DOI: 10.1002/cmdc.200600225

Comparison of 2D Fingerprint Methods for Multiple-Template Similarity Searching on Compound Activity Classes of Increasing Structural Diversity

Andrea Tovar, Hanna Eckert, and Jürgen Bajorath*[a]

We studied the similarity search performance of differently designed molecular fingerprints using multiple reference structures and different search strategies. For this purpose, nine compound activity classes were assembled that exclusively consisted of molecules with different core structures and that represented different levels of intra-class structural diversity. Thus, there was a strict one-to-one correspondence between test compounds and core structures. Analysis of unique core structures was found to be a better measure of class diversity than distributions of simplified scaffolds. On increasingly diverse classes, a trainable fingerprint using a unique search strategy performed better than others tested herein. Overall, clear preferences were detected for nearest-neighbor search strategies over fingerprint-averaging techniques. Nearest-neighbor searching that relied on selecting database compounds most similar to one of the reference structures often improved compound recovery over other averaging methods, but at the cost of decreasing the ability to detect hits that were structurally distinct from reference molecules.

Introduction

Similarity searching using two-dimensional (2D) molecular fingerprints has a long history in pharmaceutical research[1] and continues to be one of the most widely used approaches for computational screening of compound databases.[1–4] The primary goal of fingerprint searching is the identification of different molecules that have similar biological activity to known active compounds. Thus, in pharmaceutical research, fingerprints are primarily used as a hit identification tool.[3] Fingerprints are bit-string representations of molecular structure and properties that encode various types of descriptors.[2,3] Typically, each bit detects the presence or absence of a specific chemical feature or represents a value range of a property descriptor. 2D fingerprints can be calculated from 2D molecular graphs of molecules, whereas three-dimensional fingerprints capture 3D pharmacophore patterns or encode conformation-dependent descriptors.[3,4]

Similarity searching involves the calculation of fingerprints of query and database compounds and quantitative assessment of pairwise fingerprint overlap as a measure of molecular similarity. For this purpose, a variety of similarity metrics has been introduced,[1] the most popular being the Tanimoto coefficient (*Tc*), which is defined in [Eq. (1)]:

$$Tc = \frac{N_{ab}}{N_a + N_b - N_{ab}} \qquad (1)$$

Here N_a is the number of bits set on in the fingerprint of molecule a, N_b the number of bits set on in b, and N_{ab} the number of bits common to both molecules. Applying similarity measures such as *Tc*, fingerprint search calculations produce a ranking of database compounds in order of decreasing similarity to query molecules which provides a basis for compound selection.

In addition to computational efficiency, another reason for the popularity of similarity searching is that fingerprints can be applied when only single-query compounds are available, whereas other approaches such as cluster analysis or machine-learning methods require the presence of multiple active templates.[3,5] However, compared with single reference structures, fingerprint searching often becomes more effective when different query molecules are available,[4,5] owing to an increase in information content of the calculations. Accordingly, for multiple template-based fingerprint searching, a number of different approaches have been introduced including consensus[6] or averaged[7] fingerprints, scaling procedures,[8] or nearest-neighbor methods.[7,9,10]

In this study, we focus on a performance evaluation of multiple-template similarity searching using state-of-the-art 2D fingerprints for two reasons: 2D fingerprints are surprisingly effective in many search situations in comparison with more complex 3D designs[2–4] and because we engage in the development of novel 2D fingerprints. To evaluate multiple template-based searching beyond investigations available thus

[a] A. Tovar, H. Eckert, 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

far,[7–11] we include conceptually diverse fingerprints in our comparison and, importantly, use sets of active compounds that were especially designed for the evaluation of similarity search performance. The results presented herein shed light on the relationship between compound diversity and relative performance of different fingerprints and search strategies that employ multiple reference structures. They also provide a rationale for differences in the ability of nearest neighbor and fingerprint averaging or bit-frequency methods to identify structurally diverse hits.

Computational Methods

Design of compound activity classes

To compare multiple-template searching in detail, we reasoned that compound activity classes had to fulfill two major requirements. First, these classes should only consist of selected compounds such that each compound contains a unique core structure. This is done to avoid similarity searching on very similar compounds or analogue series, which would be expected to yield artificially good results. Second, they should have increasing intra-class structural diversity in order to evaluate the potential of different methods to recognize diversified and thus "difficult" structure–activity relationships. To meet these specific requirements, we investigated a number of activity classes in the Molecular Drug Data Report (MDDR)[12] and ultimately selected compounds belonging to a total of nine classes, each containing between 22 and 27 compounds, as summarized in Table 1. For our analysis, only a subset of compounds was selected from each MDDR activity class and these compounds were assembled such that each compound contained a unique core structure. Thus, for each activity class studied herein, the total number of selected compounds per class corresponds to the number of core structures reported in Table 1. The evaluation and selection process proceeded in three steps. MDDR candidate compounds were first processed with an algorithm that isolates ring-containing core structures from molecules by removal of non-ring substituents,[13] and only compounds yielding unique core structures were accepted. In the second step, core-structure diversity was assessed by exhaustive comparison of structural resemblance using MACCS structural keys[14] and calculation of average *Tc* (av*Tc*) values. Then, unique core structures were transformed into cyclic carbon scaffolds using Meqi indices.[15,16] This additional step was carried out despite the fact that unique core structures provided the primary basis for the assessment of intra-class structural diversity, because it enabled us to consider class diversity from different points of view. Carbon scaffolds represent an abstraction from core structures and are obtained by omitting heteroatoms and bond order information.

Table 1. Diversity assessment of activity classes.[a]

Activity Class[b]	NoCore[c]	NoScaff[d]	min*Tc*[e]	max*Tc*[f]	av*Tc*[g]	SD[h]	Diversity[i]
ANG	27	26	0.186	0.931	0.529	0.135	low
REN	24	24	0.083	0.944	0.516	0.179	low
HIV	24	24	0.122	0.917	0.442	0.149	low
THR	23	22	0.046	0.951	0.404	0.183	medium
IL1	23	22	0.048	1.000	0.398	0.213	medium
ETA	22	18	0.106	0.821	0.396	0.140	medium
ARI	23	17	0.030	0.857	0.327	0.157	high
LSI	23	19	0.056	0.800	0.323	0.120	high
COX	24	19	0.048	0.864	0.293	0.122	high

[a] Statistics were derived based on pairwise calculation of *Tc* values for unique core structures using MACCS keys. [b] Activity classes are abbreviated as follows: ANG, angiotensin-II antagonists; ARI, aldose reductase inhibitors; COX, cyclooxygenase-2 inhibitors; ETA, endothelin antagonists; HIV, HIV protease inhibitors; IL1, IL-1β-converting enzyme inhibitors; LSI, leukotriene synthesis inhibitors; REN, renin inhibitors; THR, thrombin inhibitors. [c] NoCore specifies the number of unique core structures per class. [d] NoScaff is the number of corresponding cyclic carbon scaffolds. [e] min*Tc* = minimum *Tc* value. [f] max*Tc* = maximum *Tc* value. [g] av*Tc* = average *Tc* value. [h] av*Tc* standard deviation. [i] Activity classes are grouped into three sets (low, medium, high) representing different diversity categories.

Fingerprints

We included six conceptually different 2D fingerprints in our comparison that are representative of currently pursued fingerprint design strategies: BCI (consisting of 1052 bits),[17,18] Molprint2D (variable bit length),[19,20] TGD (735 bits),[21,22] TGT (13824 bits),[22] MPMFP (171 bits),[23] and PDR-FP (500 bits).[24] BCI is a preeminent structural fragment-based fingerprint, and the version representing its standard fragment dictionary (1052 fragments) was used. Molprint2D is derived from atom environments of the connectivity table of a molecule and combines strings of varying size, each representing a unique atom environment. The total number of possible strings could potentially reach ~2^{50}. Of the fingerprints studied herein, Molprint2D embodies by far the highest degree of complexity. TGD and TGT represent two-point and three-point pharmacophore-type fingerprints calculated from 2D molecular graphs, respectively, which are implemented in the Molecular Operating Environment (MOE).[22] TGD encodes atom-pair-type descriptors[21] using seven-atom features and distances of up to 15 bonds, and TGT captures triplets of four-atom features using three graph distances divided into six distance ranges. MPMFP is a hybrid fingerprint consisting of 110 structural keys and 61 binary-encoded 2D molecular property descriptors. Finally, PDR-FP represents a novel fingerprint design that exploits our previous observations that activity-selective descriptor value ranges could be identified for many molecular property descriptors and compound classes.[25] PDR-FP encodes value ranges of 93 molecular descriptors that were selected on the basis of their potential to adopt class-selective value ranges for different activity classes. In PDR-FP, value ranges of descriptors are encoded using between two and seven non-overlapping intervals (bits) such that the frequency of screening database compounds falling into each interval is the same. Because of this binning procedure, the database value distribution is implicitly accounted for in PDR-FP. Bit-string settings of multiple active template molecules are summed up to create an activity-oriented search string that captures the bit frequency at each position.[24] It should be emphasized that only the bit settings of the limited number of bait molecules (five molecules per class and calculation; see below) are used for class-directed training and the generation of the class search strings. This approach gives rise to what we call the "frequency" method for multiple template-based similarity searching, as described in the following section.

Multiple-template search strategies

Two major current search techniques were investigated: the centroid[7] and nearest-neighbor[7,9] methods. In addition, the newly introduced frequency approach that is unique to PDR-FP was investigated. The centroid approach calculates an average fingerprint for all template structures and compares it to fingerprints of database compounds using the general formulation of the *Tc* for numerical values,[1] rather than its conventional form for binary vectors. Like consensus fingerprints[6] or bit-scaling techniques,[8] the centroid approach emphasizes bit positions that are conserved in a set of active compounds and likely to account for activity-relevant features. In contrast, the nearest-neighbor method separately calculates the *Tc* similarity of a database compound to all individual reference structures. Then the similarity scores of the *k* nearest neighbors are averaged (*k*-NN or SUM fusion rule) or the largest observed *Tc* value is used (1-NN or MAX fusion rule).[9] We tested both rules and carried out 5-NN and 1-NN calculations.

The recently introduced frequency method that is specifically tailored to the design of PDR-FP makes use of the activity-oriented search string described above. In contrast to PDR-FP, this search string is no longer a binary representation, but captures the bit frequency of descriptor settings of active compounds relative to the database distribution. If frequency descriptor bits for a class of active molecules are predominantly focused on only one or two bit positions, the descriptor detects a highly activity-selective feature. Thus, the generation of an activity-oriented search string corresponds to activity-class-dependent training of PDR-FP. For the frequency method, a new similarity coefficient is introduced that compares individual PDR-FP settings of database compounds to an activity-oriented search string.[24] If the fingerprint of a database compound matches bit positions with high frequency values in the templates, it matches activity-specific descriptor settings and achieves a high similarity value (SV). This is accomplished through application of the following function [Eq. (2)]:

$$SV = \frac{\sum_{i=1}^{500} x_i y_i}{NF} \tag{2}$$

This equation represents the dot product of vector interpretations of the activity-oriented search string ($\bar{x}$) and an individual fingerprint ($\bar{y}$). NF is a normalization factor to produce similarity values between 0 (no similarity) and 1 (maximal similarity).

Calculations

As a source database, a "2D unique" version of ZINC[26] was used containing ~1.44 million compounds. This database was generated by selecting all molecules from ZINC that produced unique 2D graphs. For similarity searching, all ZINC compounds were considered inactive and potential false positives (although the database might also contain hits for the activity classes studied herein). For each activity class, sets of five compounds were randomly selected as templates ("baits") for fingerprint searching, and the remaining 17–22 active molecules were added to the source database as potential hits. Thus, small numbers of available hits in a very large source database presented an overall challenging search scenario. Recovery of active compounds was monitored among the 100 top-scoring database molecules as well as the top 0.1% of 2D unique ZINC (that is, 1442 database molecules). For PDR-FP, the frequency method was applied, and for the other fingerprints, the centroid, 1-NN, and 5-NN techniques were applied. For each combination of a fingerprint, search method, and activity class, 100 bait sets were randomly selected, and the results of 100 individual search calculations were averaged.

Results and Discussion

We systematically evaluated six 2D fingerprints of different complexity and four multiple-template search strategies on a total of nine compound classes of increasing structural diversity and which consist exclusively of compounds with different core structures.

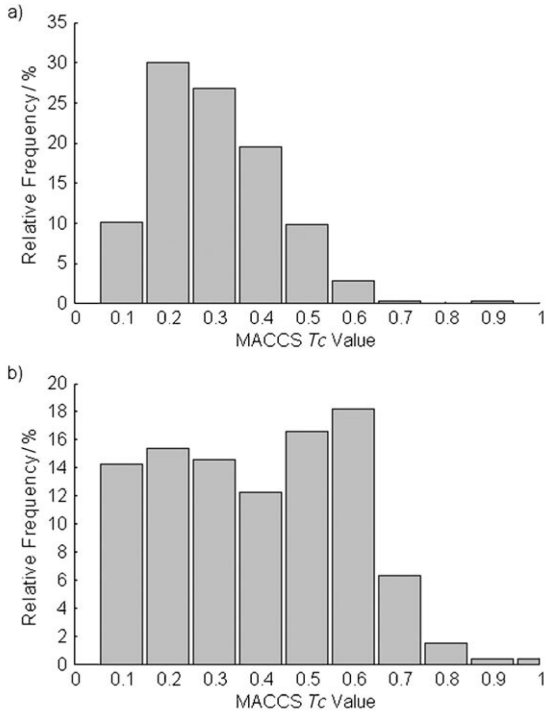

Figure 1. Distribution of *Tc* similarity values. For activity classes COX (a) and IL1 (b), the relative frequencies of pairwise MACCS *Tc* values are reported for systematic comparison of active compounds.

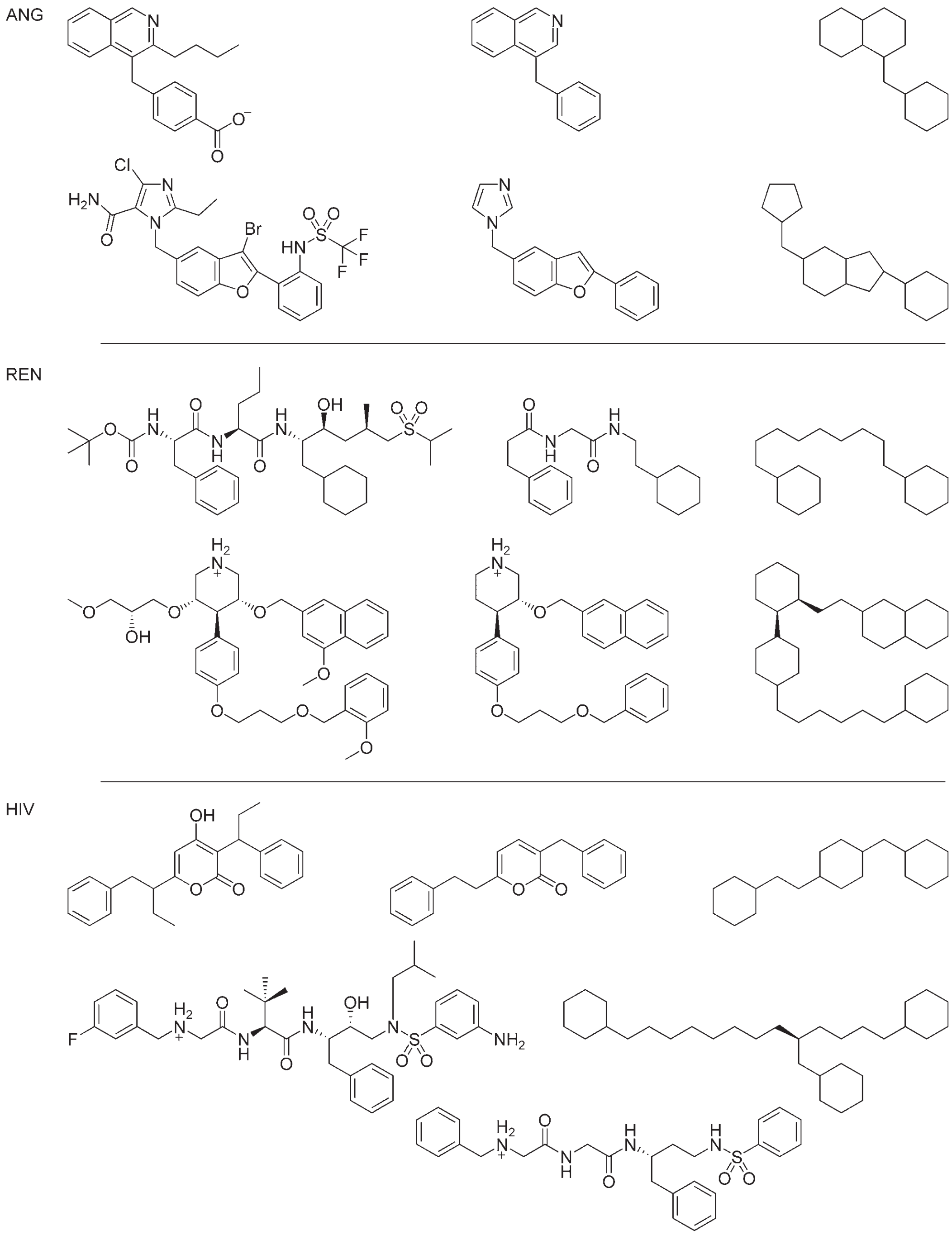

Figure 2. Active compounds, unique core structures, and cyclic carbon scaffolds for activity classes that belong to the low-diversity category; the pair of molecules with the lowest MACCS *Tc* value is shown on the left. Active compounds, their unique core structures, and the corresponding cyclic carbon scaffolds are shown from the left to right. In general, compound size decreases from the low to high diversity sets (compare with compounds shown in Figures 3 and 4).

Activity classes and relative structural diversity

On the basis of *Tc* similarity, the activity classes were selected to represent three groups of three classes, each characterized by what we consider to represent low (av*Tc*: 0.529–0.442), medium (0.404–0.396), and high (0.327–0.293) intra-class structural diversity, as reported in Table 1. Our assignment of compound classes to different diversity levels was based on the calculation of average pairwise *Tc* similarities (av*Tc*) rather than *Tc* standard deviations (SD) that we also investigated. This choice can be rationalized when comparing the intra-set diversity of activity classes IL1 and COX. The distributions of pairwise *Tc* values for these classes are reported in Figure 1. Both classes show similar minimum and maximum *Tc* values, but IL1 is considered medium diverse, as it has a larger av*Tc* value in combination with a high SD, whereas COX is considered highly diverse, even though it has a considerably smaller *Tc* SD. For the assessment of intra-set diversity, *Tc* SD can provide additional insight into the similarity distribution when considering it together with av*Tc* values. The low SD value of COX actually confirms that this class is of high structural diversity, as already indicated by its av*Tc* value. The low SD indicates that almost all compound pairs (76 %) have comparably low *Tc* similarity values of only 0.15–0.45, as shown in Figure 1 a, which closely match the av*Tc* of ~0.3. The underlying *Tc* distribution greatly challenges fingerprint methods to detect these similarity relationships. In contrast, IL1 has a higher SD value in combination with a higher av*Tc* value, indicating that there are many compound pairs (48 %) with *Tc* values greater than the av*Tc* of ~0.4, which can clearly be observed in Figure 1 b. Thus, small SD and av*Tc* values are indicators of high structural diversity, whereas larger av*Tc* and increasing SD values are consistent with the presence of broadly distributed pairwise *Tc* values and lower intra-class structural diversity. However, SD values were generally small for the compound sets designed herein and should not be taken as a diversity criteria.

A key task of our design effort has been the selection of individual molecules for a given activity class in order to satisfy our unique core structure and diversity requirements. From the MDDR, numerous activity classes can be selected that produce av*Tc* values at the compound (not core structure) level of approximately 0.7 or greater (and that are frequently used for benchmarking of similarity search methods). Thus, even the three classes we characterize herein as having low structural diversity already present rather challenging test cases.

In our analysis, we deliberately distinguished between unique core structures and corresponding cyclic carbon scaffolds that represent another level of abstraction. This was done because diversity assessments on the basis of unique core structures and simplified scaffolds did not yield equivalent results. In Table 1, the number of unique core structures and corresponding carbon scaffolds are reported for each activity class, revealing an unexpected relationship, at least at first glance. For low- and medium-diversity classes, the numbers of unique core structures and scaffolds corresponded closely, with one exception (ETA). However, for the three classes with high diversity on the basis of core structure comparison, four to six fewer scaffolds than unique core structures were produced in each case. This apparent inconsistency could be rationalized when the original molecules of different classes were compared. Representative compounds, unique core structures, and corresponding carbon scaffolds are shown in Figures 2–4. Molecules in our high-diversity classes (Figure 4) were often smaller than those in medium- (Figure 3) and low-diversity classes (Figure 2). High-diversity compounds contained more unique chemical features, but had less topological variation than larger molecules and therefore a higher probability to produce the same simplified scaffold. Thus, the evaluation of relative diversity might change with the reference frame that is applied; therefore, relying on simplified scaffold representations might be mislead-

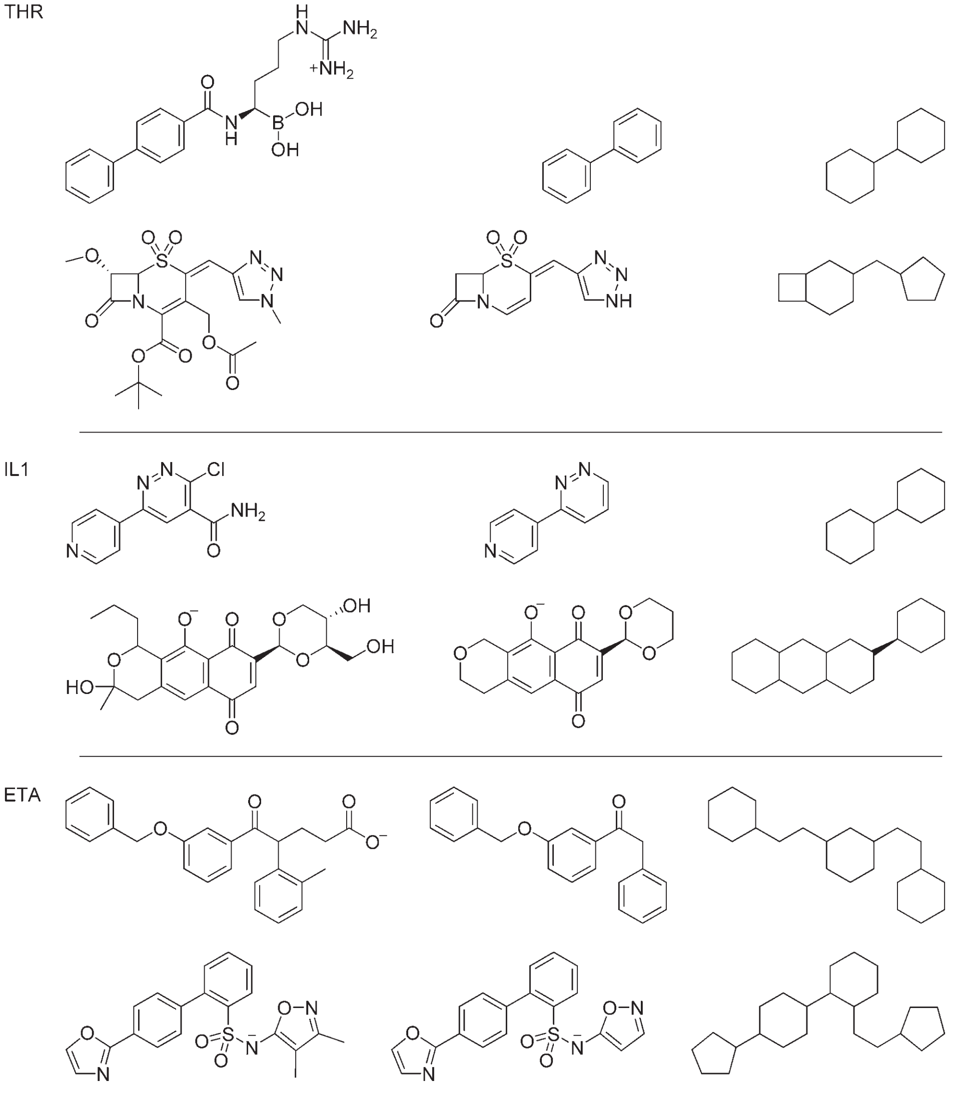

Figure 3. Active compounds, unique core structures, and cyclic carbon scaffolds for activity classes that belong to the medium-diversity category; the pair of molecules with the lowest MACCS *Tc* value is shown on the left. Active compounds, their unique core structures, and the corresponding cyclic carbon scaffolds are shown from the left to right.

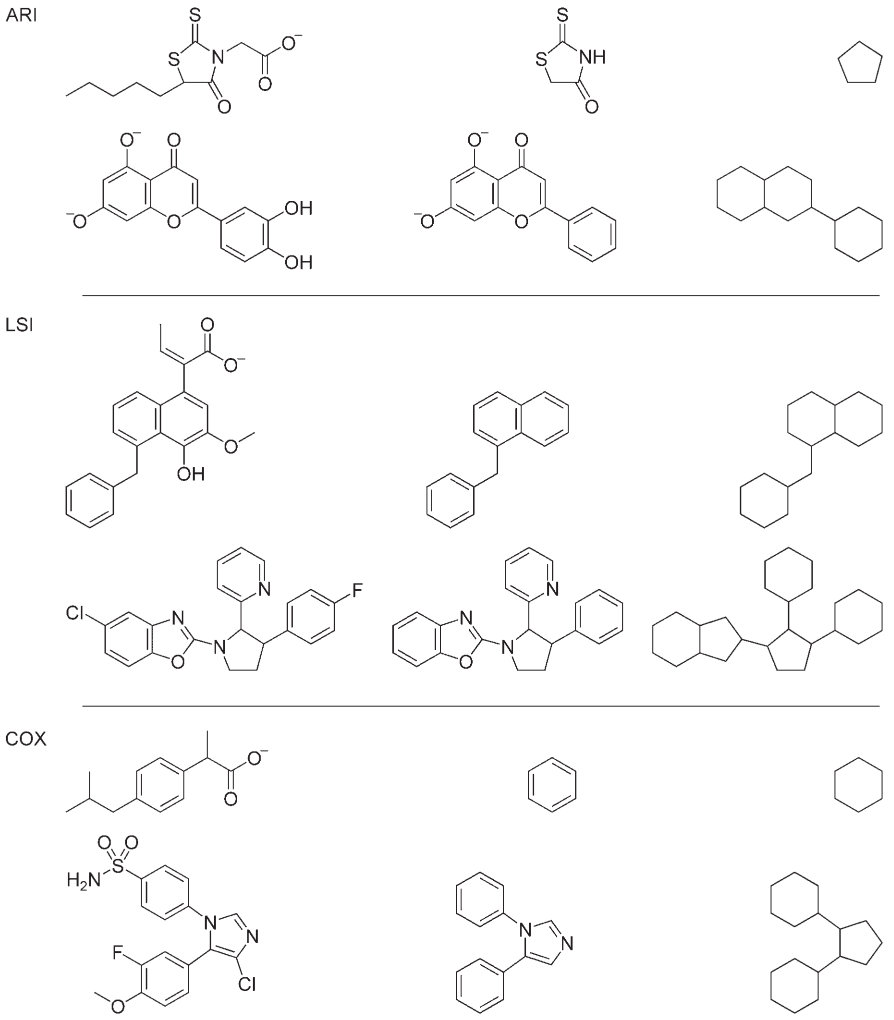

Figure 4. Active compounds, unique core structures, and cyclic carbon scaffolds for activity classes that belong to the high-diversity category; the pair of molecules with the lowest MACCS *Tc* value is shown on the left. Active compounds, their unique core structures, and the corresponding cyclic carbon scaffolds are shown from the left to right.

ing. Relative diversity assessment is further complicated by the fact that *Tc* calculations have a general tendency to produce higher values for comparison of larger molecules than smaller ones, irrespective of the fingerprints used, although such effects are often subtle.[27] The results of our search calculations, as described below, were clearly more consistent with diversity evaluation focusing on unique core structures rather than cyclic scaffolds. These findings also highlight the need to clearly define molecular scaffold representations when assessing diversity or the potential of similarity-based methods to recognize different structural classes with similar activity.[28]

Fingerprint performance

Initially, we investigated the performance of different fingerprints in the recognition of compounds with similar activity. Tables 2–4 summarize the results of our systematic search calculations. For each fingerprint and activity class, the search strategy that produced the best results is reported. For each diversity category, average results are presented in Table 5. For classes that belong to the low-diversity category (Table 2), all fingerprints produced acceptable to very good results overall. A minimum recovery rate of ~15% was observed for TGD and activity class HIV, corresponding to the presence of about three correctly identified active compounds among the 100 top-scoring molecules. A maximum recovery rate of ~88% was observed for PDR-FP and REN, corresponding to the detection of about 17 of 19 potential hits among the top 100 molecules. In many cases, the specificity of the calculations was high because compound recovery did not greatly improve when the top 0.1% of the screening database was selected. On low-diversity classes, PDR-FP and Molprint2D gave the best results overall and were similar in their performance (Table 5).

On medium-diversity classes, the performance of all fingerprints was consistently lower than on low-diversity classes, as expected, and relative differences in fingerprint performance became more distinct (Table 3). On these classes, fingerprints encoding structural fragment-type descriptors showed substantially decreased search performance, consistent with lower structural similarity between active compounds. On medium-diversity classes, PDR-FP dominated the calculations for selection sets of 100 database compounds producing recovery rates between ~23% and ~55%. On average, PDR-FP achieved 17% higher recovery rates than Molprint2D, which corresponded to the detection of about three more hits per calculation (Table 5). The superior performance of PDR-FP calculations on classes of increasing diversity might be a result of its inherent class-specific training potential, which sets its design apart from other fingerprints that cannot be adjusted to compound-class-selective features. However, considering the conceptually different designs of the fingerprints compared herein, their different degrees of complexity, and their focus on either structural features or molecular properties, this cannot be concluded with certainty. For BCI, a training effect could, in principle, also be achieved by adding signature fragments of activity classes to its fragment diction-

Table 2. Recall rates for low-diversity activity classes.[a]

Activity Class	Method	Best Approach	RR_100 [%][b]	RR_1442 [%][b]
ANG	Molprint2D	5-NN	56.7	67.7
	PDR-FP	frequency	33.3	49.0
	TGD	5-NN	30.5	49.9
	MPMFP	centroid	29.9	49.5
	BCI	5-NN	25.8	44.3
	TGT	5-NN	17.9	28.8
	av[c]		32.4	48.2
REN	PDR-FP	frequency	88.2	95.7
	Molprint2D	5-NN	77.3	88.1
	BCI	5-NN	65.7	85.6
	TGT	5-NN	59.2	71.3
	MPMFP	5-NN	45.3	63.7
	TGD	5-NN	33.8	47.9
	av[c]		61.6	75.4
HIV	PDR-FP	frequency	68.3	87.6
	Molprint2D	5-NN	50.5	69.8
	TGT	5-NN	39.2	58.9
	MPMFP	centroid	29.0	49.0
	BCI	5-NN	23.4	46.7
	TGD	1-NN	14.6	24.5
	av[c]		37.5	56.1

[a] Fingerprints are ranked according to recovery rates of active compounds among the 100 top-scoring database molecules. In each case, the multiple-template search method giving best results is reported. Activity classes are abbreviated according to Table 1. [b] Recall rates are reported for the 100 top-scoring database molecules (RR_100) and 0.1% of 2D unique ZINC (RR_1442). [c] av = average.

Table 3. Recall rates for medium-diversity activity classes.[a]

Activity Class	Method	Best Approach	RR_100 [%]	RR_1442 [%]
THR	PDR-FP	frequency	31.2	47.0
	TGT	1-NN	13.3	25.8
	Molprint2D	1-NN	11.3	15.7
	MPMFP	5-NN	7.6	18.9
	TGD	1-NN	6.3	11.2
	BCI	5-NN	5.1	14.8
	av		12.5	22.6
IL1	PDR-FP	frequency	55.2	65.6
	Molprint2D	5-NN	39.3	59.2
	TGT	1-NN	23.7	41.4
	TGD	5-NN	21.8	32.6
	MPMFP	5-NN	5.2	17.2
	BCI	1-NN	5.1	13.9
	av		25.1	38.3
ETA	PDR-FP	frequency	22.6	46.2
	TGT	1-NN	7.6	18.4
	TGD	5-NN	7.4	28.0
	Molprint2D	5-NN	7.3	16.2
	MPMFP	5-NN	4.7	12.8
	BCI	5-NN	3.8	15.4
	av		8.9	22.8

[a] Abbreviations used are as defined in the legend of Table 2.

Table 4. Recall rates for high-diversity activity classes.[a]

Activity Class	Method	Best Approach	RR_100 [%]	RR_1442 [%]
ARI	Molprint2D	1-NN	2.4	6.4
	BCI	1-NN	2.1	6.2
	TGT	1-NN	1.5	3.2
	TGD	1-NN	1.2	6.2
	MPMFP	5-NN	1.1	6.5
	PDR-FP	frequency	0.7	2.4
	av		1.5	5.2
LSI	Molprint2D	1-NN	2.7	6.1
	PDR-FP	frequency	0.7	5.7
	BCI	5-NN	0.7	2.5
	MPMFP	centroid	0.6	2.3
	TGT	5-NN	0.1	1.0
	TGD	5-NN	0.0	0.6
	av		0.6	3.0
COX	BCI	5-NN	1.4	4.0
	Molprint2D	5-NN	1.2	4.6
	TGT	1-NN	0.9	2.8
	TGD	1-NN	0.7	1.6
	MPMFP	centroid	0.2	1.3
	PDR-FP	frequency	0.1	1.4
	av		0.8	2.6

[a] Abbreviations used are as defined in the legend of Table 2.

ary, if such fragments could be identified, which would be expected to further increase search performance. The observed differences in performance between PDR-FP and Molprint2D are particularly notable when considering the differences in size and complexity between these fingerprints. Compared with Molprint2D, PDR-FP is a very simple fingerprint, which fur-

Table 5. Average performance of fingerprints at different diversity levels[a]

Diversity	Method	RR_100 [%]	RR_1442 [%]
High	Molprint2D	2.1	6.0
	BCI	1.4	4.3
	TGT	0.8	3.0
	MPMFP	0.6	3.7
	TGD	0.6	3.1
	PDR-FP	0.5	3.2
Medium	PDR-FP	36.3	52.9
	Molprint2D	19.3	30.7
	TGT	14.9	30.3
	TGD	11.8	24.4
	MPMFP	5.8	17.1
	BCI	4.7	15.7
Low	PDR-FP	63.3	77.4
	Molprint2D	61.5	75.5
	TGT	38.8	53.0
	BCI	38.3	58.9
	MPMFP	34.7	54.1
	TGD	26.3	42.1

[a] Average search results per diversity category are reported. Fingerprints are ranked according to recovery rates of active compounds among the 100 top-scoring database molecules.

ther emphasizes the potential of class-specific training as a major aspect of the PDR-FP design. The composition of our activity classes made it possible to investigate the scaffold-hopping[28] potential of the similarity search calculations in detail and also the specificity of molecular recognitions. An example is shown in Figure 5 for activity class THR. The search calculation using the bait set shown in this figure produced multiple hits that were characterized by scaffold diversity and either identified with several methods or only with one of them. For the fingerprints investigated herein, the ability to facilitate a transition from given active compounds to structurally distinct compounds could be confirmed.

On the high-diversity set we assembled, essentially all fingerprints failed to produce statistically relevant search results (Table 4). In these calculations, a maximal recovery rate of ~3% among 100 top-scoring compounds was achieved by Molprint2D on activity class LSI, which corresponded to the detection of less than one compound per search calculation. Thus, many individual search calculations failed to produce hits. It follows that the diversity of structure–activity relationships encoded in these compound sets goes beyond the search potential of currently available 2D fingerprints, which should also make these compounds excellent test sets for the evaluation of other virtual screening methods.[29] However, those individu-

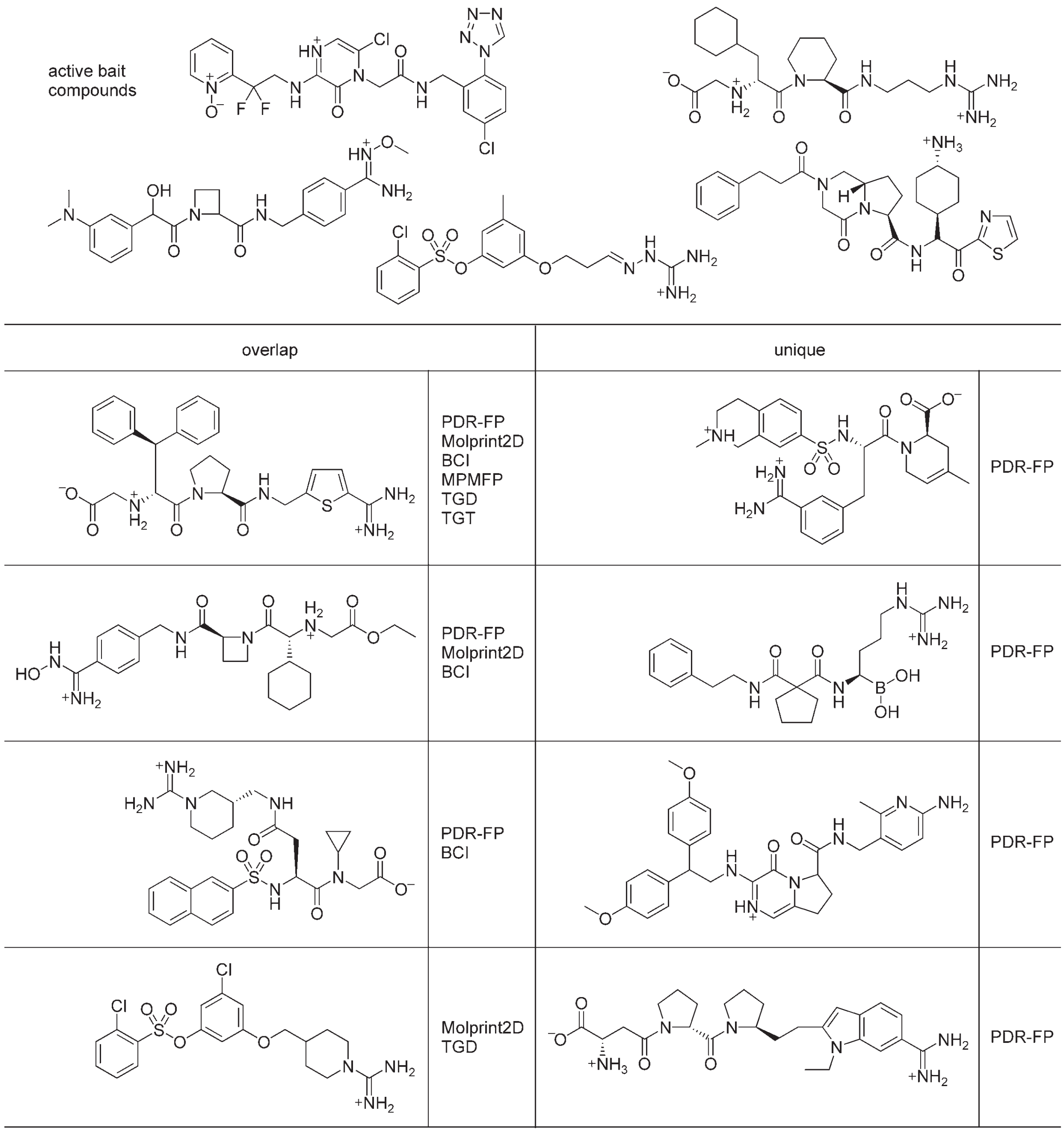

Figure 5. Diverse hits identified in a virtual screening trial. For activity class THR, all hits are shown that were correctly identified by the different fingerprint methods when selecting 100 database compounds and using a set of five randomly chosen active molecules as baits (that is, in one of a hundred individual search calculations). Seven of the eight hits have cyclic carbon scaffolds that do not occur in the reference molecules and thus represent true "scaffold hops".

al search calculations on these challenging classes that did produce hits were useful to further analyze the ability of the different search strategies tested herein to identify structurally diverse active compounds, as discussed below.

Comparison of multiple-template search methods

After analyzing the search performance of different fingerprints, we compared the multiple-template search techniques. Except for PDR-FP, for which only the frequency method could be used given its design, each fingerprint was applied in combination with the centroid, 1-NN, and 5-NN techniques, and the approach that produced the best overall results was selected for comparison. Our calculations revealed that almost all fingerprints displayed a preference for nearest-neighbor searching. Only MPMFP performed better in centroid searches for four of nine classes. Similar statistical preferences of nearest-neighbor searching were previously observed by others using different test cases.[7,24] We went a step further and attempted to relate the observed preferences for nearest-neighbor searching to structural features of test molecules. Therefore, we analyzed individual search calculations on high-diversity classes that produced hits with one or more methods and compared the structures of reference compounds and hits. Owing to the small number of hits identified in these calculations, structural similarity between hits and baits could be studied in detail and quantified using MACCS *Tc* calculations. We found that nearest-neighbor searches displayed a general tendency to produce hits that closely resembled single template molecules, especially for complex fingerprint designs, whereas hits identified using centroid or frequency searching were less similar to reference molecules. Figure 6 shows a representative example in which individual trials with the same bait set produced unique hits with different fingerprints in combination with the 1-NN, centroid, and frequency methods. The hit identified using the 1-NN closely resembles the structure of one of the reference compounds. By contrast, the hit identified with the PDR-FP frequency method is dissimilar to the baits and represents a different scaffold.

How can one rationalize such findings? Nearest-neighbor searching separately calculates the Tanimoto similarity of a database compound against all reference structures. Thus, the information provided by multiple templates is not used as a whole, but at the level of individual molecules, which favors specific compound-to-compound matches over the abstraction of activity-relevant features from structurally diverse templates. Therefore, nearest-neighbor calculations improve the probability of recognizing compounds that are similar to single reference molecules, which often provides a statistical advantage in identifying hits, but decreases the probability of recognizing structurally diverse molecules.

For compound sets that are rich in unique core structures, such as those reported herein, we would expect that the extraction of common molecular features through fingerprint calculations is more difficult than the recognition of hits using single bait molecules. Thus high-complexity fingerprints such as Molprint2D that evaluate compounds at high resolution will be increasingly difficult to use in combination with the centroid method, but will benefit from nearest-neighbor searching

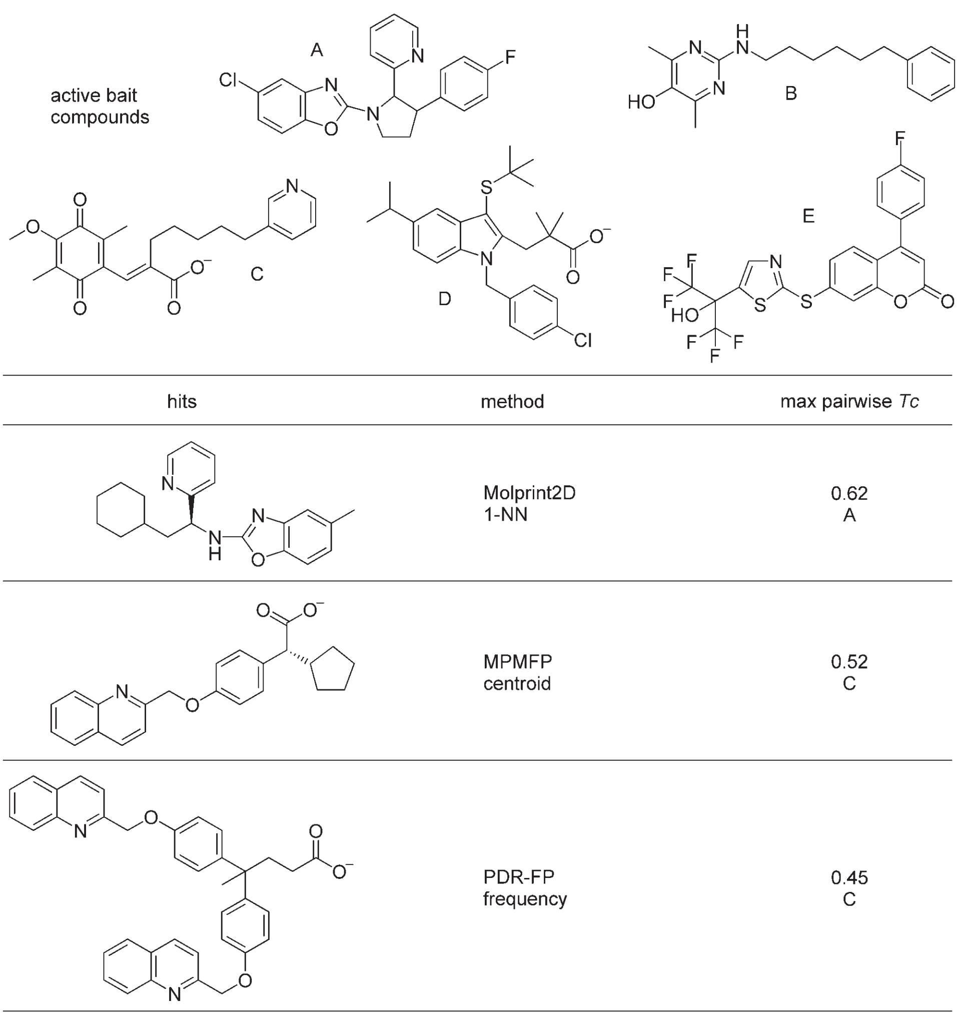

Figure 6. Structural similarity between reference molecules and hits identified with different methods. For a single bait set of activity class LSI, hits are shown that were uniquely identified with different combinations of fingerprints and multiple-template search methods. For each hit, the MACCS *Tc* value to the most similar bait compound A–E (max pairwise *Tc*) is reported.

in which most similar compounds can be selected. In contrast, fingerprints such as MPMFP, which, by design, scan molecules at medium resolution, are less affected by structural diversity of baits and more likely to benefit from averaging procedures, which is consistent with the observed preference of this fingerprint for the centroid method in several of the test cases studied herein.

When examining structural similarities between baits and hits, overall most distant similarity relationships were identified using PDR-FP (see examples in Figures 5 and 6). In Figure 5, hits that were uniquely identified with PDR-FP had consistently lower maximum MACCS *Tc* similarity to any of the reference molecules than those also identified with other methods. The ability of PDR-FP and the frequency approach to recognize more diverse structures having similar activity than other fingerprints is a direct consequence of the fact that PDR-FP combines information from all reference molecules for similarity searching and, different from the centroid approach when applied to conventional 2D fingerprints, takes the database descriptor distributions into account in order to identify highly activity-selective molecular features.

Conclusions

In the study reported herein, we investigated similarity searching using multiple reference compounds on activity classes, which, by design, consisted exclusively of molecules with different core structures. The analysis of these classes has pushed fingerprint search calculations to the limit. The results obtained for classes of increasing structural diversity highlighted differences in the performance of fingerprints of different design. In addition, our calculations revealed overall preferences for nearest-neighbor search methods over fingerprint-averaging procedures at best performance levels. Through a detailed analysis of individual search calculations, we could relate such statistical preferences to similarities between reference compounds and hits. Our findings suggest that search calculations that focus on nearest neighbors of individual templates followed by data fusion have less potential to identify diverse structures having similar activity than methods that use multiple compound information as a whole, such as the PDR-FP frequency approach.

Keywords: biologically active compounds · cheminformatics · drug discovery · in silico screening · similarity searching

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

[4] P. Willett, *J. Med. Chem.* **2005**, *48*, 4183 – 4199.

[5] F. L. Stahura, J. Bajorath, *Curr. Pharm. Des.* **2005**, *11*, 1189 – 1202.

[6] N. E. Shemetulskis, D. Weininger, C. J. Blankley, J. J. Yang, C. Humblet, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862 – 871.

[7] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391 – 405.

[8] L. Xue, J. W. Godden, F. L. Stahura, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218 – 1225.

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

[10] J. Hert, P. Willett, D. J. Wilton, *J. Chem. Inf. Model.* **2006**, *46*, 462 – 470.

[11] M. Whittle, V. J. Gillet, P. Willett, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840 – 1848.

[12] Molecular Drug Data Report (MDDR), MDL Information Systems Inc., San Leandro, CA (USA) **2005** (http://www.mdl.com).

[13] L. Xue, J. Bajorath, *J. Mol. Model.* **1999**, *5*, 97 – 102.

[14] MACCS structural keys, MDL Information Systems Inc., San Leandro, CA (USA) **2005** (http://www.mdl.com).

[15] Y. L. Xu, M. Johnson, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912 – 926.

[16] Meqi, Pannanugget Consulting LLC, Kalamazoo, MI (USA) **2006** (http://www.pannanugget.com).

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

[18] BCI, Version 7.0.1, Digital Chemistry Ltd., Leeds, (UK) **2006** (http://www.digitalchemistry.co.uk).

[19] A. Bender, Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170 – 178.

[20] A. Bender, Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708 – 1718; (also see: http://www.molprint.com).

[21] R. P. Sheridan, M. D. Miller, D. J. Underwood, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128 – 136.

[22] Molecular Operating Environment (MOE), Chemical Computing Group Inc., Montreal, QC (Canada) **2005** (http://www.chemcomp.com).

[23] L. Xue, J. W. Godden, F. L. Stahura, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151 – 1157.

[24] H. Eckert, J. Bajorath, *J. Chem. Inf. Model.* **2006**, *46*, in press.

[25] H. Eckert, J. Bajorath, *J. Med. Chem.* **2006**, *49*, 2284 – 2293.

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

[27] D. R. Flower, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379 – 386.

[28] G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem.* **1999**, *111*, 3068 – 3070; *Angew. Chem. Int. Ed.* **1999**, *38*, 2894 – 2896.

[29] Compound data sets used in this study can freely be obtained through the following URL: http://www.b-it-center.de/Wob/en/view/class211_id675.html.

Received: September 18, 2006
Revised: October 31, 2006
Published online on December 4, 2006