Two New Criteria for Choosing Sample Size in Combinatorial Chemistry

Peng-Liang Zhao,* Robert B. Nachbar, James A. Bolognese, and Kevin Chapman

Merck Research Laboratories, P.O. Box 2000 Rahway, New Jersey 07065

Received October 6, 1995

Combinatorial chemistry has gained wide appeal as a technique for generating molecular diversity.[1,2] Among the many combinatorial protocols, the "split/recombine" method is quite popular and particularly efficient at generating large libraries of compounds.[3−7] In this process, polymer beads are equally divided into a series of pools, and each pool is treated with a unique fragment. The beads are recombined, mixed to uniformity, and redivided equally into a new series of pools for the subsequent couplings (see Figure 1). When a mixture of beads is separated into groups, a statistical sampling is introduced. Thus if every compound in a combinatorial library is to be represented, the number of beads at the start of the experiment must exceed the number of compounds expected. How many beads are required? Depending on the methodology used for identifying activity from a combinatorial library, different numbers of beads will be required for execution of an experiment. Iterative resynthesis methodologies wherein the final products are cleaved from the solid supports require approximately equimolar final compound concentrations.[2] Chemical encoding strategies, on the other hand, require that each compound be represented in the library at least once if the entire library is used for one assay.[2] Here we address the problem of determining the number of polymer beads required to perform a split/recombine combinatorial synthesis for these two strategies.

Consider the standard "split/recombine" method which has m splitting and subsequant coupling steps. Step 1 splits n uniformly mixed unloaded beads into r_1 pools equally and couples each pool with only one of the r_1 fragments. Starting from step 2, each step j ($2 \le j \le m$) begins by recombining and mixing these n beads uniformly, then splits them equally into r_j pools, and finally couples each pool with only one of the r_j fragments. Figure 1 shows schematically an example for $m = 3$. The compounds generated in this way can be denoted as $\{A_{i_1}B_{i_2}...M_{i_m}\colon\ 1 \le i_1 \le r_1, 1 \le i_2 \le r_2, ..., 1 \le i_m \le r_m\}$. Thus the number of these different compounds is $R = r_1r_2...r_m$. Let us denote $X_{i_1i_2 \cdots i_m}$ = the number of obtained beads carrying compound $A_{i_1}B_{i_2}...M_{i_m}$. Ideally, all possible compounds should be present in equal amounts (i.e., $X_{i_1i_2 \cdots i_m} = (n/R)$ for all $i_1i_2...i_m$) using this strategy. Due to random splitting, uncertainty in uniform mixing, and measurement error, however, this ideal equimolar representation will not be achieved unless there is a substantial excess of beads (i.e., $n \gg R$) at the start of the experiment.[8]

The problem of choosing the number of beads required in order to cover $(1 - q)100\%$ of the desired combinatorial space (i.e., cover $(1 - q)R$ compounds of the desired R) at the 99% confidence level has been discussed.[9] Since q cannot be taken to equal zero, this

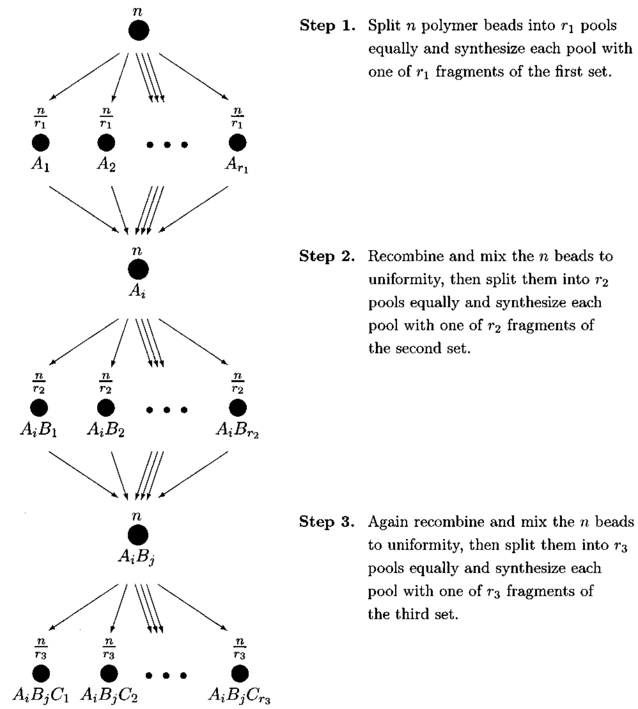

Figure 1. Split/recombine method.

criterion cannot provide statistical assurance that all compounds will be produced. We recently reported methodology for calculating the number of beads required to control the deviation from true equimolar representation in terms of overall and individual relative errors.[8] By this criterion, all compounds are adequately sampled in the combinatorial experiment. We found, however, that approaching true equimolarity requires more beads than is practical for a typical combinatorial experiment. We then asked the question: if true equimolarity is impractical, are there limits that may be placed on a combinatorial experiment such that it adequately samples all compounds in the library, and is practical to execute experimentally?

The iterative resynthesis approach to library deconvolution utilizes deductive logic to correlate chemical structure with biological activity.[2,6] In this scheme, it is assumed that all compounds be present in nearly equal amounts. Thus, for iterative resynthesis, we set out to determine the sample size n required such that, with 99% confidence, each compound will be present on at least $(1 - L)100\%$ of the ideal number of beads, (n/R), where $0<L<1$. That is, we want to find the smallest n that guarantees

$$P(X_{i_1i_2...i_m} \ge (1 - L)(n/R), \text{ for all } i_1i_2...i_m) = 0.99 \quad (1)$$

If L is taken to be close to zero, then all compounds will be present in close to equal amounts. In a practical experiment, we may require that all compounds simply be present in concentrations large enough to exert independent measurable effects on a biological assay. Thus for example, selecting $L = 0.7$ will ensure that all compounds are present to at least 30% of the ideal amount.

Chemically encoded combinatorial libraries carry structural information along with the compound on each bead.[1,2] Thus in principle, representation of each compound by a single bead is sufficient if the entire

* Corresponding author.

0022-2623/96/1839-0350$12.00/0

Table 1. Simulation Results of 500 Runs for Criterion 1 with $(1 - L) = 30\%$ Where the Total Beads n is Determined by Eq 3

						least percentage of (n/R)	
m	R	ν	$z_{((0.01)/R)}$	n (via eq 3)	n/R	minimum	1st percentile
3	10^3	972	−4.264891	36,082	36.1	33.2	38.8
4	10^4	9,963	−4.753424	459,418	45.9	34.8	39.1
5	10^5	99,954	−5.199338	5,514,424	55.1	36.2	38.9
3	20^3	7,942	−4.708130	359,279	44.9	33.4	37.8
4	20^4	159,923	−5.286029	9,119,562	57.0	36.8	38.6

Table 2. Simulation Results of 500 Runs for Criterion 2 with $K = 4$ Where the Total Beads n is Determined by Eq 4

						least number of beads	
m	R	ν	$z_{((0.01)/R)}$	n (via eq 4)	n/R	minimum	1st percentile
3	10^3	972	−4.264891	23,294	23.3	4	6
4	10^4	9,963	−4.753424	281,923	28.2	6	7
5	10^5	99,954	−5.199338	3,274,584	32.7	6	7.5
3	20^3	7,942	−4.708130	221,446	27.7	6	7
4	20^4	159,923	−5.286029	5,385,806	33.7	6	7.5

library is committed to a single assay. In practice, representation by more than one bead is desirable. Thus for chemically encoded libraries we set out to determine the sample size n required such that, with 99% confidence, each compound will be represented by at least K beads. That is, we want to find the smallest n that ensures

$$P(X_{i_1 i_2 \ldots i_m} \geq K, \text{ for all } i_1 i_2 \ldots i_m) = 0.99 \qquad (2)$$

The exact required sample size formulas for criteria 1 and 2 depend on the distribution of the smallest among all $X_{i_1 i_2 \cdots i_m}$. Since the distribution of the smallest among all $X_{i_1 i_2 \cdots i_m}$ is very difficult to obtain, the exact required sample size formulas for criteria 1 and 2 are difficult to derive. Here conservative sample size formulas for criteria 1 and 2 are developed by using the asymptotic marginal distribution of $X_{i_1 i_2 \cdots i_m}$ (assuming that $n \rightarrow \infty$ and R is fixed) and the Bonferroni method (see supporting information). The (conservative) required sample size n that guarantees eq 1 is

$$n = \nu\, z^2_{((0.01)/R)}/L^2 \qquad (3)$$

where $\nu = ((r_1 r_2 \ldots r_m) - (r_1 + r_2 + \ldots + r_m) + (m - 1))$. The (conservative) required sample size n that ensures eq 2 is

$$n = (1/4)[-z_{((0.01)/R)}\sqrt{\nu} + (\nu\, z^2_{((0.01)/R)} + 4R(K - 1))^{1/2}]^2 \qquad (4)$$

Note that $z_{((0.01)/R)}$ is the $100((0.01)/R)$th percentile of the standard normal distribution and can be obtained by using $p = ((0.01)/R)$ in eq 10 of ref 8 or using the statistical software S-PLUS.

Now our results can be stated as follows. The smallest n determined by eq 3 is the (conservative) required number of beads such that each compound will attain at least $(1 - L)100\%$ of the expected amount (n/R) at the 99% confidence level. The smallest n given in eq 4 is the (conservative) required number of beads such that each compound will be present on at least K beads at the 99% confidence level.

Monte Carlo simulations were performed to evaluate the sample size determination procedures 3 and 4 for criteria 1 and 2, respectively. The n's indicated by eqs 3 and 4 and the simulation results for several examples of m and R are listed in Tables 1 and 2. In both tables, m is the number of splitting and subsequant coupling steps used in the standard "split/recombine" synthesis and $R = r_1 r_2 \ldots r_m$ is the total number of generated compounds. Note that r_j is the number of pools in the jth splitting step ($1 \leq j \leq m$, see Figure 1 for $m = 3$). Since here we take $r_1 = r_2 = \ldots = r_m = r$, we have $R = r^m$ and $\nu = r^m - mr + (m - 1)$. For example, if we apply the standard "split/recombine" method with three splitting and subsequant coupling steps to the 20 natural amino acids to create tripeptides, then $m = 3$, $r_1 = r_2 = r_3 = 20$, $R = 8000$, and $\nu = 7942$. The $100((0.01)/R)$th percentile $z_{((0.01)/R)}$ of the standard normal distribution given in both tables is obtained from the statistical software S-PLUS.

For criterion 1, we took $(1 - L) = 0.3$ and then obtained the required total number of beads n from eq 3. With this n, all compounds are expected to attain at least 30% of (n/R) at the 99% confidence level. In Table 1, the least percentage of (n/R) for each run is the smallest value among all obtained $X_{i_1 i_2 \cdots i_m}$ divided by (n/R). The minimum and the first percentile of the least percentage of (n/R) were computed from 500 runs of simulation. The first percentile should be close to the target $(1 - L) = 30\%$ because we set the confidence level as 99%. Table 1 indicates that both the minimum and the first percentile are larger, but reasonably larger, than the target $(1 - L) = 30\%$. This is purely due to the fact that the total beads n determined by eq 3 is a conservative solution for criterion 1.

Similarly, for criterion 2, we took $K = 4$ and then obtained the required total beads n from eq 4. With this n, all compounds are expected to be present on at least $K = 4$ beads at the 99% confidence level. In Table 2, the least number of beads in each run was the smallest value among all obtained $X_{i_1 i_2 \cdots i_m}$. The minimum and the first percentile of the least number of beads were obtained from 500 runs of simulation. The results of Table 2 show that the minimum and the first percentile are just slightly larger than the target $K = 4$. Thus, the total beads n determined by eq 4 is indeed a conservative solution for criterion 2.

Since eqs 3 and 4 are conservative solutions for criteria 1 and 2, respectively, a reviewer asked whether one can use only 80% of the beads calculated by eq 3 or eq 4 and still have the first percentile of the least number of beads above $(1 - L)100\%$ of (n/R) or above K. Our simulation results listed in Tables 3 and 4

Table 3. Simulation Results of 500 Runs for Criterion 1 with $(1 - L) = 30\%$ Using $n^* = 0.8n$ Beads, Where n is Determined by Eq 3

m	R	ν	$z_{((0.01)/R)}$	$n^* = 0.8n$ (n is via eq 3)	n^*/R	least percentage of (n^*/R) minimum	least percentage of (n^*/R) 1st percentile
3	10^3	972	−4.264891	28,866	28.9	27.7	31.1
4	10^4	9,963	−4.753424	367,534	36.8	27.2	32.6
5	10^5	99,954	−5.199338	4,411,539	44.1	31.7	34.0
3	20^3	7,942	−4.708130	287,423	35.9	27.8	32.5
4	20^4	159,923	−5.286029	7,295,650	45.6	28.5	32.8

Table 4. Simulation Results of 500 Runs for Criterion 2 with $K = 4$ Using $n^* = 0.8n$ Beads, Where n is Determined by Eq 4

m	R	ν	$z_{((0.01)/R)}$	$n^* = 0.8n$ (n is via eq 4)	n^*/R	least number of beads minimum	least number of beads 1st percentile
3	10^3	972	−4.264891	18,635	18.6	3	3.5
4	10^4	9,963	−4.753424	225,538	22.6	3	4
5	10^5	99,954	−5.199338	2,619,667	26.2	4	5
3	20^3	7,942	−4.708130	177,157	22.1	3	4
4	20^4	159,923	−5.286029	4,308,645	26.9	4	4

indicate that for $(1 - L)100\% = 30\%$ and $K = 4$, using 80% of the calculated beads almost still meets criteria 1 and 2, but using 75% or 70% of the calculated beads probably will not.

In deriving the sample size formulas and performing the simulations, we have assumed that the bead splitting was perfect. In such situation, 80% of the calculated number of beads are almost adequate for coverage under criteria 1 and 2. In reality, however, there are experimental errors in bead splitting (perhaps 2−5%) and the number of beads required for adequate coverage will be somewhat higher.

In Tables 1 and 2, we also listed the required multiplicative factor n/R (i.e., the average number of beads per compound). The required multiplicative factor changes rather dramatically depending upon the degree of statistical protection desired (i.e., the choice of L in eq 1 or K in eq 2) and which criterion is used. It is clear that for both criteria, as the number of the distinct compounds increases, the required multiplicative factor increases.

Iterative resynthesis and chemical encoding place different constraints on the number of beads needed to ensure adequate representation of each chemical member of a combinatorial library. Criterion 1 provides a simple method for determining the number of beads required to execute a combinatorial synthesis such that all compounds will be within a defined percentage of the ideal. Criterion 2 provides a method for determining how many beads are required for an encoded combinatorial library. These methods should be useful for the practical execution of combinatorial experiments.

Acknowledgment. We thank the referees and Dr. Simon Kearsley for their thorough reviews and helpful comments.

Supporting Information Available: Outline of the derivation of eqs 3 and 4, and the simulation program (19 pages). Ordering information is given on any current masthead page. [It should be noted that if the encoded library is itself split and screened in many assays, an additional statistical sampling has taken place and the situation is much more complex. The methods described here work well for the standard split/recombine strategy. Complex variations of the standard split/recombine strategy are best modelled by computer simulations.]

References

(1) Gallop, M. A.; Barret, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251.

(2) Gordon, E. M.; Barret, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(3) Furka, A.; Sebestyen, F.; Asgedom, M.; Dibo, G. General Method for Rapid Synthesis of Multicomponent Peptide Mixtures. *Int. J. Pept. Protein Res.* **1991**, *37*, 487−493.

(4) Sebestyen, F.; Dibo, G.; Kovacs, A.; Furka, A. Chemical Synthesis of Peptide Libraries. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 413−418.

(5) Lam, K. S.; Salmon, S. E.; Hersh, E. M.; Hruby, V. J.; Kazmierski, W. M.; Knapp, R. J. A New Type of Synthetic Peptide Library for Identifying Ligand-binding Activity. *Nature* **1991**, *354*, 82−84.

(6) Houghten, R. A.; Pinilla, C.; Blondelle, S. E.; Appel, J. R.; Dooley, C. T.; Cuervo, J. H. Generation and Use of Synthetic Peptide Combinatorial Libraries for Basic Research and Drug Discovery. *Nature* **1991**, *354*, 84−86.

(7) Ohlmeyer, M. H. J.; Swanson, R. N.; Dillard, L. W.; Reader, J. C.; Asouline, G.; Kobayashi, R.; Wigler, M.; Still, W. C. Complex Synthetic Chemical Libraries Indexed with Molecular Tags. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 10922−10926.

(8) Zhao, P.-L.; Zambias, R. A.; Bolognese, J. A.; Boulton, D. A.; Chapman, K. Sample Size Determination in Combinatorial Chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 10212−10216.

(9) Burgess, K.; Liaw, A. I.; Wang, N. Combinatorial Technologies Involving Reiterative Division/Coupling/Recombination: Statistical Considerations. *J. Med. Chem.* **1994**, *37*, 2985−2987.

(10) Knuth, D. E. *The Art of Computer Programming: Seminumerical Algorithms*, 2nd ed.; Addison-Wesley: 1981; Vol. 2, p 139.

JM950054X