Design and Prioritization of Plates for High-Throughput Screening

Dimitris K. Agrafiotis* and Dmitrii N. Rassokhin

3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Exton, Pennsylvania 19341

Received December 31, 2000

A general algorithm for the prioritization and selection of plates for high-throughput screening is presented. The method uses a simulated annealing algorithm to search through the space of plate combinations for the one that maximizes some user-defined objective function. The algorithm is robust and convergent, and permits the simultaneous optimization of multiple design objectives, including molecular diversity, similarity to known actives, predicted activity or binding affinity, and many others. It is shown that the arrangement of compounds among the plates may have important consequences on the ability to design a well-targeted and cost-effective experiment. To that end, two simple and effective schemes for the construction of homogeneous and heterogeneous plates are outlined, using a novel similarity sorting algorithm based on one-dimensional nonlinear mapping.

I. INTRODUCTION

In recent years, combinatorial chemistry[1] has enabled pharmaceutical research companies to augment their chemical archives with hundreds of thousands of new compounds of potentially broad pharmaceutical interest. In most cases, combinatorial libraries are synthesized in 96-well plates in the form of arrays, which comprise all possible combinations of a given set of building blocks as prescribed by the reaction scheme. Once made, these libraries represent an important and lasting corporate resource, and serve as the main source of compounds for lead discovery and lead evolution. Although several effective methodologies exist for selecting an appropriate set of reagents prior to synthesis[2−20] (for recent reviews, see refs 21−24), little attention has been paid to how the compounds are distributed among the synthesis plates and eventually archived.

The storage of libraries in plate format introduces a natural grouping of compounds that has several important consequences in high-throughput screening. Despite the existence of sophisticated hardware for extracting individual samples from large chemical banks (a process known as cherry-picking), this process is laborious and time-consuming, and can only be used when the number of compounds requested is relatively small. When throughput becomes a factor, a more cost-effective strategy is to identify the plates containing the compounds of interest, replicate them, and test them in their entirety in a high-throughput assay. This paradigm becomes increasingly appealing in the post-genomics era as it maximizes efficiency and minimizes the risk of sample depletion that would result from indiscriminate testing against an ever-increasing number of biological targets.

It does, however, raise some interesting questions on experimental design. Historically, the mining of corporate files has been based on two- or three-dimensional database searching techniques, which prioritize the compounds based on some user-defined query and produce a hit list of potential screening candidates. In high-throughput mode, this approach is no longer sufficient as one needs to take into account the collective properties of all the compounds in each plate and assign a priority score to the plate itself, rather than (or in addition to) the individual compounds in it. For example, plates which contain a large number of compounds that fit the selection criteria should be assigned a higher priority score than those containing only a few interesting compounds, as do plates that contain samples of higher purity. In addition, the plates may be interdependent; i.e., the score of a particular plate cannot be evaluated without explicit knowledge of the remaining plates comprising the selection, as is the case with virtually all diversity functions, property distributions, duplication, etc. In general, the evaluation process should be flexible and able to accommodate any selection criteria that the user may wish to employ in the design of a screening experiment.

While there are a wealth of methodologies for library design, to the best of our knowledge no method has yet been reported for the selection and prioritization of plates or, more generally, compounds that are grouped together into distinct subsets by some common characteristic such as physical location, molecular scaffold, ownership, etc. In this paper we present a general methodology for selecting a subset of plates from a large collection based on the properties of their respective compounds. The method is an extension of the stochastic optimization approach that we presented in refs 10, 25, 26, 27, and 28, and permits the expedient selection of plates that simultaneously optimize multiple design objectives. We find that the distribution of compounds among the plates may have important consequences on their subsequent use, and we present two simple and effective schemes for controlling the internal diversity of each plate using a similarity sorting algorithm based on one-dimensional nonlinear mapping. While the examples used in this study assume that the plates contain a single compound per well, the method is general and can be easily extended to handle reaction mixtures.

* Corresponding author phone: (610) 458-6045; fax: (610) 458-8249; e-mail: dimitris@3dp.com.

10.1021/ci000313d CCC: $20.00
Published on Web 04/11/2001

The results presented here are based exclusively on simulation and lack experimental support. Our intention is to present a set of useful algorithms for designing experiments of this type, and stir the attention of the combinatorial chemistry and high-throughput screening communities to this important but relatively unexplored aspect of library design. Experimental validation is currently in progress, and we hope to report our findings soon. Hopefully, other research groups with mature combinatorial chemistry and HTS programs will follow suit.

II. METHODS

Selection Algorithm. Given a collection of compounds arranged in n distinct plates, the problem is to identify a subset of k plates that satisfy some user-defined criteria. As previously mentioned, the term plate is used to denote a set of compounds grouped together by some common characteristic, and is not necessarily limited to their physical location. A plate subset is defined as

$$S \subseteq P, |S| = k, |P| = n \tag{1}$$

where P represents a collection of n plates and k is the number of plates requested. When the plates are not interdependent, the task involves a linear scan through P, assignment of scores based on the properties of the plates or their respective compounds, and selection of the k highest scoring plates. When the plates are interdependent, each candidate design S must be evaluated in its entirety in order to identify the optimal subset. The problem is similar to the selection of sparse combinatorial arrays[10] and involves a large search space, whose size is given by the binomial coefficient:

$$\frac{n!}{(n-k)!k!} \tag{2}$$

Since in most practical applications the combinatorial nature of the problem precludes any form of systematic search, the approach taken here is to combine all the selection criteria in the form of an objective function, and to maximize (or minimize) that function using simulated annealing. Simulated annealing is a global optimization technique based on the Metropolis Monte Carlo search algorithm. The method starts from an initial random state and walks through the state space associated with the problem of interest by a series of small, stochastic steps. In the problem at hand, a state represents a particular subset of k plates, and a step is a small change in the composition of that set (i.e., replacement of a small fraction of plates comprising the subset). An objective function, f_o, maps each state to a real value which represents its energy or fitness. While downhill transitions are always accepted, uphill transitions are accepted with a probability that is inversely proportional to the energy difference between the two states. This probability is computed using Metropolis' acceptance criterion:

$$p = \mathrm{e}^{-\Delta E/K_B T} \tag{3}$$

or Felsentein's function:

$$p = \frac{1}{1 + \mathrm{e}^{-\Delta E/K_B T}} \tag{4}$$

The latter ensures that the transition probability never exceeds 0.5, and thus prohibits the system from performing random walks. Boltzmann's constant, K_B, is used for scaling purposes, and T is an artificial temperature factor that controls the ability of the system to overcome energy barriers. The temperature is systematically adjusted during the course of the simulation in a manner that gradually reduces the probability of high-energy transitions. To circumvent the problem of selecting an appropriate value for K_B, we use an adaptive approach in which K_B is not a true constant, but rather it is continuously adjusted during the course of the simulation based on a running estimate of the mean transition energy. In particular, at the end of each transition, the mean transition energy is updated, and the value of K_B is adjusted so that the acceptance probability for a mean uphill transition at the final temperature is some predefined small number (usually 0.1%). The temperature is reduced using a Gaussian cooling schedule with a half-width of 5−10 deviation units.

In the present study, the selections were carried out using a parallel implementation known as synchronous annealing. In this algorithm, each execution thread is allowed to follow its own independent Monte Carlo trajectory during each temperature cycle. The threads synchronize at the end of each cycle, and the best among the last states visited by each thread is recorded and used as the starting point for the next iteration. Given sufficient simulation time, this parallel algorithm produces results that are comparable to those obtained with the traditional serial implementation. Let n_T denote the number of temperature cycles and $\mathbf{T}$ the vector of temperatures in the cooling schedule, n_C the number of sampling steps per temperature cycle, $f_o(\cdot)$ the multiobjective fitness function, and $\overline{\Delta E}$ the average uphill transition energy (fitness). The algorithm proceeds as follows:

1. Initialize S with a random subset of k plates from P.
2. Set $f = f_o(S)$, $f_{min} = f_o(S)$, and $\overline{\Delta E} = 0$.
3. Perform steps 4−12 for each $t \leq n_T$.
4. Set $T = \mathbf{T}[t]$.
5. Perform steps 6−11 for each $p \leq n_P$, where n_P is the number of processors (execution threads).
6. Set $S_p = S$, $f_p = f$, $S^p_{min} = S$, $f^p_{min} = f$, and $\overline{\Delta E}_p = \overline{\Delta E}$.
7. Perform steps 8−11 for *each* $c \leq n_C$.
8. Mutate S by replacing a randomly chosen plate in S with a randomly chosen plate from $\bar{S}$, and denote the resulting state as S^*.
9. Set $f^* = f_o(S^*)$ and $\Delta E = |f_p - f^*|$.
10. Update ΔE_p and set $K^p_B = -\Delta E_p/(T \ln a)$, where a is a predefined small number in the interval [0, 1].
11. If $f^* \leq f_p$ or if $f^* > f_p$ and $r < \mathrm{e}^{-\Delta E/K_B pT}$, where r is a random number in the interval [0, 1], then
    11.1. Set $S_p = S^*$ and $f_p = f^*$.
    11.2. If $f_p < f^p_{min}$, set $f^p_{min} = f_p$ and $S^p_{min} = S_p$.
12. Set $f_{min} = \min_p f^p_{min}$, $S_{min} = S^q_{min}$: $S^q_{min} \leq S^p_{min}\ \forall\ p \neq q$, $f = \min_p f_p$ and $S = S_q$: $S_q \leq S_p\ \forall\ p \neq q$.
13. Output S_{min} and f_{min}.

The major advantage of this approach is that the search algorithm is completely independent of the performance measure, and can be applied on a wide variety of selection criteria and fitness functions. The choice of simulated annealing was based on its programmatic simplicity, the fact that the mutation function (or step) can be designed in a way that guarantees the creation of valid states (something that

requires extra care with genetic approaches), and in-house comparative studies, which demonstrated superior convergence compared to evolutionary schemes.

Selection Criteria. A selection criterion is a function that encodes the ability of a given set of compounds to satisfy a particular design objective. These functions are simple numerical indices that can be combined into a single objective function that measures the overall quality of a candidate design, S:

$$f_o(S) = f(f_1(S), f_2(S), ..., f_n(S)) \quad (5)$$

The objective function f_o can assume any desired functional form. Two types of selection criteria are considered in this work: maximum diversity and maximum similarity to a known lead. Alternative definitions and other selection criteria can be found in several publications by this[10,25,27,28] and other research groups.[5,6,7,9,15,17]

Similarity. The similarity of a given set of compounds, S, to a set of leads is defined as a function of the average distance of a compound to its nearest lead:

$$S(S) = \frac{1}{m}\sum_{i=1}^{m} f(\min_{j=1}^{l}(d_{ij})) \quad (6)$$

where m is the number of compounds in S (i.e., the number of compounds in all the plates comprising S), l is the number of leads, d_{ij} is the distance between the ith compound and the jth lead in some molecular descriptor space, and f is a user-defined function known as a *kernel*.[28] The default kernel is the identity function. Since typically a higher similarity score indicates a collection of compounds that are more distant and therefore less similar to the leads, focused libraries are obtained by minimizing S. Note that when similarity is the only criterion, the selection can be carried out by evaluating each plate independently and selecting the ones with the maximum score. This criterion is used here merely to demonstrate the breadth of designs that can be created with this methodology, and to illustrate some subtle points regarding the importance of layout when contemplating focused screening experiments (see below).

Diversity. The intrinsic diversity of a set of compounds, S, is defined as a function of the average nearest neighbor distance:[26]

$$D(S) = \frac{1}{m}\sum_{i}^{m} f(\min_{j\neq i}^{m}(d_{ij})) \quad (7)$$

where m is again the cardinality of S, d_{ij} is the Euclidean distance between the ith and jth compounds in S in some molecular descriptor space, and f is a user-defined kernel. The kernel is used to tailor the diversity of the design, and defaults to the identity function. Since typically the value of this function increases with spread, diverse libraries are obtained by maximizing D. Naively implemented, eq 7 requires $m(m-1)/2$ distance computations and scales adversely with the number of compounds selected. To reduce the quadratic complexity of the problem, D is computed using the k−d tree algorithm presented in ref 26. This algorithm achieves computational efficiency by first organizing all the points in S in a k-dimensional tree and then performing a nearest neighbor search for each point using a branch-and-bound approach. For a relatively small number of dimensions, this algorithm exhibits m log m time complexity and scales favorably with the number of compounds selected.

Note that the number of compounds in the plates need not be the same; in this case the selection criteria must be defined in a manner that does not favor the selection of plates that contain either too few or too many compounds, unless it is so desired.

Computational Details. All computations, including virtual library generation, descriptor calculation, principal component analysis, nonlinear mapping, and plate selection, were carried out using proprietary software written in the C++ programming language and based on 3-Dimensional Pharmaceuticals' Mt++ class library.[29] These programs are part of the DirectedDiversity[30] software suite, and were designed to run on all POSIX-compliant Unix and Windows platforms. Parallel execution on systems with multiple CPUs is supported through the multithreading classes of Mt++. The calculations were performed on a Dell workstation equipped with two 800 MHz Intel Pentium III processors running Windows 2000. Selections were carried out in 30 temperature cycles using a Gaussian cooling schedule and 1000 sampling steps per temperature cycle. Boltzmann's constant was determined in an adaptive manner so that the acceptance probability for a mean uphill transition at the final temperature was 0.1%.

III. RESULTS AND DISCUSSION

Data Set. The data set used in this study is a combinatorial library based on the reductive amination reaction. A set of 300 primary and secondary amines and 300 aldehydes were selected from the Available Chemicals Directory,[31] and were used to generate a virtual library of 90 000 products using the library enumeration classes of the DirectedDiversity toolkit.[29] Each compound in the library was characterized by an established set of 117 topological descriptors,[32,33] which included molecular connectivity indices, kappa shape indices, subgraph counts, information-theoretic indices, Bonchev-Trinajstic indices, and topological state indices. It has previously been shown that these descriptors exhibit proper "neighborhood behavior"[34] and are thus well suited for diversity analysis and similarity searching.[35−38]

These 117 molecular descriptors were subsequently normalized and decorrelated using principal component analysis (PCA). This process resulted in an orthogonal set of 23 latent variables, which accounted for 99% of the total variance in the data. To simplify the analysis and interpretation of results, this 23-dimensional data set was further reduced to two dimensions using a very fast nonlinear mapping algorithm developed by our group.[39−41] The projection (Figure 1) was carried out in such a way that the pairwise distances between points in the 23-dimensional principal component space were preserved as much as possible on the two-dimensional map, which had a Kruskal stress of 0.187. The PCA preprocessing step was necessary in order to eliminate duplication and redundancy in the data, which is typical of graph-theoretic descriptors.

Plate Assignment. To illustrate the convergence of the annealing algorithm and the impact of layout on the quality of the final selection, the compounds in the 90 000-member library were divided into 900 plates, each comprised of 100

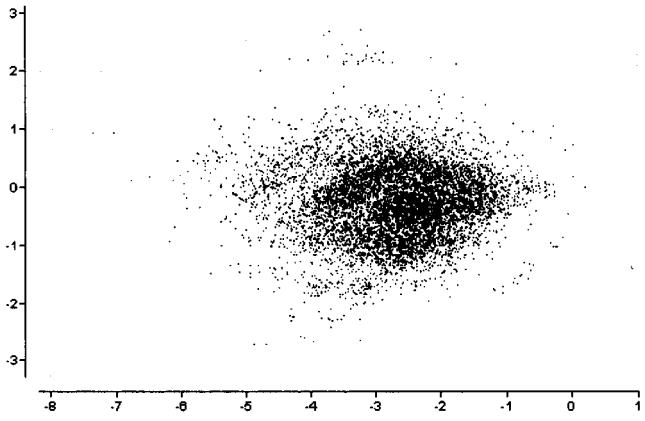

Figure 1. Two-dimensional nonlinear map of the reductive amination library. For clarity, only 10 000 randomly chosen compounds are plotted.

compounds derived from the combination of 10 R_1 and 10 R_2 inputs. (This number is chosen here for simplicity; in practice, combinatorial arrays are typically synthesized as 8 × 11 or 8 × 12 arrays.) In general, given a combinatorial library of r variation sites and a list of building blocks for each site, R_{ij}, $i = 1, ..., r$, $j = 1, ..., n_i$, the problem is to divide each list R_i into k_i groups, S_{ij}, $i = 1, 2, ..., r$; $j = 1, 2, ..., k_i$, $S_{ij} \subseteq R_i$, and combine these groups in all possible ways to assemble the plates, $P = \{S_{1i_1} \times S_{2i_2} \times ... \times S_{ri_r}, i_1 = 1, 2, ..., k_1, i_2 = 1, 2, ..., k_2, ..., i_r = 1, 2, ..., k_r\}$. The assignment of reagents to plates is carried out in a recursive manner:

```
void assign(Library& lib, Vector<int>& var, Vector<int>& beg, int i)
{
   if (i < lib.dims())
   {
      for (int j = 0; j < lib.size(i); j += var[i])
      {
         beg[i] = j;
         assign(lib, var, beg, i + 1);
      }
   }
   else
   {
      // print lib(i, j), 0 ≤ i < lib.size(), beg[i] ≤ j < beg[i] + var[i]
   }
}
```

In the preceding code written in the C++ programming language, var[i] is the number of inputs at the *i*th substitution site that are varied in each plate, beg[i] is the first input at the *i*th variation site in the current plate, and lib is a combinatorial library object supporting (among others) the methods dims( ) which returns the number of variation sites in the library, size(i) which returns the number of inputs at the *i*th site, and operator(i, j) which returns the *j*th input at the *i*th site. In practice, it is extremely rare to vary more than two *R*-sites within the same plate, which means that all but two of the k_i's (or var[i] in the code segment shown above) are equal to 1.

Similarity Sorting. The design problem is thus reduced to sorting the reagent lists R_{ij} so that the resulting plates satisfy some user-defined criteria. Three different methods are examined in the present work: random assignment, and assignment based on maximum internal similarity and maximum internal diversity. The objective of the latter is to construct plates that are either internally homogeneous (i.e., comprised of relatively similar compounds) or internally heterogeneous (i.e., comprised of dissimilar or diverse compounds). An effective strategy to achieve this goal is to sort the reagent lists according to the molecular similarity of their respective reagents. The idea is to place similar reagents close to each other in the list and separate the ones that are more structurally unrelated. This can be easily accomplished by computing the reagent similarity matrix and then embedding the building blocks into a one-dimensional space using multidimensional scaling.[42,43] In this study, the projection was carried out using a variant of Sammon's nonlinear mapping algorithm.[44] Given a set of n objects, a symmetric matrix d_{ij} of their observed dissimilarities, and a set of images on an k-dimensional display plane $\{\boldsymbol{x}_i, i = 1, 2, ..., n; \boldsymbol{x}_i \in R^k\}$, the objective is to place $\boldsymbol{x}_i$ onto the plane in such a way that their Euclidean distances $\delta_{ij} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||$ approximate as closely as possible the corresponding values d_{ij}. The embedding is carried out in an iterative fashion by minimizing an error function, E, which measures the difference between the original and projected distance matrices:

$$E = \frac{\sum_{i<j}^{k} \frac{[d_{ij} - \delta_{ij}]^2}{d_{ij}}}{\sum_{i<j}^{k} d_{ij}} \tag{8}$$

E is minimized using a steepest-descent algorithm. The initial coordinates $\boldsymbol{x}_i$ are determined at random or by some other projection technique such as PCA, and are updated using

$$x_{ij}(m+1) = x_{ij}(m) - \lambda \Delta_{ij}(m) \tag{9}$$

where m is the iteration number and λ is the learning rate parameter, and

$$\Delta_{ij}(m) = \frac{\partial E(m)}{\partial x_{ij}(m)} \Big/ \left| \frac{\partial^2 E(m)}{\partial x_{ij}(m)^2} \right| \tag{10}$$

When $k = 1$, this procedure places the objects along a projection axis in a way that preserves their pairwise relationships as closely as possible and, in effect, produces a list sorted according to the objects' similarity. This algorithm will hereafter be referred to as S-SORT.

The one-dimensional nonlinear map of the amine building blocks is illustrated in Figure 2. It is evident that the reagents are sorted in a way that is consistent with our general notion of molecular distance, which is determined mainly by molecular weight, ring character, heteroatom content, and degree of branching. Note that in general the pairwise similarities cannot always be accurately reproduced in a low-dimensional manifold, resulting in a projection that may exhibit a significant amount of distortion. In fact, in the case at hand, the Kruskal stress of the one-dimensional projection for the amines and aldehydes was 0.317 and 0.327, which is equivalent to a Pearson correlation coefficient of 0.83 and 0.81, respectively. This distortion is graphically illustrated in Figure 3, which also reveals that the mapping algorithm appears to underestimate short distances and overestimate large ones. This, however, is not a serious problem since in this case we are only interested in the relative rank-ordering of the reagents as opposed to their absolute positions.

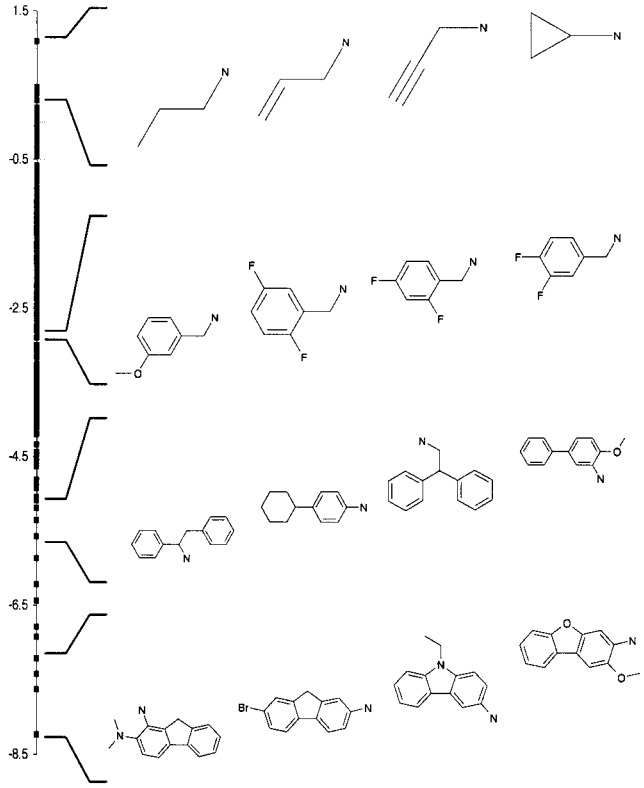

Figure 2. One-dimensional nonlinear map of the 300 amines that were used as input for the construction of the reductive amination library.

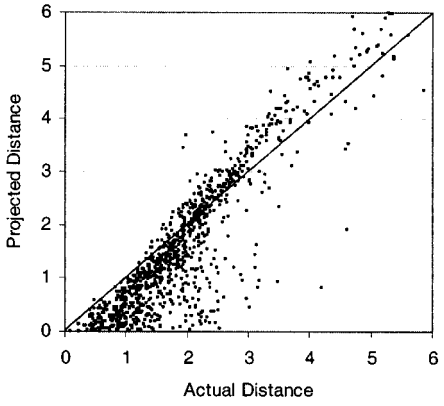

Figure 3. Preservation of molecular similarities on the one-dimensional nonlinear map for the amine building blocks.

Nevertheless, the plot does reveal several pairs whose projected distance is much shorter than their actual similarity, which results in the presence of reagents that appear to be "out of place" with respect to their neighbors. Unfortunately, this is an inevitable consequence of the high intrinsic dimensionality of the descriptor space used to define molecular similarity, and cannot be avoided no matter what projection or sorting technique is employed.

With a simple modification, this algorithm can also be used to partition the reagents into subsets that are internally heterogeneous, i.e., diverse. This is achieved by grouping the sorted reagents into disjoint subsets using the membership rule:

$$S_{ij} = \{R_{ik}, k = j, j + k_i, j + 2k_i, ...\} \qquad (11)$$

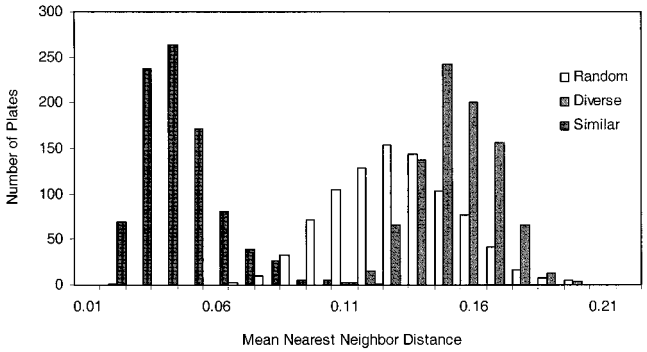

Figure 4. Histogram of diversity scores (mean nearest neighbor distances) of the 900 plates constructed with the S-SORT, D-SORT, and random shuffling algorithms.

where R_i is assumed to be presorted by S-SORT. Unlike conventional dissimilarity selection algorithms such as maxmin,[45] this approach (which will hereafter be referred to as D-SORT) has the advantage that it produces subsets of comparable diversity. Since the method is primarily aimed at creating plates for general and recurrent screening, other target-specific criteria based on, e.g., QSAR predictions or docking scores are of limited value.

When combined with the recursive plate assignment algorithm described above, reagent sorting can be very effective in controlling the diversity of the output plates. This is nicely illustrated in Figure 4, which shows the distribution of diversity scores (mean nearest neighbor distance) of the plate sets derived with S-SORT, D-SORT, and random assignment. The diversity scores of the plates produced by S-SORT exhibit a sharp distribution with a mean of 0.038 and a standard deviation of 0.015, almost 0.1 unit below the mean of the respective distribution obtained by random shuffling (0.130 ± 0.023). In contrast, the distribution of diversity scores for the plates produced by D-SORT is shifted in the opposite direction, albeit to a somewhat lesser extent (0.150 ± 0.015). In both cases, the distributions are relatively narrow compared to random, which indicates that the two algorithms are very effective in maintaining a consistent spread across the entire plate set.

Selections. Two types of selection criteria are explored in this work: (1) molecular diversity and (2) similarity to a known query (lead). To simplify the presentation of results, selections were limited to 10 plates (i.e., 1000 compounds), which amounts to a state space of 10^{23} distinct solutions. Preliminary studies have shown that the trends outlined below are true regardless of the number of plates requested, as long as that number is substantially smaller than the total number of plates available to choose from.

The analysis was based on the three different plate assignment algorithms described in the previous section. The compounds comprising the 10 most diverse plates from each of these sets as identified by the annealing algorithm are illustrated in Figure 5. The diversity scores (i.e., the mean nearest neighbor distance) of the selected plates from the random, D-SORT, and S-SORT plate sets were 0.074, 0.070, and 0.071, respectively, which indicates that the solutions are of comparable quality. However, the distribution of compounds on the nonlinear map reveals a subtle difference between the S-SORT and random/D-SORT selections in that the latter appear to be somewhat more homogeneous. In contrast, the S-SORT selection covers a wider area of

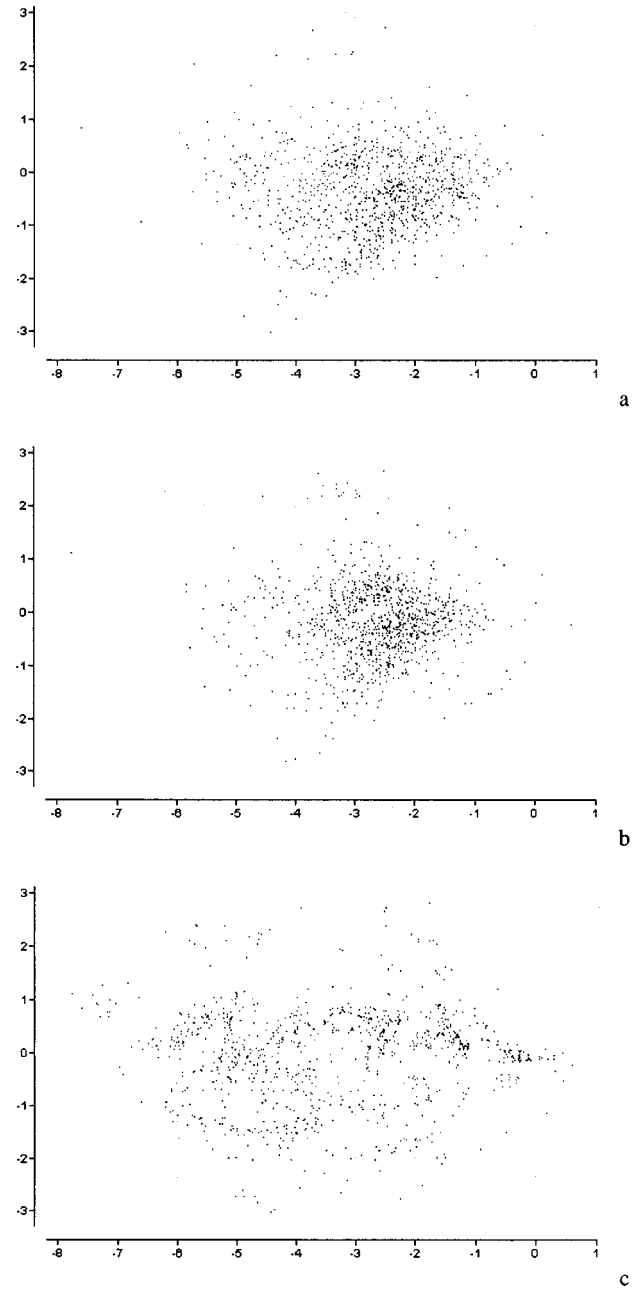

Figure 5. Selection of the 10 most diverse plates from the (a) random, (b) D-SORT, and (c) S-SORT plate sets.

chemical space but does show signs of local clustering that are consistent with the way that these plates were constructed. Despite their differences, for all practical purposes, these designs are equivalent and sample chemical space in an adequate manner. However, we should caution the reader that the diversity metric employed in this study is not very well suited for locally clustered data. Indeed, the metric is determined primarily by intracluster distances, and is not effective in discriminating the relative separation between the clusters. In such cases, use of alternative methods such as minimum spanning trees,[46] grid-based approaches,[18] and probabilistic sampling[47] may be more appropriate, but a detailed comparison of these approaches is beyond the scope of this paper.

These results suggest two different ways for exploiting molecular diversity: using diverse sets of similar compounds and using similar sets of diverse compounds. The former

J. Chem. Info. Comput. Sci., in press.

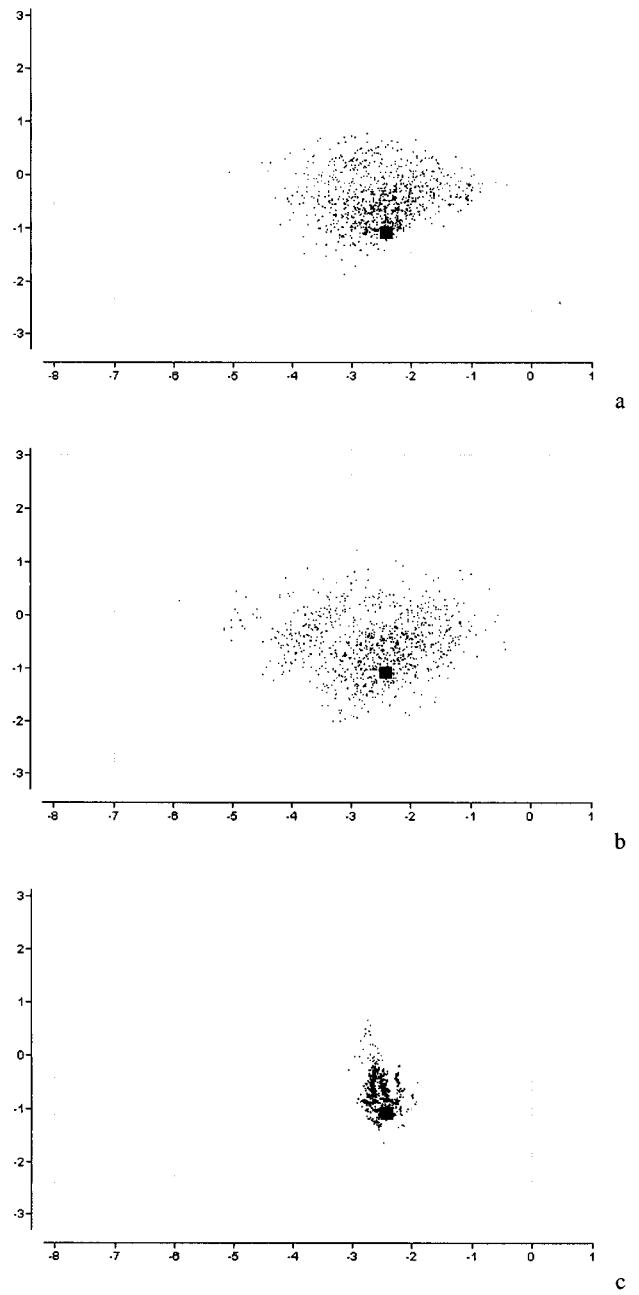

Figure 6. Selection of 10 plates that are most similar to the highlighted query from the (a) random, (b) D-SORT, and (c) S-SORT plate sets.

strategy presents an additional advantage that is related to the errors inherent in high-throughput screening, and could have important consequences during hit selection. High-throughput screening errors may be related to mechanical, procedural, temporal, or chemical factors, and are very difficult to trace without retesting and, often, resynthesis of the compounds that test positive in the high-throughput screen. The latter is a costly and time-consuming process, and the decision as to what constitutes a "true" hit is often based upon the presence of structurally related compounds that show activity in the same assay. This follows directly from the "similar property principle", which is the assumption that structurally similar compounds tend to have similar

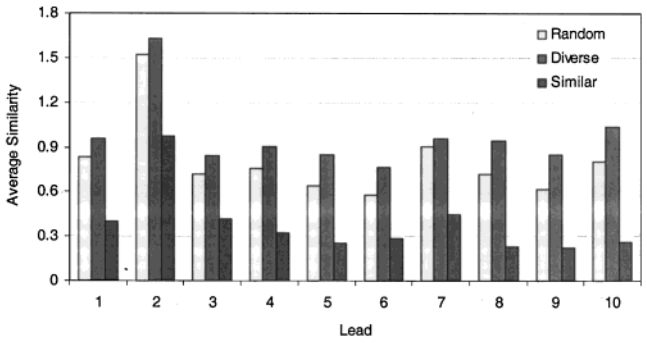

Figure 7. Average similarity scores of the 10 most similar plates from the random, D-SORT, and S-SORT plate sets for 10 randomly chosen queries.

biological properties. Thus, screening closely related compounds in the same plate can potentially minimize the impact of random errors and provide valuable information during hit validation.

The difference among the three layout strategies becomes much more pronounced in the context of focused screening. Figure 6 shows the compounds comprising the 10 plates from each of the three plate sets (S-SORT, D-SORT, and random) that are most similar to a randomly chosen query from the same library. The average similarity scores of the selected compounds were 0.893, 0.994, and 0.481 for the random, D-SORT, and S-SORT plates, respectively, which manifest a profound advantage of the S-SORT algorithm over the other alternatives. As the nonlinear maps in Figure 6a,b indicate, plates containing random or diverse arrays do not lend themselves for similarity searching, and lead to highly subotpimal experiments that involve the screening of a large number of compounds that are structurally removed from their target. As shown in Figure 7, this trend is general and independent of the structure used as a query.

V. CONCLUSIONS

Historically, computational methods for library design have focused almost exclusively on the selection of building blocks prior to synthesis, using molecular diversity, structure−activity correlation, and structure-based design as a means to prioritize experiments. However, little attention has been paid to how these experiments are organized and executed, and how these factors affect the outcome of subsequent high-throughput screening operations. In this paper, we presented two very efficient algorithms for controlling the diversity of plates constructed with parallel synthesis techniques, and a multiobjective approach for selecting a subset of plates from a large corporate collection for screening against a biological target. Although these algorithms are both novel and relevant, this paper is best seen as an attempt to raise awareness and trigger further research into this important but overlooked aspect of library design.

ACKNOWLEDGMENT

The authors thank Drs. Victor S. Lobanov, Sergei Izrailev, and Edward P. Jaeger of 3-Dimensional Pharmaceuticals, Inc., for many useful suggestions, and Dr. Raymond F. Salemme for his insightful comments and support of this work.

REFERENCES AND NOTES

(1) Thompson, L. A.; Ellman, J. A. *Chem. Rev.* **1996**, *96*, 555−600.
(2) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
(3) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. United States Patents 5,463,564, 1995; 5,574,656, 1996; 5,684,711, 1997; and 5,901,069, 1999.
(4) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(5) Sheridan, R. P.; Kearsley, S. K. Using a genetic algorithm to suggest combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.
(6) Weber, L.; Wallbaum, S.; Broger, C.; Gubernator, K. Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2280−2282.
(7) Singh, J.; Ator, M. A.; Jaeger, E. P.; Allen, M. P.; Whipple, D. A.; Soloweij, J. E.; Chowdhary, S.; Treasurywala, A. M. Application of genetic algorithms to combinatorial synthesis: a computational approach for lead identification and lead optimization. *J. Am. Chem. Soc*. **1996**, *118*, 1669−1676.
(8) Taylor, R. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.
(9) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity* **1996**, *2*, 64−74.
(10) Agrafiotis, D. K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 841−851.
(11) Chapman, D. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 501−512.
(12) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.
(13) Pickett, S.; Mason, J. S.; McLay, I. M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.
(14) Shemetulsksis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871.
(15) Brown, R. D.; Martin, Y. C. Designing combinatorial library mixtures using genetic algorithms. *J. Med. Chem*. **1997**, *40*, 2304−2313.
(16) Good, A. C.; Lewis, R. A. New methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARPcik. *J. Med. Chem*. **1997**, *40*, 3926.
(17) Gillet, V. J.; Willet, P.; Bradshaw, J.; Green, D. V. S. Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169−177.
(18) Pearlman, R. S.; Smith, R. S. *Perspectives Drug Discovery Des*. **1998**, *9*, 339−353.
(19) Agrafiotis, D. K.; Lobanov, V. S. Ultrafast algorithm for designing focused combinatorial arrays. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1030−1038.
(20) Stanton, R. V.; et al. Combinatorial library design: maximizing model fitting compounds with matrix synthesis constraints. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 701−705.
(21) Martin, E. J.; Spellmeyer, D. C.; Critchlow, R. E.; Blaney, J. M. Does combinatorial chemistry obviate computer-aided drug design? In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1997; Vol. 10, pp 75−100.
(22) Agrafiotis, D. K. The diversity of chemical libraries. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; pp 742−761.
(23) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in diversity profiling and combinatorial series design. *Mol. Diversity* **1999**, *4* (1), 1−22.
(24) Agrafiotis, D. K.; Lobanov, V. S.; Rassokhin, D. N.; Izrailev, S. The measurement of molecular diversity. In *Virtual screening of bioactive molecules*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000; pp 265−300.
(25) Agrafiotis, D. K. On the use of information theory for assessing molecular diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (3), 576−580.
(26) Agrafiotis, D. K.; Lobanov, V. S. An efficient implementatin of distance-based diversity metrics based on k−d trees. *J. Chem. Inf. Comput. Sci*. **1999**, *39* (1), 51−58.
(27) Rassokhin, D. N.; Agrafiotis, D. K. Kolmogorov-Smirnov statistic and its applications in library design. *J. Mol. Graphics Model.* **2000**, *18* (4−5), 370−384.
(28) Agrafiotis, D. K. Multiobjective optimization of combinatorial libraries, *IBM J. Res. Dev*., in press.
(29) Copyright 3-Dimensional Pharmaceuticals, Inc., 1994−2001.

(30) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. United States Patents 5,463,564, 1995; 5,574,656, 1996; 5,684,711, 1997; and 5,901,069, 1999.
(31) Marketed by MDL Information Systems, Inc., 140 Catalina Street, San Leandro, CA 94577.
(32) Hall L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Relations. In *Reviews of Computational Chemistry*; Boyd, D. B., Lipkowitz, K. B., Eds.; VCH Publishers: New York, 1991; Chapter 9, pp 367−422.
(33) Bonchev, D.; Trinajstic, N. *J. Chem. Phys*. **1977**, *67*, 4517−4533.
(34) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(35) Downs, G. M.; Willett, P. Similarity searching and clustering of chemical-structure databases using molecular property data. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1094−1102.
(36) Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(37) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1−9.
(38) Lobanov, V. S.; Agrafiotis, D. K. Stochastic similarity selections from large combinatorial libraries. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 460−470.
(39) Agrafiotis, D. K.; Lobanov, V. S. Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1356−1362.
(40) Rassokhin, D. N.; Lobanov, V. S.; Agrafiotis, D. K. Nonlinear mapping of massive data sets by fuzzy clustering and neural networks. *J. Comput. Chem.* **2001**, *22*(4), 373−386.
(41) Agrafiotis, D. K.; Rassokhin, D. N.; Lobanov, V. S. Multidimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* **2001**, *22*(5), 488−500.
(42) Kruskal, J. B. *Psychometrika* **1964**, *29*, 115−129.
(43) Torgeson, W. S. *Psychometrika* **1952**, *17*, 401−419.
(44) Sammon, J. W. *IEEE Trans. Comput.* **1969**, *C-18*, 401−409.
(45) Lajiness, M. S. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991; pp 201−204.
(46) Mount, J.; Ruppert, J.; Welsh, W.; Jain, A. N. *J. Med. Chem.* **1999**, *42*, 60−66.
(47) Agrafiotis, D. K. A constant time algorithm for assessing the diversity of large chemical libraries. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 159−167.

CI000313D