Inference of Gene Tree Discordance and Recombination

by

Yujin Chung

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

University of Wisconsin–Madison

2012

Date of final oral examination: 12/12/2012

The dissertation is approved by the following members of the Final Oral Committee:

Bret R. Larget, Associate Professor, Statistics
Cécile Ané, Associate Professor, Statistics
Michael A. Newton, Professor, Statistics
Murray K. Clayton, Professor, Statistics
Nicole T. Perna, Associate Professor, Genetics

Inference of Gene Tree Discordance and Recombination

Yujin Chung

Under the supervision of Professor Cécile Ané

At the University of Wisconsin–Madison

Abstract

The evolutionary history of a group of organisms is important for studying their genomes. Many biological processes can cause the evolutionary histories of different genes to vary, and such incongruent histories may mislead the inference of a species tree. New methods for the reconstruction of gene trees and species trees from molecular data have been enabled by recent sequencing and computing advances. I examine here the sensitivity of two such Bayesian methods, BEST and BUCKy, to the violation of their assumptions in the presence of two biological processes: incomplete lineage sorting (ILS) and horizontal gene transfer (HGT). BEST assumes gene tree discordance is due to ILS based on the coalescent model. Instead, BUCKy uses a non-parametric clustering prior distribution on gene trees. The species tree was found to be more accurately reconstructed by BUCKy than BEST in the presence of HGT. Both methods performed similarly in most other cases. A test is proposed to determine if the coalescent model alone can explain gene tree discordance. This test uses the concordance factor (CF) of clades, i.e. the proportion of genes that have the clade

in their trees, estimated by BUCKy. The test is powerful when HGT is the main source of the discordance and has low Type I error.

Ancestral recombination events can cause the underlying genealogy of a site to vary along the genome. The genomic location where the site-specific phylogenetic tree changes are called "recombination breakpoints". I propose here a Bayesian model to simultaneously infer the recombination breakpoints and the tree of each segment between breakpoints. The proposed model uses a dissimilarity measure between trees in the prior distribution on segment trees, to match empirical evidence that neighboring genomic regions have similar trees. The main obstacle to using the proposed model is the calculation of the normalizing function for the prior distribution on segment trees. An algorithm is developed to calculate this normalizing function exactly. Fast approximations are also proposed, which are very accurate with little impact on the inference of gene trees and recombination breakpoints.

To Mom and Dad

Acknowledgments

When my journey to a Ph.D started here at the University of Wisconsin - Madison, I was very excited to start this new adventure in my life. I am now at the end of this long journey and face a new journey to start ahead. When I look back at the road I have walked, I find that this journey is full of tough but fruitful and joyful events and many people made it possible to complete my journey. I cannot be more happy to share everything I achieved during this journey with them.

First and foremost, I would like to express my sincerest gratitude to my advisor Prof. Cécile Ané. Without her consistent support, both academically and emotionally, this adventure would not have been possible. Dr. Ané introduced me to the world of statistical phylogenetics and allowed me to pursue the research I was interested in. With her knowledge and creativity, she always guided me wisely through the obstacles I faced. Through working with her, I have achieved professional skills and qualities required to be a real scientist: how to look into research questions, to solve problems scientifically, to collaborate and network with other scientists and to present my work in papers and presentations. By just watching how she enthusiastically works, I could see what type of scientist I would like to become. Every experience with her was invaluable to me, and it has been a fortune for me to do research under her supervision.

I would like to give my appreciation to Prof. Bret Larget who has advised me on both my research and career since I started to work on statistical phylogenetics. His

kind and sincere support always made me smile. I would also like to express my genuine gratitude to Nicole Perna for providing Enterobacteria genomes and great suggestions on my thesis work. One of my main research projects was motivated by learning about Enterobacteria generated in Nicole Perna's lab. I would also like to sincerely thank Prof. David Baum for unstinting help and great suggestions on my graduate study. His passion and great interest in evolution and learning new approaches are a unique inspiration to me.

Additionally, I would like to thank my committee members Profs. Michael Newton and Murray Clayton for their insightful comments, suggestions, and time. My thanks also goes to Guy Plunkett III in Perna's lab for providing alignments of 33 *Escherichia* genomes and 8 *Shigella* genomes. I also thank Aaron Darling for technical assistance with the alignments. I would like to thank former and current students and postdoc in Ané's lab and Larget's Lab: Jee Yeon Kim, Lam Ho, Heejung Shim, Joungyoun Kim, Charles-Elie Rabier. I really enjoyed my discussions with them! Special thanks should go to Prof. Kjell Doksum for providing me an opportunity to learn and work on semiparametric models. His insightful guidance on the work enriched my research experience. I am thankful to Quoc Tran for not only being my colleague, but also being one of my best friends.

I also thank my many close friends, Jeeyoung Moon, Lisa Chung, Perla Reyes, Dongjun Chung and Seho Park for their friendship and emotional support and Sohee Oh for her realistic advice and cheerful support. I am thankful to all my friends who shared every moment of happiness and hardships with me through my PhD journey.

Most of all, I wish to thank my parents, Jeongja Kang and Hejun Chung, and my sisters and brothers who were always there to support me and made me feel how much

they love me. I also thank my nephews and niece for just being there and spreading happiness. This thesis is dedicated to my parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Gene tree discordance

Evolutionary relationships among a group of organisms are visualized as phylogenetic trees to examine ancestor-descendant relationships. Phylogenies can be reconstructed from any information that reveals similarity by descent. For example, some phylogenies are based on observations of morphology and development of living species (see Dayrat, 2003) or palaeontological information to reconstruct the relationships between extinct species (Sereno, 1999). Recent phylogenetic studies have used molecular data to uncover the evolutionary history of a group of organisms, from one gene to massive amounts of sequence data, in some cases using over 100 genes selected from genomes of eight species of yeast (Rokas et al., 2003) or whole genomes of human, chimpanzee, gorilla, orangutan, and rhesus (Ebersberger et al., 2007). More recently, with the advent of next-generation sequencing (NGS; or high-throughput sequencing), data accumulate even faster. These time-efficient and cost-effective technologies can be used for phylogenomic studies (McCormack et al., 2011; Ekblom and Galindo, 2010; Lerner and Fleischer, 2010). Unfortunately, however, it is challenging to apply NGS to phylogenomics and McCormack et al. (2011) specified reasons, such as the difficulty to find homologous genomic regions from many individuals or to generate whole loci from short reads.

A species tree is defined as the branching pattern of species lineages through the process of speciation (Maddison, 1997). A gene at a non-recombining locus has a single ancestral gene, so the resulting history of a recombination-free gene a tree-like descent pattern, called a gene tree. It has been recognized that evolutionary histories

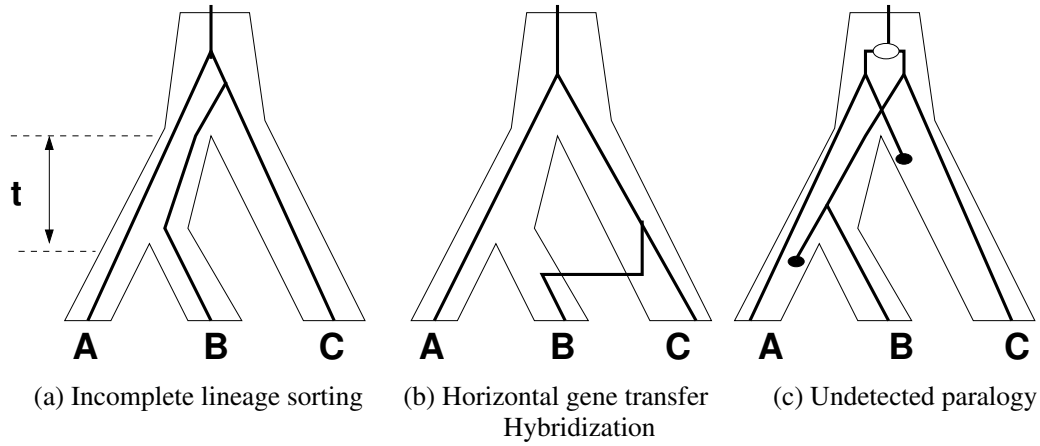

Figure 1.1: Biological processes causing gene tree discordance in a species tree ((A,B),C). (a) Failing in the coalescence of gene copies of A and B during time period t gives rise to a chance that gene copy of B coalesces with that of C, not A, which results in gene tree (A,(B,C)). (b) Genetic material from population C is transferred into population B and hence the genealogical tree of such a gene has tree (A,(B,C)). (c) Genes can be duplicated (white circle) and lost (black circle). Observing paralogous genes can result in a different gene tree (A,(B,C)).

of different genes can have conflicting branching patterns (Maddison, 1997; Degnan and Rosenberg, 2009) from the studies of eukaryotes (Chen and Li, 2001; Ebersberger et al., 2007; Takahashi et al., 2001) through prokaryotes (Zhaxybayeva et al., 2006; Galtier, 2007).

Gene tree discordance can be due to biological processes such as incomplete lineage sorting (ILS), horizontal gene transfer (HGT), hybridization, and gene duplication and extinction (Maddison, 1997). ILS is explained as either a time-forward process or a time-backward process. As time goes forward, ancestral polymorphism can be retained through several speciation events, which results in a given gene having a different tree from the species tree (Maddison, 1997). Looking backward in time, in Fig. 1.1a, ILS is described as the failure of two or more lineages in a population to coalesce, i.e., to share a common ancestor. The failure of coalescence of two closely

related lineages creates the possibility that one of the two lineages first coalesces with a distantly related lineage instead of the other lineage (Degnan and Rosenberg, 2009). For example, in Fig. 1.1a, gene copies of A and B failed to coalesce during time period t and then gene copy of B coalesces with C, not A. ILS is ubiquitous in many organisms such as eukaryotes. While ILS does not allow genes to cross species boundaries, HGT does (Fig. 1.1b). It is the process by which some genes break species lineages and move horizontally across the phylogeny (Maddison, 1997). HGT events are thought to be more common among closely related organisms. HGT is also very common in bacteria (Lawrence, 1999). When populations are not fully isolated, hybridization events between populations is likely (Brumfield, 2010). This process is similar to HGT (Fig. 1.1b) and is common in many groups of plants. By gene duplication, the duplicated genes in an organism occupies two different positions in the same genome. These two copies are paralogous. The process of gene duplication generates multiple gene lineages in a species lineage. If gene copies are then lost so as to result in a single copy in each species (as shown in Fig. 1.1c), then we have paralogous genes instead of orthologous genes but this paralogy may go undetected. The phylogenetic tree of such genes can be different from the species tree and from other gene trees (Maddison, 1997).

1.2 Inference of species trees

Gene tree discordance cannot be ignored when a species tree is inferred from particular gene trees (Degnan and Rosenberg, 2009; Kubatko and Degnan, 2007; Degnan and Rosenberg, 2006). "Total evidence" and "democratic vote" are common approaches to estimate a species tree. The total evidence approach uses a single

phylogenetic tree estimated from the concatenated genes, ignoring any discordance. The democratic vote approach uses the tree most frequently inferred from the sampled genes. However, both methods can be statistically inconsistent as more data are added (Degnan and Rosenberg, 2009).

Many methods have been developed to estimate species trees under the assumption of specific biological processes causing gene tree discordance. For example, minimizing deep coalescence (MDC) in Mesquite (Maddison, 1997; Maddison and Knowles, 2006) assumes that the incongruence is exclusively due to ILS. MDC is a parsimony-based approach for inferring species trees from gene trees by minimizing the number of extra lineages, or minimizing deep coalescenses (MDC). There are also many other methods estimating species trees in the presence of ILS and most of them are based on the coalescent model (Kubatko et al., 2009; Edwards et al., 2007; Liu, 2008; Heled and Drummond, 2010), see Table 1.1. The coalescent process is a stochastic model for the random joining of sampled gene lineages as they are followed back in time. Under the coalescent model, times between coalescent events follow an exponential distribution. The parameter of the exponential distribution depends on both the remaining number of sampled lineages and the size of the underlying population. Bayesian estimation of species tree (BEST; Edwards et al., 2007; Liu, 2008) is one method using this coalescent process. It builds on a Bayesian hierarchical model: sequence data are modeled from gene trees using standard substitution models, and gene trees are modeled from the species tree using the coalescent process. Since the model assumes that discrepancies between gene trees and the species tree are due exclusively to lineage sorting, it is not appropriate to use BEST when other biological phenomena have been at work, such as horizontal transfers or gene duplications/deletions (Liu, 2008).

Biological reason	Methods
ILS	STEM (Kubatko et al., 2009), iGLASS (Jewett and Rosenberg, 2012), STEAC (Liu et al., 2009), STAR (Liu et al., 2009), NJst (Liu and Yu, 2011), MDC (Maddison, 1997; Maddison and Knowles, 2006), BEST (Liu, 2008), *BEAST (Heled and Drummond, 2010)
HGT/Hybridization	HybTree (Gerard et al., 2011), STEM-hy (Kubatko, 2009)
none	BUCKy (Ané et al., 2007)

Table 1.1: List of methods for the estimation of species trees categorized based on their assumption of underlying biological sources for gene tree discordance.

In contrast to other methods for species tree inference, BUCKy (Bayesian untangling of concordance knots; Ané et al., 2007) does not assume any biological reason for the discordance. As a measure of the magnitude of concordance among gene trees, BUCKy estimates the concordance factor of each clade, that is, the proportion of genes in the genome that include the clade in their gene trees. The primary concordance tree is built from clades with high CFs. In order to estimate CFs and the concordance tree, BUCKy first estimates the joint distribution of gene trees in a two-stage Bayesian concordance analysis. For each gene, the posterior probability distribution of trees for that gene is obtained based on the a priori belief that genes tend to share the same trees. This is achieved using existing methods such as MrBayes (Huelsenbeck and Ronquist, 2001). Then, the joint distribution of gene trees is obtained using a Dirichlet process, which acts as a non-parametric clustering to detect which genes share the same tree. From the estimated joint distribution of gene trees, BUCKy estimates the sample-wide CF of each clade, i.e., the proportion of genes in the available sample that truly have the clade, and the genome-wide CF of each clade, i.e., the proportion of genes in the entire genome that truly have the clade.

There are many methods to estimate species trees and gene trees, but their assumptions about the underlying biological source for gene tree discordance are varied (Table 1.1). The performance of these methods can depend on the true biological source of gene tree discordance. For example, in the presence of HGT or hybridization, methods based on coalescent theory are expected to show poor performance because of the violation of assumptions. In Chapter 2, we study the sensitivity of existing methods to the violation of their assumptions. Moreover, we develop a test to determine whether the coalescent model is a reasonable explanation of discordance for the data at hand.

1.3 Study of comparing BEST and BUCKy and a test for the adequacy of using coalescent model

Most methods, including BEST, are most concerned with ILS and are based on the coalescent theory. BUCKy, on the other hand, does not assume any specific biological sources for gene tree discordance, but is based on a non-parametric clustering prior distribution. Therefore, its performance is not expected to be very sensitive to the true biological reason for discordance.

The first part of this thesis compares two Bayesian models BEST and BUCKy for gene tree and species tree estimation in Chapter 2. A simulation study is conducted to compare the performance of BEST and BUCKy in the presence of ILS and/or two kinds of HGT events. Besides the simulated data analysis, we propose a test to see if the coalescent model only is enough to explain gene tree discordance. The test uses CFs of conflicting clades to see if the data satisfies a symmetric pattern, expected

under the coalescent model. The power of the proposed test is also evaluated.

1.4 Recombination and phylogenetic studies

Recombination is one of the vital evolutionary processes for shaping the structure of genomes. Through recombination, genetic material is rearranged by breaking DNA strands and joining genomic fragments from different strands. Recombination in eukaryotes happens during mitosis or meiosis and recombined genetic material is transferred to the next generation, which is called vertical gene transfer. In contrast to vertical transmission of genetic material in eukaryotes, genetic material in prokryotes can be horizontally transferred between species and this process is called horizontal gene transfer (HGT) or lateral gene transfer. As one type of HGT events in bacteria, homologous recombination occurs through unidirectional transfer of genetic material from the donor to the recipient cell by three difference processes: transformation, conjugation and transduction. Transformation is the process by which bacteria can acquire new genes by directly taking up exogenous genetic material. Once inside the bacterial cell, the foreign genomic fragments can be integrated into the host's chromosome by homologous recombination. Through transduction, bacteria can receive genomic fragments from other bacteria or organisms by infection from a bacteriophage. When the phage attacks, it injects its genetic fragment into the chromosome of the bacterium. The viral DNA can be incorporated into the host's DNA by recombination between homologous regions. Conjugation is the process by which two bacteria are physically connected by a tube-like structure called a sex pilus. Through the pilus, genetic material is unidirectionally transferred from the donor cell to the recipient cell.

Recombination can seriously confound results obtained from molecular evolu-

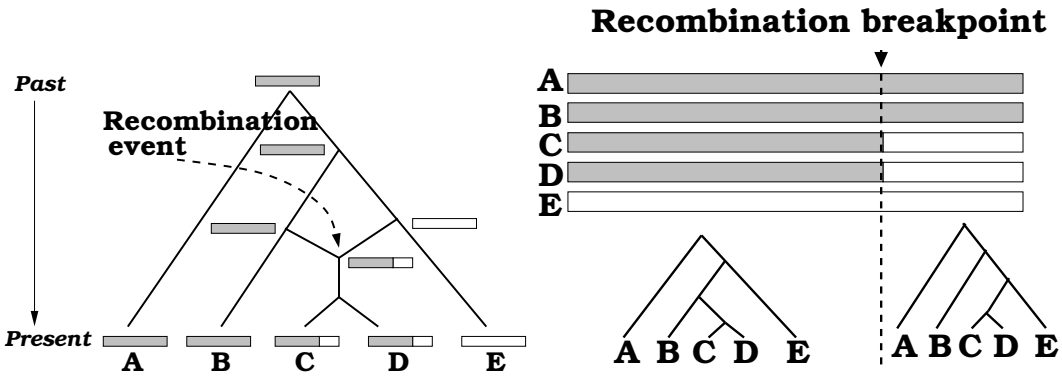

Figure 1.2: Left: A phylogenetic tree with a recombination event on five taxa. After the recombination event between two sequences, the descendants after the recombination event have recombined sequences. Right: The recombination event depicted in the left tree creates a recombination breakpoint the alignment of the five sequences, where the phylogenetic trees of the left and right genomic regions of the recombination breakpoint are different.

tion studies (Schierup and Hein, 2000a,b). The process of recombination mixes DNA sequences and thereby generates evolutionarily mosaic recombinants. When recombination goes undetected, it can mislead phylogenetic reconstruction since traditional phylogenetic methods assume vertical transmission only and infer bifurcating trees. For example, in Fig. 1.2, a recombination event between different species can create a "recombination" breakpoint in an alignment, where the trees of the left and right genomic regions of the breakpoint are different. More specifcally, the left tree contains clade (B,C,D), while the right tree contains clade (C,D,E).

1.5 Detecting recombination breakpoints on alignments

The detection of recombination from DNA sequences is important to understand evolutionary history and molecular genetics. Around 20 methods have been developed to detect recombination from alignments (Posada and Crandall, 2001). Recombination events can be categorizied by their effects on evolutionary histories and all these

methods do not target the same types of recombination events. Some events that occur between closely related individuals (through sexual reproduction in eukaryotes for instance) may not affect the gene tree or its branch lengths, as measured in number of generations between coalescent events (Hein J, 2005). Some methods aim to estimate the total recombination rate, which includes the rate of these events. Other methods do not aim to detect these recombination events, because they do not affect the underlying phylogenetic tree and do not create recombination breakpoints in the genome. Other recombination events create breakpoints in the alignment, where the neighboring genomic regions around the breakpoint have different tree topologies or simply different branch lengths on the same tree topology. For the purpose of species tree reconstruction, it is more important to detect recombination breakpoints affecting the tree topology than those affecting branch lengths only (Ané, 2011). Recombination detection methods are generally categorized in four groups: Distance methods, phylogenetic methods, compatibility methods and methods based on substitution distributions.

Distance methods search for genomic regions with abnormal genetic distance patterns among the sequences (Weiller, 1998; Lee and Sung, 2008). Since they generally use a sliding window approach and are based on distances, these methods are typically fast. PhylPro (Weiller, 1998), for example, calculates pairwise distances between sequences in a certain region and then computes the correlation between pairwise distances from two neighboring genomic regions. Low correlation is a sign of recombination between the two regions. This method is fast and able to analyze more than 1000 sequences, but the threshold to find low correlation is arbitrary and no statistical test of the presence of recombination is provided.

Compatibility approaches compare the patterns at pairs of sites and search for spatial clustering of compatible sites that are incompatible with other sites (Jakobsen and Easteal, 1996; Bruen et al., 2006). Reticulate (Jakobsen and Easteal, 1996) defines 2 sites as compatible if these 2 sites support the same parsimony tree. This method does not require the phylogeny of sequences but only needs to compare patterns of sites. The presence of recombination is tested using random permutations. This method is also fast and able to analyze large data sets. However, it is sensitive to the number of parsimony informative sites and does not provide locations of recombination breakpoints.

Approaches based on substitution distributions detect significantly similar or clustered partitions with respect to their nucleotide substitution distributions (Maynard Smith, 1992; Sawyer, 1989). In CHIMAX (Maynard Smith, 1992), for example, a sliding window approach is considered. The χ^2-test statistic is computed based on the proportion of variable or non-variable sites before and after every putative recombination breakpoint in the window. Recombination breakpoints are found with the maximum χ^2 and their statistical significance is calculated using a permutation test.

Phylogenetic approaches detect genomic regions with discordant phylogenetic relationships. RecPars (Hein, 1993) finds the most parsimonious substitution and recombination history weighing the cost of substitutions and of recombination events with a penalty for each type of event. It is fast but parsimony phylogenetic tree estimation can be inconsistent, no uncertainty of estimation is provided and the recombination cost relative to the cost of substitutions is left to the user. MDL (White et al., 2009; Ané, 2011) defines the description length (DL) of an alignment with respect to a partition as the sum of the maximum-parsimony tree lengths plus a penalty

cost for each fragment. It finds a partition with smallest DL using dynamic programming. MDL is fast enough to analyze whole mammalian genomes, but again, parsimony tree estimation can be inconsistent. PLATO (Grassly and Holmes, 1997) infers the maximum likelihood phylogenetic tree from the whole input alignment and then detects regions whose likelihood value for this tree is relatively small. Similarly, ClonalOrigin (Didelot et al., 2010) uses a hierarchical Bayesian model to estimate recombination breakpoints, based on the primary phylogenetic tree estimated by ClonalFrame (Didelot and Falush, 2007) in advance. BARCE (Husmeier and McGuire, 2003) uses a hidden Markov model (HMM) where the hidden state is the underlying tree at each site. It estimates the posterior probability of gene trees at each site. BARCE is accurate, but available to analyze up to four taxa only. DualBrothers (Minin et al., 2005) is a Bayesian model to detect genomic regions with different tree topologies and/or different substitution models. DualBrothers provides accurate estimation of recombination breakpoints and gene trees and estimates uncertainty of estimations. Also implemented in cBrother (Fang et al., 2007) for faster analysis, it can deal with up to 8 taxa. StepBrothers (Bloomquist et al., 2009) is also a Bayesian model. It separates recombined sequences into segments, so sites in the same segment share the same evolutionary history. Then each segment is augmented to form an entire individual sequence in the alignment. From the augmented alignment, it estimates a single phylogenetic tree with the the relative times of recombination events. However, this method can deal with at most 1 or 2 recombinant sequences, that is, when recombination events altered the phylogenetic placement of 2 sequences at most. Biomc2 (de Oliveira Martins et al., 2008) also uses a Bayesian framework. Its prior distribution on gene trees favors similar trees at neighboring loci. More specifically,

it assumes that the subtree prune and regraft (SPR) distance between trees at neighboring loci follows a truncated-Poisson distribution. Such model is thought to better detect recombination breakpoints between very similar gene trees (de Oliveira Martins et al., 2008). Biomc2 scales better with more taxa and longer sequences than other phylogenetic approaches. However, it miscalculates the normalizing function of the truncated Poisson distribution on gene trees.

The performance of many of the above methods has been evaluated (Smith, 1999; Brown et al., 2001; Posada and Crandall, 2001; Wiuf et al., 2001; Posada et al., 2002). Non-phylogenetic approaches are fast and able to analyze large data sets, but are sensitive to their parameter settings. Some methods do not provide phylogenetic trees and locations of recombination breakpoints. All methods do not aim to detect the same kinds of recombination breakpoints. Phylogenetic approaches are most accurate than other approaches but take longer to run typically. Bayesian phylogenetic approaches provide breakpoints and trees with measures of uncertainty. Although new phylogenetic methods recently emerged, there is no study comparing their performance.

In order to analyze large data sets on many taxa or whole genomes, biomc2 seems a proper method to use. The prior distribution on gene trees in biomc2 is the key component to accurate estimation of recombination breakpoints and gene trees, but the normalizing function of the prior distribution is miscalculated. In chapter 3, we study the issue of the normalizing function and develop a new model favoring similar trees at neighboring loci a priori.

1.6 Calculating the normalizing function for a Bayesian model to detect recombination breakpoints and infer gene trees

In chapter 3, we propose a Bayesian model to simultaneously detect recombination breakpoints and infer phylogenetic trees of genomic regions defined by the breakpoints. The model can be used when discordant gene trees are caused by HGT or hybridization involving recombination events within or between species. In order to detect recombination breakpoints between similar trees at neighboring loci, the proposed prior distribution on trees favors similar neighboring gene trees like biomc2. A real data analysis of 41 Enterobacteria confirms the correlation between trees at neighboring loci. A Gibbs distribution is used as a prior on trees and the Robinson-Foulds distance is used as a measure of dissimilarity between trees.

Computing the normalizing function of the Gibbs distribution is computationally expensive. This problem is not new (Bryant and Steel, 2009). We show that the normalizing function is miscalculated in biomc2 and quantify the impact of not using the correct normalization. Moreover, we provide an efficient algorithm to compute the correct normalizing function. For application to MCMC algorithms, we propose accurate and fast approximations to the normalizing function.

Chapter 2

Comparing Two Bayesian Methods for Gene Tree/Species Tree Reconstruction: Simulations with Incomplete Lineage Sorting and Horizontal Gene Transfer

Abstract

With the increasing interest in recognizing the discordance between gene genealogies, various gene tree/species tree reconciliation methods have been developed. We present here the first attempt to assess and compare two such Bayesian methods, Bayesian estimation of species trees (BEST) and BUCKy (Bayesian untangling of concordance knots), in the presence of several known processes of gene tree discordance. DNA alignments were simulated under the influence of incomplete lineage sorting (ILS) and of horizontal gene transfer (HGT). BEST and BUCKy both account for uncertainty in gene tree estimation but differ substantially in their assumptions of what caused gene tree discordance. BEST estimates a species tree using the coalescent model, assuming that all gene tree discordance is due to ILS. BUCKy does not assume any specific biological process of gene tree discordance through the use of a nonparametric clustering of concordant genes. BUCKy estimates the concordance factor (CF) of a clade, which is defined as the proportion of genes that truly have the clade in their trees. The estimated concordance tree is then built from clades with the highest estimated CFs. Because of their different assumptions, it was expected that BEST would perform better in the presence of ILS and that BUCKy would perform better in the presence of HGT. As expected, the species tree was more accurately reconstructed by BUCKy in the presence of HGT, when the HGT events were unevenly placed across the species tree. BUCKy and BEST performed similarly in most other cases, including in the presence of strong ILS and of HGT events that were evenly placed across the tree. However, BUCKy was shown to underestimate the uncertainty in CF estimation, with short credibility intervals. Despite this, the discordance pattern estimated by BUCKy could be compared with the signature of ILS. The resulting

test for the adequacy of the coalescent model proved to have low Type I error. It was powerful when HGT was the major source of discordance and when HGT events were unevenly placed across the species tree.

2.1 Introduction

Inferring species trees is challenging in the many situations when gene trees appear to differ from one another (Knowles, 2009). Whereas some gene-to-gene discordance may be due to stochastic error (e.g., incorrect estimation of gene trees) or technical issues (e.g., incorrect detection of true orthology), incongruent gene trees often reflect different underlying evolutionary histories (Maddison, 1997; Wendel and Doyle, 1998). One potential cause of incongruence is incomplete lineage sorting (ILS), which typically affects lineages across recently diverged species. ILS can also be retained along deep branches when these branches separate speciation events by few generations or very large populations (Takahashi et al., 2001). Empirical evidence of ILS has been documented in various organisms (e.g., Carstens and Knowles, 2007; Ebersberger et al., 2007; Heckman et al., 2007). Other sources of genealogical discordance include introgressive hybridization and horizontal gene transfer (HGT). HGT events are especially frequent in unicellular and prokaryotic organisms (e.g., Zhaxybayeva et al., 2006; Galtier, 2007) and can also occur in eukaryotes (e.g., Machado and Hey, 2003; Richardson and Palmer, 2007; Loreto et al., 2008). Nevertheless, disagreement among gene trees has long been ignored in empirical studies. This was defended because concatenation of multiple genes often leads to a single tree with high support values (Rokas et al., 2003). However, the extensive role of HGT in bacterial evolution raised some doubt that a single tree could faithfully represent the bacterial

species genealogical relationships (Doolittle and Bapteste, 2007; Galtier and Daubin, 2008). Furthermore, it is now recognized that concatenation can give misleading results, such as high support for an incorrect species tree, even when ILS is the only source of conflict (Kubatko and Degnan, 2007).

A new paradigm has now emerged, where species trees are considered separately from gene trees. New methods recognize that gene alignments do not bear on species trees directly. Instead, these methods are based on the premise that species trees affect gene trees, which then affect sequence alignments. Various parsimony based methods are available for reconciling gene trees and species trees. GeneTree (Page, 1998), DupTree (Wehe et al., 2008), and NOTUNG (Chen et al., 2000) aim to minimize the number of gene duplications/losses needed to map the gene trees onto the species tree. Alternatively, trees can be reconciled by minimizing the number of deep coalescences. This parsimony criterion is implemented in GeneTree, minimize deep coalescences (MDC) in Mesquite (Maddison, 1997; Maddison and Knowles, 2006), accommodating uncertainty in genealogies while inferring species trees (AUGIST) in Mesquite (Oliver, 2008), and MDC in PhyloNet (Than and Nakhleh, 2009), which is guaranteed to find the most parsimonious solution. Species tree estimation using maximum likelihood (STEM; Kubatko et al., 2009) is a model-based method. STEM assumes that discordance among gene trees is only caused by ILS, as modeled by the coalescent process (Kingman, 2000). Because gene trees serve as input, STEM is computationally very tractable. It has recently been expanded to test for putative hybrid speciation events placed on a known phylogeny, in the presence of ILS (Kubatko et al., 2009; Meng and Kubatko, 2009). Except for AUGIST, the methods cited above take in as input a single tree for each gene and assume that these individual gene trees are

inferred without error.

Various Bayesian methods use sequence alignments as input and treat gene trees separately from species trees. Bayesian estimation of species trees (BEST; Edwards et al., 2007; Liu et al., 2008) and *BEAST (Heled and Drummond, 2010) both assume the multispecies coalescent model. Bayesian concordance analysis (BCA), implemented in BUCKy (Ané et al., 2007), similarly uses a Bayesian approach to distinguish between gene tree estimation error and gene tree incongruence. BUCKy uses a nonparametric approach to modeling gene tree discordance, making no assumptions about the reasons for discordance. These model-based methods have been applied successfully to analyze data sets with extensive incongruence among genes (e.g., Maddison and Knowles, 2006; Carstens and Knowles, 2007; Belfiore et al., 2008; Brumfield et al., 2008; Horvath et al., 2008; Cranston et al., 2009; Leaché, 2009; Rodriguez et al., 2009; White et al., 2009). In particular, Cranston et al. (2009) compared the utility of BEST and BUCKy on a genome-wide real data set.

The primary goal of this paper is to provide the first thorough simulation-based comparison between two of these methods, BEST and BUCKy. We chose to compare these Bayesian methods because both fully account for uncertainty in individual gene tree estimation. We did not compare them with the concatenation approach because other authors have previously shown that concatenation, when consistent, is typically more powerful than supertree or consensus methods, at the cost of returning inflated support values and of being inconsistent in some cases (Kubatko and Degnan, 2007; Kubatko et al., 2009; DeGiorgio and Degnan, 2010). BEST and BUCKy may both lose power to recover the true vertical signal compared with the concatenation approach but for the sake of returning reliable statistical support values.

Both BEST and BUCKy assume free recombination between genes and no recombination within genes. The primary difference between them lies in their prior distribution on gene trees. BEST uses a uniform prior distribution on species trees with a coalescent model for gene trees given the species tree. BUCKy uses a nonparametric Dirichlet process to cluster genes into groups that share the same topology. The Dirichlet process uses an a priori level of discordance , conveniently bridging concatenation methods ($\alpha = 0$) with consensus/supertree approaches ($\alpha = \infty$). Any intermediate value of combines the information in the sequences of those genes that are inferred to be congruent. Finally, BUCKy estimates the primary concordance tree, whose clades are inferred for the largest proportion of loci in the genome (Baum, 2007).

We conducted simulations to assess the accuracy of the species tree inferred by BEST and the accuracy of the concordance tree inferred by BUCKy as an estimate of the true (known) species tree. Our simulation of gene trees included two processes of discordance: ILS and HGT. Because the coalescent process is used to model gene tree discordance in BEST, it was expected that BEST would return more accurate species trees than BUCKy when ILS was the only source of discordance. Similarly, when HGT was a significant source of discordance, thereby violating the assumptions of BEST, it was expected that the nonparametric method BUCKy would provide more accurately estimated species trees.

The second goal of this paper is to assess the accuracy of BUCKy at estimating concordance factors (CFs). CFs are measures of genomic support. The CF of a clade is the proportion of genes that truly have the clade in their trees. Ideally, one would like to know the genome-wide CF, that is, the proportion of genes in the entire genome

that truly have the clade. In practice, we need to first consider the sample-wide CF of the clade, that is, the proportion of genes in the available sample that truly have the clade. If sequence data are available for just a handful of genes, then inferring the genome-wide CF from the observed sample-wide CF may be no easy task. For instance, if only a sample of four genes is available, and exactly two of them are shown to have the clade, then it is still unclear whether more or less than 50% of the genome truly has the clade.

The genomic measure of support provided by CFs is fundamentally different from the statistical support provided by the typical bootstrap values or posterior probabilities of clades. For example, consider the 5-taxon species tree in Figure 1c. The shortest branch in this tree defines the clade formed by taxa 1 and 2. Its length of 0.4 coalescent units means that the number of generations is 0.4 times the effective population size along this branch. The true CF for this branch is 0.46, that is, only 46% of genes in the genome truly have clade (1,2). The other 54% of genes do not. This low genomic support is due to the high level of ILS along this branch. Ideally, we would want to infer the truly low genomic support for clade (1,2) and at the same time get high statistical support that this clade indeed belongs to the species tree. Statistical support for a clade to be in the species tree is obtained in BUCKy by comparing the CF of the clade with that of conflicting clades. For instance, if there is a 1.0 posterior probability that more genes have clade (1,2) than any other conflicting clade, then the statistical support for clade (1,2) to be in the species tree is 1.0. Along with this very high statistical support, the CF of clade (1,2) might be estimated to be between 0.44 and 0.48 with 99% credibility, recognizing the fact that the sister relationship of taxa 1 and 2 is not true for all the genes.

The third goal of this work is to assess the utility of BUCKy at testing the null hypothesis that the coalescent model provides an adequate explanation of the observed gene tree discordance. The tests alternative hypothesis is that some other biological process(es) contributed to gene tree discordance, along with ILS. This alternative hypothesis includes models with HGT, for example, or models with population structure where the panmictic assumption of the coalescent is violated. Because the test is not designed to focus on a specific kind of alternative model, it is called an "omnibus" test. It takes advantage of the signature left by the coalescent on CFs and aims to detect departure from this signature. The coalescent model completely determines the CF of each clade in the species tree and of each clade contradicting the species tree, based on the hypothesized population sizes and number of generations between speciation events. On 4 taxa, there are two clades that conflict with the clade in the species tree. Under the coalescent model, these two minor clades are expected to be true for a minority of the genome and to have equal CFs. If these two minor clades have significantly different CFs, then we can reject the coalescent model and infer the presence of some other source of discordance (Degnan and Rosenberg, 2009). For example, if one species is a hybrid, then two clades would have relatively high CF. These clades, placing the hybrid species with either of its parents, would each be expected to have a CF near 0.50 if the two parental lineages contributed equally to the hybrid genome. The third clade is expected to have a CF near 0, far below that of the other two clades, resulting in a pattern almost opposite to the pattern expected under the coalescent. The presence of population subdivision in ancestral species-violating the random mating assumption of the coalescent-was also shown to cause the two minor clades' CFs to differ (Slatkin and Pollack, 2008). Note that BEST cannot be directly

used to test the hypothesis that the coalescent model is correct because it assumes this coalescent model in the first place. Alternatively, BUCKy infers CFs in a way that is not constrained by ILS. CF estimates can therefore be used to test the signature of the coalescent model (Ané, 2010). In this paper, we assess the accuracy of BUCKy at estimating CFs and at testing the adequacy of the coalescent model in realistic situations when gene discordance is due to both ILS and HGT.

The concordance tree estimated by BUCKy represents the primary features in the history of a set of taxa, so that it can be used on organisms for which the concept of a species tree is controversial. Even though BUCKy does not involve an explicit population model on a species tree, it was shown by Degnan et al. (2009) that under the coalescent model, the species tree is fully recoverable from the true CFs (see also Ané, 2010). Degnan et al. (2009) showed that the concordance tree built from the 3-taxon rooted (or 4-taxon unrooted) subtrees with highest CFs provides a consistent estimate of the species tree. Therefore, there is a tight link between concordance trees and species trees when the concept of a species tree is applicable and under the coalescent. Degnan et al. (2009) also showed that the estimated concordance tree built from high-CF clades on the full-taxon set (rather than on 4-taxon sets) may not always provide a consistent estimate of the species tree. Unfortunately, the version of BUCKy that is tested in this paper uses clades on the full-taxon set to build the estimated concordance tree (but see BUCKy version 1.4.0; Larget et al., 2010). Therefore, we conducted our study outside the "too-greedy zone" identified by Degnan et al. (2009), where the current version of BUCKy is expected to be consistent.

2.2 Methods

2.2.1 Simulation of Multilocus Alignments with Gene Tree Discordance

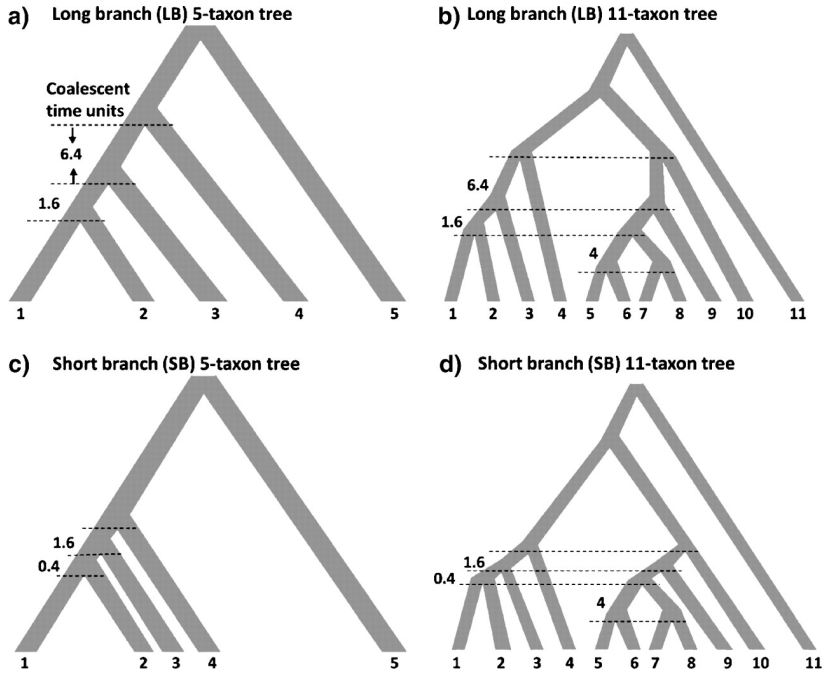

Figure 2.1: True species trees used in simulations. Top: trees with LB. Bottom: trees with SB. Left: 5-taxon trees. Right: 11-taxon trees. Numbers on left-hand side indicate coalescent units (number of generations on a branch divided by effective population size). The effective population size is 50,000 and does not change through time.

We generated DNA alignments from 5-taxon and 11-taxon species trees, as shown in Figure 2.1. An asymmetric tree topology was chosen on 5 taxa, as this was proven to be more difficult to reconstruct in the presence of gene-to-gene discordance Kubatko and Degnan (2007). Our 11-taxon tree contains two copies of our 5-taxon tree (subtree with taxa 1, 2, 3, 4 and subtree with taxa 5, 7, 9, 10, both with taxon 11 as an outgroup). In one of the two copies, taxa 6 and 8 (Fig. 2.1b,d) were added in order to detect potential effects of the number of taxa on the estimation of internal edges' CFs.

For each species tree topology, two sets of branch lengths were considered. One set had long internal branches (LB, top panels in Fig. 2.1), whereas the other set had some short internal branches (SB, bottom panels in Fig. 2.1). Species tree branch lengths were measured in coalescent units (Fig. 2.1), as obtained by dividing the number of generations by the effective population size. Under the coalescent model, branch lengths in coalescent units determine the proportion of genes that share the species tree topology and the proportion of genes that have any given conflicting topology.

In order to simulate multilocus data sets, 10, 50, or 100 unlinked gene trees were generated along the species trees in Figure 2.1 using Serial SimCoal (Anderson et al., 2005). Serial SimCoal is an extension of SIMCOAL (Excoffier et al., 2000), based on the retrospective coalescent approach. We used an effective size of 50,000 haploid individuals in each population. The numbers of generations between speciations were determined by multiplying branch lengths in coalescent units by the population size. Next, a number of HGT events and rate change events were generated on each gene tree, from which 500-bp-long DNA alignments were generated using an extension of HGTsimul (Galtier, 2007). Within this program, the number of HGT events placed on gene trees was either set to 0 or to a Poisson-distributed number with an average of 0.5 HGT events per gene. In the former case, ILS was the only source of discordance, and in the latter case, both ILS and HGT were causing gene tree discordance. In order to better compare the results with and without HGT, the same ILS simulations from Serial SimCoal were used with and without HGT. For each HGT event simulated by HGTsimul, the recipient lineage was randomly and evenly placed on the gene tree, with branches weighted by their branch lengths. The placement of the donor lineage was randomly drawn from the locations contemporary with the recipient lo-

cation. Note that a fair number of these HGT events did not actually change the tree topology. Thus, in order to control for the level of HGT, multilocus data sets were generated so that exactly 70% of the gene topologies were not affected by the simulated HGT events. In order to achieve an exact proportion of 70%, the following procedure was used iteratively. If more than 70% of gene tree topologies were unaffected by HGT, the affected gene trees were saved and the unaffected gene trees were rejected and resimulated. This process was repeated until the proportion of gene topologies unaffected by HGT dropped to 70% or less. If this proportion dropped below 70%, the unaffected genes were saved and each remaining gene was resimulated until its topology was unaffected by HGT.

HGTsimul was used to simulate a Poisson-distributed number of genomic rate change events (with a mean of three changes) on the species tree, for genomic departure from the molecular clock. Lineage-specific rates were simulated from a gamma distribution with mean 1 and shape parameter 2.0. For each gene, branch lengths obtained from Serial SimCoal were multiplied by these lineage-specific rates, then further multiplied by a common factor to obtain a randomly chosen gene diameter (uniform in 0.024 and 0.037 substitutions per site). Next, gene tree branch lengths were modified in a gene-specific manner: for each individual gene, a Poisson-distributed number of rate change events (three changes on average) were placed on the gene tree, whose branch lengths were multiplied by a gamma-distributed rate (mean 1 and shape parameter 2.0) in between these gene-specific rate change events. Finally, sequences were simulated using the Jukes-Cantor (JC) model Jukes and Cantor (1969) and no site-specific rate variation, for computational feasibility (for full details, see Galtier, 2007). In summary, our simulations included important factors that contribute to hetero-

ILS	HGT	5-taxon case		11-taxon case	
		Proportion	RF distance	Proportion	RF distance
Weak (LB)	No	0.14	1.98	0.28	2.19
	Yes	0.32	2.67	0.49	4.32
Strong (SB)	No	0.54	2.34	0.79	3.26
	Yes	0.66	2.73	0.84	4.41

Table 2.1: Discordance between simulated gene trees and species trees, in terms of the proportion of gene trees whose topology differed from the species topology and of the average RF distance between gene trees and species trees

geneity among genes, such as heterogeneity in the overall rate of evolution, departure from clock-like evolution, and topological discordance.

A total of 24 conditions were considered (5 or 11 taxa, LB or SB species tree, HGT present or absent, 10, 50, or 100 genes). For each condition, 100 replicate data sets were generated and analyzed. Additional simulations were carried out with longer DNA alignments (1000 bp instead of 500 bp) on the SB species trees, HGT present and 100 genes, so that estimation of CFs by BUCKy on DNA alignments of 500 and 1000 bp could be compared.

The level of discordance between the simulated gene trees and species trees is summarized in Table 2.1 for each type of discordance. On LB species trees, HGT was considered to be the major source of discordance, whereas strong ILS on short branches resulted in the highest levels of discordance.

A separate set of simulations was carried out in which HGT events were unevenly placed on the 5-taxon species tree. In this second simulation, all HGT events were set to originate from the edge leading to taxon 2 (Fig. 2.1). The exact event location was randomly placed along the edge. The recipient lineage was set to be the edge leading to taxon 3, 4, or 5. Each of these three types of HGT events happened on

10% of the genes, so that the overall proportion of genes whose topology was affected by HGT was maintained at 30%, like in the first simulation. ILS simulation and sequence alignment simulation were then carried out using Serial SimCoal followed by our modified HGTsimul program. Unevenly placed HGT events were thus simulated in an extra six conditions (SB and LB 5-taxon trees; 10, 50, or 100 gene sets).

2.2.2 Analysis of Simulated Alignments

BEST (version 2.2) was applied to multilocus data sets under the JC model. On all 10-gene and 50-gene data sets, 1 million generations were used for 5-taxon alignments and 3 million generations for 11-taxon alignments. For 100-gene data sets, BEST was run on a subset of the 100 replicates from each condition. On 100-gene 5-taxon alignments, 20 replicates from each condition were analyzed with 10 million generations. On 100-gene 11-taxon alignments, 10 replicates from each condition were analyzed with BEST for 1 month each, reaching an average of 9.6 million generations. All runs appeared to reach convergence. For all BEST analyses, 1 cold and 3 heated chains were used, with four simultaneous independent analyses starting from different random trees. The chain was sampled every 100th generation, and the first 10% of generations were discarded as burn-in. The default "poissonmean = 5" was used on 11-taxon alignments. This parameter determines how much the maximum tree is modified to propose a new species tree. A smaller value was chosen on 5-taxon alignments (poissonmean = 3) to increase the acceptance rate (Liu et al., 2008). Other parameters were set to their default values.

The Bayesian concordance analysis was performed by first applying MrBayes (Huelsenbeck and Ronquist, 2001) to each individual gene. The JC model was used,

with 4 chains, 4 independent runs, and 500,000 generations on 5 taxa and 1 million generations on 11 taxa. Gene trees were sampled every 100th generation, and the first 10% of generation were discarded. The results from multiple genes were then analyzed with BUCKy (version 1.3), using 4 chains and 4 independent runs of 1 million generations on 5 taxa and 2 million generations on 11 taxa. The a priori level of discordance among loci was set to $= 2.5$, so that the prior distribution of the number of distinct gene trees (mean 7.25) was intermediate between the various empirical distributions from the known true gene trees: on 100 gene trees and 5 taxa, the mean number of distinct gene trees ranged from 3.16 (LB tree, absence of HGT) to 8.73 (SB tree, absence of HGT) to 9.71 (LB tree with HGT) to 12.61 (SB tree with HGT).

To compare the performance of BEST and BUCKy, we calculated the Robinson-Foulds (RF) distance (Robinson and Foulds, 1981) between the true species tree and the species tree estimated by BEST or the concordance tree estimated by BUCKy. In the presence of HGT, the true species tree in Figure 2.1 represents the true vertical inheritance pattern, so that a low RF distance from this tree indicates a good estimation of the vertical signal. We averaged these RF distances over all replicates from each condition. To determine if the observed differences between the performances of BEST and BUCKy were significant, two-sided pair-wise Wilcoxon signed-rank tests were conducted at each of the three conditions: absence of HGT, even presence of HGT, and uneven presence of HGT.

To assess the estimation of CFs by BUCKy, the difference between the estimated sample-wide CFs and the true sample-wide CFs was calculated on each data set and for each clade. Note that different multilocus data sets could have different CFs for the

same clade, especially in the presence of HGT. The proportion of credibility intervals including the true sample-wide CFs was considered. In the case of 100 genes and strong ILS (SB), DNA alignments with 1000 bp were also simulated and the accuracy of their estimated sample-wide CFs was compared with that from the 500-bp alignments.

We used genome-wide CFs and their credibility intervals to test the adequacy of the coalescent model. Under the coalescent on the 5-taxon species tree topology shown in Figure 2.1, the true genome-wide CFs obey the following properties for all sets of branch lengths (Pamilo and Nei, 1988):

$$\begin{aligned} \text{CF for clade } 13 &= \text{CF for clade}23, \\ \text{CF for clade } 14 &= \text{CF for clade}24, \\ \text{CF for clade } 15 &= \text{CF for clade}25 \end{aligned} \tag{2.1}$$

Although the exact values of these CFs are determined by the species tree branch lengths (in coalescent units), these equalities among CFs are a signature of the coalescent process. To test the adequacy of the coalescent model, we compared the 95% credibility interval for the CF of clade 13 with that of clade 23. If these two intervals did not overlap, we rejected the null hypothesis. Similarly, we rejected the adequacy of coalescent model if the credibility intervals for the CFs of clades 14 and 24 did not overlap or if the credibility intervals for the CFs of clades 15 and 25 did not overlap. In cases when at least one of the three pairs of credibility intervals did not overlap, it was inferred that a source of discordance other than the coalescent should be invoked to explain the discordance in the data.

The test described above assumes that the true sister relationship of taxa 1 and 2 is known without error. In practice, this sister relationship may be estimated from

the same data used for testing the adequacy of the coalescent. To account for possible errors in the species tree estimation, we carried out a second test procedure. If clade 12 was indeed inferred to form a clade in the species tree, the test was carried out as described above from Equation 1. If clade 12 was not inferred to be in the species tree, then taxa 1 and 2 were replaced in Equation 2.1 by a pair of taxa inferred to be sister. If clade 13 or clade 23 was inferred to be in the species tree, then this pair of taxa was used as a substitute for 12 in Equation 2.1. For example, if taxa 1 and 3 were inferred to be sister, then we compared the 95% credibility intervals for the genome-wide CFs of 12 versus 23, 14 versus 34, and 15 versus 35. In case none of the clades 12, 13, or 23 was inferred to belong to the species tree, then taxon 4 was necessarily inferred to be sister to another taxon, and this pair was used as a substitute for 12 in Equation 2.1. Overall, this case occurred in only 11 of the 1800 replicates.

In our first simulation with evenly distributed HGT events, the equalities in Equation 2.1 were still expected to hold genome wide. For instance, HGT transfers from taxon 1 to taxon 3 were simulated with the same frequency as transfers from taxon 2 to taxon 3, so that the symmetric signature of the coalescent model was expected to be maintained in the presence of HGT in our first simulation. As a result, we carried out a second simulation with unevenly placed HGT events. The power of the test for the adequacy of the coalescent model was calculated as the proportion of rejections among each set of 100 replicates.

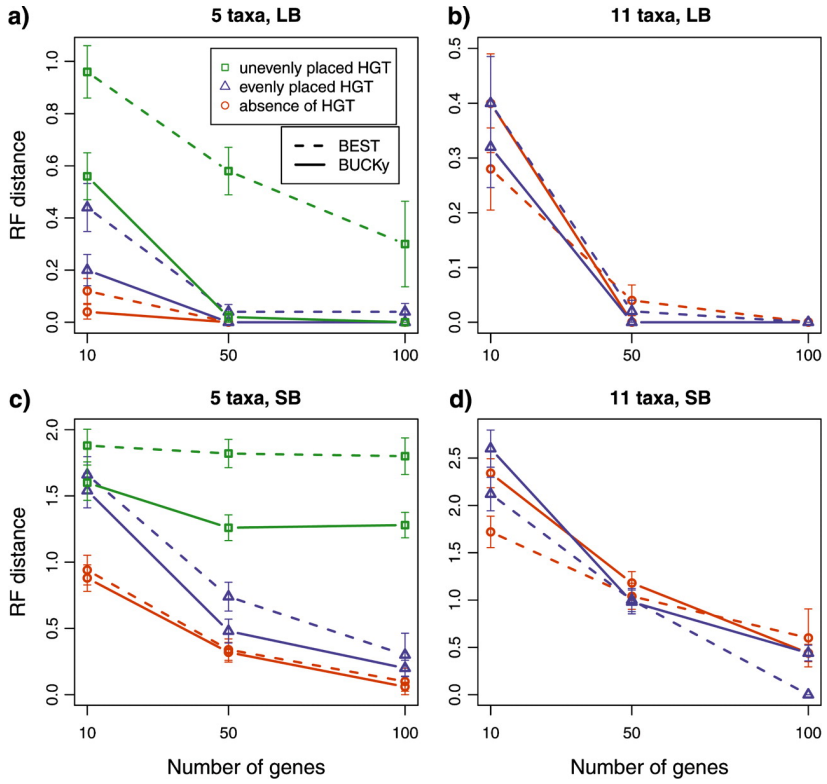

Figure 2.2: Performance of BEST and BUCKy for species tree estimation, as measured by the RF distance between the true species tree and the species tree estimated by BEST (- - -) or the concordance tree estimated by BUCKy (−). Each point represents the average across 100 simulated multilocus data sets except on 100-gene alignments analyzed with BEST. For these cases, each point represents the average of 20 replicates on 5 taxa and of 10 replicates on 11 taxa. Bars indicate standard errors. Note that with only 10 replicates (on 100 genes, 11 taxa), the standard error shown here is a poor reflection of the sampling error when all observed distances are 0, due to the discrete and skewed distribution of RF distances. Simulations included unevenly placed HGT events (green □), evenly placed HGT events (blue $\bigtriangleup$), or absence of HGT events (red ○).

2.3 Results

2.3.1 Comparison between BEST and BUCKy

Even though BEST and BUCKy do not estimate the same quantities, both account for gene tree discordance and estimate a tree to describe the history of the sample. BEST estimates the species tree under the coalescent-only model, whereas BUCKy estimates a concordance tree, featuring clades supported by the largest proportions of the genes. These two tree estimates were compared with the true species tree using the RF distance (Fig. 2.2). When HGT events were unevenly placed, BUCKy performed better than BEST in all 10- and 50-gene cases (P values pair-wise Wilcoxon signed-rank tests ranged from ¡0.0001 to 0.0167). From 100 genes, BUCKy showed a better performance than BEST, but the difference was marginally or not significant (P = 0.037 on SB tree and P= 0.14 on LB tree) due to the small sample size (20 replicates) and to the presence of ties (the concordance tree estimated by BUCKy had a 0 RF distance from the species tree in many instances), undermining the reliability of the Wilcoxon test. In the absence of HGT or in the presence of evenly placed HGT, BEST and BUCKy performed similarly on average (as determined by the Wilcoxon test), except in three cases. In the presence of strong ILS (absence of HGT on SB trees), BEST performed better than BUCKy on 11 taxa, 10 genes (P value 0.0032). In the presence of HGT with evenly placed events, BUCKy performed better than BEST on 5 taxa, 10 genes when the major source of discordance was HGT (LB tree, P value 0.0031), and on 5 taxa, 50 genes when discordance was equally due of ILS and HGT (SB tree, P value 0.0231). Both methods generally performed better

Tree	Clades in the true species tree	Clades not in the true species tree
5-taxon tree	12; 13	13; 14; 15;
		23; 24; 25; 34; 35
11-taxon tree	12; 13; 14;	13; 23; 14; 24;
	56; 78; 58;	14,11; 14,9,11;
	59	59; 69; 79; 89; 9,10;
		569; 789; 56,10; 78,10

Table 2.2: List of clades represented in Figures 2.32.5

as the number of genes increased and as the overall level of discordance decreased.

2.3.2 CF Estimation

Figures 2.32.5 summarize the accuracy of the estimated sample-wide CFs for all 10 possible clades in the 5-taxon tree and for a selected set of 22 clades in the 11-taxon tree (Table 2.2). To assess accuracy, we measured the difference between the estimated and true CFs and the proportion of times that the 95% credibility intervals included the true CFs. In a Bayesian framework, credibility intervals are used to summarize posterior distributions. They are influenced by the choice of a prior distribution and are not designed to cover the true value at a nominal (95%) level. Nevertheless, we were interested in evaluating the coverage level of credibility intervals here because it is unknown how well BUCKy's Dirichlet-based prior distribution approximates the biological process used to generate gene trees in our simulations.

Estimated CFs on 5 taxa (Fig. 2.3) and 11 taxa (Fig. 2.4) were generally accurate, although they became less so as the true level of discordance increases from low ILS and HGT presence (Figs. 2.3 and 4, top) to high ILS and HGT presence (Figs. 2.3 and 2.4, bottom). The plots for low ILS and HGT absence are not included because their accuracy was better than that for low ILS and HGT presence. For most clades,

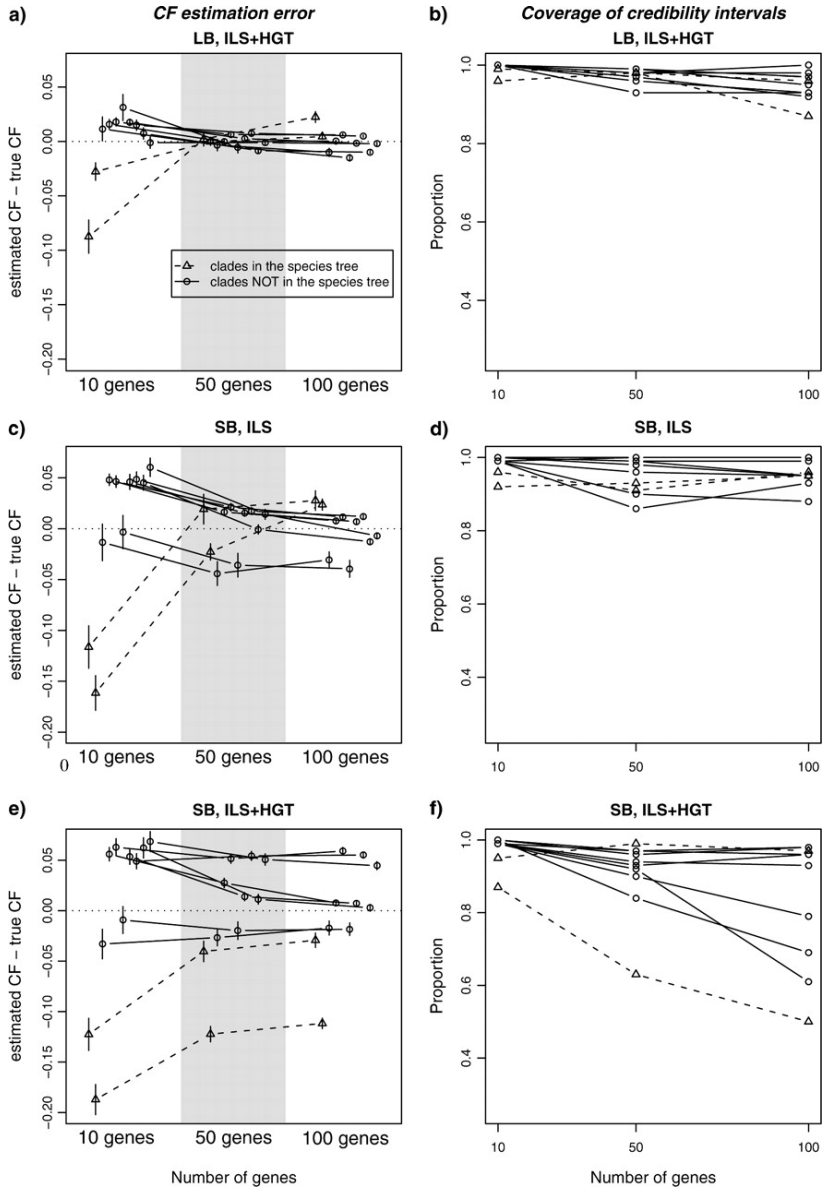

Figure 2.3: Accuracy of estimated CFs on the 5-taxon trees and for the 10 nontrivial bipartitions, in the absence/presence of evenly placed HGT. Left: difference between estimated CFs and true CFs, averaged over 100 replicates. Points were horizontally scattered around the corresponding number of genes to avoid overlap. Bars indicate standard errors. Right: proportion of credibility intervals containing the true CF, for the same 10 bipartitions. Results from the same bipartition are joined by lines ($-\circ-$), except for clades 12 and 123 ($\cdots\circ\cdots$), which belong to the true species tree.

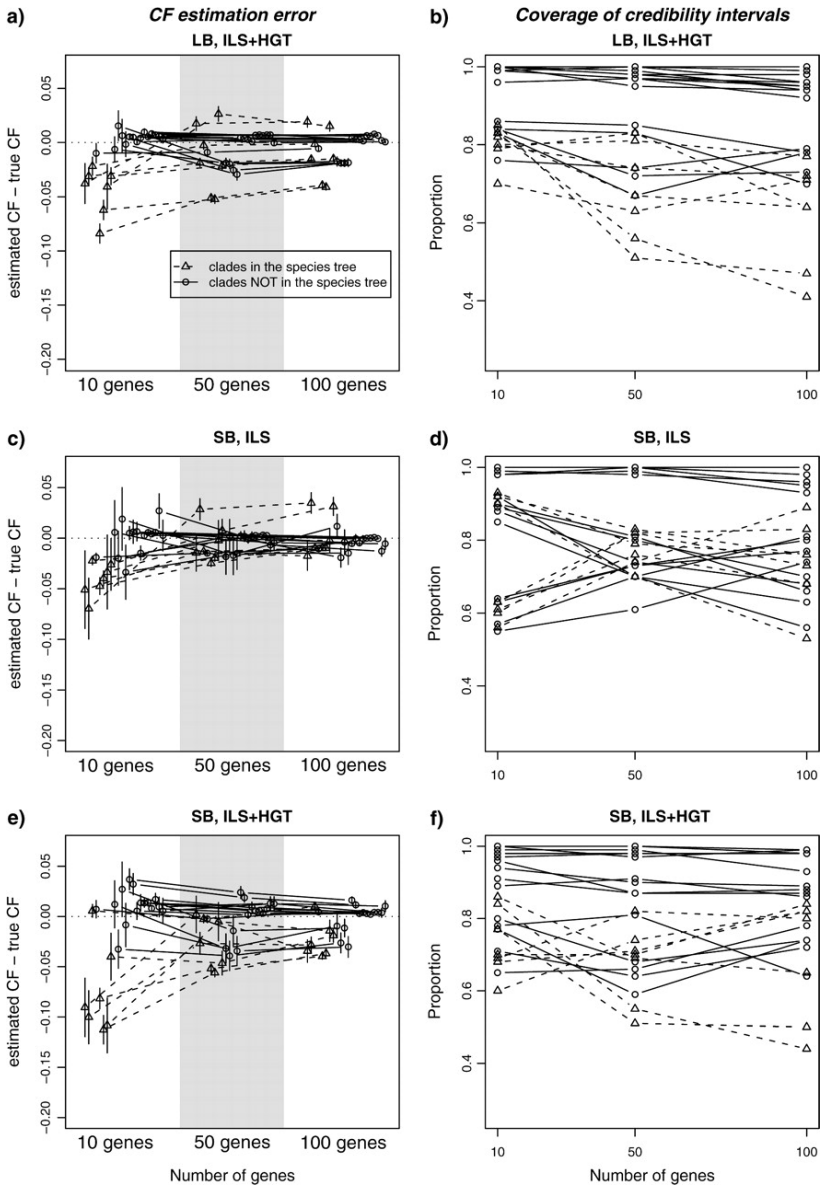

Figure 2.4: Accuracy of estimated CFs on the 11-taxon tree and for the 22 bipartitions listed in Table 2.2, in the absence/presence of evenly placed HGT. Left: average difference between estimated CFs and true CFs. Points with bars were horizontally scattered around the corresponding number of genes to minimize overlap. Bars indicate standard errors. Right: proportion of credibility intervals containing the true CF. Results from the same bipartition are joined by lines ($-\circ-$), except for the seven clades that truly belong to the species tree ($\cdots\circ\cdots$).

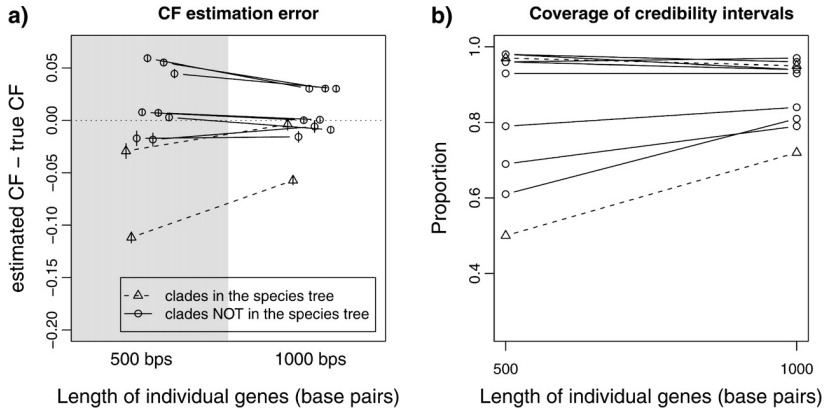

Figure 2.5: Accuracy of estimated CFs as a function of individual gene length, from 100-gene alignments on the 5-taxon SB tree (high ILS) in the presence of evenly placed HGT events. Left (a) and Right (b) as in Figure 2.3.

the estimated CFs became less biased as the number of genes increased. Estimated CFs appeared to be slightly more accurate on average on 11 taxa than on 5 taxa. In all cases however, unexpectedly small proportions of credibility intervals contained their true CFs, especially on 11 taxa. Even more surprisingly, the proportion of times the true CF belonged to the CF's credibility interval tended to decrease with the number of genes. One of the most extreme cases is that of clade 123 on 5 taxa, represented by the lowest dashed line in Figure 2.3e,f. The credibility interval for this clade's CF missed the true CF value in 50% of the replicates in the presence of strong ILS and HGT. On 11 taxa (Fig. 2.4, right), the credibility intervals contained their true CFs in only 44% of the replicates for some clades.

We conducted additional simulations to determine the effect of increasing the amount of data by doubling the length of each gene, from 500 to 1000 bp. We focused on the cases with strong discordance (SB tree in the presence of HGT) and 100 genes. On 5 taxa, increasing the length of individual genes provided a substantial improvement in the accuracy of CFs (lower bias) and of their credibility intervals (Fig. 2.5).

In particular, the coverage of the true CF for clade 123 increased from 50% to 72%. On 11 taxa, longer sequences provided smaller improvements (data not shown).

2.3.3 Omnibus Test of the Adequacy of the Coalescent Model

Credibility intervals of genome-wide CFs were used to test the null hypothesis that the coalescent model alone can explain the discordance among gene trees, using the symmetric signature of the coalescent. The results are shown in Table 2.3. When the null model was correct, the null hypothesis was rejected in very few cases, showing a low Type I error rate. This error rate was 4% or lower in all but one case. In the presence of strong ILS on 100 genes and when the test was based on sample-wide CFs, a 7% Type I error rate was observed. Using genome-wide CFs with wider credibility intervals resulted in a more conservative test (lower Type I error) than using sample-wide CFs. In the first simulation when HGT was present and evenly distributed, the test based on genome-wide CFs had no power, with a rejection rate below 3%. This result was expected because the symmetry of genome-wide CFs was preserved by an even distribution of the HGT events. The test based on sample-wide CFs, however, had some power to detect the presence of HGT: between 13% and 18% when ILS was strong and between 43% and 52% when ILS was weak. In the second simulation when HGT events were unevenly placed along the tree, the test based on genome-wide CFs showed a 99% power to detect the presence of HGT from 100 genes when ILS was weak and had little power otherwise. The test based on sample-wide CFs had moderate power in the presence of strong ILS (12% to 58% rejections) but high power when ILS was weak (80% to 100% rejections). These results were almost unchanged when the inferred species tree was used instead of the true species tree. When a

difference was observed, the proportion of rejections decreased very slightly, resulting in a slightly more conservative and slightly less powerful test.

2.4 Discussion and Conclusion

2.4.1 Power Assessment and Comparison

BEST and BUCKy are two Bayesian methods that aim to reconstruct a species or concordance tree from a set of potentially discordant genes. These methods recognize that discordance among gene trees can be real. Both account for two levels of uncertainty: uncertainty at the gene tree level due to having a limited number of gene trees to reconstruct the species tree and uncertainty at the molecular level due to having a limited number of nucleotide substitutions to reconstruct gene trees. The concatenation approach accounts for uncertainty at the molecular level but assumes no variability or no uncertainty at the gene tree level. By recognizing that only a sample of genes bear on the species tree reconstruction, BEST and BUCKy typically return lower support values for relationships in the species tree compared with the potentially misleading 100% bootstrap values often obtained from the concatenation approach (Kubatko and Degnan, 2007; Huang and Knowles, 2009; Liu et al., 2009). The decrease in support values can be especially striking when few genes are sampled and when these gene trees show strong discordance. Although recognizing the reality of gene tree variability is theoretically highly valuable, a legitimate concern is that of power. In practice, one wants to recover the species tree with high confidence from the available data. This paper takes a step at assessing the power of two gene tree reconciliation methods.

CFs used for testing	Number of genes	No HGT		Evenly distributed HGT		Unevenly distributed HGT	
		Weak ILS	Strong ILS	Weak ILS	Strong ILS	Weak ILS	Strong ILS
Genome wide	10	0	0	0	0	0	0
	50	0	0.01	0.02	0.02 (0.01)	0.04	0.05 (0.04)
	100	0	0.04	0	0.03 (0.02)	0.99	0.10 (0.09)
Sample wide	10	0	0.01	0.46 (0.45)	0.15 (0.14)	0.8	0.12
	50	0.02	0.04 (0.03)	0.52	0.18 (0.17)	1	0.54
	100	0 (0.03)	0.07	0.43	0.15 (0.13)	1	0.58 (0.52)

Table 2.3: Proportion of rejections when conducting the test for the adequacy of the coalescent model, based on Equation 2.1 when the true species tree is known

BEST and BUCKy use very different prior distributions on gene trees. Because BEST is based on the coalescent process along the species tree and BUCKy is based on a Dirichlet prior to cluster genes that have concordant topologies, we expected these two methods to perform well under different conditions. In the situation when discordance was due to ILS only, BEST was expected to reconstruct accurate species trees and BUCKy was expected to be less powerful. In the presence of HGT however, BUCKy was expected to be robust and more accurate than BEST for which the coalescent assumption was violated. Our initial expectations held only when HGT events were unevenly distributed along the species tree, in which case BUCKy was always more accurate than BEST, even in the presence of deep coalescence. The two methods proved to have surprisingly similar accuracies however, in cases when HGT events, if present, were evenly distributed across the species tree and when the Markov chain Monte Carlo (MCMC) algorithm in BEST was permitted to converge. BEST was only slightly more accurate than BUCKy when ILS was the sole cause of discordance. BUCKy was significantly more accurate than BEST only when there were few genes, few taxa, and when the HGT was the major contribution to gene tree discordance.

One basic difference that was not considered here between BEST and BUCKy is their treatment of multiple individuals from the same species. Individuals need to be preassigned to species for analysis in BEST, which then estimates a species tree with the same number of tips as the number of species. BUCKy does not need any assignment of individuals to species and returns a concordance tree where each individual corresponds to a tip. In our study, only one individual per species was simulated because comparing BEST and BUCKy on multiple individuals was beyond

the scope of this work. The two methods have different goals. BEST focuses on discovering the species relationship. Leaché (2009) showed that BEST is sensitive to incorrect assignment of individuals to species, as may happen when species limits are estimated from geography or mitochondrial DNA genealogy. BUCKy focuses on discovering groups of taxa supported with high concordance among genes. Species are expected to result in such groups, so that concordance analysis may be used as a tool to help assign individuals to species, based on all the available molecular data (Baum, 2007).

2.4.2 Computing Time and Mixing Issues

The number of 100-gene replicates that could be analyzed with BEST in our simulation study was limited by the computing time needed for the MCMC algorithm to converge. On 100-gene data sets, the BEST analysis of each replicate took an average of 5.8 days on 5 taxa (10 million generations) and 30 days on 11 taxa (9.6 million generations on average). In comparison, the computing time needed for the concordance analysis (MrBayes and BUCKy combined) was much more affordable, at 3.3 h on 5 taxa and 14.3 h on 11 taxa. On 100-gene data sets, when fewer generationsand a shorter running timewere used in BEST (1 million generations on 5 taxa and 3 million generations on 11 taxa), the resulting species tree estimates were quite inaccurate. Indeed, species trees estimated from 50 genes or even 10 genes were closer to the true species tree on average (results not shown), even though the computation time allocated to the short analysis of 100-gene sets was over 2 times and 10 times longer than the computation time allocated to the analysis of the 50-gene and 10-gene data sets. For this reason, we strongly recommend against using or publishing species trees

estimated by BEST when convergence failed (Cranston et al., 2009; Leaché, 2009). Our study shows that BUCKy is a good alternative when large numbers of genes are available because very accurate species trees were reconstructed in a low or reasonable amount of computing time.

2.4.3 Robustness of BEST to HGT Presence and to Clock Departure

The robustness of BEST that we observed in the presence of evenly placed HGT events is very encouraging. One possible explanation for this result is that simulated HGT events were randomly placed on the species tree with symmetric probabilities. For instance, the expected proportion of HGT transfers from taxon 1 to taxon 2 was equal to the expected proportion of HGT transfers in the opposition direction, from taxon 2 to taxon 1. Thus, some discordance caused by HGT in our simulations could be explained quite well by the coalescent process under the correct species tree. The extra discordance caused by HGT events might have been accounted for in BEST by somewhat overestimating population sizes or underestimating divergence times, while recovering the correct species tree. We looked at the branch lengths of two edges in the species tree estimated by BEST from 50 genes: the internal edges leading to clade (1,2) and clade (1,2,3) where most of the discordance occurred. The estimated population sizes along these edges were similar in the presence or absence of HGT. But in all situations, these edges had shorter estimated branch lengths in the presence of HGT, on average, than in the absence of HGT. Therefore, the discordance due to evenly distributed HGT might have been explained by a higher level of ILS. If HGT is favored in some directions more than in others, however, the discordance among genes may no longer be well explained by the coalescent process only. Indeed,

BEST's accuracy dropped in our second simulation when HGT events were unevenly distributed. In real data sets, it is unclear how much asymmetry is involved in HGT events or in other gene flow processes. Our results suggest that BEST will be robust to the presence of gene flow in some cases but maybe not in species groups where gene flow occurred preferentially in some directions more than in others.

Our simulations included two ways in which gene trees departed from the clock assumption in order to reflect the across-lineage rate variability that is found in most real interspecific data sets. BUCKy was expected to be robust to this rate variability because branch length information is not pooled across genes in BUCKy. However, BEST makes the assumption that each gene tree is clock like, and this is often violated in practice. Our simulations showed that the ad hoc rate smoothing performed in BEST worked very well as BEST was not affected by rate variability across lineages. Overall, the robustness of BEST to various departures from its assumptions was very positive in our simulation settings.

2.4.4 Accuracy of Estimated CFs in BUCKy

The nonparametric clustering used by BUCKy to group compatible genes is not based on a probabilistic model of a biological process, so the accuracy of the resulting estimated CFs is unknown. In our simulations, estimated CFs became less biased overall as more genes were available or as the amount of discordance among gene trees decreased. However, the credibility intervals for CFs showed very poor properties as the level of discordance went up: the true sample-wide CFs were not included in their estimated 95% credibility interval up to 50% of the time for some important clades, and this poor performance grew worse with the number of genes. With few

genes, credibility intervals were wide enough to include the true value of the CF most of the time. But as the number of genes increased, the credibility intervals became unreasonably short and no longer covered the true value with a large probability.

Several reasons could be at the heart of this poor performance. One such reason could be the inadequacy of the Dirichlet prior distribution at modeling gene tree discordance. This Dirichlet prior distribution differs from the true distribution of gene trees and may have an undue influence when many genes are not informative. Specifically, narrow credibility intervals for CFs could be caused by an inflated confidence in the estimated gene tree clustering. If the Dirichlet prior inadequacy is the reason for poor coverage of CFs, then increasing the amount of data could minimize the influence of the prior distribution and should lead to an increased accuracy. This is exactly what was observed: when longer genes were simulated, increased information on individual gene trees lead to less biased estimates of CFs and to more reliable credibility intervals of CFs.

We also investigated the adequacy of the Dirichlet prior with parameter $\alpha = 2.5$ as opposed to other choices of that parameter. The choice of a consensus approach with a priori independent gene trees ($\alpha = \infty$) resulted in a much more pronounced bias of CFs, even though the accuracy of the estimated species tree was almost unchanged (results not shown). With this consensus-like approach, information was not shared across compatible gene trees and uncertainty was confounded with discordance: CFs of clades in the true species tree tended to be strongly underestimated, whereas CFs of clades that truly had low CFs tended to be overestimated. The choice of $\alpha = 25$ or 250 did not affect the results qualitatively compared with $\alpha = 2.5$ (data not shown).

2.4.5 Impact of the Taxon Number in BUCKy

The two-step approach in BUCKy is very attractive computationally. The first step is conducted in MrBayes, when each gene is analyzed separately. In practice, these separate analyses are easily parallelized. From this first step, the whole posterior distribution of individual gene trees is retained, then used in the second step to determine if two genes are truly incompatible or not. The second step combines the separate analyses into a joint analysis of all genes and is typically much faster than the first step. The downside of this two-step approach is a concern about accuracy: any approximation error made in Step 1 is carried over into Step 2. We wanted to assess the impact of this error propagation by comparing results from 5 taxa and 11 taxa. With 5 taxa and only 15 possible unrooted trees, the posterior distribution for a gene tree consists in a list of 15 posterior probabilities, one for each topology. This is easily and accurately calculated in MrBayes for each gene. But for 11 taxa and 34, 459, and 425 possible topologies, a list of over 34 million posterior probabilities is required for each gene. This list is necessarily obtained with some estimation error from MrBayes. Therefore, we expect that propagation errors may affect BUCKy on 11 taxa but not on 5 taxa.

A second related issue may arise when the tree space is very large. The analysis in MrBayes of a gene with few informative sites may return an MCMC sample where each topology appears only once: the region of high posterior probability may be correctly identified, but this region may contain so many topologies that any representative sample is sparse. Two independent analyses of the same gene in MrBayes may each be so sparse that they may not overlap on a single topology, even though

they may correctly return approximately equal posterior probabilities of clades. The same phenomenon may occur for two genes that truly have the same tree. If their estimated posterior distributions do not overlap from Step 1, then in Step 2 BUCKy cannot recognize that these two genes are compatible and will not pool information from these two genes in order to obtain a more accurate estimate of their common tree. In effect, BUCKy would be forced to take a consensus approach without pooling information across genes, as if with the prior value $\alpha = \infty$. This issue is diagnosed easily by examining the tree files from MrBayes for sparseness. Alternatively, if the results of BUCKy are identical with $\alpha = \infty$ and $\alpha = 1$ for instance, this may indicate that samples from Step 1 are too sparse.

In summary, two issues may arise on a large tree space: potential propagation error and sparse MCMC samples that force BUCKy to use $\alpha = \infty$. By comparing BUCKy on 5 and 11 taxa, we attempted to determine the extent to which these two factors affect the concordance analysis. Very little differences were obtained: concordance tree estimates were accurate on both 5 and 11 taxa, the biases of estimated CFs were similar, and the coverage of their credibility intervals was also similaralthough better on 5 taxa. The only difference that we found was in the coverage improvement from analyzing longer genes. A large improvement was observed on 5 taxa, but the improvement was relatively disappointing on 11 taxa. The reason for this difference may come from propagation error mostly, because the sparseness of the individual gene tree distributions did not seem to be too severe in our 11-taxon simulations. The most likely topologies reached a posterior probability of 0.02 in over half the 500-bp genes (HGT and strong ILS), and such a topology would easily provide overlap between independent runs. However, a very large number of topologies were needed to describe

95% of the highposterior probability region: an average of 2594 for 500-bp genes and still 551 for 1000-bp genes. The propagation of errors in estimating the small posterior probability of each of these numerous topologies may explain why 1000-bp genes were not long enough in our setting to override the influence of the Dirichlet prior.

Our work suggests that the reliability of CF credibility intervals partly depends on the accuracy of gene tree posterior distributions from the first step of BUCKy. Adequate accuracy is expected on very low numbers of taxa or from highly informative genes. This constraint largely limits the data sets that can be reliably handled by BUCKy at this time.

2.4.6 Testing the Adequacy of the Coalescent Model Using CFs

We introduced an omnibus test for the adequacy of the coalescent model. The null hypothesis states that the coalescent model is sufficient to explain the observed gene tree discordance. The alternative hypothesis is that some other biological process(es) (e.g., HGT events, population structure not modeled by the coalescent) contributed to gene tree discordance along with ILS. On a 4-taxon set, the coalescent process implies that the two minor splits conflicting with the clade in the species tree have equal genome-wide CFs. Therefore, we used estimated genome-wide CFs in order to test the signature of the coalescent. Despite the low coverage of CF credibility intervals, the Type I error of the test was adequate. The test achieved high power when 100 genes were available and when HGT accounted for more incongruence compared with ILS. However, the power stayed low when HGT events were evenly distributed along the tree, probably because the equalities between CFs predicted by the coalescent were still approximately true.

Overall, the test was found to be conservative, with a Type I error rate well below 5% in most cases. Although 95% credibility intervals were used to test the equality of CFs, there was no theoretical basis to expect a 5% Type I error rate for two reasons. First, 95% credibility intervals have a different interpretation than confidence intervals, in that there is no expectation that they cover the true value in 95% of the experiments. Their coverage probability is expected to depend on the prior distribution. Second, we examined three equalities and rejected the null hypothesis as soon as one equality was rejected, with no correction for multiple testing. Such a correction would technically be difficult because the output from BUCKy only includes 95% and 99% credibility intervals. Despite this, the test showed an adequate Type I error rate.

Theoretically, the test should be applied to genome-wide CFs because the coalescent model predictions apply to genome-wide CFs. When we used sample-wide CFs instead, we found that the test still had appropriate Type I error rate and gained substantial power. These results are quite surprising, given the poor coverage properties of CF credibility intervals. They could be explained if the estimated CFs of minor clades were positively correlated, which would minimize the estimation error of their difference. Further studies should be conducted to confirm these results in other situations.

Although the test was based on CFs from 4 taxa and applied to 5 taxa, this test could be used on larger trees. In a bifurcating species tree, any given edge defines a set of quartets. For each of these quartets, the coalescent model predicts equal CFs for the two minor resolutions of the quartet. The credibility intervals of these two resolutions' CFs could then be compared. This approach could be helpful to test the

adequacy of the coalescent locally along specific edges. However, it seems difficult to generalize this test into a global test of the adequacy of the coalescent on large trees.

Several other methods have been developed to test the hypothesis that ILS is at the origin of all gene tree discordance. The supernetwork approach by Holland et al. (2008) builds a network from clades with highest CFs, filtering out splits based on combinatorial criteria. The ILS hypothesis is then accepted on each tree-like part of the network. Individual gene trees do not need to have all taxa, and this method is very fast. The filtering parameters defined by the user may have a strong impact of the conclusions, however. Than et al. (2007),Meng and Kubatko (2009) and Kubatko et al. (2009) developed model-based methods to compare the coalescent-only model with putative hybridization or HGT hypotheses on a known species tree, using maximum likelihood. These methods have only been thoroughly tested on small numbers of taxa containing at most two horizontal events. Joly et al. (2009) used coalescent simulations to compare gene divergence times observed from the real data with those expected under ILS, building on the work by Sang and Zhong (2000) and Holder et al. (2001). Buckley et al. (2006) similarly used coalescent simulations to determine if a specific pattern of discordance observed in their gene trees was atypical under ILS. Coalescent simulations were also used by Maureira-Butler et al. (2008) to compare the amount of discordance between two gene trees with that expected under the coalescent-only model. These various methods all aim to determine the nature of reticulation between population lineages. Each have specific strengths, but more work needs to be done for these methods to be broadly applicable. The test based on the comparison of CFs seems to be especially promising when many unlinked genes are available.

2.4.7 Future Work to Estimate CFs

Our work sheds light on strengths and weaknesses of the BCA implemented in BUCKy. This nonparametric method for estimating CFs accommodates various kinds of discordance, but the Dirichlet prior distribution should be improved to match more closely the typical biological processes responsible for gene tree discordance. The two-step approach in BUCKy has the strength of great computational tractability but at the cost of accuracy with moderate numbers of taxa. Current work is being done to develop a fast and robust method for concordance analysis.

Results in this paper are limited to the situations considered in our simulations. Our highest level of ILS caused only 46.4% (on 5 taxa) and 15.8% (on 11 taxa) of all gene trees to share the same rooted topology as the species tree. But this level of incongruence was not high enough to cause the presence of anomalous gene trees, when the most likely gene tree does not match the species tree topology (Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008). It would be interesting to conduct simulations in the too-greedy zone identified by Degnan et al. (2009). This zone contains species trees with very short branch lengths. Under the coalescent model on such species trees, the greedy consensus method misidentifies the species tree topology, even from a perfect reconstruction of CFs. Fortunately, identification of the species tree can be corrected by using CFs of quartets, rather than CFs of full-taxon clades (Degnan et al., 2009). It would be interesting to test the performance of BUCKy in the too-greedy zone, using the quartet-based method to reconstruct the population tree after CFs have been estimated.

Chapter 3

Computing the Joint Distribution of Tree Shape and Tree Distance for Gene Tree Inference and Recombination Detection

Abstract

Recombination events and other biological processes can cause the topologies of phylogenetic trees to be discordant for different genes. We consider Bayesian models for the analysis of very long sequence alignments, to simultaneously infer the number and position of recombination breakpoints and the phylogenetic tree of each block between breakpoints. The models we consider use a prior distribution on gene trees that favor similar trees at neighboring loci, through the use of a dissimilarity measure between tree topologies. Indeed, we show empirical evidence in Enterobacteria that trees at neighboring loci are more similar than trees at random locations along the alignment. The main hurdle to using models favoring similar trees at neighboring loci is the need to properly calculate the normalizing function for the prior probabilities on trees. Calculating this normalizing function as a naive sum is computationally prohibitive. In this work, we quantify the impact of approximating this normalizing function as done in biomc2. We then derive an algorithm to calculate the normalizing function exactly, for a Gibbs distribution based on the Robinson-Foulds (RF) distance between gene trees at neighboring loci. This algorithm is based on a system of generating functions. At the core is the calculation of the joint distribution of the shape of a random tree and its RF distance to a fixed tree. We also propose fast approximations to the normalizing function, which are shown to be very accurate with little impact on the inference of gene trees and recombination breakpoints.

3.1 Introduction

Recombination occurs in the genomes of many organisms leading to exchange of genetic material. In eukaryotes, recombination is reciprocal. In prokaryotic or-

ganisms, homologous recombination leads to a unidirectional flow of genetic material from a donor to a recipient, more akin to eukaryotic gene conversion. This is one type of horizontal gene transfer (HGT) that is particularly common among closely related organisms, such as within species of enterobacteria. Recombination events can complicate the analysis of the evolution of a group of organisms, as they can cause conflicting phylogenetic relationships between different regions of the genomes. Recently developed statistical methods simultaneously detect the location of recombination events along an alignment and infer phylogenetic histories of regions in the alignment defined by recombination breakpoints. These methods are based on the premise that discordant phylogenetic trees from different genomic regions are due to recombination events. RecPars (Hein, 1993) infers the most parsimonious history of substitutions on trees and recombinations, and MDL (Ané, 2011) enables a penalty parameter to control the number of breakpoints. PLATO (Grassly and Holmes, 1997) infers the maximum likelihood phylogenetic tree from the whole input alignment and then detects regions whose likelihood values for this tree are relatively small. Similarly, ClonalOrigin (Didelot et al., 2010) estimates the phylogenetic tree of the genome and recombination breakpoints in a two-stage hierarchical Bayesian framework. Hidden Markov models (HMM) assume that hidden states are the underlying trees of genetic regions (Husmeier and McGuire, 2003; Webb et al., 2009; Boussau et al., 2009). DualBrothers (Minin et al., 2005) is the first Bayesian method to infer breakpoint positions and phylogenetic trees simultaneously, but works well on only few taxa. cBrother (Fang et al., 2007) improved the computational issues of DualBrothers and StepBrothers (Bloomquist et al., 2009) further infers relative times of recombination events. Biomc2 (de Oliveira Martins et al., 2008) incorporates correlation in tree

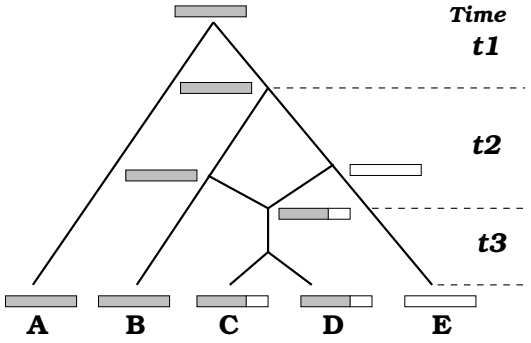

Figure 3.1: **Illustration of different tree topologies of genomic regions because of recombination event.** The phylogenetic tree of the gray region has the clade BCD, but the white region has the clade CDE.

topologies through the distance between trees at neighboring regions in a Bayesian model and is able to handle larger data sets.

Although tree topologies of regions between recombination breakpoints are different, these genomic regions share some evolutionary history before and after the recombination events. For example, in Figure 3.1, the gray genomic regions in taxa C and D have a different evolutionary history than the white genomic regions: genes in the gray region in taxa C and D are more closely related to genes in taxon B than to genes in taxon E, but genes in the white region are more closely related to genes in taxon E. However, the trees of the gray and white regions share features during time periods t_1 (A is an outgroup in both regions) and t_3 (both regions have clade CD in their trees).

Models that favor similar trees at neighboring genomic regions can detect breakpoints between very similar trees (Webb et al., 2009) and inference can be more accurate (Boussau et al., 2009). As far as we know, biomc2 is one of the few methods that take into account correlation between tree topologies: topologies at adjacent genomic regions can be different but preferably similar (but see Bloomquist et al.

(2009); Didelot et al. (2010)). Note that other methods such as cBrother and HMM uniformly prefer different topologies of adjacent genomic regions. Empirical evidence for correlation between trees at adjacent genomic regions is assessed in section 3.3.1. Biomc2 uses approximated SPR distances ($\hat{d}_{SPR}$) between tree topologies at adjacent pre-defined segments. The SPR distance is considered to have a truncated-Poisson distribution a priori, using the following probability-like function on tree topologies $\mathbf{T} = (T_1, \ldots, T_L)$:

$$\tilde{P}(\mathbf{T}|\beta, \mathbf{w}) = \frac{\prod_{l=1}^{L-1} \left\{ \frac{e^{-\beta_l} \beta_l^{\hat{d}_{SPR}(T_l, T_{l+1})}}{\hat{d}_{SPR}(T_l, T_{l+1})!} \right\}^{w_l+1}}{\tilde{\eta}(\beta, \mathbf{w}, L)} \tag{3.1}$$

where N is the number of taxa, L is the number of segments in the alignment,

$$\tilde{\eta}(\beta, \mathbf{w}, L) = \prod_{l=1}^{L-1} \left[\sum_{d=0}^{D} \left\{ \frac{e^{-\beta_l} \beta_l^d}{d!} \right\}^{w_l+1} \right] \tag{3.2}$$

and $D = N - 3$ is the number of internal edges and an upper bound for the SPR distance. The function $\tilde{\eta}$ used by de Oliveira Martins et al. (2008) is meant to normalize $\tilde{P}$ so that $\tilde{P}$ is a probability distribution: $\tilde{\eta}$ should ensure that the probabilities sum up to one. Such a function is called a normalizing function. The parameter β of the prior truncated-Poisson distribution is larger than the expected distance between neighboring tree topologies but for convenience it is interpreted here as the prior mean distance between neighboring trees. The vector $\mathbf{w}$ of non-negative weights enables the distribution (3.1) to have smaller mean and variance. Therefore, the use of (3.1) in biomc2 can give higher weights to similar trees at adjacent segments.

One difficulty for Gibbs-like distributions such as the prior distribution (3.1) used in biomc2 is that normalizing functions are easily overlooked or miscalculated. For example, Gibbs-like distributions have also been used for supertree estimation. A

model to find the maximum likelihood supertree from estimated smaller phylogenetic trees was proposed by Steel and Rodrigo (2008). Estimated gene trees can be different from the true tree on the full taxon set ("supertree") because of technical issues (e.g. incorrect orthology detection), stochastic error (e.g. estimation error), or biological processes (e.g. incomplete lineage sorting). In Steel and Rodrigo (2008), the discrepancy between gene trees T_i and the supertree $\mathcal{T}$ is modeled using the Robinson-Foulds (RF) distance (Robinson and Foulds, 1981) d and the likelihood of $\mathcal{T}$ as

$$P_{\mathcal{T}}(T_1, \ldots, T_k) \propto \prod_{i=1}^{k} \exp\left[-\beta_i d(T_i, \mathcal{T})\right], \tag{3.3}$$

where $T_1, \ldots, T_k$ are the k input gene trees estimated on taxon subsets. The "maximum likelihood supertree" proposed by Steel and Rodrigo (2008) is

$$\arg\min_{\mathcal{T}} \sum_{i=1}^{k} \beta_i d(T_i, \mathcal{T}). \tag{3.4}$$

However, as pointed out in Bryant and Steel (2009), the correct likelihood maximization should normalize the term (3.3) using

$$Z_{\mathcal{T},\beta} = \prod_{i=1}^{k} Z^{(i)}_{\mathcal{T},\beta_i}$$

with

$$Z^{(i)}_{\mathcal{T},\beta_i} = \sum_{T:\mathcal{L}(T)=\mathcal{L}(T_i)} \exp\left[-\beta_i d(T, \mathcal{T})\right] \tag{3.5}$$

where $\beta = (\beta_1, \ldots, \beta_k)$ and $\mathcal{L}(T_i)$ is the set of tip labels of gene tree T_i. In Bryant and Steel (2009), criterion (3.4) is corrected as

$$\arg\min_{\mathcal{T}} \left\{ \sum_{i=1}^{k} \beta_i d(T_i, \mathcal{T}) + \log Z_{\mathcal{T},\beta} \right\},$$

and a polynomial-time algorithm is described to calculate the distribution of the RF distance given the tree shape of $\mathcal{T}$. In biomc2, the function $\tilde{\eta}$ used in the prior distribution (3.1) on trees is *not* the actual normalizing function, that is, (3.1) is not a probability distribution because

$$\sum_{T_1} \cdots \sum_{T_L} \tilde{P}(T_1, \ldots, T_L | \beta, \mathbf{w}) \neq 1.$$

The correct normalizing function is

$$\eta(\beta, \mathbf{w}, L) = \sum_{T_1} \cdots \sum_{T_L} \prod_{l=1}^{L-1} \left\{ \frac{e^{-\beta_l} \beta_l^{d(T_l, T_{l+1})}}{d(T_l, T_{l+1})!} \right\}^{w_l+1}. \tag{3.6}$$

In other words, the numerator in (3.1) should be summed over all possible trees, not over tree distance values. The normalizing function in biomc2 has not been corrected in its implementation or in publication as far as we know. More generally, complex computation of normalizing functions makes it difficult to embed correlations among tree topologies into statistical models.

To simultaneously detect recombination breakpoints and infer phylogenetic trees of genomic regions, we propose a method with a new Gibbs prior distribution. The Gibbs probability of a random variable X having value x is

$$P(X = x) = \frac{1}{Z(\beta)} \exp\left(-\beta E(x)\right),$$

where $E(x)$ is called the energy function of the configuration x, β is a parameter called the inverse temperature (Cipra, 1987) and $Z(\beta)$ is the normalizing function, also called a partition function. We consider the sum of RF distances (interpreted as dissimilarities) between tree topologies at adjacent genomic regions as the energy of the phylogenetic histories of the genomic regions.

In section 3.2, the impact of overlooking the normalizing function in biomc2 is investigated. In section 3.3, a Bayesian model is introduced to simultaneously identify recombination breakpoints and infer phylogenetic histories. We use marginal likelihood after integrating out branch lengths on phylogenetic trees and a Gibbs distribution is used as a prior distribution on tree topologies. Section 3.4 shows that the normalizing function of the Gibbs distribution can be calculated through the number of tree topologies with a certain shape and at a certain distance away from a given tree topology. In section 3.5 we propose approximations to the normalizing function. Conclusion and discussion are in section 3.6.

3.2 Importance of the normalizing function

Not fully knowing the prior distribution might not be a problem under a Markov chain Monte Carlo (MCMC) approach, where the prior distribution needs to be known only up to a constant. Assume we want to use a prior distribution on tree topologies, $P(T_1, \ldots, T_L|\beta)$, that cannot be easily calculated. Suppose that P is replaced in the MCMC algorithm by $\tilde{P}$ where the product

$$\tilde{P}(T_1, \ldots, T_L|\beta)\tilde{f}(\beta) = P(T_1, \ldots, T_L|\beta)f(\beta),$$

is easily evaluated at each step of the MCMC. Here $f(\beta)$ is the real normalizing function for the true prior distribution P, but is difficult to calculate. Instead, $\tilde{f}(\beta)$ is a pseudo-normalizing function easier to calculate. If a fixed β is used to infer the posterior distribution of tree topologies, it is fine to use the pseudo-normalizing function $\tilde{f}(\beta)$ or to simply ignore the true normalizing function $f(\beta)$. If we assume a

hyperprior distribution $\pi(\beta)$ on β, however, then

$$\tilde{P}(T_1, \ldots, T_L|\beta)\pi(\beta) = P(T_1, \ldots, T_L|\beta)\frac{f(\beta)}{\tilde{f}(\beta)}\pi(\beta).$$

If $\tilde{P}$ is used instead of P in an MCMC approach to define the prior probability of trees given β, then the hyperprior distribution *actually used* on β is

$$\tilde{\pi}(\beta) = \frac{f(\beta)}{\tilde{f}(\beta)}\pi(\beta)\left[\int \frac{f(\beta')}{\tilde{f}(\beta')}\pi(\beta')d\beta'\right]^{-1}. \tag{3.7}$$

If the ratio $f/\tilde{f}$ is a constant, i.e. the true normalizing function is known up to a constant, then the MCMC sampling procedure is not affected by the usage of the function $\tilde{P}(T_1, \ldots, T_L|\beta)$ as a prior distribution. However, a problem arises when $f/\tilde{f}$ is not constant, since MCMC samples are not from the assumed posterior distribution in that case. The correct calculation of the normalizing function is then required.

3.2.0.1 Hyperprior used in biomc2

Biomc2 aims to estimate the trees $(T_1, \ldots, T_L)$ along an alignment with L predefined short segments. For the prior distribution on tree topologies, biomc2 considers the truncated-Poisson distribution in (3.1) parametrized by $\beta = (\beta_1, \ldots, \beta_L)$. Independent gamma hyperprior distributions are placed on β_l and w_l. To see the impact of using the pseudo-normalizing function (3.2) rather than the true normalizing function (3.6), we compare the ratio of normalizing functions $\eta(\beta, L)/\tilde{\eta}(\beta, L)$ and calculate the hyperprior distribution on β actually used as determined by (3.7). Either comparison is not easy to carry out analytically, so we consider here a simple case when all w_l's are fixed to 1 and β_l's are all equal. The function used as a prior distribution on tree topologies in (3.1) becomes:

$$\tilde{P}(T_1, \ldots, T_L|\beta) = \frac{1}{\tilde{\eta}(\beta, L)}\prod_{l=1}^{L-1}\frac{e^{-\beta}\beta^{d_{\mathrm{SPR}}(T_l, T_{l+1})}}{d_{\mathrm{SPR}}(T_l, T_{l+1})!}, \tag{3.8}$$

where d_{SPR} is the true SPR distance between tree topologies and the pseudo-normalizing function is

$$\tilde{\eta}(\beta, L) = \prod_{l=1}^{L-1} \left[\sum_{d=0}^{D} \frac{e^{-\beta}\beta^d}{d!} \right] .$$

The correct normalizing function for (3.8) is

$$\eta(\beta, L) = \sum_{T_1} \cdots \sum_{T_L} \prod_{l=1}^{L-1} \frac{e^{-\beta}\beta^{d_{\mathrm{SPR}}(T_l, T_{l+1})}}{d_{\mathrm{SPR}}(T_l, T_{l+1})!} .$$

We follow de Oliveira Martins and Kishino (2010) and choose an exponential distribution $\mathcal{E}(\lambda)$ with mean $1/\lambda$ for the distribution of β, which is a special case of the gamma distribution. The hyperprior distribution actually used is then:

$$\tilde{P}(\beta) \propto P(\beta)\frac{\eta(\beta, L)}{\tilde{\eta}(\beta, L)} = \lambda e^{-\lambda\beta} \times \sum_{T_1} \cdots \sum_{T_L} \left\{ \frac{\beta^{\sum_{l=1}^{L-1} d_{\mathrm{SPR}}(T_l, T_{l+1})}}{\prod_{l=1}^{L-1} d_{\mathrm{SPR}}(T_l, T_{l+1})!} \right\} \Big/ \left(\sum_{d=0}^{D} \frac{\beta^d}{d!} \right)^{L-1} .$$

When there are 2 candidate recombination breakpoints ($L = 3$) and $N = 5$ taxa, the ratio of the true normalizing function to the pseudo-normalizing function can be calculated exactly and it is not a constant (Figure 3.2 (a)), although the ratio converges to 4 as β increases. The hyperprior distribution actually used

$$\tilde{P}(\beta) \propto \frac{1 + 24\beta + 146\beta^2 + 24\beta^3 + \beta^4}{\left(1 + \beta + \beta^2/2\right)^2} \lambda e^{-\lambda\beta}, \tag{3.9}$$

differs from the targeted hyperprior distribution $\mathcal{E}(\lambda)$ as shown in Figure 3.2 (b)-(d) for $\lambda = 100, 1$ and 0.01. The exponential density has a mode at $\beta = 0$, but the density actually used (3.9) has very small values near $\beta = 0$ except for $\lambda = 100$. This discrepancy might partly explain why it is recommended to use a very large λ in biomc2 and even more so when there are more candidate recombination breakpoints.

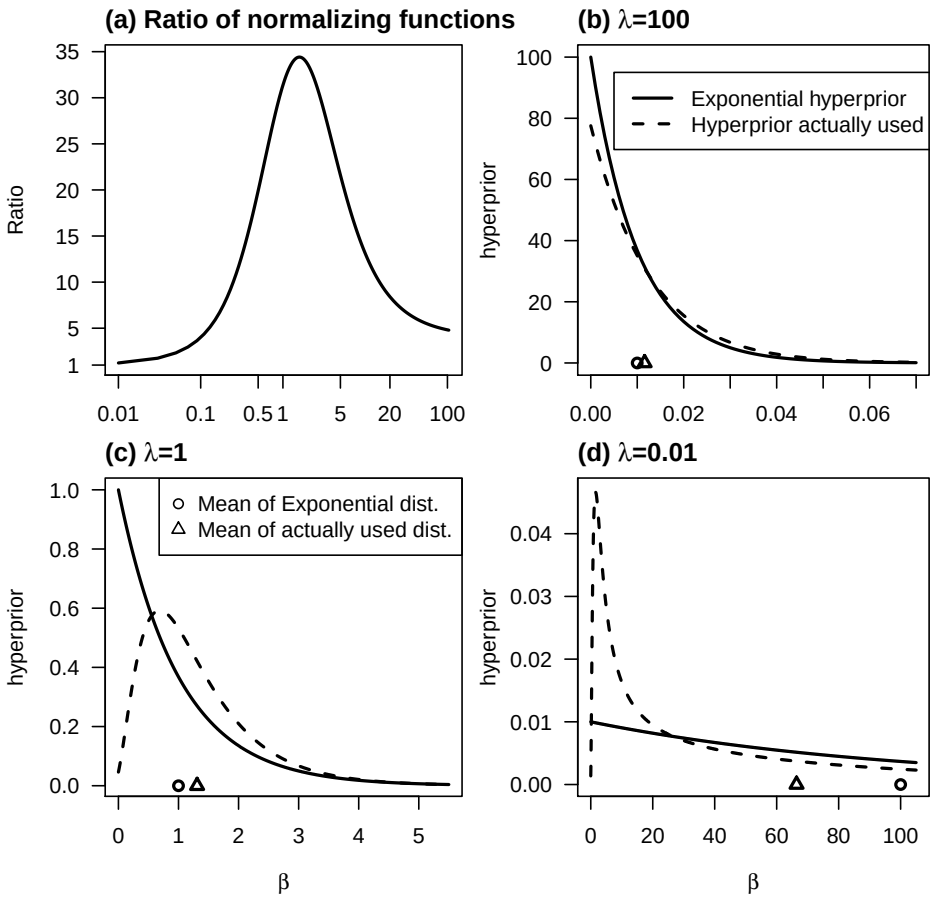

Figure 3.2: **Impact of using the pseudo-normalizing function $\tilde{P}$ in (3.8).** The alignment has 2 candidate breakpoints ($L = 3$) and $N = 5$ taxa. (a) Ratio of the true normalizing function to the pseudo-normalizing function $\eta/\tilde{\eta}$. (b)-(d) The real line indicates the exponential distribution $\mathcal{E}(\lambda)$, whose mean $1/\lambda$ is indicated with a circle ($\circ$). The hyperprior density actually used $\tilde{P}$ is indicated with a dotted line (- -) whose mean is indicated with a triangle (Δ) when $\lambda = 100, 1$ and 0.01. Note that the axis for β in (a) is on a log scale.

3.3 Model similarity among gene trees

3.3.1 Correlation among gene trees in real data

To investigate the level of spatial correlation between phylogenetic trees at neighboring loci, progressiveMauve was applied to generate alignments of 33 *Escherichia* genomes and 8 *Shigella* genomes (Darling et al., 2010). The longest alignment among those with non-empty sequences from all of the 41 taxa contained 52080 base pairs (bps). We partitioned this alignment into 103 segments of 500 bps and 1 segment of 580 bps. We excluded segments from the analysis if they shared less than 4 taxa with all other segments. 76 segments remained. We applied MrBayes to each segment independently. The HKY model (Hasegawa et al., 1985) with Gamma-distributed rates across sites was used with 4 chains, 2 independent runs and 10 million generations. Trees were sampled every 100th generation and the first 10% were discarded. We estimated the phylogenetic tree of each segment by the greedy consensus tree with posterior probabilities on internal edges.

We modified the RF distance to account for posterior probabilities of internal edges in the trees, to give lower weight to edges with high uncertainty. Our modified weighted RF (wRF) distance is also normalized to compare subtrees on identical taxon sets:

$$wRF(T_1, T_2) = \frac{\sum\limits_{\substack{c \in \mathcal{C}(T_{1|L_1 \cap L_2}) \\ c \notin \mathcal{C}(T_{2|L})}} pp_1(c) + \sum\limits_{\substack{c \notin \mathcal{C}(T_{1|L}) \\ c \in \mathcal{C}(T_{2|L})}} pp_2(c)}{\sum\limits_{c \in \mathcal{C}(T_{1|L})} pp_1(c) + \sum\limits_{c \in \mathcal{C}(T_{2|L})} pp_2(c)},$$

where $L = L_1 \cap L_2$ is the set of taxa that are common to both trees T_1 and T_2, $T_{i|L} (i = 1, 2)$ is the subtree obtained from T_i after pruning taxa whose labels are not

in the other tree, $\mathcal{C}(T)$ is the collection of all bipartitions in tree T, and $pp_i(c)$ is the posterior probability of c for tree T_i. Since the wRF distance was scaled between 0 and 1, it provides comparable distances between trees of different sizes. We computed the wRF distance between the consensus trees from all pairs of segments.

To determine if trees from nearby segments more similar (positively correlated) than expected by chance, a permutation test was conducted on the wRF distance between trees from segments located k segments apart, for each $k = 1, \ldots, 102$. We randomly shuffled the greedy consensus trees along the alignment, and then computed average wRF distances between trees located k segment away from each other. We repeated the process 100,000 times and calculated p-values by counting the number of times that the sampled average wRF distance was smaller than the observed average wRF distance.

The average wRF distance roughly increases with the physical distance between segments, as shown in Figure 3.3. Note that only a small number of segment pairs underlies the average wRF distance for segments located very far apart, leading to a large standard error.Although the average wRF distance between trees at neighboring segments is as high as 0.744, the trees were significantly more similar to each other than expected by chance. At the 1% significance level, trees from regions no more than 2kb (4 segments) away from each other were significantly similar to each other. The correlation between trees was too weak to be detected in our experiment across distances beyond 2kb.

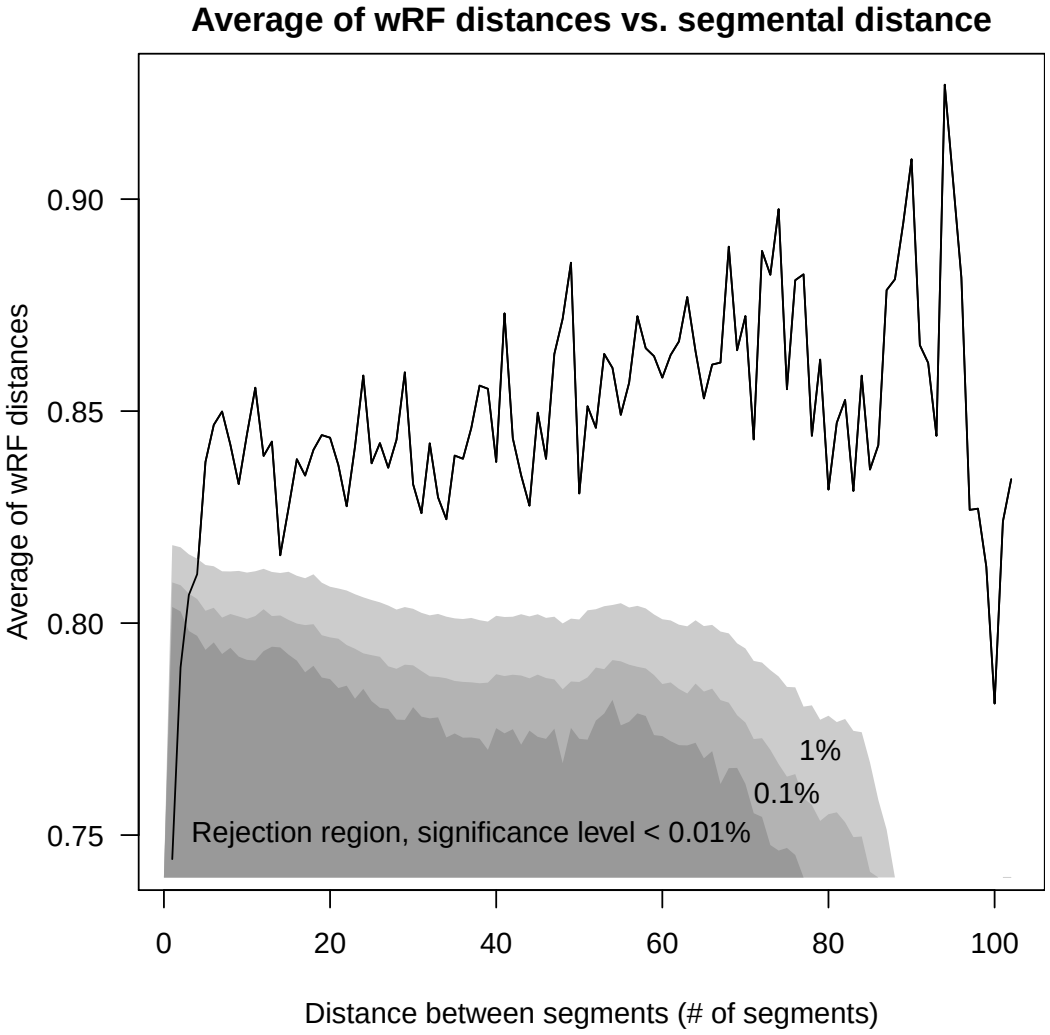

Figure 3.3: Average wRF distance between trees from 500-bp segments that are a given physical distance apart in the alignment. For each k ($k = 1, \ldots, 102$), a permutation test was conducted to determine if the wRF distance between trees of loci located k segments apart is lower than expected by chance. The test was significant (p-value < 0.01) for $k = 1, 2, 3$ and 4 segments only (i.e 2kb).

3.3.2 Model to simultaneously infer the position of recombination breakpoints and phylogenetic trees

3.3.2.1 Sequence likelihood

We describe here the sequence evolution model which motivates our work, to show how our Gibbs model on tree topologies can be used as a prior distribution for simulataneously detecting recombination and estimating phylogenetic trees. Along an alignment $\mathbf{X}$ from N taxa, each site can be a candidate recombination breakpoint, and hence sites may have different underlying phylogenetic tree topologies. If we estimate a tree topology at each individual site, however, the uncertainty in tree topology estimation can be unreasonably large. Therefore, the alignment is divided into L pre-defined arbitrary short segments, $(\mathbf{X}_1, \ldots, \mathbf{X}_L)$. Segments may have different tree topologies $\mathbf{T} = (T_1, \ldots, T_L)$ because of recombination. Within a segment, all sites are assumed to have the same phylogenetic tree. Substitutions at each site are modeled by standard continuous-time Markov processes along the tree, such as the HKY model (Felsenstein, 2004). The likelihood of the sequence alignment given the segment trees and evolutionary parameters can then be calculated efficiently. If branch lengths are allowed to be segment-independent and site-independent, they can be integrated out analytically to reduce the computational burden. This is very helpful when considering extremely long alignments (Minin et al., 2005; de Oliveira Martins et al., 2008). Recombination breakpoints and segment trees can then be estimated in a Bayesian framework, provided that a suitable prior distribution be placed on segment trees.

3.3.2.2 Prior on tree topologies: Gibbs distribution

We propose a new Gibbs prior distribution on tree topologies $(T_1, \ldots, T_L)$ to take into account the similarity of trees across consecutive segments. This similarity can be measured by the RF distance between tree topologies. The RF distance between two fully resolved unrooted trees is the number of bipartitions found in only one of the two trees. It has an even value and the distance $d(\cdot, \cdot)$ used here is one-half of the RF distance. The proposed prior probability of tree topologies is then:

$$P(T_1, \ldots, T_L) = \exp\left(-\beta \sum_{l=1}^{L-1} d(T_l, T_{l+1})\right) / Z_L(\beta), \tag{3.10}$$

where β a non-negative parameter and $Z_L(\beta)$ is the normalizing function

$$Z_L(\beta) = \sum_{t_1} \ldots \sum_{t_L} \exp\left(-\beta \sum_{l=1}^{L-1} d(t_l, t_{l+1})\right) \tag{3.11}$$

to ensure that the probabilities in (3.10) sum to one. When there is only $L = 1$ segment, $Z_1(\beta) = Z_1 = (2N-5)!!$ is the total number of tree topologies and does not depend on β.

Under this Gibbs distribution, similar trees at adjacent segments are favored. For large β, $Z_L(\beta)$ approaches Z_1 and the Gibbs distribution forces all trees to be identical:

$$P(T_1, \ldots, T_L) = \begin{cases} 1/Z_1 & \text{if } T_1 = \ldots = T_L, \\ 0 & \text{otherwise,} \end{cases}$$

as if no recombination occurred. When $\beta = 0$, $Z_L(\beta) = Z_1^L$ and the Gibbs probability $P(T_1 = t_1, \ldots, T_L = t_L) = 1/Z_1^L$ regardless of the values of $t_1, \ldots, t_L$. In other words, trees become independent, each with a uniform distribution. Between these two extreme distributions, $1/\beta$ scales with the average recombination rate per segment. We will informally call β the a priori inverse recombination rate.

The Gibbs distribution has desirable properties, such as the following Markov property, by the Hammersley-Clifford theorem (Preston, 1973):

$$\begin{aligned} & P(T_j = t | T_1, \ldots, T_{j-1}, T_{j+1}, \ldots, T_L) \\ & = P(T_j = t | T_{j-1}, T_{j+1}), \text{ for } j = 2, \ldots, L-1. \end{aligned} \tag{3.12}$$

In other words, conditional on its neighbors, a tree T_j is independent of the trees at all other segments. Moreover, the distribution is homogeneous across the alignment: $P(T_{i+1} = t_2 | T_i = t_1, T_{i+2} = t_3)$ is independent of i for $1 \leq i \leq L-2$.

The sequence of tree topologies from the Gibbs distribution (3.10) is a Markov chain, since we can easily show that, for $i = 1, \ldots, L-1$,

$$P(T_{i+1} = t | T_1, \ldots, T_i) = P(T_{i+1} = t | T_i) = \frac{\exp(-\beta d(T_i, t)) Z_{L-i,t}(\beta)}{Z_{L-(i-1),T_i}(\beta)}, \tag{3.13}$$

where $Z_{k,T_1}(\beta) = \sum_{T_2} \cdots \sum_{T_k} \exp\left\{ -\beta \sum_{j=1}^{k-1} d(T_j, T_{j+1}) \right\}$ and $Z_{1,T}(\beta) = 1$. However, this Markov chain is non-stationary, since the transition probability (3.13) may depend on location i. It may also depend on the total number of segments L. The marginal distribution of each tree topology may vary across locations:

$$P(T_i = t) = \frac{Z_{i,t}(\beta) Z_{L-(i-1),t}(\beta)}{Z_L(\beta)},$$

for $i = 1, \ldots, L$. Note that the marginal distributions are symmetric, that is, $P(T_i = t) = P(T_{L-(i-1)} = t)$ for any t.

We define a block as the collection of all the consecutive segments located between two recombination breakpoints. In other words, segments (and sites) within a block are inferred to have the same tree topology while segments in two adjacent blocks are inferred to have different tree topologies. Knowing the prior distribution

on the number of breakpoints B can be useful to choose an appropriate value for the recombination rate, or an appropriate hyperprior mean if the inverse recombination rate, β, is given a hyperprior distribution. Indeed, the following proposition (proved in Appendix A) links β to the expected number of recombination breakpoints.

Proposition 1. *Assume the Gibbs distribution (3.10). When $\beta = 0$, the number of recombination breakpoints B has a binomial distribution $\mathcal{B}(L-1, 1-1/Z_1)$ with expectation $E(B) = (L-1)(1-1/Z_1)$. If $\beta = \infty$ there is exactly one block: $B = 0$ with probability 1. In general,*

$$P(B=b) = \binom{L-1}{b} \sum_{i=0}^{b} \frac{\binom{b}{i}(-1)^{b-i} Z_{i+1}(\beta)}{Z_L(\beta)},$$

with an expected number of breakpoints

$$E(B) = (L-1)\left(1 - \frac{Z_{L-1}(\beta)}{Z_L(\beta)}\right).$$

At each segment boundary, the probability of there being a recombination breakpoint is $1-Z_{L-1}/Z_L$. This is maximum at $1-1/Z_1$ when the recombination rate is very large ($\beta = 0$). When the recombination rate is small (large β), this is approximately $2(N-3)e^{-\beta}$ (see Appendix A).

3.4 Computing the normalizing function

When tree topologies $(T_1, \ldots, T_L)$ of consecutive segments follow the Gibbs distribution in (3.10), the corresponding normalizing function $Z_L(\beta)$ in (3.11) depends on β. With this model, it is necessary to either compute the normalizing function exactly or to provide a good approximation for it if we want to place a hyperprior on β, such as an exponential distribution with mean $1/\lambda$. In this section, we develop an

algorithm to calculate $Z_L(\beta)$ exactly. We can rewrite

$$Z_L(\beta) = \sum_{S_1 \in \mathcal{S}_N} \zeta(S_1) Z_{L,S_1}(\beta) \tag{3.14}$$

where the sum goes over the set of unrooted tree shapes $\mathcal{S}_N$ on N tips, $\zeta(S) = |\{T : \mathbf{S}(T) = S\}|$, $\mathbf{S}(T)$ denotes a tree shape from tree T by discarding the terminal node labels and

$$Z_{L,S_1}(\beta) = \sum_{T_2} \cdots \sum_{T_L} \exp\left\{-\beta \sum_{l=1}^{L-1} d(T_l, T_{l+1})\right\}$$

for any fixed T_1 of shape S_1. The value of $Z_{L,S_1}(\beta)$ can be recursively computed as:

$$Z_{L,S_1}(\beta) = \sum_{S_2 \in \mathcal{S}_N} \sum_{y=0}^{N-3} \zeta_{2,S_1}(S_2, y) e^{-\beta y} Z_{L-1,S_2}(\beta) \tag{3.15}$$

where, for any fixed T of shape S,

$$\zeta_{2,S}(S', y) = |\{T' : d(T, T') = y \text{ and } \mathbf{S}(T') = S'\}|. \tag{3.16}$$

Therefore, $Z_{L,S_1}(\beta)$ in (3.15) and eventually $Z_L(\beta)$ can be recursively computed from the $\zeta_{2,S}(S', y)$ values. The rest of the section provides a way to calculate $\zeta_{2,S}(S', x)$ for all values of x and all shapes S, S'. In other words, the goal of the following subsections is to determine the joint distribution of the shape of T_2 and $d(T_1, T_2)$ conditional on T_1 (or its shape) when T_2 has a uniform distribution.

3.4.1 Computing the joint distribution of the Robinson-Foulds metric and tree shape

Computing $\zeta_{2,S}(S', x)$ in (3.16) is required to recursively compute the normalizing function $Z_L(\beta)$. We fix tree T with shape S in the rest of section 4. Then, $\zeta_{2,S}(S', x)$ is the number of tree topologies with shape S' whose distance from T is

x. In this subsection, we provide several generating functions that are linked to our target frequency $\zeta_{2,S}(S', x)$, simplified as $\zeta_S(S', x)$ here. First, we define $q_S(S', d)$ as

$$q_S(S', d) = |\{T' : \mathbf{S}(T') = S',\ T \text{ and } T' \text{ share } \textit{exactly } d \text{ bipartitions}\}|,$$
$$= \sum_{\alpha \in \mathcal{A}:|\alpha|=d} |\{T' : \mathbf{S}(T') = S',\ T \text{ and } T' \text{ share } \textit{exactly} \text{ bipartitions } \alpha\}|,$$

where $\mathcal{A}$ is the power set of all possible bipartitions from tree T, and thereby $\zeta_S(S', x)$ can be calculated through

$$q_S(S', d) = \zeta_S(S', N - 3 - d) \text{ for } d = 0, \ldots, N - 3.$$

The generating function for $q_S(S', d)$, defined as

$$Q_{S,S'}(x) = \sum_{d=0}^{N-3} q_S(S', d)x^d$$

is called the "exact" generating function by Goulden and Jackson (2004). The "at-least" generating function for the number of tree topologies with shape S' is defined as

$$U_{S,S'}(x) = \sum_{d=0}^{N-3} u_S(S', d)x^d, \tag{3.17}$$

where

$$u_S(S', d) = \sum_{\alpha \in \mathcal{A}:|\alpha|=d} |\{T' : \mathbf{S}(T') = S',\ T \text{ and } T' \text{ share partitions } \alpha \text{ and possibly others}\}| \tag{3.18}$$

and satisfies the following equation by the principle of inclusion and exclusion(Goulden and Jackson, 2004):

$$Q_{S,S'}(x) = U_{S,S'}(x - 1).$$

Therefore, if we can determine U, then we can determine Q and all $\zeta_S(S', d)$ values. The following subsections present an algorithm to compute $u_S(S', d)$.

3.4.2 Definitions and theorems

To compute $\zeta_{2,S}(S', x)$ in (3.16) through $u_S(S', d)$ in (3.18), we first define the terminology used in the following sections. First, we assume that all trees and tree shapes are in their left-light centered (LLC) form, which provides an unique representation and was used to rank all possible tree shapes (Furnas, 1984). Edges and nodes on trees or shapes in LLC form can be labeled in a unique way. To transform an unrooted tree or tree shape into its LLC form, we first determine its centroid(s). A centroid is a node that leads to no more than half of the terminal nodes. Furnas (1984) showed that any binary tree has either a single or two centroid nodes, and that these two centroids must be neighbors. If there are two centroids, a new node called the "pseudo-root" is introduced on the edge connecting the two centroids (Figure 3.4) and used to root the tree. The tree is rooted at the unique centroid node otherwise. Then, every edge should lead to an equal number of or fewer terminal nodes than any sister edge on its right, for the tree to be in LLC form. Once trees and shapes are in LLC form, edges are labeled as $1, \ldots, N-3$ following a pre-order tree traversal (root to tip then left to right, see Figure 3.4). Note that these edge labels do not correspond to bipartitions, but instead only depend on the tree shape.

We now define edge and node properties. A node is called "cherry" if it is directly connected to two leaves. Edges e and e' in a tree are *symmetric* if we can exchange the labels of e and e' by flipping subtrees at their most recent common ancestor (MRCA) and possibly at some of its descendant nodes while maintaining the

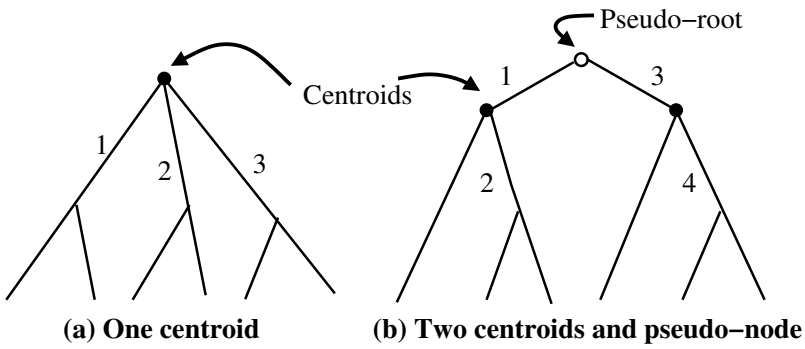

Figure 3.4: Two tree shapes in LLC form with (a) one centroid and (b) two centroids and newly introduced pseudo-root. Centroids are indicated with filled circles (•) and the pseudo-root is indicated with an empty circle (∘). Internal edge labels are defined using a pre-order tree traversal.

tree in LLC form. For example, edges labeled 2 and 4 are symmetric in Figure 3.4(b). Two nodes are *symmetric* in a tree if their parent edges are symmetric. Two nodes in a tree are *incomparable* if one is not ancestor or descendant of the other. A set of nodes $\{\nu_1, \ldots, \nu_n\}$ in a tree is an *antichain* if the nodes are pairwise incomparable. If an antichain is not a proper subset of any other antichain, then it is a *maximal antichain*.

Let $\mathbf{e}$ be a vector of internal edge labels on tree T. Define the tree forest $T\backslash_m\mathbf{e}$ as the set of subtrees derived by disconnecting edges in $\mathbf{e}$ and by adding labels as described next. Pseudo-terminal nodes are introduced where internal edges in $\mathbf{e}$ are disconnected. The edge indices are used to labels these pseudo-terminal nodes. That way, the two new terminal nodes from the same original internal edge have *matching* labels. More specifically, the two new nodes obtained from cutting e_i are both labeled m_i.

If the argument of a shape function $\mathbf{S}$ is a tree forest $T\backslash_m\mathbf{e}$, then $\mathbf{S}$ generates a forest from $T\backslash_m\mathbf{e}$ by removing the (pseudo)-root and terminal node labels but keeping

pseudo-terminal node labels. That is,

$$\mathbf{S}(T\backslash_m \mathbf{e}) = \{\mathbf{S}(F_1), \ldots, \mathbf{S}(F_{|\mathbf{e}|+1})\},$$

where the F_i's are the elements of forest $T\backslash_m \mathbf{e}$. Note that if T_1 and T_2 are different topologies but have the same shape, then $\mathbf{S}(T\backslash_m \mathbf{e}) = \mathbf{S}(T'\backslash_m \mathbf{e})$ for any edge vector $\mathbf{e}$ on T (or T').

Similarly, for any label set $\mathcal{L}$ and permutation $\boldsymbol{\sigma}_{\mathcal{L}}$ of these labels, we consider $\boldsymbol{\sigma}_{\mathcal{L}}$ as applying to trees by only permuting labels in $\mathcal{L}$. If the argument tree contains pseudo-terminal nodes with matching labels, $\boldsymbol{\sigma}_{\mathcal{L}}$ only permutes the original node labels in $\mathcal{L}$.

Key to our formulas are two equivalence relations.

Definition 1. Set Equivalence. *Let* $\mathbf{e}$ *and* $\mathbf{e}'$ *be vectors of edge labels on tree* T*. They are set-equivalent if* $\mathbf{e}$ *can be obtained from* $\mathbf{e}'$ *by permuting the order of elements in* $\mathbf{e}'$*. For each set-equivalence class, the representative edge vector* $\mathbf{e}$ *is defined as the only class member whose elements are arranged with ascending labels. The collection of all set-equivalence class representatives is denoted as* $\overset{\circ}{\mathcal{E}}(T)$*.*

Definition 2. Subtree-shape Equivalence. *Vectors of edge labels* $\mathbf{e}$ *and* $\mathbf{e}'$ *are subtree-shape equivalent if* $\mathbf{S}(T\backslash_m \mathbf{e}) = \mathbf{S}(T\backslash_m \mathbf{e}')$*. Note that this relation depends on* T *through its shape only.* $\mathbf{e} = (e_1, \ldots, e_d)$ *is defined as the representative of its subtree-shape equivalence class if it satisfies the following conditions:*

1. $e_1 \leq e'$ *for any edge* e' *symmetric with* e_1*.*

2. *For* $d > 1$*,*

(a) *sub-vector* $(e_1, \ldots, e_{d-1})$ *is the representative of its subtree-shape equivalence class,*

(b) $e_d \leq e'$ *for any* $e' \notin (e_1, \ldots, e_{d-1})$ *symmetric with* e_d *and that satisfies the following conditions: for each* $e_i \in (e_1, \ldots, e_{d-1})$*, (i)* e_d *and* e' *are descendants of* e_i *or (ii) any pairs of* e_d*,* e' *and* e_i *are incomparable and* $MRCA(e_i, e_d) = MRCA(e_i, e')$*.*

If T *has a pseudo-root, the 2 edges* e_L *and* e_R *connected to that root represent a unique edge on the unrooted tree. Therefore, for this definition, all edges (except* e_L *and* e_R*) are considered to be descendent of the left edge* e_L*.*

We prove in the Appendix that this definition identifies a unique representative of every equivalence class. $\breve{\mathcal{E}}(T)$ is defined as the collection of all subtree-shape equivalent class representatives.

For a vector $\mathbf{e} = (e_1, \ldots, e_h)$ of edges in a tree topology T, $\mathbf{S}(T/\bar{\mathbf{e}})$ is defined as the shape of the consensus tree obtained by contracting all edges but $e_1, \ldots, e_h$ on T, and by giving label c_i to the edge corresponding to e_i. Suppose that trees T and T' have shape S and S', respectively, and consider edge vectors $\mathbf{e}$ on T and $\mathbf{e}'$ on T'. Note that $\mathbf{S}(T/\bar{\mathbf{e}}) = \mathbf{S}(T'/\bar{\mathbf{e}}')$ holds precisely when there exist tree topologies T_1 and T_1' with shape S and S', respectively, such that the bipartitions defined by $\mathbf{e}$ on T_1 are the same as the bipartitions defined by $\mathbf{e}'$ on T_1'.

For the rest of this paper, we further fix a tree T' with shape S' and define ν_0 and ν_0' to be the roots of T and T' (once in LLC form). For $d \geq 0$ we define

$$\gamma_S(S', d) = \sum_{\substack{\mathbf{e} \in \mathring{\breve{\mathcal{E}}}(T), \\ |\mathbf{e}| = d}} \sum_{\substack{\mathbf{e}' \in \breve{\mathcal{E}}(T'), \\ |\mathbf{e}'| = d}} N(T' \backslash_m \mathbf{e}') \mathbb{I}_{\mathbf{S}(T/\bar{\mathbf{e}}) = \mathbf{S}(T'/\bar{\mathbf{e}'})}, \tag{3.19}$$

where $\mathbb{I}$ is the indicator function, and

$$N(T\backslash_m \mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|+1} \#\left\{F : \exists \boldsymbol{\sigma}_{\mathcal{L}_i} \text{ such that } \boldsymbol{\sigma}_{\mathcal{L}_i}(F) = F_i, F_i \in T\backslash_m \mathbf{e}\right\}. \tag{3.20}$$

Each term in the product is the number of trees obtained by permuting the original tip labels $\mathcal{L}_i$ on tree F_i in the forest $T\backslash_m \mathbf{e}$. Note that the $N(T\backslash_m \mathbf{e})$ values are easily calculated recursively (see the Appendix C). We also define the generating function

$$\Gamma_{S,S'}(x) = \sum_{d=0}^{N-3} \gamma_S(S', d)x^d. \tag{3.21}$$

The following theorem shows that γ equals u, and hence is the object of interest to eventually compute $\zeta_{2,S}(S', x)$.

Theorem 1. *$\Gamma_{S,S'}(x)$ is the "at-least" generating function for the number of tree topologies with shape S'. In other words, $\Gamma_{S,S'}(x) = U_{S,S'}(x)$ and $u_S(S', d) = \gamma_S(S', d)$ in (3.18).*

We are now ready to define the main object that our algorithm calculates recursively through the tree. Consider a vector V of p antichain nodes in tree T, arranged with ascending labels, and a vector V' of q antichain nodes in tree T'. Further, consider vectors D and K of p non-negative integers, and a vector M of p 0/1 elements. Similarly, consider vectors D', K' and M' of size q with non-negative and binary elements. Finally, H is assumed to be a set of pairs of indices, pairing elements of V

with elements of V'. The following function generalizes the γ function (3.19):

$$R(V,V',D,D',K,K',M,M',H) = \sum_{\substack{E=(\mathbf{e}_1,\ldots,\mathbf{e}_p)\\ \in \mathring{\mathcal{M}}_{V,D,K,M}}} \sum_{\substack{E'=(\mathbf{e}'_1,\ldots,\mathbf{e}'_q)\\ \in \breve{\mathcal{M}}_{V',D',K',M'}}} \sum_{G' \in \mathbb{G}_{E',V',D',M',H}} \left\{ \prod_{i=1}^{q} N(T'_{v'_i} \backslash_m \mathbf{e}'_i) \right.$$
$$\left. \times \mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', G') \right\}, \quad (3.22)$$

where all elements are described in the rest of this section, and such that $\gamma_S(S', d) = \gamma(d)$ is

$$\gamma(d) = \sum_{k=0}^{N} R\left((\nu_0), (\nu'_0), (d), (d), (k), (k), (0), (0), \emptyset\right). \quad (3.23)$$

Given $V = (v_1, \ldots, v_p)$, E, M and K of size p, $\mathcal{T}_V = (T_{v_1}, \ldots, T_{v_p})$ is a vector of subtrees of T satisfying the following conditions: (i) T_{v_i} contains all descendants of node v_i; (ii) T_{v_i} is rooted at v_i if $m_i = 0$. If $m_i = 1$, the parent edge is included in T_{v_i} as a root edge and is considered as an internal edge. We define $\mathring{\mathcal{M}}_{V,D,K,M} = \prod_{i=1}^{p} \mathring{\mathcal{M}}_{v_i,d_i,k_i,m_i}$ with

$$\mathring{\mathcal{M}}_{v,d,k,m} = \left\{ \mathbf{e} : |\mathbf{e}| = d + m, \mathbf{e} \in \mathring{\mathcal{E}}(T_v), |F_v| = k; \text{the parent edge of } v \in \mathbf{e} \text{ if } m = 1 \right\},$$

where F_v is the element of $T_v \backslash_m \mathbf{e}$ containing node v and $|F_v|$ is the number of original terminal nodes in F_v, not counting pseudo-terminal nodes. Similarly, $\breve{\mathcal{M}}_{V',D',K',M'} = \prod_{j=1}^{q} \breve{\mathcal{M}}_{v'_j,d'_j,k'_j,m'_j}$ and

$$\breve{\mathcal{M}}_{v,d,k,m} = \left\{ \mathbf{e} : |\mathbf{e}| = d + m, \mathbf{e} \in \breve{\mathcal{E}}(T_v), |F_v| = k; \text{the parent edge of } v \in \mathbf{e} \text{ if } m = 1 \right\},$$

We next consider position vectors. They will be used later to merge vectors $(\mathbf{e}_1, \ldots, \mathbf{e}_p) \in \breve{\mathcal{M}}_{V',D',K',M'}$ onto a single vector $\mathbf{e}^*$ of all elements in a specific order.

This order can be specified by a positioning $G = (\mathbf{g}_1, \ldots, \mathbf{g}_p)$, to place edge $e_{i,j}$ in position $g_{i,j}$ in $\mathbf{e}^*$, that is $e^*_{g_{i,j}} = e_{i,j}$.

Definition 3. *Given E, V, M and D of size p, the set $\mathbb{G}_{E,V,D,M}$ of permissible positionings of edges in E is defined as the set of $G = (\mathbf{g}_1, \ldots, \mathbf{g}_p)$ such that*

1. *for all i, $|\mathbf{g}_i| = d_i + m_i$;*
2. $\bigcup_{i=1}^{p} \mathbf{g}_i = \{1, \ldots, \sum_{i=1}^{p}(d_i + m_i)\}$
3. *for all i, elements in $\mathbf{g}_i$ are arranged in ascending order;*
4. *For any symmetric sibling nodes ν_1 and ν_2 in tree T and any maximal antichain W_1 in subtree T_{ν_1} and maximal antichain W_2 in subtree T_{ν_2}, if $W_1 \subset V$ and $W_2 \subset V$, say $W_1 = \{v_{i_1}, \ldots, v_{i_r}\}$ and $W_2 = \{v_{j_1}, \ldots, v_{j_s}\}$ $(i_r < j_1)$, then*

$$\min\{\mathbf{g}_{i_1}, \ldots, \mathbf{g}_{i_r}\} \leq \min\{\mathbf{g}_{j_1}, \ldots, \mathbf{g}_{j_s}\}.$$

If the pseudo-root exists and has 2 symmetric children ν_1 and ν_2, and if $e_{1,1} = 1$, then it is additionally required that

$$2nd\min\{\mathbf{g}_{i_1}, \ldots, \mathbf{g}_{i_r}\} \leq \min\{\mathbf{g}_{j_1}, \ldots, \mathbf{g}_{j_s}\}.$$

Note that $(\mathbf{e}_1, \ldots, \mathbf{e}_p) \in \mathring{\mathcal{M}}_{V,D,K,M}$ are naturally merged onto a single edge vector by concatenation. In other words, the position of edge $e_{i,r} \in \mathbf{e}_i$ is defined as $\sum_{x=1}^{i-1}(d_i + m_i) + r$. Finally, $\mathbb{G}_{E',V',D',M',H}$ in (3.22) is defined as the set of position vectors G' in $\mathbb{G}_{E',V',D',M'}$ such that for any pair $(i,j) \in H$, the position $g'_{j,r}$ corresponding to the parent edge $e'_{j,r} \in \mathbf{e}'_j$ of v'_j is different from $\sum_{x<i}(d_x + m_x) + r$.

Given $V = (v_1, \ldots, v_p)$, E, M, K and G of size p, the consensus tree $C_{\mathcal{T}_V}(E)$ is constructed by grafting the trees $\mathbf{S}(T_{v_i}/\bar{\mathbf{e}}_i)$ at their roots. Edges in $C_{\mathcal{T}_V}(E)$ are named by the positioning vector G. The shape of Consensus tree $C_{\mathcal{T}_V}(E) \setminus\setminus K$ is obtained by removing k_i tips directly connected to v_i in $C_{\mathcal{T}_V}(E)$ for all i. Then, $\mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', G')$ in (3.22) is 1 if the following conditions are satisfied, 0 otherwise:

1. $C_{\mathcal{T}_V}(E) \setminus\setminus K = C_{\mathcal{T}'_{V'}}(E') \setminus\setminus K'$,

2. $k_i = (1 - m_i)|F_{v_i}|$ and $k'_j = (1 - m'_j)|F'_{v'_j}|$.

3.4.3 Recursive equations for the algorithm

We present here the key equations for the recursive derivation of $R(V, V', D, D', K, K', M, M', H)$, which is used to calculate $\gamma_S(S', d)$ through (3.23). The first theorems initialize the R values, while theorems 6, 7 and 8 enable the decomposition of R values during the recursion through the tree. All proofs are found in Appendix C-D, if not provided here.

Theorem 2. *If* $\sum_i (d_i + m_i) = \sum_j (d'_j + m'_j) = 0$ *then*

$$R(V, V', D, D', K, K', M, M', H) = \begin{cases} \prod_j N(T'_{v'_j} \setminus_m \emptyset) & \text{if } \forall i, j, k_i = |T_{v_i}|, k'_j = |T'_{v'_j}|, \\ 0 & \text{otherwise.} \end{cases}$$

In the special case of $V = (v)$ *and* $V' = (v')$,

$$R((v), (v'), (0), (0), (k), (k'), (0), (0), \emptyset) = \begin{cases} N\left(T'_{v'} \setminus_m \emptyset\right) & \text{if } k = |T_v|, k' = |T'_{v'}| \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 3. *$R = R(V, V', D, D', K, K', M, M', H) = 0$ if $\sum_i (d_i + m_i) \neq \sum_j (d'_j + m'_j)$ and if V and V' do not contain both children of the pseudo-root. Generally, $R = 0$ if $\Delta \neq \Delta'$, where $\Delta = \sum_i (d_i + m_i)$ if V does not contain both children of the pseudo-root, $\Delta = d_1 + d_2 + m_1$ otherwise. Δ' is defined similarly.*

Theorem 4. *$R = 0$ if there exists an index i satisfying at least one of the following conditions: (1) v_i is a cherry, $d_i > 0$; (2) $d_i > 0, k_i > |T_{v_i}| - 2$; (3) $d_i = m_i = 0, k_i \neq |T_{v_i}|$; (4) $m_i = 0$, $k_i + 1$ or more tips are directly connected to v_i; or (5) $d_i > |T_{v_i}| - 2$. Similarly, if there exists an index j satisfying at least one of the analogous conditions in terms of V', D', K', M', then $R = 0$.*

Theorem 5. *Consider two trees T_ν and $T'_{\nu'}$ with the same shape. Let d be the number of internal nodes in T_ν. Then $R((\nu), (\nu'), (d), (d), (k), (k), (0), (0), \emptyset) = 1$ if k is the number of tips directly connected to ν; 0 otherwise. When trees T_ν and $T'_{\nu'}$ have different shapes, $R((\nu), (\nu'), (d), (d), (k), (k), (0), (0), \emptyset) = 0$ for all k.*

The following theorems 6 and 7 decompose R into a sum of R values, where one node in V or V' is replaced by it children.

Theorem 6 (Formula dismantling a node in T)**.** *Consider $v_x \in V$ such that $m_x = 0$ and v_x has $r(\leq 3)$ internal nodes and k_0 tips as children. Let $w_1, \ldots, w_r$ be the r internal node children of v_x. We define the following sets:*

$$\begin{aligned}
\mathcal{C} \quad &= \left\{ \tilde{\mathbf{d}}, \tilde{\mathbf{m}} \,\middle|\, \tilde{\mathbf{d}} = (\tilde{d}_x, \ldots, \tilde{d}_{x+r-1}),\ \tilde{\mathbf{m}} = (\tilde{m}_x, \ldots, \tilde{m}_{x+r-1}),\ \sum_{i=x}^{x+r-1} (\tilde{d}_i + \tilde{m}_i) = d_x; \right. \\
&\qquad \left. \tilde{m}_2 = 1 \text{ and } \tilde{d}_1 + \tilde{m}_1 + \tilde{d}_2 = d_x \text{ if } \nu_x \text{ is the pseudo-root and } \tilde{m}_1 = 1 \right\}, \\
\mathcal{K}_{\tilde{\mathbf{m}}} \quad &= \left\{ \tilde{\mathbf{k}} \,\middle|\, \tilde{\mathbf{k}} = \{\tilde{k}_x, \ldots, \tilde{k}_{x+r-1}\},\ \sum_{i=x}^{x+r-1} \tilde{k}_i^{(1-\tilde{m}_i)} = k_x - k_0;\ \tilde{k}_i = 0 \text{ if } \tilde{m}_i = 1 \right\}.
\end{aligned}$$

Then

$$R(V,V',D,D',K,K',M,M',H) \;=\; \sum_{\tilde{\mathbf{d}},\tilde{\mathbf{m}}\in\mathcal{C}}\sum_{\tilde{\mathbf{k}}\in\mathcal{K}_{\tilde{\mathbf{m}}}} R(\tilde{V},V',\tilde{D},D',\tilde{K},K',\tilde{M},M',\tilde{H})$$

where $\tilde{V}$ *is similar to* V *except that* v_x *is replaced by its children. More specifically,* $\tilde{v}_i = v_i$*, for* $i \le x-1$*;* w_{i-x+1}*, for* $x \le i \le x+r-1$*;* v_{i-r+1}*, for* $i \ge x+r$*.* $\tilde{D}'$*,* $\tilde{K}'$ *and* $\tilde{M}'$ *are defined similarly. By definition,* $\tilde{H}$ *contains* (i,j) *if* $(i,j)\in H$ *and* $i \le x-1$*;* $(i+r-1,j)$ *if* $(i,j)\in H$ *and* $i \ge x+1$*. Note that* $|H| = |\tilde{H}|$*.*

Theorem 7 (Formula dismantling a node in T')**.** *Consider* $v'_x \in V'$ *such that* $m'_x = 0$ *and* v'_x *has* $r(\le 3)$ *internal nodes and* k'_0 *tips as children. Let* $w'_1,\ldots,w'_r$ *(*$r \le 3$*) be the* r *internal node children of* v'_x*. We define the following sets:*

$$\mathcal{C} \;=\; \left\{ \tilde{\mathbf{d}}',\tilde{\mathbf{m}}' \;\middle|\; \tilde{\mathbf{d}}' = (\tilde{d}'_x,\ldots,\tilde{d}'_{x+r-1}),\tilde{\mathbf{m}}' = (\tilde{m}'_x,\ldots,\tilde{m}'_{x+r-1}),\ \sum_{i=x}^{x+r-1}(\tilde{d}'_i+\tilde{m}'_i) = d'_x;\right.$$
$$\tilde{d'}_j + \tilde{m'}_j = 0 \text{ if } \tilde{d'}_{j-1} + \tilde{m'}_{j-1} = 0, \text{and if } \tilde{w'}_{j-1} \text{ and } \tilde{w'}_j \text{ are symmetric};$$
$$\left.\tilde{m}'_2 = 1 \text{ and } \tilde{d}'_1 + \tilde{m}'_1 + \tilde{d}'_2 = d'_x \text{if } v'_x \text{ is the pseudo-root and if } \tilde{m}'_1 = 1\right\},$$

$$\mathcal{K}_{\tilde{\mathbf{m}}'} = \left\{ \tilde{\mathbf{k}}' \,\middle|\, \tilde{\mathbf{k}}' = \{\tilde{k'}_x,\ldots,\tilde{k'}_{x+r-1}\},\ \sum_{i=x}^{x+r-1} \tilde{k'}_i^{(1-\tilde{m}'_i)} = k'_x - k'_0;\ \tilde{k'}_i = 0 \text{ if } \tilde{m'}_i = 1 \right\},$$

$$\mathrm{sym}_{F_{v'_x}}(v'_x) = \begin{cases} 1 & \text{if none of } \mathbf{S}(F_{w'_i}) \text{ are the same,} \\ 2 & \text{if exactly 2 of } \mathbf{S}(F_{w'_i}) \text{ are same,} \\ 3 & \text{if } r=3 \text{ and all 3 } \mathbf{S}(F_{w'_i}) \text{ are same.} \end{cases}$$

Then

$$R(V,V',D,D',K,K',M,M',H) =$$
$$\sum_{\tilde{\mathbf{d}}',\tilde{\mathbf{m}}'\in\mathcal{C}}\sum_{\tilde{\mathbf{k}}'\in\mathcal{K}_{\tilde{\mathbf{m}}'}} \left\{ R(V,\tilde{V}',D,\tilde{D}',K,\tilde{K}',M,\tilde{M}',\tilde{H}) \frac{k'_x!}{\mathrm{sym}_{F_{v'_x}}(v'_x)!\ \prod_{i=x}^{x+r-1}(\tilde{k'}_i!)^{(1-\tilde{m}'_i)}} \right\},$$

where $\tilde{V}'$ is similar to V' except that v'_x is replaced by its children, as defined by $\tilde{v}'_i = v'_i$, for $i \leq x-1$; w'_{i-x+1}, for $x \leq i \leq x+r-1$; v'_{i-r+1}, for $i \geq x+r$. $\tilde{D}'$, $\tilde{K}'$ and $\tilde{M}'$ are defined similarly. By definition, $\tilde{H}$ contains (i,j) if $(i,j) \in H$ and $j \leq x-1$; $(i, j+r-1)$ if $(i,j) \in H$ and $j \geq x+1$. Note that $|H| = |\tilde{H}|$.

Theorem 8 (Factorization formula)**.** *Consider $v_x \in V$ such that $m_x = 1$ and assume that the partial sum $\sum_{i=1}^{x-1}(d_i + m_i) = 0$. Define $\mathcal{Z}$ as the index set of nodes v'_j in V' that can be paired with v_x to defined the same bipartition, as specified below. If V and V' contain all internal node children of roots ν_0 and ν'_0, $\mathcal{Z} = \{j | (x,j) \notin H, m'_j = 1, k'_j = 0, d_x = d'_j, |T_{v_x}| = |T'_{v'_j}|, v'_j$ has no symmetric sibling in $(v'_1, \ldots, v'_{j-1})\}$. More generally we define*

$$\mathcal{Z} = \Big\{ j \,\Big|\, (x,j) \notin H, m'_j = 1, k'_j = 0, d_x = d'_j, |T_{v_x}| = |T'_{v'_x}|; \\ \forall \nu' \leq v'_j, \text{symmetric sibling of either } v'_j \text{ or its ancestor} \\ \text{and } \forall W \text{ maximal antichain in } T'_{\nu'},\ W \nsubseteq V' \Big\}.$$

Let H^* be the augmented constraint set $H^* = H \cup \{(x,j) : j \in \mathcal{Z}\}$. Then

$$R(V, V', D, D', K, K', M, M', H) = R(V, V', D, D', K, K', M, M', H^*) \\ + \sum_{j \in \mathcal{Z}} \left[\sum_{k=0}^{|T_{v_x}|} R\left((v_x), (v'_j), (d_x), (d_x), (k), (k), (0), (0), \emptyset\right) \times \right. \\ \left. R(V_{-x}, V'_{-j}, D_{-x}, D'_{-j}, K_{-x}, K'_{-j}, M_{-x}, M'_{-j}, \tilde{H}) \right]$$

where V_{-x} contains all elements in V except for v_x and we similarly define V'_{-j} and so on. We also define $\tilde{H} = \{(i,j) : i < x$ and $(i,j) \in H$, or $i \geq x$ and $(i+1, j) \in H\}$.

3.5 Approximations to the normalizing function

Although we can calculate exact values of the normalizing function $Z_L(\beta)$ through (3.15), (3.16) and the algorithm outlined in section 3.4, its computation is usually too expensive to be repeated at each iteration of an MCMC algorithm. Therefore, we propose two approximations to this normalizing function.

3.5.1 Large-L normal approximation

Recall that L denotes the number of segments and N the number of taxa. We can write

$$Z_L(\beta) = Z_1 + \zeta_L(1)e^{-\beta} + \sum_{x=2}^{D_L} \zeta_L(x)e^{-\beta x} \tag{3.24}$$

where $D_L = (L-1)(N-3)$, $Z_1 = (2N-5)!!$ was defined previously,

$$\zeta_L(x) = \#\left\{(T_1, \ldots, T_L) : \sum_{l=1}^{L-1} d(T_l, T_{l+1}) = x\right\}$$

and $\zeta_L(1)$ is easily shown to be $\zeta_L(1) = (L-1)\ 2(N-3)Z_1$. The sum in (3.24) is approximated using the following central limit theorem.

Theorem 9. *Consider independent, uniformly distributed unrooted N-taxon trees $(T_i)_{i\geq 1}$. Let $S_L = \sum_{l=1}^{L-1} d(T_l, T_{l+1})$. Then $P(S_L \leq 1)$ goes to 0 as L goes to infinity and both $(S_L - \mu_L)/\sigma_L$ and*

$$(S_L - \mu_L)/\sigma_L\ \mathbb{I}(S_L \geq 2) \xrightarrow{d} \mathcal{N}(0,1) \quad as \quad L \to \infty$$

where $\mu_L = (L-1)E(d(T_1, T_2))$ and

$$\sigma_L^2 = (L-1)\left\{\mathrm{var}(d(T_1, T_2) + \mathrm{cov}(d(T_1, T_2), d(T_2, T_3))\right\}.$$

The proof (Appendix F.1) rests on the weak dependence of the sequence $(d(T_i, T_{i+1}))_{i\geq 1}$. The second part results in a normal approximation for the sum in (3.24), from which we obtain the *normal approximation* $Z_L(\beta) \approx \hat{Z}_{(1)}$:

$$\hat{Z}_{(1)} = Z_1 + (L-1)\zeta_2(1)e^{-\beta} + \left\{Z_1^L - Z_1 - (L-1)\zeta_2(1)\right\}$$
$$\times \left[\Phi(D_L + .5; \mu_L - \beta\sigma_L^2, \sigma_L^2) - \Phi(2 - .5; \mu_L - \beta\sigma_L^2, \sigma_L^2)\right] \exp\left[-\beta\mu_L + \frac{\beta^2\sigma_L^2}{2}\right], \quad (3.25)$$

where $\Phi(\cdot; \mu, \sigma^2)$ is the cumulative distribution function of the normal distribution with mean μ and variance σ^2.

3.5.2 Independence approximation

Our second approximation is simply obtained by ignoring the dependence between distances $d(T_{l-1}, T_l)$ and $d(T_l, T_{l+1})$, for $l = 1, \ldots, L-1$. We can write

$$\begin{aligned} Z_L(\beta) &= Z_1^L E\left(e^{-\beta \sum_{l=1}^{L-1} d(T_l, T_{l+1})}\right) \approx \hat{Z}_{(2)} \\ \hat{Z}_{(2)} &= Z_1^L E\left(e^{-\beta d(T_1, T_2)}\right)^{L-1} . \end{aligned} \quad (3.26)$$

Note that $d(T_{l-1}, T_l)$ are indeed independent when there is only one possible tree shape, i.e. when $N \leq 5$. We prove in the appendix that for all N, L and all β,

$$Z_{(2)}(\beta) \leq Z_L(\beta) .$$

3.5.3 Accuracy of approximations

The proposed approximations (3.25)-(3.26) to the normalizing function are compared with the true value $Z_L(\beta)$ for various values of β, on trees with 5 taxa and 10 taxa and when the length of the alignment varies from 10 to 1000 (Figure 3.5). The normalizing function $Z_L(\beta)$ quickly drops to Z_1 as β grows. The extent of the decline

is more profound with more segments or more taxa. Since distances between tree topologies $\{d(T_i, T_{i+1}) : i \geq 1\}$ are independent when there is only one tree shape, the independence approximation $\hat{Z}_{(2)}$ in (3.26) is exact for $N \leq 5$. The large-L normal approximation $\hat{Z}_{(1)}$ in (3.25) is a good approximation except for $\beta \in (1.5, 5)$ approximately. Note that the distribution of the sum of tree distances S_L is skewed left because its mean μ_L is approximately $(L-1)(N-3-1/8)$ Steel and Penny (1993), which is very close to its maximum value $(L-1)(N-3)$. The symmetric normal approximation to the distribution of S_L is thus expected to underestimate the true probabilities at small values. These small values of x are given more weight by the exponential term in (3.24), so $\hat{Z}_{(1)}$ is expected to underestimate the true Z_L. This is indeed what we observe in Figure 3.5. The proposed approximations showed similar accuracy on 10 taxa. In particular, the independence approximation $Z_{(2)}$ is still very close to the true normalizing function.

Figure 3.6 shows the impact of using $\hat{Z}_{(2)}$ instead of the true normalizing function in terms of hyperprior densities. Although the hyperprior actually used on β has a slightly higher density than the assumed hyperprior on small β values when $\lambda = 0.01$ (Figure 3.6 (c)), the difference is small enough to be ignored. Overall, the hyperprior actually used is very close to the assumed hyperprior.

3.6 Discussion

In this work, we first show empirical evidence that the phylogenetic trees of neighboring genomic regions are correlated, in the sense that they are more similar than expected by chance. In *Escherichia* and *Shigella* genomes, the correlation between neighboring trees was shown to span across distances of about 2 kb. This is

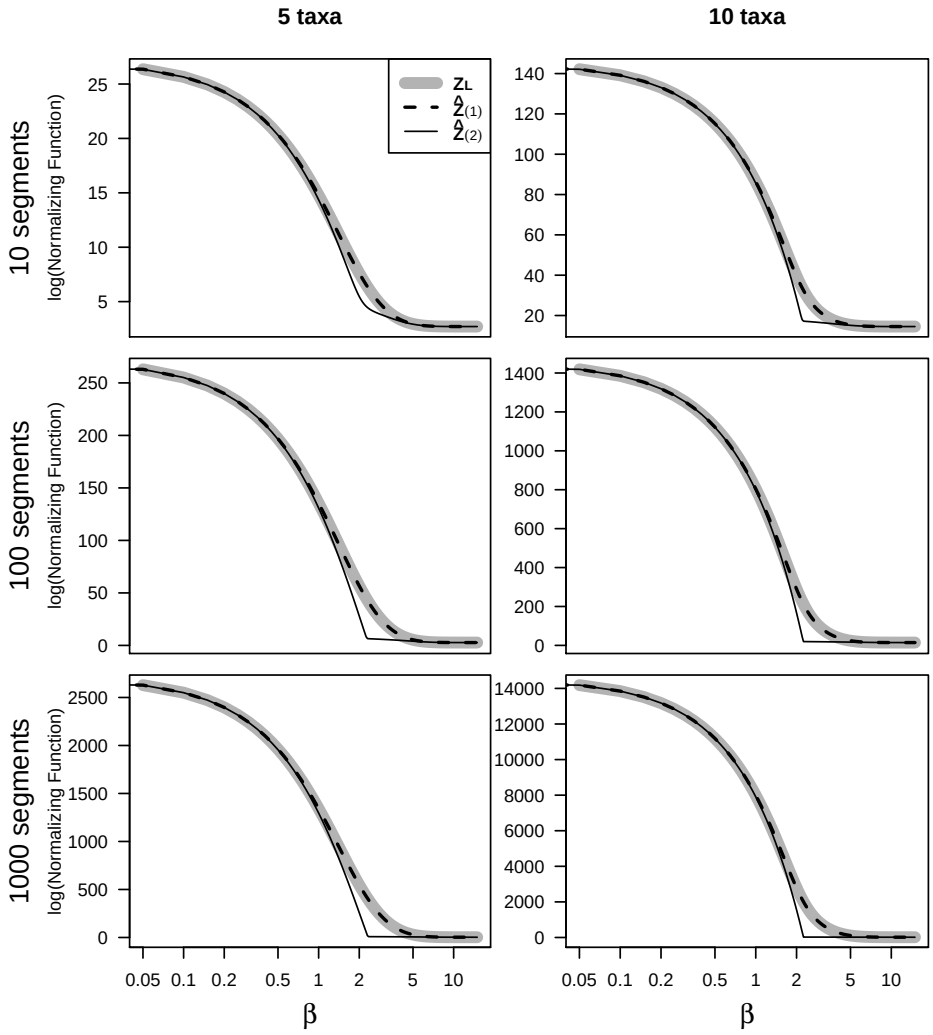

Figure 3.5: **Accuracy of approximations to the normalizing function $Z_L(\beta)$.** On 5 taxa and 10 taxa, when the number of segments is $L = 10$, $L = 100$ or $L = 1000$. The true normalizing function $Z_L(\beta)$ in the thick gray line is compared with two approximations: the normal approximation $\hat{Z}_{(1)}$ (—) and the independence approximation $\hat{Z}_{(2)}$ (- -).

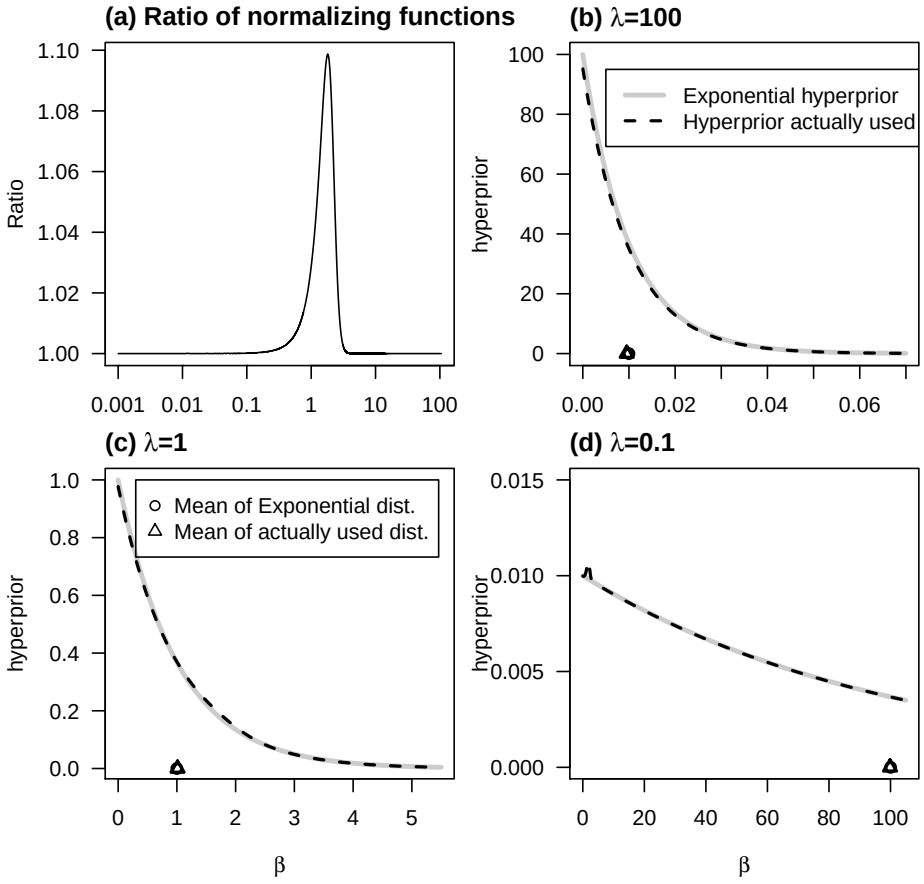

Figure 3.6: **Impact of using the independence approximation.** On 10 taxa with $L = 10$ segments. (a) Ratio of the true normalizing function to independence approximation $Z_L/\hat{Z}_{(2)}$. (b)-(d) The thick gray line indicates the exponential distribution $\mathcal{E}(\lambda)$, with mean indicated by a circle ($\circ$). The hyperprior density actually used is indicated with a dotted line (- -) with mean indicated by a triangle (Δ) when $\lambda = 100$, 1 and 0.01. Note that the axis for β in (a) is on log scale.

in support of methods that go beyond detecting gene tree discordance, towards the analysis of the dissimilarity of discordant gene trees. Leigh et al. (2011) take this approach to cluster predefined genes based on the similarity of their gene trees. We focus here on long alignments for which recombination-free loci are not predefined. We consider a Bayesian approach to simultaneously detect recombination breakpoints and phylogenetic trees based on a Gibbs prior distribution, to account for the correlation between phylogenetic trees at neighboring loci. The behaviour of the Gibbs distribution is controlled by a parameter β which scales with the inverse recombination rate per segment. The dissimilarity between tree topologies is measured by the RF distance. We show how to calculate the normalizing function of the Gibbs distribution exactly, and propose fast and accurate approximations. We thus provide the mathematical foundation for the future implementation of Gibbs-distribution based methods to simultaneously infer recombination breakpoints and the phylogenetic history of individual recombination blocks.

The RF distance may not be the ideal dissimilarity measure to quantify gene tree discordance due to recombination, because one recombination event is expected to cause the trees on the left and right side of the breakpoint to disagree by one SPR rearrangement (Song YS, 2005). However, computing the SPR distance between 2 trees is computationally heavy (Bordewich and Semple, 2005). In biomc2, the SPR distance between trees is approximated, for instance. On the other hand, computing the RF distance is fast. Additionally, there is a wide lack of tools to study the normalizing function of the Gibbs distribution based on the SPR distance. For instance, the distribution of the SPR distance between a random tree and a fixed tree, as a function of the shape of the fixed tree, is unknown. The diameter of the SPR metric space is

bounded above by $N-3$ and below by $N/2-o(N)$, where N is the number of taxa (Allen and Mike, 2001).

The core of the present work is an algorithm to calculate the joint distribution of the shape of a random tree and its RF distance to another fixed tree. This joint distribution completely determines the Gibbs distribution for the trees at 2 neighboring segments. It is then used to recursively calculate the normalizing function of the Gibbs distribution on any number of segments. The core algorithm to calculate the joint distribution of tree shape and RF distance builds on Bryant and Steel (2009), who provide the distribution of the RF distance only, based on the shape of the fixed tree. Their algorithm recursively calculates a quantity analogous to $R(v,d,k)$, where v is the root of a subtree and d relates to the RF distance between 2 subtrees. To also track the second tree shape, our algorithm needs to condition the R value on many other variables, making the algorithm much more complicated. We had to add arguments such as v', d' and k' for the other tree. To specify the shared bipartitions between two trees, additional arguments m, m' and H were introduced to avoid counting some pairs of edges multiple times, due to symmetries in tree shapes. When both trees are fixed, the complexity of the algorithm calculating $\zeta_S(S',d)$ for all RF distance values (d) depends on the shapes S and S' of the trees. If both are caterpillar trees whose shape is the most asymmetric shape a tree can have (Semple and Steel, 2003), then the algorithm runs in a polynomial time. If both trees are fully symmetric, then the algorithm has an exponential time complexity.

Two approximations to the normalizing function were proposed, and our 'independence' approximation showed excellent performance. Both approximations require the marginal distribution of the RF distance between a random tree and a fixed tree,

whose shape is known but arbitrary. This can be calculated in polynomial time (Bryant and Steel, 2009). These practical considerations are important, because the normalizing function needs to be evaluated each time a new prior inverse recombination rate β is proposed during Bayesian inference with Markov Chain Monte Carlo. Bryant and Steel (2009) also provides two approximations to their normalizing function (3.5) when β is either small or large. Their approximations cut down computing time substantially, as they do not require the distribution of the RF distance. Our attempts to use their small β and large β approximations to speed up our independence approximation resulted in large errors unfortunately, and increasingly more so as more segments were considered. Therefore, using the independence approximation instead of the normalizing function will reduce the computing time a lot without misleading the MCMC results.

Chapter 4

Discussion and Future Work

In this thesis, different aspects of gene tree discordance are studied. We first addressed the timely and important problem of distinguishing incomplete lineage sorting (ILS) from horizontal gene transfer (HGT) using multi-locus data. Using simulations, we compared the performance of two Bayesian approaches, BEST and BUCKy: BEST assumes ILS is the only reason for discordance while BUCKy uses a non-parametric clustering prior distribution. The simulation conditions are more realistic than those that are often used in this type of studies. It is the only study assessing the performance of a coalescent-based method in the presence of HGT. We found that BEST is robust to HGT events evenly placed across a species tree. We did not examine the impact of HGT on the estimated population sizes, but these are expected to be overestimated, to accommodate the extra discordance caused by HGT. When HGT is the main reason for discord and HGT events are unevenly placed, BEST showed poor performance. BUCKy was found to be robust to any biological processes of ILS or HGT. Secondly, we proposed a test to see whether the coalescent model only is enough to explain gene tree discordance or not. The test examines the overlap of the credibility intervals of CFs of two conflicting clades, which are expected to be the same under the coalescent model. The test is easy to use and powerful when the true main source of discordance is HGT unevenly placed across a species tree. When the true source is mainly ILS, or when HGT events occur evenly across a species tree, the test has low type I error or low power, respectively. Lastly, we proposed a Bayesian model to infer phylogenetic trees with recombination breakpoints along long alignments. In order to take into account the correlation between trees from neighboring genomic regions, we proposed a distance-based Gibbs distribution as a prior distribution on trees. The importance of the exact computation of the normalizing function of the

Gibbs distribution was discussed. We proposed an exact algorithm and provided fast and accurate approximations to the normalizing function.

In the rest of this chapter, we discuss future work on the Bayesian model to detect recombination breakpoints and infer gene trees.

4.1 Model for sequence evolution and tree branch lengths

We specify here the likelihood model and the prior distribution on branch lengths, which provide the basis for the tree estimation at each segment along the alignment. Along an alignment $\mathbf{X}$ from N taxa, each site can be a candidate recombination breakpoint, and hence sites may have different underlying phylogenetic tree topologies. If we estimate a tree topology at each individual site, the uncertainty in tree topology estimation can be unreasonably large. Therefore, we assume that the alignment can be divided into L pre-defined arbitrary short segments, $(\mathbf{X}_1, \ldots, \mathbf{X}_L)$, and that segments may have different tree topologies $\mathbf{T} = (T_1, \ldots, T_L)$ because of recombination. Within segment i, sites $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p_i})$ are assumed to have the same phylogenetic topology, where p_i is the number sites in the ith segment. We use a standard evolutionary model (Felsenstein, 2004), where nucleotide substitutions at a given site are explained by a continuous-time Markov chain. The evolutionary history of the alignment is described by a vector $(\mathbf{T}, \mathbf{u})$ of phylogenetic trees, where $\mathbf{u} = (\mathbf{u}_{1,1}, \ldots, \mathbf{u}_{1,p_1}, \ldots, \mathbf{u}_{L,1}, \ldots, \mathbf{u}_{L,p_L})$ is a matrix of tree branch lengths and $\mathbf{u}_{i,j} = (u_{i,j,1}, \ldots, u_{i,j,2N-3})'$ is the vector of branch lengths on T_i at the jth site of segment i. We assume an HKY rate matrix Q (Hasegawa et al., 1985) with transition/transversion ratio $\kappa = (\kappa_1, \ldots, \kappa_L)$ and stationary distribution of nucleotides $\pi = (\pi_1, \ldots, \pi_L)$. The transition/transversion ratio and the stationary nucleotide fre-

quencies are assumed constant within each segment. For the likelihood calculation, we arbitrarily choose an internal node in each tree to act as the root. The likelihood is then:

$$L(\mathbf{X}|\mathbf{T},\mathbf{u},\kappa,\pi)=\prod_{i=1}^{L}L(X_i|T_i,(\mathbf{u}_{i,1},\ldots,\mathbf{u}_{i,p_i}),\kappa_i,\pi_i)=\prod_{i=1}^{L}\prod_{j=1}^{k_i}L(X_{ij}|T_i,\mathbf{u}_{i,j},\kappa_i,\pi_i),$$

where the site likelihood is

$$L(X_{ij}|T_i,\mathbf{u}_{i,j},\kappa_i,\pi_i)$$
$$=\sum_{a(x_{i,j,N})}\cdots\sum_{a(x_{i,j,2N-2})}\pi_i(x_{i,j,2N-2})\prod_{n=1}^{2N-3}P(x_{i,j,n}|a(x_{i,j,n}),u_{i,j,n},\kappa_i,\pi_i),\quad(4.1)$$

where $x_{i,j,n}$ is a nucleotide at node n on T_i at the jth site of segment i, $x_{i,j,2N-2}$ is the arbitrarily chosen root of tree T_i, $a(x_{i,j,n})$ is the nucleotide of the ancestor of node n and $P(x_{i,j,n}|a(x_{i,j,n}),u_{i,j,n},\kappa_i)$ is the transition probability from the ancestor to node n. For convenience, under the HKY model, terminal nodes are indexed by $n=1,\ldots,N$ and internal nodes by $n=N+1,\ldots,2N-2$. In other words, $u_{i,j,n}$ represents an external branch length for $n=1,\ldots,N$; and an internal branch length for $n\geq N+1$.

Branch lengths on a phylogenetic tree are correlated across sites, but ignoring the correlation allows for an easy integration of branch lengths. Many recent papers (Minin et al., 2005; de Oliveira Martins et al., 2008) assume that branch lengths are independent across sites, which allows them to integrate branch lengths out analytically in a Bayesian framework. This reduced computational burden is particularly helpful in the context of inferring a potentially large number of trees along a single alignment. With the premise of independent branch lengths across sites in this "no-common-mechanism" (NCM) model, no information about the rate of change of a character is shared across sites (Tuffley and Steel, 1997). In a maximum likelihood

framework, the NCM model gives rise to the same tree estimate as that given by maximum parsimony, which may be inconsistent (Goldman, 1990; Tuffley and Steel, 1997). Therefore, we propose here an intermediate model that allows branch lengths to be independent across sites but we introduce more structure in the prior distribution of branch lengths in order to allow for some level of information sharing across sites. More specifically, we assume for an external branch $u_{i,j,n}$ $(n = 1, \ldots, N)$ at the jth site of segment i, that

$$u_{i,j,n} \sim \text{Exponential}(1/\mu_n)$$

and for an internal branch $u_{i,j,n}$ $(n = N+1, \ldots, 2N-3)$, we use

$$u_{i,j,n} \sim \text{Exponential}(1/\mu_I),$$

where μ_n for $n = 1, \ldots, N$ is the prior mean of the external branch length $u_{i,j,n}$ leading to taxon n across all sites and μ_I is the common prior mean of internal branch lengths. With these different prior means, we retain the computational benefit of independent branch lengths across sites, and we hope to achieve some sharing of information across sites about the length of external branches.

With this independence assumption (given prior means), branch lengths are integrated out analytically and the marginal likelihood has expected branch lengths as parameters. The probability of substitutions on a branch $P(x_{i,j,n}|a(x_{i,j,n}), u_{i,j,n}, \kappa_i, \pi_i)$ is obtained from the transition probability matrix $\exp(u_{i,j,n}Q)$, where Q is the HKY substitution rate matrix. Then the marginal transition probability becomes

$$\int P(x_{i,j,n}|a(x_{i,j,n}), u_{i,j,n}, \kappa_i, \pi_i) p(u_{i,j,n}|\mu_n) du_{i,j,n} = \widetilde{P}(x_{i,j,n}|a(x_{i,j,n}), \mu_n, \kappa_i, \pi_i)$$

where $p(u_{i,j,n}|\mu_n)$ is the density function of $\mathcal{E}(1/\mu_n)$, $\mu_n = \mu_I$ if $u_{i,j,n}$ is an internal branch length and the integrated transition probability matrix $\widetilde{P}$ is $(I - \mu_n Q)^{-1}$. We use the marginal likelihood as follows:

$$L(\mathbf{X}|\mathbf{T}, \mu, \kappa, \pi) = \prod_{i=1}^{L} \prod_{j=1}^{k_i} L(X_{ij}|T_i, \mu, \kappa_i, \pi_i), \tag{4.2}$$

where

$$L(X_{ij}|T_i, \mu, \kappa_i, \pi_i) = \sum_{a(x_{i,j,N})} \cdots \sum_{a(x_{i,j,2N-5})} \pi_i(a(x_{i,j,2N-3})) \prod_{n=1}^{2N-3} \widetilde{P}(x_{i,j,n}|a(x_{i,j,n}), \mu_n, \kappa_i, \pi_i)$$

and $\mu = (\mu_1, \ldots, \mu_N, \mu_I)$. We futher assume that the μ_i's are independent with hyperprior distributions $\mathcal{E}(1/\mu_0)$. This approach is computationally tractable despite a potentially large number of trees, because individual branch lengths do not need to be tracked during the MCMC.

Assessment of our branch length model

Several studies showed that the prior distribution on branch lengths can affect the estimation of phylogenetic tree topologies in important ways (Yang and Rannala, 2005; Wu et al., 2008; Huelsenbeck et al., 2008). Huelsenbeck et al. (2008) showed that the NCM Bayesian model can lead to inconsistent tree topology estimation, when applied to a single locus. We will use simulations to determine if our structured prior distribution on branch lengths reduces the opportunity of inconsistent topological estimation, on a single locus. Our prior assumption on branch lengths on one tree – independence across sites and branches but the same prior mean for internal branches across sites – will be compared with the prior assumption used in other models (Minin et al., 2005; Bloomquist et al., 2009; Husmeier and Mantzaris, 2008; de Oliveira Martins et al., 2008) – independence across sites and branches but the same prior mean

for all branches – in terms of the accuracy of tree topology estimation.

4.2 Implementation of the proposed Bayesian model to detect recombination breakpoints and infer gene trees

We will modify biomc2 (de Oliveira Martins et al., 2008) with the new Gibbs distribution (3.10) as our prior distribution on tree topologies, and with the our prior distribution on branch lengths. We will use the same Monte Carlo Markov Chain (MCMC) scheme.

Monte Carlo Markov Chain scheme

The posterior distribution is simulated through a Gibbs sampler where each parameter is independently and sequentially updated. The main interesting parameters in our proposed model are tree topologies of segments $T = (T_1, \ldots, T_L)$, the inverse recombination rate β and the average branch lengths $\mu = (\mu_1, \ldots, \mu_N, \mu_I)$. The average branch lengths are updated by random perturbations where the new state μ_i^* is sampled as $\mu_i^* = \mu_i e^{\xi_i(U-0.5)}$, where $U \sim \text{Uniform}(0,1)$ and ξ_i is an arbitrary tuning constant. The proposal density is $q(\mu_i^*|\mu_i) = 1/(\mu_i^* \xi_i)$. Then the proposal ratio is $q(\mu_i|\mu_i^*)/q(\mu_i^*|\mu_i) = \mu_i^*/\mu_i$ and hence the proposed μ_i^* is accepted with probability $a(\mu_i^*) = \min\{1, A(\mu_i^*)\}$, where

$$A(\mu_i^*) = \frac{L(X|T, \mu = (\mu_1, \ldots, \mu_i^*, \ldots, \mu_I), \kappa, \pi)}{L(X|T, \mu = (\mu_1, \ldots, \mu_i, \ldots, \mu_I), \kappa, \pi)} \frac{\exp(\mu_i^*/\mu_0)}{\exp(\mu_i/\mu_0)} \frac{\mu_i^*}{\mu_i}.$$

In the same way, the inverse recombination rate β^* is proposed and the acceptance probability reduces to

$$A(\beta^*) = \frac{p(\mathbf{T}|\beta^*)}{p(\mathbf{T}|\beta)} \frac{\exp(\lambda\beta^*)}{\exp(\lambda\beta)} \frac{\beta^*}{\beta},$$

when $\beta \sim \mathcal{E}(\lambda)$ a priori. If we want to allow that β changes 10% at most at each

step, the tuning parameter ξ_β should be 0.2 Therefore, the choice $\xi_i = \xi_\beta = 0.2$ is reasonable. Note that, whenever β is proposed, we need to calculate the normalizing function or its approximation.

The distances between trees and the locations of recombination breakpoints are updated as a consequence of updating the tree topologies, using several proposal types. The update of topologies is simultaneously performed in blocks of segments sharing the same topology. The first proposal type updates the topology of one entire block. Let the current partition have breakpoints before segment a and after segment b, that is, $T_a = \ldots = T_b (= T)$, $T_{a-1} \neq T_a$ and $T_b \neq T_{b+1}$. A new tree $T^* = T_a = \ldots = T_b$ is proposed by applying one SPR or nearest neighbor interchange (NNI) move at the current tree as in biomc2. Then, the proposal density is $q(T^*|T) = 1/\{2(N-3)\}$ if the new topology is sampled by applying an NNI on the current topology T, and $q(T^*|T) = 1/\{2(N-3)(2N-7)\}$ if T^* is sampled by applying an SPR to the current topology T (Allen and Mike, 2001). In both cases, the proposal ratio is 1, and the proposed tree T^* will be accepted with probability

$$A(T^*) = \frac{p(X_{a,b}|T^*, \mu, \kappa_i, \pi_i)}{p(X_{a,b}|T, \mu, \kappa_i, \pi_i)} \frac{\exp\{-\beta(d(T_{a-1}, T^*) + d(T^*, T_{b-1})\}}{\exp\{-\beta(d(T_{a-1}, T) + d(T, T_{b-1})\}} \tag{4.3}$$

where $X_{a,b} = (X_a, \ldots, X_b)$. In proposing a new topology T^* to a group of consecutive segments sharing the same tree, the proposed tree may by chance be equal to either one of the two trees T_{a-1} or T_{b+1} at the neighboring blocks. This would change the total number of breakpoints, and would prevent the move back by the same proposal type, and so this would need to be rejected. Consequently, if T_{a-1} or T_{b+1} is one NNI move or one SPR move away from T, then $q(T^*|T)$ is slightly modified.

Another proposal type will be added to shift a current recombination breakpoint,

using a uniform distribution for the new location reflected at the boundaries. This proposal conveniently has a proposal ratio of 1. A third proposal type is used for the removal of one recombination breakpoint and the addition of a new breakpoint. For the removal of a breakpoint, suppose that a block goes from segments a to b. If the current tree $T = T_a = \ldots = T_b$ is at most k NNI moves or SPR moves away from T_{a-1} (or T_{b+1}), then the new tree T^* for the block is proposed to be T_{a-1} (or T_{b+1}). Our restriction differs from the current proposal in biomc2, but is necessary for the reverse move to be possible, because the addition of breakpoints only proposes new trees that are no more than a fixed distance away from the current tree. The addition of a new breakpoint in a block from segment a to b will be proposed using a uniform distribution on the possible new breakpoint location within the block. Once a breakpoint l^* is proposed, a tree for the block either on the left or on the right of the breakpoint l^* (with the same probability) should be also proposed, using at most k NNI or SPR moves for the current tree. More details are found in de Oliveira Martins et al. (2008)

Summary of sampled recombination breakpoints

Along the MCMC simulation, we obtain sampled trees of segments which form the posterior distribution of trees. At the same time, the posterior distribution of distances between trees at neighboring loci is estimated. de Oliveira Martins and Kishino (2010) proposed three ways to summarize these posterior samples to provide point estimates of recombination breakpoints. For each candidate breakpoint, first of all, the average distance between the trees on the left and on the right is calculated over all samples. If the average distance is larger than one, then the point is indicated as a recombination breakpoint. An alternative method is based on the estimated

number of recombination breakpoints. Given either the average or the mode of the number of recombination breakpoints, say b, we can find the first b breakpoints with highest credibility of having non-zero RF distance. The third method is to find the "centroid mosaic structure", a vector of breakpoint locations along the alignment that is as similar to all sampled vectors of breakpoints as possible (de Oliveira Martins and Kishino, 2010). More specifically, let b_i be the number of breakpoints in the ith MCMC sample and $\mathbf{a}_i = (a_{i,0}, \ldots, a_{i,b_i}, a_{i,b_i+1})$ be the corresponding vector of recombination breakpoint locations, where $a_{i,0} = 0$ and a_{i,b_i+1} is the length of the alignment. Then we define the distance between two vectors a_i and a_j as follows

$$D(\mathbf{a}_i, \mathbf{a}_j) = \sum_{s=1}^{b_i} \{\min_{k \geq 0} |a_{i,s} - a_{j,k}|\} + \sum_{k=1}^{b_j} \{\min_{s \geq 0} |a_{i,s} - a_{j,k}|\}.$$

Then the centroid mosaic structure a^* is the mosaic structure that minimizes the total distance $\sum_{i=1}^{n} D(a^*, a_i)$, where n is the number of MCMC samples. Here we make the simplifying assumption that the centroid can be found among the samples, but there may be other mosaics, not present in the posterior samples, with a total distance even smaller.

4.3 Simulations to study the performance of the proposed Bayesian model

A simulation study will be conducted to compare the performance of the proposed method with other existing methods such as biomc2 (de Oliveira Martins et al., 2008), ClonalOrigin (Didelot et al., 2010) and cBrother (Fang et al., 2007). We expect that our proposed method is more accurate than cBrother and ClonalOrigin when the segment trees are similar to each other, and generally better than biomc2 because of

the miscalculated normalizing function in biomc2. Moreover, the proposed method scales well with large data sets, while cBrother is available up to eight taxa only. Real data analysis will also be conducted on whole-genome alignments of closely-related Enterobacteria generated in Nicole Perna's lab.

4.4 Inference of concordance factors and concordance trees

BUCKy estimates the joint posterior distribution of gene trees using a prior gene tree distribution based on the Dirichlet process and then estimates CFs of clades and a concordance tree from the joint posterior distribution. Since the proposed Bayesian model estimates the joint posterior distribution of gene trees, we can also estimate CFs of clades and build concordance trees. The CF of a clade here is defined as the proportion of sites (as opposed to the proportion of loci) that include the clade in their phylogenetic trees. A concordance tree can be also estimated from the CFs of clades. Then this concordance tree reflects the primary signal of vertical inheritance and CFs measure the magnitude of horizontal signal affecting a particular clade. Therefore, the concordance tree and the CFs can help summarize both the vertical history and the horizontal history of the sampled organisms.

4.5 Impact of the study of recombination detection on biological research

Detecting and estimating recombination in microbial pathogens can be used to understand pathogen evolution and to identify medically relevant loci. (Awadalla, 2003; Touchon et al., 2009; Rasko et al., 2001; Spratt et al., 2001; Yan et al., 2008). For example, we consinder sequences of pathogenic organisms where homologous se-

quences flank sequences that are absent from nonpathogenic organisms of the same or closely related species. By homologous recombination events, the recipient (nonpathogenic organisms) may acquire an insertion from another strain of unrelated bacteria (pathogenic organisms). Large insertions that have been acquired from pathogenic organisms and are absent from related strains of bacteria are called "pathogenicity islands". Core genes flanking such pathogenicity islands show higher rates of recombination (Touchon et al., 2009). The study of pathogenicity islands could be helpful using the tools developed in this work, by looking for (1) regions with a high density of breakpoints, (2) breakpoints where the RF distance between the two neighboring loci is particularly high, suggesting HGT over a long evoutionary distance, or (3) region where the segment tree is particularly discordant with the concordance tree.

Appendices

Appendix A

Number of Recombination Breakpoints and Recombination Rate

A.1 Fundamental properties of Gibbs distributions

The following lemma will be used to prove Proposition 1, which links the distribution of the number of breakpoints to the parameter β and shows that β scales with the inverse recombination rate per segment.

Lemma 1. *Consider two tree sequences of different lengths $K > L$, $\{T_l; l = 1, \ldots, L\}$ and $\{V_k; k = 1, \ldots, K\}$, both following the Gibbs distribution (9). Then*

i) $P(T_1 = t_1, \ldots, T_L = t_L) = P(V_1 = t_1, \ldots, V_L = t_L | V_L = V_{L+1} = \ldots = V_K)$.

ii) For any fixed sequence of indices $1 \leq k_1 < \cdots < k_L = K$, we have that $P(T_1 = t_1, T_2 = t_2, \ldots, T_L = t_L) = P(V_{k_1} = t_1, \ldots, V_{k_L} = t_L | V_1 = \ldots = V_{k_1}, V_{k_1+1} = \ldots = V_{k_2}, \ldots, V_{k_{L-1}+1} = \ldots = V_{k_L})$.

iii) For any subset $\{i_1, \ldots, i_l\}$ of $\{1, \ldots, L\}$ of size l, $P(T_{i_1} = T_{i_1+1}, T_{i_2} = T_{i_2+1}, \ldots, T_{i_l} =$

$T_{i_l+1}) = Z_{L-l}(\beta)/Z_L(\beta)$.

iv) $P(T_1 \neq T_2, T_2 \neq T_3, \ldots, T_{L-1} \neq T_L) = \left(\sum_{i=0}^{L-1} \binom{L-1}{i}(-1)^{L-1-i} Z_{i+1}\right)/Z_L$.

Proof. i) It is easy to see that $P(V_L = \ldots = V_K) = Z_L(\beta)/Z_K(\beta)$. Consequently,

$$\begin{aligned} P(V_1 = t_1, \ldots, V_L = t_L | V_L = \ldots = V_K) &= e^{-\beta \sum_{i=1}^{L-1} d(t_i, t_{i+1})}/Z_L(\beta) \\ &= P(T_1 = t_1, \ldots, T_L = t_L)\,. \end{aligned}$$

ii) Once again, it is easy to see that $P(V_1 = \cdots = V_{k_1}, \ldots, V_{k_{L-1}+1} = \cdots = V_{k_L}) = Z_L(\beta)/Z_K(\beta)$ and that $P(V_1 = \cdots = V_{k_1} = t_1, \ldots, V_{k_{L-1}+1} = \cdots = V_{k_L} = t_L) = e^{-\beta \sum_{i=1}^{L-1} d(t_i, t_{i+1})}/Z_K(\beta)$. Therefore, $P(V_{k_1} = t_1, \ldots, V_{k_L} = t_L | V_1 = \ldots = V_{k_1}, \ldots, V_{k_{L-1}+1} = \ldots = V_{k_L}) = e^{-\beta \sum_{i=1}^{L-1} d(t_i, t_{i+1})}/Z_L(\beta)$, which is also $P(T_1 = t_1, T_2 = t_2, \ldots, T_L = t_L)$.

iii) Let $i_1, \ldots, i_l$ be l distinct indices in $\{1, \ldots, L\}$. Then

$$\begin{aligned} &P(T_{i_1} = T_{i_1+1}, T_{i_2} = T_{i_2+1}, \ldots, T_{i_l} = T_{i_l+1}) \\ &\qquad = \sum_{t_j, j \notin \{i_1, \cdots, i_l\}} \exp(-\beta \sum_{j \notin \{i_1, \cdots, i_l\}} d(t_j, t_{j+1}))/Z_L(\beta) = Z_{L-l}(\beta)/Z_L(\beta). \end{aligned}$$

iv) Define the events $E_i = \{T_i = T_{i+1}\}$. By the principle of inclusion and exclusion, $P(E_1 \cup \ldots \cup E_{L-1}) = \sum_{j=1}^{L-1} (-1)^{j-1} \sum_{\{i_1, \ldots, i_j\} \in \{1, \ldots, L-1\}} P(E_{i_1} \cap \ldots \cap E_{i_j})$. Using $P(E_{i_1} \cap \ldots \cap E_{i_j}) = \dfrac{Z_{L-j}(\beta)}{Z_L(\beta)}$ by Lemma 1 (iii), we get that $P(T_1 \neq T_2, T_2 \neq T_3, \ldots, T_{L-1} \neq T_L) = P(E_1^c \cap \ldots \cap E_{L-1}^c) = 1 - P(E_1 \cup \ldots \cup E_{L-1}) = 1 - \sum_{j=1}^{L-1} (-1)^{j-1} \binom{L-1}{j} \dfrac{Z_{L-j}(\beta)}{Z_L(\beta)} = \left\{ \sum_{i=0}^{L-1} \binom{L-1}{i} (-1)^{L-1-i} Z_{i+1}(\beta) \right\} \Big/ Z_L(\beta)$.

□

A.2 Proof of Proposition 1

For any placement $1 \leq i_1 < \ldots < i_b < L$ of exactly b breakpoints we have

$$\begin{aligned} & P\left(T_1 = \ldots = T_{i_1} = t_1, T_{i_1+1} = \ldots = T_{i_2} = t_2, \ldots, T_{i_{b-1}+1} = \ldots = T_{i_b} = t_b, \right. \\ & \qquad \left. T_{i_b+1} = \ldots = T_L = t_{b+1}\right) \\ & = P(T_1 = t_1, T_2 = t_2 \ldots, T_b = t_b, T_{b+1} = \ldots = T_L = t_{b+1}) \\ & = e^{-\beta\{d(t_1,t_2)+\ldots+d(t_b,t_{b+1})\}}/Z_L(\beta). \end{aligned}$$

The number of ways to choose b breakpoints out of $L-1$ candidate breakpoints is $\binom{L-1}{b}$, so the distribution of the number of breakpoints is

$$\begin{aligned} P(B=b) &= \binom{L-1}{b} P(T_1 \neq T_2, \ldots, T_b \neq T_{b+1}, T_{b+1} = \cdots = T_L) \\ &= \binom{L-1}{b} \sum_{i=0}^{b} \frac{\binom{b}{i}(-1)^{b-i} Z_{i+1}(\beta)}{Z_b(\beta)} \frac{Z_b(\beta)}{Z_L(\beta)} \end{aligned}$$

by Lemma 1, which proves the first part of Proposition 1. To calculate the expected number of recombination breakpoints, we simply recognize that

$$E(B) = E\left(\sum_{i=1}^{L-1} 1_{T_i \neq T_{i+1}}\right) = (L-1)P(T_1 \neq T_2)$$

by homogeneity of the Gibbs distribution in Lemma 1 (iii), which also implies that $P(T_1 \neq T_2) = 1 - P(T_1 = T_2) = 1 - Z_{L-1}(\beta)/Z_L(\beta)$. This completes the second part of Proposition 1: $E(B) = (L-1)\{1 - Z_{L-1}(\beta)/Z_L(\beta)\}$.

A.3 Approximated recombination rate for large β

Under the Gibbs distribution, Proposition 1 shows that the recombination rate per segment is $1 - Z_{L-1}/Z_L$. When β is large, i.e. $\epsilon = e^{-\beta}$ is small, we can be rewrite

the recombination rate as

$$\begin{aligned} f(\epsilon) &= 1 - \left(\sum_{x=0}^{(N-3)(L-2)} \zeta_{L-1}(x)\epsilon^x \right) \Big/ \left(\sum_{x=0}^{(N-3)(L-1)} \zeta_L(x)\epsilon^x \right) \\ &= 1 - \frac{Z_1 + 2(L-2)(N-3)Z_1\epsilon + O(\epsilon^2)}{Z_1 + 2(L-1)(N-3)Z_1\epsilon + O(\epsilon^2)}, \end{aligned}$$

where $\zeta_L(x) = \#\{(T_1, \ldots, T_L) : \sum_{i=1}^{L-1} d(T_i, T_{i+1}) = x\}$ is as defined in the main text. Using a Taylor expansion, we get that the recombination rate per segment is $f(\epsilon) = 0 + 2(N-3)\epsilon + O(\epsilon^2) \sim 2(N-3)e^{-\beta}$.

Appendix B

Proof of Theorem 1 on the at-least Generating Function

B.1 Subtree-shape equivalence classes

We start by proving that the definition of representatives in Definition 2 identifies exactly one member of each shape-equivalence class. We show this by induction on the size d of edge vectors. Let two subtree-shape equivalent edge vectors $\mathbf{e}$ and $\mathbf{e}'$ satisfy the conditions in Definition 2. Since $\mathbf{S}(T\backslash_m \mathbf{e}) = \mathbf{S}(T\backslash_m \mathbf{e}'))$, e_n and e'_n are symmetric for each $n = 1, \ldots, d$. For $d = 1$ the conditions then require that $e_1 \leq e'_1$ and $e_1 \geq e'_1$, hence $e_1 = e'_1$. Now assume that $(e_1, \ldots, e_{d-1}) = (e'_1, \ldots, e'_{d-1})$. We want to show that $e_d = e'_d$. Since e_d and e'_d are symmetric we can apply condition (b) twice to get both $e_d \leq e'_d$ then $e'_d \leq e_d$, hence $e_d = e'_d$. Condition (b) applies indeed to $\mathbf{e}$ and $e' = e'_d$ because $\mathbf{S}(T\backslash_m \mathbf{e}) = \mathbf{S}(T\backslash_m \mathbf{e}')$ and therefore (i) if e_d is a descencent of e_i then e'_d is also a descencent of e_i, and (ii) if e_d is not comparable to e_i then e'_d is not either.

B.2 Proof of Theorem 1

The following lemma will be used to prove Theorem 1.

Lemma 2. *Let T and T' be two tree topologies, and $\mathbf{e}$ and $\mathbf{e}'$ be vectors of edge labels in T and T' respectively. If $T\backslash_m \mathbf{e} = T'\backslash_m \mathbf{e}'$, then T and T' are identical.*

Proof. T can be reconstructed from the forest $T\backslash_m \mathbf{e}$ by reconnecting pseudo-terminal nodes with the same label. Similarly, T' can be entirely reconstructed from $T'\backslash_m \mathbf{e}'$. Since both pseudo-terminal nodes corresponding to e_i are labeled by the index i, the reconstruction process will yield the same tree if $T\backslash_m \mathbf{e} = T'\backslash_m \mathbf{e}'$, which would then imply $T = T'$.

□

We now prove that $\Gamma_{S,S'}(x)$ in (3.21) is the "at-least" generating function $U_{S,S'}(x)$ in (3.17). We first decompose $u_S(S', d)$ in (3.18) as

$$u_S(S', d) = \sum_{\mathbf{e}\in\mathring{\mathcal{E}}(T),|\mathbf{e}|=d} |B_{S,\mathbf{e}}(S')|, \tag{B.1}$$

where $B_{S,\mathbf{e}}(S') = \{T' : \mathbf{S}(T') = S', T'$share the bipartitions defined by $\mathbf{e}$ on T and possibly others$\}$. In order to decompose $B_{S,\mathbf{e}}(S')$ further, we use mutually exclusive sets $\{B_{S,\mathbf{e}}(S', \mathbf{e}') : \mathbf{e}' \in \breve{\mathcal{E}}(S'), |\mathbf{e}'| = |\mathbf{e}|\}$, where

$$B_{S,\mathbf{e}}(S', \mathbf{e}') = \{T' : \mathbf{S}(T') = S', \text{bipartitions by } \mathbf{e} \text{ on } T = \text{bipartitions by } \mathbf{e}' \text{ on } T'\}.$$

To show that these sets are mutually exclusive, consider a tree $T' \in B_{T,\mathbf{e}}(S', \mathbf{e}') \cap B_{T,\mathbf{e}}(S', \mathbf{e}'')$ for $\mathbf{e}'$ and $\mathbf{e}''$ in $\breve{\mathcal{E}}(T')$. Since bipartitions defined by $\mathbf{e}'$ on tree T' are the same as those defined by $\mathbf{e}''$ on the same tree T', forest $T'\backslash_m \mathbf{e}' = T'\backslash_m \mathbf{e}''$. Then

$\mathbf{S}(T'\backslash_m \mathbf{e}') = \mathbf{S}(T'\backslash_m \mathbf{e}'')$ so $\mathbf{e}'$ and $\mathbf{e}''$ are in the same equivalence class and $\mathbf{e}' = \mathbf{e}''$ because $\breve{\mathcal{E}}(T')$ has only one representative per class.

To show that $B_{S,\mathbf{e}}(S')$ is covered by the sets $B_{T,\mathbf{e}}(S', \mathbf{e}')$ for $\mathbf{e}' \in \breve{\mathcal{E}}(T')$, we first notice that $B_{S,\mathbf{e}}(S') = \cup_{|\mathbf{e}'|=d} B_{S,\mathbf{e}}(S', \mathbf{e}')$ where the union is over all vectors $\mathbf{e}'$. For any T' in $B_{S,\mathbf{e}}(S')$, there exists $\mathbf{e}'$ on T' such that the same bipartitions are defined by $\mathbf{e}$ on T and by $\mathbf{e}'$ on T'. Let $\mathbf{e}^*$ be the representative of the subtree-shape equivalence class of $\mathbf{e}'$. Since $T'\backslash_m \mathbf{e}^*$ has shape $\mathbf{S}(T'\backslash_m \mathbf{e}')$, we can construct a tree T^* such that $T^*\backslash_m \mathbf{e}^* = T'\backslash_m \mathbf{e}'$. Since bipartitions defined by $\mathbf{e}^*$ on T^* are the same as those defined by $\mathbf{e}$ on T, T^* is in $B_{S,\mathbf{e}}(S', \mathbf{e}^*)$. By Lemma 2, T' and T^* are identical. It follows that T' is in $B_{S,\mathbf{e}}(S', \mathbf{e}^*)$ and finally $B_{S,\mathbf{e}}(S') = \cup_{\substack{\mathbf{e}' \in \breve{\mathcal{E}}(T') \\ |\mathbf{e}'|=d}} B_{S,\mathbf{e}}(S', \mathbf{e}')$, where the union is over mutually exclusive sets. Therefore, we get that

$$u_S(S', d) = \sum_{\mathbf{e} \in \mathring{\mathcal{E}}(T), |\mathbf{e}|=d} \; \sum_{\mathbf{e}' \in \breve{\mathcal{E}}(T'), |\mathbf{e}'|=d} |B_{S,\mathbf{e}}(S', \mathbf{e}')|.$$

To finish the proof of Theorem 1, we just need to show that

$$|B_{S,\mathbf{e}}(S', \mathbf{e}')| = N(T'\backslash_m \mathbf{e}') \mathbb{I}_{\mathbf{S}(T/\bar{\mathbf{e}}) = \mathbf{S}(T'/\bar{\mathbf{e}'})}$$

for any fixed tree T' of shape S', to show that $u_S(S', d)$ is equal to $\gamma_S(S', d)$ in (3.19). First, notice that the existence of a tree T^* in $B_{S,\mathbf{e}}(S', \mathbf{e}')$ implies that the relationships among the edges in $\mathbf{e}$ is the same as the relationships among the edges in $\mathbf{e}'$ in the sense that $T/\bar{\mathbf{e}} = T^*/\bar{\mathbf{e}}'$. Therefore their shapes are equal and $\mathbf{S}(T/\bar{\mathbf{e}}) = \mathbf{S}(T^*/\bar{\mathbf{e}}') = \mathbf{S}(T'/\bar{\mathbf{e}}')$. Conversely, if $\mathbf{S}(T/\bar{\mathbf{e}}) = \mathbf{S}(T'/\bar{\mathbf{e}}')$ then it is easy to see that $B_{S,\mathbf{e}}(S', \mathbf{e}') \neq \emptyset$. Assume now that $\mathbf{S}(T/\bar{\mathbf{e}}) = \mathbf{S}(T'/\bar{\mathbf{e}}')$. All we need to show is that $|B_{S,\mathbf{e}}(S', \mathbf{e}')| = N(T'\backslash_m \mathbf{e}')$, which was defined in (3.20) as $\prod_{i=1}^{|\mathbf{e}|+1} |\mathcal{F}_i|$ where $\mathcal{F}_i = \{F : \exists \boldsymbol{\sigma}_{\mathcal{L}_i}$ such that $\boldsymbol{\sigma}_{\mathcal{L}_i}(F) = F_i\}$, the F_i trees are the members of $T\backslash_m \mathbf{e}$. Is is easy to see

that $B_{S,\mathbf{e}}(S', \mathbf{e}')$ is in bijection with $\prod_i^{|\mathbf{e}|+1} \mathcal{F}_i$, so $|B_{S,\mathbf{e}}(T', \mathbf{e}')| = \prod_i |\mathcal{F}_i| = N(T' \backslash_m \mathbf{e}')$ follows and completes the proof of Theorem 1.

Appendix C

Decomposition of $N(T\backslash_m \mathbf{e})$

This section provides a recursion formula to calculate $N(T\backslash_m \mathbf{e})$ in (3.20). Denotes the root of T as v_0. Let $\tilde{v}_1, \ldots, \tilde{v}_r$ $(0 \leq r \leq 3)$ be the internal node children of the (pseudo-)root v of T. If $r = 0$, then T is a cherry and $N(T\backslash_m \mathbf{e}) = N(T) = 1$. If $r > 0$ Define $\tilde{\mathbf{e}}_i$ to be the sub-vector of $\mathbf{e}$ whose elements are descendents of $\tilde{v}_i$, for $i \leq r$. The trees in $T_{\tilde{v}_i}\backslash_m \tilde{\mathbf{e}}_i$ will be denoted as $\tilde{F}_{i,j}$. Then we can compute $N(T\backslash_m \mathbf{e}) = N(T_{v_0}\backslash_m \mathbf{e})$ through $N(T_{\tilde{v}_i}\backslash_m \tilde{\mathbf{e}}_i)$ as follows:

$$N(T_{v_0}\backslash_m \mathbf{e}) = \frac{|F_{v_0}|! \ \prod_{i=1}^{r} N(T_{\tilde{v}_i}\backslash_m \tilde{\mathbf{e}}_i)}{\mathrm{sym}_{F_{v_0}}(v_0)! \prod\limits_{\substack{i\leq r \\ \text{parent edge of } \tilde{v}_i \notin \mathbf{e}}} |F_{\tilde{v}_i}|!}, \tag{C.1}$$

where F_v is the element of $T_v\backslash_m \mathbf{e}$ containing v and $\mathrm{sym}_F(v)$ is the number of symmetries at node v in rooted tree F. To define this number of symmetries, let $w_1, \cdots, w_k$ be the children of v in F $(k \leq 3)$. Then

$$\mathrm{sym}_F(v) = \begin{cases} 1 & \text{if none of the } \mathbf{S}(F_{w_i}) \text{ are the same,} \\ 2 & \text{if exactly 2 of } \mathbf{S}(F_{w_i}) \text{ are the same,} \\ 3 & \text{if } k = 3 \text{ and all 3 } \mathbf{S}(F_{w_i})\text{'s are the same.} \end{cases}$$

We now prove (C.1). For a tree F_i in $T\backslash_m \mathbf{e}$, $|\{F : \exists \boldsymbol{\sigma}_{\mathcal{L}_i} \text{ such that } \boldsymbol{\sigma}_{\mathcal{L}_i}(F) = F_i\}|$ is the size of equivalence classes by the action of the permutation group $\{\boldsymbol{\sigma}_{\mathcal{L}_i}\}$. By Burnside

(1955),

$$\begin{aligned} |\{F : \exists \boldsymbol{\sigma}_{\mathcal{L}_i} \text{ such that } \boldsymbol{\sigma}_{\mathcal{L}_i}(F) = F_i\}| &= \frac{|\{\boldsymbol{\sigma}_{\mathcal{L}_i}(F_i)\}|}{|\{\boldsymbol{\sigma}_{\mathcal{L}_i} : \boldsymbol{\sigma}_{\mathcal{L}_i}(F_i) = F_i\}|} \\ &= |F_i|! \Big/ \prod_{v \in \overset{\circ}{V}(F_i)} \mathrm{sym}_{F_i}(v)! \end{aligned}$$

where $|F|$ is the number of original tip labels in F and $\overset{\circ}{V}(F)$ is the set of internal nodes in F. Thus we have

$$N(T \backslash_m \mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|+1} |F_i|! \Big/ \prod_{v \in \overset{\circ}{V}(F_i)} \mathrm{sym}_{F_i}(v)! \; .$$

To decompose $N(T\backslash_m \mathbf{e}) = N(T_{v_0}\backslash_m \mathbf{e})$, we compute the following ratio:

$$\frac{N(T_{v_0}\backslash_m \mathbf{e})}{\prod_{i=1}^{r} N(T_{\tilde{v}_i}\backslash_m \tilde{\mathbf{e}}_i)} = \frac{\prod_{k=1}^{|\mathbf{e}|+1} |F_k|!}{\prod_{i=1}^{r}\prod_{j=1}^{|\tilde{\mathbf{e}}_i|+1} |\tilde{F}_{i,j}|!} \frac{\prod_{i=1}^{r}\prod_{j=1}^{|\tilde{\mathbf{e}}_i|+1} \prod_{v \in \overset{\circ}{V}(\tilde{F}_{i,j})} \mathrm{sym}_{\tilde{F}_{i,j}}(v)!}{\prod_{k=1}^{|\mathbf{e}|+1} \prod_{v \in \overset{\circ}{V}(F_k)} \mathrm{sym}_{F_k}(v)!} \tag{C.2}$$

We first compute the second ratio in (C.2). Since $\overset{\circ}{V}(F_1), \ldots, \overset{\circ}{V}(F_{|\mathbf{e}|+1})$ form a partition of $\overset{\circ}{V}(T_{v_0})$,

$$\overset{\circ}{V}(T) = \bigcup_{k=1}^{|\mathbf{e}|+1} \overset{\circ}{V}(F_k) = \{v_0\} \bigcup_{i=1}^{r} \bigcup_{j=1}^{|\tilde{\mathbf{e}}_i|+1} \overset{\circ}{V}(\tilde{F}_{i,j}).$$

Consider an internal node v^* in $F_k \in T_v \backslash_m \mathbf{e}$. Let i be an index such that v^* is a descendent of $\tilde{v}_i$. Then there exists a unique subtree $\tilde{F}_{i,j}$ containing v^* and $\mathrm{sym}_{F_k}(v^*) = \mathrm{sym}_{\tilde{F}_{i,j}}(v^*)$, so we get

$$\prod_{k=1}^{|\mathbf{e}|+1} \prod_{v \in \overset{\circ}{V}(F_k)} \mathrm{sym}_{F_k}(v)! = \mathrm{sym}_{F_v}(v)! \times \prod_{i=1}^{r} \prod_{j=1}^{|\tilde{\mathbf{e}}_i|+1} \prod_{v \in \overset{\circ}{V}(\tilde{F}_{i,j})} \mathrm{sym}_{\tilde{F}_{i,j}}(v)!$$

Therefore, the second ratio in (C.2) is simply $1/\mathrm{sym}_{F_v}(v)!$. To simplify the first ratio in (C.2), we use the following properties.

1. If $\mathbf{e}$ contains a child edge e^* of v_0 and $\tilde{v}_i$ is the child node of e^*, then there exist $F_k \in T_{v_0} \backslash_m \mathbf{e}$ and $\widetilde{F}_{i,j} \in T_{\tilde{v}_i} \backslash_m \tilde{\mathbf{e}}_i$ such that $F_k = \widetilde{F}_{i,j}$ and both contain $\tilde{v}_i$.

2. If $\mathbf{e}$ contains a descendant edge e^* of $\tilde{v}_i$, then there exist $F_k \in T_{v_0} \backslash_m \mathbf{e}$ and $\widetilde{F}_{i,j} \in T_{\tilde{v}_i} \backslash_m \tilde{\mathbf{e}}_i$ such that $F_k = \widetilde{F}_{i,j}$ and both contain the child node of e^*.

These two properties imply that the first ratio in (C.2) is $\prod_{k=1}^{|\mathbf{e}|+1} |F_k|! \Big/ \prod_{i=1}^{r} \prod_{j=1}^{|\tilde{\mathbf{e}}_i|+1} |\widetilde{F}_{i,j}|! = |F_{v_0}|! \Big/ \prod_{\substack{i \le r \\ \text{parent edge of } \tilde{v}_i \notin \mathbf{e}}} |\widetilde{F}_{\tilde{v}_i}|!$. By combining these two ratios, we obtain the recursion formula in (C.1).

Appendix D

Proofs of Theorems for the recursive calculation of function R

D.1 Proof of Theorem 2

If d_i, m_i, d'_j and m'_j are equal to zero for all i and j, then

$$R(V, V', D, D', K, K', M, M', H) = \mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, \emptyset, \emptyset, K, K', M, M', \emptyset) \prod_{j=1}^{|V'|} N\left(T'_{v'_j}\right)$$

and consensus trees $C_{\mathcal{T}_V}(\emptyset)$ and $C_{\mathcal{T}'_{V'}}(\emptyset)$ are star trees. If $k_i = |T_{v_i}|$ and $k'_j = |T'_{v'_j}|$ for all i and j, then both $C_{\mathcal{T}_V}(\emptyset) \setminus\setminus K$ and $C_{\mathcal{T}'_{V'}}(\emptyset) \setminus\setminus K'$ are empty, therefore equal and $\mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', \emptyset) = 1$. If $k_i \neq |T_{v_i}|$ or $k'_j \neq |T'_{v'_j}|$, then the indicator is 0 by the definition.

D.2 Proof of Theorem 3

For any $E \in \mathring{\mathcal{M}}_{V,D,K,M}$, the number of edges in $C_{\mathcal{T}_V}(E) \setminus\setminus K$ is $\Delta = d_1 + d_2 + m_1$ if V contains both children of the pseudo-root, $\Delta = \sum_i (d_i + m_i)$ otherwise. Similarly, the number of edges in $C_{\mathcal{T}'_{V'}}(E') \setminus\setminus K'$ is obtained and denoted by Δ'. If $\Delta \neq \Delta'$, then

$C_{\mathcal{T}_V}(E) \setminus\setminus K$ and $C_{\mathcal{T}'_{V'}}(E') \setminus\setminus K'$ are different, so $\mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', G') = 0$ and $R = 0$.

D.3 Proof of Theorem 4

If there exists i satisfying at least one of five conditions, then $\mathring{\mathcal{M}}_{v_i,d_i,k_i,m_i} = \emptyset$, so $\mathring{\mathcal{M}}_{V,D,K,M} = \emptyset$ and $R = 0$.

D.4 Proof of Theorem 5

Assume here that T_ν and $T'_{\nu'}$ have the same shape. Let d be the number of internal nodes in T_ν (and in $T'_{\nu'}$). Let k_0 be the number of tips directly connected to ν in T_ν (or to ν' in $T'_{\nu'}$). If $k \neq k_0$ then $\mathring{\mathcal{M}}_{(\nu),(d),(k),(0)} = \emptyset$ and $R = 0$. Similarly, if $k' \neq 0$ the $\breve{\mathcal{M}}_{(\nu'),(d),(k'),(0)} = \emptyset$ and $R = 0$. Now consider the case $k = k' = k_0$. Then $\mathring{\mathcal{M}}_{(\nu),(d),(k),(0)}$ contains a single element $\mathbf{e} = (1, 2, \ldots, d)$. The set $\breve{\mathcal{M}}_{(\nu'),(d),(k),(0)}$ may contain more than one element, but $\mathbb{G}_{E',(\nu'),(d),(0),\emptyset}$ contains a single element $G' = (1, 2, \ldots, d)$. For each $\mathbf{e}' \in \breve{\mathcal{M}}_{(\nu'),(d),(k),(0)}$, the consensus trees $C_{T_\nu}(\mathbf{e})$ and $C_{T'_{\nu'}}(\mathbf{e}')$ have shape $\mathbf{S}(T_\nu/\bar{\mathbf{e}}) = \mathbf{S}(T_\nu) = \mathbf{S}(T'_{\nu'})$. Therefore $C_{T_\nu}(\mathbf{e}) \setminus\setminus (k) = C_{T'_{\nu'}}(\mathbf{e}') \setminus\setminus (k)$ if their edges are named in the same way, which happens for $\mathbf{e}' = \mathbf{e}$ only. In this case $\mathbb{I}(\mathcal{T}_\nu, \mathcal{T}'_{\nu'}, (\mathbf{e}), (\mathbf{e}'), (k), (k'), (0), (0), G') = 1$. This indicator is 0 for any other $e' \in \breve{\mathcal{M}}_{(\nu'),(d),(k),(0)}$, so $R = N(T_\nu \backslash_m \mathbf{e}) = 1$. If T_ν and $T'_{\nu'}$ have different shapes, then there is no $\mathbf{e}' \in \breve{\mathcal{M}}_{(\nu'),(d'),(k'),(0)}$ satisfying $\mathbb{I}(\mathcal{T}_\nu, \mathcal{T}'_{\nu'}, (\mathbf{e}), (\mathbf{e}'), (k), (k'), (0), (0), G') = 1$ so $R = 0$.

D.5 Proof of Theorem 6

Define $\mathring{\mathcal{M}}_{\widetilde{V}} = \bigcup_{\tilde{\mathbf{d}},\tilde{\mathbf{m}}\in\mathcal{C}} \bigcup_{\tilde{\mathbf{k}}\in\mathcal{K}} \mathring{\mathcal{M}}_{\widetilde{V},\widetilde{D},\widetilde{K},\widetilde{M}}$. Also consider the function $f : \mathring{\mathcal{M}}_{\widetilde{V}} \to \mathring{\mathcal{M}}_{V,D,K,M}$ such that $f(\widetilde{E}) = E$ is defined by $\mathbf{e}_i = \tilde{\mathbf{e}}_i$ for $i \le h-1$; $\mathbf{e}_h = (\tilde{\mathbf{e}}_h, \ldots, \tilde{\mathbf{e}}_{h+r-1})$ and $\mathbf{e}_i = \tilde{\mathbf{e}}_{i+(r-1)}$ if $i > h$. Then f is well-defined. That is, $f(\widetilde{E}) \in \mathring{\mathcal{M}}_{V,D,K,M}$ if $\widetilde{E} \in \mathring{\mathcal{M}}_{\widetilde{V}}$. It is easy to see that f is a bijection. Consider now $\widetilde{E} \in \mathring{\mathcal{M}}_{\widetilde{V},\widetilde{D},\widetilde{K},\widetilde{M}}$, $f(\widetilde{E}) = E \in \mathring{\mathcal{M}}_{V,D,K,M}$, $E' \in \breve{\mathcal{M}}_{V',D',K',M'}$ and $G' \in \mathbb{G}_{E',V',D',M',H}$. There exists an index i such that $k_i \neq |F_{v_i}|$ if and only if there exists an index j such that $\tilde{k}_j \neq |F_{\tilde{v}_i}|$, where F_ν is the tree in $T_\nu \backslash_m \mathbf{e}$ containing ν. By the definitions of $\mathcal{C}$, $\mathcal{K}$ and f, $C_{\mathcal{T}_V}(E) \setminus\setminus K = C_{\widetilde{\mathcal{T}}_{\widetilde{V}}}(\widetilde{E}) \setminus\setminus \widetilde{K}$. Therefore $\mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', G') = \mathbb{I}(\widetilde{\mathcal{T}}_{\widetilde{V}}, \mathcal{T}'_{V'}, \widetilde{E}, E', \widetilde{K}, K', \widetilde{M}, M', G')$. We can now rewrite $R = R(V, V', D, D', K, K', M, M', H)$ as follows:

$$R = \sum_{\widetilde{E}\in\mathring{\mathcal{M}}_{\widetilde{V}}} \sum_{E'\in\breve{\mathcal{M}}_{V',D',K',M'}} \sum_{G'\in\mathbb{G}_{E',V',D',M',H}} \prod_{i=1}^{q} N(T'_{v'_i}\backslash_m \mathbf{e}'_i)\mathbb{I}(\widetilde{\mathcal{T}}_{\widetilde{V}}, \mathcal{T}'_{V'}, \widetilde{E}, E', \widetilde{K}, K', \widetilde{M}, M', G')$$

$$= \sum_{\tilde{\mathbf{d}},\tilde{\mathbf{m}}\in\mathcal{C}} \sum_{\tilde{\mathbf{k}}\in\mathcal{K}} R(\widetilde{V}, V', \widetilde{D}, D', \widetilde{K}, K', \widetilde{M}, M', \widetilde{H}).$$

D.6 Proof of Theorem 7

The decomposition of an edge vector on T' is similar to that on T (Theorem 6), except that the order of edges matters now, and positionings need to be used to map V' onto $\widetilde{V}'$. Define $\breve{\mathcal{M}}_{\widetilde{V}'} = \bigcup_{\tilde{\mathbf{d}}',\tilde{\mathbf{m}}'\in\mathcal{C}} \bigcup_{\tilde{\mathbf{k}}'\in\mathcal{K}} \breve{\mathcal{M}}_{\widetilde{V}',\widetilde{D}',\widetilde{K}',\widetilde{M}'}$, $\mathbb{G}_{\widetilde{V}'} = \{\mathbb{G}_{\widetilde{E}',\widetilde{V}',\widetilde{D}',\widetilde{M}',\widetilde{H}} : \widetilde{E}' \in \breve{\mathcal{M}}_{\widetilde{V}'}\}$ and $\breve{\mathcal{M}}_{\widetilde{V}'} \otimes \mathbb{G}_{\widetilde{V}'} = \{\{\widetilde{E}'\} \times \mathbb{G}_{\widetilde{E}',\widetilde{V}',\widetilde{D}',\widetilde{M}',\widetilde{H}} : \widetilde{E}' \in \breve{\mathcal{M}}_{\widetilde{V}'}\}$. We consider the function $f : \breve{\mathcal{M}}_{\widetilde{V}'} \otimes \mathbb{G}_{\widetilde{V}'} \to \breve{\mathcal{M}}_{V',D',K',M'} \otimes \mathbb{G}_{V'}$ such that $f(\widetilde{E}', \widetilde{G}') = (E', G')$ is defined as follows: $e'_{i,j} = \tilde{e}'_{i,j}$ and $\mathbf{g}'_i = \tilde{\mathbf{g}}'_i$ for $i \le x-1$; $e'_{i,j} = \tilde{e}'_{i+r-1,j}$ and $\mathbf{g}'_i = \tilde{\mathbf{g}}'_{i+r-1}$ for $i \ge x+1$; $e'_{x,j} = \tilde{e}'_{a,b}$ if $\tilde{g}'_{a,b}$ is the jth smallest value in $(\tilde{\mathbf{g}}_x, \ldots, \tilde{\mathbf{g}}_{x+r-1})$ and $\mathbf{g}'_x$ is the vector all

values in $(\tilde{\mathbf{g}}_x, \ldots, \tilde{\mathbf{g}}_{x+r-1})$ sorted in ascending order. We claim that f is a one-to-one map (proved later). Then we can decompose $R = R(V, V', D, D', K, K', M, M', H)$, building on the decomposition of $N(T_v \backslash_m \mathbf{e})$ from Appendix C:

$$
\begin{aligned}
R = & \\
& \sum_{\substack{E \in \\ \mathring{\mathcal{M}}_{V,D,K,M}}} \sum_{\substack{(\widetilde{E}', \widetilde{G}') \in \\ \breve{\mathcal{M}}_{\widetilde{V}'} \otimes \mathbb{G}_{\widetilde{V}'}}} \frac{k'_x! \prod_{i=1}^{|\widetilde{V}'|} N(T'_{\tilde{v}'_i} \backslash_m \tilde{\mathbf{e}}'_i)}{\mathrm{sym}_{F_{v'_x}}(v'_x)! \prod_{i=x}^{x+r-1} (\tilde{k}'_i!)^{(1-\tilde{m}'_i)}} \mathbb{I}(\mathcal{T}_V, \widetilde{\mathcal{T}'}_{\widetilde{V}'}, E, \widetilde{E}', K, \widetilde{K}', M, \widetilde{M}', \widetilde{G}') \\
& = \sum_{\substack{E \in \\ \mathring{\mathcal{M}}_{V,D,K,M}}} \sum_{\tilde{\mathbf{d}}', \tilde{\mathbf{m}}' \in \mathcal{C}} \sum_{\tilde{\mathbf{k}}' \in \mathcal{K}} \sum_{\substack{\widetilde{G}' \in \\ \mathbb{G}_{\widetilde{E}', \widetilde{V}', \widetilde{D}', \widetilde{M}', \tilde{H}}}} \left\{ \frac{k'_x! \prod_{i=1}^{|\widetilde{V}'|} N(T'_{\tilde{v}'_i} \backslash_m \tilde{\mathbf{e}}'_i)}{\mathrm{sym}_{F_{v'_x}}(v'_x)! \prod_{i=x}^{x+r-1} (\tilde{k}'_i!)^{(1-\tilde{m}'_i)}} \right. \\
& \qquad \left. \times \mathbb{I}(\mathcal{T}_V, \widetilde{\mathcal{T}'}_{\widetilde{V}'}, E, \widetilde{E}', K, \widetilde{K}', M, \widetilde{M}', \widetilde{G}') \right\} \\
= & \sum_{\tilde{\mathbf{d}}', \tilde{\mathbf{m}}' \in \mathcal{C}} \sum_{\tilde{\mathbf{k}}' \in \mathcal{K}} R(V, \tilde{V}', D, \tilde{D}', K, \tilde{K}', M, \tilde{M}', \tilde{H})\, k'_x! \Big/ \left(\mathrm{sym}_{F_{v'_x}}(v'_x)! \prod_{i=x}^{x+r-1} (\tilde{k'}_i!)^{(1-\tilde{m}'_i)} \right).
\end{aligned}
$$

We now prove our claim that f is one-to-one. It is easy to show that f is injective. Now consider $(E', G') \in \breve{\mathcal{M}}_{V',D',K',M'} \otimes \mathbb{G}_{V'}$. Then $\mathbf{e}'_x \in E'$ can be decomposed into $(\mathbf{e}^*_1, \ldots, \mathbf{e}^*_r)$, where elements in $\mathbf{e}^*_s$ are edges adjacent to w'_s or in $T'_{w'_s}$ and keep their relative order from $\mathbf{e}'_x$. We define $(\widetilde{E}', \widetilde{G}')$ as follows: (1) $\tilde{\mathbf{e}}'_i = \mathbf{e}'_i$ for $i \leq x-1$; $\tilde{\mathbf{e}}'_i = \mathbf{e}^*_{i-x+1}$ for $x \leq i \leq x+r-1$; $\tilde{\mathbf{e}}'_i = \mathbf{e}'_{i-r+1}$ for $i \geq x+r$; (2) $\tilde{m}'_i = 1$ if $\tilde{e}'_i$ contains the parent edge of $\tilde{v}'_i$, $\tilde{m}'_i = 0$ otherwise; (3) $\tilde{d}'_i = |\tilde{\mathbf{e}}'_i| - \tilde{m}'_i$; (4) $\widetilde{\mathbf{g}}'_i = \mathbf{g}'_i$ for $i \leq x-1$; $\widetilde{\mathbf{g}}'_i = \mathbf{g}^*_i$ for $x \leq i \leq x+r-1$; $\widetilde{\mathbf{g}}'_i = \mathbf{g}'_{i-r+1}$ for $i \geq x+r$, where $g^*_{i,j} = g'_{x,s}$ if $e^*_{i,j} = e'_{x,s}$. Then, by definition, $f(\widetilde{E}', \widetilde{G}') = (E, G)$. To conclude that f is one-to-one, all we need to show is that $(\widetilde{E}', \widetilde{G}') \in \breve{\mathcal{M}}_{\widetilde{V}'} \otimes \mathbb{G}_{\widetilde{V}'}$. If $i \leq x-1$ or $i \geq x+r$, then $\tilde{\mathbf{e}}'_i \in \breve{\mathcal{M}}_{\tilde{v}'_i, \tilde{d}'_i, \tilde{k}'_i, \tilde{m}'_i}$. We now show that for $x \leq i \leq x+r-1$, $\tilde{\mathbf{e}}'_i$ is the representative of its subtree-shape

equivalence class (i.e. $\tilde{\mathbf{e}}'_i \in \breve{\mathcal{M}}_{\tilde{v}'_i,\tilde{d}'_i,\tilde{k}'_i,\tilde{m}'_i}$) as follows.

1. Consider e'' symmetric with the first element $\tilde{e}'_{i,1}$ of $\tilde{\mathbf{e}}'_i$ in $T'_{\tilde{v}'_i}$. Let s such that $\tilde{e}'_{i,1} = e'_{x,s}$. Then, none of $(e'_{x,1}, \ldots, e'_{x,s-1})$ are in $T'_{\tilde{v}'_i} = T'_{w'_s}$. For e'', $e'_{x,s}$ and $e'_{x,j} \in (e'_{x,1}, \ldots, e'_{x,s-1})$ the condition (ii) in Definition 2-2 (b) is satisfied, so $\tilde{e}'_{i,1} = e'_{x,s} \leq e''$ and $(\tilde{e}'_{i,1})$ is the representative of its class.

2. Assume that $(\tilde{e}'_{i,1}, \ldots, \tilde{e}'_{i,n-1})$ is a representative. Let s such that $\tilde{e}'_{i,n} = e'_{x,s}$. Consider $e'' \notin (\tilde{e}'_{i,1}, \ldots, \tilde{e}'_{i,n-1})$ in $T'_{\tilde{v}'_i}$ that is symmetric with $\tilde{e}'_{i,n}$. Then $(\tilde{e}'_{i,1}, \ldots, \tilde{e}'_{i,n-1}) \subset (e'_{x,1}, \ldots, e'_{x,s-1})$. Therefore, for any $\tilde{e}'_{i,j} \notin (\tilde{e}'_{i,1}, \ldots, \tilde{e}'_{i,n-1})$ that satisfies at least one of two conditions in Definition 2-2 (b), $\tilde{e}'_{i,n} = e'_{x,s} \leq e''$.

For now turn to showing that $\widetilde{G}' \in \mathbb{G}_{\widetilde{E}',\widetilde{V}',\widetilde{D}',\widetilde{M}',\widetilde{H}}$. For this, we first need to prove that $\widetilde{G}'$ satisfies the four conditions in Definition 3. Conditions 1-3 are easy to check. We now prove that condition 4 is satisfied. If W_1 or W_2 contains $(\tilde{v}'_x, \ldots, \tilde{v}'_{x+r-1})$, then the condition in Definition 3-4 is satisfied. For any symmetric sibling nodes v'_1 and v'_2 in $T'_{v'_x}$, let $W_1 = \{\tilde{v}'_{i_1}, \ldots, \tilde{v}'_{i_p}\} \subset \widetilde{V}$ maximal antichain in $T'_{v'_1}$ and $W_2 = \{\tilde{v}'_{j_1}, \ldots, \tilde{v}'_{j_q}\} \subset \widetilde{V}$ maximal antichain in $T'_{v'_2}$, where $x \leq i_1$, $i_p \leq j_1$ and $j_q \leq x + r - 1$. Let $e^*_1 = e'_{x,a}$ have the smallest position g^*_1 in $(\tilde{\mathbf{e}}'_{i_1}, \ldots, \tilde{\mathbf{e}}'_{i_p})$ and $e^*_2 = e'_{x,b}$ have the smallest position g^*_2 in $(\tilde{\mathbf{e}}'_{j_1}, \ldots, \tilde{\mathbf{e}}'_{j_q})$. Then, by Definition 2, $a \leq b$ and $g'_{x,a} \leq g'_{x,b}$. Therefore, $\min\{\tilde{\mathbf{g}}'_{i_1}, \ldots, \tilde{\mathbf{g}}'_{i_p}\} = g^*_1 = g'_{x,a} \leq g'_{x,b} = g^*_2 = \min\{\tilde{\mathbf{g}}'_{j_1}, \ldots, \tilde{\mathbf{g}}'_{j_q}\}$. Condition 4 is then satisfied and $\widetilde{G}' \in \mathbb{G}_{\widetilde{E}',\widetilde{V}',\widetilde{D}',\widetilde{M}'}$. To show that $\widetilde{G}' \in \mathbb{G}_{\widetilde{E}',\widetilde{V}',\widetilde{D}',\widetilde{M}',\widetilde{H}}$, consider $(i,j) \in \widetilde{H}$ and let $\widetilde{g}'_{j,s}$ be the position of the parent edge of $\widetilde{v}'_j$. If $j \leq x - 1$, then $(i,j) \in H$ and $\widetilde{g}'_{j,s} = g'_{j,s} \neq \sum_{y<i}(d_y + m_y) + s$. If $j \geq x + r$, then $(i, j - r + 1) \in H$ and $\widetilde{g}'_{j,s} = g'_{j-r+1,s} \neq \sum_{y<i}(d_y + m_y) + s$. Note that $\widetilde{H}$ does not include any pair (i,j) for $x \leq j \leq x + r - 1$. This completes the proof that f is bijective, and the proof of

Theorem 7.

D.7 Proof of Theorem 8

$R = R(V, V', D, D', K, K', M, M', H)$ can be factorized as follows:

$$R = R(V, V', D, D', K, K', M, M', H^*)$$
$$+ \sum_{j \in \mathcal{Z}} \sum_{E \in \mathring{\mathcal{M}}_{V,D,K,M}} \sum_{E' \in \breve{\mathcal{M}}_{V',D',K',M'}} \sum_{G' \in \mathbb{G}^{(j)}_{E',V',H}} \mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', L') \prod_{i=1}^{q} N(T'_{v'_i} \backslash_m \mathbf{e}'_i),$$

where

$$\mathbb{G}^{(j)}_{E',V',H} = \left\{ G' \,\middle|\, G' \in \mathbb{G}_{E',V',H}, \text{ position of the parent edge of } v'_j = \text{ position of the parent edge of } v_h \right\}.$$

Given $E \in \mathring{\mathcal{M}}_{V,D,K,M}$, $E' \in \breve{\mathcal{M}}_{V',D',K',M'}$ and $G' \in \mathbb{G}^{(j)}_{E',V',H}$, we first want to show that for $j \in \mathcal{Z}$,

$$\mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', G')$$
$$= \left\{ \sum_{k=0}^{|T_{v_h}|} \mathbb{I}\left((T_{v_h}), (T'_{v'_j}), (\mathbf{e}_h), (\mathbf{e}'_j), (k), (k), (0), (0), (\tilde{\mathbf{g}}_j) \right) \right\}$$
$$\times \mathbb{I}\left(\mathcal{T}_{V_{-h}}, \mathcal{T}'_{V'_{-j}}, E_{-h}, E'_{-j}, K_{-h}, K'_{-j}, M_{-h}, M'_{-j}, \tilde{G}'_{-j} \right), \quad \text{(D.1)}$$

where $\tilde{\mathbf{g}}'_j$ = order of $\mathbf{g}'_j$ and $\tilde{G}'_{-j}$ = order of G'_{-j}.

Note that the orders of parent edges of v_h in T and v'_j in T' are the same, since $L' \in \mathbb{L}^{(j)}_{E',V',H}$. Consensus tree $C_{\mathcal{T}_V}(E) \setminus\!\setminus K$ is able to be disjoint into two sub-consensus trees $C_1 = C_{T_{v_h}}(\mathbf{e}_h) \setminus\!\setminus (k_h)$ and $C_2 = C_{\mathcal{T}_{V_{-h}}}(E_{-h}) \setminus\!\setminus K_{-h}$ by disconnecting the parent edge of v_h. The ordering vector used for $C_{\mathcal{T}_V}(E) \setminus\!\setminus K$ is $L = (\mathbf{l}_1, \ldots, \mathbf{l}_{|E|})$. Similarly,

consensus trees $C_{\mathcal{T}'_{V'}}(E') \setminus\!\setminus K'$ is disjoint into two sub-consensus trees C'_1 and C'_2 and the orders of edges used in C'_1 and C'_2 are $\mathbf{l}'_j$ and L'_{-j}, respectively. Let $\tilde{C}'_1$ and $\tilde{C}'_2$ denote the consensus trees C'_1 and C'_2 after replacing order $\mathbf{l}'_j$ by $\tilde{\mathbf{l}}'_j$ and L'_{-j} by $\tilde{L}'_{-j}$, respectively. Similarly $\tilde{C}_1$ and $\tilde{C}_2$ are defined.

Assume that the left-hand side of eq. (D.1) is equal to 1. Then, $C_{\mathcal{T}_V}(E) \setminus\!\setminus K = C_{\mathcal{T}'_{V'}}(E') \setminus\!\setminus K'$ implies $C_1 = C'_1$ and $C_2 = C'_2$, and thereby $\tilde{C}_1 = \tilde{C}'_1$ and $\tilde{C}_2 = \tilde{C}'_2$. Moreover, $\tilde{\mathbf{l}}'_j \in \mathbb{L}_{\mathbf{e}'_j, v'_j, \emptyset}$ and $\tilde{L}'_{-j} \in \mathbb{L}_{E'_{-j}, V'_{-j}, H} = \mathbb{L}_{E'_{-j}, V'_{-j}, \tilde{F}}$. Since the second and third conditions to have $\mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', L')$ equal to 1 are always satisfied in $\tilde{C}_1, \tilde{C}_2, \tilde{C}'_1$ and $\tilde{C}'_2$, the right-hand side of eq. (D.1) is also equal to 1:

$$\mathbb{I}\left((T_{v_h}), (T'_{v'_j}), (\mathbf{e}_h), (\mathbf{e}'_j), (k), (k), (0), (0), (\tilde{\mathbf{l}}_j)\right) = \begin{cases} 1 & \text{if } k = |F_{v_h}| = |F'_{v'_j}|, \\ 0 & \text{otherwise,} \end{cases}$$

and $$\mathbb{I}\left(\mathcal{T}_{V_{-h}}, \mathcal{T}'_{V'_{-j}}, E_{-h}, E'_{-j}, K_{-h}, K'_{-j}, M_{-h}, M'_{-j}, \tilde{L}'_{-j}\right) = 1.$$

Similarly, if the left-hand side of eq. (D.1) is equal to 0, then at least one of two factors in the right-hand side of eq. (D.1) is equal to 0. Therefore, the eq. (D.1) holds.

Additionally, the following two functions are bijective:

$$\begin{aligned} f_1 : \mathring{\mathcal{M}}_{V,D,K,M} &\longrightarrow \mathring{\mathcal{M}}_{v_h,d_h,k_h,m_h} \times \mathring{\mathcal{M}}_{V_{-h},D_{-h},K_{-h},M_{-h}} \\ E &\mapsto f_1(E) = (\mathbf{e}_h, E_{-h}), \\ f_2 : \breve{\mathcal{M}}_{V',D',K',M'} \otimes \mathbb{L}_{V'} &\longrightarrow \left\{\breve{\mathcal{M}}_{v'_j,d'_j,k'_j,m'_j} \otimes \mathbb{L}_{v'_j}\right\} \times \left\{\breve{\mathcal{M}}_{V'_{-j},D'_{-j},K'_{-j},M'_{-j}} \otimes \mathbb{L}_{V'_{-j}}\right\} \\ (E', L') &\mapsto f_2(E', L') = \left(\mathbf{e}'_j, \mathbf{l}'_j, E'_{-j}, L'_{-j}\right) \end{aligned}$$

where $\mathbb{L}_{V'} = \{\mathbb{L}^{(j)}_{E',V',H} : E' \in \breve{\mathcal{M}}_{V',D',K',M'}\}$, $\mathbb{L}_{v'_j} = \{\mathbb{L}_{\mathbf{e}'_j,v'_j,\emptyset} : \mathbf{e}'_j \in \breve{\mathcal{M}}_{v'_j,d'_j,k'_j,m'_j}\}$ and $\mathbb{L}_{V'_{-j}} = \{L_{E'_{-j},V'_{-j},\tilde{F}} : E'_{-j} \in \breve{\mathcal{M}}_{V'_{-j},D'_{-j},K'_{-j},M'_{-j}}\}$.

Therefore, for $j \in \mathcal{Z}$,

$$\sum_{E\in\mathring{\mathcal{M}}_{V,D,K,M}} \sum_{E'\in\breve{\mathcal{M}}_{V',D',K',M'}} \sum_{L'\in\mathbb{L}^{(j)}_{E',V',H}} \left[\left\{\prod_{i=1}^{q} N(T'_{v'_i}\backslash_m \mathbf{e}'_i)\right\} \times \mathbb{I}(\mathcal{T}_V, \mathcal{T}'_{V'}, E, E', K, K', M, M', L')\right]$$

$$= \left[\sum_{k=0}^{|T_{v_h}|} \sum_{\mathbf{e}_h\in\mathring{\mathcal{M}}_{v_h,d_h,k_h,m_h}} \sum_{\mathbf{e}'_j\in\breve{\mathcal{M}}_{v'_j,d'_j,k'_j,m'_j}} \sum_{\mathbf{l}'_j\in\mathbb{L}_{\mathbf{e}'_j,v'_j,\emptyset}} N(T'_{v'_j}\backslash_m \mathbf{e}'_j) \times \mathbb{I}\left(\mathcal{T}_{v_h}, \mathcal{T}'_{v'_j}, (\mathbf{e}_h), (\mathbf{e}'_j), (k), (k), (0), (0), (\tilde{\mathbf{l}}_j)\right)\right]$$

$$\times \left[\sum_{E_{-h}\in\mathring{\mathcal{M}}_{V_{-h},D_{-h},K_{-h},M_{-h}}} \sum_{E'_{-j}\in\breve{\mathcal{M}}_{V'_{-j},D'_{-j},K'_{-j},M'_{-j}}} \sum_{L'_{-j}\in\mathbb{L}_{E'_{-j},V'_{-j},\tilde{F}}} \left\{\prod_{i\neq j} N(T'_{v'_i}\backslash_m \mathbf{e}'_i)\right\} \times\mathbb{I}\left(\mathcal{T}_{V_{-h}}, \mathcal{T}'_{V'_{-j}}, E_{-h}, E'_{-j}, K_{-h}, K'_{-j}, M_{-h}, M'_{-j}, L'_{-j}\right)\right]$$

$$= \left\{\sum_{k=0}^{|T_{v_h}|} R\left((v_h), (v'_j), (d_h), (d_h), (k), (k), (0), (0), \emptyset\right)\right\} \times R(V_{-h}, V'_{-j}, D_{-h}, D'_{-j}, K_{-h}, K'_{-j}, M_{-h}, M'_{-j}, \tilde{F})$$

Appendix E

Theorems and Proofs for Approximations

E.1 Proof of Theorem 9

When $(T_i)_{i\geq 1}$ are independent and uniformly distributed, the sequence $(d(T_i, T_{i+1}))_{i\geq 1}$ is 1-dependent, meaning that $\{d(T_i, T_{i+1}) : i \leq k\}$ and $\{d(T_i, T_{i+1}) : i \geq n + k\}$ are independent for $n \geq 2$. Consequently, this sequence is α-mixing with strong mixing coefficient $\alpha_n = 0$ for any $n \geq 2$ (Billingsley, 1995). Let $S_L = \sum_{l=1}^{L-1} d(T_l, T_{l+1})$ with mean $\mu_L = (L-1)E(d(T_1, T_2))$. Define $\sigma_L^2 = (L-1)\sigma^2$ where $\sigma^2 = \text{var}(d(T_1, T_2)) + 2\text{cov}(d(T_1, T_2), d(T_2, T_3))$. By Theorem 27.4 of Billingsley (1995), $\text{var}(S_L)/L \to \sigma^2 (> 0)$, and

$$(S_L - \mu_L)/\sigma_L \xrightarrow{d} \mathcal{N}(0, 1). \tag{E.1}$$

As L goes to infinity, $P(S_L \leq 1) = \left(Z_1 + (L-1)\zeta_2(1)\right)/Z_1^L$ goes to 0 for $N \geq 4$ (to ensure $Z_1 > 1$), and $P(S_L > 1)$ converges to 1 in probability. Moreover, the convergence of $(S_L - \mu_L)/\sigma_L \cdot \mathbb{I}_{S_L>1}$ to the standard normal distribution results from

Slutsky's theorem combined with (E.1). The sum in (3.24) is then

$$\sum_{x=2}^{D_L} \zeta_L(x) e^{-\beta x} = (Z_1^L - Z_1 - (L-1)\zeta_2(1)) \sum_{x=2}^{D_L} e^{-\beta x} \frac{\zeta_L(x)}{Z_1^L} / P(S_L \geq 2)$$

and for large L, we finally get

$$\begin{aligned}
\sum_{x=2}^{D_L} e^{-\beta x} \frac{\zeta_L(x)}{Z_1^L} &\approx \int_{2-.5}^{D_L+.5} e^{-\beta x} \phi(x; \mu_L, \sigma_L^2) dx \\
&= \exp\left(-\beta\mu_L + \frac{\beta^2\sigma_L^2}{2}\right) \int_{2-.5}^{D_L+.5} \phi(x; \mu_L - \beta\sigma_L^2, \sigma_L^2) dx \\
&= \exp\left(-\beta\mu_L + \frac{\beta^2\sigma_L^2}{2}\right) \left\{\Phi(D_L + .5; \mu_L - \beta\sigma_L^2, \sigma_L^2) - \Phi(2 - .5; \mu_L - \beta\sigma_L^2, \sigma_L^2)\right\}
\end{aligned}$$

where $\Phi(\cdot; \mu, \sigma^2)$ is the cumulative distribution function of the normal distribution with mean μ and variance σ^2. It is well approximated by Winitzki (2008):

$$\Phi\left(x; \mu, \sigma^2\right) \approx \frac{1}{2}\left[1 + \hat{\text{erf}}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right], \quad x \in \mathbb{R},$$

where $\hat{\text{erf}}(x) = \left(1 - \exp\left(-x^2 \frac{4/\pi + ax^2}{1+ax^2}\right)\right)^{1/2}$ and $a = (8(\pi-3))/(3\pi(4-\pi))$.

E.2 Independence approximation

Lemma 3. *The independence approximation $\hat{Z}_{(2)}$ in (3.26) is an underestimate, that is, $Z_L/\hat{Z}_{(2)} \geq 1$.*

Proof. Define $X_i = e^{-\beta d(T_i, T_{i+1})}$. If L is odd we can condition on the middle tree $T_{(L+1)/2}$ as follows:

$$\begin{aligned}
E\left(\prod_{i=1}^{L-1} X_i\right) &= E\left[E\left(\prod_{i=1}^{(L-1)/2} X_i \Big| T_{\frac{L+1}{2}}\right) E\left(\prod_{i=(L+1)/2}^{(L-1)} X_i \Big| T_{\frac{L+1}{2}}\right)\right] \\
&= E\left[\left\{E\left(\prod_{i=1}^{(L-1)/2} X_i \Big| T_{\frac{L+1}{2}}\right)\right\}^2\right].
\end{aligned}$$

By the CauchySchwarz inequality $E(Y^2) \geq (EY)^2$ we get

$$Z_L = Z_1^L E\left(\prod_{i=1}^{L-1} X_i\right) \geq Z_1^L \left(E \prod_{i=1}^{(L-1)/2} X_i\right)^2 = Z_{(L+1)/2}^2 \Big/ Z_1. \tag{E.2}$$

If L is even, we condition on the middle $X_{L/2}$ as follows:

$$\begin{aligned} E\left(\prod_{i=1}^{L-1} X_i\right) &= E\left[E\left(\prod_{i=1}^{L/2-1} X_i \Big| X_{L/2}\right)\ X_{L/2}\ E\left(\prod_{i=L/2+1}^{L-1} X_i \Big| X_{L/2}\right)\right] \\ &= E\left[X_{L/2}\left\{E\left(\prod_{i=1}^{L/2-1} X_i \Big| X_{L/2}\right)\right\}^2\right]. \end{aligned}$$

We apply the CauchySchwarz inequality again:

$$\begin{aligned} E\left[\frac{X_{L/2}}{E(X_{L/2})}\left\{E\left(\prod_{i=1}^{L/2-1} X_i \Big| X_{L/2}\right)\right\}^2\right] &\geq \left\{E\left[\frac{X_{L/2}}{E(X_{L/2})} E\left(\prod_{i=1}^{L/2-1} X_i \Big| X_{L/2}\right)\right]\right\}^2 \\ &= \left\{\frac{E\left(\prod_{i=1}^{L/2} X_i\right)}{E(X_{L/2})}\right\}^2, \end{aligned}$$

and hence

$$Z_L = Z_1^L E\left(\prod_{i=1}^{L-1} X_i\right) \geq Z_1^L E(X_{L/2}) \left\{\frac{E\left(\prod_{i=1}^{L/2} X_i\right)}{E(X_{L/2})}\right\}^2 = \frac{Z_{L/2+1}^2}{Z_1^2 \mu_X}, \tag{E.3}$$

where $\mu_X = E(X_1)$. We now show that $Z_L \geq \hat{Z}_{(2)} = Z_1^L \mu_X^{L-1}$ by induction. $Z_L = \hat{Z}_{(2)}$ for $L = 2$ by definition and $Z_L \geq \hat{Z}_{(2)}$ for $L = 3$ because of (E.2). Now assume $Z_L \geq \hat{Z}_{(2)}$ for all $L \leq L_0$. If $L_0 = 2p$ is even we use (E.2) and get $Z_{L_0+1} \geq Z_{p+1}^2 / Z_1 \geq \{Z_1^{p+1} \mu_X^p\}^2 / Z_1 = Z_1^{L_0+1} \mu_X^{L_0}$. If $L_0 = 2p - 1$ is odd we use (E.3) to get $Z_{L_0+1} \geq Z_{p+1}^2 / (Z_1^2 \mu_X) \geq (Z_1^{p+1} \mu_X^p)^2 / (Z_1^2 \mu_X) = Z_1^{L_0+1} \mu_X^{L_0}$, which finishes the proof. □

Bibliography

Allen, B. L. and Mike, S. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5:1–15.

Anderson, C. N. K., Ramakrishnan, U., Chan, Y. L., and Hadly, E. A. (2005). Serial simcoal: A population genetics model for data from multiple populations and points in time. *Bioinformatics*, 21(8):1733–1734.

Ané, C. (2010). Reconstructing concordance trees and testing the coalescent model from genome-wide data sets. In Knowles, L. L. and Kubatko, L. S., editors, *Estimating species trees: Practical and Theoretical Aspects*, chapter 3, pages 35–52. Wiley-Blackwell.

Ané, C. (2011). Detecting Phylogenetic Breakpoints and Discordance from Genome-Wide Alignments for Species Tree Reconstruction. *Genome Biology and Evolution*, 3:246–258.

Ané, C., Larget, B., Baum, D. A., Smith, S. D., and Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Molecular biology and evolution*, 24(2):412–426.

Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. *Nat Rev Genet*, 4(1):50–60.

Baum, D. A. (2007). Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon*, 56:417–426.

Belfiore, N. M., Liu, L., and Moritz, C. (2008). Multilocus phylogenetics of a rapid radiation in the genus thomomys (rodentia: Geomyidae). *Systematic Biology*, 57(2):294–310.

Billingsley, P. (1995). *Probability and Measure*. John Wiley & sons.

Bloomquist, E. W., Dorman, K. S., and Suchard, M. A. (2009). StepBrothers: inferring partially shared ancestries among recombinant viral sequences. *Biostat*, 10(1):106–120.

Bordewich, M. and Semple, C. (2005). On the computational complexity of the rooted subtree prune and regraft distance. *Annals of combinatorics.*, 8(4):409–423.

Boussau, B., Guguen, L., and Gouy, M. (2009). A mixture model and a hidden markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. *Evolutionary Bioinformatics*, 5:67–79.

Brown, C. J., Garner, E. C., Keith Dunker, A., and Joyce, P. (2001). The power to detect recombination using the coalescent. *Molecular Biology and Evolution*, 18(7):1421–1424.

Bruen, T. C., Philippe, H., and Bryant, D. (April 2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172(4):2665–2681.

Brumfield, R. T. (2010). Speciation genetics of biological invasions with hybridization. *Molecular Ecology*, 19(23):5079–5083.

Brumfield, R. T., Liu, L., Lum, D. E., and Edwards, S. V. (2008). Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (aves: Pipridae, manacus) from multilocus sequence data. *Systematic Biology*, 57(5):719–731.

Bryant, D. and Steel, M. (2009). Computing the distribution of a tree metric. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 6(3):420–426.

Buckley, T. R., Cordeiro, M., Marshall, D. C., and Simon, C. (2006). Differentiating between hypotheses of lineage sorting and introgression in new zealand alpine cicadas (maoricicada dugdale). *Systematic Biology*, 55(3):411–425.

Burnside, W. (1955). *Theory of Groups of Finite Order*. Dover Publications, New York.

Carstens, B. C. and Knowles, L. L. (2007). Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from melanoplus grasshoppers. *Systematic Biology*, 56(3):400–411.

Chen, F.-C. and Li, W.-H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *The American Journal of Human Genetics*, 68(2):444 – 456.

Chen, K., Durand, D., and Farach-Colton, M. (2000). Notung: A program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology*, 7(3-4):429–447.

Cipra, B. A. (1987). An introduction to the ising model. *Am. Math. Monthly*, 94:937–959.

Cranston, K. A., Hurwitz, B., Ware, D., Stein, L., and Wing, R. A. (2009). Species trees from highly incongruent gene trees in rice. *Systematic Biology*, 58(5):489–500.

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, 5(6):e11147.

Dayrat, B. (2003). The roots of phylogeny: How did haeckel build his trees? *Systematic Biology*, 52(4):515–527.

de Oliveira Martins, L. and Kishino, H. (2010). Distribution of distances between topologies and its effect on detection of phylogenetic recombination. *Annals of the Institute of Statistical Mathematics*, 62:145–159.

de Oliveira Martins, L., Leal, E., and Kishino, H. (2008). Phylogenetic Detection of Recombination with a Bayesian Prior on the Distance between Trees. *PLoS ONE*, 3(7):e2651.

DeGiorgio, M. and Degnan, J. H. (2010). Fast and consistent estimation of species trees using supermatrix rooted triples. *Molecular biology and evolution*, 27(3):552–569.

Degnan, J. H., DeGiorgio, M., Bryant, D., and Rosenberg, N. A. (2009). Properties of consensus methods for inferring species trees from gene trees. *Systematic zoology*, 58(1):35–54.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet*, 2(5):e68.

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340.

Didelot, X. and Falush, D. (March 2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics*, 175(3):1251–1266.

Didelot, X., Lawson, D., Darling, A., and Falush, D. (2010). Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics*, 186(4):1435–1449.

Doolittle, W. F. and Bapteste, E. (2007). Pattern pluralism and the tree of life hypothesis. *Proc Natl Acad Sci U S A*, 104(7):2043–2049.

Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M., and von Haeseler, A. (2007). Mapping human genetic ancestry. *Molecular biology and evolution*, 24(10):2266–2276.

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

Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-resolution species trees without concatenation. *PNAS*, 104(14):5936–5941.

Ekblom, R. and Galindo, J. (2010). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107:1–15.

Excoffier, L., Novembre, J., and Schneider, S. (2000). Computer note. simcoal: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *Journal of Heredity*, 91(6):506–509.

Fang, F., Ding, J., Minin, V. N., Suchard, M. A., and Dorman, K. S. (2007). cBrother: relaxing parental tree assumptions for Bayesian recombinatio n detection. *Bioinformatics*, 23(4):507–508.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA.

Furnas, G. (1984). The generation of random, binary unordered trees. *Journal of Classification*, 1(1):187–233.

Galtier, N. (2007). A model of horizontal gene transfer and the bacterial phylogeny problem. *Systematic Biology*, 56(4):633–642.

Galtier, N. and Daubin, V. (2008). Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512):4023–4029.

Gerard, D., Gibbs, H., and Kubatko, L. (2011). Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. *BMC Evolutionary Biology*, 11(1):291 – 302.

Goldman, N. (1990). Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of dna substitution and to parsimony analyses. *Systematic Biology*, 39(4):345–361.

Goulden, I. P. and Jackson, D. M. (2004). *Combinatorial Enumeration*. Dover Publications, Incorporated.

Grassly, N. and Holmes, E. (1997). A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol Biol Evol*, 14(3):239–247.

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

Heckman, K. L., Mariani, C. L., Rasoloarison, R., and Yoder, A. D. (2007). Multiple nuclear loci reveal patterns of incomplete lineage sorting and complex species history within western mouse lemurs (microcebus). *Molecular phylogenetics and evolution*, 43(2):353–367.

Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, 36(4):396–405.

Hein J, Schierup MH, W. C. (2005). *Gene genealogies, variation and evolution, a primer in coalescent theory*. New York: Oxford University Press.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular biology and evolution*, 27(3):570–580.

Holder, M. T., Anderson, J. A., and Holloway, A. K. (2001). Difficulties in detecting hybridization. *Systematic Biology*, 50(6):978–982.

Holland, B., Benthin, S., Lockhart, P., Moulton, V., and Huber, K. (2008). Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology*, 8(1):202.

Horvath, J. E., Weisrock, D. W., Embry, S. L., Fiorentino, I., Balhoff, J. P., Kappeler, P., Wray, G. A., Willard, H. F., and Yoder, A. D. (2008). Development and application of a phylogenomic toolkit: resolving the evolutionary history of madagascar's lemurs. *Genome research*, 18(3):489–499.

Huang, H. and Knowles, L. L. (2009). What is the danger of the anomaly zone for empirical phylogenetics? *Systematic Biology*, 58(5):527–536.

Huelsenbeck, J. P., Ané, C., Larget, B., and Ronquist, F. (2008). A bayesian perspective on a non-parsimonious parsimony model. *Systematic Biology*, 57(3):406–419.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Husmeier, D. and Mantzaris, A. V. (2008). Addressing the shortcomings of three recent bayesian methods for detecting interspecific recombination in dna sequence alignments. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

Husmeier, D. and McGuire, G. (2003). Detecting Recombination in 4-Taxa DNA Sequence Alignments with Bayesia n Hidden Markov Models and Markov Chain Monte Carlo. *Mol Biol Evol*, 20(3):315–337.

Jakobsen, I. B. and Easteal, S. (1996). A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Computer applications in the biosciences : CABIOS*, 12(4):291–295.

Jewett, E. and Rosenberg, N. (2012). iglass: an improvement to the glass method for estimating species trees from gene trees. *Journal of computational biology: a journal of computational molecular cell biology*, 19(3):293–315.

Joly, S., McLenachan, P. A., and Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, 174(2):E54–E70.

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of Protein Molecules*. Academy Press.

Kingman, J. F. C. (2000). Origins of the coalescent: 1974-1982. *Genetics*, 156(4):1461–1463.

Knowles, L. L. (2009). Estimating species trees: Methods of phylogenetic analysis when there is incongruence across genes. *Systematic Biology*, 58(5):463–467.

Kubatko, L. S. (2009). Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology*, 58(5):478–488.

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

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24.

Kuo, C.-H., Wares, J. P., and Kissinger, J. C. (2008). The apicomplexan whole-genome phylogeny: An analysis of incongruence among gene trees. *Molecular Biology and Evolution*, 25(12):2689–2698.

Larget, B. R., Kotha, S. K., Dewey, C. N., and Ané, C. (2010). Bucky: Gene tree / species tree reconciliation with bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911.

Lawrence, J. G. (1999). Gene transfer, speciation, and the evolution of bacterial genomes. *Current Opinion in Microbiology*, 2(5):519 – 523.

Leaché, A. D. (2009). Species tree discordance traces to phylogeographic clade boundaries in north american fence lizards (sceloporus). *Systematic zoology*, 58(6):547–559.

Lee, W.-H. and Sung, W.-K. (2008). Rb-finder: An improved distance-based sliding window method to detect recombination breakpoints. *Journal of Computational Biology*, 15(7):881–898.

Leigh, J. W., Schliep, K., Lopez, P., and Bapteste, E. (2011). Let them fall where they may: Congruence analysis in massive phylogenetically messy data sets. *Molecular Biology and Evolution*, 28(10):2773–2785.

Lerner, H. and Fleischer, R. (2010). Prospects for the Use of Next-Generation Sequencing Methods in Ornithology. *Auk*, 127:4–15.

Liu, L. (2008). Best: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21):2542–2543.

Liu, L., Pearl, D. K., Brumfield, R. T., Edwards, S. V., and Knowles, L. (2008). Estimating species trees using multiple-allele dna sequence data. *Evolution*, 62(8):2080–2091.

Liu, L. and Yu, L. (2011). Estimating species trees from unrooted gene trees. *Systematic Biology*, 60(5):661–667.

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5):468–477.

Loreto, E. L. S., Carareto, C. M. A., and Capy, P. (2008). Revisiting horizontal transfer of transposable elements in drosophila. *Heredity*, 100(6):545–554.

Machado, C. A. and Hey, J. (2003). The causes of phylogenetic conflict in a classic drosophila species group. *Proceedings of Biolocial Sciences*, 270(1520):1193–1202.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

Maddison, W. P. and Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55(1):21–30.

Maureira-Butler, I. J., Pfeil, B. E., Muangprom, A., Osborn, T. C., and Doyle, J. J. (2008). The reticulate history of medicago (fabaceae). *Systematic Biology*, 57(3):466–482.

Maynard Smith, J. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, 34(2):126–129.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., and Brumfield, R. T. (2011). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, (0):–.

Meng, C. and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical population biology*, 75(1):35–45.

Minin, V. N., Dorman, K. S., Fang, F., and Suchard, M. A. (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 21(13):3034–3042.

Oliver, J. C. (2008). Augist: inferring species trees while accommodating gene tree uncertainty. *Bioinformatics*, 24(24):2932–2933.

Page, R. (1998). Genetree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14(9):819–820.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular biology and evolution*, 5(5):568–583.

Posada, D. and Crandall, K. A. (2001). Evaluation of methods for detecting recombination from dna sequences: Computer simulations. *Proceedings of the National Academy of Sciences*, 98(24):13757–13762.

Posada, D., Crandall, K. A., and Holmes, E. C. (2002). Recombination in evolutionary genomics. *Annual Review of Genetics*, 36(1):75–97.

Preston, C. J. (1973). Generalized gibbs states and markov random fields. *Advances in Applied Probability*, 5(2):pp. 242–261.

Rasko, D. A., Phillips, J. A., Li, X., and Mobley, H. L. T. (2001). Identification of dna sequences from a second pathogenicity island of uropathogenic escherichia coli cft073: Probes specific for uropathogenic populations. *Journal of Infectious Diseases*, 184(8):1041–1049.

Richardson, A. O. and Palmer, J. D. (2007). Horizontal gene transfer in plants. *Journal of experimental botany*, 58(1):1–9.

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

Rodriguez, F., Wu, F., An, C., Tanksley, S. D., and Spooner, D. M. (2009). Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? *BMC Evolutionary Biology*, 9:191. ID: 198.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425:798–804.

Rosenberg, N. A. and Tao, R. (2008). Discordance of species trees with their most likely gene trees: The case of five taxa. *Systematic Biology*, 57(1):131–140.

Sang, T. and Zhong, Y. (2000). Testing hybridization hypotheses based on incongruent gene trees. *Syst Biol*, 49(3):422–434.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, 6(5):526–538.

Schierup, M. H. and Hein, J. (2000a). Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156(2):879–891.

Schierup, M. H. and Hein, J. (2000b). Recombination and the molecular clock. *Molecular Biology and Evolution*, 17(10):1578–1579.

Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press, New York, NY.

Sereno, P. C. (1999). The evolution of dinosaurs. *Science*, 284(5423):2137–2147.

Slatkin, M. and Pollack, J. L. (2008). Subdivision in an ancestral species creates asymmetry in gene trees. *Molecular biology and evolution*, 25(10):2241–2246.

Smith, J. M. (1999). The detection and measurement of recombination from sequence data. *Genetics*, 153(2):1021–1027.

Song YS, H. J. (2005). Constructing minimal ancestral recombination graphs. *J Comput Biol.*, 12(2):147–69.

Spratt, B. G., Hanage, W. P., and Feil, E. J. (2001). The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Current Opinion in Microbiology*, 4(5):602 – 606.

Steel, M. and Rodrigo, A. (2008). Maximum likelihood supertrees. *Systematic Biology*, 57(2):243–250.

Steel, M. A. and Penny, D. (1993). Distributions of tree comparison metricssome new results. *Systematic Biology*, 42(2):126–141.

Takahashi, K., Terai, Y., Nishida, M., and Okada, N. (2001). Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in lake tanganyika as revealed by analysis of the insertion of retroposons. *Molecular biology and evolution*, 18(11):2057–2066.

Than, C. and Nakhleh, L. (2009). Species tree inference by minimizing deep coalescences. *PLoS Comput Biol*, 5(9):e1000501.

Than, C., Ruths, D., Innan, H., and Nakhleh, L. (2007). Confounding factors in hgt detection: Statistical error, coalescent effects, and multiple solutions. *Journal of Computational Biology*, 14(4):517–535.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M. E., Frapy, E., Garry, L., Ghigo, J. M., Gilles, A. M., Johnson, J., Le Bougunec, C., Lescat, M., Mangenot, S., Martinez-Jhanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M. A., Pichon, C., Rouy, Z., Ruf, C. S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Mdigue, C., Rocha, E. P. C., and Denamur, E. (2009). Organised genome dynamics in the ¡italic¿escherichia coli¡/italic¿ species results in highly diverse adaptive paths. *PLoS Genet*, 5(1).

Tuffley, C. and Steel, M. (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59:581–607. 10.1007/BF02459467.

Webb, A., Hancock, J. M., and Holmes, C. C. (2009). Phylogenetic inference under recombination using Bayesian stochastic topology selection. *Bioinformatics*, 25(2):197–203.

Wehe, A., Bansal, M. S., Burleigh, J. G., and Eulenstein, O. (2008). Duptree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, 24(13):1540–1541.

Weiller, G. F. (1998). Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Molecular Biology and Evolution*, 15(3):326–335.

Wendel, J. F. and Doyle, J. J. (1998). Phylogenetic incongruence: Window into genome history and molecular evolution. In Soltis, P. and Doyle, J., editors, *Molecular Systematics of Plants II*, pages 265–296. New York.

White, M. A., Ané, C., Dewey, C. N., Larget, B. R., and Payseur, B. A. (2009). Fine-scale phylogenetic discordance across the house mouse genome. *PLoS Genet*, 5(11):e1000729.

Winitzki, S. (2008). A handy approximation for the error function and its inverse. Lecture note.

Wiuf, C., Christensen, T., and Hein, J. (2001). A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution*, 18(10):1929–1939.

Wu, J., Susko, E., and Roger, A. J. (2008). An independent heterotachy model and its implications for phylogeny and divergence time estimation. *Molecular Phylogenetics and Evolution*, 46(2):801 – 806.

Yan, S., Liu, H., Mohr, T. J., Jenrette, J., Chiodini, R., Zaccardelli, M., Setubal, J. C., and Vinatzer, B. A. (2008). Role of recombination in the evolution of the model plant pathogen pseudomonas syringae pv. tomato dc3000, a very atypical tomato strain. *Applied and Environmental Microbiology*, 74(10):3171–3181.

Yang, Z. and Rannala, B. (2005). Branch-length prior influences bayesian posterior probability of phylogeny. *Systematic Biology*, 54(3):455–470.

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., and Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Research*, 16:1099–1108.