DOI: 10.1002/cmdc.200900243

Molecular Fingerprint Recombination: Generating Hybrid Fingerprints for Similarity Searching from Different Fingerprint Types

Britta Nisius and Jürgen Bajorath*[a]

Molecular fingerprints have a long history in computational medicinal chemistry and continue to be popular tools for similarity searching. Over the years, a variety of fingerprint types have been introduced. We report an approach to identify preferred bit subsets in fingerprints of different design and "recombine" these bit segments into "hybrid fingerprints". These compound class-directed fingerprint representations are found to increase the similarity search performance of their parental fingerprints, which can be rationalized by the often complementary nature of distinct fingerprint features.

Introduction

Molecular fingerprints are bit string representations encoding structural, topological, pharmacophore, or property descriptors and are widely used as similarity search tools in chemical database mining and computer-aided hit identification.[1] Reasons for the popularity of fingerprints include their computational efficiency, ease of use, intuitive nature, and often surprising effectiveness in identifying new active molecules.[2] This applies to both 2D and 3D fingerprints that are generated from 2D and 3D molecular representations, respectively.[3] A variety of fingerprint types of different design have been introduced including, for example, substructure,[4] hashed connectivity pathway,[5] extended connectivity,[6] pharmacophore,[7,8] or molecular property fingerprints.[9,10] A systematic comparison of various fingerprint types capturing different aspects of molecular structures has recently been reported.[11]

Keyed 2D fingerprints where each individual bit position is associated with a structural feature, pharmacophore pattern, or property descriptor are particularly intuitive representations because similarity search results can often be chemically interpreted and features that produce bit settings characteristic of different compound activity classes can be identified. This can be accomplished through the application of computational methods such as consensus fingerprinting,[12] fingerprint profile analysis,[13] or reverse fingerprinting.[14]

In recent years, relatively few conceptually new fingerprint designs have been reported,[6,10,15] but much attention has been focused on developing fingerprint search strategies that effectively take into account information from multiple reference compounds. Intensely investigated approaches include consensus fingerprinting,[12] fingerprint averaging (or centroid) techniques,[16] fingerprint scaling,[17] feature scoring[14] and, in particular, data fusion methods such as nearest neighbour calculations.[16,18,19]

In addition to exploring different fingerprint designs and similarity search strategies, efforts in our laboratory have also focused on fingerprint reduction approaches in order to generate minimal fingerprint representations retaining high search performance. For this purpose, methods such as bit density reduction,[20] bit silencing,[21] fragment frequency analysis,[22] or feature filtering[23] have been developed. In addition, Kullback-Leibler (KL) divergence analysis from information theory[24] has been adopted for Bayesian fingerprint similarity searching[25] to perform fingerprint bit ranking and fingerprint reduction.[26] Fingerprint reduction analyses have revealed that typically only subsets of bit positions determine fingerprint search performance.[21,23,26]

Although individual fingerprints have been modified in order to increase their search performance and different fingerprint designs have been extensively compared in similarity searching, it has thus far not been attempted to combine features selected from different fingerprint types. Reasons for this might include that fingerprints have traditionally been utilized as standalone search tools and that it is perhaps not very intuitive to merge different fingerprint designs. However, given our findings that bit subset are often responsible for fingerprint search performance, we have set out to isolate preferred bit segments from different types of fingerprints and "recombine" them into other fingerprint representations. The methodology reported herein generates "hybrid fingerprints" for conventional similarity searching. We show that merging selected structural fragments and pharmacophore features from different fingerprints into hybrid representations further increases the search performance of the original fingerprints.

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

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cmdc.200900243.

Computational Methods

Three fingerprints of different design were utilized for our analysis including MACCS[27] structural keys and two 2D topological or pharmacophore fingerprints, TGD and TGT.[28] MACCS consists of 166 bits that represent structural fragments or patterns consisting of 1-10 nonhydrogen atoms. By contrast, TGD is a two-point topological fingerprint containing 420 bits that represent atom pairs[29] where each atom is assigned to one of seven atom types and inter-atomic distances are divided into 15 different bond distances. Furthermore, TGT is a three-point pharmacophore fingerprint containing 1,704 bits that represent atom triangles. This fingerprint encodes combinations of four different atom types and six different graph distance ranges. A single feature consists of a sextuplet of three atom types and three bond distances between those atom types.

Individual fingerprint bit positions were ranked based on KL divergence utilized as a measure of differences in bit distributions between active and database compounds. Accordingly, the KL divergence mirrors the ability of a bit position to discriminate between active and database molecules (i.e. the larger the divergence value, the more discriminatory the bit position). A detailed description of the KL divergence formalism and divergence calculations for fingerprint bit positions is provided as supplementary methods in the Supporting Information together with an exemplary bit position/feature ranking.

KL divergence calculations and similarity search calculations were carried out on a previously reported[30] set of 27 compound activity classes containing between 30 and 159 compounds, as summarized in table S1 (Supporting Information). As background database for divergence and search calculations, ~3.7 million compounds from ZINC[31] with unique 2D molecular graphs were used.

For 100 randomly selected reference sets of 20 compounds of each activity class and each fingerprint, a KL divergence analysis was carried out to compare bit settings in active and database compounds and rank bit positions according to average divergence values. Based on KL divergence ranking, the 100 top-scoring bit positions of each fingerprint were selected and combined to create 27 activity-class specific hybrid fingerprints, each consisting of 300 bits encoding structural or topological/pharmacophore features. The hybrid fingerprints were then evaluated and compared to the parental fingerprints and a fingerprint representing their complete combination (i.e. 2,290 bits) in conventional similarity search calculations using multiple reference compounds. Therefore, *k* nearest neighbour (*k*-NN)[16,18,19] searching and Tanimoto coefficient (Tc)[32] calculations were carried out. The *k* nearest neighbour method (*k*-NN) calculates the Tc of a database molecule for each individual reference compound. The resulting similarity scores are sorted, and the *k* highest values, corresponding to the *k* nearest-neighbours, are averaged to yield the final similarity score for ranking. In 1-NN calculations, only the top score is selected. We have carried out systematic 1-NN and 5-NN similarity search calculations over all activity classes.

For each activity class, 100 sets of 20 reference compounds were randomly selected for similarity searching and the remaining active compounds were added to the background database as potential hits. The performance was analyzed by calculating cumulative recall curves reporting the number of active molecules among the top-ranked database compounds averaged over all 100 search calculations.

Results and Discussion

Combining parts of different fingerprints into hybrids represents a previously unexplored fingerprint design strategy. Figure 1 schematically illustrates the fingerprint recombination approach. Preferred subsets of fingerprint bit positions might

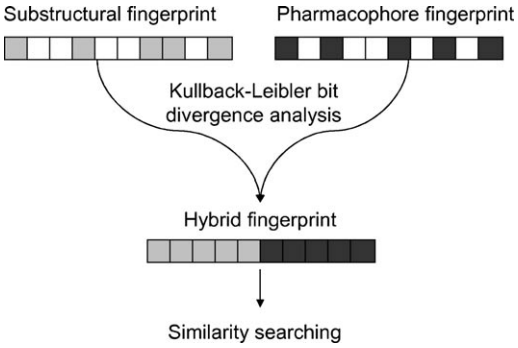

Figure 1. Fingerprint recombination. Bit subsets of two or more different types of fingerprints are selected based on divergence analysis and combined to generate a hybrid fingerprint for conventional similarity searching.

be selected by different means. Thus, fingerprint recombination is not dependent on information theoretic approaches. However, KL divergence analysis has proven to be an effective method to prioritize fingerprint bits based on their potential to differentiate between active and random database compounds.[24,25] KL divergence selection makes it possible to rank bit positions and individually add them to a new fingerprint. Tables 1 and 2 report top-ranked bit positions/features for two activity classes. The probability and corresponding divergence values reveal that these bits have a high probability to be set on in active and a very low probability to be set on in database compounds. Hybrid fingerprints are activity class-directed, of variable size, and, due to their keyed format, generally applicable in conventional similarity searching. For our study, we have generated hybrid fingerprints containing 300 bits taken from a total of 2,290 bit positions available in MACCS, TGD, and TGT. Thus, these hybrids combine features selected from a substructure fingerprint and two distinct 2D topological or pharmacophore fingerprints.

Hybrid fingerprints were compared to their parental and additional control fingerprints in systematic similarity search calculations over 27 compound activity classes. Figure 2 shows representative examples of cumulative recall curves for 1-NN calculations that mirror trends we consistently observed. The search performance of MACCS, TGD, and TGT varied in an activity class-dependent manner: MACCS was best for 16 classes, TGT for six, and TGD for five. However, in almost all cases, hybrid fingerprints produced higher recall rates than their parental fingerprints, as discussed in more detail below. Recovery rate increases were often significant and led to a substantial enrichment of active compounds in relatively small database selection sets of approximately 100 compounds. As a control, we merged MACCS, TGD, and TGT into a single fingerprint consisting of 2,290 bits and this combination was also found

Table 1. Discriminatory fingerprint features in activity class DIR: MACCS,[a] TGD,[b] TGT[c]

Bit MACCS[a]	pA	pD	KL divergence	Description
25	0.97	0.03	3.25	C bonded to $\geq$3N
53	0.97	0.07	2.47	2 QHs separated by 4 bonds
84	0.97	0.09	2.24	NH_2 groups
TGD[b]				
363	0.90	0.03	2.84	D<4>D
468	0.94	0.22	1.24	A<4>D
677	0.98	0.32	1.03	X<3>D
TGT[c]				
8107	0.90	0.01	4.60	A<2>D<4>D<2>
11570	0.97	0.03	3.35	H<3>D<4>D<3>
11575	0.97	0.03	3.33	H<2>D<4>D<4>

[a] The top-ranked MACCS bits for activity class DIR, their corresponding probabilities for active (pA) and database (pD) compounds, and the resulting KL divergences are reported. "Q" means any hetero atom except carbon. [b] The top-ranked TGD bits for DIR. The topological patterns of TGD are described using different atom types (D: hydrogen-bond donor, A: hydrogen-bond acceptor, X: neither hydrogen-bond donor or acceptor, nor acidic, basic or hydrophobic atom) and bond distances between these atoms. [c] The top-ranked TGT bits for DIR. The pharmacophore patterns of TGT are described using different atom types (D: hydrogen-bond donor, A: hydrogen-bond acceptor, H: hydrophobic atom) and bond distances between these atoms.

Table 2. Discriminatory fingerprint features in activity class INO: MACCS,[a] TGD,[b] TGT[c]

Bit MACCS[a]	pA	pD	KL divergence	Description
89	0.95	0.20	1.34	2 Os separated by 4 bonds
57	0.92	0.21	1.17	O in rings
62	0.98	0.34	0.97	Non-ring bond that connects rings
TGD[b]				
471	0.89	0.15	1.39	A<7>D
680	0.89	0.20	1.11	X<6>D
468	0.87	0.22	0.97	A<4>D
TGT[c]				
9882	0.86	0.09	1.72	H<2>D<4>D<4>
5781	0.58	0.02	1.61	A<1>H<5–9>D<4>
8379	0.86	0.11	1.54	A<4>D<5–9>A<5–9>

[a] The top-ranked MACCS bits for activity class INO, their corresponding probabilities for active (pA) and database (pD) compounds, and the resulting KL divergences are reported. [b] The top-ranked TGD bits for INO. The topological patterns of TGD are described using different atom types (D: hydrogen-bond donor, A: hydrogen-bond acceptor, X: neither hydrogen-bond donor or acceptor, nor acidic, basic or hydrophobic atom) and bond distances between these atoms. [c] The top-ranked TGT bits for INO. The pharmacophore patterns of TGT are described using different atom types (D: hydrogen-bond donor, A: hydrogen-bond acceptor, H: hydrophobic atom) and bond distances between these atoms.

to increase the search performance of individual fingerprints in many cases, as reported in figure S1 (Supporting Information). However, as shown in figure S2 (Supporting Information), these increases in compound recall were overall much smaller than those achieved with hybrid fingerprints containing only 13% of the bits of the combined fingerprint. Interestingly, as shown in figure S3 (Supporting Information), control fingerprints containing 300 randomly chosen bits also performed in many instances better than the parental fingerprints, but increases in recall rates were in almost all cases significantly higher for hybrid fingerprints consisting of bit positions selected based on KL divergence (figure S4, Supporting Information). By contrast, using only the top 100 MACCS bits often reached and sometimes exceeded the performance level of the unmodified MACCS fingerprint, whereas fingerprints only comprising the top 100 TGD or TGT bits were outperformed by their parental fingerprints (figure S5, Supporting Information). These observations were consistent with previous findings that fingerprint bit subsets are largely responsible for similarity search performance but that reduced fingerprint representations rarely exceed the performance of unmodified versions of constant-format fingerprints.[21,23,26] Figure 3 reports the gain or loss in recovery rates of hybrid fingerprints in 1-NN and 5-NN search calculations over all activity classes compared to the best performing parental fingerprint. In 1-NN calculations, hybrid fingerprints achieved top recovery rates for 24 of 27 compound classes and in 5-NN calculations for 22 classes. Increases in recovery rates of up to 40% were observed. Taken together, the findings discussed above confirm (1) the ability of the KL divergence selection scheme to identify most discriminatory bit positions/features of original fingerprints and (2) the potential of fingerprint recombination to further increase search performance.

Because features from different fingerprints are independently selected for recombination (exclusively guided by identifying most discriminatory bit positions), it is attractive to explore potential relationships between different types of features at the molecular level. Therefore, we have mapped features corresponding to top-scoring bit positions onto compound activity classes. In a number of instances, relationships were observed that help to rationalize the superior performance of hybrid fingerprints. Figure 4 shows two examples. In these cases, the most discriminatory fingerprint features reported in Table 1 and Table 2, respectively, were mapped onto the compound activity classes for which they were derived. In both cases, the top MACCS, TGD, and TGT features delineated substructures that were shared by nearly all compounds in each activity class. Thus, in these cases, the most discriminatory structural and topological or pharmacophore patterns put high emphasis on compound class-characteristic substructures when hybrid fingerprints were applied, which suggests an explanation for the higher discriminatory ability and ensuing better search performance of compound class-directed hybrid fingerprints compared to their generally applicable parents.

In conclusion, the fingerprint recombination approach reported herein represents a new concept for the design of compound class-directed fingerprints. By combining rationally selected bit positions corresponding to different types of molecular representations such as substructures and pharmacophore patterns, compound class-specific molecular features are fur-

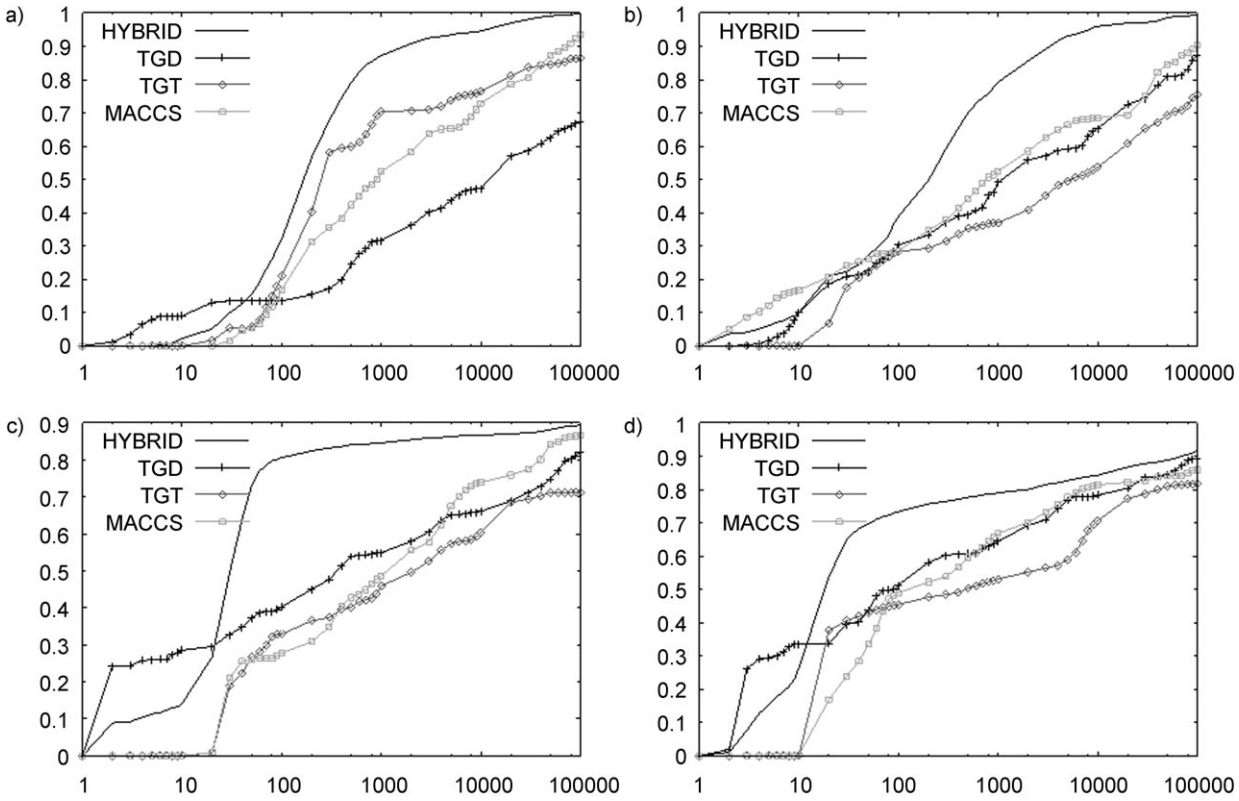

Figure 2. Recall curves. The 1-NN similarity search performance of a hybrid fingerprint is compared to the unmodified MACCS, TGD, and TGT fingerprints by calculating cumulative recall curves reporting the number of active compounds in database selection sets of increasing size. Four representative examples are shown: a) dihydrofolate reductase inhibitors (DIR); b) glycoprotein IIb-IIIa receptor antagonists (GLY); c) inosine monophosphate dehydrogenase inhibitors (INO), and d) xanthine oxidase inhibitors (XAN).

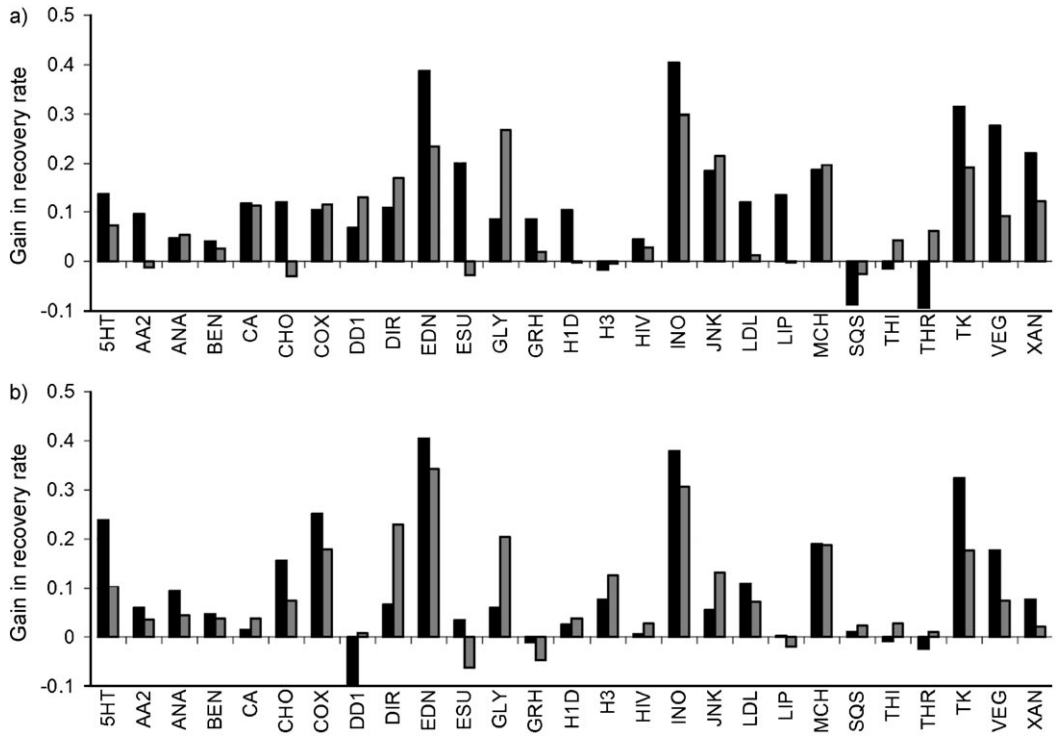

Figure 3. Hybrid fingerprint performance. For each activity class, the difference in recovery rate between the hybrid fingerprint and the best performing unmodified fingerprint (MACCS, TGD, or TGT) is reported for a) 1-NN and b) 5-NN calculations and databases selection sets of 100 (■) or 1000 (■) compounds. Positive values indicate a gain in recovery rate and negative values a loss.

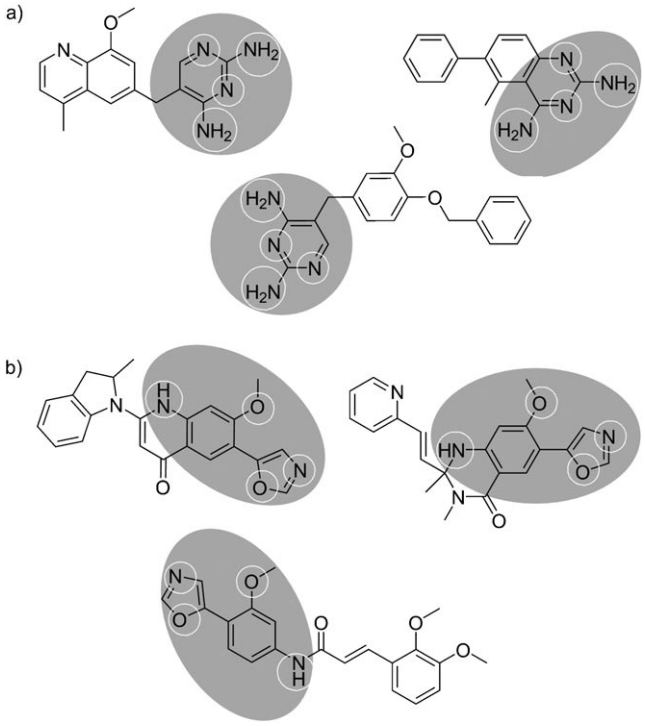

Figure 4. Mapping of preferred fingerprint features. For two activity classes, a) DIR and b) INO, substructures are highlighted that are present in the majority of active compounds (28 of 30 for DIR and 31 of 35 for INO). Top-ranked MACCS, TGD, and TGT bits/features according to Table 1 and Table 2 were mapped onto these substructures. Mapped hydrogen-bond acceptors and donors are circled.

ther emphasized. The selection of discriminatory features from different fingerprints is facilitated by KL divergence analysis. The resulting gain in chemical information leads to improved search performance of hybrid fingerprints compared to their parental fingerprints.

Acknowledgements

The authors thank Martin Vogt for helpful discussions and critical reading of the manuscript. B. N. is supported by Bayer Healthcare, Wuppertal, Germany.

Keywords: feature mapping · hybrid fingerprints · information theory · similarity searching · virtual screening

[1] P. Willett, *J. Med. Chem.* **2005**, *48*, 4183–4199.
[2] H. Eckert, J. Bajorath, *Drug Discovery Today* **2007**, *12*, 225–233.
[3] P. Willett, *Drug Discovery Today* **2006**, *11*, 1046–1053.
[4] J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.
[5] C. A. James, D. Weininger, Daylight Theory Manual. Daylight Chemical Information Systems Inc.: Aliso Viejo, CA (USA) **2008** (http://www.daylight.com).
[6] D. Rogers, R. D. Brown, M. Hahn, *J. Biomol. Screening* **2005**, *10*, 682–686.
[7] J. S. Mason, D. L. Cheney, *Pac. Symp. Biocomput.* **2000**, *5*, 576–587.
[8] E. K. Bradley, P. Beroza, J. E. Penzotti, P. D. J. Grootenhuis, D. C. Spellmeyer, J. L. Miller, *J. Med. Chem.* **2000**, *43*, 2770–2774.
[9] L. Xue, J. W. Godden, F. L. Stahura, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151–1157.
[10] H. Eckert, J. Bajorath, *J. Chem. Inf. Model.* **2006**, *46*, 2515–2526.
[11] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, J. W. Davies, *J. Chem. Inf. Model.* **2009**, *49*, 108–119.
[12] N. E. Shemetulskis, D. Weininger, C. J. Blankley, J. J. Yang, C. Humblet, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.
[13] J. W. Godden, F. L. Stahura, L. Xue, J. Bajorath, *Pac. Symp. Biocomput.* **2000**, *5*, 566–575.
[14] C. Williams, *Mol. Diversity* **2006**, *10*, 311–332.
[15] A. Bender, Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
[16] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
[17] L. Xue, F. L. Stahura, J. W. Godden, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746–753.
[18] J. Hert, P. Willett, D. J. Wilton, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
[19] J. Hert, P. Willett, D. J. Wilton, *J. Chem. Inf. Model.* **2006**, *46*, 462–470.
[20] Y. Wang, H. Geppert, J. Bajorath, *Chem. Biol. Drug Des.* **2008**, *71*, 511–517.
[21] Y. Wang, J. Bajorath, *J. Chem. Inf. Model.* **2008**, *48*, 1754–1759.
[22] Y. Hu, E. Lounkine, J. Batista, J. Bajorath, *Chem. Biol. Drug Des.* **2008**, *72*, 341–349.
[23] Y. Hu, E. Lounkine, J. Bajorath, *ChemMedChem* **2009**, *4*, 540–548.
[24] S. Kullback, *Information Theory and Statistics*, Dover Publications, Mineola, MN, **1997**.
[25] M. Vogt, J. Bajorath, *J. Chem. Inf. Model.* **2008**, *48*, 247–255.
[26] B. Nisius, M. Vogt, J. Bajorath, *J. Chem. Inf. Model.* **2009**, *49*, 1347–1358.
[27] MACCS structural keys, Symyx Software, San Ramon, CA (USA) **2008** (http://www.symyx.com).
[28] Molecular Operating Environment (MOE), Chemical Computing Group Inc., Montreal, Quebec, Canada **2008** (http://www.chemcomp.com).
[29] R. P. Sheridan, M. D. Miller, D. J. Underwood, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
[30] M. Vogt, J. Bajorath, *J. Chem. Inf. Model.* **2007**, *47*, 337–341.
[31] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
[32] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

Received: June 19, 2009
Published online on August 27, 2009

