UNIVERSITY OF CALIFORNIA

Los Angeles

Phenotypic Bayesian phylodynamics: hierarchical graph models, antigenic clustering and latent liabilities

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biomathematics

by

Gabriela Bettella Cybis

2014

Abstract of the Dissertation

Phenotypic Bayesian phylodynamics: hierarchical graph models, antigenic clustering and latent liabilities

by

Gabriela Bettella Cybis

Doctor of Philosophy in Biomathematics

University of California, Los Angeles, 2014

Professor Marc A. Suchard, Chair

Combining models for phenotypic and molecular evolution can lead to powerful inference tools. Under the flexible framework of Bayesian phylogenetics, I develop statistical methods to address phylodynamic problems in this intersection. First, I present a hierarchical phylogeographic method that combines information across multiple datasets to draw inference on a common geographical spread process. Each dataset represents a parallel realization of this geographic process on a different group of taxa, and the method shares information between these realizations through a hierarchical graph structure. Additionally, I develop a multivariate latent liability model for assessing phenotypic correlation among sets of traits, while controlling for shared evolutionary history. This method can efficiently estimate correlations between multiple continuous traits, binary traits and discrete traits with many ordered or unordered outcomes. Finally, I present a method that uses phylogenetic information to study the evolution of antigenic clusters in influenza. The method builds an antigenic cartography map informed by the assignment of each influenza strain to one of the antigenic clusters.

The dissertation of Gabriela Bettella Cybis is approved.

Christina M. Kitchen

James O. Lloyd-Smith

Kenneth L. Lange

Janet S. Sinsheimer

Marc A. Suchard, Committee Chair

University of California, Los Angeles

2014

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Marc A. Suchard for always having extremely insightful comments to my questions, be they work related or not. Thanks for teaching me so much about how good science should actually be made. I would also like to thank Janet S. Sinsheimer for always being so attentive and providing interesting perspectives on the big picture. Thank you both for being excellent and complementary mentors.

I would also like to thank the other members of my doctoral committee, Kenneth Lange, Jamie Lloyd-Smith and Christina Ramirez for helpful comments and suggestions regarding my work. In particular to Ken, for teaching an amazing course in stochastic processes, that is a foundation for so much of my work.

I would like thank my parents, for being the best role models one could hope for. To my mom, for always encouraging me in this journey, even though it meant I had to leave; and to my dad, for never failing to be my hero. And to Mariana, the sister I never had.

To Silvia R. Lopes, my first academic mentor, if you hadn't believed in me the way you did, I would never have made it to UCLA. To Phillippe Lemey and Trevor Bedford for being extremely helpful in providing the datasets so central to this work, and for answering my question about them.

To all the people that made my time in LA so special. To all the Biomath students with whom I shared the adventure of grad school. In particular, to Forrest Crawford, for being the best academic big brother; and to Kevin Keys for all the provoking conversations and endless supply of chocolate. I would also like to thank the other members of "Marc's lab" for commiseration and related philosophies on the challenges of academic life. To my Brazilian friends in LA, for being my family away from home, and to Ana Luisa Pessoa de Araujo for being the center of it. To Elias Bareinboim, for some crazy conversations about

science.

I would like to thank the Fulbright Science & Technology program, not only for funding my first three years at UCLA, but for creating incredibly motivating opportunities and for introducing me to the amazing group of grantees. I feel incredibly honoured to be a part of this community. My fourth year at UCLA was funded by the Institute for Society and Genetics, which also introduced me to a broader way of understanding science. I also received funding from the UCLA dissertation year fellowship. My stipend was also supplement by my advisor, for which I am grateful.

Vita

2010	MS, Biomathematics, UCLA
2009	MS, Mathematics, Area of concentration: probability and mathematical statistics, Federal University of Rio Grande do Sul (UFRGS)
2007	BS, Biology, Emphasis: cellular, molecular and functional biology, UFRGS
2007	BS, Biology, Emphasis: teaching, UFRGS
2013 - present	Assistant Professor, Department of Statistics, UFRGS
2012 - 2013	Teaching Assistant, Institute for Society and Genetics, UCLA
2011	Teaching Assistant, Department of Biomathematics, UCLA
2006	Intern, FK- Biotec, Brazil

Publications and Presentations

Cybis, G.B, Sinsheimer J.S, Lemey P, Suchard, M.A. (2013). "Graph hierarchies for phylogeography". Philosophical Transactions of the Royal Society B. 368, 20120206.

Cybis, G.B, Sinsheier J.S, Suchard, M.A (2012)."Family-Style Chinese Restaurant Process: Modeling Antigenic Evolution in Influenza". BSB & EBB Digital Proceedings, Campo Grande, MS, Brasil.

Cybis, G.B, Lopes, S.R.C, Pinhero, H. (2011). “Power of the Likelihood Ratio Test for models of DNA base substitution”. Journal of Applied Statistics. 38(12), 2723-2737.

Presentation, Bayesian Hierarchical Graph models for Phylogeography; Joint Statistical Meetings; San Diego. 2012

Presentation, 27 Colóquio Brasileiro de Matemática. Rio de Janeiro, Brazil. 2009

Presentation, 18o Simpósio Nacional de Probabilidade e Estatística. Estância de São Pedro, Brazil. 2009

Presentation; X Simpósio Nacional /Jornadas de Iniciação Científica do IMPA, Rio de Janeiro, Brazil. 2008

Presentations, XIX and XX Salão de Iniciação Científica UFRGS. Porto Alegre, Brazil, 2006 and 2007

CHAPTER 1

Introduction

Ever since Darwin, biologists have had interest in reconstructing the evolutionary relationships between groups of organisms. Haeckel's drawings and Darwin's notebook illustrations show early attempts to represent these relationships in the form of trees (Haeckel, 1866). In an era before molecular sequences were available, the first information used by systematists for such reconstructions were phenotypes. The first formalized methods for phylogenetic reconstructions from phenotypes were parsimony based approaches

When amino acid and later DNA sequences were made available by advances in molecular biology, molecular information became the most reliable data for phylogenetic reconstructions. The increased sample sizes, combined with computational improvements, paved the road for more reliable phylogenetic reconstructions using maximum likelihood and Bayesian methods (Felsenstein, 1981a; Rannala and Yang, 1996; Felsenstein, 2004).

Through advancing evolutionary modeling of molecular sequence data, it became possible to investigate how the evolution of a phenotypic trait relates to sequence evolution. In this dissertation, I explore this interface by developing models for phenotypic trait evolution and integrating them with Bayesian phylogenetic methods. In the projects presented here, the focus is on features such as antigenic properties of influenza, antibiotic resistance in bacteria and cell type of infection in HIV. The genetic data have a primary role in informing the evolutionary relationship between the organisms. These data are modeled using standard Bayesian phylogenetic approaches and should be viewed here as a covariate enabling better inference for the phenotypic traits.

Many of the datasets analysed in this dissertation come from rapidly evolving pathogens. These applications demonstrate rich evolutionary features and provide the biological significance that motivates the field of phylodynamics.

1.1 Phylodynamics

Diseases associated with viral infections, such as human immunodeficiency virus (HIV), hepatitis C virus (HCV) and dengue, kill millions of people every year; even seemingly benign influenza is alone responsible for 250,000 to 500,000 deaths annually (Stohr, 2002). Additionally, emergent infectious diseases are a constant threat to public health, as exemplified by the recent outbreak of swine-origin Influenza A that reached pandemic proportions in only a few months. Pheotypic traits such as those related to the interaction between the virus and host immune system and geographic distribution are important determinants of epidemic impact.

These infectious agents, primarily RNA viruses, consistently defy host immune systems and vaccination attempts, due mainly to their high mutation rates (Domingo and Holland, 1997). These rapidly evolving pathogens present a unique feature; their evolutionary dynamics occur on the same time scale as important ecological processes that determine patterns such as variability and distribution of epidemics. The burgeoning field of phylodynamics exploits these commensurate scales by integrating theoretical approaches from phylogenetics and epidemiological dynamics (Drummond et al, 2003; Grenfell et al, 2004).

Different features of epidemics have been examined trough phylodynamic approaches. The rapid mutation rates of these viruses allows for time calibration of evolution, by collecting samples at different time points (Rambaut, 2000; Drummond et al, 2002). Additionally, phylodynamics studies have used molecular sequence data to draw inference on variation in viral population sizes (Kuhner et al, 1998; Minin et al, 2008; Gill et al, 2013). Other studies have analyzed geographical spread of epidemics and viral transmission networks (Lewis, 2001; Lemey et al, 2009; Vrancken et al, in press).

1.2 Outline

This dissertation is composed of three individual projects connected by the common objective of creating new statistical methods to address phylodynamic problems at the intersection of sequence and phenotypic evolution. These projects are presented in chapters 3 through 5, and can be read as individual research articles.

All models developed in this dissertation are built under the methodological framework of Bayesian phylogenetics. For this reason, in chapter 2, I present a short overview of a basic Bayesian phylogenetic model used to estimate phylogenetic trees from molecular sequence data. Additionally, I briefly address inference through Markov chain Monte Carlo (MCMC), introducing a few concepts that are central to the efficiency of the inference techniques developed here.

Then, in chapter 3, I present a Bayesian hierarchical model for phylogeography. This project builds upon previous work that models geographical dispersion as a continuous time Markov chain along the phylogenetic tree (Lemey et al, 2009). To improve inference, the proposed hierarchical model combines information across many conditionally independent evolutionary processes that share similar geographic dispersion properties. The relevant information for these processes is summarized in a hierarchical level graph, in which the nodes represent geographical locations and edges connect the locations that are linked by migration. I present two data applications that exemplify the use of the method. They are also used to compare this method to alternative approaches for the same type of data. The work presented in chapter 3 has been published in the journal *Transactions of the royal society B* (Cybis et al, 2013), and is joint work with Janet S. Sinsheimer, Phillippe Lemey and Marc A. Suchard.

Chapter 4 presents the latent liability model for studying correlation between traits, while accounting for shared evolutionary history. The model is extremely flexible and can estimate correlation between continuous traits, discrete binary traits, discrete traits with multiple ordered or unordered states, and combinations thereof. The model considers the evolution of a continuous unobserved latent liability variable that determines the outcome of the observed traits of interest at the tips of the tree. The multivariate latent liabilities evolve along the phylogenetic tree through Brownian diffusion, and the covariance matrix of the diffusion process serves as a proxy for covariance among traits. In this chapter I discuss development of efficient MCMC transition kernels for this model and inference techniques for hypothesis testing. Additionally, I present applications of the method to Columbine flower morphology data, antibiotic resistance data in *Salmonella*, and epitope data in Influenza. This project is joint work with Janet S. Sinsheimer, Trevor Bedford, Alison Mather, Phillippe Lemey and Marc A. Suchard.

In chapter 5, I explore a nonparametric clustering method to investigate the intersection between genetic and antigenic evolution in influenza. Antigenicity in influenza can be visualized on antigenic maps, which are low dimensional representations of antigenic distances between viruses. In these maps, influenza strains naturally form clusters of similar antigenic properties. In this chapter I focus on the antigenic clusters, since they may have implications for vaccine design. The method employs a Bayesian multidimensional scaling model to create the antigenic map (Bedford et al, 2014), and adopts a novel nonparametric clustering prior on viral locations on the map. The clustering prior combines a modified version of a Chinese restaurant process mixture model and phylogenetics, using the tree structure to induce dependency in antigenic clustering. I present an application to an H1N1 influenza dataset in which the method produces estimates of an antigenic map with better resolved clusters as well as probabilities of cluster associations for individual viruses. This project is joint work with Janet S. Sinsheimer, Trevor Bedford, Andrew Rambaut, Phillippe Lemey and Marc A. Suchard.

Finally, in chapter 6, I discuss directions for future research arising from the latent liability model of chapter 4. As an additional avenue for future work, I introduce concepts for the correction of sampling bias in phylogenetic reconstructions based on single nucleotide polymorphism (SNP) data.

CHAPTER 2

Bayesian Phylogenetics

Throughout this dissertation, Bayesian phylogenetics is the backbone over which all new methodology is constructed. For this reason, I now present a short overview of these methods.

2.1 Phylogeneic models

Phylogenetic methods use sequence data $\mathbf{S}$ to estimate a phylogenetic tree F representing the evolutionary relationship between N organisms. The $N \times L$ sequence matrix $\mathbf{S} = \{s_{ij}\}$ contains N aligned DNA or RNA sequences of length L, originating from each of the organisms in the sample. The alignment process takes sequences of potentially different lengths and rearranges them by removing elements or inserting spaces ("-"), to create correspondence between sites in all sequences. All the entries in one column (site) of the aligned sequence matrix $\mathbf{S}$ are assumed to be homologous, that is, generated as one realization of the same evolutionary process on the tree. The alignment process is a statistical procedure susceptible to errors that may affect phylogenetic reconstruction. For this reason, methods that combine alignment and phylogenetic inference have been proposed (Redelings and Suchard, 2005, 2007). However, these methods can become computationally intensive, and for the purpose of this dissertation I adopt the widespread approach of disregarding potential errors in the alignment of $\mathbf{S}$.

A phylogenetic tree is an acyclic graph with N nodes of degree 1 representing the N

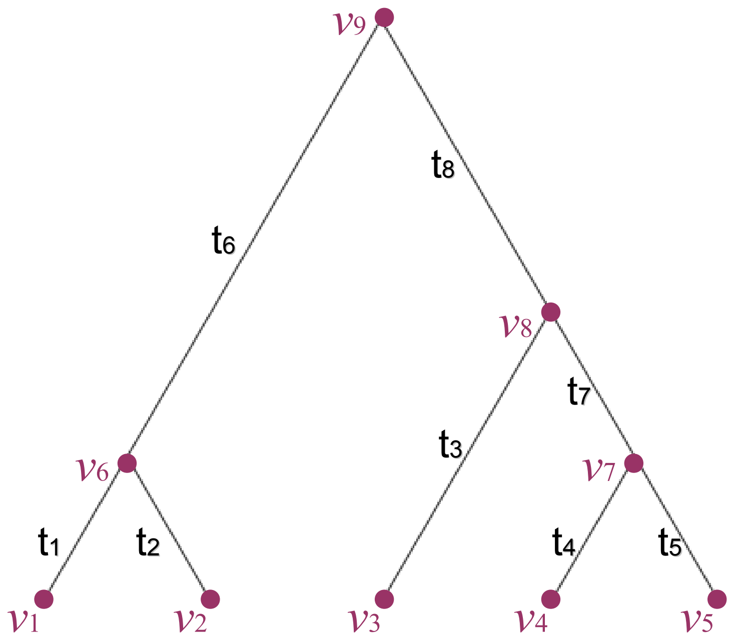


Figure 2.1: Example rooted tree with $N = 5$ tips.

organisms in the sample. These nodes are generally termed as tips, and denoted by $\nu_1, \cdots, \nu_N$. The tree also has $N - 2$ nodes of degree 3 called internal nodes, usually denoted by $\nu_{N+1}, \cdots, \nu_{2N-2}$. The internal nodes represent common ancestors to two or more organisms in the sample. Finally, the tree may have one node of degree 2 called the root and denoted by ν_{2N-1}. If this node exists, it represents the most recent common ancestor of all N organisms, and we say that the tree is rooted. The weights $\mathbf{t} = (t_1, \cdots, t_{2N-2})$ on the edges of a rooted tree represent elapsed evolutionary time between two nodes, and are generally referred to as branch lengths. When temporal information is available, trees can be calibrated, so that branch lengths represent physical time (Sanderson, 2002; Drummond et al, 2006). Figure 2.1 presents an example rooted tree with $N = 5$ tips.

In order to estimate the tree F from sequence data, we require a model for computing the probabilities of changes (mutations) in the molecular sequences over evolutionary time. For each site of the molecular sequence, this process is usually modelled as a continuous time Markov chain (CTMC), defined through the infinitesimal rate matrix **Q**. When the sequence matrix **S** is composed of nucleotide data, the CTMC has four different states {A,G,C,T/U}, corresponding to each of the four bases of DNA/RNA, and **Q** is a 4×4 matrix. The matrix $\mathbf{P}(t)$ of transition probabilities in time t can be obtained through matrix exponentiation as

$$\mathbf{P}(t) = \exp(t\mathbf{Q}). \tag{2.1}$$

Generally, the process is assumed to be reversible and have reached stationarity, although exceptions arise (Lemey et al, 2009). Biological knowledge about the base substitution process often guides parameterization of **Q**; Jukes and Cantor (1969) and Hasegawa et al (1985) provide two common examples.

The Markovian property of the base substitution process implies that, after two lineages split, their mutation processes are independent, given their most common recent ancestor. For the tree in figure 2.1, this means that

$$p(s_{1j}, s_{2j} | s_{6j}, t_1, t_2) = p(s_{1j} | s_{6j}, t_1) p(s_{2j} | s_{6j}, t_2), \tag{2.2}$$

where s_{ij} represents the base at site j of sequence i. The probability $p(s_{ij} | s_{i'j}, t_i)$ is obtained from from $\mathbf{P}(t)$. Propagation of this property throughout the tree leads to the tree likelihood for site j.

The tree likelihood $L(\mathbf{Q}, \mathbf{t}, F | \mathbf{S}) = p(\mathbf{S}_j | \mathbf{Q}, \mathbf{t}, F)$ computes the probability of the data, given the molecular evolution process on the tree. If we observed the sequences at the internal nodes of the tree, computing the likelihood for one site of the sequence would simply be a matter of multiplying the probabilities of mutational events for each branch. However, since this information is not available, we must integrate over all possible base combinations for the unobserved nodes of the tree. For the tree in figure 2.1, this yields

$$\begin{aligned} p(\mathbf{S}_j|\mathbf{Q},\mathbf{t},F) &= \sum_{s_{9j}}\sum_{s_{8j}}\sum_{s_{7j}}\sum_{s_{6j}} p(s_{9j})p(s_{6j}|s_{9j},t_6)p(s_{8j}|s_{9j},t_8)p(s_{7j}|s_{8j},t_7) \\ &\times\ p(s_{5j}|s_{7j},t_5)p(s_{4j}|s_{7j},t_4)p(s_{3j}|s_{8j},t_3)p(s_{2j}|s_{6j},t_2)p(s_{1j}|s_{6j},t_1), \end{aligned} \tag{2.3}$$

where the probability $p(s_{9j})$ of the root state is often obtained from the CTMC equilibrium distribution. Explicit dependency on the rate matrix $\mathbf{Q}$ was dropped from all probabilities on the right side of this equation for notational ease.

Naive evaluation of the expression (2.3) would be computationally prohibitive for large N, requiring the computation of all the 4^{N-1} terms in the sum. Computing this likelihood is made feasible by a pruning algorithm that traverses the tree in post order, keeping track of conditional probabilities, and evaluates the likelihood through $\mathcal{O}(N)$ operations (Felsenstein, 1981a) .

In order to obtain the likelihood for the whole matrix $\mathbf{S}$, one must assume a model for molecular evolution across sites. The simplest approach takes all sites to be independent and identically distributed, and computes the overall likelihood as

$$p(\mathbf{S}|\mathbf{Q},\mathbf{t},F) \propto \prod_{j=1}^{L} p(\mathbf{S}_j|\mathbf{Q},\mathbf{t},F). \tag{2.4}$$

Nonetheless, this independent model seems to be oversimplified in many cases. Considerable effort has gone into the development and testing of more elaborate models of molecular evolution across sites. These include, among others, the partitioning of the sites into classes with different substitution processes and models for the variation of substitution rates across the sequence (Yang, 2006).

2.2 Bayesian models in phylogenetics

Bayesian phylogenetic analyses draw inference based on the posterior probability $p(\boldsymbol{\theta}|\mathrm{S})$ of parameters $\boldsymbol{\theta}$, given the sequence data S. Here $\boldsymbol{\theta}$ collects all the phylogenetic parameters, such that for the simple model described in section 2.1 we have $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{t}, F\}$. Through Bayes theorem, the posterior can be computed as

$$p(\boldsymbol{\theta}|\mathrm{S}) = \frac{p(\mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{S})}, \tag{2.5}$$

where $p(\boldsymbol{\theta})$ is the prior distribution representing our beliefs *a priori* about $\boldsymbol{\theta}$ and the normalizing constant $p(\mathbf{S})$ is the marginal likelihood of the data S. The likelihood $L(\boldsymbol{\theta}|\mathrm{S}) = p(\mathrm{S}|\boldsymbol{\theta})$ of the molecular evolution process can be obtained through expression (2.4).

This Bayesian approach requires the definition of prior distributions for all parameters in $\boldsymbol{\theta}$, and different choices of priors are possible. Depending on the parametrization adopted for $\mathbf{Q}$, common choices for the parameters of the base substitution model are exponential and Dirichlet distributions. Prior distributions for the tree topology F and branch lengths $\mathbf{t}$ can come from models that generate tree-like structures, such as the birth-death process and the coalescent. An interesting direction arises from exploring the coalescent prior to study demographic dynamics (Kuhner et al, 1998; Minin et al, 2008).

To compute the posterior in (2.5), we would also need an expression for the normalizing constant $p(\mathbf{S})$, which can be computed as the integral

$$p(\mathbf{S}) = \int p(\mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}. \tag{2.6}$$

However, since $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{t}, F\}$, evaluating (2.6) requires integrating over the space of all possible tree topologies, as well as all possible branch length combinations and base substitution parameters. In general, there is no tractable solution for $p(\mathbf{S})$, and consequently Bayesian phylogenetic inference generally relies on Markov chain Monte Carlo (MCMC) (Sinsheimer et al, 1996; Rannala and Yang, 1996; Suchard et al, 2001).

2.3 Markov chain Monte Carlo

Monte Carlo integration is a simulation method for estimating multidimensional integrals. Suppose we wish to estimate the expected value of $h(\theta)$, then

$$E(h(\boldsymbol{\theta})|\mathbf{S}) = \int h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{S})\ \mathrm{d}\boldsymbol{\theta},$$

where $p(\boldsymbol{\theta}|\mathbf{S})$ is the posterior distribution of $\boldsymbol{\theta}$. If we cannot analytically evaluate the integral, random samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ from the distribution $p(\boldsymbol{\theta}|\mathbf{S})$ can be used to estimate $h(\boldsymbol{\theta})$ as

$$\widehat{h(\boldsymbol{\theta})} = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{\theta}^{(i)}).$$

The samples can also be used to obtain variance of the estimates and marginal distributions on individual components of $\boldsymbol{\theta}$.

However, for phylogenetic models it generally is not straightforward to generate samples from the distribution $p(\boldsymbol{\theta}|\mathbf{S})$. MCMC methods use Markov chains to generate dependent samples of the target distribution. These chains are constructed to be ergodic and have equilibrium distribution $p(\boldsymbol{\theta}|\mathbf{S})$. Consequently, the process is asymptotically guaranteed to achieve the target distribution.

The construction of ergodic Markov chains with the correct stationary distribution is central to MCMC. The two most used methods for producing these chains are the Metropolis-Hastings method (Metropolis et al, 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984). These can be thought of as recipes for constructing MCMC algorithms. Importantly, neither method requires the evaluation of the normalizing constant in expression (2.6) to generate samples from the posterior distribution.

Metropolis-Hastings algorithms rely on a two step procedure to generate consecutive posterior samples for $\boldsymbol{\theta}$. First a new state $\boldsymbol{\theta}^{\star}$ is proposed according to a proposal distribution $q_{\boldsymbol{\theta}^k}(\boldsymbol{\theta}^{\star})$, that usually depends on the current state $\boldsymbol{\theta}^{(k)}$. Then, the new state may be accepted $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{\star}$, with probability

$$A(\boldsymbol{\theta}^k, \boldsymbol{\theta}^{\star}) = \min \left\{ 1, \frac{q_{\boldsymbol{\theta}^{\star}}(\boldsymbol{\theta}^k) p(\boldsymbol{\theta}^{\star}|\mathrm{S})}{q_{\boldsymbol{\theta}^k}(\boldsymbol{\theta}^{\star}) p(\boldsymbol{\theta}^k|\mathrm{S})} \right\}, \tag{2.7}$$

or rejected $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)}$. Note that only the ratio of posterior probabilities is required for this evaluation.

Gibbs samplers divide the parameter $\boldsymbol{\theta}$ into M components $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M)$, and update each individual component $\boldsymbol{\theta}_i$ at a time. New samples for each $\boldsymbol{\theta}_i$ are drawn from the conditional distribution $p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \mathrm{S})$, where $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \cdots, \boldsymbol{\theta}_M)$ represents all other component parameters in $\boldsymbol{\theta}$. In these complex phylogenetic models, however, full conditional distributions are not available for all the parameter components. Metropolis-Hastings algorithms can be used to generate samples for individual parameter components for which a Gibbs sampler is not available. This approach produces a Metropolis-within-Gibbs sampler, in which some parameter components are updated based on full conditional probabilities, and others are updated using Metropolis-Hastings algorithms.

The phylogenetic methods presented in this dissertation exploit the flexibility of this Metropolis-within-Gibbs approach. I explore combinations of different transition kernels for the parameter components, with the goal speeding up convergence, and reducing autocorrelation between samples. A significant portion of this dissertation is dedicated to finding efficient transition kernels for parameters in the different models.

2.4 Combining models for molecular a phenotypic evolution

In this dissertation I present projects that combine models for molecular and phenotypic evolution. To jointly model these processes, I assume that the models are independent, conditional on the phylogenetic tree F. Thus, joint likelihoods can be computed as the product of the likelihoods for each evolutionary process. Additionally, MCMC transition kernels for the phylogenetic and molecular evolution parameters remain unchanged. This allows for the seamless combination of complex evolutionary models, and is one of the advantages of the Bayesian approach.

CHAPTER 3

Graph hierarchies for phylogeography[1]

Abstract. Bayesian phylogeographic methods simultaneouly integrate geographic and evolutionary modeling and have demonstrated value in assessing spatial spread patterns of measurably evolving organisms. We improve on existing phylogeographic methods by combining information from multiple phylogeographic datasets in a hierarchical setting. Consider N exchangeable datasets or strata consisting of viral sequences and locations, each evolving along its own phylogenetic tree and according to a conditionally independent geographic process. At the hierarchical level, a random graph summarizes the overall dispersion process by informing which migration rates between sampling locations are likely to be relevant in the strata. This approach provides an efficient and improved framework for analyzing inherently hierarchical datasets. We first examine the evolutionary history of multiple serotypes of dengue virus in the Americas to showcase our method. Additionally, we explore an application to intrahost HIV evolution across multiple patients.

[1] The research presented in this chapter has been published in the journal *Transactions of the royal society B* (Cybis et al, 2013), and is joint work with Janet S. Sinsheimer, Phillippe Lemey and Marc A. Suchard.

3.1 Introduction

Viral pathogens represent serious burdens on public health, as the current HIV-1 pandemic exemplifies, with an estimated 33.3 million people infected worldwide (UNAIDS, 2010). To reduce the impact of such pathogens, public health policies that aim to reduce viral dispersion and global circulation are of paramount importance. Detailed knowledge of the processes that govern epidemic evolution and spread is crucial for the determination of these policies. Recent phylogeographic studies explore the commensurate time scales of geographic dispersion and evolutionary processes to increase our understanding of the dispersal patterns of rapidly evolving pathogens (Wallace et al, 2007; Biek et al, 2007; Paraskevis et al, 2009).

Methods that integrate genetic and geographic analyses have been used to assess origins and dispersal patterns of many organisms of interest such as modern humans (Fagundes et al, 2007) and dogs (Pollinger et al, 2010). These phylogeographic studies largely benefit from the development of new methodologies (Bloomquist et al, 2010). An emerging methodological approach is the introduction of spatial diffusion processes in both discrete (Lemey et al, 2009) and continuous space (Lemey et al, 2010) to Bayesian phylogenetic analysis. This introduction allows for simultaneous reconstructions of geographic spread history, estimation of clade geographical origins and characterization of the dispersion process (Bloomquist et al, 2010; Faria et al, 2011).

When we assume discrete geographical locations, a key feature for characterizing dispersion processes in these models becomes the *migration graph*. In this graph, the vertices represent the geographic states of the model, and an edge connects two vertices only if the instantaneous migration rate between these vertex locations is nonzero. To infer the nonzero rates, a variable selection procedure controls the total number of edges to avoid overfitting. Thus, the migration graph contains the information of which migration rates are relevant for describing the overall dispersion process.

One troubling aspect of phylogeographic inference centers on the observation that we generally only have one realization of the geographic evolutionary process that produced the spatial distribution of the group under study. Thus, even for moderately small numbers of geographic states, most transitions between locations might not even be sampled. Consequently, when there is a large number of locations, phylogeographic analyses frequently have to control for poor estimates for the migration rates.

Ideally we would like to have many realizations under the same geographic process in order to improve inference. In some cases, parallel geographic realizations are available for this type of analysis. An example of such data structure that we explore in this paper concerns the four different serotypes of dengue virus in the Americas. Although the different serotypes do not have the same evolutionary history and probably arrived in the Americas at different times and in different locations, they share the same vector, host, mode of transmission and most other aspects of viral biology and ecological niches (Halstead, 2008). Thus, we reasonably assume that the factors that govern their dispersion processes, and in turn, their migration graphs are similar. Consequently, information obtained from one serotype should improve inference on the phylogeography of the other serotypes of dengue if we allow for occasional differences.

Previous studies have explored two different strategies when dealing with these parallel data structures. One alternative is to treat each phylogeographic dataset or strata independently, as Allicock et al (2012) explore for the four serotypes of dengue. This approach allows for comparisons between the different serotypes and can identify discrepancies in the migration process. Nevertheless, there is no structure for sharing information between the four separate analyses.

Alternatively, Sanmartín et al (2008), in a similar data analysis, model all strata jointly by imposing the same migration graph and rates to all strata of the analysis. This method effectively combines information from all strata to improve rate estimates. But in doing so, it disregards any stratum-specific peculiarities, and does not provide an opportunity to distinguish the processes of different strata. One reasonably assumes that inherent variability or systematic differences of scientific interest may introduce small differences in the processes.

As a middle-ground to these two extreme alternatives, we propose to model these inherently parallel datasets through a Bayesian hierarchical model. The hierarchical structure of our model effectively represents the overall dispersion process and allows for sharing information between the strata. We allow each individual stratum small variation from the hierarchical level; and large deviations from the expected pattern naturally assesses discrepancies between the strata. To accomplish this task, we introduce a novel hierarchical migration graph that summarizes the geographic dispersion process over all strata and informs which are the predominant migration rates for the individual strata. Figure 3.1 presents a schematic representation of the structure of the model.

Hierarchical models have found success in Bayesian phylogenetics for modeling the molecular sequence substitution processes within multipartite data, in contexts where overall properties of the data are of interest (Suchard et al, 2003). These hierarchical phylogenetic models (HPMs) generally have the property of reducing variability in estimates for phylogenetic parameters of individual partitions while providing a framework for assessing overall tendencies. Examples of HPMs involving the sequence substitution processes in infectious diseases include examining selective pressures in HIV intrahost data, where information pools across multiple patients (Edo-Matas et al, 2011), and estimating the time to most recent common ancestor across the multiple gene segments of influenza (Tom et al, 2010).

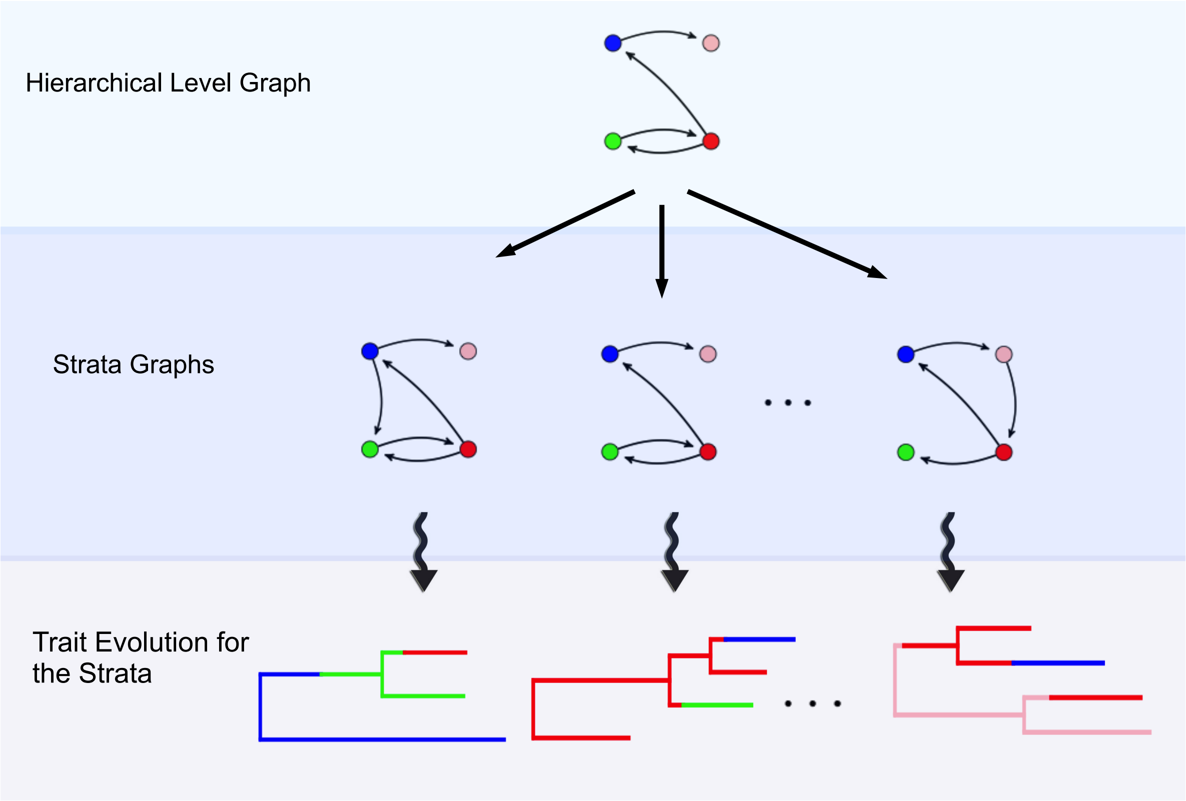


Figure 3.1: Schematic representation of the hierarchical structure of the model. At the hierarchical level, the hierarchical migration graph summarizes the overall process. For each of the parallel stratum, an analogous graph indicates the positive rates that define trait evolution along the tree for the stratum.

After presenting the details of our novel geographic HPM in section 3.2, we analyze two datasets that showcase our method. In section 3.3 we first present the joint analysis of the dataset of dengue virus in the Americas mentioned above for comparison with the previous study (Allicock et al, 2012). Although our method draws inspiration from purely geographic problems, the model is composed of Markov chains describing the evolution of any discrete trait with a single or small number of realizations per strata along phylogenetic trees. To demonstrate its application to other traits, in Section 3.3, we also present the analysis of intrahost HIV data from 14 different patients. Here, the discrete trait records the different cell compartments from which the virus was isolated, and each patient represents one parallel realization of the intrahost trait process. Finally, in Section 3.4, we conclude with some features arising from the methodology and examples.

3.2 Methods

We present a hierarchical model for the evolution of a discrete trait with a single (or small of number of) realization(s) in N similar strata. This trait typically tracks geographic location of sampling, but can also represent any discrete character with a fixed number of states, evolving through a homogeneous Markovian process. Each stratum represents a conditionally independent evolutionary history for the same trait. We assume sufficient similarity between the strata for a joint modeling approach. Our model has two levels. At the stratum level, each stratum possesses its own evolutionary process; the hierarchical level shares information across strata.

We have both sequence and trait information for each sample of each stratum. For each individual stratum, we model the sequence data through a phylogenetic tree, with sequence evolution and demographic parameters coming from standard Bayesian phylogenetic methods (Drummond et al, 2012).

We denote the trait data for each stratum i by $\mathbf{X}_i = \{X_{i1}, \ldots, X_{in_i}\}$, where the number of samples n_i may vary by strata. For each stratum we assume that a phylogenetic tree relates the samples, where the state at the root $X_{i,\text{root}}$ and internal nodes $\{X_{i,n_i+1}, \ldots, X_{i,2n_i-2}\}$ are not observed. Following the approach of Lemey et al (2009), we model trait evolution on the tree as a continuous time Markov chain (CTMC) with infinitesimal rate matrix $\mathbf{\Lambda}_i = \{\lambda_{ijk}\}$. This $K \times K$ rate matrix for stratum i is parametrized as

$$\mathbf{\Lambda}_i = \mu_i \ \mathbf{S}_i \mathbf{\Pi}_i, \tag{3.1}$$

where $\mathbf{\Pi}_i = \text{diag}(\pi_{i1}, \ldots, \pi_{iK})$ is a diagonal matrix with equilibrium frequencies for each state, μ_i is a scalar overall transition rate, and $\mathbf{S}_i = \{s_{ijk}\}$ is a $K \times K$ matrix normalized to give overall transition rate 1. When the matrix $\mathbf{S}_i$ is taken to be symmetric, then the CTMC is time reversible.

For traits with a large number of sampling states K, as is commonly the case for geography, the number of rates $K(K-1)$ is large. Since each sequence only has one sampling location, we expect *a priori* that many of the possible transitions will be very unlikely, rendering a sparse matrix $\mathbf{\Lambda}_i$. Thus we adopt the approach of Lemey et al (2009) and employ Bayesian Stochastic Search Variable Selection (BSSVS) to select a parsimonious interpretation of $\mathbf{\Lambda}_i$. This approach is instrumental in dealing with the high variances associated with this type of inference.

In BSSVS the model is augmented with indicator variables δ_{ijk}. Each indicator is placed on one directed edge of the graph connecting the states of the CTMC. When $\delta_{ijk} = 0$ the infinitesimal rate between states j and k is zero. When $\delta_{ijk} = 1$ the rate from state j to k is λ_{ijk}.

At the hierarchical level our model is composed of a migration graph, whose nodes are the sampling states of the process. A directed edge from node j to k on this hierarchical graph represents a nonzero infinitesimal rate from state j to k, and is present when the indicator $\delta_{Hjk} = 1$. The hierarchical graph is a representation of the overall evolutionary process of the trait across strata. It highlights which transitions are dominant in the model. By introducing this hierarchical level, we create a structure through which information is shared across the different strata.

Individual strata are allowed to diverge from the hierarchical graph to account for inherent variability. For each stratum, the number of differences between the hierarchical graph and the one induced by the BSSVS follows a binomial distribution

$$\sum_{j \neq k} |\delta_{ijk} - \delta_{Hjk}| \sim \text{Binomial}(\nu, p), \tag{3.2}$$

where p is a fixed error parameter, and $\nu = K(K-1)$ or $K(K-1)/2$ depending on whether the process is assumed time reversible. The binomial distribution is chosen mainly for the convenience of independence between edges of the graph. An alternative option that favors the hierarchical and stratum graphs being identical and introduces dependency between edge differences is the geometric distribution.

3.2.1 Priors

We use standard noninformative prior choices for the parameters of the infinitesimal rate matrices at the stratum level, following the suggestions of Lemey et al (2009). The overall rate parameter μ_i is taken to be Exponential(1), and the unnormalized elements of $\mathbf{S}_i$ are also assumed to be independent Exponential(1). When extra information is available, alternative prior specifications for these parameters are possible without affecting the hierarchical structure of the model. Informative priors for the transition rates can be based on geographic distances, population size, or even air traffic data. The prior distribution on the indicator variables of the BSSVS procedure is given by the hierarchical level of the model.

We must also set prior distributions on the hyper parameters of the hierarchical level graph. We assume *a priori* that each directed edge is included in the model according to a Bernoulli random variable with small success probability χ. The sum of these independent random variables $\sum \delta_{Hjk}$ is binomially distributed. In the limit when $\chi << K(K-1)$ this distribution is approximately a Poisson distribution with expected number of edges $K(K-1)\chi$.

3.2.2 Inference

Inference on this model is made by a Markov chain Monte Carlo (MCMC) procedure, where each parameter of the model is updated in turn to generate a Markov chain whose limiting distribution is the posterior. This is done in accordance with standard Bayesian phylogeographic methods (Lemey et al, 2009).

One note on this procedure is that independent updates of the edge indicator variables from the hierarchical level and all the strata may lead to poor mixing and slow convergence. This is especially the case when the number of states is high. An alternative to independent updates is to jointly update all the indicator variables by adopting a Metropolis-Hastings step with a proposal distribution that updates one edge of the graph in all strata simultaneously. Updating multiple edges at a time also improves the mixing of the chain.

The method described in this paper has been incorporated into the software package BEAST-Bayesian Evolutionary Analysis Sampling Trees (Drummond et al, 2012).

3.2.3 Bayes Factors

To verify support for a particular edge in the migration graph being included in the model, we use the Bayes factor. The Bayes factor measures how the data change the support for edge $\{jk\}$ being included in the graph relative to the change in support for it being excluded. Formally, the Bayes factor is defined as the ratio of the marginal likelihood of a model and the marginal likelihood of the alternative. For graph i, it can be computed simply as the ratio between posterior and prior odds

$$\mathrm{BF}_{ijk} = \frac{\text{Posterior Odds}}{\text{Prior Odds}} = \frac{p_{ijk}}{1-p_{ijk}} \Big/ \frac{q_{ijk}}{1-q_{ijk}}. \tag{3.3}$$

The posterior probability of the edge $\{jk\}$ is the posterior mean of the indicator δ_{ijk}. For the hierarchical graph, the prior q_{Hjk} is obtained from the Poisson distribution. For the stratum graphs, the prior probability of edge $\{jk\}$ being included in the model depends on the distribution of edge differences between hierarchical and stratum graphs. If we adopt the Binomial(ν, p) distribution for these differences, then

$$q_{ijk} = q_{Hjk}(1-p) + (1-q_{Hjk})p. \tag{3.4}$$

3.2.4 Entropy

We use the entropy as measure of uncertainty for the distribution of edge inclusions in the migration graph. The entropy of a distribution is a quantity commonly used in information theory (see for example Gray (2011)). For a discrete random variable Y assuming values

y_i it is defined as

$$H(Y) = E(-\log(p(Y))) = -\sum(p(y_i)\log(p(y_i))). \tag{3.5}$$

Entropy is used as a measure of uncertainty in probability distributions, and it attains its maximum when the distribution is uniform (ie. all outcomes have the same probability). It is especially useful for assessing variability of distributions over categorical data. When **Y** is a vector of random variables with components Y_j, then

$$H(\mathbf{Y}) \leq \sum H(Y_j) \tag{3.6}$$

with equality holding when the elements of **Y** are independent.

3.3 Results

We analyze two datasets for which we integrate information across multiple strata to obtain better representations of evolutionary and spatial processes. The first is in the context of geographical dispersion of viral pathogens, in which we analyze the migration pattern of dengue virus in the Americas by combining information from the four different serotypes of the virus. Next, we explore the use of our hierarchical model for a different type of discrete trait: cellular compartments infected by HIV. We use information from 14 different patients to study the intrahost dynamics of the virus between these compartments.

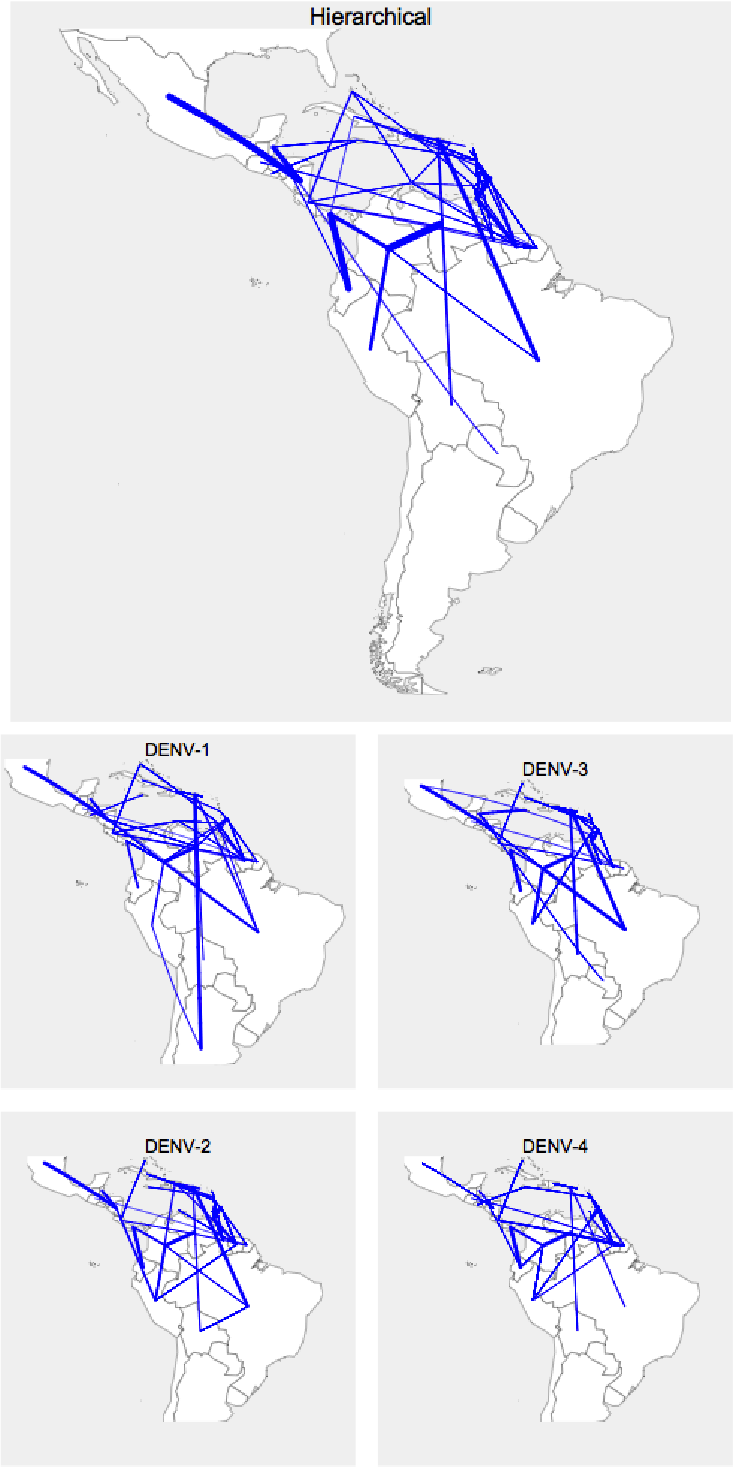
3.3.1 Dengue Virus in the Americas

With approximately 50 million infections annually, dengue is a serious public health issue in the tropical and subtropical regions where its mosquito vectors, *Aedes aegypti* and *Aedes albopictus*, are common (Guzman et al, 2010). In the Americas, the virus is distributed widely, with reported cases in all but 3 countries; and the number of cases has been steadily increasing since the 1980s (San Martín et al, 2010). There are four antigenically distinct dengue virus serotypes (DENV-1 to DENV-4), and we integrate data from all serotypes to study geographical patterns of the virus in the Americas.

We analyze a dataset consisting of 904 sequences of the envelope gene, divided between all four serotypes of the virus. The samples originated from $K = 36$ different countries in Latin America and the Caribbean, and date from 1977 to 2009. These data were previously analyzed by treating each serotype independently (Allicock et al, 2012).

Because of the biological similarities, we model geographical diffusion for the serotypes jointly to identify the overall dispersal pattern. For each serotype we have a migration graph, with nodes representing the sampling locations and edges representing which instantaneous rates between locations are nonzero. An overall migration graph summarizes this information at the hierarchical level.

Figure 3.2 presents these graphs, superimposed over maps (Bielejec et al, 2011), for the hierarchical level and each stratum. Thickness of the edges are proportional to edge support, and only those edges that have Bayes factors larger than 3 are shown. Our results agree with the previous independent analysis in that most of the significant links are between neighbouring countries (Allicock et al, 2012). An example of this are the highly supported links between Peru, Venezuela and Colombia. Additionally, our model indicates that a few countries such as Colombia, Suriname, Trinidad and Tobago, and Martinique act as centers of viral dispersion, with high number of links to other countries.



Hierarchical

DENV-1

DENV-3

DENV-2

DENV-4

Figure 3.2: Hierarchical and serotype location maps with links connecting the countries that have direct viral migration. Edge widths are proportional to posterior probability for edge inclusions, and only edges with a Bayes factor higher than 3 are shown.

An example of how the hierarchical model integrates information across the strata can be seen by comparing edge probabilities in hierarchical and stratum graphs for locations with incomplete sampling. Notice that although we do not have samples from Nicaragua for serotype 4, the hierarchical structure of our model requires that all graphs have the same nodes, and so the probability of an edges linking Nicaragua to other countries are estimated for DENV-4. Building on information from the other three serotypes, the posterior probability for an edge between Nicaragua and Mexico in the hierarchical graph is 0.93. The inclusion probability of this edge for DENV-4 is dictated by the hierarchical prior. Thus, even though the data carry no direct information on the migration of DENV-4 through Nicaragua, we have a Bayes factor of 6.1 for including the edge.

For some edges, there are notable discrepancies between hierarchical and stratum graphs; however, overall deviations conform well with the specified model for all strata. The posterior mean edge differences lie between 0.044 and 0.050 per edge for all serotypes; while the prior model specification assigned a 0.05 probability for an edge in the stratum graph being different from the hierarchical graph.

We have also analyzed this data using geographical distances to inform the prior probabilities of edge inclusions in the hierarchical graph. The results, however, were only slightly different from those presented here and the qualitative conclusions remained absolutely unchanged

3.3.2 Intrahost HIV

Treatment of HIV with highly active antiretroviral therapy (HAART) significantly suppresses viral replication in CD4^{+} T lymphocytes. In this context, alternative sites of HIV infection, such as CD8^{+} T cells, may become increasingly important. To assess the role of infection of CD8^{+} cells in patients under HAART, we study a previously published HIV-1 dataset from 14 different patients (Potter et al, 2006). In each patient, the virus was isolated from two different cell compartments, CD4$^+$ T cells, CD8$^+$ T cells, as well as from plasma. We analyze samples collected at two or three different times for each patient, the first being at the time of treatment initiation.

Under the assumption that viral migration between cell types is similar in all patients, each patient represents one stratum in our hierarchical model. Since there are only 6 edges in the migration graph for this problem, we use a binomial prior with inclusion probability 0.5. Figure 3.3 presents the hierarchical graph representing the main viral migrations between cell compartments. Directed edges and corresponding Bayes factors are only shown for links with a Bayes factor higher than 1. In this graph, the main connections are between plasma and CD4$^+$ compartments. Additionally, there is evidence for a directed edge from the CD8$^+$ to the CD4$^+$ compartment (BF=35.3). Equivalent graphs for the patients show the same overall pattern, with the eventual addition of one other edge. These graphs can be found in the supplementary material (section 3.5).

We compare the graph hierarchical model to two alternative approaches for analyzing this dataset: the consensus approach, where all patients are assumed to have the same migration matrix, and the independent approach, where patient analyses are carried out separately. We make analogous choices of prior distributions in all analyses. Table 3.1 presents a comparison between analyses of the uncertainty of graph estimates measured through the entropy of their posterior distributions for edges. The comparison of the hierarchical model and independent analysis shows lower entropy values in the hierarchical model for every patient. This indicates that combining the patient data in the hierarchical setting reduces the uncertainty in the estimates of individual patient graphs. The consensus analysis also presents higher entropy than the overall matrix of the hierarchical model.

In general the posterior probabilities of edges in the hierarchical and stratum graphs are

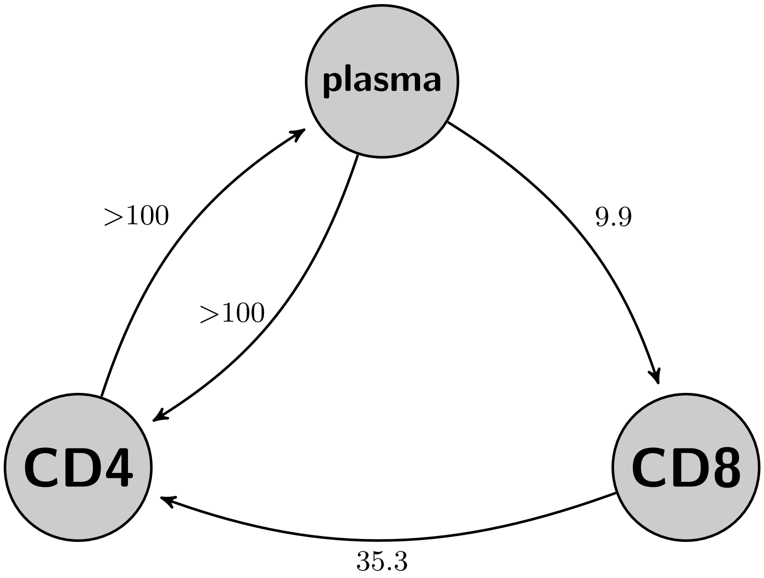


Figure 3.3: Migration graphs for the hierarchical level of the intrahost HIV data, with Bayes factor value for each edge in the graph. Only edges with a Bayes factor higher than 1 are shown.

Table 3.1: Comparison of hierarchical model, independent and consensus analysis.

	hierarchical model		independent	consensus
	pp	entropy	entropy	entropy
patient 1	0.0667	1.9260	3.0124	
patient 2	0.0929	2.0128	3.1824	
patient 3	0.0692	1.9800	3.5299	
patient 4	0.0565	1.8330	2.7376	
patient 5	0.0734	1.8557	3.0509	
patient 6	0.0631	1.8734	3.4204	
patient 7	0.0685	1.9541	2.7179	
patient 8	0.1094	1.9844	3.2912	
patient 9	0.0803	1.9516	3.4381	
patient 10	0.0944	2.0824	3.2942	
patient 11	0.0698	1.9066	3.2193	
patient 12	0.0764	1.9619	3.1608	
patient 13	0.0729	1.8894	3.2017	
patient 14	0.0549	1.8399	3.0861	
overall		1.6022		2.2937

Entropy values are computed for the posterior distributions of edge inclusions for the graphs of states, and pp represents the posterior probability of edges in the stratum graph being different from the hierarchical graph.

similar. Table 3.1 shows that, in the posterior distribution, between 0.05 - 0.1% of the stratum edges differ from their counterparts in the hierarchical graph. The discrepancy fraction observed in the sample varies among patients, and for patients 2, 8 and 10 is close to twice the expected 0.05 defined as the binomial parameter p for the edge differences.

3.4 Discussion

We present a hierarchical phylogenetic model for the evolution of a discrete trait in multiple strata. Our method integrates information across the strata to improve estimation, while allowing for inter-strata variation. At the hierarchical level, properties of the overall process are summarized through the migration graph, with edges representing nonzero instantaneous rates.

The main motivation for this method comes from phylogeographic applications, in which the trait of interest is geographic location, and we wish to study the migration process. Recently, many studies have analyzed spatial spread of epidemics using analogous Bayesian phylogeographic methods for only one sample of the geographical process (Nelson et al, 2011; Auguste et al, 2010; Allicock et al, 2012). In addition to the potential public health significance, these studies are motivated by the fact that rapidly evolving pathogens allow for viral sampling in a time frame comparable to sequence evolution, leading to reconstructions of geographic diffusion in real time units. In this context, we present an example in which we use data from multiple serotypes of dengue virus to study the viral dispersion process in the Americas. The analysis supports similar dispersion processes across the four serotypes.

In the parallel datasets for which our model is constructed, it is possible that some of the strata have incomplete sampling of state locations, as with the dengue example. In the dengue dataset, some countries do not have samples for all serotypes, yet the hierarchical structure of our model integrates information from the other serotypes to estimate migration rates for the missing strata. Our hierarchical model naturally deals with these often troubling missing data problems.

Hierarchical phylogenetic models have been used in phylodynamic studies of the intrahost behavior of HIV (Liang and Weiss, 2007; Edo-Matas et al, 2011). The long timespan of HIV infections, which sometimes last more than 10 years, combined with high mutation rates, makes phylodynamics a useful tool for assessing viral intrahost biology. Thus, hierarchical models that combine data from multiple patients are relevant. In this context, we present an example in which we use an HIV intrahost dataset to asses the role of infection in CD8$^+$ T cells for patients under HAART.

Our analysis suggests that an important component of the inter-compartment dynamics is the replenishment of CD4$^+$ T cells by viral populations from CD8$^+$ cells. Because CD4$^+$

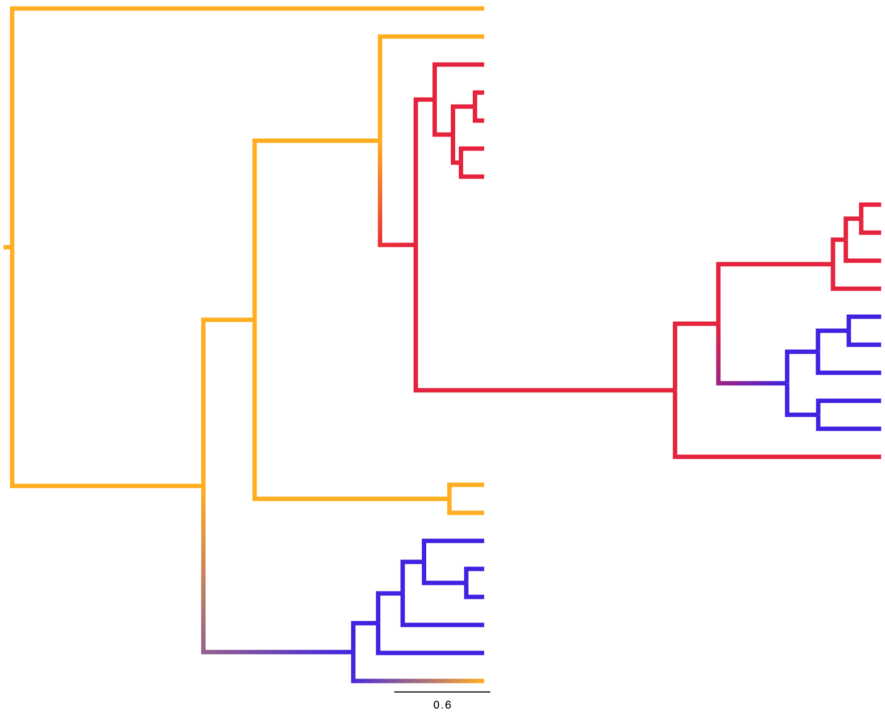


Figure 3.4: MCC phylogeny for patient 13 with branches coloured according to the most probable posterior cell compartment. Red represents CD8 T cells, blue represents CD4 T cells, and yellow represents plasma.

cells represent the main pool of infected cells, in line with the higher HIV-1 diversity observed in this compartment (Potter et al, 2006), this compartment is expected to replenish other cell compartments under a metapopulation dynamics scenario. According to the hierarchical graph, CD4$^+$ cells could still be the major source of infection for CD8$^+$ cells when seeded by plasma virus. However, since our analysis also indicates an non-negligible role for viral migration from CD8$^+$ cells to CD4$^+$ cells during HAART, as also exemplified by the reconstructed migration history in the maximum clade credibility (MCC) tree for a particular patient (Figure 3.4), the role of CD8$^+$ cells in the maintenance of HIV reservoir dynamics may require further attention.

In our method, the conformity of stratum graphs to the hierarchical graph is dependent on the choice of parameter p for the binomial distribution of edge differences. Small values of p induce a large amount of information sharing between strata, and impose a large constraint on similarity between stratum graphs. In the limit, we have the consensus analysis, in which all strata have the same graph. On the other hand, large values of p generate estimates with smaller dependency between graphs.

Even though we use a fixed value for binomial parameter p, the fraction of discrepancies in the posterior distribution may differ from the specified probability p. This may be used as an indicator of the degree of similarity between the strata, or to identify an outlier stratum. This was observed in the HIV example, where posterior discrepancies were higher in some strata than in others. In comparison, in the dengue example all strata conformed better to a common dispersion process.

We follow Lemey et al (2009) in our choice of prior distribution for the total number of edges in the hierarchical graph, by adopting the Poisson approximation for the sum of a large number of Bernoulli random variables. This choice of prior distribution has been used in a number of subsequent applications, where it has adequately controlled the total number of edges included in the graph. Other popular prior choices for graph edges consider edge inclusions as exchangeable Bernoulli trials with common success probability (Brown et al, 1998; Dobra et al, 2004). It follows that the total number of edges in the graph has a binomial prior. Additionally, Carvalho and Scott (2009) show that when the inclusion probability comes from the Beta hyperprior, the model has a strong control over the number of false edges included in the graph. Telesca et al (2012) model dependent gene expression through a graph structure similar to the one in our model, and adopt the binomial-beta model for total number of edges; they show through simulations that their model presents good control over false discovery rates.

One advantage of the Bayesian phylogenetic framework we exploit in formulating our graph hierarchical model is that the framework easily lends itself to combination of different models. These could be phylogenetic methods for demographical inference (Minin et al, 2008), methods for calibrating trees and relaxed clock models (Drummond et al, 2006). Our hierarchical phylogeographic approach can easily be associated with these existing models to provide comprehensive analysis of viral history.

The data analyses presented here highlight the strengths of the hierarchical phylogenetic model for analyzing these parallel dataset consisting of conditionally independent trait evolution processes. In particular, the entropy comparisons for the HIV data show the reduction in overall uncertainty of the hierarchical model, in comparison to the independent approach. This is obtained by sharing information over the parallel strata. On the other hand, the consensus approach does not allow for the variability between strata processes observed in both examples.

Our method paves the way for further exploration of geographic dispersion processes. Through an analysis of the migration graph, for example, we can identify structural properties of the system: fully connected subgraphs and cycles are motifs that may represent local dynamics of interest. We could also assess the reversibility of the geographic process, by testing time-reversibility on the migration matrix. Additionally, changes in the migration process over time could be assessed by generalizing our model into a dynamic model. This method could also be extended to account for more general dependency structures on the hyper-graph.

3.5 HIV supplementary figures

This section presents the figures supplementary figures for the HIV intrahost analysis of subsection 3.3.2.

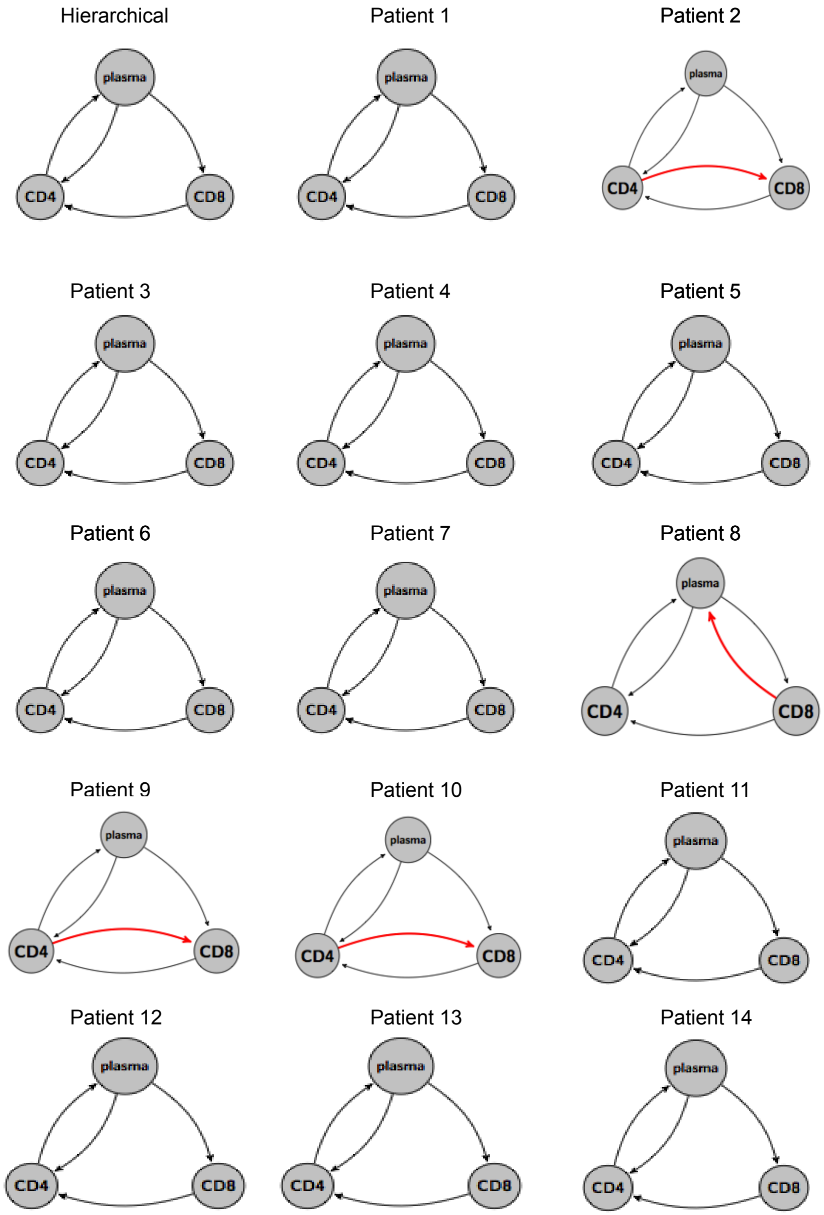


Figure 3.5: Migration graphs for the strata and hierarchical level of the intrahost HIV dataset. Edges that differ from the hierarchical graph are presented in red. Only edges with a Bayes factor higher than 1 are shown.

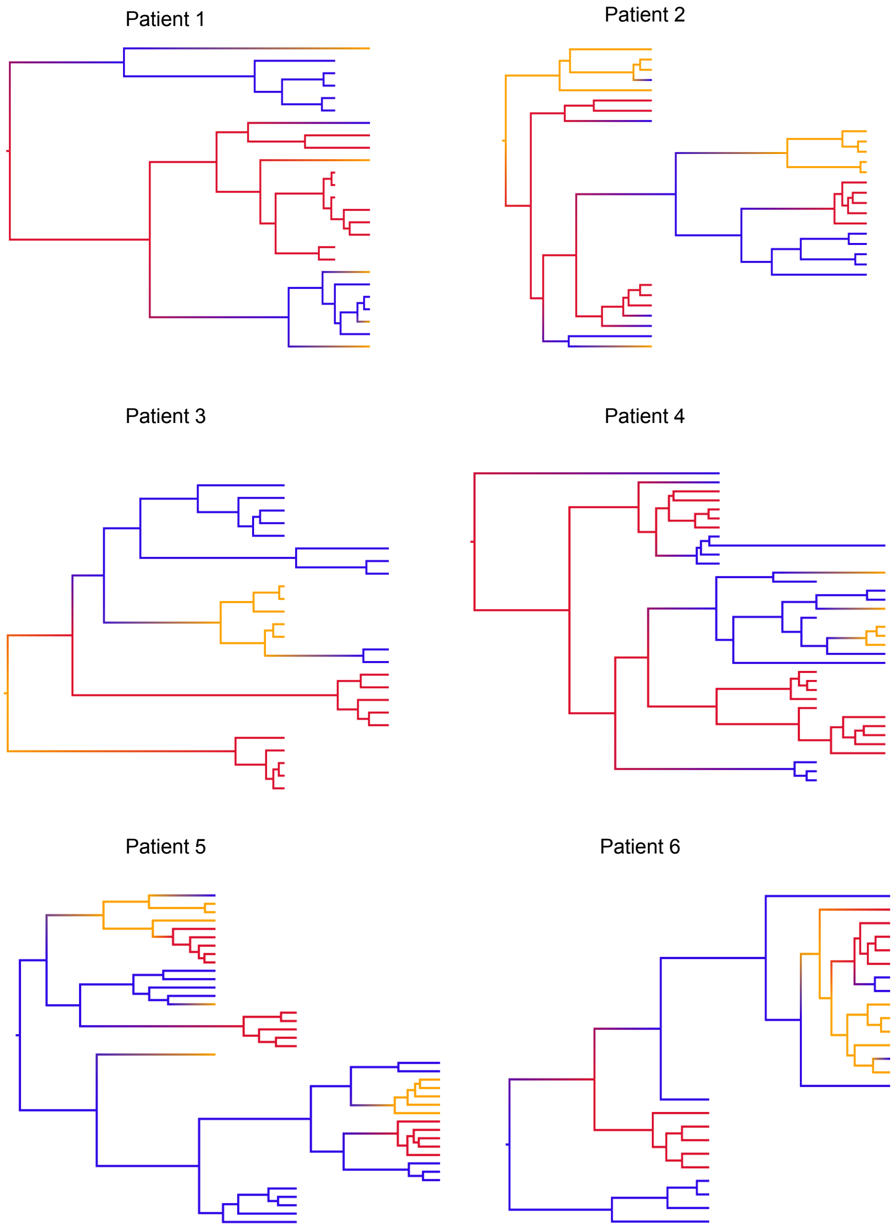


Figure 3.6: MCC phylogeny for patients 1 - 6 of the HIV dataset, with branches colored according to the most probable posterior cell compartment. Red represents CD8 T cells, blue represents CD4 T cells, and yellow represents plasma.

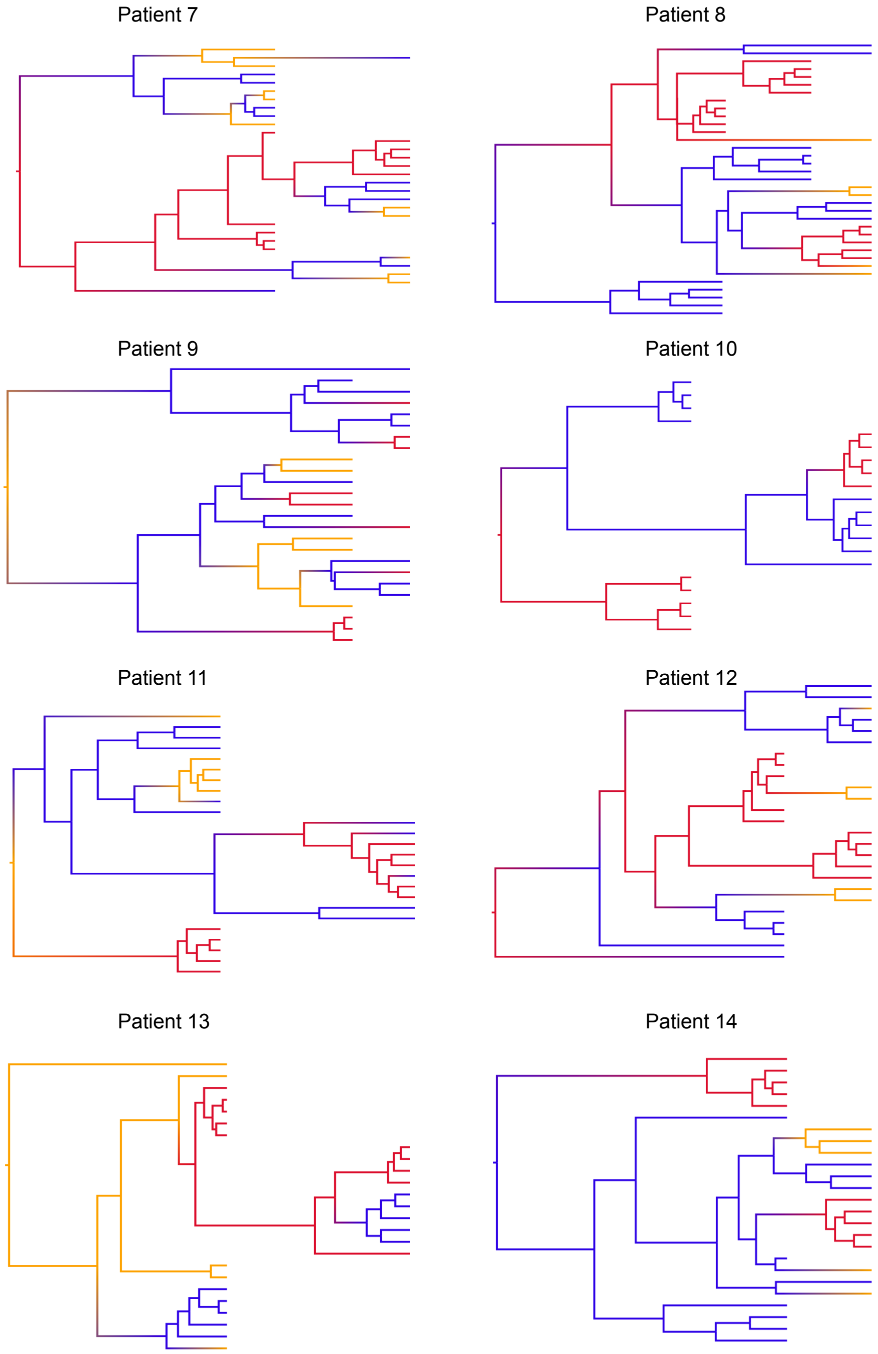


Figure 3.7: MCC phylogeny for patients 7 - 14 of the HIV dataset, with branches colored according to the most probable posterior cell compartment. Red represents CD8 T cells, blue represents CD4 T cells, and yellow represents plasma.

CHAPTER 4

Assessing phenotypic correlation through the multivariate phylogenetic latent liability model[1]

Abstract. Understanding which phenotypic traits are consistently correlated throughout evolution is a highly pertinent problem in modern evolutionary biology. Here, we propose a multivariate phylogenetic latent liability model for assessing the correlation between multiple types of data, while simultaneously controlling for their unknown shared evolutionary history informed through molecular sequences. The latent formulation enables us to consider in a single model combinations of continuous traits, discrete binary traits, and discrete traits with multiple ordered and unordered states. Previous approaches have entertained a single data type generally along a fixed history, precluding estimation of correlation between traits and ignoring uncertainty in the history. We implement our model in a Bayesian phylogenetic framework, and discuss inference techniques for hypothesis testing. Finally, we showcase the method through applications to columbine flower morphology, antibiotic resistance in *Salmonella*, and epitope evolution in influenza.

[1] This project is joint work with Janet S. Sinsheimer, Trevor Bedford, Alison E. Mather, Phillippe Lemey and Marc A. Suchard

4.1 Introduction

Biologists are often interested in assessing phenotypic correlation among sets of traits, since it can help elucidate many biological processes. These correlations may be a result of genetic correlation, in which traits are partially determined by the same or linked loci. Alternatively, they may be evidence of selective correlation, in which the same environmental pressure acts on two seemingly unrelated traits or the outcome of one trait affects selective pressure on the other. Studying these processes is one of the aims of comparative biology.

The purpose of this project is to present a statistical framework for estimating phenotypic correlation among many traits simultaneously for combinations of different types of data. We consider combinations of continuous data, discrete data with binary outcomes, and discrete data with multiple ordered and unordered outcomes. We also provide inference tools to address specific hypotheses regarding the correlation structure.

Several comparative methods have been proposed to assess the phenotypic correlation between groups of traits (Felsenstein, 1985; Pagel, 1994; Grafen, 1989; Ives and Garland, 2010). These methods estimate correlations in trait data across multiple species while controlling for shared evolutionary history through phylogenetic trees. Yet their use is generally limited to fixed phylogenetic trees, specific types of data or small datasets.

Markov chains are a natural choice to model the evolution of discrete traits, allowing for correlation between them (Pagel, 1994; Lewis, 2001). In this case, the state space of the Markov chain includes all combinations of possible values for all the traits, and correlation is assessed through the transition probabilities between states. Thus, when the number of traits and possible outcomes for each trait increase, the number of parameters to be estimated in the rate matrix scales up rapidly.

For continuous data, a common approach for assessing phenotypic correlation is the independent contrasts method that models the evolution of multiple traits as a multivariate Brownian diffusion process along the tree (Felsenstein, 1985). Correlation between traits is assessed through the precision matrix of the diffusion process. This method has been extended to account for phylogenetic uncertainty by integrating over the space of trees in a Bayesian context (Huelsenbeck and Rannala, 2003). Recent developments also increase the methods flexibility by allowing for different diffusion rates along the branches of the tree (Lemey et al, 2010) and more efficient computation of the model likelihood, and thus, larger datasets (Pybus et al, 2012).

Phylogenetic linear models and related methods can naturally consider combinations of different types of data (Grafen, 1989; Ives and Garland, 2010). Developments in this area have led to flexible and efficient methods (Faria et al, 2013; Ho and Ané, 2014). However, these models assess the effects of independent variables on a dependent variable that evolves along a tree. Although it is possible that the independent variables are phylogenetically correlated, this aspect is generally not explicitly modeled. Thus, these models are not tailored to assess correlation between sets of traits throughout evolutionary history.

An approach for assessing correlated evolution that can combine both binary and continuous data is the phylogenetic threshold model (Felsenstein, 2005, 2012). The threshold model is used in statistical genetics for traits with a discrete outcome determined by an underlying unobserved continuous variable (Wright, 1934; Falconer, 1965). Felsenstein (2005) proposed the use of this model in phylogenetics. In his model, the underlying continuous variable (or latent liability) undergoes Brownian diffusion along the phylogenetic tree. At the tips, a binary trait is defined depending on the position of the latent liability relative to a specified threshold. This non-Markovian model has the desirable property that the probability of transition from the current state to another can depend time spent in that current state.

A possible interpretation for this model is that the binary outcome represents the presence or absence of some phenotypic trait, and the underlying continuous process represents the combined effect of a large number genetic factors that affect this trait. During evolution, these factors undergo genetic drift, that is usually modeled as Brownian diffusion.

In its multivariate version, the threshold model allows for inference on the phenotypic correlation structure between a few continuous and binary traits. As with the independent contrasts method, this correlation can be assessed through the covariance matrix of the multivariate Brownian diffusion for the continuous latent liability.

In this project we build upon the flexibility of the threshold model to create a Bayesian phylogenetic model for the evolution of binary data, discrete data with multiple ordered or unordered states and continuous data. We explore recent developments in models for continuous trait evolution that improve computational efficiency, and make the joint analysis of multiple traits feasible in the presence of possible phylogenetic uncertainty (Lemey et al, 2010; Pybus et al, 2012).

Importantly, our approach estimates the between trait correlation while simultaneously controlling for the correlation induced through the traits being shared by descent.

As shown in one of our examples, failing to control for the evolutionary history can result in confounded inference of the correlation between traits, in analogy to false inference in association analysis when failing to control for population substructure or relatedness among individuals. In section 4.2 we introduce our model and discuss inference procedure and hypothesis testing. Then, in section 4.3, we present three applications that exemplify the use of the phylogenetic latent liability model to different biological problems. First, we assess phenotypic correlation in multi-drug resistance for *Salmonella* through the analysis of binary resistance data. Then we analyse a dataset with a combination of continuous and discrete traits to investigate how a series of morphological floral traits relate to shifts if pollinators for Columbine flowers. And, in the third application, we assess correlations in the evolution of epitopes of the HA protein in Influenza, through the analysis of discrete data with multiple unordered states. Finally, in section 4.4 we discuss our results, and some extensions to the model.

4.2 Methods

Consider a dataset of N aligned molecular sequences $\mathbf{S}$ from related organisms and an $N \times P$ matrix $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_N)^t$ of P-dimensional trait observations from each of the N organisms, such that $\mathbf{Y}_i = (y_{i1}, \dots, y_{iP})$ for $i = 1, \dots, N$. We model the sequence data $\mathbf{S}$ using standard Bayesian phylogenetics models (Drummond et al, 2012) that include, among other parameters $\boldsymbol{\phi}$ less germaine to our development here, an unobserved phylogenetic tree F. This phylogenetic tree is a bifurcating, directed graph with N terminal nodes $(\nu_1, \dots, \nu_N)$ of degree 1 that correspond to the tips of the tree, $N - 2$ internal nodes $(\nu_{N+1}, \dots, \nu_{2N-2})$ of degree 3, a root node ν_{2N-1} of degree 2 and edge weights $(t_1, \dots, t_{2N-2})$ between nodes that track elasped evolutionary time. Conditional on F, we assume independence between $\mathbf{S}$ and $\mathbf{Y}$, and refer interested readers to, for example, Suchard et al (2001) and Drummond et al (2012) for detailed development of $p(\mathbf{S}, \boldsymbol{\phi}, F)$.

The dimensions of $\mathbf{Y}_i$ contain trait observations that may be binary, discrete with multiple states, continuous or a mixture thereof. Importantly, to handle the myriad of different data types, we assume that the observation of $\mathbf{Y}$ is governed by an underlying unobserved continuous random variable $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N)^t$, called a latent liability, where each row $\mathbf{X}_i = (x_{i1}, \dots, x_{iD}) \in \mathbb{R}^D$ with $D \geq P$ depending on the mixture of data types. In brief, we assume that $\mathbf{X}$ arise from a multivariate Brownian diffusion along the tree F (Lemey et al, 2010) for which we provide a more indepth description shortly. At the tips of F, the realized values of $\mathbf{Y}$ emerge deterministically from the latent liabilities $\mathbf{X}$ through the mapping function $g(\mathbf{X})$.

4.2.1 Latent Liability Mappings

When column j of $\mathbf{Y}$ provides binary data, these values map from a single dimension j' in $\mathbf{X}$ following a probit-like formulation in which the outcome is one if the underlying continuous value is larger than a threshold and zero otherwise. Without loss of generality, we take the threshold to be zero, such that

$$y_{ij} = g(x_{ij'}) = \begin{cases} 0 & \text{if } x_{ij'} \leq 0 \\ 1 & \text{if } x_{ij'} > 0. \end{cases} \tag{4.1}$$

Alternatively, if column j of $\mathbf{Y}$ assumes K possible discrete states $(s_1, \ldots, s_K)$, and they are ordered so that transitions from state s_k to s_{k+2} must necessarily pass through s_{k+1}, we entertain a multiple threshold mapping (Wright, 1934). Again, column j of $\mathbf{Y}$ maps from a single dimension j' in the latent liabilities $\mathbf{X}$; however, the position of $x_{ij'}$ relative to the multiple thresholds $(a_1, \ldots, a_{K-1})$ determines the value of y_{ij} through the function

$$y_{ij} = g(x_{ij'}) = \begin{cases} s_1 & \text{if } x_{ij'} < a_1 \\ s_k & \text{if } a_{k-1} \leq x_{ij'} < a_k \quad \text{for } k = 2, \ldots, K-1 \\ s_K & \text{if } x_{ij\prime} \geq a_{K-1}, \end{cases} \tag{4.2}$$

where $a_2, \ldots, a_{\text{K}-1}$ in increasing values are generally estimable from the data if we set $a_1 = 0$ for identifiability. Let $\mathbf{A} = \{a_k\}$ track all of the non-fixed threshold parameters for all ordered traits.

When column j of $\mathbf{Y}$ realizes values in K multiple states, but there is no ordering between them, we adopt a multinomial probit model Here the observed trait maps from $K-1$ dimensions in the latent liabilities $\mathbf{X}$, and the value of y_{ij} is determined by the largest component of these latent variables, such that

$$y_{ij} = g(x_{ij'}, \ldots, x_{i,j'+K-2}) = \begin{cases} s_1 & \text{if} \quad 0 = \sup(0, x_{ij}, \ldots, x_{i,j+K-2}) \\ s_{k+1} & \text{if} \;\; x_{ik} = \sup(0, x_{ij}, \ldots, x_{i,j+K-2}), \end{cases} \tag{4.3}$$

where we have taken without loss of generality the first state s_1 to be the reference state.

Finally, if column j of $\mathbf{Y}$ returns continuous values, a simple monotonic transform from $\mathbb{R}$ suffices. For example, for normally distributed outcomes, $y_{ij} = g(x_{ij'}) = x_{ij'}$.

4.2.2 Trait Evolution

A multivariate Brownian diffusion process along the tree F (Lemey et al, 2010) gives rise to the elements of $\mathbf{X}$. This process posits that the latent trait value of a child node ν_k in F is multivariate normally distributed about the unobserved trait value of its parent node $\nu_{\text{pa}(k)}$ with variance $t_k \times \mathbf{\Sigma}$. In this manner, the unknown $D \times D$ matrix $\mathbf{\Sigma}$ characterizes the between-trait correlation and the tree F controls for trait values being shared by descent.

Assuming that the latent trait value at the root node ν_{2N-1} draws *a priori* from a multivariate normal distribution with mean $\boldsymbol{\mu}_0$ and variance $\tau_0 \times \mathbf{\Sigma}$ and integrating out the internal and root node trait values (Pybus et al, 2012), we recall that the latent liabilities $\mathbf{X}$ at the tips of F are matrix normally distributed, with probability density function

$$p(\mathbf{X} \,|\, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0) = \frac{\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\boldsymbol{\mu}_0\right)^t\left(\mathbf{V}(F)+\tau_0\mathbf{J}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}_0\right)\right]\right\}}{\left(2\pi\right)^{NP/2}\left|\boldsymbol{\Sigma}\right|^{N/2}\left|\mathbf{V}(F)+\tau_0\mathbf{J}\right|^{P/2}}, \qquad (4.4)$$

where $\mathbf{J}$ is the $N \times N$ matrix of all ones and $\mathbf{V}(F) = \{v_{ii'}\}$ is an $N \times N$ matrix that is a deterministic function of F. Specifically, let $d_F(u, w)$ equal the sum of edge weights along the shortest path between node u and node w in F. Then diagonal elements $v_{ii} = d_F(\nu_{2N-1}, \nu_i)$, the time-distance between the root node and tip node i, and the off-diagonal elements $v_{ii'} = \left[d_F(\nu_{2N-1}, \nu_i) + d_F(\nu_{2N-1}, \nu_{i'}) - d_F(\nu_i, \nu_{i'})\right]/2$, the time-distance between the root node and the most recent command ancestor of tip nodes i and i'.

We consider the augmented likelihood for the trait data $\mathbf{Y}$ and latent liabilities $\mathbf{X}$ and highlight a convenient factorization

$$p(\mathbf{Y}, \mathbf{X} \,|\, \mathbf{V}(F), \boldsymbol{\Sigma}, \mathbf{A}, \boldsymbol{\mu}_0, \tau_0) = p(\mathbf{Y} \,|\, \mathbf{X}, \mathbf{A}) \times p(\mathbf{X} \,|\, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0). \qquad (4.5)$$

The conditional likelihood $p(\mathbf{Y} \,|\, \mathbf{X}, \mathbf{A}) = \mathbf{1}_{(\mathbf{Y}=g(\mathbf{X}))}$ in factorization (4.5) is simply the indicator function that $\mathbf{X}$ are consistent with the observations $\mathbf{Y}$. Consequentially, the augmented likelihood is a truncated, matrix normal distribution.

Figure 4.1 illustrates schematic representations of the latent liability model for all four types of data. In the figure, we include trees with $N = 4$ to 6 taxa with their observed traits $\mathbf{Y}$ at the tree tips and plot potential realizations of the latent liabilities $\mathbf{X}$ values along these trees that give rise to $\mathbf{Y}$.

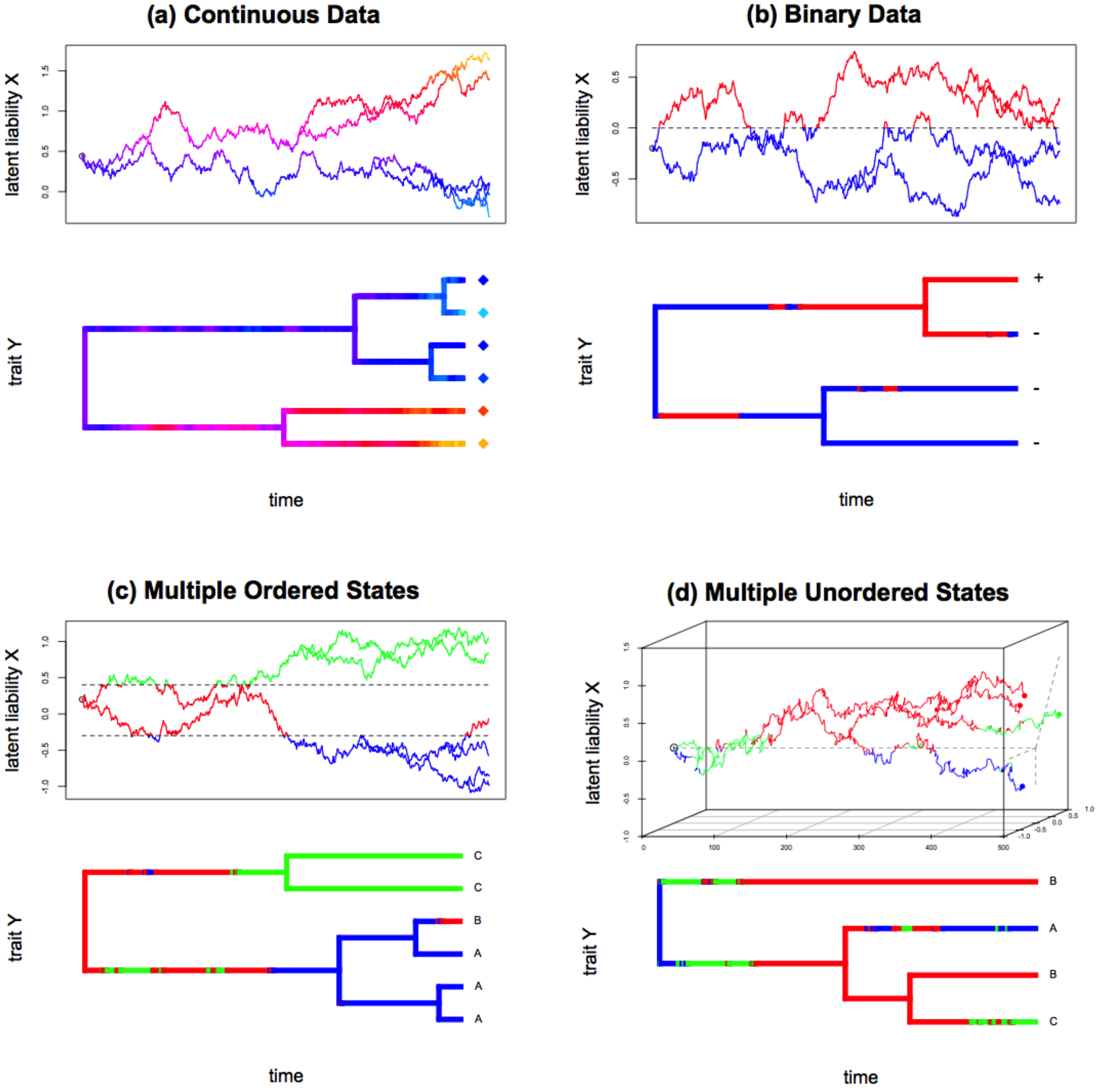


Figure 4.1: Realizations of the evolution of latent liabilities **X** and observed trait **Y** for different types of data. Both tree and Brownian motion plots are color coded according to the trait **Y**. Realization **(a)** represents a continuous trait, **(b)** represents discrete binary data, **(c)** represents discrete data with multiple ordered states, and **(d)** represents discrete data with multiple unordered states, for which the latent liabilities **X** is multivariate. **This figure was created using code modified from R package *phylotools* (Revell, 2012).

We complete our model specification by assuming *a priori*

$$\mathbf{\Sigma}^{-1} \sim \text{Wishart}(d_0, \mathbf{T}), \tag{4.6}$$

with degrees of freedom d_0 and rate matrix $\mathbf{T}$. For the non-fixed threshold parameters $\mathbf{A}$, we assume differences $a_k - a_{k-1}$ for each trait are *a priori* independent and Exponential(α) distributed, where α is a rate constant. Finally, we specify fixed hyperparameters $(\boldsymbol{\mu}_0, \tau_0, d_0, \mathbf{T}, \alpha)$ in each of our examples.

4.2.3 Inference

We aim to learn about the posterior distribution

$$\begin{aligned} p(\mathbf{\Sigma}, F, \boldsymbol{\phi}, \mathbf{A} \,|\, \mathbf{Y}, \mathbf{S}) &\propto p(\mathbf{Y} \,|\, \mathbf{\Sigma}, F, \mathbf{A}) \times p(\mathbf{\Sigma}) \times p(\mathbf{A}) \times p(\mathbf{S}, \boldsymbol{\phi}, F) \\ &= \left(\int p(\mathbf{Y}, \mathbf{X} \,|\, \mathbf{\Sigma}, F, \mathbf{A}) \mathrm{d}\mathbf{X} \right) \times p(\mathbf{\Sigma}) \times p(\mathbf{A}) \times p(\mathbf{S}, \boldsymbol{\phi}, F). \end{aligned} \tag{4.7}$$

We accomplish this task through Markov chain Monte Carlo (MCMC) and the development of computationally efficient transitions kernels to faciliate sampling of the latent liabilities $\mathbf{X}$. We exploit a random-scan Metropolis-with-Gibbs scheme. For the tree F and other phylogenetic parameters $\boldsymbol{\phi}$ involving the sequence evolution, we employ standard Bayesian phylogenetic algorithms (Drummond et al, 2012) based on Metropolis-Hastings parameter proposals. Further, the full conditional distribution of $\mathbf{\Sigma}^{-1}$ remains Wishart (Lemey et al, 2010), enabling Gibbs sampling.

MCMC transition kernels for sampling $\mathbf{X}$ are more problematic; tied into this difficulty also lies computationlly efficient evaluation of Equation (4.4). Strikingly, the solution to

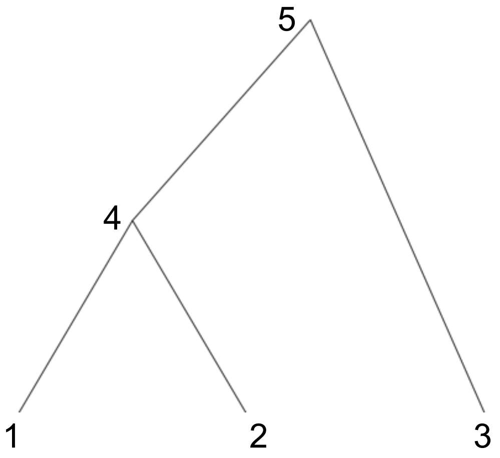


Figure 4.2: Example $N = 3$ tree to illustrate pre- and post-order traversals for efficient sampling of latent liabilities $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)^t$.

the latter points to new directions in which to attack the sampling problem. As written, computing $p(\mathbf{X} \,|\, \mathbf{V}(F), \mathbf{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ to evaluate a Metropolis-Hasting acceptance ratio appears to command the high computational cost of $\mathcal{O}(N^3)$ involved in forming $(\mathbf{V}(F) + \tau_0 \mathbf{J})^{-1}$. In general, such a cost would be prohibitive for large N when F is random, necessiating repeated inversion. This is one reason why previous work has limited itself to fixed, known F. However, we follow Pybus et al (2012), who develop a dynamic programming algorithm to evaluate density (4.4) in $\mathcal{O}(N)$ that avoids matrix inversion. Critically, we extend these algorithmic ideas in this project to construct computationally efficient sampling procedures for $\mathbf{X}$.

Pybus et al (2012) propose a post-order tree traversal that visits each node u in F, starting at the tips and ending at the root. For the example tree displayed in Figure 4.2, one possible post-order traversal proceeds through nodes $\{1 \rightarrow 2 \rightarrow 4 \rightarrow 3 \rightarrow 5\}$. Let $\mathbf{X}_u$ for $u = N+1, \ldots, 2N-1$ imply now hypothesized latent liabilities at the internal and root nodes of F. Then, at each visit, one computes the conditional density of the tip latent liabilities $\{\mathbf{X}\}_u^{\text{post}}$ that are descendent to node u given $\mathbf{X}_{\text{pa}(u)}$ at the parent node of u by integrating out the hypothesized value $\mathbf{X}_u$ at node u. For example, when visiting node $u = 4$ in Figure 4.2, one considers the conditional density of $(\mathbf{X}_1, \mathbf{X}_2) \,|\, \mathbf{X}_5$. Each of these conditional densities are proportional to a multivariate normal density, so during the traversal it suffices to simply keep track of the partial mean vector $\mathbf{m}_u^{\text{post}}$, partial precision scalar p_u^{post} and remainder term ρ_u that characterize the conditional density. We refer interested readers to the Supplementary Material in Pybus et al (2012) for further details.

Building upon this post-order traversal algorithm, we identify that it is possible and practical to generate samples from $p(\mathbf{X}_i \,|\, \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ for tip ν_i without having to manipulate $\mathbf{V}(F)$ via one additional pre-order traversal of F. This enables us to exploit $p(\mathbf{X}_i \,|\, \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ as a proposal distribution in an efficient Metropolis-Hastings scheme to sample $\mathbf{X}_i$, since the distribution often closely approximates the full conditional distribution of $\mathbf{X}_i$.

To ease notation in the remainder of this section, we drop explicit dependence on $\mathbf{V}(F)$, $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_0$, τ_0 in our distributional arguments. Further, let $\{\mathbf{X}\}_u^{\text{pre}}$ collect the latent liabilities at the tree tips that are not descendent to node u for $u = 1, \ldots, 2N-1$, such that $\{\mathbf{X}\}_u^{\text{pre}} \cup \{\mathbf{X}\}_u^{\text{post}} = \mathbf{X}$ and $\{\mathbf{X}\}_u^{\text{pre}} \cap \{\mathbf{X}\}_u^{\text{post}} = \emptyset$. Notably, $\{\mathbf{X}\}_i^{\text{pre}} = \mathbf{X}_{(-i)}$ and $\{\mathbf{X}\}_{2N-1}^{\text{pre}} = \emptyset$. With these goals and definitions in hand, we find $p(\mathbf{X}_i \,|\, \mathbf{X}_{(-i)})$ recursively.

Consider a triplet of nodes in F such that node u has parent pa$(u) = w$ that it shares with sibling sib$(u) = v$. For example, in Figure 4.2, $u = 1$, $v = 2$ and $w = 4$ is one of two choices. Because of the conditional independence structure of the multivariate Brownian diffusion process on F, we can write

$$p(\mathbf{X}_u \,|\, \{\mathbf{X}\}_u^{\text{pre}}) = \int p(\mathbf{X}_u \,|\, \mathbf{X}_{\text{pa}(u)})\, p(\mathbf{X}_{\text{pa}(u)} \,|\, \{\mathbf{X}\}_{\text{pa}(u)}^{\text{pre}}, \{\mathbf{X}\}_{\text{sib}(u)}^{\text{post}})\, \mathrm{d}\mathbf{X}_{\text{pa}(u)}, \tag{4.8}$$

where Equation (4.8) returns the desired quantity when $i = u$ and the first term of the integrand is a multivariate normal density MVN $\left(\mathbf{X}_u\,;\mathbf{X}_{\text{pa}(u)},(t_u\boldsymbol{\Sigma})^{-1}\right)$ centered at $\mathbf{X}_{\text{pa}(u)}$ with precision $(t_u\boldsymbol{\Sigma})^{-1}$. The second term requires more exploration

$$p(\mathbf{X}_{\text{pa}(u)}\,|\,\{\mathbf{X}\}^{\text{pre}}_{\text{pa}(u)},\{\mathbf{X}\}^{\text{post}}_{\text{sib}(u)}) \;=\; \frac{p(\mathbf{X}_{\text{pa}(u)},\{\mathbf{X}\}^{\text{post}}_{\text{sib}(u)}\,|\,\{\mathbf{X}\}^{\text{pre}}_{\text{pa}(u)})}{p(\{\mathbf{X}\}^{\text{post}}_{\text{sib}(u)}\,|\,\{\mathbf{X}\}^{\text{pre}}_{\text{pa}(u)})} \;\propto\; p(\{\mathbf{X}\}^{\text{post}}_{\text{sib}(u)}\,|\,\mathbf{X}_{\text{pa}(u)})\,p(\mathbf{X}_{\text{pa}(u)}\,|\,\{\mathbf{X}\}^{\text{pre}}_{\text{pa}(u)}), \tag{4.9}$$

where the normalization constant does not depend on $\mathbf{X}_{\text{pa}(u)}$ and we fortuitously have determined that $p(\{\mathbf{X}\}^{\text{post}}_{\text{sib}(u)}\,|\,\mathbf{X}_{\text{pa}(u)})$ is proportional to a MVN $\left(\mathbf{X}_{\text{pa}(u)}\,;\mathbf{m}^{\text{post}}_{\text{sib}(u)},p^{\text{post}}_{\text{sib}(u)}\boldsymbol{\Sigma}^{-1}\right)$ during the post-order traversal.

Substituting Equation (4.9) in Equation (4.8) furnishes a set of recursive integrals down the tree

$$p(\mathbf{X}_u\,|\,\{\mathbf{X}\}^{\text{pre}}_u)\propto\int p(\mathbf{X}_u\,|\,\mathbf{X}_{\text{pa}(u)})\,p(\{\mathbf{X}\}^{\text{post}}_{\text{sib}(u)}\,|\,\mathbf{X}_{\text{pa}(u)})\,p(\mathbf{X}_{\text{pa}(u)}\,|\,\{\mathbf{X}\}^{\text{pre}}_{\text{pa}(u)})\,\mathrm{d}\mathbf{X}_{\text{pa}(u)}. \tag{4.10}$$

To solve the set of integrals in (4.10), we recall that $p(\mathbf{X}_{2N-1}\,|\,\{\mathbf{X}\}^{\text{pre}}_{2N-1})=p(\mathbf{X}_{2N-1})$ is MVN $(\mathbf{X}_{2N-1}\,;\boldsymbol{\mu}_0,(\tau_0\boldsymbol{\Sigma})^{-1})$ and so define pre-order, partial mean vector $\mathbf{m}^{\text{pre}}_{2N-1}=\boldsymbol{\mu}_0$ and partial precision scalar $p^{\text{pre}}_{2N-1}=1/\tau_0$. Since the convolution of multivariate normal random variables remains multivariate normal, we identify that $p(\mathbf{X}_u\,|\,\{\mathbf{X}\}^{\text{pre}}_u)$ is multivariate normal MVN $\left(\mathbf{X}_u\,;\mathbf{m}^{\text{pre}}_u,p^{\text{pre}}_u\boldsymbol{\Sigma}^{-1}\right)$ where pre-order, partial mean vectors and precision scalars unwind through

$$\begin{aligned}\mathbf{m}^{\text{pre}}_u &= \frac{p^{\text{post}}_{\text{sib}(u)}\mathbf{m}^{\text{post}}_{\text{sib}(u)}+p^{\text{pre}}_{\text{pa}(u)}\mathbf{m}^{\text{pre}}_{\text{pa}(u)}}{\mathbf{m}^{\text{post}}_{\text{sib}(u)}+\mathbf{m}^{\text{pre}}_{\text{pa}(u)}},\ \text{and}\\ \frac{1}{p^{\text{pre}}_u} &= t_u+\frac{1}{p^{\text{post}}_{\text{sib}(u)}+p^{\text{pre}}_{\text{pa}(u)}},\end{aligned} \tag{4.11}$$

until we hit tip node i.

With a simple algorithm to compute the mean and precision of the full conditional distribution $p(\mathbf{X}_i \,|\, \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ at our disposal, we finally turn our attention toward a Metropolis-Hastings scheme to sample $\mathbf{X}_i$. The algorithm must only generate samples for the latent liabilities $\mathbf{X}_{i(-c)}$ corresponding to the discrete traits, since the map function $g(\cdot)$ fixes the latent liabilities $\mathbf{X}_{ic}$ for all the continuous traits. Thus we consider the proposal distribution $p(\mathbf{X}_{i(-c)} \,|\, \mathbf{X}_{ic}, \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$, which is obtained from $p(\mathbf{X}_i \,|\, \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ by further conditioning on the fixed liabilities $\mathbf{X}_{ic}$. This conditional distribution is MVN $(\, \mathbf{X}_{ic}\, ; \mathbf{m}_i^{\text{cond}}, p_i^{\text{pre}} \mathbf{W}_{cc})$, where

$$\mathbf{m}_i^{\text{cond}} = \mathbf{m}_{i(-c)}^{\text{pre}} - \mathbf{W}_{cc}^{-1} \mathbf{W}_{c(-c)} \left(\mathbf{X}_{i(-c)} - \mathbf{m}_{i(-c)}^{\text{pre}} \right). \tag{4.12}$$

Here the vector $\mathbf{m}_{i(-c)}^{\text{pre}} = (\mathbf{m}_{i(-c)}^{\text{pre}}, \mathbf{m}_{ic}^{\text{pre}})$ is partitioned according to correspondence to continuous traits, as is the precision matrix for the diffusion process

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \mathbf{W}_{(-c)(-c)} & \mathbf{W}_{(-c)c} \\ \mathbf{W}_{c(-c)} & \mathbf{W}_{cc} \end{pmatrix}. \tag{4.13}$$

Several approaches compete for generating truncated multivariate normal random variables, including rejection sampling (Breslaw, 1994; Robert, 1995) and Gibbs sampling (Gelfand et al, 1992; Robert, 1995) possibly with data augmentation (Damien and Walker, 2001). For the examples we explore in this manuscript, the dimension D of $\mathbf{X}_i$ can be large, ranging up to 54 with $N = 360$ tips, with occasionally high correlation in $\boldsymbol{\Sigma}$. Gibbs sampling can suffer from slow convergence in the presence of high correlation between dimensions. Consequentially, we explore an extension of rejection sampling that involves a multiple-try Metropolis (Liu et al, 2000) construction. We simulate up to R draws

$\mathbf{X}_i^{(r)} \sim p(\cdot \,|\, \mathbf{X}_{(-i)}, \mathbf{V}(F), \mathbf{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$. For draw $\mathbf{X}_i^{(r)}$, if $p(\mathbf{X}_i^{(r)} \,|\, \mathbf{Y}_i, \mathbf{A}) \neq 0$, then we accept this value as our next realization of $\mathbf{X}_i$. Appendix 4.5 demonstrates that the Metropolis-Hastings acceptance probability of this action is 1. If all R proposals return 0 density, the MCMC chain remains at its current location.

In our largest example, we briefly evaluate one approach to select R. We start with a very large $R = 10000$ and observe that most proposals that lead to state changes occur in the first 20 attempts; further, after 100 attempts, the residual probability of generating a valid sample becomes negligible. Thus, we set $R = 100$ for future MCMC simulation. As MCMC chains converge towards the posterior distribution, the probably of generating a valid sample approaches the 75 – 85% range in our examples. Finally, we employ a Metropolis-Hastings scheme to sample $\mathbf{A}$ in which the proposal distribution is a uniform window centered at the parameter's current value with a tunable length.

4.2.4 Correlation Testing and Model Selection

To assess the phenotypic relationship between two specific components of the trait vector $\mathbf{Y}$, we look at the correlation of the corresponding elements in the latent variable $\mathbf{X}$. One straight-forward approach entertains the use of the marginal posterior distribution of pair-wise correlation coefficients $\rho_{jj'}$ determined from $\mathbf{\Sigma}$. As a simple rule-of-thumb, we designate $\rho_{jj'}$ significantly non-zero if $> 99\%$ of its posterior mass falls strictly greater than or strictly less than 0.

When scientific interest lies in formal comparison of models that involve more than pair-wise effects, we employ Bayes factors. Possible examples include identifying block-diagonal structures in $\mathbf{\Sigma}$, comparing the latent liability model to other trait evolution models and, as demonstrated in our examples, state-ordering of multiple discrete traits.

The Bayes factor that compares models M_0 and M_1 can be obtained as

$$B_{01} = \frac{p(\mathbf{Y}, \mathbf{S}|M_0)}{p(\mathbf{Y}, \mathbf{S}|M_1)}, \tag{4.14}$$

in which $p(\mathbf{Y}, \mathbf{S}|M)$ is the marginal likelihood of the data under model M (Jeffreys, 1935). Computing these marginal likelihoods is not straightforward, involving high dimensional integration. We adopt a path sampling approach which estimates these integrals through numerical integration.

While estimating the Bayes factor directly by integrating along a path that goes from model M_0 to model M_1 is possibly a good strategy for comparing nested or closely related models, it does not present the same flexibility as estimating individual marginal likelihoods. Individual marginal likelihoods can be efficiently used for comparisons between multiple models. Additionally, this strategy is better suited for comparisons between models defined on different parameter spaces. For this reason, we pursue the latter.

To estimate the marginal likelihoods in (4.14), we follow Baele et al (2012) in considering a geometric path $q_\beta(\mathbf{Y}, \mathrm{S}; \mathbf{X}, \boldsymbol{\theta})$ that goes from a normalized source distribution $q_0(\mathbf{Y}, \mathrm{S}; \mathbf{X}, \boldsymbol{\theta})$ to the unnormalized posterior distribution $p(\mathbf{Y}, \mathrm{S}|\mathbf{X}, \boldsymbol{\theta})p(\mathbf{X}, \boldsymbol{\theta})$. Here both distributions are defined on the same parameter space, and $\boldsymbol{\theta} = \{\boldsymbol{\Sigma}, F, \boldsymbol{\phi}, \mathbf{A}\}$ collects all model parameters. The path sampling algorithm employs MCMC to numerically compute the path integral

$$\log(p(\mathbf{Y}, \mathbf{S}|M)) = \int_0^1 E_{q_\beta} \left[\log(q_1(\mathbf{Y}, \mathrm{S}; \mathbf{X}, \boldsymbol{\theta})) - \log(q_0(\mathbf{Y}, \mathrm{S}; \mathbf{X}, \boldsymbol{\theta}))\right] \mathrm{d}\beta. \tag{4.15}$$

A natural choice for the source distribution is the prior $p(\mathbf{X}, \boldsymbol{\theta})$. However, due to truncations in the distribution of $\mathbf{X}$ induced by the map function $g(\cdot)$, the path from the prior to the unnormalized posterior is not continuous. Since continuity along the whole path is required for (4.15) to hold, we propose here a different destination distribution that leads to a continuous path. Let

$$q_0(\mathbf{Y}, \mathbf{S}; \mathbf{X}, \boldsymbol{\theta}) = p(\mathbf{X}|\mathbf{Y}, \mathbf{A})\psi(\mathbf{X})p(\boldsymbol{\theta}), \tag{4.16}$$

where $p(\boldsymbol{\theta})$ is the prior, $p(\mathbf{X} \mid \mathbf{Y}, \mathbf{A}) = \mathbf{1}_{(\mathbf{Y}=g(\mathbf{X}))}$, and $\psi(\mathbf{X})$ is a function proportional to a conveniently chosen matrix normal distribution. The proportionality constant of $\psi(\mathbf{X})$ is selected to guarantee

$$\int p(\mathbf{X} \mid \mathbf{Y}, \mathbf{A})\psi(\mathbf{X})\mathrm{d}\mathbf{X} = 1, \tag{4.17}$$

and thus a normalized source distribution $q_0(\mathbf{Y}, \mathbf{S}; \mathbf{X}, \boldsymbol{\theta})$.

The choice of function $\psi(\mathbf{X}) = \psi^*(\mathbf{X})/Q(\mathbf{Y}, \mathbf{A})$ is central to the success of this path sampling approach. We select the matrix normal distribution $\psi^*(\mathbf{X})$ so that all entries in $\mathbf{X}$ are independent, and consequently the proportionality constant is

$$Q(\mathbf{Y}, \mathbf{A}) = \prod_{i=1}^{N}\prod_{j=0}^{P} Q(y_{ij}, \mathbf{A}) = \prod_{i=1}^{N}\prod_{j=0}^{P} \int p(\mathbf{X}_{ij^*} \mid y_{ij}, \mathbf{A})\psi^*(\mathbf{X}_{ij^*})\mathrm{d}\mathbf{X}_{ij^*}, \tag{4.18}$$

where $\mathbf{X}_{ij^*}$ are all the entries of the latent liability corresponding to y_{ij}.

For binary traits, $\mathbf{X}_{ij^*}$ is univariate, and $\psi(\mathbf{X}_{ij^*})$ is proportional to a normal distribution whose mean $\bar{X}_{ij^*}$ and variance $\bar{\sigma}^2_{ij^*}$ match those of the posterior distribution of $\mathbf{X}_{ij^*}$. Considering that the map function $g(\cdot)$ restricts $\mathbf{X}_{ij^*}$ to be larger (or smaller) than 0, and that

$\bar{X}_{ij^*}$ always belongs to this valid region, the proportionality constant for a binary trait is

$$Q(\mathbf{Y}_{ij}, \mathbf{A}) = \Phi\left(\frac{|\bar{X}_{ij^*}|}{\bar{\sigma}_{ij^*}}\right), \tag{4.19}$$

where $\Phi\left(\cdot\right)$ is the cumulative distribution function (CDF) of the standard normal distribution.

For traits with $K \geq 3$ ordered states, $\mathbf{X}_{ij^*}$ is also univariate, and we make the same choice for mean and variance parameters of $\psi^*(\mathbf{X}_{ij^*})$. The map function depends on the threshold parameters $\mathbf{A}$, that must be fixed for this analysis. If $a_l(y_{ij})$ and $a_u(y_{ij})$ denote respectively the lower and upper threshold for the valid region mapped from y_{ij}, then the proportionality constant becomes

$$Q(y_{ij}, \mathbf{A}) = \Phi\left(\frac{a_u(y_{ij}) - \bar{X}_{ij^*}}{\bar{\sigma}_{ij^*}}\right) - \Phi\left(\frac{a_l(y_{ij}) - \bar{X}_{ij^*}}{\bar{\sigma}_{ij^*}}\right). \tag{4.20}$$

When y_{ij} assumes one of the extreme states s_1 and s_K, then the normalizing constant considers the appropriate open interval.

For discrete data with $K \geq 3$ unordered states, y_{ij} maps from $K-1$ dimensions in $\mathbf{Y}$. For simplicity, $\psi^*(\mathbf{X}_{ij^*})$ is a standard multivariate normal distribution, and the proportionality constant is

$$Q(y_{ij}, \mathbf{A}) = \begin{cases} 2^{-(K-1)} & \text{if} \quad y_{ij} = s_1 \\ \frac{1-2^{-(K-1)}}{K-1} & \text{if} \quad y_{ij} = s_2, \cdots, s_K. \end{cases} \tag{4.21}$$

Finally, for continuous y_{ij} we simply have $\psi(\mathbf{X}_{ij^*}) = y_{ij}$.

Implementation

The methods described in this chapter have been implemented in the software package BEAST (Drummond et al, 2012).

4.3 Real-World Examples

We present applications of our model to three problems in which researchers wish to assess correlation between different types of traits while controlling for their shared evolutionary history.

4.3.1 Antimicrobial resistance in *Salmonella*

Development of multidrug resistance in pathogenic bacteria is a serious public health burden. Understanding the relationships between resistance to different drugs throughout bacterial evolution can help shed light on the fundamentals of multidrug resistance on the epidemiological scale.

We use the phylogenetic latent liability model to assess phenotypic correlation between resistance traits to 13 different antibiotics in *Salmonella*. We analyse 248 isolates of *Salmonella* Typhimurium DT104, obtained from animals and humans in Scotland between 1990 and 2011 (Mather et al, 2013). For each isolate, we have sequence data and binary phenotypic data indicating the strains resistance status to each of the 13 antibiotics.

To assess which resistance traits tend to be associated, we examine the correlation matrix of the latent liabilities $\mathbf{X}$. Because the trait data are binary, the underlying latent variables $\mathbf{X}_i$ for this problem are $D = 13$-dimensional, with each entry corresponding to resistance to one antibiotic. To highlight the main correlation structure of $\mathbf{\Sigma}$, Figure 4.3 presents a heatmap of the significantly non-zero pair-wise correlation coefficients. This matrix contains only positive correlations, consistent with genetic linkage between resistance traits. Additionally, the significant correlations form a block-like structure. Table 4.3 presents posterior mean and 95% BCI estimates for all correlations between resistance traits. Estimates of non-significant correlations tend to be slightly positive, with the exception of correlations involving resistance to ciprofloxacin.

Our analysis reveals a block of strong positive correlations between resistance traits to the antibiotics tetracycline, ampicillin, chloramphenicol, spectinomycin, streptomycin and sulfamethoxazole (sulfonamide), similar to those found using a simpler model (Mather et al, 2012). We estimate a posterior probability > 0.9999 for positive correlation between all these resistance traits simultaneously. This block is consistent with the *Salmonella* genomic island 1 (SGI-1), a 43-kb genomic island conferring multidrug resistance. Among the drugs considered here, SGI-1 confers resistance to these 6 antibiotics (Boyd et al, 2001).

Another pair of antibiotic resistance traits that we infer to be strongly correlated are gentamicin and netilmicin, with a 95% BCI of [0.80, 0.98]. These drugs are both aminoglycoside antibiotics, and the same genes may confer resistance to both antibiotics. These drugs also appear correlated with some of the resistance traits connected to SGI-1.

Although previous analysis of this dataset has revealed that most of the evolutionary history that these data capture was spent in human hosts, human-to-animal or animal-to-human transitions do occur across the tree (Mather et al, 2013). We investigate whether these interspecies transitions also correlate with antibiotic resistance. To do so, we include host species (animal/human) as a 14th binary trait under in latent liability model. None of the pair-wise correlations are significantly non-zero given our rule-of-thumb definition. Table 4.2 contains estimated correlations to the host trait.

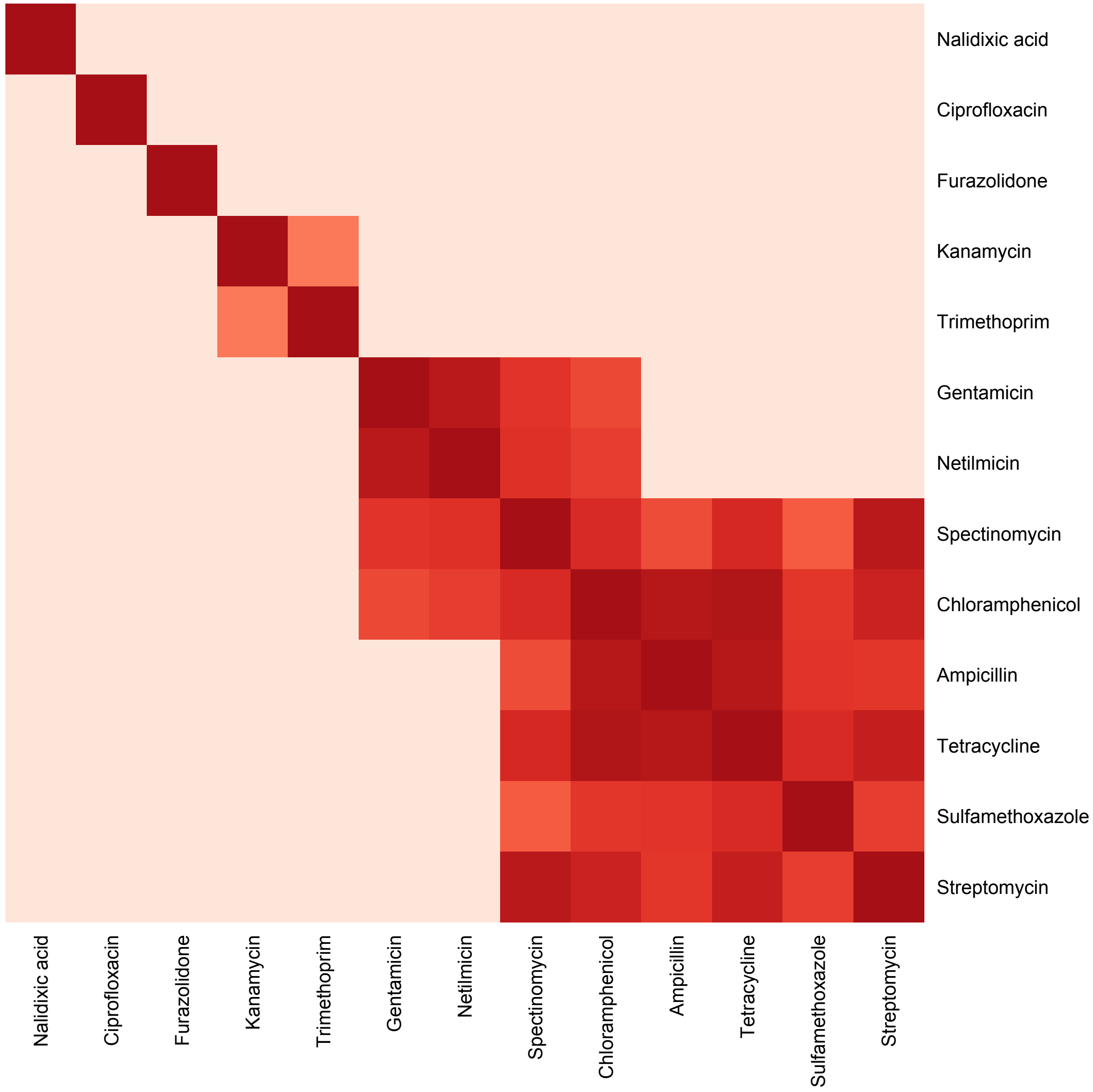


Figure 4.3: Heatmap of posterior means for significantly non-zero correlations between antibiotic resistance traits for the latent liability model. Darker colors indicate stronger positive correlation.

4.3.2 Columbine flower evolution

The flowers of columbine genus *Aquilegia* have attracted several different pollinators throughout their evolutionary history. One question that remains is the exact role the pollinators play in the tempo of columbine flower evolution (Whittall and Hodges, 2007). Since different pollinator species demonstrate distinct preferences for flower morphology and color, we investigate here how these traits correlate over the evolutionary history of *Aquilegia*.

We analyse $P = 12$ different floral traits for $N = 30$ monophyletic populations from the genus *Aquilegia*. Of these traits, 10 are continuous and represent color, length and orientation of different anatomical features of the flowers. Additionally, we consider a binary trait that indicates presence or absence of anthocyanin pigment; and another discrete trait that indicates the primary pollinator for that population. The pollinator trait was assessed through a combination of floral characters known as pollination syndrome. The prevailing hypothesis is that evolutionary transitions from bumblebee-pollinated flowers (Bb) to those primarily pollinated by hawkmoths (Hm) are obligated to pass through an intermediate stage of hummingbird-pollination (Hb) (Whittall and Hodges, 2007), we treat pollinators as ordered states, but we formally test alternative orderings. Taken together, this results in a latent liability model with $D = 12$ dimensions. As sequence data are not readily available for all the taxa included in this analysis, we consider for our analysis the same fixed phylogenetic tree used in Whittall and Hodges (2007). The ability to either condition on a fixed phylogeny F or integrate over a random F in a single framework presents a strength in a field that has traditionally focused on either genetic or phenotypic data alone and joint datasets are an emerging addition. Whittall et al (2006) and Whittall and Hodges (2007) have published the original data that are available on the Bodega phylogenetics website (`http://bodegaphylo.wikispot.org`).

To draw inference on the phenotypic correlation structure of these traits, we focus on the

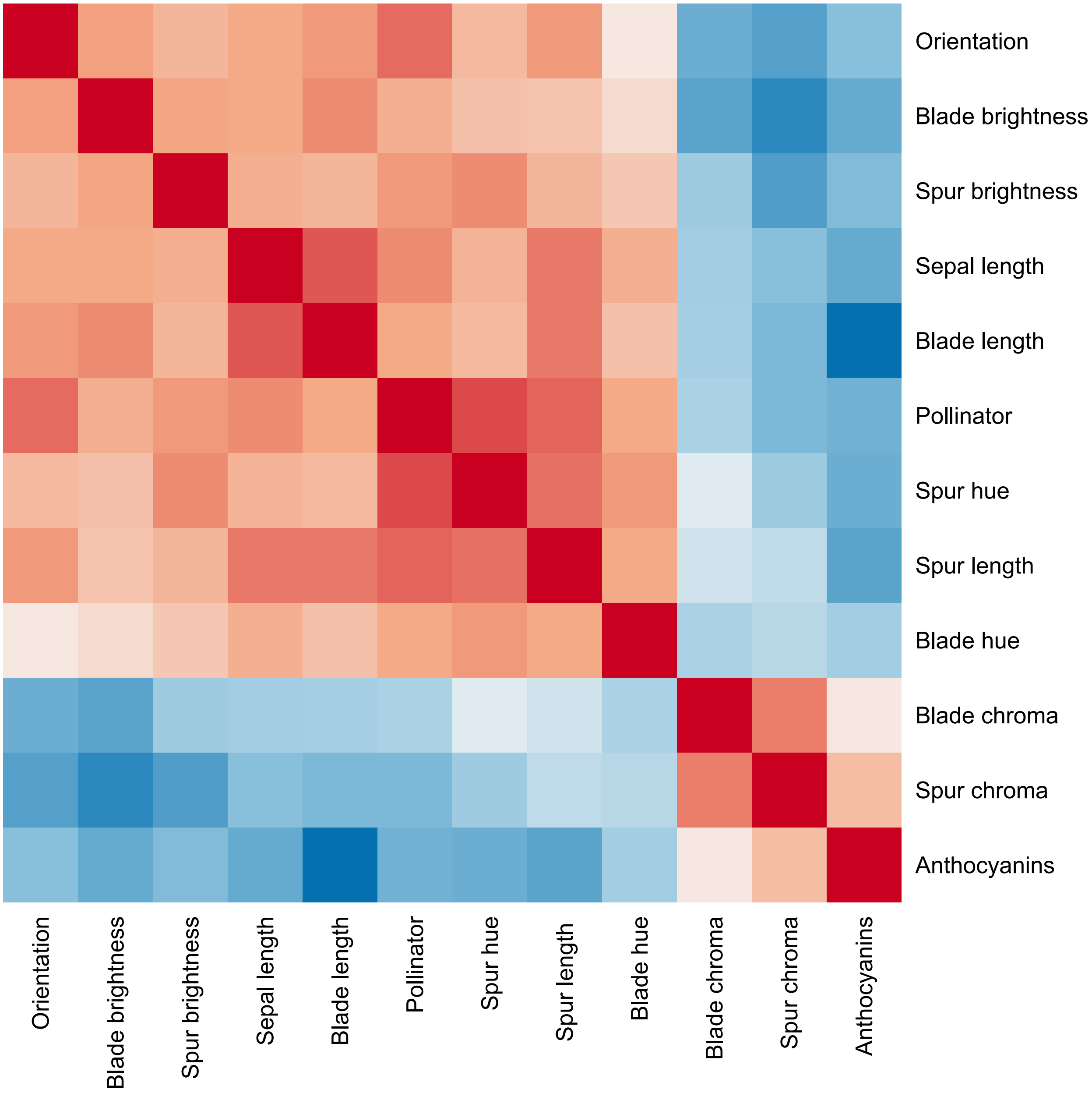


Figure 4.4: Heatmap of the posterior mean for the phenotypic correlation of columbine floral traits in the latent liability model. Darker colors indicate stronger correlations; shades of red for positive correlation and blue for negative correlation.

12×12 variance matrix $\mathbf{\Sigma}$ of the Brownian motion process that governs the evolution of $\mathbf{X}$ on the tree. We report posterior mean and BCI estimates for all pair-wise correlations in $\mathbf{\Sigma}$ in Table 4.4. Figure 4.4 presents a heatmap of the posterior means of the correlations, in which blue represents negative phenotypic correlation, and red represents positive ones. Our analysis reveals a strong block correlation structure between the floral traits. We find one block of positive correlation between chroma of both spur and blade and the presence of anthocyanins. All other color and morphological traits in the analysis form a second block of positive correlation. Additionally, phenotypic correlation between the first and second trait blocks are all negative.

Whittall and Hodges (2007) highlight the relationship between changes in pollinators and increases in the length of floral spurs. They argue that flowers with longer spurs are only pollinated by the hawkmoths, because the other pollinators with shorter tongues cannot access and feed on the nectar contained at the end of the spur. Here we estimate a positive correlation between pollinators and spur length, with a posterior mean of 0.76, and a 95% BCI of [0.60; 0.88], consistent with their findings.

The pollinator trait has $K = 3$ ordered states and, under the latent liability model, one dimension in $\mathbf{X}$ determines the trait outcome. Specifically, the position of this column relative to threshold parameters $a_1 = 0$ and a_2 determine the outcome. Consequently, our estimate of a_2 is instrumental in determining the relative probabilities of the states in our model and the inferred trait state at the root of the tree. We estimate a_2 to have a posterior mean of 3.00 with a 95% BCI of [1.14; 5.34].

The bumblebee $\leftrightarrow$ hummingbird $\leftrightarrow$ hawkmoth (Bb-Hb-Hm) orderings is only one of several, and alternative hypotheses regarding pollinator adaptation could be proposed (van der Niet and Johnson, 2012; Smith et al, 2008b). We examine whether the data support the Bb-Hb-Hm ordering, or if there is another model with a better fit. We use the Bayes factors

Table 4.1: Model selection for the ordering of bumblebee (Bb), hummingbird (Hb) and hawkmoth (Hm) pollinators in Columbine flowers.

	log Marginal	log Bayes Factor		
Order	Likelihood	Hm-Bb-Hb	Hb-Hm-Bb	unordered
Bb-Hb-Hm	-11.2	9.4	14.2	24.8
Hm-Bb-Hb	-20.6	-	4.8	15.3
Hb-Hm-Bb	-25.4	-	-	10.5
unordered	-36.0	-	-	-

to compare four different models for the pollinator trait: the Bb-Hb-Hm, Hb-Hm-Bb, Hm-Bb-Hb, and an unordered formulation. Note that there are only three possible orderings for a $K = 3$ state ordered latent liability model since, for symmetric models such as Bb-Hb-Hm and Hm-Hb-Bb, inverting the order leads to equivalent models with inverted signs for the latent traits. The unordered model leads to a bivariate contribution to latent liabiliy $\mathbf{X}$. Table 4.1 presents the path sampling estimates for the marginal likelihood of each model and the corresponding Bayes factors. These comparisons indicate that, in agreement with Whittall and Hodges (2007), the data strongly support the Bb-Hb-Hm model.

Our latent liability model estimates correlation between traits while accounting for shared evolutionary history. To evaluate the effect that phylogenetic relatedness has on our estimates of the trait correlation structure, we estimated the same correlation under a latent liability model with no phylogenetic structure. In this analysis, a star tree with identical distance between all taxa was used. Table 4.5 presents these correlation estimates and the corresponding 95% BCI. Comparing these results to the original latent liability analysis that accounts for shared evolutionary history, we noticed that most estimates were consistent between the both analyses, with a mean absolute difference for posterior means of correlation of 0.11. However, for three of the pairwise correlations (anthocianins × orientation, orientation × blade length, spur length × spur hue) the BIC‘s for the model that does not account for shared evolution did not contain the posterior mean for the evolutionary model. In particular, the evolutionary model estimates a significantly weaker correlation between spur length and spur hue (posterior mean of 0.55) than does the model that does not account for shared history, with a 95% BCI of [0.63; 0.87].

4.3.3 Correlation within and across influenza epitopes

In influenza, the viral surface proteins hemagglutinin (HA) and neuraminidase provide the antigenic epitopes to which the host immune system responds. Rapid mutation of these proteins to evade immune response, known as antigenic drift, severely challenges the design of annual influenza vaccines. The epitope regions in these genes are particularly important to the drift process (Fitch et al, 1991; Plotkin and Dushoff, 2003). In this context, we are interested in studying the phenotypic correlation among the amino acid sites of these epitopes, because the identification of correlated amino acids grants insight into the dynamics of antigenic drift in influenza.

The HA protein has five identified epitopes A-E, each containing around 20 amino acids. We focus on epitopes A and B, because these are the most immunologically stimulating for most influenza strains (Bush et al, 1999; Cox and Bender, 1995). We analyse sequence data for 180 strains of human H3N2 influenza dating from 1995 to 2012, obtained from the Influenza Research Database (`http://www.fludb.org`) and selected to promote geographic diversity. We use the amino acid information in epitope A and B for the latent liability part of the model, and the remaining sequence data in a standard phylogenetic approach to inform the tree structure.

Of the 40 amino acid sites in epitopes A and B of the HA protein, we find 17 to be variable in our sample. The number of unique amino acids in these sites varies between $K = 2$ and $K = 5$. Through a preliminary survey of a larger sample of influenza strains (900 samples) from the same period we find that all polymorphic sites for which the major allele

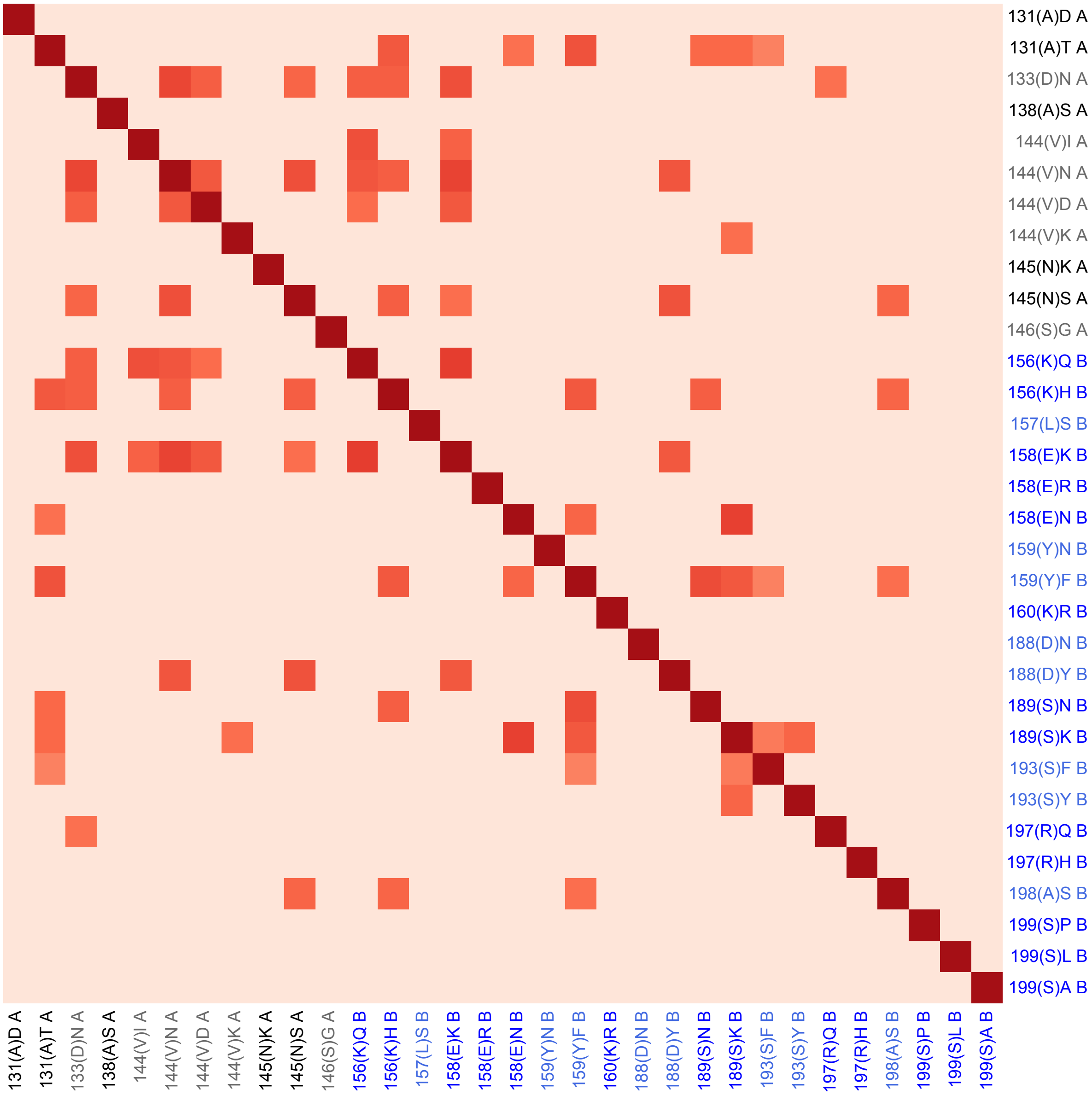


Figure 4.5: Heatmap of the posterior mean for the non-zero phenotypic correlations of amino acids in H3N2 epitopes A and B in the latent liability model. Darker colors indicate stronger correlation. We list the sites as follows: the number of the amino acid site in the aligned sequence; the one letter code for the reference amino acid for the site, in parenthesis; the code for the amino acid corresponding to the latent trait; and the epitope to which the site belongs.

frequency is $< 99\%$ are also variable in our 180 sequence sample, strongly suggesting that our limited dataset contains information about all the common variant sites in epitopes A and B during this period.

We model these data with the latent liability model for multiple unordered states. For each amino acid site, we have $K-1$ corresponding latent traits, yielding a total of $D = 32$ latent dimensions in $\mathbf{X}$. Without loss of generality, we take the amino acid observed in the oldest sequence of the sample as the reference state, and each entry of the latent liability column corresponds to one of the other amino acid variants for that site.

To assess the phenotypic correlation structure between sites in epitopes A and B, we estimate the correlation matrix associated with $\mathbf{\Sigma}$ of the latent liability $\mathbf{X}$. Figure 4.5 presents a heatmap with the pairwise correlations for the significantly non-zero estimates. The arrangement of the states follows the order of the sites in the primary amino acid sequence, even though the sites are not necessarily contiguous in folded protein-space.

Our analysis suggests a group of 10 sites that are strongly correlated with each other. This group includes all the sites identified by Koel et al (2013) as being the major determinants of antigenic drift that are polymorphic in our sample. We do not find preferential correlations within epitopes.

Table 4.6 presents a list with point estimates and 95% BCI of correlations whose credible intervals do not include zero. All correlations in this list are positive and point estimates range from 0.6 to 0.74. Since, for all sites the oldest variant was taken as the reference state, a positive correlation between two latent traits could be seen as association between novel amino acids in both sites. The strongest evidence for correlation was found between sites 158(E)K and 156(K)Q, with an estimated correlation coefficient of 0.74 (95% BIC of [0.40, 0.93]). Koel et al (2013) identified these specific mutations in both sites as being the main drivers of an event of major antigenic change that took place between 1995 and

1997. Mutations in sites 159 and 189 are another example of a pair of substitutions identified as driving an event of major antigenic change taking place in the late 1980's. Even though the oldest sequence in our sample only dates back to 1995, correlation between these two sites remains strongly supported by our analysis, with an estimated correlation coefficient between 159(Y)F and 189(S)N of 0.69 (95% BIC of [0.27, 0.92]).

4.4 Discussion

We present the phylogenetic latent liability model as a framework for assessing phenotypic correlation between different types of data. Through our three applications, we illustrate the use of our methodology for binary data, discrete data with multiple ordered and unordered states, continuous data and combinations thereof. The applications exemplify current biological problems which our method can naturally address. Additionally, we show how the model can be used to reveal the overall phenotypic correlation structure of the data, and we provide tools to test hypotheses about individual correlations and for general model testing.

The threshold structure of the phylogenetic latent liability model renders it non-Markovian for the discrete traits. Both Felsenstein (2005, 2012) and Revell (2013) argue that this is actually a valuable property for many phenotypic traits for which the probability of transitioning between states should vary depending on the time spent at that state. Based on this argument, Revell (2013) investigates ancestral state reconstruction for univariate ordered traits under the threshold model, and finds consistent reconstructions for simulated data. For our model, it would be straightforward to perform ancestral state estimation for multivariate traits of all types considered, since the inference machinery is already implemented in BEAST.

A problem with many comparative biology methods for phenotypic correlation is the requirement for a fixed tree. Through sequence data, our model can account for the uncertainty of tree estimation by integrating over the space of phylogenetic trees, as we do for the influenza epitope and antibiotic resistance examples.

As a caveat for this type of model, Felsenstein (2012) points out a general lack of power, arguing that for real size datasets confidence intervals would be too large. This issue could be magnified on discrete traits, since the correlations are an extra step removed from the data. In our applications, the size of our posterior credible intervals are relatively large for intervals constrained between -1 and 1. However, this did not prevent us from recovering general correlation patterns and identifying important correlations. Moreover, for the columbine flower example, we find no difference in average size of credible intervals for correlations including latent traits and those between two continuous traits.

Analytically integrating out continuous trait values at root and internal nodes to compute the likelihood of Brownian motion on a tree leads to significant improvement in efficiency of inference methods (Pybus et al, 2012). This strategy computes successive conditional likelihoods by a post-order tree traversal in a procedure akin to Felsenstein's peeling algorithm (Felsenstein, 1981a). Its effectiveness has been explored in similar contexts in univariate (Novembre and Slatkin, 2009; Blum et al, 2004) and multivariate Brownian motion (Freckleton, 2012) and even to estimate the Gaussian component of Lévy processes (Landis et al, 2013). A related post-order traversal approach has been used to improve computation in the context of phylogenetic regressions for some Gaussian and non-Gaussian models (Ho and Ané, 2014). Unfortunately, a similar solution is not available to marginalize the latent liability $\mathbf{X}$ at the tips of the tree in our model. Consequently this integration must be performed by MCMC. Integration for $\mathbf{X}$ is a critical part of our method, and for large datasets, mixing becomes a problem. To address this issue, we present an efficient sampler that, at each iteration, updates all components of the multivariate latent variable $\mathbf{X}$ at one tip of the tree. This algorithm builds upon the dynamic programming strategy of Pybus et al (2012) to obtain a truncated multivariate normal as the full conditional distribution of $\mathbf{X}_i$. Even though sampling from this truncated distribution still requires an accept/reject step that could be highly inefficient, we find that as the chain approaches equilibrium, rejection rates tend to become small.

In our analysis of influenza epitopes, we set the oldest amino acid observed for each site as the reference state, and for each of the remaining variants we assigned an entry in $\mathbf{X}$. For the multiple unordered states model, this choice results in a reduction of dimensionality in the problem, but is done mainly to improve identifiability. However, this procedure breaks the symmetry of the model and complicates interpretability of correlations. In fact, a correlation between two entries of the latent trait $\mathbf{X}$ cannot be directly translated as a correlation between the states they represent, since variations in an entry of $\mathbf{X}$ are linked to all other states for that trait through the reference state. Despite this caveat, general statements about the correlation structure of the data can still be made based on the latent liability $\mathbf{X}$, as we show in the influenza epitopes application.

In this context, different model choices could be used to change the interpretational links between correlations in $\mathbf{X}$ and in the data. Hadfield and Nakagawa (2010) briefly discuss a multinomial phylogenetic mixture model where a latent variable determines the probability of the multinomial outcome. They consider the common choice of constraining the latent variable to a simplex by setting the sum of its components to one. This makes the value of the latent trait immediately interpretable as probabilities, however it further complicates interpretability of the correlations. A possible alternative to address this issue is to model the evolution of $\mathbf{X}$ in the latent liability model with a central tendency such as the Ornstein-Uhlenbeck process. It remains to be investigated whether this would improve identifiability, eliminating the need to impose constraints on the model.

Lartillot and Poujol (2011) have studied the correlation between continuous traits and parameters of the molecular evolution model, such as dS/dN ratio and mutation rate, by modelling the evolution of these parameters as a diffusion process along the tree. One possible extension to our method would be to incorporate the evolution of these parameters in our model, allowing for the estimation of correlations between our continuous and discrete traits and these evolutionary parameters.

The Bayesian phylogenetic framework in which we integrate our model easily lends itself to combination of different models. These could be phylogenetic models for demographic inference (Minin et al, 2008), methods for calibrating trees and relaxed clock models (Drummond et al, 2006). Additionally, we can explore the relaxed random walk (Lemey et al, 2010) to get varying rates of trait evolution along different branches of the tree. The latent liability model can easily be associated with these existing models to provide comprehensive analyses.

4.5 Appendix: Sampling repeatedly from truncated multivariate normal and Metropolis-Hastings acceptance ratio

In this section we obtain the Metropolis-Hastings ratio for the multiple-try Metropolis algorithm pretended in subsection 4.2.3. Proposals are designed as follows: draw repeated samples from $\mathbf{X}_i^{(r)} \sim p(\cdot \mid \mathbf{X}_{(-i)}, \mathbf{V}(F), \mathbf{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$, stop when $p(\mathbf{X}_i^r \mid \mathbf{Y}_i) \neq 0$ or after R attempts. To compute the acceptance ratio for this Metropolis-Hastings algorithm, we must first obtain the proposal distributions $q(\mathbf{X}_i^* | \mathbf{X}_i^k)$. Notice, however, that $p(\cdot \mid \mathbf{X}_{(-i)}, \mathbf{V}(F), \mathbf{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ does not depend on the current state $\mathbf{X}_i^k$, thus the proposal distribution can be written as $q(\mathbf{X}_i^*)$. For the remainder of this section, let the parameter $\boldsymbol{\theta} = \{\mathbf{V}(F), \mathbf{\Sigma}, \boldsymbol{\mu}_0, \tau_0\}$.

The outcome of the proposal can be divided in two groups. In the first, we reach R'th attempt without proposing a valid value for $\mathbf{X}_i$. In this case, an invalid value will be proposed and automatically rejected (since it's likelihood is zero). In the second group, a valid value is proposed in one of the R attempts. The probability of proposing a valid value is

$$P(\text{valid}) = \sum_{j=0}^{R-1} \left[\int_{/\mathcal{V}} p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \boldsymbol{\theta}) \mathrm{d}\mathbf{X}_i \right]^j \int_{\mathcal{V}} p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \boldsymbol{\theta}) \mathrm{d}\mathbf{X}_i,$$

where $\mathcal{V}$ represents the valid region and $/\mathcal{V}$ its complement. If $\mathbf{X}_i^*$ is a valid value we have

$$\begin{aligned} q(\mathbf{X}_i^*) &= p(\text{valid}, \mathbf{X}_i^*) = p(\mathbf{X}_i^* | \text{valid}) p(\text{valid}) \\ &= \frac{p(\mathbf{X}_i^* \mid \mathbf{X}_{(-i)}, \boldsymbol{\theta})}{\int_{\mathcal{V}} p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \boldsymbol{\theta}) \mathrm{d}\mathbf{X}_i} \sum_{j=0}^{R-1} \left[\int_{/\mathcal{V}} p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \boldsymbol{\theta}) \mathrm{d}\mathbf{X}_i \right]^j \int_{\mathcal{V}} p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \boldsymbol{\theta}) \mathrm{d}\mathbf{X}_i \\ &= p(\mathbf{X}_i^* \mid \mathbf{X}_{(-i)}, \boldsymbol{\theta}) \sum_{j=0}^{R-1} \left[\int_{/\mathcal{V}} p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \boldsymbol{\theta}) \mathrm{d}\mathbf{X}_i \right]^j . \end{aligned}$$

Notice that the sum of integrals above does not depend on $\mathbf{X}_i^k$, so the Hastings ratio becomes

$$HR = \frac{q(\mathbf{X}_i^k)}{q(\mathbf{X}_i^*)} = \frac{p(\mathbf{X}_i^k \,|\, \mathbf{X}_{(-i)}, \boldsymbol{\theta})}{p(\mathbf{X}_i^* \,|\, \mathbf{X}_{(-i)}, \boldsymbol{\theta})},$$

which does not depend on the maximum number of attempts R. The acceptance ratio for the Metropolis-Hastings algorithm is then

$$\begin{aligned} AR &= \frac{q(\mathbf{X}_i^k)}{q(\mathbf{X}_i^*)} \frac{p(\mathbf{X}^*, \mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{X}_i^k, \mathbf{Y}|\boldsymbol{\theta})} \\ &= \frac{p(\mathbf{X}_i^k \,|\, \mathbf{X}_{(-i)}, \boldsymbol{\theta})}{p(\mathbf{X}_i^* \,|\, \mathbf{X}_{(-i)}, \boldsymbol{\theta})} \frac{p(\mathbf{X}_i^* \,|\, \mathbf{X}_{(-i)}, \boldsymbol{\theta}) p(\mathbf{X}_i^*|\mathbf{Y}) P(\mathbf{X}_{(-i)}, \mathbf{Y}|\boldsymbol{\theta})}{p(\mathbf{X}_i^k \,|\, \mathbf{X}_{(-i)}, \boldsymbol{\theta}) p(\mathbf{X}_i^k|\mathbf{Y}) P(\mathbf{X}_{(-i)}, \mathbf{Y}|\boldsymbol{\theta})} = 1. \end{aligned}$$

Thus, if a valid value is proposed, the Metropolis-Hastings algorithm accepts it with probability 1.

Summary

In summary, our multiple-try Metropolis algorithm generates samples for the latent liability $\mathbf{X}$ at the tips as follows:

1. Randomly select a tip i to update.
2. Obtain the conditional distribution $p(\cdot \,|\, \mathbf{X}_{(-i)}, \boldsymbol{\theta})$:
   (a) Compute the partial mean vectors $\mathbf{m}_u^{\text{post}}$ and precision scalars p_u^{post} in the post order traversal:
      - For the tips $\mathbf{m}_u^{\text{post}} = \mathbf{X}_u$ and $p_u^{\text{post}} = 1/t_u$;
      - For the internal nodes, if $\nu_{d1(u)}$ and $\nu_{d2(u)}$ represent the two child nodes to

node ν_u, then

$$\mathbf{m}_u^{\text{post}} = \frac{p_{d1(u)}^{\text{post}} \mathbf{m}_{d1(u)}^{\text{post}} + p_{d2(u)}^{\text{post}} \mathbf{m}_{d2(u)}^{\text{post}}}{\mathbf{m}_{d1(u)}^{\text{post}} + \mathbf{m}_{d2(u)}^{\text{post}}}, \text{ and}$$
$$\frac{1}{p_u^{\text{post}}} = t_u + \frac{1}{p_{d1(u)}^{\text{post}} + p_{d2(u)}^{\text{post}}}, \tag{4.22}$$

(b) Compute the partial mean vectors $\mathbf{m}_u^{\text{pre}}$ and precision scalars p_u^{pre} in the path from the root to ν_u:

- For the root $\mathbf{m}_u^{\text{pre}} = \mu$ and $p_u^{\text{pre}} = 1/\phi$;
- For the other nodes:

$$\mathbf{m}_u^{\text{pre}} = \frac{p_{\text{sib}(u)}^{\text{post}} \mathbf{m}_{\text{sib}(u)}^{\text{post}} + p_{\text{pa}(u)}^{\text{pre}} \mathbf{m}_{\text{pa}(u)}^{\text{pre}}}{\mathbf{m}_{\text{sib}(u)}^{\text{post}} + \mathbf{m}_{\text{pa}(u)}^{\text{pre}}}, \text{ and}$$
$$\frac{1}{p_u^{\text{pre}}} = t_u + \frac{1}{p_{\text{sib}(u)}^{\text{post}} + p_{\text{pa}(u)}^{\text{pre}}}, \tag{4.23}$$

(c) Obtain $p(\cdot|\mathbf{X}_{(-i)}, \boldsymbol{\theta}) = MVN(\cdot; \mathbf{m}_i^{\text{pre}}, p_i^{\text{pre}}\mathbf{P})$.

3. PROPOSAL: Generate proposal $\mathbf{X}_i^*$ according to $q(\mathbf{X}_i)$ by repeatedly generating samples from $p(\cdot|\mathbf{X}_{(-i)}, \boldsymbol{\theta})$. Stop when $P(\mathbf{X}_i^*|\mathbf{Y}_i) = 1$ or after R attempts.

4. DECISION:

- accept $\mathbf{X}_i^{k+1} = \mathbf{X}_i^*$, if $p(\mathbf{X}_i^*|\mathbf{Y}_i) = 1$,
- reject $\mathbf{X}_i^*$ if $(\mathbf{X}_i^*|\mathbf{Y}_i) = 0$, and set $\mathbf{X}_i^{k+1} = \mathbf{X}_i^k$.

4.6 Appendix: Supplementary tables form the applications

Table 4.2: Posterior mean and 95% Bayesian credible interval (BCI) estimates for pairwise correlation between the host trait (Animal/Human) and the different antibiotic resistance traits.

	Correlation	95% BCI
Ampicillin	0.0349	[-0.2357, 0.3036]
Chloramphenicol	-0.0610	[-0.3055, 0.1903]
Ciprofloxacin	0.1505	[-0.3313, 0.6099]
Gentamicin	-0.3651	[-0.6893, -0.0086]
Kanamycin	-0.1578	[-0.4641, 0.1715]
Furazolidone	0.0001	[-0.3131, 0.3098]
Nalidixic acid	-0.0967	[-0.4199, 0.2439]
Netilmicin	-0.3315	[-0.6551, 0.0145]
Spectinomycin	-0.2696	[-0.5130, 0.0009]
Streptomycin	-0.1392	[-0.4020, 0.1375]
Sulphamethoxazole	0.1399	[-0.2104, 0.4768]
Tetracycline	-0.0142	[-0.2716, 0.2471]
Trimethoprim	0.0049	[-0.2888, 0.2976]

Table 4.3: Posterior mean and 95% Bayesian credible interval (BCI) estimates for latent liability model correlations between antibiotic resistance traits in *Sallmonela*.

	Ciprofloxacin		Furazolidone		Kanamycin		Trimethoprim		Gentamicin		Netilmicin	
	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI
Nalidixic acid	0.34	[-0.07, 0.68]	0.10	[-0.31, 0.48]	0.16	[-0.22, 0.51]	0.24	[-0.10, 0.54]	0.20	[-0.17, 0.54]	0.16	[-0.21, 0.51]
Ciprofloxacin	1.00	[1.00, 1.00]	-0.20	[-0.73, 0.42]	-0.15	[-0.67, 0.45]	0.06	[-0.39, 0.50]	-0.20	[-0.75, 0.47]	-0.23	[-0.77, 0.42]
Furazolidone	-	-	1.00	[1.00, 1.00]	0.18	[-0.26, 0.56]	0.21	[-0.17, 0.55]	0.53	[0.02, 0.84]	0.57	[0.06, 0.87]
Kanamycin	-	-	-	-	1.00	[1.00, 1.00]	0.52	[0.23, 0.76]	0.41	[0.06, 0.71]	0.42	[0.06, 0.72]
Trimethoprim	-	-	-	-	-	-	1.00	[1.00, 1.00]	0.20	[-0.16, 0.53]	0.21	[-0.15, 0.53]
Gentamicin	-	-	-	-	-	-	-	-	1.00	[1.00, 1.00]	0.92	[0.80, 0.98]
Netilmicin	-	-	-	-	-	-	-	-	-	-	1.00	[1.00, 1.00]
Spectinomycin	-	-	-	-	-	-	-	-	-	-	-	-
Chloramphenicol	-	-	-	-	-	-	-	-	-	-	-	-
Ampicillin	-	-	-	-	-	-	-	-	-	-	-	-
Tetracycline	-	-	-	-	-	-	-	-	-	-	-	-
Sulphamethoxazole	-	-	-	-	-	-	-	-	-	-	-	-
Streptomycin	-	-	-	-	-	-	-	-	-	-	-	-

	Spectinomycin		Chloramphenicol		Ampicillin		Tetracycline		Sulphamethoxazole		Streptomycin	
	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI
Nalidixic acid	0.21	[-0.13, 0.51]	0.07	[-0.28, 0.41]	0.04	[-0.33, 0.38]	0.10	[-0.26, 0.44]	0.18	[-0.25, 0.55]	0.19	[-0.17, 0.52]
Ciprofloxacin	-0.13	[-0.62, 0.42]	-0.29	[-0.73, 0.23]	-0.26	[-0.72, 0.28]	-0.22	[-0.69, 0.32]	-0.02	[-0.63, 0.57]	-0.14	[-0.65, 0.42]
Furazolidone	0.66	[0.04, 0.93]	0.79	[0.20, 0.97]	0.76	[0.10, 0.97]	0.77	[0.09, 0.97]	0.62	[-0.15, 0.91]	0.70	[0.00, 0.95]
Kanamycin	019	[-0.21, 0.54]	0.28	[-0.07, 0.58]	0.23	[-0.16, 0.57]	0.20	[-0.16, 0.53]	0.19	[-0.30, 0.64]	0.17	[-0.24, 0.54]
Trimethoprim	0.10	[-0.26, 0.43]	0.20	[-0.13, 0.49]	0.18	[-0.18, 0.49]	0.16	[-0.19, 0.47]	0.30	[-0.13, 0.67]	0.13	[-0.24, 0.46]
Gentamicin	0.78	[0.28, 0.97]	0.70	[0.20, 0.98]	0.58	[-0.02, 0.95]	0.66	[0.10, 0.97]	0.46	[-0.19, 0.90]	0.72	[0.15, 0.96]
Netilmicin	0.79	[0.30, 0.97]	0.75	[0.24, 0.98]	0.63	[0.01, 0.95]	0.70	[0.12, 0.97]	0.49	[-0.17, 0.90]	0.74	[0.17, 0.97]
Spectinomycin	1.00	[1.00, 1.00]	0.82	[0.67, 0.92]	0.70	[0.50, 0.85]	0.82	[0.68, 0.92]	0.64	[0.36, 0.85]	0.93	[0.84, 0.98]
Chloramphenicol	-	-	1.00	[1.00, 1.00]	0.94	[0.86, 0.98]	0.96	[0.91, 0.98]	0.76	[0.54, 0.91]	0.86	[0.73, 0.94]
Ampicillin	-	-	-	-	1.00	[1.00, 1.00]	0.93	[0.86, 0.98]	0.78	[0.55, 0.92]	0.77	[0.59, 0.89]
Tetracycline	-	-	-	-	-	-	1.00	[1.00, 1.00]	0.82	[0.63, 0.93]	0.89	[0.78, 0.96]
Sulphamethoxazole	-	-	-	-	-	-	-	-	1.00	[1.00, 1.00]	0.74	[0.50, 0.91]
Streptomycin	-	-	-	-	-	-	-	-	-	-	1.00	[1.00, 1.00]

Table 4.4: Posterior mean and 95% Bayesian credible interval (BCI) estimates for latent liability model correlations between floral traits in *Aquilegia*.

	Orientation		Blade brightness		Spur brightness		Sepal length		Blade length		Pollinator	
	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI
Orientation	1.00	[1.00; 1.00]	0.54	[0.32; 0.73]	0.42	[0.17; 0.64]	0.50	[0.26; 0.70]	0.55	[0.33; 0.73]	0.74	[0.56; 0.87]
Blade brightness	0.54	[0.32; 0.73]	1.00	[1.00; 1.00]	0.52	[0.29; 0.71]	0.49	[0.26; 0.69]	0.60	[0.39; 0.76]	0.50	[0.24; 0.70]
Spur brightness	0.42	[0.17; 0.64]	0.52	[0.29; 0.71]	1.00	[1.00; 1.00]	0.47	[0.22; 0.67]	0.43	[0.17; 0.64]	0.60	[0.38; 0.77]
Sepal length	0.50	[0.26; 0.70]	0.49	[0.26; 0.69]	0.47	[0.22; 0.67]	1.00	[1.00; 1.00]	0.74	[0.60; 0.85]	0.64	[0.38; 0.83]
Blade length	0.55	[0.33; 0.73]	0.60	[0.39; 0.76]	0.43	[0.17; 0.64]	0.74	[0.60; 0.85]	1.00	[1.00; 1.00]	0.55	[0.27; 0.78]
Pollinator	0.74	[0.56; 0.87]	0.50	[0.24; 0.70]	0.60	[0.38; 0.77]	0.64	[0.38; 0.83]	0.55	[0.27; 0.78]	1.00	[1.00; 1.00]
Spur hue	0.40	[0.14; 0.63]	0.36	[0.09; 0.59]	0.59	[0.38; 0.76]	0.44	[0.19; 0.65]	0.40	[0.14; 0.63]	0.86	[0.74; 0.94]
Spur length	0.56	[0.34; 0.73]	0.35	[0.08; 0.59]	0.43	[0.17; 0.65]	0.65	[0.47; 0.80]	0.65	[0.47; 0.80]	0.76	[0.60; 0.88]
Blade hue	0.15	[-0.14; 0.42]	0.22	[-0.06; 0.49]	0.33	[0.05; 0.57]	0.47	[0.22; 0.67]	0.36	[0.10; 0.60]	0.54	[0.28; 0.75]
Blade chroma	-0.52	[-0.71; -0.29]	-0.58	[-0.75; -0.37]	-0.33	[-0.57; -0.06]	-0.31	[-0.55; -0.03]	-0.30	[-0.55; -0.02]	-0.30	[-0.56; -0.01]
Spur chroma	-0.59	[-0.76; -0.38]	-0.71	[-0.84; -0.55]	-0.61	[-0.77; -0.41]	-0.42	[-0.64; -0.17]	-0.46	[-0.67; -0.21]	-0.49	[-0.70; -0.24]
Anthocyanins	-0.45	[-0.67; -0.19]	-0.57	[-0.75; -0.34]	-0.47	[-0.68; -0.22]	-0.56	[-0.77; -0.30]	-0.92	[-0.98; -0.83]	-0.56	[-0.79; -0.27]

	Spur hue		Spur length		Blade hue		Blade chroma		Spur chroma		Anthocyanins	
	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI
Orientation	0.40	[0.14; 0.63]	0.56	[0.34; 0.73]	0.15	[-0.14; 0.42]	-0.52	[-0.71; -0.29]	-0.59	[-0.76; -0.38]	-0.45	[-0.67; -0.19]
Blade brightness	0.36	[0.09; 0.59]	0.35	[0.08; 0.59]	0.22	[-0.06; 0.49]	-0.58	[-0.75; -0.37]	-0.71	[-0.84; -0.55]	-0.57	[-0.75; -0.34]
Spur brightness	0.59	[0.38; 0.76]	0.43	[0.17; 0.65]	0.33	[0.05; 0.57]	-0.33	[-0.57; -0.06]	-0.61	[-0.77; -0.41]	-0.47	[-0.68; -0.22]
Sepal length	0.44	[0.19; 0.65]	0.65	[0.47; 0.80]	0.47	[0.22; 0.67]	-0.31	[-0.55; -0.03]	-0.42	[-0.64; -0.17]	-0.56	[-0.77; -0.30]
Blade length	0.40	[0.14; 0.63]	0.65	[0.47; 0.80]	0.36	[0.10; 0.60]	-0.30	[-0.55; -0.02]	-0.46	[-0.67; -0.21]	-0.92	[-0.98; -0.83]
Pollinator	0.86	[0.74; 0.94]	0.76	[0.60; 0.88]	0.54	[0.28; 0.75]	-0.30	[-0.56; -0.01]	-0.49	[-0.70; -0.24]	-0.56	[-0.79; -0.27]
Spur hue	1.00	[1.00; 1.00]	0.68	[0.50; 0.81]	0.55	[0.33; 0.73]	-0.03	[-0.31; 0.26]	-0.33	[-0.57; -0.06]	-0.55	[-0.74; -0.32]
Spur length	0.68	[0.50; 0.81]	1.00	[1.00; 1.00]	0.50	[0.26; 0.70]	-0.10	[-0.38; 0.18]	-0.19	[-0.45; 0.10]	-0.61	[-0.78; -0.39]
Blade hue	0.55	[0.33; 0.73]	0.50	[0.26; 0.70]	1.00	[1.00; 1.00]	-0.28	[-0.53; -0.00]	-0.21	[-0.47; 0.07]	-0.34	[-0.59; -0.05]
Blade chroma	-0.03	[-0.31; 0.26]	-0.10	[-0.38; 0.18]	-0.28	[-0.53; -0.00]	1.00	[1.00; 1.00]	0.63	[0.44; 0.79]	0.17	[-0.13; 0.45]
Spur chroma	-0.33	[-0.57; -0.06]	-0.19	[-0.45; 0.10]	-0.21	[-0.47; 0.07]	0.63	[0.44; 0.79]	1.00	[1.00; 1.00]	0.43	[0.16; 0.65]
Anthocyanins	-0.55	[-0.74; -0.32]	-0.61	[-0.78; -0.39]	-0.34	[-0.59; -0.05]	0.17	[-0.13; 0.45]	0.43	[0.16; 0.65]	1.00	[1.00; 1.00]

Table 4.5: Posterior mean and 95% Bayesian credible interval (BCI) estimates for latent liability correlations between floral traits in *Aquilegia* not controlling for phylogenetic relatedness.

	Orientation		Blade brightness		Spur brightness		Sepal length		Blade length		Pollinator	
	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI
Orientation	1.00	[1, 1]	0.56	[0.35, 0.71]	0.51	[0.28, 0.67]	0.59	[0.38, 0.73]	0.71	[0.55, 0.82]	0.70	[0.51, 0.82]
Blade brightness	0.56	[0.35, 0.71]	1.00	[1, 1]	0.53	[0.3, 0.68]	0.52	[0.28, 0.67]	0.60	[0.4, 0.74]	0.44	[0.17, 0.62]
Spur brightness	0.51	[0.28, 0.67]	0.53	[0.3, 0.68]	1.00	[1, 1]	0.58	[0.37, 0.72]	0.60	[0.4, 0.74]	0.60	[0.39, 0.74]
Sepal length	0.59	[0.38, 0.73]	0.52	[0.28, 0.67]	0.58	[0.37, 0.72]	1.00	[1, 1]	0.78	[0.65, 0.86]	0.74	[0.55, 0.86]
Blade length	0.71	[0.56, 0.82]	0.60	[0.4, 0.74]	0.60	[0.4, 0.74]	0.78	[0.65, 0.86]	1.00	[1, 1]	0.66	[0.44, 0.8]
Pollinator	0.70	[0.51, 0.82]	0.44	[0.17, 0.62]	0.60	[0.39, 0.74]	0.74	[0.55, 0.86]	0.66	[0.44, 0.8]	1.00	[1, 1]
Spur hue	0.39	[0.13, 0.58]	0.26	[-0.022, 0.47]	0.51	[0.28, 0.67]	0.56	[0.34, 0.7]	0.47	[0.22, 0.64]	0.89	[0.81, 0.94]
Spur length	0.56	[0.34, 0.71]	0.30	[0.027, 0.5]	0.49	[0.25, 0.65]	0.69	[0.52, 0.8]	0.70	[0.53, 0.8]	0.85	[0.74, 0.91]
Blade hue	0.22	[-0.064, 0.43]	0.24	[-0.04, 0.45]	0.38	[0.12, 0.57]	0.60	[0.4, 0.74]	0.43	[0.18, 0.61]	0.68	[0.48, 0.81]
Blade chroma	-0.46	[-0.67, -0.27]	-0.64	[-0.79, -0.5]	-0.33	[-0.57, -0.12]	-0.29	[-0.54, -0.076]	-0.32	[-0.56, -0.11]	-0.12	[-0.41, 0.11]
Spur chroma	-0.61	[-0.77, -0.46]	-0.72	[-0.84, -0.6]	-0.63	[-0.78, -0.48]	-0.43	[-0.64, -0.24]	-0.52	[-0.71, -0.35]	-0.40	[-0.63, -0.2]
Anthocyanins	-0.62	[-0.78, -0.46]	-0.54	[-0.73, -0.37]	-0.61	[-0.78, -0.46]	-0.65	[-0.82, -0.48]	-0.93	[-0.98, -0.88]	-0.70	[-0.86, -0.54]

	Spur hue		Spur length		Blade hue		Blade chroma		Spur chroma		Anthocyanins	
	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI	cor	95% BCI
Orientation	0.39	[0.13, 0.58]	0.562	[0.34, 0.71]	0.22	[-0.064, 0.43]	-0.457	[-0.67, -0.27]	-0.61	[-0.77, -0.46]	-0.62	[-0.78, -0.46]
Blade brightness	0.26	[-0.022, 0.47]	0.304	[0.027, 0.5]	0.24	[-0.04, 0.45]	-0.637	[-0.79, -0.5]	-0.72	[-0.84, -0.6]	-0.54	[-0.73, -0.37]
Spur brightness	0.51	[0.28, 0.67]	0.490	[0.25, 0.65]	0.38	[0.12, 0.57]	-0.326	[-0.57, -0.12]	-0.63	[-0.78, -0.48]	-0.61	[-0.78, -0.46]
Sepal length	0.56	[0.34, 0.7]	0.693	[0.52, 0.8]	0.60	[0.4, 0.74]	-0.288	[-0.54, -0.076]	-0.43	[-0.64, -0.24]	-0.65	[-0.82, -0.48]
Blade length	0.47	[0.22, 0.64]	0.697	[0.53, 0.8]	0.43	[0.18, 0.61]	-0.321	[-0.56, -0.11]	-0.52	[-0.71, -0.35]	-0.93	[-0.98, -0.88]
Pollinator	0.89	[0.81, 0.94]	0.846	[0.74, 0.91]	0.68	[0.48, 0.81]	-0.121	[-0.41, 0.11]	-0.40	[-0.63, -0.2]	-0.70	[-0.86, -0.54]
Spur hue	1.00	[1, 1]	0.798	[0.68, 0.87]	0.70	[0.53, 0.8]	0.150	[-0.14, 0.37]	-0.20	[-0.46, 0.022]	-0.63	[-0.79, -0.47]
Spur length	0.80	[0.68, 0.87]	1.000	[1, 1]	0.63	[0.43, 0.75]	0.061	[-0.23, 0.28]	-0.15	[-0.42, 0.076]	-0.72	[-0.85, -0.6]
Blade hue	0.70	[0.53, 0.8]	0.626	[0.43, 0.75]	1.00	[1, 1]	-0.136	[-0.41, 0.091]	-0.17	[-0.44, 0.053]	-0.45	[-0.67, -0.25]
Blade chroma	0.15	[-0.14, 0.37]	0.061	[-0.23, 0.28]	-0.14	[-0.41, 0.091]	1.000	[1, 1]	0.68	[0.5, 0.79]	0.16	[-0.14, 0.39]
Spur chroma	-0.20	[-0.46, 0.022]	-0.147	[-0.42, 0.076]	-0.17	[-0.44, 0.053]	0.677	[0.5, 0.79]	1.00	[1, 1]	0.47	[0.21, 0.64]
Anthocyanins	-0.63	[-0.79, -0.47]	-0.724	[-0.85, -0.6]	-0.45	[-0.67, -0.25]	0.159	[-0.14, 0.39]	0.47	[0.21, 0.64]	1.00	[1, 1]

Table 4.6: Posterior mean and 95% Bayesian credible interval (BCI) estimates for significant correlations between sites in Influenza epitopes A and B.

	Sites	Correlation	95% BCI
1	[156(K)Q EpB, 158(E)K EpB]	0.7432	[0.3999, 0.9284]
2	[158(E)N EpB, 189(S)K EpB]	0.7365	[0.2806, 0.9499]
3	[144(V)N EpA, 158(E)K EpB]	0.7204	[0.3469, 0.9249]
4	[133(D)N EpA, 144(V)N EpA]	0.7180	[0.2908, 0.9391]
5	[159(Y)F EpB, 189(S)N EpB]	0.6913	[0.2655, 0.9174]
6	[144(V)I EpA, 156(K)Q EpB]	0.6883	[0.2880, 0.9161]
7	[133(D)N EpA, 158(E)K EpB]	0.6849	[0.2355, 0.9360]
8	[144(V)N EpA, 145(N)S EpA]	0.6826	[0.1411, 0.9304]
9	[131(A)T EpA, 159(Y)F EpB]	0.6792	[0.1932, 0.9353]
10	[145(N)S EpA, 188(D)Y EpB]	0.6726	[0.2276, 0.9137]
11	[144(V)N EpA, 188(D)Y EpB]	0.6640	[0.0368, 0.9332]
12	[144(V)N EpA, 156(K)Q EpB]	0.6602	[0.2393, 0.9076]
13	[159(Y)F EpB, 189(S)K EpB]	0.6586	[0.1364, 0.9346]
14	[156(K)H EpB, 159(Y)F EpB]	0.6585	[0.1583, 0.9204]
15	[158(E)K EpB, 188(D)Y EpB]	0.6534	[1.503e-05, 0.9262]
16	[144(V)N EpA, 144(V)D EpA]	0.6523	[0.0478, 0.9326]
17	[144(V)D EpA, 158(E)K EpB]	0.6516	[0.1447, 0.9094]
18	[131(A)T EpA, 156(K)H EpB]	0.6500	[0.1195, 0.9357]
19	[156(K)H EpB, 189(S)N EpB]	0.6381	[0.1477, 0.9119]
20	[144(V)N EpA, 156(K)H EpB]	0.6376	[0.0889, 0.9335]
21	[133(D)N EpA, 156(K)Q EpB]	0.6343	[0.1697, 0.9142]
22	[133(D)N EpA, 156(K)H EpB]	0.6328	[0.0869, 0.9432]
23	[145(N)S EpA, 156(K)H EpB]	0.6324	[0.0278, 0.9333]
24	[133(D)N EpA, 144(V)D EpA]	0.6320	[0.0886, 0.9170]
25	[144(V)I EpA, 158(E)K EpB]	0.6291	[0.1824, 0.8998]
26	[145(N)S EpA, 198(A)S EpB]	0.6195	[0.1026, 0.9115]
27	[156(K)H EpB, 198(A)S EpB]	0.6192	[0.0813, 0.9172]
28	[158(E)N EpB, 159(Y)F EpB]	0.6192	[0.0549, 0.9204]
29	[133(D)N EpA, 145(N)S EpA]	0.6190	[0.0058, 0.9267]
30	[189(S)K EpB, 193(S)Y EpB]	0.6138	[0.0330, 0.9231]
31	[131(A)T EpA, 189(S)N EpB]	0.6047	[0.0788, 0.9113]
32	[131(A)T EpA, 189(S)K EpB]	0.6030	[0.0881, 0.9200]
33	[144(V)D EpA, 156(K)Q EpB]	0.5898	[0.0680, 0.8939]

Continues on next page

	Sites	Correlation	95% BCI
34	[159(Y)F EpB, 198(A)S EpB]	0.5780	[0.0158, 0.9013]
35	[145(N)S EpA, 158(E)K EpB]	0.5774	[0.0008, 0.8927]
36	[144(V)K EpA, 189(S)K EpB]	0.5738	[0.0089, 0.9101]
37	[131(A)T EpA, 158(E)N EpB]	0.5652	[0.0013, 0.9056]
38	[133(D)N EpA, 197(R)Q EpB]	0.5644	[0.0600, 0.8911]
39	[189(S)K EpB, 193(S)F EpB]	0.5149	[0.0128, 0.8652]
40	[131(A)T EpA, 193(S)F EpB]	0.4865	[0.0112, 0.8355]
41	[159(Y)F EpB, 193(S)F EpB]	0.4849	[0.0015, 0.8354]

-

*The code for sites is as follows: of the number of the amino acid site in the aligned sequence; the one letter code for the reference amino acid for the site in parenthesis; the code for the amino acid corresponding to the latent trait; and the epitope to which the site belongs.

CHAPTER 5

Bayesian Nonparametric Clustering in Phylogenetics: Modeling Antigenic Evolution in Influenza [1]

Abstract. Influenza is responsible for up to 500,000 deaths every year, and antigenic variability represents much of its epidemiological burden. To visualize antigenic differences across many viral strains, antigenic cartography methods use multidimensional scaling on binding assay data to map influenza antigenicity onto a low-dimensional space. In these assays, the influenza strains naturally form clusters of similar antigenicity that correlate with sequence evolution. To understand the dynamics of these antigenic groups, we present a framework that jointly models genetic and antigenic evolution by combining multidimensional scaling of binding assay data, Bayesian phylogenetic machinery and nonparametric clustering methods. We propose a phylogenetic Chinese restaurant process that modifies the Chinese restaurant process to incorporate the evolutionary dependency structure between strains in the modeling of antigenic clusters. With this method, we are able to use the genetic information to better understand the evolution of antigenicity throughout epidemics, as shown in an application of this model to a H1N1 dataset.

[1] This project is joint work with Janet S. Sinsheimer, Trevor Bedford, Andrew Rambaut and Marc A. Suchard

5.1 Introduction

Every year, 10 to 20% of the world population is affected by influenza epidemics, with a death toll of approximately half a million people. Additionally, there is an enormous disease burden associated with influenza, with economic losses reaching an estimated 87 billion dollars for seasonal influenza in the US alone (Stohr, 2002; Molinari et al, 2007; Thompson et al, 2006). The World Health Organization (WHO) carefully monitors the influenza epidemics and defines strategies regarding disease control, including vaccine design. One of the main challenges for vaccination is the constant evolution of viral immunogenic proteins to evade immune response, known as antigenic drift. To be effective, vaccines must be designed specifically for the antigenic types circulating after they are administered, and these do not necessarily coincide with those circulating at the time of design. Consequently, an understanding of how viral antigenicity evolves over time is paramount for the efficacy of future influenza vaccination campaigns. In this paper we present methodology to study this evolutionary process through the perspective of clusters of viruses with similar antigenicity, and its relation to genetic evolution.

To characterize changes in antigenicity, researchers use information on how strongly the viral proteins interact with sera that represent immune responses to specific viral strains. This information is traditionally obtained from binding assays that measure the affinity of the two main immunogenic viral proteins, Hemagglutinin (HA) and Neuraminidase (NA), to the sera. To visualize patterns across many viral strains, antigenic cartography methods use these binding assay data to map influenza antigenicity onto a low-dimensional space (with generally 2 dimensions). In these maps, points that are close together represent antigenically similar strains (Smith et al, 2004; Cai et al, 2010).

Since their introduction in 2004, antigenic cartography methods have gained significant popularity, and are currently among the tools used by the WHO for vaccine strain selection

(Smith et al, 2008a). Antigenic cartography methods have been used to study differences in antigenic evolution between NA and HA proteins, the history of seasonal influenza in the last 35 years, and vaccination strategies for avian influenza (Sandbulte et al, 2011; Smith et al, 2004; Fouchier and Smith, 2010). Although the initial motivation was seasonal influenza, antigenic cartography methods have also found applications in horse and swine influenza, as well as in other diseases such as rabies and malaria (Smith et al, 2008a).

These studies have shown that the influenza strains tend to form aggregates based on similar antigenicity. Clusters are temporally oriented, so that in most years the circulating influenza strains belong to only one or two of the clusters. Additionally, they show that antigenic evolution correlates with genetic evolution. Nevertheless, this correlation is not perfect. While genetic evolution is approximately continuous over time, antigenic evolution seems to be more punctuated (Smith et al, 2008a).

For all its potential impact, methods that explicitly model the correlation between genetic evolution and antigenicity, as measured through antigenic cartography maps, have been conspicuously absent. Bedford et al (2014) lay groundwork for this type of study by presenting a probabilistic model for the multidimensional scaling of binding assay data. They use a Bayesian phylogenetic framework to connect molecular evolution to the antigenic map. We build upon their model to create our phylogenetic antigenic clustering method. Our goal is to use the genetic information to better understand the evolution of antigenicity, the emergence of new antigenic groups and the molecular changes which give rise to new clusters.

Our model focuses on the antigenic groups. Since there is potentially an infinite number of groups and the interactions of the evolutionary processes that govern antigenic evolution are not simply defined parametrically, we opt to use a nonparametric model. A canonical choice for modeling nonparametric clustering would be the Dirichlet mixture model, where the likelihood of the data is a mixture of normal distributions with the normal components representing the clusters (Antoniak, 1974). However, in the Dirichlet mixture model, the data points are assumed to be exchangeable and there is no flexibility for a dependency structure. This is not appropriate for antigenic data, since the viral strains are related to each other through evolution. Some strains are more recently diverged than others, and these relationships are captured by the phylogenetic tree. Thus, for this problem, a method that considers the dependency structure between samples in determining the probabilities of clustering would be more adequate.

Recent developments in Bayesian nonparametics have made it feasible to account for the dependency structure required to incorporate phylogenetic data in this clustering problem. Miller et al (2012) present a variation on the Indian buffet process that incorporates a nonexchangeable prior in the form of a tree. Dahl (2008) presents a modification of the Chinese restaurant process (CRP) that considers distances between data points for computing the probabilities of cluster arrangements. Blei and Frazier (2011) present an alternative representation of the CRP that also incorporates a distance matrix, but presents a more efficient sampling scheme. Both CRP models can be reduced to the original CRP by an appropriate choice of parameters. We build upon the distance dependent Chinese restaurant process - ddCRP (Blei and Frazier, 2011) as a clustering method for defining antigenic groupings. In our phylogenetic Chinese restaurant process (pCRP), distances between data points are informed by the phylogenetic tree.

In summary, our model follows Bedford et al (2014) in generating an antigenic map from binding assay data. The virus locations in the antigenic map are parameters of their probabilistic multidimensional scaling model. We define the pCRP as a clustering prior for these location parameters. This prior assigns each viral strain to one antigenic cluster, such that probabilities of clusters are a function of phylogenetic relatedness. By jointly modeling the cartographic map, antigenic clustering and molecular evolution, we effectively incorporate uncertainty about mapping, the unobserved phylogeny, and evolutionary parameters. Therefore, we are able to jointly estimate distributions for the antigenic map and clustering, while assessing how these relate to molecular evolution.

In the following section we present our model, and the sampling scheme that allows us to draw inference from it. Then, in Section 5.3, we present an application of this model to H1N1 influenza. A discussion of the results, modeling and future directions is presented in Section 5.4.

5.2 Methods

Consider a dataset of aligned molecular sequences $\mathbf{S}$ from N influenza strains and an $N \times M$ cross-reactivity matrix $\mathbf{H} = \{h_{ij}\}$ originating from hemagglutination inhibition (HI) assays for the N viral strains and M challenging sera. With the purpose of assessing antigenic similarities between viruses, HI assays measure the reactivity of one viral strain to serum containing antibodies raised against another. These assessments are made through serial dilutions, and the cross-reactivity measure h_{ij} represents $\log_2$ of the largest dilution titer at which serum j is effective against viral strain i.

We model the sequence data $\mathbf{S}$ using standard Bayesian phylogenetics models (Drummond et al, 2012) that include, among other evolutionary parameters $\boldsymbol{\theta}$, an unobserved phylogenetic tree F that represents the evolutionary relationship between the N viruses. This phylogenetic tree is a bifurcating, directed graph with N terminal nodes of degree 1 that correspond to the tips of the tree, $N - 2$ internal nodes of degree 3, a root node of degree 2. The edge weights between nodes are termed branch lengths and track elapsed evolutionary time. Conditional on F, we assume independence between the sequence data and $\mathbf{H}$.

In our model, $\mathbf{H}$ is used to generate an antigenic cartography map of the N viruses. We then model the locations of the virus on this map as a phylogenetic Chinese restaurant process (pCRP) that clusters viruses based on antigenic and phylogenetic distances, linking $\mathbf{H}$ and $\mathbf{S}$. In order to create a 2-dimensional antigenic map that preserves the relationships represented in $\mathbf{H}$, we employ the Bayesian multidimensional scaling (BMDS) method of Bedford et al (2014).

5.2.1 Bayesian multidimensional scaling

Let $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_N)^t$ be the $N \times 2$ matrix of virus locations in the antigenic cartography map, such that $\mathbf{X}_i = (x_{i1}, x_{i2})$ for $i = 1, \cdots N$. Likewise, let $\mathbf{Y} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_M)^t$ represent the $M \times 2$ analogous matrix of antigenic coordinates for the sera, with $\mathbf{Y}_j = (y_{j1}, y_{j2})$ for $j = 1, \cdots M$. BMDS is a probabilistic approach that estimates locations $\mathbf{X}$ and $\mathbf{Y}$ by matching immunologic distances from $\mathbf{H}$ to distances in the antigenic map (Bedford et al, 2014).

If h'_{ij} represents the theoretical titer at which reactivity ceases between virus i and serum j on a continuous scale, the immunologic distance can be defined as

$$\Delta_{ij} = s_j - h'_{ij} \tag{5.1}$$

in which $s_j = \max\{h_{1j}, \cdots, h_{Nj}\}$ represents the fixed serum effect. Additionally, let $\delta_{ij} = ||\mathbf{X}_i - \mathbf{Y}_j||_2$ represent the Euclidean distance between $\mathbf{X}_i$ and $\mathbf{Y}_j$. BMDS assumes that the HI titers are normally distributed with variance φ^2 and mean such that the expected value for the immunologic distance Δ_{ij} is the map distance δ_{ij}, thus

$$h'_{ij} \sim \mathcal{N}(s_j - \delta_{ij}, \varphi^2), \tag{5.2}$$

where $\mathcal{N}(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2.

However, the observed h_{ij}'s are integer values representing the last titer in the serial dilution at which reactivity was detected, consequently the matrix $\mathbf{H}$ is censored and intervaled. To circumvent this issue, we adopt the interpretation of an observed titer h_{ij} as a threshold, which implies that reactivity has ceased somewhere between the titers h_{ij} and $h_{ij} + 1$. Thus, the likelihood of an observed titer can be computed as

$$p(h_{ij}|\mathbf{X}_i, \mathbf{Y}_j, \varphi^2) = \phi\left(\frac{h_{ij} + \delta_{ij} - s_j + 1}{\varphi}\right) - \phi\left(\frac{h_{ij} + \delta_{ij} - s_j}{\varphi}\right), \tag{5.3}$$

where $\phi(\cdot)$ is the standard normal cumulative distribution function. When a serum and virus pair is not reactive for the smallest titer in the assay, the likelihood is defined analogously as an open interval, through the lower tail probability.

Assuming independence between the observations in $\mathbf{H}$, the joint likelihood can be computed as

$$p(\mathbf{H}|\mathbf{X}, \mathbf{Y}, \varphi^2) = \prod_{(i,j)\in\mathcal{I}} p(h_{ij}|\mathbf{X}_i, \mathbf{Y}_j, \varphi^2), \tag{5.4}$$

where $\mathcal{I}$ is the set of virus/serum pairs (i, j) for which observations are available. We note that, due to experimental constraints, most of the $N \times M$ comparisons cannot be made, leading to an incomplete matrix $\mathbf{H}$ and identifiability issues in the model.

We address identifiablity by adopting the drift prior of Bedford et al (2014) on the serum locations. The drift prior assumes that locations are normally distributed, and their expected values increase linearly with date of sampling along antigenic dimension 1. Thus

$$y_{j1} \sim \mathcal{N}(\beta t_j, \sigma_y^2) \quad \text{and} \quad y_{j2} \sim \mathcal{N}(0, \sigma_y^2), \tag{5.5}$$

for $j = 1, \cdots M$, where t_i is the difference between time of sampling of serum j and that of the oldest virus or serum in the sample, β is the drift parameter and σ_y^2 is the serum prior variance. This choice of prior is motivated by the observation that antigenic distances in influenza tend to increase over time (Smith et al, 2004).

To complete specification of the BMDS model, we select gamma prior distributions for the multidimensional scaling precision $1/\varphi^2$, and the precision $1/\sigma_y^2$ of the serum drift prior. We also adopt a diffuse gamma prior on the serum drift parameter β. Finally, the prior distribution of virus locations $\mathbf{X}$ on the antigenic map is given by the pCRP, thus connecting BMDS to antigenic clustering.

5.2.2 Phylogenetic Chinese restaurant process

The CRP (Blackwell and MacQueen, 1973) is a stochastic process that can be used to generate samples from the Dirichlet process. It is usually understood through the following analogy: customers arrive in turn at a restaurant with an infinite number of tables and choose a table to sit at according to a predefined distribution. After the arrival of N customers, their distribution in this Chinese restaurant represents a random partition of customers into table groups. The ddCRP modifies the CRP to consider affinities between customers for table assignment (Blei and Frazier, 2011). We use this feature to incorporate phylogenetic distances into our model.

In our phylogenetic Chinese restaurant process (pCRP), each customer represents one of the N viral strains. Each table represents one antigenic cluster, so that all customers assigned to the same table represent viruses in the same antigenic cluster. Even though theoretically the pCRP has an infinite number of potential clusters, in one realization only a finite number K is observed.

The dependency structure between customers is represented by the phylogenetic distance matrix $\mathbf{D} = \{d_{il}\}$, in which d_{il} is computed as the sum of branch lengths separating viruses i and l on the tree F. The effect that the phylogenetic distances $\mathbf{D}$ have on antigenic clustering is modulated through a decay function $f(\cdot)$. We adopt the form $f(d) = 1/d^{\lambda}$, which for positive λ takes large distances and transforms them into low probabilities of belonging to the same cluster. The parameter λ can regulate the effects of differences in the d_{il}'s that may span many orders of magnitude, especially for asynchronous data. When $\lambda = 0$ we have no phylogenetic effect on the clustering, and the pCRP becomes that standard CRP.

Under the Chinese restaurant analogy, the customer groupings in this pCRP act on two levels. On the first level, each customer chooses another customer with whom he would like to sit, and forms one single directional link. Customer i forms a link to customer l with probability proportional to $f(d_{il})$. Alternatively, the customer might choose not to form a link with any other customer and form an auto-link instead, with probability proportional to α. If c_i represents the customer to which customer i is linked, and $\mathbf{c} = \{c_i\}$, then

$$p(\mathbf{c}|\mathbf{D}, \lambda, \alpha) = \prod_{i=1}^{N} \frac{\mathbf{1}_{\{c_i=i\}}\alpha + \mathbf{1}_{\{c_i\neq i\}} f(d_{i,c_i})}{\alpha + \sum_{\ell\neq i} f(d_{i,\ell})}. \tag{5.6}$$

On the second level, the set of customer links is converted into table assignments through the transform $z(\mathbf{c})$ that takes all connected customers and assigns them to the same table.

In our antigenic cartography setting, this translates into a set of links $\mathbf{c}$ between viruses that is resolved through $z(\mathbf{c})$ into antigenic cluster associations. The virus locations $\mathbf{X}$ in the antigenic map are modelled as a mixture of normal distributions, where the mixture components are given by the antigenic clusters. Thus, to each antigenic cluster corresponds one mean vector $\boldsymbol{\mu}_k$ and one precision matrix $\boldsymbol{\Lambda}_k$, such that, if $\mathbf{z^k}$ represents the viruses in antigenic cluster k, then

$$\mathbf{X}_{\mathbf{z^k}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k \sim \mathcal{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) \tag{5.7}$$

for $k = 1, \cdots, K$. Here, $\mathcal{MVN}(\cdot)$ represents the multivariate normal distribution.

For convenience, we use the conjugate Normal-Wishart prior for the mean and precision parameters of the mixture components, thus

$$\begin{aligned} \boldsymbol{\Lambda}_k &\sim \mathcal{W}(\mathbf{T}_0, u_0) \\ \boldsymbol{\mu}_k &\sim \mathcal{MVN}\left(\mathbf{m}_0, (\kappa_0 \boldsymbol{\Lambda}_k)^{-1}\right), \end{aligned} \tag{5.8}$$

for $k = 1, \cdots, K$. Here, $\mathcal{W}(\mathbf{T}, u)$ is the Wishart distribution with scale matrix $\mathbf{T}$ and ν degrees of freedom. For ease of notation, we collect all the hyperparameters for the cluster normal distributions in $\mathbf{G}_0 = \{\mathbf{m}_0, \kappa_0, \mathbf{T}_0, u_0\}$. Exploiting the conjugate prior, we can analytically integrate out the the cluster mean and precision parameters, and compute the density of the viral locations in antigenic cluster k as

$$p(x_{\mathbf{z^k}}|\mathbf{G}_0, \mathbf{c}) = \frac{1}{\pi^{N_k}} \frac{\Gamma_2(u_k/2)}{\Gamma_2(u_0/2)} \frac{|\mathbf{T}_k|^{u_k/2}}{|\mathbf{T}_0|^{u_0/2}} \frac{\kappa_0}{\kappa_k}, \tag{5.9}$$

for $k = 1, \cdots, K$. Here, N_k is the number of viruses in cluster k, and

$$\begin{aligned} u_k &= u_0 + N_k, \\ \kappa_k &= \kappa_0 + N_k. \end{aligned} \tag{5.10}$$

Also, the posterior scale matrix can be obtained through

$$\mathbf{T}_k^{-1} = \mathbf{T}_0^{-1} + \sum_{\mathbf{z^k}} (\mathbf{X}_i - \bar{\mathbf{X}}_{\mathbf{z^k}})(\mathbf{X}_i - \bar{\mathbf{X}}_{\mathbf{z^k}})^t + \frac{\kappa_0 N_k}{\kappa_0 + N_k} (\bar{\mathbf{X}}_{\mathbf{z^k}} - \mathbf{m}_0)(\bar{\mathbf{X}}_{\mathbf{z^k}} - \mathbf{m}_0)^t, \qquad (5.11)$$

where $\bar{\mathbf{X}}_{\mathbf{z^k}}$ represents the mean location for viruses in cluster k, and Γ_2 is the 2-dimensional Gamma function. For the further information on expression (5.9) we refer the reader to the appendix in section 5.5.

Combining expressions (5.6) and (5.9), the joint density of the virus location matrix $\mathbf{X}$ and link vector $\mathbf{c}$ can be computed as

$$p(\mathbf{X}, \mathbf{c} | \mathbf{D}, \lambda, \alpha, \mathbf{G}_0) = p(\mathbf{c} | \mathbf{D}, \lambda, \alpha) \prod_{k=1}^{K} p(\mathbf{X}_{\mathbf{z^k}} | \mathbf{G}_0, \mathbf{c}). \qquad (5.12)$$

We assume *a priori* that the tuning parameter λ of the decay function is normally distributed with zero mean, and the concentration parameter α has an exponential distribution. Figure 5.1 presents a schematic representation of the pCRP.

5.2.3 Inference

The posterior distribution for our model can be expressed as

$$p(\mathbf{c}, \mathbf{X}, \mathbf{Y}, F, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\eta} | \mathbf{H}, \mathbf{S}) \propto p(\mathbf{H} | \mathbf{X}, \mathbf{Y}, \boldsymbol{\psi}) \times p(\mathbf{X}, \mathbf{c} | F, \boldsymbol{\eta}) \times p(\mathbf{Y}, \boldsymbol{\psi}) \times p(\boldsymbol{\eta}) \times p(\mathbf{S}, \boldsymbol{\theta}, F), \qquad (5.13)$$

where $\boldsymbol{\eta} = \{\alpha, \lambda\}$ collects the parameters of the pCRP, and $\boldsymbol{\psi} = \{\varphi^2, \beta, \sigma_y^2\}$ collects parameters of BMDS. To learn about this distribution we employ Markov chain Monte

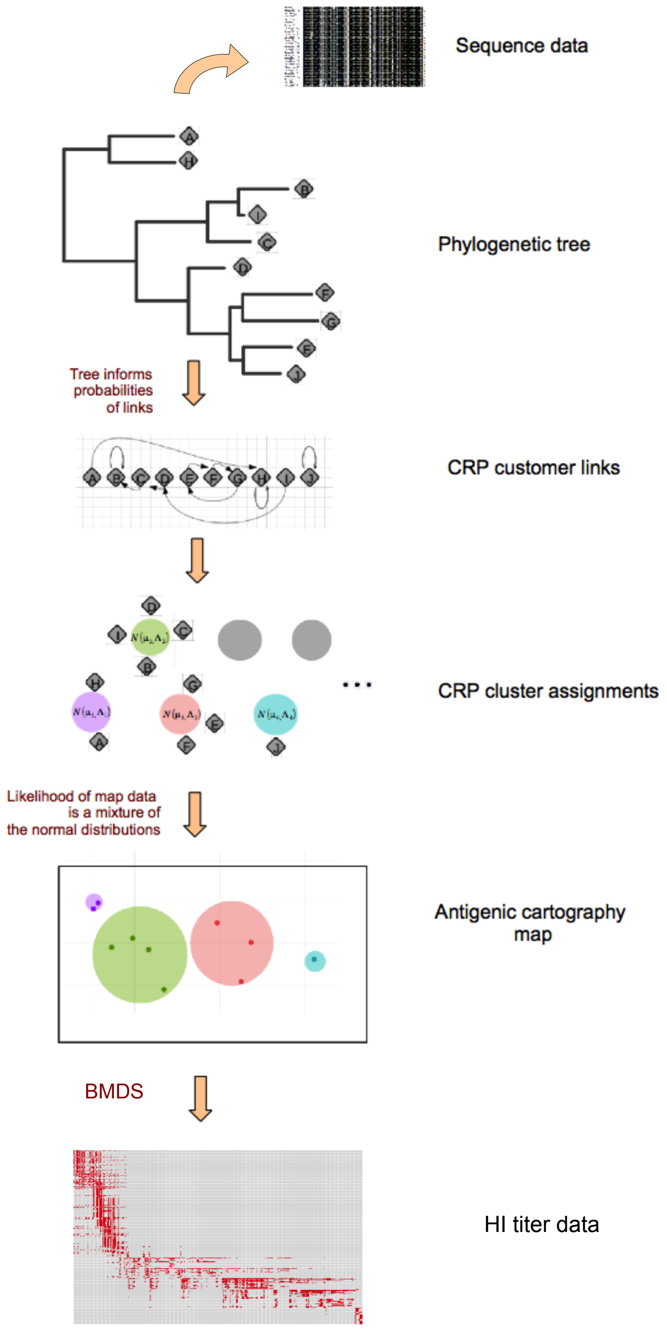

Sequence data
Phylogenetic tree
Tree informs probabilities of links
CRP customer links
CRP cluster assignments
Likelihood of map data is a mixture of the normal distributions
Antigenic cartography map
BMDS
HI titer data

Figure 5.1: Schematic representation of the pCRP: The sequence data are modelled through a phylogenetic tree F. The tree also dictates the probabilities of links between viruses. In turn, links translate into cluster assignments trough transform $z(c)$. Virus map locations $\mathbf{X}$ for each cluster are modelled as a Gaussian mixture component with mean $\boldsymbol{\mu}_k$ and precision $\boldsymbol{\Lambda}_k$. Finally, binding assay data $\mathbf{H}$ are modelled through BMDS to generate the antigenic map.

Carlo (MCMC). We exploit a random-scan Metropolis-with-Gibbs scheme. For the tree F and other phylogenetic parameters $\boldsymbol{\theta}$ modeling sequence evolution, we employ standard Bayesian phylogenetic algorithms (Drummond et al, 2012) based on Metropolis-Hastings parameter proposals. For the parameters $\mathbf{X}$, $\mathbf{Y}$ and $\boldsymbol{\psi}$ of the BMDS model, we follow Bedford et al (2014) in using Metropolis-Hastings transition kernels.

A central issue for this model is sampling for the links $\mathbf{c}$ between viruses, and consequently the cluster assignments. For this parameter we employ a Gibbs sampling scheme akin to the one of Blei and Frazier (2011) develop for the ddCRP. This Gibbs sampler explores the space of possible links between viruses by replacing at random one link at each step.

This individual link update to $\mathbf{c}$ can be understood in two steps. First, a virus i is chosen at random from a uniform distribution and the corresponding link is removed, resulting in a partition of $\mathbf{X}$ induced by the remaining links $\mathbf{c}_{-i}$. Subsequently, a new link c_i^* is created, rendering a new partition of $\mathbf{X}$ induced by $\mathbf{c}^* = (\mathbf{c}_{-i} \cup c_i^*)$.

Up to a constant, the conditional distribution of c_i^* can be computed as

$$p(c_i^*|\mathbf{c}_{-i}, \mathbf{X}, \mathbf{G}_0, \boldsymbol{\eta}, \mathbf{D}) = \frac{p(\mathbf{c}^*|\mathbf{X}, \mathbf{G}_0, \boldsymbol{\eta}, \mathbf{D})}{p(\mathbf{c}_{-i}|\mathbf{X}, \mathbf{G}_0, \boldsymbol{\eta}, \mathbf{D})} \propto \left(\mathbf{1}_{\{c_i^*=i\}}\alpha + \mathbf{1}_{\{c_i^*\neq i\}} f(d_{i,c_i^*})\right) \frac{p(\mathbf{X}|\mathbf{G}_0, \mathbf{c}^*)}{p(\mathbf{X}|\mathbf{G}_0, \mathbf{c}_{-i})}. \tag{5.14}$$

The creation of a new link i can have two possible effects on the virus partitions. First, virus i might form a new link to another virus that already belongs to the same cluster or form an auto-link: this operation does not change the data partition, and thus does not affect the $p(\mathbf{X}|\mathbf{G}_0, \mathbf{c})$. Alternatively, virus i might form a link to a virus from another cluster: this would join he two clusters, and therefore affect $p(\mathbf{X}|\mathbf{G}_0, \mathbf{c})$.

With these partition changes in mind, and noting that

$$p(\mathbf{X}|\mathbf{G}_0, \mathbf{c}) = \prod_{\ell=1}^{K} p(\mathbf{X}_{\mathbf{z}^\ell}|\mathbf{G}_0, \mathbf{c}), \tag{5.15}$$

we can obtain the full conditional distribution for the Gibbs sampler as proportional to

$$p(c_i^*|\mathbf{c}_{-i}, \mathbf{X}, \mathbf{G}_0, \boldsymbol{\eta}) \propto \begin{cases} \alpha & \text{if } c_i^* = i \\ f(d_{i\ell}) & \text{if } c_i^* = \ell \text{ does not join two clusters} \\ f(d_{i\ell}) \frac{p\left(\mathbf{X}_{{}_z k+\ell}|\mathbf{G}_0,\mathbf{c}_{-i}\right)}{p\left(\mathbf{X}_{{}_z\ell}|\mathbf{G}_0,\mathbf{c}_{-i}\right) \times p\left(\mathbf{X}_{\mathbf{zk}}|\mathbf{G}_0,\mathbf{c}_{-i}\right)} & \text{if } c_i^* = \ell \text{ joins clusters } k \text{ and } \ell. \end{cases} \tag{5.16}$$

Additionally, we use Metropolis-Hastings schemes to sample λ and α.

Even though the mean and precision of the normal components that represent the clusters have been analytically integrated out of the pCRP posterior distribution, it may be of interest to generate posterior samples for these parameters. By exploring the conjugate structure of the model, these parameters can be directly obtained from their posterior distribution given the cluster assignments $\mathbf{c}$ and hyperparameters $\mathbf{G}_0$ and antigenic locations $\mathbf{X}$. Thus we can sample directly from

$$\begin{aligned} \boldsymbol{\Lambda}_k &\sim \mathcal{W}(\mathbf{T}_k, u_k) \\ \boldsymbol{\mu}_k &\sim \mathcal{MVN}\left(\mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1}\right), \end{aligned} \tag{5.17}$$

where u_k, κ_k and $\mathbf{T}_k$ have been defined respectively in expressions (5.10) and (5.11). Additionally, we have $\mathbf{m}_k = \frac{\kappa_0 \mathbf{m}_0 + N_k \bar{\mathbf{X}}_{\mathbf{zk}}}{\kappa_0 + \mathbf{m}_0}$.

5.3 H1N1 influenza

We examine a dataset of H1N1 influenza with 115 viral strains and 77 sera obtained from (Bedford et al, 2014). We have 1882 cross reactivity measurements and nucleotide sequence data from the HA gene for each of the 115 viruses. The isolates have a wide geographic distribution and were collected between the years 1977 and 2009.

Figure 5.2 presents the maximum *a posteriori* estimate for the antigenic map, with viruses color coded according to antigenic cluster assignments. As expected, our analysis indicates a time directionality in the clusters, with older strains in the clusters of the left corner of the antigenic map, and younger strains at the right end of the plot. The strains present good correlation between the phylogenetic tree and antigenic groups. This can be seen in figure 5.3 which shows the viral phylogenetic tree with tip annotations color coded according to the antigenic clusters of figure 5.2. The fact that recent clusters have larger number of viruses is mainly a reflection of unequal temporal sampling.

For the time frame covered in the sample, we infer high posterior probability for the presence of 4 or 5 antigenic clusters (figure 5.4). For strains sampled post 1985, we clearly identified three antigenic clusters: the first, shown in yellow in figure 5.2, contains strains sampled between 1986 and 1996; the second cluster, in pink, is composed of viruses ranging from 1995 to 2009; and the third antigenic cluster, in green, has strains from 2006 to 2009. All strains in this period can be clearly associated to one of these antigenic clusters based on their posterior distribution of cluster assignments. The only noticeable exception is strain A/HongKong/1252/2000 that has similar posterior probabilities of being assigned to the pink and green clusters. These three antigenic clusters can be clearly seen in figure 5.5, which presents a heatmap of posterior probabilities of cluster co-assignments. Here, strains are not ordered temporally, but arranged to highlight cluster associations.

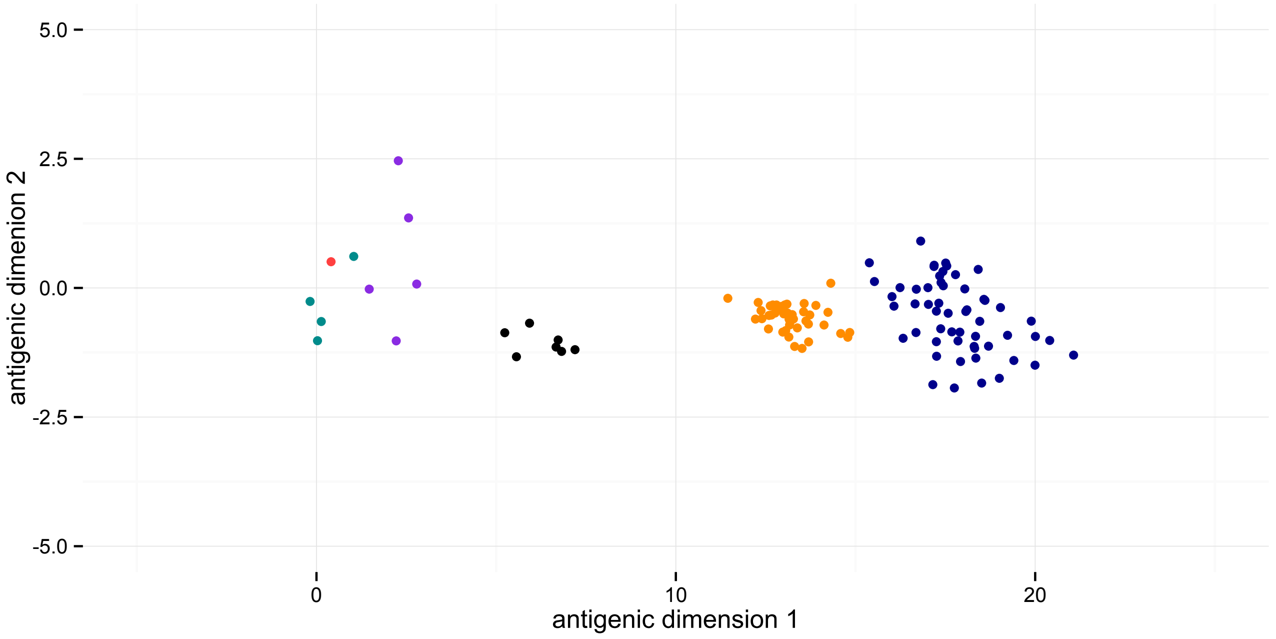


Figure 5.2: Maximum *a posteriori* (MAP) estimates of virus locations $\mathbf{X}$ in the antigenic map. Strains are color coded according to MAP cluster assignments.

The uncertainty on the number of antigenic clusters mainly reflects the mapping of viruses from 1977 to 1983. In this 7 year period, there is considerable antigenic variation that cannot be easily resolved into antigenic clusters (figure 5.2). Although these strains are clearly distinct from later clusters, there is considerable uncertainty on whether they should all be grouped together (figure 5.5). A better sampling of this time period would probably help resolve the issue.

Figure 5.6 presents the posterior distribution of cluster means. This distribution has three concentrated peaks representing the means of the most recent antigenic clusters. It also presents a more diffuse peek whose location corresponds to the viruses from 1977 to 1983. This last peak has a wider distribution, particularly along antigenic dimension 2, reflecting the clustering uncertainty in this period.

The drift parameter β of the prior distribution on serum locations can be seen as a measure of the overall linear change in antigenicity over time, since under this choice of prior most

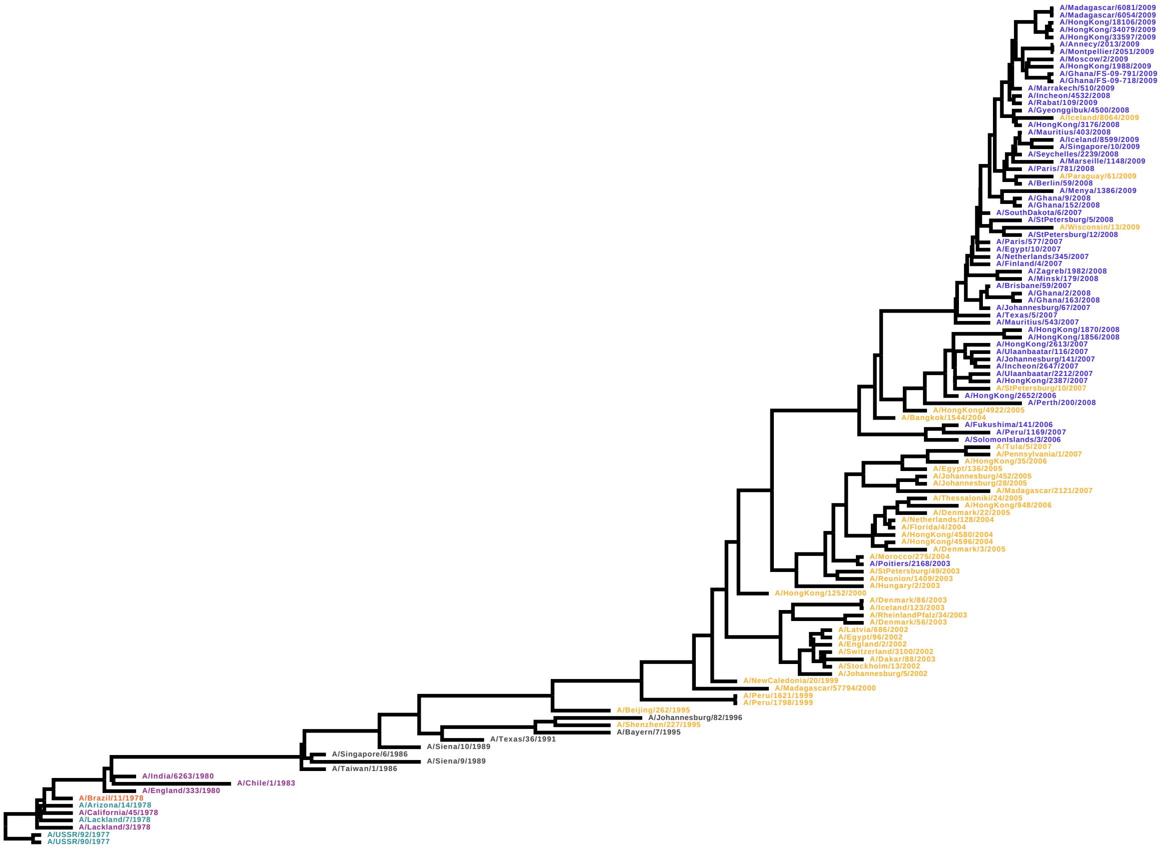


Figure 5.3: Maximum clade credibility tree for H1N1 influenza viruses. Tips are color coded according to MAP cluster assignments to match figure 5.2.

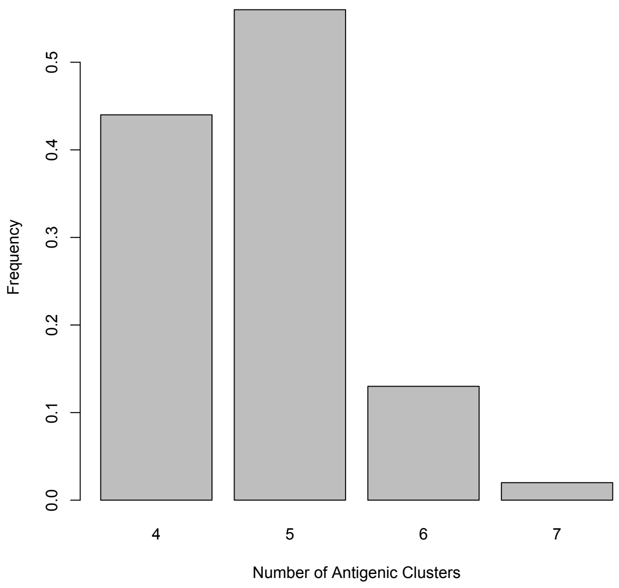


Figure 5.4: Posterior distribution of the number of antigenic clusters.

of the antigenic change happens across antigenic dimension 1. We estimate a posterior mean of 0.5112, with a 95% Bayesian credible interval of [0.4692, 0.5539] for β. These numbers are consistent with the findings of Bedford et al (2014) for their corresponding model. However, the rate of antigenic change is not constant trough time, as is evident by the uneven spacing of clusters along antigenic dimension 1 (figures 5.2 and 5.6).

We compare the posterior distribution of virus mappings for our model with that of the phylogenetic diffusion model of Bedford et al (2014). In their model, the prior distribution for the virus location $\mathbf{X}$ on the antigenic map is composed of a linear drift term and a continuous diffusion process along the phylogenetic tree. Supplementary videos (5.1) and (5.2) contain dynamic representations of the posterior distribution of the antigenic map locations $\mathbf{X}$ for both models. While both models exhibit similar distributions along antigenic dimension 1 reflecting their corresponding drift priors, the overall aggregation of strains is quite different. The drift-diffusion model generates a rather diffuse distribution

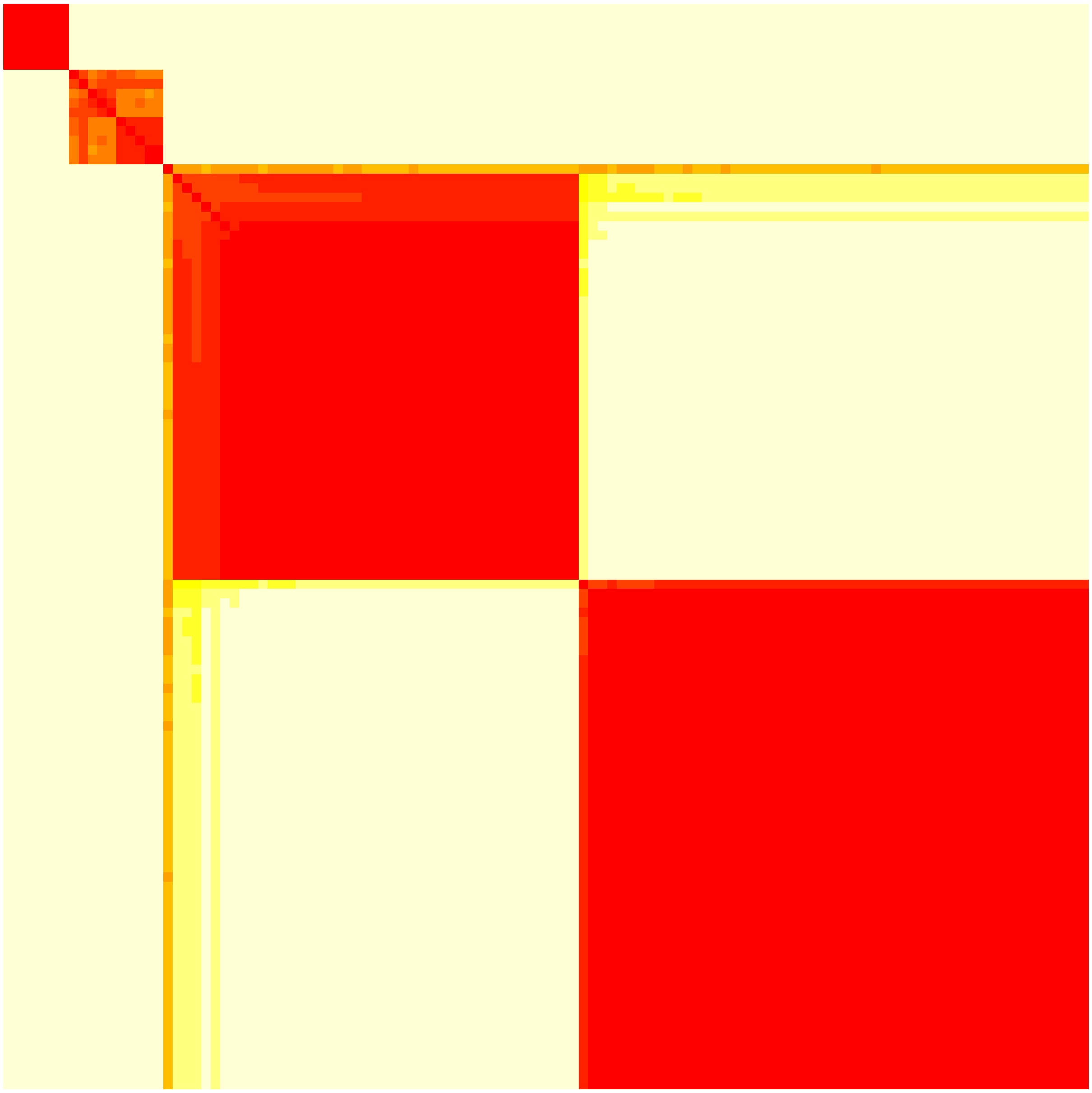


Figure 5.5: Heatmap of the probabilities of viruses being assigned to the same antigenic cluster. Dark orange represents higher probabilities. Arrangement of the order of viruses is not chronologic, but optimizes cluster visualization.

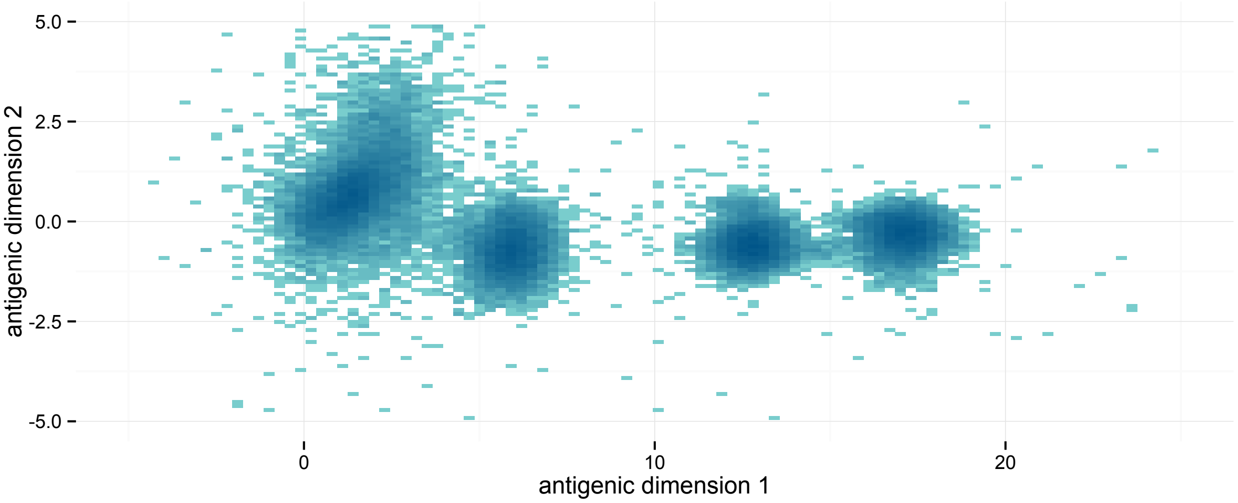


Figure 5.6: Posterior distribution of cluster means on antigenic map.

of viruses on the map, while the pCRP produces antigenic maps in which viral strains have a tendency toward aggregation. This distinction is, without doubt, a consequence of modeling choices. We argue that better defined groupings on the antigenic map, as well as explicit cluster association are important features of pCRP for the study of arising antigenic clusters.

Another important feature of the pCRP is the connection of antigenic clustering to molecular evolution. The tuning parameter λ of the decay function modulates the effect of the tree F on co-assignment probabilities. Our posterior mean estimate for λ was 0.3781, with a 95% Bayesian credible interval of [0.0184, 0.7559]. This implies that the probabilities of links between two viruses are less than inversely proportional to their phylogenetic distance. Consequently, the decay function has an attenuating effect on the large variability between phylogenetic distances induced by the trunk-like structure of the influenza phylogenetic tree. Additionally, at the 95% significance level, we infer that $\lambda \neq 0$, highlighting the impact of phylogenetic distances on antigenic clustering.

5.4 Discussion

In this paper we present a method for studying the interplay of antigenic and genetic evolution in influenza through the prism of antigenic clusters. We explicitly model the interaction between phylogenetics, antigenic clustering and the multidimensional scaling. This allows us to jointly estimate the antigenic map and cluster assignments. We also demonstrate, in an application to H1N1 influenza, that phylogenetic relatedness is an important factor in antigenic clustering. Additionally, we show that our pCRP can lead not only to antigenic maps with better defined clusters, but also high confidence in cluster assignments for most viral strains.

The purpose of the multidimensional scaling method is to represent the structure of the HI titer data in a low-dimensional space, for better interpretation. Both Smith et al (2004) and Bedford et al (2014) have compared different space dimensionalities and found the 2D plane to be optimal for their influenza data in terms of visualization and fit. For this reason, we develop our analysis with the 2-dimensional antigenic map. Nevertheless, pCRP and BMDS are not constrained to 2D maps, and can easily accommodate other map dimensions.

The drift prior for serum locations $\mathbf{Y}$ on the antigenic map, although motivated by observations of increased antigenic distance over time (Smith et al, 2004), is important to address identifiability of map locations. This choice of prior yields a biologically interpretable rate parameter representing the overall antigenic change over time, and has the effect of generating more linear antigenic maps. In Bedford et al (2014), inclusion of drift prior and a phylogenetic model not only improved identifiability of virus $\mathbf{X}$ and serum $\mathbf{Y}$ locations, it also improved model performance as measured through prediction errors.

Bayesian nonparametric modeling has found use in phylodynamics for estimating effective populations sizes of ancestral populations (Pybus et al, 2000). The nonparametric setting allows for a degree of flexibility for the shape of population size curve that would not be feasible with parametric forms. But nonparametric clustering methods have not yet been employed in phylodynamic studies of phenotypic traits such as antigenicity.

Traditional antigenic cartography applications use K-means clustering algorithms to define the antigenic groups Smith et al (2004). The pCRP, on the other hand, does not predefine the number of clusters, allowing a potentially infinite number of clusters. This is a strength of the non-parametric clustering, and it is particularly relevant if we are interested in identifying the rise of new antigenic clusters.

One desirable propriety of Dirichlet processes is marginal invariance: the marginal distribution when one observation is removed is the same as the distribution of the process without that observation. The ddCRP, and consequently our model, are not marginally invariant Blei and Frazier (2011). It is still not clear what repercussion, if any, follow from this property. However, it could be said that the existence of a strain with a particular antigenic profile could alter the antigenic landscape and the selective pressure on other strains. Thus, clustering should be different if such a strain exists or not, independent of it being sampled. Yet, the best we could expect is to have a representative sample of the antigenic landscape, since we could never observe all existing strains.

Antigenic cartography methods have been used to analyse other pathogens besides influenza, such as malaria and rabies. For these organisms, if the antigenic variability present the cluster-like structure observed in influenza, then our method could be instrumental to understanding the evolution of antigenicity and its relation to molecular evolution.

The fact that antigenic modeling is incorporated in the Bayesian phylogenetic context allows for joint estimation of the tree and the antigenic process. Demographic inference and geographic analysis are features already developed in this framework that can be jointly analysed with antigenicity, leading to a more complete representation of the evolution of influenza both genotypically and phenotypically (Minin et al, 2008; Lemey et al, 2009).

Through the posterior distribution of cluster assignments for an individual strain, we can assess the probability of recent viruses forming new antigenic clusters. Accurate detection of strains that are likely to seed new antigenic clusters could be particularly useful for influenza surveillance and vaccine design. For this purpose, the effectiveness of our method in the definition and detection of antigenic clusters should be further evaluated through applications to different strains of influenza and larger datasets. In the H1N1 example, the high confidence in cluster assignments obtained through the pCRP is particularly encouraging, even though in this example recent strains were well nested in current antigenic clusters, and we have no indication of new cluster formation.

5.5 Appendix: Normal-Wishart conjugate prior and marginal likelihood of $\mathbf{X}$

For ease of notation, in this section, we drop the explicit dependency of all densities on hyperparameters $\mathbf{m}_0$, κ_0, $\mathbf{T}_0$ and u_0. Additionally, for this section, let $\mathbf{X}$ represent all the virus locations in the antigenic map for strains belonging to cluster k. Then, these locations are all generated by the same normal component, and their density is given by

$$p(\mathbf{X}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = (2\pi)^{-rN_k/2} \left|\boldsymbol{\Lambda}_k\right|^{N_k/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{N_k} (\mathbf{X}_i - \boldsymbol{\mu}_k) \boldsymbol{\Lambda}_k (\mathbf{X}_i - \boldsymbol{\mu}_k)' \right), \tag{5.18}$$

where r is the dimension of the antigenic map. All normal components share the same priors for the mean parameter $\boldsymbol{\mu}_k$ and precision matrix $\boldsymbol{\Lambda}_k$. We adopt the conjugate Wishart Normal prior, where

$$\boldsymbol{\Lambda}_k \sim \mathcal{W}(u_0, \mathbf{T}_0)$$

and

$$\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k \sim \mathcal{MVN}(\mathbf{m}_0, (\kappa_0\boldsymbol{\Lambda}_k)^{-1}).$$

Thus, the joint prior density can be expressed as

$$\begin{aligned} p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = P(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)p(\boldsymbol{\Lambda}_k) &= \left(\frac{\kappa_0}{2\pi}\right)^{d/2} |\boldsymbol{\Lambda}_k|^{1/2} \exp\left(-\frac{\kappa_0}{2}(\boldsymbol{\mu}_k - \mathbf{m}_0)\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)^t\right) \\ &\times \quad \frac{1}{Z_0}|\boldsymbol{\Lambda}_k|^{\frac{u_0-r-1}{2}} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{T}_0^{-1}\boldsymbol{\Lambda}_k)\right) \end{aligned} \tag{5.19}$$

where

$$Z_0 = 2^{\frac{u_0 r}{2}}|\mathbf{T}_0|^{\frac{u_0}{2}}\Gamma_r(u_0/2)$$

and $\Gamma_r(\cdot)$ is the multivariate gamma function,

$$\Gamma_r(v) = \pi^{r(r-1)/4}\prod_{n=1}^{r}\Gamma(v + (1-n)/2)$$

In this conjugate model, the posterior distribution also assumes the form of a normal-Wishart distribution, and

$$\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k|\mathbf{X} \sim \mathcal{NW}(\mathbf{m}_k, \kappa_k, \mathbf{T}_k, u_k),$$

where $\mathcal{NW}(\cdot)$ represents the normal Wishart distribution with density

$$\begin{aligned} p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k|\mathbf{X}) &= \left(\frac{\kappa_k}{2\pi}\right)^{d/2} |\boldsymbol{\Lambda}_k|^{1/2} \exp\left(-\frac{\kappa_k}{2}(\boldsymbol{\mu}_k - \mathbf{m}_k)\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_k)^t\right) \\ &\times \frac{1}{Z_0}|\boldsymbol{\Lambda}_k|^{\frac{u_k - r - 1}{2}} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{T}_k^{-1}\boldsymbol{\Lambda}_k)\right) \end{aligned} \tag{5.20}$$

where

$$Z_n = 2^{\frac{u_k r}{2}} |\mathbf{T}_k|^{\frac{u_k}{2}} \Gamma_r(u_k/2)$$

and , $u_k = u_0 + N_k$, $\kappa_k = \kappa_0 + N_k$ and $\mathbf{T}_k^{-1} = \mathbf{T}_0^{-1} + \sum -N_k(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^t + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\mathbf{X}} - \mathbf{m}_0)(\bar{\mathbf{X}} - \mathbf{m}_0)^t$ and $\mathbf{m}_k = \frac{\kappa_k \mathbf{m}_0 + N_k \bar{\mathbf{X}}}{\kappa_0 + N_k}$ (DeGroot, 2005).

The posterior precision for each cluster can be samples from a $\mathcal{W}(\mathbf{T}_k, u_k)$, and $\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k, \mathbf{X} \sim \mathcal{MVN}(\mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1})$.

We now compute the marginal likelihood of the data, integrating out the mean and precision parameters. Notice, however, that

$$p(\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)}{p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k|\mathbf{X})}.$$

Combining expressions (5.18) and (5.20), and noting that $p(\mathbf{X})$ does not depend on $\boldsymbol{\mu}_k$ or $\boldsymbol{\Lambda}_k$, we can obtain the marginal likelihood as

$$p(\mathbf{X}) = \frac{(2\pi)^{-rN_k/2}\left(\frac{\kappa_0}{2\pi}\right)^{r/2}\frac{1}{Z_0}}{\left(\frac{\kappa_k}{2\pi}\right)^{r/2}\frac{1}{Z_n}} = \frac{\kappa_0^{r/2} Z_n}{(2\pi)^{rN_k/2}\kappa_k^{r/2} Z_0} = \frac{1}{\pi^{N_k r/2}} \frac{\Gamma_r(u_k/2)}{\Gamma_r(u_0/2)} \frac{|\mathbf{T}_k|^{u_k/2}}{|\mathbf{T}_0|^{u_0/2}} \left(\frac{\kappa_0}{\kappa_k}\right)^{r/2}.$$

When $r = 2$, this becomes the marginal likelihood of expression (5.9).

CHAPTER 6

Future Directions

In this dissertation I develop statistical methods for studying the intersection between phenotypic and genetic evolution under the framework of Bayesian phylogenetics. I now present some future directions arising from the latent liability model developed in chapter 4. As an additional future direction, I also outline a method for correcting ascertainment bias in phylogenetic reconstructions from single nucleotide polymorphism (SNP) data.

6.1 Symmetry of latent liability model for discrete traits with multiple unordered outcomes

In chapter 4 of this dissertation, I present the multivariate latent liability model, a non-Markovian model for the evolution of phenotypic traits. The main motivation for introducing this model is estimating the correlation structure between sets of traits while controlling for shared evolutionary history. The flexibility of the model allows for assessment of correlation between continuous traits, discrete binary traits, discrete traits with multiple ordered or unordered states, and combinations thereof.

For the discrete traits with multiple unordered outcomes, I adopt a multinomial probit function for the mapping of the latent liability $\mathbf{X}$ onto the observed trait $\mathbf{Y}$. In the notation of chapter 4, when the trait in column j of $\mathbf{Y}$ has K possible unordered states, then its outcome is a function of $K-1$ dimensions in the latent liability. Moreover, if $x_{ij'}, \ldots, x_{i,j'+K-2}$ are the latent liability entries corresponding to y_{ij}, then the largest of these elements determines the trait value,

$$y_{ij} = g(x_{ij'}, \ldots, x_{i,j'+K-2}) = \begin{cases} s_1 & \text{if} \quad 0 = \sup(0, x_{ij}, \ldots, x_{i,j+\mathrm{K}-2}) \\ s_{k+1} & \text{if} \;\; x_{ik} = \sup(0, x_{ij}, \ldots, x_{i,j+\mathrm{K}-2}). \end{cases} \tag{6.1}$$

Here, the first state s_1 is taken as reference to address identifiability. Notice that this choice of $g(\cdot)$ induces asymmetry in the model, since *a priori* the reference state s_1 has probability $2^{-(K-1)}$ and all other states have probability $\frac{1-2^{-(K-1)}}{K-1}$. In a small simulated example, the choice of reference state did not have major implications for inference. Figure 6.1 gives geometric intuition of this asymmetry for $K = 3$ unordered states.

In addition to its implications for the symmetry between states, this particular map function also has implications for the interpretability of inferred correlation. As discussed in chapter 4, other choices of map function such as the simplex might improve the symmetry but further complicate interpretation of correlation between traits.

A simpler model that simultaneously addresses both symmetry and interpretability maps the trait $\mathbf{Y}$ from K dimensions in the latent liability $\mathbf{X}$, through the function

$$y_{ij} = g^{*}(x_{ij'}, \ldots, x_{i,j'+K-1}) = s_k \qquad \text{if} \qquad x_{ik} = \sup(x_{ij}, \ldots, x_{i,j+\mathrm{K}-1}). \tag{6.2}$$

However, this model is not identifiable. Identifiability issues in general multinomial probit models have been well documented (Bunch, 1991; Weeks, 1997), and arise from the fact that only differences between the latent traits can be effectively estimated from the multinomial data. Lack of identifiability is frequently addressed by reducing the dimension of $\mathbf{X}$, as done in $g(\cdot)$.

In the latent liability model, an alternative solution would be to address identifiability

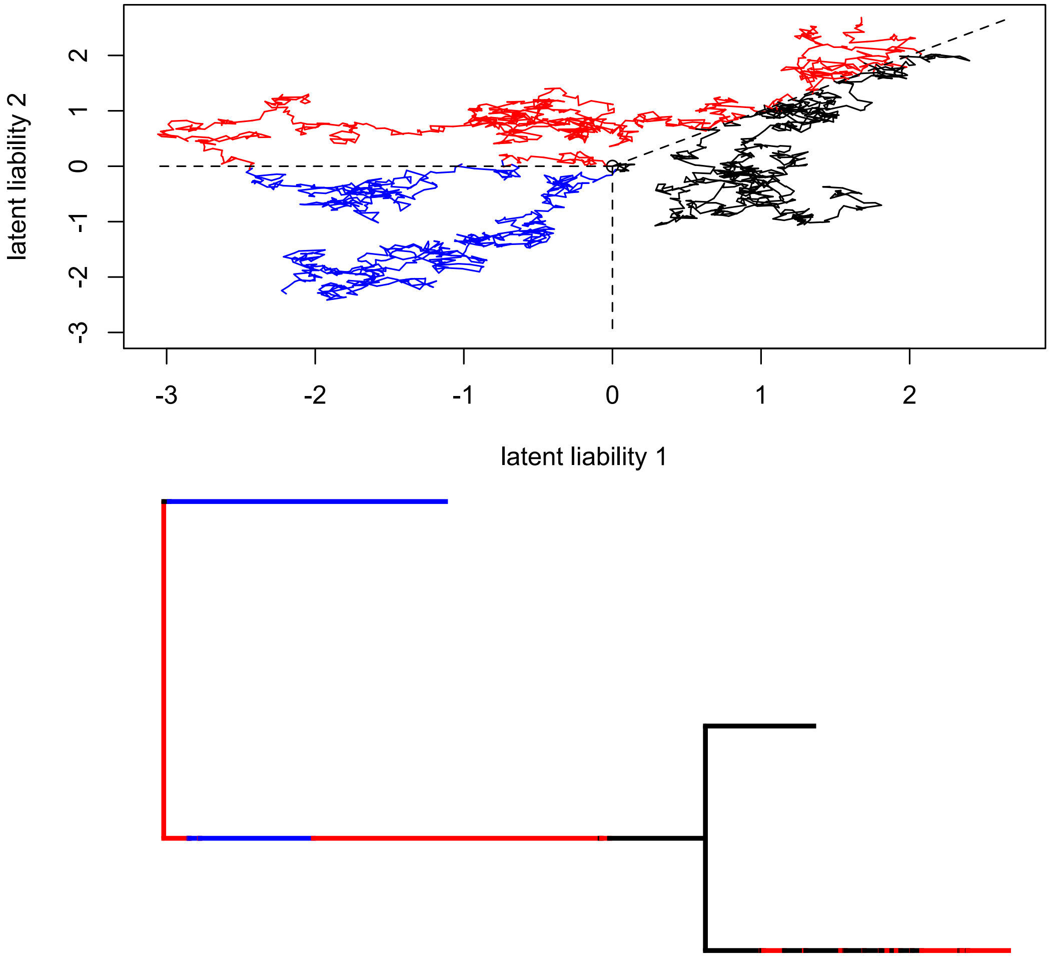


Figure 6.1: Realization of the evolution of latent liabilities $\mathbf{X}$ and observed trait $\mathbf{Y}$ for discrete data with 3 unordered states. Both tree and Brownian motion plots are color coded according to the trait $\mathbf{Y}$. ** This figure was created using code modified from R package *phylotools* (Revell, 2012).

through specification of a different diffusion model for the evolution of $\mathbf{X}$. One option is the Ornstein-Uhlenbeck (OU) model, a stationary Gaussian mean-reverting stochastic process. It can be defined from a Brownian motion process $W(t)$ through the stochastic differential equation

$$\mathrm{d}x_t = \theta(\mu - x_t)\mathrm{d}t + \sigma \mathrm{d}W(x_t), \tag{6.3}$$

for parameters $\sigma \geq 0$, $\theta \geq 0$ and μ. When $\theta = 0$ the OU process becomes Brownian motion.

The OU process is used in comparative biology to model stabilizing selection (Bartoszek et al, 2012). Importantly, while the variance of Brownian diffusion processes increases linearly over time, the variance of OU processes is bounded, and its stationary value can be obtained as $\frac{\sigma^2}{2\theta}$.

A critical issue in the latent liability model is efficient computation of the likelihood for the Brownian motion part of the model, and the efficient MCMC transition kernel derived from this result. Thus, in order to efficiently use OU for the evolution of the latent liability, similar results are required. Ho and Ané (2014) develop linear time algorithms for a more general class of linear models including OU, which could be used to compute the likelihood of the latent liability model, as well as to develop MCMC transition kernels.

An important feature of the latent liability model is the biological appeal of the non-Markovian property, as it is reasonable to expect for many biological problems that transition probabilities are affected by the time in which the process has been in the current state. This premiss motivates the work of Felsenstein (2012) and Revell (2012), however neither paper addresses the case of multiple unordered states. This highlights the importance of improving biological interpretiability through the use the map function $g^*(\cdot)$ instead of $g(\cdot)$.

6.2 Alternative Bayes factors computation in the latent liability model for large datasets

Bayes factors for comparing different latent liability models are computed in chapter 4 through a path sampling approach (Gelman and Meng, 1998). In this approach, first individual log marginal likelihoods for each model are estimated by numerically evaluating an path integral along a path between carefully chosen distributions. Then, the log Bayes factor is obtained as the difference between log marginal likelihoods.

Although there are many methods for calculating Bayes factors, their success in handling complex models is limited (Suchard et al, 2005; Dutta and Ghosh, 2013). Computing the high dimensional integrals required to estimate marginal likelihoods for these models is a challenging task. In the particular case of path sampling, the choice of the origin and destination distribution, as well as the path parametrization is particularly important for success of numerical integration. The path sampling approach of chapter 4 considers the geometric path between the potentially unnormalized distributions $q_0(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta})$ and $q_1(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta})$,

$$q_\beta(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta}) = q_0(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta})^{1-\beta} q_1(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta})^{\beta}, \tag{6.4}$$

for the path parameter $\beta \in [0;1]$. If z_0 is the normalizing constant for $q_0(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta})$, and z_1 the normalizing constant for $q_1(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta})$, then the identity of thermodynamic integration gives us

$$\log(z_1) - \log(z_0) = \int_0^1 E_{q_\beta}\left[\log(q_1(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta})) - \log(q_0(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta}))\right] \mathrm{d}\beta. \tag{6.5}$$

The path sampling algorithm employs MCMC to sample from $q_\beta(\mathbf{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta})$ along the path and estimate the integral above (Gelman and Meng, 1998).

To estimate marginal likelihoods for the latent liability model, the destination distribution is taken to be unnormalized posterior $p(\mathbf{Y}, \mathbf{S} | \mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X}, \boldsymbol{\theta})$, whose normalizing constant $z_1 = p(\mathbf{Y}, \mathbf{S})$ is the marginal likelihood. The source distribution $q_0(\mathbf{Y}, \mathbf{S}; \mathbf{X}, \boldsymbol{\theta})$ is an appropriately chosen normalized distribution defined on the same parameter space as the destination, so that $\log(z_0) = 0$. Separately estimating marginal likelihoods for each model is convenient for multiple model comparisons and allows for comparison between models defined on different parameter spaces. However, even with the carefully chosen source distribution presented in chapter 4, for large models this procedure may suffer from numerical issues.

A biologically relevant hypothesis for correlation analyses with the latent liability model is whether blocks of traits evolve independently. To address this hypothesis, one must test for a block structure in the covariance matrix. This requires a comparison between two nested models: the full model in which all covariances are allowed to be non-zero, and the block model in which the covariance between traits from different blocks is zero.

For comparisons between nested models or models defined on similar parameter spaces, an alternative path sampling approach suggests itself. Instead of individually approximating the marginal likelihoods for each model, we can directly use path sampling to estimate the Bayes factor by considering a path between the two models. In this scheme, source and destination distributions are the unnormalized posteriors under each of the models being compared. Thus, the left side of equation (6.5) becomes the Bayes factor.

Because this choice of path spans two models in similar probability spaces, it should be easier to adequately sample from $q_\beta(\mathbf{Y}, \mathbf{S}; \mathbf{X}, \boldsymbol{\theta})$ through the whole path, leading to better Bayes factor estimates. However, efficient MCMC tradition kernels that generate samples from $q_\beta(\mathbf{Y}, \mathbf{S}; \mathbf{X}, \boldsymbol{\theta})$ are required in order to employ this strategy. In particular, new versions of the Gibbs samplers should be developed.

6.3 Correction of ascertainment bias in SNP data

A single nucleotide polymorphism (SNP) is a populational variation in a single position of the genome. Each diploid individual may have 0, 1 or 2 copies of the minor allele for a particular SNP. Through microarray technology, multiple SNPs can be assessed simultaneously in a single chip, making the collection of SNP data fast and inexpensive.

By collecting data from multiple individuals in a population, SNP minor allele frequencies can be estimated and used for phylogenetic reconstruction. The use of gene frequency data to reconstruct evolutionary relationships between species dates back to early statistical phylogenetic methods (Cavalli-Sforza and Edwards, 1967). To model SNP frequency data through phylogenetics, one must first transform the frequency data onto the real line through a transform function such as the logit. Alternatively, one may employ a variance stabilizing transform such as the arcsine transform (Felsenstein, 1981b).

The $N \times M$ matrix of transformed frequencies $\mathbf{W} = (\mathbf{W}_1, \cdots, \mathbf{W}_M)$, for N taxa and M SNPs, is then modelled as M independent Brownian diffusion processes along a phylogenetic tree F. Assuming a conjugate normal distribution with mean μ_0 and variance τ_0 for the transformed frequency at the root of the tree, $\mathbf{W}_j$ becomes multivariate normal $\mathcal{MVN}(\mathbf{W}_j; \mu_0\mathbf{J}, \rho\boldsymbol{\Sigma})$. Here $\mathbf{J}_N$ is a vector of ones of length N, ρ is the diffusion variance, and

$$\boldsymbol{\Sigma} = \mathbf{V}(F) + \tau_0\mathbf{J}_{N\times N}, \tag{6.6}$$

with $\mathbf{J}_{N\times N}$ representing an $N \times N$ matrix of ones. Also, $\mathbf{V}(F)$, as defined in section 4.2.2, is a phylogenetic matrix tracking the shared evolutionary history between pairs of taxa on the tree F. Finally, the likelihood for $\mathbf{W}$ is obtained as the product of the likelihoods for the individual SNPs

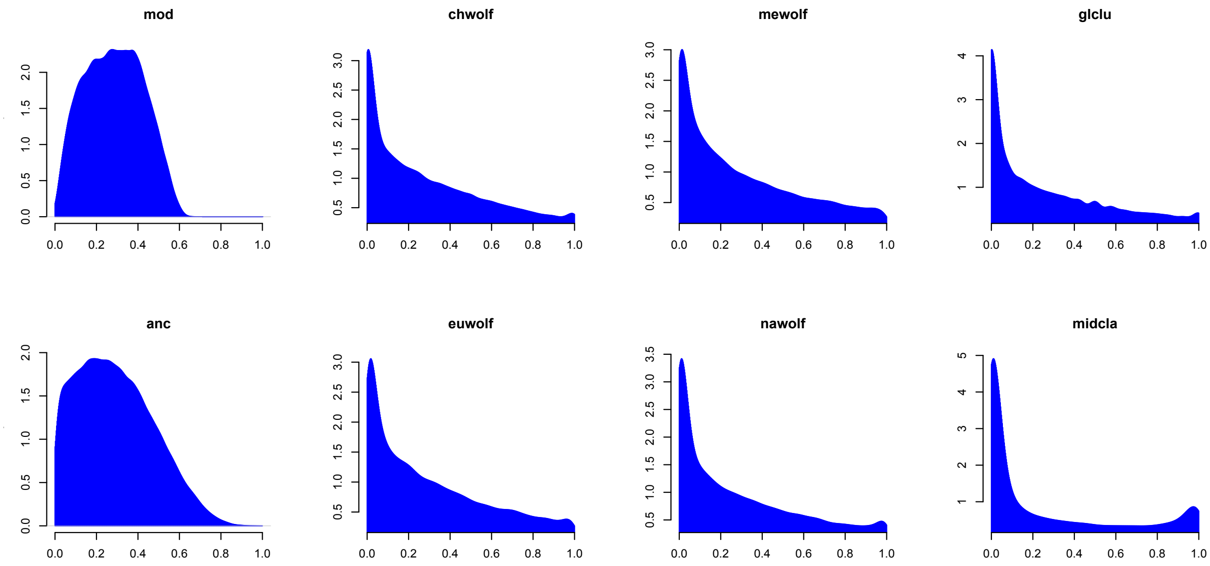


Figure 6.2: Effect of SNP ascertainment on minor allele frequencies. Kernel density estimates for the distribution of minor allele frequencies (defined in mod) for 2 dog populations (mod and anc) and 6 other canids.

$$p(\mathbf{W}|\mathbf{V}(F), \rho) = \prod_{j=1}^{M} p(\mathbf{W}_j|\mathbf{V}(F), \rho). \tag{6.7}$$

After prior specification for ρ and F, phylogenetic inference can be performed through MCMC.

However, a practical caveat complicates inference for this type of frequency data. In order to collect the SNP data for the multiple taxa, it is necessary to first select which sites in the genome will be queried. This step is carried out in a preliminary study, frequently in only one of the taxa, and may include the design of a microarray chip. SNP selection is done by identifying single nucleotide sites that are variable in the sample of the preliminary study. Using the same chip for multiple taxa, when SNP ascertainment was performed in only one of these taxa, leads to differences in the overall allele frequency distributions. These distributions may violate model assumptions and introduce bias in phylogenetic estimation.

Figure (6) exemplifies the ascertainment effect on frequency distributions through a canid dataset (Pollinger et al, 2010). In this example, a microarray chip was designed for SNP analysis in a particular dog population. In a subsequent study, the same chip was used to obtain frequency data in two populations of dogs and 6 populations of other canids (wolfs and coyotes). The figure shows considerable variability in SNP minor allele frequencies for the dog populations. For the other populations, most SNPs are invariable or have very low minor allele frequencies. A tree reconstruction on these frequency data is likely to produce an inaccurate tree F, and particularly distorted branch lengths leading to the dog populations.

One approach for correcting ascertainment bias is to condition on the taxa for which the data was ascertained. That is, if SNP selection was performed on the taxa at tip ν_1, then instead of performing inference based on (6.7), one could consider

$$p(\mathbf{W}_{(-1)j}|w_{1j}, \mathbf{V}(F), \rho) = \frac{p(\mathbf{W}_j|\mathbf{V}(F), \rho)}{p(w_{1j}|\mathbf{V}(F), \rho)}, \tag{6.8}$$

where $\mathbf{W}_{(-1)j} = (w_{2j}, \cdots w_{Nj})$ collects all transformed allele frequencies for SNP j, except w_{1j} for tip ν_1. Note that, $p(w_{1j}|\mathbf{V}(F), \rho)$ is the density of a normal distribution with mean μ_0 and variance $\tau_1\rho + \tau_0$, where τ_1 is the sum of branch lengths on F connecting tip ν_1 to the root.

BIBLIOGRAPHY

Allicock OM, Lemey P, Tatem AJ, Pybus OG, Bennett SN, Mueller BA, Suchard MA, Foster JE, Rambaut A, Carrington CV (2012) Phylogeography and population dynamics of dengue viruses in the Americas. Molecular Biology and Evolution 29(6):1533–1543

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics pp 1152–1174

Auguste AJ, Lemey P, Pybus OG, Suchard MA, Salas RA, Adesiyun AA, Barrett AD, Tesh RB, Weaver SC, Carrington CV (2010) Yellow fever virus maintenance in Trinidad and its dispersal throughout the Americas. Journal of Virology 84(19):9967–9977

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution 29(9):2157–2167

Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen TF (2012) A phylogenetic comparative method for studying multivariate adaptation. Journal of Theoretical Biology 314:204–215

Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell CA, Smith DJ, Rambaut A (2014) Integrating influenza antigenic dynamics with molecular evolution. eLife 3:e01,914

Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA (2007) A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. Proceedings of the National Academy of Sciences 104(19):7993–7998

Bielejec F, Rambaut A, Suchard MA, Lemey P (2011) SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics 27(20):2910–2912

Blackwell D, MacQueen JB (1973) Ferguson distributions via Pólya urn schemes. The Annals of Statistics pp 353–355

Blei DM, Frazier PI (2011) Distance dependent Chinese restaurant processes. The Journal of Machine Learning Research 12:2461–2488

Bloomquist EW, Lemey P, Suchard MA (2010) Three roads diverged? Routes to phylogeographic inference. Trends in Ecology & Evolution 25(11):626–632

Blum MG, Damerval C, Manel S, François O (2004) Brownian models and coalescent structures. Theoretical Population Biology 65(3):249–261

Boyd D, Peters GA, Cloeckaert A, Boumedine KS, Chaslus-Dancla E, Imberechts H, Mulvey MR (2001) Complete nucleotide sequence of a 43-kilobase genomic island associated with the multidrug resistance region of *Salmonella enterica* serovar Typhimurium DT104 and its identification in phage type DT120 and serovar agona. Journal of Bacteriology 183(19):5725–5732

Breslaw J (1994) Random sampling from a truncated multivariate normal distribution. Applied Mathematics Letters 7(1):1–6

Brown PJ, Vannucci M, Fearn T (1998) Multivariate Bayesian variable selection and prediction. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60(3):627–641

Bunch DS (1991) Estimability in the multinomial probit model. Transportation Research Part B: Methodological 25(1):1–12

Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. Science 286(5446):1921–1925

Cai Z, Zhang T, Wan XF (2010) A computational framework for influenza antigenic cartography. PLoS Computational Biology 6(10):e1000,949

Carvalho CM, Scott JG (2009) Objective Bayesian model selection in Gaussian graphical models. Biometrika 96(3):497–512

Cavalli-Sforza L, Edwards A (1967) Phylogenetic analysis: Models and estimation procedures. Evolution pp 550–570

Cox NJ, Bender CA (1995) The molecular epidemiology of influenza viruses. In: Seminars in Virology, Elsevier, vol 6, pp 359–370

Cybis GB, Sinsheimer JS, Lemey P, Suchard MA (2013) Graph hierarchies for phylogeography. Philosophical Transactions of the Royal Society B: Biological Sciences 368(1614)

Dahl D (2008) Distance-based probability distribution for set partitions with applications to Bayesian nonparametrics. JSM Proceedings Section on Bayesian Statistical Science, American Statistical Association

Damien P, Walker SG (2001) Sampling truncated normal, beta, and gamma densities. Journal of Computational and Graphical Statistics 10(2)

DeGroot MH (2005) Optimal statistical decisions, vol 82. Wiley-Interscience

Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M (2004) Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis 90(1):196–212

Domingo E, Holland J (1997) RNA virus mutations and fitness for survival. Annual Reviews in Microbiology 51(1):151–178

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161(3):1307–1320

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003) Measurably evolving populations. Trends in Ecology & Evolution 18(9):481–488

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biology 4(5):e88

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29(8):1969–1973

Dutta R, Ghosh JK (2013) Bayes model selection with path sampling: factor models and other examples. Statistical Science 28(1):95–115

Edo-Matas D, Lemey P, Tom JA, Serna-Bolea C, van den Blink AE, van't Wout AB, Schuitemaker H, Suchard MA (2011) Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intrahost evolutionary processes: efficient hypothesis testing through hierarchical phylogenetic models. Molecular Biology and Evolution 28(5):1605–1616

Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L (2007) Statistical evaluation of alternative models of human evolution. Proceedings of the National Academy of Sciences 104(45):17,614–17,619

Falconer DS (1965) The inheritance of liability to certain diseases, estimated from the incidence among relatives. Annals of Human Genetics 29(1):51–76

Faria NR, Suchard MA, Rambaut A, Lemey P (2011) Toward a quantitative understanding of viral phylogeography. Current Opinion in Virology 1(5):423–429

Faria NR, Suchard MA, Rambaut A, Streicker DG, Lemey P (2013) Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. Philosophical Transactions of the Royal Society B: Biological Sciences 368(1614)

Felsenstein J (1981a) Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17(6):368–376

Felsenstein J (1981b) Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution pp 1229–1242

Felsenstein J (1985) Phylogenies and the comparative method. American Naturalist 125:1–15

Felsenstein J (2004) Inferring phylogenies, vol 2. Sinauer Associates Sunderland

Felsenstein J (2005) Using the quantitative genetic threshold model for inferences between and within species. Philosophical Transactions of the Royal Society B: Biological Sciences 360(1459):1427–1434

Felsenstein J (2012) A comparative method for both discrete and continuous characters using the threshold model. The American Naturalist 179(2):145–156

Fitch WM, Leiter J, Li X, Palese P (1991) Positive Darwinian evolution in human influenza A viruses. Proceedings of the National Academy of Sciences 88(10):4270–4274

Fouchier RA, Smith DJ (2010) Use of antigenic cartography in vaccine seed strain selection. Avian Diseases 54(s1):220–223

Freckleton RP (2012) Fast likelihood calculations for comparative analyses. Methods in Ecology and Evolution 3(5):940–947

Gelfand AE, Smith AF, Lee TM (1992) Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. Journal of the American Statistical Association 87(418):523–532

Gelman A, Meng XL (1998) Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical Science pp 163–185

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on (6):721–741

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA (2013) Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Molecular Biology and Evolution 30(3):713–724

Grafen A (1989) The phylogenetic regression. Philosophical Transactions of the Royal Society Series B, Biological Sciences 326(1233):119–157

Gray RM (2011) Entropy and information theory. Springer

Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303(5656):327–332

Guzman MG, Halstead SB, Artsob H, Buchy P, Farrar J, Gubler DJ, Hunsperger E, Kroeger A, Margolis HS, Martínez E, et al (2010) Dengue: a continuing global threat. Nature Reviews Microbiology 8:S7–S16

Hadfield J, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. Journal of Evolutionary Biology 23(3):494–508

Haeckel EHPA (1866) Generelle Morphologie der Organismen: allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie, vol 2. G. Reimer

Halstead SB (2008) Dengue virus-mosquito interactions. Annual Review in Entomology 53:273–291

Hasegawa M, Kishino H, Yano Ta (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22(2):160–174

Hastings WK (1970) Monte carlo sampling methods using markov chains and their applications. Biometrika 57(1):97–109

Ho LST, Ané C (2014) A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Systematic Biology 3(3):397–402

Huelsenbeck JP, Rannala B (2003) Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 57(6):1237–1247

Ives AR, Garland T (2010) Phylogenetic logistic regression for binary dependent variables. Systematic Biology 59(1):9–26

Jeffreys H (1935) Some tests of significance, treated by the theory of probability. In: Proceedings of the Cambridge Philosophical Society, Cambridge Univ Press, vol 31, pp 203–222

Jukes TH, Cantor CR (1969) Evolution of protein molecules. Manmmalian Protein Metabolism pp 21–132

Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GC, Vervaet G, Skepner E, Lewis NS, Spronken MI, Russell CA, et al (2013) Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science 342(6161):976–979

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149(1):429–434

Landis MJ, Schraiber JG, Liang M (2013) Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. Systematic Biology 62(2):193–204

Lartillot N, Poujol R (2011) A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Molecular Biology and Evolution 28(1):729–744

Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Computational Biology 5(9):e1000,520

Lemey P, Rambaut A, Welch JJ, Suchard MA (2010) Phylogeography takes a relaxed random walk in continuous space and time. Molecular Biology and Evolution 27(8):1877–1885

Lewis PO (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50(6):913–925

Liang LJ, Weiss RE (2007) A hierarchical semiparametric regression model for combining HIV-1 phylogenetic analyses using iterative reweighting algorithms. Biometrics 63(3):733–741

Liu JS, Liang F, Wong WH (2000) The multiple-try method and local optimization in metropolis sampling. Journal of the American Statistical Association 95(449):121–134

Mather A, Reid S, Maskell D, Parkhill J, Fookes M, Harris S, Brown D, Coia J, Mulvey M, Gilmour M, et al (2013) Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. Science 341(6153):1514–1517

Mather AE, Matthews L, Mellor DJ, Reeve R, Denwood MJ, Boerlin P, Reid-Smith RJ, Brown DJ, Coia JE, Browning LM, et al (2012) An ecological approach to assessing the epidemiology of antimicrobial resistance in animal and human populations. Proceedings of the Royal Society B: Biological Sciences 279(1733):1630–1639

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21:1087

Miller KT, Griffiths T, Jordan MI (2012) The phylogenetic indian buffet process: A non-exchangeable nonparametric prior for latent features. arXiv preprint arXiv:12063279

Minin VN, Bloomquist EW, Suchard MA (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution 25(7):1459–1471

Molinari NAM, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, Bridges CB (2007) The annual impact of seasonal influenza in the US: measuring disease burden and costs. Vaccine 25(27):5086–5096

Nelson MI, Lemey P, Tan Y, Vincent A, Lam TTY, Detmer S, Viboud C, Suchard MA, Rambaut A, Holmes EC, et al (2011) Spatial dynamics of human-origin H1 influenza A virus in North American swine. PLoS Pathogens 7(6):e1002,077

van der Niet T, Johnson SD (2012) Phylogenetic evidence for pollinator-driven diversification of angiosperms. Trends in Ecology & Evolution 27(6):353–361

Novembre J, Slatkin M (2009) Likelihood-based inference in isolation-by-distance models using the spatial distributions of low frequency alleles. Evolution 63(11):2914–2925

Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society Series B: Biological Sciences 255(1342):37–45

Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing AM, van de Vijver DA, Albert J, Angarano G, Åsjö B, Balotta C, et al (2009) Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. Retrovirology 6(1):49

Plotkin JB, Dushoff J (2003) Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proceedings of the National Academy of Sciences 100(12):7152–7157

Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A, Reynolds A, et al (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464(7290):898–902

Potter SJ, Lemey P, Dyer WB, Sullivan JS, Chew CB, Vandamme AM, Dwyer DE, Saksena NK (2006) Genetic analyses reveal structured HIV-1 populations in serially sampled T lymphocytes of patients receiving HAART. Virology 348(1):35–46

Pybus OG, Rambaut A, Harvey PH (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155(3):1429–1437

Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, Gray RR, Arinaminpathy N, Stramer SL, Busch MP, Delwart EL (2012) Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proceedings of the National Academy of Sciences 109(37):15,066–15,071

Rambaut A (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16(4):395–399

Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution 43(3):304–311

Redelings BD, Suchard MA (2005) Joint Bayesian estimation of alignment and phylogeny. Systematic Biology 54(3):401–418

Redelings BD, Suchard MA (2007) Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. BMC Evolutionary Biology 7(1):40

Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3(2):217–223

Revell LJ (2013) Ancestral character estimation under the threshold model from quantitative genetics. Evolution

Robert CP (1995) Simulation of truncated normal variables. Statistics and Computing 5(2):121–125

San Martín JL, Brathwaite O, Zambrano B, Solórzano JO, Bouckenooghe A, Dayan GH, Guzmán MG (2010) The epidemiology of dengue in the Americas over the last three decades: a worrisome reality. The American Journal of Tropical Medicine and Hygiene 82(1):128

Sandbulte MR, Westgeest KB, Gao J, Xu X, Klimov AI, Russell CA, Burke DF, Smith DJ, Fouchier RA, Eichelberger MC (2011) Discordant antigenic drift of neuraminidase and hemagglutinin in H1N1 and H3N2 influenza viruses. Proceedings of the National Academy of Sciences 108(51):20,748–20,753

Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution 19(1):101–109

Sanmartín I, Van Der Mark P, Ronquist F (2008) Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary islands. Journal of Biogeography 35(3):428–449

Sinsheimer JS, Lake JA, Little RJ (1996) Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. Biometrics pp 193–210

Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305(5682):371–376

Smith DJ, de Jong JC, Lapedes AS, Jones TC, Russell CA, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA (2008a) Antigenic cartography of human and swine influenza A (H3N2) viruses. Novel and Re-emerging Respiratory Viral Diseases 699:32

Smith SD, Ané C, Baum DA (2008b) The role of pollinator shifts in the floral diversification of iochroma (Solanaceae). Evolution 62(4):793–806

Stohr K (2002) Influenza - WHO cares. The Lancet infectious diseases 2(9):517

Suchard MA, Weiss RE, Sinsheimer JS (2001) Bayesian selection of continuous-time Markov chain evolutionary models. Molecular Biology and Evolution 18(6):1001–1013

Suchard MA, Kitchen CM, Sinsheimer JS, Weiss RE (2003) Hierarchical phylogenetic models for analyzing multipartite sequence data. Systematic Biology 52(5):649–664

Suchard MA, Weiss RE, Sinsheimer JS (2005) Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. Biometrics 61(3):665–673

Telesca D, Müller P, Parmigiani G, Freedman RS (2012) Modeling dependent gene expression. The Annals of Applied Statistics 6(2):542–560

Thompson WW, Comanor L, Shay DK (2006) Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease. Journal of Infectious Diseases 194

Tom JA, Sinsheimer JS, Suchard MA (2010) Reuse, recycle, reweigh: Combating influenza through efficient sequential Bayesian computation for massive data. The Annals of Applied Statistics 4(4):1722–1748

UNAIDS (2010) Joint United Nations Programme on HIV/AIDS. AIDS Scorecards: Overview: UNAIDS Report on the Global AIDS Epidemic 2010. UNAIDS

Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, Van Wijngaerden E, Vandamme AM, Van Laethem K, Lemey P (in press) The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. PLoS Computational Biology

Wallace RG, HoDac H, Lathrop RH, Fitch WM (2007) A statistical phylogeography of influenza A H5N1. Proceedings of the National Academy of Sciences 104(11):4473–4478

Weeks M (1997) The multinomial probit model revisited: A discussion of parameter estimability, identification and specification testing. Journal of Economic Surveys 11(3):297–320

Whittall JB, Hodges SA (2007) Pollinator shifts drive increasingly long nectar spurs in columbine flowers. Nature 447(7145):706–709

Whittall JB, Voelckel C, Kliebenstein DJ, Hodges SA (2006) Convergence, constraint and the role of gene expression during adaptive radiation: floral anthocyanins in *Aquilegia*. Molecular Ecology 15(14):4645–4657

Wright S (1934) An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics 19(6):506

Yang Z (2006) Computational molecular evolution, vol 284. Oxford University Press Oxford