Improving Information Propagation in Phylogenetic Workflows

a dissertation presented
by
August Guang
to
The Department of Applied Mathematics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Applied Mathematics

Brown University
Providence, Rhode Island
May 2018

Improving Information Propagation in Phylogenetic Workflows

Abstract

Despite the enormous amount of biological variation and technical uncertainty in sequence data, most phylogenetic workflows propagate a single point estimate throughout the numerous analysis components, and only in a forward direction. This approach relies on three implicit assumptions: (i) the order of the analysis steps is biologically justified, (ii) a Markovian dependency structure exists between analysis components, and (iii) there is low relative entropy between results at each analysis step. There is evidence that these assumptions, in particular low relative entropy, are frequently violated in empirical studies with potential detrimental effects in phylogenetic analyses.

In this thesis, I lay out a probabilistic framework that provides a unified perspective to provide context for evaluating priorities for future developments of methods and tools. I then develop a generative model of the natural and technical processes that produce observed genomic reads within the framework that can be used to assess and validate approaches that relax the implicit assumptions. Finally, I explore two ways to accommodate and propagate more information in a phylogenetic workflow. The first way, an HMM profile-sampling approach to genome assembly, relaxes the assumption of low relative entropy in results from the genome assembly analysis component. This approach finds relevant applications to HIV transmission networks. The second way, an iterative approach to identifying and resolving transcriptome assembly errors, capitalizes on the assumption of Markovian dependence.

This dissertation by August Guang is accepted in its present form by
the Department of Applied Mathematics as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date _____ _____

Charles E. Lawrence

Recommended to the Graduate Council

Date _____ _____

Casey W. Dunn
Yale University

Date _____ _____

Paul Lewis
University of Connecticut

Approved by the Graduate Council

Date _____ _____

Yan Guo
Dean of the Graduate School

August Guang

Department of Applied Mathematics
Brown University
Providence, RI

Phone: +1 (617) 276 6176
Email: august_guang@brown.edu
ORCID iD: 0000-0003-4974-1873

Education

† *Indicates expected*

2012–2018 † Ph.D., Applied Mathematics, Brown University

Thesis Title: Improving Information Propagation in Phylogenetic Workflows

Supervisors: Casey W. Dunn and Charles E. Lawrence

2012–2014 M.Sc., Applied Mathematics, Brown University

2008–2012 B.Sc., Mathematics, Harvey Mudd College

Internships

Summer 2016 Bioinformatics and Computational Biology Graduate Intern, Genentech, South San Francisco, CA

Teaching Assistant

2017	APMA1710	Information Theory
2016	APMA1660	Statistical Inference II
2014	APMA1650	Statistical Inference I
2014	APMA0650	Essential Statistics

Selected Honours and Awards

2016–2017 Blue Waters Graduate Fellowship, National Center for Supercomputing Applications

2012–2014 Integrative Graduate Education Research Traineeship, National Sciences Foundation

2012 Outstanding Presentation Honors for ”Application of a Hill Climbing Algorithm to Parallelize Graph-based Genome Assembly,” Joint Mathematics Meetings

Publications

1. Guang, A., Howison, M., Zapata, F., Dunn, C.W. Revising transcriptome assemblies with phylogenetic information. bioRxiv, October 2017.
2. Guang, A., Zapata, F., Howison, M., Lawrence, C.E. & Dunn, C.W. Better integrating the components of phylogeneic analyses. Trends in Ecology and Evolution, February 2016.
3. Hinkelmann, F., Brandon, M., Guang, B., McNeill, R., Blekherman, G., Veliz-Cuba, A., Laubenbaccher, R. ADAM: Analysis of Discrete Models of Biological Systems Using Computer Algebra. BMC Bioinformatics, 12(1):295, July 2011.

Presentations

1. **Guang, A.**, Howison, M., Coetzer, M., Dunn, C.W., Ledingham, L., D’Antuono, M., Chan, P., Kantor, R. Preserving intra-patient variance improves phylogenetic inference of HIV transmission. Blue Waters Symposium, Sunriver, OR, USA, May 2017.
2. **Guang, A.**, Howison, M., Coetzer, M., Dunn, C.W., Ledingham, L., D’Antuono, M., Chan, P., Kantor, R. Preserving intra-patient variance improves phylogenetic inference of HIV transmission. Center For AIDS Research Seminar, University of California San Diego, San Diego, CA, April 2017.

3. **Guang, A**. Summarizing population genome variation in phylogenetic analyses. Next generation phylogenetic inference spotlight session, Evolution meetings, Austin, TX, June 2016.

4. Guang, A, Hancock, L., Hollenbeck, E., Rand, D. Untangling morphotype and latitudinal variation in Spartina alterniflora. Ecology and Evolution Biology Brown Bag Seminar, Brown University, Providence, RI, September 2015.

5. **Guang, A.** Fablast: a fast method to cluster homologs. Applied Mathematics Graduate Student Seminar, Brown University, Providence, RI, 2014.

6. **Guang, A.**, Su, F. Switching between cooperation and competition in game theory. Pacific Coast Undergraduate Math Conference, Loyola Marymount College, Los Angeles, CA, 2012.

7. Guang, A., **Gendreau, A.**, Chen, X. A hillclimbing algorithm to parallelize graph-based genome assembly. Joint Mathematics Meetings, Boston, MA 2012.

8. **Guang, A.**, Gendreau, A., Chen, X. A hillclimbing algorithm to parallelize graph-based genome assembly. Nebraska Conference for Undergraduate Women in Mathematics, Lincoln, NK, 2012.

9. **Guang, A.** Fablast: a fast method to cluster homologs. Applied Mathematics Graduate Student Seminar, Brown University, 2014.

Posters

1. **Guang, A.**, **Hancock, L.**, Hollenbeck, E., Rand, D. Untangling morphotype and latitudinal variation in Spartina alterniflora. Evolution meetings, Portland, OR, June 2017.

2. **Guang, A.**, Howison, M., Coetzer, M., Dunn, C.W., Ledingham, L., D'Antuono, M., Chan, P., Kantor, R. Preserving intra-patient variance improves phylogenetic inference of HIV transmission. HIV Dynamics & Evolution, Isle of Skye, Scotland, May 2017.

3. **Guang, A.**, Howison, M., Coetzer, M., Dunn, C.W., Ledingham, L., D'Antuono, M., Chan, P., Kantor, R. Preserving intra-patient variance improves phylogenetic inference of HIV transmission. Blue Waters Symposium, Sunriver, OR, May 2017.

4. **Guang, A.** Fablast: a fast method to cluster homologs. Center for Computational Molecular Biology Research Convening, Brown University, 2014.

5. Hinkelmann, F., **Brandon, M.**, **Guang, A.**, McNeill, R., Blekherman, G., Veliz-Cuba, A., Laubenbacher, R. ADAM: Analysis of discrete models of biological systems using computer algebra, Joint Mathematics Meetings, New Orlenas, LA, 2011.

Acknowledgments

Chip, I can only hope to aspire to your levels of generosity, support, and vision. I have learned so much from you: clarity of communication, consistency, tenacity, and a sense of where the most impactful solutions are. I will try to take your sharp insights with me into my career.

Casey, this thesis would certainly not have happened without you. You have guided me through so much, offering opportunities at every turn and coaching me through career and life decisions. You have been a phenomenal mentor in not just research practicum, but also what I think of as the distilled essence of research: seeing where the most pressing problems are, keeping perspective on the bigger picture and human impact, implementing just enough for a minimum viable product, and gratitude for those who helped you up along the way. Thank you.

Paul, thank you for agreeing to be a member of my committee and for your helpful suggestions on my thesis and timeline. I really do want to try out PHYCAS some day... and I hope to continue to learn from your ideas.

To my Chip and Dunn lab groups, you've given me much feedback over the years, all of which has been useful. More importantly, you've given me a rich research community that has made this process actually really enjoyable. Thank you for letting me bounce ideas off in email or Slack and venting over drinks. A special thank you to Mark and Felipe who have worked so much with me and imparted so much computational and phylogenetic knowledge.

PrYSM family, Providence would not have been the same without you all. I can't say enough about what you've given me. You will always be family and my political home.

Mom, Dad, Young, thank you for believing in me throughout the good and bad. Thank you especially to Mom for all the breakfasts and steamed buns that helped me maintain perspective on what's really important.

Vanessa, meeting you through Pronk (shoutout to Jenny Li) four years ago was a dream come true. Who knew there could be multiple queer and trans scientists of color in this world? Thank you for supporting me through this adventure, and I'm so excited to follow you onto the next one.

Contents

Listing of figures

0

Introduction

Molecular phylogenetics is the study of evolutionary relationships between biological sequences, often to infer the evolutionary relationships of organisms. These studies require many analysis components, i.e. specific inference processes within an analysis to estimate random variables, including sequence assembly, identification of homologous sequences, gene tree inference, and species tree inference (see 0.2 for definitions of each component). At present, each component is usually treated as a single step in a linear analysis, where the output of each component is passed as input to the next as a point estimate. Such a linear analysis makes strong assumptions about order, dependence, and relative entropy of each step. These assumptions have tended to be implicitly understood rather than explicitly described, and method developments that relax these assumptions have focused narrowly on specific

downstream analysis components.

In this thesis I take a broader look at the whole phylogenetic workflow, clarifying implicit assumptions and suggesting general methods for improved communication between analysis components. These methods are rooted in a probabilistic framework and generative models. I apply this framework to upstream components of the phylogenetic workflow, showing that biological variation and technical errors in genome assembly can respectively affect inference of transmission clusters and gene trees further downstream.

In Chapter 1, I outline a probabilistic framework and generative model that helps clarify assumptions that are implicit to phylogenetic workflows, focusing on the assumption of low relative entropy. Using the probabilistic framework, I describe three communication strategies in data analyses, two of which relax assumptions about relative entropy of point estimates made in a linear analysis. This perspective unifies currently disparate advances, and will help investigators evaluate which steps would benefit the most from additional computation and future methods development.

In Chapter 2, I describe an integrated generative model for the generation of sequencing reads from individuals infected with Human Immunodeficiency Virus (HIV) that incorporates evolutionary processes touched on in Chapter 1 as well as additional epidemiological processes representative of contact and transmission between individuals who are either susceptible to or infected with HIV. This model serves as a foundation for assessing whether consensus genomes are an appropriate summary of an infected individual's viral population in transmission cluster inference. In other words, I develop a generative model to analyze to what extent the violation of assumed low relative entropy between the consensus genome estimate and within-host viral genomes in assembly affects transmission cluster inference. In the chapter I additionally describe and provide a simulation tool, transmissim, that implements the generative model.

As described in Chapter 1, much of the work in improving information communication has been on downstream components such as gene tree inference and species tree inference. Thus, in Chapters 3 and 4, I focus on adopting strategies of information communication described in Chapter 1 between upstream components of the phylogenetic workflow to address violations of low relative entropy assumptions that affect inference.

Chapter 3 specifically focuses on violations of the assumption in the transmission inference workflow described in Chapter 2. I conclude that employing consensus genomes for HIV transmission inference can lead to erroneous results, and provide an alternative method to summarize within-host viral variation based on Hidden Markov Models. This alternative method, termed a profile-sampling approach, propagates multiple HIV genome hypotheses in one direction, employing one of the communication strategies discussed in Chapter 1. I show that the profile-sampling approach has higher phylogenetic bootstrap support on simulations generated with transmissim and was able to identify transmission clusters in a real dataset with higher resolution.

In Chapter 4, I take advantage of the Markovian dependence assumption to iteratively revise transcriptome assemblies using phylogenetic information. One of the most common transcriptome assembly errors is to mistake different transcripts of the same gene as transcripts from multiple closely related genes. It is difficult to identify these errors during assembly, but in a phylogenetic analysis these errors can be diagnosed from gene trees containing clades of tips from the same species with improbably short branch lengths. I present a tool, *treeinform*, that uses this phylogenetic information to refine transcriptome assemblies. This iterative revision has the potential to be developed into simultaneous estimation, although we do not present a plan to do so in this thesis.

The central theme throughout these chapters is a quest to understand to what extent violations of implicit assumptions in phylogenetic workflows negatively impact analysis results

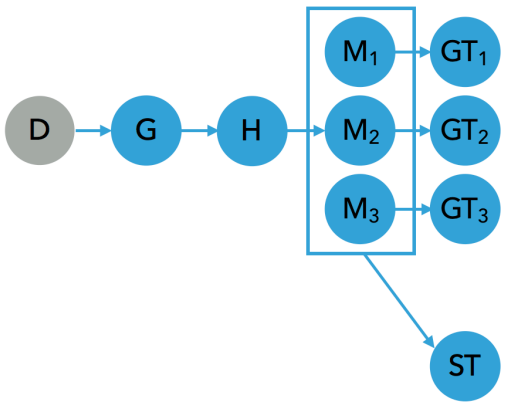
and derive solutions to improve communication between analysis steps. In particular, I focus on relaxing the assumption of low relative entropy in genome and transcriptome assembly, i.e. the assumption that a point estimate sufficiently represents all possible assembly hypotheses. Figure 1 shows graphical representations of the analysis components and random variables within the integrated perspective and will be referenced in each chapter as we progress.

The studies and analyses throughout these chapters have been written in collaboration with my advisors and colleagues, in particular Mark Howison and Felipe Zapata. Each chapter includes text and materials from either published manuscripts or preliminary drafts. At the start of each chapter I have indicated the materials I was responsible for within the chapter. In the next section I additionally provide a glossary of definitions for various components and processes, which will be referred to throughout the thesis.

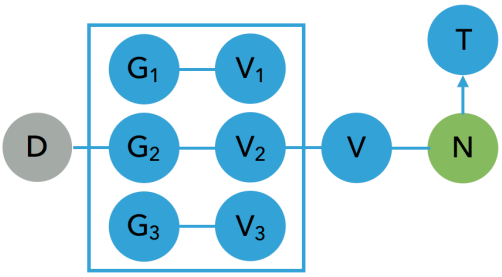
0.1 Glossary of random variables in the phylogenetic workflow

Here we provide a set of definitions for the random variables in the phylogenetic workflow, as shown in Figure 1.

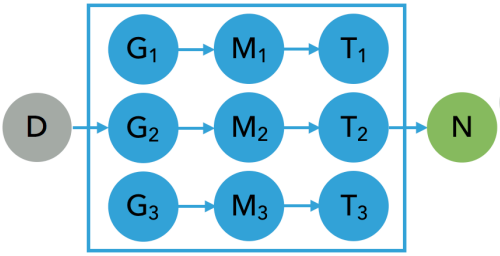
- **Data (reads):** represented as D in Figure 1. The input data of the pipeline, in the phylogenetic workflow these are usually fragments of DNA or RNA.
- **Genome:** represented as G in Figure 1a. The original DNA sequences of the set of organisms that will be placed relative to each other on the phylogenetic tree. In Figures 1b and 1d, this is further broken down into $G = \{G_1, G_2, G_3, \ldots\}$. In Ch2, $G_1, G_2, G_3, \ldots$ represents genomes from each individual i's viral phylogeny, and in Ch3 $G_1, G_2, G_3, \ldots$ represents genomes from each sample i drawn from the HMMs for each individual that will later be used for a phylogenetic analysis.
- **Transcriptome:** represented also as G in Figure 1d. The original RNA sequences, including splice and allelic variants, of the set of organisms that will be placed relative to each other on the phylogenetic tree.



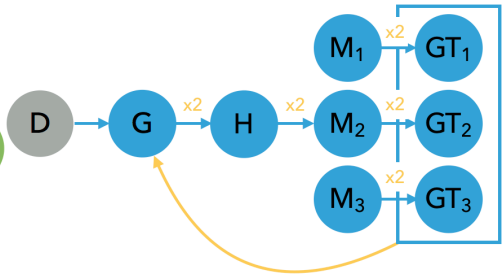
(a) Chapter 1: An integrated perspective on phylogenetic workflows



(b) Chapter 2: An integrated generative model of HIV transmission and viral evolution



(c) Chapter 3: Propagating intra-patient variation in transmission network inference



(d) Chapter 4: Revising transcriptome assemblies with phylogenetic information

Figure 1: Diagrams representing overviews of the random variables from analysis components and models, and proposed methods of information communication within each chapter. Note that while in all workflows and models there exist multiple random variables for the same component, the reasons for multiple random variables differs. In Chapters 1, 2 and 4 those random variables are present due to actual variation in either rates of gene evolution or transmission times, while in Chapter 3 those random variables represent iid samples, or hypotheses, drawn from the analysis component. Refer to Section 0.1 for a set of definitions for the random variables.

- **Homologous sequences:** represented as H in Figure 1a and 1d. Genes, or stretches of DNA, that are descendants from a common ancestral sequence.
- **Multiple sequence alignment:** A matrix consisting of the set of sequences to be analyzed, where the rows represent each individual sequence, and the columns represent the alignment of sites in each sequence. The alignment can be interpreted as a representation of molecular evolutionary processes acting on the ancestral sequence; gaps indicate insertions and deletions, and mismatches indicate substitutions. Each alignment is an intermediate step to the phylogenetic tree.
- **Gene tree:** A phylogenetic tree of a set of homologous sequences considered to belong to the same gene family.
- **Species tree:** A phylogenetic tree of a set of sampled individuals from different species.

0.2 Glossary of analyses components within the phylogenetic workflow

Here we provide a set of definitions for the analysis components of a typical phylogenetic workflow. When we refer to **analysis components,** we mean specific inference processes within an analysis workflow to estimate random variables. These are further expanded upon in Chapter 1. In a phylogenetic analysis, these components include:

- **Assembly:** the technical process of aligning and merging fragments of DNA or RNA sequences to estimate the genome or transcriptome, respectively.
- **Homology evaluation:** in this context, the technical process of identifying homologous sequences, typically through phenetic comparisons of sequence similarity as a proxy for evolutionary descent from a common ancestor.
- **Multiple sequence alignment:** the inference of site homology within sets of homologous sequences by partitioning sequence variation into indels (insertions and deletions) and site substitutions.

- **Phylogenetic inference:** the inference of a phylogenetic tree from a multiple sequence alignment, typically by assuming site independence in the alignment and computing sets of probable phylogenies for each site.
- **Gene tree inference:** phylogenetic inference from a set of homologous sequences considered to belong to the same gene family.
- **Species tree inference:** phylogenetic inference in order to determine evolutionary relationships between species. The inference may either be based on a multiple sequence alignment of concatenated gene alignments, or from multiple gene trees.
- **Transmission tree inference:** phylogenetic inference in order to identify transmission links between individuals. Individuals who share a most recent common ancestor on the phylogeny with high confidence, typically approximated with bootstrap support, are determined to share a transmission link.

1

An integrated perspective on phylogenetic workflows

This chapter reflects the work published on Trends in Ecology and Evolution, doi: http://dx.doi.org/10.1016/j.tree.2015.12.007. My role in the paper was formulating a probabilistic perspective on phylogenetic workflows rooted in generative models of the evolutionary and technical processes, identifying implicit assumptions made in phylogenetic workflows, conducting a literature review, and co-writing and editing the manuscript.

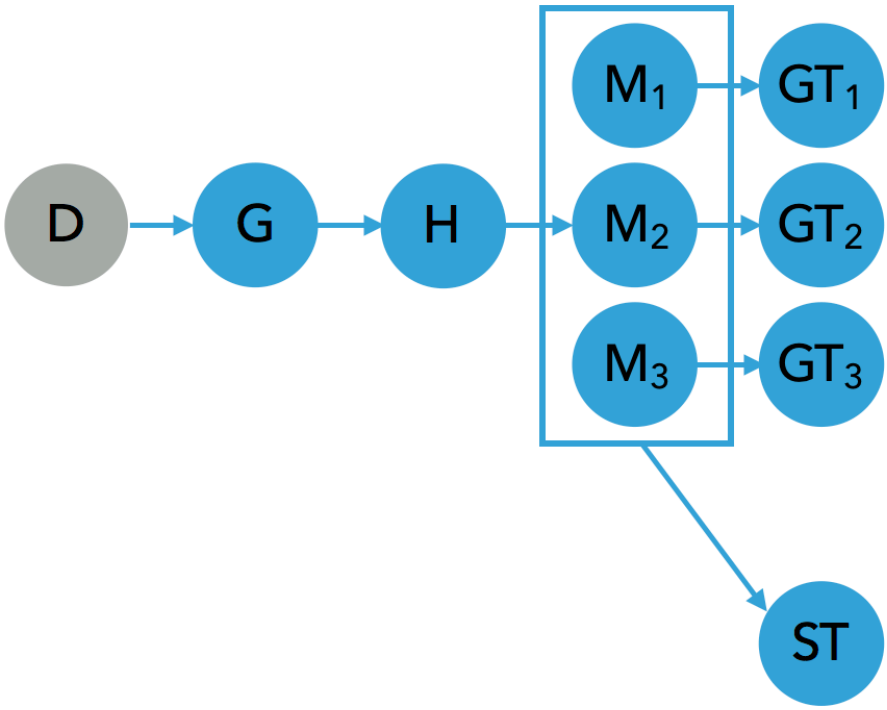


Figure 1.1: An overview of the linear phylogenetic workflow that will be described more in this chapter. The absence of edges indicates the assumption of conditional independence between any pair of random variables not connected by an arrow. This graph assumes a 1st order Markov dependence, i.e. that each state only depends on the previous state.

1.1 Molecular phylogenetics and information communication

Molecular phylogenetic analyses have multiple components, including assembly of raw sequence data into gene sequences, identification of homologous sequences across species, multiple sequence alignment, inference of gene trees, and integration of information across gene trees to infer species trees. How each of these analysis components is implemented and integrated is an important decision in designing a phylogenetic study[2]. In most phylogenetic analyses, each analysis component is treated as a separate step in a linear workflow, in which results from each component are passed as input to the next component. These results are usually communicated as a single hypothesis (i.e., a point estimate). For example, a single hypothesis on gene homology is estimated and passed on to indicate which gene sequences belong to the same gene tree.

There are multiple limitations, each of which reflects simplifying assumptions about the data and methods, with phylogenetic workflows that pass only a single hypothesis in a single direction between each analysis component. These workflows cannot accommodate the uncertainty present in the data or introduced in the inference process. In addition, a strictly stepwise workflow doesn't allow for interactions between analysis components when different stages cannot be solved independently. Biologists have long recognized these limitations in the context of particular analysis steps, including sequence homology evaluation[71], multiple sequence alignment[123], and species tree inference[50]. This has spurred important methods development to relax some of these limitations, including the use of Bayesian approaches to accommodate uncertainty when inferring phylogenetic trees from multiple sequence alignments[91] and the simultaneous estimation of gene and species trees[67,6,18,116].

While there has been much productive work on understanding and relaxing these limitations at particular points in phylogenetic analyses, there has been little work on the systematic

evaluation of these limitations across the entire phylogenetic workflow. An integrated workflow perspective can make several types of important contributions. First, a better understanding of workflow assumptions will lead to better interpretation of the results of current tools. Second, this integrated perspective can provide clear criteria for prioritizing future methods development. In particular, improving communication between largely independent analysis components can relax some simplifying assumptions. Improved communication reduces the accumulation of errors across components[5], accommodates interactions between components, provides more information to the investigator, and enables statistically grounded interpretations of results. These improvements come at the cost of increased engineering and computational costs, and this integrated perspective provides a way to better evaluate the tradeoffs between these costs and the benefits to the investigator.

Here we present a unified framework for understanding information communication in molecular phylogenetic analyses that brings together advances on particular analysis components, helps identify what methods and tools are missing, and provides a holistic context for evaluating which of these should be the highest priority for future development. This will lead to more informed decisions about how to allocate computational resources to different analysis components to maximize investigator insight. This framework is grounded in a generative model for raw sequence reads that reflects the biological and technical processes that underlie the observed data. This model provides the unified perspective needed to state the assumptions implicit to information communication between phylogenetic analysis components, and places these assumptions in the context of a general probabilistic framework that facilitates statistical interpretation of results.

1.2 A molecular phylogenetic generative model describes how unobserved processes generate observed sequence data

A generative model is an explicit hypothesis of the natural and technical processes that produce observed data. Because generative models are based in probability theory, they provide principled means to describe the entities and processes investigators seek to understand. An explicit generative model therefore clarifies the goals and role of each processes in an analysis workflow. Generative models can also be used to guide simulation, which is helpful to evaluate methods and make decisions about project design before data collection. Generative models have been employed with considerable success in many fields[4,12,15,70,61,89,34].

A generative model can be represented graphically (Figure 1.2, orange). Graphical models are a way to describe joint probability distributions. Nodes on a graphical model are random variables that describe the observed data and unobserved biological entities that the investigator would like to know about. The edges are dependency relationships between the random variables, and represent natural and technical processes that connect these entities. Edges can be directed or undirected. Undirected edges in a graph describe Markovian dependence. Directed edges in a graph describe conditional dependencies. The nodes are random variables that describe the observed data and unobserved biological entities that the investigator would like to know about. The edges are natural and technical processes that connect these entities and describe their dependence.

A generative model for molecular phylogenetics must incorporate multiple random variables as well as evolutionary, cellular, and technical processes (Figure 1.2, orange). Explicit generative models have already been described for some subcomponents of a phylogenetic workflow, such as the generation of gene sequences given a phylogenetic tree and an explicit model of molecular evolution[31], or the generation of gene trees given a species tree[116,5,21]. A

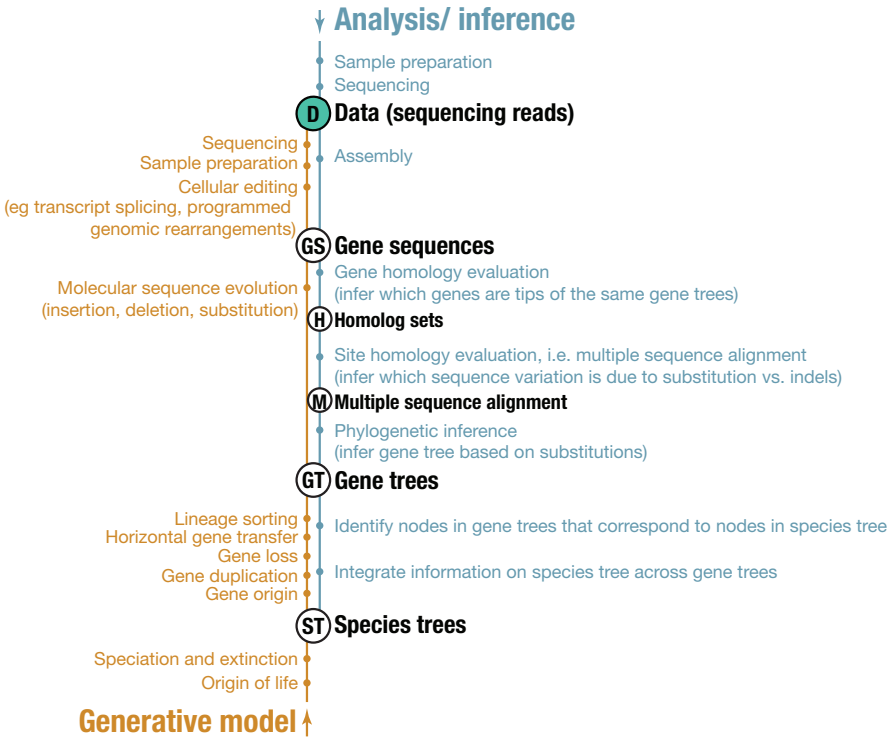


Figure 1.2: Graphical Representations of a Generative Model and Phylogenetic Analyses, under the Assumption of Markovian Dependence. Large open circles indicate biological entities, which can be described as random variables in the inference process. Edges (lines) represent processes. In the case of the generative model (orange), these are biological and technical processes that give rise to the entities. In the case of the inference process (blue), they are analysis processes that consist of one or more analysis components. Random variables that represent intermediate analysis products that are specific to inference are shown with small circles. The assumption of Markovian dependence specifies that variables depend only on neighboring variables, relaxing it would give a more general model wherein there are more connections in addition to those between adjacent variables.

generative model can be extended to an entire phylogenetic analysis, from species tree through to the generation of raw sequence data. To simplify presentation, we here assume Markovian dependence in the generative model, which gives it a linear structure. The origin of life and speciation-extinction result in species trees that describe the structure of populations of individuals (ST, Figure 1.2). The origin of new genes (e.g., gene duplication, gene loss, and horizontal gene transfer between species) results in gene trees (GT, Figure 1.2)[114]. The species trees impose structure on these gene trees through lineage sorting[21] and by defining barriers to gene flow that gene trees must obey. If lineage sorting is complete and there is no horizontal gene transfer, then nodes in the gene trees correspond directly to nodes on the species tree. Otherwise there is not a one-to- one correspondence[84]. Processes of molecular evolution, including insertion, deletion, and substitution, result in the diversity of gene sequences (GS, Figure 1.2) present at the tips of these gene trees. These sequences can be further edited within the lifespan of an organism through cellular processes such as programmed genome rearrangements[57,105] and RNA splicing in transcriptomes. When the investigator isolates molecules from the organism and prepares them for sequencing they are often further modified through fragmentation, ligation, and errors introduced during amplification. Finally, sequencing produces raw reads (D, Figure 1.2) that are estimates of the sequences in these prepared samples. Sequencing instruments introduce errors during this process and have ascertainment biases that make it more difficult to observe some sequences than others[97]. The raw reads (D) are the observed data, and the other random variables (GS, GT, and ST) are unknowns that the investigator seeks to estimate from the observed data.

Simulation tools have already been developed for most of the steps in this generative process, including raw sequence reads from gene sequences[72,49,10], gene sequences given gene trees[111,35,11,110], and gene trees given species trees[104,44,75]. Even so, there is not yet an integrated tool for simulating raw data under all the processes in a phylogenetic analysis that

biologists are regularly interested in.

1.3 A molecular phylogenetic workflow makes inferences about unobserved variables in observed sequence data

The goal of an analysis workflow is essentially to undo a generative model. While a generative model explains how observed data are generated from unobserved variables, an analysis workflow estimates unobserved variables from the observed data. With a generative model that is a 1st order Markov chain, inference of the unobserved variables from the observed data can be done with priors and in a forward step marginalized over variables within the generative model at each stage. A Markov chain of order m (where m is finite) is a process where the next state depends on the past m states. Thus, the next state in a 1st order Markov chain only depends on the previous state. In the case of linear models and workflows, which are 1st order Markov chains, there is an antiparallel relationship between the generative model (Figure 1.2, orange) and the analysis workflow (Figure 1.2, blue)[116].

Here, we describe a phylogenetic analysis workflow that is typical of many currently implemented, though often differences exist in goal and design details between studies. Different data-acquisition approaches can simplify various analysis components. Targeted enrichment, for example, greatly simplifies assembly and homology evaluation, but at the cost of only providing data for pre-selected genes[63].

The first analysis step begins by estimating gene sequences (GS, Figure 1.2) from the observed data (D, Figure 1.2). This analysis process is referred to as assembly. It identifies which sites in different reads correspond to the same sites in the original molecules, identifies the relative location of reads to each other, corrects technical errors introduced by sample preparation and sequencing, and accommodates and annotates some aspects of cellular

editing (such as splice variation)[78,82]. Changes in sequencing technology have had major impacts on assembly methods[78,7]. It is anticipated, for example, that read lengths will be much longer in the near future, even extending to the full length of the biological molecules under study[54,98,14]. This will make it much simpler to identify the location of reads, but does not obviate assembly as it is still necessary to identify which reads are derived from the same molecules and to correct technical error.

The next analysis challenge is to proceed from gene sequences (GS, Figure 1.2) to gene trees (GT, Figure 1.2). Most phylogenetic tools assume homology, so the first step in this process is to identify homologous sequences through phenetic comparisons of sequence similarity[87]. This is typically implemented in a few steps. First, pairwise sequence comparisons and clustering are used to identify homologous sequences, i.e. sequences that are tips in the same gene tree. Some sequence homology tools compare all sequences to each other[23], others compare new sequences to a curated set of pre-selected sequences[25]. Next, site homology is evaluated within each set of homologous sequences by multiple sequence alignment (MSA). MSA partitions the variation observed among homologous sequences into two categories: variation due to insertion and deletion (which results in sites placed in different columns) and variation due to substitution (which is contained within a single column)[69]. Homology evaluation and MSA generate two random variables not present in the generative model, which do not represent estimates of entities that occur in nature (H, M, Figure 1.2). With the alignments in hand, the investigator can proceed to the core of a phylogenetic study – phylogenetic inference. Given the gene sequences and a model of molecular evolution, phylogenetic inference tools evaluate alternative hypotheses regarding the evolutionary relationships between the sequences[31,96,108]. For historical and computational reasons these tools typically model only site substitution, and the insertion and deletion events identified in MSA are not evaluated.

Inferring species trees (ST, Figure 1.2) requires both the identification of gene tree fea-

tures that correspond to features of the species tree (i.e., nodes due to speciation that result in orthologs) as well as the integration of information from many gene trees to learn about their shared history constrained through speciation and lineage sorting. These steps are implemented in many different ways in different linear studies, including consensus trees[9] and matrix concatenation[19] approaches.

1.4 The implicit assumptions of phylogenetic workflows

Several implicit assumptions are usually made to simplify analyses and reduce their computational cost, resulting in one extreme in the linear type of workflow described above. Making these assumptions explicit is critical to understanding the limitations imposed on results by current methods and establishing future research priorities. Here we describe three assumptions that are central to phylogenetic workflows, explain how they allow for the linear workflow in Figure 1.2, then discuss whether these assumptions are justified and if violations of the assumption jeopardize the interpretation of the results.

1.4.1 Assumption 1. Biologically sensible ordering of steps

Biologically sensible ordering of steps justifies the direction of information propagation in typical phylogenetic workflows. The generative model is a hypothesis of the technical and biological processes of data generation. The phylogenetic analysis workflow attempts to undo the generative model. To be biologically justified, the ordering of analysis steps should therefore be the reverse of the processes in the generative model. The ordering of analysis steps therefore reflects assumptions about the generative model. The ordering of steps in the generative model presented here (Figure 1.2) is based on extensive understanding of the processes at play, so the ordering of analysis steps is well justified.

1.4.2 Assumption 2. Markovian dependency

Markovian dependency justifies the consideration at each analysis component of only the results of the preceding component. In probability theory, a Markov process is a process such that inference of the current state depends directly only on its neighboring states. For first order Markov processes this is the equivalent to the more commonly used definition, where the inference of the next state depends only on the past states. To see this, suppose we have a 1st order Markov chain $X \rightarrow Y \rightarrow Z$. By the properties of joint probabilities, we have $P(Z|X,Y)P(X|Y) = P(Z,X|Y) = P(X|Z,Y)P(Z|Y)$. Thus, $P(Z|X,Y) = P(Z|Y) \iff P(Z|Y)P(X|Y) = P(Z,X|Y) \iff P(X|Z,Y) = P(X|Y)$, meaning that inference of each step depends directly only on its neighboring states and not just the past states. A linear phylogenetic analysis workflow assumes a Markovian dependency structure because each analysis component only needs information from the previous component and/or the next component. For example, homology can be inferred from only the assembly and/or the multiple sequence alignment, and the multiple sequence alignment can be inferred from only the homology and/or the gene tree.

1.4.3 Assumption 3. Low relative entropy

Passing only a single hypothesis from one analysis component to another assumes that the result of an analysis component can be reasonably summarized by a single point estimate. This depends critically on the probability distribution of the result. Some probability distributions can be adequately summarized by a single estimate, others cannot. If the resulting distribution is unimodal and has low variance, then a single estimate such as the mean is a good approximation of other results that would be drawn from the distribution if it were resampled 1.3. If instead the resulting distribution is multimodal and has high variance, then repeat sampling

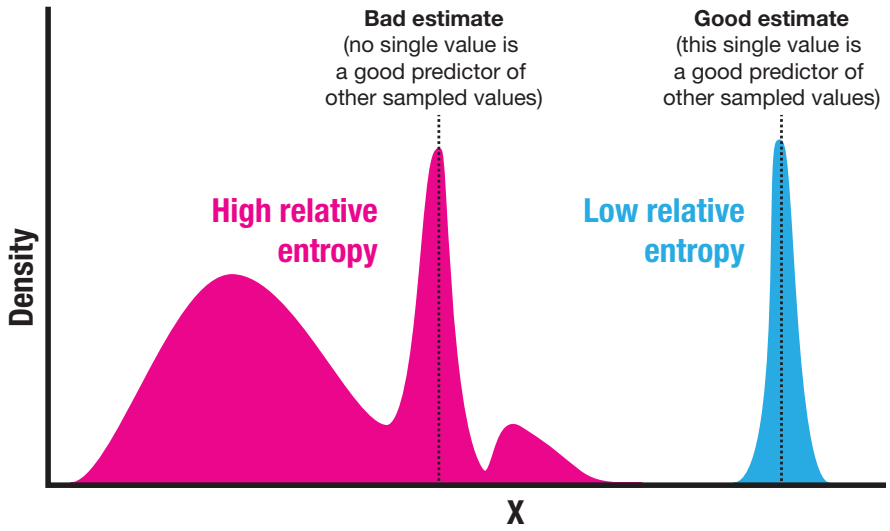


Figure 1.3: Two probability density functions representing random variables. The relative entropy between the magenta (left) distribution and the point estimate is high due to the high variance and a complex shape of the distribution. No single estimate is a good predictor of other values drawn from this distribution, but two point estimates can provide a much better approximation. The relative entropy between the cyan (right) distribution and the point estimate is low. A single estimate drawn from this distribution is a good predictor of other values.

would generate widely different results and none of them is a good description of the distribution as a whole. This has been shown for phylogenetic trees and for local sequence alignment[100,83].

In information theory, the concept of relative entropy[17,102] describes these differences in terms of how well one distribution is represented by another. It helps us understand when a point estimate is sufficient to describe a distribution and when more information is required. A low relative entropy means that the representation distribution is close to the target distribution 1.3. Current phylogenetic analysis workflows assume low relative entropy for most analysis components because they pass only a point estimate (e.g., a single multiple sequence alignment) from one component to another.

1.4.4 A probabilistic perspective on the implicit assumptions of phylogenetic workflows

The five random variables in the phylogenetic analysis workflow (Figure 1.2) - GS (gene sequences), H (homology), M (multiple sequence alignment), GT (gene trees), ST (species tree) - all concern high dimensional unknowns. A fully general model assumes that these five random variables are all directly related to one another. For example, in order to compute the probability of any random variable given D (data), say ST, we have to sum over all the values of the other unknowns:

$$\begin{aligned} P(ST|D) &= \sum_{GS}\sum_{H}\sum_{M}\sum_{GT} P(GS, H, M, GT, ST|D) \\ &= \sum_{GT} P(ST|GT, M, H, GS, D) \sum_{M} P(GT|M, H, GS, D) \sum_{H} P(M|H, GS, D) \\ &\sum_{GS} P(H|GS, D)P(GS|D) \end{aligned}$$

Current linear phylogenetic workflows make the assumption that these random variables form a one-dimensional Markov random field, i.e. a Markov chain. Specifically, this model specifies that

$$P(GS, H, M, GT, ST|D) = P(ST|GT, D)P(GT|M, D)P(M|H, D)P(H|GS, D)P(GS|D)$$

and that for any particular random variable, for example M,

$$P(M|H, GS, D) = P(M|H)$$

and

$$P(\mathcal{M}|GT, ST) = P(\mathcal{M}|GT).$$

Estimating the probability of $ST|D$ then becomes

$$P(ST|D) = P(ST|GT, D) \sum_{GT} P(GT|\mathcal{M}, D) \sum_{\mathcal{M}} P(\mathcal{M}|H, D) \sum_{H} P(H|GS, D) \sum_{GS} P(GS|D).$$

Note that in general a graphical model does not have to assume that each random variable in an ordered process is conditionally independent of all random variables except for the previous stage, however this increases the computational complexity significantly. In between these two extremes are models that are conditionally independent on some, but not all of the random variables beyond its immediate predecessor. Chapter 4 shows an example of the need for a model between the extremes, specifically with the inclusion of a direct dependence of gene trees on the transcriptome assembly. This is discussed more in Section 1.5 as well.

A Markov model is specified in the forward direction, for example the model specifies the dependence of homology on the gene sequence, $P(H|GS, D) = P(H|GS)$. However, this specification also produces a dependence in the backward direction as can be seen using Bayes rule: $P(GS|H) = \frac{P(H|GS)P(GS)}{P(H)}$. Generative inference procedures account for this bidirectional dependence using a forward algorithm to account for upstream effects and a backward algorithm to account for downstream effects[Durbin et al.]. The fact that current phylogenetic workflows employ only backward steps (e.g., gene trees depend on assembled gene sequences, but sequence assembly doesn't use any information about gene trees) imposes important limitations. For example, one of the hardest problems in sequence assembly is correctly deconvoluting repeats. One biological source of repeats is gene duplication. Since a phylogenetic workflow includes

gene tree inference, this analysis component could help the assembler correctly assemble regions that repeat due to gene duplication, but this strategy is not currently employed.

In addition, the typical phylogenetic workflow makes the strong assumption of low relative entropy, so that summing over all possible hypotheses is not required, as the probability of the point estimate is close to 1. Given these two assumptions, estimating $ST|D$ becomes

$$P(ST|D) = P(ST|GT) \max_{GT} P(GT|\mathcal{M}) \max_{\mathcal{M}} P(\mathcal{M}|H) \max_{H} P(H|GS) \max_{GS} P(GS|D) \approx 1.$$

This equation embodies the linear workflow (Figure 1.2): the simplest possible inference process of species trees from the data that passes minimal information, i.e., a point estimate, from one analysis component to the next under the assumptions that there is a biologically relevant ordering of steps, the steps form a Markov chain, and that relative entropy between the point estimate and the true distribution is low.

1.4.5 Are the implicit assumptions justified?

Explicitly stating these assumptions raises the question of whether they are routinely violated in real-world analyses and if these violations jeopardize the interpretation of the results. These are empirical as well as theoretical questions. The ordering of the analysis steps is justified inasmuch as this representation of evolution is widely accepted as credible. The Markov chain assumption could be violated in a number of ways. For example, duplication events within a gene tree are a source of repeats encountered in assembly. Hence a dependency exists between gene trees and gene sequence assembly, and the assembly process could benefit from consideration of gene tree inference. In addition, population size can simultaneously impact multiple processes, including speciation, rates of molecular evolution, and incomplete lineage sorting.

There is good reason to believe that the third assumption of low relative entropy is frequently violated in ways that negatively impact phylogenetic analyses. There is often considerable uncertainty regarding each step in a phylogenetic analysis[71,123,50,47], and discarding information about other hypotheses that also find support from the data can mislead analyses and greatly complicates their biological interpretation. This concern was in part the motivation for the field of phylogenetics moving away from single point estimates of trees toward the generation and presentation of whole distributions of results, such as Bayesian posterior distributions of trees and bootstrap replicates[46]. This makes relaxing the assumption of low relative entropy a particularly high priority, and is therefore the focus of this thesis.

Two different communication strategies can relax the assumption of low relative entropy. One is to propagate multiple hypotheses from one component to another, providing a distribution of hypotheses. Another is to jointly estimate multiple components, allowing for a comprehensive assessment of uncertainty contributed by the components. They are expanded upon below.

1.5 Three communication strategies in data analyses

Most data analyses, including phylogenetic analyses, require multiple components that each address a subproblem in a larger analysis challenge. These analysis components often consist of separate tools that were designed to work together or were repurposed in new ways so that they could be combined into novel workflows. Data and intermediate analysis results must be communicated between these components for them to work together. While much effort has been put into the effectiveness and efficiency of each of these components, how they communicate and what information they share is a major scientific and engineering challenge that often receives less technical and theoretical attention. Here we consider three specific commu-

nication approaches between analysis components in scientific computing.

The first is to communicate a minimum amount of information by propagating one hypothesis (i.e., a point estimate) such as the maximum a posteriori (MAP estimator in a single direction. Sometimes this hypothesis is chosen according to ad hoc criteria, and other times it is statistically explicit (such as the selection of a phylogenetic hypothesis by maximum likelihood). This is the most common communication strategy currently implemented in phylogenetic workflows.

The second approach is to communicate more information by propagating multiple hypotheses in a single direction. This allows uncertainty and error to be communicated from one analysis component to another. In phylogenetics, this approach has been used for inferring a species tree given a sample of gene trees[59]. This strategy has also been explored in other fields, such as cognitive psychology and hydrological modeling[56,13].

The third approach is to combine analysis components into a single component that simultaneously infers multiple hypotheses that were otherwise estimated independently and sequentially. This approach fully accommodates the uncertainty present in the data and generated during the inference process, as well as the non-independence of solutions between analysis components. Some phylogenetic species tree methods use this approach[18]. Some tools also allow for simultaneous alignment and homolog clustering to entities from a specific type of database, such as Pfam[33] for protein families.

Changing the information communicated between analysis components can require that the analysis components themselves are re-engineered. Many tools that are critical to phylogenetic analysis are built to input and output only point estimates. Assemblers output single assembly hypotheses. Multiple sequence aligners accept only a single estimate of each gene sequence and output only one possible multiple sequence alignment. The simplest approach to implementation is to iteratively run these existing tools on a distribution of inputs, and then

propagating the distribution of outputs to later steps. This assumes that the additional parameters of each step are known and fixed. This comes with minimal implementation costs since the tools can be used as-is, but is computationally quite expensive. To make improved communication tractable, tools will need to be re-engineered to intrinsically accommodate multiple hypotheses or to simultaneously infer multiple steps.

The way that hypotheses are chosen has important implications for their interpretation, regardless of how those hypotheses are communicated to other analysis steps. Hypotheses are often chosen according to ad hoc criteria, such as minimizing gap penalties in multiple sequence alignment. If they are instead selected according to statistically explicit criteria, such as an approximation of the posterior distribution under an explicit probabilistic model and subsequent sampling in proportion to their probability, their interpretation is much clearer.

1.6 Existing approaches that relax the assumption of low relative entropy

The phylogenetic analysis workflow described above (Figure 1.2, blue) represents one extreme–it communicates the minimal amount of information in the simplest possible way. Even so, it has been widely adopted due to two key practical advantages: it is the most computationally tractable (since all but one hypothesis is discarded at each step) and it is technically the most straightforward to implement (since it can be built from existing tools). Implementing this workflow as a software pipeline has varied from completely manual analyses, where information is formatted and passed to each component by hand or semiautomated tools, to fully automated workflows that also track and summarize results[23,40,85,113].

Most of the recent work on improving communication has focused on the joint estimation of gene trees and a species tree[67,6,1]. This focus is motivated in part by the fact that gene trees and species trees should not necessarily be expected to be congruent[73], and that the incongru-

ence can only be correctly accounted for by simultaneously estimating both gene and species relationships. Incomplete lineage sorting has been the primary focus of most recent work in this area, largely due to the mature mathematical and statistical foundation provided by coalescent theory[67,20,45,74]. Nevertheless, recent work now also accounts for other sources of incongruence such as gene duplication and loss[6,116,109,3,93]. An interesting recent approach to this problem is the work of Martins et al.[18], which considers a Bayesian posterior distribution of gene trees during species tree estimation.

Efforts to improve communication at earlier steps have been largely neglected. Current approaches treat homology evaluation (the identification of homologous sequences), multiple sequence alignment (the identification of homologous sites in homologous sequences), and phylogenetic tree inference (identification of phylogenetic trees that best explain the variation in homologous sites) as separate problems. Better integration of multiple sequence alignment and phylogenetic tree inference has been one focus of work[122]. Several approaches have been developed to improve communication between these two analysis components by coestimating multiple sequence alignment and gene trees inference under parsimony[117], maximum likelihood[65,66], and Bayesian approaches[94].

1.7 Joint estimation requires an integrated generative model

Under the assumption of Markovian dependence, as well as the assumption that each gene tree is conditionally independent given the species tree, we can describe a high-level integrated generative model for the unobserved processes that generate observed sequence data.

Let ST be the species tree, n be the number of gene families, GT be the collection of gene trees $GT_1, \ldots, GT_n$, G be the collection of sequence data $G_1, \ldots, G_n$, and D be the collection of read data $D_1, \ldots, D_n$. Furthermore, assume that each gene tree is conditionally independent

given the species tree, that G_i only depends on GT_i, and also that D_i depends only on G_i. Finally, let x_i represent parameters in gene tree evolution, y_i represent parameters in molecular evolution and z_i represent parameters in sequencing technology. Up to this point we had assumed the parameters of the model were known and fixed. This now gives us a model that explicitly includes the parameters as unknowns:

$$\begin{aligned} P(D, G, GT|ST) &= \prod_{i=1}^{n} P(D_i, G_i, GT_i|ST) \\ &= \prod_{i=1}^{n} P(D_i|G_i, GT_i, S)P(G_i|GT_i, ST)P(GT_i|S) \\ &= \prod_{i=1}^{n} P(D_i|G_i, \mathrm{z}_i)P(\mathrm{z}_i|G_i) \int P(G_i|GT_i, \mathrm{y}_i)P(\mathrm{y}_i|GT_i)d\mathrm{y}... \int P(GT_i, \mathrm{x}|S)d\mathrm{x} \end{aligned}$$

In order to generate the data D given the species tree S then, we generate n gene trees GT_i independently given S, then generate each set of sequence data D_i from G_i, and finally concatenate D_i together into D by species. A subset of this model limited to only two random variables is what simultaneous estimation tools like phyldog[6] and BALIphy[94] use. It would be possible to build a simultaneous estimation tool over all the unobserved processes, but such a tool would have minimal use due to computational limitations. Other simultaneous estimation schemes are possible as well; sampling is an alternative for these schemes when integrals can't be computed.

1.8 Identifying future priorities for improved communication

There are clear benefits to relaxing the assumption of low relative entropy through improved information communication between analysis components, including the ability to evaluate results according to statistically explicit criteria and compare support for alternative hypotheses.

These potential benefits, however, come with costs that must be considered when setting priorities for future work. The cost of implementation can be very high, as many tools will have to be reengineered to accommodate multiple hypotheses or information from non-neighboring analysis components. Most investigators already struggle with the computational costs of existing tools, so understanding the relative benefit of increasing the cost of any particular step is critical to making informed decisions about which investments will have the greatest benefit.

Sensitivity analyses can play an important role in evaluating the benefits of relaxing or strengthening assumptions. If the investigator relaxes an assumption by a far more computationally expensive approach but this has little impact on the result, then the cost of adopting the more general approach might not be justified.

As discussed above, recent advances have largely focused on improving information communication between gene tree and species tree inference. Recent reviews have addressed these advances in detail[116,58,30], hence we here focus on opportunities to relax the assumption of low relative entropy by improving information communication between upstream analysis components.

Relaxing the assumption of low relative entropy of assembly output will likely have great benefit to researchers[47]. Most assemblers choose a single hypothesis according to ad hoc criteria, assuming it to be an adequate point estimate of the original gene sequence (e.g., [126,103,39]). This reduces computational cost, but at the price of discarding uncertainty in the results and neglecting important biological information such as variation due to heterozygosity or somatic polymorphisms. Generating multiple assembly hypotheses for each gene and propagating them to homology evaluation would create the opportunity to assess alternative assembly hypotheses from a more informed position, where they can be compared to homologous sequences from other species. Rather than only evaluate assembly hypotheses according to information available within each species, information would be borrowed across species.

For example, chimeric gene sequences, formed by the spurious fusion of sequences from two different genes, are particularly common and problematic[124]. They confound homology evaluation by creating spurious sequence similarities between disparate gene families. These spurious similarities provide signatures during homology evaluation (e.g., high degree centrality in graphs of hypothesized homologous sequences) that make it possible to identify them and remove them[80]. Retaining multiple gene sequence hypotheses and passing them all forward would allow the analysis workflow to identify and retain non-chimeric gene sequences at this later stage.

Implementing this improved communication strategy between assembly and homology evaluation requires minimal cost when analyzing transcriptome data. Transcriptome assemblers already generate multiple assembly hypotheses per gene in order to accommodate splice variants, however some of these variants are assembly errors rather than true alternative splice variants[124]. In addition, homology evaluation tools will not require much reengineering to accept multiple assembly hypotheses, as it would just be a matter of performing the same sequence similarity comparisons but on a larger dataset. Recent advances in assembly methods will also facilitate this approach with non-transcriptome data. There has been growing interest in moving to a statistically grounded approach to sequence assembly[47,90,37]. These tools provide a distribution of gene sequences, rather than a single assembly, that captures uncertainty about the original biomolecule sequence[48,77]. These are not yet computationally viable for analyses of large real-world datasets, though they could soon be.

1.9 Next steps

As the focus on phylogenetic workflow development transitions from improving individual analysis components to better integrating those components, it will be necessary to prioritize

those developments that will have the most positive impact. This requires first an understanding to what extent violations of assumptions around order, dependence and relative entropy negatively impact analysis results. For those assumptions that do negatively impact analysis results, an assessment needs to be made of the tradeoffs between implementation costs, computational costs, and improved results.

A generative model and probabilistic framework provide the perspective needed to describe and evaluate different potential improvements. At one extreme is the current approach of passing a single hypothesis between steps in linear analyses, discarding all alternative hypotheses. This approach is simple to implement and minimizes computational demands, but relies on an assumption of low relative entropy that is routinely violated in ways that negatively impact analyses. At the other extreme is the simultaneous estimation of all unknowns, including gene sequences, gene trees, and species trees. This approach makes no assumptions about Markovian dependence, ordering of analysis steps, or low relative entropy, but is computationally prohibitive. In practice, the optimal approach to molecular phylogenetic analyses will be between these extremes. Sensitivity analyses and probability theory provide an informative guide for finding just where that optimal point will be, and for making well informed decisions about the relative benefits of allocating additional engineering and computational resources to different analysis steps.

2

An integrated generative model to assess low relative entropy assumptions within genome assembly in transmission inference

This work has not been published anywhere. I was responsible for summarizing previous phylogenetic and phylodynamic models, developing an algorithm and model for generating the epidemiological and molecular phylogenetic trees, and developing transmissim.

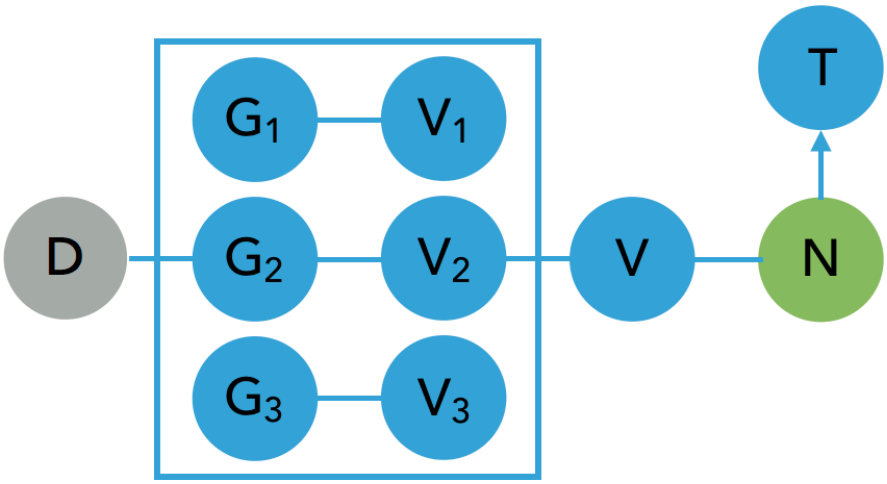


Figure 2.1: An overview of the components of the generative model that will be described in this chapter.

2.1 Background and motivation

Most current approaches to phylogenetic analysis from next-generation sequencing (NGS) data employ consensus sequences to summarize the variation observed within samples. This assumes that a consensus is an appropriate summary of biological and technical sequence variation within the sample. In the integrated perspective described in Chapter 1, this is an assumption of low relative entropy within genome assembly. When the diversity of genome sequences is very low, as for most somatic tissue samples, this assumption may hold. When the diversity of genome sequences is high, as when rapidly evolving viruses are sampled from a patient, this assumption is almost certainly violated. Attempts have been made to identify the composition of each individual viral genome within a host[99], known as phasing, in order to relax this assumption, but have largely been unsuccessful due to limitations on sequencing technology[101].

The transmission of virions between individuals results in a molecular signature that allows researchers to identify epidemiological links between infected individuals with phylogenetic

analyses[62]. It is therefore important to assess to what extent a violation of low relative entropy in genome assembly negatively impacts inference of the epidemiological phylogeny or transmission clusters on the phylogeny, defined loosely as a set of individuals who share a most recent common ancestor with some high level of confidence, typically bootstrap support values[43]. In order to do so, it is necessary to develop a detailed generative model that accounts for the multiple epidemiological, evolutionary and technical processes that generate the data. Such a generative model is necessary to develop communication strategies that relax the assumption as well. User-friendly simulation tools based on clear generative models can then provide a mathematical ground truth to validate models, assess tools, and perform sensitivity analyses. As established in Chapter 1, one of the deficiencies in existing simulation tools is an integrated way to simulate next generation sequencing reads or other observed data from multiple evolutionary processes with variable parameter values, although recently this has started to change[79].

This chapters aims to address this deficiency and answer questions about violations of low relative entropy in genome assembly of viral sequences from infected individuals through an integrated generative model and associated simulation tool. The integrated generative model is primarily based on existing models for transmission, evolutionary and technical processes, but adds a component to describe which virions are passed between individuals in a transmission event and how they evolve within a host. Transmissim is an open source Python package that simulates all of the processes described in the generative model. Transmissim has a modular structure, allowing the user to specify which processes to simulate, which parameters to vary, and which products to generate.

2.2 A high-level overview of the model

Figure 2.1 depicts the random variables generated by the model. Here we assume Markovian dependence in the generative model as well. Infection events between susceptible and infected individuals leads to a transmission network that describes the direction and times of infection between individuals (N, Figure 2.1). A surjective mapping from the transmission network along with diagnosis, sampling, and treatment times generates an epidemiological phylogenetic tree that describes transmission links and times between individuals, though not direction (T, Figure 2.1). Within-host viral evolutionary diversity and transmission dynamics between hosts leads to a molecular phylogenetic tree that describes the structure of populations of viruses (V, Figure 2.1). Because each population of viruses can be categorized by individual, V can be deterministically represented as a set of individual molecular phylogenies ($V_1, V_2, \ldots, V_m$, Figure 2.1). Processes of molecular evolution as described in Chapter 1 result in genomic sequences, again categorized by individual ($G_1, G_2, \ldots, G_m$, Figure 2.1). Finally, also as described in Chapter 1, sequencing produces raw reads (D, Figure 2.1).

Let x represent parameters in within-host evolution and transmission dynamics, y_i represent parameters in molecular evolution for individual i, and z represent parameters in sequencing technology. Under the assumption of Markovian dependence, we can as in 1.7 generate the observed sequence data and other random variables with a hierarchical model:

$$\begin{aligned} P(D, G, V|N) &= \prod_{i=1}^{n} P(D_i, G_i, V_i|N) \\ &= \prod_{i=1}^{n} P(D_i|G_i, V_i, N)P(G_i|V_i, N)P(V_i|N) \\ &= \prod_{i=1}^{n} P(D_i|G_i, \mathrm{z})P(\mathrm{z}|G_i)P(G_i|V_i, \mathrm{y}_i)P(\mathrm{y}_i|V_i)\ldots \int P(V_i, \mathrm{x}|N) \end{aligned}$$

In the following sections we describe in more detail the parameters and models for each module in the hierarchical model.

2.3 A single infection spreads on a contact network to generate a transmission network

In epidemiological models there is understood to be an interaction between the contact network and the transmission network[118,92]. Contact networks are networks of individuals that are connected to other individuals based on if those individuals have interacted. So each individual is a node in the graph, and edges exist between nodes when the individuals representing those node have interacted with each other in some way. The type of interaction is relevant to what the network looks like.

Contact networks are important in epidemiology because infectious diseases that are directly transmitted do so over these networks. An infected individual can only pass on the infection to other individuals they have come into contact with. A transmission network can thus be thought of as a sub-network of a contact network, where a connection exists between two individuals if a direct transmission link exists between them. Contact networks and transmission networks differ in that while contact networks represent contacts between individuals and possible transmission links, transmission networks represent actual transmission links. In this model we assume that the underlying contact network is a set of n fully connected individuals. The model for the transmission network comes from the R package outbreaker[52].

At time $t = 0$ there is a single infection in a population of n susceptible hosts. Let i represent the index of the infection case, starting at $i = 1$ and t_i represent the time at which case i is infected. The initial infection is thus case 1 and $t_1 = 0$. In this scenario, susceptible means not yet infected but able to become infected. No native host immunity exists in these scenarios.

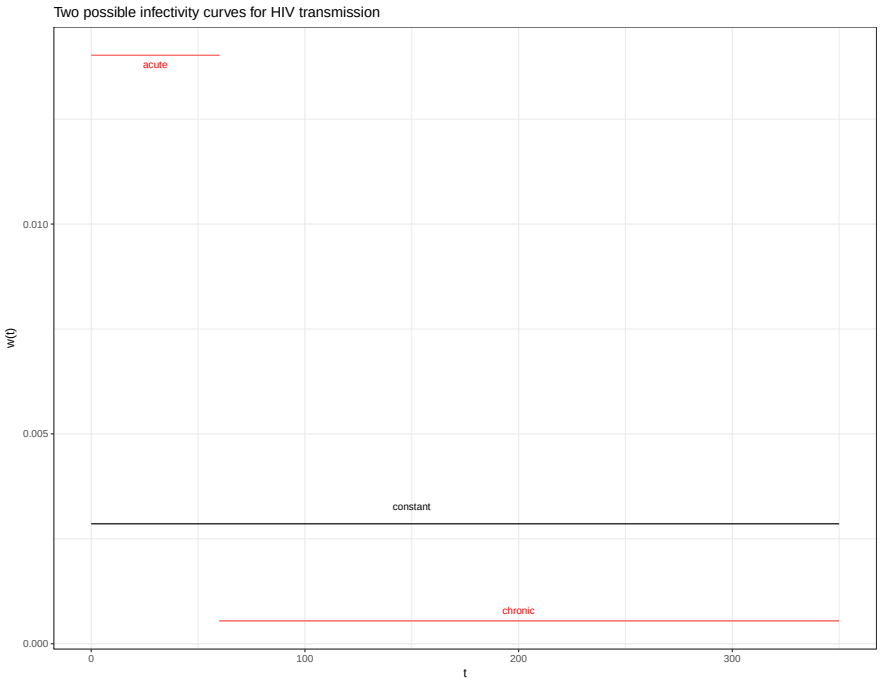


Figure 2.2: Two example infectivity curves for HIV over $[0, 350]$. The black is simply a uniform distribution, while the red is piecewise uniform, with an acute segment when infectivity is high lasting around 60 days, and a chronic segment when infectivity is low lasting the rest of the time. Infectivity ratios for the acute and chronic segments are taken from Volz et al. (2012)[118].

Additionally, the infectivity curve is presumed to be known as a probability mass (not density) function $w(t)$, with time in $[0, t_{max}]$. This is a representation of how relatively infectious an individual is at various time points. Two examples of HIV infectivity curves over $[0, 350]$, one a uniform distribution, and one a simplified representation of acute and chronic states[118], are given in Figure 2.2. The use of known infectivity curves is relatively common in the inference of disease outbreaks[52,38].

Given R_0, the basic reproduction number, the probability for a susceptible individual to become infected on day t, is thus:

$$p_t^{\text{inf}} = 1 - e^{-\sum_i R_0 w(t-t_i)/n}$$

Note that here the inf in p_t^{inf} means infection, not infinity. Given that there are S_t number of susceptible hosts at time t, with each time step the number of new cases is a binomial distribution with S_t draws and probability p_t^{inf}. The choice of source transmission individual from the pool of infected individuals is further a multinomial distribution with probabilities

$$\frac{w(t - t_i)}{\sum_i w(t - t_i)}$$

This models how much time passes between infections.

A few additional optional parameters are allowed to make the simulation scenario more complex. First, additional infections from outside sources can be imported at a constant rate, say I. These imported cases as well as the seed infection are assumed to descend from the same common ancestor $t_{ancestral}$ time ago. If $I > 0$, meaning we have imported cases, then we denote this common ancestor as case $i = 0$. We use a birth-death process from 0 to $t_{ancestral}$ to simulate t_j for imported or index case j. See 2.4.2 for more details on birth-death processes.

2.4 Transmission networks generate both epidemiological and molecular phylogenetic trees

Due to within-host evolution, transmission networks differ from the molecular phylogenetic tree where tips are viral variants from individual infection cases [101]. However, all transmission networks also have a surjective mapping to a phylogenetic tree where tips are infection cases. It is this phylogenetic tree that researchers try to infer by employing consensus sequences. Occasionally distinctions are made between the two phylogenetic trees by defining one as molecular and the other as epidemiological, but in general it is implicitly assumed that the molecular phylogenetic tree is a proxy for the epidemiological phylogenetic tree, which itself is a proxy for the transmission network. Here we explicitly define both the epidemiological

and molecular phylogenetic trees, and describe how both are generated by the transmission network.

2.4.1 Epidemiological phylogenetic trees

Transmission networks have a surjective mapping to epidemiological phylogenetic trees. This mapping is not one-to-one, as multiple transmission networks can be mapped to the same phylogenetic tree. Despite this, they are often used as a proxy for the transmission network, as public health officials are more interested in actionable epidemiological links between infected persons that can be identified from the network. Examples of actionable epidemiological links include high rates of growth for a certain risk group that may indicate an epidemic[16] or identifying high transmission rates tied to a single individual[68].

Epidemiological phylogenetic trees can be generated from the transmission network with a deterministic algorithm as follows. Suppose that there are m total infections, where $m < n$, the number of individuals in the contact network. As in 2.3, let i represent the index of the infection case and t_i represent the time at which case i is infected. Without loss of generality we assume here that $i \in \{0, 1, \ldots, m-1\}$. Furthermore let $t_{i,sample}$ represent the time at which case i is sampled, and let a_i represent the index of the source transmission individual for case i. Our model assumes that when case i is diagnosed is when they are sampled and undergo treatment, that both sampling and treatment occur with 100% probability after diagnosis, and further that treatment immediately implies that they are no longer infectious. This means that $t_{i,sample} > t_j$ for all infection cases j with $a_j = i$, and $a_i < i$, i.e. $a_i \in \{0, 1, \ldots, i-1\}$. Then:

1. For $i = 0$, create one branch with edge length $t_{0,sample}$ and taxon name 0.
2. For each $i > 0$, $i \leq m$, add two children to the leaf with taxon name a_i. Reassign one leaf with taxon name a_i and branch length $t_{a_i,sample} - t_i$ and one leaf with taxon name i and branch length $t_{i,sample} - t_i$.

3. For the new MRCA of a_i and i, adjust the branch length to $t_i - t_{a_i}$.

2.4.2 Molecular phylogenetic trees

Our model for molecular phylogenetic trees consists of two components. The first is a simple model for which virions are transmitted from the source individual to a new infection case. The second is a birth-death process for evolutionary diversity within an individual once an infection has been established.

Our model for which virions are transmitted involves just one additional parameter constant v, the number of virions transmitted. Let $K_i(t)$ be the set of virions of infection case i at time t, and let $k_i(t)$ be the size of $K_i(t)$. For each infection case i with source a_i and time of transmission t_i, the transmitted virions are drawn uniformly (without replacement) from $K_{a_i}(t_i)$ if $v < K_{a_i}(t_i)$, otherwise all virions in $K_{a_i}(t_i)$ are transmitted. Each combination of $v < K_i(t)$ virions from $K_i(t)$ thus has a probability of $\frac{1}{\binom{K_i(t)}{v}}$ of being transmitted.

Our model for evolutionary diversity within an individual is based on well-established algorithms for birth-death processes. Birth-death processes have been employed to model speciation/extinction[91] and gene duplication/loss[115,104,20] in phylogenetic trees with considerable success. It is thus a natural fit to model within-host evolution with a birth-death process as well.

Here we assume a constant linear birth-death process, although the assumption of constant birth and death is almost certainly violated[53]. This process takes as parameters a birth rate λ and death rate μ. Assume that all evolutionary processes within individuals are independent, and furthermore that rates of birth and death are constant in all individuals.

As above, let $k_i(t)$ be the size of the viral population of infection case i at time t. Following

the terminology in the previous section, for infection case i, $k_i(t)$ has a non-zero value only between $[t_i, t_{i,sample}]$. We can model the process as follows. Let the possible transitions and associated probabilities in an element of time dt be

$$\begin{aligned} k_i(t+dt) &= k_i(t)+1 && \lambda k_i(t)dt + o(dt) \\ k_i(t+dt) &= k_i(t) && 1 - (\lambda+\mu)k_i(t)dt + o(dt) \\ k_i(t+dt) &= k_i(t)-1 && \mu k_i(t)dt + o(dt). \end{aligned}$$

The rate at which an event E occurs, birth or death, is $\lambda + \mu$. The probability of an event at time t is hence $(\lambda+\mu)k(t)$. The probability of a birth event given that an event E occured is then $\frac{k_i(t)}{(\quad+\quad)k_i(t)} = \frac{\quad}{\quad+\quad}$. If there are $k(t)$ lineages at time t, then the waiting time to the next event E is drawn from an exponential distribution with parameter $k_i(t)(\lambda+\mu)$.

This allows for the following algorithm to simulate birth and death events. Let $T = t_{i,sample} - t_i$. For each individual i we can linearly transform $k_i(t)$ to $k'_i(t)$ such that $k'_i(0) = k_i(t_i)$ and $k'_i(T) = k_i(t_{i,sample})$. Let $t_{begin} = 0$. We generate $W = w$ from an exponential distribution with parameter $k'_i(t = t_{begin})(\lambda+\mu)$. If $w < T - t$ then we update t to $t = t + w$ and generate X from a uniform distribution on $[0,1]$. If $X < \frac{\quad}{\quad+\quad}$, then there is a birth event. Otherwise we have a death event. From the set of branches we uniformly select one branch with probability $\frac{1}{k'_i(t)}$ and update both the branches and the population size $k'_i(t+w)$ accordingly.

This process does not preclude the possibility that $k'_i(t)$ may go to 0 and leave the lineage extinct. This violates our requirement that $k_i(t)$ is non zero for all values in $[t_i, t_{sample}]$, as we assume that once an individual has been infected, they do not spontaneously recover. In such cases the process can be rerun.

2.5 Molecular evolution down a molecular phylogenetic tree generates genomic sequences

Several well-established models exist to describe sequence evolution along a phylogenetic tree. These models typically incorporate substitution, insertions, and deletions and site rate heterogeneity. The most generalized time-reversible substitution model is the General Time-Reversible model[127]. Here we go into just one substitution model, HKY[42], and also describe how site rate heterogeneity can be incorporated through the example of codon frequencies.

The HKY model allows for different rates of transitions and transversions as well as unequal nucleotide base frequencies. Parameters for the model are transition to transversion rate ratio and base frequencies. Codon frequencies specify for site-specific rate heterogeneity: each codon position has a different rate of substitution. This allows for a higher mutation rate at the third codon position than at the 1st and 2nd due to the 3rd codon position being more synonymous. The model is described as follows:

For a given codon position, let the probability that a given site of that position class is variable be f. We have each variable site evolving according to a Markov process with the following probability:

$$P_{ij}(dt) = P(x(t+dt) = j | x(t) = i) = \begin{cases} \alpha \pi_j dt & \text{(transition)} \\ \beta \pi_j dt & \text{(transversion)} \end{cases}$$

where π_j is the base frequency of nucleotide $j \in \{T, A, C, G\}$, α is the rate of transition, and β is the rate of transversion. (Glossary at the end of this subsection includes definition of symbols as well) The substitution probability matrix for a small interval of time dt can be written as

$$P(dt) = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} 1-(\quad_C+\quad_A+\quad_G)dt & \quad_C dt & \quad_A dt & \quad_G dt \\ \quad_T dt & 1-(\quad_T+\quad_A+\quad_G)dt & \quad_A dt & \quad_G dt \\ \quad_T dt & \quad_C dt & 1-(\quad_G+\quad_T+\quad_C)dt & \quad_G dt \\ \quad_T dt & \quad_C dt & \quad_A dt & (1-(\quad_A+\quad_T+\quad_C)dt \end{pmatrix} \end{array}.$$

This matrix can further be broken down into $P(dt) = I + \mathcal{A}dt$, where I is the identity matrix and $\mathcal{A}$ is the rate matrix. For an arbitrary time interval t, $P(t)$ satifies the Chapman-Kolmogorov equation:

$$P(t+dt) = P(t)P(dt) = P(t)(I+\mathcal{A}dt).$$

This allows us to write

$$\frac{dP(t)}{dt} = P(t)\mathcal{A}.$$

Since $P(0) = I$, this gives $P(t) = e^{t\mathcal{A}}$, and so we can compute $P_{ij}(t)$ for any given branch length t as follows. Decomposing $\mathcal{A}$ into

$$\mathcal{A} = \sum_{i=1}^{4} \lambda_i u_i v_i'$$

gives

$$e^{t\mathcal{A}} = \sum_{i=1}^{4} e^{t\ \ {}^i} u_i v_i',$$

where $|\lambda_i I - \mathcal{A}| = 0$, $\mathcal{A}u_i = \lambda_i u_i$, $\mathcal{A}'v_i = \lambda_i V_i$, $(u_i, v_j) = \delta_{ij}$ for $i,j = 1,2,3,4$. Solving

for this gives

$$\lambda_1 = 0 \qquad \lambda_2 = -\beta$$

$$\lambda_3 = -(\pi_Y\beta + \pi_R\alpha) \quad \lambda_4 = -(\pi_Y\alpha + \pi_R\beta)$$

$$v_1 = \begin{pmatrix} \pi_T \\ \pi_C \\ \pi_A \\ \pi_G \end{pmatrix} \qquad v_2 = \begin{pmatrix} c\pi_R\pi_T \\ \pi_R\pi_C \\ -\pi_Y\pi_A \\ -\pi_Y\pi_G \end{pmatrix}$$

$$v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix} \qquad v_4 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}$$

$$u_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \qquad u_2 = \begin{pmatrix} 1/\pi_Y \\ 1 \\ pi_Y \\ -1/\pi_R \end{pmatrix}$$

$$u_3 = \begin{pmatrix} 0 \\ 0 \\ \pi_G/\pi_R \\ -\pi_A/\pi_R \end{pmatrix} \qquad u_4 = \begin{pmatrix} \pi_C\pi_Y \\ -\pi_T\pi_Y \\ 0 \\ 0 \end{pmatrix}$$

where $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. All that remains then is to estimate α and β.

Let the numbers of transition-type differences $S(j_1, j_2)$ and transversion-type differences $V(j_1, j_2)$ between the j_1-th and j_2-th sequences be defined as follows:

$$S(j_1, j_2) = n_{..T_{j_1}..C_{j_2}} + n_{..C_{j_1}..T_{j_2}} + n_{..A_{j_1}..G_{j_2}..} + n_{..G_{j_1}..A_{j_2}..}$$

and

$$\begin{aligned} V(j_1, j_2) &= n_{..T_{j_1}..A_{j_2}} + n_{..A_{j_1}..T_{j_2}} + n_{..T_{j_1}..G_{j_2}..} + n_{..G_{j_1}..T_{j_2}..} \\ &= n_{..C_{j_1}..A_{j_2}} + n_{..A_{j_1}..C_{j_2}} + n_{..C_{j_1}..G_{j_2}..} + n_{..G_{j_1}..C_{j_2}..} \end{aligned}$$

where $n_{..T_{j_1}..C_{j-2}..}$ is the number of sites with T in the j_1-th sequence and C in the j_2-th sequence, and likewise for the other ns. Consider a pair of sequences separated t million years ago. Let $x(t)$ and $y(t)$ denote the states of each site of the 2 sequences. We see that for $i \neq j$

$$\begin{aligned} P(x(t) = i, y(t) = j) &= P(x(t) = i, y(t) = j | \text{variable site}) \cdot P(\text{variable site}) \\ &= f \sum_{l=T,C,A,G} \pi_l P_{li}(t) P_{lj}(t). \end{aligned}$$

Since this process is also reversible, we get

$$P(x(t) = i, y(t) = j) = f\pi_i \sum_l P_{il}(t) P_{lj}(t) = f\pi_i P_{ij}(2t),$$

hence the average numbers of transition and transversion-type differences becomes:

$$\overline{V(t)} = 2fr\pi_Y\pi_R[1 - e^{-2 \quad t}]$$

$$\begin{aligned}\overline{S(t)} = 2fr\{&(\pi_T\pi_C + \pi_A\pi_G) \\ &+ (\pi_T\pi_C\pi_R/\pi_Y + \pi_A\pi_G/\pi_R)e^{-2 \quad t} \\ &- (\pi_T\pi_C/\pi_Y)e^{-2t(\quad Y+ \quad R)} \\ &- (\pi_A\pi_G/\pi_R)e^{-2t(\quad R+ \quad Y)}\},\end{aligned}$$

where r is the number of sites in that class. (i.e. 1st, 2nd or 3rd codon position) Variance and covariance of the number of transition and transversion-type differences can be calculated as well; those details can be found in[42] and[81].

In terms of implementation, a sequence evolves down a branch using the transition probability matrices P_1, P_2 and P_3 that we calculate for the given branch length t. For each base i in the sequence, we draw from a uniform distribution on $[0, 1]$ in order to determine what i will change into based on the row of P corresponding to i. We then move on to the next branches, traversed in pre-order convention.

2.5.1 transmissim simulates networks, phylogenies, and sequences in the context of HIV transmission

The integrated generative model described in the previous sections is the foundation for transmissim, a simulation tool that can simulate networks, phylogenies, and sequences in the context of transmission of HIV and other infectious diseases. transmissim is split into three modules: networks, phylogenies and sequences. This allows for user flexibility in specifying which random variables to simulate, as well as what parameters to utilize. These specifications are inputted as a YAML file (params.yaml), although default settings exist as well. Usage is simple, consisting of running the com-

mand `PYTHON SIMULATE.PY -P PARAMS.YAML`.

transmissim is additionally built as a Python package with test-driven development and reproducibility in mind. transmissim has about 70% test coverage of written code to verify the simulations, and allows for a random seed to reproduce all simulation outputs. Below we briefly describe each of the three modules, and explain parameters that must be specified for each module.

2.5.2 Network

The transmission network in transmissim was simulated using an R package, outbreaker[52] while the birth-death process for imported cases was custom written. The user must specify the following parameters in the YAML file:

- $R_\circ$, the basic reproduction number;
- n, the number of susceptible hosts;
- t_{max}, the amount of time to run the simulation for;
- I, the rate of imported cases;
- $t_{ancestral}$, the amount of time to the most recent common ancestor of all imported cases and index case;
- w, the discrete infection curve.

2.5.3 Phylogeny

The epidemiological phylogeny is simulated using custom written code. The within-host molecular phylogeny is currently simulated with SimPhy[76], while the transmis-

sion dynamics are custom written as well as the subsequent overall molecular phylogeny. The user may choose to simulate only the epidemiological phylogeny, only the molecular phylogeny, or both. The user must specify the following parameters in the YAML file in addition to which phylogenies to simulate:

- λ, the viral birth rate;
- μ, the viral death rate.

Note that λ and μ are not relevant if only the epidemiological phylogeny is simulated. Additionally, although $t_{i,sample}$ is specified in the models for the phylogeny, at this time multiple sampling times have not been implemented in transmissim, and all sampling times are assumed to be at t_{max}.

2.5.4 Sequences

To simulate genomic sequences we use a Python package called pyvolve[106]. The package takes as input a phylogeny with branch lengths in mean number of substitutions per unit time rt. As the phylogenies generated are in units of time t, we must scale the tree by a substitution rate r. We then simulate sequencing reads with the package ART[49]. The user has the option of simulating only the genomic sequences, or both genomic sequences and sequencing reads. The user must specify the following parameters in the YAML file:

- r, the substitution rate;
- l, the read length;
- c, the per site coverage.

2.6 Use cases for transmissim

The present designed use case for transmissim is to compare between simulations with within-host evolutionary processes and simulations without. This is the reason for allowing the user to select between simulating both/either epidemiological or molecular phylogenies. However, transmissim has been written to be as flexible as possible. In particular, it has been structured so that minimal re-engineering will be required to simulate contact networks in addition to transmission networks. We discuss this possible extension in Future Directions.

The within-host evolution in transmissim can be further specified to be fast or slow through adjustments of λ and μ. This allows us to thoroughly test the assumption that a consensus sequence accurately summarizes variation observed in viral samples from individuals. In the next chapter, we show that this assumption can lead to erroneous inferences of the epidemiological phylogeny, and provide a Hidden Markov Model-based solution to relax this assumption.

3

Propagating multiple HIV genome hypotheses improves phylogenetic inference of transmission clusters

This work has not been published anywhere. It is currently in draft stages in preparation for submission to PLoS Computational Biology. I contributed ideas for the Hidden Markov Model based approach described in the chapter, its use to provide probabilities on identified transmission clusters, generation of simulated data with transmissim, and visualizations of results on real and simulated data.

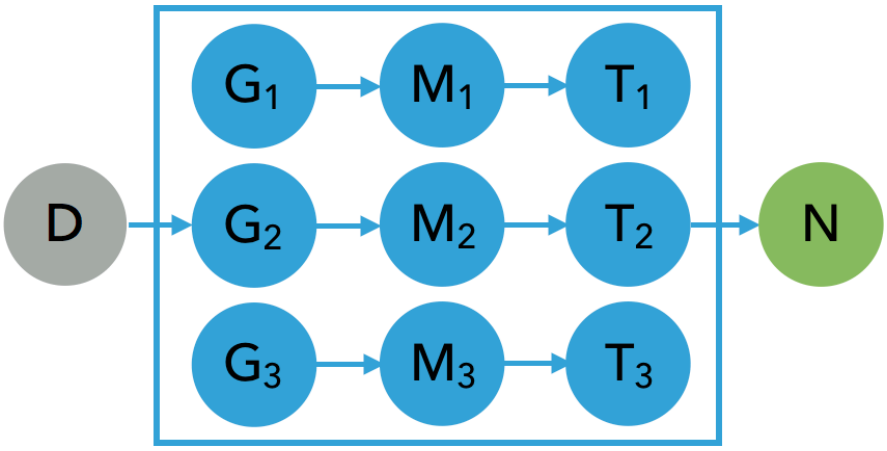


Figure 3.1: An overview of propagating multiple hypotheses within a phylogenetic workflow for the purposes of inferring transmission clusters, designed to relax the assumption of relative entropy.

3.1 Background and motivation

The HIV pandemic is a great health burden globally, nationally and locally. Ongoing transmission is largely by networks containing high-risk individuals. As described briefly in Chapter 2, phylogenetic analyses of HIV sequences across patients are used to infer features of the transmission network, such as to identify clusters. Improving phylogenetic inference of HIV transmission is therefore a high priority.

In the previous chapter we presented a generative model of the transmission network, epidemiological phylogeny, molecular phylogeny, genomic sequences, and next generation sequencing reads and associated simulation tool transmissim. This model captures relevant processes that generate observed virus sequence read data that serves as input for phylogenetic inference of transmission clusters. Most such phylogenetic analyses either assemble observed reads into a whole genome consensus summary or directly Sanger sequence only the *pol* region of the genome sequencing,

which also produces a consensus summary. Both approaches make the implicit assumption that the relative entropy between a consensus sequence summary and the overall within-host virus diversity is low. The phylogenetic inference power of the *pol* region has previously been contested[51,112], and recent evidence suggests that full-length genomes improve inference of HIV transmission clusters[125].

More recent approaches have made use of NGS methods to sequence whole HIV genomes and investigated their impact on phylogenetic inference of HIV transmission. Most of these approaches still build consensus sequences from the NGS data, neglecting the key challenge of deciding how best to summarize HIV sequence variation within patients.

In this chapter, we follow up on the questions posed in Chapter 2 and ask:

- How does within-host variation impact the inference of epidemiological phylogenies? Is the consensus summary approach sufficient for characterizing the phylogenetic tree space?
- Following the communication strategy of propagating multiple hypotheses laid out in Chapter 1, can we improve inference of HIV transmission clusters by propagating multiple HIV genome samples per individual?
- If we propagated multiple HIV genome samples per individual, will full-length genomes still improve inference of HIV transmission clusters compared to the *pol* region?

We show in this chapter that the consensus summary approach is insufficient in transmission inference problems. We present an alternative summary method, termed a profile-sampling approach, to accommodate within-host variance by relaxing this assumption. Our profile-sampling approach uses the probabilistic aligner HMMER[29,28]

to generate per-site nucleotide frequency distributions from NGS data, generate a population of sequences that sample from this distribution, generate multiple phylogenies from the population sequences, and finally identifies transmission clusters based on the frequency of their occurrence within the sampled phylogenies. This approach not only captures within-host sequence variation, but also has the benefit of quantifying the probability of a molecular transmission cluster.

On a dataset of individuals newly diagnosed with HIV in Rhode Island, USA in 2013, we compared (i) a phylogeny with one full genome consensus sequence per patient; and (ii) a set of 100 sampled phylogenies of NGS-derived full genome profile-sampled sequences per patient. We found that the profile-sampling approach was able to recover more transmission clusters for both the *pol* and whole genome sequence data, and that using whole genome sequence data instead of *pol* sequence data improves support for inferred transmission clusters. Enhanced transmission cluster detection has the potential to improve HIV transmission prevention.

3.2 The profile-sampling approach summarizes within-host genomic variation and estimates transmission cluster probabilities

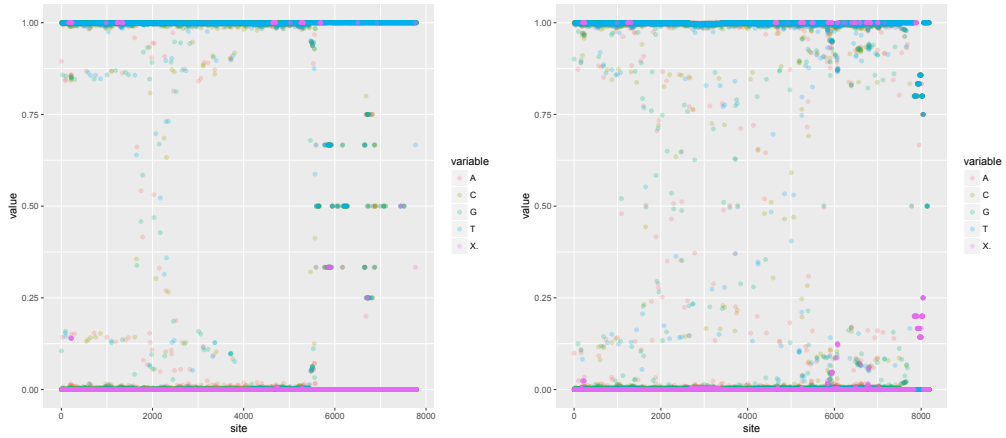
To assess our questions on the impact of within-host variation we developed a genome summary approach based on profile Hidden Markov Models (pHMM)[27] that allows us to summarize within-host variation for each individual that is diagnosed and sampled through base pair match, insertion, and deletion states at each site in the HIV genome. Using a multiple sequence alignment of HIV reference genomes from the Los Alamos HIV Sequence Database (`HTTP://WWW.HIV.LANL.GOV`), we first build

a reference pHMM. We then iteratively align whole-genome NGS reads to our reference pHMM. With this new read alignment, we build a second individual-specific pHMM from the first alignment, then perform a second and more sensitive realignment of the NGS reads to the individual-specific pHMM. This second re-alignment finally is built into a pHMM as well that summarizes the within-host variation in terms of site base pair, insertion and deletion frequencies.

From this pHMM, we can visualize the amount of polymorphism within an individual (Figure 3.2), sample genomic sequences in proportion to the within-host variation, or build a consensus summary genome. The ability to visualize the amount of polymorphism is useful because it allows us to assess first whether genomic variation exists at significantly high enough frequencies such that neither consensus genomes nor any other point estimator are sufficient summaries.

Figure 3.2 shows substantial variation exists in the HIV profiles of two individuals from the cohort of individuals diagnosed with HIV in RI 2013. For MC13, we found 2,308 sites out of 7,793 with 2 or more base pairs present, and 170 sites with 2 base pairs present in frequencies between 0.1 and 0.9. For MC14, we found 4,300 sites out of 8,188 with 2 or more base pairs present, and 251 sites with 2 base pairs present in frequencies between 0.1 and 0.9. Across all individuals in our dataset, we find that 7.86% of the sites in the HIV genome are polymorphic. This confirms that consensus genomes do not fully summarize the within-host variation.

The ability to sample genomic sequences further allows us to estimate the probability of a molecular transmission cluster given the within-host variation. To do so, we first infer a multiple sequence alignments and phylogenetic tree for each sampled ge-



(a) MC13 **(b)** MC14

Figure 3.2: Normalized site frequencies of A,C,T,G and missing data (X) from individuals (a) MC13 and (b) MC14.

nomic sequence. Then, we compute the bootstrap support[32] for each sampled tree and use a 99% bootstrap cutoff to identify transmission clusters. The bootstrap support value is the most common approach used to identify transmission clusters[43], hence our decision to apply it to our workflow. The number of times a particular transmission cluster is identified in the set of phylogenetic tree samples divided by the number of samples is the estimated probability of that transmission cluster given the within-host variation and the bootstrap cutoff. For example, if (A,B,C) are identified as a transmission cluster using a 99% bootstrap cutoff in 100 out of 1000 samples, then we would give it a value of 10. If (D,E,F) are identified as a transmission cluster in 1000 out of 1000 samples using a 99% bootstrap cutoff, then we would give it a value of 100.

We call this the *profile-sampling* approach, and will refer to it as such through the rest of the text. Additionally we will refer to the typical phylogenetic workflow that

uses a consensus genome summary, whether for *pol* or whole genome, as the *consensus* approach.

For the profile-sampling approach we use the well-established pHMM alignment tool `HMMER` 3.1b2[26] to build, align and sample from the pHMMs. For both approaches we estimated multiple alignments across individuals using `MAFFT` 7.305b[55], and maximum-likelihood phylogenies with bootstrap replicates using `RAxML` 8.0[108].

To provide some intuition into the profile-sampling approach, we give an example. Suppose there is a transmission chain consisting of A $\rightarrow$ B at time t_1 and later on, A $\rightarrow$ C at time t_2. Under our generative model from Chapter 2, A transmits v virions to B and v virions to C. Each population of viruses then evolves independently within the hosts. When we sample genomic sequences from these three populations, each subsequent tree estimate should either (1) have A and B in a (pair) clade sharing a most recent common ancestor around t_1, while C is sister to (A,B), or (2) have A and C in a (pair) clade sharing a most recent common ancestor around t_2, while B is sister to (A,C). Both of these trees represent specific paths sampled from the molecular phylogeny.

3.3 Phylogenetic tree estimates are sensitive to within-host variation

Given that there is significant within-host variation in our dataset, we investigated its impact on downstream topological variation in the phylogenetic tree (Figure 3.3). To do this, we took 200 genome, multiple sequence alignments, and phylogenetic tree samples from the profile-sampling approach as a representative sample of the topological variation given within-host variation. We additionally computed phylo-

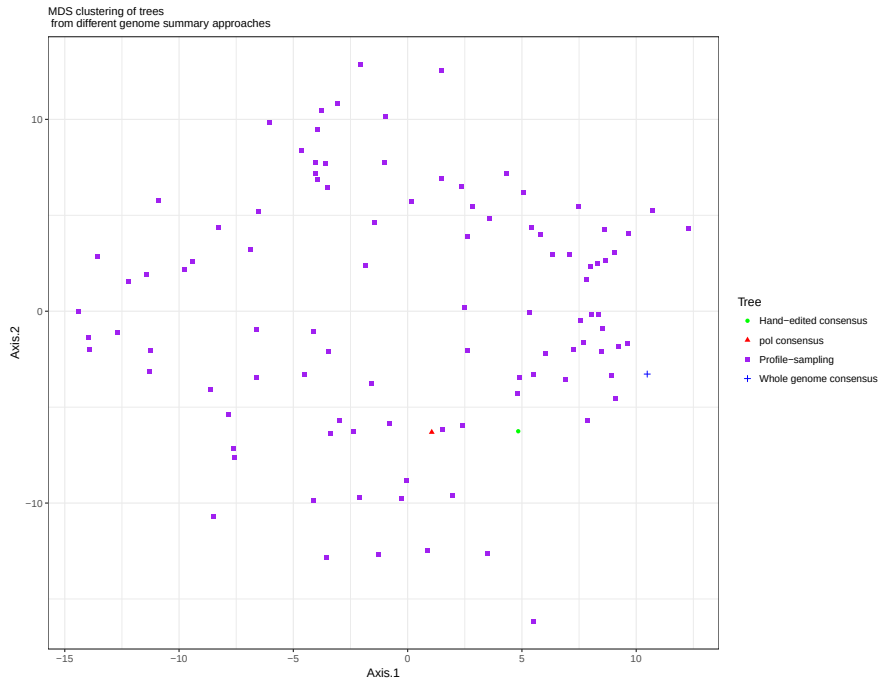


Figure 3.3: Multi-dimensional scaling of pairwise Robinson-Foulds distance among maximum-likelihood phylogenies from different consensus genome summary approaches. None of the estimated trees from genome point estimates appear to be central in the cluster.

genetic tree estimates on our dataset from the three different genome consensus estimates commonly used: *pol* Sanger sequencing, whole genome consensus sequences (emitted from the individual profile HMMs we built), and hand-edited consensus sequences based on the whole genome consensus sequences. We then computed the pairwise Robinson-Foulds (RF) distance[95] between all of the inferred topologies and performed classic multidimensional scaling on the distance matrix in order to visualize the tree space in two dimensions. The results show that trees inferred from the profile-sampled sequences have a wide spread and do not cluster around any of the trees inferred from a genome point estimate. This suggests that the phylogenetic tree estimates are sensitive to the within-host genomic variation.

It should be noted that the size of the space of phylogenetic trees also scales on a factorial order? with the number of tips, so that the prior probability of any one individual tree will be vanishingly small. Accordingly, it is more appropriate to characterize a single tree estimate by credibility or confidence limits[121]. The probability distribution of the distance from a given tree drawn from the tree distribution to the estimating tree gives an idea of how tight the limits are, as well as whether there are multiple clusters of quality trees[83]. A peak in the probability distribution is indicative of a cluster of quality trees, while the more to the left the peak is, the tighter the limit.

Here we have chosen to use RF distance to characterize only the topological distribution (Figure 3.4). We approximate the distribution of the distance from the estimating trees for the *pol* consensus approach, the whole genome consensus approach, and the hand-edited whole genome consensus approach by our 200 profile-sampled trees, using a subset of the same distance matrix from above.

We found that there appeared to be at least two distinct clusters of trees based on the number of modes. The *pol* and hand-edited consensus approach trees fell within the same cluster of trees and had equally tight limits within that cluster, while the HMM consensus approach tree fell within the other cluster of trees. Some of the additional peaks from the HMM consensus approach are most likely artifacts of the fact that Robinson-Foulds distances are only in even integers combined with the kernel bandwidth. More robust tree distance measures such as the geodesic distance[86] would provide a more interpretable overview of the phylogenetic tree space, but this result still shows that distinct clusters of phylogenetic trees exist given the within-host variation, thus no single genome estimator is adequate.

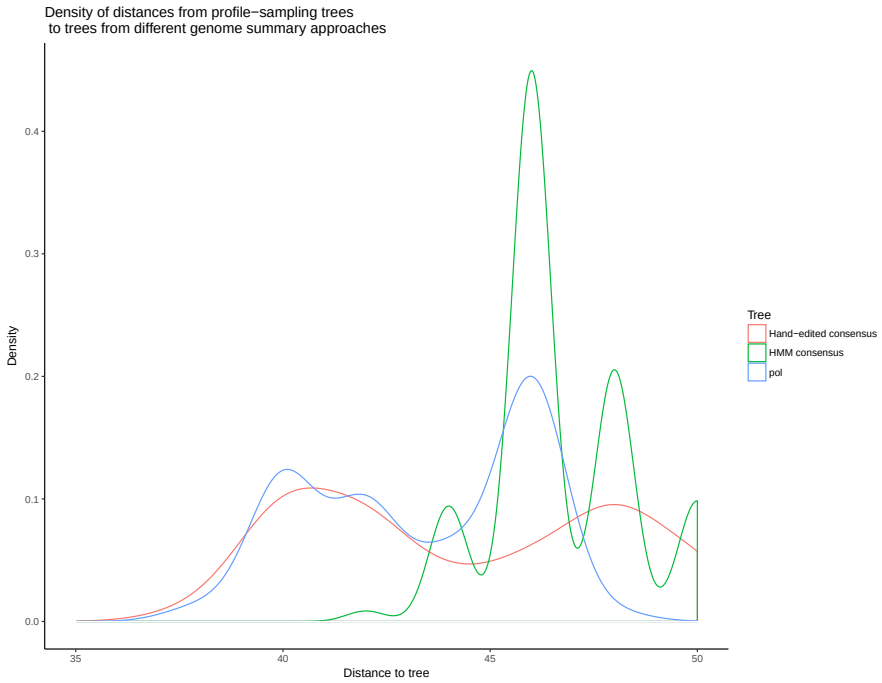


Figure 3.4: Density of Robinson-Foulds (RF) distance of profile-sampled trees to phylogenetic trees inferred from three different consensus genome summary approaches. There appears to be at least two modes in the RF distance to the trees inferred from the pol consensus and hand-edited consensus.

Our example in Section 3.2 provides some insight into why multimodality in tree topologies may occur. One peak could represent the cluster of tree topologies with ((A,B),C), whereas the other peak could represent the cluster of tree topologies with (B,(A,C)). This is of course a large oversimplification, and also not the only reason multimodality could occur.

3.4 The profile-sampling approach provides more information on inferred HIV transmission clusters

Having established that phylogenetic trees are topologically sensitive to within-host variation and likely form multiple clusters, we apply our profile-sampling approach to our Rhode Island dataset in order to estimate transmission cluster probabilities. We compared the profile-sampling approach and the consensus approach on both the whole genome NGS reads dataset, and a *pol* only dataset that consisted of both the pre-existing Sanger sequences (see Data) as well as a *pol* specific subset of our NGS reads.

3.4.1 The profile-sampling approach provides more resolution on identified transmission clusters

Figure 3.5 shows in red the transmission clusters identified with the consensus approach, and in blue the transmission clusters identified with the profile-sampling approach, as well as the associated probabilities. While most of the clusters identified were the same and with high probability between the consensus approach and the profile-sampling approach, an interesting result is the clade of five individuals C=(MC47, MC56, MC41, MC53, MC37). While the consensus approach identifies the clade as a transmission cluster, the profile-sampling approach assigns the clade only a 67% probability as a transmission cluster, and assigns 100% probability to two sub-clusters instead. This suggests that the profile-sampling approach, by estimating probabilities for clusters given within-host variation, indeed provides more informa-

tion on transmission clusters. Once within-host variation is incorporated, transmission cluster probabilities change.

The fact that most clusters identified by the consensus approach had 100% probability with the profile-sampling approach is perhaps not surprising given our dataset. Since only individuals who had been diagnosed with HIV in 2013 and had been sampled were present in this dataset, we almost certainly had several missing links between individuals. These missing links represent individuals who had been diagnosed with HIV before 2013, and their absence on the inferred phylogeny would create unresolved clades deeper in the tree. Thus only more recent transmission events between small numbers of individuals would be clustered together by bootstrap support.

3.4.2 The profile-sampling approach places higher probabilities on and identifies more phylogenetically inferred transmission clusters from whole genome sequences compared to *pol*

We additionally compared the profile-sampling approach to the consensus approach on just the *pol* region (Figure 3.6; blue indicates profile-sampling approach and red indicates consensus approach). Our findings reinforce the hypothesis that whole genome sequences provide additional information compared to just the *pol* region. The whole genome consensus approach identifies 3 additional transmission clusters compared to the *pol* consensus approach, while the whole genome profile-sampling approach identifies 2 additional transmission clusters compared to the *pol* profile-sampling approach. Probabilities for all transmission clusters identified by *pol* approaches increased to 100 on the whole genome.

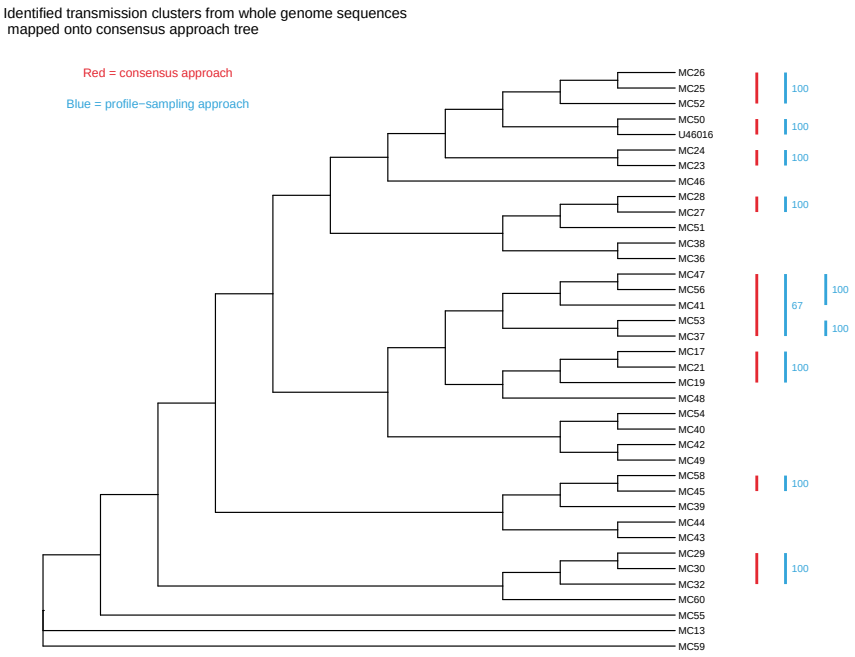


Figure 3.5: On the whole genome sequences, the profile-sampling approach recovers the same transmission clusters as the consensus approach with high probabilities except for the 5-individual cluster of (MC47,MC56, MC41, MC53, MC37). That cluster was seen only 67 times in the 100 tree samples. Additionally,the subsets of that cluster (MC47, MC56, MC41) and (MC53, MC37) were seen 100 times out of 100 tree samples, suggesting higher probabilities for the cluster subsets than for the larger cluster.

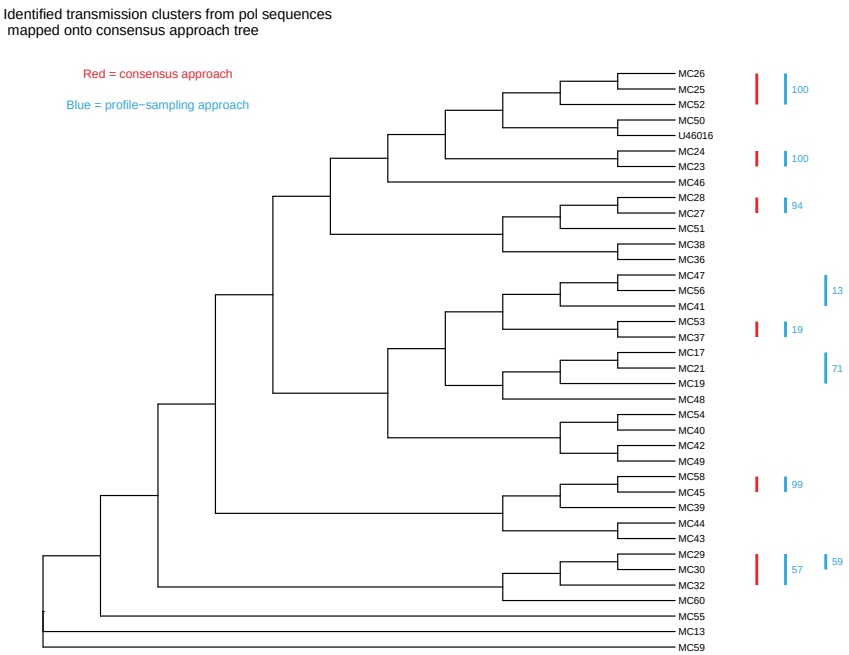


Figure 3.6: On the pol sequences, the profile-sampling approach recovers additional clusters compared to the consensus approach, including (MC17, MC21, MC19) and (MC47, MC56, MC41), clusters found when using the whole genome. Probabilities increase for all clusters found when going from pol to whole genome,suggesting that whole genome sequences provide higher resolution of transmission clusters than pol.

The results here also suggest that the profile-sampling approach provides more information on transmission clusters than the consensus approach. The profile-sampling approach found support for two additional clusters that the consensus approach did not, one of which was a high probability cluster. These two additional clusters were identified in the whole genome approaches as well with high probabilities, suggesting that they are actual transmission clusters. Combined with our comparisons on the whole genome, it seems promising that the profile-sampling approach will improve transmission cluster inference from phylogenies.

3.5 Discussion

Phylogenetic trees will continue to be used as proxies for transmission networks for quite some time. While previous studies have considered within-host evolution and transmission dynamics[119] in the transmission phylogeny, typically through coalescent based models[107], to date none have taken into account within-host viral sequence diversity as well. Here we show that the within-host viral sequence diversity impacts downstream phylogenetic inference through the existence of multiple high-probability tree clusters and provide an approach that takes advantage of the sequence diversity to estimate transmission cluster probabilities.

Our profile-sampling approach uses only well-established and optimized tools, making user implementation as easy as possible. Although we do not actually pass on a distribution for inferring phylogenetic trees, our approach is easily adaptable for any assembler that does output a distribution of results, as we simply need to sample from that distribution. A method that can infer phylogenetic trees directly from a profile HMM or other read distribution may be even more powerful as it will be able to propagate all existing variation within read populations rather than approximating such variation.

One weakness of our approach is that it does not provide any linkage information or take into account codon awareness on the sequences, thus it cannot phase haplotypes or detect recombination events that may provide information on transmission links or patterns. However, for the purposes of inferring the transmission tree, we believe our approach may be sufficient, as most standard maximum likelihood tree infer-

ence methods assume site independence and thus compute site-specific likelihoods[31]. As our approach does provide information on per-site base pair frequencies, standard tree inference methods should infer the same transmission topologies, although intra-patient branch lengths may be longer. Further approaches that can effectively incorporate linkage information into phylogenetic inference will produce trees with more accurate branch lengths and the ability to infer recombination or coinfection events.

3.6 Conclusion

In this study, we present a novel approach to HIV transmission network inference that is able to summarize genome variation present within viral reads sequenced from individuals living with HIV. We show that there is significant within-host variation in the HIV genome and that this variation leads to multiple clusters of tree topologies. Within-host variation thus impacts the inference of epidemiological phylogenies, and the consensus summary approach is not sufficient for accurately characterizing the phylogenies.

Our profile-sampling approach can characterize the tree space by propagating multiple HIV genome samples per individual which approximates within-host variation. By characterizing the tree space better we can also estimate probabilities on transmission clusters, and in some cases identify additional clusters or sub-clusters. These probabilities help determine our confidence in transmission clusters and suggest features that may be of epidemiological interest.

The proposed profile-sampling approach, which preserves the within-host variation available with longer and deeper whole-genome NGS data, better summarizes vari-

ation in topology, provides more information and a support value for inferred transmission clusters, and reinforces the growing body of evidence that whole genome sequences improves transmission cluster inference compared to just the *pol* sequences. Enhanced inference of HIV phylogenies has the potential to improve HIV transmission inference, and ultimately prevention.

3.7 Data

Our dataset consisted of patients newly diagnosed with HIV during 2013 and treated at The Miriam Hospital Immunology Center in Providence, Rhode Island, USA. Inclusion criteria were: (i) HIV infected adults $\geq$18 years; (ii) diagnosed with HIV during 2013; (iii) available *pol* sequence from routinely performed drug resistance testing. From these patients we sequenced whole genome Illumina MiSeq reads. Patient identifiers have been entirely removed.

4

Revising transcriptome assemblies with phylogenetic information

This chapter reflects the work available as a preprint on bioRxiv, doi:https://doi.org/10.1101/202416. My role in the paper was assessing the extent of transcriptome assembly errors, writing the implementation of *treeinform* (the module that revises transcriptome assemblies with phylogenetic information, determining a default threshold for the *treeinform* implementation, and validating the effectiveness of our implementation.

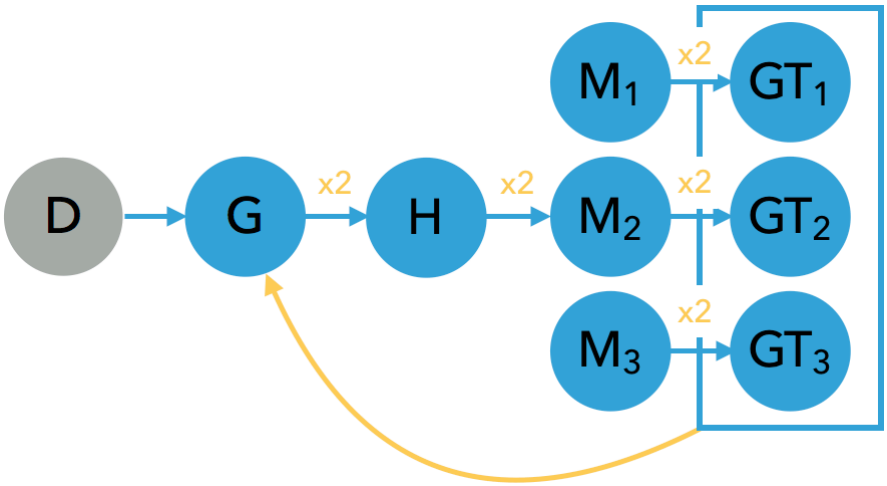


Figure 4.1: An overview of the iterative phylogenetic workflow built on the assumptions of Markovian dependence that will be described in this chapter.

4.1 Background and motivation

The development of RNA-seq has made it possible to generate large amounts of transcriptome data for a broad diversity of species, providing novel insights into the history of organisms and the study of gene evolution and function[120]. The central goal of transcriptome assembly is to use the raw reads to estimate the correct sequence of the original transcripts and assign these to their respective genes. One of the biggest challenges in transcriptome assembly is to identify whether sequence variance is due to technical factors (such as errors introduced in library prep and sequencing), splicing differences, different alleles of the same locus, or evolutionary divergence between closely related gene copies following duplication. This makes it difficult to determine if slightly different transcripts are variants of the same gene or are derived from different closely related genes. When different transcripts of the same gene are incorrectly assigned to multiple genes, downstream analyses can be compromised.

Whether two transcripts are splice variants from the same gene or are from two different genes is not the only distinction. Two transcripts can be from two different genes, but be within the same gene family (often referred to as paralogs), or be from different gene families as well. The downstream step of homology identification categorizes transcripts as being from the same gene family or from different gene families. After homology identification, multiple sequence alignments are inferred for each set of transcripts within a gene family, then phylogenetic gene trees are inferred for each multiple sequence alignment. Here we assume that homology identification is error-free, and consider only transcript assignment errors arising from transcriptome assembly.

When transcriptomes belonging to the same gene family are compared across species in a phylogenetic framework, transcripts of the same gene that have been incorrectly assigned to multiple genes will appear in gene trees as tips with improbably short branch lengths (Figure 4.2). Here we look at a test dataset of Siphonophora to explore a few questions:

- How prevalent is transcript misassignment?
- Following the ideas laid out in Chapter 1 of Markovian dependence, can we use phylogenetic information to identify misassigned transcripts and reassign them to the same gene?

After determining that transcript misassignment is common in a test dataset (see Methods and Data for more information), we present a method, *treeinform*, that uses this phylogenetic information to identify misassigned transcripts and reassign them to the same gene. Our approach uses a subtree length threshold and revises the mapping

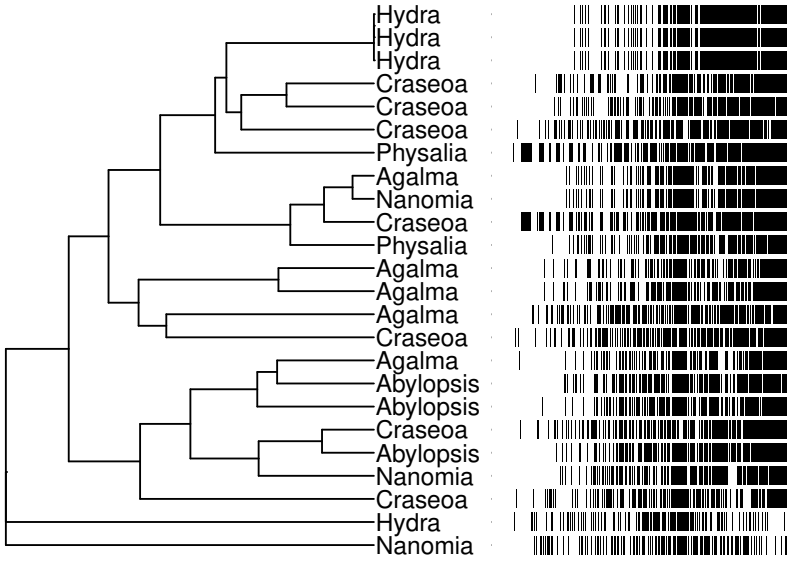


Figure 4.2: On the left is a gene tree from the test dataset before running `treeinform`. Each tip is an exemplar transcript that was initially assigned to a different gene in the assembly step but determined to be in the same gene family in the homology identification step. On the right is the multiple sequence alignment, with sequence sites ordered from highest to lowest identity to the inferred ancestral site for clarity on sequence diversity. Black indicates a difference from the ancestral sequence. The three *Hydra* transcripts in the grey box at the top of the gene tree were assigned to different genes by Trinity[39] despite having nearly indistinguishable sequences. This figure is from before *treeinform* was run. After *treeinform*, all transcripts from these three genes would be reassigned to a single gene.

of transcripts to genes accordingly. *treeinform* essentially borrows information across species to refine the assignment of transcripts to genes within species. It capitalizes on the implicit assumption of conditional independence within phylogenetic workflows, as described in Chapter 1. (Figure 4.1)

4.2 Assessing the extent of transcript assignment errors

We first examined the prevalence of transcript misassignment. For each node in each of the 5304 gene phylogenies, we calculated the length of the corresponding subtree. This is the sum of the length of all branches in the subtree defined by the node. An excess of very short subtrees would be a strong indication of assigning different transcripts of the same gene, which have very similar sequences and therefore short branches connecting them in phylogenetic trees, to different genes. This is the pattern we found (Figure 4.3).

Two issues could create a misleading impression in the histogram of subtree lengths for internal nodes (Figure 4.3). First, it considers all subtrees, including those defined by both speciation and duplication nodes. Misassigning transcripts from the same gene to multiple genes will artificially inflate only the number of duplication nodes, since variation across transcripts within a gene are essentially misassigned to gene duplication events. Examining just the duplication events in the gene trees therefore provides a more direct perspective on the problem we investigate here. Second, subtree lengths are in units of expected numbers of substitution, which depend on both rates of molecular evolution and time. Because the rates of evolution can vary within and between gene phylogenies, variation in rates could confound the interpretation of

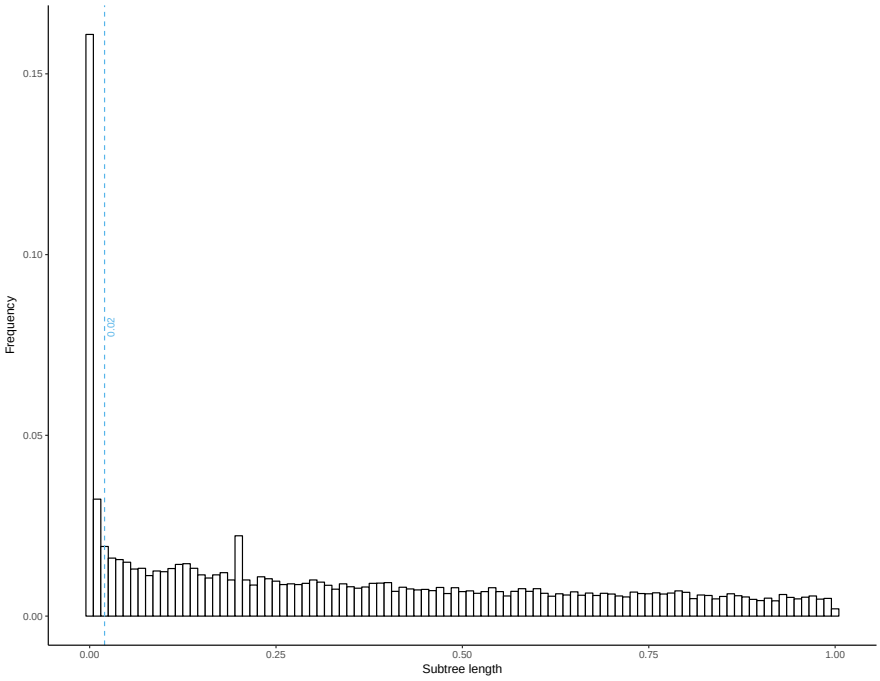


Figure 4.3: Histogram of subtree lengths for internal nodes in each Siphonophora subset gene tree from Agalma1.0 containing tip descendants from the same species. Subtree lengths greater than 1 were filtered out for clarity.

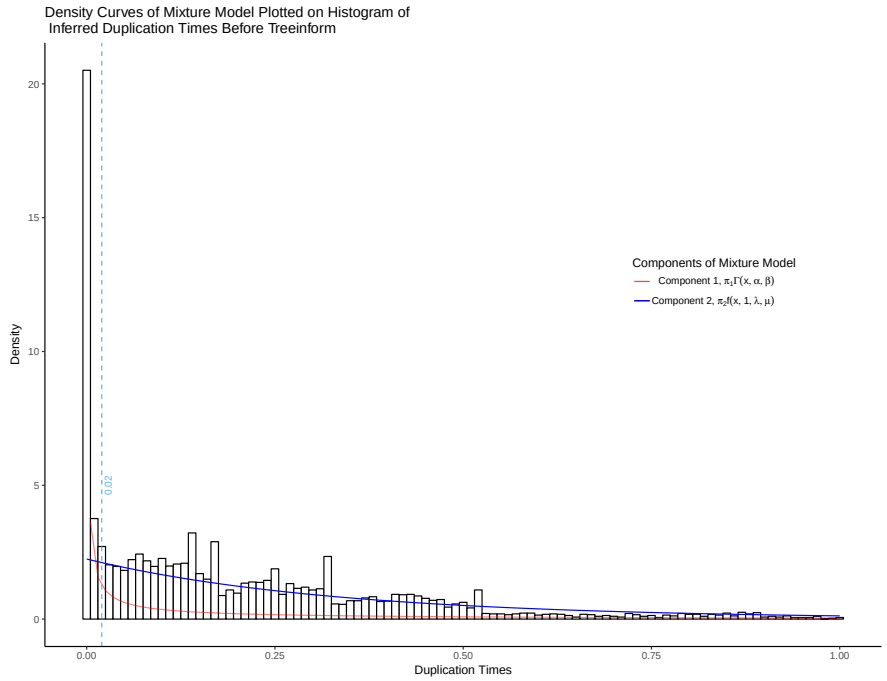


Figure 4.4: Histogram of the inferred duplication times. We first ran phyldog[6] on the Siphonophora subset multiple sequence alignments from Agalma1.0[23] and a user-inputted species tree. This provided gene trees with internal nodes annotated as duplication or speciation events. We then fitted chronograms onto these gene trees with our user-inputted species tree. In the overlaid mixture model, the intersection point between the two distribution curves was 0.009.

gene tree sublength.

We therefore performed a calibrated analysis and focused only on duplication nodes. We first created a time calibrated species tree, with all tips with age 0 and the root node with age 1. We then transformed the branch lengths of the gene trees so that each speciation node in each gene tree had the same age as the corresponding node in the species tree. A histogram of the calibrated duplication times (Figure 4.4) indicates there is a large excess of recent duplications. This provides additional evidence for the frequent misassignment of transcripts from the same gene to artefactual recent gene duplicates.

4.3 Implementation

Given that transcript misassignment is common in our test dataset, and given that it can be identified in phylogenetic trees, we implemented a module within the phylogenomic workflow Agalma1.0[23], titled *treeinform*.

The algorithm is as follows: for each internal node on a tree, subtree length is calculated as the sum of all branch lengths in the subtree defined by the node. *treeinform* then identifies subtrees with total length below a given threshold. If multiple tips (i.e., genes) belonging to a single species exist in an identified subtree, all transcripts for these multiple genes are flagged for reassignment to a single gene. *treeinform* outputs a list of transcripts for reassignment, which become input for a second run of Agalma1.0 starting from RSEMEval[64] to complete a phylogenomic analysis.

Input consists of a set of gene trees generated by the pipeline `GENETREE` in Agalma1.0 and a user-designated threshold. The default value for the threshold is 0.02, selected based on analyses of the test dataset. We describe the mixture model used to analyze the test dataset below. Although we did not test the threshold on other datasets, we suspect that threshold selection is specific to the dataset that Agalma1.0 will be applied to. Thus, in the Github repository we have also provided the user source code to run their own mixture model analysis.

4.4 Selecting a threshold for transcript reassignment

A visual inspection of the histogram of subtree lengths (Figure 4.3) suggested that 0.02 was an appropriate threshold for this particular dataset, as the frequency of sub-

tree length for internal nodes was high below such threshold but leveled out above it. In addition, 19.12% of internal nodes containing tip descendants from the same species had a subtree length less than 0.01, with an additional 2.17% having a subtree length between 0.01 and 0.02. It is unlikely that all of these clades are gene duplication events.

These observations suggest that two different processes are operating simultaneously to generate the observed subtree lengths, one for the misassigned transcripts and one for the correctly assigned transcripts. To model this pattern, we applied a mixture model to the inferred duplication times from the gene trees. Duplication times are related to subtree lengths because when transcripts from the same gene are misassigned to different genes, gene tree/species tree reconciliation programs compensate by inferring additional duplication events. Because such transcripts have almost identical sequences, the inferred duplication events will be extremely shallow, with correspondingly recent duplication times (Figure 4.5). Conversely, correctly assigned transcripts from different genes will have less similar sequences and thus older duplication times.

In our mixture model, one component modelled duplication events and associated times arising from transcripts assigned to different genes that belong to the same gene (i.e., misassigned transcripts), and the other component modelled duplication events and associated times arising from transcripts assigned to different genes that in fact do belong to different genes (i.e., correctly assigned transcripts) (Figure 4.4).

We expected the implied duplication events of transcripts of the same gene that were misassigned to different genes to have very short implied duplication times approaching 0, and thus chose to model that component (Component 1) as a gamma

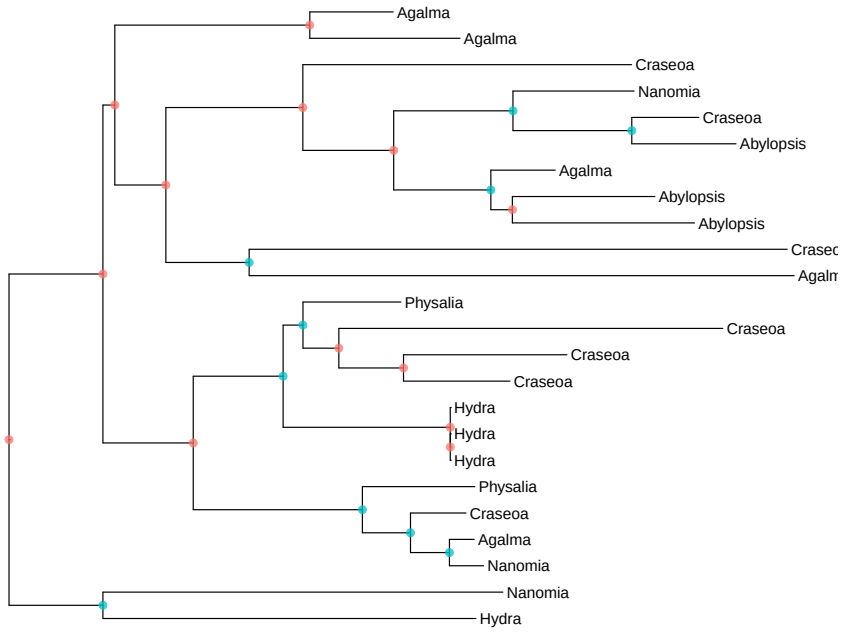


Figure 4.5: The same tree as from Figure 4.2, with branch lengths as duplication times rather than expected number of substitutions and internal nodes annotated with duplication (blue) or speciation (red) events as determined by phyldog[6]. The three Hydra transcripts assigned to different genes by Trinity[39] result in two duplication events inferred by phyldog with very small duplication times. Sets of similar transcripts are prevalent across the set of gene trees, contributing to the peak of very small duplication times seen in Figure 4.4 and small subtree branch lengths seen in Figure 4.3.

distribution with parameters shape$=\ \alpha$ and rate$=\ \beta$. To model duplication events and associated times arising from the correctly assigned transcripts (Component 2), we used a birth-death process[36], which is well studied and often applied to gene analyses of duplication and loss. The probability distribution function in the model we used has parameters birth rate λ, death rate μ, and tree time of origin t_{or}. This pdf is what we will call the *theoretical* curve, and is the pdf we compare our results against to assess the effectiveness of *treeinform*.

Because we fitted a chronogram with time of origin 1 onto the gene trees $G = G_1, G_2, \ldots, G_K$, we made the assumption that all gene tree times of origin are $t_{or} = 1$. Some gene trees have duplication events predating the first speciation event, thus when we fitted chronograms onto those gene trees they had times of origin greater than 1. We chose to filter these gene trees out of the mixture model and subsequent analyses.

If $x_{i,k}$ represents duplication times i from gene tree G_k, π_1 and π_2 denote the probability that a duplication time belongs to the 1st and 2nd component respectively, $\Gamma(x_{i,k}|\alpha, \beta)$ is the probability density function for the gamma distribution, and $f(x_{i,k}|t_{or,k} = 1, \lambda, \mu)$, then we get the expression

$$P(x_{i,k}) = \pi_1\Gamma(x_{i,k}|\alpha, \beta) + \pi_2 f(x_{i,k}|t_{or,k} = t, \lambda, \mu)$$

We used Just Another Gibbs Sampler (JAGS)[88] to perform Bayesian Gibbs Sampling in order to infer the parameters α, β, λ, and μ as well as the mixing proportions π_1 and π_2. This gave us the parameter estimates in (Table 4.1.

Table 4.1: Summary of parameter estimates from JAGS.

	Lower95	Mean	Upper95	MCerr
α	0.2271265	0.2424239	0.2580250	0.0007320
β	1.1420843	1.7861715	2.5024476	0.0184331
μ	0.0000006	0.0789256	0.2355944	0.0010044
λ	2.5932256	2.9634860	3.3280288	0.0018998
π_1	0.2373131	0.2840289	0.3337098	0.0003343
π_2	0.6662902	0.7159711	0.7626869	0.0003343

4.5 Validating the effectiveness of treeinform

We tested *treeinform* on a dataset of seven species of Siphonophores. We took three different approaches to assess the efficacy of *treeinform*. First, we spot checked the results to confirm that they were biologically and technically sensible. This provided detailed confirmation on a small fraction of the output.

Second, we plotted the percentage of reassigned genes at different thresholds to assess the performance of the default threshold value of 0.02 (Figure 4.6) Below the default value, the percentage of reassigned genes begins to plateau, while above the default value the percentage of reassigned genes increases very quickly, increasing the probability that *treeinform* reassigns transcripts from different genes to the same gene in addition to reassign transcripts from the same gene together.

Finally, we compared the density of duplication times under the model provided for Component 2 of the mixture model to the distribution of estimated duplication times for gene trees from Agalma1.0 before *treeinform*, and gene trees from Agalma1.0 after *treeinform* under 3 different thresholds: 50, 0.05, and the default value 0.02 (Figure 4.7. We again fitted chronograms with the same Siphonophora species tree onto

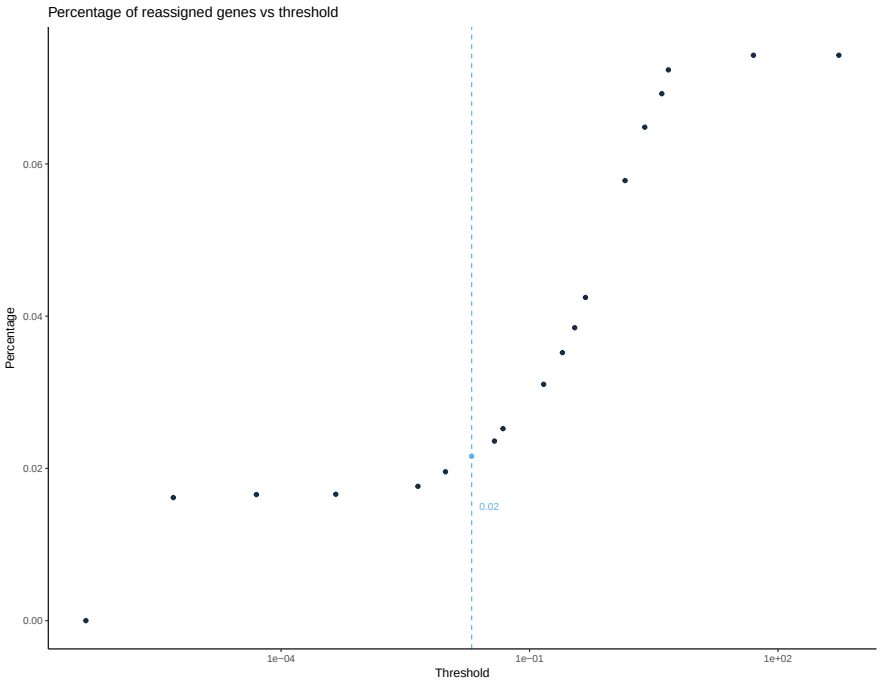


Figure 4.6: The percentage of reassigned tips is plotted above on a log scale. The original assembly had 315,041 genes, with at most 23,396 possible candidates (7.43% of genes) for reassignment. The default threshold for treeinform is highlighted in light blue. This threshold value is robust over a wide range from 0.02 down several orders of magnitude.

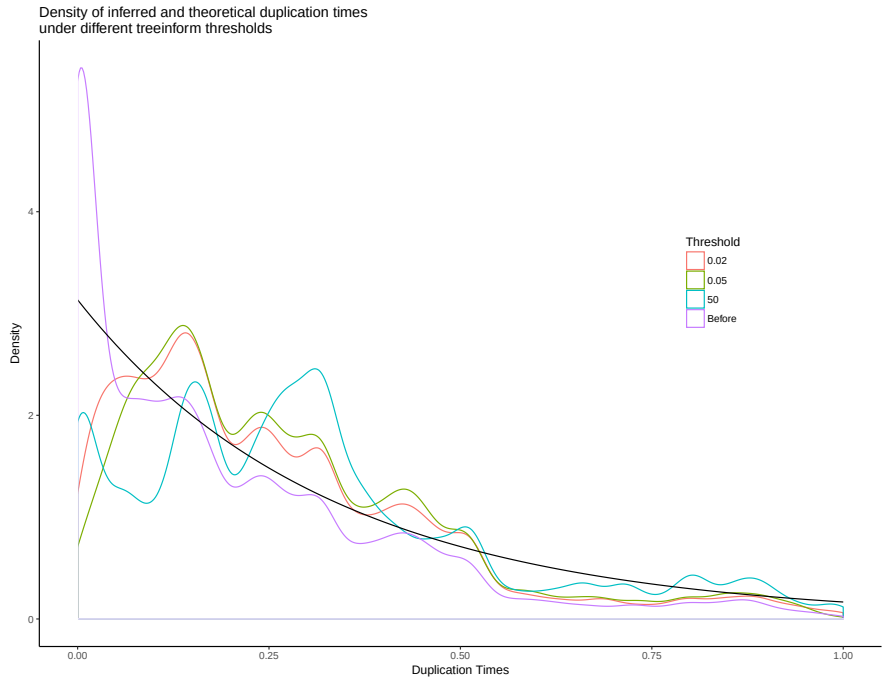


Figure 4.7: Density from theoretical and the empirical density under the 3 different thresholds as well as before treeinform was run. The distribution before treeinform has a large peak on the left that is removed by treeinform with all examined thresholds. Of the evaluated thresholds, 0.02 is closest to the expected distribution.

all gene trees from Agalma1.0 and filtered out those gene trees with time of origin greater than 1, so that duplication times were comparable between trees. Visually, the analyses with the 0.02 threshold comes closest to the theoretical.

Additionally, we computed the Kullback-Leibler divergence[60] between the distributions of duplication times under different thresholds and the theoretical distribution of duplication times (Table 4.2). The theoretical distribution is on the bottom. Kullback-Leibler distance, otherwise known as relative entropy, measures the distance between two distributions. The KL distance between the distribution of duplication times after running *treeinform* with the default threshold value of 0.02 come closest to the theo-

Table 4.2: Kullback-Leibler distances between duplication times after running treeinform with different thresholds and theoretical duplication times.

	KL.Distance
Before	0.3704129
50	0.2699556
0.05	0.1145989
0.02	0.0994612
5e-05	0.1114073

retical distribution as compared to both threshold levels below and above the default value. This indicates that *treeinform* produces more accurate gene trees with appropriate threshold selection. The multiple approaches we took to assessing *treeinform* provided an overview of the impact of the method across the entire output.

4.6 Discussion

We confirm that assignment of transcripts from the same gene to different genes is quite common in transcriptome assemblies. Using phylogenetic information, our new approach reassigns transcripts to their corresponding gene when different transcripts of the same gene are mistaken as transcripts from different closely related genes. This approach capitalizes on the Markovian dependency assumption in phylogenetic workflows[41], as not only can inferences about the gene trees be made solely from the gene assemblies, but inferences about the gene assemblies can be made solely from the gene trees as well. Analyses of *treeinform* shows that it brings estimates of duplication times much closer to theoretical expectations. *treeinform* will be useful for any studies requiring accurate gene trees, in particular accurate counts of different genes

in expression studies. It is available in Agalma1.0, an end-to-end phylogenomic workflow.

4.7 Methods and data

The code for the analyses presented here (including an executable version of this document) are available in a git repository at https://github.com/caseywdunn/ms_treeinform. The phylogenetic analyses considered here are based on a 7 taxon siphonophore[22] dataset. This dataset originated as the regression test dataset for the Agalma automated phylogenetics workflow[23] and was selected for its well-resolved species tree as well as the fact that *treeinform* is implemented in Agalma. The gene trees were built with Agalma1.0, and bash scripts for the run can be found at the Github repository.

The phylogenetic analyses in Agalma followed standard approaches with default settings. Speciation and duplication nodes were identified in the gene trees with phyldog[6]. Bash scripts and associated code for the runs can be found at the Github repository.

Agalma uses the transcriptome assembler Trinity[39]. Given the intrinsic challenges of assigning assembled transcripts to genes it is likely that the same misassignment errors are generated by other transcriptome assemblers as well.

5

Future directions and conclusion

Within each chapter several additional questions or gaps in the models were laid out. Here we address just a few of those questions as potential future directions, based on the guidelines laid out in Chapter 1 with regards to the impact of model improvements and the cost of engineering.

5.1 Sensitivity analyses and ROC curves of the profile-sampling and consensus approaches for HIV transmission inference

The profile-sampling approach developed in Chapter 3 improves inference of HIV transmission clusters, but we did not quantify its impact with sensitivity analyses.

One useful future direction is thus to perform a parameter sweep using simulations from our generative model from Chapter 2 in order to understand how the profile-sampling approach is impacted by factors such as transmission rates, infection case sampling rates, dates of infection, and within-host evolutionary dynamics. We would perform the same parameter sweep on the consensus approach as well. For each parameter sweep, we would investigate how probabilities for transmission clusters change as a function of that parameter. This will allow us to better assess the applicability of the profile-sampling approach to infer transmission clusters from NGS HIV data.

Performance measures of interest would include the Receiver Operator Curve, i.e. the ratio of true transmission cluster detection rate to false transmission cluster detection rate as a function of probability cutoffs and the Robinson-Foulds distances and clusterings including the true epidemiological tree to measure phylogenetic accuracy of our approach.

5.2 A generative model for the processes that generate contact networks, epidemiological transmission networks, and molecular transmission networks

While in Chapter 2 we formed a model for the dynamics between transmission networks and the molecular phylogeny, we did not develop an explicit definition of transmission clusters, the inference unit that clinicians and public health officials are most interested in. In part this is because of the ad hoc way in which transmission clusters have been defined, with bootstrap cutoffs in the range of 90-95% and genetic distance cutoffs in the range of 1.5-4.5% genetic distance. Larger genetic distances are selected

when looking for missing links, while smaller genetic distances are selected when looking for rapidly increasing cluster size[43]. These cutoff measures have not been explicitly modeled from epidemiological features of interest, thus there is no clear consensus on how to define the clusters[8]. This lack of explicit connection posed a problem for simulation studies based on the generative model in Chapter 2 and the profile-sampling approach in Chapter 3. Technically, every node on the simulated phylogeny represents a true transmission cluster. When we ran either the consensus approach or the profile-samping approach on the simulations, we would get differently sized but largely true transmission clusters.

An explicit model of epidemiological features of interest, such as rapid expansion in a network indicative of the beginnings of an epidemic, or superspreader dynamics, and how it maps to phylogenetic trees, will facilitate a better understanding of how to define transmission clusters in a principled way. A shared agreement on the definition of transmission clusters will not only improve detection of the relevant epidemiological dynamics that transmission clusters attempt to capture, but also allow for better comparisons between studies.

5.3 A genome summary approach for within-host variation that includes linkage information and codon awareness

The profile-sampling approach summarizes the within-host variation by modeling base pair, insertion and deletion frequencies at each site with a profile Hidden Markov Model. This approach has the advantage of being easy to use and implement with the pre-existing tool HMMer, but lacks linkage information that exists in the reads.

For example, if within an individual we have half of the reads with the sequence ATAC(rest of sequence), and the other half of reads are CAAC(rest of sequence) , then in the pHMM we will sample sequences ATAC(rest of sequence), ACAC(rest of sequence), CTAC(rest of sequence), and CAAC(rest of sequence) with equal probabilities. However, only ATAC(rest of sequence) and CAAC(rest of sequence) are actual genomic sequences. There exist genome assembly tools that model linkage[48], but these tools require additional algorithmic improvements and mathematical tuning to be applicable to HIV.

5.4 Final thoughts

In this thesis I presented a unified conceptual framework for improvements in the phylogenetic workflow that takes into consideration not only model improvements but also re-engineering cost and computational tractability. We defined implicit assumptions in the workflow, then focused on relaxing the assumption of low relative entropy between a point estimate and the space of solutions, particularly within genome assembly. Both of our improvements are grounded in an explicit generative model, one of which we developed for HIV transmission links which incorporates within-host evolution and transmission dynamics. All of our approaches benefitted from minimal implementation cost with large improvements in results of interest.

Here we have presented several future directions that incorporate the principles laid out in our conceptual framework for improvements in the phylogenetic workflow: performing sensitivity analyses with simulations to assess workflows, extensions to the generative model that allow us to define transmission clusters in a generative way

rather than using ad hoc heuristics like bootstraps, and model improvements for summarizing and propagating genomic variation. These future directions will allow us to better assess where the most impactful improvements can be made in HIV and other infectious disease transmission inference, and hopefully develop models and tools to fill those gaps.

References

[1] Akerborg, O., Sennblad, B., Arvestad, L., & Lagergren, J. (2009). Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci USA*, 106, 5714–9.

[2] Anisimova, M., Liberles, D. A., Philippe, H., Provan, J., Pupko, T., & von Haeseler, A. (2013). State-of the art methodologies dictate new standards for phylogenetic analysis. *BMC Evol. Biol.*, 13, 161.

[3] Bayzid, M. S., Mirarab, S., & Warnow, T. (2013). Inferring optimal species trees under gene duplication and loss. *Pac. Symp. Biocomput.*, (pp. 250–261).

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993–1022.

[5] Boussau, B. & Daubin, V. (2010). Genomes as documents of evolutionary history. *Trends Ecol. Evol.*, 25(4), 224–232.

[6] Boussau, B., Szollosi, G. J., Duret, L., Gouy, M., Tannier, E., & Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2), 323–330.

[7] Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. A., Ganapathy, G., Gibbs, R. A., Gnerre, S., Godzaridis, E., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., Maccallum, I., Macmanes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T. D., Paten, B., Paulo, O. S., Phillippy, A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D. S., Ruby, J. G., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C.,

Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., & Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*, 2(1), 10.

[8] Brenner, B. G. & Wainberg, M. A. (2013). Future of Phylogeny in HIV Prevention. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 63, S248–S254.

[9] Bryant, D. (2003). A classification of consensus methods for phylogenetics. *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, 61, 163–184.

[10] Caboche, S., Audebert, C., Lemoine, Y., & Hot, D. (2014). Comparison of mapping algorithms used in high-throughput sequencing: application to ion torrent data. *BMC Genomics*, 15, 264.

[11] Cartwright, R. A. (2005). DNA assembly with gaps (dawg): simulating sequence evolution. *Bioinformatics*, 21 Suppl 3, iii31–8.

[12] Chomsky, N. (1956). Three models for the description of language. *Information Theory, IRE Transactions on*, 2(3), 113–124.

[13] Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.*, 47(9).

[14] Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, 4(4), 265–270.

[15] Collins, M. (2003). Head-Driven statistical models for natural language parsing. *Comput. Linguist.*, 29(4), 589–637.

[16] Conrad, C., Bradley, H. M., Broz, D., Buddha, S., Chapman, E. L., Galang, R. R., Hillman, D., Hon, J., Hoover, K. W., Patel, M. R., Perez, A., Peters, P. J., Pontones, P., Roseberry, J. C., Sandoval, M., Shields, J., Walthall, J., Waterhouse, D., Weidle, P. J., Wu, H., & Duwve, J. M. (2015). Community Outbreak of HIV Infection Linked to Injection Drug Use of Oxymorphone — Indiana, 2015. *Mmwr*, 64(16), 443–444.

[17] Cover, T. M. & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[18] De Oliveira Martins, L., Mallo, D., & Posada, D. (2014). A bayesian supertree model for Genome-Wide species tree reconstruction. *Syst. Biol.*

[19] de Queiroz, A. & Gatesy, J. (2007). The supermatrix approach to systematics. *Trends Ecol. Evol.*, 22(1), 34–41.

[20] Degnan, J. H. & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.*, 24(6), 332–340.

[21] Degnan, J. H. & Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59(1), 24–37.

[22] Dunn, C. W. (2009). Siphonophores. *Current Biology*, 19(6), R233–R234.

[23] Dunn, C. W., Howison, M., & Zapata, F. (2013). Agalma: an automated phylogenomics workflow. *BMC Bioinformatics*, 14, 330.

[Durbin et al.] Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. 1998.

[25] Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.*, 9, 157.

[26] Eddy, S. (2011). Accelerated Profile HMM Searches. *PLOS Comput Biol*, 7, e1002195.

[27] Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, 14(9), 755–763.

[28] Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, 23(1), 205–211.

[29] Eddy, S. R. & Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Res.*, 22(11), 2079–2088.

[30] Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, 63(1), 1–19.

[31] Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17, 368–376.

[32] Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer associates Sunderland.

[33] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, 44(D1), D279–D285.

[34] Fischer, A. & Igel, C. (2012). An introduction to restricted boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Lecture Notes in Computer Science (pp. 14–36). Springer Berlin Heidelberg.

[35] Fletcher, W. & Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.*, 26(8), 1879–1888.

[36] Gernhard, T. (2008). The conditioned reconstructed process. *Journal of theoretical biology*, 253(4), 769–778.

[37] Ghodsi, M., Hill, C. M., Astrovskaya, I., Lin, H., Sommer, D. D., Koren, S., & Pop, M. (2013). De novo likelihood-based measures for comparing genome assemblies. *BMC Res. Notes*, 6, 334.

[38] Giardina, F., Romero-Severson, E. O., Albert, J., Britton, T., & Leitner, T. (2017). Inference of Transmission Network Structure from HIV Phylogenetic Trees. *PLoS Computational Biology*, 13(1).

[39] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., & Regev, A. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat Biotech*, 29(7), 644–652.

[40] Grant, J. R. & Katz, L. A. (2014). Building a phylogenomic pipeline for the eukaryotic tree of life - addressing deep phylogenies with genome-scale data. *PLoS Curr.*, 6.

[41] Guang, A., Zapata, F., Howison, M., Lawrence, C. E., & Dunn, C. W. (2016). An Integrated Perspective on Phylogenetic Workflows. *Trends in Ecology & Evolution*, 31(2), 116–126.

[42] Hasegawa, M., Kishino, H., & Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol. Evol.*, (pp. 160–174).

[43] Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J., & Esbjörnsson, J. (2017). Defining HIV-1 transmission clusters based on sequence data. *AIDS*, 31(9), 1211–1222.

[44] Heled, J., Bryant, D., & Drummond, A. J. (2013). Simulating gene trees under the multispecies coalescent and time-dependent migration. *BMC Evol. Biol.*, 13, 44.

[45] Heled, J. & Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27(3), 570–580.

[46] Holder, M. & Lewis, P. O. (2003). Phylogeny estimation: traditional and bayesian approaches. *Nat. Rev. Genet.*, 4(4), 275–284.

[47] Howison, M., Zapata, F., & Dunn, C. W. (2013). Toward a statistically explicit understanding of de novo sequence assembly. *Bioinformatics*, 29(23), 2959–2963.

[48] Howison, M., Zapata, F., Edwards, E. J., & Dunn, C. W. (2014). Bayesian genome assembly and assessment by markov chain monte carlo sampling. *PLoS One*, 9(6), e99497.

[49] Huang, W., Li, L., Myers, J., & Marth, G. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, 28, 593–594.

[50] Huelsenbeck, J. P., Rannala, B., & Masly, J. P. (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, 288(5475), 2349–2350.

[51] Hué, S., Clewley, J. P., Cane, P. A., & Pillay, D. (2004). HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS*, 18(5), 719–728.

[52] Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., & Ferguson, N. (2014). Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology*, 10(1).

[53] Joseph, S. B., Swanstrom, R., Kashuba, A. D. M., & Cohen, M. S. (2015). Bottlenecks in hiv-1 transmission: insights from the study of founder viruses. *Nat Rev Microbiol*, 13, 414–425.

[54] Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.*, 93(24), 13770–13773.

[55] Katoh, K. & Standley, D. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol*, 30(4), 772–780.

[56] Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proc. Natl. Acad. Sci. U. S. A.*, 105(31), 10687–10692.

[57] Kloc, M. & Zagrodzinska, B. (2001). Chromatin elimination–an oddity or a common mechanism in differentiation and development? *Differentiation*, 68(2-3), 84–91.

[58] Knowles, L. L. & Kubatko, L. S. (2011). *Estimating Species Trees: Practical and Theoretical Aspects*. Wiley.

[59] Kubatko, L. S., Carstens, B. C., & Knowles, L. L. (2009). STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7), 971–973.

[60] Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1), 79–86.

[61] Langmead, C. J. (2014). Generative models of conformational dynamics. *Adv. Exp. Med. Biol.*, 805, 87–105.

[62] Leitner, T., Escanilla, D., Franzén, C., Uhlén, M., & Albert, J. (1996). Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 93(20), 10864–9.

[63] Lemmon, E. M. & Lemmon, A. R. (2013). High-Throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, 44(1), 99–121.

[64] Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., & Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from rna-seq data. *Genome Biology*, 15(12), 553.

[65] Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., & Warnow, T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934), 1561–1564.

[66] Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., & Linder, C. R. (2012). SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst. Biol.*, 61(1), 90–106.

[67] Liu, L. & Pearl, D. K. (2007). Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.*, 56(3), 504–514.

[68] Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature-London-*, 438(7066), 355.

[69] Löytynoja, A. & Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U. S. A.*, 102(30), 10557–10562.

[70] Lu, W., Ng, H. T., Lee, W. S., & Zettlemoyer, L. S. (2008). A generative model for parsing natural language to meaning representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08 (pp. 783–792). Stroudsburg, PA, USA: Association for Computational Linguistics.

[71] Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., & Hein, J. (2008). Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, 18(2), 298–309.

[72] Lysholm, F., Andersson, B., & Persson, B. (2011). An efficient simulator of 454 data using configurable statistical models. *BMC Res. Notes*, 4, 449.

[73] Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46, 523–536.

[74] Maddison, W. P. & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.*, 55(1), 21–30.

[75] Maddison, W. P. & Maddison, D. R. (2007). Mesquite: a modular system for evolutionary analysis. version 2.75. 2011. *URL http://mesquiteproject. org*.

[76] Mallo, D., de Oliveira Martins, L., & Posada, D. (2015). SimPhy: Phylogenomic Simulation of Gene, Locus and Species Trees. *Syst Biol*, 65, 334–344.

[77] Maretty, L., Sibbesen, J., & Krogh, A. (2014). Bayesian transcriptome assembly.

[78] Martin, J. A. & Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet.*, 12(10), 671–682.

[79] McTavish, E. J., Pettengill, J., Davis, S., Rand, H., Strain, E., Allard, M., & Timme, R. E. (2017). Treetoreads - a pipeline for simulating raw reads from phylogenies. *BMC Bioinformatics*, 18(1), 178.

[80] Misner, I., Bicep, C., Lopez, P., Halary, S., Bapteste, E., & Lane, C. E. (2013). Sequence comparative analysis using networks: software for evaluating de novo transcript assembly from next-generation sequencing. *Mol. Biol. Evol.*, 30(8), 1975–1986.

[81] Muse, S. V. & Weir, B. S. (1992). Testing for equality of evolutionary rates. *Genetics*, 132(1), 269–276.

[82] Nagarajan, N. & Pop, M. (2013). Sequence assembly demystified. *Nat. Rev. Genet.*, 14(3), 157–167.

[83] Newberg, L. A. & Lawrence, C. E. (2009). Exact calculation of distributions on integers, with application to sequence alignment. *J. Comput. Biol.*, 16(1), 1–18.

[84] Nichols, R. (2001). Gene trees and species trees are not the same. *Trends Ecol. Evol.*, 16(7), 358–364.

[85] Oakley, T. H., Alexandrou, M. A., Ngo, R., Pankey, M. S., Churchill, C. K. C., Chen, W., & Lopker, K. B. (2014). Osiris: accessible and reproducible phylogenetic and phylogenomic analyses within the galaxy workflow management system. *BMC Bioinformatics*, 15, 230.

[86] Owen, M. & Provan, J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1), 2–13.

[87] Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinformatics*, Chapter 3, Unit3.1.

[88] Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

[89] Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2), 257–286.

[90] Rahman, A. & Pachter, L. (2013). CGAL: computing genome assembly likelihoods. *Genome Biol.*, 14(1), R8.

[91] Rannala, B. & Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.*, 43(3), 304–311.

[92] Rasmussen, D. A., Kouyos, R., Günthard, H. F., & Stadler, T. (2017). Phylodynamics on local sexual contact networks. *PLoS Computational Biology*, 13(3), 1–23.

[93] Rasmussen, M. D. & Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.*, 22(4), 755–765.

[94] Redelings, B. D. & Suchard, M. A. (2005). Joint bayesian estimation of alignment and phylogeny. *Syst. Biol.*, 54(3), 401–418.

[95] Robinson, D. F. & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2), 131–147.

[96] Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, 61(3), 539–542.

[97] Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.*, 14(5), R51.

[98] Rusk, N. (2009). Cheap third-generation sequencing. *Nat. Methods*, 6(4), 244–244.

[99] Salazar-Gonzalez, J. F., Salazar, M. G., Keele, B. F., Learn, G. H., Giorgi, E. E., Li, H., Decker, J. M., Wang, S., Baalwa, J., Kraus, M. H., Parrish, N. F., Shaw, K. S., Guffey, M. B., Bar, K. J., Davis, K. L., Ochsenbauer-Jambor, C., Kappes, J. C., Saag, M. S., Cohen, M. S., Mulenga, J., Derdeyn, C. A., Allen, S., Hunter, E., Markowitz, M., Hraber, P., Perelson, A. S., Bhattacharya, T., Haynes, B. F., Korber, B. T., Hahn, B. H., & Shaw, G. M. (2009). Genetic identity, biological

phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early hiv-1 infection. *Journal of Experimental Medicine*, 206(6), 1273–1289.

[100] Sanderson, M. J., McMahon, M. M., & Steel, M. (2011). Terraces in phylogenetic tree space. *Science*, 333(6041), 448–450.

[101] Schultz, A.-K., Zhang, M., Leitner, T., Kuiken, C., Korber, B., Morgenstern, B., & Stanke, M. (2006). A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC bioinformatics*, 7, 265.

[102] Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1), 3–55.

[103] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.*, 19(6), 1117–1123.

[104] Sjöstrand, J., Arvestad, L., Lagergren, J., & Sennblad, B. (2013). GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics*, 14, 209.

[105] Smith, J. J., Baker, C., Eichler, E. E., & Amemiya, C. T. (2012). Genetic consequences of programmed genome rearrangement. *Curr. Biol.*, 22(16), 1524–1529.

[106] Spielman, S. J. & Wilke, C. O. (2015). Pyvolve: A flexible python module for simulating sequences along phylogenies. *PLoS ONE*, 10(9).

[107] Stadler, T., Kühnert, D., Bonhoeffer, S., & Drummond, A. J. (2013). Birth – death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Pnas*, 110(1), 228–233.

[108] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313.

[109] Steel, M., Linz, S., Huson, D. H., & Sanderson, M. J. (2013). Identifying a species tree subject to random lateral gene transfer. *J. Theor. Biol.*, 322, 81–93.

[110] Stoye, J., Evers, D., & Meyer, F. (1998). Rose: generating sequence families. *Bioinformatics*, 14(2), 157–163.

[111] Strope, C. L., Abel, K., Scott, S. D., & Moriyama, E. N. (2009). Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Mol. Biol. Evol.*, 26(11), 2581–2593.

[112] Sturmer, M., Preiser, W., Gute, P., Nisius, G., & Doer, H. (2004). Phylogenetic analysis of HIV-1 transmission: pol gene sequences are insufficient to clarify true relationships between patient isolates. *AIDS*, 18, 2109–2113.

[113] Szitenberg, A., John, M., Blaxter, M. L., & Lunt, D. H. (2015). ReproPhylo: An environment for reproducible phylogenomics. *bioRxiv*.

[114] Szollosi, G. J. & Daubin, V. (2012). Modeling gene family evolution and reconciling phylogenetic discord. *Evolutionary Genomics*, 856, 29–51.

[115] Szöllosi, G. J. & Daubin, V. (2012). Modeling gene family evolution and reconciling phylogenetic discord. *Methods Mol. Biol.*, 856, 29–51.

[116] Szöllősi, G. J., Tannier, E., Daubin, V., & Boussau, B. (2015). The inference of gene trees with species trees. *Syst. Biol.*, 64(1), e42–62.

[117] Varón, A., Vinh, L. S., & Wheeler, W. C. (2010). POY version 4: phylogenetic analysis using dynamic homologies. *Cladistics*, 26(1), 72–85.

[118] Volz, E. M., Koopman, J. S., Ward, M. J., Brown, A. L., & Frost, S. D. W. (2012). Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. *PLoS Computational Biology*, 8(6), 2–11.

[119] Vrancken, B., Rambaut, A., Suchard, M. A., Drummond, A., Baele, G., Derdelinckx, I., Van Wijngaerden, E., Vandamme, A. M., Van Laethem, K., & Lemey, P. (2014). The Genealogical Population Dynamics of HIV-1 in a Large Transmission Chain: Bridging within and among Host Evolutionary Rates. *PLoS Computational Biology*, 10(4).

[120] Wang, Z., Gerstein, M., & Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1), 57–63.

[121] Webb-Robertson, B. J. M., McCue, L. A., & Lawrence, C. E. (2008). Measuring global credibility with application to local sequence alignment. *PLoS Computational Biology*, 4(5).

[122] Wheeler, W. (1996). OPTIMIZATION ALIGNMENT: THE END OF MULTIPLE SEQUENCE ALIGNMENT IN PHYLOGENETICS? *Cladistics*, 12(1).

[123] Wong, K. M., Suchard, M. A., & Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, 319(5862), 473–476.

[124] Yang, Y. & Smith, S. A. (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, 14, 328.

[125] Yebra, G., Hodcroft, E. B., Ragonnet-Cronin, M. L., Pillay, D., Brown, A. J. L., Fraser, C., Kellam, P., & de Oliveira, T. (2016). Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. *Scientific Reports*, 6(1).

[126] Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.*, 18(5), 821–829.

[127] Zwickl, D. J. & Holder, M. T. (2004). Model parameterization, prior distributions, and the general time-reversible model in bayesian phylogenetics. *Systematic Biology*, 53(6), 877–888.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under cc by-nc-nd 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive mit (x11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.