RANDOM SUBCLONING

"You cannot see the wood for the trees."

English Colloquial Saying

Peer as we might, we find ourselves in a deep entangled forest, seeing about for only a short distance. What we see is vivid, in great detail, but it is not enough. Standing in one place we cannot grasp the whole. This is the state of genome research.

Imagine visiting the Louvre with a great magnifying glass, constrained to examine each painting one tiny fleck of paint at a time, unable to step back, unable to see the art itself. As you look at the Mona Lisa, with each glance, you are only able to see a spot one twentieth of a millimeter in diameter. The textures and colors of the cracked oil pigments leap out at you. You make detailed notes, recording each observation in a computer database. Your computer churns out statistics on pigment height, color, and crack length distribution. You become frustrated by your inability to see the Mona Lisa.

The human genome, packaged into a living cell a few microns in diameter, is paradoxically too large to see. It is true that chromosomes, the structural material encoding the genome, can be seen, but not the information itself. The word "genome" defines an informational concept. Chromosomes contain DNA. A genome contains information. DNA is an ink that forms words. Genes are sentences and stories, information conveyed by DNA. The genome is often analogized as a book of life (e.g., Wills, 1991).[1]

It is not proper, however, to compare the genome to a book, for a book can be read by human eyes, while the genome cannot. A better analogy is that of a magnetic storage disk, able to release its information only with the aid of a mechanical interpreter, such as a computer. Within the confines of a living cell the genome is read with biological machines - polymerases, ribosomes, and a multitude of supporting enzymes and structures. Human

[1] Humans are by no means currently capable of understanding all of the informational content of any genome. Nor will such understanding come solely from genomics. Progress in all allied disciplines, particularly those focusing on protein function, is necessary to contribute to such understanding. Biological understanding is the result of a vast friendly collaboration between workers in an amazingly diverse array of fields.

observers use other machines – DNA sequencers – to read the genome. It is the limitations of these machines, and of the chemistries that underlie their workings, that place the genome researcher in the shoes of the magnifying-glass-toting art connoisseur.

Each machine can read only a small fraction of the genome at a time. The human genome contains three billion base pairs arrayed on twenty-four separate chromosomes. If this information were stored on a computer disk, it would occupy about a gigabyte of space. A single "sequence read" from a DNA sequencer permits the visualization of only a tiny amount of this data, currently about five hundred to one thousand base pairs.

1.1 MAPPING AND SEQUENCING LARGE GENOMES

Genomicists wish to obtain the entire sequence of the human genome. This is the goal of the Human Genome Project, which is currently a major international effort (McKusick, 1997; Rowen et al., 1997). However, the methodologies developed for the human genome will not end there, for many other genomes await. Nor will the human genome be the toughest genome ever to be sequenced. The genome of the lily *Fritillaria davisii*, for example, contains $3x10^{11}$ bp, while *Amoeba dubia* has $6.7x10^{11}$ bp (Wachtel and Tiersch, 1993; Li and Graur, 1991).

How to sequence these genomes is problematic. How to best sequence these genomes is even more problematic. Optimizing genome sequencing is not merely an academic question. The scale and expense of genome projects is such that slight differences in the efficiencies of various strategies can spell the difference between feasibility and impossibility. The first step in choosing the best strategy is to understand the consequences of pursuing any given strategy. We must ask ourselves, "What results can we expect to obtain given a certain level of expenditure of resources?" If we can decide upon our goals ahead of time, we should be able to estimate how much it will cost us to reach these goals.

Random subcloning is a popular strategy in use by genomicists. Random subcloning is simple, and is the easiest strategy to implement. The strategy iteratively generates sequences from random locations in a genome until, by chance, sequences from the entire genome have been obtained. At this point, data analysis algorithms are used to reconstruct the global picture from the random fragments. The "forest" is reconstructed from a random collage of local snapshots of the "trees."

Random subcloning is not merely a sequencing strategy. The principle of breaking down large problems into small problems is not unique to sequencing. It is possible to analyze

the genome without determining the details of its sequence. Such an analysis is termed "mapping." A genome mapping project seeks to identify and order landmarks of the genome. A physical map uses structural landmarks, such as restriction sites, or the ends of clones. A genetic map uses informational landmarks, such as genes. The methodologies of genetic mapping are quite different from those of physical mapping, but when landmarks from the different methodologies can be correlated with each other, the maps can be integrated. I will not discuss genetic mapping further in this dissertation, and will focus on physical mapping and sequencing.

It is possible to use physical mapping techniques to analyze fragments of DNA that are much longer than the length of a sequence read. It is quicker and easier to build up a global picture with such techniques, as the fragments of the puzzle can be a thousand times larger than individual sequence reads. The trade-off for speed is that the resulting picture is fuzzier, much like an impressionist painting compared to a sharp color photograph. One advantage of physical maps is that their component subclones can be used as targets for sequencing. Since subclones are smaller than genomes, they present far more tractable problems. In practice, sequencing strategies are not applied directly to a target as large as the human genome, although they are to much smaller genomes, such as those of certain bacteria.[2]

An important and recurring debate among structural genomicists is when and in what proportions to use directed strategies as opposed to random strategies. Examples of both of these types of strategies are discussed further below, but the strategies can be characterized briefly: random strategies are cheap but necessarily redundant, with an exponentially decreasing return on investment as a project progresses; directed strategies are expensive but non-redundant, with a constant rate of return. This dichotomy immediately suggests compromise. It is possible to begin a project with a random strategy and switch to a directed strategy in a "finishing" phase. Timing this switch is an important economic choice.

Until 1993, when I was able to approach the mathematics of random subcloning, little was known about the expected outcomes of random subcloning projects given certain levels of resource expenditures and parameter choices. A simple equation for expected target coverage was communicated to the genomics community by Clarke and Carbon in 1976. In

---

[2] Sequencing and mapping can be interleaved. A promising strategy is described by Venter et al. (1996) and analyzed in detail by Siegel et al. (1999). Bacterial genomes sequenced to date include those described in Fleischmann al. (1995), Fraser et al. (1995), Himmelreich et al. (1996), Bult et al. (1996), Kaneko et al. (1996), and Blattner et al. (1997).

1988, Lander and Waterman addressed several additional issues, and developed additional equations describing random subcloning projects, but these equations were valid only at low redundancies, and had little value for strategy determination. In fact, because of their flaws at high redundancies, these equations were often misleading. Most random strategies were thus conducted and evaluated empirically, but seldom could enough empirical data be collected to yield generalizable results.

The most fundamental question to ask of a random subcloning project is, "How many clones on average does one have to sequence or analyze before I completely cover my target?" Before we address this question, let us step back a bit, and discuss why a simple directed strategy, sequence walking, fails.

1.2 SEQUENCE WALKING

If one were presented with an unknown genome of $3x10^9$ bp, it would seem that the easiest and most efficient way to sequence it would be to sequence all $3x10^9$ bp in order, just one time, and be done with it. This strategy is known as "sequence walking" and it, in fact, works quite well, at least for very short genomic targets. The difficulties of sequence walking lie in the technical details of sequencing.

One cannot start a sequencing reaction, which will produce a sequence read, at any arbitrary point in the genome. Each sequencing reaction must be primed, and priming requires known sequence. A sequencing reaction can only start where the sequence is already known. The reaction can extend into an unknown region, but it must start in a known region. The required length of this known region is not set in stone, but it is about 20 bp. Each sequencing reaction starts with a primer, which is a short DNA oligonucleotide complementary to the known region.

Each primer must be unique. The need for unique primers adds to the expense of sequence walking.[3] If the known region on the target genome to which the primer binds is repeated in more than one site, the primer will bind at all these sites, and multiple sequencing

[3] The commercial cost of primer synthesis is approximately $0.70 per base, or $15 for a typical sequencing primer (e.g., Operon sales literature, 1997). Due to significant competition in the commercial primer synthesis field, this price is probably a good reflection of the actual cost of synthesizing primers. On-site primer synthesis can be done overnight; commercial synthesis incurs an additional delay due to shipping. These delays add to the net opportunity cost of sequence walking. On-site robotics drop the individual cost of primer synthesis, secondary to a large initial investment in equipment.

reactions will occur simultaneously, producing nonsensical and useless sequence reads.[4] There are 4^{20} different possible 20 bp primers (4^{20} is roughly $1x10^{12}$). However, three additional factors must be considered. First, a primer will recognize not only a site that is exactly complementary, but also many other sites that are merely similar. Secondly, the composition of genomes is not completely random, in ways that increase the probability of a given sequence occurring more than once. Thirdly, not all primers are particularly good at initiating sequencing reactions, so many of them cannot be used.[5] These factors limit the complexity of target templates used in sequencing reactions. It is hard to obtain sequence reads from templates that are longer than a few hundred kilobases. In addition to this, it is difficult to physically manipulate such templates in the laboratory.

With these factors in mind, it is not surprising that the quality of sequence data diminishes as the template length increases. So although sequence can be obtained from bacterial artificial chromosomes (BACs), the highest quality sequences are obtained from phage or plasmid templates, which are only a few thousand base pairs long.

Template complexity is only one problem. Consider now the basic strategy of sequence walking. The idea is to start at one end of the target and "walk" towards the other end.[6] Each sequence read provides known sequence that can be used to generate a primer for the next sequence. If each sequence is 500 bp, and each primer is designed from the last 20 bp of the previous sequence read, then it would take 42 sequencing reactions to cross a twenty kilobase target. The redundancy *R* of such a project would be calculated as follows:

$$R = \frac{\text{Total Sequence Obtained}}{\text{Target Length}} = \frac{nL}{G} \qquad (1.1)$$

Here, and in general, *n* is the number of fragments sequenced in a project, *L* is the length of a fragment, and *G* is the length of the target. In this case *R* is 1.05, which is very close to the ideal of exactly one. Redundancy is one measure of the efficiency of a project. If this were the only measure of efficiency, the extremely low redundancy would be a compelling argument for sequence walking, at least on targets that were short enough so as not to be too complex

[4] There are some clever tricks for doing multiple sequencing reactions in the same tube (i.e., Wiemann et al. 1996). However, none of these bypass the target sequence complexity issue discussed here.

[5] For chemical reasons, primers are usually chosen so that their G-C content is about 50%. Additionally, the last base of a primer is usually chosen to be a guanine or a cytosine.

[6] An obvious and common extension to the strategy is to start "walking" at both ends at meet at the middle.

for a sequencing reaction.

One might ask how a walking strategy is initiated. In many cases, sequencing targets are cloned DNA. A clones consists of unknown sequence embedded in a known "vector" sequence. The ends of the unknown target sequence are flanked by known vector sequence. It is thus a simple matter to initiate a walking strategy on cloned DNA. In other cases, such as mapping, a walking strategy can only be initiated from a previously analyzed region.

There is an additional problem that limits sequence walking. Regardless of the template complexity, some sequencing reactions will nevertheless "refuse" to work. When walking, each sequence read provides the data for priming the next, so if even one fails, the project halts.[7] Reactions fail for many reasons, some of them unknown, but one cause might be the presence of tertiary structure such as a hairpin loop in the target DNA. Even if such failures are rare, the longer a template is, the more likely such a problem is to occur. Failure rates vary highly. If one primer fails, a nearby primer or alternative chemistry can be tried, often with success, but occasionally with repeated failure and always with project delay. In practice, few sequence walking projects tackle targets longer than 20 kb.

One of the most significant causes of the failure of a walking iteration is not so much that the reaction has failed to work, but rather that it has worked more than once. This happens when the target sequence contains sequences that are nearly identical to each other, known as repeats. Such repeats are common in metazoans and infrequent in other organisms. The presence of such a repeat brings a sequence walk to a halt as soon as the walk attempts an iteration from within a repeat. The only solution is to fragment the target and initiate sequencing on a smaller subclone. If the target is going to have to be subcloned anyway, it would have been better to subclone it initially followed by detailed mapping of the subclones or, more likely, random sequencing.

Nevertheless, one of the largest headaches of sequence walking has nothing to do with chemistry, but rather administration.[8] An administrative step must occur after every sequencing reaction in a sequence walking strategy.[9] This step involves analyzing the data from the previous reaction, designing a new primer, synthesizing the primer, and then

---

[7] Even if the project is not terminated, a considerable amount of effort must be expended in order to continue walking, such as running sequencing reactions with alternative chemistries or using a different primer.

[8] This general observation holds for many things in life.

initiating a new sequencing reaction. Much of this administration can be implemented by computers and robots, but an unavoidable consequence is that each reaction must be done consecutively. Contemporary automated DNA sequencers have the capacity to run about fifty samples simultaneously. Well-equipped labs can process thousands of reactions per day. With a sequence walking strategy, all this capacity is wasted, as only one reaction can be done at a time.[10] Keeping track of data, primers, clones, and targets can easily be the greatest challenge and cost of a walking strategy, far exceeding any savings in efficiency gained from an ultra-low redundancy.

Furthermore, an extremely low redundancy is not desirable. Errors occur in sequencing reactions at a rate of roughly 1%.[11] The desired accuracy of most projects is about 0.01%.[12] Therefore it is not sufficient to obtain a single sequence read across each region of the target. A bare minimum is to obtain one read from each strand of the target, bringing the theoretical ideal redundancy to twofold. In practice, more than two reads are often needed for accuracy, particularly in places where the first two reads disagree. For targets that have highly similar repeats, even two sequence reads may not be enough to distinguish minor variations between the repeats. This presents significant challenges to a directed strategy or even to a low redundancy random strategy. A highly redundant random strategy will usually have relatively little problem resolving such repeats, often with no further need for experimentation. It can turn out to be less expensive to do more up-front random sequencing than to solve problems that arise from low redundancy data. Another approach to problem resolution is to use mapping data such as those provided by a pairwise project. Pairwise projects are described in

---

[9] A detailed analysis of administrative costs is beyond the scope of the present work. From an operations research viewpoint, estimating such costs is one of the most difficult aspects of analyzing strategy costs. A series of preliminary forays into this difficult field can be found in several papers by Siegel et al. (1998a; 1998b; 1999). Many costs can be reduced by automation. However, the development costs, ultimate utility, and life cycle of robots are extremely hard to predict, or even in retrospect to calculate. Failed efforts at automation (i.e., "dud" robots) should be accounted for in such calculations. In the face of these difficulties, empiricism provides powerful insight into some of these hard to calculate costs. For example, all major sequencing labs have abandoned primer walking in favor of random subcloning. This suggests either mass delusion or a consensus that the overall opportunity cost of random subcloning is lower. However, even assuming that this global shift in strategies was originally fundamentally sound, it remains ever possible that improvements in alternative strategies might tip the opportunity cost balance in another direction. In future sections, I will point out a few potentials for such balance shifts.

[10] In practice, many targets are analyzed in parallel, allowing the laboratory's full capacity to be used. However, this adds to the administrative complexity.

Chapter 2. It should be noted that as advances in technology drop the error rate of sequence reads, then directed strategies become slightly more favorable.

Not all of the drawbacks to sequence walking apply to mapping projects. The technologies and chemistries involved in mapping projects are quite different. However, the problem of administrative overhead remains. Additionally, the problem of the "rare failure" becomes much worse. The iterative step in a sequence walking strategy that can fail is a sequencing reaction. The analogous step in a physical mapping project is the identification of a clone that overlaps a known map and extends into an unknown region. This identification process can also have a high failure rate. Most mapping walking projects are doomed to long-term failure, and are thus reserved for the generation of very short maps. An excellent review of the technology and difficulties of map walking has been provided by Stubbs (1992).

The considerable drawbacks of walking strategies have fueled the exploration of alternative strategies, including random subcloning.[13]

1.3 OVERVIEW OF RANDOM SUBCLONING

The forest-and-trees analogy can be rephrased in molecular biology terms: 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. Following detailed analysis of the many, a map or sequence of the target can be reconstructed.

As I move to a more technical description of random subcloning, I wish to be clear

---

[11]The exact error rate varies as a function of position in the sequence read. Additional parameters include the sequencing chemistry, the template, the primer, the choice of automated sequencing machine (or absence thereof), and the choice of run parameters such as voltage and run time. For example, ABI sales literature claims production of 800 bp of 98.5% accurate sequence in 8 hours on a PRISM 377 DNA Sequencer using a LongRead DNA cycle sequencing standard, beginning from base twenty of the read. Many genome centers maintain statistics on error rate as a function of sequence position, and will usually provide such statistics upon request. These statistics are constantly changing however, as centers implement incremental advances in technology. The estimate in the text of roughly 1% error is a ballpark estimate of current error rates. A detailed discussion of error rates is beyond the scope of the present work. One starting point for the acquisition of additional information is the web site of the National Human Genome Research Institute (www.nhgri.nih.gov). Between 520 and 550 "high quality" bases are obtained for sequence reads from the *Pseudomonas aeruginosa* project underway at the University of Washington (Maynard Olson, personal communication). "High quality" is a statistic produced by the programs *PHRED* and *PHRAP* (Phil Green, author).

with my terminology. I mean different things by the word “fragment” when referring to mapping or sequencing. The use of the term “fragment” will allow me to discuss both methodologies concurrently and to develop a mathematical model applicable to both. Physical mapping requires fragmentation of the target. The resulting fragments are cloned into vectors. If the target was a clone, the new constructs are called “subclones.” I refer to the unknown DNA present in these clones or subclones as “fragments,” and this quite literally represents an actual physical fragment of the target. For sequencing projects I refer to each sequence read as a “fragment.” Although less literal, this definition allows me to maintain a precise analogy.

Ideally, a direct strategy is pursued by analyzing a minimum number of fragments such that a minimum tiling path is followed. Walking, described above, is an example of a minimum-tiling-path strategy. In general, the determination of a minimum tiling path requires prior knowledge of the relation of each fragment to the original target. Such information is not easily available. Sequence walking obtains this data by iterative step-by-step sequencing. An alternative to sequence walking is to physically map a large number of subclones before sequencing any of them. If the number of subclones mapped is much larger than the number needed to define a minimum tiling path, it is usually possible to choose a path of subclones to sequence that approaches a minimum tiling.

In a prototypical random subcloning sequencing project, only one end of a subcloned fragment is sequenced. This is largely a matter of convenience, as the same primer, derived

---

[12]The current standard for federally funded Genome Centers is 1 error in 10 kb (National Human Genome Research Institute guidelines). The error rate for genomic sequencing in the Hood lab from 1991 to 1994 ranged between 0.98 and 1.4 errors per 10 kb. For 1996 and 1997, the error rate was 0.16 to 0.20 errors per 10 kb (Lee Rowen, personal communication). Error rates are estimated based on discrepancies between overlapping clones derived from the same haplotype.

[13]There is a constant interplay between cost, strategy choice, and advances in technology. For example, a recent advance in primer initiation chemistry may permit sequencing walking without the necessity of synthesizing a new primer for each walking iteration (Mugasimangalam et al., 1997). This method exploits “differential extension of nucleotide subsets.” Short primers from a presynthesized library upstream from target regions lacking a particular nucleotide, then extended without that nucleotide at low temperatures. Spurious priming sites are not extended due to the missing nucleotide. Subsequently the temperature is raised and the missing nucleotide added in order to complete a sequencing reaction. Early implementations of this technique entailed a compromise in sequence read length and an increased failure rate. However, optimization might eliminate such drawbacks. In any case, the use of this technique would still involve the administrative overhead of clone tracking through iterative steps.

from known vector sequence, can be used for every sequencing reaction. The unknown DNA in a subclone is often much longer than a sequence read - perhaps a couple of thousand of base pairs compared to a read length of about 600 bp. Thus a tiling path for such a project will involve clone ends spaced less than one sequence read length apart.

The cost of producing a minimum tiling path map can be quite large. The exact costs of generating such “sequence-ready” maps are hard to determine, either empirically or theoretically. In all cases, however, these costs must be weighed against the alternative of using a less optimal map, or even no map at all. Without a map, fragments must be picked and analyzed at random. This limiting case is the strategy of random subcloning, also known as “shotgun sequencing.”

Note that if one could analyze a single target molecule at a time, additional strategies would become available. One can imagine fragmenting a single DNA target molecule, and keeping track of each fragment and where it came from. One could analyze each fragment and then immediately reconstruct the target sequence. One would reach the ideal project redundancy of onefold. It is actually be possible in some circumstances to pursue such a strategy. Optical restriction mapping promises exactly this (see, for example, Anantharaman et al., 1997). Currently, it is not possible to sequence DNA in such a fashion.

During shotgun sequencing, fragments are generated from a vast number of identical target sequences, typically about a trillion. The resulting “library” from which the fragments are selected for further analysis is thus redundant. Individual fragments may overlap in the sense that they mutually possess, in part or entirety, the same bit of target sequence. In particular, because of the effectively infinite number of fragmented target sequences, each fragment chosen at random from the resulting fragmented mixture is independent of all the other fragments. The locations from which these fragments arise can thus be considered to be uniformly distributed.[14]

One difficulty of random strategies is the problem of retrospectively determining from where a fragment came. If a fragment by chance happens to overlap known sequence, such as would occur on the boundary between the vector and the unknown target, the region of known

[14]In practice, fragments from some locations are observed less often than others. The deviations from uniformity depend on the technique used to fragment the DNA (see, for example, Deininger, 1983). However, most modern techniques, such as shearing by HPLC, tend to be quite uniform. As long as the deviations in start site uniformity are small compared to the fragment length, a uniform distribution works well as an approximation.

DNA can be extended. If another fragment overlaps the first fragment, the known region can again be extended. The process of extending the known region of the target sequence is similar to that of sequence walking, but with the methodology reversed. In walking, one first identifies an unknown fragment overlapping known sequence, and then analyzes. In shotgunning, one first analyzes a very large number of fragments, and then finds one that overlaps known sequence.

The beauty of the "assembly" stage of a shotgun project is that it is not linear. Since every fragment has been analyzed up front, they all represent known sequence. The relationships of the known sequences to the original unknown target sequence are not known, but the sequences themselves are. Every fragment is a "seed" from which longer known sequence can be "grown" by identifying overlapping fragments. Every fragment is a starting point. Shotgun assembly works like a polymerization reaction, with eventual coalescence of all the fragments into one final assembled sequence. If enough fragments have been analyzed, this final assembly will continuously cover the target sequence, and the project will be over.

If not enough fragments have been analyzed, there will be gaps in the target. This is undesirable. On the other hand, analyzing more fragments than necessary to cover the target is also undesirable. Understanding the mathematics of shotgunning permits a judicious choice of the number of fragments to be analyzed up front. If not enough fragments are analyzed at first, a trial assembly can be made, and if gaps are discovered, more fragments can be analyzed until the gaps are closed. Alternatively these gaps can be closed by other strategies, such as walking. Determining the costs of gap closure of various strategies is important in choosing between alternative approaches to gap closure. It should be emphasized, however, that iterative strategies are undesirable, as they require extra administrative overhead.[15] It is more desirable to analyze all necessary fragments at once and then make one final assembly than to analyze fewer fragments at first than necessary, assemble them, determine the need for more analysis, and then repeat the process, perhaps several times.

Assembly of analyzed fragments is a fascinating theoretical challenge. For random projects, no map exists to determine that two fragments lie adjacent to each other. Initially, decisions of adjacency must be made by pairwise comparison of fragments. Two fragments that overlap will share a portion of target sequence in common. By looking for these common sequences, adjacency can be detected. Sequence reads may contain errors, so fragments may be declared to overlap even if their common sequences do not match perfectly. Furthermore, even if two sequences match each other perfectly, or nearly so, the fragments from which they are generated may not overlap, as the target may contain two or more nearly identical

sequences. Often higher-order comparisons, with checks for consistency between three or more sequences, are necessary to resolve ambiguities in assembly. Assembly is a task best done by computers. Algorithms for assembly, such as that of *PHRAP*, are constantly improving (Phil Green, author).

As mentioned above, from a theoretical standpoint, random mapping and sequencing strategies can be treated identically. It is worth emphasizing, however, that they involve very different scales. Most physical mapping projects approach targets on the megabase scale or larger, such as entire genomes. These large targets are randomly fragmented into YAC, BAC, cosmid, or phage subclones ranging in size from tens of kilobases to several megabases. Analysis techniques include restriction mapping, 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. A schematic diagram of a random subcloning project is shown in Figure 1.1.

In the increasingly automated modern scientific laboratory, an additional appeal of random subcloning is rooted in the absence of need for prior information about particular fragments. This allows projects to be undertaken with a great deal of “blind” automation and

---

[15]The costs of iterative steps are undesirable, but not necessarily prohibitive, depending on the strategy. An example of strategically desirable iteration is illustrated by the “sequence-tagged-connector” (STC) strategy of Venter et al. (1996). There are several key differences in the nature of the STC iterations and sequence walking iterations. First, STC iterations are mapping iterations that occur with low periodicity. A laboratory must cycle through one STC iteration for each BAC sequenced. If an iterative sequencing strategy were employed on a BACs completely sequenced during the course of an STC mapping and sequencing project, the complete BAC sequencing iterations would have to occur two to three orders of magnitude more frequently than the mapping iterations. Secondly, the failure rate of STC mapping iterations is predicted to be low (Siegel et al., 1999). Furthermore, the “cost “of a rare failed STC mapping iteration will be low – in the worst case, a non-contiguous BAC will be completely sequenced, reducing the cost of future work. This cost will not be completely recouped due to inefficiencies introduced by the rogue BAC into the final sequenced tiling path. More likely than this scenario, however, would be an extremely early termination of complete BAC sequencing due to recognition of sequence inconsistencies. The STC strategy is a member of the pairwise end sequencing family of strategies, which I discuss at further length in Chapter 2.

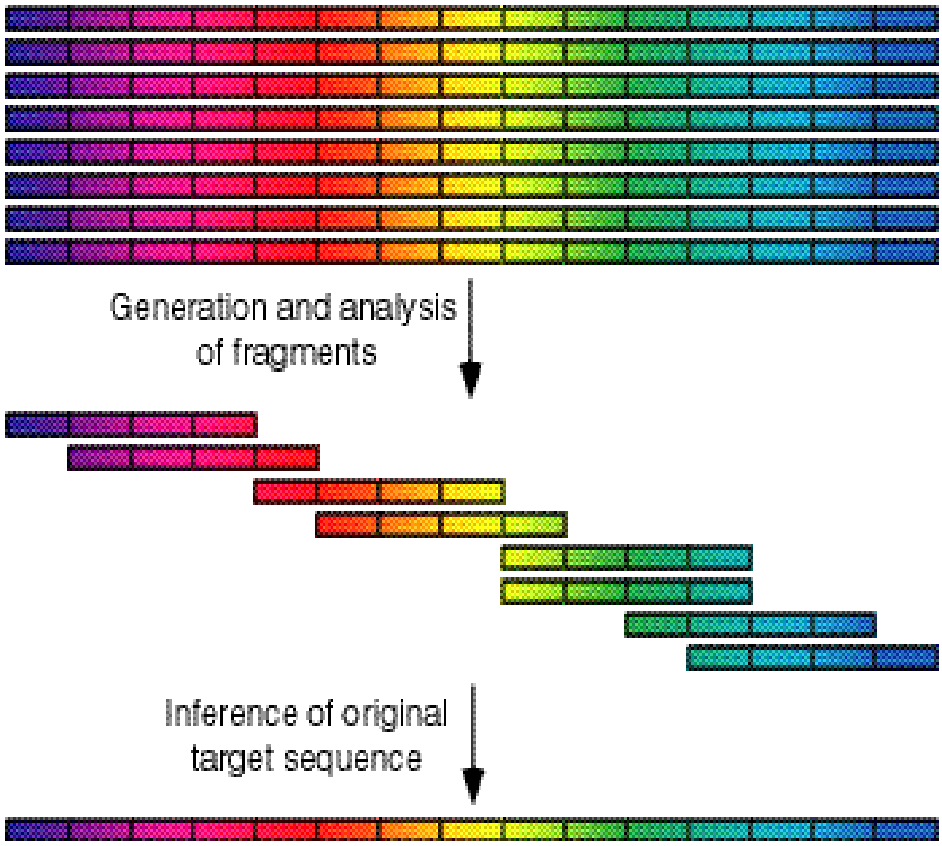

Figure 1.1. A schematic cartoon of a random subcloning project. Fragments of multiple identical copies of a target sequence (here a rainbow of colors) are analyzed and then reassembled based on bits of overlapping sequence identity.

with a decreased need for highly trained human intervention. The main drawback of 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 progressing 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.

In the following sections, I will present analytical models and simulations of random subcloning. I will illustrate these with empirical observations, where available.

1.4 MATHEMATICAL MODEL - BASIC FORMULATION

I assume a linear target of length *G*. In Section 1.13, I will modify my analysis to make it applicable to circular targets. Most sequencing targets consist of a single linear strand of DNA cloned into a circular vector, resulting in a circular construct similar to the original vector, just much bigger.

For purposes of this mathematical model these seemingly circular targets are properly considered to be linear. The reason for this is that the vector sequence is already known. No additional information is gained by analyzing fragments that arise exclusively from the vector. One method of avoiding analyzing vector sequence is to "screen" all fragments before analyzing them. A labeled probe made from vector sequences can be used to tag any fragments that contain vector sequences, allowing then to be excluded from further analysis. This is seldom done, as it involves extra labor and invites the possibility of error. In particular, unknown target sequence might be accidentally screened out by the labeled vector probe. Also, fragments that overlap both the vector and the unknown target would be screened out. This is usually undesirable, as such fragments can provide critical anchoring information during the project assembly phase.

More often, a sequencing project will not screen out vector sequences before fragment analysis begins. Thus any fragments that exclusively overlap vector will be sequenced regardless. This will represent wasted effort. However, fragments that overlap both the vector and the target contribute information to the final assembled sequence. Therefore the best way to set up a theoretical framework for the analysis of typical random shotgun sequencing projects is to ignore fragments that exclusively overlap target except for their presence as "wasted" sequencing reactions. Fragments that overlap the target by a base pair or more are not ignored, so the effective target length in such typical situations is increased by the length of a fragment (minus a base pair) on each of two ends.

Now, the total subclone length is the vector length *V* plus the target length:

$$\text{Subclone Length} = V + G \tag{1.2}$$

Fragments, each of length *L*, that overlap both the vector and the target will be included in assembly, so the total length of the linear target should be calculated as:[16]

$$G' = 2L + G - 2 \tag{1.3}$$

Furthermore, if screening is not employed then not all analyzed fragments will be usefully

included in the assembly, so the effective redundancy will be less than the actual redundancy:

$$R_{\text{effective}} = \left(\frac{G'}{V+G}\right) R_{\text{actual}} \tag{1.4}$$

I will be defining "gaps" in such a way that the uncovered target sequences, if any, at either end of the final assembled target are not counted as gaps. At the high redundancies necessary for project completion, these uncovered end regions are likely to be quite short. Furthermore, if the project is a typical shotgun sequence project as described above, these end regions will actually be vector, so it won't matter if they are not covered by analyzed fragments.

End regions do, however, become a concern if the target actually is physically linear, and not cloned into a vector. The only practical examples of this of which I am aware are genomic physical mapping projects targeted at a eukaryotic chromosomes. On average, the probability of a particular base pair of the target in such a project being covered by at least one fragment is:

$$P_{\text{coverage}} = 1 - \left(1 - \frac{L}{G}\right)^n \tag{1.5}$$

This probability drops precipitously near the ends of the target, however. In particular, the probability of the first or the last base pair being covered is:[17]

$$P_{\text{coverage}} = 1 - \left(1 - \frac{1}{G}\right)^n \tag{1.6}$$

---

[16]Note that V is assumed to be greater than $2L-2$. If $2L-2 > V \geq L-1$, then V should be used in place of $2L-2$ in equation (1.3). If $V < L-1$, then the project should be analyzed using the results for circular targets presented in Section 1.13. In this final case, gaps in the vector sequence can be ignored, so the expected number of gaps should be modified as:

$$N_{\text{expected gaps}} = \left(\frac{G}{V+G}\right) N_{\text{predicted gaps}}$$

Other equations should be modified appropriately. None of these cases is likely, as vectors are usually longer than two fragment lengths.

[17]The general equation for a base pair of a distance d away from the target end, with $0 < d < L$, is:

$$P_{\text{coverage}} = 1 - \left(1 - \frac{d}{G}\right)^n$$

These probability "edge effects" could become important if the fragment size was on the order of the target size. This does not occur in typical random subcloning projects, but a similar phenomenon occurs with genetic mapping of marker loci. Bishop et al. (1983) analyze this issue at length.

Therefore such projects need to employ special techniques to analyze the target ends. This will usually necessitate telomeric cloning, a discussion of which is outside the scope of this dissertation.

Another kind of project worth mentioning is one in which the target may consist of multiple linear segments. Most eukaryotic genomes contain multiple linear chromosomes, so this is the usual case for a eukaryotic mapping project. Universally, the length of each chromosome is much longer than the fragment length, so such projects can accurately be modeled using the present linear framework by setting the target length G to be the sum of all of the chromosome lengths. If the chromosomes are not present in stoichiometric amounts, appropriate considerations must be made for the underrepresented chromosome(s). This might occur if a mapping project were undertaken on an entire male genome, but most projects would seek to avoid this situation. With multiple chromosomes in a random subcloning project, both ends of each chromosome are likely to be uncovered, necessitating telomeric cloning or other approaches to the ends. It is also possible just to ignore the ends and allow their characterization, if necessary, to be carried out by a completely different project.

With the above considerations in mind, one can define the variables necessary for a mathematical analysis. For a given project, n fragments of constant length L are generated from the target and analyzed in some manner such that overlaps between fragments are detectable. This analysis would be sequencing for a sequencing project and might typically be restriction digestion for a physical mapping project. All fragments are generated from distinct identical copies of G.[18] No fragments may start within L-1 bases of the last, rightmost base of G as such fragments would not be entirely contained within G. Thus the effective length G_e available for fragment start sites is G-L+1. I designate the starting, or leftmost, base pair of each fragment S_k, such that S_1 is the start site of the leftmost fragment, with S_k $[1,2,3,...,G_e]$. S_n begins the rightmost, 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.

Because $G_e >> 1$, we can consider the possible range of fragment start sites to be continuous, rather than the quantum entity that it is. In the present analysis, I will switch back

[18]This statement is made to emphasize the claim that the fragments are independently and identically distributed. It is reasonable; the ratio of originally fragmented target molecules to randomly analyzed fragments is typically about 10^{12}:10^3, or 10^9. The probability that two fragments are derived form the same target molecule is negligible. Even if this were not the case, this mathematical model would likely remain quite valid.

and forth from discrete to continuous models without extensive justification. Continuous models tend to allow more elegant equations and derivations, while discrete formulations occasionally give more precise answers. The main difference between these two formulations is actually quite trivial. As noted above, $G_e=G-L+1$. However, for a continuous model, $G_e=G-L$. In all cases $G>>L$, so $G_e \approx G$. Nevertheless, without careful record keeping, the use of a continuous model will allow some asymptotic limits to converge towards slightly incorrect limits, such as $G+1$ rather than G. This is largely an aesthetic matter without much practical implication. Nevertheless, I believe that a failure of a mathematical model to converge towards an expected limit in extreme cases is a serious drawback. Therefore I will use discrete formulations when necessary to satisfy my own personal aesthetic appreciation for perfect asymptotic behavior. With this in mind, in the continuous 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 1.2.

An important assumption is that all the fragments are considered to be the same length. This is seldom, if ever, strictly the case. The actual length of fragments may vary by 20% or more. It turns out that this has very little effect on the utility of the mathematical model developed here. In some cases this can be demonstrated mathematically. This can be demonstrated quite easily for the equations for target coverage; I will leave this as an exercise for the reader. Also, the expected number of gaps in a project is unaffected by varying the fragment length. Siegel and Holst (1982) provide a formal proof of this for circular targets. Variations in fragment lengths can have subtle effects on the distributions of the number of gaps and the gap lengths, as well as a few other parameters. For practical purposes, these subtle effects are insignificant. Monte Carlo simulation, discussed further in Section 1.14, is another way to verify this assertion.

1.5 TARGET COVERAGE (1)

Two questions familiar to anyone who has taken a long family trip are “Are we there yet?” and “How much farther do we have to go?” For a random subcloning project, arrival means that there are no gaps in the coverage. The question of how much farther can be answered both in terms of the number of gaps that remain and in terms of what percentage of the target remains to be covered. It turns out that the number of gaps remaining turns out to be more useful for gauging the amount of additional effort necessary to complete a project. Nevertheless, the fractional target coverage is also an interesting quantity and, perhaps more importantly, is easy and fun to compute.

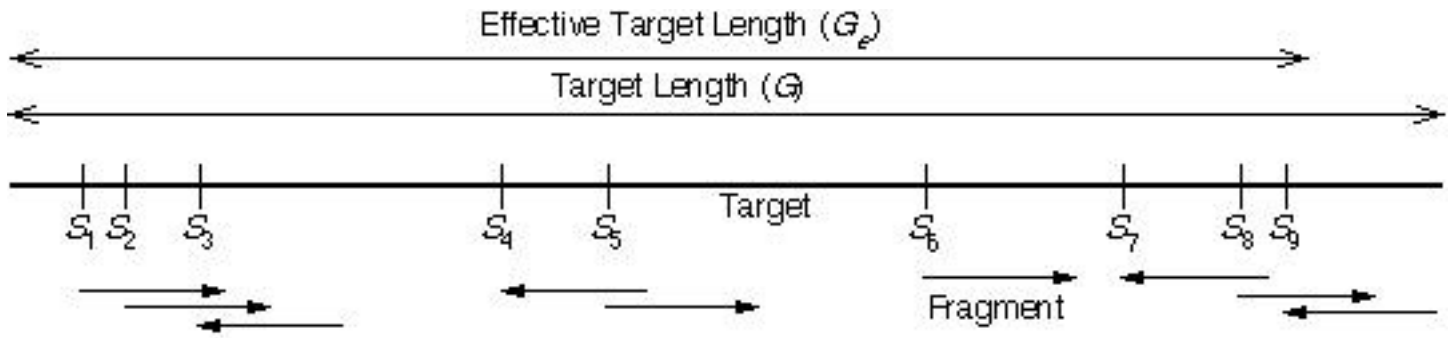

Figure 1.2. A schematic of a mathematical formulation for random subcloning. Note that the choice of orientation for the target is arbitrary. Following this arbitrary choice, the fragment orientation is ignored; the start site (S_k) of fragment k is the leftmost fragment end, whether that end is the 3' or the 5' end of the fragment.

Consider Xeno's paradox. Xeno shoots an arrow at a target. The arrow flies half way, covering half the distance. The arrow covers half the remaining distance, and then again half the remaining distance, and so on. Supposedly, the arrow never reaches its target. This is what happens during a random subcloning project. Actually, not quite, but let us examine the issue in more detail.

Each analyzed fragment in a random subcloning project can be considered to be generated sequentially. Each time a fragment is generated, it covers a random portion of the target. Each fragment is of length L. The fraction of the target covered by one fragment is thus $\frac{L}{G}$. A subsequent fragment will also randomly cover another region of the target of length L, on average proportionally distributed between already covered and uncovered target. Thus, on average, each additional fragment covers a fraction $\frac{L}{G}$ of the remaining uncovered sequence. The actual amount of additional unknown sequence covered rapidly gets smaller and smaller, much as Xeno's arrow will fly shorter and shorter distances with each iteration. Much as it would seem that the arrow can never reach its target, it might also seem that a shotgun project could never reach its goal of perfect coverage.[19]

It turns out that Xeno's arrow will actually reach its goal, but for a different reason than a shotgun project will. Xeno's arrow is helped along by the nature of time, infinity, and the convergence of a series towards its limit. Each iteration of the flight of Xeno's arrow takes place in half the time of the previous iteration.[20] Time is relentless, so the arrow hits its target. However, each generation of an analyzed fragment in a shotgun project requires the same amount of time and effort as was spent on the previous fragment. Time is not the solution to the shotgunner's dilemma.

Consider instead a blob of decaying uranium. As each half-life passes, half of the uranium decays. But not exactly. Only on average. Usually there are a vast number of uranium atoms in any blob, so an average is a fairly accurate estimate of the percent of atoms that decay during any given half-life. But consider now the case of an almost exhausted uranium blob, with only a few atoms left to decay. Quite by chance, all or none of them might decay. It is a stochastic process. Unlike Xeno's arrow driven by certainty, randomness takes charge. With only one atom left, it has a 50% chance of decaying in any given half-life. Eventually, it

---

[19]Many researchers have also had this thought during the assembly of particularly difficult targets.

[20]This assumes the arrow is not losing velocity. If it is, the arrow may very well never reach its target.

will decay. There are no paradoxes for blobs of uranium. The same is true for random subcloning. As a random process, eventually the target will be covered. There is a chance that it might not ever be covered, but this chance is infinitely small. These probabilities will be addressed in more detail later.

For now, let us return to the determination of expected target coverage. Because radioactivity has an exponential decay, and the analogy with random subcloning fits well, one expects to find an exponential "decay" equation for shotgunning. Assume for now that the target is a circle so that we can treat all base pairs identically. Our conclusions will also turn out to be excellent approximations for linear targets as well.

The probability that any given base pair is not covered is:

$$P_{\text{base not covered}} = \left(1 - \frac{L}{G}\right)^n \tag{1.7}$$

Therefore the probability that a base *is* covered is:[21]

$$P_{\text{base covered}} = 1 - \left(1 - \frac{L}{G}\right)^n \tag{1.8}$$

On average, the number of covered bases will be equal to the number of base pairs times the probability that each one is covered:

$$\text{Expected Coverage} = G\left(1 - \left(1 - \frac{L}{G}\right)^n\right) \tag{1.9}$$

Equation (1.9) is often approximated. Recall that $\frac{L}{G} << 1$. Furthermore, for small x:

$$e^{-x} \approx 1 - x \tag{1.10}$$

Therefore, we can rewrite equation (1.9) as:

$$\text{Expected Coverage} = G\left(1 - \left(e^{-\frac{L}{G}}\right)^n\right) = G\left(1 - e^{-R}\right) \tag{1.11}$$

Equation (1.11) is often referred to in genomics circles as the "Clarke-Carbon" equation. Note

---

[21]As a note of historical trivia, it is this equation that appears in Clarke and Carbon (1976). The first published use of the eponym "Clarke-Carbon" for equation (1.11) appeared in Waterman (1995). This equation without the eponym appeared in Lander and Waterman (1988). It also appeared, in a slightly different genomics context, in Lange and Boehnke (1982). However, before that, it may have appeared in a classroom lecture by Dr. Carbon (recollection communicated by a student, Stephen Lasky). To my knowledge, the first derivation of equations (1.9) and (1.11) was provided by Robbins (1944 and 1945).

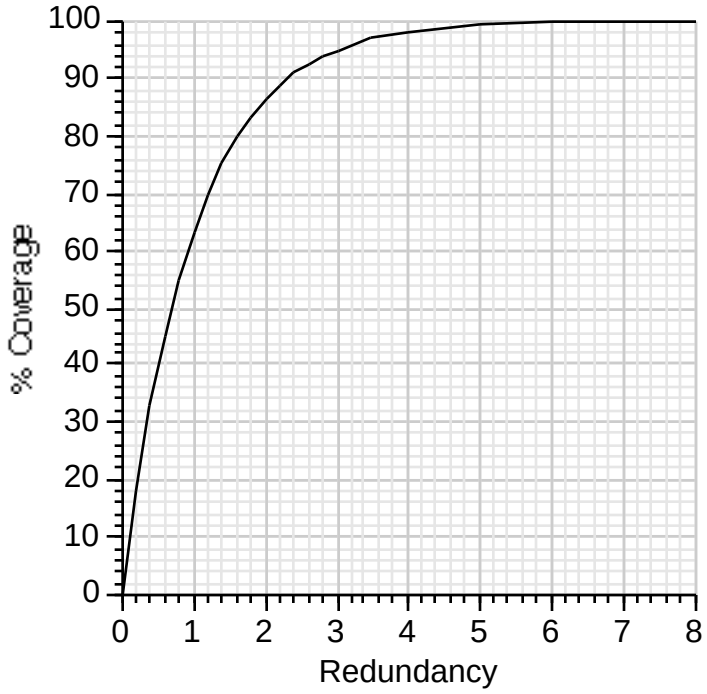

Figure 1.3. Expected coverage of a target with respect to redundancy (the Clarke-Carbon equation: Expected Fraction of Target Covered=1-e^{-R}).

that it has the form of an exponential decay, as we had anticipated. The Clarke-Carbon equation is diagrammed in Figure 1.3.

In future sections, I will return to the subject of coverage in somewhat more rigor and detail.

1.6 MATHEMATICAL MODEL - THE BETA DISTRIBUTION

Recall that the fragment start sites S_k are an ordered sample of n independently, identically, and uniformly distributed observations on the interval $(0,G_e)$. If one can specify the distribution of spacings between fragment start sites, then the number of gaps can be determined simply by counting the number of spacings greater than the length of a fragment. Furthermore, one can determine many other properties of interest by extending from this basic formulation. Before we go there, let us cover some basic definitions.

Let $D_k=S_{k+1}-S_k$ represent the distance between start sites, for $k=1,2,...,n-1$. 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.

Assume that an overlap of length at least T is necessary and sufficient to detect adjacency of two fragments. This is clearly a simplification over the complex reality of fragment assembly.[22] This assumption is very reasonable for shotgun sequencing projects, where $T<<L$, so variations in T have little effect on the mathematical model. Some mapping projects necessitate both large and varying overlaps, and in these cases, some care must be taken in interpreting the mathematical model. Note that the use of this simple assumption is quite necessary. Without it, the mathematical model quickly becomes very complicated, and in many cases intractable. When in doubt as to whether this assumption holds, computer simulations make an excellent adjuvant to the mathematical model.

Redundancy, R, is defined as $\frac{nL}{G}$. For notational ease, I also define the effective fractional coverage f_G of the target provided by one fragment as $\frac{L-T}{G_e}$, and the effective redundancy R_e as nf_G.

By genomics conventions, an island is a maximal set of fragments each of which is connected to all other island members by at least one path of fragments overlapping by T or more. A contig is an island consisting of a least two fragments (Staden, 1980). In the genomics community, the term “contig” is occasionally used loosely as a synonym for “island.” This is inconsistent with its original definition and can lead to linguistic imprecision; I discourage such usage. A contig consists of at least two overlapping fragments. An isolated fragment is not a contig; it is a singleton island.

In general, a target region not covered in any fragment is a gap. Adjacent islands are thus separated by gaps. The length of the gap between fragment start sites S_k and S_{k+1} is D_k-L, but a gap will only occur if $D_k>L$-T. If $D_k<L$-T, then the fragment that starts at S_k will extend at least T base pairs past the beginning of the fragment that starts at S_{k+1}, the overlap between these fragments will be detected, and no gap will occur in this interval. The two adjacent fragments will both be assigned to the same contig during assembly. Note that it is possible for the length of a gap to be negative. A negative gap length indicates that an overlap is present, but not detected. As mentioned previously, I am not counting the uncovered ends of

[22]Overlap is more appropriately expressed as a probability, not a certainty, and this "probability of overlap" is further affected when more than two fragments overlap at the same position. Also, repeated sequence elements in the target tend to decrease the probability of certain overlaps. These and other effects bring into question the use of the parameter T to model real projects. As long as $T<<L$, or if an “effective” T can be defined, the current formulation will result in an adequate model. An example of a project that would probably not meet this constraint would be an STS content mapping project.

the target as gaps.[23]

A simple geometric observation provides the basis and elegance for many of the equations derived from the mathematical model presented here: the domain space of the spacings D_k is the surface of the simplex $D_0+D_1+D_2+...+D_n=G_e$, and their joint probability density is constant.[24] Since $D_k \geq 0$, this surface will represent a line segment in one dimension (n=1), an equilateral triangle in two dimensions (n=2), a pyramid in three dimensions (n=3), and similar but hard to visualize symmetrical objects in higher dimensions. The n dimensional simplex is the shadow of the n+1 dimensional simplex. Executing a shotgun project with n fragments analyzed is exactly analogous to choosing a point at random from the n dimensional simplex. The Cartesian coordinates of this randomly chosen point represent D_0, D_1, D_2, ..., D_n. The symmetrical nature of all of the spacings is immediately apparent. In particular, the two end spacings, D_0 and D_n, will have the same distribution as all the other spacings. This last point is sometimes hard to visualize.

Another useful analogy to shotgun sequencing is to imagine a circular piece of string representing the genome. Take a pair of scissors and make n cuts at random places in the string. You will now have n pieces of string. These pieces of string do not represent the analyzed fragments of a shotgun project, but rather the spacings between the fragment start sites.[25] Clearly, by symmetry, the probability distribution for the length of each piece of string will be identical. This will not change if you start with a linear piece of string and make n-1 cuts. This is easy to see by recognizing that the linear piece of string may very well have just been the result of cutting a circular piece of string once. It doesn't matter if this first cut is in a non-random location, as long as all the other cuts are random.

These observations permit many probabilities of interest to be calculated by geometric considerations. Note again that in order for the model to work elegantly, G_e and D_k will usually be treated as continuous rather than discrete. This approximation is quite minor, given

[23]This is purely a semantic issue; it is very easy to modify the equations to account for a definition that defines uncovered target ends to be gaps. It is not quite sufficient just to add 2 to the number of gaps, as there is a small probability that either D_0 or D_n will equal 0, resulting in no gap at one target end or the other. Also, in a typical shotgun sequencing projects such end "gaps," at moderate to high redundancies, will actually be within vector sequence (see Section 1.4).

[24]To my knowledge, the first use of the simplex as an analogy for a problem of this nature was by Lévy (1939).

[25]Recall that all analyzed fragments have the same length.

the scope of a genome, or even a cosmid, compared with the unit of divisibility: a base pair.

The probability distribution of a single coordinate of a randomly chosen point from a simplex is well known and characterized. This distribution is called the "beta distribution."[26] I will use the beta distribution to obtain most of the results of interest to this present study. It should be noted that the beta distribution is a special case of a Dirichlet distribution. The Dirichlet distribution can characterize the joint distribution of all of the coordinates of a point on the simplex, rather than just one at a time. This can be important in some cases, because the length of one spacing can influence the length of another from the same project. For example, if we know that one of the spacings in a project is greater than half the length of the target, we know that none of the others are. In most cases however, I will be ignoring the correlation between spacing lengths. This is reasonable, since n is universally large for genome projects, so the correlation between any two given spacings is negligible.

An additional bookkeeping consideration should be mentioned. Strictly speaking, the beta distribution is defined on the interval [0,1]. However, the effective length of the target is not 1, but G_e. Some authors emphasizing mathematical purity address this issue by setting the genome size to 1, and allow the reader to retrospectively scale back results to the size of the genome. There is nothing wrong with such an approach. However, I prefer to maintain the proper proportionality throughout, as I this results in equations that are intuitively easier to grasp. Therefore, in what is to follow, I will make a few small deviations from notational orthodoxy for this purpose.

With these considerations, the density function for the beta distribution of the lengths of spacings between fragment start sites is:

$$f_{D_k}(x) = n\left(1 - \tfrac{x}{G_e}\right)^{n-1} \tag{1.12}$$

Now, the expectation of the beta distribution,

$$f_A(x) = n(1-x)^{n-1} \tag{1.13}$$

is easily verified to be:

$$E(A) = \frac{1}{n+1} \tag{1.14}$$

This makes intuitive sense, as we expect n+1 fragments when we make n cuts in a string of

[26]In particular, this is a special beta distribution, i.e., Beta(1,*n*). As a stylistic choice, I refer to the distribution employed in this paper as *the* beta distribution, rather than *a* beta distribution.

unit length. So we immediately have the expected value for the length of a spacing from equation (1.12):

$$E(D_k) = \frac{G_e}{n+1} \tag{1.15}$$

Notice that equation (1.15) specifies that the expected length of a spacing is equal to the effective target length divided by the number of spacings, which, for a linear target, is one greater than the number of fragments.

1.7 GAPS AND ISLANDS

As mentioned previously, a gap will occur following a fragment starting at S_k if and only if $D_k > L-T$. Thus, the probability of a gap, p_{gap}, following a given fragment is equivalent to the probability that $D_k > L-T$:

$$\begin{aligned} p_{\text{gap}} &= \int_{L-T}^{G_e} f_{D_k}(x)\,dx \\ &= \int_{L-T}^{G_e} n\left(1-\tfrac{x}{G_e}\right)^{n-1} dx \qquad \left(\text{now let } y = \tfrac{x}{G_e}\right) \\ &= n\int_{f_G}^{1} (1-y)^{n-1}\,dy \\ &= (1-f_G)^n \end{aligned} \tag{1.16}$$

Recall that my definition of "gap" precludes a gap from occurring in either the first or the last spacing. Therefore this "gap probability" does not apply to D_0 or D_n, although this does not alter the fact that these spacings are distributed identically to the other spacings.

Emphasizing 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 truth, there is a slight deviation from binomiality, as the occurrence of one gap will tend to inhibit the occurrence of others. Likewise, the absence of a gap in a short spacing will tend to promote the probability of a longer spacing with a gap. These effects will cause the actual distribution to have the same mean as the binomial distribution outlined here, but a smaller variance. The effects can be accurately modeled with a Dirichlet distribution.[27]

In a given project there are n-1 opportunities for a gap to occur, one between each pair

[27]The specific Dirichlet distribution is $D(1,1,1, \ldots ,1)$.

of adjacent fragment start sites. This gives us a binomial distribution for the number of gaps:

$$\begin{aligned} P\left(N_{\text{gaps}} = x\right) &\approx \binom{n-1}{x} p_{\text{gap}}{}^{x}\left(1-p_{\text{gap}}\right)^{n-1-x} \\ &= \binom{n-1}{x} \left(1-f_G\right)^{nx}\left(1-\left(1-f_G\right)^{n}\right)^{n-1-x} \end{aligned} \tag{1.17}$$

One immediately has the probability of project closure as:

$$\begin{aligned} P\left(N_{\text{gaps}} = 0\right) &\approx \left(1-p_{\text{gap}}\right)^{n-1} \\ &= \left(1-\left(1-f_G\right)^{n}\right)^{n-1} \end{aligned} \tag{1.18}$$

There is no particular need to approximate this equation, but we can if we want, again by noting that for small x, $e^{-x} \approx 1-x$. I will use this approximation twice in a row. So, continuing from equation (1.18),

$$\begin{aligned} P\left(N_{\text{gaps}} = 0\right) &\approx \left(1-e^{-R}\right)^{n-1} \\ &\approx \left(e^{-e^{-R}}\right)^{n-1} \\ &\approx e^{-ne^{-R}} \end{aligned} \tag{1.19}$$

The foregoing approximation is made primarily to draw a parallel with Siegel (1979), who provides an alternative derivation of equation (1.19) in more rigorous detail.

For the binomial distribution in equation (1.17), we have the expected number of gaps in a project as:

$$\begin{aligned} E\left(N_{\text{gaps}}\right) &= \left(n-1\right)p_{\text{gap}} \\ &= \left(n-1\right)\left(1-f_G\right)^{n} \end{aligned} \tag{1.20}$$

This is an exact equation. There is no particular reason to approximate it, except to illustrate parallels with other models, such as that of Lander and Waterman (1988). With this in mind, one could write equation (1.20) as follows:

$$E\left(N_{\text{gaps}}\right) \approx ne^{-R} \tag{1.21}$$

As noted above, the distribution for the number of gaps can be made exact with a Dirichlet distribution, which amounts to a summation of appropriate areas of an 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).[28]

[28]A perusal of the literature would be incomplete without a glance at Fisher (1940), in which there is some discussion of Stevens (1939). To my knowledge, the first genomics use of the equation of Flatto and Konheim (1962) was by Lange and Boehnke (1982).

Stevens' distribution is approximated by equation (1.17).

The number of islands will be one greater than the number of gaps, as each gap separates two adjacent islands. To write this definition as an equation,

$$N_{\text{islands}} = N_{\text{gaps}} + 1 \tag{1.22}$$

The expected number of islands is therefore:

$$E\left(N_{\text{islands}}\right) = 1 + (n-1)\left(1 - f_G\right)^n \tag{1.23}$$

1.8 THE NUMBER OF CLONES IN AN ISLAND

Any fragment start site spacing longer than *L*-*T* will create a gap. All other spacings will not. This divides the spacings into two categories: those that form gaps and those that do not. Now the order of spacings is arbitrary, so all possible orders of gap-forming spacings amongst non gap-forming spacings are equally likely. This observation will permit us to determine the distribution of the number of clones in an island as well as the island length.

Let z_m specify the number of fragments in the m^{th} island in a project. Now, 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:

$$E\left(z_m \middle| N_{\text{islands}}\right) = \frac{n}{N_{\text{islands}}} \tag{1.24}$$

This simple result is easy to obtain and whets our appetite for things to come. It also will permit us to verify that the mean of the distribution that I derive meets this expectation.

To obtain the probability distribution of z_m, formally divide the spacings $\{D_k | k=1,2,..., n\text{-}1\}$ into two subsets: those D_k>L-T and those $D_k \leq$L-T. The number of spacings in the first subset is N_{gaps}. I will refer to these spacings as long spacings. Since there are n-1 total spacings, the number of spacings in the second subset is n-1-N_{gaps}. I will refer to these as short spacings.

Each island is bounded by two long spacings. An exception may occur for the island that begins with the first fragment, starting at S_1. This is because D_0 might be a short spacing. Likewise, the last island ending with the fragment starting at S_n will be bounded on the end by a short spacing if D_n is short.

The number of fragments in an island is equal to one plus the number of short spacings between its two bounding long spacings.[29] The shorter the spacings are in an island, the more

"piled up" will be the fragments in that island. This will result in multiple coverage of areas of the island. Multiple coverage is more common at higher redundancies. In fact, the average multiplicity of coverage is exactly equal to the redundancy.[30]

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, 1897b; also Baticle, 1935), but to maintain simplicity one may employ a continuous approximation analogous to that employed previously to model spacing length in equation (1.12). This approximation will be good as long as the number of long spacings is small compared to the total number of spacings. This will be the case at higher redundancies when there are few singleton islands. The approximation should also be acceptable at lower redundancies.

During the actual assembly of a shotgun project, even at high redundancies, some analyzed fragments never get assembled into any contigs. This will be true even after the project is completed and there are no gaps. The reason for the continued existence of these singleton islands is that they represent "orphaned" fragments. These fragments may represent extremely poor quality sequence reads, a mislabeled or mishandled clone, contamination from another project, or contamination from a vector organism such as *E. coli*.[31] Since these singleton islands exist at high redundancies, it can be quite deceiving to compare the average number of fragments in islands from an actual project to the expected number of fragments predicted by a mathematical model. It is more informative to examine the distributions of the number of fragments in the larger islands. An excellent way to do this is graphically, by a bar graph, for example. I would recommend the inclusion of such a graphical comparison tool in any shotgun assembly computer program, particularly as it is relatively simple to program. A valuable use of such a tool would be to detect the presence of a significant number of orphaned clones in a project by noting a deviation from the expected number of singleton islands without deviations in other areas. An overall decrease in the number of clones in

---

[29]If the last spacing is short, then the number of fragments in the last island will be one plus the number of short spacings following its initiating long spacing. Regardless of whether or not the first spacing is short, the number of fragments in the first island will be equal to the number of spacings preceding the first long spacing other than D_0. I ignore these minor effects in the main discussion, but they may be accounted for, if desired, at the cost of a little algebra.

[30]This should not be surprising.

[31]Now that the complete genome sequence of *E. coli* is known (Blattner et al., 1997), this last source of orphaned fragments should be a bane of the past.

islands from their predicted numbers might suggest a problem with detecting overlaps. Comparisons of reality to expectations can be extremely valuable in troubleshooting problems during the course of a project.

To continue with our task at hand, we seek to know the distribution of the number of short spacings that lie between two long spacings. We will assume that there are plenty of short spacings, so that we can treat this number as a continuous variable. We immediately recognize that we are presented with the same problem of determining the lengths of pieces of string after random cuts have been made in an original piece. The cuts are the long spacings. Recall that since all orderings of spacings are equally likely, these cuts can be considered to randomly (i.e., independently and identically) distributed. The length of the string is equal to the number of short spacings. We can thus employ the beta distribution to model the number of short spacings bounded by two long spacings. There is thus a curious methodological symmetry between determining the distribution of the lengths of the spacings and determining the distribution of the number of spacings in an island.

The long spacings, or gaps, are uniformly distributed over the continuous interval [0,n-1-N_{gaps}]. As before, we will need to scale the beta distribution from its defined domain of [0,1], this time by a factor of n-1-N_{gaps}. The conditional probability density for the number of short spacings bounded by two long spacings is therefore:

$$f\left(x|N_{\text{gaps}}\right) \approx N_{\text{gaps}} \left(1 - \frac{x}{n-1-N_{\text{gaps}}}\right)^{N_{\text{gaps}}-1} \tag{1.25}$$

This is written explicitly as an approximation because it represents a continuous approximation of a discrete phenomenon. Note also that it is conditioned on the number of gaps in an assembled project. This is not a problem when comparing the state of an actual project to its expected state based on modeling, as the number of gaps in the actual project will be known. It is a problem when using the model as a predictive theoretical tool for planning a project. In this case, this distribution can be evaluated approximately by using the expected number of gaps (equation (1.20)), or at the cost of a little extra algebra it can be evaluated more precisely by employing a probability weighted summation over all possible values for N_{gaps}. At the extreme end of precision, a combinatorial approach could be used, but this would be straying quite far away from the ideal of elegancy in equations.

Note that it is the number of short spacings in an island that is beta distributed. The number of fragments in an island is one greater than the number of short spacings confined by the two long spacings that bound the island. Recall that there is a fragment associated with the

terminating long spacing of an island. So with z_m as the number of fragments in an island, z_m-1 is the number of short spacings in that island. The conditional probability density for z_m is therefore:

$$f_z\left(x\middle|N_{\text{gaps}}\right) \approx N_{\text{gaps}}\left(1-\frac{x-1}{n-1-N_{\text{gaps}}}\right)^{N_{\text{gaps}}-1} \tag{1.26}$$

An additional reason to write equation (1.26) as an approximation is that here the "edge effects" of the first and last island are ignored.[32]

Recalling equation (1.14), which gives the expected value of a beta distribution, one can calculate the expected number of clones in an arbitrary island (conditioned on the number of gaps):

$$E\left(z_m\middle|N_{\text{gaps}}\right) \approx \frac{n-1-N_{\text{gaps}}}{N_{\text{gaps}}+1}+1 = \frac{n}{N_{\text{islands}}} \tag{1.27}$$

Note that equation (1.27) was anticipated by equation (1.24).

The fraction of singleton islands expected in a project can be obtained by integrating the probability density in equation (1.26) over the range $x \in [1,2)$; the remaining islands will be contigs. As mentioned above, due to the approximation of continuity, equation (1.26) is most valid at high redundancies, where there are few singleton islands. Therefore, if just the number of singletons is sought, then it is more accurate to calculate this more simply as the probability that a spacing is long and is immediately followed by another long spacing times the total number of non-end spacings (i.e. excluding D_0, D_1, and D_n). Additionally, if D_1 is long, a singleton will occur starting at S_1, and if D_{n-1} is long, a singleton will occur starting at

---

[32]An anonymous reviewer of Roach (1995) suggested the following equation for the number of fragments in an island:

$$P\left(z \geq x\middle|N_{\text{gaps}}\right) = \prod_{b=1}^{N_{\text{gaps}}}\left(1-\frac{x-1}{n-b}\right)$$

This discrete form of equation (1.26) is precisely analogous to the continuous Beta(1,n) distribution used in the text. Considering the availability of computers, there is no reason not to use this discrete form in place of the easier-to-manipulate equation used in the main body of the text. Note that although the discrete equation is an equality, it still requires conditioning on N_{gaps}, which will be influenced to a small extent by "edge effects." The expected value of this equation can be calculated as:

$$E\left(z\middle|N_{\text{gaps}}\right) = \sum_{x=1}^{n-N_{\text{gaps}}}\prod_{b=1}^{N_{\text{gaps}}}\left(1-\frac{x-1}{n-b}\right) = \frac{n}{N_{\text{gaps}}+1}$$

Working the algebra of this last equality can be amusing.

S_n. This results in:

$$E\left(N_{\text{singletons}}\right) \approx \left(p_{\text{gap}}\right)^2 (n-2) + 2\,p_{\text{gap}} \tag{1.28}$$

The distribution of singletons can be well approximated, if desired, with binomial considerations, or by making use of the discrete equation given in the last footnote.

Some motivation 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 (1897a) shows that for a given project, if the islands are ordered by increasing number of fragments, the expected number of fragments in the x^{th} smallest island is:

$$E\left(\text{number of fragments in the } x^{\text{th}} \text{ smallest island} \middle| N_{\text{gaps}}\right) =$$

$$1 + \frac{n-1-N_{\text{gaps}}}{N_{\text{gaps}}} \sum_{1}^{x} \frac{1}{N_{\text{gaps}} - x + 1} \tag{1.29}$$

This expected value may be substituted in equation (1.32) below, and enables the prediction of the longest expected island for a project. A couple of points should be addressed, however. First, because the expectation is conditioned on N_{gaps}, to be useful in a predictive manner, a probability weighted summation would have to be employed. However, since equation (1.29) will be evaluated by a computer anyway, this extra computation will perhaps not be extremely tedious. Secondly, Whitworth's equation does not address higher moments, such as the variance. There is likely to be high variance in this statistic, at least for the longest island. Thus, perhaps the best use of equation (1.29) is as a curiosity. It is nevertheless valuable to bring to light Whitworth's historic contribution to this field.

1.9 ISLAND LENGTH

The distribution of the number of clones in an island enables the determination of the distribution of the length of that island. Each island is the union of one or more fragments starting at base pairs S_k, S_{k+1}, S_{k+2}, ... , 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 (see Figure 1.2). Thus,

$$l_m = \begin{cases} L + \sum_{x=k}^{k+z_m-2} D_x & \text{if } z_m > 1 \\ L & \text{if } z_m = 1 \end{cases} \tag{1.30}$$

Now, 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:

$$E\left(l_m \middle| z_m\right) = L + E\left(D_k\right)\left(z_m - 1\right) \tag{1.31}$$

By assuming that z_m is equal to its average value, one may approximate expected island length as:

$$\begin{aligned} E\left(l_m\right) &\approx L + E\left(D_k\right)\left(E\left(z_m\right) - 1\right) \\ &\approx L + G_e\left(\frac{1}{n+1}\right)\left(\frac{n}{E\left(N_{\text{islands}}\right)} - 1\right) \\ &= L + G_e\left(\frac{1}{n+1}\right)\left(\frac{n}{1+(n-1)\left(1-f_G\right)^n} - 1\right) \end{aligned} \tag{1.32}$$

This approximation is most valid when the relative variance of z_m is small, often the 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.

Spacings greater than L-T form gaps, so are not included in the subset of spacings that may be included in the length of an island. To proceed, one must eliminate these spacings from the distribution of D_k (equation (1.12)) by truncating and normalizing. The expected value of this truncated distribution is:

$$\begin{aligned} E\left(D_k \middle| D_k \le L - T\right) &\approx \frac{\int_0^{L-T} xn\left(1-\frac{x}{G_e}\right)^{n-1} dx}{\int_0^{L-T} n\left(1-\frac{x}{G_e}\right)^{n-1} dx} \qquad \left(\text{now let } y = \tfrac{x}{G_e}\right) \\ &= G_e \frac{\int_0^{f_G} y\left(1-y\right)^{n-1} dy}{\int_0^{f_G} \left(1-y\right)^{n-1} dy} \\ &= G_e \frac{\left[\frac{\left(1-y\right)^{n+1}}{n+1} - \frac{\left(1-y\right)^{n}}{n}\right]_0^{f_G}}{\left[-\frac{\left(1-y\right)^{n}}{n}\right]_0^{f_G}} \\ &= G_e \frac{1-\left(1+nf_G\right)\left(1-f_G\right)^n}{\left(n+1\right)\left(1-\left(1-f_G\right)^n\right)} \end{aligned} \tag{1.33}$$

Note that the only reason that this equation is written as an approximation is the assumption of continuity. Thus, it probably would also have been reasonable for me to have written it as an

equality.[33] Substituting equation (1.33) into equation (1.32), we have:

$$E(l_m) \approx L + G_e \left(\frac{1-(1+nf_G)(1-f_G)^n}{(n+1)\left(1-(1-f_G)^n\right)} \right) \left(\frac{n}{1+(n-1)(1-f_G)^n} - 1 \right) \tag{1.34}$$

The accuracy of our calcualtion of expected island length can be further improved with the aid of a computer by summing equation (1.31) over all possible values of z_m, rather than employing the expected value of z_m. Summing over all possible values of z_m, we have:

$$\begin{aligned} E(l_m) &= L + E(D_k) \sum_{i=1}^{n} P(N_{\text{islands}} = i) \sum_{m=1}^{i} (z_m - 1) \\ &= L + E(D_k) \sum_{i=0}^{n-1} P(N_{\text{gaps}} = i) \left(E(z_m | N_{\text{gaps}} = i) - 1 \right) \\ &= L + E(D_k) \sum_{i=0}^{n-1} P(N_{\text{gaps}} = i) \left(\frac{n}{1+i} - 1 \right) \\ &\approx L + G_e \frac{1-(1+nf_G)(1-f_G)^n}{(n+1)\left(1-(1-f_G)^n\right)} \sum_{i=0}^{n-1} \binom{n-1}{i} (1-f_G)^{ni} \left(1-(1-f_G)^n\right)^{n-1-i} \left(\frac{n}{1+i} - 1 \right) \end{aligned} \tag{1.35}$$

Equation (1.35) is very accurate (Figure 1.4), and is calculable in seconds with software such as *Mathematica 4.0* (Wolfram Research). However, at high redundancies, equation (1.32) or equation (1.34) offer sufficient accuracy for most genomic purposes. For historical reasons, equation (1.34) was used to calculate expected island length values presented in this thesis, except where mentioned otherwise.

1.10 TARGET COVERAGE (2)

I introduced the Clarke-Carbon equation in Section 1.5. I now consider a couple of alternative ways to derive this equation. The advantage of considering these different methodologies is primarily to gain insight. In addition, this will give me an opportunity to address the mathematical history of coverage problems. This will permit a digression on the higher moments of the coverage, such as the variance. It will also allow us to interpret

---

[33]The reader may have noticed that I spend some effort discussing whether or not these equations are exact or approximations, and why. These were points of misunderstanding by anonymous reviewers of Roach (1995), so I felt it prudent to spend the extra effort here to elucidate. An example at a ludicrous extreme makes it clear that equation (1.33) is an approximation. The reader is encouraged to explore the parameterization of $G=3$, $L=2$, $T=1$, and $n=2$. The form of the last line of the equation was also suggested to me by the anonymous reviewer. The reviewer felt this expression most effectively brought out the subtle difference between the form of $E(D_k)$ in the equation and the expression $G_e/(n+1)$.

situations where the apparent coverage is greater than the target length.

Perhaps the most obvious way to calculate target coverage is by multiplying the number of islands by their expected length:

$$\begin{aligned} \text{Coverage} &= E\left(N_{\text{islands}} E\left(l \middle| N_{\text{islands}}\right)\right) \\ &\approx E\left(N_{\text{islands}}\right) E(l) \end{aligned} \tag{1.36}$$

In order to demonstrate rough equivalence with the Clarke-Carbon equation, we can continue to make rough approximations by substituting equations (1.23) and (1.32) into equation (1.36). Equation (1.23) can be approximated as follows:

$$\begin{aligned} E\left(N_{\text{islands}}\right) &\approx \left(1 + (n-1)\left(1 - f_G\right)^n\right) \\ &\approx 1 + n\left(1 - f_G\right)^n \\ &\approx 1 + ne^{-R} \end{aligned} \tag{1.37}$$

Equation (1.32) can be approximated as:

$$\begin{aligned} E(l) &\approx L + G_e \left(\frac{1}{n+1}\right) \left(\frac{n}{1 + (n-1)\left(1 - f_G\right)^n} - 1\right) \\ &\approx \frac{G}{n} \left(\frac{n}{1 + ne^{-R}} - 1\right) \end{aligned} \tag{1.38}$$

Combining equations (1.37) and (1.38) gives us the Clarke-Carbon equation:

$$\begin{aligned} \text{Coverage} &\approx E\left(N_{\text{islands}}\right) E(l) \\ &\approx \left(1 + ne^{-R}\right) \left(\frac{G}{n}\right) \left(\frac{n - 1 - ne^{-R}}{1 + ne^{-R}}\right) \\ &= G\left(1 - \frac{1}{n} - e^{-R}\right) \\ &\approx G\left(1 - e^{-R}\right) \qquad \left(\text{if } \tfrac{1}{n} << 1\right) \end{aligned} \tag{1.39}$$

The coverage may also be calculated by subtracting the sum of the gap lengths from the total target length. This would entail the following:

$$\text{Coverage} \approx G - \left(N_{\text{islands}} - 1\right)\left(E\left(D_k \middle| D_k \geq \text{Gap length}\right) - L\right) \tag{1.40}$$

This calculation will be saved as an exercise for the reader. The necessary integral can be evaluated similarly to that of equation (1.33).

Note that an excess of negative gap lengths will result in clonal coverage in apparent excess of the total target length.[34] This is most apparent when T is large. This situation has been known to occur in some physical mapping projects. An apparent coverage in excess of the target length is a poor prognosticator for the efficiency of overlap detection.[35] If the

"actual" coverage is desired, the length of a gap should be calculated as D-L for $D>L$; alternatively, T can be set equal to zero.

Let us consider the issue of coverage a little more generally. Problems of coverage have intrigued mathematicians for some time. Perhaps the first "useful" application of such mathematics occurred in World War II. In addition to providing the genesis for the discipline of operations research, World War II stimulated interest in coverage problems and their relation to strategic bombing.[36] The exact location of bomb and shell hits was largely a random process, so the percent of the target area affected by explosions could be calculated using an approach similar to the one that I used in Section 1.5. Robbins (1944) dealt with this problem more explicitly and rigorously.[37]

In brief, 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 are calculated by Robbins, with:

$$E(f) = \frac{1}{G}\sum_{x=1}^{G}\left(1-\left(1-p_x\right)^n\right) \tag{1.41}$$

When $p_x=\frac{L}{G}$ for all x, this expected value is approximated by equation (1.11). The exact manner of coverage is not important. That is, p_x times G equals the number of base pairs sequenced in each of the n coverage iterations, regardless of whether or not the base pairs are contiguous. Therefore this equation will hold for the pairwise projects discussed in the next chapter. For linear targets p_x is not constant, and falls off near the edges, as discussed in Section 1.5. Despite this, unless L is a significant fraction of G, the Clarke-Carbon equation remains an adequate approximation to that of Robbins (equation (1.41)).

Note that if coverage is determined by the method of Robbins, or by the Clarke-Carbon equation, the expected island length can be calculated directly, rather than using the

---

[34]This is something that will not be predicted from the simple application of the Clarke-Carbon equation and is one of the advantages of the present methodology.

[35]Other conditions, such as extreme library contamination, could also produce this effect. In any case, it is not a good sign.

[36]I mention operations research here because operations research methodology has much to offer genomics. See, for example, papers by Siegel et al. (1998a; 1998b; 1999).

[37]Robbins was unaware that he was working on a genomics problem. Thus by employing his equations in this context, we are in essence beating swords into plowshares.

approach of Section 1.9. This is done merely by dividing the coverage by the expected number of islands. However, such an approach removes any insight into the distribution of island lengths.

1.11 COMPARISON WITH THE LANDER-WATERMAN EQUATIONS

In 1988, in a watershed paper, Lander and Waterman published a model which formed a cornerstone of strategic genomic analysis. Their main concern in this model was to provide a mathematical model for the early physical mapping efforts underway at that time. Many of these efforts were done at low redundancies with the goal of building partial fragmented maps. The equations were not intended to model shotgun sequencing. As a result, the equations were valid at low redundancies, but not at high redundancies. Unfortunately, many subsequent workers have misinterpreted these results and applied them erroneously to high redundancy situations such as more advanced maps or to shotgun sequencing.[38] This has led to some remarkably incorrect claims about the state, or expected state, of completion of several projects.

The Lander-Waterman (L-W) equations were reworked with more painstaking detail by Port et al. (1995); I will use this more recent paper as a reference for the comments that follow. Two relevant L-W results are:

(I.i) The expected number of islands:

$$\begin{aligned} E\left(N_{\text{islands}}\right) &= ne^{-R\left(1-\frac{T}{L}\right)} \\ &\approx ne^{-R} \qquad \left(\text{if } \frac{T}{L} << 1\right) \end{aligned} \tag{1.42}$$

(I.v) The expected length of an island:

$$\begin{aligned} E(l) &= L\left(\frac{e^{R\left(1-\frac{T}{L}\right)}-1}{R}+1-\frac{T}{L}\right) \\ &\approx L\left(\frac{e^{R}-1}{R}+1\right) \qquad \left(\text{if } \frac{T}{L} << 1\right) \end{aligned} \tag{1.43}$$

---

[38] Surprisingly, there is neither discussion of the limits of the accuracy nor simulations for the equations in Lander and Waterman (1988). It is unfortunate that several of the figures in this paper graph equations into moderate to high redundancies where inaccuracies occur, making it easier for a casual reader to be misled.

I provide the limits of automatic overlap detection (i.e., T=0) to make the following discussion clearer.

The most striking thing about the L-W equations is their behavior as the redundancy grows. To wit,

$$\lim_{R\to\infty} E\left(N_{\text{islands}}\right) = 0 \tag{1.44}$$

and

$$\lim_{R\to\infty} E(l) = \infty \tag{1.45}$$

Clearly, the expected number of islands at high redundancy is one single island that covers the target. The length of this island should be equal to the target length. An island length that approaches infinity, or that even exceeds the target length is ludicrous.[39] Therefore the L-W equations are not accurate at high redundancy. Let us determine how high one must go before they reach their limit of accuracy.

In the present model, let us approximate equation (1.23) as follows:

$$\begin{aligned} E\left(N_{\text{islands}}\right) &= 1 + (n-1)\left(1 - f_G\right)^n \\ &\approx 1 + ne^{-R} \end{aligned} \tag{1.46}$$

Now if $ne^{-R} \gg 1$, then we can claim that the L-W equation will approximate the current model. This will occur when $n \gg e^R$.

Likewise, in the present model,

$$\begin{aligned} E(l) &\approx \frac{G}{n}\left(\frac{n}{1+ne^{-R}} - 1\right) \qquad \left(\text{now let } ne^{-R} >> 1\right) \\ &\approx \frac{G}{n}\left(\frac{1}{e^{-R}} - 1\right) \\ &\approx \frac{G}{n}\left(e^R - 1\right) \end{aligned} \tag{1.47}$$

Again, the models are almost exactly equivalent as long as $n \gg e^R$. This defines an upper bound as a function of n (or R) for the accuracy of the L-W equations. This bound occurs at redundancies in excess of threefold, depending on the exact parameterization of the project. A

[39]Note that the sum of the lengths of two or more islands can exceed the target length if the overlap parameter T is large. No single island can exceed the target length. Also, it is nonsensical for T to be greater than L, so in no case should the sum of island lengths be greater than 2G. Even this would be ludicrous.

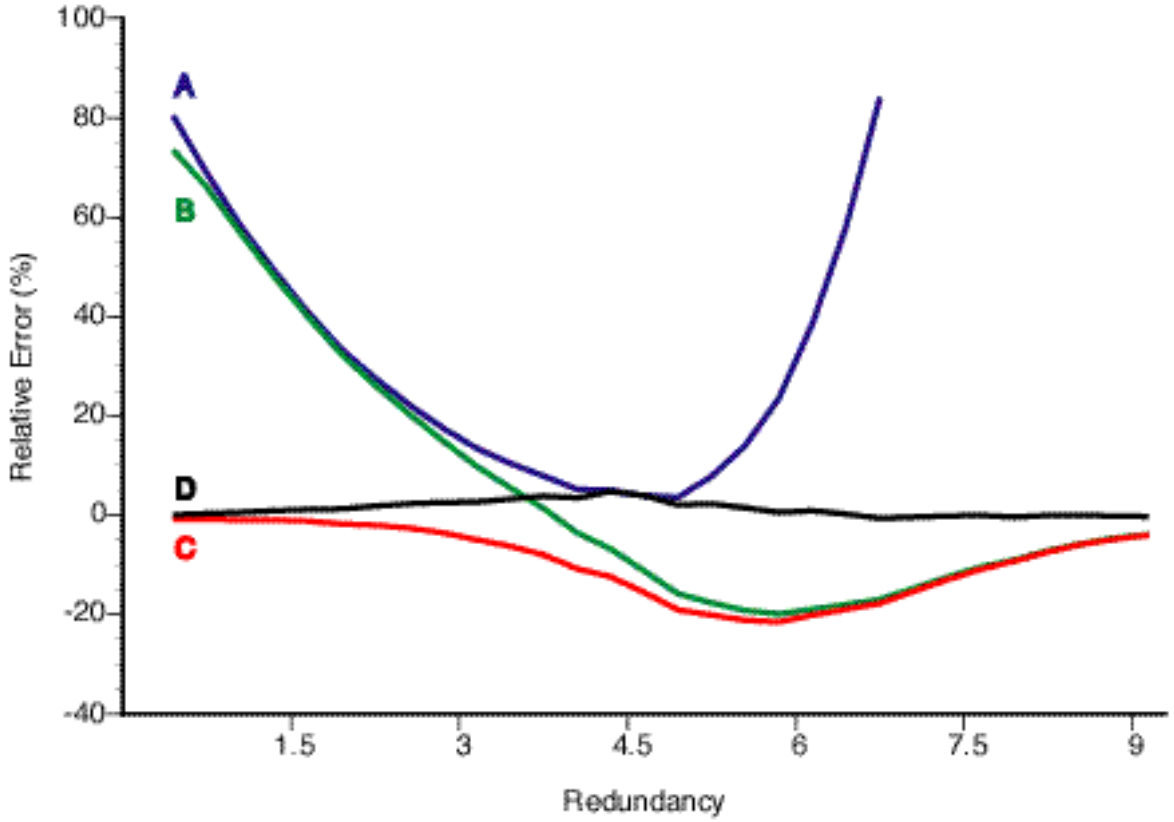

Figure 1.4. Relative error of the models. The predicted average island lengths are plotted with respect to island lengths generated from simulations (L=600, T=20, G=40000). Each simulation data point is the mean value of the mean island length from 10000 project simulations. A, Lander-Waterman; B, Equation (1.32); C, Equation (1.34); D, Equation (1.35).

graph of the percent relative error of the L-W and the model of this thesis relative to average island lengths from simulated projects is shown in Figure 1.4. Equation (1.35) is better at all redundancies, and considerably better at high redundancies. The high excess error of the L-W and equation (1.32) curves at low redundancies is largely due to failure to take into account the adjustment of D_k computed in equation (1.33). The dip of equation (1.32) and equation (1.34) at moderate redundancies is largely due to the higher relative variance of z_m at these redundancies, which is the correction provided by equation (1.35).

The exponential rise of the L-W equations at higher redundancies results from several assumptions that were necessary for the L-W model. First, the fragment start points (S_k) are assumed to follow a Poisson process. This is a deceptively facile assumption to make. In particular, a Poisson distribution looks very much like a beta distribution, so it would seem

legitimate to expect only small differences between the two models. However, the problem with using a Poisson distribution lies with certain assumptions of independence. A Poisson distribution assumes that the distance between two start sites is completely independent of the distance between any other two start sites. Thus the use of a Poisson model precludes analyzing any problems that involve fragment dependence, such as those discussed in the next two sections.

A consequence of using the Poisson distribution is that the exact number of clones in a project can not be prespecified.[40] This problem is noted by Port et al. (1995). As a result, one immediately has a feeling of uneasiness while using the model. From a biological point of view, this parameter (n) is precisely the parameter over which the investigator has the most knowledge and control. However, it is not clear that this consequence is particularly damaging.

The most damaging assumption for the L-W equations is that they assume that an island must always be bounded on its right by a long spacing ($D>L-T$). In fact, there is usually an island that does not meet this assumption. The reader is asked to visualize this island before reading the answer in the footnote.[41]

If the last spacing D_n is short, then the L-W equations will undercount the number of islands by one. This is not a very big deal if the number of predicted islands is large, but as the number of islands shrinks, an extra island becomes a significant factor. In fact, at even moderate redundancies, the L-W equations will predict less than one island in a project. This is clearly ludicrous.[42] Nor can the problem be dismissed by suggesting that a miscount by a single island is a minor deviation. Project closure is defined by the arrival at a state of one

[40]Suppose n was prespecified. Then there would be a random and independent distance between each clone start point. The sum of these distance would not necessarily add up to G. It would be close, but not quite. A few clones would have to be added to or subtracted from the mathematical model, bringing it out of synch with the reality of the project. It is not immediately clear, however, that this particular objection will cause major consequences to the accuracy of any derived equations.

[41]The rightmost island in a project is bounded by the spacing D_n, which at high redundancies is more likely to be short than long.

[42]An individual at a scientific meeting once gave a lecture in which they claimed that, because the L-W equations predicted less than one island for their then-underway genome mapping project, they were therefore nearly certain to have no gaps in their project. The importance of making the limitations of mathematical models clear to the end user cannot be overemphasized.

island. If one cannot predict when this state occurs, then this most important question becomes unanswerable.

Furthermore, the prediction of island length is profoundly altered by the miscount of islands. Island length is, in essence, calculated as the reciprocal of the number of islands times the coverage. Thus, the difference of a miscount of one in the estimated number of islands, say between 0.1 and 1.1, results in a tenfold overestimate of expected island length. This miscount of one island is noted in Port et al. (1995), but no solution is offered, nor are the bounds of island length accuracy determined.

It might be naively suggested that the L-W equations can be "fixed" merely by adding one to the number of islands predicted. Unfortunately, the L-W equations are only off by *exactly one* in the limit as redundancy tends towards infinity. Below this limit, they are off by slightly less than one. Therefore, the determination of island length and project closure probabilities would remain open questions. Thus, what appears to be a subtle problem is actually somewhat difficult to fix, and requires that a Poisson model be discarded so that the interdependence of fragment spacings can be accounted for. What appear to be mere "edge effects" shake the foundation of the model. Nor does it turn out that increasing the genome size will diminish these "edge effects." The reason for this is that in real projects the number of fragments n increases proportionally with the genome size to keep the redundancy R constant. In order for the "edge effects" to be negligible, the number of islands must be large, which occurs only at low redundancies, regardless of the genome size.[43]

One additional note: The L-W equations were developed for linear targets. Nevertheless, as we shall see in Section 1.13, it turns out that they are more accurate when applied to circular targets (i.e., bacterial genomes).[44]

1.12 ISLAND CO-DEPENDENCY

One must be careful 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

---

[43]It might be argued that I have shrugged off some "edge effects" during the course of the development of the "beta" model. This was done with due consideration for their effects on the relevant variables. The reader is encouraged to verify that the impact of these particular "edge effects" on calculations of interest to genomics is negligible.

[44]Some intuition for this at first surprising result can be gained by re-reading the last few paragraphs. In some ways, a circular target *is* an infinite target, and has no "edge effects."

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). There are several calculations where it is no longer sufficient to assume that the sizes of the spacings, gaps, and islands are independent of each other. These calculations are somewhat more arcane than what has been presented hitherto, and tend to have more of a niche utility. Nevertheless, from time to time they can be useful. The calculations also provide a powerful illustration of the mathematical model constructed here, as it is extremely difficult to accurately adapt other models to the same ends.

For example, one may wish to consider the probability that a project contains at least *i* islands of length greater than a certain critical length *C*. This sort of calculation may be useful in evaluating performance of fragment-assembly algorithms or in planning projects with limited goals. Such a limited goal might be to sequence or map until at least one contig greater than a fixed length has been obtained, perhaps as part of an exploratory or preliminary survey of a target or genome.

Let N_{long} be the number of long spacings. Let g_m be the number of short spacings contained between adjacent long spacings. Now,

$$l_m \approx L + E(D)g_m \tag{1.48}$$

Thus:

$$P(l_m > C) \approx P\left(g_m > \frac{C - L}{E(D)}\right) \tag{1.49}$$

The number of short spacings preceding the first long spacing is g_0; the number following the last long spacing is $g_{N_{\text{long}}}$.[45] Now, since the distribution of long spacings is uniform among the set of all spacings, the distribution of g is defined on the simplex

$$\frac{g_0 + g_1 + g_2 + \ldots + g_{N_{\text{long}}}}{n + 1 - N_{\text{long}}} = 1 \tag{1.50}$$

with all points of the simplex equiprobable. The reader should be by now familiar with the previous invocations of this analogy. To continue, we might employ the Dirichlet distribution instead of the beta distribution. However, rather than to invoke its full complexity here, I will borrow only a result from Stevens (1939), who was the first to address this particular aspect of the distribution.

---

[45]Since 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 $N_{\text{gaps}} >> 2$. At high redundancies $n >> N_{\text{long}}$, so it is unlikely that either D_0 or D_n is long. If greater accuracy is desired, an appropriate summation can be made.

Let R be the number of islands exceeding length *C*, let $c = \frac{C - L}{E(D)\left(n + 1 - N_{\text{long}}\right)}$, and let *k* be the greatest integer less than $\frac{1}{c}$. Then one has directly from Stevens:

$$P\left(R \geq i \middle| N_{\text{gaps}}\right) \approx \sum_{j=1}^{k} (-1)^{j-1} \frac{\left(N_{\text{gaps}} + 1\right)!}{\left(N_{\text{gaps}} + 1 - j\right)!(j-i)!(i-1)!} \frac{(1 - jc)^{N_{\text{gaps}}}}{j} \tag{1.51}$$

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

$$\begin{aligned} P\left(\text{one contig} > \tfrac{G}{2}\right) &\approx \sum_{v=0}^{n} P\left(N_{\text{gaps}} = v\right) P\left(\text{one contig} > \tfrac{G}{2} \middle| N_{\text{gaps}} = v\right) \\ &\approx \sum_{v=0}^{n} \binom{n-1}{v} \left(1 - f_G\right)^{vn} \left(1 - \left(1 - f_G\right)^{n}\right)^{n-1-v} \frac{v+1}{2^v} \end{aligned} \tag{1.52}$$

At moderate to high redundancies, only the first few terms of this last sum are necessary, as the subsequent terms rapidly diminish. This formula is best evaluated by a computer.

1.13 CIRCULAR TARGETS

The equations presented in the foregoing sections were designed explicitly with linear targets in mind. It is somewhat simpler, however, to write similar equations for circular targets. The choice to begin with equations for linear targets was motivated by the fact that the vast majority of targets are linear. Even targets that are seemingly circular, such as cosmids and BACs, need to be treated as linear due to the presence of the vector sequence. In current practice, only bacterial genomes would be modeled as circular targets.[46]

I will briefly recap the discussion of my mathematical model, with appropriate modifications for circular targets. Note that G_e=G, so that S_k [1,G]. D_k=S_{k+1}-S_k is the distance between start sites, for k=1,2,..., n-1. D_n is the distance between S_n and S_1. We therefore have the underlying beta distribution (cf. equation (1.12)):[47]

$$f_{D_k}(x) = \begin{cases} (n-1)\left(1 - \frac{x}{G}\right)^{n-2} & n > 1 \\ \delta(G) & n = 1 \end{cases} \tag{1.53}$$

---

[46] Organellar genomes are also circular, but due to their small size they are seldom targets for random subcloning.

[47] $\delta(n)$ is the Dirac-delta function. In this case, it implies that with only one fragment in a project, then the sole spacing length is G.

The expected spacing length for the circle is (cf. equation (1.15)):

$$E(D_k) = \frac{G}{n} \tag{1.54}$$

The gap probability for the circle is (cf. equation (1.16)):

$$p_{\text{gap}} = (1 - f_{\text{G}})^{n-1} \tag{1.55}$$

The gap distribution for the circle is (cf. equation (1.17)):

$$\begin{aligned} P(N_{\text{gaps}} = x) &\approx \binom{n}{x} p_{\text{gap}}{}^{x} (1 - p_{\text{gap}})^{n-x} \\ &= \binom{n}{x} (1 - f_{\text{G}})^{(n-1)x} \left(1 - (1 - f_{\text{G}})^{n-1}\right)^{n-x} \end{aligned} \tag{1.56}$$

The approximation for the circle closure probability is (cf. equation (1.18)):

$$\begin{aligned} P(N_{\text{gaps}} = 0) &\approx (1 - p_{\text{gap}})^{n} \\ &= \left(1 - (1 - f_{\text{G}})^{n-1}\right)^{n} \end{aligned} \tag{1.57}$$

The expected number of gaps in a circle is (cf. equation (1.20)):

$$E(N_{\text{gaps}}) = n(1 - f_{\text{G}})^{n-1} \tag{1.58}$$

Note that the calculation in equation (1.58) for the expected number of gaps in a circular target is exact. This is because there are no "edge effects" for a circle.[48]

The number of islands will equal the number of gaps, unless there are zero gaps, in which case there will still be one island. So (cf. equation (1.22)),

$$N_{\text{islands}} = \begin{cases} N_{\text{gaps}} & \text{if } N_{\text{gaps}} > 0 \\ 1 & \text{if } N_{\text{gaps}} = 0 \end{cases} \tag{1.59}$$

and the expected number of islands can be calculated as (cf. equation (1.23)):

$$\begin{aligned} E(N_{\text{islands}}) &= P(N_{\text{gaps}} = 0) + \sum_{x=0}^{n} xP(N_{\text{gaps}} = x) \\ &\approx \left(1 - (1 - f_{\text{G}})^{n-1}\right)^{n} + n(1 - f_{\text{G}})^{n-1} \end{aligned} \tag{1.60}$$

A few subtle changes must also be accounted for in order to determine the distribution of the number of clones in an island. The number of long spacings is N_{gaps}. The number of short spacings is n-N_{gaps}. Now the number of fragments in an island is equal to one plus the number of short spacings between its bounding long spacing(s), or simply n if there are no

[48]David Gordon and Phil Green independently brought this to my attention.

long spacings. The conditional probability density for z_m is therefore (cf. equation (1.26)):[49]

$$f_z\left(x|N_{\text{gaps}}\right) \approx \begin{cases} \left(N_{\text{gaps}}-1\right)\left(1-\dfrac{x-1}{n-N_{\text{gaps}}}\right)^{N_{\text{gaps}}-2} & N_{\text{gaps}} > 1 \\ \delta(n) & N_{\text{gaps}} \le 1 \end{cases} \tag{1.61}$$

The expected value of z_m is therefore (cf. equation (1.27)):

$$E\left(z_m|N_{\text{gaps}}\right) \approx \begin{cases} \dfrac{n}{E\left(N_{\text{gaps}}\right)} = \dfrac{n}{E\left(N_{\text{islands}}\right)} & N_{\text{gaps}} > 0 \\ n = \dfrac{n}{E\left(N_{\text{islands}}\right)} & N_{\text{gaps}} = 0 \end{cases} \tag{1.62}$$

The definition of island length can be expressed as (cf. equation (1.30)):

$$l_m = \begin{cases} \begin{cases} L + \sum_{k}^{k+z_m-2} D_x & z_m > 1 \\ L & z_m = 1 \end{cases} & N_{\text{gaps}} > 0 \\ G & N_{\text{gaps}} = 0 \end{cases} \tag{1.63}$$

We can approximate expected island length as (cf. equation (1.32)):

$$\begin{aligned} E\left(l_m\right) &\approx L + E\left(D_k\right)\left(E\left(z_m\right)-1\right) \\ &\approx L + E\left(D_k\right)\left(\frac{n}{E\left(N_{\text{islands}}\right)}-1\right) \\ &\approx L + G\left(\frac{1}{n}\right)\left(\frac{n}{\left(1-\left(1-f_{\text{G}}\right)^{n-1}\right)^{n}+n\left(1-f_{\text{G}}\right)^{n-1}}-1\right) \\ &\approx G\left(\frac{1}{n}\right)\left(\frac{n}{\left(1-\left(1-f_{\text{G}}\right)^{n-1}\right)^{n}+n\left(1-f_{\text{G}}\right)^{n-1}}-1\right) \\ &\approx G\left(\frac{1}{n}\right)\left(\frac{n}{\left(1-\left(1-f_{\text{G}}\right)^{n-1}\right)^{n}+n\left(1-f_{\text{G}}\right)^{n-1}}\right) \\ &= \frac{G}{\left(1-\left(1-f_{\text{G}}\right)^{n-1}\right)^{n}+n\left(1-f_{\text{G}}\right)^{n-1}} \\ &\approx \frac{G}{E\left(N_{\text{islands}}\right)} \end{aligned} \tag{1.64}$$

[49] $\delta(n)$ is the Dirac-delta function. In this case, it implies that if the number of gaps is one or zero, then the number of clones in an island is certain to be n.

One can choose whichever successive approximation one finds most comfortable. The last approximation can also be made for the linear case. This approximation is very intuitive, and can be arrived at quickly as a “back-of-the-envelope” scribble, perhaps multiplied by $(1-f_g)^n$, or maybe $(1-e^{-R})$.[50] Again, the use of $E(D_k)$ 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 (similar to that done in the linear case). As usual, the accuracy of this approximation can be improved with the aid of a computer by summing over all possible values of z_m, rather than employing the expected value of z_m, and by taking into account the alternative cases in the distributions fed into equation (1.61).

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, for the limit as $T \to 0$, and where B is the greatest integer smaller than $\frac{G}{L}$:

$$E(n \text{ needed for closure}) = 1 - \sum_{k=1}^{B} (-1)^k \frac{\left(1 - k\frac{L}{G}\right)^{k-1}}{\left(k\frac{L}{G}\right)^{k+1}} \qquad (1.65)$$

One can also obtain a reasonable estimate for this value from the probability of project closure, equation (1.18), by assuming that the redundancy required to obtain a 50% chance of closure is roughly equal to the expected redundancy necessary for closure.

Some results for the circle are also 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. The utility of considering such cases is evident, since in actuality both the lengths of the clones and the amount of overlap necessary for detection will vary. The effects do not have to be considered separately, but can be combined into a single new parameter (*L*') equal to *L*-*T*. Such equations will nevertheless get complicated very quickly, and it may be that rather than pursue such a route, computer simulations would be a superior option. Unless, and perhaps even if, *L*' varies greatly, the assumption of a constant *L*' is reasonable.

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 provided for coverage of a circle. Variation of the expected number of gaps on a

[50]At high redundancies the coverage $1-e^{-R}$ is very close to one, and can be so approximated.

finite line due to variation of fragment length is thus expected only due to "edge effects" and is predicted to be small. Computer simulations confirm this prediction (data not shown). Although the expected number of gaps remains constant in the face of varying fragment lengths, the distribution of island sizes will change. The probability of project closure will also be affected. Except in extreme cases, these effects will be slight

1.13.1 Exact Expectations of Island Length for Circular Targets[51]

It is possible to use the general strategy of the mathematical approach of the mathematical model of this thesis to obtain equations of any accuracy desired, as I have alluded at several points. I present one example of such equations in this section. The only approximation I use here is that of continuity, and even that can be discarded by exchanging the integrals for sums.

A singleton contig occurs if and only if two adjacent start-site spacings are both long. The probability that a fragment start site is at the left end of two long spacings is therefore the probability that a singleton contig begins at the start site immediately to the right of the start site in question. There are *n* start sites, so the expected number of singletons is *n* times this probability.

Likewise, the probability of a doubleton equals the probability of a long spacing followed by a short spacing followed by another long spacing. The probability of a tripleton is the probability of a long spacing, follwed by a short spacing follwed by another short spacing followed by a long spacing.

The probabilities of occurence of either a long or short spacing depend on the sum of the lengths of the spacings already observed, so for these exact equations, we must consider each probability separately. Each separate spacing distribution can be calculated as the cumulative distribution of a beta distribution. Each subsequently considered spacing is derived from a beta distribution with one less start site splitting a target that is reduced by the sum of the lengths of the already considered spacings. A target reduced below the effective length of a clone has zero probability of having a long spacing, so we must truncate integration bounds where appropriate.

Now, the shorthand for the beta distribution and the cumulative distribution of the beta

---

[51] This subsection is new in the third printing.

distribution is

$$B_x(1,n) = n(x-1)^{n-1}$$
$$I_a(1,n) = 1-(1-a)^n \tag{1.66}$$

Therefore, we may write exact equations for the expected number of contigs with a specified number of clones as follows:

$$E(N_{\text{singletons}}) = n\int_{f_G}^{1} B_x(1,n-1)\left(1 - I_{\min\left(\frac{f_G}{1-x},1\right)}(1,n-2)\right)dx$$

$$E(N_{\text{doubletons}}) = n\int_{f_G}^{1} B_x(1,n-1)\int_{\min\left(\frac{f_G}{1-x},1\right)}^{1} B_y(1,n-2)\, I_{\min\left(\frac{f_G}{(1-x)(1-y)},1\right)}(1,n-3)\, dydx \tag{1.67}$$

$$E(N_{\text{tripletons}}) = n\int_{f_G}^{1} B_x(1,n-1)\int_{\min\left(\frac{f_G}{1-x},1\right)}^{1} B_y(1,n-2)\int_{0}^{\min\left(\frac{f_G}{(1-x)(1-y)},1\right)} B_z(1,n-3)\, I_{\min\left(\frac{f_G}{(1-x)(1-y)(1-z)},1\right)}(1,n-4)\, dzdydx$$

Equation (1.67) can be extended to any arbitray multiplicity of clones in a contig, and to predict expected contig length. In the somewhat-standardized form of equation (1.67), the first two integrals compute the probability of long spacings, while the remaining integrals compute the probability of the sandwhiched short spacings. In general, the resulting integrals are hard to compute, as they must be evaluated numerically. *Mathematica 4.0* handles embedded triple integrals well, but has difficulty with numerical integration of quadruple integrals. Simulations suggest that equation (1.67) is extremely precise (data not shown).

1.14 SIMULATIONS AND DATA

A large number of Monte Carlo simulations of projects can be generated quickly with a computer. They provide a useful comparison to the mathematical models, and can either support them or point out areas of weakness. The *JASON Report*, an independent review of the U.S. Department of Energy's contribution to the Human Genome project, calls for solicitation and support of detailed Monte Carlo computer simulations of the complete mapping and sequencing process (MITRE Corporation, 1997). Such simulations have been very useful for modeling other large-scale scientific efforts, encompassing areas such as particle physics, astronomy, and oceanography. The *JASON Report* has been summarized by Koonin (1998).

Computer simulations nicely demonstrate the accuracy of the present model, as shown in Figure 1.5. Computer simulations are particularly valuable, as many approximations necessary for mathematical tractability are easily incorporated into simulations. For example,

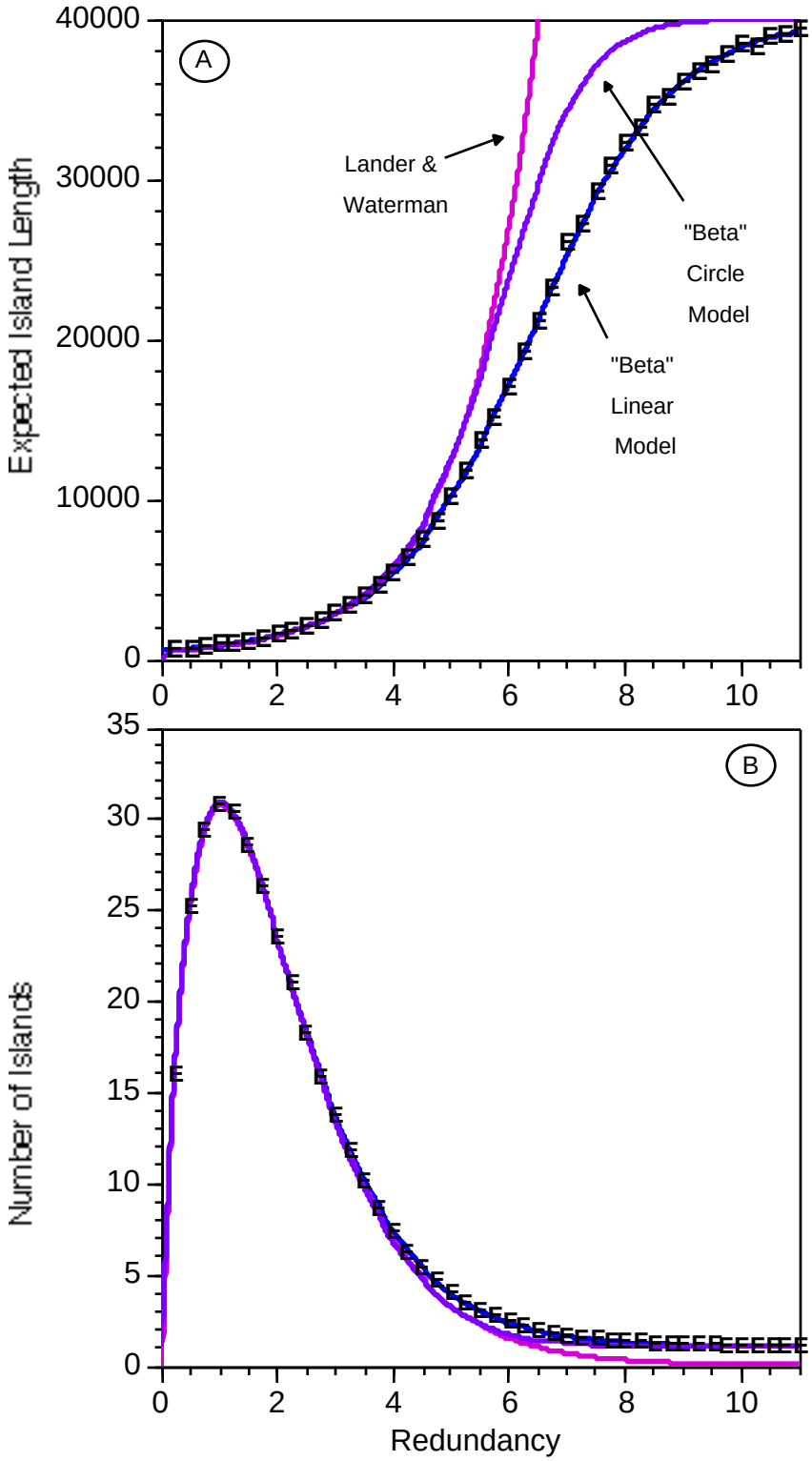

Figure 1.5. A graph of data points from computer simulations. Theoretical curves from Lander-Waterman model and the model of this thesis are provided for reference. Each data point represents the average of 1000 independent Monte Carlo simulations. (**E**) simulated data from a project with a linear target. (*G*=40 kb; *L*=500 bp; *T*=20 bp)

fragment lengths and other variables can be modeled as distributions rather than a constant. An exploration, allowing such parameters to vary, can help validate the approximations of mathematical model, such as the "beta" model described here. In particular, modeling the fragment length as a square pulse of 100 bp width, rather than as a constant, has almost no effect on the statistics of Figure 1.5 (data not shown).

The model presented here also agrees with experimental data, where such data is available. There are few compilations of robust statistical data taken during intermediate project assembly points, particularly because compiling a statistically significant set of such data is burdensome. Nevertheless, experience in our laboratory is that cosmid sequencing projects require redundancies around sevenfold for closure (Rowen and Koop, 1994). This is also what the mathematical model predicts. Results from other laboratories support this view (Davison, 1991; Bodenteich et al., 1994, Martin-Gallardo et al., 1994).

1.15 EXAMPLES

I will briefly present a few examples to illustrate the use of a few of the equations developed in previous sections.

1.15.1 MAPPING THE HUMAN GENOME WITH YACS

Suppose one wishes to map the human genome with restriction digested YACs. The target *G*, in this case the entire human genome, has a length of 3×10^9 bp. The average YAC fragment length *L* has an average size of around 2.5×10^5 bp. The minimum detection overlap *T* for restriction mapping is around 3×10^4 bp. These numbers are approximate, and would be subject to the exact implementation of the strategy.

If a project were to undertake the analysis of 2.4×10^5 YACs, one would obtain a redundancy of twenty. From the Clarke-Carbon equation (1.9), there would be an average of 6.18 uncovered bases. From equation (1.18), the probability of closure would be 99.46%. From equation (1.21), the expected number of gaps would be 5.44×10^{-3}. From equation (1.32), the average island length would be 2.98×10^9 bp. This would be a robust project.

1.15.2 SHOTGUN SEQUENCING A BAC

Consider the task of shotgunning a BAC with a target length *G* of 1.5×10^5 bp. Imagine sequencing it to a redundancy of seven, which is typical for shotgunning cosmids. Assume a typical sequence read length *L* of 650 bp. Assume *T* to be 20 bp.

Sevenfold redundancy will entail sequencing 1.62×10^3 fragments. The Clarke-Carbon

equation predicts an average of 135 uncovered bases. The probability of closure will be 17.40%. The expected number of gaps will be 1.75. The average island length will be 5.48×10^4. From equation (1.52), the probability of obtaining an island greater than half the target length will be 78.21%.

If we were to sequence a 3.5×10^4 bp cosmid target to a redundancy of sevenfold, our probability of closure would be 70.40%. Obtaining analogous project goals for a BAC requires higher redundancies than for a cosmid.

1.15.3 The Choice of Phage Clones, BACs, or Cosmids as Sequencing Targets

Planners of high-throughput genome sequencing projects are faced with the decision of what type of clone to use as targets for shotgun sequencing. One can address the sequencing costs associated with different choices of clones with the aid of the equations presented here. To consider a simple example, hypothesize that a mapping protocol has produced an approximation to a minimum tiling path of sequencing targets, with adjacent targets overlapping by 5 kb. Assume that each clone is sequenced to the expected redundancy needed for closure, and then any remaining gaps are closed by directed sequencing. Continue to assume L=650 bp and T=20 bp. Consider an overall project goal of sequencing a 100 Mb genome (but the results can easily be scaled to an arbitrary genome size).

If the targets are λ clones with G=20 kb, then the expected number of fragments for λ closure is 177 (equation (1.65)), which is a redundancy of 5.75. The expected number of gaps per λ clone (equation (1.20)) will be 0.50. The average gap length will be 126 bp.[52] Considering the 5 kb target overlaps, one needs to sequence 6667 target λ clones to span 100 Mb. The total number of fragments sequenced will be 1.18×10^6. Each gap has a one third chance of being covered by sequence from an overlapping λ clone, so a total of (0.50)(6667)(0.66)=2231 gaps need to be closed.

If the targets are cosmids with G=35 kb, then the expected number of fragments for cosmid closure is 345, which is a redundancy of 6.42. The expected number of gaps per cosmid will be 0.58. The average gap length will be 99 bp. Considering the 5 kb target overlaps, one needs to sequence 3334 target cosmids to span 100 Mb. The total number of

[52]See the discussion in Section 1.10 for approaches to calculating gap length. A useful quick approximation for the expected gap length is:

$$E(\text{gap length}) = \frac{Ge^{-R}}{E(N_{\text{gaps}})}$$

fragments sequenced will be 1.15×10^6. Each gap has a 17% chance of being covered by sequence from an overlapping cosmid, so a total of (0.58)(3334)(0.83)=1598 gaps need to be closed.

If the targets are BACs with *G*=150 kb, then the expected number of fragments for BAC closure is 1869, which is a redundancy of 8.10. The expected number of gaps per BAC will be 0.69. The average gap length will be 66 bp. Considering the 5 kb target overlaps, one needs to sequence 690 target BACs to span 100 Mb. The total number of fragments sequenced will be 1.29×10^6. Each gap has a 3.5% chance of being covered by sequence from an overlapping BAC, so a total of (0.69)(690)(0.965)=460 gaps need to be closed.

Assuming the cost of generating all three maps is equal, then clearly cosmids are a better choice than λ clones, as there are fewer overall sequence reads with fewer gaps to be closed. If BACs are employed instead of cosmids, then the required number of sequence reads increases by 1.38×10^5, but the number of gaps to be closed by directed sequencing drops by 1138. So if the price of closing a gap by directed sequencing is less than 121 times the price of a single random sequence read, then a cosmid strategy would be cheaper. In reality, the cost of constructing a BAC map will be less than constructing a cosmid map, which will bias the choice of sequencing targets towards BACs.

Note that the above analysis assumes that the projects stop at the expected redundancy for closure and then move to directed gap closure. The choice of when to stop is another important parameter in strategy choice.[53] Consider the next example.

1.15.4 When To Stop Random Subcloning and Start Directed Gap Closing

From the last example, one can recognize that there may be an optimal juncture at which to stop random subcloning and start directed sequencing. This juncture will depend on the relative costs of random and directed sequencing. Let us consider several possible cost ratios for the case of BAC sequencing: 10, 20, 75, 300, and 1000. A directed to random cost ratio of 10 might represent a scenario with negligible primer synthesis costs and completely automated clone selection and primer design for gap closures. A ratio of 1000 might represent a scenario with high personnel and primer synthesis costs.[54]

The expected costs to close a BAC are graphed in Figure 1.6. The parameters from the previous example are employed. As the cost of directed sequencing rises relative to an

[53]Knowing when to stop is important in many things: cars, genomics, and life activities in general.

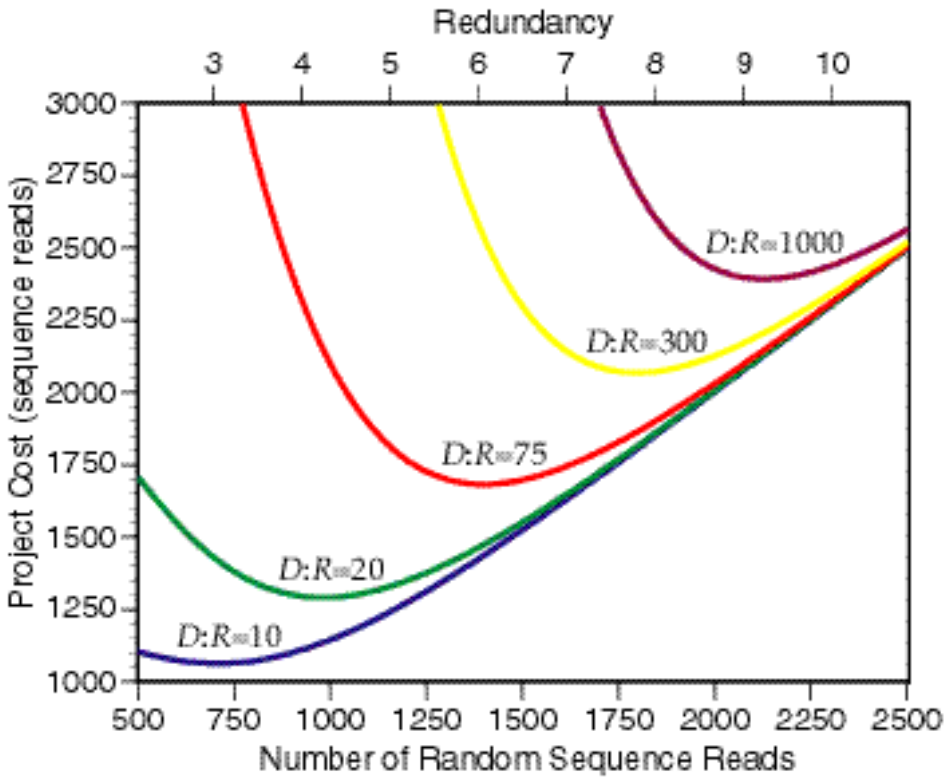

Figure 1.6. The expected cost of closure for a BAC shotgun sequencing project with directed finishing. The ratio *D*:*R* is the relative cost of closing one gap versus executing a single random sequence read. Cost is standardized to the cost of a random sequence read [Cost=n+(D:R)E(N_{gaps})]. The abscissa represents the number of random reads obtained before directed sequencing begins. Note that all curves are asymptotic to a purely random strategy that has produced closure. (G=150 kb; L=650 bp; T=20 bp)

arbitrarily fixed random sequence read cost, clearly the overall project cost rises as well. The optimal stopping point for random sequencing occurs at a progressively higher redundancy as the cost of directed gap closure increases. Note that this example assumes that when a fixed number of sequence reads is obtained, random sequencing stops regardless of the number of gaps, and then gaps are closed by a directed strategy.[55]

A slight decrease in expected cost might be obtained by altering the strategy in a manner that involves iterative feedback between assembly and shotgun sequencing. This

[54]A December, 1997, price quote from a commercial sequencing company gives approximately $25 as the price of generating a single sequence read and approximately $600 as the cost of closing a gap (Genome Systems, price quote). Costs in a genome center setting are considerably lower: approximately $8-$10 to generate a single read (Stephen Lasky and Maynard Olson, personal communications).

would entail shotgun sequencing until a fixed number of gaps (e.g., two) was obtained, and then closing the gaps in a directed manner. The administrative costs in executing a project in this fashion are slightly more intangible than described above. Nevertheless, this is an extremely common actual implementation of the shotgun strategy, with the number of sequence reads obtained in batches of several hundred between assemblies. Shotgunning stops when one of these incremental assemblies reduces the number of gaps to a point at which directed sequencing can begin.

1.15.5 The Rise in Cost as the Target Length is Increased

A higher redundancy is needed to reach a state of expected closure for a longer target. Thus all other things being equal, shotgunning each clone in a minimum tiling path of mapped subclones of a target is cheaper than shotgunning the whole target. This is analogous to arguments that a divide-and-conquer strategy for STS-mapping the human genome chromosome by chromosome is cheaper than mapping all markers simultaneously.[56] However, creating such a tiling path has a cost. Additionally, the resulting "minimum" tiling path invariably has considerable overlap, increasing the total sequencing redundancy by perhaps 30%. Thus the decision to "divide and conquer" or to "brute force" the target must be made by comparing the decrease in shotgunning cost per base pair for shorter targets with the increase in mapping cost. This is a common dilemma in structural genomics. A typical example is a decision whether to subclone a BAC into cosmids or to directly shotgun the BAC. Another example would be the choice of subcloning a bacterial genome or shotgunning it directly.

We can determine the exact expected redundancy for closure of a circular target by

---

[55]An additional complication to consider is that some gaps are harder to close by a directed strategy than others. Short gaps that are spanned by known sequencing templates are straightforward to close with a simple walking iteration. Gaps without templates must generally be closed by first generating a PCR sequencing template, which involves extra cost and can increase the sequencing error rate. At high redundancies, particularly with a pairwise strategy (discussed in Chapter 2), all gaps are highly likely to be spanned by known sequencing templates, so the approximation of constant cost per directed gap closure is reasonable. However, at low redundancies, the average cost of a directed gap closure may rise due to increasing frequency of gaps not spanned by known templates. Figure 1.6 can be modified for this effect by assigning different gap closure costs to each type of gap and assigning the relative frequencies for gap type according to a model for the strategy in use. In this case, both the clone length and the sequence read length must be incorporated, as well as any mapping data, such as that obtained form a pairwise project. Incorporating these details is not difficult, but does depend on the exact strategy parameterization.

employing the equation of Flatto and Konheim (1962). Alternatively, we can approximate the expected redundancy for closure by numerically integrating the weighted derivative of equation (1.57) (or, for a linear target, equation (1.18)). Such numerical integration is best carried out by a mathematical analysis package such as *Mathematica 3.0* (Wolfram Research), which is what I used for this purpose.[57] With respect to the expected redundancy for closure of a circular target, and for parameterizations of interest to genomics, the relative error of the numerical integration with respect to the Flatto-Konheim summation was between one and three percent. The use of numerical integration allows one to calculate probabilities not provided by Flatto and Konheim. For example, one can compute the expected redundancy necessary for a project to reach a state of three or fewer gaps as follows:

$$E(n) = \int_{W}^{\infty} n \frac{\partial\left(P\left(N_{\text{gaps}} \le 3\right)\right)}{\partial n}\, dn \tag{1.68}$$

Here, W is the minimum number of clones with which it is possible to span the target (i.e., the number of clones necessary for onefold redundancy). If the lower bound W is not used, then semantic difficulties arise as to exactly what is meant by a project with three gaps. Technically, most projects with three clones also have three gaps. We do not wish to include such cases in our integration, so we reasonably but somewhat arbitrarily specify that a completed project must have at least onefold redundancy. The probability of a project having three or fewer gaps can be approximated by summing appropriate parameterizations of equation (1.56) for the circle or equation (1.17) for the linear case:

$$P\left(N_{\text{gaps}} \le 3\right) = P\left(N_{\text{gaps}} = 0\right) + P\left(N_{\text{gaps}} = 1\right) + P\left(N_{\text{gaps}} = 2\right) + P\left(N_{\text{gaps}} = 3\right) \tag{1.69}$$

The resulting equation becomes quite bulky but is easily handled by *Mathematica*. By employing this technique, I have calculated the expected redundancy to reach a state of three or fewer gaps for a variety of target sizes, as well as the expected redundancy to reach a state of six or fewer gaps (Figure 1.7). The Flatto-Konheim predicted redundancies for closure are

---

[56]This is discussed by Lange and Boehnke (1982). Note that these authors mistakenly exaggerate the cost of mapping 24 discrete linear chromosomes relative to the cost of mapping one hypothetical chromosome with length equal to the sum of the lengths of the 24 chromosomes. Presumably, this results from a specification in their computer simulations that the extreme telomeric ends of the chromosomes must be within a specified distance of a marker. A more thorough treatment of this issue is provided by Bishop et al. (1983).

[57]It should also be noted that the Flatto-Konheim summation employs alternating differences of ratios of very small numbers. This requires a numerical precision of approximately 400 digits for some parameterizations of relevance to genomics. The sum does not converge quickly, and cannot be approximated by a truncation.

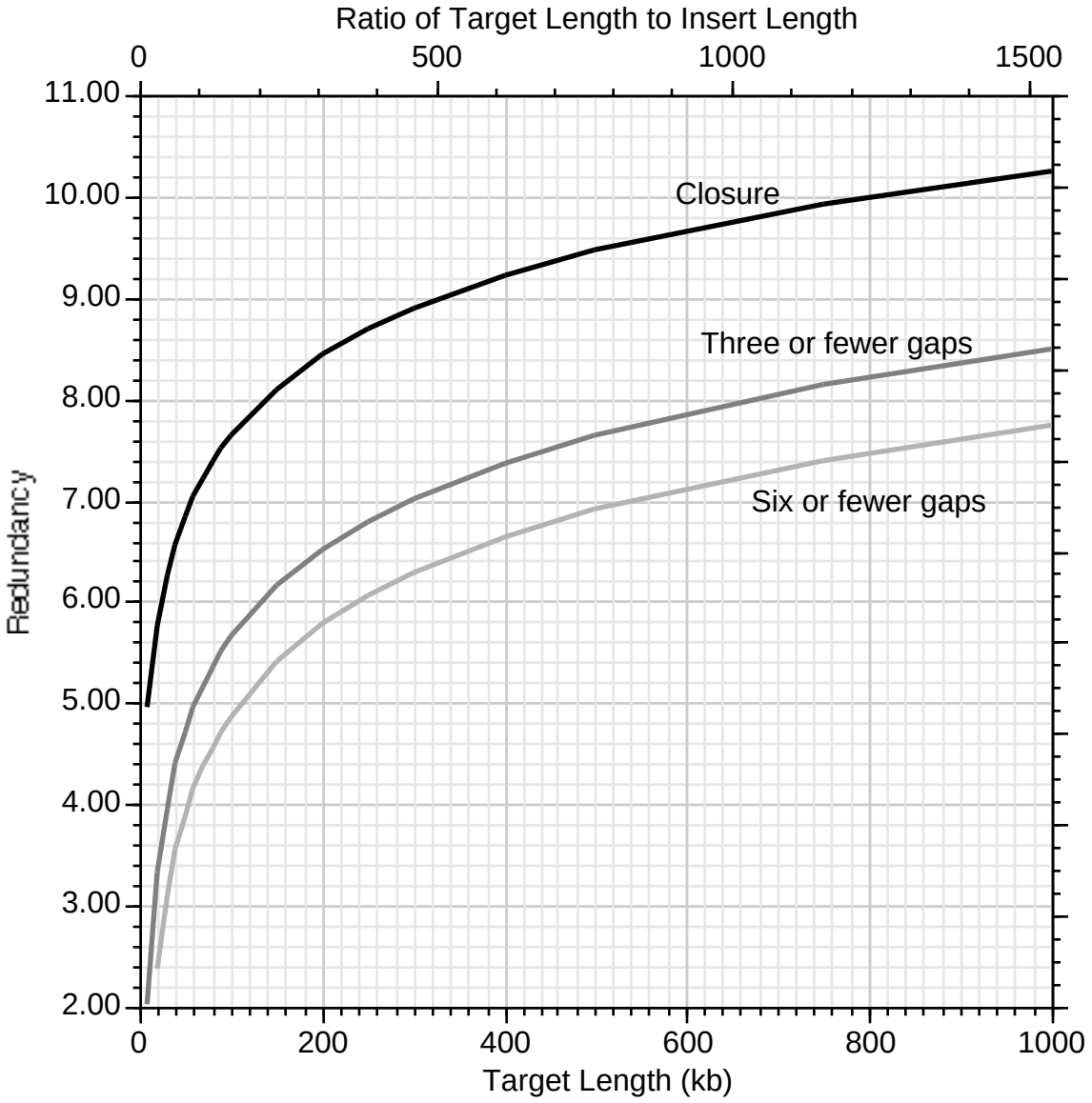

Figure 1.7. The expected cost of reaching a project state of three or fewer gaps graphed versus target size. The cost for six or fewer gaps is also graphed. These costs were calculated by numerical integration (see text). The expected costs of closure calculated by the equation of Flatto and Konheim are shown as a reference. A circular target is assumed. (L=650 bp; T=20 bp)

shown for reference. Note that this calculation is invariant to scale; G, L, and T can be altered proportionately with no change in redundancy costs to reach an expected number of gaps.

One can see that per-base-pair costs rise roughly logarithmically with respect to target length. This occurs when either closure or a fixed number of gaps is sought. However, if one allows projects to stop at a stage with a number of gaps proportional to the target length, then

costs will rise at a slower rate, and in fact are almost invariant, at least with respect to this choice of parameters. For example, sevenfold redundancy will take a 600 kb project to a state of six gaps, while the same redundancy will take a 300 kb project to a state of three gaps. This observation tends to support the conclusion that longer is better.

1.16 RANDOM CLOSING REMARKS

The general approach to modeling finite genomes presented here may be useful when applied to other mapping and sequencing strategies, such as those based on random transposon insertion. Such strategies represent a genomics implementation of the well established theory of "coverage processes." Hall (1988) provides a nice entry into some of the relevant mathematics literature. Solomon (1978) provides a good overview of the problem of random arcs on the circumference of a circle.

The equations derived here have many applications.[58] 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 closure at a given redundancy (equation (1.18)). These results are consistent with the expected redundancy needed for closure (equation (1.65)). I have combined results from these two equations in Figure 1.8.

Figure 1.8 highlights some of the most useful results to arise from the "beta" model for random subcloning. The figure shows that longer targets have a higher cost in redundancy to close. This means that, all other factors being equal, it is cheaper to shotgun two halves of a target separately than to do both at once. A practical application of this observation might be to partition a multiple chromosome genome into its individual chromosomes before commencing a shotgun project (but see also Section 1.15.5).

In addition, Figure 1.8 shows that, in general, fewer longer fragments are more desirable than proportionally more shorter fragments. There are situations in sequencing projects where longer reads can be obtained, but at higher cost. The trade-off of increased cost per read versus decreased redundancy needed for closure can be analyzed. The payoff is more than linear as fragment length increases. This result, in particular, is completely unanticipated by the Clarke-Carbon formula, which predicts the same amount of coverage at a given redundancy, regardless of fragment length.[59]

---

[58]A reader interested in quick access to some simple Java implementations of some of the equations presented in this paper might wish to explore Andrei Grigoriev's web tools at www.embl-heidelberg.de/~toldo/JaMBW.

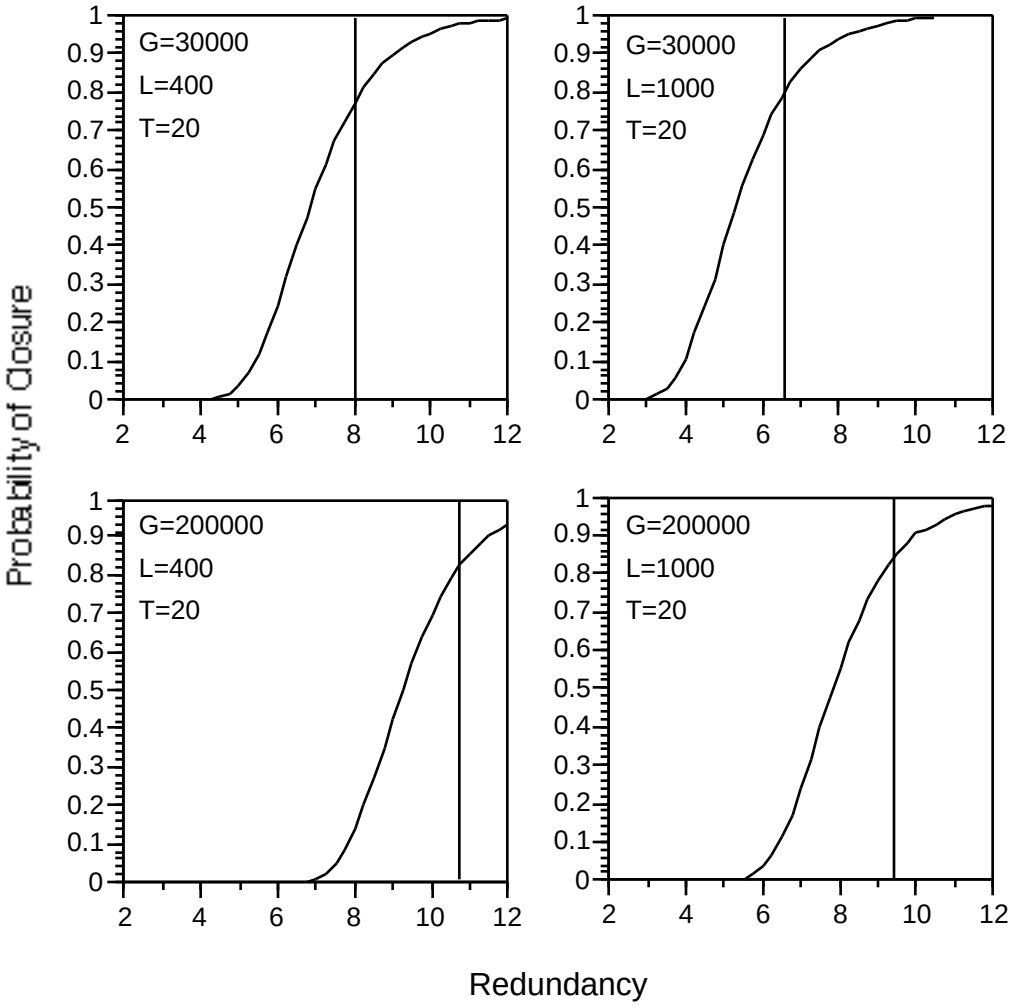

Figure 1.8. The probability of project completion with respect to redundancy, calculated using the exact equation of Stevens (1939). This equation is approximated in the text by equation (1.18). Four parameterizations are shown. The vertical lines intersect the expected redundancy necessary for closure, calculated using the exact equation of Flatto and Konheim (1962), with their parameter a set to (*L*-*T*)/*G*.

At high shotgun redundancies, the cost of directed sequencing is roughly constant per gap, no matter how long the gap is. This is because at high redundancies gaps are almost universally shorter than a sequence read length. Efforts to close such a gap will be equally expensive to administrate and execute whether the length of the gap is one base pair or *L*-*T*

[59]To rephrase: there are fewer gaps in projects with longer fragments, but these gaps are longer. Thus the total uncovered area remains constant as long as redundancy stays the same.

base pairs.[60]

Should one choose to close gaps by continuing random subcloning, there will be an exponentially increasing cost in redundancy to close gaps as a project proceeds. 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. This cost can be compared with the cost of directed sequencing to determine if and when directed methodology is appropriate for a project. A graph of gap closure cost is shown in Figure 1.9.[61] The more gaps there are in a project, the cheaper it is to close them by shotgun sequencing. The cost of closing gaps rises exponentially with each successive gap closed.

One potential objection to any mathematical model for DNA cloning is that there may be regions of the target that are nearly impossible to clone. For example, some regions of the HIV genome are genetically unstable and thus absent from subclone libraries. This results in a large local deviation from the uniformity of fragment start site distribution. Such large variations in uniformity tend to be target idiosyncratic and very difficult to model. This by no means dooms the utility of mathematical models. In fact, such cases are precisely where mathematical models shine. Deviations in the actual target parameters from predicted values indicate subcloning problems. Once detected, such problems can be addressed and corrected. Also, once knowledge is obtained of an unclonable or unanalyzable region, the model can be appropriately modified to reflect the new constraints.

---

[60]A gap can be closed by primer walking along an existing template known to cross the gap. See Chapter 2 for discussion of one method to determine template positioning relative to a gap. In the absence of an existing sequencing template, a gap can be closed with PCR methodology.

[61]There is a subtlety here worth elucidating. Before a project starts, the redundancy cost of a project with x expected gaps is easily calculated from equation (1.20). However, once a project is underway, and a preliminary assembly has been made, the information gained from that assembly affects the prediction of how many gaps will be present in a future state of the project. In most cases this effect will be minor, particularly if the actual number of gaps is close to the expected number of gaps. However, the effect cannot be predicted ahead of time. Therefore the cost of closing a gap in Figure 1.9 is calculated by determining the redundancy necessary for an uncommenced project to reach a state with an expected number of gaps that is one less than the expected number of gaps of a project executed at the redundancy graphed on the abscissa.

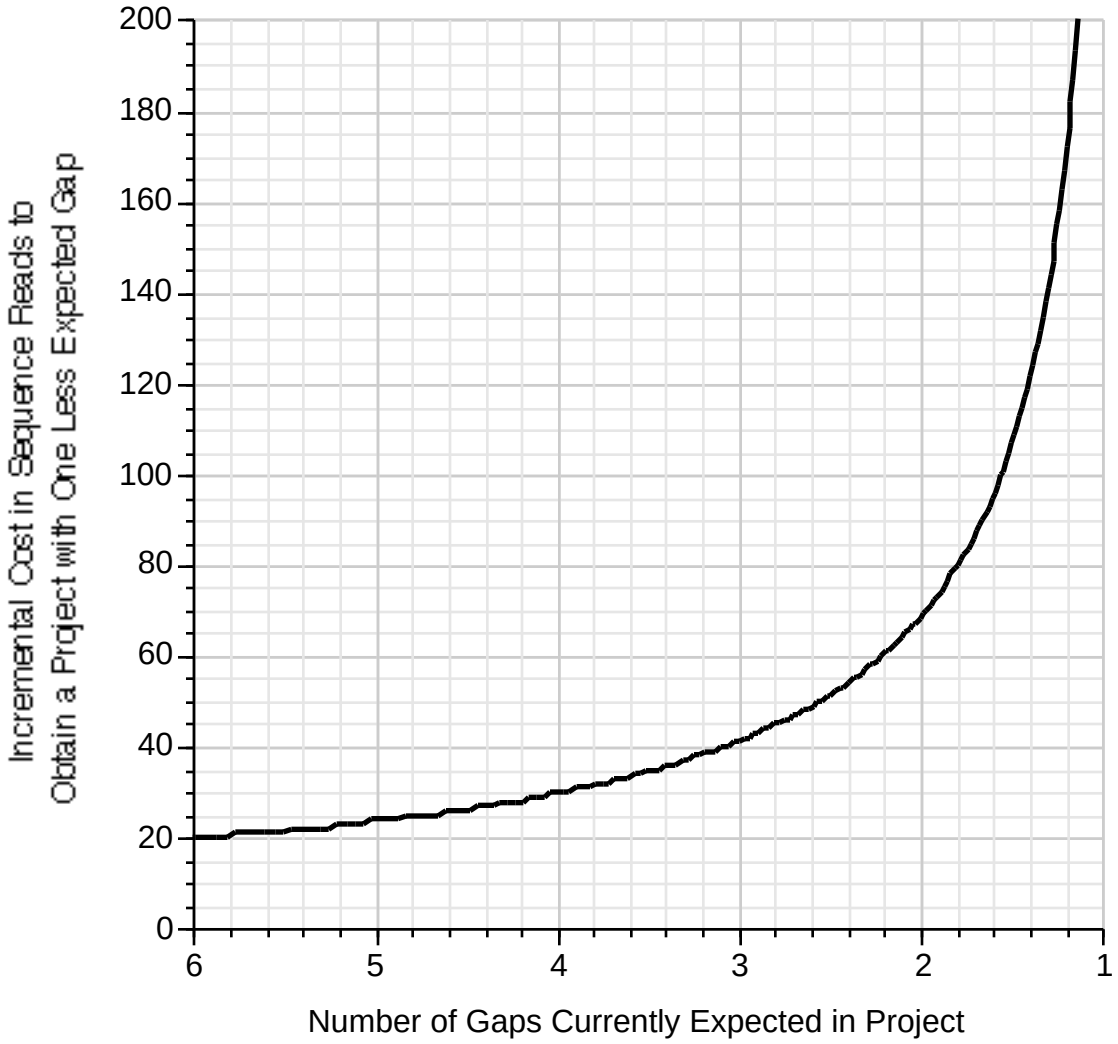

Figure 1.9. The incremental cost of closing one gap. This is calculated from the number of expected gaps in a project with no knowledge of a prior state of that project (see equation (1.20)). Note that it is impossible to plan a project with zero expected gaps, as gaps always remain a small but finite possibility. (G=40 kb; L=500 bp; T=20 bp)

There must be a constant interplay between theory and practice. Each serves to refine the other. Neither exists alone, nor is valuable without the other.