PKC) and three search strategies (ACF BDM, 1NN Tc, and 10NN Tc).

BDM was applied without ACF filtering, database selection sets were typically very large, often containing thousands of molecules, and produced low recall of active compounds. This was simply a consequence of large numbers of non-ACF features that were present in database compounds and hence increased feature density in an activity-class-independent manner. As shown in Table 4, the application of the Tc metric to ACF-filtered fingerprints produced consistently lower recovery rates than standard *k*NN Tc calculations. This is due, at least in part, to the fact that ACF-filtered fingerprints are only of small size, as discussed above, which statistically limits Tc calculations. Furthermore, due to ACF filtering, database compounds have a general tendency to produce higher Tc values (see the Tc equation in the Methods section; after ACF filtering *b* decreases, whereas *a* and *c* remain constant, thus resulting in a systematic increase in Tc values for database compounds with decreased feature numbers). This might also make it more difficult to preferentially detect active compounds. However, the increase in compound recall of ACF BDM calculations over *k*NN Tc calculations was not only consistent, producing higher recovery rates for 21 compound classes without exception, but also substantial. For ECFP_4, average recovery rates over 21 classes were 26.3% for ACF BDM and 16.0% for 1NN Tc; for ECFP_6, the corresponding rates were 42.3% for ACF

Table 4. Similarity search trials.[a]

AC	ACF BDM		ACF Tc				Tc			
	Size	RR	1NN	5NN	10NN	20NN	1NN	5NN	10NN	20NN
ECFP_4:										
5HT	91.2	20.6	13.6	6.8	2.4	0.2	16.9	13.4	10.7	6.8
ACAT	106.4	36.4	14.0	5.2	1.9	0.3	17.7	11.8	9.7	7.2
ARI	71.8	23.7	8.8	3.3	1.1	0.1	11.5	8.3	6.6	4.9
COX	42.1	12.5	3.9	1.6	0.7	0.1	5.0	3.7	2.9	1.6
COX2	55.3	27.2	15.8	7.4	4.1	0.9	20.8	15.9	13.9	10.8
FPT	77	55.5	21.9	12.4	5.5	1.6	26.3	19.1	13.9	8.5
FXa	81.8	56.3	20.3	15.7	11.6	5.8	25.7	27.4	28.2	25.8
IL1	90.7	27.7	14.4	5.9	2.5	0.4	18.9	10.6	7.81	5.4
LKA	94.5	30.1	14.6	7.5	3.8	1.2	16.8	11.7	9.4	8.2
LPI	73.8	10.0	7.7	3.4	1.2	0.4	9.3	7.4	5.3	4.5
LSI	111.3	20.2	12.2	6.0	2.3	0.2	17.0	12.7	8.6	5.1
MRI	107.9	20.5	11.1	4.6	0.9	0.2	13.7	9.9	6.3	2.7
NOS	49.3	18.5	12.4	7.2	3.5	1.1	17.0	13.9	11.5	7.3
PA2	90	30.4	18.8	8.5	2.6	0.2	23.0	18.0	12.7	6.6
PAF	78.8	24.2	8.0	4.4	1.5	0.4	10.4	8.6	7.0	5.8
PDE	70.9	25.0	13.1	5.4	2.2	0.1	19.9	10.5	9.5	6.8
PKC	68.7	31.3	14.9	5.5	3.0	1.0	17.8	10.8	9.2	6.7
RTI	66.1	15.9	8.3	5.7	2.1	0.3	10.1	9.9	7.9	4.8
TNF	51	17.2	5.8	2.4	1.7	0.8	7.5	4.3	4.1	3.6
TPK	93.3	18.2	7.9	3.4	1.8	0.6	10.2	6.0	4.3	3.0
TSI	79.5	30.3	16.6	9.7	4.9	2.2	20.1	18.0	15.9	14.5
ECFP_6:										
5HT	103.5	32.3	12.1	5.6	1.7	0.4	16.5	13.1	11.7	9.0
ACAT	94.3	46.4	11.9	4.4	1.9	0.4	16.5	13.0	10.2	8.8
ARI	82.2	36.5	9.0	3.6	1.1	0.1	14.3	10.0	7.9	6.1
COX	89.9	27.8	6.7	2.1	1.1	0.1	10.3	7.2	5.7	3.5
COX2	71.6	62.4	17.4	7.8	3.8	0.7	25.4	20.8	18.2	14.5
FPT	84.3	74.8	21.7	11.8	5.2	2.1	26.2	20.1	16.6	11.3
FXa	105.7	84.6	20	13.3	8.7	3.3	27.0	33.3	31.8	28.6
IL1	89.3	49.0	15.2	5.5	2.6	0.3	19.9	12.3	10.1	8.4
LKA	93	36.7	13.9	5.6	2.6	0.8	15.7	11.4	9.5	8.1
LPI	86	32.3	7.3	3.1	1.1	0.1	9.7	8.7	7.3	6.0
LSI	88	23.8	9.2	4.4	2.0	0.3	13.0	9.9	8.3	6.0
MRI	72.4	25.2	7.4	3.6	1.0	0.2	11.2	8.0	6.6	4.2
NOS	91.7	50.2	14.8	9.4	4.0	0.8	26.3	22.2	18.9	12.8
PA2	90.8	43.4	15.8	8.5	3.4	0.2	22.2	22.4	17.6	10.2
PAF	104.6	34.6	9.3	4.7	1.8	0.3	13.2	12.2	9.7	7.7
PDE	79.9	36.2	11.7	5.0	1.9	0.3	20.5	12.2	11.1	8.1
PKC	96.9	47.8	15.5	7.2	4.2	1.0	20.9	18.8	15.3	10.6
RTI	84.4	29.7	8.7	5.5	1.6	0.1	12.3	11.7	10.3	7.8
TNF	87.4	32.1	7.4	3.0	1.7	0.8	10.8	6.8	6.6	6.4
TPK	89.2	27.8	7.2	3.0	1.6	0.4	11.1	7.0	5.8	4.4
TSI	118.8	54.9	19.5	10.4	4.2	2.1	26.8	22.9	20.9	19.3

[a] Average recovery rates from ten independent trials are reported for ACF BDM, ACF Tc, and standard Tc calculations. The selection set sizes for standard Tc and ACF Tc calculations are equal to those of ACF BDM trials.

BDM and 17.6% for 1NN Tc. Thus, whereas compound recall in standard calculations was similar for both fingerprints, ACF BDM substantially increased average recovery rates by ~10% for ECFP_4 and by ~25% for ECFP_6. In the latter case, the recovery rates for 19 of 21 activity classes doubled, or more than doubled, when ACF BDM was applied.

We also carried out systematic search calculations using ACF BDMRank and monitored the results in cumulative recall curves. Figure 3 shows ACF BDMRank recall curves for the same compound classes shown in Figure 2 compared with 1NN calculations. Furthermore, Figure 4 reports the average cumulative recall of active compounds over all 21 activity classes for ACF BDMRank calculations and the preferred 1NN similarity search strategy. Comparison of the results shows that differences in recovery rates were already significant for small selection sets, consistent with our ACF BDM results, and that a nearly optimal increase in search performance was generally achieved when selecting approximately 200 database compounds (Figure 4). Furthermore, the results illustrate that the increase in ACF BDM and ACF BDMRank search performance over standard calculations was significantly larger for ECFP_6 than for ECFP_4, although both fingerprints produced very similar average recovery rates in 1NN Tc calculations. This relative increase was likely due to the fact that ECFP_6 produced

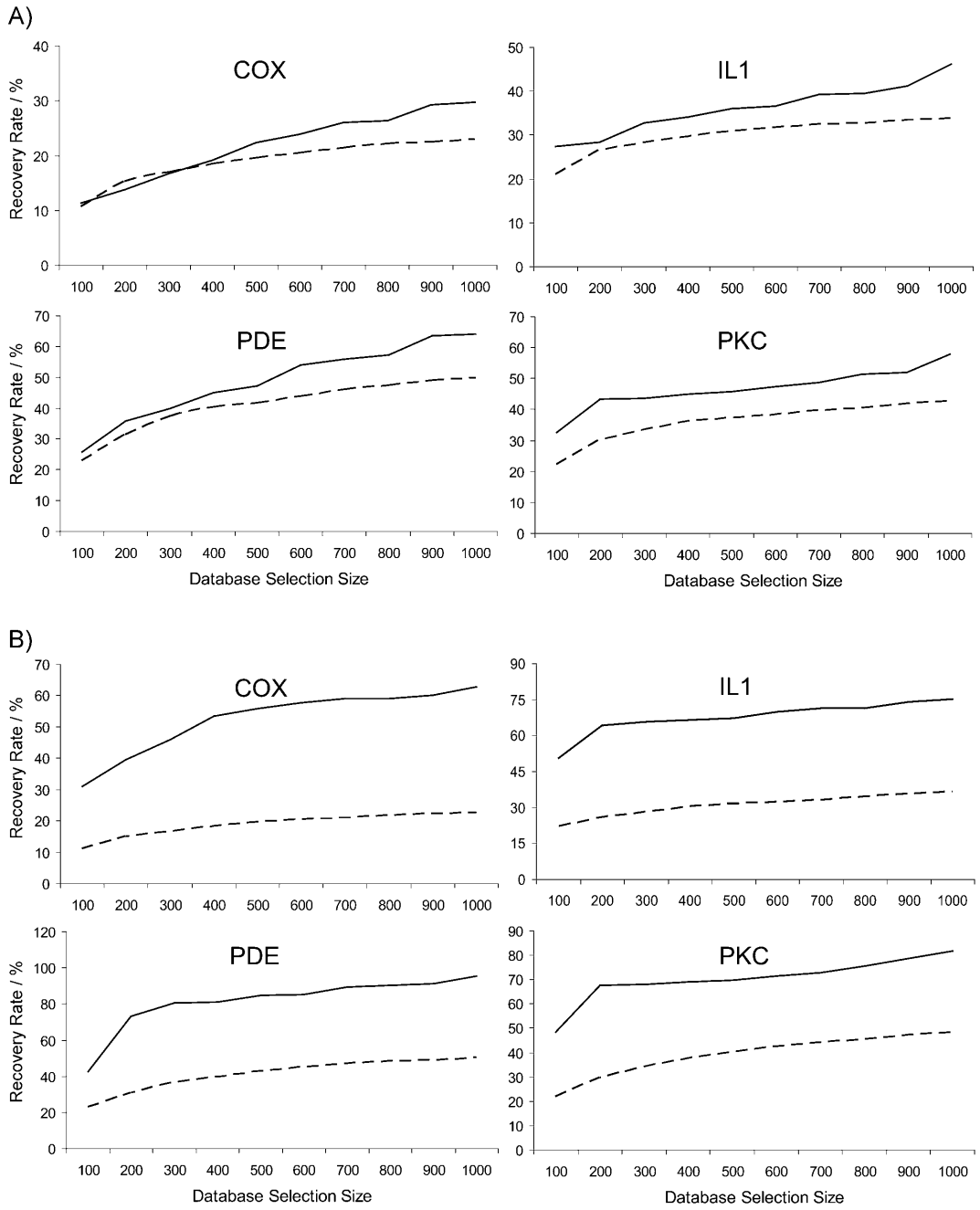

Figure 3. Exemplary cumulative recall curves: For A) ECFP_4 and B) ECFP_6, cumulative recall curves are shown for four representative activity classes (COX, IL1, PDE, and PKC; corresponding to Figure 2) obtained by ACF BDMRank (——) and 1NN Tc (-----) calculations.

more ACF than ECFP_4, as reported in Table 2. Overall, the performance improvement over standard ECFP similarity searching was highly significant; for database selection sets of 200 to 300 compounds, ACF BDMRank average recovery rates achieved with ECFP_6 nearly tripled relative to 1NN Tc calculations.

Conclusions

A similarity search approach has been introduced that is specifically designed for fingerprints with flexible and adjustable formats. Extended connectivity fingerprints are molecule-oriented fingerprints that generate structural features from individual compounds. Therefore, these fingerprint representations can be easily modified depending on the requirements of specific applications. We focused ECFP similarity searching on specific compound classes by taking into account only class-directed features collected from active reference molecules. For significantly decreased feature numbers, the calculation of Tanimoto similarity was found to be much less effective than the bit density metric. The combination of ACF filtering, BDM, and BDMRank calculations consistently improved the search perfor-

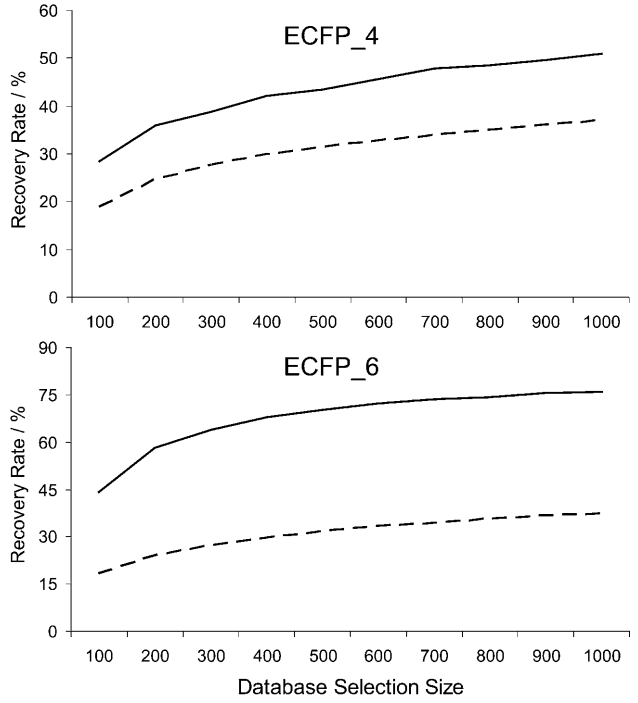

Figure 4. General search performance: For ECFP_4 and ECFP_6, the average cumulative recall of active compounds over all 21 activity classes is reported for ACF BDMRank (——) and 1NN Tc (-----) calculations.

mance of standard ECFP calculations for all activity classes. For both ECFP_4 and ECFP_6, the increase in the recall of active compounds was substantial. Control calculations revealed that feature filtering is critical for achieving high recovery rates. From a methodological point of view, attractive aspects of the ACF BDM(Rank) approach include its intuitive nature and simplicity of the calculations. From a practical perspective, the consistently better similarity search performance of ACF BDM(Rank) calculations relative to data fusion methods suggests that this approach provides a meaningful alternative to currently preferred search strategies and that it should have significant potential to identify active compounds in virtual screening applications.

Keywords: biological activity · chemoinformatics · extended connectivity fingerprints · feature filtering · similarity metrics

[1] G. W. Adamson, S. E. Creasey, M. F. Lynch, *J. Chem. Doc.* **1973**, *13*, 158–162.
[2] A. Feldman, L. Hodes, *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 147–152.
[3] J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.
[4] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
[5] C. Merlot, D. Domine, C. Cleva, D. J. Church, *Drug Discovery Today* **2003**, *8*, 594–602.
[6] P. Willett, *J. Med. Chem.* **2005**, *48*, 4183–4199.
[7] P. Willett, *Drug Discovery Today* **2006**, *11*, 1046–1053.
[8] H. Eckert, J. Bajorath, *Drug Discovery Today* **2007**, *12*, 515–521.
[9] C. A. James, D. Weininger, *Daylight Theory Manual*, Daylight Chemical Information Systems Inc.: Aliso Viejo, CA (USA) **2008**, http://www.daylight.com (accessed January 26, 2009).
[10] C. Williams, *Mol. Diversity* **2006**, *10*, 311–332.
[11] BCI, Digital Chemistry Ltd.: Leeds (UK) **2008**, http://www.digitalchemistry.co.uk (accessed January 26, 2009).
[12] MACCS Structural Keys, Symyx Software, San Ramon, CA (USA) **2008**, http://www.symyx.com (accessed January 26, 2009).
[13] J. L. Durant, B. A. Leland, D. R. Henry, J. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
[14] A. Bender, Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
[15] D. Rogers, R. D. Brown, M. Hahn, *J. Biomol. Screening* **2005**, *10*, 682–686.
[16] Scitegic Pipeline Pilot, Accelrys, Inc.: San Diego, CA (USA) **2008**, http://accelrys.com/products/scitegic (accessed January 26, 2009).
[17] J. Batista, J. Bajorath, *ChemMedChem* **2008**, *3*, 67–73.
[18] R. C. Glen, A. Bender, C. H. Arnby, L. Carlsson, S. Boyer, J. Smith, *IDrugs* **2006**, *9*, 199–204.
[19] R. C. Glen, S. E. Adams, *QSAR Comb. Sci.* **2006**, *25*, 1233–1242.
[20] H. Eckert, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 2515–2526.
[21] A. Tovar, H. Eckert, J. Bajorath, *ChemMedChem* **2007**, *2*, 208–217.
[22] L. Xue, F. L. Stahura, J. W. Godden, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746–753.
[23] L. Xue, J. W. Godden, F. L. Stahura, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218–1225.
[24] N. E. Shemetulskis, D. Weininger, C. J. Blankley, J. J. Yang, C. Humblet, C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.
[25] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
[26] J. Hert, P. Willet, D. J. Wilton, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
[27] J. D. Holliday, C.-Y. Hu, P. Willett, *Comb. Chem. High Throughput Screening* **2002**, *5*, 155–166.
[28] N. Salim, J. D. Holliday, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435–442.
[29] Y. Hu, E. Lounkine, J. Batista, J. Bajorath, *Chem. Biol. Drug Des.* **2008**, *72*, 341–349.
[30] Molecular Drug Data Report (MDDR), version 2005.02, Symyx Software: San Ramon, CA (USA) **2005**, http://www.symyx.com (accessed January 26, 2009).
[31] Xue, J. Bajorath, *J. Mol. Model.* **1999**, *5*, 97–102.
[32] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 177–182.

Received: November 24, 2008
Revised: January 6, 2009
Published online on March 4, 2009

