Learning from the Data: Mining of Large High-Throughput Screening Databases

S. Frank Yan,*,[†] Frederick J. King,[†,‡] Yun He,[†] Jeremy S. Caldwell,[†] and Yingyao Zhou[†]

Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive,
San Diego, California 92121, and Novartis Institutes for BioMedical Research, 250 Massachusetts Avenue,
Cambridge, Massachusetts 02139

Received March 20, 2006

High-throughput screening (HTS) campaigns in pharmaceutical companies have accumulated a large amount of data for several million compounds over a couple of hundred assays. Despite the general awareness that rich information is hidden inside the vast amount of data, little has been reported for a systematic data mining method that can reliably extract relevant knowledge of interest for chemists and biologists. We developed a data mining approach based on an algorithm called ontology-based pattern identification (OPI) and applied it to our in-house HTS database. We identified nearly 1500 scaffold families with statistically significant structure−HTS activity profile relationships. Among them, dozens of scaffolds were characterized as leading to artifactual results stemming from the screening technology employed, such as assay format and/or readout. Four types of compound scaffolds can be characterized based on this data mining effort: tumor cytotoxic, general toxic, potential reporter gene assay artifact, and target family specific. The OPI-based data mining approach can reliably identify compounds that are not only structurally similar but also share statistically significant biological activity profiles. Statistical tests such as Kruskal−Wallis test and analysis of variance (ANOVA) can then be applied to the discovered scaffolds for effective assignment of relevant biological information. The scaffolds identified by our HTS data mining efforts are an invaluable resource for designing SAR-robust diversity libraries, generating in silico biological annotations of compounds on a scaffold basis, and providing novel target family specific scaffolds for focused compound library design.

INTRODUCTION

High-throughput screening (HTS) plays a central role in modern small-molecule drug discovery. Typically, an extremely large number of chemical compounds, either synthesized via combinatorial chemistry or derived from natural products, is screened in a single experimental setting against a biological target or pathway that is believed to be implicated in a certain disease. Candidates selected from HTS ultimately can be developed into drugs through the multistep optimization and development process. Thanks to recent advances in screening technologies and readily available large-scale compound libraries, it is now possible to screen several million compounds in a matter of days. Furthermore, the completion of Human Genome Project[1,2] provides a great number of potentially therapeutically relevant molecular targets that may be amendable to small-molecule intervention. Indeed, over recent years, numerous HTS campaigns throughout the pharmaceutical industry have already accumulated a large amount of data, substantially covering both the compound and target space. For example, the HTS database at the Genomics Institute of the Novartis Research Foundation contains a data matrix of over 1.6 million compounds across approximately 200 HTS assays. Data obtained from individual HTS campaigns typically are analyzed to identify interesting compounds that lead to tractable starting points for medicinal chemistry development. This hit-to-lead process has been assisted by a wide range of available cheminformatics tools developed over the years.[3−7] However, despite the general belief that the potential value of a corporate HTS database should exceed the sum of all individual data sets, only a limited number of studies that systematically mine large HTS databases in their entirety have been published. This is probably due to the fact that it is difficult to extrapolate methods that were developed for analysis of individual HTS data sets (e.g., recursive partitioning) to the results from an array of assays. The activity profile across various available assays is often merely used as a visualization component to help scientists filter out nonselective primary hits (by eye) with no application of more sophisticated data mining techniques. Beyond serving as a central data repository, knowledge discovery, defined as "...the nontrivial extraction of implicit, unknown, and potentially useful information from data...",[8] from the ever-growing corporate HTS databases has become one of the major cheminformatics challenges in modern day drug discovery.

The rapidly evolving field of chemogenomics[9,10] aims to study the large-scale relationships between chemical compounds and biological targets. In an effort to derive relationships between anticancer compounds and biological targets, Weinstein and colleagues employed a clustered image map (CIM) approach based on compound and activity profile clustering.[11] The color-blocked regions in the resultant CIM indicate potential correlations between compounds and targets. However, the CIM method lacks rigorous statistical control, and the putative relationships derived from such

* To whom correspondence should be addressed. Phone: +1-858-812-1896; fax: +1-858-812-1570; e-mail: syan@gnf.org, sfrank.yan@gmail.com.

[†] Genomics Institute of the Novartis Research Foundation.

[‡] Novartis Institutes for BioMedical Research.

10.1021/ci060102u CCC: $33.50
Published on Web 09/28/2006

clustered images are susceptible to being simply caused by the randomness in the data. It also often leads to fragmented compound clusters, i.e., compounds belonging to the same chemical scaffold are scattered around the map. This approach was further extended to include molecular descriptors or structural fragments as the basis of compound characterization—by pooling together compounds that share the same structural feature or substructure.[12] This modification greatly enhances the sensitivity and has led to the identification of two interesting anticancer scaffolds. We believe, however, that the limitation of this approach lies in its summation of the contributions over all the compounds that share a similar structural feature without taking into account the probabilistic nature of a structure−activity relationship (SAR), that is, not all the structurally similar compounds are expected to share the same biological activity profile. Without being able to identify and eliminate these SAR-related outliers, the above approaches can only have limited discovery power when applied to large-scale HTS database mining.

Among the limited number of published work on multiassay HTS data analysis, a proof-of-concept study (42 targets × 584 compounds) carried out by Horvath and colleagues demonstrated the concept of generalized neighborhood behavior (i.e., structurally similar compounds may have similar biological profiles).[13] The authors developed the overall optimality criterion and consistency criterion, in conjunction with various descriptor-based structural similarity measures, to characterize structure−profile relationship (SPR) behavior in a multiassay setting.[13,14] Successful applications of similar analysis of compound HTS profiles across multiple assays have also been reported by others.[15,16] However, in these studies, every compound family is still treated in the same way without taking into account the fact that the strength of neighborhood behavior can also be scaffold-dependent.[7] We believe, on the other hand, that a successful HTS data mining method should be able to process each scaffold family individually based on its observed SAR/SPR robustness in the biological space rather than treating all the compounds as a single large collection.

Attempts have been made to mine the HTS data collection in order to decipher relationships between compound scaffolds and target families.[17−23] For example, the recently proposed homology-based similarity searching method is able to identify potential ligands for a new, homologous target based on both compound structural similarity and target homology similarity.[24] The self-organizing map (SOM) approach also has been applied to correlate chemical structures and biological responses in the case of anticancer drug screening.[23] A more comprehensive review can be found in the recently published book edited by H. Kubinyi and G. Müller.[9] It should be mentioned however that in most previous studies special attention was often paid to the relationships between compounds and “drugable” protein target families, such as G-protein coupled receptors (GPCRs), kinases, and proteases.[25,26] In contrast, few studies have focused on the correlations among compounds, assay formats, and screening readouts that may cause artifactual results. This is especially pertinent to a modern drug discovery HTS, where the specific target may not be known and the HTS data can be intrinsically noisy.[27]

In this study we attempted to identify compound scaffolds that appear to give technology-related screening artifacts or demonstrate target family specific activities from mining the corporate HTS database using rigorous statistical method. This topic is of great importance to the drug discovery program for two key reasons. First, general toxic compounds may show consistently high activities in many cell-based assays, while compounds that are known to form aggregates may also display misleadingly high activities in enzyme inhibition assays.[28] Biological and chemical artifacts cannot be easily pinpointed from individual assays; however, multiassay HTS databases allow for the possibility to identify and eliminate such promiscuous or unwanted compounds through data mining. Based on the “fail early, fail cheap” dictum in modern drug discovery, we intended to construct a knowledge base that can be applied as an effective filter for undesirable scaffolds during the lead identification phase. Second, because the methodology used to identify assay-related artifacts is also directly suitable for studying target class specific scaffolds, we carried out a proof-of-concept investigation to learn whether such compound scaffolds can be systematically identified by exploring the chemical space and associated HTS profiles covered in an HTS database. This type of characterization will be extremely helpful to guide the design of more focused screening libraries.

Recently, we have developed a novel ontology-based pattern identification (OPI) data mining algorithm that originated from research on gene function prediction using microarray gene expression profiles and the guilt by association (GBA) principle.[29] In our previous study, we further modified and extended this approach to HTS data analysis and successfully applied it to independent HTS data sets for primary hit selection based on the principle of SAR.[7] In the case of single HTS data analysis, the OPI-based method is able to automatically determine the subgroup of structurally similar compounds that also have close activity values based on a rigorous statistical framework and rank them with a probability score.[7] The capability of identifying compounds that not only are structurally similar but also demonstrate strong neighborhood behavior[30] makes OPI an ideal tool to leverage both the probabilistic nature and the scaffold specific nature of the SAR principle, which have been key limitations in the existing computational studies as discussed above. In the context of this large-scale HTS data mining exercise, the HTS activities against individual targets are replaced by activity profiles against a battery of assays. This novel OPI algorithm is shown to be particularly useful to automatically determine a subset of structurally similar compounds that also demonstrate the generalized neighborhood behavior,[13] i.e., identify a subset of similar compounds that show similar biological profiles. In this way, statistically significant correlations between compound scaffold and biological profile can be established. Then, well established statistical tests were applied to the biological activity profiles to separate the promiscuous hitters from the selective ones (in terms of screening technology) and to distinguish the target specific scaffolds from the target independent ones. Statistical tests such as the familiar two-group Student’s *t*-test are incorporated in many popular cheminformatics methods including e.g. recursive partitioning; on the other hand, multigroup statistical tests are generally found their applications in bioinformatics for identifying differentially expressed

Table 1. List of the High-Throughput Screens Used in This Study

classification	category	inhibition	induction
assay format	enzyme activity	12	1
	proliferation (cellular)	19	0
	reporter gene (cell-based assay)	19	23
readout	fluorescence	10	1
	alamar blue (fluorescence)	17	0
	luciferase	21	25
target type	GPCR	5	3
	kinase	10	1
	nuclear receptor	3	6
	protease	3	0

genes and their applications to cheminformatics problems have been rare. We demonstrated here that the mature statistical methods such as analysis of variance (ANOVA) and Kruskal−Wallis test can also be applied to HTS data mining to evaluate the statistical significance of observed differences among various assay formats, readouts, and target types, respectively, without having to design ad hoc approaches.

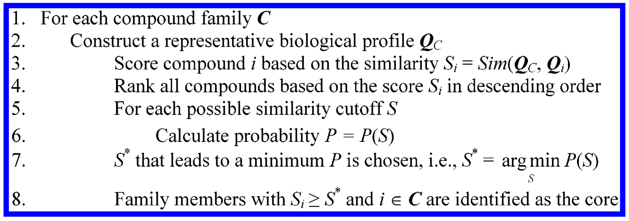
In summary, we present here a novel HTS data mining method and show how the large-scale HTS databases can serve as an invaluable source for systematic identification of SAR-reliable scaffold families that demonstrate general cellular toxicity, tumor cell cytotoxicity, potential reporter gene assay artifact, and target family specific activity. The extracted knowledge base, not easily obtainable by standard HTS analysis on individual assays, may benefit future compound acquisition, HTS assay development, and focused library design, as well as furthering our understanding on biological pathways. In the following sections, we first explain our corporate HTS database and the idea of the OPI algorithm. The key advantages of the OPI data mining method are illustrated by an example. Statistically reliable compound scaffolds showing signs of artifact due to assay format and/or readout as well as potential selectivity among target families are then presented.

METHODS

At the Genomics Institute of the Novartis Research Foundation, over 200 ultrahigh-throughput screening (uHTS) campaigns have been undertaken, spanning a wide range of disease areas and molecular targets. A variety of functional assays that are based on interrogating particular biological pathways using cultured cells were used in many of the uHTS campaigns. The resultant compound activities are not limited to a specific target. In particular, many screens employed functional assays based on cell proliferation (e.g., in oncology) and luciferase expression regulated by an autologous promoter. Therefore, fluorescence- and luciferase-based readouts were typically utilized in these screenings. Given the circumstances it becomes particularly important to analyze these HTS data with caution and statistical rigor, because of the intrinsically noisy nature of the data stemming from assay formats and readouts.

HTS Data Collection and Preprocessing. Every HTS campaign in our company database is annotated by its assay format, readout, target type, and expected signal direction of desired compounds. Each of these annotation fields can take values within a set of controlled vocabularies (see Table 1). A total of 74 HTS data sets, which have passed quality control and also maintain reasonable library coverage, were selected for this study. Table 1 shows the various classifications and lists the number of screens under the key categories. We also distinguish the screens based on inhibition and induction assays (i.e. signal direction). It should be mentioned, however, that not all the screens have complete annotation, and particular categories that contain too few screens are not shown (e.g., protein secretion assay format).

Chart 1. Outline of the OPI Algorithm to Identify the Core Members of Each Compound Family



```
For each compound family $\boldsymbol{C}$
    Construct a representative biological profile $\boldsymbol{Q}_C$
        Score compound $i$ based on the similarity $S_i = Sim(\boldsymbol{Q}_C, \boldsymbol{Q}_i)$
        Rank all compounds based on the score $S_i$ in descending order
        For each possible similarity cutoff $S$
            Calculate probability $P = P(S)$
        $S^*$ that leads to a minimum $P$ is chosen, i.e., $S^* = \arg\min_S P(S)$
        Family members with $S_i \geq S^*$ and $i \in \boldsymbol{C}$ are identified as the core
```

The activity measurement for each compound in an HTS was normalized and given a score between 0 and 1—the score 0 is assigned to compounds with no activity in the particular assay, and the score 1 is associated with the most potent compounds. For an inhibition assay, the percent inhibition value is directly taken as the activity score, while for an induction assay, the score is defined as 1−1/fold induction. Based on this definition, a compound with a 2-fold induction is assigned a score of 0.5, and the negative controls, i.e., no fold induction, have a score of 0.

From over 1 million compounds that are routinely screened in each of our HTS campaigns, we chose 33 107 compounds, each of which shows a score greater than 0.8 in at least one assay and has activity values for at least 80% of the abovementioned 74 assays (due to either quality control or logistic factors at the time of the screening). All 33 107 compounds were then clustered by an in-house clustering program developed based on the sphere exclusion clustering algorithm.[31] The compound structural similarity is measured by Tanimoto coefficient using Daylight fingerprints,[32] and a cutoff value of 0.85 was applied as suggested by previous studies.[33] As a result, 18 856 clustered compound families were obtained, among which 3834 were not singletons (containing more than one compound) with sizes varying from 2 to 116. Considering that HTS campaigns are subjected to various potential biological, chemical, and mechanical artifacts, singletons were not analyzed in this exercise due to their lack of statistical power.

The Ontology-Based Pattern Identification (OPI) Algorithm. The OPI algorithm, previously developed for gene functional annotation based on microarray expression profiles[29] and subsequently modified and applied to identify reliable primary hits from HTS,[7] was adapted to find large-scale correlations between chemical scaffolds and biological profiles from the large HTS data collection discussed above (74 assays × 33 107 compounds). Chart 1 outlines the OPI algorithm. Specifically, for a given compound family $\boldsymbol{C}$, it constructs a representative biological profile $\boldsymbol{Q}_C$ and subsequently identifies the optimal profile similarity cutoff S^* in a way that the family members sharing similar biological profile [i.e., profile similarity comparing to $\boldsymbol{Q}_C$ is greater than the cutoff value, $Sim(\boldsymbol{Q}_C, \boldsymbol{Q}_i) \geq S^*$] are statistically most enriched compared to all the compounds with profile similarity beyond that cutoff. In this way, this subset of compounds taken from the compound family ***C*** possesses similarities not only in their chemical structures but also in their biological activity profiles; to some extent, it is akin to the concept of "true similar" suggested by Horvath and colleagues,[13] while here the SAR outliers are automatically excluded based on a maximum statistical likelihood criterion (see below). The initial representative biological profile $\boldsymbol{Q}_C$ is either the profile of an individual family member that is most similar to the profiles of the rest family members or constructed simply as the average profile of all family members. As shown in Chart 1, both biological profiles can be employed as the query profile, whichever yields a better probability score is eventually used as the final representative for ***C***.

Assuming that there are a total of N compounds in the entire data set and given a compound family ***C*** of interest with N_C members, for an arbitrary similarity cutoff S based on the query profile $\boldsymbol{Q}_C$, we find n compounds that have profile similarities above the cutoff (i.e. $S_i \geq S$), among which n_C compounds are in the current family ***C***. Then, the odds of at least n_C family members are found among a list of n randomly selected compounds from a pool of size N is calculated by the accumulative hypergeometric score:

$$P = \sum_{i=n_C}^{\min(N_C,n)} \frac{\binom{N_C}{i}\binom{N-N_C}{n-i}}{\binom{N}{n}}$$

For example, given a pool of 33 107 (N) compounds and a family with five (N_C) members, if we choose six (n) compounds that share a similar profile and three (n_C) of them are found to belong to this family, the P-value is then 10^{-10}. The smaller the P-value, the less likely that the n_C member compounds happen to share a similar initial profile $\boldsymbol{Q}_C$ by chance, which implies the neighborhood behavior we observe for the n_C core members is most likely to be genuine. The optimal cutoff S^* is determined when the P-value reaches a global minimum. If one takes a cutoff below S^* and subsequently includes more family members into the core group, the P-value increases, indicating that the additionally recruited members are likely to be SAR outliers. The same argument also applies to the case when the cutoff values S is set to be greater than S^*—the P-value increases because of missing some of the "true similar" compound members that are both structurally and biologically similar to the existing core members. In this way, the OPI algorithm can identify the optimal subset of compounds that best demonstrate the neighborhood behavior by minimizing the probability of random enrichment. Furthermore, due to the iterative nature of this type of statistical test, we carried out permutation simulations by randomizing the compound activity profiles and repeating the above algorithm to avoid the "multiple test problem".[34] In the following section, we further illustrate how the OPI algorithm captures the probabilistic nature of the generalized neighborhood behavior, i.e., the principle of structure−profile relationship.

In addition, for any large-scale HTS campaign, the false positives can arise due to mechanical, biological, and/or chemical artifacts. For instance, artifacts caused by pipetting error or bacterial contamination usually show extremely high activities that may result in very low confirmation. In a previous study, we have shown how the OPI approach can be used to alleviate this problem by taking advantage of the built-in chemical redundancy in the HTS library.[7] Here we extended the OPI algorithm to work with a corporate HTS database rather than an individual assay, as the fundamental idea underlying the OPI algorithm is apt for both bioinformatics and cheminformatics applications individually as well as synergistically.

Kruskal−Wallis Analysis and Analysis of Variance (ANOVA). For any given compound family, its representative biological profile consists of an array of normalized activity scores across 74 assays, among which approximately 50 are inhibition assays. We grouped the activity scores from the inhibition assays according to the categories based on each type of classification (Table 1). For example, we grouped the compound inhibition activity scores into three categories according to assay format: enzyme activity, proliferation, and reporter gene (Table 1). ANOVA and Kruskal−Wallis tests are simply the multigroup version of the familiar parametric *t*-test and nonparametric Wilcoxin test, respectively.[35] The *t*-test and Wilcoxin test are only applicable in the case of two sample groups, while ANOVA and Kruskal−Wallis tests address multiple categories as in the current case. The null hypothesis here is that the compound families do not show differential activities among various categories, and a low probability value (usually defined as <0.01) rejects the null hypothesis, which indicates statistically significant differential activities among different categories. ANOVA test is a parametric test, which assumes the underlying data (the activity scores) follow a normal distribution within each category. Kruskal−Wallis test, on the other hand, is a nonparametric test that only relies on the rank information and is immune to whether the activity score distribution is normal or not. In this study, we required both probability values assigned by ANOVA and Kruskal−Wallis tests to be below our threshold in order to minimize the false correlations (insignificant correlations due to randomness) between compounds and biological profiles. We used the standard box plot, which shows median, lower, and upper quartiles information in a succinct manner, for visualization. It offers an effective visual tool to further distinguish the behaviors of the compound families in each category of interest.

RESULTS AND DISCUSSION

Identify Statistically Significant Correlations between Compound Scaffolds and Biological Profiles by the OPI Approach. To effectively identify compound families with strong neighborhood behavior (i.e. SPR), three major challenges need to be addressed: (1) Whether the correlation in biological profile demonstrated by the member compounds is simply because of statistical randomness or actually being driven by their structural commonality. An extreme example is that any two compounds could easily show perfect Pearson correlation (or anticorrelation) if the profile only contains two assays. In this case, the observed SPR is merely a random artifact. (2) Neighborhood behavior or SPR is a probabilistic rule, which means that a portion of the compound family members may not necessarily share the same similar profile with the rest, despite their structural commonalties. Therefore, it is often desirable, albeit difficult,

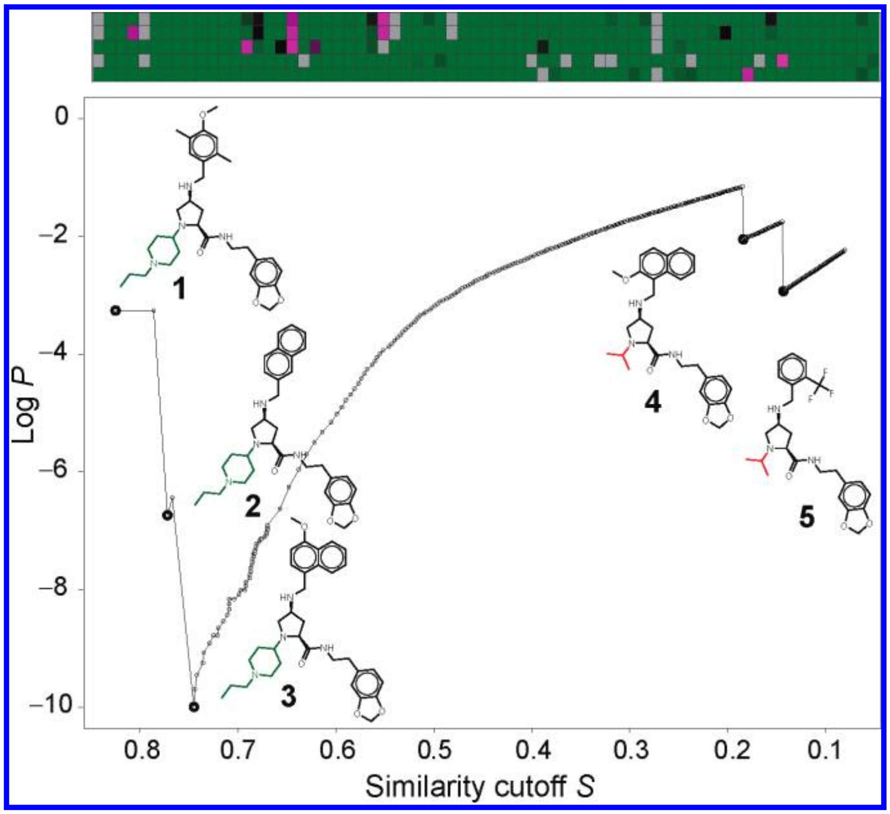


Figure 1. Illustration of the OPI algorithm. The compound family contains five members, among which three core members share a similar activity profile as well as similar chemical structure. The other two outlier compounds, albeit structurally similar, are effectively identified and excluded from subsequent analysis. Heatmap depicts the normalized assay activities of member compounds in the 74 assays. The row corresponds to compound, and the column is for assay. Magenta in the heatmap corresponds to a normalized activity score close to 1 (active), green is for a score close to 0 (inactive), and gray is for missing values.

to exclude such outliers and identify the core family members that preserve the SPR. (3) Most data mining algorithms employ some clustering analysis based on, individually, either compound structural similarity or biological profile, but not both. Clustering analysis such as the popular CIM method[11] is such an example. However, we found that compounds sharing similar scaffolds are often widely distributed in the clustered tree when the biological profiles are used as the clustering metric and the results are often too noisy to locate reliable, meaningful correlations between scaffolds and their biological effects. Others attempted to apply both structural and biological activity similarities to identify common tree components based on visual inspection.[36] It remains an open question as how to take advantage of both chemical and biological data simultaneously in knowledge extraction. We show here how the OPI algorithm addresses the above challenges and demonstrates that it is able to mine both biologically and chemically reliable data from the large-scale, often noisy HTS database.

As a simple illustration of the OPI approach, Figure 1 shows a compound cluster family with five members (compounds **1**−**5**). The heatmap shows the biological profiles of these five compounds across the 74 screens used in our study. If one uses the average profile of all five compound family members as the initial representative pattern $\boldsymbol{Q}_C$ and ranks all 33 107 compounds based on the profile similarity compared to $\boldsymbol{Q}_C$, the five compounds of interest are ranked at the second, fourth, sixth, 5162nd, and 6606th positions, respectively. As shown in Figure 1, the logarithmic *P*-value varies when more compounds are selected by effectively lowering the profile similarity cutoff value *S* (also see Chart 1 for detail). Specifically, if we choose two most similar compounds based on the biological profile compared to $\boldsymbol{Q}_C$, only compound **1** is included and the *P*-value is $10^{-3.5}$. If we continue to select four most similar compounds, two compounds **1** and **2** from this particular family are now included, and the *P*-value decreases to $10^{-7.0}$. Furthermore, when an effective cutoff of 0.745 is applied, six compounds whose profiles best correlate with $\boldsymbol{Q}_C$ are selected, three of which (compounds **1**−**3**) belong to our compound family. This enrichment is signified by a very low *P*-value of 10^{-10}. However, to include compound **4** from this family, we have to lower the cutoff value to 0.184 and select a total of 5162 compounds whose biological profiles, compared to $\boldsymbol{Q}_C$, are at least as similar as that of compound **4**. The corresponding *P*-value of 10^{-2} indicates that simply too many compounds outside this family of interest also have more similar biological profiles compared to the query pattern than compound **4** (a similar result also applies to compound **5** which ranks even lower at the 6606th). Therefore, these two members are considered as potential outliers. In this case, based on the minimum *P*-value the OPI algorithm automatically determines a family specific cutoff of 0.745 and selects compounds **1**−**3** as the core members of this scaffold family; their average profile is then passed on to the next stage of analysis.

This result is consistent with the biological heatmap shown in Figure 1. From a structural point of view, the three core compounds are different from the two outliers primarily in the groups attached to the nitrogen of the central pyrrolidine ring [marked in green (propylpiperidine) and red (isopropane), respectively, in Figure 1]. Although this difference is too subtle for the fingerprint-based clustering program to separate the two compound sets, with additional data from the biological activity profiles they are clearly distinguished using the OPI algorithm. The information we extracted here is that this scaffold family indeed shows a strong neighborhood behavior (P-value = 10^{-10}). On the other hand, for compounds that are more than 85% structurally similar to the core scaffold, only three out of five actually share the same representative activity profile. The OPI algorithm thus works aptly with the probabilistic nature of the SAR/SPR principle to identify the core subset of a compound family which shares bona fide, reliable structure−profile relationship within a group of seemingly similar compounds. Since the SAR principal that structurally related compounds share similar activity is the underpinning of any quantitative structure−activity relationship (QSAR) study and it has been recognized that more data do not necessarily help develop a more accurate QSAR model,[37] it will then be highly desirable if we can identify a subset of structurally similar compounds that do share bona fide SAR from mining the HTS data and further use them to develop possibly more reliable QSAR models.

Furthermore, diversity-oriented compound library design has been a popular approach for companies that wish to cover broad chemical space with limited screening effort.[38,39] This is also important for assays that are cost prohibitive or not amenable to high-throughput formats. It is a basic requirement that compounds in a diversified collection should well represent a lead island in the chemical space. As discussed above, the core members of a scaffold family identified by the OPI algorithm belong to a subset of compounds that carry statistically reliable SAR/SPR. We suggest that these representative core member(s) selected from each scaffold family identified by our OPI approach might better serve as candidates for constructing such diversity-oriented libraries, because of their capability in best capturing the SAR information with minimum structural redundancy.

Multigroup Statistical Tests of Categories in Different Assay Classifications. The compound activity represented by a fold induction value may not be directly related to its efficacy (e.g. in many reporter gene assays). Therefore, although the activity scores we used in the OPI analysis have been normalized, those derived from induction assays are not directly comparable across different assays either of inhibition or induction nature. In order to eliminate such inconsistency, we excluded all induction assays from the multigroup statistical tests. For any given compound family previously determined by the OPI algorithm to be statistically significant (a total of 1391 families), we first generated the average biological profile by taking the median of the normalized activity scores of all core members in each (inhibition) assay. We then grouped these median scores according to the categories in a certain assay classification (e.g. assay format, see Table 1). As shown in Figure 2, the activity scores of this compound family [GNF11742 (also see Table 3)] can be grouped, e.g. based on assay format, into three categories—enzyme activity assay, proliferation assay, and reporter gene assay. The activity scores in each category are represented by a box plot (Figure 2), where the central heavy line indicates the median activity value and the bottom and top of the box correspond to the 25 and 75 percentiles of the activity, respectively. The whiskers represent the minimum and maximum scores, and the outliers are shown as circles. The P-value calculated by either Kruskal−Wallis test or ANOVA test evaluates the null hypothesis that the core compounds from a scaffold have similar inhibition scores across different categories. In general, if the boxes vertically overlap significantly with one another, it indicates lack of selectivity among these categories; on the other hand, if the boxes are well distant from one another, some level of preference to certain categories does exist and is typically assigned a low P-value (by rejecting the null hypothesis). In this example, the logarithmic Kruskal−Wallis and ANOVA P-values are −7.80 and −9.45, respec-

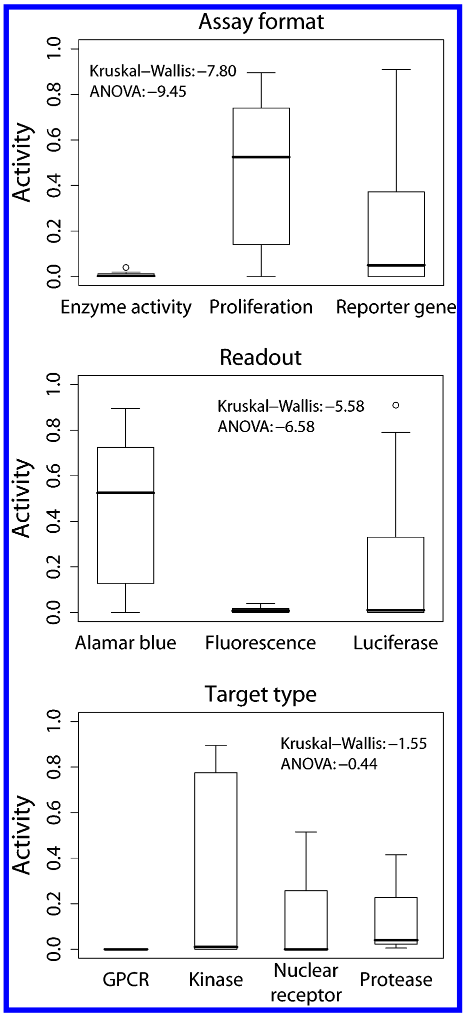


Figure 2. Box plots of categories in different assay classifications for compound family GNF11742. Assay format: enzyme activity, proliferation, and reporter gene; readout: alamar blue, fluorescence, and luciferase; target type: GPCR, kinase, nuclear receptor, and protease. The activity scores from the inhibition assays were used in the multigroup Kruskal−Wallis and ANOVA tests, and the logarithmic P-values are shown. The central heavy line in the box plot represents the median value, and the bottom and top of the box correspond to the 25 and 75 percentiles, respectively. The whiskers represent the minimum and maximum scores, and the outliers are shown as circles.

Table 2. Contingency Table for Assay Format and Readout

	readout		
assay format	alamar blue	fluorescence	luciferase
proliferation	**17 (4.1)**	0 (1.9)	0 (10.9)
enzyme activity	0 (2.7)	**8 (1.3)**	3 (7.1)
reporter gene	0 (10.2)	0 (4.8)	**42 (27.0)**

Table 3. Example Compound Scaffolds and WDI Annotations Identified by Data Mining

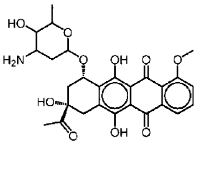
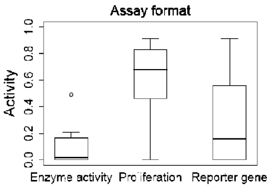
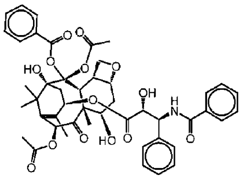
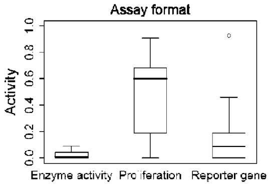
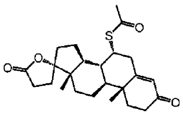
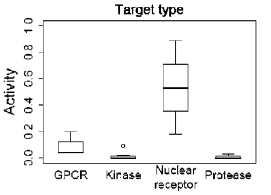
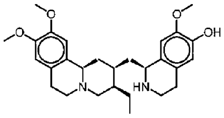
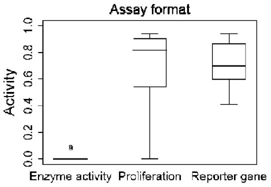
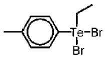
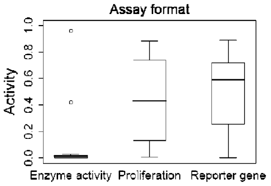
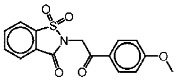
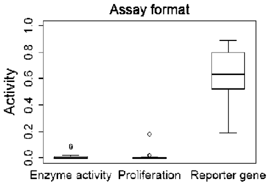
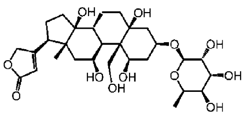
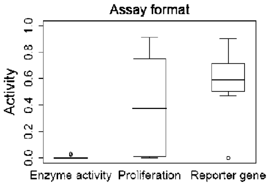
scaffold family	assay classification	log *P* (K.–W.)	log *P* (ANOVA)	WDI annotation[a]	WDI ID[b]
 GNF08103		−7.21	−9.89	Antibiotics; antimitotic; apoptosis-inducers; cell-proliferation-inhibitor; cytostatic-anthracycline; cytostatics; DNA-synthesis-inhibitor; inhibits nucleic-acid-synthesis by inserting into the DNA double helix; (see also Figure 4, panel A)	ACLARUBIC MA-144-M1 DR0121612 DR9605483 SPARTANAA ACLACINOB CINERUBIB DR9601547 CG-10 ACLACINOM
 GNF08454		−9.74	−10.9	Cytostatics; TNF-antagonist; synergists; apoptosis-inducer; antiarteriosclerotics; induces abnormal microtubule formation throughout the cell cycle; p-glycoprotein-inhibitor; MDR-modulator; prodrug; microtubule-stabilizer; prevents mitosis via promotion of microtubule formation and stability; (see also Figure 4, panel B)	DR9604300 DR9502824 DR0113859 DR9807280 DR0052446 DEACOTA10 NSC647752 DR0002550 DR0009666 DR9504919
 GNF08670		−4.91	−8.70	Aldosterone-antagonists; binds aldosterone-receptors of the distal tubule; diuretics; protein-synthesis-inhibitor; (see also Figure 5, panel C)	SC-26519 SC-26962 SPIRONOLA ZK-114082 SC-24813 HOSPIR11B SPIROXASO
 GNF09366		−12.4	−29.1	Acts primarily in bowel wall and liver; dopamine-antagonist; amebicides; antitussives; expectorants; inhibits polypeptide chain elongation in parasitic-cells and in mammalian-cells; anthelmintics; neuroleptics; analgesics	ACETARSEM EMETINHCL CEPHAELIN RO-1-9564 DIHTETRAB TETRABENA DR0115471 BENOLIZIM QUILLIFOL P-2565
 GNF28317		−5.49	−4.54	(see Figure 5, panel A)	
 GNF05008		−15.6	−36.6	(see Figure 5, panel B)	
 GNF09786		−8.55	−11.2	Antidiarrheics; cardiants; cardioglycosides; cytostatics; increases intracellular calcium-ion concentration; Na-K-ATPase-inhibitor; sedatives; apoptosis-inducers; vasoconstrictors; reported inotropic-agent	DIGITOXIN ACDIGOX12 ACDIGOXIN DESLANOSI DR9602976 UZARIN ACDIGITOX ACDIGOXAL DR0053329 DR0102321

Table 3 (Continued)

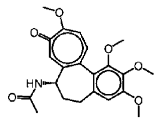
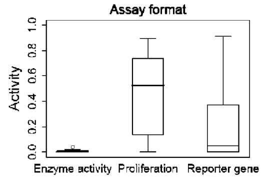
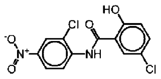
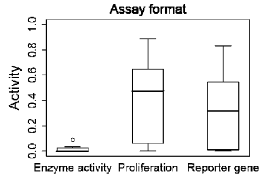
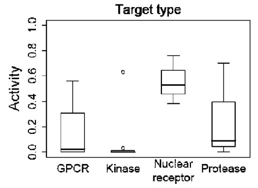
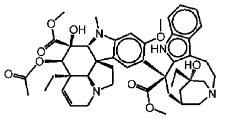
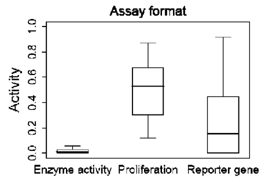
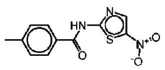
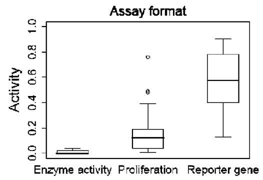
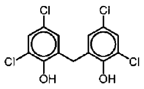
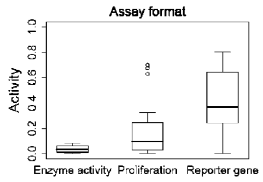
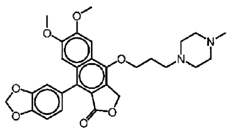
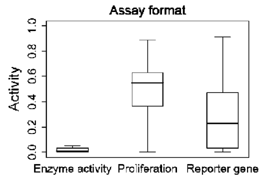
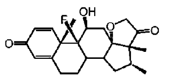
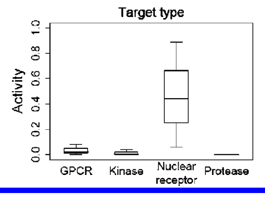
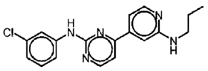
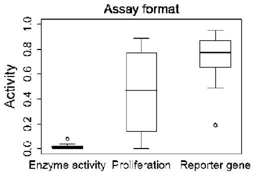
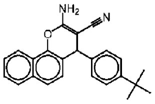
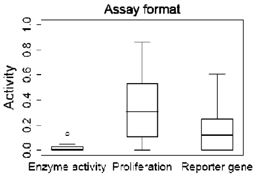
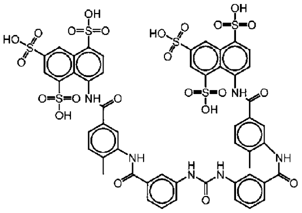
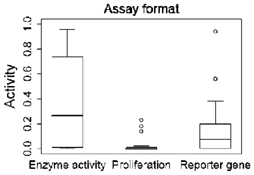
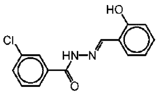
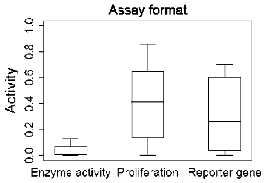
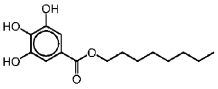
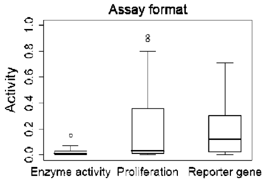
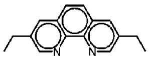
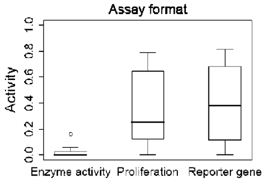
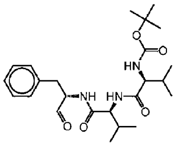
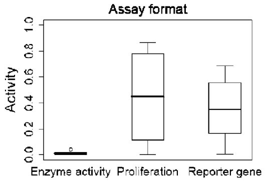
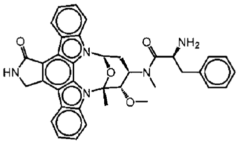
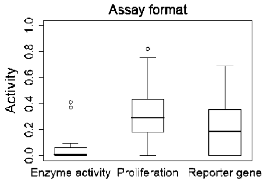
scaffold family	assay classification	log *P* (K.–W.)	log *P* (ANOVA)	WDI annotation[a]	WDI ID[b]
 GNF11742		−7.80	−9.45	Antiinflammatories; cytostatics; antirheumatics; reported microtubulin-inhibitor; immunosuppressive; interferes with kinin formation; leukocyte-migration-inhibitor; angiogenesis-inhibitors; antigouts	COLCHICIN COLCHICEI COLCHICID COLCHICEN DEMECOHCL DR0018984 DR9801624
 GNF13420		−5.86	−6.65	Anthelmintics; molluscicides; insulin-agonist; mitochondrial oxidative-phosphorylation-inhibitor in cestodes; pancreas-hormone; scolex and proximal segments are killed on contact	NICLOSAMI DR0036112 WR-39958 S-13
		−4.72	−4.62		
 GNF29061		−9.15	−10.6	Arrests metaphase so blocking mitosis; binds specifically to cellular microtubules of the mitotic apparatus; cytostatics; microtubule-inhibitor; leads to inability of the dividing cell to correctly segregate chromosomes; antimitotic; apoptosis-inducer; reported cell-proliferation-inhibitor; vinca-alkaloid; angiogenesis-inhibitors	VINBLAHCL VINBLASTI VINCRISTI VINEPIDIN VINROSIDI ANHVINBLA LY-203728 VINDESINE CATHARINI NAPAVIN
 GNF61834		−15.3	−19.2	Protozoacides	TOLAMIZOL DR9607986
 GNF45065		−8.17	−8.88	Activity against *Staphylococcus* and other gram-positive bacteria; antiseptics; bacteriostatic; anthelmintics	HEXACLPHE TRICLOPHE DICLPHEN CLOROFENE
 GNF43019		−8.54	−11.6	Antiinflammatories; virucides; antiaggregants; cytostatics; antiseptics	NEOJUSTIA DIPHYLLIN JUSTICIDA DR9701394 TAIWANINE DR9605338 DEHPODOPH
 GNF57395		−3.37	−5.82	Glucocorticoid-agonist; corticosteroid-antagonist; corticosteroids	DR0010115 DEXAMETHO

Table 3 (Continued)

scaffold family	assay classification	log *P* (K.–W.)	log *P* (ANOVA)	WDI annotation[a]	WDI ID[b]
 GNF63619		−12.4	−19.1	Caspase-inhibitor; protein-kinase-C-inhibitor; tau-protein-kinase-inhibitor; cytostatics	DR9712282
 GNF97290		−6.21	−5.99	Angiogenesis-inhibitor; antiarteriosclerotics; cytostatics; myocyte-proliferation-inhibitor; synergists	DR9507525 DR9507528
 GNF97986		−4.62	−5.96	Cytostatics; purine-antagonists; reverse-transcriptase-inhibitors; virucides; active against *Trypanosoma gambiense*; adenosine-antagonist; angiogenesis-inhibitor; anthelmintics; G-protein-antagonist; vasodilators; protozoacides	DR9506320 NF-037 DR9506403 DR0027076 DR0004198 DR0086451 DR9800511 DR9605294 NF-150 DR9800504
 GNF11743		−5.76	−6.18	Cytostatics; leukotriene-antagonists; chelator; peptide-hydrolase-inhibitor; protozoacides	DR9704464 DR0018837 DR9508229 ZIMET6185 DR9803137
 GNF90316		−3.12	−2.73	Antioxidant; antiseptics	DR0026655 GALLATEDE GALLATEBU DR9900876 AMYLPARAE
 GNF26372		−7.87	−6.70	Antiseptics	MANCO
 GNF37706		−6.98	−7.47	Substance-P-antagonists	DR9500016
 GNF78404		−3.75	−3.64	Antiaggregants; antibiotics; antiinflammatories; cytostatics; immunostimulants; protein-kinase-C-inhibitors; apoptosis-inducer; fungicides; hypotensives; radiosensitizers; tissue-factor-antagonist; tyrosine-kinase-inhibitor; immunosuppressives	DR0034518 DR0007510 DR9603493 OXOSTAUR7 CGP-41251 CGP-42700 UCN-01 DR0051400 DR9505817 TAN-999

[a] The WDI database used in this study is the first quarter release in 2005. [b] For simplicity only the 10 most similar WDI compounds are shown (Tanimoto similarity ≥ 0.85).

tively, for the assay format classification, indicating a statistically significant difference in activity for this scaffold among the three different types of assays, where the proliferation assay seems to have the highest activity score (Figure 2). On the other hand, the same respective *P*-values for the target type classification are only −1.55 and −0.44 for this scaffold (Figure 2), suggesting the core compounds from this group do not have specific activities that associate with the four target families considered here.

We believe these statistical tests provide a convenient tool to identify profiles of certain biological interest in a statistically rigorous fashion. These results also provide useful in silico annotations for the scaffolds in our database—the knowledge derived from mining the HTS database may become valuable to help scientists select better lead scaffolds, avoid generally toxic scaffolds (active in both proliferation and reporter gene assays), eliminate the ones that cause assay-related artifacts, and design new focused compound libraries. However, it should be mentioned that additional studies, such as literature searches, should be carried out before one can accept the in silico annotation, since the statistical tests only describe the data (by rejecting or accepting the null hypothesis) instead of providing a causality explanation.

Correlations of Various Assay Classifications. Table 2 shows a typical contingency table for the various assay formats and readouts. The frequencies of different categories were compiled for all the inhibition assays consistent with the above-mentioned statistical tests. The number of assays that fall into a given two-way category (e.g., proliferation assays that also use alamar blue as readout) is shown with the expected counts (in parentheses); the expected counts are based on the assumption that the two properties are independent. The diagonal counts in Table 2 clearly deviate from the expected values. The computed χ^2 is 117.5 with degrees of freedom of 4; therefore, the *P*-value is 10^{-12}. This extremely low *P*-value indicates the assay format and readout have a strong interdependence. Indeed, in our ANOVA and Kruskal−Wallis analyses, there is a strong correlation pattern that the compound that is active in proliferation assays also has higher scores in alamar blue assays—the same also applies to enzyme activity−fluorescence and reporter gene−luciferase pairs (Figure 2). Because of the strong correlation between assay format and readout, we focused on analyzing assay formats in our results, and the same conclusion for readouts is automatically implied. Based on the biological knowledge, we believe that for most cases compound preference shown in assay format probably is the cause and that in the readout may be just an outcome due to its association with an assay format. However, further experimental studies are required if one would like to understand the causality relationship between the two.

We repeated the similar χ^2 analyses for other classifications (Table 1) and found an additional unusual correlation: kinase assays tend to be associated more often with enzyme activity assays (8 vs 3.7 expected) and never with reporter gene assays. The *P*-value is $10^{-2.5}$. Nonetheless, since we have few kinase specific profiles as discussed below, this correlation does not affect our analyses presented here. In all, except for the readout, our biological classifications of assay format and target type can be considered as independent knowledge based on the above statistical χ^2 analyses.

Scaffold Families Identified by Data Mining with Statistically Reliable Neighborhood Behavior and World Drug Index (WDI) Cross-Annotation. By applying the OPI algorithm to the 3834 scaffold families that contain more than one member, we identified a total of 1391 compound families with significant statistical *P*-value scores that pass the permutation test. To gain additional knowledge about the scaffold families, we used the World Drug Index database (Derwent World Drug Index 2005/01, Derwent Information Ltd.) as an information source to annotate these chemical scaffolds. Specifically, we applied the core member compounds of each of the 1391 scaffolds as query structures to search the entire WDI database based on chemical structural similarity. The Daylight fingerprints[32] and a Tanimoto similarity coefficient cutoff of 0.85 were used. The available mechanism of action information of the similar WDI compounds was then used to cross-annotate the differential inhibition patterns of the corresponding scaffold family. Table 3 shows examples of the compound scaffolds obtained from our data mining efforts, together with their Kruskal−Wallis and ANOVA *P*-values, the box plots from the multigroup statistical tests, the WDI annotations, and the most similar WDI analogues.

Characterization of Compound Scaffolds Based on Assay Classification. As discussed above, we carried out the multigroup Kruskal−Wallis and ANOVA tests for all the 1391 statistically significant scaffolds according to the various assay classifications shown in Table 1. To effectively annotate the compound scaffolds with information derived from these analyses, we defined four types of compound scaffolds of our interest based on the observed box plots. The definitions are as follows: type **A** or tumor cytotoxic—the scaffolds that are primarily active in the proliferation assays; type **B** or general toxic—active in both proliferation and reporter gene assays; type **C** or potential reporter gene assay artifact—active primarily in the reporter gene assays; and type **D** or target family specific—active in a specific target family, e.g., nuclear receptor or GPCR. Because the majority of the proliferation assays in our database were utilized in the oncology disease area and oftentimes based on the tumor cell cytotoxicity studies, we characterized, in a broad sense, the scaffolds that are primarily active in the proliferation assays as tumor cytotoxic (type **A**). It should be mentioned that this type of annotation is just a general characterization of compound scaffolds, largely for the sake of filtering. If one were to pursue any scaffold in later steps of lead identification, it is obvious that detailed examination of the scaffold has to be carried out. Moreover, since many nonspecific tumor cytotoxic compounds, such as most chemotherapy agents, are also toxic to normal cells, we consider a compound scaffold may be generally cytotoxic given that it is highly active in many proliferation assays as well as cell-based reporter gene assays (type **B**). In addition, the reporter gene assays were widely used in our screening efforts, encompassing different molecular targets in various disease areas such as diabetes and metabolism, autoimmunity and transplantation, etc. We deem a scaffold that is active in numerous cell-based reporter gene assays as a potential reporter gene assay artifact (type **C**). However, we note that these four types of scaffold categories described here are only loosely defined. They merely suggest a property of a compound scaffold and certainly should not be regarded as

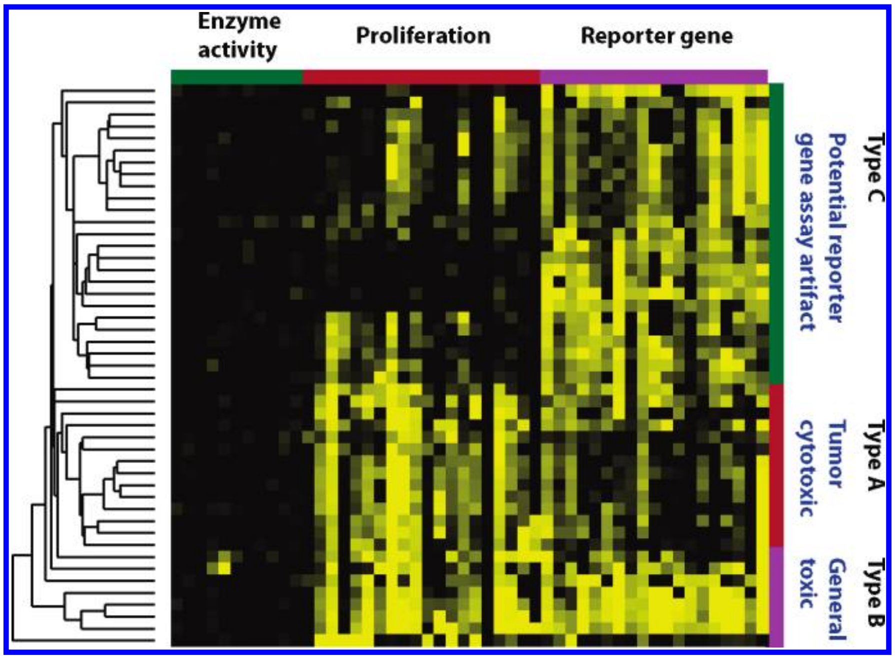


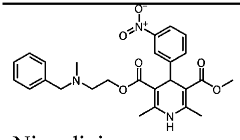
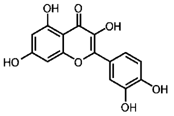
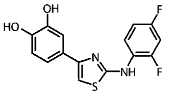
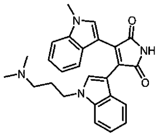
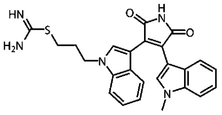
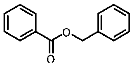
Figure 3. Overview of representative biological profiles of selected compound scaffold families. Each row in the heatmap represents the median assay activity profile for a differentially behaved scaffold family. The assays are sorted according to their formats; strong inhibition is represented by yellow and weak inhibition is in black. The hierarchical clustering revealed three generic patterns for compound scaffold based on assay format—general toxic (type **B**); tumor cytotoxic (type **A**); potential reporter gene assay artifact (type **C**).

definitive proof of their biological effects without further investigation. It is used as a convenient tool to annotate the corporate compound library on a scaffold level based on the information gathered from mining the HTS database, which cannot be otherwise easily obtained by examining individual compound HTS profiles. This kind of knowledge may help facilitate the HTS lead discovery efforts by filtering out undesirable scaffolds (e.g. general toxic ones), focusing on specific scaffolds for a particular target of interest (e.g. nuclear receptors), designing targeted library for a particular disease area (e.g. oncology), and more. Some of the proof-of-concept examples of this scaffold-based library annotation are discussed below.

Profiling of the HTS Data on a Scaffold Level. Based on the assay format, we selected a subset of 49 scaffold families that demonstrate statistically significant differential inhibition patterns in the enzyme activity, proliferation, and reporter gene assays—the scaffolds that are annotated as tumor cytotoxic, general toxic, or potential reporter gene assay artifact according to the above-mentioned scaffold characterization. We then hierarchically clustered them according to the HTS activity profiles based on the Pearson correlation coefficient. As shown in Figure 3, these selected scaffolds (i.e., types **A**, **B**, and **C**) are clearly grouped together with distinct activity profiles. It should be mentioned that in our analysis, each row in the heatmap is a representative activity profile of a scaffold family rather than that of an individual compound used in many previous studies such as those based on the CIM approach. In this way, the representative scaffold activity profile obtained by the OPI approach is supported by the multiple core member compounds—this arguably increases the robustness of the observed correlation between two different scaffold families. Following the identification of scaffolds with statistically significant neighborhood behavior using the OPI approach, this type of clustering analysis also provides an intuitive way to locate structurally diversified scaffold families that share similar biological profiles. It provides an opportunity to suggest multiple alternative scaffolds for a particular biological pattern that is of interest, e.g., antiproliferation.

Specifically, in the case of the general toxic scaffolds that show activity against both proliferation and reporter gene assays (both of which are cell-based), many compounds belonging to those scaffolds were selected repeatedly as primary hits in various HTS campaigns. In many occasions these compounds failed to demonstrate similar activity in the follow-up experiments. It is a significant waste of resources and those compounds should never have been selected or even put into HTS library that is intended to be used in the cell-based assays. Based on our data mining efforts across a large panel of assays in the HTS database, these general toxic scaffolds and their member compounds can now be flagged and excluded from either the cell-based screening library or hit-picking system. Moreover, as shown in Figure 3, the tumor cytotoxic scaffolds are active in the proliferation assays but not cell-based reporter gene assays. Presumably, the compounds from these scaffolds may inhibit the growth of tumor cells as most proliferation assays use tumor cell lines, but they do not kill the normal cells that are used in the reporter gene assays. These scaffolds may be used as potential starting point for an anticancer therapy (see also below). In addition, the compounds/scaffolds that show broad activity in reporter gene assays are considered as potential artifacts to this type of assay technology. As shown in Figure 3, this group of scaffolds can be further divided into two subtypes, that is, one with selective tumor cytotoxicity and one without. Artifacts in reporter gene assays can be caused by many reasons. For example, the compound family may interfere with the luciferase—luciferin interaction, as the reporter gene assay technology is strongly correlated with luciferase readout (see also Table 2). Also, if a compound interacts

Table 4. Aggregator Compounds[28,40,41] Similar to In-House Compound Scaffolds[a]

aggregator	similarity to in-house compound[b]	assay (μM)			
		β-lactamase	chymotrypsin	MDH[c]	β-galactosidase
 Nicardipine	0.89	20	175	50	
 Quercetin	1.0	4	100	6	220
 Gyrase inhibitor	0.90	18	100		320
 Bisindolyl-maleimide I	1.0	60	200	>400	
 Bisindolyl-maleimide IX	0.91	5	20	45	
 Benzyl benzoate	0.86	90	250	125	

[a] Daylight fingerprints[32] and Tanimoto similarity coefficient cutoff of 0.85 were used in the similarity search. [b] Similarity is measured based on the Tanimoto coefficient between the aggregator compound and the most similar in-house compound. [c] MDH stands for malate dehydrogenase.

somehow with the upstream pathway of the reporter gene system, in which the same reporter genes are often used in multiple assays from assay development, it will likely show a broad interaction pattern for many targets rather than the specific target of interest. Furthermore, a compound could also simply be toxic to a normal cell line but not to the highly mutated tumor cell lines. In all, the exact molecular mechanisms that cause the artifacts in reporter gene assays are of an extremely complicated nature and an interesting research topic. In our study, we simply annotated these scaffolds as potential reporter gene assay artifact and flagged them as undesirable chemical scaffolds.

It should be mentioned that we did not observe any scaffolds that show a specific inhibition pattern broadly against enzyme activity assays. It has been demonstrated by Shoichet and colleagues that some compounds, so-called aggregators, tend to form large particles and show promiscuous binding and faulty potency measurements in many assays.[40] It was also argued that the IC_{50} values of these compounds may change with respect to temperature.[40] We carried out similarity-based structural search in our HTS compound collection using the previously determined aggregators as query structures[28,40,41] and a Tanimoto similarity coefficient cutoff of 0.85. The aggregators that bear structural similarity to our compound scaffolds are listed in Table 4, together with their reported IC_{50} values.[28,40,41] Interestingly, most of these aggregators are weak inhibitors with IC_{50} values of over 100 μM. However, in our organization all HTS campaigns are carried out with an assay compound concentration of 10 μM; therefore, these promiscuous compounds most likely would have shown no (or very weak) inhibition activities. On the other hand, for HTS databases that contain fold inhibition data obtained at higher compound concentrations, we believe our data mining method can also identify such aggregator candidates.

Some Proof-of-Concept Examples of the Chemical Scaffold Annotation. *Two Known Anticancer Chemotypes, Daunorubicin and Paclitaxel and Their Derivatives, Are Correctly Annotated as Tumor Cytotoxic.* As shown in Figure 4, the compound cluster families GNF08103 and GNF08454 are annotated as tumor cytotoxic (type **A**) as they primarily show high activities in the proliferation assays. Specifically, a close examination of cluster GNF08103 revealed that this six-membered group shares the same core anthracycline scaffold to the known anticancer drug daunorubicin (Figure 4, panel A). In fact, one of the family members is also a related known anthracycline drug called aclarubicin (Figure 4, panel A). It is well-known[42] that daunorubicin and its derivatives, produced by *Streptomyces peucetius* and related bacteria, are toxic antineoplastic and used to treat leukemia and other cancers—this is consistent with our annotation of this scaffold as tumor cytotoxic. Moreover, as shown in Figure 4, panel B the compounds in the cluster GNF08454 are diastereoisomers to the popular anticancer drug paclitaxel, which induces apoptosis of tumor cells by disrupting microtubule dynamics via hyperstabilization.[43] The annotation of this

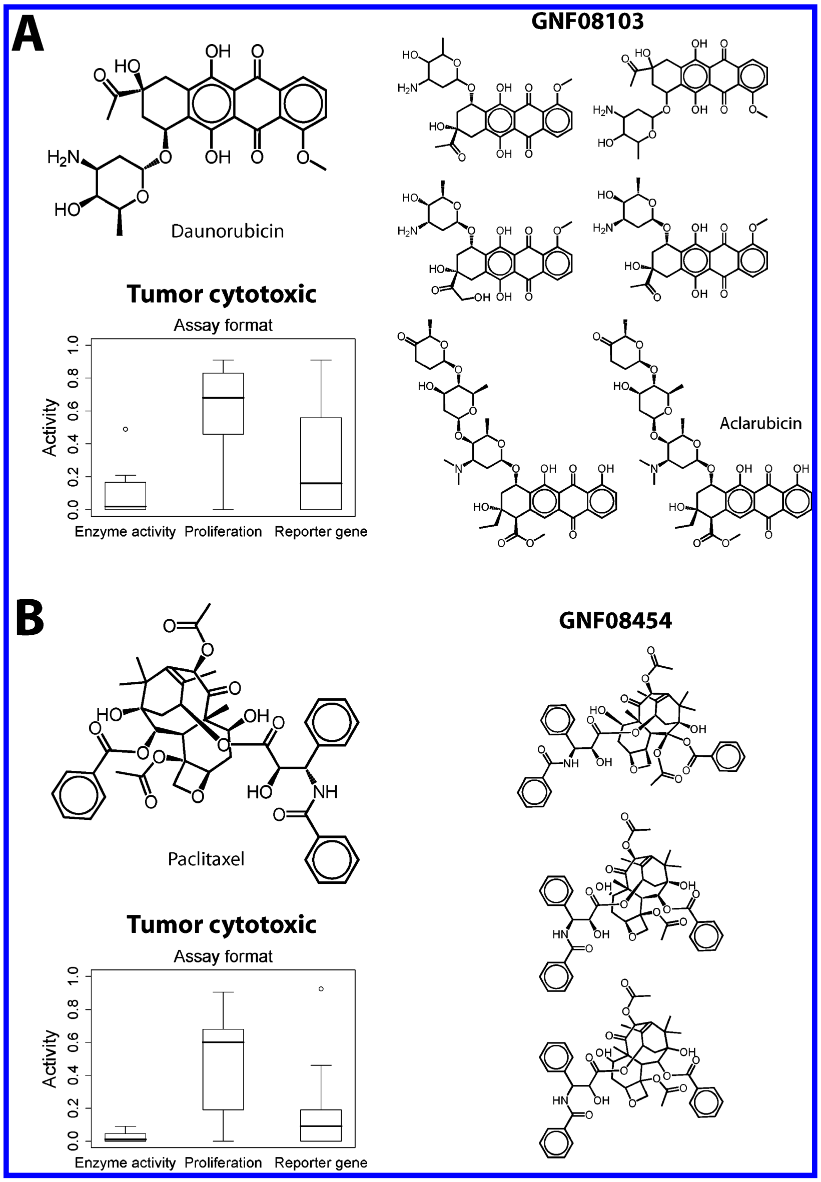


Figure 4. Two tumor cytotoxic scaffold families structurally similar to the known anticancer drugs (A) daunorubicin and (B) paclitaxel, respectively. The core member compounds from each scaffold family are shown.

scaffold as tumor cytotoxic (type **A**) based on HTS profiles is also in line with the mechanism of action of this type of anticancer compounds.

A Tellurium-Containing Scaffold Shows General Toxicity in Both Proliferation and Reporter Gene Assays. Members of the compound cluster family GNF28317 contain a heavy tellurium atom in their chemical scaffold and are annotated as general toxicity (Figure 5, panel A). We believe it is consistent with the knowledge that tellurium compounds are generally considered as toxic, and the high activities observed by this group of compounds in both proliferation and reporter gene cell-based assays are potentially caused by this general toxicity of heavy metal atom rather than the actual activity against the specific molecular targets/pathways of interest designed in these assays. Indeed, the representative compound (on the left) was assayed for its cytotoxicity against three B-cell follicular lymphoma cell lines. The LD_{50} for one member of the panel is 0.16 μM and about 10-fold higher for the other two cell lines.

The Benzisothiazolone Scaffold: A Potential Reporter Gene Assay Artifact. The compound cluster GNF05008, a benzisothiazolone scaffold, is annotated as a potential reporter gene assay artifact (type **C**), demonstrated by its unusual, if not exclusive, high activity in a large number of different reporter gene assays (Figure 5, panel B). The heatmap in Figure 5 also shows the HTS profiles of these five member compounds. This unusual pattern of high activity across many reporter gene assays suggests that this scaffold might interfere with the reporter gene detection technology and cause an assay specific artifact.

A Spironolactone-like Scaffold Demonstrates Target Family Specific Activities Against Nuclear Receptor Family. Figure 5, panel C shows a scaffold (GNF08670) that demonstrates nuclear receptor specific activities in many HTS campaigns. We carried out structural similarity searches of these compounds in the WDI database and revealed that they are structurally similar to the known drug spironolactone (Figure 5, panel C). In fact, three out of these five compounds are diastereoisomers of spironolactone. Spironolactone has been described as an aldosterone antagonist by virtue of its association with the mineralocorticoid receptor.[44] Our annotation of this group of compounds as nuclear receptor

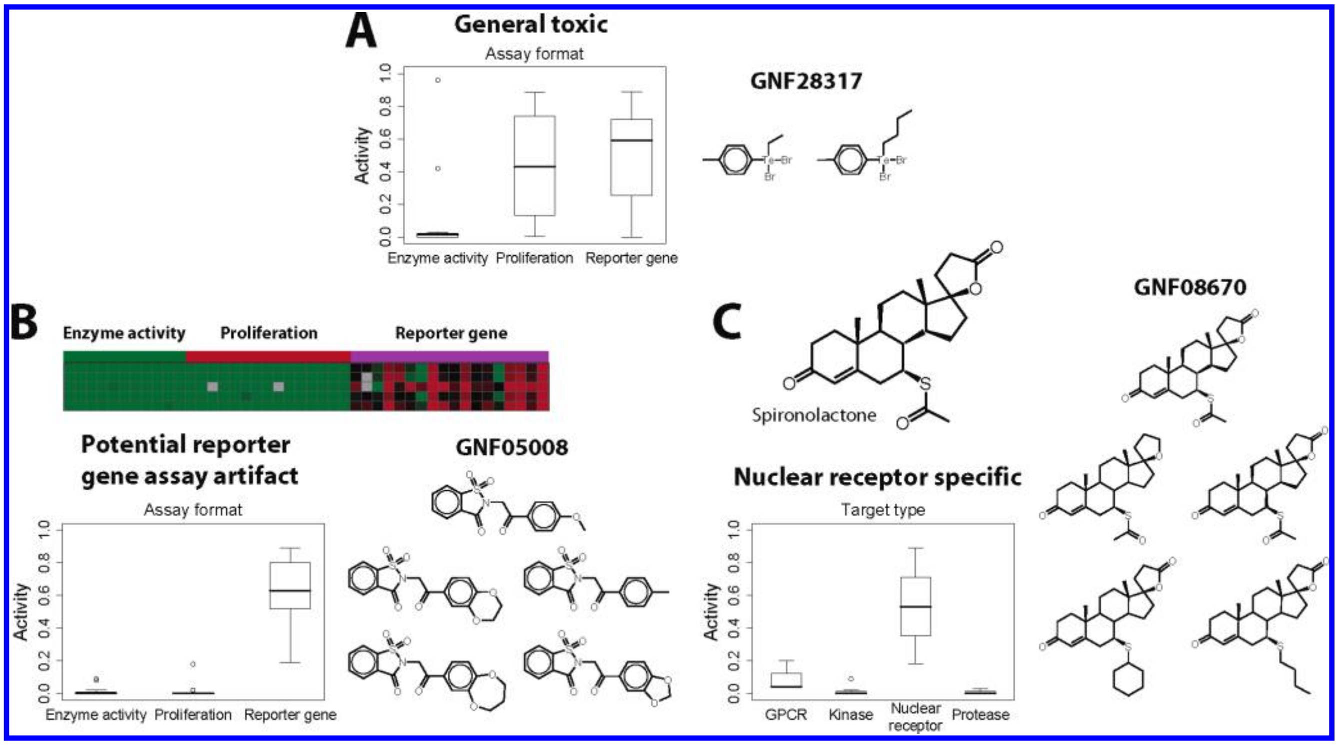


Figure 5. (A) Tellurium-containing general toxic scaffold family. (B) Benzisothiazolone scaffold as potential reporter gene assay artifact. (C) Nuclear receptor specific scaffold family structurally similar to the known aldosterone antagonist, spironolactone. The core member compounds from each scaffold family are shown.

specific is apparently consistent with their putative function as aldosterone antagonist.

Here we presented proof-of-concept examples of our OPI-based approach in mining the large-scale HTS database. Reasonable compound scaffold annotations, characterized as tumor cytotoxic, general toxic, potential reporter gene assay artifact, and target family specific, were observed, demonstrating its ability to extract useful information and knowledge from the raw HTS profiling data. In particular, we were able to annotate a large corporate compound library on a scaffold basis based on not only the biological target information but also the assay format such as proliferation and reporter gene. In addition, we have set up an internal Web-based interface for accessing the knowledge base of all the 1391 compound scaffolds with detailed heatmaps, WDI annotations, and the capability for chemists/biologists to contribute additional information regarding compounds of interest.

CONCLUSIONS

In this report we proposed a novel data mining method based on the ontology-based pattern identification algorithm and applied it to our large-scale in-house HTS database. About 1500 compound scaffold families with statistically significant structure−profile relationships were discovered. The Kruskal−Wallis and ANOVA statistical tests were further applied to the discovered scaffolds for functional annotation. These rigorous statistical test controls have been extensively applied to prevent overinterpretation of the observed selectivity in HTS profiles. Four types of compound scaffolds, namely tumor cytotoxic, general toxic, potential reporter gene assay artifact, and target family specific, were defined to annotate the HTS library on a scaffold basis. Several proof-of-concept examples have been shown to be correctly annotated, demonstrating the applicability of our method in mining this type of large-scale HTS database. The scaffolds identified by our HTS data mining efforts can be a valuable resource for choosing SAR-reliable diversity libraries, generating in silico biological annotations of compound scaffolds, and providing novel target family specific scaffolds for focused compound library design.

ACKNOWLEDGMENT

We thank Dr. Gang Xia for critical reading of the manuscript.

REFERENCES AND NOTES

(1) Collins, F. S.; Morgan, M.; Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **2003**, *300*, 286−290.

(2) Guttmacher, A. E.; Collins, F. S. Realizing the promise of genomics in biomedical research. *JAMA* **2005**, *294*, 1399−1402.

(3) Parker, C. N.; Schreyer, S. K. Application of chemoinformatics to high-throughput screening: practical considerations. *Methods Mol. Biol.* **2004**, *275*, 85−110.

(4) Young, S. S.; Hawkins, D. M. Using recursive partitioning analysis to evaluate compound selection methods. *Methods Mol. Biol.* **2004**, *275*, 317−334.

(5) Harper, G.; Pickett, S. D.; Green, D. V. Design of a compound screening collection for use in high throughput screening. *Comb. Chem. High Throughput Screen.* **2004**, *7*, 63−70.

(6) Willett, P. Evaluation of molecular similarity and molecular diversity methods using biological activity data. *Methods Mol. Biol.* **2004**, *275*, 51−64.

(7) 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.

(8) Frawley, W. J.; Piatetsky-Shapiro, G.; Matheus, C. J. Knowledge discovery in databases: an overview. *Knowledge Discovery in Databases*; AAAI/MIT Press: Cambridge, MA, 1991; pp 1−30.

(9) Kubinyi, H.; Müller, G. *Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective*; Wiley-VCH: Weinheim, 2004.

(10) Stockwell, B. R. Exploring biology with small organic molecules. *Nature* **2004**, *432*, 846−854.

(11) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J., Jr.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W. W.; Monks,

A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997**, *275*, 343−349.

(12) Blower, P. E.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Yu, L.; Richman, S.; Weinstein, J. N. Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data. *Pharmacogenomics J.* **2002**, *2*, 259−271.

(13) Horvath, D.; Jeandenans, C. Neighborhood behavior of *in silico* structural spaces with respect to in vitro activity spaces − a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680−690.

(14) Horvath, D.; Jeandenans, C. Neighborhood behavior of *in silico* structural spaces with respect to *in vitro* activity spaces − a benchmark for neighborhood behavior assessment of different *in silico* similarity metrics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691−698.

(15) Froloff, N. Probing drug action using in vitro pharmacological profiles. *Trends Biotechnol.* **2005**, *23*, 488−490.

(16) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261−266.

(17) Bredel, M.; Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262−275.

(18) Jacoby, E.; Schuffenhauer, A.; Popov, M.; Azzaoui, K.; Havill, B.; Schopfer, U.; Engeloch, C.; Stanek, J.; Acklin, P.; Rigollier, P.; Stoll, F.; Koch, G.; Meier, P.; Orain, D.; Giger, R.; Hinrichs, J.; Malagu, K.; Zimmermann, J.; Roth, H. J. Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection. *Curr. Top. Med. Chem.* **2005**, *5*, 397−411.

(19) Fischer, H. P.; Heyse, S. From targets to leads: the importance of advanced data analysis for decision support in drug discovery. *Curr. Opin. Drug Discov. Dev.* **2005**, *8*, 334−346.

(20) Vieth, M.; Sutherland, J. J.; Robertson, D. H.; Campbell, R. M. Kinomics: characterizing the therapeutically validated kinase space. *Drug Discov. Today* **2005**, *10*, 839−846.

(21) Böcker, A.; Schneider, G.; Techentrup, A. Status of HTS data mining approaches. *QSAR Comb. Sci.* **2004**, *23*, 207−213.

(22) Root, D. E.; Flaherty, S. P.; Kelley, B. P.; Stockwell, B. R. Biological mechanism profiling using an annotated compound library. *Chem. Biol.* **2003**, *10*, 881−892.

(23) Covell, D. G.; Wallqvist, A.; Huang, R.; Thanki, N.; Rabow, A. A.; Lu, X. J. Linking tumor cell cytotoxicity to mechanism of drug action: an integrated analysis of gene expression, small-molecule screening and structural databases. *Proteins* **2005**, *59*, 403−433.

(24) 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.

(25) Lowrie, J. F.; Delisle, R. K.; Hobbs, D. W.; Diller, D. J. The different strategies for designing GPCR and kinase targeted libraries. *Comb. Chem. High Throughput Screen.* **2004**, *7*, 495−510.

(26) Whittaker, M. Discovery of protease inhibitors using targeted libraries. *Curr. Opin. Chem. Biol.* **1998**, *2*, 386−396.

(27) Fishman, M. C.; Porter, J. A. Pharmaceuticals: a new grammar for drug discovery. *Nature* **2005**, *437*, 491−493.

(28) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.* **2003**, *46*, 4477−4486.

(29) Zhou, Y.; Young, J. A.; Santrosyan, A.; Chen, K.; Yan, S. F.; Winzeler, E. A. *In silico* gene function prediction using ontology-based pattern identification. *Bioinformatics* **2005**, *21*, 1237−1245.

(30) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(31) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.* **1997**, *15*, 372−385.

(32) James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual: Daylight Version 4.9*; Daylight Chemical Information Systems, Inc.: Mission Viejo, CA, 2005.

(33) Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469−474.

(34) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **1995**, *57*, 289−300.

(35) Zar, J. H. *Biostatistical Analysis*; Prentice Hall: Upper Saddle River, NJ, 1999.

(36) Kibbey, C.; Calvet, A. Molecular Property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. *J. Chem. Inf. Model.* **2005**, *45*, 523−532.

(37) Martin, Y. C. Challenges and prospects for computational aids to molecular diversity. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 159−172.

(38) Goodnow, R. A., Jr.; Guba, W.; Haap, W. Library design practices for success in lead generation with small molecule libraries. *Comb. Chem. High Throughput Screen.* **2003**, *6*, 649−660.

(39) Webb, T. R. Current directions in the evolution of compound libraries. *Curr. Opin. Drug Discov. Dev.* **2005**, *8*, 303−308.

(40) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **2002**, *45*, 1712−1722.

(41) McGovern, S. L.; Shoichet, B. K. Kinase inhibitors: not just for kinases anymore. *J. Med. Chem.* **2003**, *46*, 1478−1483.

(42) Minotti, G.; Menna, P.; Salvatorelli, E.; Cairo, G.; Gianni, L. Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. *Pharmacol. Rev.* **2004**, *56*, 185−229.

(43) Jordan, M. A.; Wilson, L. Microtubules as a target for anticancer drugs. *Nat. Rev. Cancer* **2004**, *4*, 253−265.

(44) Menard, J. The 45-year story of the development of an anti-aldosterone more specific than spironolactone. *Mol. Cell. Endocrinol.* **2004**, *217*, 45−52.

CI060102U