Novel Statistical Approach for Primary High-Throughput Screening Hit Selection

S. Frank Yan,* Hayk Asatryan, Jing Li, and Yingyao Zhou

Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121

Received July 8, 2005

The standard activity threshold-based method (the "top X" approach), currently widely used in the high-throughput screening (HTS) data analysis, is ineffective at identifying good-quality hits. We have proposed a novel knowledge-based statistical approach, driven by the hidden structure−activity relationship (SAR) within a screening library, for primary hit selection. Application to an in-house ultrahigh-throughput screening (uHTS) campaign has demonstrated it can directly identify active scaffolds containing valuable SAR information with a greatly improved confirmation rate compared to the standard "top X" method (from 55% to 85%). This approach may help produce high-quality leads and expedite the hit-to-lead process in drug discovery.

INTRODUCTION

A modern small-molecule drug discovery project usually starts by screening a large collection of compounds against a biological target that is believed to be associated with a certain disease, with the goal of identifying interesting, tractable starting points for medicinal chemistry. Despite the fact that primary screening of huge libraries containing over 1 million compounds can now be accomplished in a matter of days in pharmaceutical companies, the throughput of secondary confirmation screening is still largely limited to several thousand compounds,[1,2] and the number of compounds that eventually enter the medicinal chemistry phase of lead optimization is further reduced to only a couple of hundred at best.[3,4] This presents a great challenge to the early hit-to-lead process of drug discovery in selecting the most promising hits from the high-throughput screening (HTS) and/or ultrahigh-throughput screening (uHTS) results.[3−8] In the current HTS data analysis, an activity cutoff value is usually set to select a certain number of compounds whose tested activities are greater than this threshold. These are called primary HTS hits and are subject to further retesting for structure and activity validation. Secondary confirmation screening, typically by a detailed dose−response study, is then carried out. The compounds that pass the secondary screening are designated as validated hits or HTS actives,[1] which then are often grouped into scaffold families. Based upon further evaluation and/or additional chemical exploration, the scaffolds that exhibit a certain degree of structure−activity relationship (SAR), advantageous patent status, amenability to chemical modification, favorable physicochemical and pharmacokinetic properties, etc. are selected as lead series for subsequent optimization. However, the current cutoff-based primary HTS hit selection process, often dubbed as the "top X" approach,[2] has two major weaknesses: First, the confirmation rate is rather low, often in the range of ∼40%[4] mainly due to the noisy and error-prone nature of single-dose HTS/uHTS.[9] Second, no knowledge-based analyses, e.g. SAR examination, are considered in this very first hit-picking process. This is undesirable, considering medicinal chemists are often willing to trade (pending secondary confirmation of course) a potent scaffold with weak SAR for a scaffold that possesses better SAR but slightly weaker activity (e.g., 1−2 orders of magnitude less active), as the latter oftentimes has a better chance to become a good starting point for optimization. Being the first step of a drug discovery project, this seemingly simple "cherry-picking" step has fundamental far-reaching effects on later processes.[8] The current cutoff-based method is clearly ineffective and may contribute to the disappointing fact that high-throughput technologies have not yet lived up to the high expectations set for them.[10−14] A novel approach that can effectively address these challenges in primary HTS hit selection is therefore urgently needed.

A variety of cheminformatics methods have been developed for HTS data analysis for the past decades.[2,9,15−22] Most notably, it includes recursive partitioning (RP), clustering analysis, similarity and nearest-neighbor searching, cell-based analysis, pharmacophore modeling, data shaving, and many others (see the chapter by Parker and Schreyer[9] for a more comprehensive review). It should be noted however that most of these methods were designed primarily for modeling the HTS data with application in the screening paradigm dubbed as sequential screening or smart screening.[15,23,24] In this paradigm a relatively small-sized, diverse set of compounds are first selected from the company library and screened. Based on the initial results, cheminformatics methods such as those mentioned above are applied to derive predictive (quantitative) structure−activity relationship [(Q)SAR] models for the tested compounds. These models are in turn used to guide selecting additional new compounds from the library for screening. New data often are incorporated back into the existing results in order to refine the predictive (Q)SAR models. This process iterates which ultimately produces a set of candidate compounds that advance into the next phase of lead discovery. Although some pharmaceutical companies may still adopt this sequential screening philosophy in early

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

10.1021/ci0502808 CCC: $30.25
Published on Web 10/14/2005

drug discovery, in recent years many companies have embraced a screening practice that screens all the available compounds (often over a million compounds that are readily available for testing from compound management) in one, usually, uHTS campaign. This is due to the rapid development in HTS technology, significantly reduced cost associated with modern uHTS, greatly alleviated logistic burden, and the general awareness of the values provided by the inactive hits.[1,9] With the increasing deployment of such uHTS platform, the main purpose of primary HTS/uHTS data analysis has shifted from building statistical models (from a small set of tested compounds) in order to select new compounds for additional screening to, instead, picking the most promising, active compounds that have the potential to become good lead series from a complete, extremely large experimental screening data set. Most of these cheminformatics methods, albeit valuable for improving the understanding of the HTS data, are not directly applicable for this type of data analysis (e.g. similarity and nearest-neighbor searching) and/or do not address the major challenge of low confirmation rate encountered in uHTS.[2] In addition, while some of the methods allow using a continuous response in modeling the HTS data such as RP,[25] many of them still apply a rather arbitrary activity cutoff to designate the "active" and "inactive" compounds in order to build training data sets required for the predictive models for new compound selection, for example in the case of the cell-based analysis.[16]

Due to the great advance in HTS technologies,[9] a large-scale uHTS campaign has become common in early discovery projects, accompanied by the resultant massive amount of data, for which effective data mining techniques are apparently needed to rationally select the compounds that are most likely to be validated (in the secondary screening) and then to become the lead series. However, the historic, simple "top X" method still dominates the hit selection process, relying on one activity threshold that is oftentimes determined arbitrarily depending on for example the capacity of the secondary screening, experience of the assigned scientist, or even logistic reasons such as compound availability.[2,4,5,8] To overcome the limitations of the "top X" method, several techniques have been proposed to consider other factors in the HTS hit selection process besides the assay activity alone.[2] One such example is the SCAM (statistical classification of the activities of molecules) program from GlaxoSmithKline, which utilizes a large number of binary descriptors and the RP algorithm to partition the tested compounds into classes.[25] In addition, it has been proposed to filter out compounds with undesirable and/or uninteresting properties (such as physicochemical properties, pharmacokinetics, pharmacodynamics, toxicity, etc.) from the screening collection based on chemical structure in order for relatively weaker but more interesting compounds to be selected.[2,3] Similarly, the leadlikeness concept proposed by Oprea and colleagues is also used to filter out compounds that clearly lack the characteristics of promising lead series from the screening set.[1,4] Moreover, attempts have been made to establish a statistical data model of the uHTS results, hoping to help guide the hit selection process.[7,26−30] Most notably, the Z' score[26] suggested by these studies is now commonly applied for quality evaluation of HTS assays.[3] Despite these ongoing efforts, a very limited number of statistically rigorous methods (the SCAM program[25] being one) have been designed specifically to help pick and rank HTS hits from the vast amount of raw data generated by the uHTS campaign. The current uHTS primary hit selection process still heavily, if not exclusively, relies on the screening activity value.

Furthermore, although it is widely recognized that SAR is an important characteristic of lead compounds, this type of structure-related information is rarely utilized in selecting hits from raw uHTS data. Indeed, only limited attempts have been made toward this direction.[7,10,16,22,30] One key premise for such consideration is that by grouping compounds into clusters based on structure, the cluster that is enriched by active compounds with reasonable SAR should be given greater consideration in the hit selection process.[2,16,22,30,31] For example, based on ring hashcodes[32] clustering and binomial distribution, McFadyen et al. proposed a cumulative significant probability score that quantifies the enrichment of active compounds within a cluster.[2] The cell-based analysis proposed by Lam et al. utilizes the BCUT descriptors[33] to group compounds into cells; a probability score based on hypergeometric distribution and a binomial hit rate lower confidence limit were introduced to characterize the enrichment against random selection.[16,34,35] The phylogenetic-like tree (PGLT) algorithm hierarchically clusters compounds into classes based on common substructure and extracts activity and SAR information for each node.[22] However, many of these methods still largely use a somewhat arbitrarily predefined, "static" activity threshold, often based on a single-dose test, to determine the active and inactive compounds before carrying out the statistical analysis.[8] In this way, the SAR principle is merely used as a hit-reporting/summarizing tool, instead of being employed to reduce the false positive and false negative rates in the hit-picking step as introduced in this study.

Here we have proposed a novel HTS primary hit identification method by integrating SAR information into the hit selection process with statistical rigor. This approach not only is able to choose HTS hits of much higher confirmation rate (in secondary confirmation screening) but also can identify scaffolds with sufficient SAR directly from the noisy primary uHTS data. Our method is based on two important ideas: First, almost all large compound libraries used in pharmaceutical uHTS campaigns have built-in chemical redundancy.[36−39] Even though each compound is typically screened only once, they are often co-screened with several other structurally similar neighbors. This makes the SAR principle directly applicable in pooling the HTS results that belong to a compound family as a whole; an effective statistical test can then be employed to dynamically pick an active scaffold family with much greater confidence than simply hit-picking individual compounds, which is often error-prone due to the inherently noisy nature of HTS. This idea is best illustrated by an intuitive example—some of the most active compounds from an HTS campaign are often artifacts largely due to experimental accidents such as pipetting errors[28,40] or because of promiscuous aggregates formation.[41,42] Assuming a compound is observed as the only active one amid a decent-sized scaffold family, it is not difficult to single it out as a potential false positive given that all its neighbors are inactive. Therefore, with careful design the large amount of inactive compound data provided

by an HTS campaign could be exploited to help identify hits of improved quality. Second, it is possible to develop a rigorous statistical ranking score for the selected scaffolds, which takes into account both the assay activity criterion and the chemical redundancy information of a scaffold. This has been demonstrated by the previous studies.[2,16,30] Recently, Zhou et al. have developed an ontology-based pattern identification (OPI) method and applied it successfully to the prediction of gene functions based on microarray gene expression data and the guilt by association (GBA) principle.[43,44] This method provides a sound statistical framework of scoring each biological process (comprised of multiple genes) using the expression level measured for each gene.[43] In this report we modified and adopted a similar idea to dynamically score each compound scaffold family based on the uHTS assay activity measured for each member compound without having to predefine a discriminating activity cutoff. The OPI score probabilistically takes into account both the compound family size and their individual member activities. Also, when the compound collection is totally diversified (all singletons), our OPI-based approach degenerates and becomes equivalent to the activity cutoff-based method.

METHODS

Statistical Framework of the OPI Methodology. The ontology-based pattern identification method was originally developed for gene expression profiling data analysis, and we have described in detail the mathematical framework and implementation of the OPI algorithm in our recent publication.[43] Here we outlined the key ideas of this method in the context of high-throughput screening data analysis, particularly the primary HTS hit selection.

An ontology represents the knowledge that several compounds belong to a compound (scaffold) family, which can be obtained computationally by any clustering analysis. In the current case, the compound families were generated based on chemical similarity by an in-house clustering program using Daylight fingerprints[45] with a Tanimoto similarity coefficient threshold of 0.85. Each compound family is prone to have distinctive activity potency and SAR strength (i.e., how likely similar compounds share similar activities); this, we believe, is the main reason the "top X", one-size-fits-all cutoff-based method fails to deliver high-quality HTS hits. The goal of the OPI method is to automatically determine the cutoff value for each individual family and then select the primary hits in a family specific fashion with improved confirmation rate. Assuming all the tested compounds are sorted based on activity (from potent to weak) and because there are usually both active and inactive compounds in a scaffold family, we expect to see two clusters—active member compounds tend to cluster on the top of the list, while inactive ones are scattered sparsely in the bottom. The key advantage of the OPI method is to be able to automatically derive such a binary partition that the active members show an optimal statistical agreement under the SAR principle.

Assume an entire compound collection of size N has been screened in an HTS/uHTS campaign and one particular scaffold family $\boldsymbol{K}$ of interest consists of n compounds. If one sets a cutoff value c and then obtains m compounds as hits (among which m' compounds happen to be in that family

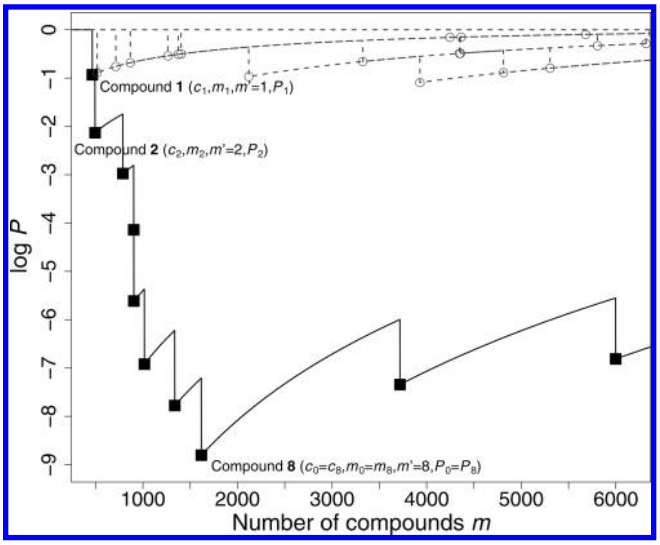

Figure 1. Plot of the computed logarithmic P-value versus the number of selected compounds for a compound scaffold. The black solid line with solid squares is for the actual calculation of the compound scaffold with 15 members, and the gray dashed lines with circles are for the permutations runs of this scaffold.

$\boldsymbol{K}$), the optimal partition defined by OPI is the one with the least expected enrichment of m' among the m compounds. The odds of picking m' members in a randomly selected list of m compounds is known to be quantified by the following probability score P based on the accumulative hypergeometric distribution:[46]

$$P(N,n,m,m') = \sum_{k=m'}^{\min(m,n)} \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$$

According to this statistical framework, OPI is an activity-based scaffold-driven global minimization algorithm[43] that searches for the optimal subset of m' unusually active compounds out of any compound family $\boldsymbol{K}$, so that these m' accepted hits not only share structural similarity but also share similar high HTS activities, a tribute of interesting lead series pending further validations.

HTS Hit Selection Using the OPI Approach. Based on the OPI algorithm a novel HTS hit selection process is carried out as follows: (1) All N tested compounds are grouped into scaffold families (by an in-house clustering program) based on their chemical structures. (2) Compounds are then ranked according to their screening activities from potent to weak just as in the standard, cutoff-based method. (3) For a given compound family of size n (as illustrated in Figure 1), its members have the ranks m_1, m_2, ..., m_n according to the activity-sorted compound list, respectively. We first take m_1 most active compounds using the HTS activity value of the most active member (compound **1** and the m_1th compound in the list), c_1, as the cutoff. (4) We then assign a probability score P_1 to this list

$$P_1 = P(N,n,m_1,m'=1)$$

which represents the probability that at least one compound belongs to the family of n compounds, when one randomly selects m_1 compounds out of a total collection of N compounds. (5) We then expand this list to include the

second most active compound from this family (compound **2** and the m_2th compound in the list); this is equivalent to decreasing the cutoff value to c_2 (the activity of compound **2**) and increasing the size of the list to m_2 (Figure 1). Similarly, we assign another probability score P_2 to this expanded list

$$P_2 = P(N,n,m_2,m'=2)$$

which represents the odds that two or more compounds happen to belong to the same scaffold family of size n, while randomly sampling m_2 out of the N compounds. (6) We repeat step 5 until all the n member compounds from this scaffold family are included as hits, which yields the largest possible list for this family. (7) Among all the n lists generated above, we then only choose the list with the lowest P-value, in which compounds from the scaffold family of interest are most statistically enriched in a smallest possible list size. Specifically, in the case illustrated in Figure 1, the globally minimized probability score $P_0 = P_8$; therefore, the optimal number of compounds selected from this family $m' = 8$ (compounds **1**−**8**), and the activity of compound **8** (c_8) is the optimal activity cutoff c_0. (8) The above steps 3−7 are iteratively applied to all scaffold families; the selected compounds from each family are then prioritized by the family P_0-value first and second by the screening activity. Furthermore, to avoid the "multiple test problem" in such iterative statistical tests,[47] we randomly permute the compound activities and repeat the above algorithm to estimate the likelihood of the original P_0-value that occurs by chance simply because of the iterative nature of this method. For example, the dashed lines in Figure 1, representing several such permutation runs, indicate that the lowest P_0-value obtained by OPI algorithm using the real data set is statistically robust against "multiple tests" for this scaffold family. Our optimal P_0-value sufficiently quantifies both the activity potency and the SAR strength of a compound family, that is, the lower the P_0-value the more desirable that scaffold family is to be hit-picked.

RESULTS AND DISCUSSION

We applied this knowledge-based hit-picking method to an in-house cell-based uHTS campaign using an internal corporate library, whereby the assay was validated with a Z' score of 0.5. Following quality control and normalization, which eliminated obvious artifacts and outliers, we obtained single-dose activity data for ~1.1 million compounds. Here we selected 50 000 most active compounds to be analyzed by the OPI approach. Since the activities of many of those compounds are already well within the low activity region that would be certainly considered as inactive, we believe this should provide sufficient data points for us to pick the primary HTS hits for the subsequent confirmation test, which is often limited to about a couple of thousand.[1,2] The compounds were first clustered into scaffold families by an in-house clustering algorithm based on Tanimoto similarity coefficient and Daylight fingerprints[45] using a threshold value of 0.85.[48] Figure 2 shows the plots of the confirmation rate (i.e., the ratio between number of confirmed actives over number of selected compounds) versus the number of compounds selected using both the cutoff-based and our methods. It is noted that the two different methods selected

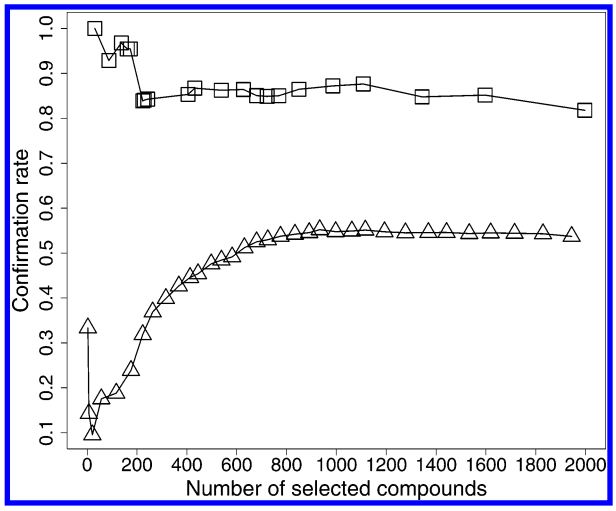

Figure 2. Plot of the confirmation rate versus the number of compounds selected by two different primary HTS hit selection methods. The squares are for our scaffold-based OPI method, and the triangles are for the standard threshold-based "top X" method. The confirmation rate is computed as the ratio between the number of confirmed actives over the number of selected compounds.

two distinctive sets of compounds. As shown in Figure 2, when a small number of compounds (~150) is selected, the confirmation rate is quite low (~20%) using the "top X" method, most likely due to the presence of experimental artifacts with erroneously high activities despite the preliminary quality control. The confirmation rate, however, increases as more compounds are included until a maximum of about 55% is reached, when nearly 1000 compounds are selected (Figure 2). This type of behavior is often observed in HTS/uHTS data analysis.[28] In contrast, our approach performs significantly better. As shown in Figure 2, it can achieve a high confirmation rate of over 95% when only ~150 compounds are selected, demonstrating its ability to pick the most promising compounds with high accuracy by effectively eliminating potent false positives. A high confirmation rate (~85%, squares in Figure 2) remains largely stable with increased number of selected compounds, which is consistently much higher than that using the cutoff-based method (~55%, triangles). We also repeated the same analysis using a similarity threshold value of 0.7 instead of 0.85 in the aforementioned clustering process, which resulted in a decreased number of scaffolds but generally an increased cluster size. Results similar to Figure 2 were obtained (data not shown), indicating the robustness of our method against different ways of compound clustering.

In addition, to further assess our method compared to the simple threshold-based approach, we carried out additional experiments to retest those compounds that were ranked high based on our P_0-value probability score but had been considered as inactive by the "top X" method (i.e., compounds that are not among the previously selected most active ones and potential false negatives). For the first 1108 compounds selected by our OPI method, 825 were originally considered as inactive based on the activity threshold. We selected 202 of these "inactive" compounds for retesting (due to compound availability); 144 of them were determined to be actual actives in the secondary assay (i.e., IC_{50} values range from high nanomolar to low micromolar), yielding a confirmation rate of ~71%, which is even higher than the ~55% confirmation rate of the "active" compounds deter-

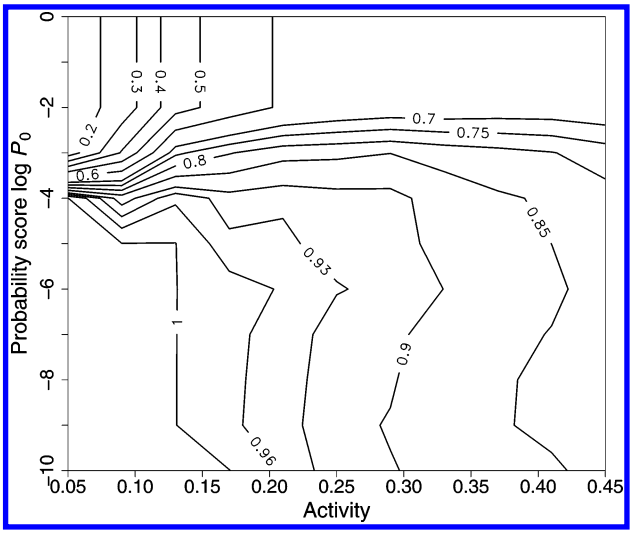

Figure 3. Confirmation rate contour plot of compounds selected based on both OPI score and activity. Compounds are selected when their scaffold-based log P_0 scores are smaller than a threshold (more likely to be true actives) and their activity values are smaller than an activity threshold (more active).

mined by the "top X" method. This demonstrates the ability of the OPI method to identify the potential false negatives characterized by the simple cutoff-based approach. Furthermore, the P_0-value scoring scheme used by our hit-picking method is nonparametric, i.e., it does not require any a priori statistical model for the primary HTS data, as in contrast to many previous studies in which the data were often modeled by a known statistical distribution such as uniform distribution,[28] normal distribution,[26,27] log-normal distribution,[30] or some other complex formulas that are still open for discussion. This suggests that our results presented here, based on a typical (in-house) uHTS campaign, are likely to represent the performance of this novel hit-picking approach in general.

Compared to the widely used activity threshold-based method for HTS hit selection, this novel knowledge-based statistical approach has many advantages. First, we introduced a new, computationally determined probability score P_0-value which can be used, together with the assay activity criterion, to identify promising HTS hits with an improved accuracy. Figure 3 shows the confirmation rate contour plot of selected compounds based on both activity and P_0 score. When the activity criterion is applied alone (i.e., ignore the OPI score by fixing log P_0 at 0 in Figure 3), the confirmation rate actually decreases when increasing the activity threshold (the smaller the activity value, the more active the compound is) (also see Figure 2). This seemingly abnormal behavior is commonly observed in HTS, which oftentimes indicates the existence of a high proportion of potent false positives despite the initial quality control efforts.[28,40] The abnormally low confirmation rate at high activity cutoffs also illustrates the inability of the "top X" method in identifying such false positives. On the other hand, by additionally applying our new probability score P_0 criterion, e.g., log P_0 set at -6, the majority of the false positives can be eliminated, and the confirmation rate is improved significantly even when a marginal activity threshold, e.g. 0.4, is applied (Figure 3). Moreover, the probability score alone also appears to be a much better selector for the true active compounds than the assay activity criterion alone, as illustrated by the high confirmation rate (over 80%) when this score is set at a low value (e.g., log $P_0 \leq -4$) regardless of the activity threshold (Figure 3).

However, it should be noted that for a completely diversified compound collection where all compound families are singletons (this is most likely hypothetical), the hypergeometric P_0-value score becomes equivalent to the activity score. That is, if the screening compound set is extremely diversified with a large portion of the clusters being singleton, there is no difference to use the activity threshold or the probability score as the selector. Our new approach, in this case, degenerates to the simple cutoff-based approach, as there is no compound cluster information/knowledge to guide the hit selection process. Nonetheless, since for any typical uHTS compound library there often exists some level of structural redundancy, by clustering the compound collection before the hit-picking process, our scaffold-based method, as demonstrated above, can effectively eliminate experimental artifacts (particularly those in the high-activity region) from the selected hits and therefore substantially increase the selection accuracy. Furthermore, analyzing compounds by their families also makes it convenient to apply the variable structural filters as mentioned above to eliminate the undesirable scaffolds, e.g., to eliminate all the steroids in a nuclear hormone receptor project.

The new OPI score takes great advantage of the structural redundancy in the screening library. For any compound cluster that is not a singleton (i.e., a cluster containing more than one structurally similar compounds), the statistical OPI P_0 score can effectively rank the compound scaffold for hit-picking based on the enrichment of compounds with high activities. For a simple example, given a cluster containing 5 compounds, if there are 4 member compounds from this group found to be among the first 200 most active ones (of a total of 50 000 screened compounds), the log P_0 score for this cluster is -8.9, which is a very desirable score for this cluster to be selected since there is a clear enrichment of active compounds in this scaffold. In contrast, if there is only one compound from this five-membered cluster found in the first 200 ones, the log P_0 score becomes -1.7. We would significantly deprioritize this scaffold, because without SAR support this only "active" compound may likely be a false positive.

One potential concern of the OPI score may be raised that the hypergeometric score appears to penalize the singletons (i.e., compound cluster that contains only one compound). For example, if a marginal singleton is ranked the 100th based on activity among the 50 000 compounds, this compound may be picked by the "top X" method, given that the hit rate is set at 0.2%. (In real uHTS data analysis, it is unlikely to set such a high hit rate.) The log P_0 score for this singleton is only -2.7, which may concern some readers that this compound might become a false negative according to the new approach. However, as far as our new scaffold-based method is concerned, this is reasonable because such a singleton, despite its marginally acceptable screening activity, does not have any compound family members for statistical support to earn a more favorable score compared to a decent-sized scaffold family of slightly weaker average activities—this is exactly what we aim to achieve as explained above. Two key points also need to be mentioned: (1) The OPI score provides statistical rankings rather than making an arbitrary active/inactive call; therefore, one can always

lower the P_0-value threshold to include those aforementioned questionable singletons as hits, if sufficient resources are available for the confirmation test. (2) Besides the structural redundancy, the OPI score also takes the activity into account. For example, if the above singleton had a high potency, say ranked 10th, the OPI score would then have been −3.7. Having a superior score than some modestly active compound family with multiple members, a truly active singleton will not be easily missed in our hit-picking process.

Moreover, this scaffold-based HTS hit selection approach is in essence driven by the SAR knowledge. In order for it to effectively identify truly active compounds from the often noisy uHTS data, it relies on a presumption that chemically similar active compounds within a scaffold possess a certain level of SAR. By effectively taking advantage of the SAR information embedded in each compound scaffold family, this method is able to pick promising active scaffolds, instead of individual and unrelated compounds, based on a rigorous statistical model. It has always been extremely challenging to make effective use of the SAR information, because SAR strength among a family of compounds depends not only on the chemical structure similarity but also on many other factors such as intended biological target, specific HTS assay, or particular chemotype, most of which are not known a priori. Another related challenge is that SAR is also probabilistic, which means only a fraction of the members in a compound family may show similar activities. Nonetheless, our method is able to provide an individually tailored activity cutoff value c_0 and a probability score P_0 for each compound scaffold using a rigorous statistical test,[43] in contrast to the one-threshold-fits-all approach employed by the cutoff-based "top X" method. In addition, the hits identified by this novel approach contain considerably more information than those obtained from the standard method, including statistical significance, scaffold information, and SAR profiles. Therefore, our new method can improve the quality of HTS hits from the very beginning of the drug discovery process and may then help facilitate discovering lead series with high information content,[5] as information such as scaffolds and SAR derived in this early HTS hit selection step are considered as favorable characteristics for promising lead series.[4,5,8] For example, Figure 4 shows some of the chemical scaffolds discovered by our approach; significant chemical diversity among the scaffolds and favorable SAR among compounds from the same chemotype were observed.

Furthermore, while the compound clustering step currently used in our method relies on fingerprints-based chemical structure similarity, it is also possible to use other properties such as various molecular descriptors to group the compounds into families[16] and then carry out a similar analysis of the HTS results using our method. For example, the PGLT algorithm mentioned above, which in essence is a knowledge-based clustering algorithm, can be applied first to group compounds into classes.[22] The resultant compound clusters can then be used as input for the OPI method to select primary HTS hits. It will also be interesting to see which clustering algorithm on what molecular properties works the best with our OPI method in terms of producing the best hit-picking results; this is clearly one future direction to further extend the OPI approach, working in synergy with compound clustering algorithms, in HTS data analysis. Moreover, additional knowledge, other than simple SAR, regarding the nature of compound bioactivity and its relationship with molecular structure may be extracted from the vast HTS/uHTS data when a set of meaningful, interpretable molecular descriptor is used.

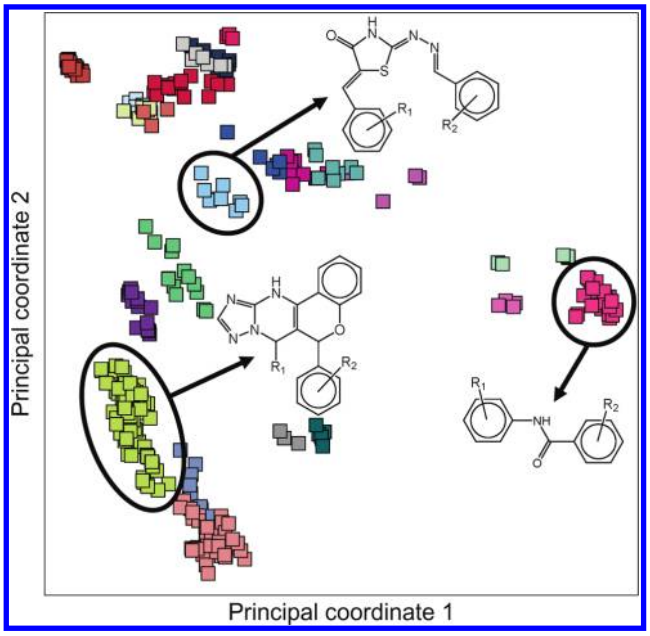

Figure 4. Various compound scaffolds discovered by our scaffold-based HTS hit selection method. The compounds are represented by their first two principal coordinates determined by principal coordinate analysis of structural similarity using Tanimoto coefficient and Daylight fingerprints.[45] Different colors represent structurally distinctive compound scaffolds.

CONCLUSIONS

In this report we proposed a novel knowledge-based statistical method for HTS/uHTS primary hit selection. Our approach not only significantly improves the hit confirmation rate compared to the widely used simple activity cutoff-based method, from 55% to 85% in a validation data set, but also allows direct identification of active scaffolds with a certain level of SAR from the often noisy HTS data. Valuable scaffold and SAR information learned at this very early stage of the hit-to-lead process can greatly facilitate the identification of high-quality lead series in later steps and may contribute to the overall success rate of the drug discovery process.

REFERENCES AND NOTES

(1) Oprea, T. I. Chemoinformatics in lead discovery. *Chemoinformatics in Drug Discovery*; Wiley-VCH: Weinheim, 2005; pp 25−41.
(2) McFadyen, I.; Walker, G.; Alvarez, J. Enhancing hit quality and diversity within assay throughput constraints. *Chemoinformatics in Drug Discovery*; Wiley-VCH: Weinheim, 2005; pp 143−173.
(3) Walters, W. P.; Namchuk, M. Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* **2003**, *2*, 259−266.
(4) Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255−263.
(5) Alanine, A.; Nettekoven, M.; Roberts, E.; Thomas, A. W. Lead generation—enhancing the success of drug discovery by investing in the hit to lead process. *Comb. Chem. High Throughput Screen.* **2003**, *6*, 51−66.
(6) Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325−334.

(7) Fogel, P.; Collette, P.; Dupront, A.; Garyantes, T.; Guedin, D. The confirmation rate of primary hits: a predictive model. *J. Biomol. Screen.* **2002**, *7*, 175−190.
(8) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* **2003**, *2*, 369−378.
(9) Parker, C. N.; Schreyer, S. K. Application of chemoinformatics to high-throughput screening: practical considerations. *Methods Mol. Biol.* **2004**, *275*, 85−110.
(10) 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.
(11) Kubinyi, H. Drug research: myths, hype and reality. *Nat. Rev. Drug Discov.* **2003**, *2*, 665−668.
(12) Lahana, R. Who wants to be irrational? *Drug Discov. Today* **2003**, *8*, 655−656.
(13) Smith, A. Screening for drug discovery: the leading question. *Nature* **2002**, *418*, 453−459.
(14) Drews, J. Drug discovery: a historical perspective. *Science* **2000**, *287*, 1960−1964.
(15) Young, S. S.; Hawkins, D. M. Using recursive partitioning analysis to evaluate compound selection methods. *Methods Mol. Biol.* **2004**, *275*, 317−334.
(16) Lam, R. L.; Welch, W. J.; Young, S. S. Cell-based analysis of high-throughput screening data for drug discovery; Research Report RR-02-02; Institute for Improvement in Quality and Productivity, University of Waterloo, 2002.
(17) Xue, L.; Stahura, F. L.; Bajorath, J. Cell-based partitioning. *Methods Mol. Biol.* **2004**, *275*, 279−290.
(18) Willett, P. Evaluation of molecular similarity and molecular diversity methods using biological activity data. *Methods Mol. Biol.* **2004**, *275*, 51−64.
(19) van Rhee, A. M. Use of recursion forests in the sequential screening process: consensus selection by multiple recursion trees. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 941−948.
(20) Schreyer, S. K.; Parker, C. N.; Maggiora, G. M. Data shaving: a focused screening approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 470−479.
(21) Hopfinger, A. J.; Duca, J. S. Extraction of pharmacophore information from high-throughput screens. *Curr. Opin. Biotechnol.* **2000**, *11*, 97−103.
(22) 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.
(23) Engels, M. F. M.; Venkatarangan, P. Smart screening: approaches to efficient HTS. *Curr. Opin. Drug Discov. Dev.* **2001**, *4*, 275−283.
(24) Young, S. S.; Lam, R. L.; Welch, W. J. Initial compound selection for sequential screening. *Curr. Opin. Drug Discov. Dev.* **2002**, *5*, 422−427.
(25) Rusinko, A., III.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 1017−1026.
(26) Zhang, J. H.; Chung, T. D.; Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* **1999**, *4*, 67−73.
(27) Zhang, J. H.; Chung, T. D.; Oldenburg, K. R. Confirmation of primary active substances from high throughput screening of chemical and biological populations: a statistical approach and practical considerations. *J. Comb. Chem.* **2000**, *2*, 258−265.
(28) Fay, N.; Ullmann, D. Leveraging process integration in early drug discovery. *Drug Discov. Today* **2002**, 7, S181−S186.
(29) Brideau, C.; Gunter, B.; Pikounis, B.; Liaw, A. Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* **2003**, *8*, 634−647.
(30) Engels, M. F. M.; Wouters, L.; Verbeeck, R.; Vanhoof, G. Outlier mining in high throughput screening experiments. *J. Biomol. Screen.* **2002**, *7*, 341−351.
(31) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. LeadScope: software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.
(32) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database diversity assessment: new ideas, concepts, and tools. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 447−452.
(33) Pearlman, R. S.; Smith, K. M. Novel software tools for chemistry diversity. *Perspect. Drug Discov. Des.* **1998**, *9*, 339−353.
(34) Lam, R. L.; Welch, W. J.; Young, S. S. Uniform coverage designs for molecule selection. *Technometrics* **2002**, *44*, 99−109.
(35) Welch, W. J.; Lam, R. L.; Young, S. S. Cell-based analysis of high throughput screening data for drug discovery, PCT Int. Appl. WO 02/12568 A2, 2002.
(36) Golebiowski, A.; Klopfenstein, S. R.; Portlock, D. E. Lead compounds discovered from libraries. *Curr. Opin. Chem. Biol.* **2001**, *5*, 273−284.
(37) Golebiowski, A.; Klopfenstein, S. R.; Portlock, D. E. Lead compounds discovered from libraries: part 2. *Curr. Opin. Chem. Biol.* **2003**, *7*, 308−325.
(38) Oprea, T. I. Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **2002**, *6*, 384−389.
(39) Rose, S.; Stevens, A. Computational design strategies for combinatorial libraries. *Curr. Opin. Chem. Biol.* **2003**, *7*, 331−339.
(40) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882−894.
(41) 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.
(42) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **2005**, *1*, 146−148.
(43) 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.
(44) Young, J. A.; Fivelman, Q. L.; Blair, P. L.; de la Vega, P.; Le Roch, K. G.; Zhou, Y.; Carucci, D. J.; Baker, D. A.; Winzeler, E. A. The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol. Biochem. Parasitol.* **2005**, *143*, 67−79.
(45) James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual*: *Daylight Version 4.9*; Daylight Chemical Information Systems, Inc.: Mission Viejo, CA, 2005.
(46) Zar, J. H. *Biostatistical Analysis*; Prentice Hall: Upper Saddle River, NJ, 1999.
(47) 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.
(48) 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.

CI0502808