This article has been modified from the original printed version. Originally the caption for Figure 4 on page 684 had incorrect colors as the key for the figure. They are now correct.
Practical Approaches to Efficient Screening:
Information-Rich Screening Protocol

PRABHA S. KARNACHI and FRANK K. BROWN

The past approach of high-throughput screening of everything in the corporate collection has been shown to be very expensive in terms of reagents cost, disposal cost, and compound collection depletion. It is well known that screening campaigns produce several hits, of which only 50% confirm on average. More efficient ways of screening can provide an informative structure-activity relationship (SAR), which in turn can be used to build mathematical models for further probing the activity space and directing chemical synthesis. The authors report new methods and insights to extract the maximum possible information from a screening experiment and find most of the possible hits in the corporate collection while screening as few compounds as possible. (Journal of Biomolecular Screening 2004:678-686)

Key words: iterative focused screening, in silico screening, sequential screening, cluster-based follow-up screening

INTRODUCTION

Lead candidates in drug discovery are routinely identified by screening large corporate collections that are supplemented by combinatorial chemistry synthesis and external compounds available through a wide network of suppliers. In corporate high-throughput screening (HTS) environments, as the number of targets and compounds increases, costs escalate dramatically. To use the limited resources of proteins and compounds from inventory more efficiently, one needs to increase the focus and efficiency of screening operations. An alternate screening strategy, termed sequential screening, partners in silico and in vitro screening in an integrated screening process. Sequential screening is usually preferred over complete screening in cases in which the capacity of the assay is limited or when sequential screening is more cost-effective. The sequential screening process starts with the screening of a relatively small initial set of compounds in the first HTS run to locate the active regions of chemical space within the collection. Based on these results, the entire collection is screened in silico, and additional compounds are identified for a second round of screening. Several cycles of HTS and in silico screening are carried out, leading to identification of more active compounds and refinement of the structure-activity relationship (SAR) model(s). The screening runs are performed on the focused portions of the chemical space; this in turn leads to biological testing of a smaller number of compounds compared to the complete screening of the compound collection. The advantage of having a smaller subset of the collection to screen is that one can either test the samples in duplicate or retest the hits from screening rounds and get better quality data. This leads to more precise SAR information, making it easier to pick a lead series for synthetic exploration.

At the end of the sequential screening rounds, one is able to identify both the active and inactive molecules and build a “predictable” SAR model from the data. The SAR model can be used to identify additional compounds with a greater likelihood of being active from either the same compound collection (previously not screened) or from a virtual combinatorial library or external compound collection. In addition, 3-dimensional pharmacophore models can be built, based on the data, to further mine the collection for active compounds. This is an iterative process in which the SAR model is refined as the sequential screening process progresses and the paradigm moves from lead identification to lead optimization.

If the cost of screening a single compound is high in a given assay, then screening a small subset of the collection in the sequential screening experiment may be economical. The initial representative subset can be screened rapidly against several targets in drug discovery in the first sequential screening round, such that several projects will be able to efficiently use the screening resources. Based on the hits from the first screening round, compounds for the subsequent screening rounds are selected. The logistics and the cost associated with this “cherry-picking” process depend on factors such as the nature of storage of compound collection (plates or individual solubilized or neat stock) and the ease of custom plating the follow-up compounds for the subsequent screening rounds.
Several computational approaches to sequential screening for rapid in silico screening and data analysis are reviewed in the literature.\textsuperscript{16,7,10-14} Cluster analysis is one such approach to sequential screening and is discussed in this article. Clustering techniques are based on the basic assumption in medicinal chemistry that similar compounds exhibit similar biological activity.\textsuperscript{15} Hence, if a representative member of the cluster is tested and found to be active, then the probability of other members of the cluster to be active is high.

In this article, we present a retrospective analysis of an in-house HTS data set using clustering analysis for sequential screening. The study addresses the following issues: (1) hit rate in sequential screening compared to that of the complete screening campaign, (2) method of selection of the initial sample for the first round of screening, (3) choice of descriptor set(s) for defining the chemical space, (4) the initial sample size and its effect on hit rate, (5) coverage of structural classes of actives, and (6) mathematical model building for further screening of new compound collections based on the results from the sequential screening rounds.

**MATERIALS AND METHODS**

Hierarchical clustering method was used in this study to define the initial sample for screening.\textsuperscript{16,17} This method produces small homogeneous clusters, in which for a given cluster, the members are similar to each other and are dissimilar to members of other clusters. In particular, Ward’s method\textsuperscript{18} has been shown to be good at separating similar and dissimilar molecules into different clusters.\textsuperscript{19,20} The BCI (Barnard Chemical Information) implementation\textsuperscript{21} of Ward’s algorithm is an agglomerative clustering method. The BCI implementation calculates a reciprocal nearest neighbor list for each molecule and finds and merges the 2 most similar molecules into different clusters.\textsuperscript{21} When merging the 2 clusters, Ward’s method maximizes intercluster variance while minimizing intrACLuster variance. The data set can be partitioned into any number of clusters using the reciprocal nearest neighbor list. Once the data set is partitioned into homogeneous clusters, a representative from each cluster is selected. The representative is usually a member closest to the cluster centroid (geometrical mean of the cluster).\textsuperscript{6,20} Alternatively, a random member of the cluster was also selected as a representative.

In the sequential screening approach, cluster analysis is used to select a representative subset of the compound collection for the initial screening round, followed by selecting members of the clusters with the active representative for the second round of screening. Based on the actives derived from the first and subsequent screening rounds, different structural classes of the hits and the relevant structural features essential for activity are identified. As a follow-up tool, SAR models are built based on 2-dimensional descriptors.

The initial sample size selected for screening depends on the logistics of screening, assay costs, and HTS assay variability.\textsuperscript{22} The hierarchical clustering produces a dendrogram that shows the relationship between each individual molecule and builds up to each individual cluster and then to multiple clusters. Different cuts in the cluster hierarchy produce a corresponding initial sample for screening. Kelley et al.\textsuperscript{23} have introduced a penalty function used to determine the most suitable cut through the cluster hierarchy to obtain maximum homogeneity within a cluster while optimizing the total number of clusters that a given data set should be divided up into. The Optclus program of the BCI software calculates the Kelley penalty function.\textsuperscript{24} The penalty value for different levels in the cluster hierarchy is calculated, and the cut with the corresponding smallest penalty value is selected as the reference level. This reference level will be henceforth referred to as the *optimal cluster level*. Abt and coworkers have suggested 5000 to 10,000 compounds (~7%-15% of the data set) as a sufficient starting point for a sequential screening experiment.\textsuperscript{10}

Various 2-dimensional descriptors such as Daylight fingerprints\textsuperscript{25} and MACCS (Molecular ACCess System)\textsuperscript{26} keys, in combination with Ward’s clustering algorithm, have been shown to be effective in separating the actives from inactives.\textsuperscript{19} On the other hand, the exact combination of descriptor type and clustering algorithm may depend on the data set used and the measured biological activity. Our approach to this issue was, rather than a priori selection of descriptors, to use 3 types of 2-dimensional descriptors to represent the chemical space of our data set. The 2-dimensional descriptors used were the hashed Daylight fingerprints,\textsuperscript{25} MACCS keys,\textsuperscript{26} and electrotopological descriptors.\textsuperscript{27,28} The data set was clustered using Ward’s algorithm, as implemented in the BCI software for each descriptor type separately.\textsuperscript{21} A representative (centroid or a random member of each cluster) for each descriptor type was selected separately as the initial sample for the first round of screening. The representatives were subjected to the first round of biological testing. The active representatives were identified, and the corresponding cluster members were then subjected to the second round of screening, leading to further identification of active molecules in the data set. This was done for each of the 3 types of descriptors separately, as schematically shown in Figure 1.

The chemical space of the data set, as described by the 3 types of descriptors, is shown in Figure 1. The active representatives (X) and their active cluster members (a) represent the active chemical space, as shown in Figure 1. For a given set of molecules, the distribution of the molecules in chemical space may be different based on the type of descriptor used to define the chemical space. For a given descriptor, the initial sample for the first sequential screening round should adequately represent the compound collection (i.e., the chemical space), as defined by the molecules in the collection. When a single descriptor is used to define a chemical space in a sequential screening experiment, there is a possibility that initial and follow-up samples lie in a very narrow region of the chemical space. We term this the local minimum effect. By combining the representation of chemical space based on 3 separate descriptor types, we avoid being trapped in a “local minimum.” This is illustrated in Figure 2. In the case in which only 1 descriptor type is used and the representative is inactive, all the active members of
The third screening round.

of cluster representative for the Daylight fingerprint descriptor. The presence of "a" in the areas of intersection between the descriptor types leads to "active". The active representatives and the corresponding active cluster members were harvested and activity checked in the second screening round. For the combined method, the representatives from each of the descriptor methods were pooled together in the first screening round. For the active representatives, irrespective of the source of the descriptor type they originated from, the corresponding cluster members from all of the descriptor types were collected, redundant compounds were removed, and activity was checked in the second screening round. The hit rates for the first and second rounds were calculated. In addition, the percentage of the actives retrieved and the percentage of the database screened at the end of the screening rounds (first and second rounds) were calculated.

For further testing the rationale of combining 3 types of descriptors, as shown in Figure 2, a third round of screening was initiated. The goal of the third screening round was to identify actives that are lost due to an inactive cluster representative. The active representatives and the corresponding active cluster members for any 2 types of descriptors taken separately, identified from the first and second rounds of screening, were pooled together. The clusters for the remaining third type of descriptor containing the above-mentioned actives were selectively isolated. If the representatives of these clusters for the third type of descriptor were inactive, then these clusters would not have been selected for the second round of screening. However, the third round of screening isolated these clusters as "previously untested" and screened for activity. These newly identified active cluster members would have been lost because the representative was found to be inactive in the first round of sequential screening.

At the end of the sequential screening rounds, we screened a certain percentage of the data set to identify the active and inactive compounds. The value of identifying not only the active compounds but also the inactive compounds is in the building of "predictable" mathematical models. These models can now be used to identify additional compounds with a higher probability of being active from the previously untested part of the data set or a virtual library or an external compound collection. In this study, we use the QSAR-like methodology termed binary QSAR, which builds

that cluster would be lost. The inactive representative may be a "true" inactive or a false negative. The combined method may lead to an overrepresentation of the activity space and, consequently, to more molecules being screened in the first screening round. However, this over-representation can be used to thoroughly explore the activity space and allow for cluster hopping. Because the method allows for cluster hopping, the active compounds in the cluster with the "inactive" centroid can be identified in the third round of screening. The flowchart of sequential screening rounds described in this article is shown in Figure 3.

APPLICATION

A data set of 147,474 compounds was screened in an enzyme-based high-throughput assay for an in-house target. The primary HTS activity was measured as percent inhibition at 50 µM concentration and rated from 1 to 5. The definitions of the rating system are as follows: 5 (highly active), 4 (moderately active), 3 (weakly active), and 1 and 2 (inactive). The compounds with ratings 3 to 5 were repeated in the HTS assay to confirm their biological activity, and 791 compounds retained the rating of 3 to 5 (active). The remaining data set was termed as inactive. The descriptors used were Daylight fingerprints, 166 MACCS keys, and 46 electrotopological descriptors calculated using Cerius² software. These descriptors were calculated for the entire data set, and the data set was clustered separately for the 3 types of descriptors using the BCI implementation of Ward’s algorithm. For each of the descriptor types, the total number of clusters was set at the optimal cluster level. The optimal cluster level was different for each descriptor space. A representative was selected from each cluster as the initial sample for the first round of screening, using each descriptor type separately. The representative selected as the initial sample was either the centroid or a random member of the cluster. If the representative was found to be active, then the members of the corresponding clusters for the corresponding type of descriptor were harvested and activity checked in the second screening round. For the combined method, the representatives from each of the descriptor methods were pooled together in the first screening round. For the active representatives, irrespective of the source of the descriptor type they originated from, the corresponding cluster members from all of the descriptor types were collected, redundant compounds were removed, and activity was checked in the second screening round. The hit rates for the first and second rounds were calculated. In addition, the percentage of the actives retrieved and the percentage of the database screened at the end of the screening rounds (first and second rounds) were calculated.

FIG. 1. The chemical space of the dataset as described by the three types of descriptors is shown. The circles represent the cluster boundaries. Xa is the inactive representative from the first round of screening and "a" is the active cluster member from the second round for each of the descriptor type.

FIG. 2. The chemical space of the dataset as described by the three types of descriptors is shown. The circles represent cluster boundaries. Xa is the active representative from the first round of screening and "a" is the active cluster member from the second screening round. X is the inactive cluster representative for the Daylight fingerprint descriptor. The presence of "a" in the areas of intersection between the descriptor types leads to "active". The active representatives and the corresponding active cluster members were harvested and activity checked in the second screening round. For the combined method, the representatives from each of the descriptor methods were pooled together in the first screening round. For the active representatives, irrespective of the source of the descriptor type they originated from, the corresponding cluster members from all of the descriptor types were collected, redundant compounds were removed, and activity was checked in the second screening round. The hit rates for the first and second rounds were calculated. In addition, the percentage of the actives retrieved and the percentage of the database screened at the end of the screening rounds (first and second rounds) were calculated.

For further testing the rationale of combining 3 types of descriptors, as shown in Figure 2, a third round of screening was initiated. The goal of the third screening round was to identify actives that are lost due to an inactive cluster representative. The active representatives and the corresponding active cluster members for any 2 types of descriptors taken separately, identified from the first and second rounds of screening, were pooled together. The clusters for the remaining third type of descriptor containing the above-mentioned actives were selectively isolated. If the representatives of these clusters for the third type of descriptor were inactive, then these clusters would not have been selected for the second round of screening. However, the third round of screening isolated these clusters as "previously untested" and screened for activity. These newly identified active cluster members would have been lost because the representative was found to be inactive in the first round of sequential screening.

A data set of 147,474 compounds was screened in an enzyme-based high-throughput assay for an in-house target. The primary HTS activity was measured as percent inhibition at 50 µM concentration and rated from 1 to 5. The definitions of the rating system are as follows: 5 (highly active), 4 (moderately active), 3 (weakly active), and 1 and 2 (inactive). The compounds with ratings 3 to 5 were repeated in the HTS assay to confirm their biological activity, and 791 compounds retained the rating of 3 to 5 (active). The remaining data set was termed as inactive. The descriptors used were Daylight fingerprints, 166 MACCS keys, and 46 electrotopological descriptors calculated using Cerius² software. These descriptors were calculated for the entire data set, and the data set was clustered separately for the 3 types of descriptors using the BCI implementation of Ward’s algorithm. For each of the descriptor types, the total number of clusters was set at the optimal cluster level. The optimal cluster level was different for each descriptor space. A representative was selected from each cluster as the initial sample for the first round of screening, using each descriptor type separately. The representative selected as the initial sample was either the centroid or a random member of the cluster. If the representative was found to be active, then the members of the corresponding clusters for the corresponding type of descriptor were harvested and activity checked in the second screening round. For the combined method, the representatives from each of the descriptor methods were pooled together in the first screening round. For the active representatives, irrespective of the source of the descriptor type they originated from, the corresponding cluster members from all of the descriptor types were collected, redundant compounds were removed, and activity was checked in the second screening round. The hit rates for the first and second rounds were calculated. In addition, the percentage of the actives retrieved and the percentage of the database screened at the end of the screening rounds (first and second rounds) were calculated.

For further testing the rationale of combining 3 types of descriptors, as shown in Figure 2, a third round of screening was initiated. The goal of the third screening round was to identify actives that are lost due to an inactive cluster representative. The active representatives and the corresponding active cluster members for any 2 types of descriptors taken separately, identified from the first and second rounds of screening, were pooled together. The clusters for the remaining third type of descriptor containing the above-mentioned actives were selectively isolated. If the representatives of these clusters for the third type of descriptor were inactive, then these clusters would not have been selected for the second round of screening. However, the third round of screening isolated these clusters as "previously untested" and screened for activity. These newly identified active cluster members would have been lost because the representative was found to be inactive in the first round of sequential screening.

At the end of the sequential screening rounds, we screened a certain percentage of the data set to identify the active and inactive compounds. The value of identifying not only the active compounds but also the inactive compounds is in the building of "predictable" mathematical models. These models can now be used to identify additional compounds with a higher probability of being active from the previously untested part of the data set or a virtual library or an external compound collection. In this study, we use the QSAR-like methodology termed binary QSAR, which builds
predictive binary models and calculates the probabilistic distribution for active (activity = 1) and inactive (activity = 0) compounds in a training set. The compounds screened in the sequential screening rounds were used as the training set in developing the binary QSAR model. The actives (confirmed rating 3-5) from the sequential screening rounds were given a value of 1 for biological activity, and the remaining compounds were assigned a value of 0. The predictive binary QSAR model was built using the MACCS key occurrence count as a descriptor. The MACCS key occurrence count is a descriptor that encodes not only the presence or absence of a MACCS key but also the occurrence count of the keys that are present. Thus, the MACCS keys that are present and occur more frequently (higher occurrence count) are given more weight than the keys that are present but occur less frequently in the model-building process. The model was then used to identify previously unscreened compounds from the data set with a good likelihood of being active. The statistical model can be used to design and virtually screen focused combinatorial libraries or select potentially active compounds from external compound collection.

The percentage of the actives recovered depends on the initial sample size used in the first round of screening. The minimum in Kelley’s penalty function provides a point of reference for the cut in the cluster hierarchy. Various cuts corresponding to different values of the penalty function provide a different initial sample size. This was tested for the combined method. The cluster representatives for each descriptor type, based on the corresponding cut in the cluster hierarchy, were pooled for the combined method. The Optclus program of the BCI software calculates the optimal number of clusters for the data set for each descriptor type. This is the reference point (optimal cluster level) in the dendrogram. The data set was clustered at 0.5, 0.75, 1.5, and 2 times the optimal cluster level. At 0.5 times the optimal cluster level, the number of clusters...
is half as much as that at the optimal level, and the cluster membership is comparatively large. On the other hand, at 2 times the optimal cluster level, the number of clusters is twice that at the optimal level, and the cluster membership is relatively small. The cluster representatives were screened at each cluster level, and cluster members of active representatives were also screened. The total percentage of the actives recovered and the corresponding percentage of the database screened were calculated.

The next exercise is to explore the level of information from the sequential screening rounds through coverage of the structural classes (chemotypes) of the actives. Using the MACCS keys occurrence counts as descriptors, the 791 actives were clustered at the optimal cluster level. The assumption here is that each cluster represents a structural class (chemotype) of the actives. Clusters containing at least 1 active, identified from the screening rounds, were isolated. The more the number of such isolated clusters, the better coverage of the chemotypes of the actives. A study by van Rhee and coworkers reported that by using the recursive partition, they were able to obtain ~75% of the actives by screening <20% of a collection.

### RESULTS AND DISCUSSION

The hit rates at the end of the first and second rounds were determined at the optimal cluster level for the 3 types of descriptors separately and for the combined method. The optimal cluster level is the lowest in Kelley’s penalty function, as determined by the Optclus program of the BCI software for each descriptor type. The hit rates were determined when the representative was either the centroid or a random member of the cluster. The compounds were clustered separately for the 3 types of descriptors. The results are shown in Table 1. The electrotopological descriptors, MACCS keys, and Daylight fingerprints were optimally clustered into 8735, 15,807, and 17,103 number of clusters, respectively. A cluster representative was selected for each of the 3 methods and checked for activity. In the case when the representative was the centroid (Table 1), the percent hit rates for the first round of screening were 1.36%, 0.56%, and 0.6% for the electrotopological descriptors, MACCS keys, and Daylight fingerprints, respectively. The cluster members of the active centroid were then collected and checked for activity. The percent hit rates for the second round for the electrotopological, MACCS keys, and Daylight fingerprints were 19.3%, 31.3%, and 31.46%, respectively. In this study, 31% to 33% of the actives were recovered on screening 6% to 12% of the database at the end of 2 rounds of sequential screening based on the 3 descriptor types. When the cluster representative was a random member of the cluster (Table 1), 33% to 36% of the actives were recovered on screening 6% to 12% of the database at the end of the 2 rounds of sequential screening based on the 3 descriptor types.

For the combined method, the cluster representatives from each of the 3 types of descriptors were pooled together and screened.

---

### Table 1. Results at the End of the First and Second Rounds of the Sequential Screening Experiment at the Optimal Cluster Level

<table>
<thead>
<tr>
<th></th>
<th>Electrotopological Properties</th>
<th>MACCS Keys</th>
<th>Daylight Fingerprints</th>
<th>Combined Method (Union)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centroid Random Member</td>
<td>Centroid Random Member</td>
<td>Centroid Random Member</td>
<td>Centroid Random Member</td>
</tr>
<tr>
<td>High-throughput screening assay (% hit rate)</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>Number of cluster representatives</td>
<td>8735</td>
<td>8735</td>
<td>15,807</td>
<td>15,807</td>
</tr>
<tr>
<td>Active representative (% hit rate, first round)</td>
<td>1.36</td>
<td>1.44</td>
<td>0.56</td>
<td>0.7</td>
</tr>
<tr>
<td>Number of follow-up samples</td>
<td>658</td>
<td>607</td>
<td>644</td>
<td>767</td>
</tr>
<tr>
<td>Active follow-up sample (% hit rate, second round)</td>
<td>19.3</td>
<td>25.2</td>
<td>31.3</td>
<td>22.55</td>
</tr>
<tr>
<td>% of total actives</td>
<td>31.1</td>
<td>35.3</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>% of data set screened</td>
<td>6.4</td>
<td>6.4</td>
<td>11</td>
<td>11.25</td>
</tr>
</tbody>
</table>

MACCS, Molecular ACCess System.

### Table 2. Comparison of Results of the First and Second Rounds of Screening for Similar Percentages of the Database Screened for the Individual Descriptor Types and the Combined Method

<table>
<thead>
<tr>
<th></th>
<th>Electrotopological Properties</th>
<th>MACCS Keys</th>
<th>Daylight Fingerprints</th>
<th>Combined Method (Union)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centroid Random Member</td>
<td>Centroid Random Member</td>
<td>Centroid Random Member</td>
<td>Centroid Random Member</td>
</tr>
<tr>
<td>% of total actives</td>
<td>42</td>
<td>50</td>
<td>46</td>
<td>53</td>
</tr>
<tr>
<td>% of database screened</td>
<td>23</td>
<td>26</td>
<td>22</td>
<td>27</td>
</tr>
</tbody>
</table>

MACCS, Molecular ACCess System.
The active representative was selected, and cluster members of the active representative from all 3 types of descriptors were screened. At the end of the second round for the combined method, 58% of the actives were recovered on screening 24% of the database, when the cluster representative was the centroid of the cluster (Table 1). When the representative was a random member of the cluster, at the end of the first follow-up round for the combined method, 70.3% of the actives were recovered upon screening 28.5% of the database (Table 1).

The hypothesis of increasing the hit rate in the follow-up rounds by combining descriptor spaces was tested. The question was as follows: is the hit rate better for the combined method simply because more compounds were included, or did the different descriptor spaces provide additional information? To make such a comparison, the percentage of the database screened at the end of the second screening round for the individual descriptors and the combined method should be the same. The descriptors were separately clustered at a level such that at the end of the second screening round, around 24% of the data set would be screened. The results of this hypothesis testing are shown in Table 2. For the centroid as the representative, by screening ~24% of the database, the percentage of the actives recovered for each of the 3 types of descriptor is 12% to 16% less than that recovered by the combined method. When the representative was a random member of the cluster, by screening ~28% of the database, the percentage of the actives recovered for each of the 3 types of descriptor is 17% to 20% less than that recovered by the combined method. The percentage yield of the actives in the sequential screening protocol was higher than randomly picking the same number of compounds from the database. The combination of the descriptor space retrieves a higher percentage of the actives in the data set while screening a similar percentage of the database compared to the any of the individual descriptor spaces.

The third screening round (Fig. 2) allows for retrieval of additional active compounds from clusters with inactive representatives for a given descriptor. These active compounds are close to the active chemical space, as identified by the active compounds from either one or both of the remaining descriptor types. For the optimal cluster level and the combined method, when the centroid was the cluster representative, an additional 3% of the database was screened to retrieve 11% more actives (Table 3). Thus, at the end of the third round of sequential screening, we had screened 27% of the database to recover 69% of actives. For the random member as cluster representative (Table 3), the third round did not provide additional significant retrieval of actives. However, it must be noted that the percentage of actives recovered was higher (70.3%) at the end of the second round (Table 2).

Screening of clusters or a “chemical family” in a sequential screening experiment allows for SAR around the hits retrieved from the first round of screening. On the other hand, if a random set of compounds is screened and actives and inactives identified, the presence or absence of activity for individual compounds may be difficult to attribute to any trends in SAR. A predictable SAR model was built using binary QSAR from the active and inactive compounds identified from the sequential screening rounds described above. For the optimal cluster level with centroid as the representative, the optimal binary QSAR model was obtained by a combination of a principal number of components of 50 and a smoothing factor of 0.2 for the MACCS keys occurrence counts. For this combination, the non-cross-validated accuracy was 75% on active compounds, 88.3% on inactive compounds, and 86% for all the compounds. The binary QSAR model was cross-validated using the leave-one-out method. The cross-validated accuracy was 69% on active compounds, 88.3% on inactive compounds, and 82% for all the compounds. The predictive binary QSAR model was used to identify additional compounds for the fourth round of screening. An additional 11% of the actives were identified on screening 2% more of the database (Table 3). Thus, at the end of the 4 rounds of screening, we were able to retrieve 80% of the actives by screening 29% of the database. Because the percentage of the actives retrieved in the third round was not significantly more than the second round with a random member of the cluster as the representative, the binary QSAR model was developed at the end of the second screening round. The binary QSAR model identified an additional 8% of the actives by screening 2.5% more of the database (Table 3). The binary QSAR model can be refined iteratively as more actives are identified.

The number of sequential screening rounds, the number of compounds screened per round, and the total number of compounds to be screened depend on the logistics and the cost-effectiveness of the HTS assay. The first decision that had to be made was the size of the initial sample for starting the sequential screening experiment. The initial sample size was varied by providing different cuts in the cluster hierarchy and using the lowest value in the penalty function (optimal cluster level) for a given type of descriptor as a reference point. The data set was clustered at 0.5,

<table>
<thead>
<tr>
<th>Table 3. Identification of Additional Active Compounds from the Third and Fourth Rounds of Screening for Clusters at the Optimal Cluster Level Using the Combined Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>First and Second Rounds</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>% of total actives</td>
</tr>
<tr>
<td>% of data set screened</td>
</tr>
</tbody>
</table>

The combination of the descriptor space retrieves a higher percentage of the actives in the data set while screening a similar percentage of the database compared to the any of the individual descriptor spaces.
0.75, 1.5, and 2 times the optimal cluster level and the cluster representatives pooled for the combined method. The hit rate for the first screening round was 0.6% to 0.9% and was independent of the size of the initial sample and the method of selection of the cluster representative (Table 4). With the centroid as the cluster representative for the combined method, as we go from half the optimal cluster level ($0.5\times$) to the optimal cluster level, the gain in the total percentage of the actives recovered for the percentage of the database screened was not substantial. An additional 3.6% of the actives were recovered on screening 6.8% more of the database at the end of the second round (Table 4). Because the gain in the third round was not significant, the binary QSAR model building (fourth and fifth rounds) was done after the second round (Table 5). The low yield in the third round can be rationalized by the logic that at 0.5 times the optimal cluster level, the cluster memberships are larger, and hence the overlap in the chemical space for the 3 types of descriptors is large. At the end of the fifth round (which included the 2 SAR model-building steps), for the 0.5 times the optimal number of clusters, 81% of the actives were recovered on screening 31% of the database (Table 5). This was comparable to the results in Table 3, which shows that at the optimal cluster level, at the end of the fourth round, 80% of the actives were recovered on screening 29% of the database. On the other hand, at the 2 times the optimal number of cluster level ($2\times$), 79.4% of the actives were recovered by screening 34% of the database at the first follow-up round of screening (Table 4). Thus, when the centroid was selected as the cluster representative, based on the size of the initial sample, the number of rounds of screening varied to obtain a similar percentage of actives recovered for the percentage of the database screened. When the initial sample size was small, the additional rounds of screening were needed to get to the same end point. Similar results were obtained when the cluster representative was a

### Table 4. Effect of Initial Sample Size in Retrieving Actives from the First and Second Rounds of Sequential Screening for the Combined Descriptors Method

<table>
<thead>
<tr>
<th></th>
<th>0.5x Optimal Cluster Level</th>
<th>0.75x Optimal Cluster Level</th>
<th>Optimal Cluster Level</th>
<th>1.5x Optimal Cluster Level</th>
<th>2x Optimal Cluster Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centroid</td>
<td>Random Member</td>
<td>Centroid</td>
<td>Random Member</td>
<td>Centroid</td>
</tr>
<tr>
<td>% Hit rate (first run)</td>
<td>0.74</td>
<td>0.65</td>
<td>0.75</td>
<td>0.75</td>
<td>0.7</td>
</tr>
<tr>
<td>% Hit rate (second run)</td>
<td>5.06</td>
<td>5.3</td>
<td>5.25</td>
<td>5.1</td>
<td>15</td>
</tr>
<tr>
<td>% of total actives</td>
<td>54.4</td>
<td>51</td>
<td>62.7</td>
<td>63.1</td>
<td>58</td>
</tr>
<tr>
<td>% of database screened</td>
<td>17.2</td>
<td>16.9</td>
<td>23.2</td>
<td>23.3</td>
<td>24</td>
</tr>
</tbody>
</table>

### Table 5. Identification of Additional Active Compounds from the Third, Fourth, and Fifth Rounds of Screening for the 0.5x Optimal Cluster Level for the Combined Method

<table>
<thead>
<tr>
<th></th>
<th>First and Second Rounds</th>
<th>Third Round (Binary QSAR Model)</th>
<th>Fourth Round (Binary QSAR Model)</th>
<th>Fifth Round (Binary QSAR Model)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centroid</td>
<td>Random Member</td>
<td>Centroid</td>
<td>Random Member</td>
</tr>
<tr>
<td>% of total actives</td>
<td>54.4</td>
<td>51</td>
<td>56.4</td>
<td>52.2</td>
</tr>
<tr>
<td>% of data set screened</td>
<td>17.2</td>
<td>16.9</td>
<td>18</td>
<td>17.5</td>
</tr>
</tbody>
</table>

**FIG. 4.** Cluster representation of the structural classes of the actives. [ ] = cluster with at least one active member; [ ] = cluster with no active member; [ ] = cluster with no active member from sequential screening and binary QSAR model. [ ] = new cluster with at least one active member as identified by binary QSAR model.
random member (Tables 4 and 5). Other studies have shown that the percentage of the actives recovered depends on the size of the initial sample and on the type of descriptors used.\textsuperscript{32} The next question that we were interested in was the following: did we cover all the structural classes of the actives by retrieving 80% of the actives? This exercise was performed for the centroid as the cluster representative at the optimal cluster level for the combined method at the end of the fourth round of screening (Tables 1 and 3). For this purpose, the 791 actives were clustered into 76 clusters (Optclus level) using the MACCS key occurrence count. This meant that the 76 structural classes of compounds (as defined by the MACCS key occurrence count descriptor) were present in 791 actives based on the assumption that Ward’s clustering method produces small homogeneous clusters, such that the cluster members are similar to each other and dissimilar to members of other clusters. Of the 76 clusters, 74 had at least 1 active identified from 4 rounds of screening (Fig. 4). Thus, 97% of structural classes of the actives were identified. The number of structural classes of actives identified at the end of the sequential screening rounds provides the chemist with the option of selecting the “interesting” chemotypes both from follow-up synthetic chemistry and a patentability perspective. On the other hand, the need to identify all the structural classes of actives present in the data set may be “overkill” because chemistry efforts for any given project may be pursued on only a few different series of compounds.\textsuperscript{22} The quality of the biological data and synthetic logistics for the chemical series determines the transition from screening to chemistry.

CONCLUSIONS

It has been clearly shown that iterative-focused screening identifies most of the actives while screening about one-third of the collection. This would save a considerable amount of cost and allows for more assays to be run. However, the best value of this approach is the level of information that is produced. This iterative approach results in a predictable mathematical model that can be used to find additional hits and drive the first round of a synthetic chemistry program. This is not true for an HTS process that screens everything in single points. At the end of this type of approach, one is left with questionable data quality due to the error rate and not enough information to build a SAR.

After identification of active molecules from the first round of screening, traditionally follow-up compounds are selected as neighbors to the active molecules in the descriptor space (similarity search). The clustering approach used in this article is in fact based on a near-neighbor calculation. The advantage of using the cluster membership rather than doing an on-the-fly similarity (nearest neighbor) search is that the similar cluster members are in a lookup table that reduces the resources used to do such a search. The results of the 2 approaches should be comparable. We have also shown that the selection of neither the cluster representative (centroid or random member of the cluster) nor the number of clusters (cluster size) is critical. This is an important practical leverage of this approach because the representative may be depleted in the collection, and another member can then be substituted. This will not violate the utility, and the database does not need to be reclustered due to the need for picking a new representative. The most important aspect of this approach is the use of 3 methods that can be combined to eliminate being trapped by a false negative. Cluster hopping is a key feature of the approach and is unique compared to previous methods. It is also reassuring that the exact approach is not as important as the simple performance of an iterative sampling chemical space that allows for lead optimization, as the information around hits increases with each round of screening. Thus, screening processes must be geared toward the information they render rather than the number of compounds screened.

This iterative approach of repeated cherry picking of the collection, to create follow-up test rounds, shifts the cost of screening to the dispensing of compounds. This cost is not trivial but would be paid for by the savings in cost realized by not screening the entire collection in each HTS campaign. Also to be considered is the cost of the cherry-picking round for confirming actives in the traditional HTS program. However, the investment in a good “hay-stack” is more advantageous than spending the same amount of money on large screening campaigns. This is because data from HTS are hard to mine, reducing the overall value of this type of data, whereas the haystack can be used over and over for HTS and non-HTS screening. In our opinion, investing in a reusable technology that can be used in multiple discovery programs is a better use of resources.

The sequential screening protocol described in this article is a technology that has been in use for the past 2 years. It has worked on most targets ranging from kinases to G-protein-coupled receptors (GPCRs). However, it does fail for targets in which there are a very small number of possible hits in our screening collection. For example, this methodology did not fare well for a few GPCR targets in which the natural ligand was a large peptide. However, screening the entire compound collection for these targets did not yield any hits. The same was true for many antibacterial targets. This method is purely based on a statistical approach that needs to have a few compounds in the compound collection with reasonable potency for the receptor.

ACKNOWLEDGMENTS

We thank Michael Engels for his helpful discussion on the clustering using Daylight fingerprints and ideas on enrichment rates. We thank Mary Jo Wildey for discussion on the screening protocols and the practicality of the approach presented. We also thank the reviewers for their insightful comments.

REFERENCES

lar Modeling and Prediction of Bioactivity, Copenhagen, Denmark, August 1998.
25. MDL Information Systems. San Leandro, CA.

Address reprint requests to:
Prabha S. Karnachi, Ph.D.
Johnson & Johnson Pharmaceutical Research and Development, LLC
1000 Route 202, P.O. Box 300
Raritan, NJ 08869
E-mail: pkarnachi@prdus.jnj.com