BAYESIAN STATISTICAL MODELS FOR HIV EVOLUTION

Alexander Fredric Braunstein

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2009

__________________________
Supervisor of Dissertation

__________________________
Graduate Group Chairperson
Statistical models provide an important mechanism for describing and understanding sequence evolution, such as the escape response of a viral population under a particular therapy. We present a new hierarchical Bayesian model that incorporates spatially varying mutation and recombination rates into a coalescent framework for sequence evolution. Focusing on evolutionary responses to therapy, we maintain separate parameters for treatment and control groups, which allows us to estimate treatment effects explicitly. Our approach is used to investigate sequence evolution at the nucleotide level of HIV populations exposed to a recently developed antisense gene therapy, as well as a more conventional drug therapy. Detection of biologically relevant signals in both studies and recovery of true mutation and recombination rates in extensive simulation studies demonstrate the effectiveness of our method.
Contents

Title Page i

Abstract ii

Table of Contents iii

List of Figures vi

Acknowledgements ix

1 Introduction and Biological Background 1

1.1 Biological Definitions 2

1.2 Biology of HIV 5

2 Previous Models for Sequence Evolution 12

2.1 Phylogenetic Trees 13

2.2 The Coalescent 15

2.2.1 The Coalescent with Mutation and Recombination 18
2.3 Approximate Likelihood Methods ............................. 22
2.4 OMEGAMAP .................................................. 24

3 PICOMAP: A Hierarchical Coalescent Model ................. 32
  3.1 Coalescent-based Likelihood ............................... 33
  3.2 Hierarchical Prior for Spatially-varying Parameters $\Theta$ .......................... 35
  3.3 MCMC Implementation .................................... 38

4 Model Validation ................................................ 47
  4.1 Simulation Evaluation ................................. 47
  4.2 Monitoring of Convergence ............................... 52

5 Applications to HIV therapies .................................. 60
  5.1 Enfuvirtide Drug Therapy .................................. 60
  5.2 VIRxSYS Antisense Gene Therapy ......................... 64

6 Comparison to Simpler Approaches ............................ 68
  6.1 EDA ......................................................... 69
  6.2 Regression Approaches ....................................... 72

7 Other Applications of the Coalescent ......................... 79
  7.1 A Generative Model for the Coalescent ..................... 80
  7.2 The Coalescent with Recombination and Genealogies .............. 82
  7.3 Likelihood Calculations and Message Passing Algorithms ......... 87
7.4 MCMC Moves Using Coalescent with Recombination and Genealogies Graphs ......................... 91

8 Discussion and Future Work .......... 94

References ........................................ 98
List of Figures

1.1 Central dogma of molecular biology ........................................ 3
1.2 Point mutation and recombination ........................................... 5
1.3 Deletion and insertion ........................................................... 6
1.4 Drug and gene therapies .......................................................... 8
1.5 Two-sample data generation ....................................................... 10

2.1 A phylogenetic tree ................................................................. 15
2.2 MRCA plot ............................................................................. 17
2.3 Coalescent with recombination ................................................ 19
2.4 OMEGAMAP blocking structure ............................................. 25

3.1 PICOMAP parameter hierarchy ............................................... 36
3.2 PICOMAP example blocking structure .................................... 36

4.1 Simulation treatment effects .................................................. 49
4.2 Experiment 1 coverage plot ..................................................... 49
4.3 Experiment 2 coverage plot ..................................................... 51
Acknowledgements

The mentorship, knowledge, and friendship of Professor Shane Jensen made this dissertation possible. Thank you for imparting on me the wisdom of the Bayesian paradigm, the importance of applied research, specifically biological applications, and both the joy and heartbreak of baseball. I could not have asked for a better advisor.

My coauthor and former Professor Jon McAuliffe also had a sizable impact on this dissertation, particularly the C++ implementation. In addition, I would like to thank the other two members of my dissertation committee, Professor Ed George and Professor Dylan Small, and the entire Statistics department at the Wharton School of the University of Pennsylvania. I am grateful to the VIRxSYS and Doms labs for access to their data, and the Center for AIDS Research (CFAR) for providing me with support during my graduate education.

I am extremely fortunate to have contemporaneously been a student in the Wharton Statistics department with Sivan Aldor-Noiman, Mike Baiocchi, Blakely McShane, James Piette, Sathya Anand, Shaun Lysen, and Frank Yoon. Whether
doing problem sets, helping each other code, or just hanging out in Philly, you have
made my graduate school experience an enjoyable one.

Far and away the best summer of graduate school was spent at Google NYC
with Samantha Cook, Jeremy Shute, Phong Pham, and David Agraz. Thanks for
getting me a job, making me a much better programmer, and helping me put the
ejew back into jiu jitsu.

I resided at Harrison College House my 4 years of graduate school. While itself
an enriching experience, the people I met while a GA have been the most impor-
tant part. Theresa Gue, Jamie Doyle, Shahbaz Alam, Niva Kramek, and Joshua
Markman, I will miss you all very much.

I would like to thank Kelly Kleinert, Carmen Barron, Veronika Dubajova, Alex
Kalkstein, Justin Zorn, Aline Normoyle, and Lt. Joshua Swartsel. Each of your
friendships have enriched my life.

Last, but certainly not least, I would like to thank my Mother, Father, Sister,
Brother, unofficial big sister Anita, and dog Lacey. Even though I am moving to
Siberia, I love you all very much. Your constant love and support helped get me
through the rigors of graduate school.
Chapter 1

Introduction and Biological Background

Acquired Immundeficiency Syndrome (AIDS), the disease caused by the progression of the HIV virus, has killed more than 25 million worldwide [Joint United Nations Programme on HIV/AIDS 2006]. Significant progress has been made using antiretroviral drug and gene therapy treatments to decrease the mortality and morbidity of the disease, by slowing or arresting the progression of HIV to AIDS. This dissertation studies two such therapies and models the resulting evolution of the HIV virus. We posit and implement a hierarchical Bayesian coalescent model that measures spatial sequence rate changes between treatment and control HIV sequences, so that scientists can understand the mechanisms and relative magnitudes of escape responses. This information could lead to improved design of future
1.1 Biological Definitions

DNA, RNA, and proteins are the biological molecules discussed in this dissertation. The basic unit of data analysis is DNA, which is composed of paired strings of nucleotides: adenine, cytosine, guanine, and tyrosine (abbreviated as A, C, G, and T). The two strings of nucleotides are reverse complimentary, meaning that nucleotides at each location in the DNA are paired A-T and C-G. This form is incredibly stable, because it allows the formation of strong hydrogen bonds between the pairs of nucleotides and a double helix to form between the pairs of DNA sequences.

The RNA molecule is a single string of nucleotides, which is inherently less stable than the double helix. In RNA, the nucleotide tyrosine is replaced by uracil (abbreviated as U). Created in the relatively safe nucleus, the molecule moves into the cytosol, where it delivers messages or interacts with the cell’s ribosomes or endoplasmic reticulum. The less stable molecule must quickly accomplish its task in the harsh environment of the cytosol, or it will be degraded and broken down by the cell.

Strings of amino acids make up proteins, which fold into three dimensional shapes called their tertiary structure. The ordering of the amino acids determines the structure, which has a strong impact on the function of the molecule. Proteins
are involved in nearly every cell function.

The three molecules are related through the central dogma of molecular biology, which dictates that small sections of the DNA called genes are used as a blueprints for creation of RNA through a process called transcription. RNA molecules are temporary intermediaries until they are used as blueprints for proteins, the workhorse of the cell. In translation, triplets of nucleotides in RNA code for one of twenty amino acids, which make up the protein. The process is illustrated in Figure 1.1. Several classes of organisms, including HIV, have RNA, rather than DNA, as their initial blueprint.

![Central Dogma Diagram](image)

Figure 1.1: Pictorial illustration of the central dogma of molecular biology.

DNA within each somatic cell of an organism is identical. The diversity in cell structure and function within individual organisms is caused by gene regulation, the process by which the activity of particular genes is increased or decreased. The primary mechanism for regulation is changing which genes are expressed and how often they are transcribed from DNA to RNA.

Errors can occur during the creation of a protein, causing it to be different than
intended. We discuss four types of errors

1. point mutation - the change of a single nucleotide

2. insertion of an extra nucleotide

3. deletion of a nucleotide

4. recombination - the process by which two sequences create a single mosaic sequence

The distinct sequences involved in recombination can be referred to as haplotypes. Point mutation and recombination are illustrated in Figure 1.2.

Some mutations are silent; the nucleotide triplets TTT and TTC both code for the amino acid phenylalanine. However, mutating TTT to ATT, CTT, and GTT code for isoleucine, leucine, and valine, respectively. Mutations that change the amino acid coded for by the nucleotide triplet can cause proteins to fold in very different ways, changing the function of the protein, rendering it inert, or preventing the transcription process from completing [Harvey, Champe, and Fisher 2001].

Insertion and deletion are illustrated in Figure 1.3. To represent deletion, we use the gap symbol —. In both cases, we observe that insertions and deletions can have profound impact on the resulting amino acid sequence by shifting the reference frame. In the remainder of the dissertation, we refer to point mutation, insertion and deletion as mutation.
Figure 1.2: Pictorial representation of sequence point mutation (left) and recombination (right). On the left side we see the third nucleotide of the sequence mutating C to T. On the right side, we see a black and blue line, which each represent a sequence of nucleotides (such as the one in the mutation figure on the left). In HIV, the virus particle contains two RNA genomes. While one is being transcribed, the particle transcribing the genome can jump between the two, producing a hybrid child. In this case, the top (black) sequence is ancestral to the child’s 5’ (left) end, while the blue sequence is ancestral to the child’s 3’ (right) end.

1.2 Biology of HIV

HIV is a lentivirus, which is 120nm in diameter, about 60 times smaller than a red blood cell and roughly spherical [Harvey, Champe, and Fisher 2001]. The genome consists of 9 genes (gag, pol, env, tat, rev, nef, vif, vpr, vpu, and tev), which encode 19 proteins. Three genes, gag, pol and env, contain the information necessary for the production of new virus particles. The remaining genes code for regulatory proteins that control its ability to infect cells [Kuiken 2008]. The reproduction and
Figure 1.3: Pictorial representation of deletion (left) and insertion (right). On the left we see the fourth nucleotide is deleted. On the right side, we see a the insertion of a T at the end of the sequence.

infection cycles of HIV work in the following way. The HIV virus attaches to the target cell, where it releases its viral RNA. Reverse transcriptase then transcribes the viral RNA to create double stranded DNA, which is integrated into the host cell’s DNA. When the host cell’s DNA is transcribed to RNA then translated into proteins, the portions inserted by HIV are transcribed and then translated as well, allowing the virus to repackage itself and infect other cells.

The escape response of HIV to therapies is particularly effective for two reasons. First, HIV generates between $10^9$ and $10^{10}$ virions per day, which is large relative to other viruses. Even for organisms with relatively low mutation rates, the more particles created, the larger the pool of HIV viruses that may contain an evolutionary advantage. Second, the virus has a mutation rate of $3 \times 10^{-5}$ per nucleotide base per cycle of replication, one of the highest mutation rates of any organism [Robertson et al 1995]. The result is a virus which is both incredibly resilient and virulent.
Recombination exacerbates the problem by creating an additional mechanism for virions to share beneficial mutations between individuals in a population.

We focus primarily on the gene $env$, which codes for the envelope proteins. After several important findings concerning the replication cycle of HIV [Kwong et al 1998], many labs began to focus on this region for drug and gene therapies. The two proteins encoded by the $env$ region, gp120 and gp41, control the ability of HIV to produce new viral particles.

The virus is typically identified by type and subtype, which are determined by differences in $env$ region in gp120 [Thomson et all 2002]. The distinction of subtype becomes particularly important when studying drug resistant mutations, which occur as an escape response to drug therapies. When the same beneficial mutation develops in more than one subtype, it is usually the result of a recombination event transferring the mutation rather than two separate HIV viruses generating the exact same drug resistant mutation.

This dissertation examines two categories of therapies for treating HIV infection, a drug therapy and an antisense gene therapy. Though the goal of both therapies is to disrupt the replication and infection cycles of HIV, they do so via vastly divergent mechanisms. Drug therapies attack the virions containing the viral load before they reach potential host cells, and the antisense gene therapy attacks the viral load once it has been released into the cytosol of host cells. Figure 1.4 demonstrates both processes.
Figure 1.4: Pictorial representation of drug (left) and gene (right) therapies attacking the HIV virion and HIV viral payload respectively.

Drug therapies operate on the actual protein products coded in the HIV genome. Proteins surrounding the HIV virion are inhibited by a drug therapy, preventing the virus from binding to host cells and delivering its viral load. Drugs attach themselves to the binding locations, so they cannot bind with potential host cells. Here the tertiary structure of the protein is very important and small changes in the shape of these proteins, caused by mutations, may prevent drug therapies from
binding to the new tertiary structures of the HIV proteins as effectively or at all.

Antisense gene therapies bind directly to the HIV genome, creating double stranded RNA, which is not a naturally occurring molecule in human cells. The double stranded RNA will be degraded by normal cellular activity in the cytosol, even more quickly than a normal RNA molecule. Rather than inhibiting the virion from delivering its viral load, this approach neutralizes the viral load once delivered and degrades it before the RNA can be incorporated into the host cell’s DNA. The neutralization of the viral load prevents infection of the cell, and consequently, the creation of additional HIV virions.

Despite the sophisticated approach of both types of therapies, HIV is still capable of escaping treatment through the process of evolution: changes in its genomic sequence. Understanding the different mechanisms of each of these therapies is important, because it underscores the necessity for methods that recognize the structure of the biology underlying the problem. A model designed to measure recombination in sequences that have been subjected to a drug therapy, which attacks the protein products created by the virus rather than the virus itself, may not be suitable for the same inference tasks on sequences subjected to an antisense drug therapy, where the viral RNA is under selective pressure.

In many biological applications we study a population of sequences, rather than individual sequences. When HIV infection occurs, it is generally with several distinct strains of the virus. The evolutionary escape response of the population, not
the individual, will dictate clinical outcomes, so it is natural to model the population rather than individual level changes.

Figure 1.5 demonstrates the two-sample data generation process. Whether the study is *in vitro* or *in vivo* or a drug or gene therapy, will determine the exact meaning of the treatment and sequence arrows.

It is also instructive to discuss the evolution of HIV via mutation and recombination in the context of censoring. As a population of HIV evolves in the absence of selective pressure (due to a drug or gene therapy), individual strains are selected as more “fit” than others. Strains not selected are censored by the evolutionary process. In the applications of this dissertation, we measure differences in mutation and recombination rates as proxies for the way in which HIV strains under selective
pressure are censored differently than neutrally evolving strains.

The remainder of the dissertation will be organized as follows. In chapter 2, we briefly survey the previous literature on the topic. We discuss phylogenetic trees and the coalescent, two models for quantifying sequence evolution, and focus on several applications of coalescent models, as they are most similar to our application. Our Bayesian hierarchical model will be explained in chapter 3. This includes specification of our prior, an explanation of the likelihood calculations, and details of our MCMC implementation. We validate our model in chapter 4, running a battery of simulations and investigating some sampler results to ensure the sampler converges. Results from our two applications will be the focus of chapter 5. One of our data sets is from a drug therapy study, and the other is from an antisense gene therapy study. Our model addresses the inference issues that arise in an antisense gene therapy clinical trial. Chapter 6 compares our model to less sophisticated approaches. The methodologies we use for posterior inference are computationally expensive, so we demonstrate that the extra computational burden is justified and adds important understanding and precision to the analysis of the results. In chapter 7, we take a closer look at the coalescent, the model underlying the majority of the calculations for our model, and consider its application to several clustering problems. Chapter 8 concludes the dissertation with a summary and suggested future directions for the research.
Chapter 2

Previous Models for Sequence Evolution

We discuss four models in this section

1. phylogenetic trees

2. the coalescent [Kingman 1982]

3. an approximation to the coalescent by [Li and Stephens 2003]

4. OMEGAMAP [Wilson and McVean 2006] - a Bayesian hierarchical model that generalizes the approach of Li and Stephens.

Each method makes assumptions about the underlying mutation or recombination process, which induce simplifications or approximations to the calculations of the
model. To assist our discussion, we define several quantities and a general computational framework into which all four methods fit.

Let \( \Theta \) be a generic set of evolutionary parameters. As an example, \( \Theta \) usually includes parameters for mutation and recombination rates, but we do not restrict the definition in order to remain as general as possible. Additionally, let \( H = (h_1, \ldots, h_n) \) be the set of sequences which we will use to estimate \( \Theta \). Traditional approaches either estimate \( \Theta \) by maximum likelihood (looking at: \( \sup_{\Theta} p(H|\Theta) \)) or take a Bayesian approach, estimating a posterior distribution \( p(\Theta|H) \propto p(H|\Theta)p(\Theta) \). Though Bayesian models typically use Markov Chain Monte Carlo (MCMC) methods to obtain samples from the posterior, maximum likelihood approaches to obtain parameter estimates can yield identical or nearly identical solutions when using flat priors. Each of the four models we discuss will be a variant on this generic setup; we will enumerate the set of parameters in and priors on \( \Theta \) as well as the particular methodologies used for parameter estimation.

### 2.1 Phylogenetic Trees

Defined formally, a phylogenetic tree is either a rooted or unrooted bifurcating tree which shows the relationship between sequences in terms of evolutionary distance. The first illustration of a tree expressing the relationship between species, in the modern sense of the word, dates back to *The Origin of Species* [Darwin 1859], and the idea has become so important and ingrained in all aspects of biology, that an
entire issue of Science was dedicated to the topic in 2003. Phylogenetic trees are the traditional method of analysis for estimation of mutation rates and evolutionary distance between species. Assuming no recombination, each nucleotide evolves independently by the process of mutation. This assumption of independence makes MLE estimation of trees, conditional on a mutation model, a relatively straightforward task from a statistical perspective. The parameters of the particular mutation model along with the structure of a phylogenetic tree are the parameters contained in $\Theta$. After aligning species, it is straightforward to calculate summary statistics of pairwise differences between species. Research in the area generally focuses on how to use this information to construct a tree, including branch lengths and internal nodes. Approaches range from relatively simple parsimony approaches [Kolaczkowski and Thornton 2004], to more difficult and computationally intensive approaches, such as MAVID, which uses polytope calculations from algebraic topology [Bray and Pachter 2004]. Most approaches are frequentist, and we can think them in terms of our setup with a flat Bayesian prior, $p(\Theta) \propto 1$. We show an example phylogenetic tree in Figure 2.1.

Unfortunately, phylogenetic trees have been shown to produce badly biased results for organisms with high recombination rates, such as viruses [Schierup and Hein 2000]. The bias is a consequence of recombination creating differing ancestral trees spatially across the sequences, which is very likely to occur within our applications. By excluding recombination rates from $\Theta$, these methods are able
to assume independence of nucleotides, which leads to massive simplifications in
the likelihood calculations. When this assumption is true, phylogenetic trees are
invaluable models for parameter estimation. However, when this assumption is not
met, the model will provide incorrect results. Consequently, phylogenetic trees are
an inappropriate modeling tool in the context of our datasets.

2.2 The Coalescent

The coalescent was first proposed in [Kingman 1982], and then generalized to include
recombination soon thereafter [Hudson 1983]. The coalescent describes a generative
model for a population of individuals, creating generation $g$ from generation $(g - 1)$. To create the next generation moving forward from the current generation of $N$
individuals, we sample individuals with replacement. It is conventional, and will
be the case in this dissertation, to maintain a constant population size, though this
can be generalized [Hein et al 2005]. Given this generative framework, we can solve
for many quantities of interest, which provide some intuition about the model

- $P(2 \text{ items find an ancestor 1 generation back}) = 1/(N)$

- $P(k \text{ items have } k \text{ different ancestors 1 generation back})$
  
  $= \left(\frac{N-1}{N}\right) \left(\frac{N-2}{N}\right) \ldots \left(\frac{N-k+1}{N}\right)$

  $= \prod_{i=1}^{k-1} \left(1 - \frac{i}{N}\right)$

  $= 1 - \left(\frac{k}{2}\right) \frac{1}{N} + O\left(\frac{1}{N^2}\right)$

- This tells us that $P(\text{coalescent event in a given generation}) \approx \left(\frac{k}{2}\right) \frac{1}{N}$

Some derivations of the Coalescent will use $2N$ instead of $N$ individuals, reflecting
the difference between species that are diploid (possessing two complete sets of
chromosomes) and haploid (possessing one complete set of chromosomes). The
most recent common ancestor (MRCA) is the first individual that is an ancestor
to all individuals selected out of a given population. Identification of the MRCA is
illustrated in Figure 2.2.

Generalizing to generations in continuous time leads to the result that the time
between coalescent events has an exponential distribution with rate $\left(\frac{k}{2}\right)$, where $k$
is the current number of individuals. This rate provides the previously mentioned
intuition concerning the number of sequences and coalescent times. When the
Figure 2.2: An illustration of the basic coalescent. Arrows from an earlier generation indicate that the individual is reproduced in the subsequent generation and is ancestral to the last generation. Multiple arrows indicate multiple reproduction of a single individual. The solid red individual in generation $t-3$ at position 4 is the MRCA for individuals 1, 3, and 5 from generation $t$. The unfilled circle in generation $t-2$ at position 3 is the MRCA for individuals 1, 3, and 4 from generation $t$.

When the number of sequences is large, sequences coalesce very quickly. For most values of $k$, even very large values, the majority of generations are spent waiting for the last two sequences to coalesce, rather than the first $k-1$. The only parameter in the basic coalescent is the population size.
2.2.1 The Coalescent with Mutation and Recombination

Though an interesting generative model, the coalescent is too simplistic to address the actual dynamics of a population because it does not include sequence changes via mutation or recombination. The coalescent is a model relating ancestries of sequences, not a model of mutation. We include mutation and recombination by modifying the generative process of the simple coalescent. For mutation, after choosing each sequence of the next generation we mutate the sequence with probability $p$ at one location chosen uniformly at random. Mutation can have many different meanings in this context, which in turn effects the size and relative difficulty of estimating $\Theta$. The infinite alleles [Ewens 1972], infinite sites [Kimura 1969], Jukes-Cantor [Jukes and Cantor 1969], Kimura [1980], etc. are all examples of mutation models that can be used in conjunction with the coalescent. As an example, the infinite alleles model has an exact parametric form and the number of sequences and their identities have direct analogies to the number of tables and individuals at each table in the Chinese restaurant process/Dirichlet process [Teh, Jordan, Beal and Blei 2006]. Unfortunately, these relatively simple forms are not relevant for the 5 symbol mutation model we outline in the biological background section.

Some quantities are still easily calculated, for example the time until the first mutation event (of any sort) for $k$ sequences. In a derivation similar to the coalescent times, we find that this quantity has an exponential distribution with rate $\frac{k\mu}{2}$, where $\mu = 4Np$, and the constants $N$ and $k$ are respectively the length of the sequence
and current number of sequences.

For recombination, when choosing each sequence of the next generation, a second sequence is chosen uniformly at random from the remaining $N - 1$ sequences. With probability $r$ we make the sequence of the next generation a hybrid of these two sequences, where the split point is chosen uniformly at random. As illustrated in Figure 1.2, this process creates a novel haplotype. Figure 2.3 illustrates the generative process. It is the analogue of Figure 2.2 for the coalescent with recombination. When considering a sample of $k$ sequences in continuous time, a recombination event increases the number of sequences by 1.

![Figure 2.3: An illustration of creation of one generation of the coalescent with recombination. We see the first sequence in the second generation is a recombinant of the first two sequences from the previous generation. The second sequence of the second generation is an exact copy of the third sequence from the first generation, etc.](image)

The time until a recombination event also has an exponential distribution with rate $\frac{\rho k}{2}$ where $\rho = 4Nr$. Again, $N$ is the length of the sequence and $k$ is the current number of sequences. The recombination and mutation processes occur
independently. Looking backwards in time from our current sample of sequences, these dynamics create a birth-death process with birth rate $\frac{\rho_k}{2}$ (recombination) and death rate $\binom{k}{2}$ (coalescence). We reach the MRCA, our absorbing state, when $k = 1$.

Recombination adds a substantial wrinkle to the coalescent as we can have regions of the sequences having coalesced to the MRCA, while the other regions have not. In addition, sections of individuals in the sample may not be from the ancestral sequence at all. The dynamics of the coalescent with recombination create a graph structure, referred to as the Ancestral Recombination Graph (ARG), that describes the entire genealogy including all recombination events [Griffiths and Marjoram 1997]. The more detailed structure significantly complicates inference and estimation of recombination and mutation rates. Combinatorially the basic coalescent has $\frac{n(n-1)!}{2^{n-1}}$ possible graphs [Hein et al 2005]. Even for a relatively small number of sequences, such as 100, there are approximately $1.37 \times 10^{284}$ graphs. This number is a lower bound for the coalescent with recombination; adding recombination events significantly increases the number of possibilities. Though ARGs reach a MRCA in finite time with probability 1, the number of events is unbounded and supported on the set of non-negative integers. This large number of ARGs significantly increases the difficulty of inference.

Let $\Theta$ be our set of parameters for mutation and recombination rates of our ARG. As described earlier, traditional approaches either estimate $\Theta$ by maximum likelihood (looking at: $\sup_{\Theta} p(H|\Theta)$) or use Bayesian methods to estimate
the posterior distribution $p(\Theta|H) \propto p(H|\Theta)p(\Theta)$. The large number of ARGs, which we represent as $G$, are integrated out of both approaches. Searching over $\sup_{\Theta, G} p(H|\Theta, G)$ or integrating over $p(\Theta|H) \propto \sum_{G} p(H|\Theta, G)p(\Theta)p(G)$ are both very difficult tasks, considering the size of the sample space for $G$. Conditioning on a particular configuration of the ARG significantly reduces the complexity of estimating the recombination and mutation rates, $\Theta$.

[Fearnhead and Donnelly 2001] take advantage of the conditional simplicity provided by conditioning on the ARG and take steps to tackle the relatively large sample space of ARGs consistent with the data. They derive near optimal proposal distributions for the set of ARGs and implement importance sampling over ARGs to estimate recombination and mutation rates, which are constant across the genome. The exact calculations are computationally intractable for even small sequences, so they implement a dynamic programming algorithm to approximate the likelihood. Though this approach is a successful first pass at simultaneous estimation of mutation and recombination rates, it does not scale well in the number and size of the sequences, mainly due to the superexponential growth in the size of the ARG space. Still, the motivation for conditioning on ARGs and the simplifications in the ARG proposals distributions serve as important motivation for future work.
2.3 Approximate Likelihood Methods

[Na Li and Matthew Stephens 2003] investigate a different approach to estimation of $\Theta$ in the presence of a large number of possible ARGs $\mathcal{G}$. Rather than sampling and conditioning on ARGs to relax the difficulty of the problem, they focus on the conditional distributions

$$p(h_1, h_2, \ldots, h_n|\Theta) = p(h_1|\Theta)p(h_2|h_1, \Theta) \cdots p(h_n|h_{n-1}, \ldots, h_1, \Theta)$$

These conditional distributions are easily calculated under some assumed models of sequence evolution, such as infinite alleles [Ewens 1972]. However, the conditional distributions are not computationally tractable for the more sophisticated models of sequence evolution that account for both mutation and recombination processes. Consequently, the authors develop a series of approximations to these conditional distributions

$$p(h_1, h_2, \ldots, h_k|\Theta) \approx \hat{\pi}(h_1|\Theta)\hat{\pi}(h_2|h_1, \Theta) \cdots \hat{\pi}(h_k|h_{k-1}, \ldots, h_1, \Theta) \equiv \text{PAC}(H|\Theta)$$

The approximation utilizes a Markov chain, where the approximate conditional distribution $\hat{\pi}(h_k|h_{k-1}, \ldots, h_1, \Theta)$ is the probability of observing sequence $k$ given the first $k - 1$ sequences and current set of parameters $\Theta$. The distributions $\hat{\pi}(h_k|h_{k-1}, \ldots, h_1, \Theta)$ can be calculated as

$$\hat{\pi}(h_k|h_{k-1}, \ldots, h_1, \Theta) = \sum_{x=1}^{k} \alpha_S(x)$$  \hspace{1cm} (2.3.1)$$

where $S$ is the length of the sequence and the $\alpha$ terms can be computed recursively.
using the equation
\[
\alpha_{j+1}(x) = \gamma_{j+1}(x) \left( p_j \alpha_j(x) + (1 - p_j) \frac{1}{k} \sum_{x' = 1}^{k} \alpha_j(x') \right) \quad (2.3.2)
\]
Additionally, we define \( p_j = \exp(-\rho_j/k) \), and \( \gamma_{j+1}(x) = P(h_{k+1,j} = a|X_j = x, h_1, \ldots, h_k) \), the probability of observing \( a \) at site \( j \) in sequence \( k + 1 \) given our observation of \( x \) at site \( j \) in the same sequence, is defined
\[
\gamma_{j+1}(x) = \begin{cases} 
\frac{k}{k+\tilde{\theta}} + \frac{\tilde{\theta}}{2(k+\theta)} & h_{x,j} = a \\
\frac{\tilde{\theta}}{2(k+\theta)} & h_{x,j} \neq a 
\end{cases} \quad (2.3.3)
\]
where
\[
\tilde{\theta} = \left( \sum_{m=1}^{n-1} \frac{1}{m} \right)^{-1} \quad (2.3.4)
\]
is a frequentist estimator of the global mutation rate. Their method uses the infinite sites model, which assumes that only one mutation can ever occur at each site. An implication of this model is that we can represent mutations at each site as 0 or 1, implying ancestral or mutated state respectively. For this particular set of assumptions, it only matters if the nucleotides are different, without regard to their individual identities. Although it holds in some circumstances where mutation rates are relatively low, most data sets violate the infinite sites assumption.

Li and Stephen’s model provides an important leap forward for two reasons. First, the approximation to the coalescent allows the authors to consider a larger number of sequences. Second, it generalizes the recombination parameter, \( \rho \) to include a spatial component. In the definition of \( p_j \) below equation 2.3.2 we can define a value of \( \rho \) for each nucleotide.
OMEGAMAP [Wilson and McVean 2006] integrates the PAC likelihood of Li and Stephens into a Bayesian model and reversible jump MCMC scheme (RJMCMC)[Green 1995] to perform spatial inference on protein change and recombination, which they denote as $\omega$ and $\rho$ respectively. Their goal is to estimate the posterior distribution of their parameter set $\Theta$ given the set of sequences $H$.

$$P(\Theta|H) \propto P(H|\Theta)P(\Theta) \quad (2.4.1)$$

First we discuss the structure on their priors $P(\Theta)$, then address their modifications to the work of Li and Stephens for calculating $P(H|\Theta)$.

Wilson and McVean place independent exponential priors on the rates for recombination and protein change for the sequences. They break the sequences into blocks and estimate the recombination and protein change rates for each block as constant, which is illustrated in Figure 2.4.

A binomial prior is placed on the number of blocks for both sets of parameters, which we denote as $B_\omega$ and $B_\rho$. Setting notation for the nucleotide locations of the boundaries of the blocks

$$s^B_\omega = (s_0, s_1, \ldots, s_{B_\omega+1}) = (1, s_1, \ldots, s_{B_\omega}, L) \quad (2.4.2)$$

where $L$ is the length of the sequences and the rates for each of the $\omega$ blocks

$$\omega^{B_\omega} = (\omega_1, \ldots, \omega_{B_\omega}) \quad (2.4.3)$$
we have the prior for this set of parameters

$$p(B_\omega, s^{B_\omega}, \omega^{B_\omega}) \propto p_\omega^{B_\omega+1}(1 - p_\omega)^{L - B_\omega - 1}\lambda_\omega^{B_\omega}\exp(-\lambda_\omega(\omega_1 + \cdots + \omega_{B_\omega}))$$  (2.4.4)

$\lambda_\omega$ is a prior on the rate on the $\omega$ rates and $p_\omega$ is a prior which controls the binomial number of blocks; both are set by the user. The expression above is a product of the binomial prior on the number of blocks and $B_\omega$ individual exponential distributions for each of the $\omega$ rates. An identical set of parameters exists for the recombination rates. $\kappa$, the transition/transversion ratio, is a parameter that controls the tendency of the nucleotide pairs A-T and C-G, to mutate between each other as opposed to other nucleotides. It has an exponential prior with mean $1/\lambda_\kappa$, which is also set by the user. All parameters are mutually independent, so the full prior takes the form

$$P(\Theta) = p(B_\omega, s^{B_\omega}, \omega^{B_\omega})p(B_\rho, s^{B_\rho}, \rho^{B_\rho})\lambda_\kappa\exp(-\lambda_\kappa(\kappa))$$  (2.4.5)

This prior has the interesting property that it depends only on the number of blocks.
and rates within blocks, not their locations. The property implies exchangeability over all blocking configurations with equal values of $B_\rho$ and $B_\omega$.

Next, we turn to Wilson and McVean’s modification of the PAC likelihood equations from [Li and Stephens 2003], which allow us to calculate the conditional likelihood $P(H|\Theta)$. The same basic equations for the likelihood calculations hold

$$\alpha_{j+1}(x) = \gamma_{j+1}(x) \left( p_j \alpha_j(x) + (1 - p_j) \frac{1}{k} \sum_{x' = 1}^{k} \alpha_j(x') \right)$$ (2.4.6)

however, their mutation model is much different than that of Li and Stephens. We again have that $p_j = \exp(-\rho_{j}/k)$ and $\gamma_{j+1}(x) = P(h_{k+1,j} = a|X_j = x, h_1, \ldots, h_k)$, but the interpretation of $\gamma_{j+1}(x)$ is slightly modified. Instead, the expression is the probability of observing $a$ at codon site $j$ in sequence $k+1$ given our observation of $x$ at codon site $j$ in the same sequence. They assume a codon-based mutation model, as they are concerned with selectively beneficial protein change, rather than the existence of nucleotide-level mutations. Continuing with the example from section 1.1, the codon TTT mutating to TTC, still yields phenylalanine, however, mutating TTT to ATT, yields isoleucine. The latter case is an example of selective diversification of a beneficial mutation. Unlike the previous application which simply kept track of two states, this model keeps track of all possible triplets of nucleotides. This creates a space 61 states and a $61 \times 61$ transition matrix. The three states corresponding to “stop codons” are not included, because they are absorbing states that only occur at the end of each sequence. The transition matrix probabilities
are defined

\[ P(a, b|t) = \delta_{ab}\pi_{a}p_{ab}^{(2t)} \]  

(2.4.7)

where

\[ \delta_{ab} = \begin{cases} 1 & \text{if } a = b \\ 2 & \text{if } a \neq b \end{cases} \]  

(2.4.8)

The transition probabilities \( p_{ab}^{(2t)} \) are defined by Nelson and Yang [1998], modified to include insertions and deletions. \( \pi_a \) is the base frequency for codon \( a \). Solving numerically for codon probabilities

\[ P(a, b|t) = \delta_{ab}\pi_{a}\sum_{c \in \mathcal{C}} \nu_{ac}\nu^{-1}_{cb} \exp(2d_{c}t) \]  

(2.4.9)

\( \mathcal{C} \) is the space of possible codons, \( \nu \) is a matrix of eigenvectors of the mutation rate matrix, \( \nu^{-1} \) is its inverse and \( d_{c} \) is the vector of eigenvalues. Because we only observe the leaves of the ARG, we need to integrate time out of this expression. Doing so yields

\[ \gamma_{j+1}(a) = P(a, b) = \int_{0}^{\infty} P(a, b|t)P(t)dt = \delta_{ab}\pi_{a}\sum_{c \in \mathcal{C}} \nu_{ac}\nu^{-1}_{cb} \frac{k}{k - 2d_{c}} \]  

(2.4.10)

Plugging this expression into the equations of Li and Stephens allows solution of the likelihood recursions.

Now that we have specified both \( P(H|\Theta) \) and \( P(\Theta) \), we enumerate the MCMC moves of the OMEGAMAP sampler, which will provide a more direct comparison to the MCMC moves of our own model later in the dissertation. The acceptance probabilities for the moves are
1. Blocking structure MCMC moves

(a) Choose an \( \omega \) block uniformly to split or merge with a neighbor

We split or merge omega blocks, with split to merge ratio

\[
\frac{s_B}{m_B} = \frac{\min(1, P(B + 1)/P(B))}{\min(1, P(B - 1)/P(B))}
\]

For a split move, draw \( U \sim \text{Uniform}[-1,1] \), then let

\[
\omega'_j \omega'_{j+1} = \omega_j
\]

and

\[
\frac{\omega'_{j+1}}{\omega'_j} = \frac{1 - U}{U}
\]

\[
R = \frac{P(H|\Theta') p_\omega e^{-\lambda_\omega(\omega'_j + \omega'_{j+1})} m_{B+1}(L - B + 1) (\omega'_j + \omega'_{j+1})^2}{P(H|\Theta) (1 - p_\omega) e^{-\lambda_\omega \omega_j} s_B(B + 1)}
\]

For a merge move, let

\[
\omega'_j = \omega_j \omega_{j+1}
\]

then

\[
R = \frac{P(H|\Theta') (1 - p_\omega) e^{-\lambda_\omega \omega'_j} s_{B-1} B}{P(H|\Theta) p_\omega e^{-\lambda_\omega(\omega_j + \omega_{j+1})} m_B(L - B) (\omega_j + \omega_{j+1})^2}
\]

(b) Choose an \( \omega \) block boundary uniformly at random and move to the left or right. The number of codons to move the boundary is controlled by an exponential distribution, with rate chosen by the user.

If the move crosses another boundary, it is automatically rejected.

\[
R = \frac{P(H|\Theta')}{P(H|\Theta)}
\]
(c) Choose a $\rho$ block uniformly to split or merge with a neighbor.

We split or merge rho blocks, with split to merge ratio

\[
\frac{s_B}{m_B} = \min\left(1, \frac{P(B + 1)/P(B)}{\min(1, P(B - 1)/P(B))}\right)
\]

For a split move, draw $U \sim \text{Uniform}[-1, 1]$, then let

\[
\rho'_j \rho'_{j+1} = \rho_j
\]

and

\[
\frac{\rho'_{j+1}}{\rho'_j} = \frac{1-U}{U}
\]

\[
R = \frac{P(H|\Theta') p_{\rho} e^{-\lambda_{\rho}(\rho'_j + \rho'_{j+1})} m_{B+1}(L - B + 1) (\rho'_j + \rho'_{j+1})^2}{P(H|\Theta) (1 - p_{\rho}) e^{-\lambda_{\rho}\rho_j} s_B(B + 1) \rho_j}
\]

For a merge move, let

\[
\rho'_j = \rho_j \rho_{j+1}
\]

then

\[
R = \frac{P(H|\Theta') (1 - p_{\rho}) e^{-\lambda_{\rho}\rho'_j} s_{B-1}B}{P(H|\Theta) p_{\rho} e^{-\lambda_{\rho}\rho_j} m_{B}(L - B) (\rho_j + \rho_{j+1})^2}
\]

(d) Choose a $\rho$ block boundary uniformly at random and move to the left or right. The number of codons to move the boundary is controlled by an exponential distribution, with rate chosen by the user.

If the move crosses another boundary, it is automatically rejected.
2. Protein change rate MCMC moves

(a) **Sample** $\omega_i$ **values for each block.**

Let $\omega_{\text{prop}} = \exp(u)\omega_{\text{cur}}$, where $u \sim \text{Uniform}[-1,1]$. We accept with probability

$$R = \frac{P(H|\Theta') e^{-\lambda_\omega(\omega_{\text{prop}})}}{P(H|\Theta) e^{-\lambda_\omega(\omega_{\text{cur}})}} \frac{\omega_{\text{prop}}}{\omega_{\text{cur}}}.$$ 

3. Recombination rate MCMC moves

(a) **Sample** $\rho_i$ **values for each block.**

Let $\rho_{\text{prop}} = \exp(u)\rho_{\text{cur}}$, where $u \sim \text{Uniform}[-1,1]$. We accept with probability

$$R = \frac{P(H|\Theta') e^{-\lambda_\rho(\rho_{\text{prop}})}}{P(H|\Theta) e^{-\lambda_\rho(\rho_{\text{cur}})}} \frac{\rho_{\text{prop}}}{\rho_{\text{cur}}}.$$ 

4. Transition/transversion ratio MCMC moves

(a) **Sample transition/transversion ratio** $\kappa$.

Let $\kappa_{\text{prop}} = \exp(u)\kappa_{\text{cur}}$, where $u \sim \text{Uniform}[-1,1]$. We accept with probability

$$R = \frac{P(H|\Theta') e^{-\lambda_\kappa(\kappa_{\text{prop}})}}{P(H|\Theta) e^{-\lambda_\kappa(\kappa_{\text{cur}})}} \frac{\kappa_{\text{prop}}}{\kappa_{\text{cur}}}.$$ 

Though a sophisticated approach to sequence evolution, OMEGAMAP is limited by several aspects. First, Wilson and McVean are only concerned with protein changes so they use a codon substitution model. If therapies do not operate on the codon
level, the model may not be appropriate. Second, they use a frequentist and con-
stant estimator of the mutation rate at the nucleotide level, which does not allow
for spatial heterogeneity. Third, the exponential distributions that serve as priors
for \( \omega \) and \( \rho \) parameters are limiting. Rather than a single parameter that controls
both the mean and variance of each block, it would be beneficial to have a more
robust distribution, such as a gamma or log-normal, where the relative values of the
mean and variance are not fixed. Finally, this model relies very directly on param-
eters set by the user. The authors take some care to show the relative robustness
of the model to different starting parameter values, however, they do not take full
advantage of the structure of a Bayesian hierarchical model, which can easily esti-
mate these parameters from the data. A more sophisticated hierarchical model will
also allow us to estimate differences between treatment and control populations.

OMEGAMAP can only be used to analyze a single population.

In the next chapter, we extend these previous approaches to a more sophisticated
novel spatial Bayesian hierarchical model. Our approach addresses the specialized
nature of evolutionary escape responses to therapy and improves on previous meth-
ods.
Chapter 3

PICOMAP: A Hierarchical Coalescent Model

Our model [Braunstein et al 2008] generalizes previous approaches in three distinct ways. First, we focus on spatial modeling of recombination and mutation rates directly, instead of the protein change focus of omegamap [Wilson and McVean 2006]. Second, our mutation model is nucleotide-based rather than codon-based. These two changes are critical for applications to antisense gene therapies, which operate on the nucleotide rather than codon/protein level. Finally, our model explicitly measures a spatial treatment effect for both recombination and mutation through a hierarchical prior structure, linking rates between treatment and control sequences.

Two components specify our model: (1) a coalescent model which defines the
likelihood equations $P(H|\Theta)$ and (2) a hierarchical prior distribution, $P(\Theta)$, that allows spatial variation in our mutation and recombination rates. The hierarchical nature of our priors are important as they explicitly link the mutation and recombination rates of blocks of sequence in the genome, allowing for a two-sample comparison between treatment and control. We implement our model using MCMC techniques, allowing estimation of the full posterior distribution.

### 3.1 Coalescent-based Likelihood

Unlike previous approaches, which have assumed either infinite sites or codon mutation models, we model mutation between the 5 symbols in our sequence alignments $\{A, C, G, T, -\}$. The nucleotide substitution model containing gaps [McGuire, Denham, Balding 2001], is initially studied in a phylogenetic context, where one calculates an emission matrix for mutation probabilities based on time. Our calculations are concerned only with the terminal nodes of the graphs, so we integrate out time to obtain the mutation probabilities between the 5 symbols. We approximate the likelihood as

$$p(H|\Theta) = p(H^T|\Theta^T)p(H^C|\Theta^C) \approx \text{PAC}(H^T|\Theta^T)\text{PAC}(H^C|\Theta^C)$$

(3.1.1)

where $H^T$ and $H^C$ are the treatment and control population sequences and $\Theta^T$ and $\Theta^C$ are the treatment and control parameters. For both sets of sequences and
parameters we let

$$PAC(H|\Theta) \equiv \hat{\pi}(h_1|\Theta)\hat{\pi}(h_2|h_1, \Theta) \cdots \hat{\pi}(h_k|h_{k-1}, \ldots, h_1, \Theta)$$

however, we define the conditional distributions differently than the previous literature. Similar to the approach of Li and Stephens, we utilize a Markov chain where the approximate conditional distribution $\hat{\pi}(h_k|h_{k-1}, \ldots, h_1, \Theta)$ is the probability of observing sequence $k$ given the first $k-1$ sequences and current set of parameters $\Theta$. The distributions $\hat{\pi}(h_k|h_{k-1}, \ldots, h_1, \Theta)$ can be calculated as

$$\hat{\pi}(h_k|h_{k-1}, \ldots, h_1, \Theta) = \sum_{x=1}^{S} \alpha_S(x)$$

where $S$ is the length of the sequence and the $\alpha$ terms can be computed recursively

where

$$\alpha_{j+1}(x) = \gamma_{j+1}(x) \left( p_j \alpha_j(x) + (1 - p_j) \frac{1}{k} \sum_{x'=1}^{k} \alpha_j(x') \right)$$

We let $p_j = \exp(-\rho_j/k)$ and define $\gamma_{j+1}(x)$, the transition probabilities between our 5 symbols as

$$\gamma_{j+1}(x) = p(x|y) = \left( 1 - \frac{k}{k + 2\beta_j} \right) \pi_x + \left( \frac{k}{k + 2(\alpha_j1[x \neq -] + \beta_j)} \right) 1[x = y] + \left( \frac{k}{k + 2\beta_j - \frac{k}{k + 2(\alpha_j + \beta_j)}} \right) \left( \frac{\pi_x}{\pi_x + \pi_y} \right) 1[(x, y) \in \{(A, G), (C, T)\}]$$

where $\pi_x$ and $\pi_y$ are the underlying probabilities of nucleotides $x$ and $y$. We define $\alpha_j$ and $\beta_j$ as follows:

$$\beta_j = \frac{\mu_j}{2\pi_R\pi_Y\kappa + 1 - \pi_R^2 - \pi_Y^2 - \pi_{DEL}^2}, \quad \alpha_j = \frac{\pi_R\pi_Y\kappa - (\pi_A\pi_G + \pi_C\pi_T)\beta_j}{\pi_R + \pi_C\pi_T}$$

(3.1.5)
and let $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$. Again, $\kappa$ represents the transition-transversion ratio.

Using equations 3.1.3, 3.1.4, and 3.1.5 we can calculate $\alpha_1(x)$ for $x = 1, \ldots, k$, and then calculate the values of $\alpha_2(x), \ldots, \alpha_S(x)$ recursively.

As before, the forward portion of the forward-backward algorithm [Rabiner 1989] is used to calculate the full approximation. The ordering of sequences in the PAC likelihood alters the value of the approximation, so we average the approximate likelihood over different orderings.

### 3.2 Hierarchical Prior for Spatially-varying Parameters $\Theta$

Drug and gene therapies target subsections of the genome, encouraging localized mutation and recombination as an escape response. To capture the local variation, we induce a prior that allows for spatially varying mutation and recombination rate parameters using a piecewise constant blocking structure.

Taking mutation rates as an example, each block has a central rate parameter $\mu_i$, around which we sample treatment and control rates $(\mu_i^T, \mu_i^C)$. Each of the central rate parameters $\mu_i$ are sampled around a grand central rate parameter $\mu_0$. There are $B_\mu$ blocks. The mutation structure and corresponding recombination structure is illustrated in Figure 3.1. An example blocking structure resulting from
a realization of the generative process outlined in Figure 3.1 is shown in Figure 3.2.

Figure 3.1: An illustration of mutation and recombination parameter hierarchy.

Figure 3.2: An example of PICOMAP mutation and recombination parameter blocking structure.

The hierarchical model and priors for our parameters are

1. Blocking Structure
(a) Number of mutation blocks \( B_\mu \sim Binomial(L - 2, p_\mu) \), where \( p_\mu \) is set by the user and \( L \) is the length of the sequence (number of nucleotides). This prior implies exchangeability over the sets of boundaries \((S_\mu)\) of identical size.

(b) Number of recombination blocks \( B_\rho \sim Binomial(L - 2, p_\rho) \), where \( p_\rho \) is set by the user. This prior implies exchangeability over the sets of boundaries \((S_\rho)\) of identical size.

2. Mutation Rates

(a) Mutation rates for blocks in the treatment and control groups:
\[
(\log \mu_1^C, \log \mu_1^T) \mid \mu_i, \sigma^2_\mu \overset{\text{iid}}{\sim} N(\log \mu_i, \sigma^2_\mu), \quad i = 1, \ldots, B_\mu
\]

(b) Central mutation rates for each of the blocks:
\[
\log \mu_i \mid \mu_0, \sigma^2_{\mu_0} \overset{\text{iid}}{\sim} N(\log \mu_0, \sigma^2_{\mu_0}), \quad i = 1, \ldots, B_\mu
\]

(c) We have a normal prior for \( \mu_0 \) and Wishart precision priors for \( \sigma^2_{\mu_0} \) and \( \sigma^2_\mu \).

3. Recombination Rates

(a) Recombination rates for blocks in the treatment and control groups:
\[
(\log \rho_1^C, \log \rho_1^T) \mid \rho_i, \sigma^2_\rho \overset{\text{iid}}{\sim} N(\log \rho_i, \sigma^2_\rho), \quad i = 1, \ldots, B_\rho
\]

(b) Central recombination rates for each of the blocks:
\[
\log \rho_i \mid \rho_0, \sigma^2_{\rho_0} \overset{\text{iid}}{\sim} N(\log \rho_0, \sigma^2_{\rho_0}), \quad i = 1, \ldots, B_\rho
\]
(c) We have a normal prior for $\rho_0$ and Wishart precision priors for $\sigma_{\rho_0}^2$ and $\sigma_{\rho}^2$. 

4. Transition/transversion ratio $\kappa \sim exp(\kappa_0)$, where $\kappa_0$ is set by the user.

The PAC likelihood and prior distributions on $\Theta$ jointly specify our model, which we call PICOMAP, for PIEcewise COntant MAP per of spatially varying recombination and mutation rates. Our model presents several important improvements over previous methodologies. In addition to addressing the three deficiencies of OMEGAMAP for measuring treatment effects in gene therapy studies, a codon-based mutation model, spatially constant mutation rates, and lack of a two-sample framework, PICOMAP also fixes the issue of a constant mean/variance ratio in rate estimates. Rather than setting this ratio at one, as implied by the exponential priors for the rate parameters in [Wilson and McVean 2006], our log-normal priors allow the model to infer this relationship.

### 3.3 MCMC Implementation

We estimate the posterior distribution of our parameter set with an MCMC implementation, which includes Gibbs, Metropolis-Hastings, and reversible jump moves. The moves for PICOMAP are enumerated in an identical manner to those of OMEGAMAP given in section 2.4 for ease of comparison. The calculations for each step are
1. Blocking MCMC moves

We split or merge blocks, with split to merge ratio

\[
\frac{s_B}{m_B} = \min\left(1, \frac{P(B + 1)/P(B)}{\min(1, P(B - 1)/P(B))}\right)
\]

where \(P(B)\) is the probability of \(B\) blocks under the binomial prior. We plug in \(B_\rho\) and \(B_\mu\) for recombination and mutation calculations, respectively.

(a) Choose a \(\mu\) block uniformly to split or merge with a neighbor

i. For a mutation split move, choose a \(\mu\) block uniformly to split or merge with a neighbor. We let \(\mu_j = \mu_j', \mu_{j+1}'\), and draw \(U \sim \text{Uniform}[0, 1]\) then 

\[
\frac{\mu_j' + 1}{\mu_j} = 1 - U \quad \text{and accept with probability:}
\]

\[
R = \frac{P(H|\Theta')} {P(H|\Theta)} \left(1 - p_\mu\right) m_B (B + 1)^r_1 r_2 r_3
\]

where:

\[
r_1 = \frac{(\mu_j' + 1)^2}{\mu_j} \frac{e^{-((\ln \mu_j' - \ln \mu_j)^2 + (\ln \mu_j + 1 - \ln \mu_j')^2)/2\sigma_{\mu}^2}} {e^{-((\ln \mu_j' - \ln \mu_j)^2 + (\ln \mu_j + 1 - \ln \mu_j')^2)/2\sigma_{\mu}^2}}
\]

\[
r_2 = \frac{(\mu_j' + 1)^2}{\mu_j} \frac{e^{-((\ln \mu_j' - \ln \mu_j)^2 + (\ln \mu_j + 1 - \ln \mu_j')^2)/2\sigma_{\mu}^2}} {e^{-((\ln \mu_j' - \ln \mu_j)^2 + (\ln \mu_j + 1 - \ln \mu_j')^2)/2\sigma_{\mu}^2}}
\]

\[
r_3 = \frac{(\mu_j' + 1)^2}{\mu_j} \frac{e^{-((\ln \mu_j' - \ln \mu_j)^2 + (\ln \mu_j + 1 - \ln \mu_j')^2)/2\sigma_{\mu}^2}} {e^{-((\ln \mu_j' - \ln \mu_j)^2 + (\ln \mu_j + 1 - \ln \mu_j')^2)/2\sigma_{\mu}^2}}
\]

ii. For a mutation merge move let \(\mu_j \mu_{j+1} = \mu_j'\)

We accept with probability

\[
R = \frac{P(H|\Theta')} {P(H|\Theta)} \left(1 - p_\mu\right) m_B (B - 1)^r_1 r_2 r_3
\]
\[ r_1' = \frac{\mu_{T,j}'}{\mu_{T,j} + \mu_{T,j+1}} e^{-\frac{(\ln \mu_{T,j}' - \ln \mu_{j}')^2}{2\sigma^2}} - \frac{\mu_{T,j} + 1}{\mu_{T,j}'} e^{-\frac{(\ln (\mu_{T,j} + 1) - \ln \mu_{j}')^2}{2\sigma^2}} \frac{\mu_{T,j} + 2\pi \sigma^2}{\mu_{T,j}' \sqrt{2\pi \sigma}} \]

\[ r_2' = \frac{\mu_{C,j}'}{\mu_{C,j} + \mu_{C,j+1}} e^{-\frac{(\ln \mu_{C,j}' - \ln \mu_{j}')^2}{2\sigma^2}} - \frac{\mu_{C,j} + 1}{\mu_{C,j}'} e^{-\frac{(\ln (\mu_{C,j} + 1) - \ln \mu_{j}')^2}{2\sigma^2}} \frac{\mu_{C,j} + 2\pi \sigma^2}{\mu_{C,j}' \sqrt{2\pi \sigma}} \]

\[ r_3' = \frac{\mu_{j}'}{(\mu_{j} + \mu_{j+1})^2} e^{-\frac{(\ln \mu_{j}' - \ln \mu_{0})^2}{2\sigma_0^2}} - \frac{\mu_{j} + 1}{\mu_{j}'} e^{-\frac{(\ln (\mu_{j} + 1) - \ln \mu_{0})^2}{2\sigma_0^2}} \frac{\mu_{j} + 2\pi \sigma_0^2}{\mu_{j}' \sqrt{2\pi \sigma_0}} \]

(b) Choose a \( \mu \) block boundary uniformly at random and move to the left or right - For the mutation block boundary move, we choose a \( \mu \) block boundary uniformly at random and move to the left or right. Let \((1, \mu_2^b, \ldots, \mu_{B^\mu}^b, L)\) be the current set of boundaries. We choose one of the \( B^\mu - 1 \) movable boundaries uniformly. We draw a value from a geometric distribution with mean 5.3, and move it left or right with equal probability. If the proposed move crosses another boundary or creates a block of size less than two it is automatically rejected. Otherwise, we accept with probability:

\[ R = \frac{P(H|\Theta')}{P(H|\Theta)} \frac{p(\mu|\Theta')}{p(\mu|\Theta)} \]

(c) Choose a \( \rho \) block uniformly to split or merge with a neighbor

i. For a recombination split move, choose a \( \rho \) block uniformly to split or merge with a neighbor. We let \( \rho_j = \rho_j' \rho_{j+1}' \), and draw \( U \sim \text{Uniform}[0, 1] \) then \( \frac{\rho_{j+1}'}{\rho_j'} = 1 - U \) and accept with probability:

\[ R = \frac{P(H|\Theta')}{P(H|\Theta)} \frac{p(\rho|\Theta')}{p(\rho|\Theta)} \frac{p_{\rho s_{B+1}}(L - B - 1)}{(1 - p_{\rho})m_B(B + 1)^2} r_1 r_2 r_3 \]
Choose a block of size less than two it is automatically rejected. Otherwise, we move it left or right with equal probability. If the proposed move crosses another boundary or creates a geometric distribution with mean 5.3, and move it left or right with equal probability. Let \( B \), \( \rho \) be the current set of boundaries. We choose \( \rho_j \rho_{j+1} = \rho_j^\prime \) and move to the left or right uniformly at random.

We accept with probability:

\[
R = \frac{P(H|\Theta^\prime)}{P(H|\Theta)} (1 - p_\rho) m_B - \frac{B}{p_\rho s_B (L - B)} r_1' r_2' r_3'
\]

where:

\[
r_1' = \frac{\rho_j^\prime \rho_{j+1}^\prime}{\rho_j^\rho_{j+1}} e^{-(\ln \rho_j^\prime - \ln \rho_j)^2/2\sigma^2} e^{-(\ln \rho_{j+1}^\prime - \ln \rho_{j+1})^2/2\sigma^2} \frac{\rho_j^\prime \rho_{j+1}^\prime 2\pi \sigma^2}{\rho_j \rho_{j+1} \sqrt{2\pi} \sigma}
\]

\[
r_2' = \frac{\rho_j^\prime \rho_{j+1}^\prime}{\rho_j^\rho_{j+1}} e^{-(\ln \rho_j^\prime - \ln \rho_j)^2/2\sigma^2} e^{-(\ln \rho_{j+1}^\prime - \ln \rho_{j+1})^2/2\sigma^2} \frac{\rho_j^\prime \rho_{j+1}^\prime 2\pi \sigma^2}{\rho_j \rho_{j+1} \sqrt{2\pi} \sigma}
\]

\[
r_3' = \frac{\rho_j^\prime \rho_{j+1}^\prime}{\rho_j^\rho_{j+1}} e^{-(\ln \rho_j^\prime - \ln \rho_j)^2/2\sigma^2} e^{-(\ln \rho_{j+1}^\prime - \ln \rho_{j+1})^2/2\sigma^2} \frac{\rho_j^\prime \rho_{j+1}^\prime 2\pi \sigma^2}{\rho_j \rho_{j+1} \sqrt{2\pi} \sigma}
\]

(d) Choose a \( \rho \) block boundary uniformly at random and move to the left or right - For the recombination block boundary move, we choose a \( \rho \) block boundary uniformly at random and move to the left or right. Let \((1, \rho_2^b, \ldots, \rho_{B_\rho}^b, L)\) be the current set of boundaries. We choose one of the \( B_\rho - 1 \) movable boundaries uniformly. We draw a value from a geometric distribution with mean 5.3, and move it left or right with equal probability. If the proposed move crosses another boundary or creates a block of size less than two it is automatically rejected. Otherwise, we...
accept with probability

$$R = \frac{p(H|\Theta')}{p(H|\Theta)}$$

2. Mutation rate MCMC movies

(a) **Sample treatment and control mutation rates** ($\mu_{i}^{T}, \mu_{i}^{C}$) for each block - Let ($\log \mu_{i,\text{cur}}^{T}, \log \mu_{i,\text{cur}}^{C}$) be the current values for the logarithm of ($\mu_{i}^{T}, \mu_{i}^{C}$). We sample proposal values $\log \mu_{i,\text{prop}}^{T}$ and $\log \mu_{i,\text{prop}}^{C}$ from

- $\log \mu_{i,\text{prop}}^{T} \sim \text{Normal}(\log \mu_{i,\text{cur}}^{T}, \tau_{2}^{2})$
- $\log \mu_{i,\text{prop}}^{C} \sim \text{Normal}(\log \mu_{i,\text{cur}}^{C}, \tau_{2}^{2})$

We accept new values with probability

$$R = \frac{p(H|\mu_{i,\text{prop}}^{T}, \mu_{i,\text{prop}}^{C})}{p(H|\mu_{i,\text{cur}}^{T}, \mu_{i,\text{cur}}^{C})} \cdot \frac{p(\mu_{i,\text{prop}}^{T}, \mu_{i,\text{prop}}^{C}|\mu_{i})}{p(\mu_{i,\text{cur}}^{T}, \mu_{i,\text{cur}}^{C}|\mu_{i})}$$

where:

$$\frac{p(\mu_{i,\text{prop}}^{T}, \mu_{i,\text{prop}}^{C}|\mu_{i})}{p(\mu_{i,\text{cur}}^{T}, \mu_{i,\text{cur}}^{C}|\mu_{i})} = \frac{\mu_{i,\text{cur}}^{T} \mu_{i,\text{cur}}^{C} e^{-(\log \mu_{i,\text{prop}}^{T} - \log \mu_{i,\text{cur}}^{T})^2 + (\log \mu_{i,\text{prop}}^{C} - \log \mu_{i,\text{cur}}^{C})^2/2\sigma_{2}^{2}}}{\mu_{i,\text{prop}}^{T} \mu_{i,\text{prop}}^{C} e^{-(\log \mu_{i,\text{prop}}^{T} - \log \mu_{i,\text{cur}}^{T})^2 + (\log \mu_{i,\text{prop}}^{C} - \log \mu_{i,\text{cur}}^{C})^2/2\sigma_{2}^{2}}}$$

The symmetry in our proposal distribution accounts for part of the cancellation of the MH acceptance probability.

(b) **Sample central mutation rate** $\mu_{i}$ for each block

$$\log \mu_{i}|\mu_{i}^{C}, \mu_{i}^{T}, \mu_{0}, \sigma_{\mu_{0}}^{2}, \sigma_{\mu}^{2} \sim \text{Normal} \left( \frac{2 \log \mu_{i}^{T} + \log \mu_{i}^{C} + B_{\mu} \log \mu_{0}}{\sigma_{\mu}^{2} + \frac{2}{\sigma_{\mu_{0}}^{2}} + \frac{B_{\mu}}{\sigma_{\mu_{0}}^{2}}}, \frac{1}{\sigma_{\mu}^{2} + \frac{2}{\sigma_{\mu_{0}}^{2}} + \frac{B_{\mu}}{\sigma_{\mu_{0}}^{2}}} \right)$$

(c) **Sample grand central mutation rate** $\mu_{0}$

$$\log \mu_{0}|\mu_{1}, \ldots, \mu_{B_{\mu}}, \sigma_{\mu_{0}}^{2} \sim \text{Normal}(\sum_{i=1}^{R} \log \mu_{i}/B_{\mu}, \sigma_{\mu_{0}}^{2}/B_{\mu})$$
(d) **Sample variance of treatment and control mutation rates** $\sigma^2_\mu$

$$(\sigma^2_\mu)^{-1}|\mu_1, \ldots, \mu_{B_\mu}, \mu^T_1, \ldots, \mu^T_{B_\mu}, \mu^C_1, \ldots, \mu^C_{B_\mu} \sim \text{Gamma} \left( B_\mu + 1, \left[ \frac{1}{2} \sum_{i=1}^{B_\mu} (\log \mu^T_i - \log \mu_i)^2 + (\log \mu^C_i - \log \mu_i)^2 \right]^{-1} \right)$$

(e) **Sample variance of central mutation rates** $\sigma^2_{\mu_0}$

$$(\sigma^2_{\mu_0})^{-1}|\mu_1, \ldots, \mu_{B_\mu}, \mu_0 \sim \text{Gamma} \left( \frac{B_\mu}{2} + 1, \left[ \frac{1}{2} \sum_{i=1}^{B_\mu} (\log \mu_i - \log \mu_0)^2 \right]^{-1} \right)$$

3. **Recombination rate MCMC moves**

(a) **Sample treatment and control recombination rates** $(\rho^T_i, \rho^C_i)$ **for each block** - Let $(\log \rho^T_{i,\text{cur}}, \log \rho^C_{i,\text{cur}})$ be the current values for the logarithm of $(\rho^T_i, \rho^C_i)$. We sample proposal values $\log \rho^T_{i,\text{prop}}$ and $\log \rho^C_{i,\text{prop}}$ from

- $\log \rho^T_{i,\text{prop}} \sim \text{Normal}(\log \rho^T_{i,\text{cur}}, \tau^2)$
- $\log \rho^C_{i,\text{prop}} \sim \text{Normal}(\log \rho^C_{i,\text{cur}}, \tau^2)$

We accept new values with probability

$$R = \frac{p(H|\rho^T_{i,\text{prop}}, \rho^C_{i,\text{prop}})}{p(H|\rho^T_{i,\text{cur}}, \rho^C_{i,\text{cur}})} \cdot \frac{p(\rho^T_{i,\text{prop}}, \rho^C_{i,\text{prop}}|\rho_i)}{p(\rho^T_{i,\text{cur}}, \rho^C_{i,\text{cur}}|\rho_i)}$$

where

$$\frac{p(\rho^T_{i,\text{prop}}, \rho^C_{i,\text{prop}}|\rho_i)}{p(\rho^T_{i,\text{cur}}, \rho^C_{i,\text{cur}}|\rho_i)} = \frac{\rho^T_{i,\text{cur}} \rho^C_{i,\text{cur}} e^{-((\log \rho^T_{i,\text{prop}} - \log \rho_i)^2 + (\log \rho^C_{i,\text{prop}} - \log \rho_i)^2)/2\sigma^2}}{\rho^T_{i,\text{prop}} \rho^C_{i,\text{cur}} e^{-((\log \rho^T_{i,\text{prop}} - \log \rho_i)^2 + (\log \rho^C_{i,\text{cur}} - \log \rho_i)^2)/2\sigma^2}}$$

The symmetry in our proposal distribution accounts for part of the cancel- cellation of the MH acceptance probability.
(b) Sample central recombination rate $\rho_i$ for each block
\[
\log \rho_i \mid \rho_i^C, \rho_i^T, \rho_0, \sigma_{\rho_0}^2, \sigma_{\rho}^2 \sim \text{Normal}\left(\frac{\frac{1}{2} \log \rho_i^T + \log \rho_i^C}{\frac{1}{2} \sigma_{\rho}^2 + \mu_{\rho}}, \frac{1}{\frac{1}{2} \sigma_{\rho}^2 + \mu_{\rho}}\right)
\]

(c) Sample grand central recombination rate $\rho_0$
\[
\log \rho_0 \mid \rho_1, \ldots, \rho_B, \sigma_{\rho_0}^2 \sim \text{Normal}(\sum_{i=1}^{B} \log \rho_i / B, \sigma_{\rho_0}^2 / B)
\]

(d) Sample variance of treatment and control recombination rates $\sigma_{\rho}^2$
\[
(\sigma_{\rho}^2)^{-1} \mid \rho_1, \ldots, \rho_B, \rho_1^T, \ldots, \rho_B^T, \rho_1^C, \ldots, \rho_B^C \sim
\Gamma\left(B_\rho + 1, \left[\frac{1}{2} \sum_{i=1}^{B_\rho} (\log \rho_i^T - \log \rho_i)^2 + (\log \rho_i^C - \log \rho_i)^2\right]^{-1}\right)
\]

(e) Sample variance of central recombination rates $\sigma_{\rho_0}^2$
\[
(\sigma_{\rho_0}^2)^{-1} \mid \rho_1, \ldots, \rho_B, \rho_0 \sim
\Gamma\left(B_\rho + 1, \left[\frac{1}{2} \sum_{i=1}^{B_\rho} (\log \rho_i - \log \rho_0)^2\right]^{-1}\right)
\]

4. Transition/transversion ratio MCMC moves

(a) Sample transition/transversion ratio $\kappa$

Let $\kappa_{\text{prop}} = \exp(u) \kappa_{\text{cur}}$, where $u \sim \text{Uniform}[-1, 1]$. We accept with probability
\[
R = \frac{P(H|\Theta') e^{-\lambda(\kappa_{\text{prop}})}}{P(H|\Theta) e^{-\lambda(\kappa_{\text{cur}})}} \frac{\kappa_{\text{prop}}}{\kappa_{\text{cur}}}
\]

There are several tuning parameters, including the variance of the proposals for $(\mu_i^T, \mu_i^C)$ and $(\rho_i^T, \rho_i^C)$ around their central rates $\mu_i$ and $\rho_i$, which we specify as $\tau^2$ and the prior probability for the binomial distribution on the number of mutation
and recombination blocks $p_\mu$ and $p_\rho$. We discuss selection of these parameters later in the dissertation in section 4.2 on monitoring convergence.

Parameter estimation in Bayesian hierarchical models is computationally intensive. PICO\textsc{map} is no exception. In addition to the computation normally associated with an increased number of parameters in a Bayesian framework, every likelihood calculation requires fitting an HMM as described in section 3.1. To ensure that our sampler runs in a reasonable amount of time, we implement it in object oriented C++. Each set of parameters, recombination $\rho$, mutation $\mu$, and transition/transversion ratio $\kappa$, have their own C++ objects, and we define functions on these objects for each step of the sampler enumerated earlier in the section. The software design provides tremendous flexibility for the user. Simply changing or commenting out a few of the more than five thousand lines of source code allows the user to modify steps, such as initializing their own non-random blocking structure, running PICO\textsc{map} without sampling the blocking structure, or fixing specific parameters.

Despite the sizable number of parameters and relatively large data sets, the overwhelming majority (>80%) of the computational cost comes from our likelihood evaluation. More details concerning the profiling of PICO\textsc{map} can be found at: http://stat.wharton.upenn.edu/~braunsf/profile.txt, which contains the output from gprof, the gnu C++ profiler.
This heavy computational load is offset by parallelization of computation, where we run several copies of our model at once. For each of the two applications we ran our model on 8 nodes of the Wharton Grid cluster. Later in the dissertation we discuss possible ways to reduce this computational load by further approximation of $P(H|\Theta)$. 
Chapter 4

Model Validation

Model validation will be addressed in two sections. In section 4.1 we implement and evaluate a set of simulations which demonstrate the success of the PICOMAP model. Our goals for the simulations are 1) recovery of true parameter values 2) coverage of true parameter values by HPD intervals in a high percentage of runs relative to the % of the intervals (i.e. 90% or 95% HPD intervals) 3) establish robustness of PICOMAP, in the sense that it is able to accomplish the first two goals under a variety of circumstances. In section 4.2, we explore convergence of the sampler and discuss tuning parameter selection.

4.1 Simulation Evaluation

The coalescent simulator MS [Hudson 2002] and SEQ-GEN [Rambaut and Grassly 1997] were used to generate sequences with spatially varying mutation and recom-
bination rates. We consider three simulation settings where our PICOMAP model is able to correctly recover these spatial patterns with very good HPD coverage.

1. We test the ability of PICOMAP to detect differences in spatial differences in $\mu$ without the presence of recombination ($\rho = 0$).
   
   **Treatment** - $\mu = 1.0 \in [1,200]$ and $[401,600]$, $\mu = 3.0 \in [201,400]$, $\rho = 0.0 \in [1,600]$.  
   
   **Control** - $\mu = 1.0 \in [1,600]$, $\rho = 0 \in [1,600]$.

2. We test the converse, the ability of PICOMAP to detect differences in recombination without differences in mutation ($\mu^T - \mu^C = 0$).
   
   **Treatment** - $\mu = 1.0 \in [1,600]$, $\rho = 3.0 \in [1,600]$.
   
   **Control** - $\mu = 1.0 \in [1,600]$, $\rho = 1.0 \in [1,600]$.

3. We vary both mutation and recombination between treatment and control sequences.
   
   **Treatment** - $\mu = 1.0 \in [1,200]$ and $[401,600]$, $\mu = 3.0 \in [201,400]$, $\rho = 1.0 \in [1,600]$.
   
   **Control** - $\mu = 1.0 \in [1,600]$, $\rho = 3.0 \in [1,600]$.

For each experiment replication we have 20 treatment and 20 control sequences, each of length 600. Twenty replications are run in each experiment. We illustrate the spatial treatment effects for each of our three simulation settings in Figure 4.1.
Figure 4.1: Spatial patterns of treatment effects in simulated datasets for experiments 1, 2, and 3.

For each of the three experiments, we examine the number of times at each nucleotide location our 95% HPD interval covers the true value. Under ideal conditions, we hope to cover the true values 95% of the time.

Figure 4.2: Coverage for mutation and recombination rates in experiment 1, with superimposed mutation treatment effect.
Figure 4.2 shows the coverage for simulation experiment 1. We expect our recombination coverage will be higher than our mutation coverage, as only the mutation treatment effect is not 0. Our recombination rate estimates cover the true value 89% of the time and the mutation rate estimates cover the true value 87% of the time. Two patterns in coverage emerge. First, we see dips in coverage for recombination around the boundaries of increased mutation. At these boundaries, the model must delineate between mutation or recombination and discover which is responsible for the sequence evolution. Consequently, these regions are the most difficult in terms of inference. Second, we observe an area of lower coverage, corresponding to the region of increased mutation in the treatment sample. Our model is being conservative. For each experiment, we see dips in coverage along the edges of the sequence, which is typical in biological data sets. Practitioners usually include extra sequence flanking the areas of interest to ensure proper coverage in the regions of interest. In each experiment, excluding 50 nucleotides on each end of the sequence, raises coverage for both mutation and recombination by approximately 1%.

Figure 4.3 shows the coverage for simulation experiment 2. We expect to see the opposite coverage pattern of experiment 1, as this experiment has a non-zero recombination treatment effect and no mutation treatment effect. In this case, the difference between our coverage for the two sets of parameters is much more pronounced. The recombination rate estimates cover the true value 80% of the time...
and the mutation rate estimates cover the true value 94% of the time. Our mutation rate coverage stays high, however, we see a dip in recombination rate. As previously mentioned, recombination is relatively difficult to estimate. The dip in coverage is explained by our model’s tendency to be conservative estimating differential rates and the difficulty of recombination estimation.

Figure 4.3: Coverage for mutation and recombination rates in experiment 2, with superimposed mutation treatment effect.

Figure 4.4 shows the coverage for simulation experiment 3. The recombination rate estimates cover the true value 87% of the time and the mutation rate estimates cover the true value 90% of the time. The final set of experiments is the most difficult, yet we see no degradation in overall coverage. We observe a dip in coverage in the region of increased mutation, as our model is being conservative in estimating differences in mutation. However, the interaction between mutation and recombination in this sample helps the coverage of recombination relative to experiment 2.
As a summary of the simulation experiments, the posterior means and medians are good estimators of the true parameter values, however, coverage is slightly less than 95%. The nature of the biological data and the relative difficulty of measuring recombination makes this fact unsurprising. Our model performs consistently well across the variables parameter patterns in all three experiments. Coverage is near 90% in almost all cases and is never below 80%. These results give us confidence in our estimates for our biological applications, where the rates are unknown.

### 4.2 Monitoring of Convergence

In this section, we discuss the details of the MCMC implementation, monitoring the sampler for convergence, thinning of samples, and selection of proper tuning parameters to ensure good mixing properties. Identical procedures were used for both the VIRxSYS antisense gene therapy and the enfuvirtide drug therapy, so
we only present pictorial evidence and discuss the convergence of the VIRxSYS sampler.

The hierarchical priors in **picomap** are designed to allow information sharing and explicitly link parameters spatially. This design imposes a high degree of autocorrelation in rates at the individual nucleotide locations. The phenomena is not unique to our model. Similar levels of parameter autocorrelation were also found in **omegamap**.

We tune $\tau^2$, the variance of the proposals for $(\rho^T_i, \rho^C_i)$ and $(\mu^T_i, \mu^C_i)$ around their current values. If $\tau^2$ is too small, the sampler will accept proposed values often, but could stay in a local maxima, preventing adequate exploration of the parameter space. If $\tau^2$ is too large, the sampler will rarely accept proposed values, leading to high autocorrelation, which requires gathering many more samples. Selection of good tuning parameter values ensures the sampler will adequately explore the parameter space, prevent it from getting stuck in local maxima, and help eliminate autocorrelation, simplifying the thinning of samples post-convergence. To achieve this goal, we choose a value with relatively high acceptance rates and good mixing properties. We test values in the range [.001,2] and display acceptance ratios and autocorrelation plots for three values, .05, .1, and .5.

For our first choice we see an acceptance ratio near 40%. Figure 4.5. plots the acceptance ratio for $\rho$ block moves as a function of the number of samples thinned by 100. Additionally, we observe very high autocorrelation in Figure 4.6; at several
nucleotide locations, the autocorrelation is not below the independence threshold at lag 500.

In Figure 4.7, with $\tau^2 = .1$, we see a lower acceptance ratio that it still reaches...
Figure 4.7: Acceptance ratio plot, $\tau^2 = .1$. 

Figure 4.8: Autocorrelation plots for $\tau^2 = .1$. 

35%. The autocorrelation in the variables decreases faster than before. In the corresponding autocorrelation plot, Figure 4.8, we find no variables with significant autocorrelation levels at lag 500. 

Our final choice of $\tau^2$, yields much lower acceptance rates, around 22%. However,
Figure 4.9: Acceptance ratio plot, $\tau^2 = .5$.

Figure 4.10: Autocorrelation plots for $\tau^2 = .5$.

Figure 4.10 shows drastically reduced levels of autocorrelation, relative to other values of $\tau^2$. Further increasing $\tau^2$ drastically reduces acceptance rates. They drop under 1% for the highest value in the range. Improvements to the autocorrelation plots also bottom out in the upper range of $\tau^2$ values. The rarity of jumps is the
culprit. We choose $\tau^2 = .1$, because it is the value with the highest acceptance rate that also eliminates autocorrelation at lag 500.

Other fixed parameters in PICOMAP are the binomial probabilities controlling the expected number of mutation and recombination blocks, $p_\mu$ and $p_\rho$. We initialize each randomly in a range implying between 30 and 60 blocks. The number of blocks, shown in Figure 4.11, is insensitive to the parameter values.

Figure 4.11: The top plot shows the value of the number of mutation blocks for each sampler at thinned iterations. The bottom plot shows the value of the number of recombination blocks for each sampler at thinned iterations. In both cases, each color represents a separate sampler.

Even with some variability between the number of blocks, the likelihood of each chain (Figure 4.12) finds nearly equal values. In the case of mutation rates, the
sampler finds the correct block locations and stays very constant, while the recombination rate blocking structure is noisier. Though we will wait until Chapter 5 to fully enumerate the results of our applications, mutation is the mechanism by which the wt-HIV escaped the VIRxSYS antisense gene therapy treatment, with no recombination blocks significantly different than 0. The clear mutation signal is very easy for the model to pick up, while the noisy and insignificant recombination signal leads the model to a less stable number of blocks.

Finally, we explore convergence of the sampler by plotting the log(likelihood) for each of 8 runs. In each case, the chain converges to an almost identical level. Figure 4.12 demonstrates the similarity of each chain.

![Figure 4.12: Plots of log(likelihood) for each of the 8 runs. On the left we have the thinned iterations. On the right we zoom into thinned iterations 10 to 35 to show the scale of the log(likelihood) differences. Each chain is a different color on the plots.](image)

The results of our simulation study chapter give us confidence applying PICOMAP
in real data applications, because of its success with regards to simulated data. In
the simulation studies, coverage is near 90% in almost all cases and is never below
80%. Several parameters of the MCMC algorithm are adjusted so the sampler will
have both high acceptance ratios and relatively low autocorrelation.
Chapter 5

Applications to HIV therapies

We apply our PICOMAP methodology to HIV sequence data from two different studies. In the first study, several HIV-infected patients are exposed to a drug therapy called enfuvirtide [Wild, Greenwell, and Matthews 1993]. In the second study, the HIV virus is exposed \textit{in vitro} to a novel antisense gene therapy developed by the VIRxSYS corporation [Lu, Yu, Binder, et al 2004]. In both cases, our analysis extracts biologically relevant features of the evolutionary response of HIV to these therapeutic challenges.

5.1 Enfuvirtide Drug Therapy

The enfuvirtide study consists of five patients. Blood samples are taken before and after treatment with the drug enfuvirtide, also known as fuzeon or T-20 [Wild, Greenwell, and Matthews 1993]. Sequences of the envelope (\textit{env}) region of the
HIV genome are generated from each of these blood samples. Pooling across these patients, we have 28 pre-exposure env sequences, which we label as the control sample, and 29 post-exposure env sequences, which we label as the treatment sample. Combining the sequences is necessary, because of the small sample sizes for several of the patients. The treatment effect of the therapy is quantified by calculating the posterior mean and 95% highest-posterior-density (HPD) intervals of the difference in recombination rates $\rho^T - \rho^C$ and mutation rates $\mu^T - \mu^C$ at each position of the genomic sequence.

The very existence of a post-exposure HIV population in these patients indicates that evolution of sequence changes occurred that must have conferred resistance to the action of the drug enfuvirtide. Resistance-conferring mutations are known a priori to occur at nucleotide locations 1639-1668 in the env sequence. The posterior estimate of the spatial treatment effect ($\mu_T - \mu_C$) for mutation on the env sequence is shown in Figure 5.1. Confirming our a priori knowledge, the entire 95% HPD interval of the mutation rate treatment effect is above zero for nucleotide positions 1590-1680, which suggests our model is able to detect elevated levels of mutation in the resistance-conferring region among sequences in the treatment sample.

Another observation from this study is that both the pre-exposure and post-exposure sequences are mixtures of several different HIV subtypes. Subtype identity is specified by the V3 loop subsequence of the env sequence, which corresponds to nucleotide positions 887-995. We observed that the same mutations were present in
Figure 5.1: Posterior estimate of the treatment effect of enfuvirtide drug therapy on mutation rates. The blue line is posterior mean, black lines are 95% highest-posterior-density (HPD) intervals, and the purple line is the posterior median. In the left plot we mark the location of drug resistant mutations with a bold red line.

different subtypes. Since it is unlikely that the same resistance-conferring mutations developed independently in each subtype, we suspect that the resistance-conferring mutations are passed between different subtypes via recombination, which is the primary means by which drug resistance is transferred \textit{in vivo} between strains of HIV. Recombination at these locations involving drug resistant strains may allow successful transfer of the resistance-conferring mutations between subtypes of HIV.

Figure 5.2 shows the spatial posterior estimate of the treatment effect $(\rho_T - \rho_C)$ on recombination. We see two areas of increased recombination, one from nucleotide positions 1020-1170 and another from nucleotide positions 1900-2200. In context of the mixture of HIV subtypes with drug resistant mutations, he first area of increased
recombination provides a putative location for the recombination for the transfer of drug resistant mutations between subtypes. It is not necessary that the 95% HPD is entirely above 0 for this region. Drug resistant mutations must occur in very a specific location of the genome, however, any recombination event between the V3 loop and the drug resistant mutations, which is approximately 600 nucleotides in length, will suffice. The result is a less concentrated, weaker signal. Additionally, mutations can be detected at specific locations, whereas recombination events occur between long stretches of sequence and are harder for the model to specifically
locate. This finding mirrors of our simulation study with regards to recombination signals.

5.2 VIRxSYS Antisense Gene Therapy

The VIRxSYS vector is a 937 bp antisense RNA gene therapy which directly targets the HIV genome sequence. Using BLAST [Altschul et al 1990] to align the antisense therapy with wild type HIV (a generic strain of HIV used in clinical tests), reveals the target portion of the sequence is nucleotides 1325-2249 of the env gene. Our VIRxSYS data set was generated in vitro, by exposing one set of wild type HIV to the antisense therapy and letting another set of wild type HIV evolve neutrally. The experimental design is shown in Figure 5.3.

Figure 5.3: Illustration of sequence generation process for VIRxSYS study [Lu, Yu, Binder, et al 2004]. The top sequences are allowed to evolve neutrally, while the bottom sequences have the antisense gene therapy.

The control sample consists of 19 env sequences from a control HIV population.
that was allowed to evolve neutrally in cell culture, along with 48 \textit{env} treatment sequences sampled from an HIV population evolving in cell cultures that were transfected with the VIRxSYS antisense vector treatment [Lu, Yu, Binder, et al 2004]. Unlike drug therapy treatments, whose effect can be nullified by just one or two well placed mutations, a relatively large number of mutations are required to escape the antisense gene therapy. We again quantify the treatment effect of exposure to the antisense vector by calculating the posterior mean and 95\% HPD interval of the difference in recombination rates $\rho^T - \rho^C$ and mutation rates $\mu^T - \mu^C$ at each position of the \textit{env} sequence.

Figures 5.4 and 5.5 show the posterior estimates of the treatment’s effect on mutation and recombination, respectively. The most striking feature of Figure 5.4 plots is the area of significantly elevated mutation in the treatment sequences. The leftmost region of the highest plateau corresponds to nucleotide position 1325, the 5’ boundary of the antisense target region. This area of heightened mutation overlaps with the target region for around 425 nucleotides in the 3’ direction.

We see fewer differences in the recombination rate (Figure 5.5), suggesting that mutation is the primary mechanism of evolutionary response to the antisense vector. The only putative observable pattern is slightly lower recombination rates in the target region of the treatment sequences relative to the control sequences. At all spatial points, the 95\% HPD for the treatment effect contains 0, implying a lack of significance for this difference.
Figure 5.4: Posterior estimates for the treatment effect of VIRxSYS antisense gene therapy on mutation rates. The blue line is posterior mean, and the black lines are 95% posterior intervals. The red line is the target region of the antisense therapy.

In both applications, our hierarchical Bayesian model has captured biologically relevant and meaningful signals. In the enfuvirtide study, our model locates drug-resistant mutations which are known to exist \textit{a priori}, and we identify a putative region for the recombination which transferred these mutations between HIV subtypes. In the VIRxSYS study, we successfully identify the location and magnitude of the mutational escape response, which significantly overlaps with the target region of the antisense gene therapy.
Figure 5.5: Posterior estimates for the treatment effect of VIRxSYS antisense gene therapy on recombination rates. The blue line is posterior mean, and the black lines are 95% posterior intervals. The red line is the target region of the antisense therapy.
Chapter 6

Comparison to Simpler Approaches

A direct comparison of omegamap [Wilson and McVean 2006] and picomap is not possible since omegamap lacks spatially varying mutations rates, the ability to measure treatment effects in a two-sample setting, and a nucleotide based mutation model. We consider two simpler comparisons that illustrate aspects of our model fit. First, we perform some detailed exploratory data analysis, which demonstrates the ability of our model to capture mutation patterns at segregating sites between treatment and control groups in the VIRxSYS data. Second, we use the BART and LASSO procedures to detect sequences features that correlate with the treatment vs. control sequence labels. Sequence features which are predictive of the treatment group memberships, are an alternate way of summarizing the sequence evolution of
HIV in response to therapies.

6.1 EDA

Segregating sites are nucleotide locations where at least one sequence has a different nucleotide than the other sequences. Figure 6.1 provides an example of segregating sites from a collection of sequences.

Figure 6.1: An illustration of segregating sites in a group of five sequences. The columns highlighted in blue are segregating sites. The nucleotide locations of the segregating sites are 5, 10, 18, and 19.

The use of segregating sites for estimation of global mutation rates is commonplace. The Watterson estimator \( \hat{W} \) [Watterson 1975] and Tajima’s D statistic - \( \hat{D} \) [Tajima 1989] are the most well known and frequently utilized [Hein et al 2005] estimators.
of this type.

\[ \hat{W} = \frac{\sum_{i=1}^{L} S_i}{\sum_{i=1}^{n-1} \frac{1}{i}}, \quad \hat{D} = \frac{\sum_{i\neq j} \pi_{ij}}{\sum_{i=1}^{n-1} \frac{1}{i}} \]  

(6.1.1)

\( S_i \) is the number of mutations at nucleotide site \( i \) over all sequences and \( \pi_{ij} \) is the number of pairwise differences between sequences \( i \) and \( j \). When scaled for sequence length, both are unbiased estimators of the global mutation rate \( \mu \). The global recombination rate \( \rho \) does not effect the expectation of these estimators, however, it does increase the variance [Hein et al 2005].

Figure 6.2: Fraction of treatment (blue) and control (red) sequences with mutations at each segregating site overlayed on VIRxSYS mutation posterior treatment effect estimates.

Figure 6.2 plots the segregating sites\(^1\) spatially overlaid on the spatial mutation plot (Figure 5.4) from our VIRxSYS results section. Areas with more segregating

\(^1\)One of our control sequences is identical to wild type HIV, so our segregating sites are equivalent to the set of sites at which at least one sequence differs from wild type HIV.
sites labeled blue correspond to regions with elevated levels of treatment sequence
mutation. An increased number of segregating sites also tends to imply larger
95% HPD intervals. The large spike at location 1618 is somewhat atypical: at
this location we observe a very large spike in sequences differing from wild type
HIV, but a relatively small bump in the treatment effect on mutation rate $\mu$. The
way we calculate the PAC likelihood for treatment sequences and control sequences
explains this phenomenon. A large majority of sequences have a common base at
this location, so this spike of nearly 1.0 is actually equivalent to a spike equidistant
from 0.0 from the viewpoint of the likelihood calculation. Since the model does not
see the initial state at each nucleotide, it cannot distinguish between this large spike
resulting from all but a few sequences mutating from intense selective pressure and
another nucleotide location at which only a few sequences differ from the initial
state. In both instances the model recognizes a handful of sequences differ from the
most common nucleotide.

Though we are able to see patterns between frequency of mutations at segre-
gating sites and our mutation treatment effect estimates, we do not have a way of
bridging the two. Although our exploration of segregating sites provides an inter-
esting first pass and could be used to identify putative locations for increased mu-
tation in the treatment sample, we need our full model-based approach to estimate
mutation and recombination rates as well as summarize uncertainty via posterior
intervals.
6.2 Regression Approaches

In this section, we use regression techniques to predict membership in the treatment or control group using the segregating site data. In the previous section, we are able to identify general regions with non-zero spatial treatment effects, but this simpler regression approach allows us to identify tiny regions or specific segregating sites that correlate with the treatment vs. control labels. There are 189 segregating sites, and at each location individual sequences either possess the wild type nucleotide or do not. We create a set of 189 indicator variables for each segregating site position which correspond to the presence or absence of the wild type nucleotide. We also create a set of 51 “pooled” variables consisting of the number of nucleotides differing from the wild type sequence in blocks of 50 nucleotides each. Since the number of predictors in the individual segregating site case is larger than our number of sequences, no solution exists when using standard regression. Instead, we focus on the BART [Chipman, George, McCulloch 2008] and LASSO [Tibshirani 1996] procedures which can provide estimates when the number of predictors is larger than the number of data points.

Bayesian Additive Regression Trees (BART) [Chipman, George, and McCulloch 2008] is a Bayesian sum of trees model, where each tree in the ensemble is constrained by a Bayesian tree prior to be a weak learner. It is similar to Random Forests [Breiman 2001] in that each regression tree uses relatively few variables and a bootstrap sample of the data to construct shallow tree classifiers. We use the
recent extension of the model for classification

\[ P[Y = 1|x] = \Phi[G(x)] \]  \hspace{1cm} (6.2.1)

where

\[ G(x) \equiv \sum_{j=1}^{m} g_j(x) \]  \hspace{1cm} (6.2.2)

\( \Phi[\cdot] \) is the standard normal CDF and the \( g_j(x) \) are individual regression trees.

Our primary interest is which covariates predict the treatment vs. control labels. Therefore, we summarize our BART models by keeping track of the number of times that each variable is included in a tree from our Bayesian ensemble.

The structure of BART ensures that variables correlating more strongly with the treatment label (our outcome variable) will appear in the ensemble more often than those that do not. We summarize the BART results in Figure 6.3, which is a spatial plot of the varcount proportion, the ratio of the number of times a particular variable is included in the ensemble to the total number of variables included in the ensemble. The top plot of Figure 6.3 summarizes the results from our data set of individual segregating sites. Regions where PICOMAP measures increased mutation, which is indicated by the red line at the top of the plot, tend to have higher variable counts. Though individual segregating sites in this region have lower than average counts, the majority of the highest variable counts exist within this region. The bottom plot summarizes the results from our “pooled” data set and exhibits a similar pattern.
Figure 6.3: Spatial plots of the variable counts from individual (top) segregating site data and BART for the pooled (bottom). The varcount proportion is the ratio of the number of times a particular variable is included in the ensemble to the total number of variables included in the ensemble. The red line in each plot corresponds to the area of increased mutation in Figure 5.4.
Region 33, which contains the segregating site at nucleotide position 1618, has the highest variable count.

In summary, BART successfully identifies the region of increased mutation found by our PICOMAP model for both the “pooled” and individual segregating sites data. The procedure successfully identifies segregating sites correlated with treatment vs. control sequence membership, but does not provide estimates of mutation or recombination rates.

The LASSO [Tibshirani 1996] is an alternative regression model that uses an $L_1$ penalty on regression coefficients for variable selection. Coefficients are standardized to prevent scale from influencing the results. In our case, the LASSO will assign non-zero coefficients to “pooled” regions or individual segregating site mutations that are most predictive of treatment vs. control labels. The LASSO model is specified as

$$Y = X\hat{\beta}_{\text{LASSO}} + \epsilon$$  \hspace{1cm} (6.2.3)

where

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left[ \sum(Y - X\hat{\beta}_{\text{LASSO}})^2 + \lambda \sum |\beta_i| \right]$$  \hspace{1cm} (6.2.4)

We view the influence of variables for the LASSO by varying the constant $\lambda$ which penalizes the sum of the absolute value of the coefficients. For large $\lambda$ we fit every player with the group mean. Decreasing the constraint to $\lambda = 0$ allows the LASSO solution to correspond to the OLS solution (when it exists) and otherwise a simple stepwise regression.
The identity of coefficients that enter the model and their magnitudes provide us with an excellent idea of the relative influence and importance of each dependent variable. Vertical lines in Figure 6.4 denote new variables entering the LASSO solution at a particular value of s on the x-axis. The colored lines track the values of the standardized regression coefficients for different values of the constraint s. At the right edge of each plot, the variable name for each of the colored lines is listed. As we can see by the identity of these variables, the LASSO variable selection plot echoes the conclusions of the BART analysis. In the left plot, we select region 33 first and it has the highest standardized coefficient at any given value of s, indicating it is the most influential. In the right plot, we select segregating site 1618 early and keep it as the only variable in the model for some time. From our overlay plot of the segregating sites (Figure 5.4), we expect the variables that correspond to nucleotide locations 600 - 1300 to be quite active. Were it not for the activity at location 1618, this would be the case.

The Lasso results are heavily influenced by the segregating site at location 1618 and region 33, which contains the segregating site 1618 in our “pooled” data. Though nucleotide location 1618 is within the region of increased mutation identified by PICOMAP, the estimated posterior mutation treatment effect at this location is low compared to other areas. The high proportion of treatment sequences with mutations at nucleotide location 1618 is the culprit. The LASSO solution is an optimization problem and lacks the flexibility of the BART procedure for exploration of
the parameter space and model averaging, so it quickly selects the variables related to segregating site 1618 in both the “pooled” and individual segregating site data which alone correctly classifies a large fraction of the data. The remaining effort of the LASSO procedure is placed on those handful of results incorrectly classified by variables related to segregating site 1618.

Though both procedures identify “pooled” or individual segregating sites corresponding to areas of increased posterior mutation treatment effect, BART is most successful. Unlike the LASSO, which is “fooled” by the behavior of segregating site 1618, the ensemble nature of the BART procedure is able to average over individual trees with variables including segregating site 1618 and provide a more holistic view of prediction of treatment vs. control labels that is more consistent with the posterior mutation treatment effect estimates of PICOMAP.

In this chapter we have investigated spatial patterns in segregating sites, and two regression models, BART and the LASSO, as alternatives to our methodology. Though each of these simpler methods provides us with information about sequence changes, they are no substitute for a full parametric model which provides parameter sharing, increased resolution, and inference about statistical significance via posterior interval.
Figure 6.4: LASSO variable selection plots for aggregated (left) and individual (right) segregating site results. The vertical black lines denote a variable entering the model. On the x-axis is $s$, the constraint on the absolute value of the regression coefficients. $s$ is scaled to $[0,1]$, with the upper boundary corresponding to the unconstrained solution.
Chapter 7

Other Applications of the Coalescent

In the main application of this dissertation, we measure parameters of ARGs and coalescent trees (recombination $\rho$ and mutation $\mu$), rather than the actual tree structure and branch lengths. The ARG structure is completely integrated out of PICOMAP. However, many applications are interested in the structure of ARGs and coalescent trees themselves. As an example, tree structures are often used for hierarchical clustering. Kingman’s coalescent [Kingman 1982] induces a prior distribution on these trees, but allows for limited interactions between the leaves of the trees which is often unrealistic. We describe an extension of the Kingman coalescent model, the coalescent with recombination and genealogies, which relaxes these unrealistic assumptions. After describing the generative process for the extension, we
discuss some theoretical aspects of the model and enumerate several MCMC moves for sampling tree genealogies. Our discussion includes three novel contributions:

1. we demonstrate how existing clustering applications are a special case of the coalescent with recombination and genealogies,

2. we generalize the likelihood calculations of a previous coalescent-based clustering model to a more general case,

3. we enumerate 3 MCMC moves on the coalescent with recombination and genealogies graphs.

Together, these contributions allow for a richer set of clustering applications that provide a more realistic model for interactions between individuals in a population.

### 7.1 A Generative Model for the Coalescent

A coalescent tree can be generated in the following way:

1. Start with $n$ sequences, so the current population size is $k = n$.

2. Simulate a coalescent time as exponential with rate $\frac{k(k-1)}{2}$.

3. Choose two sequences to coalesce uniformly. Decrease $k$ by 1.

4. Repeat steps 2 and 3 until $k = 1$, at which point you have reached the MRCA.
The likelihood \( p(k, k-1, \delta_i) \) of two particular sequences in the process coalescing in time \( \delta_i \), given \( k \) current sequences, can be calculated as

\[
p(k, k-1, \delta_i) = \left( \frac{k(k-1)}{2} \right) e^{-\left( \frac{k(k-1)}{2} \right) \delta_i} \frac{1}{k(k-1)} = e^{-\left( \frac{k(k-1)}{2} \right) \delta_i}.
\]

Equation 7.1.1 is best interpreted in two pieces, a coalescent time distributed exponentially with rate \( \left( \frac{k}{2} \right) \) and a constant accounting for the probability of choosing a particular pair of sequences out of \( k \) possibilities uniformly. We can then find the likelihood of the entire coalescent tree beginning from \( n \) sequences

\[
P(\pi_n) = \prod_{i=1}^{n} e^{-\left( \frac{k(k-1)}{2} \right) \delta_i}.
\]

This distribution is fully exchangeable over the order in which items coalesce. In the paper “Bayesian Agglomerative Clustering with coalescents” [Teh et al 2007], the authors use the coalescent as a prior distribution on trees and are interested in calculating a posterior distribution on different tree configurations with different coalescent times (branch lengths).

\[
P(\pi_n|X) = \frac{P(X|\pi_n)P(\pi_n)}{P(X)}
\]

They apply this method to the “modeling science” data set to find a tree structure on topics as well as using this method to cluster languages. In both of these applications, the basic tree structure of the coalescent is too simplistic. It is easy to imagine a scenario when one would create a hybrid topic or language, but that is not allowed by their approach. A more general model, the coalescent with recombination and genealogies allows nodes to interact in a way that allows this behavior.
7.2 The Coalescent with Recombination and Genealogies

The addition of recombination complicates considerably the generation of a coalescent. We describe this process below:

1. Start with \( n \) sequences, so the current population size is \( k = n \).

2. Simulate an exponential with rate \( \frac{k(k-1)}{2} + \frac{kp}{2} \).

3. With probability \( \frac{k-1}{k-1+\rho} \) we coalesce two sequences chosen uniformly and decrease \( k \) by 1. With probability \( \frac{\rho}{k-1+\rho} \) we recombine two sequences chosen uniformly and increase \( k \) by 1.

4. Repeat steps 2 and 3 until \( k = 1 \), at which point you have reached the MRCA.

The generative process is equivalent to a birth-death process with birth rate \( \frac{kp}{2} \), death rate \( \frac{k(k-1)}{2} \), and absorbing barrier at 1. Given \( k \), we have that the probability of a recombination event is

\[
p_{k,k+1} = \frac{\rho}{k - 1 + \rho} \quad (7.2.1)
\]

The probability of a coalescent event is

\[
p_{k,k-1} = \frac{k - 1}{k - 1 + \rho} \quad (7.2.2)
\]

The number of steps until the sequences coalesce to the MRCA can also be viewed as a random walk starting at \( n \), the initial number of sequences, with absorbing barrier
at 1 and transition probabilities as defined in equations 7.2.1 and 7.2.2 [Griffiths and Ethier 1990a]. Starting with $n$ sequences there are infinitely many ways to reach the MRCA. We can have $n - 1$ deaths and 0 births, $n$ deaths and 1 birth, $n + k - 1$ deaths and $k$ births, etc., which complicates the calculation and simulation of $P(\pi_n)$ considerably relative to the simple coalescent.

Let $R_n$ be a random variable for the number of recombination events before a coalescent with recombination and genealogies process reaches the MRCA (after starting with $n$ individuals). To provide intuition about the number of recombination events, we break up the calculation of $P(\pi_n)$ into two pieces: the conditional likelihood $P(\pi_n|R_n = k)$ of the graph starting with $n$ individuals given the number of recombination events in the sample’s history and the probability $P(R_n = k)$ of observing $k$ recombination events starting with $n$ individuals. Summing over all possibilities for $k$ we have

$$P(\pi_n) = \sum_{k=0}^{\infty} P(\pi_n|R_n = k)P(R_n = k) \quad (7.2.3)$$

The calculation of $P(R_n = k)$ is addressed first. Defining $\pi_{n,k} \equiv P(R_n = k)$, we have the recursive relationship

$$\pi_{n,k} = p_{n,n+1}\pi_{n+1,k-1} + p_{n,n-1}\pi_{n-1,k} \quad (7.2.4)$$

The first term of equation 7.2.4 is the scenario where a recombination event occurs, which increases the number of sequences from $n$ to $n + 1$ and then we observe the
compound event $\pi_{n+1,k-1}$, reaching the MRCA in $k - 1$ recombination events from $n + 1$ sequences. The second term is the scenario where a coalescent event occurs, leaving us with $n - 1$ sequences and then we observe the compound event $\pi_{n-1,k}$, reaching the MRCA with $k$ recombination events from $n - 1$ sequences.

Using the initial conditions $\pi_{1,0} = 1$ and $\pi_{i,-1} = 0$ for $i = 1, 2, 3, \ldots$, we can solve for all values of $\pi_{n,k}$ recursively. Plugging in $\rho = 0$ reduces expression 7.2.4 to the calculations for the simple coalescent. Figure 7.1 demonstrates the recursion for different values of $n$ and $\rho$.

The recursion for $\pi_{n,k}$ also allows us to solve for expectation of $R_n$,

$$E(R_n) = \sum_{i=0}^{\infty} k\pi_{n,k} = \sum_{i=0}^{\infty} k (p_{n,n+1}\pi_{n+1,k-1} + p_{n,n-1}\pi_{n-1,k})$$

$$= \sum_{i=0}^{\infty} kp_{n,n+1}\pi_{n+1,k-1} + \sum_{i=0}^{\infty} kp_{n,n-1}\pi_{n-1,k}$$

$$= \sum_{i=0}^{\infty} (k - 1)p_{n,n+1}\pi_{n+1,k-1} + \sum_{i=0}^{\infty} p_{n,n+1}\pi_{n+1,k-1}$$

$$+ \sum_{i=0}^{\infty} kp_{n,n-1}\pi_{n-1,k}$$

$$= p_{n,n+1}(E(R_{n+1}) + 1) + p_{n,n+1}E(R_{n-1})$$

This implies that

$$E(R_{n+1}) = \frac{n - 1 + \rho}{\rho} E(R_n) - \frac{n - 1}{\rho} E(R_{n-1}) - 1 \quad (7.2.5)$$

Letting $E(R_1) = 0$ and $E(R_2) = e^\rho - 1$, we find that for $n \geq 2$,

$$E(R_n) = \sum_{i=0}^{n-2} \frac{i!}{\rho^i} e^\rho - \sum_{i=0}^{n-2} \frac{1}{\rho^i} \sum_{k=i}^{n-2} \frac{k!}{(k - i)!} \quad (7.2.6)$$

84
Figure 7.1: Predicted (blue) and empirical (red) distributions of the number of recombination events before reaching the MRCA for several values of $n$, the beginning number of sequences, and $\rho$, the recombination rate. Each plot contains 100,000 simulated genealogies.

In Figure 7.2, we demonstrate the agreement of our simulated $R_n$ to equation 7.2.6 for set of parameters. Independently, Griffiths and Either investigated this model in [Griffiths and Either 1990b] and found an identical value. They also found an analytical solution
Figure 7.2: Plot of empirical (red) and predicted (blue) values of $E(R_n)$ for different values of $\rho$. Each plot has 100,000 simulated genealogies for each value on the x-axis.

for the variance

$$Var(R_n) = E(R_n) - (E(R_n))^2 + 2 \sum_{k=1}^{\infty} \rho^k \left( \sum_{m=0}^{n-2} \frac{m!}{(m+k)!} E(R_{m+k+2}) \right)$$  \hspace{1cm} (7.2.7)

Finally, we discuss the calculation $P(\pi_n|R_n = k)$ in equation 7.2.3. The calculation of $P(\pi_n|R_n = k)$ falls into two cases, either the $\pi_n$ has $k$ recombination events or it does not. In the latter, case $P(\pi_n|R_n = k) = 0$. In the former case, $\pi_n$ is a subset of possible graphs with $k$ recombination events. Letting $K$ be the number
of recombination events in $\pi_n$, equation 7.2.3 simplifies

$$P(\pi_n) = \sum_{k=0}^{\infty} P(\pi_n | R_n = k) P(R_n = k)$$

$$\quad = P(\pi_n | R_n = K) P(R_n = K) + \sum_{k\neq K} P(\pi_n | R_n = k) P(R_n = k)$$

$$\quad = \frac{P(\pi_n \cap (R_n = K))}{P(R_n = K)} P(R_n = K) + \sum_{k\neq K} 0 \times P(R_n = k)$$

$$\quad = P(\pi_n \cap (R_n = K))$$

Because $\pi_n \in (R_n = K)$, we have that $P(\pi_n \cap (R_n = K)) = P(\pi_n)$, verifying the equality. $P(R_n = k)$ drops out of the above equations, however, the number of recombination events play a large role in the unconditional calculation of $P(\pi_n)$.

In sections 7.3 and 7.4 we discuss the use of $\pi_n$ and $R_n$ in the calculation of the likelihood and as clustering structures.

### 7.3 Likelihood Calculations and Message Passing Algorithms

After describing the application of [Teh et al 2007], we discuss its generalization to address the more complicated structure of coalescent with recombination and genealogies graphs. In addition, we investigate the calculation of the probability of the data, given the coalescent tree or graph $P(H | \pi_n)$. Our goal is to estimate the posterior distribution $P(\pi_n | H) \propto P(H | \pi_n) P(\pi_n)$ for the richer set of graph structures generated by the coalescent with recombination and genealogies. Un-
fortunately, the likelihood calculations of [Teh et al 2007] do not generalize to the more general graph structure, so we discuss but do not implement two alternative approaches for calculation of $P(H|\pi_n)$. To facilitate the discussion, we show one of the trees from [Teh et al 2007] in Figure 7.3.

![Tree Diagram](image)

**Figure 7.3:** Example coalescent clustering tree from the application of [Teh et al 2007]. This tree clusters together languages to show how they are historically related.

The data set used to generate Figure 7.3 is a series of unordered binary indicator variables, which describes each language. The VIRxSYS and enfuvirtide data sets are nucleotide sequences, which can be modeled by ordered multinomial
vectors. Though Teh et al do address the multinomial case in their paper, they assume attributes (nucleotides) evolve independently, implying unordered data. All of their applications use unordered data, which is assumed by the coalescent with recombination and genealogies.

The likelihood of their data is calculated using message passing algorithms [Winn 2003], where parameters are estimated at the leaves of the coalescent tree and passed up the internal nodes. These parameter estimates are then used to cluster internal nodes and complete the coalescent tree structure. The calculations outlined in [Teh et al 2007] rely heavily on the structure of the problem, and parameter estimates of internal nodes depend only on nodes lower in the tree. No cycles exist in coalescent trees from the simple coalescent, which is not the case in the coalescent with recombination and genealogies. As a result, the normal message passing framework does not work for coalescent with recombination and genealogies graphs, significantly complicating the inference problem. There are both sampling and variational approaches for calculation of the conditional likelihood $P(H|\pi_n)$.

Sampling sequences at internal nodes reduces the complexity of likelihood calculations while introducing sampling variation. Starting with $n$ sequences there are at least $n - 1$ internal nodes, each of which must be simulated. Because internal nodes aggregate messages from lower in the tree, sampling error will grow as the tree approaches the MRCA. A successful implementation of this flavor requires derivation of optimal or near optimal proposal densities for importance sampling of
these structures. In addition to the computational burden of importance sampling, the procedure introduces yet another form of sampling variation and error into the problem.

Variational message passing algorithms [Winn and Bishop 2005] perform approximate inference on more complicated graph structures, but they are very computationally intensive. Rather than passing parameter estimates in one direction up the graph, this method allows for parameter estimates to be passed through nodes in any direction. The method is an iterative process, reaching a solution via convergence of either likelihood or estimated parameter values. Though interesting, these algorithms are tangential to the focus of this chapter. To that end, we do not discuss how they might be implemented and end our discussion of likelihood calculations at our previous quick summary of the main ideas and references to implementation specifics.

In summary, we have described two algorithms for calculation of $P(H|\pi_n)$ when $\pi_n$ is a graph generated by the coalescent with recombination and genealogies. In the next section, we discuss the calculation of $P(\pi_n)$ in the more general case, which will allow us to estimate the posterior distribution $P(\pi_n|H)$ when $\pi_n$ is a graph generated by the coalescent with recombination and genealogies.
7.4 MCMC Moves Using Coalescent with Recombination and Genealogies Graphs

The sample space of recombination with coalescent and genealogies graphs is large, making calculation of $P(\pi_n|H)$ for every graph impossible. Instead, we sample graphs to find either the maximum a posteriori (MAP) estimate or a collection with high posterior probabilities. Rather than sampling graphs entirely randomly, we hope to generate several good starting graphs for each chain and then explore the local space. To accomplish this, we can imagine altering a graph and the events composing it in three basic ways.

- add a recombination event
- delete a recombination event
- change the ordering of the sequence of events

The transition probabilities in the remainder of the section reference the definitions in equations 7.2.1 and 7.2.2.

Add a recombination event

When we observe the structure $p_{n,n+1}p_{n+1,n}$, we can add the events $p_{n+1,n+2}p_{n+2,n+1}$ or $p_{n+1,n}p_{n,n+1}$. To the structure $p_{n,n-1}p_{n-1,n}$, we can add the events $p_{n-1,n}p_{n,n-1}$ or $p_{n-1,n-2}p_{n-2,n-1}$. Denoting these four possibilities as $\pi_{n,1}^{\text{prop}}, \pi_{n,2}^{\text{prop}}, \pi_{n,3}^{\text{prop}},$ and $\pi_{n,4}^{\text{prop}}$.
then $\pi_n^{\text{orig}}$ as the probability of the tree we are proposing changes to, we have

$$\frac{\pi_{n,1}^{\text{propa}}}{\pi_n^{\text{orig}}} = p_{n+1,n+2}p_{n+2,n+1} \frac{P(H|\pi_{n,1}^{\text{propa}})}{P(H|\pi_n^{\text{orig}})}, \quad \frac{\pi_{n,2}^{\text{propa}}}{\pi_n^{\text{orig}}} = p_{n+1,n}p_{n,n+1} \frac{P(H|\pi_{n,2}^{\text{propa}})}{P(H|\pi_n^{\text{orig}})},$$

$$\frac{\pi_{n,3}^{\text{propa}}}{\pi_n^{\text{orig}}} = p_{n-1,n}p_{n,n-1} \frac{P(H|\pi_{n,3}^{\text{propa}})}{P(H|\pi_n^{\text{orig}})}, \quad \frac{\pi_{n,4}^{\text{propa}}}{\pi_n^{\text{orig}}} = p_{n-1,n-2}p_{n-2,n-1} \frac{P(H|\pi_{n,4}^{\text{propa}})}{P(H|\pi_n^{\text{orig}})}.$$  

**Delete a recombination event**

When we observe the structures $p_{n,n+1}p_{n+1,n}$ or $p_{n,n-1}p_{n-1,n}$, we can simply delete the pairs of events. Denoting these two possibilities as $\pi_n^{\text{propd}}$ and $\pi_n^{\text{propd}}$, we have

$$\frac{\pi_{n,1}^{\text{propd}}}{\pi_n^{\text{orig}}} = \frac{1}{p_{n,n+1}p_{n+1,n}} \frac{P(H|\pi_{n,1}^{\text{propd}})}{P(H|\pi_n^{\text{orig}})}, \quad \frac{\pi_{n,2}^{\text{propd}}}{\pi_n^{\text{orig}}} = \frac{1}{p_{n,n-1}p_{n-1,n}} \frac{P(H|\pi_{n,2}^{\text{propd}})}{P(H|\pi_n^{\text{orig}})}.$$  

**Change the ordering of a sequence of events**

Again we look for the structures $p_{n,n+1}p_{n+1,n}$ or $p_{n,n-1}p_{n-1,n}$, but we simply switch among the two in this case. Denoting these two possibilities as $\pi_n^{\text{props}}$ and $\pi_n^{\text{props}}$, we have

$$\frac{\pi_{n,1}^{\text{props}}}{\pi_n^{\text{orig}}} = \frac{p_{n,n-1}p_{n-1,n}}{p_{n,n+1}p_{n+1,n}} \frac{P(H|\pi_{n,1}^{\text{props}})}{P(H|\pi_n^{\text{orig}})}, \quad \frac{\pi_{n,2}^{\text{props}}}{\pi_n^{\text{orig}}} = \frac{p_{n,n+1}p_{n+1,n}}{p_{n,n-1}p_{n-1,n}} \frac{P(H|\pi_{n,2}^{\text{props}})}{P(H|\pi_n^{\text{orig}})}.$$  

We define only these three moves, because any other desired move can be constructed from the three. In each of the cases, one must check that the number of sequences does not hit the absorbing barrier, which is simple to implement.

An additional wrinkle is the consideration of branch lengths, which for clarity of presentation is suppressed in MCMC acceptance ratios earlier in the section.
Remembering the process by which we simulate this tree, the equivalent of equation 7.1.2, would be

\[ P(\pi_n) = \prod_{k=1}^{n+2R_n-1} \left( \frac{n_k(n_k - 1)}{2} + \frac{n_k\rho}{2} \right) e^{-\left( \frac{n_k(n_k - 1)}{2} + \frac{n_k\rho}{2} \right)\delta_k} \frac{1}{n_k(n_k-1)} p_{n_k, n_k+1} \] (7.4.1)

Letting \( n_k \) be the number of items before event \( k \), there are two major changes. First the rate of the exponential has been changed, adjusting for the two competing processes, recombination and coalescence. Second, we add the \( p_{n_k, n_k+1} \) term to accommodate the relative probabilities of recombination and coalescent events. This alters the calculations of the acceptance ratios for the MCMC moves outlined earlier in the chapter. Rather than write out each of the probabilities from each again, we simply overload the notation

\[ p_{k,k+1}(\delta_i) = \left( \frac{n_k(n_k - 1)}{2} + \frac{n_k\rho}{2} \right) e^{-\left( \frac{n_k(n_k - 1)}{2} + \frac{n_k\rho}{2} \right)\delta_i} \frac{1}{n_k(n_k-1)} \frac{\rho}{n_k - 1 + \rho} \] (7.4.2)

\[ p_{k,k-1}(\delta_i) = \left( \frac{n_k(n_k - 1)}{2} + \frac{n_k\rho}{2} \right) e^{-\left( \frac{n_k(n_k - 1)}{2} + \frac{n_k\rho}{2} \right)\delta_i} \frac{1}{n_k(n_k-1)} \frac{n_k - 1}{n_k - 1 + \rho} \] (7.4.3)

Plugging these expressions into the MCMC ratios and correctly accounts for branch lengths and yields the proper ratios. In summary, we have defined several MCMC moves on a graph structure, defined a prior on the structure, and detailed algorithms for calculation of its conditional likelihood, which completes our extension of the Bayesian hierarchical clustering model of [Teh et al 2007] to the coalescent with recombination and genealogies.
Chapter 8

Discussion and Future Work

This dissertation introduces a sophisticated, coalescent-based Bayesian hierarchical model for estimation of evolutionary escape response in a population exposed to therapeutic challenge. The escape response is quantified by estimation of mutation and recombination rate parameters. Our method allows for spatial heterogeneity in these rates, and it estimates differences between treatment and control sample parameters, with parameter values encouraged to be similar between the two populations except where the data suggests otherwise. We apply our procedure to sequence data from two different HIV therapy studies, a drug therapy and an antisense gene therapy.

In chapter 1 we summarize the replication cycle of HIV and other biological background. The summary serves two purposes: it allows to highlight the different mechanisms through which drug and antisense gene therapies attack the HIV
genome and it provides a basis for gauging the appropriateness of various assumptions when discussing models later in the dissertation.

In chapter 2, we discuss previous approaches to modeling evolution of sequences. The two main models for relating sequences are phylogenetic trees and the coalescent. We focus on two inference methods for the coalescent, the approximate likelihood methods of [Li and Stephens 2003], and another Bayesian hierarchical model \textsc{omegamap} [Wilson and McVean 2006].

The VIRxSYS data set is unlike most data sets of its type, because we know the ancestral sequence. Using this fact, it may be possible to create more efficient importance sampling schemes using the work of [Fearnhead and Donnelly 2001], conditioning on the MRCA, rather than integrating over all possibilities. Though interesting, these approaches are more focused on estimation of mutation and recombination rates for an entire ARG, rather than locally or spatially as in \textsc{picomap} or \textsc{omegamap}.

In chapter 3 we formally state our \textsc{picomap} model. After explaining our generalization of the approach of [Li and Stephens 2003] for calculation the conditional likelihood of our sequences $P(H|\Theta)$, we motivate our hierarchical priors $P(\Theta)$ which allow us to capture spatial treatment effects.

In chapter 4 we validate our model with an extensive battery of simulations and detailed discussion concerning the convergence of the sampler. We generate sequences under three experimental conditions. The first experiment varies only
mutation between treatment and control, the second varies only recombination, and the third varies both mutation and recombination. In most cases, coverage of true parameter values was near 90% and in no case did coverage dip below 80%. We also discuss the convergence of the sampler, tuning parameter selection, and general computational difficulties associated with model implementation. The reversible jump MCMC methods we use for estimation of the posterior distribution are computationally expensive. To that end, it would be interesting and useful to develop algorithms for approximate Bayesian inference via variational inference for the PICOMAP model. A variational approach is the most important future research direction.

In chapter 5 we apply the picomap methodology to a drug therapy, enfuvirtide, and VIRxSYS antisense gene therapy. In both cases we are able to identify biologically relevant features.

Chapter 6 compares our methodology to alternative simpler approaches. We discuss the use of mutation frequency at segregating sites and several frequentist estimators of the global mutation rate. After engaging in some thorough exploratory data analysis comparing spatial mutation frequency to our spatial estimates, we use the segregating site data in two regression procedures, BART and the LASSO, to predict treatment vs control sample membership. We are able to visually see how patterns in segregating sites relate to our estimates and our regression techniques are able to select important regions of the sequences. However, none of the simpler
approaches are able to provide estimates of mutation rates or Bayesian posterior intervals, which establish the significance of our results. This information is critical for any potential application to design of new therapies.

In chapter 7 we consider the coalescent in other research contexts. After explaining the use of coalescent trees for hierarchical clustering applications, we suggest an extension, the coalescent with recombination and genealogies, which is a model for a more general coalescent structure. We discuss how to generalize the message passing likelihood calculations of [Teh et al 2007] and generalize their MCMC moves to coalescent with recombination and genealogies clustering structures.

Although virological problems motivated the creation of our approach, our coalescent-based Bayesian hierarchical model applies more generally to two-sample data sets of nucleic acid sequences drawn from any population. The model is particularly relevant for populations in which the recombination rate is a substantial fraction of the mutation rate, since simpler models which ignore recombination can produce seriously misleading results [Schierup and Hein 2000]. Our model successfully identifies biologically relevant features in both data sets, and has excellent HPD coverage in simulation studies.
Bibliography


