Compound Set Enrichment: A Novel Approach to Analysis of Primary HTS Data

Thibault Varin,*,† Hanspeter Gubler,† Christian N. Parker,† Ji-Hu Zhang,‡ Pichai Raman,‡ Peter Ertl,† and Ansgar Schuffenhauer†

Novartis Institutes for BioMedical Research, Novartis Pharma AG, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland, and 250 Massachusetts Avenue, Cambridge, Massachusetts 02139

Received May 21, 2010

The main goal of high-throughput screening (HTS) is to identify active chemical series rather than just individual active compounds. In light of this goal, a new method (called compound set enrichment) to identify active chemical series from primary screening data is proposed. The method employs the scaffold tree compound classification in conjunction with the Kolmogorov−Smirnov statistic to assess the overall activity of a compound scaffold. The application of this method to seven PubChem data sets (containing between 9389 and 263679 molecules) is presented, and the ability of this method to identify compound classes with only weakly active compounds (potentially latent hits) is demonstrated. The analysis presented here shows how methods based on an activity cutoff can distort activity information, leading to the incorrect activity assignment of compound series. These results suggest that this method might have utility in the rational selection of active classes of compounds (and not just individual active compounds) for followup and validation.

INTRODUCTION

Primary HTS. High-throughput screening (HTS) has become widely used in the pharmaceutical industry to deliver chemistry entry points for drug discovery programs.[1] Primary HTS is the phase of a screening campaign where as many compounds as possible are tested to identify starting points for drug discovery projects.[2,3] Hits are usually defined as compounds whose activity exceeds a certain threshold value in a given assay.[4] However, the main goal of high-throughput screening is actually the identification of active chemical series, rather than just individual active compounds. The importance of identifying compound classes that show a range of potencies against the initial target is also reinforced when one considers that, during lead optimization (for either chemical genetic or drug development), properties such as solubility, attachment of linker groups, oral bioavailability, or other ADME (absorption, distribution, metabolism, and excretion) properties also need to be optimized.

Despite the high costs associated with high-throughput screening, analysis of the structure−activity data generated at this initial stage of the drug discovery process is still very restricted. In a typical pharmaceutical setup, a primary screen is conducted against the available screening collection at a single concentration, often just once. In many cases, only a simple activity cutoff is used for the selection of primary HTS hits[4−6] for which the activity is then confirmed and quantified by measurement of the concentration−response curve (CRC). Typically, compounds containing known undesirable structural features are removed from the hit list before CRC determination.[7] Usually, only very little structure−activity relationship (SAR) analysis of the chemical structures is conducted before the CRC measurement. Only after determination of the CRC are the compounds grouped into chemical classes in order to allow preliminary SAR analysis (see Figure 1). However, if the aim of high-throughput screening is the identification of chemical classes with a range of biological activity, then the decision of which compounds to allow to progress for validation and CRC determination should make use of chemical class information, even at this early stage. This strategy should maximize the number of active classes identified by HTS rather than just the number of individual active compounds.

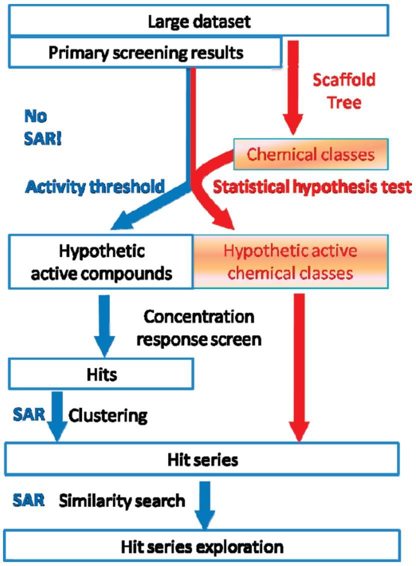
Large dataset
Primary screening results
Scaffold Tree
No SAR!
Chemical classes
Activity threshold
Statistical hypothesis test
Hypothetic active compounds
Hypothetic active chemical classes
Concentration response screen
Hits
SAR
Clustering
Hit series
SAR
Similarity search
Hit series exploration

Figure 1. Comparison of the analysis strategies commonly used (in blue) and that proposed in this article (in red) for the analysis of primary screening data.

Active Series Identification: From Gene Sets to Compound Set Enrichments. *Gene Set Enrichment and Other*

* Corresponding author e-mail: thibault.varin@novartis.com.
† Novartis Institutes for BioMedical Research, Basel.
‡ Novartis Institutes for BioMedical Research, Cambridge.

10.1021/ci100203e © 2010 American Chemical Society
Published on Web 11/12/2010

Methods of Class Enrichment. A number of methods have been described that seek to improve the possibility of selecting active series using modeling methods. The basic strategy of using the average activity of a set of related samples has been used extensively in analysis of gene expression studies, commonly referred to as "gene set enrichment", reviewed by Curtis et al.[8] This approach can be transferred to the domain of HTS data analysis by replacing gene sets with classes of chemical compounds that share some common structural features.

Identification of active chemical classes in primary HTS data requires performing two tasks. First, compounds in the data set need to be classified according to their chemical structures. Second, a determination needs to be made, for each chemical class, as to whether membership in the class increases the likelihood of a compound being active (see Figure 1).

Task 1: Chemical Classifications. The grouping of chemical structures can be accomplished in a number of different ways; in general, these methods can be described as clustering or classification. Clustering methods based on either chemical fingerprint similarity or maximal common substructure methods are often used.[9] Typical clustering methods depend on the whole data set; thus, even small changes in the composition of the data set also change the clustering outcome. Moreover, the computing time required for clustering does not scale linearly with the size of the data set, and therefore, this approach can often have prohibitive computational costs. Nevertheless, many HTS analysis applications published so far are based on clustering.[10,11]

Alternative classification methods are rule-based methods, which scale linearly with the size of data set and classify each structure individually in a deterministic way, focusing typically on a common chemical core scaffold. The best-known of these methods is the grouping by molecular frameworks (also known as "Murcko scaffolds"),[12] which are obtained by pruning all terminal side chains from the molecules. The molecular frameworks can be abstracted further by leaving out information about chemical elements and bond types to obtain the MEQI[13] keys, which have been used in the "data-shaving" method[14] of HTS data analysis or abstracted topology graphs.[15] Because of their high level of abstraction, these are chemically less straightforward to interpret than the original molecular frameworks. The LeadScope software[16,17] organizes the chemical data by a hierarchical classification using a predefined dictionary of structural features familiar to chemists, such as rings, ring systems, and functional groups. Each structure is typically assigned to several features.

The scaffold tree (ST) is a hierarchical classification of chemical scaffolds.[18] The molecular frameworks, as described by Murcko, form the leaf nodes in the hierarchy trees. By an iterative removal of rings, scaffolds forming the higher levels in the hierarchy tree are obtained. Prioritization rules ensure that less characteristic, peripheral rings are removed first. All scaffolds in the hierarchy tree are well-defined chemical entities, making this classification scheme intuitive to medicinal chemists.

The rational for using a scaffold-based classification scheme is the importance of the concept of molecular scaffolds for drug−target interactions and in the drug discovery process. Scaffolds contribute to several molecular properties that are fundamental for drug−target interactions. Furthermore, scaffolds define the basic shape of a molecule, determining whether a molecule is rigid or flexible. The compound scaffold also acts to position the molecule's substituents in their defined positions. In many bioactive molecules, scaffolds are directly involved in interactions with their target, either through heteroatoms forming hydrogen bonds with appropriate protein residues or through hydrophobic interactions.[19] Screening hits that share a common scaffold are therefore of high interest, with such HTS results possibly revealing initial structure−activity relationships. The identification of a class of compounds sharing a common scaffold has implications for synthetic expansion of the compound series, as well as for intellectual property rights.[20] Thus, the activity assessment of compounds based on their scaffold from the beginning of the hit-finding process should enhance the selection of hits from screening campaigns.

Task 2: Statistical Test for the Activity Hypotheses. The second task of SAR analysis is the evaluation of the hypothesis that a given structural feature, for example, a common scaffold S, is linked to the biological activity. If this hypothesis H_1 (alternative hypothesis) is true, then one would expect that the subset of compounds defined by the presence of S has a different distribution of activity readouts than the complete screening set. If this is not the case, then the null hypothesis H_0 is true. In this case, the activity distribution of the subset of compounds resembles the activity distribution of a randomly picked subset of the screening set. A wide range of statistical tests exist to calculate the probability (p value) that the null hypothesis is true based on the distributions of the activity data. The smaller the p value is for a given scaffold, the more significant the link between scaffold and activity on the target. The statistical tests can be grouped into two types: tests using binary data resulting from the categorization of compounds into "active" and "inactive" sets and tests using continuous activity data. When tests are performed on continuous activity data, one has the choice to use either tests assuming that the activity data follow a specific type of underlying distribution, such as the normal distribution, or so-called "nonparametric" tests that do not make such an assumption.

Analysis of Binary Activity Data. The use of primary activity data categorized into "active" and "inactive" sets is common practice in HTS data analysis. Traditionally, this was achieved using an arbitrary cutoff (say 50% of inhibition) often determined by capacity of confirmation test. Statistical measures have also been used for setting the cutoff, such as three standard deviations from the mean activity of the samples tested, which makes the assumption that the data are normally distributed. Based on binary activity data, structure−activity hypotheses can be evaluated with a binomial test. This test evaluates the hypothesis (H_0) that there is no difference in the proportion of active compounds for compounds having the scaffold S (ps) and the proportion of active compounds for the full data set (pd). Based on these statistics (ps and pd) and the number of compounds having this scaffold (n), the p value is computed using the formula given in Figure 2B.2. At a constant proportion of active compounds per scaffold, larger subset sizes will lead to smaller p values, which reflects the higher confidence gained by the increased subset size.

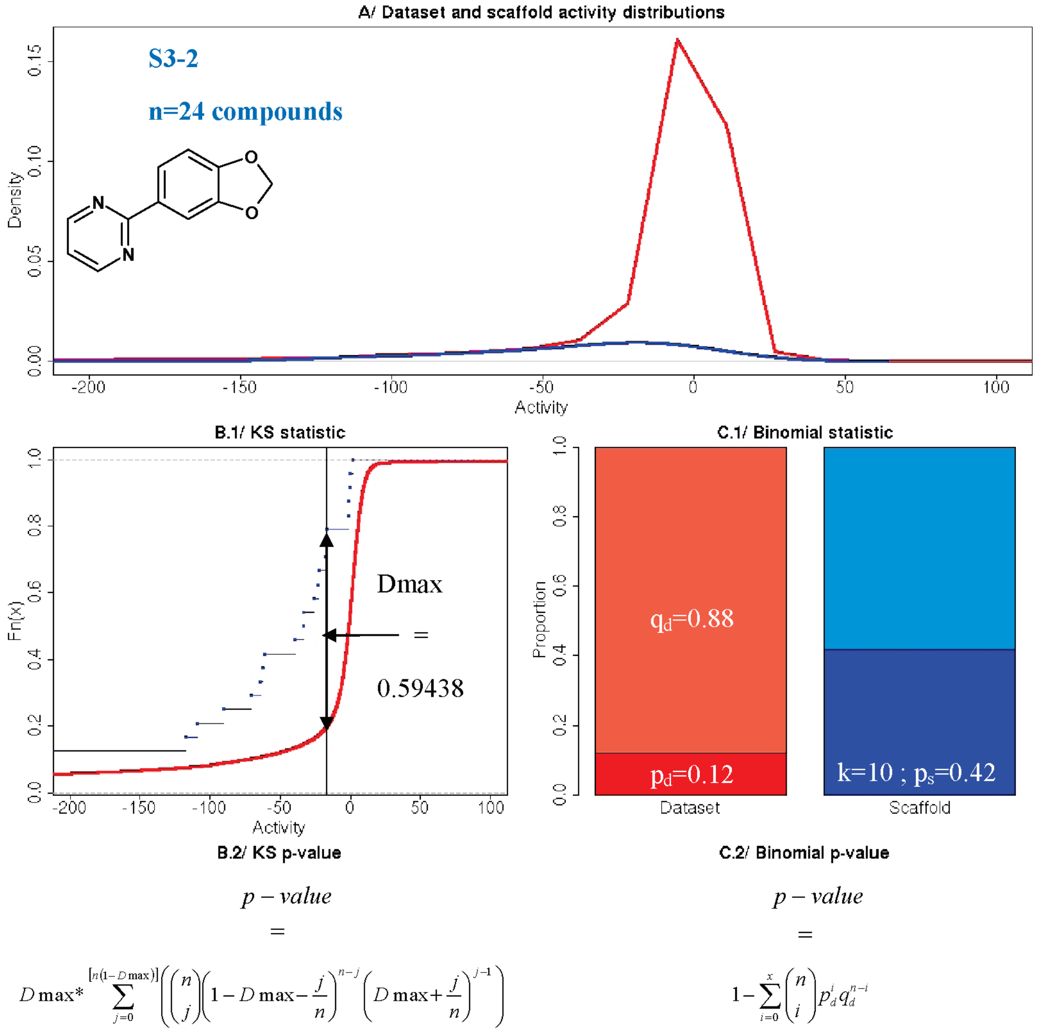

Figure 2. Application of the KS and binomial hypothesis test computation: statistics and p values. In these graphics, information relative to the full data set and scaffold S3-2 are presented in red and in blue, respectively. (A) Activity distributions for the full data set (red) and for the scaffold (blue). (B.1) Illustration of how the Kolmogorov−Smirnov (KS) statistic is calculated, showing the empirical cumulative distribution functions extracted for the full data set and the scaffold activity distribution. The activity for which the maximal distance (Dmax) between the bioassay and the scaffold empirical cumulative distribution function is observed is indicated by a vertical black line. (C.1) Illustration of how the binomial statistic used in this work is calculated. Dark colored areas represent the proportion of active compounds (compounds having a percentage of activity less than or equal to −50) for the full data set and for the scaffold activity distribution. Areas shown in light colors represent the proportion of inactive compounds. (B.2, C.2) Formulas used to compute the p value with the KS and the binomial method, respectively, where n is the number of compounds having the scaffold S3-2; $[n(1 - \mathrm{Dmax})]$ is the greatest integer contained in $n(1 - \mathrm{Dmax})$; k is the number of active compounds in the sample; x is the greatest integer less than k; and p_d and q_d are the proportions of active and inactive compounds, respectively, in the full data set.

Categorized activity data have been used successfully for applications such as naïve Bayesian classifiers and recursive partitioning.[21] However, reducing the actual quantitative screening readouts to a binary activity category results in a loss of information, because the differences in the activity readouts within the categories are discarded. Therefore, it would be preferable to asses the activity of a chemical class by comparing the complete activity distribution of the class members with the activity distribution of the remaining compounds.

Nonparametric Statistics. Assuming that the HTS activity readout values have a normal distribution, then the t test (or the z score as in the LeadScope software[16,17]) could be used to evaluate structure−activity hypotheses. Unfortunately, activity distributions of primary HTS results cannot generally be assumed to be Gaussian. It has even been suggested that primary screening results be described as a mixture with a Gaussian distribution for the inactive compounds and a gamma distribution for active compounds,[6] but it is not clear how such an approach can be generalized.

In this article, the use the Kolmogorov−Smirnov (KS) statistic[22,23] for the assessment of the activity of a chemical class is evaluated. The Kolmogorov−Smirnov statistic is a parameter-free method to empirically compare two determined distributions. In our application case, the null hypothesis (H_0) to be evaluated is that there is no difference in the activity distribution defined by compounds having the scaffold S and the background distribution. To obtain the p value in the KS statistic, the first step is to transform the scaffold and the background activity distribution (see Figure 2A) into empirical cumulative distribution functions (ecdf's) (see Figure 2B.1). Then, the maximal difference, Dmax, between the two ecdf's is determined. Based on the Dmax value and the number of compounds having this scaffold, the p value is computed using the formula given in Figure 2B.2. This is the formula for the one-sided KS test, where

Table 1. Data Sets Used in the Experiments

AID[a]	target	column[b]	P(AC)[c]	NC[d]	NAC[e]	NS[f]
893	hydroxysteroid (17-beta) dehydrogenase 4	28	0.0764	73919	5650	16136
900	caspase-1	52	0.0003	73919	20	16136
1634	pyruvate kinase	22	0.0006	263679	154	50936
411	luciferase (counterscreen)	31	0.022	71303	1571	13156
1379	luciferase (counterscreen)	52	0.0028	199080	565	41846
883	cytochrome P450 2C9	28	0.1358	9389	1275	1487
884	cytochrome P450 3A4	30	0.2629	13082	3439	2353

[a] PubChem assay ID. [b] Field number of the PubChem bioassay used to simulate the primary screening. [c] Proportion of active compounds. [d] Number of compounds. [e] Number of active compounds (according to the PubChem annotation). [f] Number of scaffolds with at least two compounds.

one looks for the shift in one direction. The direction of activity is typically known in an HTS scenario, and the two-sided test is usually less powerful than a one-sided test,[24] so the one-sided test will typically be the appropriate variant of the KS test for HTS data analysis. The higher the Dmax value and the larger the subset size n, the lower the p value will be. Thus, the increased confidence gained by an increased subset size is reflected in the p value. It is worth noting that the p value does not depend only on the activity value at which Dmax is found, making the test insensitive against any monotonic scaling operation of the activity values.

Level of Significance: Multiple Hypothesis Test Correction. The probability of having at least one false positive for a family of c tests is given by Šidák's equation,[25] $\alpha[\mathrm{PF}] = 1 - (1 - \alpha[\mathrm{PT}])^c$, where α[PF] is the critical significance level per family and α[PT] is the critical significance level per test. This means that, for $\alpha[\mathrm{PT}] = 0.01$, if one tests 100 true negative hypotheses, one will have a probability of $1 - 0.99^{100} = 0.63$ of obtaining at least one apparently significant false positive result.

To be protected from this effect, one strategy is to apply a multiple hypothesis correction in order to adjust the α level when performing multiple tests. A simple and popular multiple hypothesis testing correction is the Bonferroni correction, which suggests that the critical significance level, α[PT], be divided by the number of tests in the family to account for the number of tests being conducted.[25] This correction assumes that each individual test in the family is independent of all the others. Otherwise, the adjustments of α[PT] are likely to be too stringent.

Study Objectives. In this article, the scaffold tree compound classification method in conjunction with the KS statistic, for the comparison of each compound class activity distribution to the activity distribution of the whole data set, is described. The application of this method to seven data sets from the PubChem database is presented, and the ability of this method to identify compound classes with only weakly active compounds is demonstrated.

METHODS

Data Sets. We used seven data sets in our experiments, as detailed in Table 1. These data sets were downloaded from the PubChem database at http://pubchem.ncbi.nlm.nih.gov/.

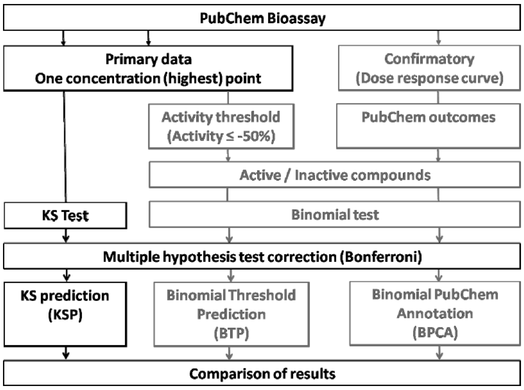

Figure 3. Illustration of the methods used to predict the scaffold activity from the primary data (KS prediction or KSP, binomial threshold prediction or BTP) and to evaluate the class activity from the PubChem annotation (binomial PubChem annotation or BPCA). Benchmark methods (BTP and BPCA) are represented in gray.

Concentration−response curves are available for all compounds tested in these bioassays. For these curves, the individual data points, the fitting parameters, and the activity categories (reported in the PubChem field "Outcome" for the bioassay activity outcome) are available. With these data, it is possible to simulate compound selection in a typical screening process consisting of a primary screen of the full screening collection at a single concentration followed by determination of the CRCs for selected primary hit compounds. The data points at one concentration were thus treated as a single-concentration primary screen. Based on these data, the activity of a compound class was predicted. The results of these predictions were then compared with the assay outcome derived from the CRCs as reported in PubChem. To detect weakly active compounds during HTS, compounds are usually tested at a high concentration. For this reason, our analysis focused on using the results from the highest concentration for which data were available. Activity measurements at concentrations at which only individual compounds had been tested were ignored. The molecules of the data set were processed to remove salts and to standardize charges and stereochemistry. The preprocessed data sets were classified with the scaffold tree.

Determining the Activity of Classes. *Protocol Overview.* The probability of a compound class being active was calculated in three different ways (see Figure 3). First, for each class, the activity p value using the single-sided KS statistics on the reported activity values at the highest screening concentration (KS prediction or KSP) was calculated. As a first benchmark for comparison with typical data analysis methods based on categorized activity data, the primary activity data were then categorized into active and inactive compounds. Use a statistical cutoff for this categorization such as the mean activity minus three standard deviations, which is common practice in many screening experiments, was initially considered. Using such a cutoff makes the underlying assumption that the screening data are normally distributed. To validate this assumption, the distribution of the activity values obtained at the highest screening concentration was tested for normality using the Lilliefors hypothesis test,[26] a modification of the KS test. According to this test, the activity values were not normally distributed ($p < 10^{-15}$ in each of the seven bioassays). These results suggest that the use of a statistical cutoff (such as three standard deviations from the mean) based on an assumed normal distribution of the activity values is not

Table 2. Determination of the Critical Significance Level for the Sixth First Levels According to the Bonferroni Correction

class level	number of scaffolds (NS)	α[PF][a]	α[PT][b] = α[PF]/NS
1	562	0.01	1.7794×10^{-5}
2	3380		2.9586×10^{-6}
3	6886		1.4522×10^{-6}
4	4033		2.4795×10^{-6}
5	1025		9.7561×10^{-6}
6	163		1×10^{-4}

[a] Critical level of significance per family (for the corresponding scaffold tree level). [b] Critical level of significance per test (per scaffold).

recommended, at least in the data sets studied here. Consequently, we used a cutoff of −50% for the activity classification with the binomial test (BTP) in all assays.

Based on this classification, a binomial hypothesis test was used to predict the scaffold activity (binomial threshold prediction or BTP). As a second benchmark, we calculated the scaffold activity *p* values based on the assay outcome annotated in PubChem, whereby all compounds not explicitly designated as active were considered to be inactive. Again, a binomial hypothesis test (binomial PubChem annotation or BPCA) was used.

This protocol enables the evaluation of two main hypotheses: first, comparison of results obtained from methods based on primary data (KSP and BTP) to results obtained from confirmatory data (BPCA) allows evaluation of the hypothesis that, within the primary HTS screening data, structure−activity relationships (SARs) are apparent and can be used to aid in the selection of active compound classes. Second, comparison of results obtained from KSP (dealing with continuous data) and BTP (dealing with binary data and thus requiring an activity cutoff) allows for the evaluation of the hypothesis that methods based on an activity cutoff distort the activity information, leading to the incorrect assignment of active series of compounds.

For each statistical test, the Bonferroni multiple hypothesis testing correction was applied. In the scaffold tree, the presence of a scaffold at a given tree level implies the presence of other, more general scaffolds at lower tree levels. Therefore, the scaffolds at different tree levels cannot be considered as independent. For this reason, the Bonferroni correction has been applied separately for each scaffold tree level. (For each level, the critical level of significance is equal to 0.01 divided by the number of scaffolds at this level; see Table 2.)

Software. The scaffold tree classification was precomputed for each molecule and stored in a relational database table together with all activity data obtained from PubChem. The activity data for the subsets defined by a scaffold were loaded into the R statistical system with the ROracle package (both available from http://www.r-project.org/). The statistical tests were then evaluated with the R statistics software using the respective functions in the R base package. The empirical cumulative distribution function for the complete activity data set of each assay serving as background distribution was calculated once using the ecdf, which was then used in each evaluation of the KS statistics for the assay.

RESULTS

The results of our study are described in detail for one example, the hydroxysteroid (17-beta) dehydrogenase 4 assay (AID 893). However, similar results were obtained for the six other bioassays and are given in the Supporting Information.

For bioassay 893, activity values correspond to a percentage of activity, and the lower this value, the more active the compound (inhibitor). For the other bioassays, please refer to the PubChem database at http://pubchem.ncbi.nlm.nih.gov/.

An evaluation of the scaffold population (number of compounds per scaffold) of the different data sets is provided in the Supporting Information (part I, Figures S1−S6).

Comparison of KSP and BTP Predictions. The overlap between the significantly active scaffolds according to the single-concentration data (KSP and BTP) and according to the PubChem activity annotation based on the concentration−response curve (BPCA) was analyzed for each bioassay (see Table 3). A more precise analysis was also done level by level (see Tables S1−S7 in part II of the Supporting Information).

The comparison of KSP2 and BTP2 with BPCA1 in Table 3 shows that both methods, KSP and BTP, generally perform well. They predict most of the BPCA significantly active classes as active, except for bioassays 900 and 1634 (comparison of BPCA1 to KSP2 and BTP2). To understand the differences between KSP and BTP, the difference, Δ2, between the number of BPCA-active scaffolds retrieved by each of the methods needs to be analyzed. Here, it can be seen that, for all assays except 1379, KSP recognizes at least as many BPCA- as BTP-active scaffolds.

Considering the total number of scaffolds predicted as active, KSP predicts more classes to be active than BTP (Δ1). The major contribution to Δ1 comes not from the additional scaffolds under Δ2, but from additional scaffolds predicted as active that are not active according to BPCA (Δ3). These scaffolds merit some further analysis. For this purpose, we examined the classification algorithm applied according to the assay description in PubChem for the concentration−response curves.

According to the automated outcome classification used in assay 893, the concentration−response curves were classified by the maximal inhibitory activity reached and the completeness of the curve. Curves were considered as complete if the two asymptotes were observed. Based on this criterion, the compounds were classified as active, inconclusive, or inactive. Especially when no full concentration−response curve with two asymptotes is observed, the maximal inhibitory activity is used as a classification criterion, which can be somewhat problematic. Typically, weakly active compounds showing only little maximal inhibition (activity < −80) and having only a partial curve are classified as "inconclusive".

It is instructive to visually inspect the superimposed concentration−response curves for all compounds of each KSP3 class. In Figure 4, we show these curves for four KSP3 classes. The full set of dose−response curve panels for the KSP3 classes of assay 893 can be found in the Supporting Information (part III.A), along with examples of such curves

Table 3. Number of KSP (KS Prediction), BTP (Binomial Threshold Prediction), and BPCA (Binomial PubChem Annotation) Significantly Active Scaffolds for the Seven Bioassays

	total				BPCA significantly active			BPCA nonsignificantly active		
bioassay	KSP1	BTP1	Δ1[a]	BPCA1	KSP2	BTP2	Δ2[a]	KSP3	BTP3	Δ3[a]
893	330	231	+99	199	183	168	+15	147	63	+84
900	331	114	+217	5	2	2	0	329	112	+217
1634	12	4	+8	12	3	3	0	9	1	+8
411	67	12	+55	15	13	11	+2	54	1	+53
1379	178	48	+130	41	32	35	−3	146	13	+133
883	58	33	+25	34	34	31	+3	24	2	+22
884	121	64	+57	60	60	53	+7	61	11	+50

[a] Δ = KSP − BTP significantly active scaffolds.

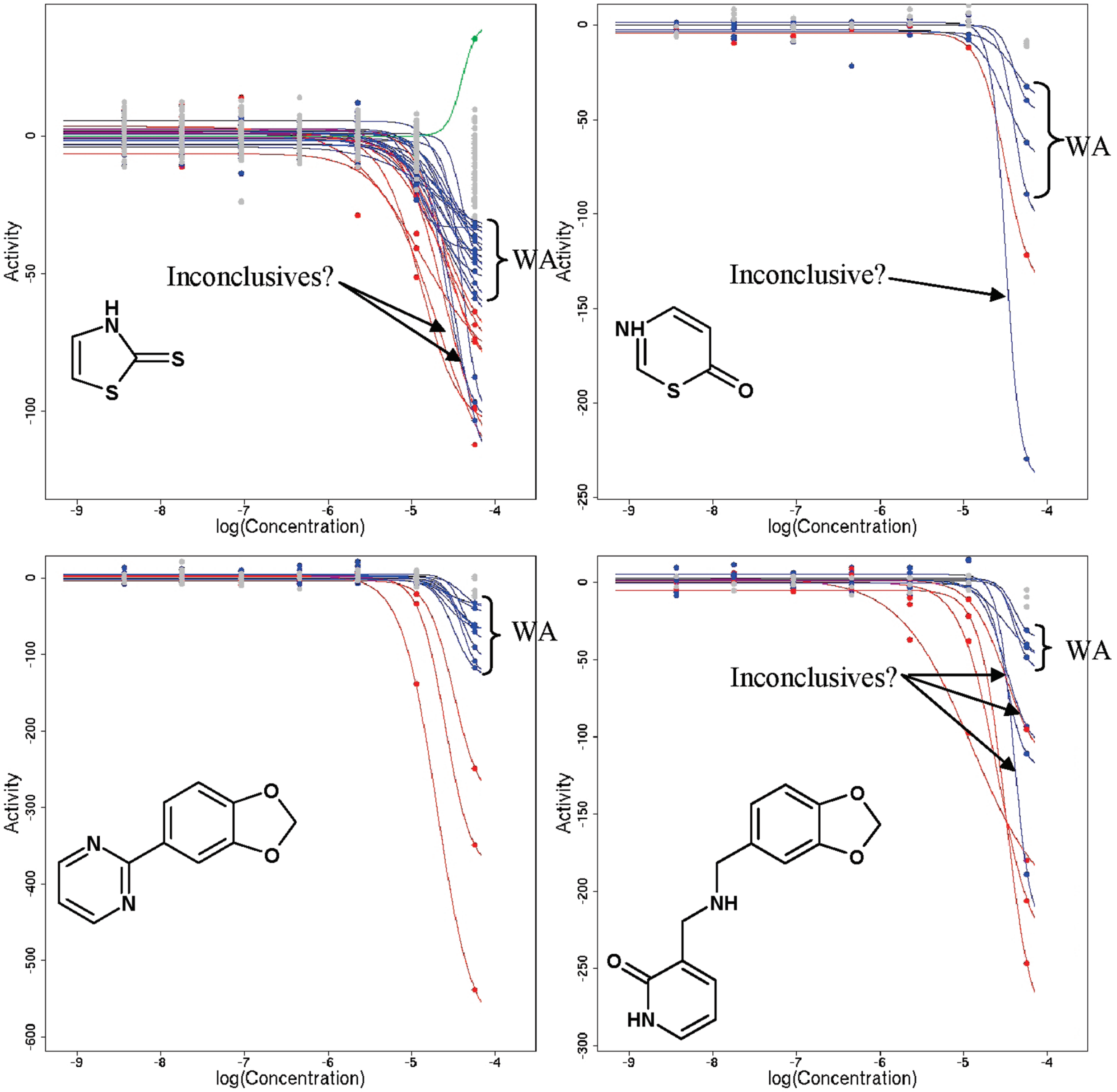

Figure 4. Concentration−response curves (CRCs) of compounds having the scaffold represented on each graph. Each of these scaffolds is significantly active according to the KS prediction but not significantly active according to the binomial PubChem annotation (BPCA). Active compounds are in red, inconclusive in blue, inactive in green, and all others for which we have no information on the CRCs are in gray (compound activity according to the PubChem annotation). The curves are plotted using the fitting parameters as reported in PubChem. Weakly active (WA) compounds are indicated by braces. Concentration is in molar concentration.

for the other assays (part III.B). In most KSP3 scaffolds, concentration−response effects can be observed for many of the scaffolds. The curves are often only partial, with low maximal inhibitory efficacy, indicating that the compounds are only weakly active. This activity pattern is usually observed consistently throughout the class.

In another type of KSP3 class, only very few compounds with a very high inhibitory activity are found. Because the

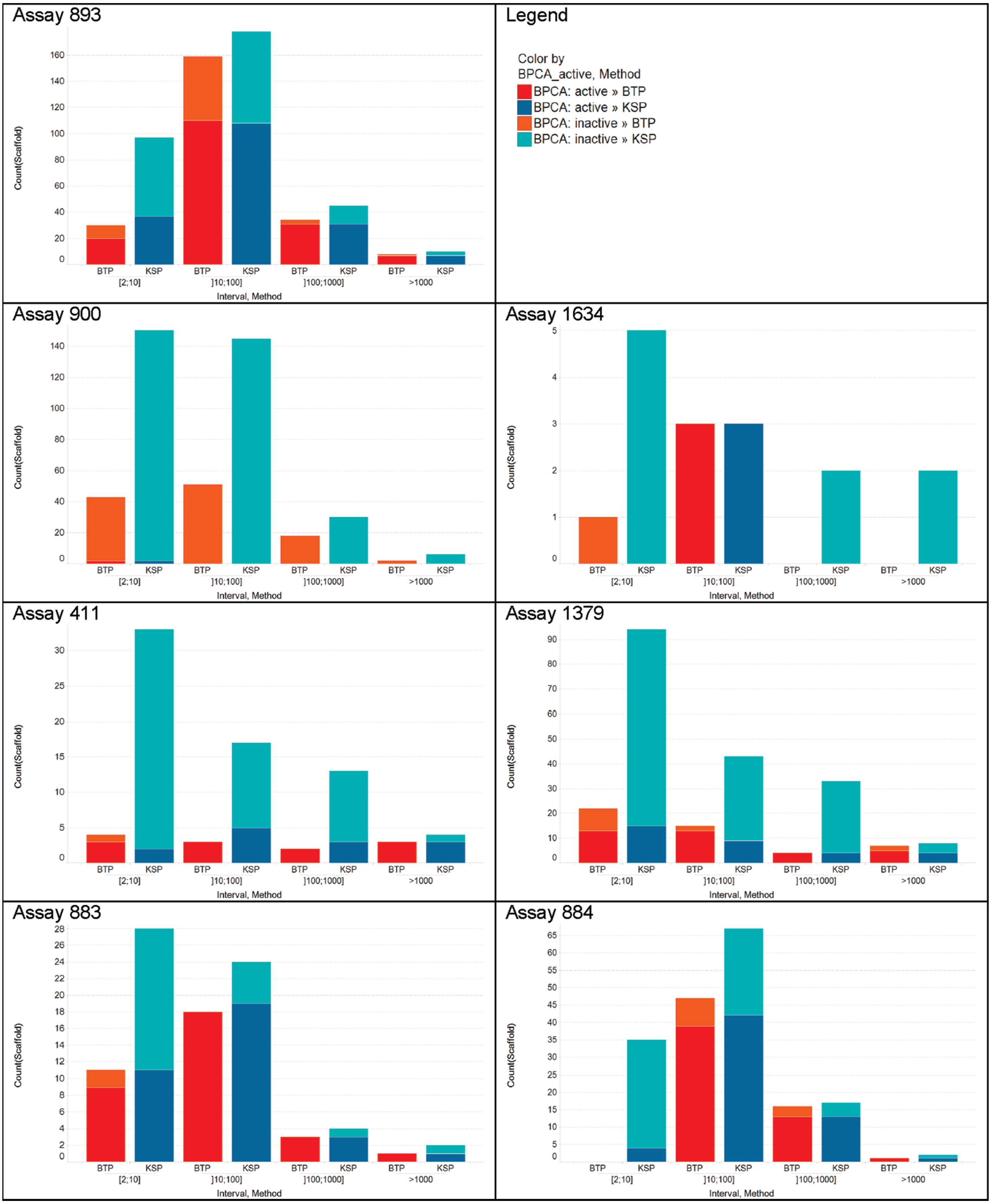

Figure 5. Number of classes predicted as active after Bonferroni correction by the KS prediction (KSP) and the binomial threshold prediciton (BTP) depending on the class size and the binomial PubChem annotation (BPCA).

number of compounds in this class is very small, the *p* value for this class is above the Bonferroni-corrected threshold of the α value. As in the BTP and BPCA, it matters only that the compound's efficacy is above the set threshold and that the compound has been classified as active, respectively; there is no influence of the actual activity on the *p* value. The situation in KSP, however, is different: When comparing classes with the same number of compounds, classes with compounds having a higher inhibitory activity give lower *p* values than those with only moderate activity. In Figure 5, the size distributions of the classes predicted as active by KSP and BTP are shown for each assay. It can be seen that KSP predicts more of the smaller classes as active, with up to 10 compounds as active, than BTP does. The same is true to a lesser extent for the class size range from 10 to 100. Whereas most of the classes in the size range of 2−10 are

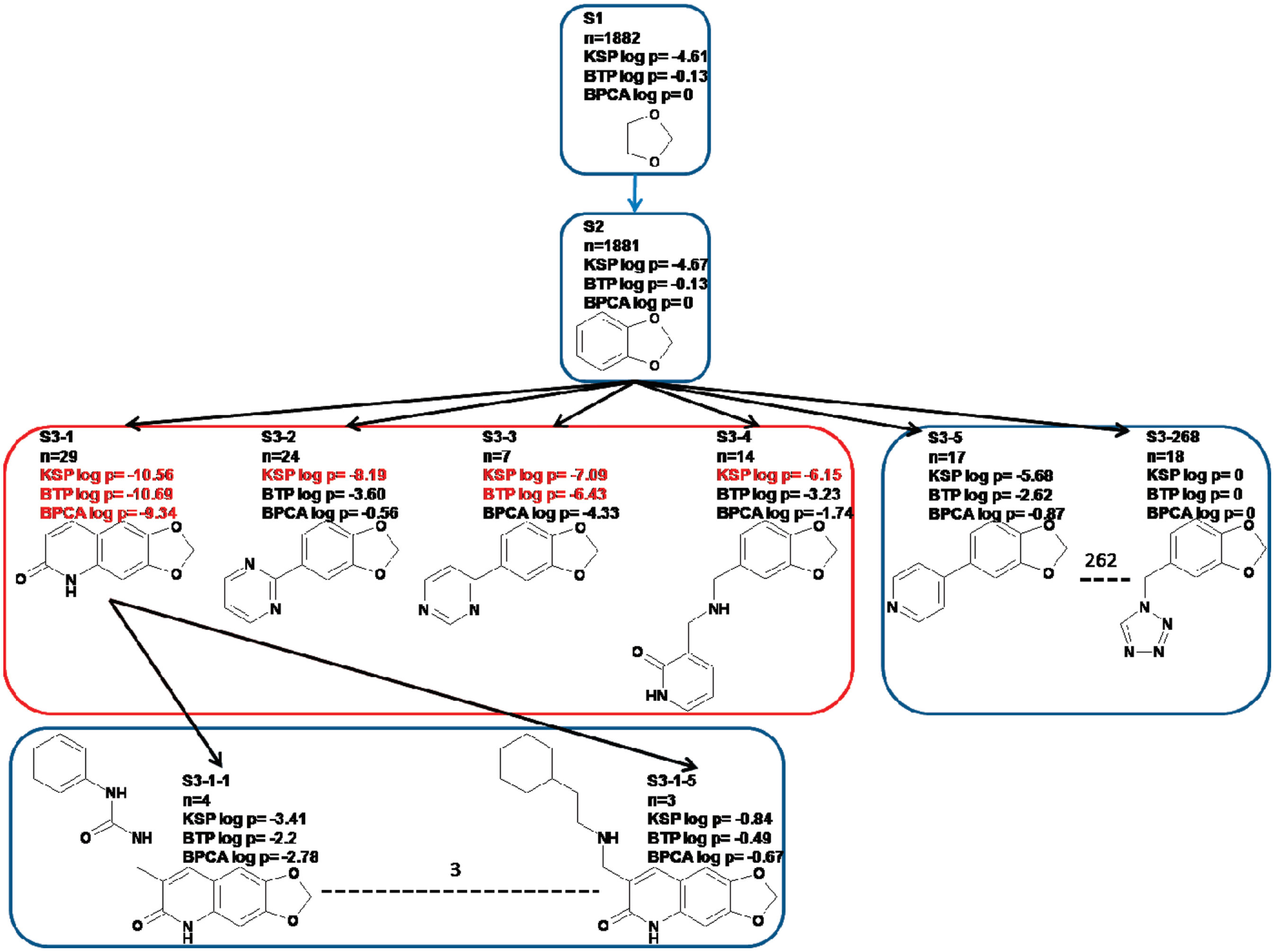

Figure 6. Illustration of the scaffold tree activity exploration using the *p* value computed by KSP (KS prediction), BTP (binomial threshold prediction), and BPCA (binomial PubChem annotation). For each class, the numbers of compounds having this scaffold and the logarithms of the KSP, BTP, and BPCA *p* values are indicated. For each level, scaffolds are ordered according to KSP *p* value by increasing order from left to right. The numbers above the horizontal dashed lines indicate the numbers of scaffolds between the corresponding left and right scaffolds. The red box indicates the KSP significantly active classes (level of significance equal to 0.01 have been adjusted for each level according to the Bonferroni correction), whereas blue boxes indicate inactive classes. Significant *p* values after Bonferroni correction are highlighted in red.

predicted as active by KSP but are not predicted as active according to the PubChem curve annotation (BPCA), usually, more of the BPCA-active classes in the size range of 2−10 are predicted as active by KSP than by BTP.

Scaffold Tree Activity Exploration. The scaffold tree is a hierarchical classification based on the compounds' core substructures. BTP, KSP, and BPCA *p* values were computed for each scaffold, allowing the evolution of the activity toward the different branches to be calculated. To illustrate this approach, one of the data set branches is presented in Figure 6.

We observe that the branch starts to fork at level 2, into many smaller branches (268 scaffolds at level 3) having logarithmic KSP *p* values ranging from −10.56 to 0. In this branch, after the Bonferroni correction, four scaffolds have a significant activity according to KSP (S3-1, S3-2, S3-3, and S3-4), two scaffolds according to BTP (S3-1 and S3-3), and only one scaffold according to BPCA (S3-1). Moreover, even without the Bonferroni correction, these scaffolds are local *p*-value minima on the tree branch. These four scaffolds are critical nodes in this branch and can serve as starting hypotheses to develop active series of compounds. Choosing a more generic scaffold makes the class definition fuzzier and leads to the inclusion of too many inactive compounds, thus highlighting the ability of this tool to help identify and refine structure−activity relationships from primary screening data.

In Figure 6, the S3-1 sub-branches have been included to show that compound classes not showing significant activity (from S3-1-1 to S3-1-5) can merge into a class that is significantly active. There can be two reasons for this: If the classes are too small, the low number of compounds leads to reduced significance in KSP despite a large Dmax value. Another reason is that new compounds can be inserted directly at level 3. S3-1 consists of 21 compounds belonging to classes from S3-1-1 to S3-1-5 and 8 compounds for which the last level in the scaffold tree is the class S3-1 because they do not have any rings in addition to scaffold S3-2. Of course, both phenomenon can occur at the same time. Being the scaffold with the locally minimal *p* value on its branch of the scaffold tree makes S3-1 a critical node that should be explored by selecting new compounds in its sub-branches.

DISCUSSION

Over the past decade, several approaches from library design to statistical analysis of high-throughput screening results have been tested by the pharmaceutical industry to maximize the number of hits identified during HTS campaigns. Often, only individual compounds are selected through an activity threshold for activity confirmation. Instead of waiting for confirmatory data to extract SARs and identify active series of compounds, this report shows that preliminary SARs can be extracted from primary screening data. As a consequence, improved allocation of resources can be made even at this early stage of the screening process. This study is focused on the identification of active series of compounds defined by a common scaffold. The classification of compounds according to a common scaffold is justified by the importance of compound scaffolds for drug−target interactions and for hit-to-lead optimization efforts (SAR exploration, assessment of the intellectual property restrictions). At each subsequent phase of the drug discovery process, active series are eliminated because of intellectual property, ADME, efficacy, or toxicological issues. Thus, increasing the number of active series identified at an early stage is fundamental, as it can compensate at least partially for the attrition at later stages.

One argument used to justify not using primary data to extract SARs is the high proportion of false positives and negatives in such data sets. When considering the activities of compounds individually, this is a strong caveat; however, this concern is greatly reduced when several compounds of the same series have been tested. In fact, the pooling of compounds into defined groups can compensate for the lack of replicates in a primary HTS. This, of course, reduces the error in the scaffold activity evaluation in proportion to the number of compounds per scaffold. It has been proposed that imprecision is reduced by $100(1 - 1/\sqrt{N})\%$ with N the number of replicates.[27] Thus, if there are four replicates, the imprecision would be reduced by nearly 50%. Clearly, this is one advantage of aggregating compounds and can actually serve to reduce error due to imprecision introduce by process errors during screening.

Both KSP and BTP enable one to identify significantly active series of compounds using the primary data, but the KSP method is more efficient at identifying additional series of active compounds (330 and 231 active series of compounds were identified by KSP and BTP, respectively). Analysis of these results shows that KSP enables one to identify most of the BPCA-active classes and most of the BTP-active classes, but also many others. The Bonferroni correction, focusing attention on the critical level of significance per family (scaffold tree level) instead of per test (individual scaffold), dramatically reduces the number of series predicted as active. (A more detailed analysis of the effects of the Bonferroni correction for both KSP and BTP is provided in part IV of the Supporting Information.) However, as HTS generates huge amounts of data, it is still highly interesting to focus on the most active series. From the assay (number 893) described in this article, 330 active series were identified by KSP, and 74% of the total number of active compounds identified belong to these classes. Therefore, according to both the number of active series and the number of active compounds in each series, it appears that this correction is not as drastic as we had first thought.

The KS-based activity detection can be seen as an intermediate solution between the binomial activity detection and the selection of individual compounds based on their activity readouts. In the binomial test, the quantitative activity readouts are discarded after an activity category has been assigned, and only the number of active compounds in the structural class is of importance. On the other hand, the individual assessment of each compound based on its activity readout alone ignores the results of structurally closely related compounds that can increase or decrease the confidence in the individual activity readout. The KS test, however, makes use of the number of compounds in the classes, as well as their quantitative activity readouts. This enables the KS method to also identify classes with a only few compounds designated as active. Its main advantage, however, is that there is no need to define any activity cutoff for the individual compounds. This helps to overcome the difficult problem of fixing an appropriate activity threshold for an assay. This is especially challenging when no reference compounds for full inhibition or agonism are known that can be used in the assay. This method also allows the detection of classes with systematic weak activity, which are lost when only binary activity data are used and many compounds are below the activity cutoff.

There is the legitimate question of whether each of the classes identified as active by KS shows genuine activity toward the target. Especially in the cases of small classes with an unusually high or low activity readout, one might think that such classes might be false positive compounds interfering with the assay system in an unwanted, but still systematic way, for example, as cytotoxic, autofluorescent, or fluorescence-quenching compounds often do. There is no way by just analyzing the assay readout data to determine whether an assay readout is caused by a genuine interaction of the compound with the target or by an unwanted interference with the assay system. Also, having full concentration−response curves typically does not help in making this distinction. This problem needs to be addressed experimentally by conducting appropriate counter screens or orthogonal assays or by evaluating multiple readouts in high-content assay systems providing multiparametric readouts. Even if compounds are interfering with the assay system, it is also of value to detect them as early as possible in the screening workflow, because then, there is still the chance to adjust the screening procedure to mitigate the effects of unwanted interference, especially if such classes can already be detected during the assay development phase.

Even if a scaffold is suitable for interaction with a particular target, this is clearly not sufficient, as other substituents (side chains and functional groups) will be needed for optimal interactions with the corresponding target. Thus, it might also be necessary to have other chemical features arranged around this active scaffold for optimal activity. Even though more than one million compounds might be tested against a particular target (as happens routinely in the pharmaceutical industry), this number is small compared to the size of the virtual library space that is covered by current synthesis protocols. This virtual library space has sizes on the order of magnitude of $10^{11}-10^{12}$ compounds.[28,29] Therefore, the chances that an active

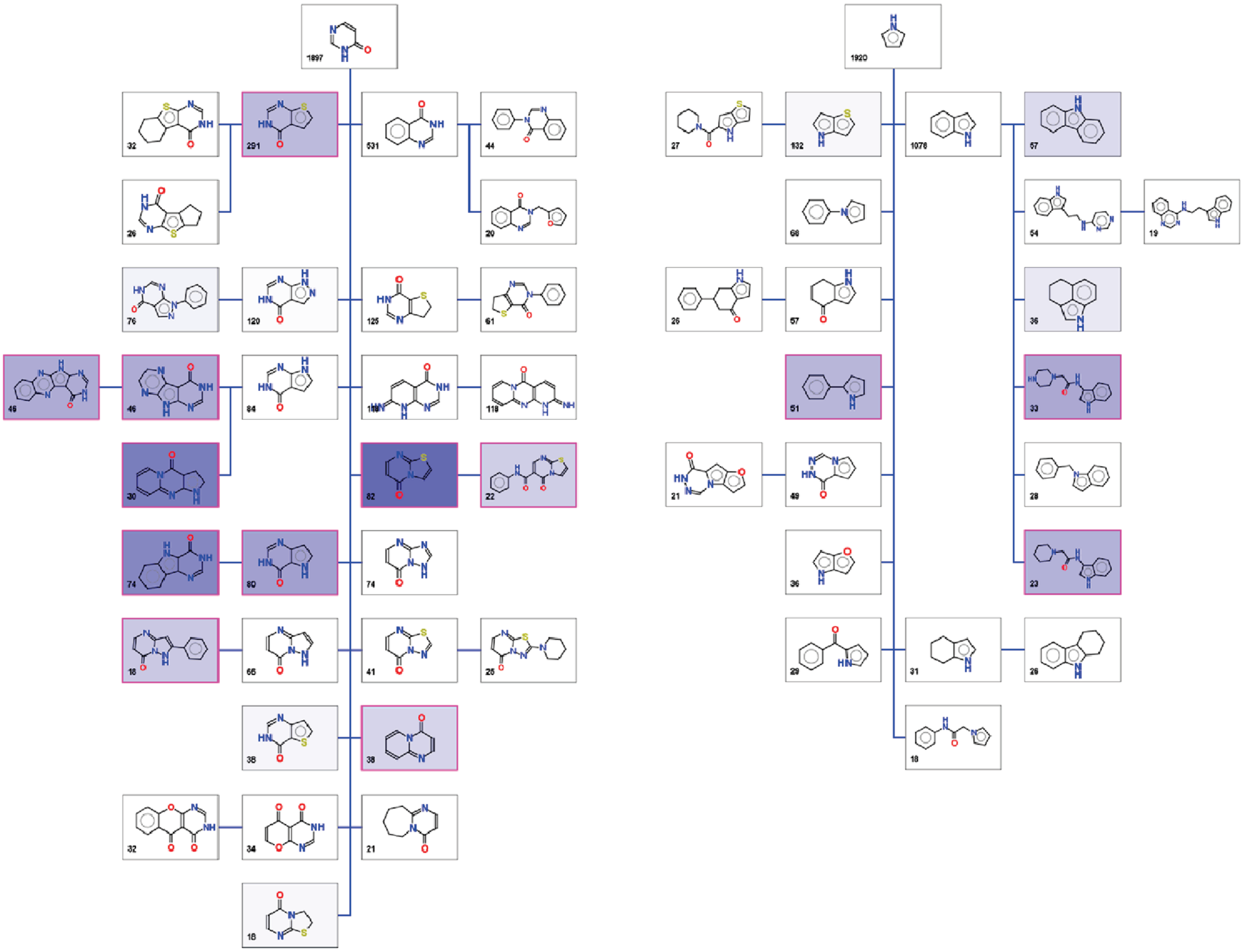

Figure 7. Visualization of the scaffold structures and activities of two scaffold tree branches (bioassay 893). Each scaffold having a *p* value less than or equal to 0.01 is represented in a cell with a blue background. The smaller the *p* value, the darker blue the background. Scaffolds significantly active with the Bonferroni correction are highlighted with a red border around the cell. The number indicated in each cell corresponds to the number of compounds having this scaffold. Only scaffolds having more than 15 compounds are represented.

compound scaffold with the optimal arrangement of side groups has been tested will be limited even with the careful design of compound screening libraries.[30] Scaffolds represented only by compounds with a suboptimal selection of side chains can be considered as “latent hits”.[31] Gaps in the bioactivity annotation of the scaffold tree constructed from structure−activity literature data[32] also indicate that some scaffolds in this tree mght have been sampled only by such latent-hit compounds. This study shows that the KS test identifies both type of classes, the ones with a high proportion of active compounds and the ones also having weaker active compounds but an activity distribution significantly different from the background distribution. These compound classes containing weaker actives are potentially latent hits that, with further optimization, could yield hits of the desired potency and selectivity.

In the past several years, visual tools to analyze screening data have become increasingly popular. Often, these tools display scaffold classes as a hierarchical tree and allow the users to color code the nodes of the tree by some property chosen by the user. Examples for such tools are Scaffold Hunter[33] or the hierarchy viewer recently published by Cho and Sun.[34] During analysis of structure−activity relationships, the property used to color the visualization is typically some sort of activity score for the scaffold. Sometimes, only the fraction of active compounds for each scaffold is used. This manuscript introduces the KS-based score as an activity score for a scaffold, which can be systematically calculated for all scaffolds in the screening set. These scores can be used for the color coding of activity in the tools mentioned above, during analysis of primary screening data. The large size of a primary screening data set typically requires some pruning of the tree to obtain a manageable visualization. Here, it has turned out to be practical only to display scaffolds that have a *p* value below a set significance threshold and have a minimum number of compounds. In addition, the scaffolds chosen according to these criteria and ancestor scaffolds of these selected scaffolds need to be displayed as well, even if they do not fulfill the significance criterion themselves, to obtain a well-connected tree. Although interactive visualizations of such trees are usually preferred in practice, we can show in this article only a static image of a branch of the scaffold tree encoded by the activity in assay 893 (Figure 7) generated by our internal scaffold tree visualization tool. More complete trees for the assays are given in the part V of the Supporting Information (Figure

S28 for bioassay 893 and Figures S29−S34 for the other bioassays). These visualization tools allow easy identification of critical nodes and provide intuitive SARs even at the primary screening stage.

CONCLUSIONS

This report describes the implementation of a method for the interrogation of primary screening data. By analogy to gene set enrichment used for microarray analysis, this method has been called compound set enrichment and allows the identification of classes of compounds with an activity profile that differs from the bulk of the other compounds screened. This method enables the identification of active compound classes even if the compound class does not contain compounds that are sufficiently active to be designated as such using a simple activity threshold. As a consequence, this method allows the identification of latent or hidden hits that can be further elaborated to show improved activity.

The method described here is based on previous reports describing the classification of compounds using a robust, reproducible, and chemically intuitive compound classification scheme (scaffold tree) and methods to use the aggregate activity of compounds or genes to identify sets of interest. However, this report describes the use of the KS statistic to compare the activity distribution of the compound class, and as we have shown, it is much more sensitive than other methods comparing the average activity of the compounds being studied. The other advantage of this method is that it does not require the researcher to set an arbitrary activity threshold when selecting hits for further analysis, making it simpler, less hazardous, and more efficient. It should be noted, of course, that this method is not merely an alternative but can also be used to complement traditional HTS hit picking, as described in Figure 1. Therefore, the common fear of missing singletons can be allayed.

By prioritizing groups of active compounds for followup and evaluation, it is hoped that effort will be focused on following compound classes that might offer the opportunity to identify novel SARs rather than just on individual active compounds (as can happen when considering just the most active compounds).

ACKNOWLEDGMENT

We thank the NIBR Education Office for the NIBR postdoctoral fellowship for T.V. We thank Martin Beibel, Sebastian Bergling, and Denis Levrat for insightful discussions on statistical topics. We thank also Meir Glick for the discussion on the concept of latent hits. The helpful advice of and discussions with the entire Molecular Library Informatics group are gratefully acknowledged.

Supporting Information Available: The Supporting Information is divided in five parts. The structure of the data sets is given in part I (number of classes according to the scaffold tree level for classes with at least two compounds and 10 compounds). The number of KSP, BTP, and BPCA significantly active scaffolds at each scaffold tree level and for each bioassay is provided in part II. The evaluation of the activity of scaffolds predicted as significantly actives by KSP but not by BPCA according to the CRC of compounds having these scaffolds is given in part III. Part IV describes the effects of the Bonferroni correction on KSP and BTP results. Part V contains scaffold tree activity visualization maps for the seven bioassays. This information is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Macarron, R. Critical Review of the Role of HTS in Drug Discovery. *Drug Discovery Today* **2006**, *11*, 277–279.

(2) Mayr, L. M.; Fuerst, P. The Future of High-Throughput Screening. *J. Biomol. Screening* **2008**, *13*, 443–448.

(3) Inglese, J.; Johnson, R. L.; Simeonov, A.; Xia, M.; Zheng, W.; Austin, C. P.; Auld, D. S. High-Throughput Screening Assays for the Identification of Chemical Probes. *Nat. Chem. Biol.* **2007**, *3*, 466–479.

(4) Zhang, J. H.; Chung, T. D. Y.; Oldenburg, K. R. A Simple Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J. Biomol. Screening* **1999**, *4*, 67–73.

(5) Gubler, H. Methods for Statistical Analysis, Quality Assurance and Management of Primary High-Throughput Screening Data. In *High-Throughput Screening in Drug Discovery*, 1st ed.; Hüser, J., Ed.; Methods and Principles in Medicinal Chemistry Series; Wiley-VCH: Weinheim, Germany, 2007; Vol. 35, pp 151−205.

(6) Buxser, S.; Chapman, D. L. Use Of Mixture Distributions to Deconvolute the Behavior of “Hits” and Controls in High-Throughput Screening Data. *Anal. Biochem.* **2007**, *361*, 197–209.

(7) Rishton, G. M. Reactive Compounds and in Vitro False Positives in HTS. *Drug Discovery Today* **1997**, *2*, 382–384.

(8) Curtis, R. K.; Oresic, M.; Vidal-Puig, A. Pathways to the Analysis of Microarray Data. *Trends Biotechnol.* **2005**, *23*, 429–435.

(9) Downs, G. M.; Barnard, J. M. Clustering Methods and Their Uses in Computational Chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1–40.

(10) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-Like Trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1069–1079.

(11) Yan, S. F.; Asatryan, H.; Li, J.; Zhou, Y. Novel Statistical Approach for Primary High-Throughput Screening Hit Selection. *J. Chem. Inf. Model.* **2005**, *45*, 1784–1790.

(12) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(13) Xu, Y. J.; Johnson, M. Using Molecular Equivalence Numbers to Visually Explore Structural Features that Distinguish Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912–926.

(14) Schreyer, S. K.; Parker, C. N.; Maggiora, G. M. Data Shaving: A Focused Screening Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 470–479.

(15) Pollock, S. N.; Coutsias, E. A.; Wester, M. J.; Oprea, T. I. Scaffold Topologies. 1. Exhaustive Enumeration up to Eight Rings. *J. Chem. Inf. Model.* **2008**, *48*, 1304–1310.

(16) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302–1314.

(17) Cross, K. P.; Myatt, G.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Blower, P. E. Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.* **2003**, *46*, 4770–4775.

(18) Schuffenhauer, A.; Ertl, P.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree−Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

(19) Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the Rings. In Silico Exploration of Ring Universe To Identify Novel Bioactive Heteroaromatic Scaffolds. *J. Med. Chem.* **2006**, *49*, 4568–4573.

(20) Webber, P. M. Protecting Your Inventions: The Patent System. *Nat. Rev. Drug Discovery* **2003**, *2*, 823–830.

(21) Xia, X.; Maliski, E. G.; Galliant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463–4470.

(22) Kirkman, T. W. Statistics to Use, 1996; http://www.physics.csbsju.edu/stats/KS-test.html (accessed Aug 13, 2008).

(23) Birnbaum, Z. W.; Tingey, F. H. One-Sided Confidence Contours for Probability Distribution Functions. *Ann. Math. Stat.* **1951**, *22*, 592–596.

(24) Siegel, S.; Castellan, N. J. The Use of Statistical Tests in Research. In *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed.; McGraw-Hill: New York, 1988; pp 6−17.

(25) Abdi, H. The Bonferonni and Šidák Corrections for Multiple Comparisons. In *Encyclopedia of Measurement and Statistics*, 1st ed.; Salkind, N. J., Ed.; Sage Publications: Thousand Oaks, CA, 2007;

Vol. 1, pp 103−107 (available at http://www.utdallas.edu/∼herve/Abdi-Bonferroni2007-pretty.pdf, accessed Aug 13, 2008).

(26) Lilliefors, H. On the Kolmogorov−Smirnov Test for Normality with Mean and Variance Unknown. *J. Am. Stat. Assoc.* **1967**, *62*, 399–402.

(27) Malo, N.; Hanley, J. H.; Cerquozzi, S.; Pelletier, J.; Nadon, R. Statistical Practice in High-Throughput Screening Data Analysis. *Nat. Biotechnol.* **2006**, *24*, 167–175.

(28) Boehm, M.; Wu, T. Y.; Claussen, H.; Lemmen, C. Similarity Searching and Scaffold Hopping in Synthetically Accessible Combinatorial Chemistry Spaces. *J. Med. Chem.* **2008**, *51*, 2468–2480.

(29) Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching Fragment Spaces with Feature Trees. *J. Chem. Inf. Model.* **2009**, *49*, 270–279.

(30) Schuffenhauer, A.; Popov, M.; Schopfer, U.; Acklin, P.; Stanek, J.; Jacoby, E. Molecular Diversity Management Strategies for Building and Enhancement of Diverse and Focused Lead Discovery Compound Screening Collections. *Comb. Chem. High Throughput Screening* **2004**, *7*, 771–781.

(31) Mestres, J.; Veeneman, G. H. Identification of “Latent Hits” in Compound Screening Collections. *J. Med. Chem.* **2003**, *46*, 3441–3444.

(32) Renner, S.; van Otterlo, W. A. L.; Dominguez Seoane, M.; Moecklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-Guided Mapping and Navigation of Chemical Space. *Nat. Chem. Biol.* **2009**, *5*, 585–592.

(33) Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive Exploration of Chemical Space with Scaffold Hunter. *Nat. Chem. Biol.* **2009**, *5*, 581–583.

(34) Cho, S. J.; Sun, X. Visual Exploration of Structure−Activity Relationship Using Maximum Common Framework. *J. Comput.-Aided. Mol. Des.* **2008**, *22*, 571–578.

CI100203E