Design of Diverse and Focused Combinatorial Libraries Using an Alternating Algorithm

S. Stanley Young*

National Institute of Statistical Sciences, RTP, North Carolina 27709

Marcia Wang

Department of Statistics and Actuarial Science, 200 University Avenue W., University of Waterloo,
Waterloo, Ontario, Canada, N2L 3G1

Fei Gu

Infinity Pharmaceuticals Inc., 780 Memorial Drive, Cambridge, Massachusetts 02139

Received June 20, 2003

There is considerable research in chemistry to develop reaction conditions so that any of a very large number of reactants will successfully form new compounds, e.g. for two components, $A_i + B_j \Rightarrow A{-}B_{ij}$. The numbers of A's and B's usually make it impossible to make all the possible products; with multicomponent reactions, there could easily be millions to billions of possible products. There is a need to identify subsets of reagents so that the resulting products have desirable predicted properties. Our idea is to select reactants sequentially and iteratively to optimize the evolving candidate library. The new Alternating Algorithm, AA, can be used for diversity, a space-filling design, or for a focused design, using either a near neighborhood or structure−activity relationship, SAR. A diversity design seeks to select compounds different from one another; a focused design seeks to find compounds similar to an active compound or compounds that follow a structure activity relationship. The benefit of the method is rapid computation of diversity or focused combinatorial chemical libraries.

INTRODUCTION

The key feature of combinatorial synthesis is that the reactions are typically orthogonal. Suppose that the reaction is $A_i + B_j \rightarrow A- B_{ij}$. If 10 A's are selected, they are reacted with each B. Each selected B is reacted with each selected A. So if there are 12 B's, then the design is $10 \times 12 = 120$ products. Conducting reactions in an orthogonal manner makes for easier logistics. There is also some advantage in having the design roughly "square" as the number of products is a maximum for a fixed number of building blocks. For example, with 20 building blocks, 10 × 10 gives 100 products, whereas 15 × 5 gives only 75 products. Cost comes in so that if A's are relatively more expensive than B's, then a more "rectangular" design is the way to go.

For a given reaction or sequence of reactions, the number of possible products can be astronomical. For example, Tan et al.[1] executed a multistep reaction starting with 18 tetracyclic scaffolds and combined these with 30 alkynes, 62 primary amines, and 62 carboxylic acids giving a total of 2.18 million distinct final products. For the simple amide bond forming reaction, there are thousands of amines, A_i-NH_2, and carboxylic acids, HOOC-B_j, that can be purchased. We arbitrarily selected 87 amines and 127 carboxylic acids giving a total of 11 049 products to illustrate our procedure. So in this situation, like most, it is impossible to make all possible products. Even with a two-component reaction with a modest number of building blocks, it generally does not make economic sense to build all the products, as there may be more than can be afforded and many may be largely redundant. A "diversity" sublibrary can be designed that covers the chemical space of the entire library. If a target has been selected and there is something known about active compounds against that target, then one would want to densely cover specific regions of chemical space with a "focused" library. One can think of the diversity library as a representative sample of the full "virtual" library.

How do diversity and focused libraries fit into the drug discovery process? Consider Figure 1. First, the number of compounds, real and virtual, available for screening is enormous. Much as we would like to, it just is not possible to "screen everything". The inner panel lays out the process from a statistical point of view. A set of compounds is screened (Diversity Library coming in from the left), and the screening results are subjected to statistical analysis. Historically, screeners would sort the compounds by potency and pass on to chemists the most active compounds. Now scientists are also looking for active compounds with consistent patterns of molecular features; computational chemists and statisticians use statistical methods combined with numerical features for pattern recognition. This effort gives rise to a statistical model of important molecular features, the upper right box in the inner panel. The structure−activity rules can be used to design a focused library, lower right box in the inner panel. These compounds are synthesized and screened. The screening data set is added to the initial data to build a better statistical model of important features.

Turn now to the boxes outside the inner panel. At the top of Figure 1 is a box indicating building blocks. In our example, the building blocks are amines, A_i-NH2, and

* Corresponding author e-mail: young@niss.org.

10.1021/ci034125+ CCC: $25.00
Published on Web 10/14/2003

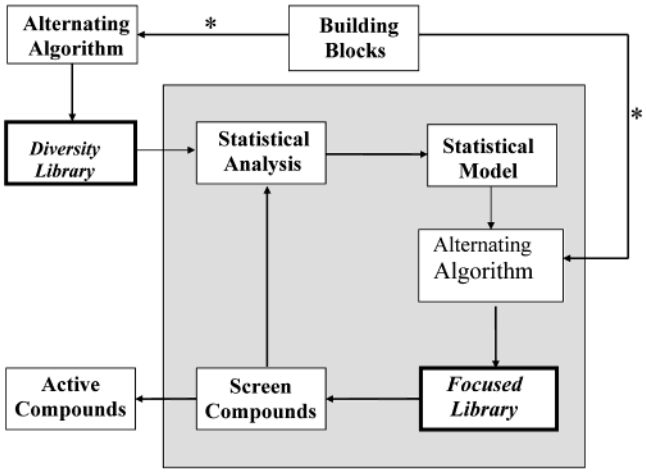

Figure 1. Diagram of sequential screening. First, building blocks are selected using the Alternating Algorithm to build a Diversity Library. This library is screened and a Statistical Model is built. The model can be used by the Alternating Algorithm to design a Focused Library. This library is screened and a new analysis is computed. The process iterates. Filters can be applied at the points noted with ∗.

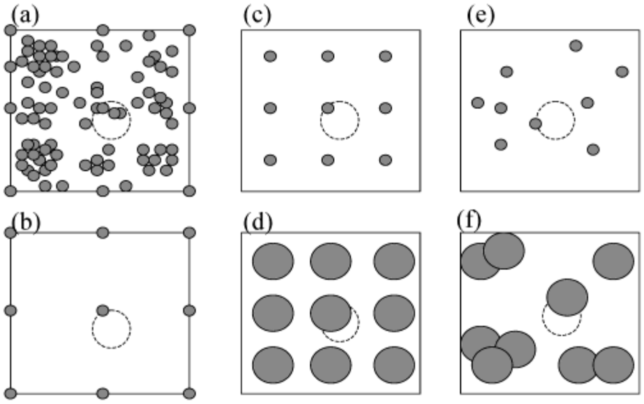

Figure 2. Examples of compounds covering a 2-D descriptor space. A solid circle represents the coverage of a compound. The dotted circle represents an active region. (a) A collection of compounds. (b) Nine compounds that “spread” away from one another. (c) Nine compounds that “cover” the space. (d) If the compounds cover more space, uniform spacing is useful. (e) If the compounds cover relatively little space, random selection leads to overlapping coverage.

carboxylic acids, HOOC-B_j. In practice, more complicated reactions or sequences of reactions and other chemical building blocks would be used. The subject of this paper is an algorithm to select building blocks so that the resulting compounds have good properties. This Alternating Algorithm, AA, described later, can be used to select building blocks so the resulting compounds are structurally diverse, see the box called Diverse Library on the left of Figure 1, and the resulting library is orthogonal. When little is known of the biological target, it is generally believed that starting with a diverse set of compounds offers the best chance of finding active compounds. Diversity sets of compounds are also useful to identify both the protein targets and their small molecule regulators.[2] It is the contrast of the features of active and inactive compounds that allows statistical analysis to proceed.

Once a statistical model of the structure−activity relationship is obtained, it can be fed into the AA to help selecting building blocks. The compounds in the resulting Focused Library satisfy the SAR and hence give an improved chance of increased activity. So, from the top, building blocks and a reaction are selected, and the AA is used to select building blocks to create a Diversity Library. This library is screened, and the data are statistically analyzed to give a model. The statistical model can be fed into the alternating algorithm to guide the selection of building blocks for a Focused Library. These compounds are synthesized and screened, which provides data that should improve the initial statistical model. Eventually, the process either fails or leads to active compounds being identified, Active Compounds box, on the lower left side of Figure 1.

Parenthetically, we say that if there is information on the target, then using that information to make an initial focused library makes sense; we will discuss focused libraries later.

There are a number of recent reviews[3−7] of combinatorial library design, so we will not review the literature in this paper.

Space Filling and Diversity. It is useful to briefly discuss chemical diversity. Consider Figure 2. Solid circles in panel (a) represent compounds that are described by two numbers. The diameter of the small circles is used to indicate the coverage of each compound. If there is a region where biological activity occurs, represented here as a dashed circle, then to find it you need compounds whose diameter overlaps the region of biological activity. A general search strategy used historically has been to screen as many compounds as possible, the screen-everything strategy. This strategy can be inefficient, as compound collections tend to have dense regions where synthetic activity has been great; chemists synthesize compounds around known active compounds. It is not so likely that a new target will center in a region of a previous target. Redundant effort will likely be misplaced by screening everything. What is needed is some sort of even spacing of compounds. Figure 2, panels (b) and (c), shows spread and coverage spacing, respectively. In a spread design, the goal is to place selected compounds as far apart from each other as possible, whereas in a coverage design, the goal is to have each nonselected point as near a design point as possible. Spread designs are less computer-intensive as you need to keep track of only the design points; coverage designs need to compute the distance from every candidate point to the nearest design point, which can be very computer intensive. The terms spread and coverage, used in the statistics literature, correspond to diverse and representative in the computational chemistry literature.[8,9] Panels (c) and (e) show that if the design points do not cover much of the space, then random designs should be as good as carefully constructed designs. Panels (d) and (f) point to the fact that if the number of points is dense or the coverage of each point is great, then careful spacing should improve the design.

The good qualities of a coverage design can be approximated by cell-based designs, Figure 3. A point in a cell is taken to cover the other points in the cell, and a massive number of interpoint distances do not need to be computed. The space is divided into small cells, and a fixed number of points are selected from each cell.[10] There are problems with the original formulation of cell-based designs. Each dimension needs to be divided into a large number of divisions so with even a modest number of dimensions the number of cells can be excessive. For 6 dimensions and 100 divisions,

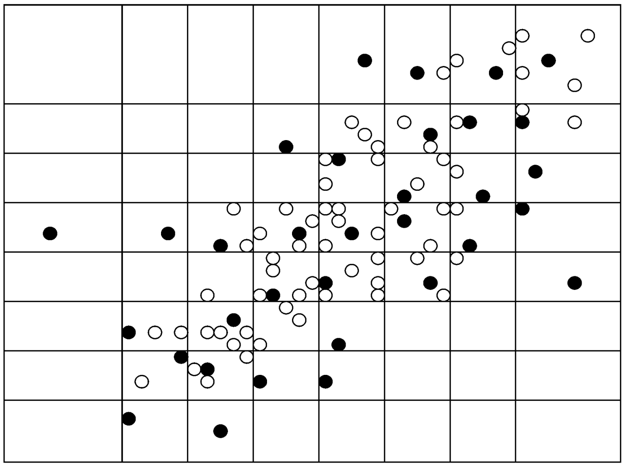

Figure 3. Example of dividing a 2D chemical space; see LWY. In the compound dense regions, bins can be of uniform width. On the edges of the space, where compound density can be low, it is useful to make the bins wider. The filled circles are for selected compounds. The Alternating Algorithm attempts to find one compound for each occupied cell in each subspace.

you would have 10^{12} cells. Lam, Welch, and Young,[11] hereafter LWY, proposed, among other things, covering only the low dimension subspaces. They also proposed that the number of cells remain constant over the subspaces. So if we divide the 1D space into 729 bins, we divide the 2D spaces into $27 \times 27 = 729$ bins and the 3D spaces into $9 \times 9 \times 9 = 729$ cells. They used an exchange algorithm to optimize the coverage of a design set of compounds selected from a candidate set. They used their uniform coverage criterion, UCC, which is the sum of the square deviations of the number of compounds in a cell from the target number of compounds in the cell, usually one. The observation is that it is essentially impossible to cover densely a high-dimensional space. For example, 6D with 100 bins leads to 100^6 or 10^{12} cells. Covering this many cells is too expensive, so LWY decided to cover 1D, 2D, and 3D subspaces densely.

Data Set and Descriptors. We choose a simple example to demonstrate our method, the amide bond formation from 87 amines and 127 carboxylic acids. Our goal is to select 20 of each such that the resulting products are uniformly spaced in the chemical space as computed by six BCUTs, described next.

Burden[12] proposed a graph-theoretic numerical measure of a molecule as a potential identification number; in computer searches of large chemical databases, it would be very useful for exact match searches to have a unique number for each compound. Researchers at Chemical Abstract Services noted that two compounds with very similar Burden numbers often were structurally similar. Pearlman and Smith[13] generalized the Burden number. These numbers can be used for the prediction of molecular properties[13] and the activity of compounds.[14] The general strategy for making a Burden number is to construct a square matrix placing some atom property on the diagonal (hydrogens are usually suppressed) and some measure of connectivity on the off-diagonal elements. Eigen values of the resulting matrix become the Burden numbers, called BCUTs by Pearlman and Smith.[15]

So the question is as follows. Products live in a 6D BCUT space, but orthogonal designs require selection of reagents. How do we get optimal or near optimal coverage in the 6D space, within the restriction of orthogonal synthesis? Or in the case of a focused design, how do we select compounds similar to an active compound or compounds that are expected to follow a SAR within the restriction of orthogonal synthesis?

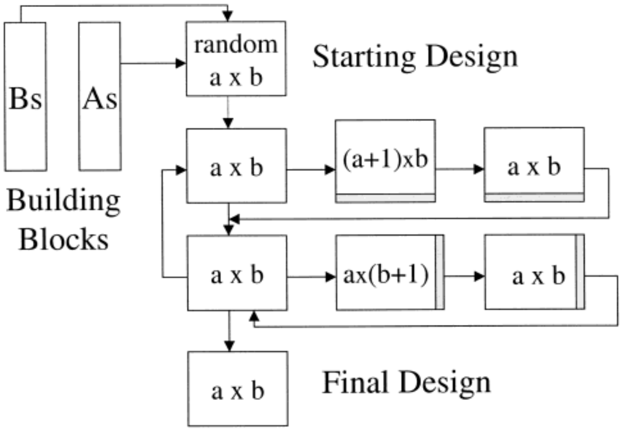

Figure 4. Alternating algorithm for a two-component chemical library. Select a As and b Bs at random. Add an A component to the library that improves the design and then remove the A that hurts the coverage the least. Now add and remove a B component. Iterate until the design does not improve.

Alternating Algorithm. A flow diagram of the Alternating Algorithm is given in Figure 4 for a two-component library. There are inventories of building blocks, here A's and B's. For a fixed size library, a A's and b B's, the starting A's and B's can be selected at random. If there are a large number of A's and B's, then the initial building blocks can be selected using some clustering or filtering method to reduce the number of reagents under consideration. Subsets of specific A's and/or B's can be fixed in the design, and the algorithm will optimize the remaining reagents. Either with random selection or some method of diversity selection the result is an a × b Starting Design. Next we focus on one of the reagents, say the A's, and move through the A's not in the current design seeing which A improves the performance of the current design. We can examine all the A reagents and select the best, or we can add a new A reagent when we find one that is sufficiently good. See LWY for the use of this strategy in a similar problem. The performance can be diversity, e.g. uniform coverage criterion, UCC, or predicted potency or some index that combines various aspects of design quality. The number of A reagents increases from a to a+1. Next the a+1 A reagents are examined, and one A reagent is selected to be removed such that it will decrease the performance of the current design as little as possible. The removal of the A reagent brings the design back to size a × b. The next type reagent is treated the same way in turn. The number of B's goes from b+1 and then back to b. Any number of reagent types can be treated. Note that we could expand the design to a+1 and b+1 and then remove an A and a B to get back to the desired a × b design size. After the final reagent has undergone an exchange of reagents, the process is started again unless the performance of the design fails to improve. When there is no more improvement or the improvement becomes smaller than a threshold, the process stops and the design is considered the

Final Design. The process can be stopped after a fixed number of iterations.

The process is stochastic, and it is not expected that one running of the algorithm will lead to “the optimal design”. Several random starts are used, and the selected design is the Final Design that has the best characteristics.

For some library design situations there may be a very large number of reagents. For example there are thousands of carboxylic acids and amines available commercially. With a very large number of reagents, it is computationally advantageous to reduce the number before executing the Alternating Algorithm. A large number of reagents can be reduced using a filter and/or clustering. If the criterion to be optimized is diversity, then one could cluster the reagents and select one compound from each cluster, or one could use a space-filling algorithm, LWY, to reduce the number of reagents. Alternatively, the criterion to be optimized can be predicted potency. Reagent reduction is more difficult in this situation. Suppose there is a known active drug. It can be divided into parts and the reagents reduced by eliminating any reagents not similar to parts of the target compound.[3] Suppose there is a known SAR, then, to the extent possible, reagents incompatible with the SAR can be eliminated. An approximate SAR can be determined using recursive partitioning.[14]

RESULTS

We applied the Alternating Algorithm to the selection of 20 A’s and 20 B’s from 87 amines and 127 carboxylic acids. The reagents were numerically characterized with six BCUTs. Here we are seeking a diversity library so we optimize UCC and also track the percent of cells covered. Over all subspaces, the percent of cells covered can be taken to the number of selected cells divided by the number of cells that are occupied by the candidates, times 100. We give several benchmarks to evaluate the performance of the algorithm: random selection of A’s and B’s; diversity selection of A’s and B’s; and finally we select 400 compounds from the product space without the 20 × 20 orthogonal restriction. For this example problem, the algorithm converged in 25 iterations. UCC decreased from ∼2500 to ∼1200. The percent of the 1D and 2D cells covered increases from ∼50% to about 70%. Figure 5 gives Box-and-Whisker glyphs for six of the 21 subspaces.

The two glyphs in each subfigure are for the random designs versus designs constructed using the AA. The number of cells of the 512 available that are occupied by at least one candidate point is given at the top of each figure. Box-and-Whisker glyphs are used to display the distribution of 20 designs made using each method. Looking at Figure 5, it is clear from the six selected subfigures that the ranges for the random designs and the AA designs largely do not overlap and the AA designs are clearly superior in number of cells covered. Figure 6 gives the distributions of UCC for three situations: Random designs are clearly inferior to designs selected using AA; finally, if the orthogonal restriction is removed, there is additional improvement in the number of cells covered.

DISCUSSION

Library design is complex, and there are many things to consider. If the number of each type of reagent is large, then the resulting number of candidate compounds can be enormous. One way to speed up the alternating algorithm for a diversity library is to select a smaller, diverse candidate set for each reagent and then use the alternating algorithm to make the final selections. A diversity selection on each reagent would obviously be much faster than a search over all or the restricted list of products. Use of the alternating algorithm on the selected candidates should allow special, nonadditive effects of products to come into play.

It is easy to generalize the alternating algorithm to more than two reagents. Each reagent is treated in turn. The reagents can be treated in a specified order, or the order can be randomized within each cycle. The use of several random starts for the entire process and the random reordering of

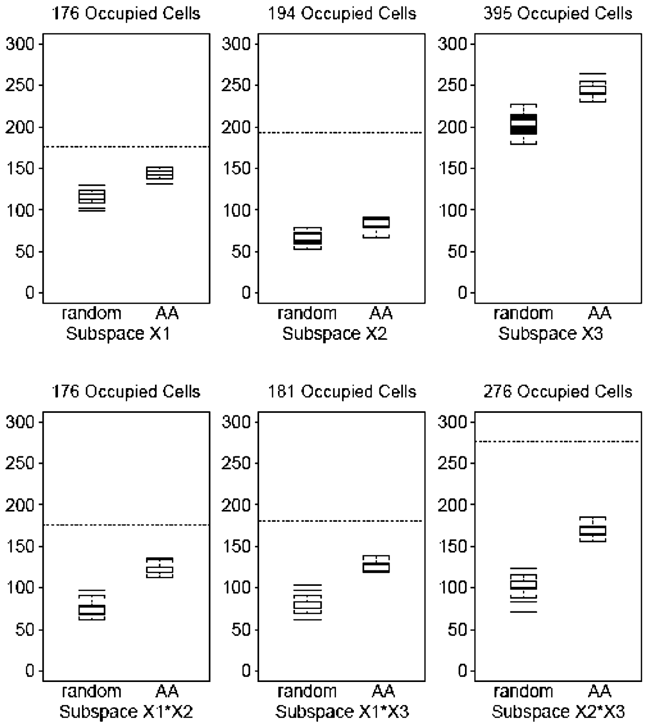

Figure 5. Twenty designs were selected using the Alternative Algorithm. The distribution of cell occupancy for random and AA designs is given for six of the 21 subspaces.

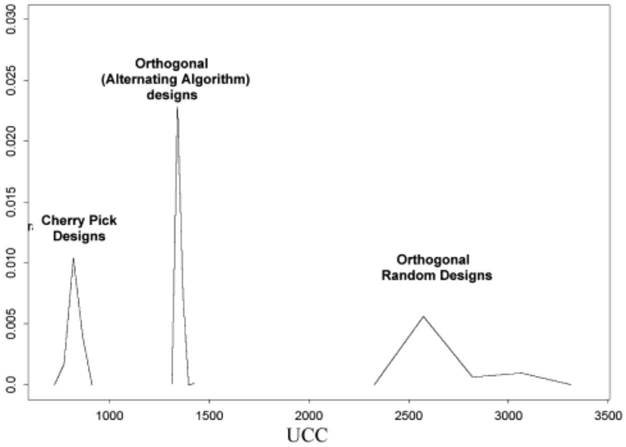

Figure 6. Distribution of UCC for 20 designs from three methods. For UCC, smaller is better. The Orthogonal Random Designs were constructed by randomly selecting building blocks. The Orthogonal Designs were selected by the Alternating Algorithm. The Cherry Pick Designs were selected from the entire candidate set without requiring orthogonal building blocks.

the reagents within a cycle allow the exploration of the variability over starting designs and reagent orderings. For this reaction and these reagents, this final design variability is much smaller than the variability of starting designs, Figure 6, and this is to be expected in general.

The experimenter needs to select the number and type of molecular descriptors. An enormous number of molecular descriptors have been developed. The Dragon software[16] has over 1400 descriptors. Our diversity calculators require a relatively small number of descriptors, 10 or fewer. The number of 1D, 2D, and 3D subspaces increases dramatically with the number of descriptors, $s = (k/6)(5+k^2)$, the sum of the numbers of n things taken one, two, and three at a time; when $k = 6$ there are 41 subspaces and when $k = 10$ there are 175, LWY. Various strategies can be used for variable reduction/selection, ranging from expert opinion to principal components. Our purpose is to present a new algorithm, not benchmark descriptors, so we selected variables, BCUTs, that are small in number and have proven useful in practice.[16,17] Other variable sets, e.g. logP, Surface Area, Polar Surface Area, Hydrogen Bond Donors, Hydrogen Bond Acceptors, # Rotatable Bonds, should also be effective for diversity calculations as they have been useful for QSAR studies.[18] It is not unexpected that different variable sets would be effective as there are high correlations among chemical descriptors. Variable selection for focused sets is likely to be more important. Fortunately, when much is known about a target and active compounds for that target, this information can be used for variable selection. One strategy would be the use of forward stepwise regression; another would be the use of multiple-tree recursive partitioning, selecting those variables that are used in many trees.[19]

How do we benchmark the diversity of a diversity library design? In our case we use uniform coverage criterion; we want to cover each cell with a target number of compounds, usually one. As we sum the squared deviation from this target number, the smaller the UCC the better will be the design. We can compare the resulting Alternating Algorithm diversity design UCC to several benchmark numbers, a random design, an orthogonal design made from diverse reagents, and finally a nonorthogonal design of the same size. Selecting reagents at random requires very little computational effort. Selecting diverse reagents, for example by clustering, requires relatively little computational effort and improves the UCC over a random design. The Alternating Algorithm requires more computational effort but produces a better UCC. It is useful to know how well we could do if we did not restrict the design to be orthogonal. When we remove the orthogonal design requirement, we improve the UCC by a modest amount, relative to the improvement we achieve over a random design.

There are a number of ways we could use external data to judge the quality of a design. We could count the number of active compounds or chemical classes found by the design. Alternatively we could judge a design method (descriptors and selection algorithm) by the predictive power of the resulting statistical model. For example, what would the number of actives be among a fixed number of compounds predicted to be most active?

There are a number of ways to consider the cost of library designs. There is the human time to execute the design. If computers take excessive time to enumerate libraries, compute descriptors, and make design selection, then we are slower to market and the payoff from the work. Finally, if our design is not as good as it could be and we find fewer active compounds, there is an opportunity cost. Expert human costs are 50−100 dollars per hour. Computer time is remarkably inexpensive, 0.30 dollars per hour. The value of a day of lost time early in the discovery process is 80−100 thousand dollars. It is fairly clear that discovery time cost should be the driving factor, and we should seek to reduce design construction time.

The time and complexity of library design can be reduced if the selection is made over reagents rather than products. So the methods of Andrews and Cramer[3] might produce a greater benefit as more potential compounds can be evaluated using the same effort. It comes down to expectations. Andrews and Cramer can examine the product of the number of components, R×S compounds, by looking at only R+S items. A product design can look at only r×s compounds, where $r \ll R$ and $s \ll S$. What are the hit rates for a reagent design and a product design? We know of no benchmarking hit rate numbers in the literature.

In summary, the Alternating Algorithm can select orthogonal designs giving uniform coverage or optimize some other property such as expected potency based upon a SAR model. The computational speed is reasonable; we selected a 20 × 20 library from 10k, 20k, and 274k candidates in 10, 15, and 86 s on a desktop Intel machine. The algorithm is stochastic and does not guarantee a global optimum. Our limited empirical examination of random restart indicates that the quality of the final designs does not differ very much.

Availability of Data. Structure files (SD/smiles) are available for the amines, carboxylic acids, and the resulting products. Also files giving the BCUTs for the three data sets are available.

REFERENCES AND NOTES

[1] Tan, D. S.; Foley, M. A.; Shair, M. D.; Schreiber, S. L. Stereoselective synthesis of over two million compounds having structural features both reminiscent of natural products and compatible with miniaturized cell-based assays. *J. Am. Chem. Soc.* **1998**, *120*, 8565−8566.

[2] Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **2000**, *287*, 1964−1969.

[3] Andrews, K. M.; Cramer, R. D. Toward general methods of targeted library design: Topomer shape similarity searching with diverse structures as queries. *J. Med. Chem.* **2000**, *43*, 1723−1740.

[4] Beno, B. R.; Mason, J. S. The design of combinatorial libraries using properties and 3D phamacophore fingerprints. *Drug Discovery Today* **2001**, *6*, 251−258.

[5] Drewry, D. H.; Young, S. S. Approaches to the design of combinatorial libraries. *Chemom. Intell. Lab. Syst.* **1999**, *48*, 1−20.

[6] Gillet, V. J. Reactant- and product-based approaches to the design of combinatorial libraries. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 371−380.

[7] Zheng, W.; Cho, S. J.; Tropsha, A. Rational combinatorial library design. 1. Focus-2D: A new approach to the design of targeted combinatorial chemical libraries. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251−258.

[8] Johnson, M. E.; Moore, L. M.; Ylvisaker, D. Minimax and maximin distance designs, *J. Stat. Planning Inference* **1990**, *26*, 131−148.

[9] Tobias, R. *SAS QC Software;* Usage and Reference, SAS Institute: Cary, NC, 1995; Vol. 1.

[10] Cummins, D. J.; Andrews, C. W.; Bentley J. A.; Cory, M. Molecular diversity in chemical databases: Comparison of medicial chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

[11] Lam, R. L. H.; Welch, W. J.; Young, S. S. Uniform coverage designs for molecule selection. *Technometrics* **2002**, *44*, 99−109.

[12] Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.

(13) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(14) 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.

(15) Dragon. See www.disat.unimib.it/chm.

(16) Stanton, D. T. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.

(17) Yi, B.; Hughes-Oliver, J. M.; Zhu, L.; Young, S. S. A factorial design to optimize cell-based drug discovery analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1221−1229.

(18) Burden, F. R.; Winkler, D. A. A quantitative structure−activity relationships model for the acute toxicity of substituted benzenes to *Tetrahymena pyriformis* using Bayesian-regularized neural networks. *Chem. Res. Toxicol.* **2000**, *13*, 436−440.

(19) Lambert, C. G. ChemTree Manual. See www.GoldenHelix.com. 2003.

CI034125+