Shannon Entropy-Based Fingerprint Similarity Search Strategy

Yuan Wang, Hanna Geppert, and Jürgen Bajorath*

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

Received May 5, 2009

For fingerprint searching using multiple active reference compounds, an information entropy-based similarity method is introduced as an alternative to conventional similarity coefficients and search strategies. The approach involves the determination of the fingerprint bit pattern entropy of a compound reference set and recalculation of the entropy following the addition of individual test compounds. If a database compound shares similar bit patterns with reference set molecules, adding this compound to the reference set only produces a small change in system entropy. By contrast, inclusion of a compound having a dissimilar fingerprint leads to a notable increase in entropy. Thus, database compounds can be screened for candidate molecules that do not cause significant changes in reference set fingerprint entropy. Compared to nearest neighbor methods, this approach has the computational advantage that it extracts reference set information only once prior to similarity searching. Test calculations on different compound data sets, fingerprints, and screening databases reveal that the ability of our entropy-based method to detect active compounds is often superior to data fusion techniques and Tanimoto similarity calculations.

1. INTRODUCTION

The basic idea of fingerprint similarity searching is to assess the molecular similarity of screening database compounds to a set of known active reference molecules or an individual reference compound.[1,2] Following the similarity property principle,[3] database compounds with the highest similarity values relative to reference compound(s) are assumed to have a high probability to be active. In order to calculate molecular similarity, fingerprints are computed for reference and screening database compounds, and their bit settings are quantitatively compared[2,4] using similarity functions or metrics such as the popular Tanimoto coefficient (Tc).[1] *k*-NN nearest neighbor searching (*k*-NN) has in recent years become a widely applied data fusion approach to utilize the information provided by multiple reference compounds.[4−7] In *k*-NN calculations, pairwise similarity values between a database compound and each reference molecule are calculated, and the *k* highest values are then averaged to produce the final similarity score. Thus, in this case, database compounds are compared to reference molecules on a compound-by-compound basis. By contrast, other search approaches such as the use of consensus fingerprints,[8] fingerprint profiling and scaling,[9−11] or fingerprint centroids[6] merge the information provided by reference set molecules, but they have proven to be often less effective in detecting active compounds than nearest neighbor methods.[5,6,12]

Here, we introduce a new fingerprint search strategy that also combines reference compound information prior to similarity assessment and that is based on the Shannon entropy (*SE*) concept. This concept was introduced in 1948 by Shannon[13] in information theory and was originally applied to assess the information content of messages transmitted through different channels. In this context, messages with high information content (high *SE*) display few or no recognizable patterns, whereas those having low information content (low *SE*) exhibit regular patterns that correspond to information redundancy.[14]

The SE concept is readily transferable to molecular fingerprints when bit positions are considered to be individual channels that are capable of transmitting binary signals, i.e. by setting bit positions on (to 1) or off (0). Accordingly,

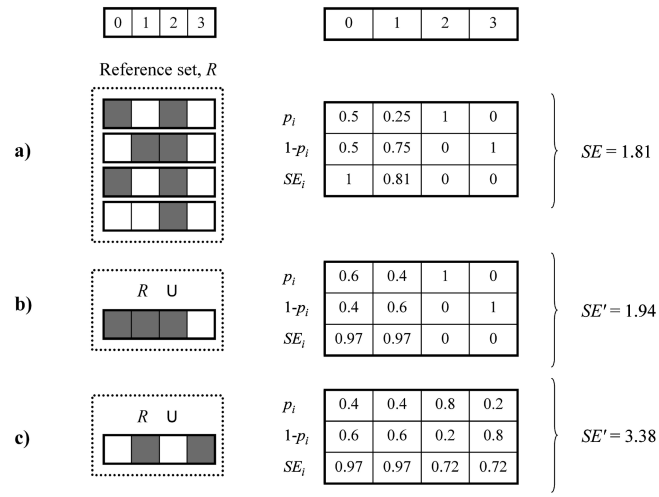

Figure 1. Calculation of fingerprint Shannon entropy. A hypothetical four-bit fingerprint is used to illustrate the calculation of Shannon entropy (*SE*) of individual bit positions and complete fingerprints for a set of molecules. "1" and "0" bits are represented using gray and white cells, respectively. In **a)**, bit strings of a reference set *R* of four molecules are shown. In **b)** and **c)**, an additional molecule (bit string) is added to *R*. For the three different compound sets, the probability p_i for a "1" bit, the probability 1-p_i for a "0" bit, and the corresponding Shannon entropy (SE_i) are reported for each bit position. Resulting Shannon entropies for complete fingerprints (SE or *SE*′) are given on the right.

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

10.1021/ci900159f CCC: $40.75
Published on Web 07/07/2009

Table 1. Activity Classes and Active Database Compounds[a]

code	activity class	active DB compounds for MACCS	active DB compounds for TGD
ACE	angiotensin-converting enzyme inhibitors	30	20
ADR	aldose reductase inhibitors	70	200
CAM	cell adhesion molecule antagonists	10	20
CLG	collagenase inhibitors	20	20
FXA	factor Xa inhibitors	40	10
PA2	phospholipase A2 inhibitors	100	100
PKC	protein kinase C inhibitors	70	100
SST	squalene synthetase inhibitors	40	100

[a] For each activity class, the number of molecules extracted from the MDDR as active database compounds (potential database hits) is reported. Compound sets were specifically assembled to have MACCS or TGD fingerprint bit densities comparable to compound averages in the two test databases. For each class, 20 unique reference compounds with corresponding bit densities were also selected.

compound (reference) sets whose fingerprints share similar bit patterns produce low SE values. By contrast, if there is only little bit pattern resemblance, high SE values are obtained. Moreover, if "0" and "1" bits are randomly distributed, the SE value of the system is maximal. Accordingly, given the premise that chemically and biologically similar molecules should yield similar fingerprint representations, ensembles of compounds having similar activity should produce low fingerprint SE values. Then, by adding a compound of unknown activity to the reference set and recalculating the SE for the expanded fingerprint ensemble, we can directly assess the similarity of a test compound to the reference set. If there is only a small change in the resulting SE value, the fingerprint of the test compound is similar to the reference set, and the compound is thought to have similar properties. In systematic similarity test calculations, the performance of the fingerprint SE approach was comparable to or better than *k*-NN nearest neighbor searching.

2. METHODOLOGY

2.1. Shannon Entropy of Binary Fingerprints. Given a compound set R and an arbitrary binary fingerprint representation $\mathbf{X}$ consisting of n bit positions, the SE value of a single bit position $i \in \{1, ..., n\}$ in the set R is calculated as[13]

$$SE_i(R) = -p_i \log_2(p_i) - (1 - p_i) \log_2(1 - p_i) \quad \text{with } p_i = \sum_{A \in R} x_{iA}$$

Here, p_i represents the relative frequency of "1" bits at fingerprint position i in R. In the case of $p_i = 0$ or $p_i = 1$, $p_i \log_2(p_i)$ or $(1 - p_i) \log_2(1 - p_i)$ become 0. The Shannon entropy of the complete fingerprint of R is the sum of the individual SE_i values obtained for each bit position i

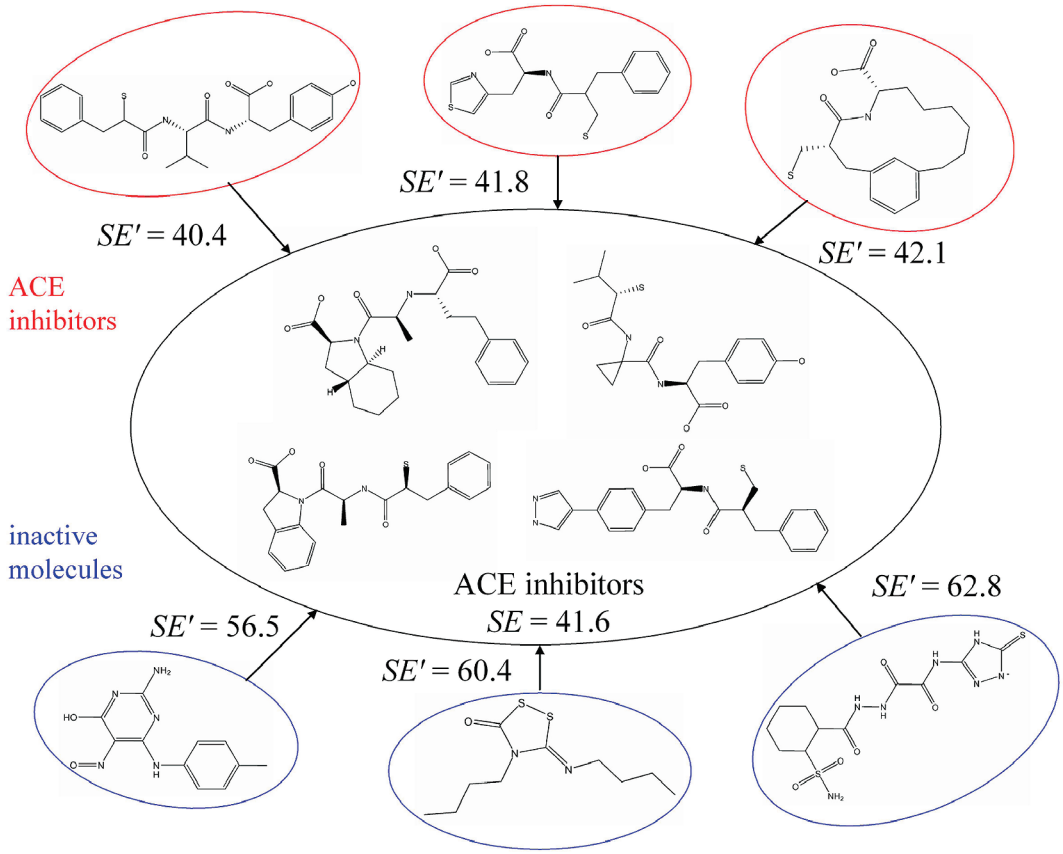

Figure 2. Shannon entropy-based fingerprint similarity. The Shannon entropy (*SE*) of a reference compound set of four ACE inhibitors (shown in the circle) is reported for the MACCS fingerprint. Three additional ACE inhibitors (shown in red ovals at the top) are separately added to the reference set, and SE values are recalculated (*SE′*). For comparison, three compounds randomly selected from the NCI database (in blue ovals at the bottom) are also separately added to the reference set and SE values are updated.

$$SE(R) = \sum_{i=1}^{n} SE_i(R)$$

Figure 1a shows an exemplary SE calculation using a hypothetical four-bit fingerprint.

2.2. Database Ranking Using SE Values. Given a set R of reference molecules and its calculated SE value, this value typically changes when adding another compound A to R. The magnitude (and algebraic sign) of the change indicates whether or not A matches a potential common bit pattern of R, as illustrated in Figure 1. Two compounds are separately added to the reference set R shown in Figure 1a, and the SE values are recalculated. The molecule introduced in Figure 1b slightly decreases or increases SE_i at bit positions 0 to 1, respectively, and matches the "1" and "0" consensus bits of R at bit positions 2 to 3, respectively, so that SE_2 and SE_3 remain 0. The overall SE value only slightly increases from SE = 1.81 to $SE' = 1.94$. By contrast, the compound shown in Figure 1c does not match this pattern (SE_2 and SE_3 become 0.72) so that the overall SE value significantly increases to $SE' = 3.38$. Hence, departure from consensus bit positions and patterns in R is associated with a significant entropy penalty. Monitoring such changes in SE values when adding individual test compounds to reference sets makes it possible to sort database compounds in the order of increasing SE' values corresponding to decreasing molecular similarity and produces a database ranking. Absolute SE values depend on the bit structure of different fingerprints and the composition of the reference sets R and can thus not be transferred or interpreted *a priori*. For a given fingerprint and reference set, the SE' ranking of database compounds is the major readout.

2.3. Test Calculations. Two databases were used for simulated similarity search calculations, the NCI anti-AIDS database[15] containing 42,687 compounds and a set of 500,000 randomly selected ZINC[16] molecules. Eight compound activity classes were assembled from the MDL Drug Data report (MDDR),[17] as reported in Table 1. In order to ensure that active compounds had physicochemical properties comparable to compounds of the screening databases, they were filtered applying the filter rules of the ZINC database, i.e., maximum molecular weight of 600 Da, logP values between −2 and +6, no more than 18 rotatable bonds, and between 1 and 10 hydrogen bond donors and acceptors. Furthermore, to avoid potential bias through the inclusion of analog series, a scaffold analysis algorithm was used to exclusively select active compounds having distinct core structures.[18]

Two molecular fingerprints were used to test the SE approach, MACCS structural keys consisting of 166 bit positions[19] and the TGD fingerprint available in the Molecular Operating Environment[20] that codes for typed graph distances and consists of 420 bit positions. From each activity class, two compound subsets were selected having MACCS or TGD fingerprint bit densities comparable to the screening databases (Table 1), and these compound subsets were used as potential database hits. Furthermore, for each compound class and fingerprint, reference sets of 20 active compounds were selected that also had fingerprint bit densities comparable to the screening databases. Bit density analysis and density-based compound selection were carried out prior to similarity searching in order to balance fingerprint complexity effects that can substantially bias similarity calculations.[21,22]

Table 2. Recovery Rates for Different Similarity Search Strategies[a]

	SE		centroid		20-NN		1-NN	
	100	1000	100	1000	100	1000	100	1000
				a)				
ACE	**83**	**90**	73	**90**	57	**90**	60	80
ADR	10	39	9	17	6	17	**21**	**44**
CAM	**40**	**40**	20	**40**	20	**40**	30	**40**
CLG	**45**	55	40	45	40	45	**45**	**75**
FXA	**20**	**65**	5	40	5	25	10	35
PA2	3	14	3	12	3	12	**8**	**16**
PKC	**16**	**47**	7	26	4	20	14	21
SST	23	**43**	23	30	20	28	**35**	**43**
average	**30**	**49**	23	38	19	35	28	44
				b)				
ACE	**47**	**83**	40	73	27	57	30	57
ADR	3	6	3	6	0	4	**13**	**26**
CAM	**20**	**30**	0	20	0	0	**20**	**30**
CLG	**35**	**40**	**35**	**40**	20	**40**	25	**40**
FXA	**5**	**8**	0	5	0	0	3	3
PA2	**3**	3	**3**	3	**3**	3	2	**4**
PKC	**4**	**13**	1	4	0	4	3	**13**
SST	20	20	20	20	20	20	**28**	**40**
average	**17**	25	13	21	9	16	15	**26**
				c)				
ACE	**50**	**65**	45	**65**	20	55	5	45
ADR	**4**	**8**	3	7	2	5	**4**	**8**
CAM	**10**	**15**	0	**15**	0	**15**	0	5
CLG	**25**	**45**	5	35	5	30	0	25
FXA	**10**	10	0	10	0	10	0	**20**
PA2	12	22	**13**	19	11	17	12	**25**
PKC	12	27	14	27	18	27	**22**	**34**
SST	8	38	**10**	**40**	9	28	7	12
average	**16**	**30**	12	28	9	24	9	25
				d)				
ACE	**25**	**45**	5	**45**	0	20	0	5
ADR	**1**	**3**	**1**	**3**	**1**	2	0	1
CAM	0	0	0	0	0	0	0	0
CLG	0	**20**	0	5	0	5	0	0
FXA	0	0	0	0	0	0	0	0
PA2	**7**	8	7	**11**	**7**	**11**	1	6
PKC	4	6	5	7	7	12	**10**	**17**
SST	**6**	**13**	**6**	10	5	9	3	7
average	**5**	**12**	3	10	3	7	2	5

[a] Recovery rates (in %) are reported for four different similarity search strategies (SE, centroid, 20-NN, 1-NN) and different combinations of fingerprints and test databases: a) MACCS and NCI, b) MACCS and ZINC, c) TGD and NCI, and d) TGD and ZINC. For each activity class, results are compared for selection sets of 100 and 1000 molecules, and the search strategies producing the highest recovery rates are boldface.

Systematic similarity search calculations were conducted for the combination of each activity class, screening database (NCI or ZINC), and fingerprint (MACCS or TGD), resulting in a total of 32 test calculations, and the recovery of active database compounds was monitored for different selection set sizes. The SE approach was compared to three standard similarity search strategies, 1-NN, 20-NN, and centroid calculations. In 20-NN calculations, the average of all 20 pairwise Tc values yielded the final similarity score, and, in 1-NN calculations, the largest of the 20 individual values was taken. For the centroid method, an average bit string was derived from the 20 active reference compounds and compared to database molecules in Tc calculations.

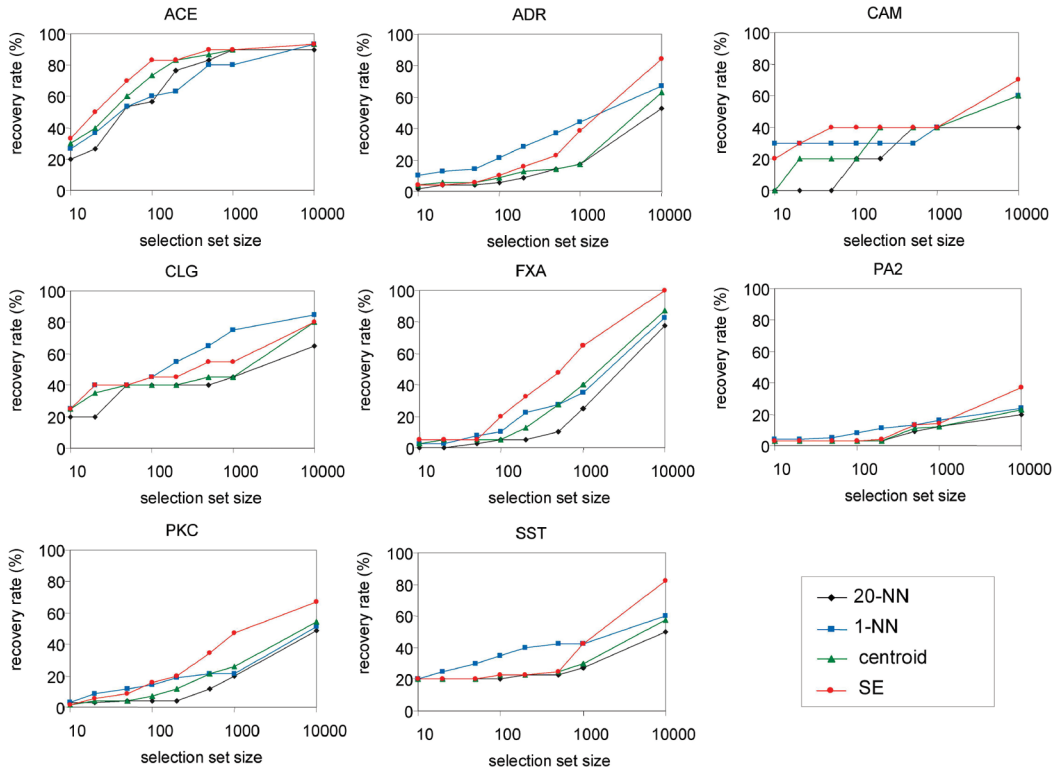

Figure 3. Comparison of recovery rates. Recovery rates (in %) for the four different similarity search strategies SE (red), 20-NN (black), 1-NN (blue), and centroid (green) using the MACCS fingerprint and NCI database are compared for selection sets of increasing size (shown on a logarithmic scale).

3. RESULTS AND DISCUSSION

3.1. Fingerprint Shannon Entropy of Compound Sets. In Figure 1, it is illustrated that the SE approach can in principle be used to detect and quantify changes in bit patterns for model bit strings. To investigate whether the SE approach can also distinguish between active and inactive compounds using conventional fingerprint representations, we first analyzed a small compound set consisting of four reference and six test molecules. Figure 2 shows the molecular graphs of these compounds and reports the SE values for the MACCS fingerprint. The four reference molecules shown in the center belong to class ACE and produce an SE value of 41.6. Separately adding three other ACE inhibitors as candidate molecules (depicted in red ovals at the top) changes the SE value of the expanded compound set only very little. Addition of the upper-left molecule actually leads to a small SE reduction ($SE' = 40.4$), separate addition of the compound in the middle results in $SE' = 41.8$ and of the upper-right molecule in $SE' = 42.1$, although these compounds are structurally distinct. By contrast, when separately adding three compounds randomly taken from the NCI database (shown at the bottom), SE values significantly increase to 56.5, 60.4, and 62.8, respectively. Thus, in this case, the three active candidate compounds were effectively separated from three inactive ones on the basis of fingerprint SE calculations.

3.2. Entropy-Based Similarity Searching. In light of these findings, we further evaluated the SE approach in similarity search calculations using eight different activity classes, two fingerprints, and two screening databases (NCI and ZINC) in comparison to the standard 1-NN, 20-NN, and centroid similarity search strategies. Recovery rates for selection sets of 100 and 1000 compounds are reported in Table 2 when using 20 active reference molecules. Results of the best-performing similarity search approach are highlighted in boldface for each trial and selection set size. The results in Table 2 reveal that SE performed consistently better than 20-NN and centroid calculations and that it was overall comparable to or better than 1-NN. Summarizing over the 32 different trials and selection sets of 100 database compounds, SE produced the highest recovery rates in 20 cases, 1-NN in 10, centroid in seven, and 20-NN in three. Furthermore, for a selection set size of 1000 compounds, SE performed best in 18 cases, 1-NN in 16, centroid in nine, and 20-NN in five. Figure 3 shows cumulative recall curves for the eight test calculations using the MACCS fingerprint and the NCI database. These curves further illustrate that SE was generally superior to centroid and 20-NN calculations and that it frequently also performed better than the 1-NN strategy.

For fingerprint similarity searching, the SE approach is computationally less complex than nearest neighbor methods. Nearest neighbor methods require the determination of pairwise similarity values between a database molecule and each reference compound (e.g., 20 calculations per database molecule in this case). By contrast, SE (and also centroid searching) utilizes the information of the whole reference set only once to generate a bit frequency profile (or centroid vector). Then, during similarity searching, a database compound is compared to the frequency profile (or centroid vector) in a single calculation. Thus, while SE leads to comparable or better search results than nearest neighbor methods, it also accelerates similarity searching, especially when large numbers of active reference compounds are available.

4. CONCLUSIONS

Fingerprint-based similarity searching using sets of active reference compounds requires the application of multiple-template search strategies such as nearest neighbor methods or the centroid technique. While nearest neighbor methods rely on pairwise compound comparisons and do not use the information provided by a reference set as a whole, they have often performed best in comparative benchmark studies. Both the centroid and nearest neighbor methods depend on the calculation of similarity coefficients. We have developed an information entropy-based similarity search strategy for binary fingerprints that implicitly captures whether or not a database compound shares bit patterns that occur in a reference set. The approach conceptually differs from other search strategies and similarity metrics and has low computational complexity. In systematic test calculations, we have shown that the SE method introduced herein can further improve the similarity search performance of current state-of-the-art approaches.

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) Stahura, F. L.; Bajorath, J. New Methodologies for Ligand-Based Virtual Screening. *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.
(3) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.
(4) Willett, P. Similarity-based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
(5) Hert, J.; Willett, 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.
(6) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
(7) 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.
(8) 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.
(9) 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.
(10) 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. Inf. Comput. Sci.* **2001**, *41*, 746–753.
(11) 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.
(12) Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D Fingerprint Methods for Multiple-Template Similarity Searching on Compound Activity Classes of Increasing Structural Diversity. *ChemMedChem* **2007**, *2*, 208–217.
(13) Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623−656.
(14) Chaitin, G. J. Goedel's Theorem and Information. *Int. J. Theor. Phys.* **1982**, *21*, 941–954.
(15) The publicly available NCI anti-AIDS database contains structural and activity data for compounds screened by the AIDS antiviral screening program of the National Cancer Institute, 1999. http://dtp.nci.nih.gov/docs/aids/aids_data.html (accessed Feb 15, 2007).
(16) 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.
(17) *MDL Drug Data Report (MDDR), version 2005.2*; Symyx Software: San Ramon, CA, 2005.
(18) Xue, L.; Bajorath, J. Distribution of Molecular Scaffolds and R-groups Isolated from Large Compound Databases. *J. Mol. Model.* **1999**, *5*, 97–102.
(19) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2002.
(20) *MOE (Molecular Operating Environment)*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2007.
(21) Wang, Y.; Eckert, H.; Bajorath, J. Apparent Asymmetry in Fingerprint Similarity Searching is a Direct Consequence of Differences in Bit Densities and Molecular Size. *ChemMedChem* **2007**, *2*, 1037–1042.
(22) Wang, Y.; Bajorath, J. Balancing the Influence of Molecular Complexity on Fingerprint Similarity Searching. *J. Chem. Inf. Model.* **2008**, *48*, 75–84.

CI900159F