Introduction of an Information-Theoretic Method to Predict Recovery Rates of Active Compounds for Bayesian in Silico Screening: Theory and Screening Trials

Martin Vogt and Jürgen Bajorath*

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

Received September 26, 2006

We present the first method to predict compound recovery rates from descriptor statistics. A log-odds function is designed that models probability distributions of descriptor values of active and inactive molecules in chemical space and used to determine the likelihood of database compounds to exhibit a specific activity. The divergence of probability models for active and inactive compounds is applied to evaluate the ability of the log-odds likelihood function to recover active compounds from a background database. The divergence measure, which is closely related to the Kullback−Leibler distance, is strongly correlated with recovery rates of Bayesian virtual screening calculations. It has thus been possible to predict compound recovery rates for different activity classes. Prior to practical virtual screening trials, one can also estimate how likely it would be to recover active compounds from a given screening database.

INTRODUCTION

Ligand-based virtual screening methods are typically evaluated and benchmarked in a retroactive manner. Regardless of the types of methods that are investigated, this is most often done by dividing sets of known active compounds into “baits” and “hits”, adding potential hits to a virtually formatted screening database, and attempting to recover them using baits for training or as screening templates.[1] Conventionally, hit and recovery rates of active compounds are determined as performance measures in retrospective virtual screening trials.[1,2] Hit rates are defined as the number of correctly identified active molecules in database selection sets of a given size (e.g., 100, 500, etc.). In biological screening, hit rates refer to the fraction of active molecules within an experimentally tested compound collection. In virtual screening, recovery rates are defined as the percentage of all available active molecules in a database contained in a compound selection set. In addition, the so-called “cumulative recall” is a popular measure of virtual screening performance.[2,3] It monitors the fraction of potential hits that are retrieved in database increments of increasing size. In many instances, cumulative recall curves are plotted for up to about 1% of a source database as a graphically intuitive measure of screening performance.[3] However, such cumulative recall curves often make virtual screening calculations artificially “look good”, if we take into consideration that 1% of a source database of a current standard size of about 1 million molecules or more would amount to 10 000 molecules and that any computational screening method having at least some predictive value should be capable of enriching active molecules in selection sets of such large size. For practical virtual screening applications, where database compounds are preselected for experimental evaluation, the calculation of hit rates is straightforward, but recovery rates could only be determined if all database compounds were assayed; otherwise, it would of course not be known how many active molecules the database contains, if any. Thus, in contrast to hit rates, recovery rates can rarely be determined to evaluate the performance of “real-life” virtual or sequential screening projects.[1] However, in benchmark situations, determining compound recall is essential to evaluate the sensitivity and specificity of virtual screening calculations, that is, the ability to select active compounds and distinguish them from false positives.

Given the importance of compound recall as a criterion for computational screening, we have investigated the question whether compound class-specific recovery rates could also be predicted by comparing chemical features of active and database compounds. We reasoned that this might be possible since compound collections are typically projected into chemical descriptor spaces of varying dimensionality for virtual screening applications.[4] Thus, large arrays of descriptor values of test compounds are available as a source of information, and the ability to recover active compounds from screening databases ultimately depends on the detection of systematic differences in value distributions of hits and decoys.

In previous studies,[5,6] we have developed distance-based methods for mining chemical descriptor spaces in order to predict the activity of compounds on the basis of their feature distributions. The DACCS method used a simple Euclidean-like distance measure[5] to navigate chemical reference spaces and considered the distance from the center of a subspace populated by an activity class as a criterion for similarity of database compounds; increasing distance from the center of an “active subspace” correlated with decreasing similarity between active and database molecules.[5] Subsequently, the BDACCS function[6] was developed to further refine the distance function approach on the basis of Bayesian principles.[7] The BDACCS function was interpreted as a “log-odds” measure for compound activity because it related the likelihood of descriptor values for active and inactive compounds to each other. DACCS distances were trans-

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

10.1021/ci600418u CCC: $37.00
Published on Web 02/16/2007

formed into a probability of a compound to be active; the larger the distance from an active subspace, the lower the likelihood of activity. Like DACCS, BDACCS operates in arbitrary high-dimensional descriptor spaces and produces an activity-likelihood ranking for database compounds. In benchmark calculations, we found in many cases that BDACCS further increased compound recovery rates achieved with DACCS and also performed better than calculation of Tanimoto similarity using standard 2D fingerprints.[6]

Since BDACCS models probability distributions of descriptor values for active and inactive compounds, we decided to use this approach as a methodological framework for the prediction of recovery rates. Principles of information theory[8,9] were applied to utilize the divergence of probability distributions for active and inactive compounds for the prediction of recovery rates. A large number of different compound activity classes were used to establish a linear relationship between probability distribution divergence and recovery rates observed in Bayesian database screening trials, which could be used to successfully predict recovery rates for other activity classes.

METHODOLOGY

The DACCS and BDACCS methods operate in unmodified chemical descriptor spaces of high dimensionality to rank database molecules according to their similarity to active compounds. The derivation of the BDACCS function is described in detail in the Supporting Information. In estimating the probability distributions of descriptor values, two simplifying assumptions are made: (i) descriptor values of test compounds follow a Gaussian distribution, and (ii) values of different descriptors are independently distributed. Both assumptions are likely to introduce approximations. These assumptions have the advantage that distributions can be estimated from the mean and standard deviation for each descriptor. If no further information about the descriptor distributions is taken into consideration, these assumptions are most unbiased in the sense that the Shannon entropy of the resulting Gaussian distribution is known to be maximal over all distributions when only first and second moments are given.[10] The soundness of this approach is consistent with the observed high predictive ability of the BDACCS method, as revealed by our previous benchmark calculations.[6] The BDACCS function was subjected to extensive benchmark calculations against three reference methods and on a total of 52 different compound activity classes.[6] In this paper, we base the development of the Kullback−Leibler (KL) variant of our Bayesian model on the previously implemented BDACCS function because we use the same approximations with respect to descriptor value statistics for the prediction of recovery rates. However, we would also like to note some possibilities for further study. As descriptor correlation effects can frequently be observed in chemical space design,[4] taking the correlation of descriptor values into consideration might be a worthwhile improvement of the method design so that correlated descriptors do not dominate the similarity measure. Furthermore, instead of estimating a continuous distribution, discrete distributions may be used on the basis of histograms of observed frequencies. In a separate follow-up study, we will investigate such variants of the current BDACCS function and compare them to other methods, for example, distance-based classifiers.

The BDACCS scoring scheme that was derived on the basis of the assumptions discussed above assigns a log-odds score to a vector $\mathbf{x} = (x_i)_i$ of descriptor values for a compound:

$$\log R(\mathbf{x}) = \log \prod_{i=1}^{n} \frac{p(x_i|\mathrm{A})}{p(x_i|\mathrm{B})} = \sum_{i=1}^{n} [\log p(x_i|\mathrm{A}) - \log p(x_i|\mathrm{B})]$$

Here, $p(x_i|\mathrm{A}) = 1/(\sqrt{2\pi\sigma_i^2})\exp-[(x_i - \mu_i)^2/2\sigma_i^2]$ and $p(x_i|\mathrm{B}) = 1/(\sqrt{2\pi\tau_i^2})\exp-[(x_i - \nu_i)^2/2\tau_i^2]$ are Gaussian distributions where μ_i and σ_i are the mean and standard deviation for descriptor i for active compounds and ν_i and τ_i are the mean and standard deviation for descriptor i for inactive compounds. From this formulation, we can obtain a BDACCS "distance" measure:

$$D_{\mathrm{BDACCS}}(\mathbf{x}) = \sum_{i=1}^{n} \frac{(x - \mu_i)^2}{\sigma_i^2} - \frac{(x - \nu_i)^2}{\tau_i^2}$$

Given this function, we can establish a statistical measure that serves as an indicator of the discriminatory power and ability to detect active compounds. It should be noted that the distance function is based upon the ratio of two probability distributions. This means that this function will be increasingly discriminatory the more the distributions for active and inactive compounds differ.

In information theory, a well-known measure for the divergence of two probability distributions is the KL function.[9] The joint distributions of descriptor vectors $\mathbf{x}$ for active (A) and inactive (B) compounds are

$$p(\mathbf{x}|\mathrm{A}) = \prod_{i=1}^{n} p(x_i|\mathrm{A})$$

$$p(\mathbf{x}|\mathrm{B}) = \prod_{i=1}^{n} p(x_i|\mathrm{B})$$

When these joint distributions are considered, the KL divergence is given as

$$D[p(\mathbf{x}|\mathrm{A})||p(\mathbf{x}|\mathrm{B})] = \int p(\mathbf{x}|\mathrm{A}) \log \frac{p(\mathbf{x}|\mathrm{A})}{p(\mathbf{x}|\mathrm{B})} \,\mathrm{d}\mathbf{x}$$

$$D[p(\mathbf{x}|\mathrm{A})||p(\mathbf{x}|\mathrm{B})] = \int p(\mathbf{x}|\mathrm{A}) \log R(\mathbf{x}) \, d\mathbf{x} = E[\log R(\mathbf{x})|\mathrm{A}]$$

and can be expressed as a function of our distance measure.

$$D[p(\mathbf{x}|\mathrm{A})||p(\mathbf{x}|\mathrm{B})] \propto E[-D_{\mathrm{BDACCS}}(\mathbf{x})|\mathrm{A}] + \mathrm{const}$$

Given this formulation, KL divergence directly corresponds to the BDACCS score expected for active compounds: the larger the KL divergence, the smaller the expected "distance" from active compounds.

Under assumptions of descriptor independence and the presence of normal distributions, the KL divergence can be calculated analytically from the means and standard deviations of the descriptor values for active and inactive compounds:[5]

$$D[p(\mathbf{x}|\mathrm{A})||p(\mathbf{x}|\mathrm{B})] = \sum_{i=1}^{n} D[p(x_i|\mathrm{A})||p(x_i|\mathrm{B})]$$

$$D[p(\mathbf{x}|\mathrm{A})||p(\mathbf{x}|\mathrm{B})] = \sum_{i=1}^{n} \log\frac{\tau_i}{\sigma_i} + \frac{\sigma_i^{\,2} - \tau_i^{\,2} + (\mu_i - \nu_i)^2}{2\tau_i^{\,2}}$$

Thus, the divergence only depends on the means and standard deviations of descriptor values for the set of active compounds and the set of database compounds.

If the screening database contains a small—yet unknown—number of actives, the KL divergence correlates with the percentage of active compounds among database compounds producing the best BDACCS scores (i.e., smallest distances). Thus, given a learning set of active compounds whose chemical features are similar to those of potential hits and given the means and standard deviations of arbitrary numbers of descriptors for active and database compounds, we can predict the recovery rate of active compounds. Since the approach combines KL divergence and the BDACCS function, it is termed KL-BDACCS.

In the above derivation, we used the directed Kullback—Leibler divergence. Alternatively, one might consider the symmetric formulation:

$$J[p(\mathbf{x}|\mathrm{A}),p(\mathbf{x}|\mathrm{B})] = D[p(\mathbf{x}|\mathrm{A})||p(\mathbf{x}|\mathrm{B})] + D[p(\mathbf{x}|\mathrm{B})||p(\mathrm{x}|\mathrm{A})]$$

Incorporating the BDACCS distributions yields

$$J[p(\mathbf{x}|\mathrm{A}),p(\mathbf{x}|\mathrm{B})] \propto E[D_{\mathrm{BDACCS}}(\mathbf{x})|\mathrm{B}] - E[D_{\mathrm{BDACCS}}(\mathbf{x})|\mathrm{A}]$$

One might expect this measure to more accurately reflect the recovery ability of the BDACCS method, but test calculations have shown that the symmetric Kullback—Leibler measure has considerably more variation. This may be ascribed to the fact that in the formula

$$D[p(\mathbf{x}|\mathrm{B})||p(\mathbf{x}|\mathrm{A})] = \sum_{i=1}^{n} \log\frac{\sigma_i}{\tau_i} + \frac{\tau_i^{\,2} - \sigma_i^{\,2} + (\mu_i - \nu_i)^2}{2\sigma_i^{\,2}}$$

σ_i^2, the variance of descriptor *i* for the active set, appears in the denominator. Typically, active sets will be small in size, so the variance will show high fluctuations which give rise to unreliable values for $D[p(\mathbf{x}|\mathrm{B})||p(\mathbf{x}|\mathrm{A})]$.

CALCULATIONS

For KL-BDACCS calculations, a pool of 141 1D and 2D descriptors implemented in the Molecular Operating Environment (MOE)[11] was used. Calculations were carried out on an in-house generated subset of the ZINC[12] database that contained ~1.44 million compounds having unique 2D graphs. A total of 40 different compound activity classes were used to study the relationship between KL divergence and compound recovery rates in BDACCS virtual screening trials. The composition and origins of these activity classes are reported in Supplementary Table 1 (Supporting Information). For each activity class, 100 randomly selected sets of 10 active compounds were used as baits for 100 independent search trials, and in each case, the remaining 4−149 active molecules were added to the background database as potential hits, while all ZINC compounds were considered inactive. In order to make recovery rates statistically independent of varying numbers of potential hits per class, the BDACCS scoring ranges of the 10, 100, and 1000 top-scoring ZINC compounds were identified for each calculation, and the numbers of active molecules falling into these scoring ranges were determined, and actual recovery rates were calculated.

Table 1. Compound Activity Classes Used to Predict Recovery Rates

class designation	biological activity	number of compounds
THI	thiol protease inhibitors[13]	34
5HT	5-HT serotonin receptor ligands[14]	71
JNK	C-jun N-terminal kinase inhibitors[15]	36
EDN	endothelin ETA antagonists[13]	32
SQS	inhibitors of squalene synthetase[13]	42
THB	thrombin inhibitors[13]	35
HIV	HIV protease inhibitors[14]	48

For each trial, the KL divergence was calculated from the mean and standard deviations of the descriptor value distributions of available active molecules and ZINC compounds. On the basis of these data, a linear regression model was derived for average recovery rates versus the logarithm of the corresponding KL divergence. This function was used to predict recovery rates from calculated KL divergence for seven other activity classes not included in the calibration set, as summarized in Table 1. For these classes, 100 independent virtual screening trials were carried out and average recovery rates for differently sized compound selection sets were determined, as described above. Once descriptor values, standard deviations, and means are calculated for a large compound database (such as the one used here), the screening calculations are fast and take minutes on a standard PC.

RESULTS

First, we analyzed the relationship between recovery rates of BDACCS virtual screening calculations and corresponding KL divergence for a calibration set of 40 activity classes and selection sets of 10, 100, and 1000 database compounds. Figure 1 shows results obtained for a selection set of 100 molecules, and Supplementary Figures 1a and 2a (Supporting Information) show corresponding results for selection sets of 10 and 1000 compounds, respectively. The obtained results were similar for the differently sized compound selection sets, indicating the presence of a nearly linear relationship between recovery rates and the logarithm of the KL divergence. Reasonable linear models could be derived by regression analysis, yielding a correlation coefficient of 0.79 for selection sets of 100 compounds (Figure 1) and an average absolute error of 16%. Regression models for selection sets of 10 and 1000 compounds produced correlation coefficients of 0.81 and 0.74, respectively. Thus, on the basis of these findings, we concluded that it was possible to calibrate a linear function relating recovery rates observed for 40 different activity classes to corresponding log(KL) divergence values.

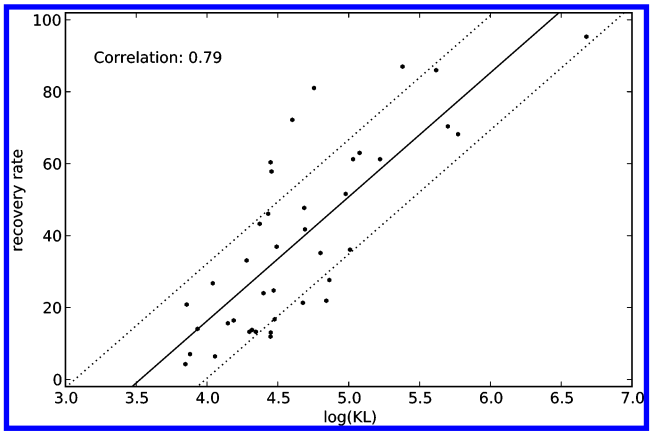

Figure 1. Relationship between recovery rates and KL divergence. For 40 different compound activity classes, average recovery rates from 100 individual trials were calculated and plotted against the logarithm of the corresponding average KL divergence. For the calculation of recovery rates, the number of active compounds falling within the scoring range of 100 database compounds was determined. The solid line represents a linear regression model yielding a correlation coefficient of 0.79, and dashed lines mark the averaged standard deviation of the virtual screening trials.

Table 2. Predicted and Measured Recovery Rates[a]

class designation	avTc	measured average recovery rate	predicted recovery rate	absolute error
THI	0.49	11.6%	33.7%	22.1%
5HT	0.67	27.9%	23.2%	4.7%
JNK	0.44	36.6%	31.6%	5.0%
EDN	0.64	50.4%	49.3%	1.1%
SQS	0.50	63.7%	81.5%	17.8%
THB	0.57	75.6%	73.1%	2.5%
HIV	0.72	88.5%	83.7%	4.8%

[a] Average recovery rates were determined over 100 individual trials and for selection sets of 100 ZINC compounds and predicted from KL divergence. In order to assess intraclass structural diversity, average values of the Tanimoto coefficient[16] (avTc) were calculated for pairwise comparison of compounds using a set of 166 publicly available MACCS structural keys.[17]

Then, we investigated whether the linear model could be used to predict recovery rates from KL divergence for compound activity classes outside the calibration set. For this purpose, activity classes were arbitrarily chosen; our only requirement was that they should span a spectrum ranging from low to high recovery rates (i.e., representing different degrees of difficulty for compound retrieval). For the seven activity classes reported in Table 1 that were selected as test cases, the number of potential hits (added to more than 1 million ZINC compounds) ranged from 22 to 51, and as shown in Table 2, their measured BDACCS recovery rates ranged from approximately 12% to 88%. Figure 2a summarizes the results of 100 individual predictions for different bait sets and selection sets of 100 ZINC compounds. Most individual predictions closely mapped the linear function established in Figure 1, yielding a correlation coefficient of 0.82 and an average absolute prediction error of 16%. Figure 2b reports the average results of these calculations and confirms the overall quality of the predictions (which also indicates that 100 individual trials with bait sets of different compositions provided a statistically relevant sample). For mean predictions, a correlation coefficient of 0.90 was observed, and the average error over seven classes was 13%. Again, similar results were obtained for the predictions on selection sets of 10 and 1000 compounds, as reported in Supplementary Figures 1b and 2b (Supporting Information), respectively. For 10 and 1000 database compounds, the correlation coefficients for predicted and measured recovery rates were 0.93 and 0.89, respectively. Thus, accurate predictions were obtained for differently sized compound selection sets, consistent with the finding that the logarithm of the KL divergence and recovery rates were directly proportional.

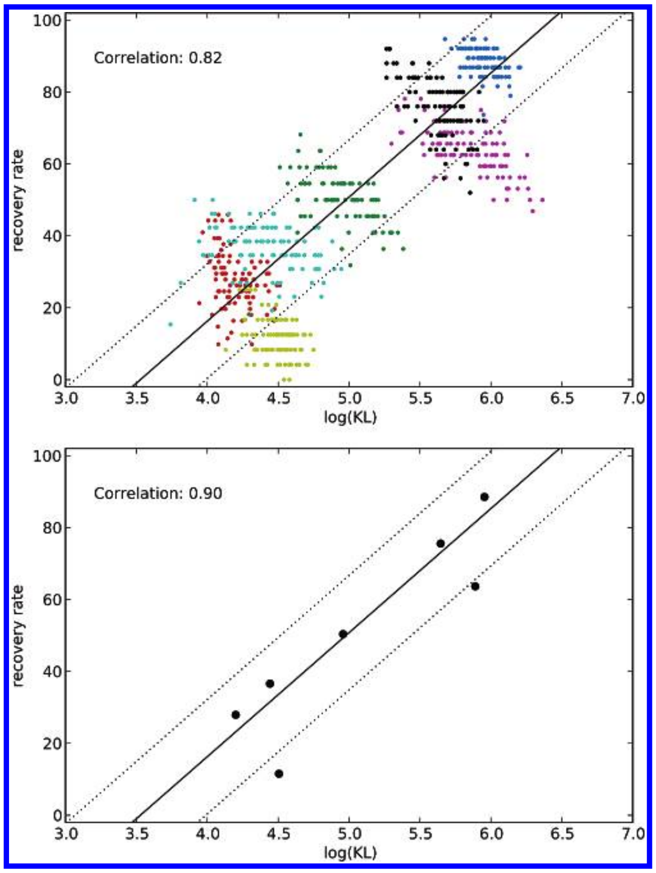

Figure 2. Prediction of recovery rates. For seven activity classes, recovery rates were predicted from KL divergence. The presentation is according to Figure 1. In a, 100 individual predictions are shown for each activity class, color-coded as follows: yellow, THI; red, 5HT; cyan, JNK; green, EDN; magenta, SQS; black, THB; and blue, HIV. In b, the mean values of the predicted recovery rates are reported for each class.

Table 2 compares the predicted and measured BDACCS recovery rates and reveals accurate predictions over the entire range. For five of seven classes, absolute errors were smaller than 5%, and their observed recovery rates ranged from 28% to 88%. The largest absolute prediction error of 22% was detected for class THI where the predicted recovery rate was too high, which could already be seen in Figure 2a. Nevertheless, even calculations producing the largest errors for the test cases studied here matched the magnitude of compound recall (low to high). Moreover, for other activity classes such as EDN or THB, recovery rates were predicted with high accuracy. Taken together, these findings indicated that the KL divergence had high predictive value for the outcome of our screening calculations.

DISCUSSION

We have introduced and tested a novel approach that combines Bayesian modeling with concepts from information theory in order to predict the recall of active compounds from screening databases on the basis of descriptor value distributions. Key aspects of the methodology are that (i) the BDACCS function can be expressed using only means and standard deviations of descriptor distributions, (ii) the KL divergence between descriptor distributions of known active and database compounds is proportional to the BDACCS score expected for active compounds, and (iii) the KL divergence correlates with recall rates of active compounds.

The assessment of compound recall is of critical importance for evaluating virtual screening methods. Applying the KL-BDACCS approach, we can estimate the limitations of test calculations that are due to descriptor value distributions. This should also help to evaluate the chances of success for other similarity-based methods that depend on chemical descriptor spaces and the evaluation of descriptor value differences. However, the potential of KL-BDACCS goes beyond benchmarking. For practical applications, we can predict expected compound recovery rates in light of chemical feature distributions of known active and database compounds. In other words, prior to practical virtual screening trials, it is possible to estimate how likely it would be to recover active compounds from a screening database, given its chemical features.

For such estimations, an important underlying assumption is that descriptor value distributions of sets of known active compounds mirror those of potential hits available in the screening database. This is generally true for structurally homogeneous activity classes. For structurally diverse classes, different sets of active compounds can be used in KL-BDACCS calculations to estimate the likelihood to recover diverse molecules having similar activity. As reported in Table 2, the activity classes tested here had considerable intraclass structural diversity, as reflected by low average MACCS Tc values. However, recovery rates could be well predicted, which emphasizes the ability of the KL-BDACCS approach to accurately capture feature distributions in structurally diverse compound sets.

It should also be noted that, through the incorporation of the Kullback−Leibler function, KL-BDACCS introduces an information-theoretic concept to the virtual screening field. Previously, another concept from information theory, the Shannon entropy formalism,[10] was adapted and extended[18] for feature profiling of different chemical databases.[19]

CONCLUSIONS

The KL-BDACCS approach is designed to predict compound recovery rates for Bayesian screening of unmodified high-dimensional descriptor spaces and can be more generally applied to estimate the probability of recovering active compounds from screening databases using molecular property descriptor-based approaches. We have presented the underlying theory and derivation of the KL-BDACCS approach and evaluated it on different activity classes. For five of seven classes used as test cases, KL-BDACCS calculations accurately predicted recovery rates from a 2D unique subset of the ZINC database with an absolute error of less than 5%, regardless of the magnitude of these recovery rates. For all seven classes, the magnitudes were well-predicted. We used a total of 40 different activity classes for calibrating recovery rate predictions; fewer would have been sufficient, considering the achieved correlation. Furthermore, the methodology performed well on differently sized compound selection sets and did not need to be optimized for specific sets. Our findings suggest that KL-BDACCS analysis should be a useful methodology to predict the outcome of virtual screening trials for benchmarking and practical applications. KL-BDACCS calculations might also help to further evaluate the relative performance of different methods by identifying test cases that represent varying degrees of difficulty in recovering active compounds.

Supporting Information Available: The derivation of the BDACCS function is provided. The 40 activity classes used to derive the function for the prediction of recovery rates are reported in Supplementary Table 1. Supplementary Figures 1 and 2 show functions calibrated for database compound selection sets of different sizes and the corresponding results of recovery rate predictions using these functions. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

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

(2) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Mol. Graphics Modell.* **2000**, *18, 343−357.*

(3) Sheridan, R. P.; Kearsley, S. K. Why Do We Need So Many Chemical Similarity Search Methods? *Drug Discovery Today* **2002**, *7*, 903−911.

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

(5) Godden, J. W.; Bajorath, J. A Distance Function for Retrieval of Active Molecules from Complex Chemical Space Representations. *J. Chem. Inf. Model.* **2006**, *46*, 1094−1097.

(6) Vogt, M.; Godden, J. W.; Bajorath J. Bayesian Interpretation of a Distance Function for Navigating High-Dimensional Descriptor Spaces. *J. Chem. Inf. Model.* **2007**, *47*, 39−46.

(7) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, 2000; pp 20−83.

(8) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley-Interscience: New York, 1991; pp 224−238.

(9) Kullback, S. *Information Theory and Statistics*; Dover Publications: Mineola, MN, 1997; pp 1−11.

(10) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana and Chicago, IL, 1963; pp 29−125.

(11) *Molecular Operating Environment (MOE)*, version 2005.06; Chemical Computing Group Inc.: Montreal, Quebec, Canada. http://www.chemcomp.com (accessed Nov 1, 2005).

(12) 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.

(13) *Molecular Drug Data Report (MDDR)*; Elsevier MDL: San Leandro, CA. http://www.mdl.com (accessed Sept 1, 2006).

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

(15) Godden, J. W.; Florence, F. L.; Bajorath, J. Anatomy of Fingerprint Search Calculations on Structurally Diverse Sets of Active Compounds. *J. Chem. Inf. Model.* **2005**, *45*, 1812−1819.

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

(17) *MACCS Structural Keys*; Elsevier MDL: San Leandro, CA. http://www.mdl.com (accessed Sept 1, 2006).

(18) Godden, J. W.; Bajorath, J. Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060−1066.

(19) Godden, J. W.; Bajorath, J. Chemical Descriptors with Distinct Levels of Information Content and Varying Sensitivity to Differences between Selected Compound Databases Identified by SE-DSE Analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 87−93.

CI600418U