Statistical Methods in Physical Mapping

David O. Nelson
(Ph.D. Thesis)

May 1995
DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from the Office of Scientific and Technical Information
P.O. Box 62, Oak Ridge, TN 37831
Prices available from (615) 576-8401, FTS 626-8401

Available to the public from the National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd., Springfield, VA 22161

Work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
Statistical Methods in Physical Mapping

David O. Nelson
(Ph.D. Thesis)

Manuscript date: May 1995

LAWRENCE LIVERMORE NATIONAL LABORATORY
University of California • Livermore, California • 94551

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER
Statistical Methods in Physical Mapping

by

David Oscar Nelson

B.A. (University of California at Santa Barbara) 1974
M.A. (University of California at Berkeley) 1977

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

GRADUATE DIVISION

of the

UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor Terence P. Speed, Chair
Professor Steven N. Evans
Professor Gerald R. Rubin

1995
Abstract

Statistical Methods in Physical Mapping

by

David Oscar Nelson

Doctor of Philosophy in Statistics
University of California at Berkeley

Professor Terence P. Speed, Chair

One of the great success stories of modern molecular genetics has been the ability of biologists to isolate and characterize the genes responsible for serious inherited diseases like fragile X syndrome, cystic fibrosis and myotonic muscular dystrophy. This dissertation concentrates on one important class of activities in this process of large-scale gene isolation: constructing high-resolution physical maps. It demonstrates how probabilistic modeling and statistical analysis can aid molecular geneticists in the tasks of planning, execution, and evaluation of physical maps of chromosomes and large chromosomal regions.

The dissertation is divided into six chapters. Chapter 1 provides an introduction to the field of physical mapping, describing the role of physical mapping in gene isolation and in past efforts at mapping chromosomal regions. The next two chapters review and extend known results on predicting progress in large mapping projects. Such predictions help project planners decide between various approaches and tactics for mapping large regions of the human genome. Chapter 2 shows how probability models have been used in the past to predict progress in mapping projects. In preparation for the third chapter, these results have been recast into the unifying vocabulary of stationary point and coverage processes. Chapter 3 presents new results, based on stationary point process theory, for progress measures for mapping projects based on directed mapping strategies.

The next two chapters concentrate on analyzing data from mapping projects. Chapter 4 describes in detail the construction of an initial high-resolution physical
map for human chromosome 19. This chapter introduces the probability and statistical models involved in map construction in the context of a large, ongoing physical mapping project. Chapter 5 concentrates on one such model, the trinomial model. This chapter contains new results on the large-sample behavior of this model, including distributional results, asymptotic moments, and detection error rates. In addition, it contains an optimality result concerning experimental procedures based on the trinomial model.

The last chapter explores unsolved problems and describes future work.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>List of Figures</strong></td>
<td>v</td>
</tr>
<tr>
<td><strong>List of Tables</strong></td>
<td>vii</td>
</tr>
<tr>
<td><strong>Preface</strong></td>
<td>ix</td>
</tr>
<tr>
<td><strong>1 An Overview of Physical Mapping</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 DNA, Genes and Maps</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Constructing Physical Maps</td>
<td>5</td>
</tr>
<tr>
<td>1.2.1 Producing a Library of Cloned Fragments</td>
<td>6</td>
</tr>
<tr>
<td>1.2.2 Producing “Fingerprint” Data for each Clone</td>
<td>7</td>
</tr>
<tr>
<td>1.2.3 Assembling Contigs and Obtaining Closure</td>
<td>9</td>
</tr>
<tr>
<td>1.2.4 Other Issues</td>
<td>14</td>
</tr>
<tr>
<td>1.3 Finding the Gene for Myotonic Dystrophy</td>
<td>16</td>
</tr>
<tr>
<td><strong>2 Predicting Progress in Random Mapping Projects</strong></td>
<td>19</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>19</td>
</tr>
<tr>
<td>2.2 Pure Lander-Waterman</td>
<td>20</td>
</tr>
<tr>
<td>2.2.1 How Many Islands</td>
<td>20</td>
</tr>
<tr>
<td>2.2.2 The Number of Clones in an Island</td>
<td>22</td>
</tr>
<tr>
<td>2.2.3 The Length of an Island</td>
<td>22</td>
</tr>
<tr>
<td>2.2.4 Distances Between Islands</td>
<td>23</td>
</tr>
<tr>
<td>2.2.5 Total Coverage</td>
<td>24</td>
</tr>
<tr>
<td>2.2.6 Overlap Detection</td>
<td>24</td>
</tr>
<tr>
<td>2.3 STS Content Mapping</td>
<td>25</td>
</tr>
<tr>
<td>2.3.1 How Many Anchored Clones</td>
<td>26</td>
</tr>
<tr>
<td>2.3.2 How Many Anchored Islands</td>
<td>26</td>
</tr>
<tr>
<td>2.3.3 The Number of Clones in an Anchored Island</td>
<td>27</td>
</tr>
<tr>
<td>2.3.4 The Length of an Anchored Island</td>
<td>28</td>
</tr>
<tr>
<td>2.3.5 Total Coverage</td>
<td>29</td>
</tr>
</tbody>
</table>
# Predicting Progress in Directed Mapping Projects

3 Predicting Progress in Directed Mapping Projects 31

3.1 Introduction .......................................................... 31
3.2 A “Finite” Analysis for Clones of Constant Length .......... 33
3.3 The Stationary Process Approach .................................. 40
3.4 Discussion ............................................................ 44
3.5 Details of Proofs ....................................................... 49
  3.5.1 The Finite Case .................................................. 49
  3.5.2 Closed Form Approximations for the Finite Case .......... 53
  3.5.3 Stationary Process Approach .................................. 54

# Constructing High-Resolution Physical Maps

4 Constructing High-Resolution Physical Maps 57

4.1 Introduction .......................................................... 57
4.2 Creating Fingerprints ............................................... 59
4.3 Detecting Overlap ..................................................... 63
  4.3.1 The Simple Trinomial Model ................................ 65
  4.3.2 Towards More Realistic Assumptions ....................... 67
4.4 Assembling Contigs .................................................. 73
4.5 Discussion ............................................................ 77

# An Analysis of the Trinomial Model

5 An Analysis of the Trinomial Model 79

5.1 The Basic Setup ...................................................... 79
  5.1.1 A Bayesian Decision Problem ................................ 79
  5.1.2 The Trinomial Model ........................................... 82
5.2 Asymptotics of the Likelihood Ratio Statistic ................. 85
  5.2.1 Asymptotics Under No Overlap ............................... 85
  5.2.2 Asymptotics Under Positive Overlap ....................... 91
5.3 Large-Sample Power of the Likelihood Ratio Test .......... 92
5.4 The Performance of the Trinomial Model ......................... 93
  5.4.1 Performance Against Specific Alternatives ............... 94
  5.4.2 Performance Against General Alternatives ............... 96
5.5 Approximating the Posterior Odds ............................... 97
  5.5.1 The Maximum Likelihood Estimate of Overlap .......... 98
5.6 The Effect of Errors ................................................. 101

# Future Work

6 Future Work .......................................................... 105

Bibliography ............................................................. 109
List of Figures

1.1 Typical diagram of a part of a double strand of DNA. .................. 2
1.2 Sample overlap configuration with STSS. ............................... 9
2.1 Typical situation in STS-content mapping. .................. 25
3.1 Spacings resulting from directed mapping. ............................ 34
3.2 Computing vacancy. ............................................. 38
3.3 Computing current and excess life. ............................... 42
3.4 \( r_2 \) for Palazzolo’s case. ........................................ 45
3.5 Error resulting from using \( r_0 \) instead of \( r_0 \). .................. 46
3.6 Simulation vs. theory—contigs. ................................... 47
3.7 Simulation vs. theory—vacancy. ..................................... 48
3.8 Simulation vs. theory—average contig length. ......................... 49
3.9 Probability that a seed clone subsumes no other clone. .......... 50
4.1 Main steps in constructing the map of chromosome 19. .............. 58
4.2 Creating fingerprints using multiple restriction digests. .......... 60
4.3 Resulting signal for one clone from an electrophoresis run. ..... 62
4.4 Typical configuration where two clones overlap. .................. 64
4.5 Proportion of fingerprints with positive probes as a function of fragment size \( \ell \). .................................................... 68
4.6 Smoothed estimates of error rates \( \pi_{\ell_0} \) and \( 1 - \pi_{\ell_1} \). ...... 70
4.7 Example of a restriction map. ...................................... 72
4.8 Power of Posterior odds compared to number of mismatches to detect overlap. ............................................. 76
4.9 \( \beta \) error vs. \( \alpha \) error for various \( \theta \) and \( p \) (200 probes). .... 77
4.10 Observed versus expected for various statistics for the D19S11 region. .................. 78

5.1 Two equal-length, overlapping clones. \( \Theta \) is defined to be the proportion of shared DNA in the two clones. .................. 79
5.2 The probability triples \( (p_0, p_1, p_2) \) all lie on a simplex \( S \). \( \gamma \) defines a parameterized surface on \( S \), the image of which we denote by \( M \). .................. 83
5.3 The coordinate curves induced by the parameterization in Eqn. (5.2). The straight lines with arrows correspond to \( \gamma(\cdot, r) \) curves for various fixed \( r \), the arrow indicating the direction of motion as \( \theta \) runs from 0 to 1. The curved lines correspond to \( \gamma(\theta, \cdot) \) curves for various \( \theta \).

5.4 The prior \( w(\lambda, r) \), considered as a function of \( r \) with \( \lambda \) fixed at 1.

5.5 Four examples of contour surfaces for the MLE as a function of \( y \), corresponding to \( r = 0.2, 0.4, 0.6, 0.8 \). The values of the surface that exceed one show the value of the unrestricted MLE. To obtain the restricted MLE, one substitutes the value 1.0 for the restricted estimate.

5.6 The effect of a measurement error model with \( \pi_0 = 0.05 \) and \( \pi_1 = 0.1 \), viewed as a vector field on the simplex \( S \). The overall effect is to shrink the manifold \( M \) towards its fixed point. The curve \( \theta = 0 \) is an integral curve for this field.

5.7 Four examples of the effect of measurement errors on the resulting expectation surface \( M \).
List of Tables

5.1  The loss function for a simple two-decision problem. ...............  80
Preface

One of the great success stories of modern molecular genetics has been the ability of biologists to isolate and characterize the genes responsible for serious inherited diseases like fragile X syndrome, cystic fibrosis and myotonic muscular dystrophy. This dissertation concentrates on one important class of activities in this process of large-scale gene isolation: constructing high-resolution physical maps. It demonstrates how probabilistic modeling and statistical analysis can aid molecular geneticists in the tasks of planning, execution, and evaluation of physical maps of chromosomes and large chromosomal regions.

The dissertation is divided into six chapters. Chapter 1 provides an introduction to the field of physical mapping, describing the role of physical mapping in gene isolation and in past efforts at mapping chromosomal regions. The next two chapters review and extend known results on predicting progress in large mapping projects. Such predictions help project planners decide between various approaches and tactics for mapping large regions of the human genome. Chapter 2 shows how probability models have been used in the past to predict progress in mapping projects. In preparation for the third chapter, these results have been recast into the unifying vocabulary of stationary point and coverage processes. Chapter 3 presents new results, based on stationary point process theory, for progress measures for mapping projects based on directed mapping strategies.

The next two chapters concentrate on analyzing data from mapping projects. Chapter 4 describes in detail the construction of an initial high-resolution physical map for human chromosome 19. This chapter introduces the probability and statistical models involved in map construction in the context of a large, ongoing physical
mapping project. Chapter 5 concentrates on one such model, the trinomial model. This chapter contains new results on the large-sample behavior of this model, including distributional results, asymptotic moments, and detection error rates. In addition, it contains an optimality result concerning experimental procedures based on the trinomial model.

The last chapter explores unsolved problems and describes future work.

Acknowledgements

While dissertations are supposed to be the original independent work of a single author, it is clear that, like all so-called one-man shows, nothing would ever happen without the efforts of a great many other people. I would like to acknowledge a few of them here, and ask forgiveness of those many more whose names I omit solely for the lack of a sufficient printing budget. They fall into three categories: academic, professional, and personal.

On the academic side, I would especially like to thank my committee, consisting of Terry Speed, Steve Evans, and Gerry Rubin, for their guidance and patience throughout this effort. I hope that at least some of their insight and experience, along with their commitment to high standards of research and rigorous prose, have been transferred to me. Terry’s initial interest in my work and confidence in my abilities made my decision to return to the rigors of graduate study an easy one. In addition, I would like to thank Bin Yu and John Rice for helpful discussions along the way.

On the professional side, some of the most satisfying moments of my professional life have come courtesy of the Genome Group at LLNL. They have also cheerfully put up with my all-too-frequent absences and distractions. Among those who stand out are Elbert Branscomb and Tony Carrano. Elbert’s observation that, from the point of view of the Genome Project, my going back to graduate school was simply “good land-use management” will always stand out in my mind as a ringing vote of confidence. And of course, Tony Carrano’s blessing made it all possible.

On the personal side, my parents have been waiting patiently for this event for over twenty years. I appreciate their love and support, and I am happy that they never
gave up hope. In addition, I would like to thank the many students, faculty, and staff of the UCB Statistics Department who made my stay there enjoyable and fulfilling. Especially important to the success of this dissertation were Bonnie Hutchings and Sara Wong, whose understanding of the conundrum that is Sproul Hall made my life much easier than it otherwise would have been.

Finally, and most importantly of all, there’s Peggy Reynolds. The most poignant dedication I’ve ever read was that of a (perhaps apocryphal) mathematics professor who dedicated a book to “... my loving spouse, whose lack of interest in my work has been a continuing source of sorrow to me.” I suffer from no such sorrow. On the contrary, it was Peggy’s enthusiasm and gentle prodding over a span of more than ten years that finally convinced me to return to the life of a graduate student. And while a student, her understanding and support were instrumental in enabling me to get through the process. Quite truthfully, without her continuing interest in my work, this dissertation would never have happened.

This work was performed under the auspices of the U. S. Department of Energy by Lawrence Livermore National Laboratory under contract number W-7405-ENG-48, with additional support from NSF grant DMS-91-13527.
Chapter 1

An Overview of Physical Mapping

1.1 DNA, Genes and Maps

The genetic complement of a human being (a "human genome") consists of twenty-three pairs of chromosomes [69]. Each chromosome is composed of a single large molecule of DNA, as well as other supporting structures. Each molecule of DNA is itself composed of two complementary polymers called strands, bound together by hydrogen bonds between the monomers in the two strands. Each strand is composed of a long sequence of four different monomers containing bases denoted by A, T, G, and C. The instructions and data necessary to transform a fertilized egg in utero into a developed human being are encoded in the sequence of bases making up the two complementary strands of the DNA in the zygote's twenty-three pairs of chromosomes.

The strands are called complementary because the configuration of hydrogen bonds between the bases ensures that, under normal conditions, base A in one strand is only paired with base T in the other strand. Likewise, base G in one strand is only paired with base C in the other strand. Thus, a portion of a DNA molecule is often diagrammed as in Figure 1.1. The short vertical lines represent the hydrogen bonds between the strands, and the horizontal arrows represent bonds between the monomers within a strand. The bonds between the monomers are drawn as arrows because each strand has a direction, and the messages encoded in the bases are to be read from the 5' end to the 3' end of a strand. (This notation for the ends of a strand
Figure 1.1: Typical diagram of a part of a double strand of DNA.

of DNA is standard, and is based upon molecular labeling conventions.) Most messages are coded in the form of genes, which are sections of DNA containing instructions for the synthesis of proteins.

Each of the twenty-three human chromosomes differ in size, and the total number of pairs of bases in one set of twenty-three chromosomes is approximately $3 \times 10^9$. The Human Genome Project is an ambitious (some would say audacious) "... international effort to develop genetic and physical maps and determine the DNA sequence of the human genome and the genomes of several model organisms" [19].

What are physical maps? The answer is not as precise as one would like. To understand this, we must first understand something about recombinant DNA techniques as well as current limitations in how regions of DNA can be analyzed by molecular geneticists. (See [13] for a readable introduction to recombinant DNA techniques and genetic analysis.) The twin overriding facts of life are that

- current methods of chemically analyzing substantial stretches of DNA require a sample containing a large number of identical molecules, typically produced by recombinant DNA amplification; however

- the maximum size of a region that can be amplified by current techniques is orders of magnitude smaller than even the smallest human chromosome.

For example, the size of the longest contiguous fragment of DNA that can be reliably amplified by a recombinant DNA process called "cloning" (see below) ranges from around $4 \times 10^4$ to $1 \times 10^6$, depending on the vector and host. Similarly, the longest stretch of DNA that can be reliably amplified by a purely chemical technique known
as polymerase chain reaction (PCR) is approximately $1 \times 10^3$ bases. In contrast, the twenty-two human autosomes range in size from around $3 \times 10^8$ bases for chromosome 1 down to about $5 \times 10^7$ bases for chromosome 21. Because of this mismatch in sizes, producing enough DNA to permit biochemical analyses currently requires a process called cloning, in which

- a large number of identical chromosomes are broken randomly into fragments by one or more of a class of enzymes known historically as restriction enzymes,
- individual fragments of appropriate size are incorporated by biological or chemical mechanisms into the DNA of host organisms such as *E. coli* or yeast,
- the individual hosts are separated from each other and allowed to grow into colonies, with the fragment in each host being replicated along with the DNA of the host during cell division (*mitosis*).

In this way, the natural DNA replication machinery of the host organism is exploited to replicate the fragment along with the host’s chromosomes. After enough mitoses, each host colony can be harvested. The result of this process is a library of cloned chromosome fragments, where each fragment is present in large enough quantities to permit isolation and purification of the fragment and subsequent biochemical analyses. Unfortunately, the library contains no information about the relative positions of the fragments along the chromosome. Physical maps are data structures which provide the necessary information to enable the order and distance among fragments to be deduced. Hence, they are essential if a collection of overlapping cloned chromosome fragments (a contig) is to be treated as though it were a contiguous region of DNA.

Sometimes the ordering information about a contig is exactly that: the relative order of the cloned fragments themselves. In this case, the physical map is intimately related to the clone library from which it arose. However, this need not be the case. Olson et al [58] propose a “common language” for physical mapping based on the notion of a sequence-tagged site, or STS. An STS is a small sequence of DNA that occurs precisely once in the human genome and can be reliably assayed by PCR. In their proposal, a physical map would consist of a linear sequence of STSSs, along
with the instructions necessary to construct reliable assays for each STS. In this case, the map would not be tied to a particular set of clones, but could be used to order any subsequently generated library. The needed resolution, or distance between STSS, would depend on the average size of cloned fragments to be ordered: "The main practical requirement is that the resolution should be high enough to make regeneration of cloned coverage of any region straightforward" [58]. Indeed, even if a physical map is constructed for a particular library, one can then create a map of STSS from the ordered library to produce an STS map for subsequent general use.

As the above paragraph indicates, the critical requirement for an STS map to be useful is that the distance between neighboring STSS be suitably matched to the average size of a cloned fragment: the larger the cloned fragments, the farther apart the STSS can be. Not surprisingly, then, most of the maps published to date which use STS techniques are for libraries containing very large fragments (see Section 1.2 below).

In summary, the main purpose of a physical map is to enable a set of DNA fragments to be treated as though it were a contiguous stretch of a chromosome. Currently published physical maps provides order information for a particular set of clones. These maps typically contain STSS, which can, in principle, be used to generate contigs from other libraries. However, the density of STSS in these maps is such that they require large cloned inserts. Despite that, an ultimate goal is to produce a detailed set of abstract landmarks which can be maintained in a database and not be associated with any particular library at all.

Finally, knowing the order among clones also allows one to construct another kind of map known as a restriction map, which can in turn help verify the ordering information. (See Figure 4.7 for an example of a restriction map.) Restriction maps document the order and distance along the genome between sequences of DNA known as restriction sites. A restriction site is the location of a sequence, typically four to six bases long, where a particular restriction enzyme will cut the DNA. Different enzymes recognize and cut different sequences. The distances between restriction sites in a cloned segment of DNA can be determined by digesting the cloned segment with the enzyme and observing the sizes of the resulting pieces by a process known as gel elec-
trophoresis. Since overlapping cloned segments must share restriction sites, ordering information can be used to generate restriction maps, and conversely, restriction maps can be used to validate orderings. Restriction maps can also be vital in identifying features known as polymorphisms, as we shall see below. (Polymorphisms are regions of DNA which tend to be different in different individuals.)

1.2 Constructing Physical Maps

In this section, we describe in more detail some of the issues involved in constructing a large-scale physical map. Some early efforts to explore these issues include those by Coulson et al [20], Olson et al [59], and Kohara et al [40]. Constructing a map of a chromosome, or of a large region on a chromosome, can be broken down into several basic steps:

1. create a library of cloned fragments as described above;
2. produce a data “fingerprint” for each cloned fragment; and
3. use the information in each fragment’s fingerprint to assemble a physical map of the region of interest.

By fingerprinting a clone, we mean performing one or more experiments on that clone, the results of which depend in some way on the underlying DNA sequence. Hence, the results of these experiments can help identify or characterize the clone. Cloned fragments which overlap, i.e., share a portion of the genome, may produce fingerprints more similar to one another than clones which do not overlap. We then use the similarity between fingerprints as a measure of similarity between clones.

Differences between maps and map-making methods boil down to making different choices in the above steps. For instance, in STS-content maps [31], a “fingerprint” consists of an enumeration of the STSs contained in a cloned fragment. Hence two clones overlap whenever they share an STS, but may very well overlap without sharing any STSs, if the resolution is too coarse. On the other hand, in the high-resolution map of chromosome 19 being produced by Lawrence Livermore National Laboratory
(LLNL), a fingerprint is a list of observed fragment sizes resulting from a complicated digestion of the cloned fragment. In this situation, statistical methods are used to compute a posterior probability of overlap, given the data from the fingerprint. (We will describe LLNL’s approach in much more detail in Chapter 4 below.) Let us now look at each of the steps enumerated above in a little more detail.

1.2.1 Producing a Library of Cloned Fragments

From the point-of-view of subsequent analyses, a library is just a “random” sample of overlapping DNA intervals. However, the choices made at this step are critical in determining the subsequent uses to which the map can be put.

Molecular cloning requires that the DNA to be cloned — the insert — be joined to another DNA molecule — the vector — that can replicate in host cells. This joining is carried out in vitro and the resulting recombinant DNA molecules are then introduced into the host cells. Common hosts include the bacterium E. coli and baker’s yeast S. cerevisiae. Vectors which are used with E. coli include naturally occurring genetic elements known as plasmids, of which pBR322 is perhaps the most well known, and bacteriophages (bacterial viruses) such as the phage λ. The LLNL library we discuss below consists of cosmids, hybrid vectors which replicate like plasmids but can be packaged in vitro into λ coats. There are also naturally occurring yeast plasmids, but more important are yeast artificial chromosomes, or YACs, which consist of the insert joined to a specially designed piece of DNA which functions as a synthetic yeast chromosome.

The importance of all of this is that different hosts and methods of incorporation can vary widely in the average size, variability in size, and subsequent stability of the incorporated fragment. For instance,

- YACs in yeast may have DNA fragments ranging from about 0.1–1 × 10⁶ bases in length;

- cosmids in E. coli may have fragments ranging from about 3.5–4.5 × 10⁴ bases in length;
- YACs are much more prone than cosmids to *chimerism*, where two or more fragments are combined in the same clone;

- some host strains can exhibit *cloning bias*, preferentially amplifying fragments from certain regions of the genome over others.

How important each of these characteristics is depends on how the map will be used. One can see that, as far as efficiency of coverage is concerned, the order-of-magnitude difference between the size of an insert in YACs and other methods make YAC-based libraries appealing. However, YAC-based maps have their weaknesses, too. For instance, YAC-based maps are, by themselves, too coarse-grained for many purposes, such as sequencing. Consequently, if the map is to be used to guide a large-scale sequencing effort, individual YACs must themselves be sub-cloned into smaller pieces of a size suitable for sequencing. In addition, the high rate of chimerism in YAC libraries can create considerable complications in map assembly. On the other hand, YAC-based maps have proved immensely useful in guiding the construction of higher-resolution maps. For instance, Baxendale et al [6] and Hoheisel et al [36] have produced cosmid-based maps of the Huntington region \((2 \times 10^6 \text{ bases})\) and the *S. pombe* genome \((1.4 \times 10^7 \text{ bases})\), respectively. In both of these cases, they had a pre-existing YAC map to guide the construction of a higher-level map. One can imagine that the other “first-generation” YAC-based maps, such as those described in [7, 16, 28, 18], will prove similarly valuable in guiding and focusing efforts towards higher resolution maps of regions of interest, including whole human chromosomes.

### 1.2.2 Producing “Fingerprint” Data for each Clone

Recall that, given a library of clones, our task is to determine the ordering relationships between them so they can be treated as though they were a single, contiguous piece of DNA. In most cases, one can imagine this task proceeding in two logical steps:

- analyze each clone to produce a vector of data which depend on its DNA content; then

- use this information to assemble contigs.
We now examine some of the kinds of data produced in mapping projects. Most types of data record the result of probing clones for the presence of particular sequences of DNA. Such probing is usually carried out by PCR, restriction digestion, or a process called *hybridization*, a kind of chemical analogue of the process of aligning strings of letters so they match [34]. In a hybridization experiment, the similarity between two single strands of DNA (a *probe* and a *target*) is measured by observing the extent to which the two single strands can form a duplex of complementary base-pair sequences at a particular temperature, pH, etc. Generally speaking, strands which are perfectly complementary will remain bound into a duplex at measurably higher temperatures than strands containing base-pair mismatches. By measuring the amount of probe that sticks to a target, one can (imperfectly) estimate whether or not a probe and a target share DNA.

Some data represent events that should happen at most once in the genome. For instance, one can assay clones for the presence of an STS. Because of the uniqueness of the sequence represented by the STS, if two clones both test positive for the STS, then we can conclude (barring experimental error) that the two clones overlap. Let us call this kind of data *unique sequence* data. Other data represent events that can happen more than once in the genome. For instance, we can assay clones for the presence of a particular small sequence like GGAATTCC. For any given clone, barring experimental error, the sequence either will or will not appear. However, as the sequence is likely to occur by chance in many different places in the genome, we cannot conclude that two clones overlap simply because they both test positive for this sequence. Such an event *does* provide some evidence for overlap. Let us call this kind of data *repetitive sequence* data. Finally, we can digest the clone using one or more restriction enzymes and observe the sizes of the resulting fragments. Again, the two clones will both contain a fragment of size $k$ whenever they both have two consecutive restriction sites that are $k$ bases apart. In this case, some experimental error is inevitable, as fragment sizes can only be measured with a relative error of several percent at best. As with the repetitive sequence data, two clones containing a fragment of a given length provides some positive evidence that the clones overlap, but is by no means conclusive proof.
Other types of data include combinations of the above approaches. For instance, Stallings et al [64] describe an approach in which fragments from restriction enzyme digestions are hybridized by repetitive sequence probes. The fact that two clones both contain a fragment of size $k$ which contains a repetitive sequence provides more compelling evidence that the two clones overlap than either event by itself.

1.2.3 Assembling Contigs and Obtaining Closure

Once we have data on individual clones, we use that data to assemble the clones into contigs. How we go about that depends to a great extent upon the type of data we have. In the case of constructing maps from unique sequence probes such as STSS, the problem is largely algorithmic. We must find an ordering of STSS and an associated configuration of clonal overlaps that is consistent with the data.

If there were no errors, this problem could be easily solved by testing whether an incidence matrix summarizing the data has the “consecutive ones property.” An example will make things clear. Consider the overlap configuration in Figure 1.2. The horizontal lines represent clones, and the vertical arrows represent STSS. Given this configuration, we can construct an incidence matrix describing which unique sequence probes are positive for which clones. This matrix will have a 1 in row $i$ and column $j$ if clone $i$ contains unique sequence $j$, and a 0 otherwise. Such a configuration would
produce the incidence matrix on the left hand side of Eqn. (1.1).

<table>
<thead>
<tr>
<th>STSS</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>STSS</th>
<th>b</th>
<th>a</th>
<th>e</th>
<th>d</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1.1) CLONES</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>⇒</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

If we permute the columns of the matrix on the left hand side of Eqn. (1.1) into the order \((b, a, e, d, c)\), we will produce the matrix on the right hand side of Eqn. (1.1). In this permuted matrix, all the 1s in each row appear consecutively. If we visualize each row of 1s as a clone, this permuted matrix shows us an overlap relationship which is consistent with the data. Note that not all overlaps are captured by the matrix. For instance, in our example, clones 4 and 5 overlap, but this fact is hidden from us by a lack of unique sequence probes in the overlap region. Such *cryptic overlaps* are quite common. However, it’s the best we can do with the data at hand, as the density of probes in the example is too coarse to detect all overlaps.

An incidence matrix like the incidence matrix in our example has the *consecutive ones property for rows* if its columns can be permuted so as to make all the 1s in each row appear consecutively. The matrix on the left hand side of Eqn. (1.1) has the consecutive ones property for rows, as seen by its permuted version on the right hand side of Eqn. (1.1). It is immediately clear that incidence matrices corresponding to error-free probings of clones with unique sequence probes

- have the consecutive ones property;
- the permutation operation corresponds to finding an ordering of STSSs consistent with the data; and
- the structure of the permuted matrix provides a consistent overlap configuration for the clones.
It is easy to determine if a matrix has the consecutive ones property. Booth and Lueker [8] describe linear time algorithms for determining if a matrix has the consecutive ones property. Furthermore, their algorithms also provide a compact description of all possible consistent permutations in the form of a special data structure called a “PQ-tree.” Hence, the problem is completely solvable in linear time when the data are error-free.

Unfortunately, real datasets are never error free. Because of occasional experimental failures, frequent clone chimerism, mislabeling, PCR or hybridization error, measurement error, etc., real data sets of any size will inevitably contain some erroneous entries. In this case, the relationship between consistent orderings and the consecutive ones property is destroyed, and the problem of map assembly from unique sequence data ceases to be well-posed. One can continue to use approaches based on PQ-trees, but now one must use heuristics to search for a nearby matrix which has the consecutive ones property. Instead, the problem is typically recast as a combinatorial optimization problem: search through the space of all linear orderings of the objects (probes or clones) to find one that minimizes some user-defined penalty function. The underlying data structure is the same: an incidence matrix of clone-probe hits. The differences now lie in what is being ordered (rows for clones or columns for probes), how the penalty function is defined, and the search technique used to generate new permutations.

In nearly all published algorithms, the penalty function is defined to be the sum of pairwise discrepancies between adjacent objects. In this case, the problem is formally identical to the Traveling Salesman Problem [44] on a complete graph. The nodes on this graph correspond to the objects to be ordered, either clones or probes. An edge between nodes $a$ and $b$ is weighted by a measure $d(a, b)$ of the discrepancy between objects $a$ and $b$. The goal is to find a path visiting every node exactly once which minimizes the sum of the discrepancies along the edges in the path. Despite this formal similarity, most efforts to date do not exploit the considerable body of research in efficient solution techniques for the Traveling Salesman Problem (see [57] for an exception). Perhaps this is because simpler heuristics have proved successful on problems encountered to date.
For example, Mott et al [54] order probes by defining the discrepancy between probes \(a\) and \(b\) as

\[
d(a, b) = 1 - \frac{\text{Number of clones positive for } a \text{ and } b}{\text{Number of clones positive for } a \text{ or } b}
\]

They then use simple reordering rules to generate new permutations and simulated annealing to determine whether or not to accept the newly generated permutation. Once probes have been ordered, clones are then ordered with respect to the probes by an algorithm that maximizes a measure of fit between the probe data for that clone and the list of ordered probes. This approach has proved successful in constructing maps of the Huntington region [6] and genome of the fission yeast \(S.\) pombe [36]. In addition, they also describe a different, heuristic ordering procedure which attempts to edit the clone-probe matrix to restore consistency. In addition to the work of Mott et al, Cuticchia et al [22] describe efforts to construct maps from simulated clone-probe matrices via simulated annealing. In this case, clones were ordered using the \(\ell_1\) distance between the rows of the clone-probe matrix as a measure of discrepancy.

The combinatorial optimization approach to finding an optimal ordering described above is based on a simple, two-part strategy: compute a measure of discrepancy between all pairs of objects, and then search through the space of all orderings for a satisfactory solution. An assumption underlying this is that the data collection process is largely independent of the map assembly process. Palazzolo et al [60] describe an entirely different, “directed” approach, in which the overlap information gathered at step \(n\) directs the experimental procedure at step \(n + 1\). The method proceeds roughly as follows, on a library of \(N\) clones . . .

1. Begin with all \(N\) clones unlabeled.

2. Repeat until all clones are labeled:

   (a) Choose an unlabeled clone at random; call this clone a “seed” clone.

   (b) Create STSS from the two ends of the seed clone, and use them to find all clones that overlap the ends of the seed clone; call this set \(S\).
(c) If one or more clones in \( S \) is already labeled, merge contigs by relabeling all the clones in \( S \), as well as the seed clone, with the smallest label among the labeled clones in \( S \).

(d) If no clone in \( S \) is labeled, add a new contig by labeling all the clones in \( S \), as well as the seed clone, with the next available contig number.

When complete, all \( N \) clones will be labeled with a contig number. Note that, as described, no mention is made of ordering probes or clones. However, evolving maps of contigs can be created and maintained as the process evolves. PQ-tree approaches can be used to edit and validate the evolving contigs. Merged contigs whose incidence matrices do not possess the consecutive ones property are symptomatic of an error, which can be localized and corrected. This approach, the success of which depends critically on the ability to do large numbers of high-accuracy probings economically, has been used successfully to map the \( S. pombe \) genome in cosmids by Mizukami et al [53].

Finally, recall that the clones in any library can be considered as a random sample of segments from some underlying genome. Hence, the map will actually consist of a collection of contigs separated by regions of DNA not covered by any clone. As noted by Lander and Waterman [43], for a library of \( N \) clones of average length \( L \) taken from a genome of length \( G \), the expected proportion of the genome left uncovered is approximately \( \exp(-NL/G) \). The above approaches provide an ordering, but do not unambiguously define contig boundaries.

Now let us examine how maps such as Bellarme-Chantelot [7] and Stallings et al [63] have been assembled from repetitive sequence data, restriction fragment data, or a combination of the two. In this situation, one can follow the same basic program as above: use the data to obtain a similarity measure between clones, and then order and assemble the clones into contigs using some optimization approach. However, in this situation the random, yet repetitive nature of the data lends itself to a probabilistic interpretation and to statistical approaches to detecting and evaluating potential clone configurations. Unfortunately, because of the complexity of the underlying models, most current efforts confine statistical decision making to detecting pairwise overlap,
and then retreat to combinatorial heuristics to assemble the map.

Statistical approaches to detecting pairwise overlap, as described in [52, 11, 4, 30], and others, all begin with a probability model for the data generating process, given an overlap configuration between two clones. Most then compute an integrated likelihood ratio or posterior probability of overlap, which they use as a similarity measure, although Fu et al [30] cast the problem as a hypothesis testing problem. The approach described in [4] has been used in [7] and [63], while the approach first outlined in [11] will be discussed in detail in Chapter 4.

Alizadeh et al [1], and Nelson et al [56] describe different statistical approaches to evaluate overlap configurations involving more than two clones. Nelson et al recast the problem of determining overlap as a Bayesian decision problem, and examine solutions from an information-theoretic point of view. Alizadeh et al examine the case of error-free data generated by repetitive probes against constant length clones. They propose methods to evaluate overlap configurations among many clones. Unlike the STS case, there is no simple mapping between a configuration of clones and a permuted version of the clone-probe matrix. Probes will occur more than once along the genome, and part of the problem is to infer a probe ordering where each probe may occur more than once. In this way, the problem once again becomes one of combinatorial optimization. Alizadeh et al describe efficient algorithms to approximate the relative likelihood of any sequence of probes, given a particular overlap configuration. They then use techniques developed for the Traveling Salesman Problem to guide the search through the space of clone overlap configurations.

1.2.4 Other Issues

As one can surmise from the size of the genomes and libraries involved, constructing a map of a large region or chromosome is a major, multi-year project. As a consequence, such efforts are often preceded by a considerable amount of analysis and simulation, designed to determine the feasibility, duration, cost, and likelihood of success of various mapping approaches.

Suppose the goal of a mapping project were to cover $P$ percent of a chromosome
in $K$ or fewer cosmid contigs. Project managers will pose questions like . . .

“We can construct a system that will detect fifty percent overlap reliably. However, constructing a system that will detect ten percent overlap reliably will be much harder to accomplish and be more expensive per fingerprint to use. On the other hand, maintaining large libraries is also expensive. How many more clones will I need to analyze if I use a cheap, fifty percent overlap detector rather than an expensive, ten percent overlap detector?”

or

“After we finish we will have around $K$ contigs. We will have to close the gaps between these contigs by very expensive methods. How many of the gaps between contigs are actually cryptic overlaps that we will be able to close with one well-selected probe, and how many are real?”

or

“Suppose we have to stop after we’ve analyzed 5,000 clones (using whatever method we’ve chosen). How close to completion will we be? In other words,

- How many contigs will there be?
- What will be the average size of a contig?
- How much of the genome will be covered?”

and so on. Of course, the answers to these questions depend on the strategy used to construct the map. Lander and Waterman [43] addressed these questions for libraries constructed by repetitive sequence fingerprinting of random clones. The models used were similar to those derived by Feller [26] and Smith [62] for analyzing the behavior of Geiger counters. Several authors, using several different approaches, have produced results allowing one to predict progress in constructing STS-content maps [2, 5, 25, 32, 66]. Finally, we describe new results in Chapter 3 which allow one to predict progress in directed mapping projects.
1.3 Finding the Gene for Myotonic Dystrophy

To gain some appreciation of the work involved in isolating a gene, and the role that physical maps play, let us look at the sequence of events leading up to the discovery of the gene associated with Myotonic Dystrophy ("DM"). Myotonic Dystrophy [35] is the most common form of adult muscular dystrophy, with prevalence estimates ranging from two to fourteen cases per 100,000 individuals. The disease allele is strongly associated with a wide range of disorders, from myotonia and other neurological defects to cataracts. The age at onset and disease severity can also vary widely: some individuals remain asymptomatic as adults, while others present severe symptoms at birth.

As early as 1983, the locus for DM had been mapped by linkage analyses to chromosome 19. Six years later, Korneluk et al [41] published an analysis of markers which localized the DM gene to a ten centimorgan region on the long or q-arm of chromosome 19, denoted by 19q, between the DNA excision repair gene ERCC1 and marker D19S50. Two years later, Tsilfidis et al [67] further localized the DM locus to a two centimorgan stretch on the band of 19q designated 19q13.3 by finding a recombination event in a pedigree which placed a polymorphic marker D19S51 just distal to the DM locus. Assuming that one centimorgan corresponds, on average, to $1 \times 10^6$ bases of DNA, this finding localized the DM locus to a $2 \times 10^6$ region on chromosome 19 which was flanked by known markers: ERCC1 and D19S51.

Up to this point, efforts to isolate DM had been focused on finding flanking markers close enough to make molecular analyses feasible. The next step would be to construct a high resolution physical map, consisting of

- a set of overlapping cosmid and YAC clones which would span the region between the two flanking markers, coupled with

- restriction maps of the region.

These maps could then be used to isolate and analyze potential genes in the region.

By the end of 1991, as the result of an intensive international effort, the region was completely cloned and mapped, and a putative defect was observed [3]. The putative
defect was a length variant observed in one of the restriction fragments among affected members. (For a detailed description of the mapping effort, see [14, 38, 61].) Shortly thereafter, Brook et al [12] published evidence of the molecular basis of the length polymorphism: an unstable expansion of a three base repeat (CTG) at the 3' end of a new gene. Within a month Mahadevan et al [48] published results indicating that the unstable region was in an untranslated region of the gene. By the end of 1992, the gene containing the repeat had been identified as a type of protein kinase and completely sequenced [49]. Finally, by the end of 1993, diagnostic probes for DM had become commercially available, even though the mechanism by which the mutation resulted in a DM phenotype remained unknown. The mechanism remains unknown today.

Note that most of the time and effort involved in locating and characterizing DM was spent finding suitable markers bracketing a small enough stretch of DNA to make molecular analysis feasible, and much of the subsequent effort involved constructing physical maps of the localized region. Once such an infrastructure was in place, isolating and characterizing candidate genes could proceed quickly. This division of labor is not unusual: see [6] and [47] for details concerning the corresponding search for the gene for Huntington’s disease. Once produced, high resolution physical maps provide the infrastructure needed to enable efficient

- discovery of new polymorphic markers that more tightly bracket the region of interest;
- screening for genes using other libraries of candidate sequences called cDNAs; and
- sequencing of candidate genes.

Thus, high resolution physical maps provide a critical resource to molecular geneticists interested in understanding our genetic makeup.
Chapter 2

Predicting Progress in Random Mapping Projects

2.1 Introduction

In 1988, Lander and Waterman [43] published research which provided approximations to parameters of interest to biologists designing physical mapping projects based on fingerprinting random clones. In their analysis, they restricted their attention to first moment approximations for the case of clones of constant length $L$. In this chapter, we summarize and clarify their results, providing variance estimates and explicit forms for distributions when possible. Also, when possible we describe the more general result for clones of varying length.

The model we use is that of a a simple marked Poisson process the real line $\mathbb{R}$. That is, we have a random set of sites $\mathcal{P} := \{\xi_i : -\infty < i < \infty\}$, which are laid down according to a homogeneous Poisson process with rate $\lambda$. Associated with each site $\xi_i$ is a length $L_i$ defining the length of the clone. The intervals are IID with a distribution function $F$ with $\mathbb{E}L_i = L$, as well as being independent of $\mathcal{P}$. Usually, the site will be the beginning of the clone, but it doesn’t really matter: if it’s convenient, we can just as well let the site be the right end of the clone. The analysis is identical.

We want to know things about the way the intervals segregate into islands in an arbitrary interval $(0, G)$. One parameter we will use a lot will be the coverage
\[ c := \lambda L. \] We call \( c \) the coverage because if we lay down \( N \) segments of mean size \( L \) in the interval \((0, G)\), then the number of clones covering a site, on average, will be \( \frac{NL}{G} \). At the same time, we estimate the Poisson intensity by \( \frac{N}{G} \), so we have
\[ c := \lambda L \approx \frac{NL}{G} = \frac{NL}{G}. \]

We haven’t specified anything about where the origin is in relation to the \( \xi_i \), but that’s no problem. The point process is stationary, so we can translate the origin around to suit our purposes. Sometimes we will assume an island starts at the origin, and other times we will assume an island ends at the origin.

Some of the estimates we want include deriving moment estimates for

- the number of islands in \((0, G)\),
- the number of clones in an island,
- the total length of an island,
- the distance between adjacent islands, and
- the proportion of \((0, G)\) covered by islands.

### 2.2 Pure Lander-Waterman

This section examines the above questions in the simplest case: bottom-up fingerprinting of random clones. We can use (at least) two approaches to computing probabilities in this situation. We can try for “exact” results on \((0, G)\), or we can convert the problem into one involving stationary Poisson processes to get simpler asymptotic results.

#### 2.2.1 How Many Islands

An island begins whenever a site \( \xi_i \) is not covered by an interval from some earlier site. There are a number of ways to derive an expression for the expected number of islands. The basic idea is to derive an expression for \( p(t) \), the probability that \( t \)
will be uncovered, and then observe that the beginnings of islands can be modeled by a point process whose intensity process has expectation at \( t \) of \( \lambda p(t) \). Thus, the expected number of islands of one or more clones in \((0, G)\) will be \( \int_0^G \lambda p(t) \, dt \).

We compute \( p(t) \) by a thinning argument: fix \( t \), then keep only those clones that overlap \( t \), and compute the probability than no clones remain in \((0, t)\). That is, we keep \( \xi_i \) whenever \( 0 < \xi_i < t \) but \( \xi_i + L_i > t \). Since the thinning probability for \( \xi_i \) depends only on \( \xi_i \) and \( L_i \), this thinned process is just a thinned Poisson process, with retention probability \( \Pr\{L_i > t - \xi_i\} \). Hence, the probability of \( t \) being retained is just the probability of no events in \((0, t)\):

\[
(2.1) \quad p(t) = \exp \left( -\lambda \int_0^t (1 - F(t - s)) \, ds \right) = \exp \left( -\lambda \int_0^t (1 - F(s)) \, ds \right)
\]

and the expected number of contigs in \((0, G)\), denoted \( \mathbb{E}N_c(G) \) is

\[
\mathbb{E}N_c(G) = \int_0^G \lambda \exp \left( -\lambda \int_0^t (1 - F(s)) \, ds \right) \, dt
\]

In general, though, this integral has no closed form solution. So, we retreat to the asymptotic case involving a stationary Poisson process on \( \mathbb{R} \). In this case, the probability that \( t \) is uncovered is

\[
(2.2) \quad p(t) = \exp \left( -\lambda \int_{-\infty}^t (1 - F(t - s)) \, ds \right) = \exp \left( -\lambda \int_{0}^\infty (1 - F(s)) \, ds \right)
\]

Hence, the expected number of islands in \((0, G)\) is just

\[
\mathbb{E}N_c(G) \approx \int_0^G \lambda p \, dt = Ne^{-\lambda}
\]

We can also get an asymptotic variance in the general case using renewal theory. We know from renewal theory that, as \( t \to \infty \),

\[
\mathbb{E}N_c(t) \sim \frac{t}{\mu_1} \quad \text{and} \quad \text{Var} \, N_c(t) \sim t \frac{\mu_2 - \mu_1^2}{\mu_1^3}
\]

where \( \mu_1 \) and \( \mu_2 \) are the first two moments of the inter-arrival distribution for our contig process. We just need those first two moments, and Smith [62] shows us how.
He shows that, for alternating renewal processes where the second process is Poisson with rate $\lambda$,

$$
\lim_{t \to \infty} p(t) = \frac{1}{\lambda \mu_1}
$$

where $p(t)$ is the probability that $t$ is in an interval associated with the second process. This is just the $p(t)$ described by Equation 2.1, whose limit is described in Equation 2.2. In addition, Smith proves that $\mu_2 = 2\mu_1^2(1 + \beta)$, where

$$
\beta = \lambda \int_0^\infty \left( p(u) - \frac{1}{\lambda \mu_1} \right) du
$$

Since in our case we have $\lim_{t \to \infty} p(t) = e^{-c} = 1/\lambda \mu_1$,

$$
\Var N_c(G) \sim G \frac{2\mu_1^2(1 + \beta) - \mu_1^2}{\mu_1^3} = \frac{G}{\mu_1} (1 + 2\beta) = Ne^{-c}(1 + 2\beta)
$$

### 2.2.2 The Number of Clones in an Island

The number of clones in an island is called its order. Computing the distribution of the order of an island is trivial, if all the clones are the same size. In that case, an island ends and a new island begins whenever the interarrival time between two sites exceeds $L$, an event which occurs with probability $e^{-\lambda L} = e^{-c}$. Thus, the number of clones in an island is geometrically distributed with a success probability of $e^{-c}$, and

$$
\Pr\{\text{Number of clones} = k\} = \left(1 - e^{-c}\right)^{k-1} e^{-c}
$$

with expectation $e^c$.

When the clones aren’t the same size, the problem gets much more complicated. However, if we retreat to asymptopia, we see that the average number of clones per island is the ratio of the clone intensity to the island intensity: $\lambda/(\lambda e^{-c}) = e^c$.

### 2.2.3 The Length of an Island

The distribution of the length of an island, even in the simple case of equal length clones, is hard. Hall [33] derives the Laplace-Stieltjes transform $\gamma(s)$ of the distribu-
tion $C$ of the length of an island:

$$
\gamma(s) = \int_0^\infty e^{-sx} dP(C \leq x) = 1 + \frac{s}{\lambda} - \left[ \lambda \int_0^\infty \exp \left( -st - \lambda \int_0^t \{1 - F(x)\} dx \right) dt \right]^{-1}
$$

From this we can derive the expected length of an island

$$
\mathbb{E}C = \frac{1}{\lambda} (e^c - 1)
$$

as well as its variance

$$
\text{Var}(C) = \frac{2}{\lambda^2} e^c \int_0^\infty \left[ \exp \left( \lambda \int_0^t \{1 - F(x)\} dx \right) - 1 \right] dt - \frac{1}{\lambda^2} (e^c - 1)^2
$$

When all the segments are the same length $L$, $F(x)$ becomes a step function, and the variance simplifies to

$$
\text{Var}(C) = \frac{1}{\lambda^2} (e^{2c} - 2ce^c - 1)
$$

As for the distribution of the length of an island, the inverse Laplace-Stieltjes transform will not generally be computable. However, Hall gives some results for the case where all the segments are of length $L$. In that case, the distribution for $C$ has an atom of size $e^{-c}$ at $L$ and an absolutely continuous component on $(L, \infty)$ of the form

$$
f(x) = \frac{\lambda}{e^c - 1} \left[ 1 + \sum_{j=1}^{\lfloor x/L \rfloor - 1} \frac{(-1)^j}{j!} \{\lambda(x - (j+1)L)\}^{j-1} e^{-jc} \{\lambda(x - (j+1)L) + j\} \right]
$$

where $\lfloor x \rfloor$ is the integer portion of $x$.

For large coverage $c$, the island length is approximately exponentially distributed with mean $\lambda^{-1}(e^c - 1)$. On the other hand, for small coverage $c$, the continuous component is approximately uniform on $(L, 2L)$.

### 2.2.4 Distances Between Islands

To deduce the distribution of the inter-island distances $D$, we place the origin at the end of an island. Then the distribution of the distance to the next island is just
the distribution of interarrival times for $P$: exponential with parameter $\lambda$. Therefore,

$$\Pr\{D > d\} = e^{-\lambda d}$$

$$\mathbb{E}D = \frac{1}{\lambda}$$

$$\text{Var}(D) = \frac{1}{\lambda^2}$$

### 2.2.5 Total Coverage

Finally, how much of the interval $(0, G)$ will not be represented in clones? Lander and Waterman don’t discuss this, but it’s an important question whose answer is not that hard. The proportion left uncovered is just $V = \frac{1}{G} \int_0^G v(t) \, dt$ where

$$v(t) = \begin{cases} 
1 & \text{if for all } i, t \not\in (\xi_i, \xi_i + L_i), \\
0 & \text{otherwise}
\end{cases}$$

To find the expected vacancy $\mathbb{E}V$, use Fubini’s theorem to get

$$\mathbb{E}V = \frac{1}{G} \int_0^G \mathbb{E}v(t) \, dt = \frac{1}{G} \int_0^G p(t) \, dt$$

where $p(t)$ is defined as in Equation 2.1.

In the limit, $p(t) \to e^{-c}$, so the proportion left vacant approaches $e^{-c}$.

### 2.2.6 Overlap Detection

The above story was about the underlying process of contig formation. The reality of fingerprinting is that overlaps are not detected perfectly. Lander and Waterman used a simple model of overlap detection in their paper. They assumed that overlap was detected if and only if two clones overlapped by more than a proportion $\theta$ of their length, where $\theta$ is some unknown constant. In that case (and also assuming constant clone length), the formulas above go straight through, except that the fixed length $L$ is replaced by $(1 - \theta)L$. 
2.3 STS Content Mapping

After Lander and Waterman, a number of authors derived similar results for the more complex story of STS content mapping [2, 5, 25, 32, 66]. Of these, Arratia [2] is the most complete. The basic setup is as follows.

We have a genome of length $G$, and a library of $N$ clones whose lengths $\{L_i\}$ are IID from some distribution whose CDF we denote by $F(l)$. We have scaled $G$ and $F(l)$ so that $E(L) = 1$. In addition to the clones, we also have $M$ STSS we call “anchors.” We know the clones that contain any given STS, and we know which STSS are contained in any clone. We build contigs by anchoring clones with STSS, noting that whenever two clones share an anchor, they must overlap. The following diagram shows a typical situation.

![Typical situation in STS-content mapping.](image-url)

Note that two separate contigs can actually overlap, and that unanchored clones don’t enter into the contig picture at all.

We model the above situation using two independent homogeneous Poisson processes on $\mathbb{R}$: one for the clones and one for the anchors. The clone process $\mathcal{P}$ has rate $\lambda = N/G$, and each point $\xi_i$ is marked with its length $L_i$. This is the same process as in section 2.2. As before, the point can represent the left or right end of the clone, as need dictates. The anchor process is an independent Poisson process with rate $\mu = M/G$. 
2.3.1 How Many Anchored Clones

Not all clones will be anchored, and the anchored clone process will be a thinned version of the original clone process. The probability that a clone is not anchored, given \( L = l \), is just \( e^{ul} \). Hence, the probability that a given clone is not anchored is

\[ q_1 = \mathbb{E} e^{-\mu L} \]

and the process of anchored clones is a stationary process with intensity \((1 - q_1)\lambda\). By the ergodic theorem, as \( G \to \infty \), the number of anchored clones \( C_G \) behaves like

\[ \frac{C_G}{G} \xrightarrow{a.s.} (1 - q_1)\lambda \]

and the expected number of anchored clones is just \( \mathbb{E} C_G = N(1 - q_1) \).

2.3.2 How Many Anchored Islands

We determine the number of anchored islands by thinning all clones that aren’t the first clone in an anchored island, and looking at what is left. First, observe that for a Poisson process with rate \( \lambda \), the probability that an interval of length \( x \) is not covered by a single clone, a value we denote by \( J(x) \), is just

\[ J(x) = \exp \left( -\lambda \int_x^\infty \Pr\{L > l\} \, dl \right) \]

We can see this by looking, WLOG, at the interval \((0, x)\). Thin \( \mathcal{P} \) by retaining a clone at \( t \) whenever \( L + t > x \). The interval is not covered by a common clone whenever no clones remain in \((-\infty, 0)\), an event with probability

\[ \exp \left( -\lambda \int_0^0 \Pr\{L > x - t\} \, dt \right) = J(x) \]

With that in mind, a clone is the left hand end of an anchored island whenever

- it has an anchor, and
- no clones starting to its left has the same anchor.
Since \( \mathcal{P} \) is a homogeneous Poisson process, let the clone we’re looking at start at 0 and have length \( L \). Also, let \( V \) denote the location of the first anchor after 0. Then the above conditions are equivalent to the event that

- \( L > V \), and
- \([0, V]\) is not covered by any other clone.

For a fixed \( V = v \), this is just \( \Pr\{L > v\} \, J(v) \). After taking expectations with respect to \( V \), we see that the process of anchored left hand island ends is just a stationary process with intensity \( p_1 \lambda \), where

\[
 p_1 = \mathbb{E}_V \Pr\{L > V\} \, J(V)
\]

Again, by the ergodic theorem, as \( G \to \infty \), the number of anchored islands \( K_G \) behaves as

\[
 \frac{K_G}{G} \xrightarrow{a.s.} p_1 \lambda
\]

and the expected number of anchored islands is just \( \mathbb{E}K_G = Np_1 \).

### 2.3.3 The Number of Clones in an Anchored Island

As one might expect, the average of the number of clones in an anchored island (its “order”) converges to the ratio of the anchored clone intensity to the anchored island intensity, as \( G \) gets large. To see this, let \( \{M_i \mid i \in \mathbb{Z}\} \) be the stationary process of anchored island orders, numbered so that island 1 overlaps the origin, i.e., \( \{M_1, \ldots, M_{K_G}\} \) denotes the number of clones in each of the \( K_G \) anchored islands whose right ends occur in \((0, G)\). Arratia et al define their “average number of clones per anchored island” as a limit

\[
 \lim_{G \to \infty} \frac{1}{K_G} \sum_{j=1}^{K_G} M_j
\]

whose value they derive as follows.
First, they note that the sequence \( \{M_1, \ldots, M_{K_G}\} \) is not stationary, due to larger than average size of \( M_1 \) (the "waiting time paradox"). Let \( M'_G \) denote the "excess" clones in \( M_1 \):

\[
M'_G = \sum_{j=1}^{K_G} M_j - C_G
\]

where, recall, \( C_G \) is the number of anchored clones in \((0, G)\). Then

\[
\frac{1}{K_G} \sum_{j=1}^{K_G} M_j = \frac{G}{K_G} \left[ \frac{M'_G}{G} + \frac{C_G}{G} \right] \xrightarrow{a.s.} \frac{1}{p_1 \lambda} \left[ 0 + (1 - q_1) \lambda \right] = \frac{1 - q_1}{p_1}
\]

by Equations 2.3 and 2.5.

### 2.3.4 The Length of an Anchored Island

As in the last section, the average length of an anchored island is defined by a limit. Now, let \( \{S_i \mid i \in \mathbb{Z}\} \) denote the stationary process of contig sizes, numbered as above. That is, \( \{S_1, \ldots, S_{K_G}\} \) denote the length in each of the \( K_G \) anchored islands whose right ends occur in \((0, G)\). Arratia et al define their "average length of an anchored island" as

\[
\lim_{G \to \infty} \frac{1}{K_G} \sum_{j=1}^{K_G} S_j
\]

whose value they derive as follows.

Let \( r_i \) denote the probability that a point is covered by exactly \( i \) anchored islands. Since a point can fall in at most two anchored islands, we know that \( r_0 + r_1 + r_2 = 1 \). Also,

\[
\frac{1}{G} \sum_{j=1}^{K_G} S_j \xrightarrow{a.s.} r_1 + 2r_2
\]

as \( G \to \infty \). Hence,

\[
(2.6) \quad \frac{1}{K_G} \sum_{j=1}^{K_G} S_j = \frac{G}{K_G} \left( \frac{1}{G} \sum_{j=1}^{K_G} S_j \right) \xrightarrow{a.s.} \frac{r_1 + 2r_2}{p_1 \lambda} = \frac{1 - r_0 + r_2}{p_1 \lambda}
\]

Estimating \( r_0 \) and \( r_2 \) is easy. First, look at \( r_0 \), the probability of a point \( x \) not being covered by an anchored island. Let \( V \) denote the distance back to the last anchor, and let \( W \) denote the distance forward to the next anchor. The \( x \) is not in an anchored island whenever
• any clone starting to the left of $x - V$ ends before $x$, and
• any clone starting in $(x - V, x]$ ends before $x + W$.

We can compute this probability by the thinning trick used earlier:
• retain the clone at $t \in (-\infty, x - V]$ whenever $t + L > x$, and
• retain the clone at $t \in (x - V, x]$ whenever $t + L > x + W$.

Then the probability of no points in $(-\infty, x]$, conditional on $(V, W)$, is just
\[
\exp \left( -\lambda \left[ \int_{-\infty}^{x-V} \Pr\{L > x - s\} \, ds + \int_{x-V}^{x} \Pr\{L > x + W - s\} \, ds \right] \right) = \frac{J(V)J(W)}{J(V + W)}
\]
where $J(\cdot)$ is defined in Equation 2.4. Hence,
\begin{equation}
(2.7) \quad r_0 = \mathbb{E}_{(V,W)} \frac{J(V)J(W)}{J(V + W)}
\end{equation}

As for $r_2$, a point $x$ is covered by precisely two anchored islands whenever
• at least one clone starts in $(-\infty, x - V)$ and ends in $(x, x + W)$,
• no clone starts in $(-\infty, x - V)$ and ends in $(x + W, \infty)$, and
• at least one clone starts in $(x - V, x)$ and ends in $(x + W, \infty)$.

When the dust settles, the probability of this event is equal to
\begin{equation}
(2.8) \quad r_2 = \mathbb{E}_{(V,W)} \left( 1 - \frac{J(V)}{J(V + W)} \right) J(V + W) \left( 1 - \frac{J(W)}{J(V + W)} \right)
\end{equation}

Combining the results in Equations 2.6, 2.7, and 2.8 and simplifying results in
\[
\frac{1}{K_G} \sum_{j=1}^{K_G} S_j \xrightarrow{a.s.} 1 + \mathbb{E}J(V + W) - 2\mathbb{E}J(V) \xrightarrow{p_1} \frac{1}{p_1 \lambda}
\]

\subsection{2.3.5 Total Coverage}

We computed the total coverage $1 - r_0$ in the last section during the derivation of the length of an anchored island. To recapitulate, total coverage is just
\[
1 - r_0 = 1 - \mathbb{E}_{(V,W)} \frac{J(V)J(W)}{J(V + W)} = 1 - \int_0^\infty \int_0^\infty \mu^2 e^{-\mu(u+v)} \frac{J(u)J(v)}{J(u + v)} \, du \, dv
\]
Chapter 3
Predicting Progress in Directed Mapping Projects

3.1 Introduction

Mizukami et al and Palazzolo et al have recently reported physical mapping efforts using so-called “directed” approaches to construct their maps [53, 60]. Roughly speaking, directed approaches proceed sequentially by repeatedly

- selecting an unmapped clone (a “seed clone”) and constructing STSS from its ends, and then
- using these STSS to find the set of overlapping clones, typically by PCR assay.

In each iteration in the above process, only those clones which have not been previously identified as overlapping are eligible to be seed clones. The process continues until all clones have been either selected for STS production or identified as overlapping some selected clone. The purpose of this chapter is to describe an approach to computing measures of progress for mapping efforts of this kind.

Most, but not all, published methods to model progress in physical mapping projects have focused on random approaches, such as bottom-up fingerprinting [43] and STS-content mapping [2, 5, 25, 32, 66]. In addition, reported efforts to model di-
rected approaches [51, 71] used an approximation technique from statistical mechanics which required assuming that all insert lengths were the same.

In this chapter, we derive approximations that do not rely on the assumption of equal length. First, however, we give a simple alternative "finite" analysis based on the distribution of spacings between the seed clones. This analysis, when simple approximations to integrals are used, can give results nearly identical to those by Zhang and Marr [71]. Next, following the lead of Arratia et al [2] and using properties of stationary processes, we derive simple asymptotic formulae which apply equally to constant and variable clone lengths. We also give explicit, closed-form formulae for the case of constant length clones. Finally, in the case of constant length clones, we note that these results are equivalent to, and extend, previous published results of [71]. Simulations show these methods provide estimates well within the limits of uncertainty inherent in any mapping project.

Recall the directed approach to contig construction, initially described on page 12. One starts out with a library of \( N \) clones randomly selected from a chromosome and can build contigs in two separate ways: "single-end" probing and "double-end" probing. In both methods, the basic algorithm consists of drawing clones at random and labeling them with contig numbers. The algorithm goes like this:

1. Begin with all \( N \) clones unlabeled.

2. Repeat until all clones are labeled:
   
   (a) Choose an unlabeled clone at random; call this clone a "seed" clone.
   
   (b) Find all clones that overlap with the seed clone; call this set \( S \).
   
   (c) If one or more clones in \( S \) is already labeled, relabel all the clones in \( S \), as well as the seed clone, with the smallest label among the labeled clones in \( S \).
   
   (d) If no clone in \( S \) is labeled, label all the clones in \( S \), as well as the seed clone, with the next available contig number.

When complete, all \( N \) clones will be labeled with a contig number.
The only difference between the “single-end” and “double-end” methods is in what determines an overlap between a seed clone and the remaining clones. In the single-end method, one end of each seed clone is chosen at random, and remaining clones “overlap” that seed clone whenever they overlap the chosen end. In the double-end method, remaining clones “overlap” a seed clone whenever they overlap either end. For this chapter, we’ll confine ourselves to the double-end approach.

Note that if we assume constant length clones, then clones overlap the seed clone in the double-end method whenever they overlap any part of the seed clone. However, if we assume variable length clones, it is possible for a seed clone to cover another clone completely. In this case, end probes from the seed clone would not detect this subsumed clone, and the subsumed clone would remain eligible for subsequent selection as a seed clone. Such a situation would be easy to detect, because if a subsumed clone is selected as a seed clone, both end probes will probe positive against the subsuming seed clone. Consequently, the only deleterious effect of this situation would be to increase the number of steps required to reach any degree of completeness in a mapping project. In the final section of this chapter we describe how serious a problem this might be. For most of this chapter, however, we will ignore the effects of subsumed clones.

Some questions for which we derive approximate answers include

- How many contigs will we have at step $n$?
- How much of the genome will be covered by contigs at step $n$?
- What is the average size of a contig at step $n$?

In the following sections we will outline two approaches to answering these questions.

### 3.2 A “Finite” Analysis for Clones of Constant Length

In this section, we will derive expressions for some measures of progress in a directed mapping effort involving a library of $N$ clones, each clone having a fixed
length $L$. We will approximate the expected number of contigs after step $n$, as well as the proportion of the genome vacant after step $n$. We will use an approach involving an explicit representation of the space between $n$ seed clones.

What is the state of affairs after step $n$? At each step, a seed clone is selected randomly from the library, subject only to the constraint that it does not overlap any of the previously selected seed clones. Let us therefore model the starting points of the set of seed clones after step $n$ as $n$ points chosen randomly from the line segment $[0, G - L]$, subject only to the constraint that each of the $n$ points is separated from its neighbors by a distance of at least $L$. If we order the seed clones by their position on the genome, we obtain a random vector of spacings, or gaps between the seed clones. Figure 3.1 shows the situation for $n = 4$.

![Diagram](image)

Figure 3.1: Spacings resulting from directed mapping.

When the points are chosen randomly in this way, the vector $Y$ of spacings between the ordered seed clones will follow a uniform Dirichlet distribution, but scaled by a constant $a$ representing the area in $[0, G]$ which is not covered by seed clones. That is, let $Y = (Y_0, \ldots, Y_n)$ be the set of spacings between the $n$ seed clones. Then the vector $Y$ has the same distribution as $aS = (aS_0, \ldots, aS_n)$, where $S$ is a "uniform Dirichlet vector" with joint and marginal densities

$$p(s) = n! \mathbf{1}(\forall i, 0 \leq s_i \leq 1, s_+ = 1),$$

$$p(s_i) = n(1 - s_i)^{n-1} \mathbf{1}(0 \leq s_i \leq 1),$$

where $\forall$ is the traditional symbol in logic meaning "for all", $s_+ = \sum_i s_i$, and $\mathbf{1}(\cdot)$ is the indicator function

$$\mathbf{1}(A) = \begin{cases} 
1 & \text{if } A \text{ is true, and} \\
0 & \text{otherwise.}
\end{cases}$$
In this model, we see that the distribution of spacings $Y$ for any step $n$ is tractable. In particular, the marginal distribution of each $Y_i$ is that of a Beta$(1, n)$, but scaled by $a$, the size of the genome not covered by seed clones.

After step $n$, we have used $n$ clones as seed clones, and have $m = N - n$ remaining clones. In our calculations, we will need to approximate the probability of finding none of the $m$ remaining clones in intervals of interest. This probability may be quite difficult to calculate exactly. However, if we make the quite pragmatic assumption that the remaining $m$ clones are approximately uniformly distributed over the region $[0, G]$ and independent of each other, the problem becomes easier. If one draws $m$ points independently from a uniform distribution over $[0, G]$, the probability of seeing no points in a collection of disjoint subintervals of $[0, G]$ of total length $\ell$ is

$$\left(1 - \frac{\ell}{G}\right)^m \approx e^{-\lambda \ell}$$

where $\lambda = m/G$. Conceptually, then, the uniformity assumption is equivalent to approximating the remaining clone process by a Poisson process with rate $\lambda$. Indeed, the distribution of the order statistics for a Poisson process on $[0, G]$, conditional on observing $n$ points, is precisely that of $n$ independent draws from a uniform distribution on $[0, G]$.

To summarize, our basic strategy will be to derive results conditional on getting to step $n$, assume the starting points of the remaining clones form a Poisson process on $[0, G]$ with rate $\lambda$, and describe the random variables of interest as functions of the vector of spacings $Y$.

Our nomenclature so far:

$G$: the length of the genome;

$L$: the length of a clone; and

$N$: the total number of clones.

(For Palazzolo et al's simulation, these numbers were $G = 10^8$, $L = 10^5$, $N = 5000$.)

After completing step $n$: 
$n$: the number of "seed" clones;

$m$: the number of "remaining" clones ($m + n = N$); and

$\lambda$: the rate for the remaining clones ($\lambda = m/G$).

Finally, we need several parameters which measure various aspects of coverage after step $n$:

$a$: the area not covered by seed clones ($a = G - nL$);

$b$: the ratio of the size of the genome covered by seed clones to the size not covered by seed clones ($b = nL/a$);

$c$: the coverage for the remaining clones ($c = \lambda L$);

$p$: the proportion of the genome covered by seed clones ($p = nL/G$); and

$r$: the ratio of the size of a clone to the size of a genome ($r = L/G$).

One small warning to the reader: we will abuse notation somewhat by using the same symbol to refer both to a region of the chromosome as well as a random variable describing that region’s length. For instance, $Y_i$ will denote both the $i$th inter-clone gap and a random variable describing that gap’s length. Context will make the particular situation clear.

**The Number of Contigs**

We can easily count the number of contigs by identifying the last seed clone in each contig: seed clone $n$ always ends a contig, and seed clone $i$ ($1 \leq i < n$) ends a contig whenever one of the following two mutually exclusive events occurs:

**Event** $A_i$ the inter seed clone gap $Y_i$ exceeds $L$, or

**Event** $B_i$ the inter seed clone gap $Y_i$ doesn’t exceed $L$, but no remaining clone bridges the gap.
Hence, the number of contigs is just the random variable

\[ N_c(n) = 1 + \sum_{i=1}^{n-1} 1(A_i \lor B_i) \]

where \( A_i \lor B_i \) corresponds to the event where either \( A_i \) or \( B_i \) occurs. The expectation of this random variable is

\[ E(N_c(n)) = 1 + \sum_{i=1}^{n-1} \Pr\{A_i \lor B_i\} = 1 + (n - 1)(\Pr\{A_1\} + \Pr\{B_1\}) \]

since all the \( Y_i \) have the same distribution.

**Vacancy**

Let us define three numbers: \( r_0, r_1, \) and \( r_2 \). For each \( i, r_i \) is the expected proportion of the genome covered by precisely \( i \) contigs. Thus, \( r_0 \) is the expected proportion of the genome not covered by contigs, \( r_1 \) is the expected proportion of the genome covered by exactly one contig, and \( r_2 \) is the expected proportion of the genome covered by two overlapping contigs that have not been merged into one contig. The situation where two contigs overlap, but have not been merged into one contig, is known as “cryptic overlap”. Note that \( r_0 + r_1 + r_2 = 1 \), since our method of constructing and merging contigs ensures that no point in the genome will be covered by more than two contigs. The number we want for vacancy is just \( r_0 \).

How much of the genome is left uncovered after \( n \) steps? The exact situation is complicated, even in our simplified model:

- \( nL \) of the genome has been covered with seed clones;
- the remaining clones form a marked Poisson process with rate \( \lambda \); and
- only the clones which overlap seed clones contribute to the coverage.

Hence, even with an infinite number of other clones (\( c = \infty \)), the random variable corresponding to the proportion of the genome left vacant (call this random variable \( V \)) is at least \( 1 - 3p \). The diagram in Figure 3.2 shows the situation after step four. The top line represents the genome to be covered. Below it are the four seed clones.
and the set of spacings \(Y\). Below \(Y\) is a series of lines representing a particular subregion of \(Y\): the “coverable region” of the genome, which we denote by \(R\). Even with an infinite number of remaining clones, only the region represented by \(R\) can be covered, because a covering clone must overlap a seed clone. Hence, the coverable region extends at most \(L\) bases to the left of the leftmost seed clone and at most \(L\) bases to the right of the rightmost seed clone. Between each pair of seed clones, the coverable region extends at most \(L\) bases inwards from each member of the pair.

Below the coverable region \(R\) is another region \(T\) represented by a series of double-arrowed lines. This “target” region defines the feasible starting locations for clones which can both overlap a seed clone and cover a region \(R_i\). That is, any remaining clone will help cover the genome only if it starts in a target region. Hence, we can partition the region not covered by seed clones into \(n + 1\) coverable regions \(R = \{R_i \mid i = 0, \ldots, n\}\), and we can associate with each coverable region \(R_i\) a target region \(T_i\), wherein covering clones must begin to overlap \(R_i\) and a seed clone.

Given this setup, how do we compute the vacancy \(V\)? With a fixed set of remaining clones \(\{X_j\}\),

\[
V(n \mid Y, \{X_j\}) = \frac{1}{G} \int_0^G 1(x \text{ uncovered}) \, dx = \frac{1}{G} \sum_i \int_{Y_i} c_i(x) \, dx
\]

where

\[
c_i(x) = \begin{cases} 
1 & \text{if for all } X_j \in T_i, \ x \not\in [X_j, X_j + L], \\
0 & \text{otherwise}
\end{cases}
\]

Taking expectations with respect to the homogeneous Poisson process \(\{X_j\}\) and using Fubini’s theorem results in the following expression for vacancy, as a function of the

Figure 3.2: Computing vacancy.
spacings $Y$:

$$V(n|Y) = \frac{1}{G} \sum_{i=0}^{n} \int_{Y_i} \Pr\{\forall X_j \in T_i, x \notin [X_j, X_j + L]\} \, dx$$

If we now take expectations with respect to $Y$, we get an expression for expected vacancy.

We could go on and obtain similar results for the length of contigs, and other statistics of interest using the same approach. However, there is an easier way, which we will describe in the next section. These results will serve mainly as “exact” results, against which to compare other approximations. To summarize,

**Proposition 3.2.1** Let $S$ be distributed as Beta(1, $n$). Then

(a) The expected number of contigs after step $n$ is given by

$$1 + (n-1) \left[ \Pr\{aS > L\} + \mathbb{E}e^{-\lambda(L-aS)}1(aS \leq L) \right]$$

(b) The expected proportion $r_0$ of the genome left uncovered after step $n$ is

$$r_0 = 2p_0 + (n-1)p_1$$

where

$$p_0 = \frac{1}{G} \left[ \mathbb{E}1(aS > L) \left\{ aS - L + \frac{1}{\lambda} \left( 1 - e^{-c} \right) \right\} + \mathbb{E}1(aS \leq L) \left\{ \frac{1}{\lambda} \left( 1 - e^{-\lambda aS} \right) \right\} \right]$$

and

$$p_1 = \frac{1}{G} \left[ \mathbb{E}1(aS \leq L) aS e^{-c} + \mathbb{E}1(aS > 2L) \left\{ aS - 2L + \frac{2}{\lambda} \left( 1 - e^{-c} \right) \right\} + \mathbb{E}1(aS \in (L, 2L)) \left\{ \frac{2}{\lambda} \left[ e^{-\lambda(2L-aS)} - e^{-c} \right] + (2L - aS)e^{-\lambda(2L-aS)} \right\} \right]$$

and all expectations are taken with respect to $S$.

See Section 3.5.1 for a detailed discussion of the these calculations.

To get closed form approximations for the above expectations, we use a simple approximation to $p(s)$:
Corollary 3.2.1 If we approximate \((1-s)^{n-1}\) by \(e^{-ns}\) in the expression for the density of \(s\) in Eqn. (3.1), the above expressions simplify as follows:

(a) The expected number of contigs after step \(n\) is approximately

\[
n \left[ \frac{c - b}{c - b} e^{-b} - \frac{b}{c - b} e^{-c} \right]
\]

(b) the expressions for \(p_0\) and \(p_1\) become

\[
p_0 = e^{-b} \left[ \frac{1}{n+1} - r \right] + r \left[ \frac{1 - e^{-(b+c)}}{b + c} \right]
\]

and

\[
p_1 = \frac{1}{n+1} \left\{ e^{-c} \left[ (1 - p) - e^{-b} \right] + e^{-2b} \left[ 1 + \frac{2(p + r)}{c} (1 - e^{-c}) - (p + 2r) \right] + e^{-(b+c)} \frac{2(p + r)}{c} \left[ \frac{b}{b - c} (1 - e^{-(b+c)}) - (1 - e^{-c}) \right] + e^{-2b} \frac{b(p + r)}{(b - c)^2} \left[ e^{b-c} (b - c - 1) + 1 \right] \right\}
\]

See Section 3.5.2 for a detailed derivation of the closed form calculations.

3.3 The Stationary Process Approach

Arratia et al [2] address the problem of predicting progress building STS content maps by using techniques involving stationary processes. Our Poisson process of remaining clones is already stationary, with rate \(\lambda\). Can we approximate the spacings by a stationary process and use the same techniques?

If we look more closely at the density of a spacing \(Y_i\), we see

\[
p(y_i) = \frac{n}{a} \left( 1 - \frac{y_i}{a} \right)^{n-1} 1(0 \leq y_i \leq a)
\]

\[
\approx \frac{n}{a} \left( 1 - \frac{y_i}{a} \right)^n 1(0 \leq y_i \leq a)
\]

\[
\approx \frac{n}{a} \exp \left( -\frac{n}{a} y_i \right)
\]
That is, in the case of fixed length clones, \( Y_i \) is approximately distributed as an Exponential random variable with parameter \( \nu \), where \( \nu = n/a \).

Let's model the seed clone process as a stationary point process \( \{U_i \mid i \text{ an integer}\} \), with inter-arrival times \( U_{i+1} - U_i = L_i + Y_i \). Here \( L_i \) is the length of seed clone \( i \), and now the clone lengths are not constant, but are independent, identically distributed draws from a non-negative random variable with distribution \( F(.) \) and finite mean \( L \). As before, \( Y_i \) is the distance between the end of seed clone \( i \) and the beginning of seed clone \( i + 1 \), but now let the \( \{Y_i\} \) be independent, identically distributed draws from an Exponential(\( \nu \)) distribution, where \( \nu = n/a \), and \( a = G - nL \), as before. Finally, let us further denote the counting process associated with the \( \{U_i\} \) by \( N(t) \), where \( t \in \mathbb{R} \). As is traditional, we will normalize \( N(t) \) so that \( N(0) = 0 \).

With this setup, the expected value of each \( U_{i+1} - U_i \) will be \( L + \frac{a}{n} \). The intensity for the resulting seed process \( \{U_i\} \), which we will call \( \mu \), is simply the reciprocal of the mean inter-arrival time:

\[
\mu = \frac{1}{L + \frac{a}{n}} = \frac{n}{nL + a} = \frac{n}{G}
\]

resulting in an expected number of seed clones in \([0, G]\) of \( n \), as required. As a cautionary note, remember that this model will not account for the situation where one seed clone completely covers another, and we are ignoring that situation.

We also need to define two other functions. First, following [2], we need to define the function \( J(\cdot) \), where \( J(x) \) is the probability that an interval of length \( x \) will not be covered by a single remaining clone:

\[
J(x) = \exp \left\{ -\lambda \int_x^\infty (1 - F(t)) \, dt \right\}
\]

Also, we will need to define the joint density \( p(v, w) \) of current and excess life for the seed clone process \( \{U_i\} \), as shown in Figure 3.3.

From, for instance, [39, page 193] we see that

\[
\Pr\{V \geq v, W \geq w\} = \mu \int_{v+w}^\infty (1 - G(t)) \, dt
\]

where \( G(\cdot) \) is the distribution of the inter-arrival times \( \{U_{i+1} - U_i\} \) and \( \mu \) is the reciprocal of the mean inter-arrival time. In our case, \( G \) is just the convolution of \( F \)
with an exponential distribution with parameter $\nu = n/a$:

$$G(x) = \int_0^x F(x - t)\nu e^{-\nu t} \, dt$$

We then obtain $p(v, w)$ by partial differentiation. When all the clones are the same length $L$, this all reduces to a joint density for $(V, W)$ of

$$p(v, w) = \mu \nu e^{-\nu(v+w-L)}1(v + w \geq L)$$

With these definitions in mind, we can follow the basic line of reasoning outlined in [2] to arrive at

**Proposition 3.3.1** Let $F(\cdot)$ denote the distribution of clone length with finite mean $L$, let $p(\cdot, \cdot)$ denote the joint density of current and excess life, and let $J(\cdot)$ be defined as in Eqn. (3.2). Then

(a) The expected number of contigs at step $n$ is approximately $np_1$, where $p_1$ is the probability that a seed clone starts a contig:

$$p_1 = \int_0^\infty J(y)e^{-\nu y} \, dy$$

(b) The proportion of the genome left uncovered at step $n$ is approximately $r_0$:

$$r_0 = \int_0^\infty \left[ \int_0^\infty \int_0^\infty \frac{J(v - l)J(w)}{J(v + w - l)} p(v, w) \, dw \, dv \right] dF(l)$$
(c) The proportion of “cryptic overlap” at step n, \( r_2 \), is approximately \( r_0 - r'_2 \), where \( r'_2 \) is given by

\[
r'_2 = \int_0^\infty \left[ \int_0^\infty (J(v - l) + J(w) - J(v + w - l))p(v, w)\, dv \, dw \right] \, dF(l)
\]

(d) The average size of a contig is approximately

\[
G \frac{1 - r_0 + r_2}{np_1} = G \frac{1 - r'_2}{np_1}
\]

See Section 3.5.3 for a detailed derivation of the stationary process-based approximations.

If we assume all the clones are the same size \( L \), then the above expressions all have closed form solutions.

**Corollary 3.3.1** Assume the clones have constant length \( L \). Then the formulae in Proposition 3.3.1 reduce as follows.

(a) The expected number of contigs at step n is approximately \( np_1 \), where

\[
p_1 = \left[ \frac{c}{c - b} e^{-b} + \frac{b}{c - b} e^{-c} \right]
\]

(b) The proportion of the genome left uncovered at step n is approximately

\[
r_0 \approx (1 - p)e^{-2b} + 2pe^{-2b} \left( \frac{1 - e^{b-c}}{c - b} \right) + e^{-c}(1 - p - e^{-b})
\]

\[
+ e^{-(b+c)}p \frac{b}{c - b} \left( \frac{e^{-b} - 1}{c - b} - 1 \right)
\]

(c) The term \( r'_2 \) is in the expression for “cryptic overlap” becomes

\[
r'_2 = (1 - p)e^{-b} - pe^{-b} \left[ 1 + \frac{b}{c - b} - \left( 1 + \frac{c}{c - b} \right) \left( \frac{1 - e^{b-c}}{c - b} \right) \right]
\]

(d) The average size of a contig remains \( G(1 - r'_2)/(np_1) \), substituting the appropriate closed form expressions for \( p_1 \) and \( r'_2 \).
3.4 Discussion

Other Approximations

Marr et al [51] and Zhang and Marr [71] have recently published expressions for the expected number of contigs and vacancy in the case of constant length clones. See, for example, Equations (6) and (8) in [71]. These expressions were derived using rather complex statistical mechanics concepts and approximations based on “nonuniform hard rod fluids” [70]. Although they were parameterized slightly differently, our Equations (3.4) and (3.5) in Corollary 3.3.1, derived using simple properties of stationary stochastic processes, turn out to be identical to Zhang and Marr’s expressions. They also derived an expression for average contig size, based on the assumption that the amount of “cryptic overlap” is zero. As our methods take cryptic overlap into account, the results, naturally enough, are slightly different. Just how different they are for a regime of interest such as that described by Palazzolo et al ($G/L = 1000, c = 5$) can be seen by examining the graph in Figure 3.4.

This graph shows the theoretical value for $r_2$, as well as observed values for ten simulations. We see that simulations indicate that by step 600, approximately six percent of the genome will be covered by cryptic overlaps, while our theory estimates $r_2$ at a little under five percent. Replacing $r_2$ by zero in Equation (3.3) underestimates average contig size by $100 \times r_2/(r_1 + 2r_2)$ percent. That is,

$$G \left[ \frac{1 - r_0 + r_2}{np_1} \right] \left[ 1 - \frac{r_2}{r_1 + 2r_2} \right] = G \left[ \frac{1 - r_0}{np_1} \right]$$

For the regime described above, this represents about a six percent error.

The expression for $r_0$ in Equation (3.5) is rather complex. However, we can see that for many problems of interest the bulk of the contribution to Equation (3.5) comes from the first two terms. Hence, $r_0$ can be closely approximated in regimes of interest by the simpler

$$r'_0 = e^{-2b} \left[ 1 - p + 2p \left( \frac{1 - e^{b-c}}{c-b} \right) \right]$$

The absolute error resulting from that approximation is shown in the graph in Figure 3.5 for the case of $G/L = 1000$. 
From that graph, we see that the absolute error $r_0 - r'_0$ will not exceed .03 as long as the total coverage $NL/G$ is at least three.

**Comparison with Simulations**

We used Splus [65] to simulate the directed mapping process described in the introduction. The simulations examined the case of constant length clones only, and used parameters corresponding to the situation outlined in Palazzolo et al: $G = 10^8$, $L = 10^5$, and $N = 5000$. The simulation was a continuous one: clones were represented as independent, identically distributed draws from a uniform distribution on $[0, G - L]$.

The following figures compare the above theory with the results of ten randomly chosen simulations for number of contigs, vacancy, and average contig length.

Interestingly, despite the fact that the estimate for number of contigs $N$ shown in
Error Using $r_0'$ Instead of $r_0$

Figure 3.5: Error resulting from using $r_0'$ instead of $r_0$.

Figure 3.6 is bit low, and the estimate for the vacancy $r_0$ shown in Figure 3.7 is a bit high, these two errors cancel in the estimate of average contig length

$$G \frac{1 - r_0 + r_2}{N}$$

shown in Figure 3.8. These results indicate that, for moderate coverage libraries at least, the stationary approach to representing the state of affairs after $n$ steps provides a simple, accurate method for estimating parameters of interest in directed mapping and sequencing projects.

**Clone Subsumption**

Until now, we have been ignoring the possibility of one clone completely covering another. What is the probability that a given seed clone completely contains another clone? Suppose, as before, that clone length can be described by a distribution $F(\cdot)$
Figure 3.6: Simulation vs. theory—contigs.

with finite mean $L$. For the moment, let us also assume that our given seed clone has length $R$. Then, using stationarity, the probability that no remaining clone $X_i$ will be subsumed by our seed clone is

$$\Pr\{\forall X_i \in [0, R], \ X_i + L_i > R\}$$

where $L_i$ is the length of the clone starting at $X_i$. We can compute this probability by thinning all $X_i \in [0, R]$ for which $X_i + L_i > R$ and computing the probability that no remaining clones are left. In this case, the retention probability at $x$ is just $F(R - x)$. Hence, the probability that a seed clone of size $R$ subsumes no other clone is $\exp(-\lambda \int_0^R F(R - x) \, dx)$. Taking averages over $F(\cdot)$, we see that the probability that a seed clone subsumes no other clone is

$$\int_0^\infty \exp \left(-\lambda \int_0^r F(r - x) \, dx\right) \, dF(r)$$
When clones lengths are uniformly distributed over the interval \([a, b]\), this expression reduces to a simple form. In this case, the integral we must evaluate is

\[
\frac{1}{b-a} \int_a^b \exp \left( -\frac{\lambda(x-a)^2}{2(b-a)} \right) \, dx = \sqrt{\frac{2\pi}{\lambda(b-a)}} \left[ \Phi \left( \sqrt{\lambda(b-a)} \right) - .5 \right]
\]

where \(\Phi(\cdot)\) is the cumulative distribution function for a standard normal random variable:

\[
\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp \left( -\frac{t^2}{2} \right) \, dt
\]

If we represent \(a\) and \(b\) as \(L(1-\delta)\) and \(L(1+\delta)\) respectively, so that \(\delta = (b-a)/2L\), Figure 3.9 shows what the above expression evaluates to, as a function of coverage and \(\delta\). We see that even for library coverage less than 5 and \(\delta\) less than 10%, around 20% of the seed clones will subsume some other clone. For larger coverages and more variable clone lengths, the probability of subsumption can become very large indeed.
3.5 Details of Proofs

3.5.1 The Finite Case

Number of Contigs. We need to compute

$$\mathbb{E}N_c(n) = 1 + \sum_{i=1}^{n-1} \Pr\{A_i \vee B_i\} = 1 + (n - 1) (\Pr\{A_1\} + \Pr\{B_1\})$$

First, \(\Pr\{A_1\}\) is just \(\Pr\{Y_1 > L\}\). Second, consider the event \(B_1\), where \(Y_1 \leq L\). A remaining clone can bridge a gap of size \(Y_1\) if and only if it its origin is in a region of size \(L - Y_1\) extending backwards from the right end of the prior seed clone. The probability of no remaining clone starting in that region is just \(\exp(-\lambda(L - Y_1))\). Hence,

$$\Pr\{B_1\} = \mathbb{E}1(Y_1 \leq L)e^{-\lambda(L-Y_1)}$$
and the expected number of contigs is
\[1 + (n - 1) \left[ \Pr\{Y_1 > L\} + \mathbb{E}\mathbf{1}(Y_1 \leq L)e^{-\lambda(L-Y_1)} \right]\]
an easily computable value.

**Vacancy.** We need to compute \( \mathbb{E}V(n \mid Y) \), where
\[
V(n \mid Y) = \frac{1}{G} \sum_{i=0}^{n} V(Y_i) = \frac{1}{G} \sum_{i=0}^{n} \int_{Y_i} \Pr\{\forall X_j \in T_i, x \notin [X_j, X_j + L]\} \ dx
\]
The probability expression in the above integral can be simplified using our approximation:
\[
V(Y_i) = \Pr\{\forall X_j \in T_i, x \notin [X_j, X_j + L]\} = \Pr\{\forall X_j \in T_i, X_j \notin [x - L, x]\}
= \exp(-\lambda \| [x - L, x] \cap T_i \|)
\]
where \( \| \cdot \| \) corresponds to the total length of a disjoint collection of intervals.

Because the distribution of each \( Y_i \) is the same, we only need to calculate expectations for one end spacing \( (Y_0 \text{ or } Y_n) \) and one interior spacing \( (Y_1 \text{ or } \ldots \text{ or } Y_{n-1}) \) and add. With this simplification, the value of the expected vacancy \( r_0 \) is

\[
(3.6) \quad r_0 = \mathbb{E}V(n|Y) = \frac{1}{G} \sum_{i=0}^{n} \mathbb{E}V(Y_i) = \frac{2}{G} \mathbb{E}V(Y_0) + \frac{n-1}{G} \mathbb{E}V(Y_1)
\]

Let’s look at the two terms in the above expression.

First, the end regions \( Y_0 \) and \( Y_n \). Consider two cases. If \( Y_0 > L \), then the target region has length \( L \) and we can translate the origin to the start of the target region to obtain

\[
V(Y_0) = Y_0 - L + \int_{0}^{L} \exp(-\lambda \| [x - L, x] \cap [0, L] \|) \, dx
\]

\[
= Y_0 - L + \frac{1}{\lambda} \left( 1 - e^{-c} \right)
\]

On the other hand, if \( Y_0 \leq L \), then the target region is all of \( Y_0 \) and

\[
V(Y_0) = \int_{0}^{Y_0} \exp(-\lambda \| [x - L, x] \cap [0, Y_0] \|) \, dx
\]

\[
= \frac{1}{\lambda} \left( 1 - e^{-\lambda Y_0} \right)
\]

Putting the two together, we obtain

\[
V(Y_0) = 1(Y_0 > L) \left\{ Y_0 - L + \frac{1}{\lambda} \left( 1 - e^{-c} \right) \right\} + 1(Y_0 \leq L) \left\{ \frac{1}{\lambda} \left( 1 - e^{-\lambda Y_0} \right) \right\}
\]

A symmetrical argument for \( Y_n \) results in values for \( V(Y_n) \) identical in distribution to those for \( V(Y_0) \).

Now, let’s look at what happens in an arbitrary interior region \( Y_i \). Here, we break up the analysis into three cases: \( Y_i \leq L \), \( Y_i \in (L, 2L] \), and \( Y_i > 2L \).

When \( Y_i < L \), the situation simplifies into an ordinary vacancy problem of the type described by Hall [33, page 127]. Hence, for \( Y_i < L \), we have \( V(Y_i) = Y_i e^{-c} \).
On the other hand, when $Y_i > 2L$,

$$V(Y_i) = Y_i - 2L + \int_0^L \exp(-\lambda \|x - L, x\| [0, L]) dx$$

$$+ \int_{Y_i-L}^{Y_i} \exp(-\lambda \|x - L, x\| [Y_i - L, Y_i]) dx$$

$$= Y_i - 2L + 2 \int_0^L e^{-\lambda x} dx$$

$$= Y_i - 2L + \frac{2}{\lambda} \left(1 - e^{-\lambda}ight)$$

Finally, the case where $Y_i \in (L, 2L]$. Here, the interaction between $R_i$ and $T_i$ is a little more complex, because in the region of $R_i$ between $Y_i - L$ and $L$, clones which overlap either neighboring seed clone can contribute to the coverage. The integral must be split into three parts:

$$V(Y_i) = \int_0^{Y_i-L} \exp(-\lambda \|x - L, x\| [-L, 0]) dx$$

$$+ \int_{Y_i-L}^L \exp(-\lambda \|x - L, x\| [-L, 0] \cup [Y_i - L, Y_i]) dx$$

$$+ \int_L^{Y_i} \exp(-\lambda \|x - L, x\| [-L, 0] \cup [Y_i - L, Y_i]) dx$$

$$= \int_0^{Y_i-L} \exp(-\lambda (L - x)) dx$$

$$+ \int_{Y_i-L}^L \exp(-\lambda (2L - Y_i)) dx$$

$$+ \int_L^{Y_i} \exp(-\lambda (x - (Y_i - L))) dx$$

which simplifies in the end to

$$V(Y_i) = \frac{2}{\lambda} \left[e^{-\lambda(2L-Y_i)} - e^{-\lambda} \right] + (2L - Y_i)e^{-\lambda(2L-Y_i)}$$

Hence, the vacancy for an interior region is

$$V(Y_i) = \begin{cases} 1(Y_i \leq L)Y_i e^{-c} + 1(Y_i > 2L) \left\{ Y_i - 2L + \frac{2}{\lambda} \left(1 - e^{-\lambda} \right) \right\} \\
+ 1(Y_i \in (L, 2L]) \left\{ \frac{2}{\lambda} \left[e^{-\lambda(2L-Y_i)} - e^{-\lambda} \right] + (2L - Y_i)e^{-\lambda(2L-Y_i)} \right\}
\end{cases}$$
3.5.2 Closed Form Approximations for the Finite Case

For the expected number of contigs, we make the simple approximation

\[ \mathbb{E}N_c(n) = 1 + (n - 1)(\Pr\{A_1\} + \Pr\{B_1\}) \approx n(\Pr\{A_1\} + \Pr\{B_1\}) \]

Computing the above probabilities is easy.

\[ \Pr\{A_1\} = \Pr\{Y_1 > L\} = \int_{L/a}^{1} p(s) \, ds \approx e^{-b} \]

where we have used the approximation \((1 - s)^{n-1} \approx e^{-ns}\). As for the event \(B_1\), where \(Y_1 \leq L\),

\[ \Pr\{B_1\} = \int_0^{L/a} e^{-\lambda(L-as)}p(s) \, ds \]
\[ \approx \int_0^{L/a} ne^{-\lambda(L-as)}e^{-ns} \, ds \]
\[ = \frac{ne^{-\lambda L}}{\lambda a - n} \left[ \exp \left( \frac{\lambda L - nL}{a} \right) - 1 \right] \]
\[ = \frac{n}{\lambda a - n} \left[ \exp \left( -\frac{nL}{a} \right) - \exp(-c) \right] \]

resulting in

\[ \Pr\{A_1\} + \Pr\{B_1\} \approx \frac{\lambda a}{\lambda a - n} e^{-b} - \frac{n}{\lambda a - n} e^{-c} = \frac{c}{c-b} e^{-b} - \frac{b}{c-b} e^{-c} \]

and

\[ \mathbb{E}N_c(n) \approx n \left[ \frac{c}{c-b} e^{-b} - \frac{b}{c-b} e^{-c} \right] \]

As for the vacancy, taking expectations with respect to \(Y_0\) (and likewise for \(Y_n\)) and simplifying results in an expression for \(p_0\), the expected proportion of the genome left vacant at either end.

\[ p_0 = \frac{1}{G} \mathbb{E}V(Y_0) = e^{-b} \left[ \frac{1}{n+1} - r \right] + r \left[ \frac{1 - e^{-(b+c)}}{b+c} \right] \]

To get an expression for the contribution of interior spacings to overall vacancy, we must take the expectation of Equation (3.7) with respect to \(Y_i\). The end result
is that, after simplifying, we obtain an expression for $p_1$, the expected proportion of the genome left vacant between any two seed clones.

\[
(3.9) \quad p_1 = \frac{1}{G} \mathbb{E} V(Y_i) = \frac{1}{n+1} \left\{ e^{-c} \left[ (1 - p) - e^{-b} \right] \\
+ e^{-2b} \left[ 1 + \frac{2(p + r)}{c} \left( 1 - e^{-c} \right) - (p + 2r) \right] \\
+ e^{-(b+c)} \frac{2(p + r)}{c} \left[ \frac{b}{b-c} \left( 1 - e^{-(b+c)} \right) - (1 - e^{-c}) \right] \\
+ e^{-2b} \frac{b(p + r)}{(b - c)^2} \left[ e^{b-c}(b - c - 1) + 1 \right] \right\}
\]

To obtain $r_0$, the expected proportion vacant, we simply combine Equations (3.6), (3.8) and (3.9) to get

\[
r_0 = 2p_0 + (n - 1)p_1
\]

and we are done.

### 3.5.3 Stationary Process Approach

**The Number of Contigs.** The basic idea here is to consider the stationary process that results from thinning the seed clone process by retaining only those seed clones which start a contig. This resulting process is clearly stationary. Recall that seed clone $i + 1$ begins a new contig whenever no single clone spans the space between it and the previous seed clone, a distance of $Y_i$. For a fixed $Y$, this probability is just $J(Y)$. Hence, the intensity for the contig process is just $\mu \mathbb{E} J(Y)$, where $Y$ is distributed as Exponential($\nu$), and the expected number of contigs in $(0, G)$ is simply $n \mathbb{E} J(Y)$.

**Vacancy.** Recall our definitions of $r_0$, $r_1$, and $r_2$ in Section 3.2. When is a point $t$ not in a contig? Suppose $t \in (U_i, U_{i+1})$. Then $t$ is not in a contig whenever

- $V > L$,
- no remaining clone starting in $(-\infty, t - V + L]$ extends beyond $t$, and
no remaining clone starting in \((t - V + L, t]\] extends beyond \(t + W\),

where the subscripts on \(V\), \(W\), and \(L\) have been omitted for simplicity. The probability of this event for a fixed \((V, W, L)\) can be computed to be

\[
\frac{\int (V - L) - (W) J(W)}{\int (V + W - L) - 1} \mathbf{1}(V > L)
\]

and \(r_0\), the expected proportion of the genome left uncovered by contigs, is simply the expectation of this expression with respect to the joint distribution of \((V, W, L)\).

**Average Contig Length.** Let \(N(t)\) denote the stationary counting process of right-hand-ends of contigs, and let \(A(t)\) denote the stationary process whose value at time \(t\) is 0, 1, or 2, according to whether \(t\) is covered by 0, 1, or 2 contigs. Then we can define the average size of a contig in the interval \((a, b)\) as the ratio

\[
\frac{\int_a^b A(t) \, dt}{\int_a^b dN(t)}
\]

As \(b - a\) gets large, this expression will approach the ratio of the intensities for \(\int A(t) \, dt\) and \(\int dN(t)\):

\[
\frac{r_1 + 2r_2}{\mu \mathbb{E}J(Y)} = \frac{1 - r_0 + 2r_2}{\mu \mathbb{E}J(Y)}
\]

It is this ratio that we use to approximate the average size of a contig. Now, we saw above that

\[
r_0 = \mathbb{E}_{(V, W, L)} \frac{\int (V - L) - (W) J(W)}{\int (V + W - L) - 1} \mathbf{1}(V > L)
\]

What about \(r_2\), the proportion of cryptic overlap? A point \(t \in (U_i, U_{i+1})\) is covered by two different contigs whenever

- \(V > L\),

- at least one remaining clone starts in \((-\infty, t - V + L)\) and ends in \((t, t + W)\),

- at least one remaining clone starts in \((t - V + L, t)\) and extends beyond \(t + W\), and

- no remaining clone starts in \((-\infty, t - V + L)\) and extends beyond \(t + W\),

and
where, as before, we have dropped the subscripts for simplicity. The probability of this event for a fixed \((V,W,L)\) is just

\[
\left[ 1 - \frac{J(V - L)}{J(V + W - L)} \right] J(V + W - L) \left[ 1 - \frac{J(W)}{J(V + W - L)} \right] \mathbf{1}(V > L)
\]

which reduces to

\[
\left[ \frac{J(V - L)J(W)}{J(V + W - L)} - (J(V - L) + J(W) - J(V + W - L)) \right] \mathbf{1}(V > L)
\]

and we see that

\[
r_2 = r_0 - \mathbb{E}(V,W,L) \left[ J(V - L) + J(W) - J(V + W - L) \right] \mathbf{1}(V > L)
\]
Chapter 4

Constructing High-Resolution Physical Maps

4.1 Introduction

Let us now look more closely at one particular project. In this chapter, we describe aspects of LLNL’s current effort to produce a high-resolution physical map of chromosome 19. Figure 4.1 shows a diagram of the basic steps. Further details may be found in [15]. Chromosome 19 is one of the smaller human chromosomes, containing approximately sixty million bases of DNA. The map is based on cosmids, with each insert approximately forty thousand bases long. Hence, around 1500 clones would cover the chromosome, if laid end-to-end. However, since the cloned inserts are randomly cut from the chromosome, many more than 1500 clones are needed to ensure nearly complete coverage of the chromosome. Consequently, the high-resolution map is being constructed from more than 10,000 clones from several different cosmid libraries. Despite the fact that we are using several libraries, we will describe the process of map construction as though the clones were derived from a single library.

LLNL is taking a “bottom-up” approach to building their map. The main steps include
Figure 4.1: Main steps in constructing the map of chromosome 19.

1. creating DNA fingerprint data for each clone by restriction digestion and electrophoresis,

2. computing the posterior odds of overlap between each pair of clones, based on the similarity between the pair of fingerprints for the two clones, and

3. using these posterior odds values to assemble initial contigs.

After constructing initial contigs by this method, other methods must be used to close the gaps between contigs and to associate known markers with contigs.
4.2 Creating Fingerprints

LLNL begin its Chromosome 19 mapping effort by spending many months analyzing concurrent efforts by other laboratories at constructing large-scale physical maps. In addition, Branscomb and coworkers performed extensive computer simulations designed to optimize their fingerprinting strategy [11]. As a result of these simulations and analysis, they chose a bottom-up approach to fingerprinting based on restriction digestion: cutting up the clone with enzymes and measuring the sizes of the resulting fragments. The method differs from normal restriction digestion in one very important respect. In ordinary restriction digests, the fragments from the digest partition the clone. In other words, the sum of the sizes of the fragments equals the size of the clone. In LLNL's method, on the other hand, the fragments eventually visualized by a fingerprint are a subset of fragments produced by a two-step multi-enzyme restriction digest. To see how this happens, let's examine the experimental protocol in a little more detail.

The digestion process proceeds as outlined in Figure 4.2. The first step is a complete double digestion with two six-base restriction enzymes, EcoRI and BglII. The result of this first step is a collection of fragments tagged on both ends with a fluorochrome dye. The DNA from this first digestion is then separated into three aliquots. Each of the three aliquots is further digested completely with a different four-base restriction enzyme: HinfI, HaeIII, or DdeI. As a result of this second digestion, each fragment from step one has been further digested in each aliquot into a number of sub-fragments, only one or two of which contain color linkers, and hence will be visible. The visible fragments are just those between one of the six-base recognition sites cut in step one and one of the four-base recognition sites cut in step two. Of course, if a fragment from step one contains no four-base recognition site between its two six-base sites, the entire fragment will be visible. The three aliquots are then combined into a single solution called a DNA prep for subsequent electrophoresis. Hence, each fragment produced by step one results in at most six visible fragments in the final sample.

DNA fragments are currently visualized by electrophoresis on ABI373 sequencers,
resulting in a list of detected fragment sizes for each clone. Because of limitations in the size standard used and the speed of electrophoresis, only fragments with sizes between 30 and 540 bases will be detected. Each sequencer run currently produces fingerprint data on up to forty-eight separate DNA preps: two different DNA preps can be loaded in each of twenty-four lanes. In addition to the two DNA preps from cloned inserts, each lane also contains a standard prep containing fragments of known sizes (a size standard). Detecting three different preps in a single lane is possible because the ABI373 a four dye-per-lane sequencer, designed to detect up to four separate sequencing reactions simultaneously in each lane, with each reaction using a different
colored fluorochrome dye. Hence, if the three preps were produced with three different dyes, the reactions will be able to be captured concurrently.

Peaks are detected by a combination of locally developed software and software provided by the sequencer manufacturer, Applied Biosystems, Incorporated (ABI). The initial signal extraction software is currently provided by ABI. It produces a vector time series

\[ \{ (X_0(k), X_1(k), X_2(k)) \mid k = 1, \ldots, 6000 \} \]

for each lane, with approximately ten vector samples taken per minute. The two component series \( X_1 \) and \( X_2 \) contain sampled fluorescence intensities for the two DNA preps, while the series \( X_0 \) contains sampled fluorescence intensities for the size standard. The laser-induced fluorescence intensities are measured 25 centimeters downstream from where the DNA prep is loaded into the gel. Thus the time series measures time-to-passage for a fragment, rather than distance traveled in a fixed amount of time, as is the case with radioactivity-based measurement of electrophoresis. Finally, the time series has already been backgrounded and run through a linear filter to extract the three components. The graph in Figure 4.3 shows a typical component time series.

LLNL's current approach to peak detection breaks the analysis into two steps: finding peaks, and estimating fragment length. The present software examines each component series independently to extract the locations of peaks corresponding to DNA fragments. LLNL has tried a number of approaches to extracting peak locations from signals like that shown in Figure 4.3. No method has proved entirely successful [55]. The current (moderately successful) approach assumes the series can be modeled as an AR(1) series contaminated by outliers (here the peaks are the outliers). It relies on robust filtering algorithms provided by Splus [65].

The approach proceeds in three steps. First, an AR(1) model is fit to the series using a generalized M-estimator. Second, a model-based “filter-cleaner” is applied to the series to classify each data point as “normal” or “outlier.” Finally, peak locations are found by locating large local maxima in the portion of the series classified in step two as “outlier.” See [65] for details on the filtering algorithms.
Figure 4.3: Resulting signal for one clone from an electrophoresis run

The process just described produces a list of peak locations \( \{k_1, k_2, \ldots \} \). To be useful for comparison between inserts, these peak locations must be translated into a standard coordinate system. The size standard in each lane is used for this coordinate transformation. LLNL's current size standard consists of a collection of fragments of known size, constructed by digesting pBR322 and SV40 with HaeIII. The standard coordinate system is constructed by associating the peak locations of the size standard with known fragments between 30 and 540 bases in length and performing a monotone spline interpolation \([\text{as}]\). The result is a function which associates a standardized fragment length \( \ell_j \) with each peak location \( k_j \).

Although it is known that fragments of the same length may vary in their elution time by up to 3%, the same fragment will migrate at approximately the same rate under standardized conditions. Indeed, experiments suggest that LLNL has been able to transform the host vector fragments to the standard coordinate system with a
precision of ±1 base out to 400 bases.

4.3 Detecting Overlap

The result of fingerprinting the library is a list of standardized fragment lengths \( \{\ell_j\} \) for each clone in the library. The next step is to compare all pairs of clones for overlap. For a library of \( n \) clones, we compare the fingerprints of all \( \binom{n}{2} \) pairs of clones, resulting in a separate comparison vector \( D \) for each pair of clones. This comparison vector details which integer fragment lengths occurred in each clone. Based on the comparison vector, we then compute the posterior odds of overlap, given the data \( D \), by Bayes rule, as suggested in [52] and [11].

To create a comparison vector \( D \) from a pair of standardized fragment length lists, LLNL currently constructs a bipartite graph describing a best match between pairs of fragment lengths, one from each clone, differing by no more than one base. This graph is then used to decide which fragment lengths have occurred in each fingerprint. As mentioned above, the resulting integer fragment lengths \( \ell \) can range from 30 to 540 bases. We choose a subset of these potential fragment sizes to use as data for detecting overlap. For the purposes of this discussion, let us assume the fragment sizes actually used to detect overlap run from \( N_0 \) to \( N_1 \) bases, where \( 30 \leq N_0 < N_1 \leq 540 \).

The comparison vector \( D \), corresponding to a comparison of clones A and B, consists of a sequence \( (d_{N_0}, d_{N_0+1}, \ldots, d_{N_1}) \) of comparison outcomes, where each comparison outcome is of the form

\[
d_\ell = \begin{cases} 
00 & \text{if a fragment of length } \ell \text{ is observed in neither clone,} \\
10 & \text{if a fragment of length } \ell \text{ is observed in clone } A \text{ only,} \\
01 & \text{if a fragment of length } \ell \text{ is observed in clone } B \text{ only, and} \\
11 & \text{if a fragment of length } \ell \text{ is observed in both clones.}
\end{cases}
\]

Here 1 and 0 represent the presence and absence of a fragment, respectively.

We can approximate the probability of the above comparison outcomes for a given fragment size \( \ell \) by arguing as follows. A fragment of length \( \ell \) occurs whenever a six
base restriction site and an appropriate four base restriction site are $\ell$ bases apart. If the restriction sites involved in the digestion process are distributed across the genome in a roughly uniform manner, then $\nu$, the average number of fragments generated along an $M$ base stretch of the genome, should be roughly proportional to $M$; that is, $\nu \approx \alpha M$ for some $\alpha$. Let us further suppose that the presence or absence on the gel of fragments of different lengths are approximately mutually independent events. Then the probability of seeing no fragments of size $\ell$ in a region of length $M$ is approximately

$$(1 - \pi_\ell)^\nu \approx e^{-\lambda_\ell M}$$

where $\pi_\ell$ is the probability that any single fragment (in the size range) is of length $\ell$, and the fragment intensity $\lambda_\ell$ is defined by $\lambda_\ell = \pi_\ell \alpha$. Of course we do not expect the assumptions leading to this probability to be more than approximately valid, if that, but they do lead to a simple and tractable model, and furthermore, one whose utility can be evaluated.

Consider then the situation diagramed in Figure 4.4, in which clone A has length $L_1$, clone B has length $L_2$, the two clones overlap by an amount $\theta$, and the intensity for fragments of size $\ell$ is $\lambda_\ell$. We immediately compute the probabilities of the above

![Figure 4.4: Typical configuration where two clones overlap.](image-url)
comparison outcomes in terms of $q = e^{-\lambda t}$ as

\begin{align}
  p_{00}(\theta) &= q^{L_1+L_2-\theta} \\
  p_{01}(\theta) &= q^{L_1}(1 - q^{L_2-\theta}) \\
  p_{10}(\theta) &= q^{L_2}(1 - q^{L_1-\theta}) \\
  p_{11}(\theta) &= 1 - q^{L_1} - q^{L_2} + q^{L_1+L_2-\theta}
\end{align}

4.3.1 The Simple Trinomial Model

For the current discussion, let us assume further that all clones are the same length $L$, and all fragment lengths have the same intensity $\lambda$. Without (any more) loss in generality, we may also choose a coordinate system in units of $G/L$, where $G$ is the length of the genome, so that all clones have length one and $\lambda$ becomes a per clone intensity. To simplify discussion further, let us call potential fragment sizes “probes,” mimicking the vocabulary used in hybridization experiments.

Under these conditions, for any two given clones $A$ and $B$, we can reduce the comparison vector $D$ by cross-classifying each of the $n$ probes according to which of the clones they hit, forming a four-fold table of counts:

\begin{align}
  \begin{array}{c|cc}
    & 0 & 1 \\
    \hline
    0 & n_{00} & n_{01} \\
    1 & n_{10} & n_{11} \\
  \end{array}
\end{align}

\begin{align}
  (4.2)
  n_{00} & n_{01} & n_{0+} \\
  n_{10} & n_{11} & n_{1+} \\
  n_{+0} & n_{+1} & n
\end{align}

where 1 means that a clone contains a probe, and 0 means that a clone does not contain a probe. The joint distribution for the cells in this table will depend upon the proportion $\theta$ of overlap between clones $A$ and $B$. Since we assumed the probes were independent, for any $\theta \in [0, 1]$, the data in the four-fold table follow a multinomial distribution with probabilities as shown below:

\begin{align}
  \begin{array}{c|cc}
    & 0 & 1 \\
    \hline
    0 & p_{00}(\theta) & p_{01}(\theta) \\
    1 & p_{10}(\theta) & p_{11}(\theta) \\
    \end{array}
\end{align}

\begin{align}
  (4.1)
  q & 1 - q & 1
\end{align}
In this case, Eqn. (4.1) reduces to

\[
\begin{align*}
p_{00}(\theta) &= q^{2-\theta} \\
p_{01}(\theta) &= q(1 - q^{1-\theta}) \\
p_{10}(\theta) &= q(1 - q^{1-\theta}) \\
p_{11}(\theta) &= 1 - 2q + q^{2-\theta}
\end{align*}
\]

We see that for all \( \theta \), \( p_{01}(\theta) = p_{10}(\theta) \). Hence, the data in the four-fold table can be reduced without loss of information to a trinomial \((n_0, n_1, n_2)\), where \( n_0 = n_{00}, n_1 = n_{01} + n_{10}, \) and \( n_2 = n_{11} \). In this case, the component probabilities reduce to

\[
\begin{align*}
p_0(\theta) &= p_{00}(\theta) = q^{2-\theta} \\
p_1(\theta) &= 2p_{01}(\theta) = 2q(1 - q^{1-\theta}) \\
p_2(\theta) &= p_{11}(\theta) = 1 - 2q + q^{2-\theta}
\end{align*}
\]

Furthermore, unless \( p = q = 1/2 \), no further reduction is possible without some loss of information. Let us abuse notation a bit and call this trinomial vector \( D \) as well.

We need to decide whether any two clones \( A \) and \( B \) overlap based on their trinomial data \( D \), and then use those decisions to build contigs. Our approach to deciding overlap and contig building is to consider the proportion of overlap \( \theta \) to be a random variable and compute the *posterior odds of overlap* \( \Pr\{\theta > 0 \mid D\} / \Pr\{\theta = 0 \mid D\} \) for every pair of clones.

Now, to compute the posterior odds of overlap for any pair of clones, we notice that

\[
\frac{\Pr\{\theta > 0 \mid D\}}{\Pr\{\theta = 0 \mid D\}} = \frac{\Pr\{D \mid \theta > 0\}}{\Pr\{D \mid \theta = 0\}} \times \frac{\Pr\{\theta > 0\}}{\Pr\{\theta = 0\}}
\]

and \( \Pr\{\theta > 0\} / \Pr\{\theta = 0\} \) is the same for all pairs of clones. Therefore, we need only to compute the ratio \( \Pr\{D \mid \theta > 0\} / \Pr\{D \mid \theta = 0\} \) for each pair of clones. Also, since

\[
\Pr\{D \mid \theta > 0\} = \int_{\theta \in [0,1]} \Pr\{D \mid \theta\} \ dP(\theta \mid \theta > 0)
\]

and \( (\theta \mid \theta > 0) \sim \text{Uniform}(0,1) \), the computation reduces to determining the integrated likelihood ratio

\[
L(D) = \frac{\Pr\{D \mid \theta > 0\}}{\Pr\{D \mid \theta = 0\}} = \int_0^1 \frac{\Pr\{D \mid \theta = t\}}{\Pr\{D \mid \theta = 0\}} \ dt
\]
for each pair of clones. Recalling that \( D = (n_0, n_1, n_2) \) is distributed as a trinomial with probabilities given by Eqn. (4.3), for this simple case Eqn. (4.4) becomes

\[
L(D) = \int_0^1 \left[ \frac{q^2 - \theta}{q^2} \right]^{n_0} \left[ \frac{1 - q^1 - \theta}{1 - q} \right]^{n_1} \left[ \frac{1 - 2q + q^2 - \theta}{1 - 2q + q^2} \right]^{n_2} \, d\theta
\]

4.3.2 Towards More Realistic Assumptions

The above description of the simple trinomial model made a number of simplifying assumptions: equal length clones, equal intensities, and perfect fragment size detection. We may need to relax some of these assumptions to produce a usable method for overlap detection. In this chapter, we will confine our attention to examining the effects on \( L(D) \) of relaxing assumptions about equal intensities and perfect detection.

**Different Intensities.** So far we have assumed that each probe is a Poisson process with intensity \( \lambda \), so that the probability of hitting a region of size \( L \) is \( 1 - \exp(-\lambda L) \). However, the graph in Figure 4.5, showing the proportion of fingerprints with positive probes for each fragment size between 30 and 462 bases, indicates that this value is not constant over the range of available fragment sizes, ranging from about 0.05 to about 0.25. This graph was produced by analyzing over 10,000 fingerprints. The conspicuous spikes in the graph are the fragment sizes associated with the cloning vector, which should appear in every fingerprint, and which we ignore.

Given an estimate for the \( \lambda_\ell \) associated with each fragment size \( \ell \), it is easy to incorporate differing intensities in an expression for \( L(D) \). However, in this case the comparison vector \( D \) described on page 63 cannot be summarized as a trinomial. The expression for \( L(D) \) in Eqn. (4.5) becomes a product of ratios over each of the potential fragment sizes:

\[
L(D) = \int_0^1 \prod_{\{\ell | d_\ell = 00\}} \frac{q_\ell^{2-\theta}}{q_\ell^{2}} \prod_{\{\ell | d_\ell = 10 \lor 01\}} \frac{1 - q_\ell^{1-\theta}}{1 - q_\ell} \prod_{\{\ell | d_\ell = 11\}} \frac{1 - 2q_\ell + q_\ell^{2-\theta}}{1 - 2q_\ell + q_\ell^{2}} \, d\theta
\]

where \( q_\ell = e^{-\lambda_\ell} \).

**Fragment Size Detection Error.** In reality, peak detection, and hence fragment size detection, is far from perfect. Consequently, we performed an experiment to
evaluate how often peaks were missed. We wanted to estimate two sets of parameters: \( \pi_{\ell0} \), the probability of mistakenly finding a fragment of length \( \ell \) when none was there, and \( \pi_{\ell1} \), the probability of correctly finding a fragment of length \( \ell \) when one actually is there. For this experiment, the values \( \ell \) ran from 30 bases to 462 bases.

We obtained a convenience sample of forty clones, created for a prior quality analysis experiment. Each clone had been fingerprinted from three to five times. We first used a clustering algorithm to cluster peaks from different samples from the same clone into a collection of putative real peaks. This produced a data set consisting of a 433 \( \times \) 40 matrix of counts, where each row corresponds to a fragment size from 30 to 462 bases, and each column corresponds to a clone. We modeled the distribution
of the count for fragment length $\ell$ for clone $j$, denoted by $Y_{\ell j}$, as

\begin{equation}
Y_{\ell j} \sim \begin{cases}
\text{Binomial}(n_j, \pi_{e0}) & \text{if clone } j \text{ contains no fragment of length } \ell \\
\text{Binomial}(n_j, \pi_{e1}) & \text{if clone } j \text{ contains a fragment of length } \ell
\end{cases}
\end{equation}

where $n_j$ is the number of samples for clone $j$.

One can reformulate the problem as a missing data problem and estimate the \{\pi_{e0}\} and \{\pi_{e1}\} using the EM algorithm [24]. The complete data would consist of a vector $(X, Y) := (X_1, X_2, \ldots, X_{40}, Y_1, Y_2, \ldots, Y_{40})$ for each fragment length $\ell$. The elements of $X$ are independent, identically distributed Bernoulli random variables, where $X_j$ takes on the value 1 whenever clone $j$ actually has a fragment of length $\ell$. The elements of $Y$ are independent of each other, and each $(q_{1X})$ is distributed as in Eqn. (4.7).

With this setup, the complete data are distributed as an exponential family, and we observe only $Y$. Hence, the maximum likelihood estimates for our parameters are quite simple:

$$
\hat{\pi}_1 = \frac{\sum_j Y_j X_j}{\sum_j n_j X_j} \quad \hat{\pi}_0 = \frac{\sum_j Y_j (1 - X_j)}{\sum_j n_j (1 - X_j)}
$$

where we have, as usual, elided the $\ell$. To complete the EM algorithm, we must compute expected values for the sufficient statistic $(\sum_j Y_j X_j, \sum_j n_j X_j)$, conditional on $Y$ and the current values for $\pi_1$ and $\pi_0$:

$$
\mathbb{E} \left[ \sum_j Y_j X_j \mid Y \right] = \Pr\{X = 1 \mid Y\} \sum_j Y_j
$$

$$
\mathbb{E} \left[ \sum_j n_j X_j \mid Y \right] = \Pr\{X = 1 \mid Y\} \sum_j n_j
$$

and $\Pr\{X = 1 \mid Y\}$ can easily be computed by Bayes rule, given a value for the prior probability $\Pr\{X = 1\}$.

We used the histogram in Figure 4.5 to estimate the prior probability of finding a fragment for any length $\ell$ and estimated the probabilities \{\pi_{e0}, \pi_{e1} \mid \ell = 30, \ldots, 462\} as described above. Figure 4.6 shows a lowess-smoothed estimate of the error rates $\pi_{e0}$ and $1 - \pi_{e1}$. From this we clearly see that the false positive rate ($\pi_{e0}$) is relatively constant over $\ell$—a little higher for the shorter fragments, a little lower for the bigger
fragments. However, the false negative rate \((1 - \pi_{\ell_1})\) begins to grow after about 300 bases. In any event, it is clear that fragment size detection is far from perfect.

Perhaps the simplest way to address the problem of errors, other than ignoring it completely, is to construct a model with

- a fixed probability \(\pi_0\) of falsely calling a peak,

- a fixed probability \(1 - \pi_1\) of dropping an existing peak, and

- independent errors among different fragment sizes.

In light of the graph in Figure 4.6, these assumptions seem reasonable if we confine our attention to fragments less than about 300–350 bases long. In this situation, however, the expressions for \(L(D)\) are no longer simple functions of \(q\) as in Eqn. (4.5) and Eqn. (4.6).
The new component probabilities for the comparison vector $D$ depend upon the error probabilities $\pi = (\pi_0, \pi_1)$, as well as $\theta$ and, in the case of differing intensities, the specific intensity $\lambda_\ell$. These component probabilities can be expressed as

\[
\begin{bmatrix}
    p_0(\theta, \pi) \\
    p_1(\theta, \pi) \\
    p_2(\theta, \pi)
\end{bmatrix} =
\begin{bmatrix}
    (1 - \pi_0)^2 & (1 - \pi_0)(1 - \pi_1) & (1 - \pi_1)^2 \\
    2\pi_0(1 - \pi_0) & \pi_0(1 - \pi_1) + \pi_1(1 - \pi_0) & 2\pi_1(1 - \pi_1) \\
    \pi_0^2 & \pi_0\pi_1 & \pi_1^2
\end{bmatrix}
\begin{bmatrix}
    q^{2-\theta} \\
    2q(1 - q^{1-\theta}) \\
    1 - 2q + q^{2-\theta}
\end{bmatrix}
\]

Here the dependence of $q$ on $\ell$ in the case of differing intensities has been suppressed. Note that this reduces to Eqn. (4.3) when $\pi_0 = 1 - \pi_1 = 0$. Incorporating errors produces the following general expression for $L(D \mid \pi)$

\[
L(D \mid \pi) = \int_0^1 \prod_{\{\ell: d_\ell = 00\}} \frac{p_{\ell 0}(\theta, \pi)}{p_{00}(0, \pi)} \prod_{\{\ell: d_\ell = 10\lor 01\}} \frac{p_{\ell 1}(\theta, \pi)}{p_{11}(0, \pi)} \prod_{\{\ell: d_\ell = 11\}} \frac{p_{\ell 2}(\theta, \pi)}{p_{11}(0, \pi)} d\theta
\]

which reduces in the obvious way for the trinomial case.

To estimate $\pi_0$ and $\pi_1$, we could perform a similar analysis on the $433 \times 40$ matrix described above, except that we would confine our attention to fragment sizes under 350 base pairs and insist that $\pi_{\ell k} = \pi_k$ for $\ell = 30, \ldots, 350$, and $k = 0, 1$. However, such an approach would be limited by, among other things, the adequacy of the clustering algorithm to identify matching peaks. Instead, we used restriction maps generated during the mapping process to estimate $\pi_0$ and $\pi_1$, as described below.

Once preliminary contigs have been assembled by methods like those outlined above or in Section 4.4, they can be verified by constructing a complete restriction map of the contig. To build a restriction map of a contig, each clone in the contig is digested by a restriction enzyme such as EcoRI. The resulting digested clone is then electrophoresed on an ABI362 GENESCANNER. As outlined in [42], the ABI362 can resolve fragments from about 350 bases long to over 22,000 bases long with a relative error of less than three percent. This wide range and accurate resolution ensures that, except for extraordinary situations, virtually all of the fragments produced by a six-base enzyme such as EcoRI will be resolved.

The result is a list of fragment sizes for each clone. The clones and their fragments are then laid out and permuted and so that identical fragments from overlapping
clones are aligned vertically, as shown in Figure 4.7. The resulting map provides an estimate of the size of each clone in the contig, as well as the amount of overlap between any two clones in the contig. These maps complement our posterior odds data about pairs of clones, and can help us evaluate our assumptions and estimate unknown parameters. To this end, we extracted data of the form \( \{ (\theta, n_0, n_1, n_2) \} \) for 680 overlapping clones taken from a restriction map of 117 clones encompassing a region of Chromosome 19 known as the D19S11 region, and used this data to compute Maximum Likelihood estimates for \( \pi \).

The mean clone length was 40 kilobases. The \( \theta \) for each overlapping pair was determined by dividing the amount of overlap by the mean clone length of all clonal inserts in the map. To determine the fragment lengths we would use as probes, we chose fragment lengths under 350 bases which had a hit probability between 0.1 and 0.3, resulting in 240 probes with an average hit probability of 0.18. Thus our data...
looked like \( \{(\theta_i, n_{i0}, n_{i1}, n_{i2}) \mid i = 1, \ldots, 680\} \) where
\[
(n_{i0}, n_{i1}, n_{i2}) \sim \text{Trinomial}(240, p_0(\theta_i, \pi), p_1(\theta_i, \pi), p_2(\theta_i, \pi))
\]
and the \( p_j(\theta_i, \pi) \) are defined as in Eqn. (4.8). The resulting maximum likelihood estimates were \( \hat{\pi}_0 \approx .03 \) and \( \hat{\pi}_1 \approx .05 \).

**Evaluating** \( L(D) \). We have described several alternative expressions for the integrated likelihood ratio \( L(D) \) (Equations 4.5, 4.6, and 4.9). We must evaluate \( L(D) \) \( \binom{n}{2} \) times, where \( n \) is the size of the library. In the case of chromosome 19, we have \( n \approx 10,000 \), resulting in approximately \( 5 \times 10^7 \) evaluations of Eqn. (4.9), the vast majority of which will be less than one. To speed up this evaluation process, LLNL screens all \( 5 \times 10^7 \) pairs of clones by computing \( L(D) \) for the simple trinomial model described in Eqn. (4.5). Only those pairs with sufficiently high value for this simple model are further evaluated using the full model (Eqn. (4.9)). Even with this screening procedure, computing a value for \( L(D) \) for all \( 5 \times 10^7 \) pairs takes a weekend’s worth of effort by more than twenty workstations.

### 4.4 Assembling Contigs

LLNL currently assembles clones into an initial set of contigs by a sequential merging algorithm designed by Thomas Slezak of the center’s Informatics group. This algorithm is similar to a hierarchical clustering algorithm known as single-linkage clustering [50, page 370]. Recall that, with merging hierarchical algorithms, one begins with each clone belonging to its own individual cluster. At each step in the algorithm, one merges the two clusters that are the most similar into a new, larger cluster. Differences between algorithms boil down to differences in what it means for two clusters to be similar. With single linkage clustering, the similarity between cluster \( A \) and cluster \( B \) is defined to be the maximum similarity between all pairs \((a, b)\) of clones, where \( a \in A, b \in B \). The eventual amount of clustering is defined by a similarity threshold: clustering stops when no two clusters have a similarity measure that exceeds that threshold.
LLNL's clustering algorithm uses log posterior odds as a similarity measure between clones. As in single-linkage clustering, merging stops when the log posterior odds drops below a user-defined threshold. As the log posterior odds and \( \log L(D) \) differ only by a constant, the algorithm actually uses \( \log L(D) \) as its similarity measure. However, Slezak has extended the basic algorithm described above in a number of ways to take advantage of the biological context in which the clustering is taking place. Most important of these extensions is the construction and maintenance of minimal tiling paths for each cluster. These minimal tiling paths provide a best guess of a minimal overlapping set of clones which span the cluster. When two clusters are selected to be merged, their respective tiling paths are examined. If the tiling paths of the two clusters cannot be consistently merged, then the merge of the two clusters is disallowed. On the other hand, if the two tiling paths are consistent, the two clusters are merged and the tiling path for the combined cluster is computed from the tiling paths of the two merged clusters. Here, “consistency” is defined in a very weak sense, involving a heuristic measure of similarity between tiling path members. When the clustering stops, the algorithm has produced a set of contigs and a set of minimal tiling paths for each contig.

To a first approximation, the algorithm can be viewed as \( \binom{n}{2} \) applications of a Bayesian decision rule, where for each pair of clones, we must decide whether or not \( \theta > 0 \), based on fingerprint data. That is, whenever the loss function is independent of actual amount of overlap, the rule “Decide overlap whenever \( \log L(D) > K \)” for an appropriate \( K \) is a Bayes rule for that loss function. The actual value of \( K \) will depend on the prior and the losses entailed by incorrect decisions.

In the case of contig building, the critical issue is to avoid so-called “false joins.” The consequences of falsely asserting that two contigs of clones should be joined into one contig are much more serious than that of failing to join two contigs that actually overlap. Thus, to a first approximation, we can analyze the decision rule in the classical sense of hypothesis testing: examine the power of the test for a given probability of falsely asserting two clones overlap. In our case, we will be examining the behavior of the decision rule over a range of alternatives \( \theta > \theta_0 \). Consequently, the power we explore will be the power with respect to the marginal distribution of \( D \),
averaged over the range of alternatives. As is traditional, let us denote the probability of falsely deciding two clones overlap by $\alpha$, and likewise denote the probability of falsely deciding that two clones don’t overlap by $\beta$. Note that all the analyses that follow use $L(D)$ as defined by the simple trinomial model without errors (Eqn. (4.5)).

What value of $\alpha$ is reasonable? Given a library of $n$ clones, we expect to have \( \binom{n}{2} \alpha \approx \frac{n^2}{2} \alpha \) false positives. In LLNL’s case, $n \approx 10^4$, so we expect around $5 \times 10^7 \alpha$ false positives. The fact that the expected number of false positives increases as the square of the library size forces us to insist upon values for $\alpha$ very much lower than one typically sees, say on the order of $10^{-6}$. With this harsh reality in mind, let us examine the performance of $L(D)$.

First, we examine the effect of the probability of a positive probe hit $p$ on the power to detect any overlap: $\theta = 0$ vs. $\theta > 0$. Fu et al [30] analyzed hybridization approaches to overlap detection from a hypothesis testing point of view, and showed results indicating that $p$ in the range $(0.4, 0.6)$ was optimal for detecting overlap when the test statistic was simply the total number of agreements $n_{00} + n_{11}$, which we denote by $T(D)$. In Figure 4.8, we compare the power of the integrated likelihood ratio $L(D)$ to that of $T(D)$ for $\alpha = 10^{-2}$, $10^{-3}$, $10^{-4}$, and $10^{-5}$. In all cases, the number of probes was 120.

Several features stand out. First, the integrated Likelihood Ratio decision rule $L(D)$ seems to outperform $T(D)$ for any $\alpha$ and $p$. This is not unexpected, since $L(D)$ is the Bayes rule for any “all-or-nothing” loss function. Also, the probability $p$ that performs the best for $L(D)$ seems to be somewhat lower than that for $T(D)$: between 0.2 and 0.3. Finally, the two approaches have equal power at $p = 0.5$. This is also not unexpected, for in that situation, the trinomial probabilities in Eqn. (4.3) reduce to binomial probabilities, the number of matches $n_{00} + n_{11}$ becomes a sufficient statistic, and the two approaches are identical. To summarize, then, if we use 120 probes and choose $L(D) > K$ as our decision rule, the optimal values for $p$ seem to range between about 0.2 and 0.3.

In the above paragraphs, we examined the power to detect an overlap averaged over all $\theta > 0$. Detecting small overlaps may be too difficult. Now, let us see how the procedure detects overlaps above a certain amount: $\theta = 0$ vs. $\theta > \theta_0$. 
In Figure 4.9, we see a form of ROC curves for detecting overlaps of various sizes ($\theta > 0$, .2, .5, .8) and three probabilities of positive probe hits ($p = .1, .3, .5$). All curves are for 200 probes. Note that these curves differ from traditional ROC curves only in their vertical axis, which shows the probability of a Type II error rather than power.

In addition, we note that these graphs show that the power relationship between $p$ and $\theta$ is not uniform. As our requirements loosen, and the minimum amount of overlap we are required to detect grows larger, the higher probability probes become much more informative.
4.5 Discussion

It is instructive to examine the observed versus expected distribution of different statistics assuming various amounts of overlap. In Figure 4.10, we see scatter plots of all 680 values of \( n_0, n_1, n_2, \) and \( \log L(D | \hat{x}) \), plotted against the estimated amount of overlap. Superimposed upon the scatter plots are the expected 10%, 50%, and 90% percentiles for the statistics under the trinomial model. These percentiles were separately computed for overlaps of 5, 10, \ldots, 40\,\text{kilobases} by simulating random draws of overlapping clones, computing the appropriate trinomial probabilities, and then simulating a trinomial random variable. For each amount of overlap, 500 clone lengths were drawn as independent, identically distributed random values from a Gamma distribution with mean of 40 and variance of 25, subject only to the condition that they exceeded the amount of overlap.

One feature is immediately apparent: the values of all statistics are tremendously
over-dispersed, relative to their expected distributions, even after compensating for unequal clone lengths. In addition, there is a slight bias upwards in $L(D)$. Some of this may be due to non-independence between fragment sizes. We are continuing to investigate the exact source of this unexpected overdispersion. The D19S11 region is unusual in that it contains multiple copies of a repeated DNA sequence. Consequently, more fragments will match than would otherwise be the case. However, this slight bias is still evident when other restriction maps are similarly evaluated.

In addition to attempting to understand the problem with overdispersion, we are planning other future analyses aimed at further optimizing the approach. For instance, it is clear the the completely general calculation in Eqn. (4.9) is quite involved. Is it worth the effort, or will a simple trinomial model do just as well? In addition, we are exploring likelihood calculations for configurations involving more than two clones [56] to see if they can improve the contig-building process.
Chapter 5

An Analysis of the Trinomial Model

5.1 The Basic Setup

5.1.1 A Bayesian Decision Problem

We approach the problem of overlap detection from a Bayesian perspective. Suppose we choose two clones $C_A$ and $C_B$ at random from a library of randomly-generated, equal length clones. This choice induces a random variable $\Theta \in [0, 1]$ which describes the proportion of overlap between $C_A$ and $C_B$, as shown in Figure 5.1. We assume

![Figure 5.1: Two equal-length, overlapping clones. $\Theta$ is defined to be the proportion of shared DNA in the two clones.](image)

that $\Theta = 0$ with a known probability $\pi_0$, and that $\mathcal{L}[\Theta \mid \Theta > 0]$ can be described by a density $w(\theta)$. Furthermore, we assume that we can extend $w(\cdot)$ to $\Theta = 0$ in such
a way that \( w(0) = \lim_{\theta \to 0^+} w(\theta) \) makes sense. For instance, if we model the process of selecting two clones at random from a library as that of choosing two random segments of length \( L \) from an interval of length \( G \), assume that \( G \gg L \), and ignore end effects, we can represent \( L[\Theta | \Theta > 0] \) as that of a random variable which is uniform on the interval \([0, 1]\).

We want to decide whether or not the two clones overlap. Let us therefore define a Bernoulli random variable \( X \equiv 1(\Theta > 0) \) which is 1 precisely when the two clones overlap, and 0 otherwise. Then the prior probability that \( X = 0 \) is \( \pi_0 \), and the prior probability that \( X = 1 \) is \( 1 - \pi_0 \equiv \pi_1 \). We perform an experiment on \( C_A \) and \( C_B \), resulting in a random variable \( Y \) that is (hopefully!) informative about the status of \( X \). The outcome \( y \) from this experiment induces a posterior distribution \( F_{X|Y=y} \) on \( X \). Our task is to use the observed value \( y \) to decide whether or not \( X = 1 \), based on the posterior distribution \( F_{X|Y=y} \) for \( X \). That is, we must construct a function \( d(y) \) producing an estimate \( d \in \{0, 1\} \) for each possible datum \( y \).

In the usual Bayesian decision theory setup, we also posit a loss function \( L(d, X) \), which describes the penalty for incorrect decisions. For this simple situation, any loss function can be described (up to a constant of proportionality) by Table 5.1, where \( k \) is a positive constant. (The case of a maximum a posteriori decision rule (a MAP rule)

\[
\begin{array}{c|cc}
  & X = 0 & X = 1 \\
\hline
  d = 0 & 0 & 1 \\
  d = 1 & k & 0 \\
\end{array}
\]

Table 5.1: The loss function for a simple two-decision problem.

corresponds to setting \( k = 1 \).)

Recall that a Bayes decision rule is one that minimizes the average loss for each \( y \):

\[
d = \begin{cases} 
0 & \text{if } \mathbb{E}L(0, X) < \mathbb{E}L(1, X) \\
1 & \text{if } \mathbb{E}L(0, X) > \mathbb{E}L(1, X) 
\end{cases}
\]
where all expectations are taken with respect to $F_{X|Y=y}$. Since

$$\mathbb{E}L(0, X) = 1 \times \Pr\{X = 1 \mid y\} = \pi_1 \Pr\{y \mid X = 1\} / \Pr\{y\}$$

and

$$\mathbb{E}L(1, X) = k \times \Pr\{X = 0 \mid y\} = k \pi_0 \Pr\{y \mid X = 0\} / \Pr\{y\}$$

the optimal decision rule is the familiar likelihood ratio, choosing $d = 1$ if and only if

$$\Lambda(y) = \frac{\Pr\{y \mid X = 1\}}{\Pr\{y \mid X = 0\}} = \frac{\int_0^1 \Pr\{y \mid \Theta = \theta\} w(\theta) d\theta}{\Pr\{y \mid \Theta = 0\}} > k \frac{\pi_0}{\pi_1}.\quad(5.1)$$

The risk for our rule $d(y)$ when $X = x$ is defined (e.g., by Ferguson [27]) to be the average loss over the random variable $Y$ when $X = x$:

$$r_d(x) = \mathbb{E}_{Y \mid X=x} L(d(Y), x).$$

Finally, its Bayes risk is defined to be the average of the risk $r_d(x)$ over $X$:

$$\mathbb{E}_X r_d(X) = \mathbb{E}_X \mathbb{E}_{Y \mid X} L(d(Y), X)$$

$$= \pi_0 \mathbb{E}_{Y \mid 0} \left[ k \mathbbm{1}(d(Y) = 1) \right] + \pi_1 \mathbb{E}_{Y \mid 1} \left[ \mathbbm{1}(d(Y) = 0) \right]$$

$$= k \pi_0 \Pr\{Y \not\in A \mid X = 0\} + \pi_1 \Pr\{Y \in A \mid X = 1\},$$

where $A$ is the acceptance set for this decision rule:

$$A \equiv \left\{ y : \frac{\Pr\{y \mid X = 1\}}{\Pr\{y \mid X = 0\}} \leq k \frac{\pi_0}{\pi_1} \right\}.$$ 

We are interested in the behavior of the Bayes risk as well as its two components $\Pr\{Y \not\in A \mid X = 0\}$ and $\Pr\{Y \in A \mid X = 1\}$.

Note that in the two clone case, the Bayes decision rule is equivalent to a likelihood ratio test of a simple null hypothesis against a simple alternative. The major complicating factor is that the distribution of $Y$ when $X = 1$ is not that of a sum of independent, identically distributed random variables. Rather, it takes the form of a mixture over $\Theta$ of $\Pr\{Y \mid \Theta\}$. We will denote this mixture distribution on $Y$ by $m(Y) = \int \Pr\{Y \mid \theta\} w(\theta) d\theta$. Despite this complication, we can still discuss the performance of the test either from the Bayesian point-of-view, minimizing the overall Bayes risk, or from the more traditional hypothesis-testing point-of-view, minimizing Type II error while holding Type I error constant.
5.1.2 The Trinomial Model

In our physical mapping context, the data for any given clone consists of a sequence of \( n \) independent Bernoulli outcomes \( \{U_i\} \), where \( \Pr\{U_i = 0\} = r_i \). (Here \( U_i = 1 \) if and only if the clone probes positive for probe \( i \), and \( r_i \in \mathcal{R} = (0, 1) \) is the probability that probe \( i \) misses a randomly chosen clone.) When we compare the above data from two clones, we obtain a sequence of \( n \) independent vector outcomes \( V_1, V_2, \ldots, V_n \), where

\[
V_i = \begin{cases} 
(1, 0, 0) & \text{if probe } i \text{ misses both clones;} \\ 
(0, 1, 0) & \text{if probe } i \text{ hits precisely one clone; and} \\ 
(0, 0, 1) & \text{if probe } i \text{ hits both clones.}
\end{cases}
\]

For the bulk of this discussion, we will make the further simplifying assumption that the miss probabilities are the same: that is, \( r_1 = r_2 = \cdots = r_n \). In this situation, we can define a vector of three probabilities \((p_0, p_1, p_2)\):

\( p_0 \): the probability that a probe misses both clones;

\( p_1 \): the probability that a probe hits precisely one clone; or

\( p_2 \): the probability that a probe hits both clones.

We can then model the outcome of \( n \) probes on a pair of clones as a trinomial random variable. That is, we have a vector outcome \( V = V_1 + \cdots + V_n = (n_0, n_1, n_2) \), representing the number of times out of \( n \) that a probe hits neither, one, or both clones, respectively, and

\[
\Pr\{n_0, n_1, n_2 \mid \theta\} = \binom{n}{n_0, n_1, n_2} p_0(\theta)^{n_0} p_1(\theta)^{n_1} p_2(\theta)^{n_2},
\]

where \( \theta \in \Theta \) is the proportion of clonal overlap.

Under suitable assumptions, this probability triple \((p_0, p_1, p_2)\) is given by

\[
\begin{align*}
p_0 &\equiv p_0(\theta) \equiv p_0(\theta, r) = r^{2-\theta} \\
p_1 &\equiv p_1(\theta) \equiv p_1(\theta, r) = 2r - 2r^{2-\theta} \\
p_2 &\equiv p_2(\theta) \equiv p_2(\theta, r) = 1 - 2r + r^{2-\theta}.
\end{align*}
\]
For convenience, we will often write the probability density of \( V = (n_0, n_1, n_2) \) in terms of \( Y = V/n \), any instance of which we will denote by \( y^n \):

\[
\Pr\{n_0, n_1, n_2 \mid \theta\} = p_n(y_0, y_1, y_2 \mid \theta) \propto e^{n\ell(\theta | y^n)}
\]

where

\[
\ell(\theta \mid y^n) = y_0 \log p_0(\theta) + y_1 \log p_1(\theta) + y_2 \log p_2(\theta).
\]

In terms of this probability model, the likelihood ratio in Eqn. (5.1) becomes

\[
(5.3) \quad \Lambda_n = \Lambda(y^n) = \frac{m_n(y^n)}{p_n(y^n \mid 0)} = \int_{\Theta} e^{n\{\ell(\theta | y^n) - \ell(0 | y^n)\}} w(\theta) d\theta.
\]

**Different parameterizations.** The three probabilities \((p_0, p_1, p_2)\) must be non-negative and add to one. Thus, any particular probability triple corresponds to a point on the simplex

\[
\mathcal{S} = \{(p_0, p_1, p_2) : p_i \geq 0, p_0 + p_1 + p_2 = 1\}.
\]

As diagrammed in Figure 5.2, Eqn. (5.2) defines a parameterized surface in \( \mathcal{S} \) which we will denote by \( \gamma : \Theta \times \mathcal{R} \to \mathcal{S} \). In addition, let \( \mathcal{M} \) denote the image of \( \Theta \times \mathcal{R} \) under the map \( \gamma \).

![Figure 5.2: The probability triples \((p_0, p_1, p_2)\) all lie on a simplex \( \mathcal{S} \). \( \gamma \) defines a parameterized surface on \( \mathcal{S} \), the image of which we denote by \( \mathcal{M} \).](image)

For any given \( r \), the map \( \gamma(\cdot, r) : \Theta \to \mathcal{M} \) defines a one-dimensional curve in \( \mathcal{M} \). Likewise, for any fixed \( \theta \), the map \( \gamma(\theta, \cdot) : \mathcal{R} \to \mathcal{M} \) defines another curve in \( \mathcal{M} \).
Figure 5.3: The coordinate curves induced by the parameterization in Eqn. (5.2). The straight lines with arrows correspond to \( \gamma(\cdot, r) \) curves for various fixed \( r \), the arrow indicating the direction of motion as \( \theta \) runs from 0 to 1. The curved lines correspond to \( \gamma(\theta, \cdot) \) curves for various \( \theta \).

Figure 5.3 shows these curves on the simplex \( S \) for various \( r \) and \( \theta \). We see that \( \mathcal{M} \) is strictly contained in \( S \). Moreover, the curve corresponding to \( \theta = 0 \) is in the interior of \( S \). Consequently, although the point \( \theta = 0 \) may be on the boundary of \( \Theta \), it is not on the boundary of the natural parameter space for the exponential family defined by \( e^{n(t \theta) y^n} \). Hence, we may expect the large-sample behavior of \( \Lambda_n \) to be reasonably well-behaved.

In addition, we note the coordinate lines for \( \Theta \) are not symmetric around \( r = 0.5 \). Hence, with a uniform distribution on \( \Theta \) for any \( r \), we should expect asymmetries in the behavior of those properties such as Bayes risk that depend on the joint values of \( r \) and \( \theta \).

Finally, the coordinate lines for \( \gamma(\cdot, r) \) in Figure 5.3 suggest another, more useful parameterization. If we look at Equation (5.2), we see that \( p_1(\theta) = 2r - 2p_0(\theta) \). Hence, the paths traced by \( \gamma(\cdot, r) : \Theta \rightarrow \mathcal{M} \) are straight lines running from \((r^2, 2r(1-\)
\( r \), \((1 - r)^2\) when \( \theta = 0 \) down to \((r, 0, 1 - r)\) when \( \theta = 1 \). Thus, for any \( r \), we can reparameterize our curved exponential family directly in terms of the line \( \lambda \mapsto \lambda(r^2, 2r(1 - r), (1 - r)^2) + (1 - \lambda)(r, 0, 1 - r) \), where \( \lambda \in [0, 1] \). Note that in this parameterization, no overlap corresponds to \( \lambda = 1 \). If we let \( s \) denote \( r(1 - r) \), then our probabilities, in terms of \( \lambda \), become

\[
\begin{align*}
p_0(\lambda) &\equiv p_0(\lambda, r) = r - \lambda s \\
p_1(\lambda) &\equiv p_1(\lambda, r) = 2\lambda s \\
p_2(\lambda) &\equiv p_2(\lambda, r) = 1 - r - \lambda s.
\end{align*}
\]

This parameterization, which brings the parameter out of the exponent, will simplify matters greatly in the following sections.

Finally, note that in parameterization defined by Eqn. (5.4), we have \( p_0(\lambda, r) = p_3(\lambda, 1 - r) \). Hence, the \( \lambda \) coordinate lines will be symmetric around \( r = 0.5 \), and the asymmetry seen in Figure 5.3 disappears. In fact, this asymmetry has been transferred to the prior distribution on \( \lambda \). Recall that the prior distribution for \((\theta \mid \theta > 0)\) was that of a random variable uniformly distributed over \([0, 1]\). By change of variables, the prior for \((\lambda \mid \lambda < 1)\) becomes

\[
w(\lambda) = \left[ \frac{1}{-\log r} \right] \left[ \frac{s}{r - \lambda s} \right] 1(\lambda \in [0, 1])
\]

### 5.2 Asymptotics of the Likelihood Ratio Statistic

We now examine the asymptotic moments and distribution of our Bayesian detector \( \Lambda_n \), defined in Eqn. (5.3). We will build on proofs described by Clarke and Barron [17]. In this section, we will let \( \Theta = [\theta_0, \theta_1] \), with \( \theta_0 = 0 \) corresponding to the configuration with no overlap.

#### 5.2.1 Asymptotics Under No Overlap

Clarke and Barron provide Bayesian asymptotics for the expectation of \( \log \Lambda_n \) as well as its limiting distribution in the case of a general one-parameter family
\{p_n(y^n | \theta) : \theta \in \Theta\}$, where $\Theta = [\theta_0 - \delta_0, \theta_0 + \delta_1]$, and $w(\cdot)$ is continuous and positive at $\theta_0$. Our case is somewhat different, in that $\theta_0$ is on the boundary of $\Theta$. However, since the point on $S$ corresponding to no overlap is in the interior of the natural parameter space of our exponential family, we can modify their proofs, using one-sided Taylor expansions around $\theta_0$ to get a slightly different asymptotic result. If we do so, we get the following:

**Theorem 5.2.1** Let $(n_0, n_1, n_2)$ be a random vector describing the number of probes hitting neither, one, or both clones respectively, and let $y^n = (y_0, y_1, y_2)$, where $y_i = n_i/n$ and $n = n_0 + n_1 + n_2$. Denote the family of conditional densities of $(y^n | \theta)$ by \{p_n(y^n | \theta) : \theta \in \Theta\}, where $\Theta = [\theta_0, \theta_1]$ is a closed interval on the real line, and $\theta_0$ corresponds to the no-overlap situation. Let $w(\theta)$ denote a prior density on $\Theta$ that is continuous from the right and positive at $\theta_0$. Let

$$
\Lambda_n = \frac{\int_{\theta_0}^{\theta_1} p_n(y^n | \theta)w(\theta) \, d\theta}{p_n(y^n | \theta_0)}, \quad \text{and}
$$

$$
D_n = \frac{1}{2}S_n^2 + \log \Phi(S_n) - \frac{1}{2} \log \left( \frac{nI(\theta_0)}{2\pi w(\theta_0)^2} \right).
$$

where $I(\theta)$ is the Fisher information in one observation under $\theta$, $\Phi(z) = \Pr\{Z \leq z\}$ when $Z$ is a standard normal random variable, and

$$
S_n = \frac{1}{\sqrt{nI(\theta_0)}} \frac{d}{d\theta} \log p_n(y^n | \theta),
$$

evaluated at $\theta_0$. Then, under $\theta_0$,

$$
\lim_{n \to \infty} \mathbb{E}[\log \Lambda_n - D_n] = 0.
$$

**Corollary 5.2.1** Under the conditions of Theorem 5.2.1,

$$
\mathbb{E}[\log \Lambda_n | \theta_0] = -\frac{1}{2} - \frac{1}{2} \log \left( \frac{nI(\theta_0)}{2\pi w(\theta_0)^2} \right) + o(1),
$$

and

$$
\mathcal{L} \left[ \log \Lambda_n + \frac{1}{2} \log \left( \frac{nI(\theta_0)}{2\pi w(\theta_0)^2} \right) \right] | \theta_0 \to \mathcal{L} \left[ \frac{1}{2}Z^2 + \log \Phi(Z) \right].
$$
Note that the corollary can be rewritten to state the Kullback-Leibler distance from \( p_n(\cdot | \theta_0) \) to \( m_n(\cdot) \) is given by

\[
D(p_n(\cdot | \theta_0) \| m_n(\cdot)) = \log \frac{ne I(\theta_0)}{2\pi w(\theta_0)^2} + o(1).
\]

**Proof.** The proof is a slight adaptation of Clarke and Barron’s proof of their Theorem 2.1 to cover asymptotics at the boundary. Recall from Eqn. (5.3) that we can rewrite \( \Lambda_n \) as

\[
\Lambda_n = \int_{\theta_0}^{\theta_1} e^{n\ell(\theta) - \ell(\theta_0)} w(\theta) \, d\theta
\]

since \( p_n(y^n | \theta) \propto e^{\ell(y^n | \theta)} \), where

\[
\ell(\theta | y^n) = y_0 \log p_0(\theta) + y_1 \log p_1(\theta) + y_2 \log p_2(\theta),
\]

and the \( p_i(\theta) \) are defined as in Eqn. (5.2). We will henceforth drop the \( y^n \) from \( \ell(\cdot) \) and its derivatives with respect to \( \theta \), the first two of which we denote by \( \ell'(\theta) \) and \( \ell''(\theta) \). Let \( J(\theta) \equiv -\ell''(\theta) \) denote the mean observed information. Note that, in this notation, \( S_n = \sqrt{n/I(\theta_0) \ell'(\theta_0)} \).

The first step is to bound \( \Lambda_n \). Fix an \( \epsilon > 0 \) and \( \delta > 0 \), and let \( N_\delta = [\theta_0, \theta_\delta] \), where \( \theta_\delta = \theta_0 + \delta/\sqrt{I(\theta_0)} \). Let

\[
A_n = \left\{ y^n : \int_{\theta_\delta}^{\theta_1} e^{n\ell(\theta) - \ell(\theta_0)} w(\theta) \, d\theta \leq \epsilon \int_{\theta_0}^{\theta_\delta} e^{n\ell(\theta) - \ell(\theta_0)} w(\theta) \, d\theta \right\}
\]

\[
B_n = \left\{ y^n : 1 - \epsilon \leq \frac{J(\theta)}{I(\theta_0)} \leq 1 + \epsilon, \ \forall \theta \in N_\delta \right\}
\]

\[
C_n = \left\{ y^n : \frac{\ell'(\theta_0)}{\sqrt{I(\theta_0)}} \leq \delta \right\}
\]

The following lemma bounds \( \Lambda_n \) on \( A_n \cap B_n \) and \( B_n \cap C_n \):

**Lemma 5.2.1** On the set \( A_n \cap B_n \), we have

\[
\Lambda_n \leq (1 + \epsilon) w(\theta^*) \exp \left\{ \frac{S_n^2}{2(1 - \epsilon)} \right\} \sqrt{\frac{2\pi}{n(1 - \epsilon) I(\theta_0)}} Q(-S_n) \left( 1 + c_1 \left( \frac{1}{\sqrt{1 - \epsilon}} - 1 \right) \right)
\]

where \( w(\theta^*) = \sup_{\theta \in N_\delta} w(\theta) \), \( S_n = \sqrt{n/I(\theta_0) \ell'(\theta_0)} \), \( Q(x) = 1 - \Phi(x) \) is the probability that a standard normal random variable exceeds \( x \), and \( c_1 \) is a positive constant independent of \( y^n, \delta, \) and \( \epsilon \).
On the set $B_n \cap C_n$, we have

$$
\Lambda_n \geq w(\theta^*) \exp \left\{ \frac{S_n^2}{2(1 + \epsilon)} \right\} \sqrt{\frac{2\pi}{n(1 + \epsilon)I(\theta_0)}} Q\left(-S_n, \left(1 - c_2 \left(1 - \frac{1}{\sqrt{1 + \epsilon}}\right)\right) \right)
\times \left(1 - c_3 \exp \left\{ -\frac{n\delta^2 \epsilon^2}{4(1 + \epsilon)} \right\} \right)
$$

where $w(\theta^*) = \inf_{\theta \in N_\delta} w(\theta)$ and $c_2$ and $c_3$ are positive constants independent of $y^n$, $\delta$, and $\epsilon$.

**Proof of upper bound.** First, restrict our attention to $A_n \cap B_n$ and expand $\ell(\theta)$ around $\theta_0$: for some $\tilde{\theta} \in N_\delta$,

$$
\Lambda_n \leq (1 + \epsilon) \int_{N_\delta} e^{n\{\ell'(\theta_0)(\theta - \theta_0) - \frac{1}{2} I(\theta_0)(\theta - \theta_0)^2\}} w(\theta) \, d\theta
\leq (1 + \epsilon) w(\theta^*) \int_{N_\delta} e^{n\{\ell'(\theta_0)(\theta - \theta_0) - \frac{1}{2} (1 - \epsilon) I(\theta_0)(\theta - \theta_0)^2\}} \, d\theta
$$

Now, complete the square:

$$
\ell'(\theta_0)(\theta - \theta_0) - \frac{1}{2} (1 - \epsilon) I(\theta_0)(\theta - \theta_0)^2 = \frac{1}{2} (1 - \epsilon) I(\theta_0)(\theta - u)^2 + \frac{1}{2} \left[\ell'(\theta_0)\right]^2
$$

where $u = \theta_0 + \frac{\ell'(\theta_0)}{(1 - \epsilon) I(\theta_0)}$ depends on $y^n$. Substituting back and expanding the region of integration to $[\theta_0, \infty)$ gives

$$
\Lambda_n \leq (1 + \epsilon) w(\theta^*) \exp \left\{ \frac{S_n^2}{2(1 + \epsilon)} \right\} \int_{\theta_0}^{\infty} e^{-\frac{1}{2} n(1 - \epsilon) I(\theta_0)(\theta - u)^2} \, d\theta
= (1 + \epsilon) w(\theta^*) \exp \left\{ \frac{S_n^2}{2(1 + \epsilon)} \right\} \sqrt{\frac{2\pi}{n(1 - \epsilon) I(\theta_0)}} Q\left(-S_n/\sqrt{1 - \epsilon}\right)
$$

To complete the process, we must bound $Q\left(-S_n/\sqrt{1 - \epsilon}\right)$ in terms of $Q(-S_n)$. This can easily be done, using the well-known upper bound $Q(x) \leq \phi(x)/x$ whenever $x > 0$, where $\phi(\cdot)$ is the density for a standard normal. Indeed, if $a > 1$, we can find a constant $k$ so that for all $x$, $Q(ax) \leq Q(x)(1 + k(a - 1))$. 

Proof of the lower bound. We take the same approach, making the necessary changes. Restricting our attention to $B_n$ for the moment, for some $\tilde{\theta} \in \mathbb{N}_\delta$,

$$
\Lambda_n \geq w(\theta_\ast) \int_{N_\delta} e^n \left\{ \ell'(\theta_\ast) (\theta - \theta_0) - \frac{1}{2} J'(\theta - \theta_0)^2 \right\} d\theta \\
\geq w(\theta_\ast) \int_{N_\delta} e^n \left\{ \ell'(\theta_\ast) (\theta - \theta_0) - \frac{1}{2} (1 + \epsilon) I(\theta_\ast)(\theta - \theta_0)^2 \right\} d\theta \\
= w(\theta_\ast) \exp \left\{ \frac{S_n^2}{2(1 + \epsilon)} \right\} \int_{N_\delta} e^{-\frac{1}{2} n(1 + \epsilon) I(\theta_\ast)(\theta - v)^2} d\theta
$$

(5.7)

where this time we complete the square to obtain

$$
\ell'(\theta_\ast)(\theta - \theta_0) - \frac{1}{2} (1 + \epsilon) I(\theta_\ast)(\theta - \theta_0)^2 = -\frac{1}{2} (1 + \epsilon) I(\theta_\ast)(\theta - v)^2 + \frac{1}{2} \frac{[\ell'(\theta_\ast)]^2}{(1 + \epsilon) I(\theta_\ast)}
$$

and $v = \theta_\ast + \frac{\ell'(\theta_\ast)}{(1 + \epsilon) I(\theta_\ast)}$ depends on $y^n$.

We now seek an upper bound for the final integral in Eqn. (5.7). For $\theta > \theta_\delta$, we can see that

$$
\sqrt{I(\theta_\ast)} |\theta - v| = \sqrt{I(\theta_\ast)} \left| \theta - \theta_\ast - \frac{1}{1 + \epsilon} \frac{\ell'(\theta_\ast)}{I(\theta_\ast)} \right| \\
\geq \delta - \frac{1}{1 + \epsilon} \frac{\ell'(\theta_\ast)}{\sqrt{I(\theta_\ast)}} \\
\geq \delta \left( 1 - \frac{1}{1 + \epsilon} \right) = \frac{\delta \epsilon}{1 + \epsilon}
$$

as long as we further restrict our attention to $C_n$. Hence, on $B_n \cap C_n$, $\theta > \theta_\delta$ implies

$$
\frac{n}{2} (1 + \epsilon) I(\theta_\ast)(\theta - v)^2 \geq \frac{n}{4} \delta^2 \frac{\epsilon^2}{(1 + \epsilon)} + \frac{n}{4} (1 + \epsilon) I(\theta_\ast)(\theta - v)^2
$$

and, expanding the region of integration, we have

$$
\int_{\theta_\delta}^{\infty} e^{-\frac{1}{2} n(1 + \epsilon) I(\theta_\ast)(\theta - v)^2} d\theta \leq e^{-\frac{n}{4} \frac{\delta^2 \epsilon^2}{(1 + \epsilon)}} \int_{\theta_\ast}^{\infty} e^{-\frac{1}{2} \left( \frac{n}{2} (1 + \epsilon) I(\theta_\ast)(\theta - v)^2 \right)} d\theta
$$

Substituting back into Eqn. (5.7), we have

$$
\Lambda_n \geq w(\theta_\ast) \exp \left\{ \frac{S_n^2}{2(1 + \epsilon)} \right\} \sqrt{\frac{2\pi}{n(1 + \epsilon) I(\theta_\ast)}} Q \left( -\frac{S_n}{\sqrt{1 + \epsilon}} \right) \\
\times \left( 1 - \exp \left\{ -\frac{n\delta^2 \epsilon^2}{4(1 + \epsilon)} \right\} \frac{Q \left( \sqrt{2} \frac{\sqrt{n}}{\sqrt{1 + \epsilon}} \right)}{Q \left( \frac{S_n}{\sqrt{1 + \epsilon}} \right)} \right)
$$
The $Q(\sqrt{2} \cdot)/Q(\cdot)$ is bounded, and one can show that if $0 < a < 1$, there is a constant $k$ so that for all $x$, $Q(ax) \geq Q(x)(1 - k(1 - a))$.

**Proof of the theorem.** Let

$$R_n = \log \Lambda_n - \left[ \frac{1}{2} S_n^2 + \log \Phi(S_n) - \frac{1}{2} \log \left( \frac{nI(\theta_0)}{2\pi w(\theta_0)^2} \right) \right].$$

We must show that $E|R_n| = E(R_n)^+ + E(R_n)^- \to 0$ under $\theta_0$, where $(x)^+ = \max(0, x)$ and $(x)^- = \max(0, -x)$. Observing that

$$\frac{S_n^2}{2(1 + \epsilon)} = S_n^2 - S_n^2 \left[ \frac{\epsilon}{1 + \epsilon} \right] \quad \text{and} \quad \frac{S_n^2}{2(1 - \epsilon)} = S_n^2 + S_n^2 \left[ \frac{\epsilon}{1 - \epsilon} \right],$$

we have $R_n \leq U_n$ on $A_n \cap B_n$, where

$$U_n = \log(1 + \epsilon) + \log \frac{w(\theta_*)}{w(\theta_0)} + \frac{S_n^2}{2} \left[ \frac{\epsilon}{1 - \epsilon} \right] + \frac{1}{2} \log \frac{1}{1 - \epsilon} + \log \left[ 1 + c_1 \left( \frac{1}{\sqrt{1 - \epsilon}} - 1 \right) \right].$$

Likewise, $R_n \geq -L_n$ on $B_n \cap C_n$, where

$$-L_n = \log \frac{w(\theta_*)}{w(\theta_0)} - \frac{S_n^2}{2} \left[ \frac{\epsilon}{1 + \epsilon} \right] + \frac{1}{2} \log \frac{1}{1 + \epsilon} + \log \left[ 1 - c_2 \left( 1 - \frac{1}{\sqrt{1 + \epsilon}} \right) \right]$$

$$+ \log \left[ 1 - c_3 \exp \left\{ -\frac{n\delta^2 \epsilon^2}{4(1 + \epsilon)} \right\} \right].$$

Note that $U_n, L_n \geq 0$. Let us denote the events $(A_n \cap B_n)^c$ by $A_n$ and $(B_n \cap C_n)^c$ by $B_n$. Then, by the non-negativity of $U_n, L_n, S_n^2$, and $-\log \Phi(S_n)$, we have for $n$ sufficiently large,

$$(R_n)^+ \leq U_n + (\log \Lambda_n)^+ 1(A) + \left( \frac{1}{2} \log \frac{nI(\theta_0)}{2\pi w(\theta_0)^2} - \log \Phi(S_n) \right) 1(A) \quad \text{and}$$

$$(R_n)^- \leq L_n + (\log \Lambda_n)^- 1(B) + \left( \frac{1}{2} S_n^2 \right) 1(B).$$

The rest of the proof consists of examining the expectation of each of the terms in the above display as $n \to \infty$. Clarke and Barron show that $\Pr\{A \mid \theta_0\} \to 0$ and $\Pr\{B \mid \theta_0\} \to 0$ quickly enough to annihilate all of the terms in which they appear, save the one involving $\log \Phi(S_n)$. However, the uniform integrability of $\{\log \Phi(S_n)\}$ allows us to conclude that $E\{\log \Phi(S_n) 1(A)\} \to 0$ as well.
Hence, for all $\epsilon$ and $\delta$, we have $\limsup (R_n)^+ \leq \limsup U_n$ and $\limsup (R_n)^- \leq \limsup L_n$. An examination of $U_n$ and $L_n$ shows that
\[
\lim_{\epsilon, \delta \to 0} \limsup_n U_n = \lim_{\epsilon, \delta \to 0} \limsup_n L_n = 0
\]
and we're done.

### 5.2.2 Asymptotics Under Positive Overlap

In this section, we look at the behavior of $\mathcal{L} [\log \Lambda_n | \theta']$, when $\theta'$ is in the interior of $[\theta_0, \theta_1]$. The situation turns out to be rather simple.

**Theorem 5.2.2** Define $\Lambda_n$ and $[\theta_0, \theta_1]$ as in Theorem 5.2.1, and let
\[
p(\theta) = (p_0(\theta), p_1(\theta), p_2(\theta)) \quad \text{and} \quad \Sigma(\theta) = \text{Diag}(p(\theta)) - p(\theta)p(\theta)^t
\]
denote the parameter vector and covariance matrix under the trinomial model for $\Theta = \theta$. Let
\[
\delta(\theta) = \left( \log \frac{p_0(\theta)}{p_0(\theta_0)}, \log \frac{p_1(\theta)}{p_1(\theta_0)}, \log \frac{p_2(\theta)}{p_2(\theta_0)} \right),
\]
and let $D(\theta||\theta_0) = p(\theta) \cdot \delta(\theta)$ denote the Kullback-Leibler distance from $p(\theta)$ to $p(\theta_0)$.

Then under $\theta'$ in the interior of $[\theta_0, \theta_1]$, \[
n^{-1/2} \left[ \log \Lambda_n - \left( nD(\theta'||\theta_0) - \frac{1}{2} \log n \right) \right] \xrightarrow{d} \mathcal{N}(0, \delta^t \Sigma \delta)
\]

**Proof.** Note that
\[
\log \Lambda_n = \log \int_{\theta_0}^{\theta'} p_n(y^n | \theta) w(\theta) d\theta = \log \left[ \int_{\theta_0}^{\theta'} p_n(y^n | \theta) w(\theta) d\theta \right] + \log \left[ \frac{p_n(y^n | \theta')}{p_n(y^n | \theta_0)} \right].
\]
Under $\theta'$, the first term in brackets, which we will denote by $\Lambda_n'$, satisfies the requirements of Clarke and Barron's Theorem 2.1, while the second term in brackets, which we will denote by $\lambda_n'$, is a simple likelihood ratio. Hence, $\log \lambda_n'$ dominates $\log \Lambda_n'$, because after centering, $\log \Lambda_n'$ converges to a non-degenerate distribution, while $\log \lambda_n'$ must be scaled by $n^{-1/2}$ to so converge.

Specifically, in our model we have
\[
\log \lambda_n' = n (\ell(\theta' | y) - \ell(\theta_0 | y)) = ny \cdot \delta(\theta').
\]
Eliding the dependence of \( p(\theta') \) and \( \Sigma(\theta') \) on \( \theta' \) for a moment, we observe that \( \sqrt{n}(y-p) \xrightarrow{d} \mathcal{N}(0,\Sigma) \) implies that
\[
n^{-1/2} \left[ \log \Lambda'_n - np \cdot \delta \right] \xrightarrow{d} \mathcal{N}(0, \delta'\Sigma \delta).\]
By Slutsky’s theorem, then, under \( \theta' \),
\[
n^{-1/2} \left[ \log \Lambda'_n + \log \lambda'_n - \left( np \cdot \delta - \frac{1}{2} \log n \right) \right] \xrightarrow{d} \mathcal{N}(0, \delta'\Sigma \delta)
\]
as well.

To summarize, then, for large \( n \), \( \mathcal{L}[\log \Lambda_n | \theta] \) depends upon whether \( \theta = 0 \) or \( \theta > 0 \). When \( \theta = 0 \), \( \log \Lambda_n - \mathbb{E} \log \Lambda_n \) has a non-degenerate distribution described by Corollary 5.2.1. On the other hand, when \( \theta > 0 \), we see that \( n^{-1/2} \log \Lambda_n \) is asymptotically normal, with expectation linear in \( D(\theta\|0) \) and an asymptotic variance equal to the value of the quadratic form \( \delta(\theta)\Sigma(\theta)\delta(\theta) \).

### 5.3 Large-Sample Power of the Likelihood Ratio Test

Theorem 5.2.1 and its corollary provide some insight into the asymptotic power of the likelihood ratio test in Eqn. (5.1), which asserts overlap whenever
\[
\log \Lambda_n(y^n) = \log \frac{m(y^n)}{p_n(y^n | 0)} > k \frac{\pi_0}{\pi_1}.
\]
If we let \( D_n \) denote the Kullback-Leibler distance from \( p_n(\cdot | 0) \) to \( m_n(\cdot) \), then \( \log \Lambda_n + D_n \) has a non-degenerate, continuous asymptotic distribution under \( \Theta = 0 \). Hence, we can define a threshold \( t_\alpha \) so that
\[
\lim_{n \to \infty} \Pr\{ \log \Lambda_n + D_n \leq t_\alpha | \Theta = 0 \} = 1 - \alpha.
\]
Let \( A_n \) denote the acceptance set \( \{ y^n : \log \Lambda_n(y^n) + D_n \leq t_\alpha \} \), and let \( \Pr\{ A_n | 0 \} = 1 - \alpha_n \). Then \( \lim_n \alpha_n = \alpha \), and the Type II error \( \beta_n \) associated with \( A_n \) can be bounded by
\[
\beta_n = \int_{A_n} m_n(y^n) \, dy^n \leq e^{-D_n + t_\alpha} \int_{A_n} p_n(y^n | 0) \, dy^n = e^{-D_n + t_\alpha} (1 - \alpha_n).
\]
On the other hand, let \( \{B_n\} \) be any sequence of acceptance regions with size \( \alpha_n \) and asymptotic size \( \alpha \). We can define a second threshold \( t'_\alpha \) so that

\[
\lim_{n \to \infty} \Pr\{\log \Lambda_n + D_n \geq t'_\alpha \mid \Theta = 0\} = 1 - \alpha,
\]

and an acceptance set \( C_n = \{y^n : \log \Lambda_n + D_n \geq t'_\alpha\} \) with acceptance probability

\[
1 - \alpha'_n.
\]

Then the asymptotic Type II error for \( B_n \) can be bounded below by

\[
\beta_n \geq \int_{B_n \cap C_n} m_n(y^n) \, dy^n \geq e^{-D_n + t'_\alpha} \int_{B_n \cap C_n} p_n(y^n \mid 0) \, dy^n
\]

\[
= e^{-D_n + t'_\alpha} \left( 1 - \int_{B_n \cup C_n} p_n(y^n \mid 0) \, dy^n \right)
\]

\[
\geq e^{-D_n + t'_\alpha} (1 - \alpha_n - \alpha'_n).
\]

We combine the above with Eqn. (5.8) and Eqn. (5.6), we have the following theorem:

**Theorem 5.3.1** Let the sequence of tests defined by \( \log \Lambda_n \geq t_n \) have asymptotic size \( \alpha \) under \( \Theta = 0 \). Then

\[
\limsup_{n \to \infty} \sqrt{n} \beta_n \leq \frac{w(0)}{\sqrt{I(0)}} f(\alpha)
\]

and

\[
\liminf_{n \to \infty} \sqrt{n} \beta_n \geq \frac{w(0)}{\sqrt{I(0)}} g(\alpha)
\]

where \( f(\alpha) \) and \( g(\alpha) \) are functions of \( \alpha \) alone. Hence

\[
\beta_n = O \left( n^{-1/2} \right).
\]

### 5.4 The Performance of the Trinomial Model

The previous sections have described the behavior of the trinomial model. However, the trinomial model is a much simplified version of the situation. In most physical mapping experiments, each probe will have its own particular probability of hitting a clone, so we will have \( r_1 \neq r_2 \neq \cdots \neq r_n \). In this section, we examine the performance of an overlap detector based on the trinomial model. We will see that,
when attempting to detect any specific amount of overlap, choosing all probes with the same miss probability \( r \) produces a better experiment, on average, than varying the probe miss probabilities around some mean value \( r \). This is true irrespective of the mean value \( r \).

### 5.4.1 Performance Against Specific Alternatives.

Recall that our data consists of a sequence of \( n \) vector outcomes \( \{V_i\} \), where

\[
V_i = \begin{cases} 
(1, 0, 0) & \text{if probe } i \text{ misses both clones;} \\
(0, 1, 0) & \text{if probe } i \text{ hits precisely one clone;} \text{ and} \\
(0, 0, 1) & \text{if probe } i \text{ hits both clones.}
\end{cases}
\]

Let us now consider situation which is perhaps more realistic. Let us assume that at each step \( i \), the experimenter uses a probe selected at random from a large collection of possible probes. Let us further assume that the space of miss probabilities corresponding to this collection of possible probes has a mean of \( \mu \) and a variance of \( \sigma^2 \geq 0 \). The trinomial model corresponds to \( \sigma^2 = 0 \).

Our data will then consist of a sequence of independent, identically distributed pairs \( \{(V_i, R_i)\} \), where \( R_i \) is the miss probability for probe \( i \). We will evaluate the effect of varying probe miss probability by analyzing the behavior of the marginal sequence \( \{V_i\} \). The first thing to notice is that the marginal sequence \( \{V_i\} \) is still a sequence of independent, identically distributed vectors, and hence \( V = V_1 + \cdots + V_n \) is still a trinomial random vector, albeit with a different mean vector. In terms of the \( \lambda \) parameterization in Eqn. (5.4):

\[
\begin{align*}
\tilde{p}_0(\lambda) &= \mathbb{E}p_0(\lambda, R) = \mu - \lambda(\mu - \mu^2 - \sigma^2) = \mu - \lambda\lambda^*\mu(1 - \mu) \\
\tilde{p}_1(\lambda) &= \mathbb{E}p_1(\lambda, R) = 2\lambda(\mu - \mu^2 - \sigma^2) = 2\lambda\lambda^*\mu(1 - \mu) \\
\tilde{p}_2(\lambda) &= \mathbb{E}p_2(\lambda, R) = 1 - \mu - \lambda(\mu - \mu^2 - \sigma^2) = 1 - \mu - \lambda\lambda^*\mu(1 - \mu)
\end{align*}
\]

where \( \lambda^* = 1 - \frac{\omega^2}{\mu(1 - \mu)} \). In other words, in terms of the \( \lambda \)-parameterization, the effect of a random miss probability \( R \) with mean \( \mu \) and non-zero \( \sigma^2 \) is to replace the curve of
distributions defined by \( \gamma(\lambda, \mu) = (p_0(\lambda, \mu), p_1(\lambda, \mu), p_2(\lambda, \mu)) \) in \( \mathcal{M} \) with the curve defined by \( \gamma(\lambda^*, \mu) \), which will also be in \( \mathcal{M} \).

Consider now the problem of detecting an overlap of \( \lambda_0 < 1 \). In hypothesis testing terms, this is a problem of discriminating between \( H: \lambda = 1 \) versus \( K: \lambda = \lambda_0 \). According to Stein's lemma [21, Theorem 12.8.1], the best achievable Type II error exponent, when Type I error is held fixed, is the Kullback-Leibler distance \( D(1\|\lambda_0) \) from \( \lambda = 1 \) to \( \lambda = \lambda_0 \), and this error exponent is achieved by the likelihood ratio test. That is,

- the asymptotic Type II error for our "simple versus simple" hypothesis testing problem for a given \( \alpha \) is completely characterized by the Kullback-Leibler distance \( D(1\|\lambda_0) \),
- the larger the Kullback-Leibler distance, the lower the error rate, and
- the effect of a varying probe probability is to replace the problem of testing \( H: \lambda = 1 \) versus \( K: \lambda = \lambda_0 \) with the problem of testing \( H: \lambda = \lambda^* \) versus \( K: \lambda = \lambda^*\lambda_0 \).

The relationship between \( D(1\|\lambda_0) \) and \( D(\lambda\|\lambda\lambda_0) \) is clarified by the following lemma:

**Lemma 5.4.1** Let \( \gamma \) denote the parameterization defined by Eqn. (5.4). Then for any \( r, \lambda, \lambda_0 \in (0, 1) \), we have

\[
D(\gamma(1, r)\|\gamma(\lambda_0, r)) > D(\gamma(\lambda, r)\|\gamma(\lambda\lambda_0, r)).
\]

**Proof.** Fix \( r \), and let \( d(x, y) = D(\gamma(x, r)\|\gamma(xy, r)) \). We need to show that \( d(x, y) < d(1, y) \). In fact, we show that \( \frac{\partial}{\partial x} d(x, y) > 0 \). Since

\[
d(x, y) = (r - sx) \log \frac{r - sx}{r - sxy} + 2sx \log \frac{2sx}{2sxy} + (1 - r - sx) \log \frac{1 - r - sx}{1 - r - sxy}
\]

we have

\[
\frac{\partial}{\partial x} d(x, y) = s \left[ \frac{r(y - 1)}{r - sxy} - \log \frac{r - sx}{r - sxy} - 2 \log y + \frac{(1 - r)(y - 1)}{1 - r - sxy} - \log \frac{1 - r - sx}{1 - r - sxy} \right]
\]
and we need to show the quantity in brackets exceeds 0. We can use the fact that 
\(-\log t \geq 1 - t\) to reduce the quantity in brackets to

\[
\frac{y - 2 + (1 - r)x}{1 - xy(1 - r)} + 4 - 2y + \frac{y - 2 + rx}{1 - rxy}
\]

and note that this quantity is symmetric in \(r\) around \(r = 0.5\). So, let \(r \leq 0.5\). Since 
\(f(r) = \frac{y - 2 + rx}{1 - rxy}\) is increasing in \(r\), we can bound the above quantity from below by

\[
4 - 2y + 2 \left[ \frac{y - 2 + rx}{1 - rxy} \right] \geq 2 [(2 - y)(1 - rxy) + y - 2 + rx] = 2rx(1 - y)^2 > 0
\]

and we are done.

Hence, the following theorem:

**Theorem 5.4.1** For any Type I error \(\alpha\), the asymptotically best achievable Type II error exponent for detecting a fixed overlap of \(\theta > 0\) among all experiments involving sampling random probes with mean miss probability \(r\) and variance \(\sigma^2\) is obtained precisely when \(\sigma^2 = 0\), i.e., under the trinomial model.

### 5.4.2 Performance Against General Alternatives

The above discussion focused on detecting a specific amount of overlap \(\theta\). In that situation, the log likelihood ratio is simply

\[
\sum_{i=1}^{n} \log \frac{p_n(V_i | \theta)}{p_n(V_i | 0)}
\]

and Stein’s lemma for independent, identically distributed random variables applies. Hence, the Type II error decreases exponentially in \(n\) with an exponent equal to the Kullback-Leibler distance from \(p(0)\) to \(p(\theta)\).

If we look at average power over all \(\Theta > 0\), weighted by the prior on \(\Theta\), the situation is different. In this case, the Type II error decreases only as \(n^{-1/2}\). We conjecture that the actual asymptotic Type II error rate (not just the lim sup and lim inf) will involve the ratio \(w(0)/\sqrt{I(0)}\) and a term involving \(n^{-1/2}\). In this case, the average error rate will involve averaging \(w(0)/\sqrt{I(0)}\) over the distribution for \(R\), the probe miss probabilities. Let us look at that ratio, but in the \(\lambda\) parameterization, where the situation becomes clear.
The Fisher information for $\lambda$ in one observation is given by

$$I_r(\lambda) = -\mathbb{E} \frac{d^2}{d \lambda^2} \log L(y) = \mathbb{E} \left[ y_0 \left( \frac{s}{r-\lambda s} \right)^2 + y_1 \left( \frac{1}{\lambda} \right)^2 + y_2 \left( \frac{s}{1-r-\lambda s} \right)^2 \right]$$

$$= \left[ \frac{s^2}{r-\lambda s} + \frac{2s}{\lambda} + \frac{s^2}{1-r-\lambda s} \right]$$

$$= \left[ \frac{2-\lambda}{\lambda} \right] \left[ \frac{s}{1-\lambda+\lambda^2 s} \right].$$

Note that $I_r(1) = 1$, irrespective of $r$. Hence we can concentrate on the prior on $\lambda$, given by Eqn. (5.5), but considered as a function of $r$, with $\lambda$ fixed at 1.

Figure 5.4 shows, as can be verified, that the prior distribution, considered as a function of $r$ and fixed at $\lambda = 1$, is convex in $r$. Hence, by Jensen’s theorem, the average Type II error for a random selection of probe probabilities with mean $\mu$ will exceed that for the trinomial model with $r = \mu$.

Figure 5.4: The prior $w(\lambda, r)$, considered as a function of $r$ with $\lambda$ fixed at 1.

5.5 Approximating the Posterior Odds

The above discussion has focused on theoretical results involving the trinomial random variable $y^n$ and the log-likelihood ratio $\log \Lambda_n$. When implementing our
overlap detector, we still must compute \( \log \Lambda_n \) for any given \( y^n \) we may encounter. From Eqn. (5.3), and the reparameterization described by Eqn. (5.4) and Eqn. (5.5) we see that

\[
\log \Lambda(y^n) = \log \int_0^1 e^{nh(\lambda | y^n)} w(\lambda) d\lambda,
\]

where \( h(\lambda | y^n) = \ell(\lambda | y^n) - \ell(1 | y^n) \).

Such integrals can be approximated using Laplace’s method, as described by de-Bruijn [23, Chapter 4.21. Let \( \hat{\lambda} \) denote the location of the maximum of \( h(\lambda) \) in \([0, 1]\). Then

\[
\int_0^1 e^{nh(\lambda | y^n)} w(\lambda) d\lambda = \left| \frac{2\pi}{nh''(\hat{\lambda})} \right|^{1/2} e^{nh(\hat{\lambda})} w(\hat{\lambda}) \left( 1 + O(n^{-1}) \right)
\]

whenever \( \hat{\lambda} \in (0, 1) \), and

\[
\int_0^1 e^{nh(\lambda | y^n)} w(\lambda) d\lambda = \left| \frac{1}{nh'(\hat{\lambda})} \right| e^{nh(\hat{\lambda})} w(\hat{\lambda}) \left( 1 + O(n^{-1}) \right)
\]

whenever \( \hat{\lambda} \in \{0, 1\} \). These integrals are easily evaluated, once we have \( \hat{\lambda} \), as

\[
h(\lambda | y^n) = y_0 \log \frac{r - \lambda s}{r^2} + y_1 \log \lambda + y_2 \log \frac{1 - r - \lambda s}{(1 - r)^2},
\]

\[
h'(\lambda | y^n) = -y_0 \frac{s}{r - \lambda s} + y_1 \frac{1}{\lambda} - y_2 \frac{s}{1 - r - \lambda s}, \text{ and}
\]

\[
h''(\lambda | y^n) = -y_0 \left( \frac{s}{r - \lambda s} \right)^2 - y_1 \left( \frac{1}{\lambda} \right)^2 - y_2 \left( \frac{s}{1 - r - \lambda s} \right)^2.
\]

Hence, we must compute \( \hat{\lambda} = \arg \max_{\lambda \in [0, 1]} \ell(\lambda | y^n) \), the maximum likelihood estimate of \( \lambda \).

### 5.5.1 The Maximum Likelihood Estimate of Overlap

Given \( y^n = (y_0, y_1, y_2) \), we need to compute the maximum likelihood estimate for \( \lambda \). Note that if we need the MLE for the proportion of overlap \( \theta \), we can compute the MLE for \( \lambda \), and then use the relation \( \hat{\theta} = 2 - \log(r - \hat{\lambda} s)/\log r \) to obtain the MLE for \( \theta \). Recall that the log likelihood for \( \lambda \) in terms of \( y \) is given by

\[
\ell(\lambda | y) = y_0 \log p_0(\lambda) + y_1 \log p_1(\lambda) + y_2 \log p_2(\lambda).
\]
If we take derivatives and rearrange, we see that the MLE can be obtained by solving

\[ \lambda^2 - \lambda \left( \frac{y_0}{r} + \frac{y_1}{s} + \frac{y_2}{1-r} \right) + \frac{y_1}{s} = 0 \]

for \( \lambda \). The following lemma clarifies the situation:

**Lemma 5.5.1** Let

\[ b = \frac{y_0}{r} + \frac{y_1}{s} + \frac{y_2}{1-r}. \]

Then the solutions to

\[ p(\lambda) = \lambda^2 - b\lambda + \frac{y_1}{s} = 0 \]

have the following properties:

(a) all solutions are real and non-negative;

(b) the solution given by \( \lambda = \frac{b + \sqrt{b^2 - 4y_1/s}}{2} \) always exceeds 1.

(c) the solution given by \( \lambda = \frac{b - \sqrt{b^2 - 4y_1/s}}{2} \) \( \geq 1 \) iff \( \frac{y_0}{r} + \frac{y_2}{1-r} \leq 1 \).

**Proof.**

**Part (a).** Since \( p(0) \geq 0 \) and \( p'(0) < 0 \), any real solutions will be nonnegative. Real solutions will exist whenever \( b^2 - 4y_1/s \geq 0 \). Rearranging this condition results in our needing to show that

\[ (y_0(1-r) + y_1 + y_2r)^2 \geq 4y_1s \]

for all \((y_0, y_1, y_2) \in S \) and \( r \in (0,1) \). Fix \( y_1 \) for a moment, and note that we can assume, without loss of generality, that \( r \leq 0.5 \). In that case, the left hand side of the above inequality is minimized over nonnegative \((y_0, y_2) \) subject to \( y_0 + y_2 = 1 - y_1 \) precisely when \( y_2 = 1 - y_1 \) and \( y_0 = 0 \). Hence we must show that

\[ (y_1 + (1-y_1)r)^2 \geq 4y_1s \]

which can be done by expanding the left hand side and rearranging.
**Part (b) and (c).** We show that there is at most one distinct solution in the interval $[0, 1]$. Since $p(0) \geq 0$ and $p'(0) < 0$, there will be two solutions in $[0, 1]$ whenever $p(\cdot)$ is nonnegative and increasing at 1; that is, whenever $p(1) \geq 0$ and $p'(1) > 0$. In terms of $p(\cdot)$, we will have two solutions whenever $b \leq 1 + y_1/s$ and $b < 2$. So we must show that it is never the case that $b \leq 1 + y_1/s$ and $b < 2$ simultaneously.

Since the discriminant is always nonnegative, $b \geq 2$ whenever $y_1/s \geq 1$. Hence, we can restrict our attention to the case where $y_1/s < 1$, i.e., $y_0 + y_2 > 1 - s$. We will show that whenever $y_0 + y_2 > 1 - s$, we have $b > 1 + y_1/s$, that is, $y_0(1-r) + y_2r > s$.

So fix $r \leq 0.5$ and consider the linear program

$$\min_{y_0, y_2 \geq 0 \quad 1 - s \leq y_0 + y_2 \leq 1} y_0(1-r) + y_2 r.$$

By the linear programming theorem, the minimum will occur at one of the four vertices of the simplex defined by the program. If we enumerate the vertices, and assume $r \leq 0.5$, then the minimum will be $(1-s)r$ and will occur at $(0, 1 - s)$. Since $(1-s)r > (1-r)r = s$, we are done.

To summarize the above discussion, we have the following theorem:

**Theorem 5.5.1** A unique maximum likelihood estimate for the likelihood function

$$\ell(\lambda \mid y) = y_0 \log p_0(\lambda) + y_1 \log p_1(\lambda) + y_2 \log p_2(\lambda)$$

always exists, and is given by

$$\hat{\lambda} = \min \left\{ 1, \frac{1}{2} \left( b - \sqrt{b^2 - 4y_1/s} \right) \right\}.$$

If we need an estimate for $\theta$, we can backsolve to get $\hat{\theta} = 2 - \log(r - \hat{\lambda}s)/\log r$.

We can gain some insight into the MLE by looking at $\hat{\lambda}$ as a function of $y$. Figure 5.5 shows four contour plots of the MLE estimate as a function of $(y_0, y_1)$, corresponding to four different values for $r$. We see (and the formula for the MLE verifies) that the contours form a collection of non-parallel straight lines. Thus, for any $y^x$, we simply project along the contour line to the place where it intersects the expectation curve, and take the value of $\lambda$ corresponding to that intersection as the MLE for $\lambda$. The
contour does not intersect the expectation curve if and only if the unrestricted MLE exceeds 1, corresponding to an MLE estimate for $\theta$ of $\hat{\theta} = 0$.

### 5.6 The Effect of Errors

So far we have assumed a perfect experiment: a probe is declared to positive for a clone if and only if it actually is positive for that clone. We now consider the effect of measurement error on the trinomial model. Perhaps the simplest way to address the problem of errors, other than ignoring it completely, is to construct a model with

- a fixed probability $\pi_0$ of a false positive;

- a fixed probability $\pi_1$ of a false negative; and
independent errors for different probes.

In this situation, the resulting sampling distribution is still a trinomial, but with a different mean vector \( \hat{\mathbf{p}}(\theta) \). For instance, a probe is observed as missing both clones whenever

- it actually misses both clones, and both probes are read correctly, or
- it hits one clone and misses the other, but the hit clone is misread while the missed clone is read correctly, or
- it hits both clones but both are misread.

That is,

\[
\hat{p}_0 = (1 - \pi_0)^2 p_0 + (1 - \pi_0)\pi_1 p_1 + \pi_1^2 p_2.
\]

Similar equations, summarized in the matrix equation Eqn. (5.10), apply to \( \hat{p}_1 \) and \( \hat{p}_2 \).

\[
\begin{bmatrix}
\hat{p}_0 \\
\hat{p}_1 \\
\hat{p}_2
\end{bmatrix}
= 
\begin{bmatrix}
(1 - \pi_0)^2 & (1 - \pi_0)\pi_1 & \pi_1^2 \\
2\pi_0(1 - \pi_0) & \pi_0\pi_1 + (1 - \pi_0)(1 - \pi_1) & 2\pi_1(1 - \pi_1) \\
\pi_0^2 & \pi_0(1 - \pi_1) & (1 - \pi_1)^2
\end{bmatrix}
\begin{bmatrix}
p_0 \\
p_1 \\
p_2
\end{bmatrix}
\] (5.10)

Thus a pair of error probabilities \((\pi_0, \pi_1)\) define a linear map on the simplex \( \mathcal{S} \). Figure 5.6 shows this map as a vector field on \( \mathcal{S} \).

The average effect on \( \mathcal{M} \) is summarized by the eigenstructure of the transformation matrix defined above. One can verify that the three eigenvalues of the transformation matrix are \((1 - \pi_+)^2\), \((1 - \pi+)\), and 1, where \( \pi_+ \) is the sum of the error probabilities \( \pi_0 \) and \( \pi_1 \). Of these three eigenvalues, the only eigenvector with all positive components (and hence in \( \mathcal{S} \)) is

\[
\frac{1}{(1 + \rho)^2}
\begin{bmatrix}
\rho^2 \\
2\rho \\
1
\end{bmatrix},
\] (5.11)

where \( \rho = \pi_1/\pi_0 \). Furthermore, this eigenvector corresponds to the eigenvalue 1.0. Hence Eqn. (5.11) describes the fixed point of the map, and the map shrinks by a
Figure 5.6: The effect of a measurement error model with \( \pi_0 = 0.05 \) and \( \pi_1 = 0.1 \), viewed as a vector field on the simplex \( S \). The overall effect is to shrink the manifold \( M \) towards its fixed point. The curve \( \theta = 0 \) is an integral curve for this field.

factor equal to its determinant, which is \( (1 - \pi_+)^3 \). Figure 5.7 shows four examples of the effect of this transformation on \( M \). We may conclude that the variance of the estimate for \( \lambda \) scales as the sixth power of the sum of the measurement errors, indicating a strong effect of measurement error on our ability to detect overlap.
Figure 5.7: Four examples of the effect of measurement errors on the resulting expectation surface $\mathcal{M}$. 

Effect of $\pi_0=0.05$, $\pi_1=0.95$

Effect of $\pi_0=0.1$, $\pi_1=0.95$
Chapter 6

Future Work

The field of physical mapping is evolving extraordinarily rapidly. From less than a decade ago, when the first tentative steps toward large-scale high-resolution physical mapping were being taken, map construction techniques have progressed to the point where recently Cohen et al. [18] published a coarse, “first-generation” physical map in YACs of the entire human genome. During that time, popular approaches to mapping have evolved in response to advances in technology. New cloning vectors such as YACS made cloning larger fragments of DNA possible, and still more recent vectors like PACs [37] raise the prospect of cloning large inserts without chimerism and deletion. New probing technologies like PCR have made possible the reliable detection and construction of unique segments of DNA from extremely small samples. Recently, Fluorescence In Situ Hybridization (FISH) techniques [46, 45, 91] have made possible the direct microscopic analysis of ordering and distance relationships among clones [10]. Such advances in technology bring with them new problems in experimental design and data analysis, and new opportunities for statistics to play a role. We now cite just a few examples.

High-resolution mapping using FISH is a very recent development. In this method, two or more fluorescent probes are hybridized to chromosomes, at a particular stage in the cell cycle, in many different cells. The distance (in microns) between the fluorescent dots in each cell is measured. (Another use of FISH involves mapping a probe to a chromosomal band. In this case, only one probe is used, and the chromosome is
stained to reveal its bands.) There are many variations on this basic theme, involving the number and type of different probes, the number of different fluorescent dyes, and the stage in the cell cycle. The data from a FISH experiment consists of a series of distance measurements between two or more probes. Data from multiple experiments involving multiple probes must be integrated into a set of ordering relationships among the probes. How should one model the randomness in this data? VandenEngh et al [68] have suggested that, under certain conditions, the path between probes can be modeled as a Brownian motion. However, other more recent data [10] strongly suggests that this model does not hold for very high resolution data. If statistical approaches are to be used to plan mapping experiments and evaluate the resulting uncertainty in FISH maps, good analytic models for the configuration of chromosomes in nuclei must be developed.

More generally, methods are required which recognize the pervasiveness of experimental error, and can quantify the resulting imprecision in generated maps. In all of the methods outlined in Section 1.1, statistical modeling is applied only to the simplest situation. Can one construct effective models that can be used to evaluate more complex structures such as contigs and entire maps? We would need realistic models, good approximations to the likelihood for such models, and some method such as Markov Chain Monte Carlo to explore the posterior distribution of contigs or maps, given the data. In ordinary hybridization experiments, for instance, the error processes are quite complex. Cosmids are typically arrayed on filters, with perhaps $10^4$ cosmids per filter. The filters are exposed to a complicated experimental protocol involving the desired probe. The result of the experiment is typically a gray-scale image of the entire filter, with dark dots corresponding to cosmids which hybridized successfully to the probe, and light (or non-existent) dots otherwise. This image is then usually reduced to a clone-probe incidence matrix for use by map assembly routines. Error models for the hybridization process would enable a suitable map assembly program to take account of the quality of the data in constructing the map. In this way, published maps could consist of not only the “best” map, but annotations indicating weaknesses in the map, as well as some indication of likely alternatives. Such additional data would be useful to subsequent researchers who need to know
how much they can depend on particular attributes of a published map.

In summary, methods for physical mapping have progressed to the point where a variety of approaches have been applied with varying degrees of success to moderate-to-large regions of the human genome, at a number of different resolutions. Future issues will revolve not around feasibility, but rather designing efficient and economical mapping techniques in the face of a rapidly changing technology and emerging low-resolution genomic maps. Future challenges to statisticians will more likely focus on evaluating and integrating different mapping alternatives, tailoring an approach to match the goals of the project and the strengths and interests of the laboratory. In addition to this experimental design role, statistics is uniquely suited to provide needed analytic techniques to integrate information from various maps as well as provide map consumers with some measure of uncertainty about the maps they use.
Bibliography


