RESEARCH

Random Subcloning

Jared C. Roach

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195

Random subcloning strategies are commonly employed for analyzing pieces of DNA that are too large for direct analysis. Such strategies are applicable to gene finding, physical mapping, and DNA sequencing. Random subcloning refers to the generation of many small, directly analyzable fragments of DNA that represent random fragments of a larger whole, such as a genome. Following analysis of these fragments, a map or sequence of the original target may be reconstructed. Mathematical modeling is useful in planning such strategies and in providing a reference for their evaluation, both during execution and following completion. The statistical theory necessary for constructing these models has been developed independently over the last century. This paper brings this theory together into a statistical model for random subcloning strategies. This mathematical model retains its utility even at high subclone redundancies, which are necessary for project completion. The discussion here centers on shotgun sequencing, a random subcloning strategy envisioned as the method of choice for sequencing the human genome.

Random subcloning is a common tool for large-scale physical mapping and DNA sequencing. Large DNA targets are intractable to direct analysis and must be broken down into smaller fragments before techniques such as restriction mapping or sequencing can be employed. A map or sequence of the original target can be deduced following analysis of the derived fragments, termed subclones. Ideally, a direct strategy is pursued by analyzing a minimum number of fragments such that a shortest tiling path is followed. This requires prior knowledge of the relation of each fragment to the original target. However, such information is not necessarily or easily available. Thus, many strategies resort to picking and analyzing fragments at random.

In random subcloning strategies, fragments are generated from a vast number of identical target sequences, so the resulting library from which they are selected for further analysis is redundant. Therefore, individual fragments may overlap in the sense that they mutually possess some bit of target sequence. The presence of such overlaps allows retrospective determination of which fragments represent adjacent target sequences. When enough overlapping fragments have been analyzed, the original sequence or map may be deduced (Fig. 1).

From a theoretical standpoint, random mapping and sequencing strategies can be treated identically, albeit on a different scale. Most phys-

ical mapping projects employ targets on the megabase scale or larger. Such targets are randomly fragmented into yeast artificial chromosome (YAC), bacterial artificial chromosome (BAC), cosmid, or phage subclones ranging in size from tens of kilobases to several megabases. Analysis techniques include restriction mapping, sequence-tagged site (STS) content mapping, in situ hybridization, and many others. Sequencing projects employ both smaller targets and smaller subclones. In particular, the targets of sequencing projects are often the subclones of mapping projects. Fragments of these subclones (i.e., subclones of subclones) are small enough to be employed as sequencing templates for automated DNA sequencing machines. The effective fragment size for sequencing projects is the sequence read length, which is the amount of sequence that can be read from one fragment by a sequencing machine. This length currently ranges from several hundred to 1000 bases.

The appeal of random subcloning strategies lies in the absence of need for prior information about particular subclones. This allows projects to be undertaken with a great deal of automation and with a decreased need for highly trained human intervention. The drawback of such "shotgun" strategies is their dependence on overdetermination of information, with a need to generate several times as much raw data as an ideal directed strategy would. Accordingly, actual strategies may be a mix of both random and directed approaches, beginning with random and pro-

E-MAIL roach@u.washington.edu; FAX (206)685-7301.

RANDOM SUBCLONING

Figure 1 An example of a shotgun strategy, with a bar code length employed as an analogy to a DNA sequence. A large number of identical but unknown target sequences are randomly fragmented. These fragments are analyzed and aligned based on unique overlapping sequences. When enough fragments have been analyzed, the original target sequence may be deduced. The number of fragments is typically much larger than depicted here.

gressing to directed when the cost of choosing and sequencing directed subclones is judged to be less than the cost of continued shotgunning. Such decisions are predicated on the ability to determine such costs. Experience, simulations, and analytical models are the tools for this analysis.

A useful analytic model of the random shotgun strategy has been developed previously and is in common use (Lander and Waterman 1988). This model is accurate at low subclone redundancies. However, most sequencing projects and many mapping projects employ higher redundancies, which are necessary to approach complete coverage and closure of targets such as cosmids and BACs (for sequencing) or the human genome (for mapping), where closure is defined as the absence of gaps in the knowledge of the target sequence.

The present analysis initially specifies the distribution of the lengths of the spacings between fragment start sites. With this distribution, the number of gaps can be determined simply by

counting the number of spacings greater than the length of a fragment. All orderings of these spacings among spacings not forming gaps are equally likely, and because spacing lengths are independent of their ordering, this permits the determination of the distribution of the number of clones in an island as well as island length. Extensions permit additional probabilities of interest to be calculated.

Formulation

A linear discrete target of length G , such as a genome, is assumed. For a given project, n fragments of constant length L are generated from the target and analyzed in a manner in which overlaps between fragments are detectable. All fragments are generated from distinct identical copies of G . No fragments may start within $L-1$ bases of the last, right-most base of G , as such fragments would not be contained entirely within G . Thus, the effective length G_e available for fragment start sites is $G - L + 1$. The starting, or left-most, base pair of each fragment is designated S_k , such that S_1 is the start site of the left-most fragment, with $S_k \in [1, G_e]$. S_n begins the right-most, or last fragment of G . The start site may be either the 5' or 3' base pair of the Crick strand of the fragment it begins, depending on fragment orientation relative to the target. In this model the S_k are an ordered sample of n independently, identically, and uniformly distributed observations on the interval $(0, G_e)$. The formulation is drawn schematically in Figure 2. Let $D_k = S_{k+1} - S_k$ represent the distance between start sites, for $k = 1, 2, \dots, n-1$. D_0 represents the length of the uncovered target region before the first fragment, and equals S_1 . D_n is the distance $G_e - S_n$. An assumption is made that an overlap of length of at least T is necessary and sufficient to detect adjacency of two fragments. Redundancy, R , is defined as nL/G . For notational ease, the effective fractional coverage f_G of the target provided by one fragment is defined as $(L - T)/G_e$, and the effective redundancy R_e is defined as nf_G . The notation and symbols used here are summarized in Table 1.

By genomic conventions, an island is a maximal set of fragments, each of which is connected to all other island members by one or more paths of fragments overlapping by T . A contig is an island consisting of at least two fragments (Sta-

ROACH

Figure 2 Schema of notation.

spacing lengths is ignored. However, because n is universally large for genome projects, this correlation approaches zero. With these considerations, the density function for the lengths of spacings between start sites is

den 1980). In general, a target region not covered in any fragment is a gap. Adjacent islands are thus separated by gaps. The length of a gap is $D_k - L$, for $D_k > L - T$. A negative gap length indicates that an overlap is present, but not detected, so this is still considered a gap.

A simple geometric observation underlies many of the equations presented here: The domain space of the spacings D_k is the surface of the simplex $D_0 + D_1 + D_2 + \dots + D_n = G_e$, and their joint probability density is constant. This observation permits many probabilities of interest to be calculated by geometric considerations (Levy 1939). Here, G_e and D_k will usually be treated as continuous rather than discrete. This approximation is quite minor, given the scope of a genome, or even a cosmid, compared with the unit of divisibility, a base pair. Thus, the normalized length of a single spacing is described by a beta distribution, which is defined on the interval $[0,1]$ and is related here to the effective target length by the factor G_e . By employing the beta distribution in this case, the correlation between

$$f_{D_k}(x) = n \left(1 - \frac{x}{G_e}\right)^{n-1} \quad (1)$$

$$\text{and } E(D_k) = \frac{G_e}{n+1}.$$

Distribution of the Number of Gaps and Islands in a Project

A gap will occur following a fragment starting at S_k if and only if $D_k > L - T$. A spacing less than this will result in a detectable overlap between the two fragments; a greater spacing will result in no overlap, or an undetectable one. Thus, the probability of a gap following a given fragment is equivalent to the probability that $D_k > L - T$ (note the change of variable $y = x/G_e$):

$$\begin{aligned} P_{\text{gap}} &\approx \int_{L-T}^{G_e} f_{D_k}(x) dx = \int_{L-T}^{G_e} n \left(1 - \frac{x}{G_e}\right)^{n-1} dx \\ &= n \int_{f_G}^1 (1-y)^{n-1} dy = (1 - f_G)^n \end{aligned} \quad (2)$$

Table 1. Notation and Symbols

G , target length
$G_e = G - L + 1$, effective target length
L , fragment length
n , number of fragments in a project
T , bases necessary to determine overlap
$R = \frac{nL}{G}$, redundancy
$R_e = \frac{n(L-T)}{G_e}$, effective redundancy
$f_G = \frac{L-T}{G_e}$ effective fractional target coverage per fragment
S_k , starting base of the k th leftmost fragment
$D_k = S_{k+1} - S_k$ spacing between adjacent fragment start sites
z_{mv} number of fragments contained within the m th leftmost island
l_{mv} length of the m th leftmost island

Employing again the assumption that the lengths of the spacings between each S_k are independent, the distribution for the total number of gaps in a project is binomial. Again, this assumption is reasonable when there are a large number of spacings, the usual case for genome projects. In a given project there are $n - 1$ opportunities for a gap to occur, one between each pair of adjacent fragment start sites. Thus,

$$P(N_{gaps} = x) \approx \binom{n-1}{x} p_{gap}^x (1 - p_{gap})^{n-1-x} \quad (3)$$

$$\approx \binom{n-1}{x} (1 - f_G)^{nx} [1 - (1 - f_G)^n]^{n-1-x}$$

One immediately has the probability of project closure as

$$P(N_{gaps} = 0) \approx (1 - p_{gap})^{n-1} \approx [1 - (1 - f_G)^n]^{n-1} \quad (4)$$

Also, the expected number of gaps in a project is

$$E(N_{gaps}) \approx (n - 1)(1 - f_G)^n \quad (5)$$

As noted above, the distribution for the number of gaps can be made exact by summing appropriate areas of $n + 1$ dimensional simplex. This somewhat awkward but nevertheless elegant distribution is provided by Stevens (1939) and in slightly different form by Flatto and Konheim (1962). Stevens' distribution is approximated by equation 3. Siegel (1979) provides an alternate derivation of equation 4 in slightly different form.

Not counting the two ends of the target as gaps, the number of islands will be one greater than the number of gaps, as each gap separates two adjacent islands. So by definition,

$$N_{islands} = N_{gaps} + 1, \text{ and} \quad (6)$$

$$E(N_{islands}) \approx 1 + (n - 1)(1 - f_G)^n$$

Distribution of the Number of Clones in an Island

The total number of fragments in a project is n , and the total number of islands is $N_{islands}$, so the expected number of fragments z_m in an arbitrary island is clearly $E(z_m | N_{islands}) = n / N_{islands}$. To obtain the probability distribution of z_m one may divide the spacings $\{D_k | k = 1, 2, \dots, n - 1\}$ into two subsets: those spacings $D_k > L - T$ that are

long and contain a gap, and those $D_k \leq L - T$ that are short and do not. The number of spacings in the first subset is N_{gaps} . By elimination, the number in the second subset is $n - 1 - N_{gaps}$. The number of fragments in an island is equal to one plus the number of short spacings between its two bounding long spacings. The last island might end with a short spacing, which would reduce its number of fragments by one. This minor effect is ignored here, but it may be accounted for, if desired, at the cost of a little algebra. Now all orderings of long and short spacings are equally likely, as the D_k are exchangeable. The probability distribution for z_m can be analyzed combinatorically (see approaches to similar problems by Whitworth 1897a; also see Baticle 1935), but to maintain simplicity a continuous approximation analogous to that employed previously to model spacing length may be employed (equation 1). In short, this determines the probability that no long spacings occur before the z_m th spacing. If there are enough short spacings, z_m may be treated as a continuous variable. This approximation of continuity is valid when n is large, the usual case for genome projects. Employing a uniform distribution of gaps over the continuous interval $[0, n - 1 - N_{gaps}]$ and scaling by a factor of $n - 1 - N_{gaps}$, a beta distribution is obtained as before. The conditional probability density for z_m is therefore

$$f_z[x | N_{gaps}] \approx N_{gaps} \left(1 - \frac{x - 1}{n - 1 - N_{gaps}} \right)^{N_{gaps} - 1} \quad (7)$$

Note that the number of short spacings in an island is beta distributed. An island will contain one additional spacing (and a fragment to go with it) because of its terminating long spacing. So with z_m the number of fragments in an island, $z_m - 1$ is the number of short spacings in that island. The expected number of clones in an arbitrary island (conditioned on the number of gaps) is, as expected,

$$E(z_m | N_{gaps}) = \frac{n - 1 - N_{gaps}}{N_{gaps} + 1} + 1 = \frac{n}{N_{islands}} \quad (8)$$

The fraction of singleton islands expected in a project can be obtained by integrating the probability density in equation 7 over the range $x \in [1, 2]$; the remaining islands will be contigs. Also if desired, the necessity for conditioning on N_{gaps} can be dropped by performing a weighted summation over all possible values of N_{gaps} .

The distribution of the number of clones in

ROACH

an island enables the determination of the distribution of the length of that island. Some motivation also exists to predict the length of the longest island resulting from a project, as it is a readily identifiable feature of a work in progress. In particular, a failure to achieve islands of predicted length is often an indication of a technical inability to detect overlaps, and thus points to a problem that needs to be addressed. Whitworth (1897b) shows that for a given project, if the island are ordered by increasing number of fragments, the expected number of fragments in the x th island is

$$E(\text{no. of fragments in } x\text{th smallest island} | N_{\text{gaps}}) = 1 + \frac{n-1-N_{\text{gaps}}}{N_{\text{gaps}}} \sum_{i=1}^x \frac{1}{N_{\text{gaps}} - i + 1} \quad (9)$$

This expected value may be substituted in equation 11 below, and enables the prediction of the longest expected island for a project.

Expected Island Length

Each island is the union of one or more fragments starting at base pairs $S_k, S_{k+1}, S_{k+2}, \dots$, and S_{K+z_m-1} . The total length l_m of an island with S_k beginning its first fragment is the sum of the spacings between its fragment start sites plus the entire length of the last fragment in the island (Fig. 2):

$$l_m = \begin{cases} L + \sum_{k=1}^{K+z_m-1} D_k & \text{if } z_m > 1 \\ L & \text{if } z_m = 1 \end{cases} \quad (10)$$

Spacings are exchangeable in that the joint distribution of all D_k is unchanged under any permutation of subscripts. Or rephrased, the lengths of the spacings are independent of their order. Expected island length conditioned on z_m is therefore

$$E(l_m | z_m) = L + E(D_k)(z_m - 1) \quad (11)$$

And so one may approximate expected island length as

$$E(l_m) \approx L + E(D_k)(E(z_m) - 1) \quad (12)$$

$$\approx L + G_e \left(\frac{1}{n+1} \right) \left(\frac{n}{1 + (n-1)(1-f_g)^n} - 1 \right) \quad (13)$$

This approximation is most valid when the relative variance of z_m is small, the usual case for genome projects. Note that the use of $E(D_k)$ as calculated above constitutes an additional approximation, as not all spacings can be included in islands. To account for this, a modification to $E(D_k)$ must be made (Appendix A). Also, the accuracy of this approximation can be improved with the aid of a computer by summing equation 11 over all possible values of z_m , rather than employing the expected value of z_m . Based on evidence from computer simulations, however, equation 13 appears to offer enough accuracy for most purposes.

The expected fraction, f , of the target covered by fragments can be calculated directly: $f = E(l)E(N_{\text{islands}})$. This fraction can also be calculated independently (Appendix B).

Under low redundancy conditions, where $n \gg e^R$, $E(z) \approx e^R$. With this approximation $E(l) \approx [L(e^R - 1)]/R$ is obtained, identical to that of Lander and Waterman (1988) for the limit as T approaches zero. This defines an upper bound as a function of n (or R) for the accuracy of the Lander and Waterman equations. This bound occurs at a redundancy of approximately threefold. In particular, above this limit, a geometric distribution ceases to be a good model for the expected number of clones in an island. At the redundancies of six to eightfold used in common practice, geometric approximations also have the nonsensical disadvantage of predicting a fractional number of islands in conjunction with an average island length longer than the target. This arises when a significant fraction of the clones in a project are contained within a single island and the stopping probability of a geometric approximation increases. The result is that the “lack-of-memory” necessary for a Poisson analysis fails. Thus, such equations work best when a large number of islands is expected and an approximation of island length is independent.

Probability of an Island Greater Than a Critical Length

Care must be used when considering more than one island from a given project, as their lengths are not independent. For example, if there are two islands in a project, it might be that either one is greater than half the length of the target, but it is certain that both of them are not (barring undetected overlap). Now, the probability that a project contains at least i islands of length

greater than a certain critical length, C , may be considered. Such a probability finds importance in evaluating performance of fragment assembly algorithms or in planning projects with limited goals.

Let N_{long} be the number of long spacings $D_k > L - T$. Let g_m be the number of short spacings contained between adjacent long spacings. Now $l_m \approx L + E(D)g_m$. Thus,

$$P(l_m > C) \approx P\left[g_m > \frac{C - L}{E(D)}\right].$$

The number of short spacings preceding the first long spacing is g_0 ; the number following the last long spacing is $g_{N_{long}}$. Because either D_0 or D_n may be long, N_{long} may be up to two greater than N_{gaps} . N_{gaps} is nevertheless a good approximation to N_{long} . This is true at low redundancies where $N_{gaps} \gg 2$. At high redundancies $n \gg N_{long}$, so it is unlikely that either D_0 or D_n is long. If greater accuracy is desired, an appropriate summation can be made. Because the distribution of long spacings is uniform among the set of all spacings, the distribution of g is defined on the simplex

$$\frac{g_0 + g_1 + g_2 + \dots + g_{N_{long}}}{(n + 1 - N_{long})} = 1$$

with all points of the simplex equiprobable. Let R be the number of islands exceeding length C , let

$$c = \frac{C - L}{E(D)(n + 1 - N_{long})}$$

and let k be the greatest integer less than $1/c$. Then, directly from Stevens (1939) one has

$$P(R > i | N_{gaps}) \approx \sum_{j=1}^k (-1)^{j-1} \frac{(N_{gaps} + 1)!}{(N_{gaps} + 1 - j)!(j - i)!(i - 1)!} \frac{(1 - jc)^{N_{gaps}}}{j} \quad (14)$$

In particular, the probability of having a contig in a project greater than half the length of the target can be approximated (i.e., $c = 1/2$) as follows:

$$\begin{aligned} P\left(\text{one contig} > \frac{G}{2}\right) &= \sum_{v=1}^n P(N_{gaps} = v) P\left(\text{one contig} > \frac{G}{2} | N_{gaps} = v\right) \quad (15) \\ &\approx \sum_{v=1}^n \binom{n-1}{v} (1 - f_G)^v [1 - (1 - f_G)^n]^{n-1-v} \frac{v+1}{2^v} \end{aligned}$$

At moderate to high redundancies, only the first few terms of this last sum are necessary.

Circular Targets

In the foregoing, considerations of a linear target forced somewhat awkward equations or mild approximations to obtain simpler equations. With circular targets, many of these considerations are unnecessary. In particular, $G_e = G$. Modifications to equations 1 and 3 result in

$$f_{D_k}(x) = (n - 1) \left(1 - \frac{x}{G}\right)^{n-2} \quad (1')$$

$$\begin{aligned} P(N_{gaps} = x) &= P(N_{islands} = x) \\ &\approx \left(\frac{n}{x}\right) p_{gap}^x (1 - p_{gap})^{n-x} \end{aligned} \quad (3')$$

Other results follow appropriately. Note that these equations are not significantly different from those for a linear target. The literature provides an exact formula for the expected number of fragments needed for closure of a circular target (Flatto and Konheim, 1962). Begging "edge effects," this equation can also be applied to the line. It is, as T approaches zero, and where B is the greatest integer smaller than G/L ,

$E(n \text{ needed for closure}) =$

$$1 - \sum_{k=1}^B (-1)^k \frac{\left(1 - k \frac{L}{G}\right)^{k-1}}{\left(k \frac{L}{G}\right)^{k+1}} \quad (16)$$

Also, some results are available for the case of a varying parameter $L - T$, where its distribution is known (Siegel and Holst 1982). These results include the distribution for the number of gaps and its corollary, the probability of project closure. These results may be applied also as approximations to the linear case. Conversely, the linear results may also be applied as approximations to the circular case.

Simulations and Experimental Data

A large number of Monte Carlo simulations of projects can be generated quickly with a computer. They provide a useful comparison with the mathematical models (Fig. 3) and demonstrate the accuracy of the present model. Agreement with experimental data, where available, is good also. There are few compilations of robust statis-

ROACH

Figure 3 (A) Expected island length modeled as a function of redundancy for a typical cosmid sequencing project. Project parameters are G , 40,000; L , 500; T , 20. Data points are the average of 1000 independent Monte Carlo simulations (○). The Lander and Waterman model (1988) is shown as a reference. (B) The expected number of islands modeled with the same parameters.

tical data taken during intermediate project assembly points, particularly because compiling a statistically significant set of such data is burdensome. Nevertheless, the experience in our laboratory is that cosmid sequencing projects require redundancies around sevenfold for closure (Rowen and Koop 1994). Results from other laboratories support this view (Davison 1991; Bodenteich et al. 1994; Martin-Gallardo et al. 1994).

DISCUSSION

The major utility of the mathematical approach presented here stems from an initial determina-

tion of the distribution of the spacings between adjacent fragment start sites and the subsequent use of geometric probability. This general approach to modeling finite genomes may be useful for other mapping and sequencing strategies, such as those based on random transposon insertion. Such strategies represent a genomics implementation of the well-established "coverage process" theory. Hall provides a nice entry into some of the relevant mathematics literature (1988).

The equations derived here have many applications. To begin with, a strategist is interested in the amount of work necessary to complete a project. This can be expressed as the probability of project completion at a given redundancy, where project completion is defined as having closed all gaps. This is given by equation 4 and is graphed in Figure 4. These results are consistent with the expected redundancy needed for closure (equation 16), which is also indicated in Figure 4. Not surprisingly, longer targets have a higher cost in redundancy to close. Also, in general, fewer longer sequences are more desirable than a proportionally greater number of shorter sequences (or their mapping equivalent).

The cost of directed sequencing is roughly constant per gap, no matter how long the gap is. However, there is an exponentially increasing cost in redundancy to close gaps as shotgun projects proceed. Therefore, choosing whether and at which point to stop shotgunning and begin directed sequencing is a fundamental economic question. For this purpose it is useful to calculate the incremental redundancy cost of shotgun projects per gap expected to be closed (Fig. 5). This cost can be compared with the cost of directed sequencing to determine if and when directed methodology is appropriate for a project.

Several simplifying assumptions were employed in the present work. In particular, uniform distribution of fragment start sites may not be the case for some shotgun projects (see Deininger 1983). However, when variations in uniformity are local, the uniform distribution is an excellent approximation. This is valid for most projects, particularly with modern DNA shearing techniques. Larger variations in uniformity tend to be target idiosyncratic and difficult to model. Such variation may result from target regions that are genetically unstable and thus absent from the subclone library. This points out one particular utility of mathematical models: Unexpected deviations from predicted values serve as

RANDOM SUBCLONING

Figure 4 The probability of project completion graphed vs. redundancy, calculated using the exact equation of Stevens (1939). This equation is approximated in the text by equation 4. Four parameterizations are shown. The vertical lines intersect the expected redundancy necessary for closure, calculated using the exact equation of Flatto and Konheim (1962), with $\alpha = (L - T)/G$.

an indication of subcloning problems. Then, such problems can be addressed and corrected.

Although held constant in the present model, both fragment length (L) and necessary overlap (T) may vary in practice. Additionally, overlap may be expressed as a probability, not a certainty, and this “probability of overlap” is affected further when more than two fragments overlap at the same position. Repeated sequence elements in the target tend to decrease the probability of certain overlaps. Such considerations can be incorporated at the price of complexity into the present model by employing distributions for the values L and T , which can be combined into a distribution for L' , where $L' = L - T$. However, variations in fragment length should not cause much concern for the average genomicist. Siegel and Holst (1982) provide a proof that the expected number of gaps is dependent only on the expected fragment length, conditioning on the number of fragments. This proof is pro-

vided for coverage of a circle and also applies to the infinite line. Variation of the expected number of gaps on a finite line caused by variation of fragment length thus is expected only because of edge effects and is predicted to be small. Computer simulations confirm this prediction (data not shown). Although the expected number of gaps remains constant with varying fragment lengths, the distribution of island sizes will change. The probability of project closure will also be affected. In practice, these effects are likely to be small because L does not vary greatly, particularly in sequencing projects. Additionally, in most cases $T \ll L$ and can be approximated as 0.

Computer simulations were employed in the present work to verify that the approximations reported in this paper were useful for modeling genomic projects. In general, because of the complexities involved with modeling even the simplest random sequencing strategies, computer simulations should be consid-

Figure 5 The incremental cost of closing one gap. This is calculated from the number of expected gaps in a project with no knowledge of a prior state of that project (see equation 5). Note that it is impossible to plan a project with zero expected gaps, because gaps always remain a small but finite possibility. G , 40,000; L , 500; T , 20.

ROACH

ered as an essential adjuvant to planning any large-scale project. Directed aspects of a project, variations in parameters, and boundary conditions are easily included in such simulations. Simplifying assumptions can be avoided. The value of simulations cannot be overemphasized. Mathematical models such as the one presented here are useful in conjunction with such simulations.

In summary, the model presented here should have utility for planning both sequencing and mapping projects, as well as for choosing realistic endpoints for such projects. In particular, the model may provide useful benchmarks for evaluating the progress of large-scale genomic projects, such as those contemplated for the human genome.

ACKNOWLEDGMENTS

I thank Alan Blanchard for inspiration, Nicholas Lovejoy for stimulating and enabling contributions, and Andrew Siegel for kind and critical commentary on the manuscript. I owe a tremendous debt to the reviewers for their many valuable comments, and to Leroy Hood, who has provided constant support. I am also grateful for a grant from the Life & Health Insurance Medical Research Fund.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Batelle, E. 1935. Le problème de la répartition. *Comptes Rendus* **201**: 862–864.
- Bodenteich, A., S. Chisoe, Y.-F. Wang, and B.A. Roe. 1994. Shotgun cloning as the method of choice to generate templates for high-throughput dideoxynucleotide sequencing. In *Automated DNA sequencing and analysis* (ed. M. Adams, C. Fields, and J. Venter), pp. 42–50. Academic Press, New York, NY.
- Clarke, L. and J. Carbon. 1976. A colony bank containing synthetic Col E1 hybrid plasmids representative of the entire *E. coli* genome. *Cell* **9**: 91–99.
- Davison, A.J. 1991. Experience in shotgun sequencing a 134 kilobase pair DNA molecule. *DNA Sequence* **1**: 389–394.
- Deininger, P.L. 1983. Random subcloning of sonicated DNA: Application to shotgun DNA sequence analysis. *Anal. Biochem.* **129**: 216–223.
- Flatto, L. and A.G. Konheim. 1962. The random division of an interval and the random covering of a circle. *SIAM Rev.* **43**: 211–222.
- Hall, P. 1988. *Introduction to the theory of coverage processes*. John Wiley & Sons, New York, NY.
- Lander, E.S. and M.S. Waterman. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2**: 231–239.
- Lévy, P. 1939. Sur la division d'un segment par des points choisis au hasard. *Comptes Rendus* **208**: 147–149.
- Martin-Gallardo, A., J. Lamardin, and A. Carrano. 1994. Shotgun sequencing. In *Automated DNA sequencing and analysis* (ed. M. Adams, C. Fields, and J. Venter), pp. 37–41. Academic Press, New York, NY.
- Roach, J.C., C. Boysen, K. Wang, and L. Hood. 1995. Pairwise end sequencing: A unified approach to genomic mapping and sequencing. *Genomics* **26**: 345–353.
- Robbins, H.E. 1944. On the measure of a random set. *Ann. Math. Stat.* **15**: 70–74.
- Rowen, L. and B.F. Koop. 1994. Zen and the art of large-scale genomic sequencing. In *Automated DNA sequencing and analysis* (ed. M. Adams, C. Fields, and J. Venter), pp. 167–174. Academic Press, New York, NY.
- Siegel, A.F. 1979. Asymptotic coverage distributions on the circle. *Ann. Probab.* **7**: 651–661.
- Siegel, A.F. and L. Holst. 1982. Covering the circle with random arcs of random sizes. *J. Appl. Probab.* **19**: 373–381.
- Staden, R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* **8**: 3673–3694.
- Stevens, W.L. 1939. Solution to a geometrical problem in probability. *Ann. Eugen.* **9**: 315–320.
- Whitworth, W.A. 1897a. *DCC exercises in choice and chance*. Cambridge University Press, Cambridge, UK.
- . 1897b. *Choice and chance*. Cambridge University Press, Cambridge, UK.

Received October 11, 1995; accepted in revised form November 21, 1995.

APPENDIX A: TRUNCATED SPACING DISTRIBUTION

Spacings greater than $L - T$ form gaps, and so are not included in the subset of spacings that may be included in the length of an island. To proceed, these spacings must be eliminated from the distribution of D_k (equation 1) by truncating and normalizing. The expected value of this truncated distribution is (note the change of variable $y = x/G_e$):

$$E(D_k | D_k \leq L - T) = \frac{\int_0^{L-T} xn \left(1 - \frac{x}{G_e}\right)^{n-1} dx}{\int_0^{L-T} n \left(1 - \frac{x}{G_e}\right)^{n-1} dx} = \quad (17)$$

$$G_e \frac{\int_0^{f_G} y(1-y)^{n-1} dy}{\int_0^{f_G} (1-y)^{n-1} dy} = G_e \left[\frac{1 - (1 + nf_G)(1 - f_G)^n}{(n+1)[1 - (1 - f_G)^n]} \right]$$

This value should be used in place of $E(D_k)$ in equation 12.

APPENDIX B: TARGET COVERAGE

The nondegenerate fraction of the target f represented in fragments is

$$f \approx 1 - e^{-R} \quad (18)$$

This is a characteristic equation for the coverage derived from a random cloning strategy, and in genomics literature is attributed to Clarke and Carbon (1976). A similar derivation is presented below, with comments to extend its applicability, particularly to pairwise projects where both ends of a fragment, but not the center, are characterized (see Roach et al. 1995).

Let a shotgun strategy be executed such that p_x is the probability of coverage of base pair x by any given fragment. The probability that the base pair is covered by at least one fragment is thus $1 - (1 - p_x)^n$. The expected value and higher moments of f follow from the results of Robbins (1944), with

$$E(f) = \frac{1}{G} \sum_{x=1}^G [1 - (1 - p_x)^n] \quad (19)$$

When $p_x = L/G$ for all x , this expected value is approximated by equation 18. Note that for pairwise projects L is replaced by the characterized length of each fragment. For linear targets p_x is not constant, and falls off near the edges. Despite this, unless L is a significant fraction of G equation 18 remains an adequate approximation to equation 19.

Using the approach of the present paper, f may be obtained also by subtracting the sum of the gap lengths from the total target length. This may be employed as an alternative approach to calculating expected island length. Note that an excess of negative gap lengths will result in clonal coverage in apparent excess of the total target length. This is most apparent when T is large. If the actual coverage is desired, the length of a gap should be calculated as $D - L$ for $D > L$.

Random subcloning.

J C Roach

Genome Res. 1995 5: 464-473

Access the most recent version at doi:[10.1101/gr.5.5.464](https://doi.org/10.1101/gr.5.5.464)

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Focused on your science.

USC
SCIENTIFIC
SIDE BY SCIENCE

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>