JOURNAL OF
MEDICINAL
CHEMISTRY

© Copyright 1994 by the American Chemical Society

Volume 37, Number 19 September 16, 1994

Communications to the Editor

Combinatorial Technologies Involving Reiterative Division/Coupling/ Recombination: Statistical Considerations

Kevin Burgess*

Department of Chemistry, Texas A & M University, College Station, Texas 77843

Andy I. Liaw and Naisyin Wang

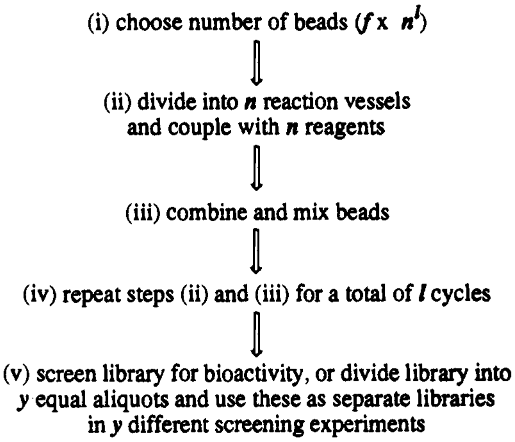
Department of Statistics, Texas A & M University, College Station, Texas 77843

Received June 28, 1994

Biological and chemical methods for production and screening of huge arrays of chemical structures provide an exciting alternative to synthesis and testing of compounds sequentially.[1-5] Notable among these are reiterative division/coupling/recombination syntheses for producing arrays of beads (or other units of unique solid support sites) each displaying single, but different, organic molecules (Scheme 1).[6-8] The beads are screened in one reaction vessel with the substrates on the solid phase, or "plated out" into small groups (*e.g.*, each of 10 beads), so that the substrates can be liberated into solution for testing. Such technologies facilitate convenient, rapid, screening for biologically significant intermolecular interactions.[9]

The chemical literature offers no guidelines on the number of beads that should be used to prepare reiterative division/coupling/recombination libraries. Statistical considerations are critical in this regard, but they are also nontrivial. There are three important factors in this type of experiment (see Scheme 1 for some definitions of algebraic terms): (i) the number of possibilities (*i.e.*, the extent of the "combinatorial space", n^l), (ii) the fraction of that number that is likely to be produced given the random nature of the experiment (*i.e.*, $1 - q$, where q is the fraction of the combinatorial space that is likely to be missed), and (iii) the level of confidence with which the coverage of combinatorial space can be asserted (*i.e.*, the degree of certainty associated with this value q, expressed as $1 - p$, where p is the fraction of uncertainty). This paper relates the number of beads used to the extent of combinatorial space and the variables p and q. For the purposes of this analysis, it is assumed that all the coupling reactions are 100% efficient; hence, each possible combination of molecular entities is equally likely to form.

Scheme 1. Outline of Typical Divide, Couple, and Recombine Technologies



(i) choose number of beads (f x n^l)

(ii) divide into n reaction vessels and couple with n reagents

(iii) combine and mix beads

(iv) repeat steps (ii) and (iii) for a total of l cycles

(v) screen library for bioactivity, or divide library into y equal aliquots and use these as separate libraries in y different screening experiments

Two ways were chosen to address the statistical problem outlined above, and the first of those was via computer simulations. In every experiment the number of beads used should be greater than the extent of combinatorial space (*i.e.* > n^l) by some "multiplicative factor", f. A FORTRAN program was devised to simulate the library, then to narrow down the range of possible f values using a trial-and-error approach (see supplementary material). Table 1 displays f values simulated (*i.e.*, f_{sim}) for a confidence level of 99 % (*i.e.*, $1 - p = 0.99$). The data are expressed at three different levels of coverage (where coverage = $\{1 - q\} \times 100\%$). Thus, for a library formed by a divide, couple, and combine sequence involving 10 different reagents ($n = 10$) and five cycles ($l = 5$), it was shown that 95% (*i.e.*, $q = 0.05$) of the combinatorial space (*i.e.*, 0.95×10^5) was covered with 99% confidence if a multiplicative factor of 3.026 were applied (*i.e.*, 3.026×10^5 beads were used).

This simulation method provides a statistically ac-

0022-2623/94/1837-2985$04.50/0 © 1994 American Chemical Society

Table 1. Computer Simulated (sim) and Calculated (calc) f Factors for a Library of 10 Reactions (*ie* $n = 10$) Applied for Three to Five Cycles of Divide, React, and Combine (*ie* $l = 3-5$) at the 99% Confidence Level (*ie* $p = 0.01$)[a]

	factor f (where no. beads required $= f \times n^l$)		
% coverage [$\{1 - q\} \times 100\%$]	$l = 3$ (sim/calc)	$l = 4$ (sim/calc)	$l = 5$ (sim/calc)
90	2.470/2.525	2.370/2.372	2.322/2.324
95	3.270/3.317	3.090/3.097	3.026/3.027
99	5.300/5.323	4.840/4.840	4.678/4.678

[a] Definitions: f = the multiplicative factor, the product of this factor and the combinatorial space (n^l) represents the number of beads to be used; n = number of reagents; l = number of reiterative cycles of divide, couple, and combine; q = acceptable fraction of combinatorial space likely to be missed; p = acceptable fraction of uncertainty. The percent coverage is defined as $\{1 - q\} \times 100\%$.

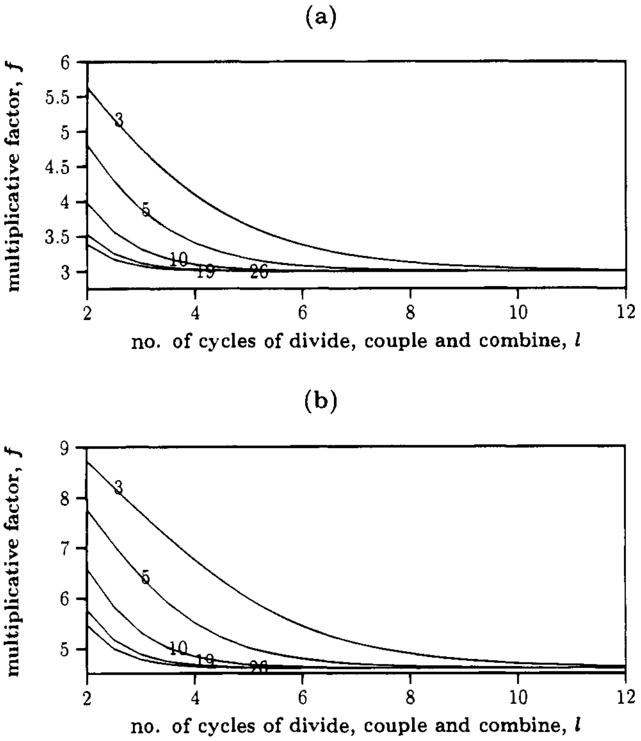


Figure 1. Multiplicative factors, f_{calc}, calculated from eq 1, and expressed as functions of the number of cycles of divide, couple, and combine, l, for (a) 95% coverage {*i.e.*, $(1 - q) \times 100\% = 95\%$} and (b) 99% coverage. The curves, from top to bottom, correspond to n = 3, 5, 10, 19, and 26, respectively, where n is the number of reagents.

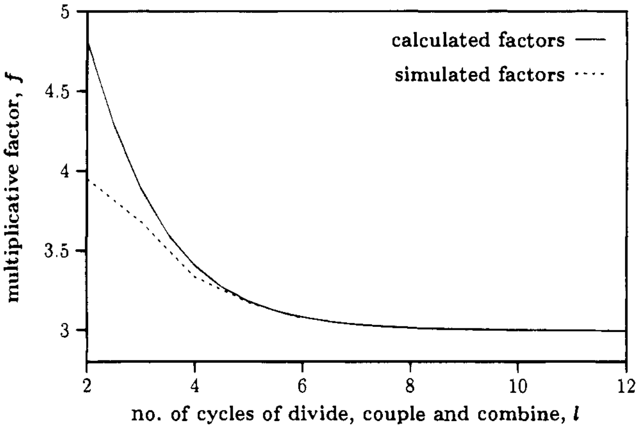


Figure 2. Simulated and calculated multiplicative factors, f_{sim} and f_{calc}, as functions of the number of cycles of divide, couple, and combine, l. The number of reagent n in each cycle was five, the coverage $\{(1 - q) \times 100\%\}$ was 95%, the fractional uncertainty p was 0.01.

curate representation of library composition, but the volume of data that must be stored makes it computationally expensive. This issue became important, for instance, when simulations were attempted for libraries involving 19 different reagents in each cycle. Thus it was possible to calculate that $f = 2.370$ for a library for which $l = 3$, $n = 19$, $q = 0.10$ (*i.e.*, 2.370×19^3 beads would be required to cover 90% of combinatorial space with 99% confidence), and that the f value was only slightly less ($f = 2.32$) for an analogous library involving four cycles ($l = 4$). However, the corresponding calculation for a library formed by using five cycles became too computationally expensive to be handled conveniently without a supercomputing resource.

Relatively large libraries are important in combinatorial chemistry. For example, a typical objective might be to prepare a library of all possible pentapeptides using all the protein amino acids except Cys, for which the combinatorial space is 19^5. The limitations of computer simulations therefore led us to explore an algebraic approximation that could be used for large libraries. Consequently, eq 1 was developed to approximate the reiterative divide and combine experiment for libraries involving much combinatorial space.[10] Briefly, eq 1 was derived using an asymptotic ap-

$$f = -\log\left\{\frac{2q + h - \sqrt{(2q + h)^2 - 4(1 + h)q^2}}{2(1 + h)}\right\} \quad (1)$$

proximation of the joint distribution function of the observed numbers for each sequence (see supplementary material). The value of h in equation 1 is $z_p{}^2/n^l$, and z_p is the "$(1 - p)$ quantile of standard normal distribution" which can be found by locating $(1 - p)$ in "standard normal probability tables", which can be found in almost any collection of standard statistical tables.[11]

Table 1 compares the f values simulated computationally, with values calculated using eq 1. The match is good for these relatively small libraries, and better for larger ones. For example, in another calculation it was shown that for a library for which $n = 19$, $l = 3$, and $q = p = 0.01$, the simulated and calculated f values are $f_{\mathrm{sim}} = 4.880$, and $f_{\mathrm{calc}} = 4.884$, respectively. Significantly, when the number of reagents and/or the number of cycles is large, the value of h approaches 0 and eq 1 collapses to eq 2. Consequently, for very large libraries

$$f = -\log\{q\} \quad (2)$$

the factor f is simply related to the required degree of coverage ($1 - q$; hence, $f = 2.9957$ for 95% coverage, and 4.6052 for 99%). This can also be seen from Figure 1: the factors f approach a minimum value when the extent of combinatorial space is large (an f value of *ca.* 3.0 in Figure 1a, and *ca.* 4.6 in Figure 1b). In general terms, the data in Figure 1 shows that the multiplicative factor f is relatively high for smaller libraries (small numbers of reagents n and/or cycles l). For larger libraries f converges to a minimum value which is then governed by the required level of coverage and degree of confidence (functions of q and p, respectively).

Figure 2 shows f_{sim} and f_{calc} for a library for which $n = 5$, $q = 0.05$, and $p = 0.01$. This graph shows eq 1 does not give an accurate approximation for very small libraries. In fact, the calculated multiplicative factor

f_{calc} is always higher than the "true" value f_{sim}. The difference between f_{calc} and f_{sim} is a function of n^l and only becomes significant when this is a relatively small value. Consequently, simulation via computation is the preferred approach for small libraries, but eq 1 is still useful for calculating a conservative value of the absolute minimum multiplicative factor that should be applied.

Finally, we used computer simulations to analyze the situation in which a library is prepared then split into y equal portions to use in y different experiments. This scenario is of interest because when many libraries are to be tested they might be prepared in one big batch, then split. If no coverage is lost in the final division, $y \times f \times n^l$ beads will be required. Conversely, if some coverage is lost then the number of beads required will be $c \times y \times f \times n^l$, where c is come correction factor greater than unity. The simulated c values were 1.003 for three libraries for which $n = l = 5$, $q = 0.05$, $p = 0.01$, and $y = 2$, 3, or 4 ($f_{\text{sim}} = 3.152$ throughout). Thus it seems that for combinatorial space of 5^5 or more, only marginally more beads are required to maintain the same percent coverage at the same confidence level when several libraries are prepared in one operation.

Acknowledgment. K.B. thanks The National Institutes of Health for supporting this work (GM50772-01A1) and for an NIH Career Development Award; he also thanks The Alfred P. Sloan Foundation for a scholarship and The Robert A. Welch Foundation. N.W. thanks The National Cancer Institute (CA57030) and The Texas Advanced Research Program (010366-051) for funding.

Supplementary Material Available: The FORTRAN programs used to simulate the libraries and key steps in the derivation of eq 1 (8 pages). Ordering information is given on any current masthead page.

References

(1) Zuckerman, R. N. The Chemical Synthesis of Peptidomimetic Libraries. *Curr. Op. Struct. Biol.* **1993**, *3*, 580.

(2) Geysen, H. M.; Mason, T. J. Screening Chemically Synthesized Peptide Libraries for Biologically-Relevant Molecules. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 397.

(3) Pavia, M. R.; Sawyer, T. K.; Moos, W. H. The Generation of Molecular Diversity. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 387.

(4) Gallop, M. A.; Barrett, 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.

(5) Gordon, E. M.; Barrett, 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.

(6) 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.

(7) Lam, K. S.; Hruby, V. J.; Lebl, M.; Knapp, R. J.; Kazmierski, W. M.; Hersh, E. M.; Salmon, S. E. The Chemical Synthesis of Large Random Peptide Libraries and Their Use for the Discovery of Ligands for Macromolecular Acceptors. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 419.

(8) Furka, A.; Sevestyen, F.; Asgedom, M.; Dibo, G. General Method for Rapid Synthesis of Multicomponent Peptide Mixtures. *Int. J. Peptide Res.* **1991**, *37*, 487.

(9) Bunin, B. A.; Ellman, J. A. A General and Expedient Method for the Solid-Phase Synthesis of 1,4-Benzodiazepine Derivatives. *J. Am. Chem. Soc.* **1992**, *114*, 10997.

(10) This equation is also only valid for $p < 0.5$.

(11) Ott, L. In *An Introduction to Statistical Methods and Data*; Duxbury Press: Belmont, CA, 1992. The log values are natural logarithms throughout. For example, to locate z_p for $p = 0.01$, $1 - p = 0.99$: look for the closest value to 0.99 in the standard normal probability table, which in this case is 0.9901; this is in row 2.3 and column 0.03 hence z_p is $2.3 + 0.03 = 2.33$. The z_p for $p = 0.05$ is exactly between 1.64 and 1.65, so the value usually taken is 1.645.