A Novel Frequency Distribution Selection Method for Efficient Plate Layout of a Diverse Combinatorial Library

Edward T. Graham,* Stephen P. Jacober, and Mario G. Cardozo

Department of Information Technology and Department of Medicinal Chemistry, Boehringer Ingelheim Pharmaceuticals, R&D, 175 Briar Ridge Road, Ridgefield, Connecticut

Received April 23, 2001

A deterministic method (frequency distribution method) for selecting compounds from a partitioned virtual combinatorial library for efficient synthesis is presented here. The method is based on reagent frequency analysis and can be applied to any library of molecules distributed in any given partitioned chemical space (cluster, cell-based, etc.). Compound selection by reagent frequency distribution can produce a unique, diverse set of molecules that adequately represents the library while requiring the least amount of compounds to be synthesized and minimizing the number of different reagents that must be used. This method also provides a practical solution to the configuration of plate layout. Because the method essentially identifies "expensive" regions in the chemical space to synthesize for a desired diversity or similarity coverage, decisions concerning the necessity to synthesize these compounds can be addressed. Minimum compound generation and efficient plate layout results in savings both in time of synthesis and cost of materials. This method always results in a discrete solution, which can be used for any given library size as well as any combination of reagents and is also readily adaptable to robotic automation.

INTRODUCTION

It is now possible with the emerging technologies of combinatorial chemistry to generate potentially enormous chemical libraries of structurally diverse molecules. Combinatorial chemistry assembles selected sets of reagents in combinatorial arrangements, using appropriate chemical reactions, into a diverse library of related compounds. However, because of the size of these libraries, it is not commercially practical to synthesize all of the potential molecules and test them for biological activity. Therefore, anyone attempting to test such libraries for biological activity is faced with the problem of devising a method to perform a selection of a subset of compounds from a large combinatorial library, such that the subset possesses maximum chemical diversity[1−3] while maintaining practical size limitations in order to obtain meaningful data or structure activity relationships. Additionally, it is also desirable to obtain a subset that can be synthesized efficiently using a minimum number of reagents.

There are many strategies to creating combinatorial libraries[4] and selecting compounds for testing. Some subset-selection methods are based on algorithms that measure library similarity/dissimilarity.[5−10] Other methods, based on genetic algorithms,[11] neural networks,[12,13] or random sampling do not produce unique solutions, which can be desirable for initiating further investigations. Another approach to the subset selection utilizes various empirical information hopefully generating a propensity of "virtual hits". For example, LEGEND[14] and LUDI[15] aid in structure generation based on knowledge of the receptor site. PRO_SELECT[16] identifies prospective substituents based on knowledge of the active site and known synthetic routes. RECAP[17] uses knowledge obtained by fragmenting biologically active molecules around bonds to select important building blocks. Still other methods have been devised based upon drug-like characteristics (e.g. DLI,[18] SELECT[19]).

Only a few methods have directly attempted to take into consideration the practicality of automated chemical synthesis (LiBrain,[20] HARPick[21]) or resources (e.g. inventory and cost[22]). SPROUT[23,24] (like LEGEND and LUDI) generates a virtual library based upon physical constraints of a pharmacophore hypothesis but offers clustering options which take into consideration the ease of synthesis and availability of starting materials. PLUMS[25] expands on the concept of optimizing a library based on physical constraints by subsequently removing monomers from the collection of "virtual hits" resulting in a smaller library which balances effectiveness and efficiency. FOCUS-2D[26] searches a library and tries to identify compounds similar to lead molecules and then uses monomer frequency analysis to minimize the library by selecting common "building blocks" of these "virtual hits".

Although many of these methods can produce maximally diverse or specific focused subsets, the subset selected still may not be as useful even though it may be considerably smaller than the original library. The reason for this is that typically these methods do not take into account a combinatorial constraint as part of the subset selection, requiring too many reagents to prepare the selected subset of the chemical library. It may be that many of these reagents have to be used for only a few, specific combinations, or, in some cases, only once, leading to an excess in the number of reagents. This situation increases the manipulation of reagents as well as requires complex and less practical robotic operations. Although all of the previously mentioned methods

* Corresponding author phone: (203)798-4404; e-mail: egraham@rdg.boehringer-ingelheim.com.

10.1021/ci0100393 CCC: $20.00
Published on Web 09/27/2001

could be used for initial clustering or selecting members of the virtual library, none of these methods specifically address the problem of an efficient combinatorial arrangement of experimental blocks (i.e. plate layout).

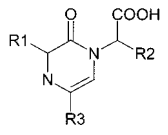
We have developed a deterministic method for selecting compounds from a partitioned virtual combinatorial library for efficient synthesis called the frequency distribution method. The method is based on reagent frequency analysis and can be applied to any library of molecules distributed in any given partitioned chemical space (cluster, cell-based, etc.). Compound selection by reagent frequency distribution can produce a unique, diverse set of molecules that adequately represents the library while requiring the least amount of compounds to be synthesized and minimizing the number of different reagents that must be used. This method also provides a practical solution to the configuration of plate layout. Because the method essentially identifies "expensive" regions in the chemical space to synthesize for a desired diversity or similarity coverage, decisions concerning the necessity to synthesize these compounds can be addressed. Minimum compound generation and efficient plate layout results in savings both in time of synthesis and cost of materials. This method always results in a discrete solution, which can be used for any given library size as well as any combination of reagents and is also readily adaptable to robotic automation.

METHODS

A typical procedure in combinatorial library design is to partition the library into regions of "similarity" by any method and to whatever degree that is desired.[27−30] For example, chemical descriptors, physical properties, pharmacophore, or any other criterion or group of criteria that provides a rational basis for grouping or placing compounds into discrete sets or clusters could be used. The sensitivity of our method to select a diverse set of compounds is directly based upon the ability to partition the library into well-separated clusters. However, it is not the intent here to present a discussion on clustering techniques.

Substituent frequency analysis of molecular fragments (reagents) used to prepare the whole molecule is applied to this partitioned chemical library resulting in ordered lists of fragments based on their frequency across the clusters. These lists, ordered by decreasing frequency and limited to one list for each location, are then further organized to minimize resources required for synthesis (e.g., number of reagents and/or number of reaction plates). When this approach is used to guide reagent selection for the synthesis of a chemical library, the subset generated by these lists will produce the minimum number of combinations for a corresponding degree of molecular diversity or target similarity using a minimum number of resources.

Consider the following general chemical structure (structure A) with three substituent group locations: R1, R2, and R3.



Structure A

R1, R2, and R3 represent the geographic locations for substitution on the common core of structure A.

A virtual chemical library based on structure A is constructed and then partitioned into regions of similarity, e.g., "clusters". The frequency distribution method is described as follows with reference to structure A.

(I) Generation of Rank Ordered Frequency Distribution Lists. Frequency distribution lists of substituents rank ordered from high to low are generated for each substituent group location (i.e., R1, R2, and R3). This is accomplished by performing a series of steps described below, see see Scheme 1. (Substituent and reagent are used interchangeably in the description below.)

(A) The first step is to determine the most frequent substituent within a specific geographic location (e.g. R1) across the clusters, i.e., what is the specific substituent that is present in the most number of clusters. This substituent is considered to provide the most chemical diversity for that location when combined with the other substituents at other locations (i.e., R2 and R3) because it is represented in more different clusters than any other substituent for that group location.

(B) The second step is to eliminate all of the clusters and their contents (all other compounds in those clusters) where the substituent determined in step A is present. This elimination is based on the rationale that all compounds within any given cluster are assumed to be equivalent for representation of that chemical space. Therefore, it is most efficient to have only one compound from each cluster, rather than several compounds from the same cluster.

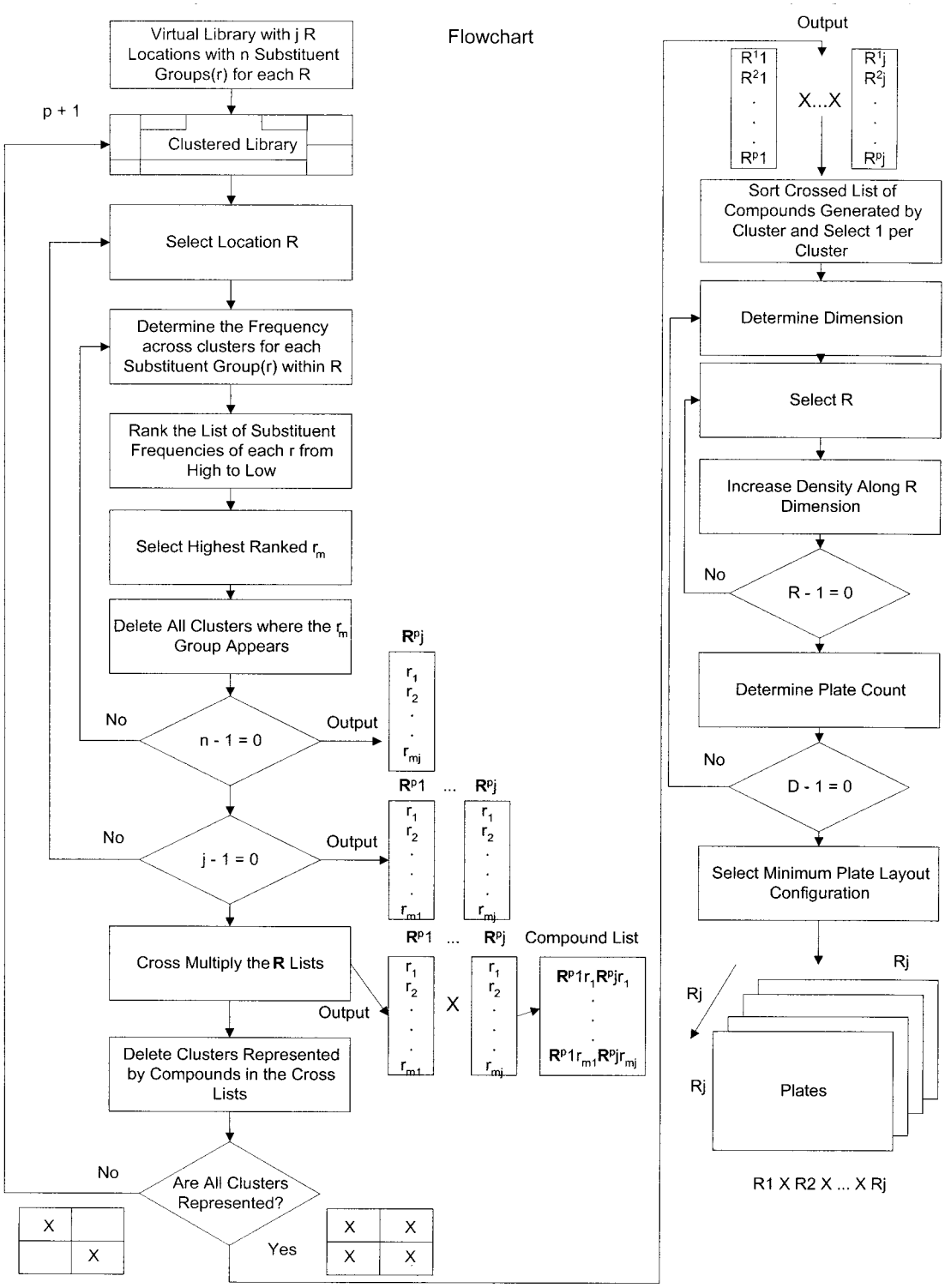
(C) At this point, steps A and B are repeated to determine the second most frequent substituent, the third most frequent, etc. This produces an ordered list, $\mathbf{R}^1 1(r_1, r_2, r_3, \ldots r_m)$, where the r_1's are ranked from most frequent to least frequent. Note that when "R" is implied to contain member substituents (r's), it will be denoted as a vector **R**. As noted before, R1 represents a specific location on the parent molecule. The superscript to the vector **R** denotes the pass through the library (see below). The resultant frequency list can have any number of members from 1 to m, the number of substituents required to cover all of the clusters (m will depend on the diversity of the substructures (reactants) at a specific location and will be less than the total number of reactants at that position). The frequency count for all of the individually selected substituents from the list will sum to the total number of clusters (N). For example, $\sum_{i=1}^{m} \text{freq}(\text{R1}(r_i)) = N$, the summation of the frequency counts for the select substituents at the R1 locations will equal the total number of clusters. This will be true for each location.

(D) Repeat this entire procedure (steps A−C) for each "R" group substituent location (e.g., R2 and R3) to generate an rank ordered frequency list for each.

(E) The cross product of the lists is constructed (e.g., $\mathbf{R}^1 1 \times \mathbf{R}^1 2 \times \mathbf{R}^1 3$), and the clusters covered by the compounds formed are removed from the virtual chemical library leaving a new smaller library that contains only compounds from the clusters not yet represented. The cross product will result in $m1 \times m2 \times m3$ compounds.

(F) Repeat steps A−E on the new library subset generating lists for the second, third, etc. pass through the library until

Scheme 1



all clusters are covered (i.e., after cluster elimination in step E there are no clusters remaining). Note: see also the section on frequency distribution method variations.

Example 1. Referring to Figure 1, **R**1 consists of the following reagents: "star", "circle", and "diamond". **R**2 consists of the following reagents: "cross", "triangle", and "square". There are four clusters shown, cluster 1 having one member, cluster 2 having two members, cluster 3 having three members, and cluster 4 having three members. Looking at R1 and using step A, the most frequent substituent for R1 is a "star" (i.e. looking at the clusters, R1 is a "star" in clusters 1, 3, and 4). Using step B, those clusters are eliminated. Step A is then used to determine the second most frequent substituent. Since only cluster 2 remains, the second most frequent substituent for R1 is a "diamond". Then, using step B, cluster 2 is eliminated, and since all clusters have been eliminated for R1, no further action is necessary. The ranking for R1 is a "star" first (present in three (3) clusters) and a "diamond" second (present in one (1) cluster), $\mathbf{R}^1 1 = \{\bigstar, \blacklozenge\}$. Note that the number of clusters represented by star and diamond is 4 which is equal to the total number of clusters.

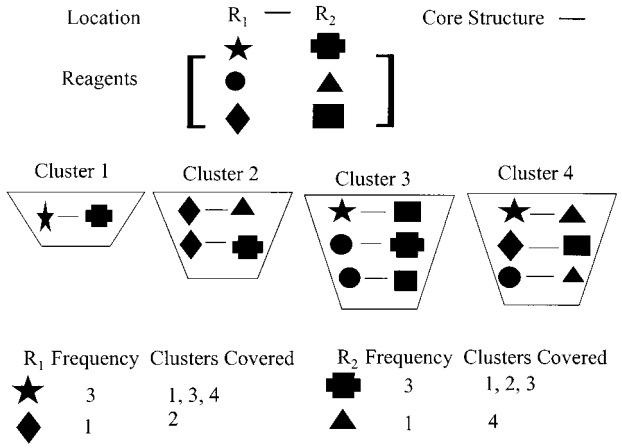


Figure 1. Pictorial example of the frequency distribution method.

Steps A−C are now performed at R2. After step A, a "cross" is the most frequent substituent for R2 found in three clusters (cluster 1, cluster 2, and cluster 3). Those clusters are eliminated. In the remaining cluster, cluster 4, the most frequent reagent for R2 is triangle. All clusters have now been eliminated, and the ranking for R2 is a "cross" first and a "triangle" second

$$\mathbf{R}^1 2 = \{\blacksquare, \blacktriangle\}$$

Note that the list for R1 has two elements, and the list for R2 also has two elements. The cross product will have four elements.

However, even though the cross product based on reagent frequency analysis has four elements

$$(\mathbf{R1} \times \mathbf{R2} = \{\bigstar - \blacksquare, \bigstar - \blacktriangle, \blacklozenge - \blacksquare, \blacklozenge - \blacktriangle\})$$

these do not produce sufficient combinations to cover all of the clusters (cluster 3 with three (3) elements is not covered). Because the "R" groups frequency is determined independently of each other, it is possible that all of the clusters would not be represented when the cross product is generated.

Accordingly, a new data set is created whose contents are the clusters not covered by the compounds produced by the crossed frequency lists (in the above example, cluster 3).

Steps A and B from above are then repeated for the new data set to create a secondary frequency list: [$\mathbf{R}^2 1 = \{\bullet\}$, $\mathbf{R}^2 2 = \{\blacksquare\}$).

If necessary, the above process is continually repeated (to generate tertiary, quaternary, etc. lists) until one of the following three criteria are met.

(1) All of the clusters are covered. [Note: It may not be possible to generate sufficient lists that result in all clusters being covered. Because clusters are deleted and not R group substituents from the clusters remaining after the list created by the cross product is generated, the same frequency lists can be continually repeated (loop) because no new combinations are being generated. Alternative methods can be used to generate lists that cover all clusters. These lists may not produce as an efficient robotic solution for synthesis (as determined by number of plates to cover clusters, method described below). See also the section on frequency distribution method variations.]

Table 1. Example: Rank Data for Eight Compounds with Two R Groups (R1 and R2)

	freq ranking for reactants at			freq ranking for reactants at	
compd ID	R1	R2	compd ID	R1	R2
1	1	2	5	3	4
2	2	1	6	3	5
3	2	4	7	4	3
4	3	3	8	4	4

(2) The frequency counts of the R groups are low (e.g. the most frequent R group has a count of 2 or 3 or any arbitrary number selected as a cutoff).

(3) The total number of members in the frequency list equals some number of *r* substituents that is desired (due to cost of materials, reagent availability, or cluster coverage).

(II) Rank the Substituents Across Lists. The "R" substituent groups across lists are ranked from the most frequently occurring (1) to least frequently occurring (*p*).

Suppose for example that a combinatorial library has 281 substituents at R1, 219 at R2, and 150 at R3. Therefore, the total number of compounds possible to make is ($281 \times 219 \times 150 = 9\,230\,850$). If after frequency analysis R1 has a primary list ($\mathbf{R}^1 1$) containing 50 substructures, a secondary list ($\mathbf{R}^2 1$) of 30, and a tertiary list ($\mathbf{R}^3 1$) of 16 (note: $50 + 30 + 16 = 96$), then arrange the frequency ranks from the primary list as 1−50, the secondary list 51−80, and the tertiary list 81−96. Therefore for the R1 location p1 = 1−96. Now let us suppose R2 has 65 substructures in the primary list ($\mathbf{R}^1 2$), 20 in the secondary ($\mathbf{R}^2 2$), and 11 in the tertiary ($\mathbf{R}^3 2$), and R3 has 45 substructures in the primary list ($\mathbf{R}^1 3$). Generate the cross product of all the "R" group lists to create a list of compounds to synthesize, as given below.

$$\{(\mathbf{R}1 \text{ primary} + \text{secondary} + ...) \times (\mathbf{R}2 \text{ primary} + \text{secondary} + ...) \times (\mathbf{R}3 \text{ primary} + \text{secondary} + ...)\}$$

The example above would result in ($96 \times 96 \times 45$) or 414 720 potential compounds (or 4.5% of the original library) to cover all clusters. While this is substantially less than the total number of compounds in the virtual library (9 230 850), it is still quite a large number.

(III) Selection of Representative Compounds. The list of compounds generated in step II are sorted hierarchically, first by cluster, second by rank-R1 (p1), third by rank-R2 (p2), fourth by rank-R3 (p3), etc. Select the *first* observation in every cluster in the sorted list. This compound is assumed to be representative of that cluster and has the property of being composed of reagent groups that are represented in more different clusters. The number of compounds required to represent the entire library now equal the number of clusters generated as determined by the partitioning method. The chemical space that the 414 720 potential compounds covered can now be represented by *N* (the number of clusters) compounds. And, if one could "cherry pick" these compounds to synthesize, then this is all that would be needed. Unfortunately, "cherry picking" is not an efficient synthesis method. Therefore, a method to minimize the compounds required to synthesize the *N* compounds is desirable.

(IV) Optimizing Synthesis. If a plot is produced using as the coordinates the ranks of the substituent groups: (e.g. R1 1−96, R2 1−96, R3 1−45) of the compounds selected in step III, this would represent the plate layout needed to

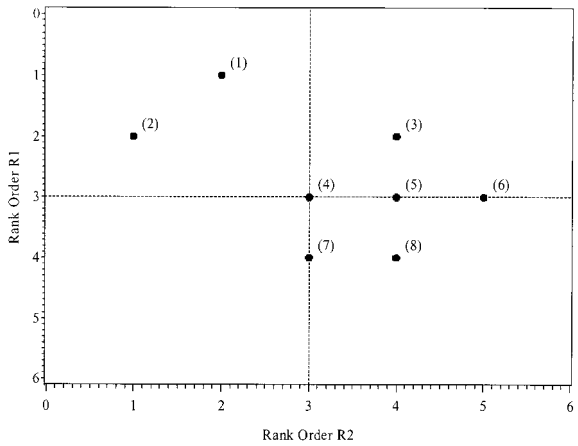


Figure 2. Density map of rank order data (R1 vs R2). Note: three compounds in quadrant 1 (upper left 1, 2, 4), three compounds in quadrant 2 (upper right 3, 5, 6), one compound in quadrant 3 (lower left 7), and one compound in quadrant 4 (lower right 8).

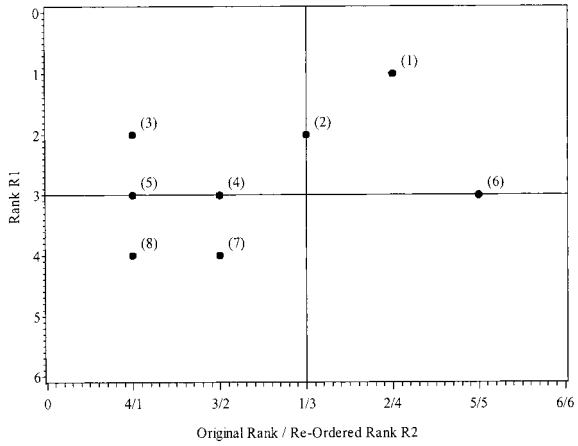


Figure 3. Column reduced density map (R1 vs remapped R2). Note: four compounds in quadrant 1 (upper left 2, 3, 4, 5), two compounds in quadrant 2 (upper right 1, 6), two compounds in quadrant 3 (lower left 7, 8), and zero compounds in quadrant 4 (lower right).

generate all of the compounds necessary to synthesize the selected representative compounds. For more than two substituent groups the plate layout would be represented by 2-D slices through this higher dimensional space. If one could maximize the density to the smallest area of space, this would be the optimum condition for the most efficient synthesis of compounds. This can be accomplished by reordering all of the coordinate ranks such that the remapped axis result in the compounds occupying as small a region as possible in the dimensional space.

Example 2. Rank data for compounds generated from a parent compound with two substituent locations (R1 and R2) are listed in Table 1 and plotted as a density map in Figure 2. Determine the number of compounds in each column/row and reorder the axis values (frequency ranks) from high to low based on the number of compounds in that column/row. Determine the number of compounds in each row/column and reorder (transform the axis) as well from high to low. The transformed axis results in a new density map (column reduced in Figure 3 and column and row reduced in Figure 4, Table 2) that increases the density to the maximum allowable for this particular set of selected compounds. Note that the maximum density may result by

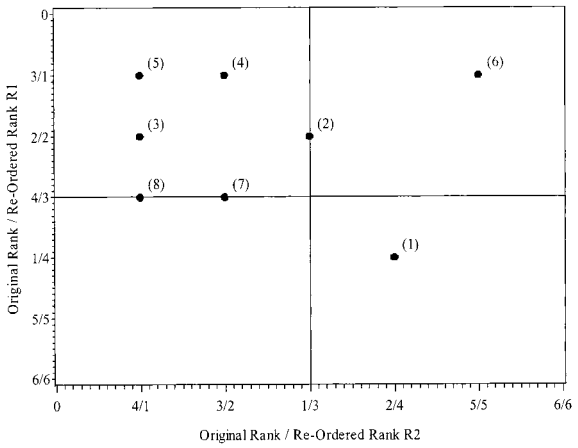


Figure 4. Column and row reduced density map (remapped R1 vs remapped R2) Note: six compounds in quadrant 1 (2, 3, 4, 5, 7, 8), one compound in quadrant 2 (6), zero compounds in quadrant 3, and one compound in quadrant 4 (1).

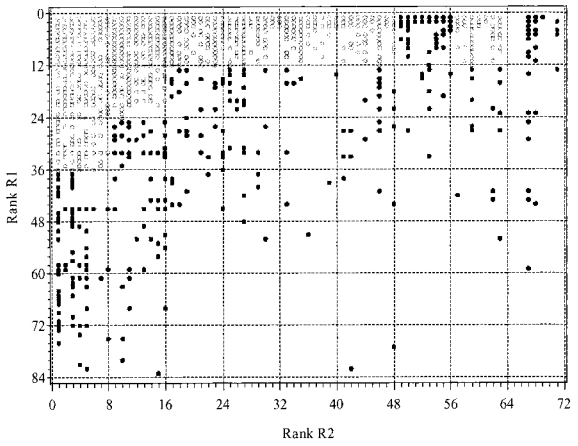


Figure 5. Frequency distribution method 1 (no plate optimization). Density map of 886 clusters ranked R1 versus ranked R2 reactants (substructures). R1 contains 84 reactants and R2 contains 72 reactants (first two lists). The 10 most dense (12 × 8) plates (open circles ○) cover 597 clusters. It would require 46 plates to cover all of the 886 clusters. Redundant compounds represented by empty spaces.

Table 2. Remapped (Axis Transformed) R Group Data

	freq reranking for reactants at			freq reranking for reactants at	
compd ID	R1	R2	compd ID	R1	R2
1	4	4	5	1	1
2	2	3	6	1	5
3	2	1	7	3	2
4	1	2	8	3	1

reordering the rows first and the columns second. Both scenarios should be checked. The scenario that produces the highest density region using the minimum number of reagents should be selected.

(V) Determine Plate Layout. After the maximum density has been determined, the map should be divided into grids of desired dimensions. For example, the density map can be divided into grids containing 96 "wells" {12 × 8, 8 × 12} and the number of clusters (compounds) in each grid is counted. The grids are then sorted from high to low by cluster coverage. This will give the minimum number of plates required (least number of compounds) to cover a specified number of clusters.

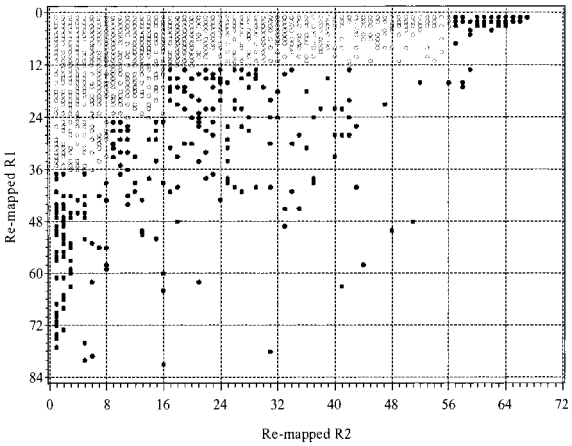


Figure 6. Frequency distribution method 1. Density map of 890 clusters reranked R1 versus reranked R2 reactants. R1 contains 84 reactants and R2 contains 67 reactants (three lists). The 10 most dense (12 × 8) plates (open circles ○) cover 636 clusters. There would be 42 plates required to cover 890 clusters. Note: Although the R1 substituent count is 92 and the R2 substituent count is 80 for the total of the three lists, because substituent groups can be reselected in method 1, the actual number of unique substituents required to cover the 890 clusters is 84 for R1 and 67 for R2. Redundant compounds represented by empty spaces.

As noted above, to guarantee the optimum solution, both R1 × R2 and R2 × R1 are required to be evaluated. When there are greater than two R group locations, the above procedures will have to be repeated for each pair of possible R group combinations to guarantee the optimum maximum density. The number of possible combinations can be calculated from the following formula

$$2\binom{m}{2} = \left(\frac{2m!}{2!(m-2)!}\right)$$

where m = the number of R group locations. For a 3-R group substituent problem, six density maps would have to be evaluated (R1 × R2 by R3, R2 × R1 by R3, R2 × R3 by Rl, R3 × R2 by R1, R1 × R3 by R2, and R3 × R1 by R2). Sometimes the number of possibilities can be reduced due to a preferential condition, i.e., if one of the substituent locations only had a few reagents selected. If R3 only had five possibilities, then it would make sense that only the R1 × R2 by R3 and R2 × R1 by R3 be considered.

FREQUENCY DISTRIBUTION METHOD VARIATIONS

A parent compound has two substituent positions (R1 and R2). R1 has 281 substituents and R2 has 219 substituents. All together there are 61 539 possible combinations (R1 × R2). The 61 539 possible structures are partitioned into 898 regions of similarity (clusters). Three different variations based on the frequency distribution method for each of the two substituent locations are performed on the 898 clusters. The variations will be noted as method 1, method 2, and method 3.

Method 1 is as described above. No improvement in number of clusters covered resulted after the tertiary list is generated. Figure 5 (no optimization performed) and Table 3 show the results for method 1. Figure 6 and Table 4 show method 1 with the three list iterations and density map optimization. Note the improvement in the partitions covered with plate number for the density map optimizations is readily apparent. Table 5 shows a summary of the results using method 1.

Table 3. Frequency Distribution Method (No Plate Optimization, Two Lists)[a]

no. of plates	plate no.	freq count	% coverage	no. of clusters covered
1	2	89	10.0	89
2	1	83	9.3	172
3	3	71	8.0	243
4	10	67	7.6	310
5	4	56	6.3	366
6	6	48	5.4	414
7	11	48	5.4	462
8	5	47	5.3	509
9	19	46	5.2	555
10	8	42	4.7	597
11	7	35	4.0	632
12	20	25	2.8	657
13	28	21	2.4	678
14	46	20	2.2	698
15	9	18	2.0	716
16	37	18	2.0	734
17	12	15	1.7	749
18	13	13	1.5	762
19	21	11	1.2	773
20	29	10	1.1	783
21	15	9	1.0	792
22	38	9	1.0	801
23	24	8	0.9	809
24	16	7	0.8	816
25	17	7	0.8	823
26	30	7	0.8	830
27	55	6	0.7	836
28	14	5	0.6	841
29	22	5	0.6	846
30	18	4	0.4	850
31	31	4	0.4	854
32	47	4	0.4	858
33	26	3	0.3	861
34	33	3	0.3	864
35	35	3	0.3	867
36	36	3	0.3	870
37	56	3	0.3	873
38	25	2	0.2	875
39	27	2	0.2	877
40	32	2	0.2	879
41	60	2	0.2	881
42	23	1	0.1	882
43	40	1	0.1	883
44	41	1	0.1	884
45	44	1	0.1	885
46	45	1	0.1	886

[a] Plate definition (12 × 8). Plates are numbered 1−63 left to right, top to bottom.

Method 2 is identical to method 1 except that when the new data set of clusters not covered by the previous list(s) is generated, in addition to eliminating the clusters that are covered, those structures that include substituent reagents already included in the frequency list are also eliminated from the clusters that remain. Accordingly, the entire structure is eliminated, which eliminates other substituent locations from further frequency analysis also. This method always results in all clusters being covered when all of the reagent lists are crossed for R1 and R2. Figure 7 and Table 6 shows the density plate optimization for method 2. Table 7 has a summary of the results for method 2.

Method 3 is also as outlined above by method 1, except that when creating the new data set of clusters not covered by the previous list(s), in addition to eliminating the clusters that are covered, only those substituents for a particular location (R1 or R2) are eliminated from the frequency lists.

Table 4. Frequency Distribution Method 1 (Plate Optimized, Three Lists)[a]

no. of plates	plate no.	freq count	% coverage	no. of clusters covered
1	1	91	10.2	91
2	2	86	9.7	177
3	3	80	9.0	257
4	4	68	7.6	325
5	10	68	7.6	393
6	5	59	6.6	452
7	11	55	6.2	507
8	19	53	6.0	560
9	6	42	4.7	602
10	7	34	3.8	636
11	12	31	3.5	667
12	28	26	2.9	693
13	8	24	2.7	717
14	37	23	2.6	740
15	20	22	2.5	762
16	13	18	2.0	780
17	46	12	1.3	792
18	21	11	1.2	803
19	55	9	1.0	812
20	29	8	0.9	820
21	14	7	0.8	827
22	22	7	0.8	834
23	30	6	0.7	840
24	31	6	0.7	846
25	9	5	0.6	851
26	15	5	0.6	856
27	23	5	0.6	861
28	32	5	0.6	866
29	38	4	0.4	870
30	17	3	0.3	873
31	24	3	0.3	876
32	16	2	0.2	878
33	42	2	0.2	880
34	56	2	0.2	882
35	33	1	0.1	883
36	34	1	0.1	884
37	41	1	0.1	885
38	47	1	0.1	886
39	48	1	0.1	887
40	51	1	0.1	888
41	53	1	0.1	889
42	58	1	0.1	890

[a] Plate definition (12 × 8). The plates are numbered 1−63 left to right, top to bottom.

Table 5. Method 1

list	no. of reagents per freq list R1	no. of reagents per freq list R2	clusters covered (cumulative)	% of total clusters covered (cumulative)	required compds to synthesize (cumulative)
1	73	61	844	93.9	4453
2	11	11	886	98.6	6048
3	8	8	890	99.1	7360
4	6	5	890	99.1	8330
5	6	5	890	99.1	8330[a]

[a] List 5 is a repeat of list 4 with the same reagents and will continue to repeat ad infinitum.

The entire structure is not eliminated as in method 2. Figure 8 and Table 8 shows the density plate optimization for method 3. Table 9 has a summary of the results for method 3.

DISCUSSION

When a large collection of compounds are partitioned by any procedure based on some criteria, there will likely be a wide distribution in the number of compounds populated per partition (cell, cluster, etc.). Since compounds are distributed according to their similarity, partitions with small populations are the least like other compounds in the collection. For example, the most uncommon compounds will occur alone or with only a few other compounds in separate partitions. These partitions are selected last during the frequency

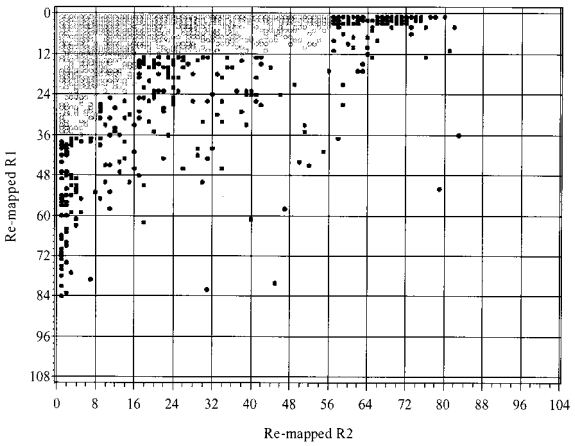


Figure 7. Frequency distribution method 2. Density map of 898 clusters reranked R1 versus reranked R2 reactants. R1 contains 123 reactants and R2 contains 100 reactants. The 10 most dense (12 × 8) plates (open circles ○) cover 620 clusters. There would be 49 plates required to cover all of the 898 clusters. Redundant compounds represented by empty spaces.

Table 6. Frequency Distribution Method 2 (Plate Optimized, Six Lists)[a]

no. of plates	plate no.	freq count	% coverage	no. of clusters covered
1	2	90	10.0	90
2	1	88	9.8	178
3	14	73	8.1	251
4	3	72	8.0	323
5	4	68	7.8	391
6	5	55	6.1	446
7	15	47	5.2	493
8	6	46	5.1	539
9	27	44	4.9	583
10	7	37	4.1	620
11	16	31	3.4	651
12	8	29	3.2	680
13	9	25	2.8	705
14	53	22	2.5	727
15	40	20	2.2	747
16	28	16	1.8	763
17	29	14	1.6	777
18	17	12	1.3	789
19	10	11	1.2	800
20	18	11	1.2	811
21	66	11	1.2	822
22	41	10	1.1	832
23	79	10	1.1	842
24	19	7	0.8	849
25	31	5	0.6	854
26	43	5	0.6	859
27	54	5	0.6	864
28	21	4	0.4	868
29	30	3	0.3	871
30	46	3	0.3	874
31	11	2	0.2	876
32	20	2	0.2	878
33	32	2	0.2	880
34	33	2	0.2	882
35	42	2	0.2	884
36	22	1	0.1	885
37	23	1	0.1	886
38	34	1	0.1	887
39	37	1	0.1	888
40	44	1	0.1	889
41	47	1	0.1	890
42	55	1	0.1	891
43	56	1	0.1	892
44	58	1	0.1	893
45	62	1	0.1	894
46	68	1	0.1	895
47	70	1	0.1	896
48	82	1	0.1	897
49	84	1	0.1	898

[a] Plate definition (12 × 8). Plates are numbered 1−117 left to right, top to bottom.

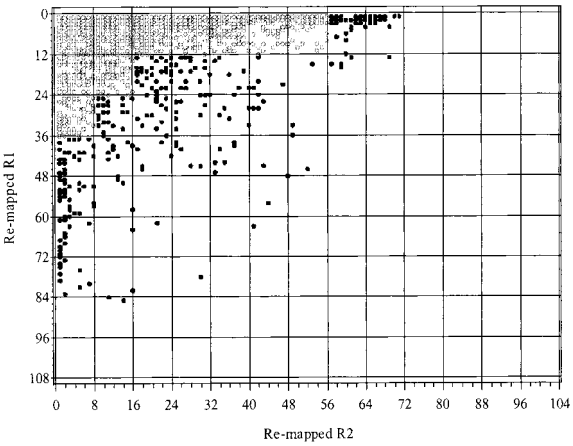


Figure 8. Frequency distribution method 3. Density map of 894 clusters reranked R1 versus reranked R2 reactants. R1 contains 96 reactants and R2 contains 78 reactants. The 10 most dense (12 × 8) plates (open circles ○) cover 632 clusters. There would be 43 plates required to cover 894 clusters. Redundant compounds represented by empty spaces.

Table 7. Method 2

list	no. of reagents per freq list R1	no. of reagents per freq list R2	clusters covered (cumulative)	% of total clusters covered (cumulative)	required compds to synthesize (cumulative)
1	73	61	844	93.9	4453
2	19	12	890	99.1	6716
3	12	9	893	99.4	8528
4	8	8	896	99.7	10 080
5	6	6	896	99.7	11 328
6	4	3	898	100	12 078
7	1	1	898	100	12 300
8	0	0			

distribution method(s). These compounds are the most expensive to synthesize in terms of unique reagents required and computer time to locate. (These unique reagents will have low frequencies and therefore will be selected in the later lists. Also, they will likely be located on plates which have higher redundant cluster information, i.e., cluster duplication from previous plates.) Note that in all three variations of the frequency distribution method described above, plates populated first cover more partitions (clusters) than plates toward the end. To cover all of the partitions, a high price in terms of compounds required to be synthesized is paid to represent these clusters containing more unique compounds. For example, in method 1 as shown in Figure 6 (Table 4), after 33 plates there are 880 clusters that are represented. To cover the next 10 clusters requires nine more plates (or 864 additional compounds to be synthesized). A determination would need to be made on whether it would be more work for a robot to generate these 864 compounds (and subsequent testing) or have a chemist synthesize 10 compounds to represent these clusters. Also note from Figure 7 (Table 6) that method 2 also covers 880 clusters after 33 plates and requires an additional 16 plates (or 1536 compounds) to cover all of the 898 clusters (18 more clusters than 33 plates cover).

All three variations of the frequency distribution method produce similar results (see Table 10). While method 1 allows reagents to be repeatedly selected as long as it represents the maximum frequency (most common reagent across clusters), method 2 and method 3 will select unique reagents as the iteration process continues. Method 2 will always eventually generate lists sufficient to cover all clusters where methods 1 and 3 cannot necessarily be driven to cover all clusters. There are several criteria that should be considered when determining which method to use: cluster coverage (only method 2 guarantees that all clusters will be covered); resources (reagents, plates, etc.); and number of compounds required to be synthesized (and tested).

In general, although all of the methods give similar results for early iterations, method 1 will produce lists with the

Table 8. Frequency Distribution Method 3 (Plate Optimized, Three Lists)[a]

no. of plates	plate no.	freq count	% coverage	no. of clusters covered
1	1	90	10.1	90
2	2	86	9.6	176
3	3	79	8.8	255
4	4	67	7.5	322
5	14	67	7.5	389
6	5	59	6.6	448
7	15	55	6.2	503
8	27	52	5.8	555
9	6	42	4.7	597
10	7	35	3.9	632
11	16	31	3.5	663
12	8	25	2.8	688
13	53	24	2.7	712
14	28	22	2.5	734
15	40	22	2.5	756
16	17	18	2.0	774
17	9	13	1.4	787
18	29	12	1.3	799
19	66	12	1.3	811
20	41	9	1.0	820
21	79	9	1.0	829
22	18	7	0.8	836
23	30	7	0.8	843
24	44	6	0.7	849
25	19	5	0.6	854
26	31	5	0.6	859
27	42	5	0.6	864
28	43	5	0.6	869
29	21	4	0.4	873
30	32	3	0.3	876
31	54	3	0.3	879
32	33	2	0.2	881
33	45	2	0.2	883
34	80	2	0.2	885
35	20	1	0.1	886
36	22	1	0.1	887
37	46	1	0.1	888
38	58	1	0.1	889
39	67	1	0.1	890
40	68	1	0.1	891
41	71	1	0.1	892
42	82	1	0.1	893
43	93	1	0.1	894

[a] Plate definition (12 × 8). The plates are numbered 1−117 left to right, top to bottom.

Table 9. Method 3

list	no. of reagents per freq list R1	no. of reagents per freq list R2	clusters covered (cumulative)	% of total clusters covered (cumulative)	required compds to synthesize (cumulative)
1	73	61	844	93.9	4453
2	18	12	891	99.2	6643
3	5	5	894	99.5	7488
4	0	0			

Table 10. Methods Summary Statistics of Plate Optimized Selections

method	no. of reagents at R1	no. of reagents at R2	max. clusters covered	no. of plates required to cover max.	no. of freq lists generated	compds required to be synthesized (no. of plates X 96)
1	84	67	890	42	3	4032
2	123	100	898	49	6	4704
3	96	78	894	43	3	4128

minimum number of reagents required and thus be less expensive in terms of resources and required number of compounds that must be synthesized for a given number of plates. However, if all clusters are required to be represented and robotic automation is available, method 2 will guarantee total coverage. Method 3 will cover slightly more clusters than that obtained by method 1 but at a price of using more reagents and synthesizing more compounds.

CONCLUSIONS

It is now possible with combinatorial chemical synthesis to generate potentially enormous chemical libraries that far exceed the capacity for testing. Many methods have been developed to select a subset from a large combinatorial library that represent the library while trying to maintain practical size limitations. The frequency distribution method described above not only satisfies these conditions but also provides several other advantages toward minimizing resources. The combinatorial arrangement of the reagents optimizes plate configuration. As a result fewer compounds need to be synthesized to represent the desired chemical space. In addition, decisions concerning sufficient representation balanced by the efficient use of resources required for this representation can be readily made since the method identifies "expensive" regions of the library. The frequency distribution method can be used for any given library size, the solution is always unique and is readily adaptable to efficient robotic automation.

PATENTS

There are patents pending that cover this research.

REFERENCES AND NOTES

(1) Warr, W. W. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.
(2) Gillet, V. J.; Wild, D. J.; Willett, P.; Bradshaw, J. Similarity and Dissimilarity Methods for Processing Chemical Structure Databases. *Comput. J.* **1998**, *41*, 547−558.
(3) Potter, T.; Matter, H. Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem.* **1998**, *41*, 478−488.
(4) James, E. A.; Hassan, M.; Waldman, M. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63−70.
(5) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.
(6) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.
(7) 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.
(8) Higgs, R. E.; Bemis, K. G.; Watson, I. A.; Wikel, J. H. Experimental Designs for Selecting Molecules from Large Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 861−870.
(9) Hassan, M.; Bielawski, J. P.; Hemple, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Molecular Diversity* **1996**, *2*, 64−75.
(10) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *Quantum Struct.-Act. Relat.* **1996**, *14*, 501−506.
(11) Brown, R. D.; Martin, Y. C. Design Combinatorial Library Mixtures Using a Genetic Algorithm. *J. Med. Chem.* **1997**, *40*, 2304−2313.
(12) Sadowski, J. Optimization of Chemical Libraries by Neural Networks. *Curr. Opin. Chem. Biol.* **2000**, *4*, 280−282.
(13) Sadowski, J.; Wagener, M.; Gasteiger, J. Assessing Similarity and Diversity of Combinatorial Libraries by Spacial Autocorrelation Functions and Neural Networks. *Angew. Chem., Int. Ed. Engl.* **1996**, *34*, 2674−2677.
(14) Nishibata, Y.; Itai, A. Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation. *Tetrahedron* **1991**, *47*, 8985−8990.
(15) Bohm, H. J. The Computer Program LUDI: A New Method for the de novo Design of Enzyme Inhibitors. *J. Comput.-Aided Mol. Design* **1992**, *6*, 61−78.
(16) Murry, W. C.; Clark D. E.; Auton, T. R.; Firth M. A.; Li, J.; Sykes, R. A.; Waszkowycz, B.; Westhead, D. R.; Young, S. C. PRO_SELECT: Combining Structure-Based Drug Design and Combinatorial Chemistry for Rapid Lead Discovery. 1. Technology. *J. Comput.-Aided Molecular Design* **1997**, *11*, 193−207.
(17) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP−Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.
(18) Xu, J.; Stevenson, J. Drug-like Index: A New Approach to Measure Drug-like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.
(19) Gillet, V. J.; Willett, 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.
(20) Polinsky, A.; Feinstein, R. D.; Shi, S.; Kuki, A. LiBrain: Software for Automated Design of Exploratory and Targeted Combinatorial Libraries. In *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; Conference Proceeding Series; Chaiken, I. M., Janda, K. D., Eds.; American Chemical Society: 1996; pp 219−232.
(21) Good, A. C.; Lewis, R. A. New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick. *J. Med. Chem.* **1997**, *40*, 3926−3936.
(22) Young, S. S.; Sheffield, C. F.; Farmen, M. Optimum Utilization of a Compound or Chemical Library for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 892−899.
(23) Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: A Program for Structure Generation. *J. Comput.-Aided Molecular Design* **1993**, *7*, 127−153.
(24) Johnson, A. P. *Challenges and Progress in Structure-Based Ligand Design*; 214th ACS National Meeting, Las Vegas, NV, September 7−11, 1997; American Chemical Society: Washington, DC, 1997.
(25) Bravi, G.; Green, D. V. S.; Hann, M. M.; Leach, A. R. PLUMS: a Program for the Rapid Optimization of Focused Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1441−1448.
(26) 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.
(27) Farley, C.; Raftery, A. E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *Comput. J.* **1998**, *41*, 578−588.
(28) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(29) Hudson, B. D.; Hyde, R. M.; Raha, E.; Wood, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quantum Struct.-Act. Relat.* **1996**, *15*, 285−289.
(30) Willett, P. Evaluation of Relocation Clustering Algorithms for the Automatic Classification of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 29−33.

CI0100393