Design and Evaluation of a Novel Class-Directed 2D Fingerprint to Search for Structurally Diverse Active Compounds

Hanna Eckert and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

Received July 18, 2006

Recent attempts to increase similarity search performance using molecular fingerprints have mostly focused on the evaluation of alternative similarity metrics or scoring schemes, rather than the development of new types of fingerprints. Here, we introduce a novel 2D fingerprint design (property descriptor value range-derived fingerprint or PDR−FP) that involves activity-oriented selection of property descriptors and the transformation of descriptor value ranges into a binary format such that each fingerprint bit position represents a specific value interval. The design is tailored toward multiple-template similarity searching and permits training on specific activity classes. In search calculations on 15 compound classes of increasing structural diversity, the PDR fingerprint performed better than other state-of-the-art 2D fingerprints. Among the structurally diverse classes were six compound sets with peptide character, which represent a notoriously difficult chemotype for 2D similarity searching. In these cases, PDR−FP produced promising results, whereas other fingerprint methods mostly failed. PDR−FP is specifically designed for search calculations on structurally diverse compounds, and these calculations are not influenced by molecular size effects, which represent a general problem for similarity searching using bit string representations.

1. INTRODUCTION

Chemical similarity searching[1,2] is a long-established yet active area of research that has its foundations in the similar property principle[3] formulated in the late 1980s. One of the reasons for the current interest in similarity searching is its widespread application in virtual screening, where reference structures with known bioactivity are used as templates to mine large compound databases for novel active molecules.[4,5] Molecular fingerprints are among the most popular similarity search tools because they transform chemical compounds into a linear format, usually a bit string, and thus permit computationally efficient compound comparisons. Fingerprints have the attractive feature that they can be applied in situations where only single reference structures are available, which sets them apart from compound classification or machine learning techniques that rigorously depend on the availability of sets of multiple active compounds.[5] However, a number of investigations have shown that fingerprint search performance usually further increases when multiple instead of single active compounds are employed.[6]

Despite the variety of available fingerprint designs,[2,5] 2D fingerprints are often the method of choice for ligand-based virtual screening because of their computational efficiency and independence of hypothetical compound conformations. In intuitive fingerprint designs,[7−9] each bit accounts for the presence or absence of a specific molecular feature. Popular among these "keyed" representations are fragment-based fingerprints such as publicly available MACCS keys[8,10] that consist of 166 structural fragments (and bits). By contrast, in more complex designs such as Daylight fingerprints[11] that typically consist of 2048 bits, different molecular features are mapped to overlapping bit segments. These fingerprints are "hashed"[11] and produce highly characteristic bit patterns but are less intuitive than keyed designs because single bit positions can no longer be associated with specific compound features. Going beyond 2D fingerprints, the currently most complex representations are 3D pharmacophore fingerprints.[12−14] They consist of up to several million bits and monitor the presence or absence of specific pharmacophore arrangements[13] in molecules that are explored by exhaustive conformational search calculations.

In situations where only one active reference compound is available, the bit string representations of the template and each database compound are compared using various similarity metrics,[1] the most prominent being the Tanimoto coefficient (Tc).[1] However, as mentioned above, search performance can significantly increase if more than one reference structure is available. Several approaches have been introduced to utilize multiple reference compounds in fingerprint searching including the calculation of consensus fingerprints,[15,16] fingerprint profiling[17] and scaling,[18,19] nearest neighbor methods,[16,20] or a centroid technique.[20] The latter method adopts the generation of protein sequence profiles[21] for detecting distantly related sequences for chemical similarity searching. An average fingerprint vector is calculated for all template structures and compared to database compounds using the general formulation of the Tc for numerical values[1] instead of binary vectors. The centroid approach and also consensus fingerprints[15,16] or fingerprint scaling[18,19] emphasize bit positions that are conserved in a set of active compounds and hence are likely to account for activity-relevant features. In contrast, the nearest neighbor method belongs to the category of data fusion techniques[6] and

* Corresponding author tel.: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci600303b CCC: $33.50
Published on Web 10/12/2006

separately calculates the Tanimoto similarity of a database compound against all reference structures. Then, the similarity scores of the *k* nearest neighbors are averaged (*k*-NN or SUM fusion rule)[16,20] or, alternatively, only the maximal Tc value (1-NN or MAX fusion rule)[16,20] is used instead, which often produces the overall best results.[16,20,22] However, data fusion has the potential drawback that the information provided by multiple templates is not utilized as a whole, which hinders the abstraction of activity-relevant features from structurally diverse compounds. However, several studies have shown that the application of nearest neighbor techniques can further increase statistical search performance compared to the centroid or other weighted methods when using conventional 2D fingerprint designs.[16,20] In summary, recent attempts to increase fingerprint search performance have largely focused on the evaluation of alternative scoring schemes, rather than the development of new types of fingerprints.

In light of this situation, we have concentrated on the design of a new fingerprint that is especially suited for identifying and emphasizing activity-class-specific features and that can be trained on multiple reference structures to enable the recognition of structurally diverse active compounds. The recognition of diverse structures, or different chemotypes, having similar activity, a process often referred to as lead hopping,[23] represents one of the primary goals of ligand-based virtual screening.[5] Initially, a critical question has been which types of molecular properties to consider, and we decided to omit structural fragment-type descriptors, despite their popularity and significant predictive value, and exclusively focus on molecular property descriptors[24] because they typically put less emphasis on structural resemblance. The feasibility of this strategy was suggested by our previous observations that activity-selective value ranges of property descriptors could be determined for many compound classes.[25]

Our fingerprint design involved three stages, (1) the selection of property descriptors on the basis of a comparison of their value distribution in different classes of active compounds and a large screening database, (2) value range encoding of each selected descriptor, and (3) activity-oriented bit frequency analysis. Thus, for similarity searching, bit patterns of multiple reference structures were combined in order to identify significant deviations between descriptor value distributions in active and database compounds. This information was then used to emphasize specific bit positions when searching for novel active compounds. Because the value range of each selected descriptor was divided into a number of intervals to which a corresponding number of bits was assigned, this fingerprint type is designated "*P*roperty *D*escriptor value *R*ange-derived *F*inger*p*rint" (PDR−FP). This new fingerprint format conceptually differs from currently available 2D fingerprint and also three-dimensional pharmacophore-type fingerprints.

PDR−FP was tested on a number of activity classes of increasing structural diversity and compared to other 2D fingerprints employing different search techniques. Its performance was best on structurally diverse classes. Among these classes were peptidelike molecules, which provided a rather challenging search scenario because, in these cases, the results of 2D similarity searching were previously found to be dominated by the amide backbone of these molecules and heavily biased toward the recognition of other peptides, regardless of their biological activity.[26] However, when applying PDR−FP to search for compounds belonging to six peptidelike classes with different activities, using only five active reference structures, promising results were obtained, in contrast to other 2D fingerprint methods.

2. METHODOLOGY

2.1. PDR−FP Design. A pool of 184 1D and 2D descriptors implemented in the Molecular Operating Environment (MOE)[27] provided the basis for the design of PDR−FP. To identify descriptors having a tendency to systematically respond to activity-relevant molecular features, while keeping the design as simple as possible, three rounds of descriptor selection were carried out. First, the descriptor scoring function of the previously reported DynaMAD algorithm[28] was applied to preselect descriptors from the pool that displayed a detectable tendency to respond to features of compounds belonging to different activity classes. The DynaMAD scoring function was designed to relate descriptor scores to mapping probabilities *P* of their activity class value ranges (i.e., the probability of database compounds to map a given range):

$$\text{score} = [1 - P(\text{classMin} \leq X \leq \text{classMax})]100$$

In this formulation, classMin is the minimum value seen within a class and classMax the maximum value. This scoring function produces scores between 0 (corresponding to no selectivity) and 100 (optimal selectivity).

Importantly, during the first round, we intended to search for descriptors that would not only be sensitive to one or a few activity classes but to a panel of different ones. This means we were searching for descriptors displaying a general tendency to distinguish active molecules from other database compounds. Therefore, for each descriptor, the score was calculated for a set of 26 different compound activity classes[9,28] summarized in Table 1 that were added to a background database containing more than 1.4 million molecules (2D-ZINC; described in section 2.4.1). Descriptors consistently having a score < 50 for all 26 classes were discarded. These descriptors adopted value ranges in given sets of active compounds that were matched by more than 50% of the background database molecules and thus show no class specificity at all. Following the first selection step, 131 property descriptors remained for further analysis. In the second round, 18 descriptors were removed that were not suitable for value range subdivision because of significantly skewed or erratic value distributions in 2D-ZINC. All of these descriptors were also found to have low scores for most of the activity classes. In the final round, strongly correlated descriptors were identified by the calculation of a matrix of pairwise correlation coefficients, and from each pair of highly correlated descriptors, the one having the lower overall score was removed. We attempted to control descriptor correlation effects in order to balance the composition of PDR−FP. Correlation analysis led to the removal of 20 more descriptors. Therefore, the final selection set included a total of 93 descriptors that are listed in Supplementary Table 1 (Supporting Information). These descriptors provided the basis for the generation of PDR−FP.

Table 1. Activity Classes Used for Descriptor Score Calculation[a]

biological activity	NC
angiotensin II antagonists	45
antiadrenergic (β-receptor)	16
antihypertensive (ACE inhibitors)	17
benzodiazepine receptor ligands	22
carbonic anhydrase II inhibitors	22
cholesterol esterase inhibitors	30
cyclooxygenase-2 (Cox-2) inhibitors	17
dihydrofolate reductase inhibitors	30
dopamine D1 agonists	30
endothelin antagonists	32
estrone sulfatase inhibitors	35
factor Xa inhibitors	14
glucocorticoid analogues	14
histamine H3 antagonists	21
HIV protease inhibitors	18
inosine monophosphate dehydrogenase inhibitors	35
κ agonists	25
β-lactamase inhibitors	29
LDL receptor upregulators	30
melatonin agonists	25
protein kinase C inhibitors	15
serotonin receptor ligands	21
thiol protease inhibitors	34
thromboxane antagonists	33
tyrosine kinase inhibitors	20
xanthine oxidase inhibitors	35

[a] For the PDR−FP design process, 26 compound activity classes[9,28] were used to evaluate the suitability of descriptors, as described in the text. "NC" stands for the number of compounds.

For the transformation of property descriptors into a binary format, the value range of each descriptor was divided into nonoverlapping intervals such that the *frequency* of database compounds falling into each interval was the same. We call this process *equifrequent binning*. To limit the size of PDR−FP to approximately 500 bits, we attempted to use six bins per descriptor, which could be done for 66 of the 93 descriptors. Depending on their individual value distributions, other descriptors were equifrequently binned into two, three, four, five, or seven intervals (for two, eleven, five, eight, or one descriptor, respectively). Interval boundaries for all descriptors are reported in Supplementary Table 2 (Supporting Information). Because one bit position is assigned to each interval, PDR−FP consists of a total of 500 bits.

It should be noted that interval boundaries were determined using the value distributions in 2D-ZINC. If another source database would be used for virtual screening, these interval boundaries might not lead to an exact equifrequent binning. However, such discrepancies are usually small for large numbers of database compounds and have no significant effect on the results. In addition, interval boundaries can be easily readjusted to obtain descriptor intervals with equal frequencies for a variety of databases.

2.2. Compound Representation. To generate the binary representation of a compound, its values for the 93 PDR−FP descriptors are calculated, and for each descriptor (represented by n bits), it is determined into which of the predefined n intervals the compound descriptor value falls. The associated bit is then set to 1 (on); all other $n-1$ bits are set to 0 (off). When this scheme is followed, the bit string representation of an arbitrary compound has exactly 93 bits that are set on. As an example, the descriptor "diameter" (reporting the largest value in the distance matrix of a compound) is encoded by five bits with corresponding five intervals [0,10], [11,12], [13,13], [14,15], and [16,∞]. If a compound has a diameter value of nine, its bit string representation is *10000* because nine matches the first interval [0,10] assigned to the first diameter bit. If the descriptor value for the diameter was 14, the resulting bit string representation would thus be *00010*.

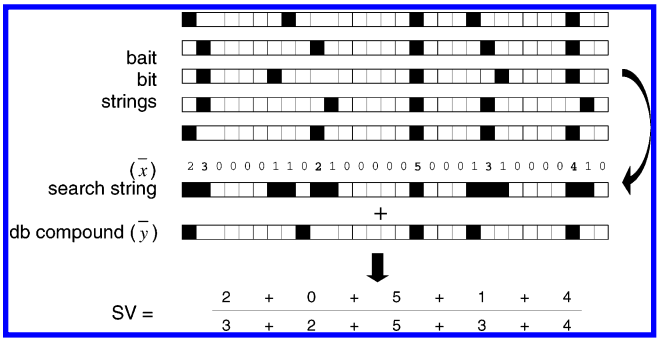


Figure 1. Multiple-template similarity searching using PDR−FP. The schematic illustrates the specifics of the PDR−FR approach. For a bait set consisting of five compounds, an activity-oriented search string $\bar{x}$ is generated by summing up the bit frequencies in the bait bit strings. To calculate the similarity value (SV) of a database compound, the dot product of the search string $\bar{x}$ and the compound bit string $\bar{y}$ is divided by a normalization factor. This factor is the sum of the maximum frequency values per descriptor (shown in bold in the search string).

2.3. Similarity Assessment. For multiple active reference compounds (baits), the bit string representations are combined to create an *activity-oriented search string* (vector $\bar{x}$) that results from the bit frequency (the sum over all bits) at each fingerprint position, as illustrated in Figure 1. Thus, the *activity-oriented search string* is no longer a binary representation but reflects the value distribution of each descriptor in the bait set and enables the identification of significant deviations with respect to descriptor values occurring in the database. Importantly, because of our e*quifrequent binning* procedure, the database value distribution is implicitly accounted for in PDR−FP. Therefore, if all bits of a descriptor have the same frequency in a bait set, the descriptor displays no selectivity with respect to the particular class under investigation, because it has an identical value distribution in the database and the active compound set. By contrast, if the frequency of descriptor bits varies and is predominantly focused on only one or two adjoining bits, the descriptor has a highly activity-selective setting. Thus, the generation of an *activity-oriented search string* corresponds to activity-class-oriented *training* of PDR−FP on the basis of a particular bait set.

The similarity value for a database compound is determined from its individual bit string representation (vector $\bar{y}$) and the *search string* $\bar{x}$. If the database compound matches the fingerprint positions with high-frequency values in the bait set, it displays the activity-specific descriptor setting and should achieve a high similarity value. Otherwise, the similarity value should be low. This is accomplished by summing up all frequency values in the search string $\bar{x}$ that correspond to bits set on in the compound bit string $\bar{y}$. The following expression is the equivalent vector-theoretic notation (dot product of $\bar{x}$ and $\bar{y}$) and represents the similarity coefficient for PDR−FP calculations (termed PDR coefficient). It applies an additional normalization factor NF to place the resulting similarity value SV in the interval [0, 1]:

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

The normalization factor NF is defined as the sum of the maximal bit frequency values occurring in the bait set for each of the 93 descriptors. Let values$_j(\bar{x})$ be the values of the search string $\bar{x}$ that are assigned to descriptor j, then NF is calculated as

$$NF = \sum_{j=1}^{93} \max[\text{values}_j(\bar{x})]$$

Thus, a database compound achieves the maximal similarity value of 1 if it matches the bit with the highest frequency in the bait set for every descriptor. By contrast, a database compound is assigned the similarity value 0 if it has for every descriptor a bit set on that is set off in all of the bait compounds.

The strategy to determine an activity-class-specific search string for multiple templates is related to the calculation of consensus fingerprints,[15] scaling factors,[17,18] or centroid fingerprints.[20] These approaches have in common that bit frequencies in known active compounds are used to assign weighting factors to different bit positions. The similarity coefficient introduced here to relate compound bit strings to class-directed search strings is only applicable to the PDR−FP design. However, the results correspond to those produced when evaluating fingerprint matches to a centroid fingerprint using the general formulation of Tc. This means that identical similarity rankings of database compounds relative to a bait set are produced, even though the actual values of these coefficients differ. The reason for this is that the number of bits set on in PDR−FP is constant (i.e., 93) for every database compound so that the general Tc only depends on the dot product between the bit string of a database compound and the centroid. For multiple-template searching using PDR−FP, the coefficient introduced here is applied because it is straightforward to interpret with respect to the comparison of value distributions between a database and an active compound set. In addition, the PDR coefficient generates similarity values between 0 and 1 for every search string, whereas for centroid fingerprints, calculation of the general Tc does not produce scaled similarity values. For the classes and fingerprint designs studied here, for example, top Tc values of only ∼0.5 were obtained on average.

2.4. Calculations. 2.4.1. Activity Classes and Database Compounds. As the source database for our studies, we generated a "2D-unique" version of the publicly available ZINC[29] database that consists of ∼2.01 million molecules in predicted 3D conformational states. For each ZINC compound, we calculated the values for our pool of 184 1D and 2D descriptors and identified compound entries with duplicate descriptor settings. Only one of these entries was retained, which led to the removal of ∼0.57 million ZINC compounds. Thus, this subset of ZINC that we call 2D-ZINC contains ∼1.44 million molecules. In our calculations, all 2D-ZINC compounds were considered inactive and thus potential false positives, although it is anticipated that the database might contain hits for at least some of the activity classes we studied here.

The PDR−FP performance was evaluated on a total of 15 different activity classes that we separated into two subsets. The first set includes nine classes with an increasing degree of intrinsic structural diversity that are reported in Table 2. Three of these classes display low structural diversity and were originally assembled from the literature for partitioning analysis.[30] The six remaining classes of molecules were extracted from the Molecular Drug Data Report[31] (MDDR). To ensure that these classes were structurally diverse, preselected MDDR compounds sharing the same activity were clustered using the publicly available set of 166 MACCS structural fragment descriptors[10] and a fingerprint clustering routine implemented in MOE that utilizes a similarity matrix for the grouping of compounds. Molecules from nonsingleton clusters were randomly selected in order to avoid the inclusion of analogue series. The degree of structural diversity within each of the nine activity classes was assessed by the systematic pairwise comparison of all compounds using MACCS keys and the calculation of minimum, maximum, and average Tc values. Three of these classes displayed average Tc values between 0.6 and 0.7 and maximum Tc values above 0.9 and were classified as low-diversity classes. Three other classes had average Tc values between 0.45 and 0.56 and maximal Tc values in 0.8−0.9 and were considered medium diverse. The remaining three activity classes displayed high structural diversity with average Tc values below 0.44, maximal Tc values below 0.82, and minimum Tc values below 0.14.

The second set of activity classes consisted of six compound sets having peptide character, as summarized in Table 3. Compound classes were considered to be peptidelike if they contained a varying number of amide bonds mimicking a peptide backbone. These classes were also extracted from the MDDR, and compounds were selected, as discussed above. The peptide-reminiscent compound sets display a rather high degree of structural diversity, with average Tc values ranging from 0.46 to 0.58 and no single Tc value above 0.85. For comparison, corresponding Tc statistics on all compound classes were also calculated using PDR−FP and are provided in Supplementary Table 3 (Supporting Information). Figure 2 shows representative examples for each of the 15 activity classes studied here, illustrating the presence of different degrees of structural diversity and peptide character.

2.4.2. Virtual Screening Trials and Performance Measures. For multiple-template similarity searching, 100 sets of five compounds each were randomly selected from each activity class and taken once as the "bait set" for the determination of the activity-oriented search string. The remaining compounds (between 10 and 43) were added to the database as potential hits, while all other database compounds were considered inactive. Thus, for each activity class, 100 different search calculations were carried out, and database compounds were ranked according to their similarity values. Search performance was analyzed by calculating cumulative recall rates that report the number of active molecules among the top-ranked 10, 50, and 100 database compounds averaged over all 100 search calculations.

To compare the performance of PDR−FP in virtual screening trials with other methods, reference calculations were carried out with four conceptually different 2D fingerprints: BCI[32,33] on the basis of its standard fragment

Table 2. Activity Classes with Increasing Degree of Structural Diversity[a]

code	biological activity	NC	min Tc	max Tc	av Tc	stdDev Tc	structural diversity
ACE	antihypertensive (ACE inhibitors)	15	0.427	0.960	0.700	0.123	low
COX	cyclooxygenase-2 (Cox-2) inhibitors	16	0.409	0.984	0.675	0.143	low
HIV	HIV protease inhibitors	15	0.420	0.922	0.638	0.132	low
BK2	bradykinin BK2 antagonists	22	0.257	0.897	0.558	0.114	medium
ETA	endothelin ETA antagonists	28	0.184	0.795	0.480	0.119	medium
SQS	squalene synthetase inhibitors	29	0.222	0.855	0.453	0.131	medium
GLU	glucagon receptor antagonists	33	0.135	0.788	0.437	0.128	high
ULD	upregulators of LDL Receptor	21	0.100	0.758	0.427	0.157	high
SQE	squalene epoxidase inhibitors	25	0.067	0.815	0.395	0.162	high

[a] Reported are nine activity classes used as test set 1 in our studies, their class codes ("code"); number of compounds per class ("NC"); intraclass similarity statistics, "min Tc", "max Tc", and "av Tc", which stand for the minimum, maximum, and average Tc value, respectively, produced in complete pairwise comparisons with MACCS structural fragment-type fingerprints;[10] "stdDev" reports the standard deviation of the Tc values. Taken together, these values indicate the degree of structural diversity within each class (low, medium, or high). Classes are sorted by increasing diversity.

Table 3. Activity Classes with Peptide Character[a]

code	biological activity	NC	min Tc	max Tc	av Tc	stdDev Tc
CAM	cell adhesion molecule antagonists	25	0.185	0.756	0.478	0.098
F7I	factor VIIa inhibitors	23	0.202	0.847	0.464	0.115
GLY	glycoprotein IIb/IIIa receptor antagonists	25	0.308	0.836	0.572	0.106
NK2	neurokinin NK2 antagonists	25	0.353	0.786	0.583	0.083
REN	renin inhibitors	22	0.337	0.787	0.560	0.096
TBI	thrombin inhibitors	48	0.140	0.794	0.526	0.122

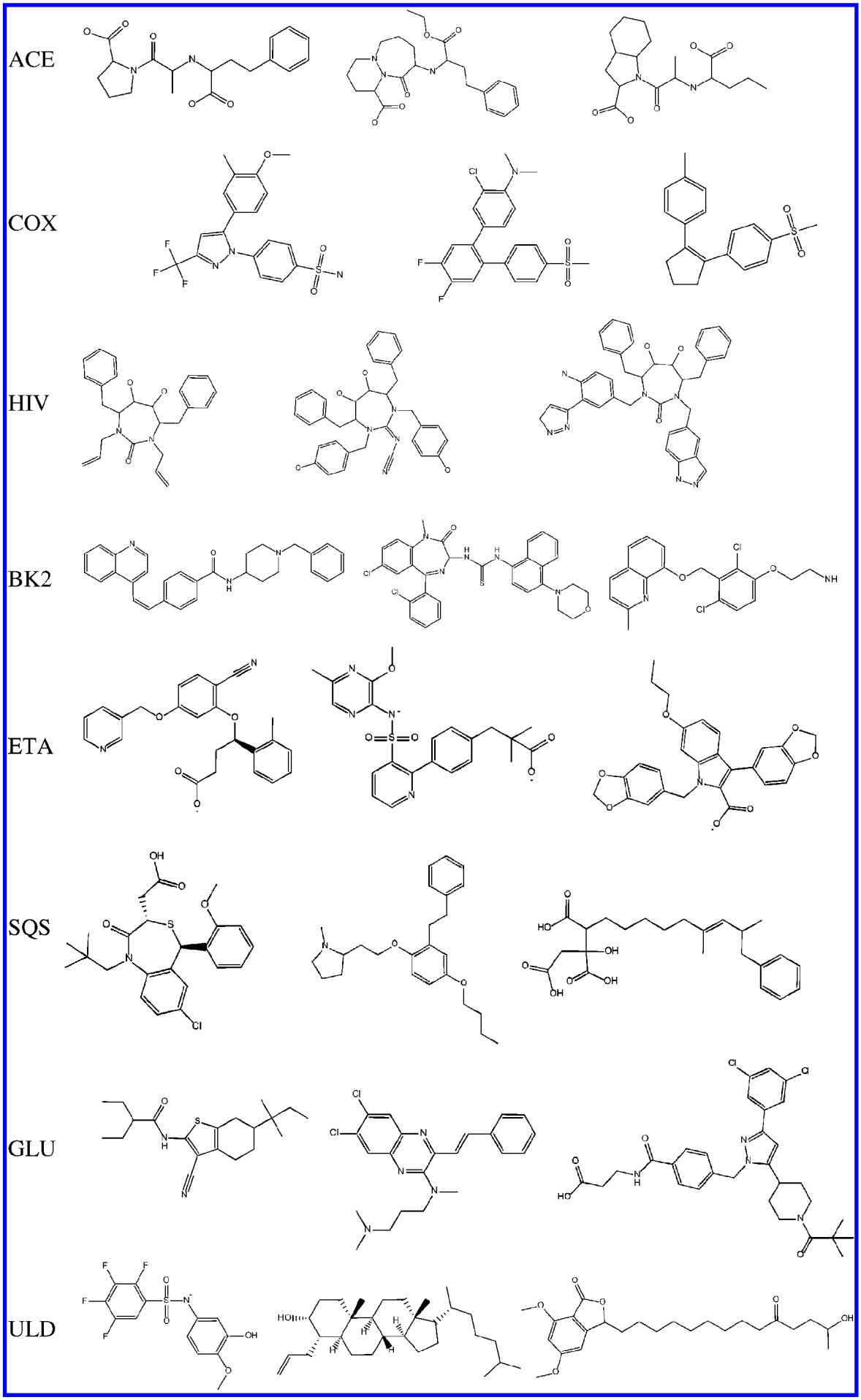
[a] Activity classes with varying peptide character used as test set 2 in simulated virtual screening trials. Abbreviations are used as in Table 2. The Tanimoto similarity statistics confirm that these peptidelike classes are structurally diverse.

dictionary consisting of 1052 bits; the publicly available version[34] of MOLPRINT 2D[35,36] derived from atom environments of the connectivity table of a compound; GpiDAPH3,[27] a three-point pharmacophore-based fingerprint calculated from the 2D molecular graph representation consisting of ~262 000 bits; and TGD,[27] a two-point pharmacophore fingerprint also calculated from the 2D molecular graph representation and consisting of 735 bits. MOLPRINT 2D combines strings of varying sizes, each representing a particular atom environment. The total number of potential strings is difficult to estimate but could reach on the order of 2^{50}. Multiple-template searches with reference fingerprints were carried out applying the 1-NN nearest neighbor method and the centroid approach. The latter technique was also selected because its averaging procedure resembles previously reported fingerprint weighting or scaling techniques.[15,17,18] For comparison with PDR−FP, results of the best-performing search method were used in each case.

3. RESULTS

3.1. PDR−FP Performance and Comparison to Reference Methods. Results of the test calculations on the first set of nine activity classes are reported in Table 4. In each case, five compounds were taken as search templates and between 10 and 28 active compounds were available as potential hits within ~1.44 million 2D-ZINC compounds. Recall rates are reported for selection sets of increasing size. The reference methods displayed a clear preference for the 1-NN search method over the centroid approach, which is well in accord with previous observations made by others.[16,20] Taking the largest set of 100 compounds (i.e., ~0.007% of 2D-ZINC) as a reference point, the overall performance of the five different fingerprint search methods compared here varied across the activity classes and displayed a clear tendency to depend on the degree of structural diversity. Consistently high recall rates were obtained for the three classes of low structural diversity (ACE, COX, and HIV). For ACE, all methods reached recall rates between 80% and 93% and MOLPRINT 2D performed best. In contrast, BCI was the best method for COX and HIV. With the exception of TGD, which produced recovery rates between 42% and 47%, overall similar rates between 72% and 88% were observed for all search tools and methods including PDR−FP, although it does not contain any structural fragment descriptors.

Next, we considered the search results for classes of medium structural diversity (BK2, ETA, and SQS). Here, the search performance of all methods decreased, but PDR−FP produced the overall highest recall rates between 25% and 38%, although differences to reference methods were subtle for class BK2. For the remaining three activity classes of high structural diversity (GLU, ULD, and SQE), this trend became more distinct. On high diversity classes, PDR−FP produced recall rates between 23% and 38% comparable to those obtained for the medium diversity classes, while recall rates for reference methods were further reduced. For GLU and ULD, reference methods recovered only between 4% and 14% of the potential hits. Similar differences were observed for smaller selection sets of 50 or 10 compounds (Table 4). Thus, on structurally homogeneous compound sets, most search tools produced very good results, but for classes of increasing structural diversity, PDR−FP produced gener-


ACE
COX
HIV
BK2
ETA
SQS
GLU
ULD

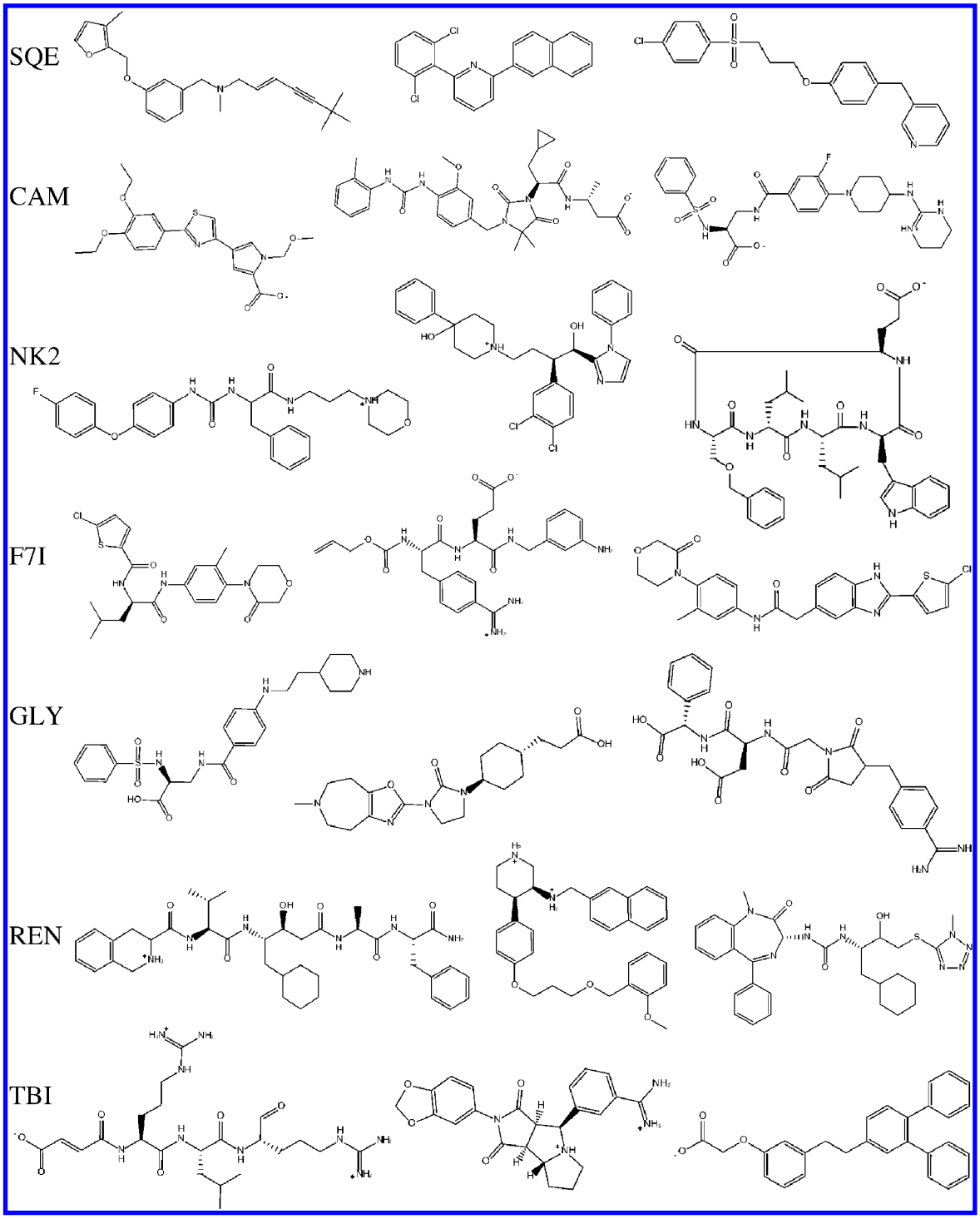


Figure 2. Representative structures. For each of the 15 activity classes studied here, three representative structures are shown side-by-side.

ally superior results. These findings are thought to be a consequence of the class-directed training capacity inherent in PDR−FP.

3.2. Activity-Specific PDR−FP Bit Patterns for Classes with Peptide Character. We further investigated the ability of PDR−FP to successfully recognize compounds belonging to structurally diverse classes. The PDR−FP design enables the comparison of descriptor value distributions in active and database compounds by use of bit frequency analysis (see Methods). If descriptor values of active compounds map to a single or a few adjoining bits, their distribution distinguishes active molecules from background compounds. Thus, we analyzed the bit patterns in our second set of six structurally diverse classes with peptidelike character and determined the spread of bits set on in bait sets of five compounds and averaged the values over 100 different trials per class. In principle, the bit spread can range from 1 (when all baits fall into the same interval) to uniform coverage of the descriptor bit segment (representing the entire value range of the descriptor). Table 5 shows that characteristic bit patterns that differed from each other emerged for all six activity classes. Thus, many descriptors encoded in PDR−FP responded to compound class-sensitive features. Averaged over all classes, 31 of the 93 PDR−FP descriptors had an ideal bit spread of 1. On the basis of the PDR−FP design, these bit settings represent consensus positions having the

Table 4. Recall Rates for PDR−FP and Reference Methods[a]

activity class	fingerprint	best approach	RR [%], 10 cpds	RR [%], 50 cpds	RR [%], 100 cpds
ACE	BCI	1-NN	72.05	81.20	83.40
	MOLPRINT 2D	1-NN	76.43	90.62	93.20
	GpiDAPH3	1-NN	83.10	85.80	86.60
	TGD	1-NN	64.60	75.60	79.70
	PDR		71.00	78.73	80.71
COX	BCI	1-NN	69.82	84.91	88.43
	MOLPRINT 2D	1-NN	54.77	69.65	72.29
	GpiDAPH3	1-NN	56.85	82.02	85.41
	TGD	1-NN	36.73	37.64	42.09
	PDR		51.82	67.78	73.05
HIV	BCI	1-NN	80.10	82.50	83.40
	MOLPRINT 2D	1-NN	70.37	75.72	76.30
	GpiDAPH3	1-NN	67.80	72.60	73.00
	TGD	1-NN	34.60	45.60	46.50
	PDR		53.73	71.38	76.67
BK2	BCI	1-NN	15.71	23.94	28.43
	MOLPRINT 2D	1-NN	12.33	20.91	25.78
	GpiDAPH3	1-NN	20.65	29.88	34.75
	TGD	1-NN	7.47	11.31	13.89
	PDR		18.09	31.13	37.76
ETA	BCI	1-NN	3.86	6.59	8.30
	MOLPRINT 2D	1-NN	4.26	8.01	10.12
	GpiDAPH3	1-NN	4.78	5.91	7.70
	TGD	1-NN	6.09	10.82	14.54
	PDR		10.21	19.56	25.25
SQS	BCI	1-NN	14.10	19.12	21.04
	MOLPRINT 2D	1-NN	15.32	18.78	20.32
	GpiDAPH3	1-NN	19.75	25.67	29.83
	TGD	centroid	3.88	5.17	5.71
	PDR		22.40	32.97	37.11
GLU	BCI	1-NN	1.64	3.15	4.26
	MOLPRINT 2D	1-NN	1.96	5.89	7.64
	GpiDAPH3	centroid	5.71	10.07	12.21
	TGD	centroid	4.36	7.98	10.00
	PDR		16.04	27.54	32.72
ULD	BCI	1-NN	0.56	0.90	6.75
	MOLPRINT 2D	1-NN	4.52	10.59	13.70
	GpiDAPH3	1-NN	5.81	9.50	10.25
	TGD	centroid	5.31	7.69	8.62
	PDR		10.77	18.61	22.60
SQE	BCI	centroid	5.55	9.40	11.75
	MOLPRINT 2D	1-NN	8.20	15.60	19.82
	GpiDAPH3	1-NN	11.70	18.75	22.70
	TGD	1-NN	3.20	3.35	3.35
	PDR		23.38	34.63	38.43

[a] "RR" stands for recall rate and "cpds" for the number of compounds per selection set. The "best approach" shows which multiple-template search strategy (1-NN or centroid; see text) produced the best results, as reported in this table. Activity classes are designated according to Table 2.

greatest ability to distinguish between active molecules and background compounds. When a bit spread of maximal 2 is considered, on average, more than half (∼53) of the descriptors displayed this setting. When 13 of 93 PDR−FP descriptors that are implemented using segments of only two or three bits are omitted from this analysis, 44 of the 80 remaining descriptors still displayed activity-sensitive settings. By contrast, descriptors displaying no detectable class sensitivity were rarely seen. As reported in Table 5, only on average four of the 67 PDR−FP descriptors encoded using six or seven bits were found to have a bit spread of six in our bait sets. Thus, almost all PDR−FP descriptors contributed to bit settings that had the potential to discriminate between active and inactive compounds.

3.3. Performance of PDR−FP and Reference Methods on Peptidelike Classes. Results of systematic search calculations on the six activity classes with peptide character are reported in Table 6. On these classes with distinct structural diversity and varying peptide character, PDR−FP was found to consistently perform better than the reference methods and achieved recovery rates between ∼25% and 77%. For reference calculations, the centroid approach was here more often the method of choice than was observed for the other activity classes, but 1-NN still dominated. With the exception of REN (where compound recovery was overall substantially higher than that for the other classes), the discrepancy in performance between PDR−FP and the reference fingerprints was further increased compared to the structurally diverse classes in the first set. For three of the six peptidelike classes, F7I, NK2, and TBI, PDR−FP recall

Table 5. Bit Distribution in Compound Classes with Peptide Character[a]

bit spread	1	2	3	4	5	6	7
CAM	21.2	25.1	18.7	11.0	9.1	7.9	0.0
F7I	25.6	26.5	19.7	11.1	6.3	3.8	0.0
GLY	16.4	25.5	27.3	14.6	7.1	2.1	0.0
NK2	41.9	14.6	13.1	12.1	7.3	4.0	0.0
REN	50.9	15.7	12.3	6.5	4.5	2.7	0.4
TBI	29.2	23.9	17.9	10.2	7.3	4.4	0.1

[a] The "bit spread" reports the number of descriptor bit positions spread out over one to seven adjacent intervals. The "1" means that all active compounds have the same bit set on (consensus position), whereas "6" (or "7") means that bits spanning the entire value range are occupied. For each activity class, the number of descriptors is reported that showed the particular bit spread. Averages were calculated over 100 random sets of five baits.

Table 6. Recall Rates for Peptidelike Activity Classes[a]

activity class	fingerprint	best approach	RR [%], 10 cpds	RR [%], 50 cpds	RR [%], 100 cpds
CAM	BCI	centroid	1.10	2.25	2.70
	MOLPRINT 2D	1-NN	4.05	9.05	11.37
	GpiDAPH3	1-NN	11.10	17.90	20.80
	TGD	1-NN	4.85	7.88	10.40
	PDR		10.50	19.38	24.50
F7I	BCI	1-NN	1.78	1.94	2.39
	MOLPRINT 2D	1-NN	7.12	10.65	12.05
	GpiDAPH3	1-NN	8.94	11.72	14.06
	TGD	1-NN	0.56	1.20	3.55
	PDR		16.38	25.67	28.75
GLY	BCI	centroid	2.30	4.85	5.95
	MOLPRINT 2D	centroid	6.70	14.10	18.05
	GpiDAPH3	1-NN	7.15	10.00	10.95
	TGD	centroid	15.00	18.82	21.05
	PDR		17.56	29.15	34.35
NK2	BCI	1-NN	5.45	6.95	7.35
	MOLPRINT 2D	1-NN	6.85	10.75	11.44
	GpiDAPH3	1-NN	13.15	18.55	21.95
	TGD	1-NN	1.25	1.65	2.00
	PDR		13.65	30.20	41.34
REN	BCI	centroid	12.65	26.24	34.24
	MOLPRINT 2D	1-NN	23.27	36.61	44.58
	GpiDAPH3	1-NN	40.65	58.00	62.53
	TGD	centroid	9.65	19.71	23.88
	PDR		53.20	72.70	76.66
TBI	BCI	1-NN	1.13	3.79	5.31
	MOLPRINT 2D	1-NN	1.78	5.70	8.39
	GpiDAPH3	1-NN	6.12	10.91	13.67
	TGD	centroid	6.77	9.40	10.87
	PDR		17.53	34.17	40.48

[a] Recall rates are reported for similarity searching using five active reference structures. Abbreviations are used according to Tables 3 and 4.

rates were approximately two or three times higher than those for the next best fingerprint (GpiDAPH3). These recall rates at the 30−40% level corresponded to the presence of approximately five to 17 active compounds within the top 100 database molecules, which represents a significant finding for activity classes of high structural diversity. For PDR−FP, we observed a direct correlation between the number of consensus bits per class and recall rates per class; the more consensus bits there were, the higher the recall rates. Thus, consensus bits were highly discriminatory bit settings, as we had anticipated. Figure 3 compares the structures of bait molecules and hits that were identified with PDR−FP and illustrates the level of structural diversity among these compounds.

4. DISCUSSION

4.1. PDR−FP Training Potential. A key aspect of the design of PDR−FP was to encode the ability for training on different activity classes. This was accomplished by the extraction of consensus bit settings from sets of bait molecules. The results of our test calculations confirmed that PDR−FP could be tuned toward the recognition of activity-class-selective features, even when structural diversity among active compounds was high, which was a prerequisite for its lead- or scaffold-hopping potential. We intended to support scaffold-hopping potential by deliberately excluding structural fragment descriptors from PDR−FP. For activity classes containing different core structures, fingerprints that put much emphasis on structural resemblance often have an intrinsic disadvantage, consistent with the results of our test calculations. For the 15 activity classes studied here, PDR−FP was capable of capturing activity-specific properties and successfully utilizing them in similarity searching. Importantly, only five bioactive reference structures per class were sufficient for the determination of many activity-relevant bit positions.

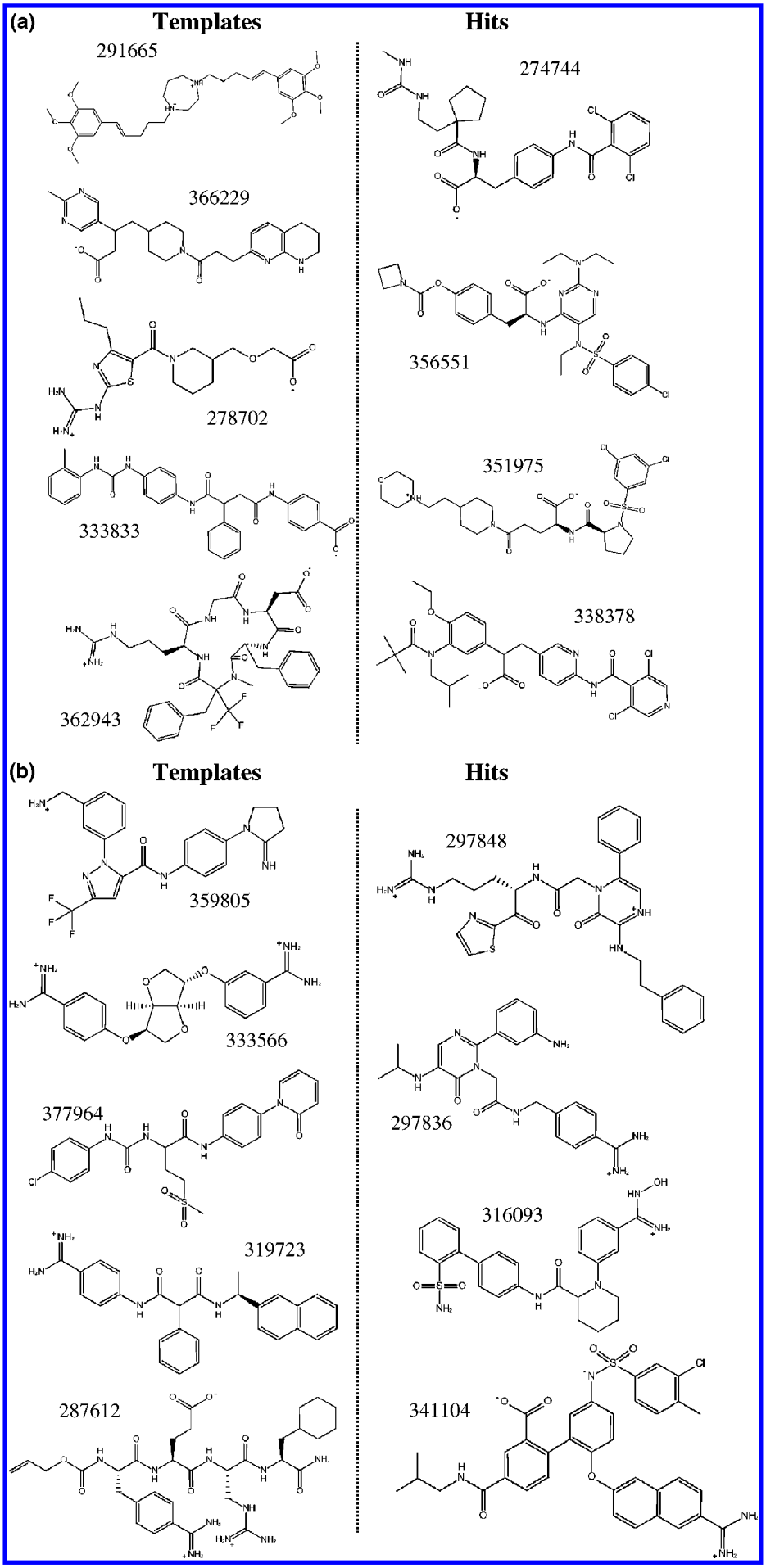
4.2. Focusing on Structural Diversity. Relative fingerprint performance was clearly dependent on the degree of structural diversity among compounds having similar activity. For structural homogeneous classes, reference fingerprints performed as well as PDR−FP or slightly better, but for increasingly structurally diverse classes, PDR−FP consistently performed best and significant differences were observed. Among activity classes of high structural diversity were six sets of compounds with peptide character. We deliberately focused on these classes because their chemical composition is known to create problems for similarity-based methods.[26] Recurrent amide bonds in these structures often favor the recognition of other peptidelike molecules, irrespective of their biological activity. This is reflected by the results we obtained on class REN where all fingerprints produced high recovery rates. This was due to the fact that REN was the most peptidelike class where many compounds shared an amide backbone, despite considerable overall diversity. Thus, REN-based search calculations recovered those compounds with high specificity. In other classes we studied, in particular, CAM and F7I, the extent of peptide character varied strongly from compound to compound, which presented an additional complication. On these six activity classes, PDR−FP consistently produced the best results and differences in performance relative to other fingerprint methods were greatest. We attribute these findings to the property-descriptor-based training effect, consistent with the results of our PDR−FP bit distribution analysis on these classes.

4.3. Multiple-Template Searching. For reference fingerprints, irrespective of their design and complexity, the 1-NN method was mostly superior to the centroid approach. It should be noted that the centroid approach corresponds to the calculation of the similarity coefficient of PDR−FP. However, PDR−FP differs from any of the 2D fingerprints in that it takes descriptor database distributions into account, which provides the basis for its trainability. The overall performance of PDR−FP illustrates the predictive value of activity-class-directed training. On the other hand, the preference of reference methods for 1-NN means that separately utilizing the information provided by each bait compound instead of combining it produced the best results. Thus, in these cases, multiple-template searching is more akin to identifying the most suitable single template. By contrast, the PDR−FP design combines information from different reference molecules, which provides an explanation for its improved scaffold-hopping potential.

4.4. Fingerprint Complexity and Discriminatory Features. Despite its predictive ability, the complexity of PDR−FP is much lower than that of other fingerprints such as GpiDAPH3 or MOLPRINT 2D. It consists only of 500 bits, and compounds are constantly represented by 93 bits set to 1. Each bit that is set on represents exactly one interval of a property descriptor value range matched by a given compound. The technique to divide property descriptor value ranges into intervals with equal relative frequency of occurrence within a source database provides the basis for the training of PDR−FP. Combinations of multiple activity-relevant bit positions and especially consensus positions become highly discriminatory against database compounds because single bits are only matched by a small database fraction (e.g., one-sixth if a descriptor is encoded by six bits). By contrast, in conventional fingerprint designs, bit positions do not take into account information about database value distributions. There is also a general reason to limit the size of fingerprint representations that rely on training and signature bit settings. In complex fingerprints that produce extremely long binary representations of molecules, a limited number of consensus positions from bait sets would have statistically only a minor influence on the resulting similarity values. Importantly, the constant setting of 93 bits in PDR−FP for all compounds, irrespective of their chemical nature, represents a critical advantage compared to conventional fingerprint designs that permit variable bit densities: quantitative similarity calculations become independent of molecular size effects that are known to produce inaccurate Tanimoto similarity values.[37]

5. CONCLUSIONS

With PDR−FP, we have introduced a new fingerprint design that involves the transformation of value ranges of 1D and 2D property descriptors into groups of bits. The fingerprint creates relatively short binary representations consisting of only 500 bit positions. Distributions of database descriptor values are implicitly implemented by equifrequent binning of their descriptor value ranges, which provides the


(a)
Templates
Hits
291665
366229
278702
333833
362943
274744
356551
351975
338378
(b)
Templates
Hits
359805
333566
377964
319723
287612
297848
297836
316093
341104

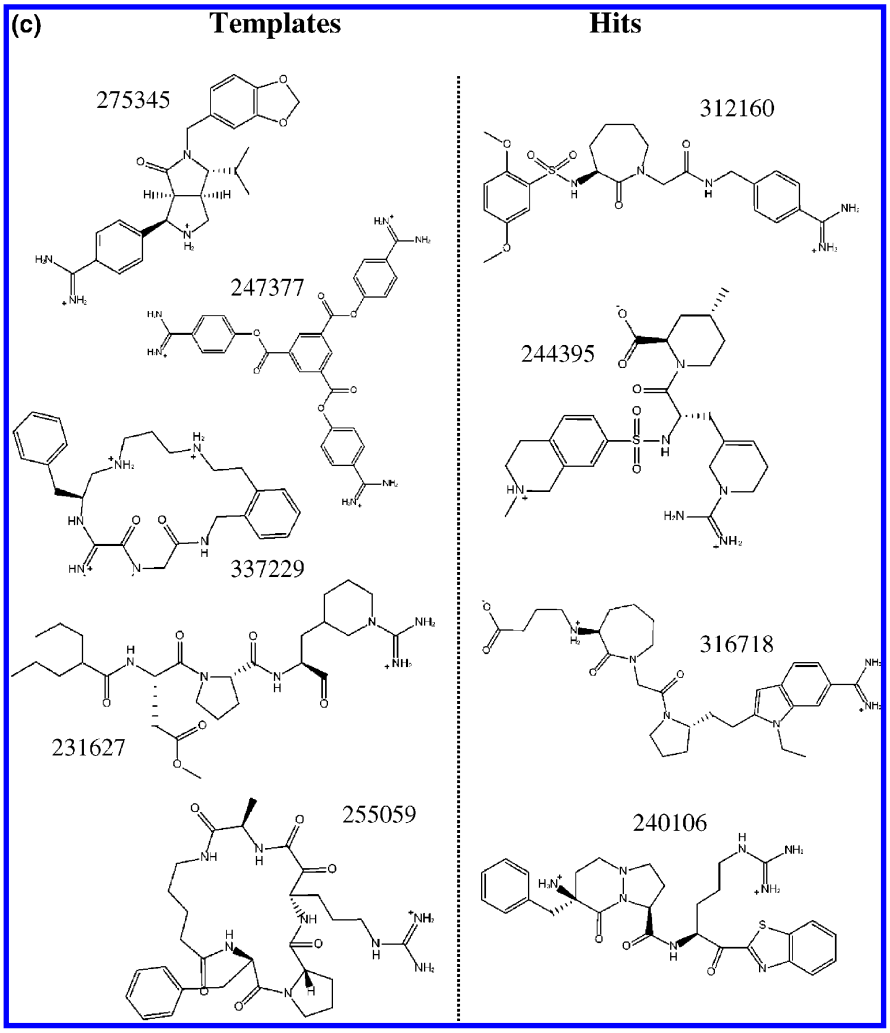


Figure 3. Diverse hits identified in virtual screening trials. For three different activity classes, (a) CAM, (b) F7I, and (c) TBI, examples of hits are shown that were correctly identified with PDR−FP in 2D-ZINC when using five template compounds. Each compound is labeled with its MDDR external registry number.

basis for systematic comparisons of descriptor settings in classes of active compounds and database molecules. Strong deviations are reflected by bit positions having a high frequency among active molecules. Database distributions were derived from a large sample of compounds from medicinal chemistry sources, in total more than 1.4 million molecules, and the resulting binning scheme is likely to be transferable to other source databases without introducing substantial inaccuracies. However, equifrequent binning can be easily carried out for other databases in order to adjust the PDR−FP binning scheme. Similarly, the PDR−FP design can be easily modified to include additional or alternative descriptors. In test calculations on 15 activity classes of increasing structural diversity and different chemical characters, PDR−FP produced promising results and performed better than reference methods when structural diversity was high. For each activity class, PDR−FP training captured a significant number of consensus bit positions. During similarity searching, combinations of activity-sensitive bit settings become highly discriminatory, which is a major determinant of PDR−FP performance. Thus, on the basis of our findings, we conclude that PDR−FP adds a novel prototype to the spectrum of currently available similarity search tools with significant potential for virtual screening applications targeting structurally diverse active compounds.

ACKNOWLEDGMENT

We are grateful to John M. Barnard and Julian Hayward of Digital Chemistry Ltd. for making the BCI fingerprint available to us and Andreas Bender for MOLPRINT 2D.

Supporting Information Available: Encoded descriptors, interval boundaries, and PDR−FP-based Tc statistics of compound activity classes are provided as Supplementary Tables 1−3. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(2) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.

(3) Johnson, M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.

(4) Stahura, F. L.; Bajorath, J. New Methodologies for Ligand-Based Virtual Screening. *Curr. Pharm. Des.* **2005**, *11*, 1189−1202.

(5) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.

(6) Hert, J.; Willett, P.; Wilton, D. J. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46*, 462−470.

(7) Xue, L.; Godden, J. W.; Bajorath, J. Mini-Fingerprints for Virtual Screening: Design Principles and Generation of Novel Prototypes Based on Information Theory. *SAR QSAR Environ. Res.* **2002**, *14*, 27−40.

(8) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(9) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151−1157.

(10) *MACCS Structural Keys*; MDL Information Systems Inc.: San Leandro, CA, 2002.

(11) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: Irvine, CA.

(12) Mason, J. S.; Cheney, D. L. Library Design and Virtual Screening Using Multiple 4-Point Pharmacophore Fingerprints. *Pac. Symp. Biocomput.* **2000**, *5*, 576−587.

(13) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview over the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(14) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A Rapid Computational Method for Lead Evolution: Description and Application to α_1-Adrenergic Antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.

(15) 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.

(16) Hert, J.; Willet, P.; Wilton, D. J. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(17) Godden, J. W.; Stahura, F. L.; Xue, L.; Bajorath, J. Searching for Molecules with Similar Biological Activity: Analysis by Fingerprint Profiling. *Pac. Symp. Biocomput.* **2000**, *5*, 566−575.

(18) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar Activity in Virtual Screening Calculations. *J. Chem. Comput. Sci.* **2001**, *41*, 746−753.

(19) 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.

(20) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Protein. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.

(21) Gribskov, M.; McLachlan, A. D.; Eisenberg, D. Profile Analysis: Detection of Distantly Related Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 4355−4358.

(22) Whittle, M.; Gillet, V. J.; Willett, P. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840−1848.

(23) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead Hopping". Validation of Topomer Similarity as a Superior Predictor of Biological Activities. *J. Med. Chem.* **2004**, *47*, 6777−6791.

(24) Livingstone, D. J. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(25) Eckert, H.; Bajorath, J. Determination and Mapping of Activity-Specific Descriptor Value Ranges for the Identification of Active Compounds. *J. Med. Chem.* **2006**, *49*, 2284−2293.

(26) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. M. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395−1406.

(27) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2005.

(28) Eckert, H.; Vogt, I.; Bajorath, J. Mapping Algorithms for Molecular Similarity Analysis and Ligand-Based Virtual Screening: Design of DynaMAD and Comparison with MAD and DMC. *J. Chem. Inf. Model.* **2006**, *46*, 1623−1634.

(29) 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.

(30) Xue, L.; Bajorath, J. Accurate Partitioning of Compounds Belonging to Diverse Activity Classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757−764.

(31) *Molecular Drug Data Report (MDDR)*; MDL Information Systems Inc.: San Leandro, CA, 2005.

(32) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141−142.

(33) *BCI*, version 7.0.1; Digital Chemistry Ltd.: Leeds, U. K.

(34) *MOLPRINT 2D*. URL for the publicly available molecular fingerprint: http://www.molprint.com (accessed Sept 2006).

(35) Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

(36) Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.

(37) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

CI600303B