Department of Mathematics and Statistics
University of Canterbury

Stochastic tree models and probabilistic modelling of gene trees of given species networks

A thesis submitted to the University of Canterbury in fulfilment of the requirements for the degree of Doctor of Philosophy

by

Sha (Joe) Zhu

June 29, 2013

Supervisors:
Dr. James Degnan
Prof. Mike Steel

谨以此文献给我的父亲和母亲.

To my father and mother.

Abstract

In the pre-genomic era, the relationships among species and their evolutionary histories were often determined by examining the fossil records. In the genomic era, these relationships are identified by analysing the genetic data, which also enables us to take a close-up view of the differences between the individual samples. Nevertheless, these relationships are often described by a tree-like structure or a network. In this thesis, we investigate some of the models that are used to describe these relationships.

This thesis can be divided into two main parts. The first part focuses on investigating the theoretical properties of several neutral tree models that are often considered in phylogenetics and population genetics studies, such as the Yule–Harding model, the proportional to distinguishable arrangements and the Kingman coalescent models.

In comparison to the first part, the other half of the thesis is more computationally oriented: we focus on developing and implementing methods of calculating gene tree probabilities of given species networks, and simulating genealogies within species networks.

Acknowledgements

Firstly I would like to thank my supervisor, James Degnan, for taking me as his first PhD student and supporting me from his Marsden fund. I have been working closely with James since 2009: an honours project 2009 entitled "Models for zero inflation data with their application to the Department of Conservation New Zealand", and the University of Canterbury summer research project 2009–2010 entitled "Effect of taxon sampling on constructing species trees from gene trees ". James is the first person who introduced me to phylogenetics, which I still enjoy very much after three years of struggling and fighting with it. His enthusiasm for research has influenced me immensely. I have enjoyed all of our hour-long meetings and have learnt an enormous amount from James.

A big thank-you goes to my other supervisor, Prof Mike Steel. Mike is always busy, but constantly checks up on me and keeps my work on schedule. Mike has set a great example for me: he is extremely hard-working and thorough with his work, which are the two keys to become a successful researcher. I am grateful that Mike has taken me under his wing, and has taken me to many conferences to learn the most cutting-edge methods and theorems, as well as to expose my work to the others in the field. At these meetings, I have learnt how influential Mike is in the field, and yet he is still very humble. Everybody loves talking to him.

On a very personal matter, after the Institute for Pure and Applied Mathematics Workshop in Los Angeles November 2011, because of some mechanical problems with the aircraft, I was stuck in Hawaii for a day. The delay caused me to miss another two flights and potentially ruined my schedule and summer internship at Beijing Genomics Institute. When I was worrying in the hotel room, I received messages from Mike offering help. Even today, I am still immensely grateful for hearing from Mike.

Both of my supervisors are patient, funny and friendly, which made my PhD life easy and smooth. I can not imagine who could be better supervisors than Mike and James. Thank you both for providing me such a wonderful opportunity to pursue my doctoral degree.

I also thank my other mentors and collaborators for valuable suggestions and comments: David Aldous for suggesting we consider the stochastic properties of RTP$'$ trees relative to their centroids; David Bryant for suggesting that I should use the extended Newick format for networks; both of my supervisors and Peter Humpheries for proof-reading my work; Taoyang Wu, Cuong Than, Bjarki Eldon, Sharyn Goldstien, Elizabeth

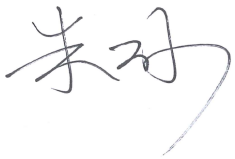
Allman and John Rhodes for collaborating and bring new insights into our projects.

I thank all the researchers whom I have cited. Without their vision and accomplishments, this project would have nothing to stand on. I also thank the UC library, Google and Wikipedia for making it easy to access to all the brilliant works on the internet.

I also thank the Marsden Fund, the Department of Mathematics and Statistics and the Allan Wilson Centre for Molecular Ecology and Evolution for supporting me in completing my PhD studies and for attending many national and international conferences.

Once, a postgraduate friend told me that "the more you can drink, the greater the mathematics you can do." Even though the truthfulness of this statement has never been verified, I have certainly enjoyed the company of our postgraduate students — the group of people who have shared the same experiences as me, such as the bitterness before deadlines, and the joys and sweets we share at the postgraduate talks. I sincerely hope our postgraduate talks will carry on after my departure. I will miss you all, especially Peter Jaksons, Nick Brettell, Gloria Teng, Peter Humpheries, Daniel Lond, Scott Graybill, Anna MacDonald, Rachael Tappenden and Timothy Candy. Also, a big thank-you goes to Penelope Goode, Steve Gourdie and Paul Brouwers for all the administration work and technical computer support. While pursuing my PhD, I have also taken some tutoring work at the Department of Mathematics and Statistics, where Irene David and Hilary Seddon have helped and supported me enormously, for which I am grateful.

The last but not the least, I would like to thank my family and all my friends who have always believed in me, especially Uncle Fu, Auntie Zhou, Frank Fu, Yanhui Li, Elva Chan, James Thompson, Nathan Winkins, Pete Fairbairn, Richard Wall, Lawrence Teo, Jonie Chang and many other close friends whose names have not been listed here. Without your unconditional love, continuing encouragement and support, I could never have made it this far.



Sha Zhu
June 29, 2013

Contents

Abstract iii

Acknowledgements iv

List of abbreviations x

1 Introduction 1
- 1.1 Phylogenetics, gene trees and species trees . . . 1
- 1.2 Background and motivation . . . 2
  - 1.2.1 Relationships between gene trees and species trees . . . 2
  - 1.2.2 Incomplete lineage sorting . . . 3
  - 1.2.3 Hybridization . . . 4
  - 1.2.4 Clade probabilities . . . 5
  - 1.2.5 Multiple merger coalescent models . . . 6
  - 1.2.6 Quartet puzzling . . . 6
- 1.3 Overview . . . 8
  - 1.3.1 Thesis objectives . . . 8
  - 1.3.2 Thesis structure . . . 8

2 Preliminaries 11
- 2.1 Sets . . . 11
- 2.2 Graphs . . . 11
  - 2.2.1 Binary trees . . . 11
  - 2.2.2 Generating random trees . . . 13
  - 2.2.3 Polytomy . . . 14
  - 2.2.4 Compatibility . . . 15
  - 2.2.5 Networks . . . 15
- 2.3 Newick format . . . 15
  - 2.3.1 Extended Newick format . . . 16
- 2.4 Probabilities . . . 16
  - 2.4.1 Conditional probabilities . . . 18
  - 2.4.2 Random variables . . . 18
  - 2.4.3 Probability distribution functions . . . 19
  - 2.4.4 Conditional expectation . . . 19
  - 2.4.5 Stochastic processes . . . 19
  - 2.4.6 Martingales . . . 19
- 2.5 Polyá urn model . . . 20
  - 2.5.1 Extended Polyá urn model . . . 22
  - 2.5.2 An example of the EPU model . . . 23

3 The RTP process and the YH process **24**
3.1 Introduction . . . 24
3.2 Formalised conjectures . . . 24
3.3 RTP is similar to YH when n is large . . . 26
3.3.1 Distribution of edge weights . . . 26
3.3.2 New leaves rarely attach to interior edges . . . 27
3.3.3 The mean and variance of the number of cherries in the RTP tree . . . 28
3.4 Is RTP the same as YH? . . . 30
3.4.1 Polyá urns and the centroid of a tree . . . 30
3.4.2 A modified RTP process . . . 32
3.5 Further discussion . . . 34

4 Clade and clan probabilities in the YHK model **37**
4.1 Introduction . . . 37
4.2 The YHK process . . . 37
4.3 Clade probabilities . . . 39
4.3.1 Properties of the YHK process . . . 40
4.3.2 One clade . . . 40
4.3.3 Pairs of clades . . . 43
4.3.4 Correlation between two clades . . . 45
4.3.5 Computing the probability of k clades recursively . . . 48
4.4 Clan probabilities . . . 49
4.4.1 Extensions of the clan condition (I) . . . 52
4.4.2 Extensions of the clan condition (II) . . . 52
4.5 Further discussion . . . 52

5 Clade and clan probabilities in the PDA model **54**
5.1 Introduction . . . 54
5.2 Notation . . . 55
5.2.1 Properties of the PDA model . . . 56
5.3 Clade probabilities . . . 56
5.3.1 Clade probability under the PDA model (I) . . . 58
5.3.2 A comparison between YHK and PDA . . . 60
5.3.3 Clade probability under the PDA model (II) . . . 61
5.3.4 Clade probability under the PDA model (III) . . . 62
5.3.5 Correlation results under the PDA model (I) . . . 63
5.4 Extension to unrooted trees . . . 64
5.4.1 Clan probability under the PDA model (I) . . . 64
5.4.2 Clan probability under the PDA model (II) . . . 65
5.4.3 Correlation results under the PDA model (II) . . . 66
5.5 Further discussion . . . 67
5.5.1 Size of a randomly selected internal node . . . 67
5.5.2 Expected value of the Sackin index . . . 67

6 Probabilities of gene trees of a given species network **70**
6.1 Introduction . . . 70
6.2 Methodology . . . 71
6.2.1 Modified post-order traversal method . . . 71
6.2.2 Decomposition operations . . . 72
6.2.3 Simplifying a network recursively . . . 76

6.2.4 Example . . . 77
6.2.5 Multiple lineages sampled from each population in the present . . . 79
6.3 hybrid_coal . . . 81
6.4 Discussion and future work . . . 81
6.4.1 Identifiability (I) . . . 81
6.4.2 Identifiability (II) . . . 82
6.4.3 Identifiability (III) . . . 83
6.4.4 Future work . . . 85

7 Simulating genealogies **87**
7.1 Introduction . . . 87
7.2 Method . . . 88
7.3 Simulating coalescent time in coalescent units . . . 90
7.3.1 Simulating lineage coalescent events in a time interval . . . 90
7.3.2 Simulating lineage coalescent events above the root . . . 90
7.3.3 Simulating lineage coalescent events at a hybrid node . . . 92
7.4 Simulating coalescent time in number of generations . . . 92
7.5 Segregating site data . . . 93
7.5.1 Expected number of mutations . . . 93
7.5.2 Simulating segregating site data . . . 94
7.6 hybrid-Lambda . . . 94
7.6.1 Features . . . 95

8 Concluding comments **97**

Bibliography **99**

A Symbols used **107**

B Technical details **109**

C hybrid_coal user manual **118**
C.1 Introduction . . . 118
C.2 Download and installation . . . 118
C.3 Notation . . . 119
C.3.1 Coalescent parameters . . . 119
C.3.2 Input/output formats . . . 119
C.3.3 Method . . . 120
C.4 Commands: . . . 120
C.4.1 Generating a list of all gene tree topologies in a taxa set . . . 120
C.4.2 Calculating gene tree probabilities of a given species network . . . 121
C.4.3 Generating Maple script for the gene tree probabilities . . . 121
C.4.4 Generating coalescent histories for the gene tree probabilities . . . 121
C.4.5 Commands for other features: . . . 122
C.5 Summary of command line options . . . 122

D hybrid-Lambda user manual **124**
D.1 Introduction . . . 124
D.2 Download and installation . . . 124
D.3 Notation . . . 125
D.3.1 Coalescent parameters . . . 125

D.3.2 Input/output formats . . . 125
D.4 Commands for simulation: . . . 126
D.4.1 Simulating gene trees . . . 126
D.4.2 Gene tree output options and user-defined mutation rate . . . 127
D.4.3 User-defined population sizes . . . 127
D.4.4 Simulating multiple samples per species . . . 128
D.4.5 Simulating gene trees with multiple merger coalescents . . . 128
D.4.6 Commands for other features: . . . 129
D.5 Summary of command line options . . . 130

List of abbreviations

C	caterpillar
DNA	deoxyribonucleic acid
EP	exchangeability property
EPU	extended Polyá urn
GE	group elimination
i.i.d.	independently and identically distributed
MP	maximum parsimony
ML	maximum likelihood
MRCA	most recent common ancestor
MT	maximum tree
NC	non-caterpillar
PDA	proportional to distinguishable arrangements
QP	quartet puzzling
RTP	random tree-puzzle
SC	sampling consistency
YH	Yule–Harding
YHK	Yule–Harding–Kingman

Chapter 1

Introduction

1.1 Phylogenetics, gene trees and species trees

In phylogenetic studies, trees are used to describe evolutionary histories. In particular, a species tree, also known as a *phylogeny*, represents population divergences; a gene tree, also known as a *genealogy*, indicates the times when genes started to differentiate within populations.

Two primary goals of phylogenetic studies are to determine the historical relationships among species or lineages within the same population, using phylogenies and genealogies respectively. To reconstruct these trees from molecular data, there are three main approaches:

- distance methods, such as *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA) (Sokal and Michener, 1958), *Neighbour–Joining* (NJ) (Saitou and Nei, 1987), BIONJ (Gascuel, 1997) and Weighted Neighbour–Joining (Bruno *et al.*, 2000);
- maximum parsimony approaches (Fitch, 1971; Sankoff, 1975; Sankoff and Rousseau, 1975);
- probabilistic approaches, such as maximum likelihood (ML) (Felsenstein, 1981), PhyML (Guindon and Gascuel, 2003; Guindon *et al.*, 2010), Quartet Puzzling (Strimmer and von Haeseler, 1996), Tree Puzzle (Schmidt *et al.*, 2002), MrBayes (Huelsenbeck and Ronquist, 2001), BEAST (Drummond *et al.*, 2012).

This thesis focuses on the problems that are raised by considering the probabilistic models, such as

- algorithms of reconstructing phylogenies from deoxyribonucleic acid (DNA) sequences,
- the statistical significance of determining whether a group of organisms share a common ancestor,
- probabilistic models of gene trees and
- genealogy simulations.

1.2 Background and motivation

1.2.1 Relationships between gene trees and species trees

This thesis treats the species trees from a population perspective. In other words, a branch of the phylogeny is not seen as a single line, but as a population that contains several individuals, which allows us to fit in multiple lineages. Thus, gene trees can be easily shown within the phylogeny (see Figure 1.1). When two gene tree lineages merge within a species tree branch, it is called a coalescent event. For convenience, this thesis uses capital letters to denote species and the corresponding lower case letters to denote the lineages sampled from these species (see Figure 1.1).

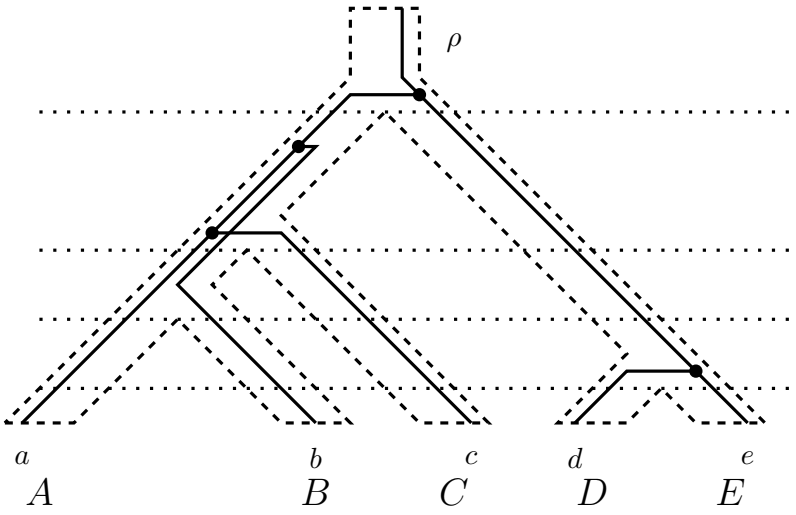


Figure 1.1: A gene tree (solid lines) within a species tree (dashed lines). The horizontal dotted lines indicate speciation times. This is an example of lineage sorting: the lineage in population A is more closely related to the lineage from population C than the one from population B, even though the separation between populations A and B occurred more recently than that between AB and C.

Even though speciation is driven by gene mutations, using a single gene tree to infer the species tree is not ideal, as the actual species tree topology may differ from the gene tree topology (see Figure 1.1). Common causes of the conflict between gene trees and species trees include gene duplication (Guigo *et al.*, 1996; Maddison, 1997), horizontal gene transfer (Maddison, 1997), incomplete lineage sorting (Degnan and Salter, 2005), and hybridization (Holland *et al.*, 2008).

As sequencing techniques and tools are improving rapidly, more data are becoming available, which may lead to better phylogenetic estimations. Chen and Li (2001) and Rokas *et al.* (2003) claimed that concatenated data, which are made by connecting short pieces of segmented data, lead to a strongly supported phylogenetic estimate, but Kubatko and Degnan (2007) suggested otherwise. Because of the variable evolutionary rate across DNA sites (Darlu and Lecointre, 2002; Leigh *et al.*, 2008), using data from multiple loci to infer phylogenetic histories appears to be more convincing than concatenating sequences,

which may result in observing more conflicts between the gene trees and species trees.

1.2.2 *Incomplete lineage sorting*

If speciation events occur close together in time (see Figure 1.1), it becomes more likely that the lineages for two leaves in a gene tree, a and b, do not coalesce in the most recent common ancestral population for species A and B. Such events are examples of *incomplete lineage sorting*, which leads to the possibility of gene tree topologies differing from the species trees. With the presence of incomplete lineage sorting, gene tree probabilities can be calculated under the *coalescent process* (Degnan and Salter, 2005).

The coalescent process starts from the bottom of a species tree (representing the present), and traces the gene history backwards in time. For example, Figure 1.1 shows that one lineage is sampled from each population A, B, C, D and E. A solid circle indicates the event when gene lineages coalesce. One particular coalescent process, known as the *Kingman coalescent* process, assumes that only two lineages can coalesce at one time; and the time that it takes for two lineages to coalesce is referred to as the coalescent time, which is an exponential random variable with a rate of 1.

As the number of lineages increases, the numbers of possible coalescent events increases rapidly, which results in a complex distribution of the coalescence time. In the 1980s, several authors (Takahata and Nei, 1985; Tavaré, 1984; Watterson, 1984) derived the formula that computes the probability of u lineages coalescing into v lineages within a time interval t.

These results were extended further to compute the gene tree topology distribution given any species tree (Degnan and Salter, 2005). However, the method introduced by Degnan and Salter (2005) is time-consuming and is inadequate for trees with a large number of taxa.

Wu (2012) used a recursive approach of ancestral configurations to improve the efficiency of computing gene tree probabilities, which reduces the computation time for some specific cases. These probabilistic models can be used as a basis for ML (Liu, 2008; Meng and Kubatko, 2009; Wu, 2012) or Bayesian estimation (Heled and Drummond, 2010; Liu, 2008) of species relationships.

Previous studies have researched methods of constructing species trees from gene trees with the presence of lineage sorting: the 'minimise deep coalescent' (Maddison and Knowles, 2006; Than and Nakhleh, 2009), average ranks of coalescence times (Liu *et al.*, 2009), average coalescence times (Liu *et al.*, 2009), and maximum tree (Liu *et al.*, 2010). In particular, the maximum tree, average ranks of coalescence times and average coalescence times trees are all consistent estimators of the species trees. These methods have been developed with the assumption that incomplete lineage sorting is the only source of gene tree conflict.

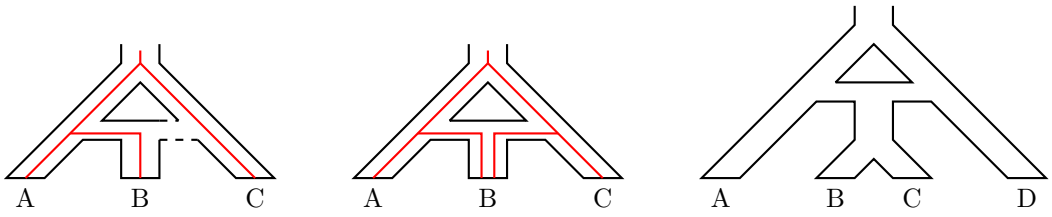
1.2.3 Hybridization

Hybridization refers to interbreeding between species or genetically distinct populations. Offspring that carry genes from both parental species then reproduce and form a new species. For many closely related species, however, both lineage sorting and hybridization are likely to occur, particularly for plants, such as New Zealand's alpine *Ranunculus* (Joly *et al.*, 2009). Other examples include the avian genus *Manacus* (Brumfield and Carling, 2010) and New Zealand alpine cicadas *Maoricicada* (Buckley *et al.*, 2006). Recent estimates indicate that hybridization events are common in nature: approximately 25% of plants and 10% of animals hybridize (Mallet *et al.*, 2007). Thus, for some organisms, it may be crucial to consider scenarios in which hybridization may have occurred.

In the past, incongruent gene trees have often been used to detect hybridization events. Simulation studies have been able to distinguish gene tree incongruence that arises from lineage sorting versus hybridization (Joly *et al.*, 2009; Sang and Zhong, 2000). Moreover, some models can even estimate hybridization in the presence of incomplete lineage sorting (Holland *et al.*, 2008; Kubatko, 2009; Meng and Kubatko, 2009).

However, these approaches have been limited to cases in which there is only one lineage sampled from a hybridized population. For example, the method of Meng and Kubatko (2009) decomposes the network into two species trees (see Figure 1.2a), with some probability γ, which is the probability that the lineage goes to one of the parental populations. Unfortunately, this model is not adequate for sampling multiple lineages from a single population.

Kubatko (2009) allows more than one hybridization event in a network as long as the hybridization events do not interact (i.e. are not nested) and there is only one population descended from each hybridized population.



(a) (b) (c)

Figure 1.2: Illustration of the hybrid speciation model proposed by Meng and Kubatko (2009). Population B indicates the hybrid species. If only one lineage is sampled from B, we can simplify the network into a tree with probability γ, which is the probability that the lineage goes to one of its parental populations. This model is not feasible for (b) cases sampling multiple lineages from a hybrid species or (c) networks in which the species split after hybridization.

Moreover, hybridization events may happen between hybrid species as well, which would cause more complex structures such as nested networks (this will be formally defined as a *level-k network* in Chapter 2). There are many interesting and open problems in

inferring and constructing nested networks, and these are often accomplished using *rooted triplets* (van Iersel *et al.*, 2008; Huynh *et al.*, 2005), which are rooted binary trees with a leaf set size of three.

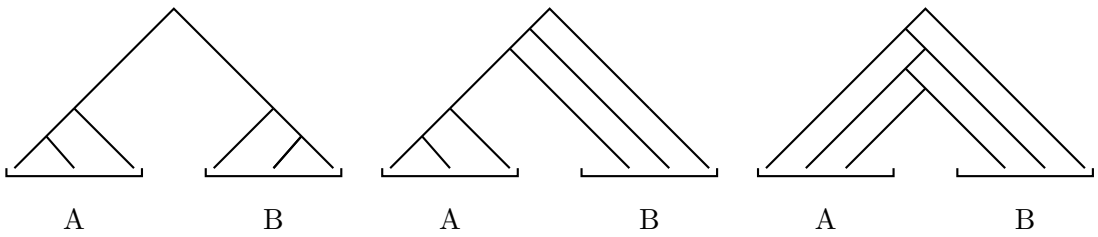
Without knowing the exact probabilities of the genealogies, the package `ms` (Hudson, 2002) can simulate gene trees within a general species network. However, the input of `ms` is difficult to automate when the network is sophisticated or is generated from other software. Other simulation studies using species networks have either used a small number of network topologies coded individually, as in `phylonet` (Than *et al.*, 2008), or have assumed that gene trees have evolved on species trees embedded within the species network (Holland *et al.*, 2008; Kubatko, 2009; Meng and Kubatko, 2009).

1.2.4 Clade probabilities

A species is commonly defined as a group of organisms that are all more closely related to each other than they are to any organisms outside the group (Shaw, 1998). Thus, genealogical shapes play a significant role in identifying species.

A *monophyletic group* refers to a fixed set of individuals for which the most recent common ancestor does not have any other descendants. For example, Figure 1.1 shows that the genes sampled from A and B are not monophyletic because their most recent common ancestor has gene c as a descendant.

When gene trees are estimated from multiple lineages taken from two or more populations, there is an increased chance that the lineages within each population form monophyletic groups compared to sampling multiple lineages from a single population. This observation has led to the adoption of a null hypothesis that a set of lineages belongs to a single population or taxonomic group, when asking whether a particular group of lineages came from a taxonomically distinct population (Cummings *et al.*, 2008; Rosenberg, 2007). A monophyletic group is also known as a *clade* in rooted trees.



(a) Populations A and B are both monophyletic; this is also known as reciprocal monophyly.

(b) Population A is monophyletic; B is paraphyletic with respect to A.

(c) Populations A and B are both polyphyletic.

Figure 1.3: Demonstration of monophyletic, polyphyletic and paraphyletic groups.

For two groups of individuals, if both groups are monophyletic, this is called reciprocal monophyly (see Figure 1.3a). Statistical tests for reciprocal monophyly between two sister taxa can then be developed to test the null hypothesis that a set of lineages belongs to a single population or taxonomic group (Hudson and Coyne, 2002; Rosenberg, 2003).

Alternatively, if group A is monophyletic but group B is not, then B is called paraphyletic with respect to A (see Figure 1.3b). Otherwise, they are called polyphyletic.

Reciprocal monophyly is central to the genealogical species concept. According to this concept, two groups come from different species if they form distinct monophyletic groups (de Quieroz, 2007; Hudson and Coyne, 2002). Gene trees from lineages sampled from one or more populations are typically estimated, and the monophyly (or lack of monophyly) of these groups can be observed from the clades of the gene tree. Statistical tests for whether observed levels of monophyly provide sufficient evidence to conclude that a group is taxonomically distinct can be performed, given a probabilistic model for the clades on a tree (Rosenberg, 2007).

In order to examine the significance of the taxonomic distinctiveness of several groups of lineages statistically, it is more powerful to perform the hypothesis test only once, instead of testing whether each group is monophyletic. Thus, methods for calculating the joint probability of monophyletic groups are desirable.

1.2.5 Multiple merger coalescent models

Species trees describe ancestral relationships among species. Gene trees describe the random ancestral relationships of alleles sampled within species. Gene trees and species trees are often assumed to be bifurcating (Degnan and Salter, 2005; Hudson, 1990; Kingman, 1982). However, for organisms exhibiting sweepstakes reproduction, such as oysters and other marine organisms (Árnason, 2004; Beckenbach, 1994; Eldon and Wakeley, 2006; Eldon, 2011; Hedgecock *et al.*, 1982; Hedgecock, 1994; Sargsyan and Wakeley, 2008), the Kingman coalescent may not be appropriate, as it allows only binary mergers of lineages. Species trees may also fail to be bifurcating due to either polytomies or hybridization events.

Thus, it is crucial to consider models that allow more than two lineages to coalesce simultaneously in the gene trees, that is multiple merger coalescent models, also known as Λ-coalescent models (Donnelly and Kurtz, 1999; Pitman, 1999; Sagitov, 1999). The reciprocal monophyletic concordance probabilities between multiple merger gene genealogies and a species tree of two species are investigated by Eldon and Degnan (2012). However, the probability of multiple merger gene trees becomes messy when the number of sampled individuals increases. Thus, studies of multiple merger coalescence for many individuals can only be undertaken by simulations. I have implemented a program which can simulate such genealogies that is not available in other software to my knowledge.

1.2.6 Quartet puzzling

The ML approach (Felsenstein, 1981; Guindon and Gascuel, 2003; Guindon *et al.*, 2010) is generally considered to be a reliable way of estimating phylogenies from DNA sequences. However, ML is not always feasible for large numbers of species because of the intensive computation required. Methods that use 'four-point subsets' (Dress *et al.*, 1986) reduce

the complexity of the problem and have assisted numerous studies (Daubin and Ochman, 2004; Nieselt-Struwe and von Haeseler, 2001; Strimmer *et al.*, 1997; Strimmer and von Haeseler, 1996).

The four-point subtree is known as the quartet tree. *Quartet puzzling* (QP) (Strimmer and von Haeseler, 1996) is an algorithm for inferring a tree on n taxa by using the quartet trees derived from DNA sequences. It first computes the likelihood of all $\binom{n}{4}$ quartets. As there are three possible topologies for any four taxa, the quartet tree which returns the greatest ML value is used (any ties are broken uniformly at random). At the puzzling step, the order of inserting new leaf nodes is randomised. A seed tree is built from the first four elements of the ordered leaf node sequence. From this point on, leaves are attached sequentially by the following procedure. When a new leaf x is to be attached to the existing tree T, quartet trees are built from the quartets formed from x and all subsets of size three chosen from the existing leaf set. If the ML quartet tree of $\{i, j, k, x\}$ is $ij|kx$, then a weight of 1 is added to the edges on the path in T connecting the two leaves i and j. This process is repeated for all such quartet trees, and x is then attached to the edge which has the lowest weight. An example is given in Figure 1.4.

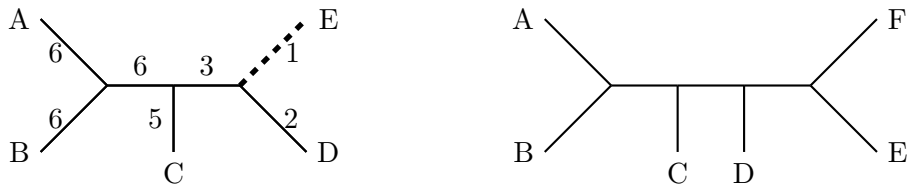


Figure 1.4: Suppose leaf F is to be attached to the five-taxon tree on the left, and the ML trees of $\{i, j, k, F\}$ are: $AB|CF$, $AC|EF$, $BC|DF$, $AC|DF$, $AB|DF$, $AD|EF$, $AB|EF$, $BC|EF$, $BD|EF$ and $CE|DF$. The external edge leading to E returns the minimal weight, so F is attached to this edge, leading to the six-taxon tree shown shown on the right.

Since the order of adding leaves is randomised, this can lead to variations in the resulting tree topologies, and so a consensus tree of numerous replicates is used as the output tree. The program *Tree-puzzle* (TP) (Schmidt *et al.*, 2002) is a parallel version of QP, which performs independent puzzling steps simultaneously.

The trees generated by either the QP or TP process depend on the biological sequences we have for the taxa. To investigate how the TP process behaves on randomized quartets, Vinh *et al.* (2011) performed a simulation study on a so-called *random tree-puzzle* (RTP) process. This assumes that no prior molecular information is given. Therefore, for the same quartet set, all three tree topologies are equally likely. The authors compare the empirical probabilities of tree topologies against the theoretical probabilities from the *proportional to distinguishable arrangements* (PDA) model and the *Yule–Harding* (YH) model.

Table 1 from Vinh *et al.* (2011) reveals that the RTP's empirical probabilities are very close to the YH theoretical probabilities (indeed, there are two cases where these

probabilities are identical). As it seems that the differences between the empirical and theoretical probabilities decrease as the number of taxa increases, Vinh *et al.* (2011) suggest that the RTP process converges to the YH process as n (the number of taxa) grows. The authors provided further evidence for their conjecture by comparing some properties of RTP trees with YH trees. A *cherry* in a binary tree is a pair of leaves that are adjacent to the same vertex. Vinh *et al.* (2011) found that the mean and variance of the number of cherries were similar for the RTP simulation and the theoretical values under the YH process (McKenzie and Steel, 2000). Although Vinh *et al.* (2011) provided evidence to suggest that the two distributions appear to become very similar as n grows, they did not provide a formal statement or proof of their claim that the two distributions converge.

1.3 Overview

1.3.1 Thesis objectives

This thesis aims to solve the following problems:

1. verify whether or not the RTP process converges to the Yule process;
2. compute the probability of $k \geq 1$ reciprocally monophyletic groups;
3. determine the correlation between two monophyletic groups;
4. compute gene tree probabilities of given species networks with the presence of incomplete lineage sorting;
5. simulate bifurcating genealogies from a species network;
6. simulate multifurcating genealogies;
7. determine whether level-k networks are identifiable by genealogy frequency in some cases.

1.3.2 Thesis structure

In answering the questions in the previous section, Chapter 2 starts by introducing some of the basic concepts of graphs, probability theory and stochastic processes that one needs in order to understand some of the mathematical expressions in this thesis.

Chapter 3 first formalises the conjecture that an RTP process leads to a YH distribution as the number of taxa becomes large, and provides a result that shows that leaves are highly likely to attach to pendant edges from some point, which suggests that the conjecture may be true. However, Chapter 3 presents evidence that, while the two distributions are close, RTP appears to converge to a different distribution than YH.

Chapter 4 formalises an identity between the Yule process and the coalescent process, and introduces the *Yule–Harding–Kingman* (YHK) model. It then extends earlier work, using the *exchangeability property* (EP) and the *group elimination* (GE) property of Yule

trees, to derive exact formulas for computing the probabilities of clades in rooted trees. The clade probabilities are then used to derive the probabilities of *clans*, which refers to monophyletic groups in unrooted trees. In the same chapter, some of the proofs are expanded in detail.

Since the PDA model also shares the EP and GE properties, clade probabilities are investigated in the PDA model, and comparisons are made between the YHK process and the PDA model, which will be discussed in Chapter 5.

Chapter 6 develops a novel coalescent model to calculate gene tree probabilities when a complex species network has been given. This model will apply a similar idea to that of Meng and Kubatko (2009) and Yu *et al.* (2012) to decompose a network into its equivalent trees and to compute the gene tree probabilities based on the species trees (Degnan, 2010; Degnan and Salter, 2005).

The algorithm for simulating gene trees from networks is investigated in Chapter 7. Several coalescent processes will be used during simulation. Both gene trees and species trees are unconstrained by a bifurcation assumption. This method supports both the Kingman coalescent and the multiple merger coalescent in species networks.

Conclusions and future work will be discussed in Chapter 8.

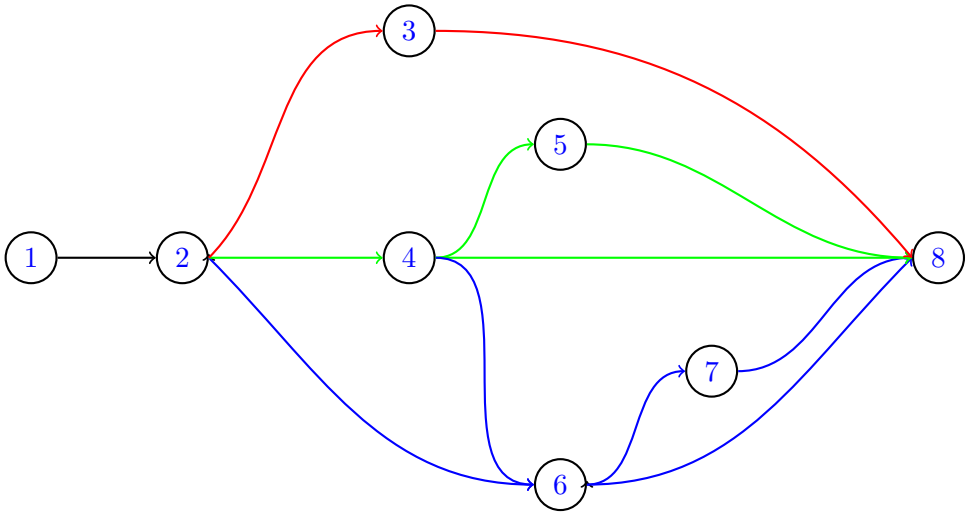


Figure 1.5: Reading guide.

Thesis outcomes

An earlier paper suggested that the so-called random tree-puzzle process converges to the Yule–Harding process with an increased number of leaves in a tree. In Chapter 3, Prof. Mike Steel and I investigate this claim and find that the reasons behind this conjecture might be true, but we presented evidence to suggest that the two processes are different. The results have been submitted to and will be published in *Mathematical Biosciences*, under the title of "Does Random Tree Puzzle produce Yule–Harding trees in the many-taxon limit?" (Zhu and Steel, 2013).

In Chapter 4, we use the EP and the GE properties of the Yule trees to derive exact formulas for computing the probabilities of clades in rooted trees. We then used the clade probabilities to derive the clan probabilities in the unrooted cases. These results were published in *Theoretical Population Biology* volume 79 (page 220–227) under the title of "Clades, clans and reciprocal monophyly under neutral evolutionary models" (Zhu *et al.*, 2011a), as a result of joint work with Prof Mike Steel and Dr James Degnan. This paper was ranked in the 25 hottest articles in *Theoretical Population Biology* from April to June 2011.

The results in Chapter 4 were also presented at Phylomania (the University of Tasmania theoretical phylogenetics meeting) 2011 and the 16th Annual New Zealand Phylogenetics Meeting 2012, along with some partial results of Chapter 5, which extends some of the results in Chapter 4 to the PDA model. A manuscript is in preparation by Dr Taoyang Wu, Dr Cuong Than and myself.

The algorithm described in Chapter 6 was supervised by Dr James Degnan. This novel method of computing the gene tree probabilities of given species network was presented in the form of a poster (Zhu *et al.*, 2011b) at the Institute for Pure and Applied Mathematics Workshop III: Evolutionary Genomics Program, at the Allan Wilson Centre annual meeting, at the New Zealand Statistical Association 2011 Conference, and at the Phylogenetics: New data, New Phylogenetic Challenges Follow-up Meeting. The software `hybrid_coal` was developed to calculate the gene tree probabilities of given species networks and to enumerate the analytical probabilities for theoretical analysis. `hybrid_coal` is now available on Google Code (`http://code.google.com/p/hybrid-coal/`).

The software `hybrid-Lambda` is an implementation of the method described in Chapter 7, which will help automate simulation studies of hybridization, allowing for a large number of species network topologies and allowing gene trees to evolve directly within the network. `hybrid-Lambda` is now available on Google Code (`https://code.google.com/p/hybrid-lambda/`). A manuscript describing `hybrid-Lambda` is in preparation by Dr James Degnan, Dr Bjarki Eldon, Dr Sharyn Goldstien and myself.

Chapter 2

Preliminaries

2.1 Sets

Definition 1. In mathematics, a *set* is a collection of distinct objects. The *cardinality* (or the *size*) of a set A is the number of elements in A, denoted as $|A|$. The *empty set* is denoted as $\emptyset$, that is $|\emptyset| = 0$.

For example, $A = \{\text{human, chimpanzee, gorilla}\}$ is a set of primate names and $|A| = 3$. Throughout this thesis, sets will be used in various cases, to refer to the collection of vertices, edges of a graph or tree topologies and so on.

Suppose that there are two sets, A and B. The *union* and *intersection* of A and B are $A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$ and $A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}$ respectively. Sets A and B are *mutually exclusive* if they have no element in common, that is, $A \cap B = \emptyset$. The *complement* of A is $A^c = \{\omega : \omega \notin A\}$. If every element of A is also an element of B, then A is a subset of B, denoted as $A \subset B$; if also $A \neq B$, then A is a proper subset of B, denoted as $A \subsetneq B$.

In phylogenetics and population genetics, we use a *taxon* (plural, *taxa*) to denote a group of organism(s) (de Quieroz and Gauthier, 1990). A monophyletic group (Rosenberg, 2003) is a subset of taxa for which the *most recent common ancestor* (MRCA) is not ancestral to any other taxa, which is a crucial concept in *species delimitation* (de Quieroz, 2007). More details on this, and how to calculate the probabilities of clades in evolutionary models, will be discussed in Chapter 4 and Chapter 5.

2.2 Graphs

2.2.1 Binary trees

Definition 2. A *graph* G is a collection of nodes (vertices) V and a collection of edges E that connect the nodes in V, denoted as $G = (V, E)$. G is a connected graph if all nodes in V are connected.
A *tree* $T = (V, E)$ is a connected graph with no cycles.

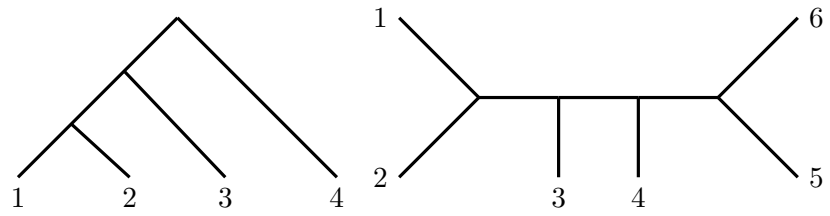
A tree is *rooted* if there is a distinct interior node from which all the other nodes are descended (for example, see Figure 2.1a); otherwise, it is an *unrooted* tree (Figure 2.1b). A tree is *binary* if the degrees of the non-root interior nodes of the tree are, at most, three. For rooted binary trees, the root has degree 2. A *star tree* has only one interior node that is connected to all the other nodes.

If all the nodes are labelled, this tree is called a *labelled tree* (see Figure 2.4b); if only the tip nodes are labelled, it is a semi-labelled tree (see Figures 2.1a).

Note that in this thesis, we use the word *shape* to refer the layout of an unlabelled graph, and use *topology* for a labelled one.

Definition 3. An X-tree T_X is a semi-labelled binary tree with the leaf set X, where $X = \{1, 2, \ldots, n\}$.

Note that some authors (Barthélemy and Guenoche, 1991; Semple and Steel, 2003) define the X-tree differently; our definition of the X-tree is consistent with the *phylogenetic X-tree* by Semple and Steel (2003).



(a) (((1,2),3),4) (b) ((((1,2),3),4),5,6)

Figure 2.1: Example of a rooted tree (a) and an unrooted tree (b).

In phylogenetics and population genetics, the root of a tree refers to the MRCA of all the descendants (Semple and Steel, 2003). A rooted phylogenetic tree is a graph with a direction, which is often accompanied by a time scale, to reflect the historical relationships among the taxa. The concept of a rooted binary tree is vital for this thesis. For example, the *Yule trees*, the *Kingman coalescent trees* and the PDA trees in Chapters 4 and 5 are all rooted trees. The species trees and gene trees or genealogies in Chapter 6 and Chapter 7 are all rooted trees, but are not necessarily binary.

Unlike a rooted tree, an unrooted tree does not need to assume any ancestry relationship among taxa, but only considers how much they differ from each other. Therefore, most of the software packages reconstruct phylogenies from DNA sequences (Felsenstein, 1981; Guindon and Gascuel, 2003; Strimmer and von Haeseler, 1996), build an unrooted tree first, and then use the *outgroup*, which is a taxon (or taxa) that is (are) already known to be less related to the other taxa, to determine the location of the root.

One of the biggest challenge in phylogenetics is that when the number of taxa increases, the number of trees grows exponentially, which makes inference increasingly difficult. The

number of rooted binary trees on n taxa (Felsenstein, 1978) can be computed by:

$$\varphi(n) = \prod_{k=2}^{n}(2k-3) = (2n-3)!!. \tag{2.1}$$

A rooted binary tree of size n is effectively an unrooted binary tree of size $n+1$, by removing one pendant edge. Therefore, the number of unrooted binary trees of size n is $(2n-5)!!$.

2.2.2 Generating random trees

Here we will introduce three different ways of generating a rooted X_n-tree.

The YH model (Yule, 1925; Harding, 1971) is often used to illustrate the speciation process in phylogenetics. One can construct an n-taxon Yule tree by starting from a 2-taxon unlabelled binary tree and then:

1. randomly choosing one of the tip nodes, and split it into two;
2. repeating step 1, until there are n tip nodes; and
3. labelling the tip nodes randomly from $\{1, 2, \ldots, n\}$.

The Kingman coalescent model (Kingman, 1982) is often considered in population genetics to trace the ancestry. One can build an n-taxon Kingman tree by starting from a set of leaves $L = \{1, 2, \ldots, n\}$ and then:

1. randomly choosing two leaves l_1 and l_2 from L;
2. introducing a new leaf node l_{12} in L, which is the MRCA of l_1 and l_2, and then connecting l_1 to l_{12} and l_2 to l_{12};
3. removing l_1 and l_2 from L;
4. repeating steps 1 to 3 until there is only one leaf node left in L.

There might exist multiple ways of generating a tree under the Yule model or the Kingman model, which may result in the probability of one n-taxon tree being different from the other trees.

The PDA model is also known as the 'uniform model'. Unlike the other two models, trees with n taxa are equally likely under this scheme. To construct an n-taxon PDA tree, one needs to first choose the order in which to add the new leaves, then start adding nodes to the 2-taxon labelled tree, by:

1. randomly attaching the new leaf to one of the edges or the root; and
2. repeating step 1 until all leaves are attached to the tree.

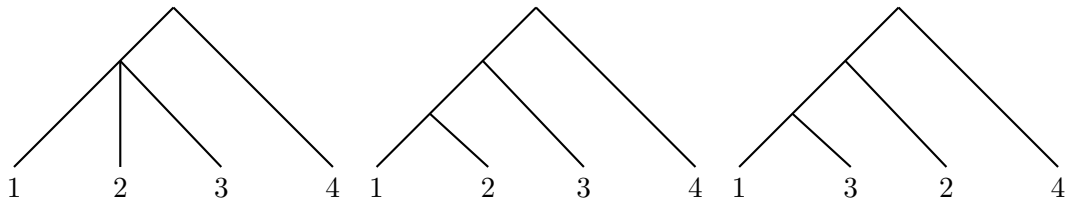
Even though one can simply obtain an unrooted tree by removing the root of a rooted tree, this is not how random unrooted trees are generated. Here, we will discuss how to generate random Yule trees and PDA trees.

One can construct an unrooted n-taxon Yule tree by starting from a 3-taxon unlabelled star tree and then following steps 1, 2 and 3 for constructing an n-taxon rooted Yule tree. Similarly, for an unrooted n-taxon PDA tree, one needs to first choose the order in which to add the new leaves and then start adding nodes to the 3-taxon labelled star tree following steps 1 and 2 for constructing an n-taxon rooted PDA tree.

2.2.3 Polytomy

Genealogies are often assumed to be binary. However, such an assumption may not apply for marine organisms, such as oysters (Beckenbach, 1994), in which a few individuals can produce a massive number of offspring, which leads to a skewed distribution of the offspring number (Eldon and Wakeley, 2006). In this case, one should consider multifurcating genealogies which contains *polytomy* nodes.

Definition 4. A *polytomy* is a node in a rooted tree which has more than two immediate descending branches. In an unrooted tree, a polytomy is a node that is connected to more than three other nodes.



(a) ((1,2,3),4) (b) (((1,2),3),4) (c) (((1,3),2),4)

Figure 2.2: The tree ((1,2,3),4) which has a polytomy node summarises the ancestral histories of the trees (((1,2),3),4) and (((1,3),2),4).

A polytomy is often found in a *consensus tree* , which reflects the common features shared among a set of trees (Semple and Steel, 2003). For example, the trees in Figures 2.2b and 2.2c are distinct, but nodes 1, 2 and 3 are monophyletic in both trees. Thus, we can use the tree shown by Figure 2.2a to summarise the ancestral histories of the other two. This technique is used to describe phylogenies when there are conflicts among gene trees (Felsenstein, 2004). This thesis will also consider phylogenies with polytomies in Chapters 6 and 7.

2.2.4 Compatibility

In a rooted X-tree, let A and B be the two subsets of X. Sets A and B are *compatible* if they satisfy the *compatibility* condition:

$$A \cap B \in \{A, B, \emptyset\}. \tag{2.2}$$

An edge of an unrooted X-tree divides the leaf set X into two non-empty sets, A and A^c. We refer this partition as a *split*, denoted as $A|A^c$. A pair of splits $A|A^c$ and $B|B^c$ are *compatible* if at least one of the sets $A \cap B$, $A \cap B^c$, $A^c \cap B$, or $A^c \cap B^c$ is the empty set .

Throughout this thesis, a split of an unrooted X-tree T is equivalent to a split of X; we use $\Sigma(T)$ to denote the collection of all the splits of T. Two trees T_1 and T_2 are compatible if all of the splits in $\Sigma(T_1)$ are compatible with all the splits in $\Sigma(T_2)$; otherwise, they are *incompatible*.

Unrooted genealogy samples of multiple species are often incompatible, particularly when two of the speciation times are close. Mutations between the lineages may not have occurred before the populations separated. Holland and Moulton (2003) suggested a way of constructing a so-called *consensus split network* from the collection of all splits to visualise the conflicting signals of phylogenies (Holland *et al.*, 2004).

2.2.5 Networks

Huson *et al.* (2010) defines a *phylogenetic network* as *any graph used to represent evolutionary relationships (either abstractly or explicitly) among a set of taxa that labels some of its nodes (usually the leaves).*

In our setting, we consider rooted networks only, which are essentially rooted trees that reflect reticulation events between two edges. For example, if we connect any two edges of a tree, then this graph becomes a network. In the next a few chapters, we are interested in *hybridization networks* (Huson *et al.*, 2010), which describe the evolutionary history with a mixture of speciation and recombination events. If hybridization events happen between hybrid species, this will result in more complicated structures, such as the level-k network , which has, at most, k biconnected reticulation nodes (Huson *et al.*, 2010). In this thesis, we will focus on investigating how these events reflect the probabilities of the genealogies.

2.3 Newick format

Newick formatted strings (Olsen, 1990) are commonly used by many software packages, such as APE (Paradis *et al.*, 2004), for inputting or outputting tree structures. In this scheme, trees are represented in a parenthetic format which labels all the tip nodes, and uses brackets to denote an internal node whose child nodes are separated by a comma. The interior nodes are indicated by complete parentheses, which are not necessarily labelled. For example, the tree in Figure 2.1a is written as `(((1,2),3),4)`.

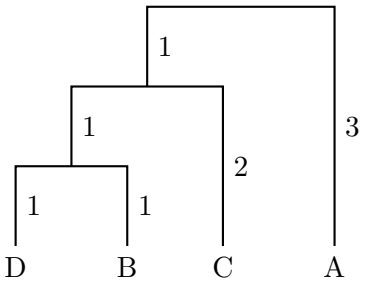


Figure 2.3: Example of a rooted four-taxon binary tree and the associated Newick format string `(((D:1,B:1):1,C:2):1,A:3)`.

Following each node, one can use a colon and numbers to denote the distance between this node and its parent node, which is also known as the branch length. For example, the tree in Figure 2.3 is written as `(((D:1,B:1):1,C:2):1,A:3)`. The branch lengths of a Newick string usually indicate the waiting time for a mutation to occur, or the population divergence time between species. In this thesis, it may also represent the population size or some other quantity. For details, see Appendix D.3.2. If the branch lengths are not specified, the Newick string represents only the topology of a tree.

Note that Newick strings of polytomies can be used to denote unrooted trees. For example, the tree in Figure 2.1b is written as `((((1,2),3),4),5,6)`. In subsequent chapters, it will be stressed when Newick strings are used to denote unrooted trees.

2.3.1 Extended Newick format

Species networks are represented using the *extended Newick format* (Cardona *et al.*, 2008; Huson *et al.*, 2010). This method labels all the internal nodes, in addition to labelling all the tip nodes. This scheme essentially uses an *MUL-tree* to represent a network (see Figure 2.4b) and merges the multi-labelled nodes, which are known as hybrid nodes and are marked with # for convenience. In the network string, the descendants of a hybrid node are recorded before the hybrid node the first time the hybrid node appears. Otherwise, it is written as a tip node. For example, the species network in Figure 2.4a is denoted as `((((B:1.C:1)s_1:1)h_1#H1:1,A:3)s_2:1,(h_1#H1:1,D:3)s_3:1)r`. More examples of the extended Newick formatted strings can be found in Appendices C and D.

In the rest of the thesis, we will use Newick strings to represent trees and extended Newick strings to represent networks. In particular, for the Newick strings and extended Newick strings in Chapters 6 and 7, capital letters are used to denote species and the corresponding lower case letters are used to denote the lineages sampled from these species.

2.4 Probabilities

The *sample space* Ω is the collection of all possible outcomes of a well-defined experiment. For example, if a coin is tossed twice, then $\Omega = \{(Head, Head), (Tail, Head), (Head, Tail), (Tail, Tail)\}$. An *event* $\mathcal{E}$ is a subset of the sample space, with the

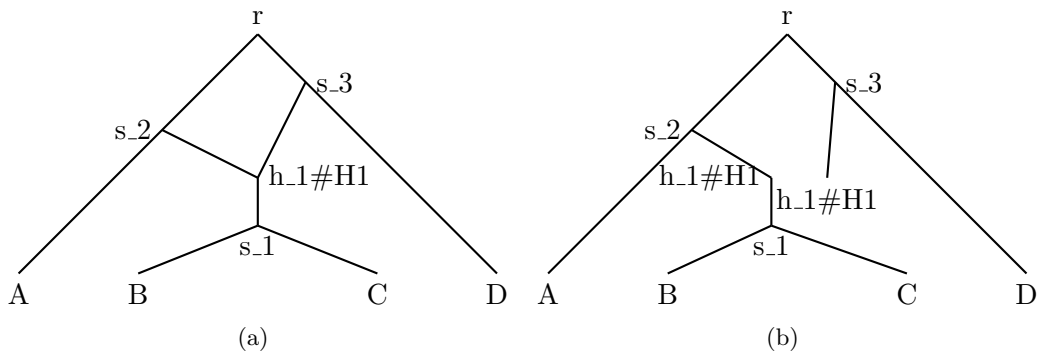


Figure 2.4: Example of a rooted four-taxon network with one hybrid node (a) and the *MUL-tree* (b) which is expressed by the same extended Newick format string `((((B:1.C:1)s_1:1)h_1#H1:1,A:3)s_2:1,(h_1#H1:1,D:3)s_3:1)r` .

probability denoted as $\mathbb{P}(\mathcal{E})$. For example, if $\mathcal{E}$ is the event that at least one tail appears when a conin is tossed twice, then $\mathcal{E} = \{(Tail, Head), (Head, Tail), (Tail, Tail)\}$ and $\mathbb{P}(\mathcal{E}) = \frac{3}{4}$.

Two events, $\mathcal{E}_1$ and $\mathcal{E}_2$, are *independent* of each other if and only if:

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \mathbb{P}(\mathcal{E}_1) \cdot \mathbb{P}(\mathcal{E}_2). \tag{2.3}$$

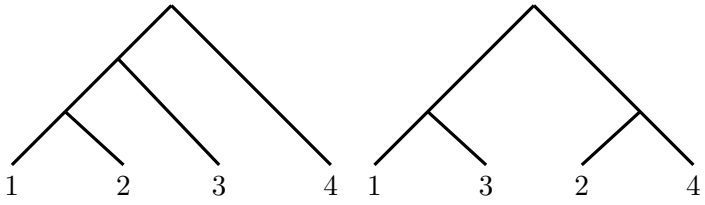


Figure 2.5: Rooted binary tree topologies on 4 taxa.

To relate the concept of probability to our scenario, we are is interested in finding the probability of tree topologies. For the four-taxon trees, T_1 and T_2, in Figure 2.5, under the PDA model:

$$\mathbb{P}(T = T_1) = \mathbb{P}(T = T_2) = \frac{1}{15}.$$

However, under the Yule model and the Kingman model, the probabilities are:

$$\mathbb{P}(T = T_1) = \frac{1}{18} \quad \text{and} \quad \mathbb{P}(T = T_2) = \frac{1}{9}.$$

The probabilities of tree topologies are the same under these two processes, for reasons that will be discussed in detail in Chapter 4.

Note that we occasionally use $\mathbb{P}(T_1)$ to denote $\mathbb{P}(T = T_1)$ in this thesis.

2.4.1 Conditional probabilities

In phylogenetic and population genetic analysis, we often consider scenarios that evolve sequentially, such as the Yule process and the coalescent process. The outcome in the previous step often affects the likelihood of the observation that follows immediately afterwards. For the events $\mathcal{E}_1$ and $\mathcal{E}_2$, with $\mathbb{P}(\mathcal{E}_1) > 0$, we define the *conditional probability* of the event $\mathcal{E}_2$ given $\mathcal{E}_1$ as:

$$\mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) = \frac{\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)}{\mathbb{P}(\mathcal{E}_1)}, \tag{2.4}$$

which is effectively the portion of $\mathcal{E}_2$ in $\mathcal{E}_1$. Note that the events $\mathcal{E}_2$ and $\mathcal{E}_1$ do not have to happen in a sequence. By rearranging Expression (2.4), we can denote the probability of the interaction of the two events $\mathcal{E}_1$ and $\mathcal{E}_2$ as:

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) \cdot \mathbb{P}(\mathcal{E}_1). \tag{2.5}$$

By combining Equations (2.3) and (2.5), we can say that two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent if $\mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) = \mathbb{P}(\mathcal{E}_2)$.

2.4.2 Random variables

Definition 5. A random variable is a function that maps every element of the sample space to a real number.

1. A random variable is *discrete* if the number of the real numbers that this random variable can take, is finite or countably infinite. Examples include:

   - A *binomial random variable* is the number of successes of an experiment that has only two outcomes (success/fail) repeated independently n times. Each time, the probability of success is p. A simple example is to toss a coin several times; the number of heads is then a binomial random variable. In Chapter 3, we find that the edge weights in the RTP process are binomial random variables. If we assume that mutations occur independently with small and equal probabilities for a long DNA sequence, the number of mutations can be approximated by a Poisson random variable.
   - A *Poisson random variable* is the number of events occurring in a time interval at some given rate λ. An example of a Poisson random variable in our case would be the number of mutations occurring during a time period. Details will be discussed in Chapter 7 for simulating the data of *segregating sites*.

2. A random variable is *continuous* if it take uncountab infinite values within an interval.

   - An *exponential random variable* describes the waiting time between two events. For example, it can refer to the time between two speciation events; or in our

case, the waiting time between two coalescent events between $k > 1$ lineages (Chapters 6 and 7).

2.4.3 Probability distribution functions

Here, we define the probability distribution function as:

$$F_X(x) = \mathbb{P}(X \leq x),$$

and the density function as follows:

- If the random variable X is a discrete random variable: $f_X(x) = \mathbb{P}(X = x)$.
- If X is a continuous random variable: $f_X(x) = \dfrac{d}{dx} F_X(x)$.

For random variables $X_1, X_2, \ldots, X_n$, we define the joint distribution function as

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n).$$

Proposition 1. *Two random variables X_1 and X_2 are independent if and only if, for any values of x_1 and x_2:*

$$F_{X_1}(x_1)F_{X_2}(x_2) = F_{X_1,X_2}(x_1, x_2).$$

The *expected value* (or *mean*) of a random variable is the average of all possible values, and is denoted as $\mathbb{E}[X]$. The *variance*, $\text{var}(X)$, of a random variable X is a measure of the spread of the random variable, derived by taking the expected value of the squared difference between X and the mean: $\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

2.4.4 Conditional expectation

The conditional expectation of a random variable conditional on an event, e.g. $\mathbb{E}[X|Y = y]$, is a number depending only on y. The conditional expectation of a random variable conditional on another random variable is not a number, but is also a random variable. Thus, there exists expectation for this random variable, specifically $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

2.4.5 Stochastic processes

A stochastic process refers to a sequence of random variables $X_1, X_2, \ldots, X_n$, that is used to describe the change in some random variable over time or space. The Yule model, the PDA model and the Kingman coalescent model mentioned in the previous section are all examples of discrete stochastic models.

2.4.6 Martingales

A martingale is a special case of a stochastic process, and is defined as follows:

Definition 6. A sequence of random variables $X_1, X_2, \ldots, X_n$ is a martingale if:

1. $\mathbb{E}\left[|X_n|\right] < \infty$;

2. $\mathbb{E}[X_n | X_1, \ldots, X_{n-1}] = X_{n-1}$.

Even though a martingale is a random process, with prior information, one can make some predictions regarding the expected value of the next outcome in the sequence. This idea is often used in betting strategies. In the next section, I will illustrate an example of a martingale: the Polyá urn model, which will assist us in verifying whether the RTP process is or is not the same as the Yule process in Chapter 3.

2.5 Polyá urn model

In practice, we often consider physical models and simulations to observe how random processes behave in order to understand the underlying probabilistic model. Here, we introduce the *Polyá urn* model, which simply is an urn with balls of different colours. By picking and replacing balls with a certain strategy, one can observe how the relative frequencies of the balls change.

Suppose there is an urn containing balls of k different colours. Randomly pick one ball at a time, and place it and another b balls of the same colour into the urn. Let α_i be the initial frequency of balls with colour i, and let the initial total number of balls in the urn be $s_0 = \sum_{i=1}^{k} \alpha_i$. We will use X_n^i to denote the number of balls of colour i when another $n \times b$ balls are placed in the urn. Thus the frequencies of balls is $\mathbf{X}_n = (X_n^1, \ldots, X_n^k)$, and $s_n = \sum_{i}^{k} X_n^i$ denotes the total number of balls when this process is repeated at n times. Let $\mathbf{Z}_n$ be the relative frequencies of balls, i.e. $\mathbf{Z}_n = (Z_n^1, \ldots, Z_n^k)$ and $Z_n^i = \frac{1}{s_n} X_n^i$. Thus, $\mathbf{Z}_n = \frac{1}{s_n}\mathbf{X}_n$ and $\sum_{i=1}^{k} Z_n^i = 1$. When $n \to \infty$, $\mathbf{Z}_n$ is a Dirichlet distribution with the parameter vector $\mathbf{Z}_0$, i.e. $(\frac{\alpha_1}{s_0}, \frac{\alpha_2}{s_0}, \ldots, \frac{\alpha_k}{s_0})$ (Blackwell and MacQueen, 1973). In particular, let $a = \alpha_1 = \alpha_2 = \cdots = \alpha_k$. When $a = b$, $\lim_{n\to\infty} \mathbf{Z}_n$ is uniformly distributed over the vector space $[0, 1]^k$. When $a > b$, $\mathbf{Z}_n$ has a unique mode, and when $a < b$, $\mathbf{Z}_n$ has more than one mode.

Lemma 1. *$\mathbf{Z}_n$ is a martingale (Mahmoud, 2008).*

Proof. For $\mathbf{Z}_n$ to be a martingale, it needs to satisfy:

1. $\mathbb{E}[|\mathbf{Z}_n|] < \infty$;

2. $\mathbb{E}[\mathbf{Z}_n | \mathbf{Z}_0, \ldots, \mathbf{Z}_{n-1}] = \mathbf{Z}_{n-1}$.

1. As $\mathbf{Z}_n = (Z_n^1, \ldots, Z_n^k)$ and $Z_n^i = \frac{1}{s_n} X_n^i$, since $\sum_i^k X_i = s_n$ and X_i are non-negative, we have $0 \leq \frac{X_n^i}{s_n} \leq 1$, i.e. $0 < Z_n^i < 1$. Thus $\mathbb{E}[|\mathbf{Z}_n|] < 1 < \infty$.

2. Let $\mathbf{X}_n$ be the frequency of balls when n balls are placed in the urn. Thus $\mathbf{X}_n = \mathbf{X}_{n-1} + \mathbf{Y}_{n-1}$, where $\mathbf{Y}_{n-1}$ is a random unit vector of length k with probabilities of the relative frequency of balls when an additional $n \times b$ balls are placed, i.e.:

$$\mathbf{Y}_{n-1} = \begin{cases} (b, 0, \ldots, 0) & w.p.\ Z_{n-1}^1; \\ (0, b, \ldots, 0) & w.p.\ Z_{n-1}^2; \\ & \vdots \\ (0, 0, \ldots, b) & w.p.\ Z_{n-1}^k. \end{cases}$$

As $\mathbf{Z}_n = \frac{1}{s_n}\mathbf{X}_n$, $\mathbf{Z}_n = \frac{1}{s_n}(\mathbf{X}_{n-1} + \mathbf{Y}_{n-1})$ and:

$$\mathbb{E}[\mathbf{Z}_n | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}] = \frac{1}{s_n}(\mathbb{E}[\mathbf{X}_{n-1} | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}] + \mathbb{E}[\mathbf{Y}_{n-1} | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}]),$$

where:

$$\begin{aligned} \mathbb{E}[\mathbf{X}_{n-1} | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}] &= \mathbb{E}[s_{n-1}\mathbf{Z}_{n-1} | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}] \\ &= s_{n-1}\mathbb{E}[\mathbf{Z}_{n-1} | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}] \\ &= s_{n-1}\mathbf{Z}_{n-1}, \end{aligned}$$

and $\mathbb{E}[\mathbf{Y}_{n-1} | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}] = b\mathbf{Z}_{n-1}$.
Therefore, $\mathbb{E}[\mathbf{Z}_n | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}] = \frac{s_{n-1} + b}{s_n}\mathbf{Z}_{n-1} = \mathbf{Z}_{n-1}$.

□

Lemma 2. *If $\mathbf{Z}_n$ is a martingale. then $\mathbb{E}[\mathbf{Z}_n]$ is constant. In particular:*

$$\mathbb{E}[\mathbf{Z}_n] = \mathbb{E}[\mathbf{Z}_{n_0}] = (\frac{\alpha_1}{n_0}, \frac{\alpha_2}{n_0}, \ldots, \frac{\alpha_k}{n_0}).$$

Proof. Since $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$, for $\mathbb{E}[\mathbf{Z}_n | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}]$ we have:

$$\mathbb{E}[\mathbb{E}[\mathbf{Z}_n | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}]] = \mathbb{E}[\mathbf{Z}_n].$$

As $\mathbf{Z}_n$ is a martingale, $\mathbb{E}[\mathbf{Z}_n | \mathbf{Z}_{n_0}, \ldots, \mathbf{Z}_{n-1}] = \mathbf{Z}_{n-1}$. Thus $\mathbb{E}[\mathbf{Z}_n] = \mathbb{E}[\mathbf{Z}_{n-1}]$ and:

$$\mathbb{E}[\mathbf{Z}_n] = \mathbb{E}[\mathbf{Z}_{n_0}] = (\frac{\alpha_1}{n_0}, \frac{\alpha_2}{n_0}, \ldots, \frac{\alpha_k}{n_0}).$$

□

2.5.1 Extended Polyá urn model

Consider the following extended Polyá urn (EPU) model. At time $t = 0$, there are b blue balls and r red balls in the urn, where $b \geq 0$ and $r \geq 0$. At each time step, one ball is picked at random from the urn. If the ball is blue, an additional c blue balls and d red balls are placed in the urn; if the ball is red, an additional e blue balls and f red balls are placed in the urn. The variables c, d, e, f can also take negative values, in which case, instead of placing new balls in the urn, the number of balls of the respective colour are withdrawn. We use X_n to denote the number of blue balls at the nth step, and s_n to denote the total number of balls. The following generating matrix describes this process:

$$A = \begin{bmatrix} c & d \\ e & f \end{bmatrix}.$$

This scheme also requires the following assumptions:

1. $c + d = e + f = s \geq 1$ (Bagchi and Pal, 1985; Mahmoud, 2008).
2. At time $t \geq 1$, $0 \leq X_n \leq s_n$ (Bagchi and Pal, 1985).
3. The generator matrix has one real positive principal eigenvalue (Mahmoud, 2008; McKenzie, 2000; McKenzie and Steel, 2000).
4. The components of the principal eigenvector are all strictly positive (Mahmoud, 2008; McKenzie, 2000; McKenzie and Steel, 2000).
5. All eigenvectors are linearly independent (McKenzie, 2000; McKenzie and Steel, 2000).

Bagchi and Pal (1985) show that, for $s_n = b + r + ns$, the probabilities can be defined as follows:

$$\mathbb{P}(X_{n+1} = X_n + c|X_n) = \frac{X_n}{s_n}, \qquad \mathbb{P}(X_{n+1} = X_n + e|X_n) = 1 - \frac{X_n}{s_n}. \tag{2.6}$$

By combining these together, we have:

$$\mathbb{P}(X_{n+1} = k) = \mathbb{P}(X_{n+1} = k|X_n = k - c)\mathbb{P}(X_n = k - c) + \mathbb{P}(X_{n+1} = k|X_n = k - e)\mathbb{P}(X_n = k - e). \tag{2.7}$$

Let λ be the principal eigenvalue of the generating matrix A, and $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ be the normalized eigenvector associated with λ. A classical result (Mahmoud, 2008; Smythe, 1996) shows that, when $n \to \infty$, $\frac{s_n}{n} \xrightarrow{a.s.} \lambda$, and thus $X_n/n \xrightarrow{p} \lambda v_1$. Moreover, it has been shown that $\frac{X_n - \lambda v_1 n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$ (Mahmoud, 2008; Bagchi and Pal, 1985). The initial conditions on b and r do not play any significant role in these convergences.

2.5.2 An example of the EPU model

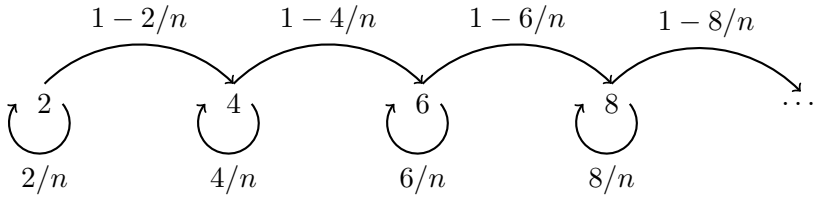
Here, we consider the following example: If a blue ball is picked, place this ball back in the urn with one additional red ball. Otherwise, if a red ball is picked, remove this red ball, but place another two blue balls in the urn. Thus, the generating matrix is:

$$A = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}.$$

We will ignore the initial number of the red balls, but we require the initial number of blue balls to be an even number. We use X_n to denote the number of blue balls, where n is the total number of balls, i.e. $s_n = n$. Since the number of blue balls is always an even number, when we apply Equation (2.6), we have:

$$\mathbb{P}(X_{n+1} = k | X_n = k) = \frac{k}{n}, \qquad \mathbb{P}(X_{n+1} = k+2 | X_n = k) = 1 - \frac{k}{n}.$$

Therefore, the state of X_n is $\{2, 4, 6, \dots\}$, and this Markov chain is as follows:



Therefore, in the limit, this chain goes to infinity when n grows large. Regardless of the initial state, the limiting distribution of this chain is a null vector that takes the value of 1 at infinity, and at all other states takes 0.

Moreover, using Equation (2.7), we have

$$\mathbb{P}(X_{n+1} = 2k) = \mathbb{P}(X_{n+1} = 2k | X_n = 2k)\mathbb{P}(X_n = 2k) + \mathbb{P}(X_{n+1} = 2k | X_n = 2k-2)\mathbb{P}(X_n = 2k-2),$$

which leads to:

$$\mathbb{P}(X_{n+1} = 2k) = \frac{2k}{n}\mathbb{P}(X_n = 2k) + (1 - \frac{2k-2}{n})\mathbb{P}(X_n = 2k-2).$$

In Chapter 3, this example is related to the Yule process to show that the RTP process is different from the Yule process.

Chapter 3

The RTP process and the YH process

Abstract

It has been suggested that an RTP process leads to a YH distribution, when the number of taxa becomes large. In this study, we formalise this conjecture, and we prove that the two tree distributions converge for two particular properties, which suggests that the conjecture may be true. However, we present statistical evidence that, while the two distributions are close, the RTP appears to converge on a different distribution than does the YH.

3.1 Introduction

Although Vinh *et al.* (2011) provided evidence to suggest that the RTP and YH distributions appear to become very similar as n grows, they did not provide a formal statement or proof of their claim that the two distributions converge. In this chapter, we investigate the RTP process further using mathematical and statistical methods. Our results demonstrate that certain properties near the 'periphery' of the tree (i.e. near the leaves) converge under the two distributions; however, the 'deep' structure of the trees (how the tree is broken up around its centroid) appears to retain a trace that distinguishes the two models as the trees become large.

3.2 Formalised conjectures

Given two discrete probability distributions p and q on a finite set Y, the *total variational distance* between p and q is defined as:

$$d_{\mathrm{VAR}}(p,q) = \max_{A \subseteq Y} |\mathbb{P}_p(A) - \mathbb{P}_q(A)|\,,$$

where $\mathbb{P}_p(A) = \sum_{y \in A} p(y)$ and $\mathbb{P}_q(A) = \sum_{y \in A} q(y)$ are the probabilities of event A under the distributions p and q respectively. Thus $d_{\text{VAR}}(p, q)$ is the largest possible probability difference of any event under the distributions p and q. A well-known elementary result is that $d_{\text{VAR}}(p, q) = \frac{1}{2} \sum_{y \in Y} |p(y) - q(y)|$, and thus the two distributions are the same if $d_{\text{VAR}}(p, q) = 0$.

A tree with the leaf set $X_n = \{1, 2, \ldots, n\}$ is called an *X_n-tree*. In the rest of this chapter, all X_n-trees referred to are binary trees, where the interior nodes have degrees of three. We use T_n to denote a labelled X_n-tree topology, and t_n to denote an unlabelled X_n-tree shape. Vinh *et al.* (2011) suggest that when the number of taxa (n) becomes large, the RTP distribution converges to the YH distribution. In this chapter, we consider the total variational distance between the two probability distributions on X_n-trees generated by the RTP and the YH processes, and we formalise the conjecture from Vinh *et al.* (2011). This formalisation states that the variational distance between the two tree distributions converges to zero as the number of taxa grows. We first note that it makes no difference to the truth of this conjecture whether the trees are labelled or unlabelled.

Lemma 3. *Let $\mathcal{T}(n)$ and $\mathcal{S}(n)$ be the set of labelled and unlabelled X_n-trees respectively. For $T_n \in \mathcal{T}(n)$ and $t_n \in \mathcal{S}(n)$, let $\Delta_n := \sum_{T_n \in \mathcal{T}(n)} |\mathbb{P}_{\text{YH}}(T_n) - \mathbb{P}_{\text{RTP}}(T_n)|$ and let $\delta_n := \sum_{t_n \in \mathcal{S}(n)} |\mathbb{P}_{\text{YH}}(t_n) - \mathbb{P}_{\text{RTP}}(t_n)|$. Then $\Delta_n = \delta_n$, and in particular, $\lim_{n \to \infty} \Delta_n = 0 \iff \lim_{n \to \infty} \delta_n = 0$.*

Proof. Let $\nu(t_n)$ be the number of X_n-trees T_n that have the shape t_n. Then, for $\mathbb{P}_{\text{YH}}(T_n) = \frac{\mathbb{P}_{\text{YH}}(t_n)}{\nu(t_n)}$ and $\mathbb{P}_{\text{RTP}}(T_n) = \frac{\mathbb{P}_{\text{RTP}}(t_n)}{\nu(t_n)}$, and we have:

$$\begin{aligned}
\Delta_n &= \sum_{T_n \in \mathcal{T}(n)} |\mathbb{P}_{\text{YH}}(T_n) - \mathbb{P}_{\text{RTP}}(T_n)| \\
&= \sum_{t_n \in \mathcal{S}(n)} \sum_{\substack{T_n \in \mathcal{T}(n) \\ T_n \text{ has shape } t_n}} |\mathbb{P}_{\text{YH}}(T_n) - \mathbb{P}_{\text{RTP}}(T_n)| \\
&= \sum_{t_n \in \mathcal{S}(n)} \nu(t_n) \left| \frac{\mathbb{P}_{\text{YH}}(T_n)}{\nu(t_n)} - \frac{\mathbb{P}_{\text{RTP}}(T_n)}{\nu(t_n)} \right| \\
&= \sum_{t_n \in \mathcal{S}(n)} |\mathbb{P}_{\text{YH}}(t_n) - \mathbb{P}_{\text{RTP}}(t_n)| \\
&= \delta_n.
\end{aligned}$$

Note that here we have applied the *exchangeability property* (Aldous, 1995) of the Yule tree, which states that if two leaf labelled trees have the same shape, i.e. removing the leaf labels leads to the same unlabelled tree, then these two trees are equally probable. Details and applications of this property will be discussed further in Chapters 4 and 5. □

Thus, we formalise the conjecture from Vinh *et al.* (2011) as follows:

Conjecture (strong version):
With $\Delta_n = \delta_n$ defined as above, $\lim_{n\to\infty} \Delta_n = 0$.

Note that, in the YH process, new leaves are only ever attached to pendant edges, and each pendant edge is selected with equal probability. We say that such leaves are attached to *uniformly selected pendant edges*. By contrast, the RTP process can attach new leaves to any edge, although RTP has an increasingly strong preference to attach leaves to pendant edges as the tree grows (Vinh *et al.*, 2011). These authors also suggested that as the tree grows, the number of cherries of an RTP tree follows the same limiting distribution as the number of cherries of a YH tree, which is normally distributed. We summarise these two claims as follows:

Conjecture (weak version)

1. Let $\mathcal{E}_m$ be the event that *all* leaf attachments under the RTP process beyond the first m leaves are to uniformly selected pendant edges. Then $\mathbb{P}(\mathcal{E}_m) \to 1$ as m tends to infinity.

2. The distribution of cherries converges to the same (asymptotic) normal distribution as the YH model.

In this chapter, we prove the two parts of the weak conjecture, and present statistical evidence that the strong conjecture is not true.

3.3 RTP is similar to YH when n is large

To verify Part 1 of the weak conjecture, we need to establish that the probability that a new leaf attaches to a pendant edge converges to 1 sufficiently quickly as the number of leaves increases. This requires that the pendant edges carry less weight than the interior edges. In addition, when the new leaf is added, all pendant edges must be equally likely to be chosen. Thus we must check the edge weight distribution during the puzzling step of the RTP process.

3.3.1 Distribution of edge weights

Let E_n^{P} denote the set of pendant edges of the current X_n-tree T_n and let E_n^{I} be the set of interior edges. For any edge e of T_n, we let $W(e)$ denote the random variable edge weight during the quartet puzzling step. Suppose that the edge e has k leaves of T_n on one side and $n-k$ leaves of T_n on the other side. The following result is established in Appendix B.

Lemma 4. *$W(e)$ is a binomial random variable with the parameters $\dfrac{k(n-k)(n-2)}{2}$ for the number of trials and $\dfrac{2}{3}$ for the probability of success on each trial.*

The parameter k takes the value 1 or $n-1$ for a pendant edge; for an interior edge, k lies between 2 and $n-2$. Next, we show that for a fixed pendant edge and a fixed interior edge,

the probability that the interior edge has lower weight converges to zero exponentially as n increases. More precisely, for any $e'' \in E_n^{\mathrm{P}}$ and any $e' \in E_n^{\mathrm{I}}$, we establish the following result in Appendix B:

$$\mathbb{P}\left(W_n(e'') \geq W_n(e')\right) \leq 2\exp(-\frac{1}{576}n). \tag{3.1}$$

This result is for a fixed pair of pendant and interior edges, but it easily implies that the probability that the smallest weight in the tree is on a pendant rather than an interior edge converges quickly to 1 as n increases. This is formalised in the following inequality, which is also proved in Appendix B:

$$\mathbb{P}\left(\min_{e\in E_n^{\mathrm{P}}}\{W_n(e'')\} \leq \min_{e'\in E_n^{\mathrm{I}}}\{W_n(e')\}\right) \geq 1 - 2n^2\exp(-\frac{1}{576}n). \tag{3.2}$$

Thus a new leaf is almost certain to be added to a pendant edge. Moreover, as noted above, each pendant edge has an equal probability of being attached to.

3.3.2 New leaves rarely attach to interior edges

Theorem 1. *Suppose that $T_m \in \mathcal{T}(m)$, and let $\mathcal{E}_m$ be the event that, under RTP, all leaves beyond T_m are attached to selected pendant edges. Then, for the constants $a, b > 0$:*

$$\mathbb{P}(\mathcal{E}_m) \geq 1 - ae^{-bm}.$$

Moreover, all pendant edges are attached with equal probabilities.

Proof. Let B_k be the event that the $(k+1)$-st leaf is not attached to any pendant edge of T_k. Then we have $1 - \mathbb{P}(\mathcal{E}_m) = \mathbb{P}\left(\bigcup_{k=m}^{\infty} B_k\right)$. By Boole's inequality, we have $\mathbb{P}\left(\bigcup_{k=m}^{\infty} B_k\right) \leq \sum_{k=m}^{\infty}\mathbb{P}(B_k)$. By Inequality (3.2), $\mathbb{P}(B_k) \leq 2k^2\exp(-\frac{1}{576}k)$. We now use the following general inequality, the proof of which is given in Appendix B.

If $Q_m = \sum_{k=m}^{\infty} k^2\exp(-ck)$, where $c \geq \frac{4\log k}{k}$ and $k > 1$, then for $m \geq m_0$:

$$Q_m \leq \frac{\exp(-cm_0/2)}{1-\exp(-c/2)}. \tag{3.3}$$

Thus

$$\begin{aligned} 1 - \mathbb{P}(\mathcal{E}_m) &\leq \sum_{k=m}^{\infty} 2k^2\exp(-\frac{1}{576}k) \\ &\leq \frac{2}{1-\exp(-\frac{1}{576}\times\frac{1}{2})}\exp(-\frac{1}{576}\times\frac{1}{2}m). \end{aligned}$$

Rearranging this inequality establishes the inequality in the theorem. The uniformity

follows by Lemma 4. □

3.3.3 The mean and variance of the number of cherries in the RTP tree

Table 3 in Vinh *et al.* (2011) reveals that the mean and variance of the number of cherries on trees generated under the RTP process and under the YH process are similar. In order to provide a formal proof that they converge to the same limiting distribution, we need to recall the EPU model from Section 2.5.1.

The EPU model and attaching new edges only to pendant edges

We relate the Yule process to the EPU model as follows: Consider the set of cherry edges to be a collection of blue balls, and the non-cherry edges to be a collection of red balls. When a new edge is attached to a pendant edge, if it is attached to a cherry edge, the number of cherry edges remains the same, but the number of non-cherry edges increases by one. If a new edge is added to an non-cherry edge, then the non-cherry edge becomes a cherry edge, and the new edge is also a cherry edge. Thus, the generating matrix is:

$$A = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}.$$

Notice that A has a row sum equal to 1 and one real positive eigenvalue λ, as required.

Let C_n be the number of cherries in a YH tree. Then, as n tends to infinity, $Z_n := (C_n - n/3)/\sqrt{2n/45}$ converges in distribution to a standard normal distribution (i.e. $Z_n \xrightarrow{\mathcal{D}} N(0,1)$), by Corollary 3 of McKenzie and Steel (2000). We now show that the same holds for the distribution of cherries in an RTP tree.

Theorem 2. *Let C_n^* be the number of cherries in an RTP tree and let $Z_n^* = (C_n^* - n/3)/\sqrt{2n/45}$. Then $Z_n^* \xrightarrow{\mathcal{D}} N(0,1)$.*

Proof. We need to show that for any $\epsilon > 0$, and for all sufficiently large values of n and all positive real values of x:

$$|\mathbb{P}(Z_n^* < x) - \mathbb{P}(Z < x)| \leq \epsilon, \tag{3.4}$$

where Z is a standard normal random variable.

As before, let $\mathcal{E}_m$ be the event that after m leaves have been attached to the starting tree by RTP, all further additions are to pendant edges, and let $\mathcal{E}_m^c$ be the complement of $\mathcal{E}_m$. For $n > m$, by the law of total probability, we have:

$$\mathbb{P}(Z_n^* < x) = \mathbb{P}(Z_n^* < x|\mathcal{E}_m)\mathbb{P}(\mathcal{E}_m) + \mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c). \tag{3.5}$$

If we now subtract $\mathbb{P}(Z_n^* < x|\mathcal{E}_m)$ from both sides of Equation (3.5), we obtain:

$$\begin{aligned} & \mathbb{P}(Z_n^* < x) - \mathbb{P}(Z_n^* < x|\mathcal{E}_m) \\ = \; & \mathbb{P}(Z_n^* < x|\mathcal{E}_m)(\mathbb{P}(\mathcal{E}_m) - 1) + \mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c). \end{aligned} \tag{3.6}$$

By the triangle inequality $|a+b| \leq |a| + |b|$, we have:

$$\begin{aligned} & |\mathbb{P}(Z_n^* < x|\mathcal{E}_m)(\mathbb{P}(\mathcal{E}_m) - 1) + \mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c)| \\ \leq \; & |\mathbb{P}(Z_n^* < x|\mathcal{E}_m)(\mathbb{P}(\mathcal{E}_m) - 1)| + |\mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c)|. \end{aligned} \tag{3.7}$$

Combining Equation (3.6) and Inequality (3.7) produces the following:

$$\begin{aligned} & |\mathbb{P}(Z_n^* < x) - \mathbb{P}(Z_n^* < x|\mathcal{E}_m)| \\ \leq \; & |\mathbb{P}(Z_n^* < x|\mathcal{E}_m)(\mathbb{P}(\mathcal{E}_m) - 1)| + |\mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c)|, \\ \leq \; & |\mathbb{P}(Z_n^* < x|\mathcal{E}_m)||(\mathbb{P}(\mathcal{E}_m) - 1)| + |\mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)||\mathbb{P}(\mathcal{E}_m^c)|. \end{aligned} \tag{3.8}$$

Theorem 1 tells us that $\mathbb{P}(\mathcal{E}_m) \geq 1 - ae^{-bm}$, which tends to 1 as m grows for $a, b > 0$. Now, since $\mathbb{P}(\mathcal{E}_m^c) \to 0$ as m tends to infinity, we can select a sufficiently large value of m that $\mathbb{P}(\mathcal{E}_m^c) \leq \epsilon/4$ and $\mathbb{P}(\mathcal{E}_m) \geq 1 - \epsilon/4$. Thus, $\mathbb{P}(\mathcal{E}_m) - 1 \geq -\epsilon/4$, and $|\mathbb{P}(\mathcal{E}_m) - 1| \leq \epsilon/4$. Since $0 \leq \mathbb{P}(Z_n^* < x|\mathcal{E}_m)$, $\mathbb{P}(Z_n^* < x|\mathcal{E}_m^c) \leq 1$, Inequality (3.8) gives:

$$|\mathbb{P}(Z_n^* < x) - \mathbb{P}(Z_n^* < x|\mathcal{E}_m)| \leq \epsilon/4 + \epsilon/4 = \epsilon/2, \tag{3.9}$$

for all sufficiently large values of m, and all $n \geq m$ and $x > 0$.

Now we consider the sequence of Z_n^* conditional on $\mathcal{E}_m$. By conditioning on this event, all the new leaves are attached to uniformly selected pendant edges. Because the EPU argument established that the convergence of the sequence Z_n (the normalisation of the number of cherries in a YH tree) does not depend on the initial number of cherries for any $\epsilon > 0$ and every m, there exists an integer n_0 so that for all $n \geq n_0$ and all $x > 0$:

$$|\mathbb{P}(Z_n^* < x|\mathcal{E}_m) - \mathbb{P}(Z_n < x)| \leq \epsilon/2. \tag{3.10}$$

Then, by the triangle inequality $|a+b| \leq |a| + |b|$, if we add Inequalities (3.9) and (3.10), we have:

$$|\mathbb{P}(Z_n^* < x) - \mathbb{P}(Z_n < x)| \leq \epsilon,$$

and since Z_n converges in distribution to a standard normal distribution, this establishes (3.4). □

Theorem 2 shows that the number of cherries on the RTP trees has a limiting normal distribution with the same asymptotic mean and variance as that for the YH distribution.

We have also shown that, from some point onward, new leaves will always be added to pendant edges, which verifies the weak conjecture. While these two results may be regarded as providing some weak evidence in favour of the strong conjecture, they do not

constitute any formal justification of it. In the next section, we will provide an analysis that suggests that the variational distance between the two distributions remains bounded away from zero as n grows, and this makes these two process distinct in the limit.

3.4 Is RTP the same as YH?

Consider the following scenario where we perform the YH process on some starting tree with more than three leaves, where v is one of the interior nodes. At a node v, the graph is divided into three subtrees (see Figure 3.1). We let L_i $(i = 1, 2, 3)$ denote the leaf sets of these subtrees and let $l_i = |L_i|$ $(i = 1, 2, 3)$ denote the number of leaves in the sets. We normalise the l_i values by the total number of leaves n. Clearly, the sequences l_i/n change as new leaves are gradually added to the whole tree.

3.4.1 Polyá urns and the centroid of a tree

Adding new leaves on to the tree under the YH process ensures that each new leaf is always added into one of the leaf sets L_i, $(i = 1, 2, 3)$. The probability that l_i increases by 1 is proportional to the number of leaves in the subtree relative to the number of leaves in the full tree. This is similar to the Polyá urn problem (Karr, 1993) involving balls of three different colours.

Suppose that one ball is picked randomly at each step, and replaced in the urn along with another ball of the same colour. Let F_n^i be the relative frequency of the ith colour ball when n balls are present, and let $\mathbf{F}_n = (F_n^1, F_n^2, F_n^3)$. $\mathbf{F}_n$ converges (as $n \to \infty$) to a Dirichlet distribution (Kotz *et al.*, 2000) with the parameter vector $\mathbf{F}_{n_0}$, where n_0 is the total initial number of balls. Different initial values in the urn produce different distributions when n balls are present in the urn, and this difference in distribution does not converge to zero as n grows. This result suggests that the YH process on different initial X-trees may well lead to different distributions of the resulting trees. However, if the final tree shape is the only information we are given, then it will be impossible to identify the position of the original vertex v in the final tree with certainty. Thus the frequencies $\mathbf{F}_n$ cannot be clearly measured from the final tree alone. However, we can partly ameliorate this problem by considering a particular vertex that we can easily identify in the final tree, namely its centroid (Jordan, 1869; Mitchell, 1978).

Definition 7. A vertex v of a tree $T = (V, E)$ is a *centroid* if each component of the disconnected graph $T \backslash v$ has, at most, $(1/2)|V|$ vertices.

A well-known property of centroids states that a tree has either a single centroid or two adjacent centroids; in the latter case, $|V|$ is even (Kang and Ault, 1975). To keep the problem simple, we only consider trees with a single centroid. However, because T is a binary tree, $|V|$ is always even, so this does not guarantee a unique centroid. Fortunately, the following lemma shows that a binary tree with an odd number of leaves always has a unique centroid.

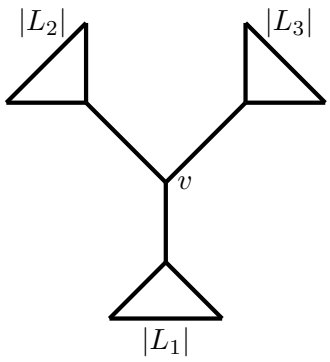


Figure 3.1: The centroid of a tree.

Lemma 5. *Let T be an unrooted binary X_n-tree. Then:*

1. *A vertex v of T is a centroid of T if and only if v satisfies $l_1, l_2, l_3 \leq \frac{n}{2}$, where l_i is the number of leaves in the three subtrees of $T\backslash v$.*

2. *If n is odd, then T has a unique centroid.*

Proof. (1) Suppose that v is an interior vertex of T. Consider the vertex sets V_1, V_2 and V_3 of the connected components of $T\backslash v$. Let l_i be the number of leaves in V_i. Considering the rooted binary tree on V_i, we have:

$$|V_i| = 2l_i - 1. \tag{3.11}$$

Also, since T is an unrooted binary tree, we have:

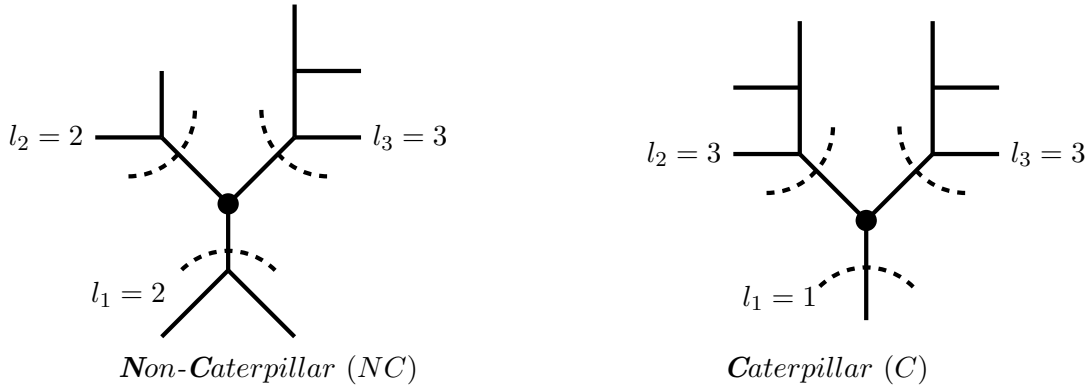
$$|V| = 2n - 2. \tag{3.12}$$

Thus, $|V_i| \leq \frac{1}{2}|V|$ if and only if $2l_i - 1 \leq \frac{1}{2}(2n-2)$ and this holds precisely when $l_i \leq n/2$. Thus, the condition for v to be a centroid (namely that $|V_i| \leq \frac{1}{2}|V|$ for $i = 1, 2, 3$) is precisely the same as that stated in the lemma.

(2) Suppose that v is a centroid of T. At v, we let L_i $(i = 1, 2, 3)$ denote the leaf sets of the subtrees T_i and let l_i denote the size of these leaf sets, ordered so that $l_j \leq l_3 \leq \frac{|X|}{2}$, $(j = 1, 2)$. Since n is odd, we have $l_3 < \frac{n}{2}$.
Suppose another centroid d exists. We use L_i' to denote the complement of L_i. Then there is a subtree H of T rooted at d, with the leaf set L_H, where $L_H \supseteq G'$, and $G' \in \{L_1', L_2', L_3'\}$. Since $l_j \leq l_3 < \frac{n}{2}$, where $j \in \{1, 2\}$, we then have $|L_H| \geq |G'| > \frac{n}{2}$. Therefore, d cannot be a centroid.

□

We now relate the centroid back to the Polyá urn problem. First, notice that tree shapes only start to differentiate when there are more than five leaves. Therefore, in the



***N**on-**C**aterpillar* (NC) ***C**aterpillar* (C)

Figure 3.2: The two tree shapes for binary trees with seven leaves

following scenario, we perform the YH process for initial trees with seven leaves (we start with trees with seven rather than six as we wish to restrict our attention to trees with an odd number of leaves, and which therefore have a unique centroid). Suppose that a tree X is either the non-caterpillar (NC) or the caterpillar (C) tree shown in Figure 3.2. We will use X as the initial tree to construct some tree t_n. At the centroid of t_n when $n = 7$, the sequences l_i/n are $(2/7, 2/7, 3/7)$ and $(1/7, 3/7, 3/7)$ for $t_7 = NC$ and $t_7 = C$ respectively. Now, let us consider only the number of leaves l_1 in the smallest subtree of t_n for all odd values of $n \geq 7$ (henceforth, all values of n in this section are odd to guarantee a unique centroid, and the limits as n tends to infinity are also over just the odd values of n). We define the ratio between l_1 and the number of leaves n to be $\pi_n^X = \dfrac{l_1}{n}$. For $\gamma \in (0, 1)$, let Π^X be the limiting probability of the event $\pi_n^X \geq \gamma$. In other words, $\Pi^X = \lim_{n\to\infty} \mathbb{P}(\pi_n^X \geq \gamma)$. To test the null hypothesis that $\Pi^{NC} = \Pi^C$, we investigate the ratio π_n^X under the YH process starting with a tree with seven leaves with the shape $X \in \{C, NC\}$. An additional 2000 leaves are attached to the starting trees of shape NC and C under the YH process, with 1000 replicates in each case. Using the initial tree of shape NC or C, we found that the probability that π_n^X is greater than $\gamma = 0.19$ does not appear to converge for the two choices of X (NC or C) (see Figure 3.3). Figure 3.3 indicates the 95% confidence interval for the proportions of events for which $\pi_n^X \geq 0.19$, which suggests the following strict inequality:

$$\Pi^{NC} > \Pi^C. \tag{3.13}$$

3.4.2 A modified RTP process

To provide evidence that the RTP and the YH processes are not exactly the same, we define a new process **RTP′**, which is equivalent to the RTP process up to $n = 7$. From this point onward, it proceeds according to the YH process. Therefore, the initial probabilities for constructing X_n-trees from NC and C under the RTP′ process are different from those

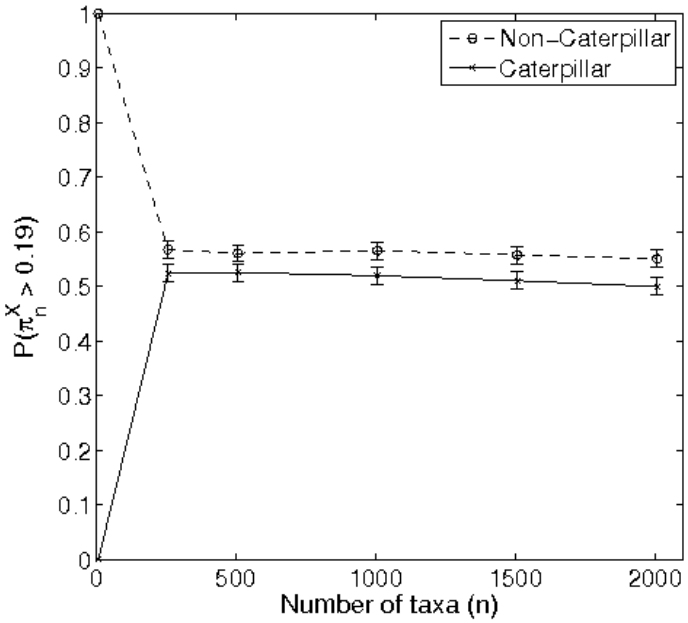


Figure 3.3: Empirical probabilities and the 95% confidence interval for the proportion of events with $\pi_n^X \geq 0.19$. The dashed line is for the initial tree of the non-caterpillar seven-taxon tree; the solid line is for the caterpillar seven-taxon tree.

for the YH process. We use the probabilities of the starting tree NC and C under the RTP process as the probabilities under the RTP$'$ process. Vinh *et al.* (2011) estimated by simulation that the probabilities for the seven-taxon non-caterpillar tree is 0.4607 under the RTP process and 0.4667 under the YH process, which gives us the following inequality:

$$\mathbb{P}_{\mathrm{YH}}(t_7 = NC) - \mathbb{P}_{\mathrm{RTP}'}(t_7 = NC) > 0. \tag{3.14}$$

Theorem 3. *If* (3.13) *holds, then:*

$$\lim_{n\to\infty} d_{\mathrm{VAR}}(\mathbb{P}_{\mathrm{RTP}'}(t_n), \mathbb{P}_{\mathrm{YH}}(t_n)) \neq 0.$$

Proof. Let $\mathcal{S}(n)$ be the set of unlabelled X_n-trees and let:

$$\delta' := \sum_{t_n \in \mathcal{S}(n)} |\mathbb{P}_{\mathrm{YH}}(t_n) - \mathbb{P}_{\mathrm{RTP}'}(t_n)|. \tag{3.15}$$

For a tree generated by YH or RTP$'$, consider the event Σ_n that $\pi_n \geq \gamma$, where $\pi_n = l_1/n$ is the proportion of leaves of the tree with n leaves that lie in the smallest subtree(s) incident with the centroid. Then:

$$\mathbb{P}_{\mathrm{YH}}(\Sigma_n) = \sum_{X \in \{NC, C\}} \mathbb{P}_{\mathrm{YH}}(\Sigma_n | t_7 = X) \mathbb{P}_{\mathrm{YH}}(t_7 = X),$$

and:

$$\mathbb{P}_{\mathrm{RTP}'}(\Sigma_n) = \sum_{X\in\{NC,C\}} \mathbb{P}_{\mathrm{RTP}'}(\Sigma_n|t_7 = X)\mathbb{P}_{\mathrm{RTP}'}(t_7 = X).$$

If we now subtract Equation (3.4.2) from Equation (3.4.2) and substitute $\mathbb{P}_*(t_7 = C)$ in $1 - \mathbb{P}_*(t_7 = NC)$, we have:

$$\begin{aligned} & \mathbb{P}_{\mathrm{YH}}(\Sigma_n) - \mathbb{P}_{\mathrm{RTP}'}(\Sigma_n) \\ = \; & (\mathbb{P}_{\mathrm{YH}}(t_7 = NC) - \mathbb{P}_{\mathrm{RTP}'}(t_7 = NC))\,(\Pi^{NC} - \Pi^{C}). \end{aligned} \tag{3.16}$$

Thus, if we apply Inequalities (3.14) and (3.13) in Equation (3.16), we obtain $\mathbb{P}_{\mathrm{YH}}(\Sigma_n) - \mathbb{P}_{\mathrm{RTP}'}(\Sigma_n) > 0$. Consequently, $\delta' > 0$ in Equation (3.15), and so $\lim_{n\to\infty} d_{\mathrm{VAR}}(\mathbb{P}_{\mathrm{RTP}'}(t_n), \mathbb{P}_{\mathrm{YH}}(t_n)) \neq 0$, as claimed. □

It is important to be clear about what we have established. We have not formally shown that RTP does not converge to YH, nor even that RTP′ fails to converge to YH. Rather, we have provided evidence that a certain property of RTP′ holds and, if so, this implies (Theorem 3) that RTP′ does not converge to YH. Then, since RTP′ is a hybrid of YH and RTP, this suggests that RTP does not converge to YH either.

3.5 Further discussion

In phylogenetic studies, trees are inferred from DNA sequences using various methods. It is also pertinent to ask what sort of trees these methods would produce given entirely random data. This is one of the motivations of the study by Vinh *et al.* (2011). In the following discussion, we use an n by k matrix D to denote a sequence of k independent characters on n taxa. Note that all the characters have the same state space S. The term 'random data' can refer to any one of the following three schemes:

(R1). State x is assigned to taxon i in character j by an independent identically distributed (i.i.d.) process with the probability $p_j(x)$ for $x \in S$.

When the probabilities of state x are the same for all characters (i.e. if $p_j(x) = p(x)$ for all j), we obtain a stronger notion as follows:

(R2). For every entry of the matrix D, D_{ij} is assigned to state x with probability $p(x)$.

If all states are equally likely (i.e. if $p(x) = 1/|S|$), we arrive at an even stronger notion as follows:

(R3). For all entries of D, all states have equal probabilities.

Vinh *et al.* (2011) suggest that random data imply that quartet trees are equally likely and are independent from each other, stating:

> *In our setting, we assume no phylogenetic information in the data. This is equivalent to the assumption that each of the three topologies for a quartet is*

> *equally likely and that the tree topology for each quartet is independent of the other quartets. ... Hence,* $3^{\binom{n}{4}}$ *possible combinations of quartet trees will serve as input to TP.*

For any of the models (R1)–(R3), it is certainly true that random sequence data provide equal support for all three possible topologies for any four taxa. However, this does not necessarily imply that the inferred quartet trees are entirely independent. Rather than pursue this question here, we will consider the behavour of TP under a model in which quartet trees are i.i.d. and uniform, as in Vinh *et al.* (2011).

While the RTP process appears to converge close to the YH distribution, it is instructive to note that another tree reconstruction method, *maximum parsimony* (MP), when applied to random data, appears to converge to a quite different distribution. Here we discuss two interesting observations of MP tree on random data. Let $B(n)$ be the set of unrooted binary trees on the leaf set $\{1, 2, \ldots n\}$.

Theorem 4. *For data D generated under random model* **(R3)** *with two states:*

1. *Each tree $T \in B(n)$ has an identical distribution for its parsimony score on D. In particular, all trees in $B(n)$ have the same expected parsimony score on D.*

2. *For each fixed n, there is a unique MP tree for D with the probability converging to 1 as k grows.*

Proof. Let $w(D, T)$, $T \in B(n)$, denote the parsimony score of T on random data D. By Theorem 7.1 of Steel (1993), the number of ways to colour the leaves of a binary tree T with n leaves using two colours, and so that the resulting colour scheme has a parsimony score of k for T depends only on n and not otherwise on the tree T. Hence, for all $T \in B(n)$, the probability $\mathbb{P}(w(D, T) = l) = f(l)$ is the same for all binary trees with a given number of leaves. The second claim in Theorem 4.1 then follows, by summing over the products of the scores and their probabilities.

Let $E_k(T, T')$ be the event that T and T' have exactly the same parsimony score. Notice that the difference in the parsimony score of T and T' on D is a sum of k i.i.d. random variables under model **(R3)** (notice that each of these k random variables is the difference of two dependent random variables – the score of the character on T and on T' – but these differences are independent across the characters). Thus, by the Central Limit Theorem, the probability that the difference in scores is exactly 0 (i.e. $\mathbb{P}(E_k(T, T'))$ tends to zero as k grows).

Let E be the event that the MP tree for D is unique and let E^c be the complement of this event, namely that there are at least two trees which have the same parisimony score for D. Note that E^c is a subset of the union of the events $E_k(T, T')$ over all T, T' (distinct). Although these events are not independent, Boole's inequality still gives us that:

$$1 - \mathbb{P}(E) \leq \mathbb{P}(\bigcup_{T,T'} E_k(T, T')) \leq \sum_{T,T'} \mathbb{P}(E_k(T, T')) \to 0.$$

Thus $\mathbb{P}(E) \to 1$ as $k \to \infty$, as required. □

Note that Theorem 4 does not establish that the maximum parsimony tree for random data generated under **(R3)** with two states is exactly the PDA distribution (or even asymptotically the PDA distribution as the number k of independent characters tends to infinity). However it suggests the distribution may at least be close to the PDA (and possibly converge to it as k grows). Investigating this further would be an interesting exercise for future work.

Chapter 4

Clade and clan probabilities in the YHK model

Abstract

In this chapter, we derive exact formulae for the probability of a clade and the joint probabilities of $k \geq 2$ clades for a random Yule/coalescent gene tree under the conditions that the k clades are mutually exclusive, and are either exhaustive (all leaves of the gene tree occur in one of the k clades) or form only a subset of the leaves of the gene tree.

In addition, we extend the results to unrooted trees by giving the probabilities of 'clans' (sets of leaves that are all on one side of a split (Wilkinson *et al.*, 2007)), as well as the joint probability of $k > 1$ clans, on Yule/coalescent trees which have been unrooted.

4.1 Introduction

The Yule model and the coalescent model are two neutral stochastic models for describing macro-evolution and micro-evolution respectively. One can use these models to generate random trees: a rooted Yule tree describes the speciation from the root of the tree; a coalescent tree models lineages coalescing back in time from the present. Although, these two models are quite different, they lead to exactly the same distributions of tree topologies.

4.2 The YHK process

Recall the algorithms for constructing a Yule tree and a Kingman coalescent tree in Chapter 2. We show that the two processes produce an identical probability distribution for a rooted binary tree.

Lemma 6. *Let $T(n)$ be the set of labelled rooted X_n-trees. We use $\mathbb{P}_{\mathrm{YH}}$ and $\mathbb{P}_{\mathrm{K}}$ to denote the tree probabilities under the YH model and the Kingman model respectively. For $T \in T(n)$, we have:*

$$\mathbb{P}_{\mathrm{YH}}(T) = \mathbb{P}_{\mathrm{K}}(T).$$

Proof. Let $s(\tau)$ be the number of symmetric nodes of tree shape τ. For an interior vertex v in τ, we use λ_v to denote the number of interior vertices of τ that are descendants of v. Since τ is binary, λ_v is one less than the number of leaves that lie below v. According to Semple and Steel (2003)'s theorem 2.5.2, we have:

$$\mathbb{P}_{\text{YH}}(\tau) = \frac{2^{n-1-s(\tau)}}{\prod_{v \in \mathring{V}} \lambda_v}, \tag{4.1}$$

where $\mathring{V}$ is the set of interior vertices of τ.

We use $\nu(\tau)$ to denote the number of X_n-trees T that have the shape τ. Recall the algorithm for constructing a Yule tree, which implies that any X-tree is equally likely if their unlabelled tree topologies are the same. In particular, step 3 of the algorithm implies that, for n leaves, there are $n!2^{-s(\tau)}$ ways to label the tip nodes of an unlabelled Yule tree randomly, which has also been shown by Semple and Steel (2003)'s corollary 2.4.3.

Thus, we divide the probability of an unlabelled tree (Equation (4.1)) by the number of trees that have the same shape τ, and then obtain the probability of a YH tree as:

$$\mathbb{P}_{\text{YH}}(T) = \mathbb{P}_{\text{YH}}(\tau)/\nu(\tau) = \frac{2^{n-1}}{n!} \prod_{v \in \mathring{V}} \frac{1}{\lambda_v}. \tag{4.2}$$

This result was also derived by Brown (1994) (Equation (4)).

For the Kingman coalescent tree, two nodes are randomly picked from the set of the remaining nodes at each time step. In total, there are the following number of coalescent sequences:

$$\prod_{i=2}^{n} \binom{i}{2} = \frac{n!(n-1)!}{2^{n-1}}. \tag{4.3}$$

However, many of these sequences results in the same tree topology. Suppose that a_j is the number of coalescent events prior to the jth coalescent event. In other words, suppose that the interior vertices of a Kingman coalescent tree are labelled $\{1, 2, \ldots, n-1\}$. Let a_j be the number of interior vertices that are below node j. Then there are:

$$c_T = (n-1)! \prod_{j=1}^{n-1} \frac{1}{1+a_j} \tag{4.4}$$

sequences of coalescences producing the same tree topology (Degnan and Salter, 2005).

Thus if we divide the number of coalescent sequences that form the same Kingman tree (Equation 4.4) by the total number of coalescent sequences among n leaves (Equation (4.3)), we have the probability of tree T under the Kingman process:

$$\mathbb{P}_{\text{K}}(T) = (n-1)! \prod_{j=1}^{n-1} \frac{1}{1+a_j} \times \frac{2^{n-1}}{n!(n-1)!}. \tag{4.5}$$

Since $|\mathring{V}| = n-1$, for $v \in \mathring{V}$, labelling v by j, $j \in \{1, 2, \ldots, n-1\}$, we have $a_j + 1 = \lambda_v$.

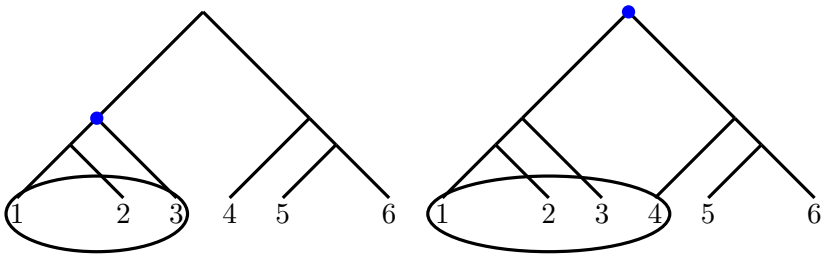
By combining Equations (4.2) and (4.5), we have $\mathbb{P}_{\mathrm{YH}}(T) = \mathbb{P}_{\mathrm{K}}(T)$, as required. □

As we have been shown that the probability distributions under the YH and the Kingman processes are equivalent, from this point on, we generalise these two processes as the YHK process. The probability measure under the YHK process is denoted as $\mathbb{P}$ (short for $\mathbb{P}_{\mathrm{YHK}}$) in this chapter.

We use X_n or X to denote a set of taxa of size n. T_X or T is used to denote a labelled rooted binary tree on X, i.e. X is the leaf set of T. Suppose that ρ is the root of T. We will use $T^{-\rho}$ to denote the unrooted X-tree that is induced from T by deleting ρ. In this chapter, we use $\mathcal{T}_X$ or $\mathcal{T}$ to denote a randomly generated tree on X under the YHK process.

4.3 Clade probabilities

Apart from the tree topologies, the Yule model and the Kingman model also produce the same probabilities for monophyletic groups, also known as clades, which are defined as follows:



(a) $\{1, 2, 3\}$ is a clade. (b) $\{1, 2, 3, 4\}$ is not a clade.

Figure 4.1: An example of a clade. The MRCA of $\{1, 2, 3\}$ only has $\{1, 2, 3\}$ as its descendants. The MRCA of $\{1, 2, 3, 4\}$ is the root, which $\{5, 6\}$ are also the descendants of.

Definition 8. A clade of an X-tree T is a subset of X that corresponds to the set of leaves that are descended from an internal vertex. Suppose that A is a clade of T. If $A \subsetneq X$, then A is a *proper clade* of T. The root of T divides X into two sets, both of which are known as *maximal clades*.

For example, as Figure 4.1a shows, tips 1, 2 and 3 share the same MRCA, of which tips 1, 2 and 3 are the only descendants. Therefore, the set $\{1, 2, 3\}$ is a clade. Moreover, as the MRCA node of $\{1, 2, 3\}$ is a child node of the root, this implies that the set $\{1, 2, 3\}$ is also a maximal clade.. On the other hand, in Figure 4.1b, the MRCA of tips 1, 2, 3 and 4 is the root, which has another two descendants, 5 and 6. Thus, the set $\{1, 2, 3, 4\}$ is not a clade.

In the rest of this section, unless stated otherwise, we use A, B and A_i to denote clades; a, b and a_i are the cardinality of clades A, B and A_i respectively.

4.3.1 Properties of the YHK process

For an X-tree T, we use $T_{X|X'}$ to denote the restricted subtree of T for $X' \subseteq X$. Here, we introduce two properties of the YHK process. Suppose that $\mathcal{T}$ is a random X-tree generated under the YHK process, we use $\mathcal{T}_{X|X'}$ to denote the restricted subtree of $\mathcal{T}$ for $X' \subseteq X$. We have:

> The exchangeability property **(EP)**: If T' is obtained from T by permuting the leaves, then
> $$\mathbb{P}(\mathcal{T} = T') = \mathbb{P}(\mathcal{T} = T).$$
>
> The group elimination **(GE)** property: Let $\mathcal{C}(\mathcal{T})$ denote the collection of clades of the tree $\mathcal{T}$. For any proper (and non-empty) subset A of X, and any rooted binary phylogenetic tree T with the leaf set $X - A$:
> $$\mathbb{P}(\mathcal{T}_{X|(X-A)} = T | A \in \mathcal{C}(\mathcal{T})) = \mathbb{P}(\mathcal{T}_{(X-A)} = T).$$

The EP property (Aldous, 1995) requires that the probability of a particular phylogenetic tree depends only on its shape and not on how its leaves are labelled. This is also known as 'label-invariance' in Steel and Penny (1993).

In the previous section, the EP property is used in the proof of Lemma 6 when deriving the probabilities of labelled trees from the probabilities of unlabelled trees by dividing by the number of labelled trees that have the same shape.

The GE property from Aldous (1995) states that, conditional on A forming a clade in the tree, the tree structure on the remaining taxa is also described by the YHK process. In turn, GE implies the following the *Sampling Consistency* (SC) property (Aldous (1995)) (**SC**):

> For any rooted binary tree T with the leaf set $A \subseteq X$, we have:
> $$\mathbb{P}(\mathcal{T}_{X|A} = T) = \mathbb{P}(\mathcal{T}_A = T).$$

Note that A is not necessarily a clade in this case. To derive the SC property from the GE property, one can gradually remove the leaves that belong to A.

4.3.2 One clade

Lemma 7. *Randomly choose one of the maximal clades of a YHK tree with n leaves. Let U be a random variable that indicates the number of leaves in that clade. We have:*

$$\mathbb{P}_n(U = u) = \begin{cases} \dfrac{1}{n-1}, & u \in \{1, 2, \ldots, n-1\}; \\ 0, & \textit{otherwise.} \end{cases} \textit{, where } n \geq 2. \tag{4.6}$$

This result was derived by Slowinski (1990). For completeness, we give the proof of Lemma 7 by induction in Appendix B.

Lemma 8. *Let $\mathcal{E}_r$ be the event that the maximal clades are of size $(r, n-r)$ in a random generated YHK tree $\mathcal{T}$ of n leaves, where $r \leq \frac{n}{2}$. Then we have:*

$$\mathbb{P}(\mathcal{E}_r) = \begin{cases} \dfrac{2}{n-1}, & 1 \leq r < \dfrac{n}{2}; \\ \dfrac{1}{n-1}, & r = \dfrac{n}{2}. \end{cases} \tag{4.7}$$

Proof. See Appendix B. □

Lemma 9. *Let A_M be the event that A is a maximal clade of a randomly generated YHK tree $\mathcal{T}$ with n leaves. Then:*

$$\mathbb{P}(A_M) = \frac{2}{n-1}\binom{n}{a}^{-1}. \tag{4.8}$$

Proof. The EP property implies that:

$$\mathbb{P}(A \text{ is a clade in } \mathcal{T}|\mathcal{T} \text{ that has } k \text{ clades of size } a) = k\binom{n}{a}^{-1}. \tag{4.9}$$

Let $\mathcal{E}_a$ denote the event that the maximal clades of T have size a and $n-a$. We now have $\mathbb{P}(A_M) = \mathbb{P}(A_M|\mathcal{E}_a) \times \mathbb{P}(\mathcal{E}_a)$, where:

$$\mathbb{P}(A_M|\mathcal{E}_a) = \begin{cases} 2\binom{n}{a}^{-1}, & a = \dfrac{n}{2}; \\ \binom{n}{a}^{-1}, & \text{otherwise.} \end{cases} \tag{4.10}$$

Suppose that $a \leq n-a$. By replacing r in Equation (4.7) with a, and combining Equations (4.7) and (4.10), we show that $\mathbb{P}(A_M) = \dfrac{2}{n-1}\binom{n}{a}^{-1}$ as required. □

Lemma 10. *Let $X_n(a)$ be the number of proper clades of size a in a random YHK tree $\mathcal{T}_X$. Then we have:*

$$\mathbb{E}[X_n(a)] = \frac{2n}{a(a+1)}, \quad 1 \leq a \leq n-1. \tag{4.11}$$

Proof. See Appendix B. □

For a set $A \subsetneq X$, let $p_n(A)$ be the probability that A is a proper clade of $\mathcal{T}_X$. From the EP property it is clear that this probability depends only on $a = |A|$ and n, i.e. the elements of the sets X and A have no effect on the probability $p_n(A)$. Thus, we write $p_n(a)$ for this probability. From Rosenberg (2003), we have the following lemma:

Lemma 11.

$$p_n(a) = \begin{cases} \dfrac{2n}{a(a+1)}\binom{n}{a}^{-1}, & \text{if } 1 \leq a \leq n-1; \\ 0, & \text{otherwise.} \end{cases}$$

The proof of this result from Rosenberg (2003) relies on a combinatorial identity to sum a series. Here, we point out how Lemma 11 follows directly from Lemma 10.

Proof. This is similar to the proof of Lemma 9.

$$\mathbb{P}(A \text{ is a clade in } \mathcal{T}) = \sum_{k\geq 0} \mathbb{P}(A \text{ is a clade in } \mathcal{T} | \mathcal{T} \text{ that has } k \text{ clades of size } a)\mathbb{P}(X_n(a) = k).$$

From Equation (4.9), we have:

$$\begin{aligned} p_n(a) &= \sum_{k\geq 0} k \binom{n}{a}^{-1} \mathbb{P}(X_n(a) = k) \\ &= \sum_{k\geq 0} \binom{n}{a}^{-1} [k\mathbb{P}(X_n(a) = k)] = \binom{n}{a}^{-1} \sum_{k\geq 0} k\mathbb{P}(X_n(a) = k), \end{aligned}$$

where $\sum_{k\geq 0} k\mathbb{P}(X_n(a) = k) = \mathbb{E}[X_n(a)]$. From Lemma 10, we then have:

$$p_n(a) = \frac{2n}{a(a+1)} \binom{n}{a}^{-1}. \tag{4.12}$$

□

From Lemma 10, it is clear that the expected number of clades with a given size a decreases as a increases. However, the function p_n in Equation (4.12) (the clade probability of a given subset with size a) is 'unimodal', as the following corollary suggests.

Corollary 1. *For $n \geq 3$, we have $p_n(a+1) \leq p_n(a)$ for all $a \leq \Delta(n)$, and $p_n(a+1) \geq p_n(a)$ for $a \geq \Delta(n)$, where $\Delta(n)$ is defined as:*

$$\Delta(n) := \sqrt{n + \left(\frac{n-3}{4}\right)^2} + \frac{n-3}{4}. \tag{4.13}$$

Proof. For $1 \leq a \leq n-2$,

$$\frac{p_n(a+1)}{p_n(a)} = \frac{a(a+1)\binom{n}{a}}{(a+1)(a+2)\binom{n}{a+1}} = \frac{a(a+1)}{(a+2)(n-a)}, \tag{4.14}$$

which is less than or equal to 1 if and only if

$$a(a+1) \leq (a+2)(n-a) \iff 2a^2 - (n-3)a - 2n \leq 0.$$

Solving $2a^2 - (n-3)a - 2n \leq 0$ for $1 \leq a \leq n-2$, we have $p_n(a+1) \leq p_n(a)$ if $a \leq \Delta(n)$. Similarly, one can show that $p_n(a+1) \geq p_n(a)$ for $a \geq \Delta(n)$. □

Note that Equation (4.13) suggests that when n goes to infinity, $\Delta(n) \approx \frac{n}{2}$, which means the minimum of $p_n(a)$ is at $a = \frac{n}{2}$.

4.3.3 Pairs of clades

Let sets A and B be two clades of T. If the two sets are equal, or one is a strict subset of the other, or the two sets are disjoint, then A and B satisfy the compatibility condition (Equation 2.2).

For a pair A, B of disjoint subsets of X, let $\hat{p}_n(A, B)$ be the probability that A and B are *sister clades* of $\mathcal{T}_X$ (i.e. A, B and $A \cup B$ are clades of $\mathcal{T}_X$). Again, by the EP property, this probability depends only on $a = |A|, b = |B|$ and n, and so we denote it as $\hat{p}_n(a, b)$.

First, we conside the special case where A and B are maximal clades, in which case $n = a + b$. From Brown (1994)'s equation (6) (see also Rosenberg (2003)), the probability of this event is the same as in Lemma 9 and can be given as follows:

Lemma 12. *For $1 \leq a \leq n$, we have:*

$$\hat{p}_n(a, n-a) = \frac{2}{n-1}\binom{n}{a}^{-1}.$$

Lemma 12 is the same as Lemma 9, because the maximal clades partition the leaf set of a tree. For a fixed set of leaves; knowing the size of one maximal clade is equivalent to knowing the sizes of both maximal clades.

Corollary 2. *$\hat{p}_n(a, n-a) > \hat{p}_n(a+1, n-a-1)$ for $a < \frac{n-1}{2}$, and $\hat{p}_n(a, n-a) \leq \hat{p}_n(a+1, n-a-1)$ for $a \geq \frac{n-1}{2}$.*

The proof of Corollary 2 follows by the unimodal property of the binomial coefficients:

$$\frac{\hat{p}_n(a+1, n-a-1)}{\hat{p}_n(a, n-a)} = \frac{\binom{n}{a}}{\binom{n}{a+1}} = \frac{a+1}{n-a}, \tag{4.15}$$

which suggests that $\hat{p}_n(a, n-a)$ is minimal if $a = \lceil n/2 \rceil$.

We generalise Lemma 12 slightly as follows:

Lemma 13. *Let $k = a + b \leq n$. Then we have:*

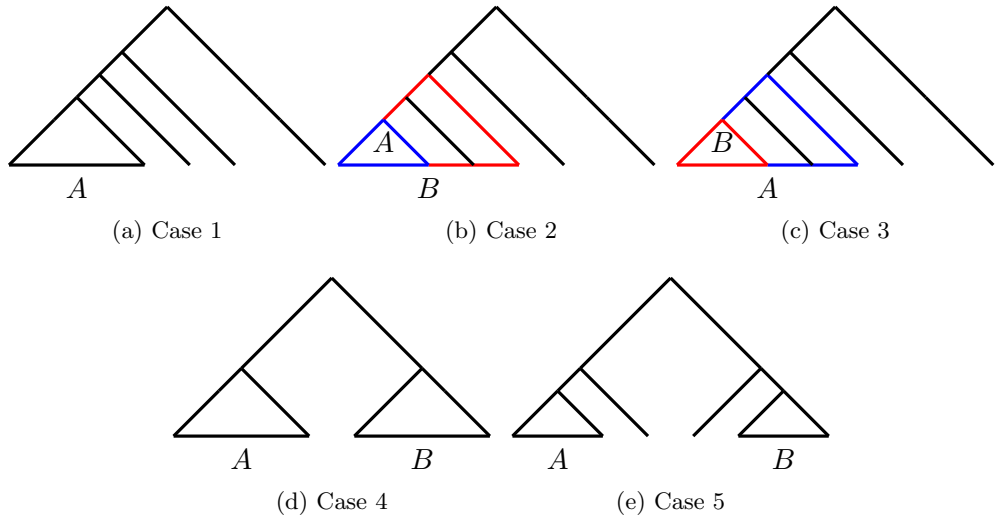
$$\hat{p}_n(a, b) = \frac{4a!b!(n-k)!}{(n-1)!k(k^2-1)}.$$

Proof.

$$\hat{p}_n(A, B) = \mathbb{P}(A \cup B \in \mathcal{C}(\mathcal{T}_X)) \cdot \mathbb{P}\left(A \in \mathcal{C}(\mathcal{T}_{X|A\cup B}) | A \cup B \in \mathcal{C}(\mathcal{T}_X)\right).$$

Applying Lemma 11 to the first term, and property SC and Lemma 12 to the second term, we have:

$$\hat{p}_n(A, B) = \frac{2n}{(a+b)(a+b+1)}\binom{n}{a+b}^{-1} \cdot \frac{2}{a+b-1}\binom{a+b}{a}^{-1},$$



(a) Case 1 (b) Case 2 (c) Case 3

(d) Case 4 (e) Case 5

Figure 4.2: The five cases of compatible clade pairs in Theorem 5.

from which the result follows. □

Now, for any two arbitrary subsets A, B of $X_n = \{1, \ldots, n\}$, let $p_n(A, B)$ be the probability that a YHK tree $\mathcal{T}$ on X_n has A and B as proper clades.

Theorem 5.

$$p_n(A,B) = \begin{cases} p_n(a) & \text{if } A = B \quad [\text{case 1}]\ ; \\ R_n(a,b), & \text{if } A \subsetneq B \quad [\text{case 2}]\ ; \\ R_n(b,a), & \text{if } B \subsetneq A \quad [\text{case 3}]\ ; \\ \hat{p}_n(a, n-a), & \text{if } A \cap B = \emptyset, A \cup B = X_n \ [\text{case 4}]\ ; \\ r_n(a,b), & \text{if } A \cap B = \emptyset, A \cup B \subsetneq X_n \ [\text{case 5}]\ ; \\ 0, & \text{otherwise} \quad [\text{case 6}]\ ; \end{cases}$$

where:

$$p_n(a), \text{ and } \hat{p}_n(a, n-a) \text{ are given by Lemmas 11 and 12,}$$

$$R_n(a,b) := \frac{4n}{a(a+1)(b+1)} \binom{n}{b}^{-1} \binom{b}{a}^{-1},$$

$$r_n(a,b) := \frac{4a!b!(n-a-b)!}{(n-1)!} G_n(a,b), \text{ and where}$$

$$G_n(a,b) := \frac{n}{ab(a+1)(b+1)} - \frac{a(a+1)+b(b+1)+ab}{ab(a+1)(b+1)(a+b+1)} + \frac{1}{(a+b)((a+b)^2-1)}.$$

Proof. Cases 1 and 4 are given by Lemmas 11 and 12, respectively. For the second case ($A \subsetneq B$), similar to the proof of Lemma 13, we have

$$p_n(A,B) = \mathbb{P}(B \in \mathcal{C}(\mathcal{T}_X)) \cdot \mathbb{P}(A \in \mathcal{C}(\mathcal{T}_X) | B \in \mathcal{C}(\mathcal{T}_X)).$$

Since $A \subsetneq B$, we can apply the SC property and Lemma 11 to deduce that the first term in this product is $\frac{2b}{a(a+1)}\binom{b}{a}^{-1}$, while the second term is $\frac{2n}{b(b+1)}\binom{n}{b}^{-1}$, from which the result follows. Case 3 follows by an analogous argument. For Case 5, consider the following two pairs of events:

- $\mathcal{E}_1 : A, B \in \mathcal{C}(\mathcal{T}_X)$,
- $\mathcal{E}_2 : A \cup B, B \in \mathcal{C}(\mathcal{T}_X)$, and
- $\mathcal{F}_1 : A \in \mathcal{C}(\mathcal{T}_{X|(X-B)})$,
- $\mathcal{F}_2 : B \in \mathcal{C}(\mathcal{T}_X)$.

$$\mathbb{P}(\mathcal{E}_1) = \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) - \mathbb{P}(\mathcal{E}_2). \tag{4.16}$$

$$p_n(A, B) = \mathbb{P}(\mathcal{E}_1) = \mathbb{P}(\mathcal{F}_1|\mathcal{F}_2) \cdot \mathbb{P}(\mathcal{F}_2) - \mathbb{P}(\mathcal{E}_2) + \hat{p}_n(A, B). \tag{4.17}$$

Now, by the GE property:

$$\mathbb{P}(\mathcal{F}_1|\mathcal{F}_2) = \mathbb{P}(A \in \mathcal{C}(\mathcal{T}_{X-B})) = p_{n-b}(a), \tag{4.18}$$

and:

$$\mathbb{P}(\mathcal{E}_2) = \mathbb{P}(A \cup B \in \mathcal{C}(\mathcal{T}_X)) \cdot \mathbb{P}(B \in \mathcal{C}(\mathcal{T}_X)|A \cup B \in \mathcal{C}(\mathcal{T}_X)) = p_n(a+b) \cdot p_{a+b}(b). \tag{4.19}$$

$$p_n(A, B) = p_{n-b}(a) \cdot p_n(b) - p_n(a+b) \cdot p_{a+b}(b) + \hat{p}_n(a, b).$$

Case 5 now follows from Lemmas 11 and 13. Case 6 follows from the compatibility condition (2.2) for clades. □

4.3.4 *Correlation between two clades*

We now ask whether the events 'A is a clade' and 'B is a clade' are positively or negatively correlated under the YHK process. Let X_A (respectively X_B) be the Bernoulli (0,1) random variable that takes the value 1 if A (respectively B) is a clade of a YHK tree $\mathcal{T}$ on X_n, and let $\rho_n(A, B)$ denote the correlation coefficient of these two random variables, which is given by:

$$\rho_n(A, B) = \frac{p_n(A, B) - p_n(A)p_n(B)}{\sqrt{p_n(A)(1 - p_n(A))p_n(B)(1 - p_n(B))}}.$$

Corollary 3. *For any two strict subsets A, B of X, the correlation $\rho_n(A, B)$ is:*

- *strictly negative, if A, B are not compatible, and undefined if $|A| = 1$ or $|B| = 1$.*
- *strictly positive, otherwise.*

Proof. If A and B are not compatible, then $p_n(A, B) = 0$, but both $p_n(A)$ and $p_n(B)$ are greater than zero, and so $\rho_n(A, B) < 0$. If $|A| = 1$, then $p_n(A) = 1$ and $p_n(A, B) = p_n(B)$ (regardless of whether A is a subset of B or is disjoint from B). Thus the numerator and denominator of $p_n(A, B)$ are both zero. A similar argument holds if $|B| = 1$.

In the remaining cases, we consider the ratio $p_n(A, B)/(p_n(A)p_n(B))$, and show that it is strictly greater than one. For details, see Appendix B.

□

Figure 4.3 illustrates the correlation coefficient $\rho_n(A, B)$ for $n = 25$ in Cases 2, 4 and 5 from Figure 4.2. Notice that the correlation is typically less in Cases 2 and 5 than in Case 4.

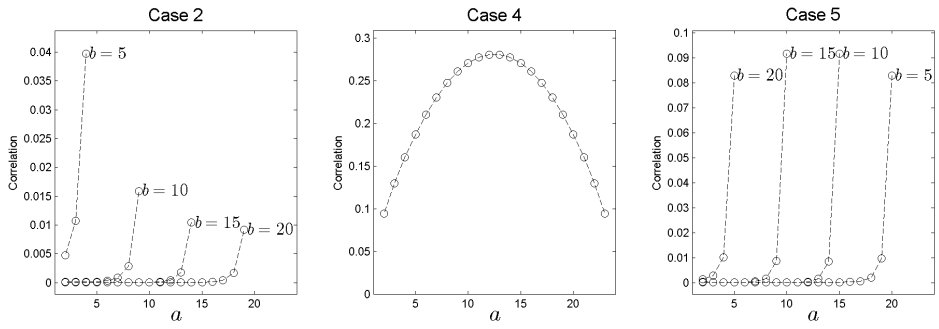


Figure 4.3: Correlation of compatible clades A and B $\rho'_n(A, B)$ for $n = 25$, in Cases in Cases 2, 4 and 5 under the YHK model, with $a = |A|$ and $b = |B|$.

For a rooted YHK tree $\mathcal{T}$, and a rooted phylogenetic tree T_k with leaf set $\{1, \ldots, k\}$, let $p(a_1, \ldots, a_k; T_k)$ be the probability that $A_1, A_2, \ldots, A_k$ are clades of $\mathcal{T}$ and that T_k is the tree obtained from $\mathcal{T}$ by replacing each clade A_i by a single leaf labelled i, for $i = 1, \ldots, k$. Let $\mathcal{I}(T_k)$ denote the set of interior vertices of T_k.

Theorem 6. *For $k > 1$, we have:*

(i)

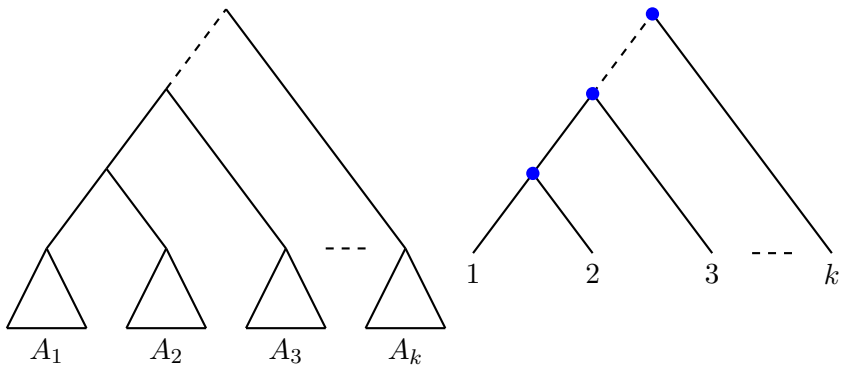
$$p(a_1, \ldots, a_k; T_k) = \frac{2^{k-1} \prod_{i=1}^{k} a_i!}{n!} \prod_{v \in \mathcal{I}(T_k)} \left(\frac{1}{\sum_{i=1}^{k} a_i I_v(A_i) - 1} \right), \tag{4.20}$$

where $I_v(A_i)$ is the indicator variable that takes the value of 1 if A_i lies below v in T_k and 0 otherwise.

(ii)

$$p(a_1, \ldots, a_k) = \sum_{T_k \in RB(k)} p(a_1, \ldots, a_k; T_k), \tag{4.21}$$

where the summation is over all distinct rooted binary phylogenetic trees on the leaf set $\{1, \ldots, k\}$. $RB(k)$ denotes the set of rooted binary trees with leaf set $\{1, 2, \ldots, k\}$.



(a) T_n (b) T_k

Figure 4.4: Disjoined sets A_is partition the leaf set $X = \{1, 2, \ldots, n\}$. The sets A_i are clades of T_n. Replace the MRCA of A_i by i and remove A_i. We then obtain the tree T_k with the leaf set $\{1, 2, k\}$.

Proof. We prove the result by induction on k. For $k = 2$, Lemma 12 gives $p(a_1, a_2; T_2) = \hat{p}_n(a_1, a_2) = \dfrac{2}{n-1}\binom{n}{a}^{-1}$, where $n = a_1 + a_2$, which agrees with the expression given in part (i) with $k = 2$.

Now suppose that part (i) holds whenever k is less or equal to $m \geq 2$. We will show that it also holds when $k = m+1$. Suppose that we have a collection $C = \{A_1, \ldots, A_{m+1}\}$ that partitions X and we also have a rooted binary phylogenetic tree T_{m+1} on the leaf set $\{1, \ldots, m+1\}$. T_{m+1} then has a cherry (two leaves adjacent to the same vertex). Without loss of generality (by re-ordering the sets if necessary), we may suppose that these two leaves are m and $m+1$. Consider the collection of m sets obtained from C by replacing A_m and A_{m+1} by their union, and let T' be the tree obtained from T_{m+1} by deleting the leaves m and $m+1$, along with their incident edges, and labelling the exposed vertex with m. Notice that T' is a rooted binary phylogenetic tree that has the leaf set $\{1, \ldots, m\}$. By the EP property and the GE property (via SC) properties we have, for $a'_m := a_m + a_{m+1}$, the following identity:

$$p(a_1, \ldots, a_{m+1}; T_{m+1}) = p(a_1, \ldots, a'_m; T') \cdot \hat{p}_{a'_m}(a_m, a_{m+1}),$$

where $\hat{p}_{a'_m}(a_m, a_{m+1})$ is the probability that a Yule tree on the leaf set $A_m \cup A_{m+1}$ has A_m and A_{m+1} as sister (and thus maximal) clades. Applying the induction hypothesis for the first term on the right-hand side of this equation, namely $p(a_1, \ldots, a'_m; T')$, then applying Lemma 12 to the second term and collecting terms, leads to the expression in Part (i) for $k = m + 1$ and thereby justifies the induction step.

Part (ii) follows by observing that each tree $\mathcal{T}$ that has $A_1, \ldots, A_k$ as clades has one (and only one) associated tree T_k, and so these trees provide a partition of the event for which the probability is given by $p(a_1, \ldots, a_k)$. □

As an illustration of Theorem 6, we have the following result for $k = 3$:

$$p(a_1, a_2, a_3) = \frac{4a_1!a_2!a_3!}{n!(n-1)} \left[\sum_{i=1}^{3} \frac{1}{n - a_i - 1} \right],$$

where $n = a_1 + a_2 + a_3$. This result is consistent with Brown (1994) Equation (7); similary, we can show Equation (9) (Brown, 1994) for $k = 4$.

We note that, as well as being a generalisation of Lemma 12 to $k > 2$, Theorem 6(i) also generalises the result of Equation (4.2), which computes the probability that a YHK tree $\mathcal{T}$ has a given tree topology T_k. By setting $a_1 = a_2 = \cdots = a_k = 1$ in Theorem 6(i), we have:

$$\frac{2^{n-1}}{k!} \prod_{v \in \mathcal{I}(T_k)} \left(\frac{1}{n_v - 1} \right),$$

where n_v is the number of leaves of T_k below v (see Brown (1994) or Semple and Steel (2003)).

4.3.5 *Computing the probability of k clades recursively*

Note that when using Equations (4.20) and (4.21) to compute the probability of k clades, one needs to enumerate all the tree topologies for rooted binary trees with the leaf set $\{1, 2, \ldots, k\}$, which the results may seem messy, and this is also agreed by Brown (1994).

Rearraging Equation (4.21), we have:

$$p(a_1, \ldots, a_k) = \frac{2^{k-1} \prod_{i=1}^{k} a_i!}{n!} g_k(a_1, a_2, \ldots, a_k),$$

where $g_k(a_1, a_2, \ldots, a_k) = \sum_{T_k \in RB(k)} \prod_{v \in \mathcal{I}(T_k)} \left(\frac{1}{\sum_{i=1}^{k} a_i I_v(A_i) - 1} \right)$.

In this section, we will introduce a method to compute $g_k(a_1, a_2, \ldots, a_k)$ recursively. We consider all the binary rooted trees with k leaves. The root of a tree T_k separates the leaf set into two sets. First, we enumerate all possibe bipartitions of k leaves. Thus, all possibilities for the maximal clades are consided. To do so, we first build two matrices $\mathbf{P}$ and $\mathbf{P}^c$, in which the pair in the jth rows are the bipartitions (non-empty) of an array of ones with length k. The element $\mathbf{P}_{ji}$ of the matrix $\mathbf{P}$ and the element $\mathbf{P}^c_{ji}$ of the matrix $\mathbf{P}^c$ are defined in Algorithm 4.1.

Moreover, each interior node of T_k divides the subset of leaves, into another two sets. Again, all possible bipartitions are consided, which enables us to apply Algorithm 4.1 at each interior node.

Let $\mathbf{a}$ denote the set of clade sizes, i.e. $\mathbf{a} = \{a_1, a_2, \ldots, a_k\}$, so that $|\mathbf{a}| = k$. We denote the jth pair of bipartitions (non-empty) of $\mathbf{a}$ as $\mathbf{a}_j$ and $\mathbf{a}^c_j$, where $\mathbf{a}_j = \{a_i : \text{for all } i \in \{1, \ldots, k\} \text{ such that } \mathbf{P}_{ji} = 1\}$ and $\mathbf{a}^c_j = \{a_i : \text{for all } i \in \{1, \ldots, k\} \text{ such that } \mathbf{P}^c_{ji} = 1\}$.

Algorithm 4.1 Recursive algorithm for constructing bipartition matrcies $\mathbf{P}$ and $\mathbf{P}^c$.

1: **for** j in (1 to $2^{k-1}-1$) **do**
2: Convert j to its binary form, then
$j = \sum_{i=1}^{k} \mathbf{P}_{ji} \times 2^{i-1}$, and
$2^{k-1} - j = \sum_{i=1}^{k} \mathbf{P}^c_{ji} \times 2^{i-1}$.
3: **end for**

Thus, we compute $g_k(a_1, a_2, \ldots, a_k)$ using the function $f(\mathbf{a})$, which is defined as:

$$f(\mathbf{a}) = \begin{cases} 1, & \text{if } k = 1, \\ m(\mathbf{a}) \sum_j f(\mathbf{a}_j) \times f(\mathbf{a}^c_j), & \text{otherwise,} \end{cases}$$

where $m(\mathbf{a}) = \dfrac{1}{\sum_{i=1}^{k} a_j - 1}$.

For example, for $\mathbf{a} = \{a_1, a_2, a_3, a_4\}$, $k = 4$, we have:

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \text{ and } \mathbf{P}^c = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

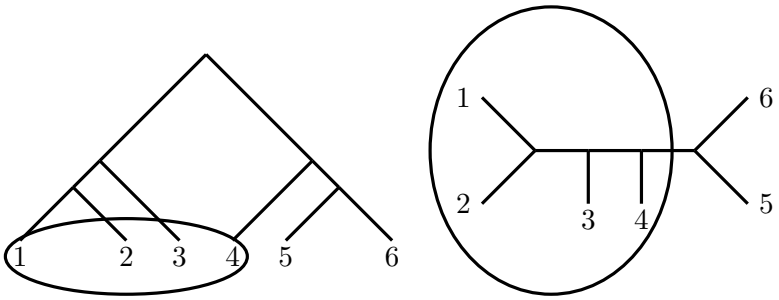
If $j = 3$, then $\mathbf{a}_3 = (a_3, a_4)$ and $\mathbf{a}^c_3 = (a_1, a_2)$.

4.4 Clan probabilities

If we suppress the root ρ of a rooted binary X-tree T (see Figure 4.5a for example), we obtain an unrooted binary X-tree, which is denoted as $T^{-\rho}$ (as shown in Figure 4.5b). For an unrooted binary X-tree Y, following Wilkinson *et al.* (2007) and Lapointe *et al.* (2010), we say that a subset A of X is a *clan* of Y if $A|X - A$ is a split of Y. Note that any clade of the rooted tree T becomes a clan of $T^{-\rho}$. However, this latter tree also has additional clans that do not correspond to clades of T. The precise relationship is given as follows:

Lemma 14. *Given a rooted binary X-tree, T, a set A is a clan of $T^{-\rho}$ if and only if either A is a clade of T or $X - A$ is a clade of T.*

Now suppose the rooted phylogenetic tree T is generated under the YHK process.



(a) T (b) $T^{-\rho}$

Figure 4.5: An example of a clan. $\{1, 2, 3, 4\}$ is not a clade of rooted tree T. $\{1, 2, 3, 4\}$ is a clan of unrooted tree $T^{-\rho}$.

We then obtain an induced probability for the unrooted tree $T^{-\rho}$. Note that the same unrooted tree can arise from different rootings. This probability distribution on unrooted phylogenetic trees can also be described directly as a Yule-type process on unrooted trees in which, at each stage, a leaf is selected uniformly at random and a new leaf (with a random label) is attached to its incident edge (see Steel and Penny (1993)).

For a strict non-empty subset A of X_n, let $q_n(A)$ be the probability that A is a clan of an unrooted YHK tree on the leaf set X_n. By the EP property, this depends only on $a = |A|$ and n so we will also write this probability as $q_n(a)$.

Lemma 15.

$$q_n(a) = 2n\left[\frac{1}{a(a+1)} + \frac{1}{b(b+1)} - \frac{1}{(n-1)n}\right]\binom{n}{a}^{-1},$$

where $a = |A|, b = n - a$.

Proof. By Lemma 14, we have:

$$q_n(A) = p_n(A) + p_n(X - A) - p_n(A, X - A).$$

Applying Lemmas 11 and 12, and noting that $p_n(A, X - A) = \hat{p}_n(A, X - A)$, we obtain the given equation. □

Now consider two disjoint subsets A and B of X, and let $q_n(A, B)$ be the probability that A and B are both clans of an unrooted YHK tree on the leaf set X_n. By the EP property, this probability depends only on $a = |A|, b = |B|$ and n, and so we will denote it as $q_n(a, b)$.

Theorem 7. *(i) If* $a + b = n$, *then:*

$$q_n(a, b) = q_{a+b}(A) = \frac{2a!b!}{(a+b-1)!}\left[\frac{1}{a(a+1)} + \frac{1}{b(b+1)} - \frac{1}{(a+b)(a+b-1)}\right].$$

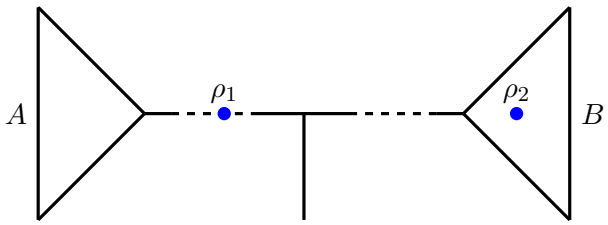


Figure 4.6: An illustration that A and B are clans of $T^{-\rho}$. If the root of T is at position ρ_1, the clan probability follows by the clade probability that the sets A and B are maximal clades. If the root of T is at position ρ_2, then the probability that A and B are clans is equal to the probability that A and $X - B$ are clades of T , but B is not a clade of T.

(ii) If $a + b < n$ then:

$$q_n(a,b) = r_n(a,b)+R_n(a,n-b)+R_n(b,n-a)-\hat{p}_n(b,n-b)p_{n-b}(a)-\hat{p}_n(a,n-a)p_{n-a}(b),$$

where the first three quantities are given in Theorem 5 (Cases 2, 3 and 5), while the last two terms are given by Lemmas 11 and 12.

Proof. Part (i) follows from Lemma 15, noting that $n = a + b$.

For part (ii), Lemma 14 implies that A and B are clans of $T^{-\rho}$ precisely if one of the following three events occur:

$\mathcal{E}_1$: A and B are clades of T;

$\mathcal{E}_2$: A and $X - B$ are clades of T, but B is not a clade of T;

$\mathcal{E}_3$: B and $X - A$ are clades of T, but A is not a clade of T.

Note that $X - A$ and $X - B$ cannot both be clades of T by the compatibility condition (2.2), since $(X - A) \cap (X - B) \neq \emptyset$ by the assumption that $a + b < n$, and since $X - A$ neither contains nor is contained in $X - B$.

Moreover, the three events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ are mutually exclusive by virtue of the assumption that A, B are disjoint and their union is a strict subset of X. Consider an unrooted tree Y and the clans A and B of Y shown in Figure 4.6. The events $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ can be generalised into two cases: the root of T is inside one of the clans, or the root of T is outside the clans A and B.

Suppose that the root of T is outside A and B. The probability of the event $\mathcal{E}_1$ is $r_n(a, b)$. Otherwise, if the root of T is in B, then we have the event $\mathcal{E}_2$, which has the probability $R_n(a, n - b) - \hat{p}_n(b, n - b)p_{n-b}(a)$, since the first term is the probability that A and $X - B$ are clades of $\mathcal{T}$, and $\hat{p}_n(b, n - b)p_{n-b}(a)$ is the probability that $A, X - B$ and B are clades of $\mathcal{T}$. Similarly, $R_n(b, n - a) - \hat{p}_n(a, n - a)p_{n-a}(b)$ is the probability of the event $\mathcal{E}_3$. The result now follows by adding the probabilities of these three mutually exclusive events. □

4.4.1 Extensions of the clan condition (I)

For a pair A, B of disjoint subsets of X, a weaker condition than requiring that A and B are both clans of $\mathcal{T}^{-\rho}$ is simply to require that at least one edge of this tree separates A from B. Let $Q_n(A, B)$ be the probability of this event for an unrooted YHK tree on the leaf set X_n. Then we have the following result, which follows from the SC property applied in the unrooted setting:

$$Q_n(A, B) = q_{a+b}(A), \tag{4.22}$$

where $q_{a+b}(A)$ is given by Theorem 7(i).

4.4.2 Extensions of the clan condition (II)

We now describe a second extension. Suppose that $A_1, A_2, \ldots, A_k$ partition X, and, as usual, let $a_i = |A_i|$. For an unrooted YHK tree $\mathcal{T}$, let $q(a_1, \ldots, a_k)$ be the probability that $A_1, A_2, \ldots, A_k$ are clans of $\mathcal{T}$.

Theorem 8. *Let $n = a_1 + a_2 + a_3$. Then:*

$$q(a_1, a_2, a_3) = \frac{4a_1!a_2!a_3!}{(n-1)!}\left[\sum_{i=1}^{3}\frac{1}{(n-a_i)((n-a_i)^2-1)}\right]. \tag{4.23}$$

Proof. The event that A_1, A_2 and A_3 (which partition X) are clans of $\mathcal{T}^{-\rho}$ is the union of three disjoint events $\mathcal{E}_{jk}$ over the three choices of $\{j, k\} \in \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$, where $\mathcal{E}_{jk}$ is the event that the union of two of the sets – say A_j and A_k – must be a clade of $\mathcal{T}$, and that this clade has the maximal clades A_j and A_k. The EP and GE conditions then give:

$$q(a_1, a_2, a_3) = \mathbb{P}(\mathcal{E}_{12}) + \mathbb{P}(\mathcal{E}_{13}) + \mathbb{P}(\mathcal{E}_{23}) = \sum_{i=1}^{3} p_n(n-a_i)\cdot \hat{p}_{a_j+a_k}(a_j, a_k),$$

where $\{a_i, a_j, a_k\} = \{1, 2, 3\}$ in the term on the right-hand side of this last equation. By Lemmas 11 and 12, this gives:

$$q(a_1, a_2, a_3) = \sum_{i=1}^{3}\frac{2n}{(n-a_i)(n-a_i+1)}\frac{(n-a_i)!a_i!}{n!}\cdot\frac{2}{(n-a_i-1)}\frac{a_j!a_k!}{(n-a_i)!},$$

which simplifies to the expression given in Equation (4.23).

□

4.5 Further discussion

Most of the results in this chapter rely heavily on the application of the EP and the GE properties of the YHK model. Besides the YHK model, the PDA and comb models also

employ these properties. Aldous (1995) has suggested that these three models are the only models which produces a probability distribution on rooted binary trees that satisfy both the EP and GE properties. Under the PDA model, all tree topologies on the same taxon set are equally likely. In the following chapter, we use similar approaches of to those used to derive Lemmas 11 and 12 to show corresponding results under the PDA model. Under the comb model (McKenzie, 2000), this only produces the caterpillar trees, which is not a realistic model in practice, and we do not pursue it further.

Chapter 5

Clade and clan probabilities in the PDA model

Abstract

The Yule model and the PDA model are two neutral evolutionary models that are often used in phylogenetic studies. These models provide different prior probability distributions on tree topologies for Bayesian analysis of tree reconstruction. Many bio-mathematical articles have investigated the properties of the Yule model and the PDA model, giving comparisons between them. In this chapter, we extend the clade probability results under the Yule model, and use an approach similar to the ones introduced by Zhu *et al.* (2011a) to derive formulas for computing clade probabilities under the PDA model.

5.1 Introduction

The Yule model is the most famous and widely used evolutionary model for phylogenetic studies (Blum *et al.*, 2006; Pinelis, 2003). The Yule model is also known as the equal-rates Markov model (Pinelis, 2003), which assumes that the speciation process develops with a constant pure-birth rate. Under the Markov process, the probability of a given topology differs among trees.

The Kingman coalescent model (Kingman, 1982) is often used in population genetics studies, and uses a tree to trace the ancestral histories of individuals backwards in time. Interestingly, the Yule process and the Kingman process produce the same probability distributions on tree topologies, as well as the probabilities that sub-groups of the leaf set form monophyletic groups, which are also known as clades. Zhu *et al.* (2011a) generalized the two processes, calling it the YHK process, in which branch lengths are ignored.

In comparison, biologists consider a null model, namely the PDA model, which assumes that all leaf labelled tree topologies on the same leaf set are equally likely. Therefore, the PDA model is also called the uniform model and is considered to be the simplest stochastic model for phylogenetic studies (Aldous, 2001). It is often used in Bayesian analysis (Li

et al., 2000) and programs such as MrBayes (Huelsenbeck and Ronquist, 2001) for prior probabilities on tree topologies.

It appears natural to consider both the Yule and the PDA models when conducting phylogenetic studies. Numerous researchers have investigated and compared the Yule and PDA processes in various scenarios. Both models can provide prior probabilities on tree topologies when Bayesian approaches are taken (Li *et al.*, 2000; Rannala and Yang, 1996). McKenzie and Steel (2000) show the asymptotic probability distributions of cherries in the phylogenetic trees; Steel (2012) discusses the root location in a random Yule tree or PDA tree; and Blum *et al.* (2006) derives formule for the mean, variance and covariance of the Sackin (Sackin, 1972) and Colless (Colless, 1982) indices in the limit for both the Yule and PDA trees. These indices are implemented in the `R` package `apTreeshape` (Bortolussi *et al.*, 2006) for measuring the balance of phylogenetic trees.

To our knowledge, studies of clade probabilities under the PDA model are still insufficient. Many questions remain unanswered, such as how to calculate the probability that a subset of leaves forms a clade, how to calculate the expected value of the number of clades with a given size, and so on. We attempt to answer these questions in this chapter.

At the end of the previous chapter, we discussed that, besides the YHK model, the PDA model also satisfies the EP property , the GE property and the SC property (Aldous, 1995). This enables us to produce most of the results for clade and clan probabilities demonstrated by Zhu *et al.* (2011a) and in Chapter 4, under the PDA model instead of the YHK model.

5.2 Notation

In this chapter, we continue using most of the notation from the previous chapter. We use X_n or X to denote a set of taxa of size n. T_X or T is used to denote a labelled and rooted binary tree on X, where X is the leaf set of T. Y is used to denote a labelled and unrooted binary tree on X.

The PDA model assumes that all the tree topologies on the same leaf set are equally likely. Therefore, the PDA model is also known as the uniform model. Recall that the number of rooted binary trees with n leaves, according to Equation (2.1) is:

$$\varphi(n) = \frac{(2n-2)!}{2^{n-1}(n-1)!}. \tag{5.1}$$

We use $\mathbb{P}_{\text{PDA}}(\mathcal{E})$ to denote the probability of the event $\mathcal{E}$ under the PDA process. Thus, the probability of a rooted X-tree T is:

$$\mathbb{P}_{\text{PDA}}(T) = \frac{1}{\varphi(n)}. \tag{5.2}$$

Since a rooted binary tree of size n is effectively an unrooted binary tree of size $n+1$ by removing one pendant edge, under the PDA model, the probability of an unrooted

X-tree Y is $1/\varphi(n-1)$.

We use the same definitions for a clade, a proper clade and a maximal clade as in the previous chapter. Recall that a clade of an X-tree T is a subset of X that corresponds to the set of leaves that are descended from an internal vertex. We use A, B and A_i to denote clades, and a, b and a_i to denote the cardinality of clades A, B and A_i respectively.

5.2.1 Properties of the PDA model

For an X-tree T, recall that $T_{X|X'}$ is the restricted subtree of T for $X' \subseteq X$. Let $\mathcal{T}_X$ or $\mathcal{T}$ denote a randomly generated PDA tree on X. Recall the two properties of the YHK process: the EP and the GE properties, which the PDA process also satisfies:

EP: If T' is obtained from T by permuting its leaves, then

$$\mathbb{P}_{\mathrm{PDA}}(\mathcal{T} = T') = \mathbb{P}_{\mathrm{PDA}}(\mathcal{T} = T).$$

GE: Let $\mathcal{C}(\mathcal{T})$ denote the collection of clades of a tree $\mathcal{T}$. For any proper (non-empty) subset A of X, and any rooted binary phylogenetic tree T with the leaf set $X - A$:

$$\mathbb{P}_{\mathrm{PDA}}(\mathcal{T}_{X|(X-A)} = T | A \in \mathcal{C}(\mathcal{T})) = \mathbb{P}_{\mathrm{PDA}}(\mathcal{T}_{(X-A)} = T).$$

We apply the EP and the GE properties to show most of the results in this chapter. For completeness of the PDA model properties, we list the SC property as follows:

For any rooted binary tree T with leaf set $A \subseteq X$, we have

$$\mathbb{P}_{\mathrm{PDA}}(\mathcal{T}_{X|A} = T) = \mathbb{P}_{\mathrm{PDA}}(\mathcal{T}_A = T).$$

5.3 Clade probabilities

We first show a property of clade probabilities that is true for any clade under any probability measure on rooted binary trees. Let $T(n)$ denote the set of rooted and labelled binary trees on X, where $|X| = n$. For any rooted binary tree $T \in T(n)$, and any set $A \subseteq X$, we define the indicator function of a clade as:

$$\mathcal{I}_T(A) = \begin{cases} 1 & \text{if } A \in \mathcal{C}(T), \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

By this definition, the probability that A is a clade of a randomly generated tree $\mathcal{T}$ is the expected value of $\mathcal{I}_{\mathcal{T}}(A)$:

$$\mathbb{P}(A) = \mathbb{E}[\mathcal{I}_{\mathcal{T}}(A)] = \sum_{T \in T(n)} \mathcal{I}_T(A)\mathbb{P}(\mathcal{T} = T), \tag{5.4}$$

where $\mathbb{P}(\mathcal{T} = T)$ denotes the probability that a randomly generated tree $\mathcal{T}$ has the topology $T \in T(n)$.

Proposition 2. *For any probability measure $\mathbb{P}$ on $T(n)$, we have:*

$$\sum_{A \in \mathcal{C}(\mathcal{T})} \mathbb{P}(A) = 2n - 1. \tag{5.5}$$

Proof. By Equation (5.4), we have:

$$\sum_{A \in \mathcal{C}(\mathcal{T})} \mathbb{P}(A) = \sum_{A \in \mathcal{C}(\mathcal{T})} \sum_{T \in T(n)} \mathcal{I}_T(A)\mathbb{P}(\mathcal{T} = T) = \sum_{T \in T(n)} \mathbb{P}(\mathcal{T} = T) \sum_{A \in \mathcal{C}(T)} \mathcal{I}_T(A).$$

Note that $\sum_{A \subseteq X} \mathcal{I}_T(A)$ counts the number of clades of T, including all trivial clades. Each clade of T is induced by a node of T. Since T is binary and rooted, we have

$$\sum_{A \in \mathcal{C}(T)} \mathcal{I}_T(A) = 2n - 1.$$

It is clear that $\sum_{T \in T(n)} \mathbb{P}(\mathcal{T} = T) = 1$. The proposition then follows. □

Let $X_n(a)$ be the number of proper clades of size a in a randomly generated tree $\mathcal{T}$ on X and $|X| = n$. We have:

$$X_n(a) = \sum_{\substack{A \in \mathcal{C}(\mathcal{T}) \\ |A| = a}} \mathcal{I}_{\mathcal{T}}(A). \tag{5.6}$$

Proposition 3. *Let a be a positive integer no greater than n. If the probability distribution $\mathbb{P}(\mathcal{T})$ on $T(n)$ has the EP property, then:*

$$\mathbb{E}[X_n(a)] = \binom{n}{a} \mathbb{P}(A), \tag{5.7}$$

where $\mathbb{P}(A)$ is the probability that the subset A of X is a clade.

Proof. From Equations (5.4) and (5.6), we have:

$$\mathbb{E}[X_n(a)] = \mathbb{E}\left[\sum_{\substack{A \in \mathcal{C}(\mathcal{T}) \\ |A| = a}} \mathcal{I}_{\mathcal{T}}(A)\right] = \sum_{\substack{A \in \mathcal{C}(\mathcal{T}) \\ |A| = a}} \mathbb{E}[\mathcal{I}_{\mathcal{T}}(A)] = \sum_{\substack{A \in \mathcal{C}(\mathcal{T}) \\ |A| = a}} \mathbb{P}(A),$$

where the sums run over all a-subsets of X. As noted previously, if $\mathbb{P}(\mathcal{T})$ has the EP property, then $\mathbb{P}(A)$ is the same for every subset A of X that has a elements. Furthermore, there are exactly $\binom{n}{a}$ subsets of X of size a. The lemma now follows. □

Note that according to Proposition 3, we can show that Lemma 10 follows from Lemma 11.

5.3.1 Clade probability under the PDA model (I)

In this section, we derive the probability that a set A is a clade of a randomly generated PDA tree. In order to distinguish this from the YHK model, we use $p'_n(A)$ to denote this probability. Since the PDA trees have the EP property, this suggests that the probabilty that A is a clade depends only on the size of a. Therefore, the probabilty that A is a clade is a function of a is also written as $p'_n(a)$.

Similarly, let $\hat{p}'_n(A, X-A)$ denote the probability that the mutually exclusive sets A and $X-A$ are the maximal clades of a PDA tree. Again, according to the EP property, this probability only depends on the size of A. Thus it can be expressed as $\hat{p}'_n(a, n-a)$.

Theorem 9. *Under the PDA model, for a positive integer $a \leq n-1$ we have:*

(i) $$p'_n(a) = \frac{\varphi(a)\varphi(n-a+1)}{\varphi(n)} = \binom{n-1}{a-1}\binom{2n-2}{2a-2}^{-1}.$$

(ii) $$\hat{p}'_n(a, n-a) = \frac{\varphi(a)\varphi(n-a)}{\varphi(n)} = \frac{1}{(2n-2a-1)}\binom{n-1}{a-1}\binom{2n-2}{2a-2}^{-1}.$$

Proof. To show Theorem 9(i), it is sufficient to show that there are $\varphi(a)\varphi(n-a+1)$ PDA trees in the set $T(\mathcal{A}) = \{T : T \in T(n), \text{ such that } \mathcal{I}_T(A) = 1\}$. Without loss of generality, we can assume that $X = \{1, 2, \cdots, n\}$ and $A = \{n-a+1, \cdots, n\}$. Let $X' := (X-A) \cup \{n-a+1\}$. Then each tree in $T(\mathcal{A})$ can be generated by the following two steps: Generate a random PDA tree $\mathcal{T}_{X'}$ on X', of which there are $\varphi(n-a+1)$ equally likely trees (Equation (2.1)). Replace the leaf labelled $n-a+1$ by a random PDA tree $\mathcal{T}_A$ on A, of which there are $\varphi(a)$ trees. In addition, a different choice of tree in the first step or in the second step will result in a different tree in $T(\mathcal{A})$. Since there are $\varphi(n-a+1)$ possible choices in the first step and $\varphi(a)$ choices in the second step, we can conclude that the number of trees in $T(\mathcal{A})$ is $\varphi(a)\varphi(n-a+1)$. Since the probability of each PDA tree in $T(n)$ is $1/\varphi(n)$, by substituting Equation (5.1) for simplicity, Theorem 9(i) then follows.

Similarly, for Theorem 9(ii), let $T(\mathcal{A}^*)$ be the subset of trees in $T(X)$ that contain both A and $X-A$ as clades. We now have $T(\mathcal{A}^*) = \{T : T \in T(n), \text{ such that } \mathcal{I}_T(A) = 1, \text{ and } \mathcal{I}_T(X-A) = 1\}$. A tree in $T(\mathcal{A}^*)$ is uniquely determined by a random PDA tree $\mathcal{T}_A$ on A, and a random PDA tree $\mathcal{T}_{X-A}$ on $X-A$, with these as the two maximal clades. This implies that the number of trees in $T(\mathcal{A}^*)$ is $\varphi(a)\varphi(n-a)$. Hence we have:

$$\begin{aligned}\hat{p}'_n(a, n-a) &= \frac{\varphi(a)\varphi(n-a)}{\varphi(n)} = \frac{1}{(2n-2a-1)}p'_n(a) \\ &= \frac{1}{(2n-2a-1)}\binom{n-1}{a-1}\binom{2n-2}{2a-2}^{-1}.\end{aligned}$$

□

We now show that both $p'_n(a)$ and $\hat{p}'_n(a, n-a)$ are unimodal.

Corollary 4. *For $1 \leq a \leq n-1$, we have:*

(i) *$p'_n(a+1) \geq p'_n(a)$ when $a \geq n/2$, and $p'_n(a+1) \leq p'_n(a)$ when $a \leq n/2$.*

(ii) $\hat{p}'_n(a+1, n-a-1) \geq \hat{p}'_n(a, n-a)$ *when* $a \geq (n-1)/2$, *and* $\hat{p}'_n(a+1, n-a-1) \geq \hat{p}'_n(a, n-a)$ *when* $a \leq (n-1)/2$.

Proof. For $a \leq n-1$, we have:

$$\frac{p'_n(a+1)}{p'_n(a)} = \binom{n-1}{a}\binom{2n-2}{2a}^{-1}\binom{n-1}{a-1}^{-1}\binom{2n-2}{2a-2} = \frac{2a-1}{2n-2a-1}, \tag{5.8}$$

which is greater than or equal to 1 when $2a-1 \geq 2n-2a-1$ or, equivalently, when $a \geq n/2$. On the contrary, Equation (5.8) is less than 1 when $a < n/2$. Hence, Corollary 4(i) holds.

Similarly, to show Corollary 4(ii), we have:

$$\frac{\hat{p}'_n(a+1, n-a-1)}{\hat{p}'_n(a, n-a)} = \frac{2n-2a-1}{2n-2a-3}\frac{p'_n(a+1)}{p'_n(a)} = \frac{2a-1}{2n-2a-3}, \tag{5.9}$$

which is greater than or equal to 1 when $2a-1 \geq 2n-2a-3$ or, equivalently, when $a \geq (n-1)/2$. On the contrary, Equation (5.9) is less than 1 when $a < (n-1)/2$. □

Remark: Note that the clade probabilities under the YHK process also show similar results to Corollary 4:

- Corollary 1 suggests that the probability that a set A, where $|A| = a$, is a clade of a random YHK tree takes a mimimum at $a = \dfrac{n}{2}$ when n is large.
- Corollary 2 suggests that the probability that a pair of maximal clades takes a minimum at $a = \lceil n/2 \rceil$, which is the same as Corollary 4(ii).

Note that combining the results of Theorem 9(i) and Proposition 3 yields the following corollary.

Corollary 5. *Let* $X_n(a)$ *be the number of proper clades of size* a *in a randomly generated PDA tree, for* $1 \leq a \leq n-1$. *Then the expected value of* $X_n(a)$ *is:*

$$\mathbb{E}[X_n(a)] = \binom{n}{a}\binom{n-1}{a-1}\binom{2n-2}{2a-2}^{-1}. \tag{5.10}$$

Recall that under the YHK model, $\mathbb{E}[X_n(a)]$ is a decreasing function of a. However, under the PDA model, the above result implies that $\mathbb{E}[X_n(a)]$ is decreasing for $a < \dfrac{3n-1}{4}$ and increasing for $a \geq \dfrac{3n-1}{4}$. To show this, one can compare 1 with the ratio below:

$$\frac{\mathbb{E}[X_n(a+1)]}{\mathbb{E}[X_n(a)]} = \frac{\binom{n}{a+1}}{\binom{n}{a+1}}\frac{p'_n(a+1)}{p'_n(a)}.$$

According to Equation (5.8), we have:

$$\frac{\mathbb{E}[X_n(a+1)]}{\mathbb{E}[X_n(a)]} = \frac{(n-a)}{(a+1)}\frac{(2a-1)}{(2n-2a-1)},$$

which is greater than or equal to 1 when $(n-a)(2a-1) \geq (a+1)(2n-2a-1)$ or, equivalently, when $a \geq \dfrac{3n-1}{4}$.

5.3.2 A comparison between YHK and PDA

Corollary 6. *For $n > 3$, there exists a number $\kappa(n)$ in $(2, n-1)$ such that $p_n(a) > p'_n(a)$ for $2 \leq a < \kappa(n)$, and $p_n(a) < p'_n(a)$ for $\kappa(n) < a \leq n-1$.*

Proof. Let $g_n(a) = \dfrac{p_n(a)}{p'_n(a)}$. From Equations (4.14) and (5.8), we have:

$$\frac{g_n(a+1)}{g_n(a)} = \frac{a(a+1)(2n-2a-1)}{(a+2)(2a-1)(n-a)},$$

which is less than 1 if $a(a+1)(2n-2a-1) < (a+2)(2a-1)(n-a)$ or, equivalently, when $a > \dfrac{2n}{n+3}$. Hence $g_n(a) > g_n(a+1)$ for $2n/(n+3) < a \leq n-2$. Since $2n/(n+3) < 2$, for $a \geq 2$, we have $g_n(a) > g_n(a+1)$.

It is easy to see that, for $n \geq 3$:

$$g_n(2) = \frac{2(2n-3)}{3(n-1)} \geq 1$$

and:

$$g_n(n-1) = \frac{2(2n-3)}{n(n-1)} \leq 1.$$

This and the fact that $g_n(a)$ is strictly decreasing on $[2, n-1]$ implies the existence of the number $\kappa(n)$ in the theorem. □

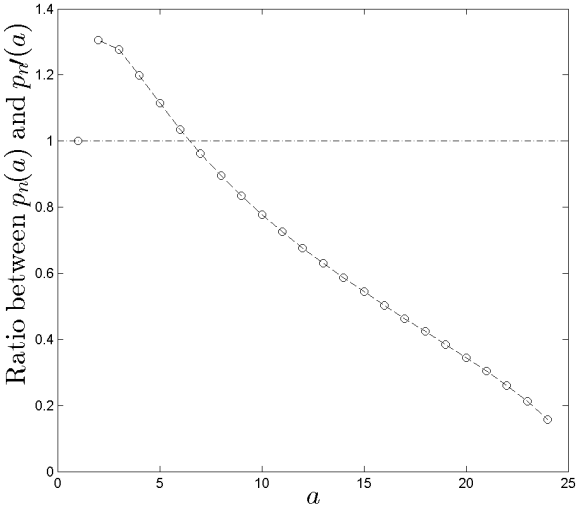


Figure 5.1: The ratio between the clade probabilities in a 25-taxon YHK tree and a 25-taxon PDA tree. It appears that the clade probability under the YHK model is greater than the PDA model when the clade sizes are between two and six for trees with 25 taxa.

Corollary 7. *For $n > 3$, there exists a number $\lambda(n)$ in $(1, \frac{n+1}{2})$ such that $p_n(a, n-a) < p'_n(a, n-a)$ for $1 \le a \le \lambda(n)$, and $p_n(a, n-a) > p'_n(a, n-a)$ for $\lambda(n) < a \le (n+1)/2$.*

Proof. Let $h_n(a) = \dfrac{\hat{p}_n(a, n-a)}{\hat{p}'_n(a, n-a)}$. From Equations (4.15) and (5.9), we have:

$$\frac{h_n(a+1)}{h_n(a)} = \frac{(a+1)(2n-2a-3)}{(2a-1)(n-a)},$$

which is greater than or equal to 1 if $(a+1)(2n-2a-3)-(n-a)(2a-1) = 3(n-2a-1) \ge 0$, which holds for $1 \le a \le (n-1)/2$. Notie that both $\hat{p}_n(a, n-a)$ and $\hat{p}'_n(a, n-a)$ are symmetrical and unimodal at $a = \dfrac{n-1}{2}$. Thus, $\dfrac{h_n(a+1)}{h_n(a)} \le 1$ if $(n-1)/2 \le a \le (n-1)$. Since $h_n(1) = \dfrac{2(2n-3)}{n(n-1)} \le 1$ and $h_n((n-1)/2) \ge 1$, Corollary 7 follows.

□

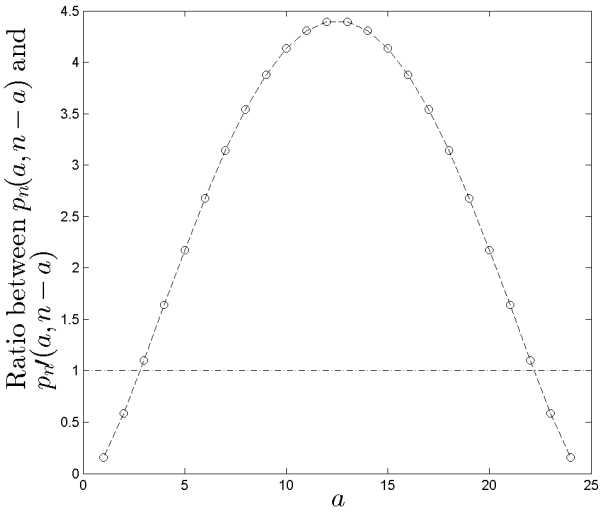


Figure 5.2: The ratio between the maximal clade probabilities in a 25-taxon YHK tree and a 25-taxon PDA tree. It appears that when one of the maximal clades has a size of between 3 and 22, the maximal clade probability under the YHK model is greater than the PDA model for trees with 25 taxa.

5.3.3 Clade probability under the PDA model (II)

Theorem 10. *Let $A_1, \ldots, A_k$ be k disjoint (non-empty) subsets of X, with $|A_i| = a_i$ and $m = \sum_{i=1}^{k} a_i$. We use $p'_{(m,n)}(A_1, \ldots, A_k)$ to denote the probability that a random PDA tree $\mathcal{T}_X$ has $A_1, \ldots, A_k$ as clades. Since the EP property holds for random PDA trees, this probability only depends on the clade sizes. Therefore, it can be expressed as:*

$$p'_{(m,n)}(a_1, \ldots, a_k) = \frac{\varphi(n-m+k)\prod_{i=1}^{k}\varphi(a_i)}{\varphi(n)}. \tag{5.11}$$

Proof. Since each PDA tree, $\mathcal{T}_X$ on X has a probability of $1/\varphi(n)$, it remains to compute the number of trees that have $A_1, \ldots, A_k$ as clades. Such a tree can be constructed in two steps:

1. Build a tree on $\left(X \setminus \bigcup_{i=1}^{k} A_i\right) \cup \{x_1, \ldots, x_k\}$, where $x_1, \ldots, x_k$ are leaves not in X and serve as "placeholders" in the second step.
2. Replace each x_i with a tree in $\mathcal{T}_{A_i}$.

There are $\varphi(n-m+k)$ different choices of tree in the first step, and there are $\prod_{i=1}^{k} \varphi(a_i)$ different ways to replace $x_1, \ldots, x_k$ with trees from $\mathcal{T}_{A_1}, \ldots, \mathcal{T}_{A_k}$. The claim then follows. □

Remark. Theorem 10 is more powerful than Theorem 6 for the YHK model. In the above theorem, $A_1, \ldots, A_k$ do not necessarily form a partition of X.

Corollary 8. *If $A_1, \ldots, A_k$ form a partition of X then the result of Theorem 10 leads to the following:*

$$p'_n(a_1, \ldots, a_k) = \frac{\varphi(k) \prod_{i=1}^{k} \varphi(a_i)}{\varphi(n)}. \tag{5.12}$$

Note that for $k = 2$, Corollary 8 generalises Theorem 9(ii).

5.3.4 Clade probability under the PDA model (III)

Theorem 11. *Let A and B be two subsets of X with $a = |A|$ and $b = |B|$. Then the probability that A and B are clades of a random PDA tree is:*

$$p'_n(A,B) = \begin{cases} p'_n(a) & \text{if } A = B \quad \text{[Case 1]} ; \\ R'_n(a,b), & \text{if } A \subsetneq B \quad \text{[Case 2]} ; \\ R'_n(b,a), & \text{if } B \subsetneq A \quad \text{[Case 3]} ; \\ r'_n(a,b), & \text{if } A \cap B = \emptyset, A \cup B \subseteq X_n \quad \text{[Case 4]} ; \\ 0, & \text{otherwise} \quad \text{[Case 5]} ; \end{cases}$$

where:

$$p'_n(a) \text{ is given by Theorem 9 (i)},$$

$$R'_n(a,b) = \frac{\varphi(a)\varphi(n-b+1)\varphi(b-a+1)}{\varphi(n)},$$

$$r'_n(a,b) = \frac{\varphi(a)\varphi(b)\varphi(n-a-b+2)}{\varphi(n)}.$$

Case 1 is given by Theorem 9 (i). To show Cases 2 and 3, one can apply Theorem 9 (i) twice to obtain the probability that the subset A of B is a clade of $\mathcal{T}_B$, given that a subset B of X is a clade of $\mathcal{T}_X$, and vice versa. Case 4 follows from Theorem 10 for $k = 2$.

5.3.5 Correlation results under the PDA model (I)

The following technical lemma is useful in our study.

Lemma 16. *Let m, n, m' and n' be positive numbers with $(m-m')(n-n') \geq 0$. Then we have:*

$$\varphi(m'+n')\varphi(m+n) \geq \varphi(n'+m)\varphi(m'+n). \tag{5.13}$$

In particular, if $a \leq b \leq b' \leq a'$ are positive numbers with $a + a' = b + b'$, we then have $\varphi(a)\varphi(a') \geq \varphi(b)\varphi(b')$.

Proof. To establish the first claim, we may assume that $m \geq m'$ and $n \geq n'$, as the proof of the other case, $m \leq m'$ and $n \leq n'$, is similar. Noting that $m \geq m'$ and $n \geq n'$, we have $(2m+2n-3)(2m+2n-5)\cdots(2m+2n'-1) \geq (2m+2n'-3)(2m+2n'-5)\cdots(2m'+2n'-1)$, because each side of this equation has $n-n'$ factors, and, clearly, each factor on the left-hand side is greater than or equal to the corresponding factor on the right-hand side. This leads to:

$$\frac{\varphi(m+n)}{\varphi(n'+m)} \geq \frac{\varphi(m'+n)}{\varphi(m'+n')},$$

from which Equation (5.13) follows.

The second assertion follows from the first by considering $m' = n' = \frac{a}{2}$, $m = b - \frac{a}{2}$ and $n = b' - \frac{a}{2}$. □

Let A and B be two compatible subsets of X, i.e. $A, B \subsetneq X$, with $|A|, |B| \geq 2$.

Theorem 12. *The events 'A is a clade' and 'B is a clade' are positively correlated under the PDA process.*

Proof. Recalling Corollary 3, let X_A (respectively X_B) be the Bernoulli (0,1) random variable that take the value 1 if A (respectively B) is a clade of the X-tree, and let $\rho'_n(A, B)$ denote the correlation coefficient between these two random variables, which is given by:

$$\rho'_n(A,B) = \frac{p'_n(A,B) - p'_n(A)p'_n(B)}{\sqrt{p'_n(A)(1-p'_n(A))p'_n(B)(1-p'_n(B))}}.$$

To show that A and B are positively correlated is equivalent to showing that:

$$p'_n(A,B) \geq p'_n(A)p'_n(B). \tag{5.14}$$

Consider the five cases listed in Theorem 11. For Case 1, it is easy to show that clades A and B are positively correlated, as $p'_n(A,B) = p'_n(A) = p'_n(B)$, and $0 < p'_n(A) \leq 1$. Therefore, Equation (5.14) holds.

For Case 2 to 4, it suffices to show that:

$$\frac{\varphi(a)\varphi(b-a+1)}{\varphi(b)}\frac{\varphi(b)\varphi(n-b+1)}{\varphi(n)} \geq \frac{\varphi(b)\varphi(n-b+1)}{\varphi(n)}\frac{\varphi(a)\varphi(n-a+1)}{\varphi(n)} \text{ for Cases 2 and 3;}$$

and:

$$\frac{\varphi(a)\varphi(b)\varphi(n-a-b+2)}{\varphi(n)} \geq \frac{\varphi(b)\varphi(n-b+1)}{\varphi(n)}\frac{\varphi(a)\varphi(n-a+1)}{\varphi(n)} \text{ for Case 4.}$$

These are equivalent to $\varphi(n)\varphi(b-a+1) \geq \varphi(b)\varphi(n-a+1)$ for Case 2 and 3, and $\varphi(n-a-b+2) \geq \varphi(n-b+1)\varphi(n-a+1)$ for Case 4, which certainly hold by Lemma 16.

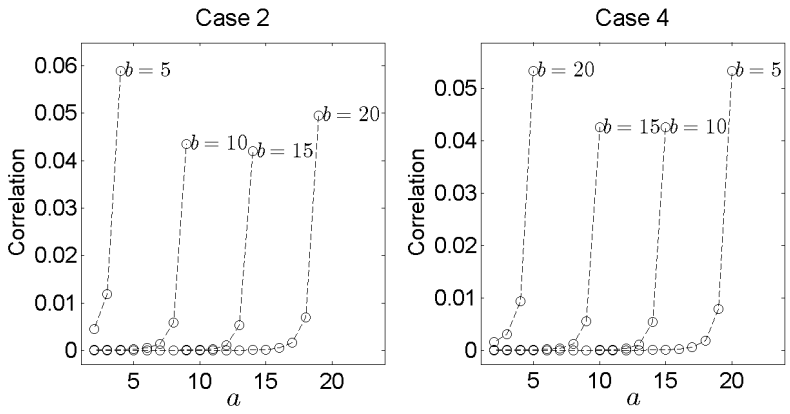


Figure 5.3: Correlation of compatible clades A and B $\rho'_n(A, B)$ for $n = 25$, in Cases 2 and 4 under the PDA model, with $a = |A|$ and $b = |B|$.

□

5.4 Extension to unrooted trees

If we suppress the root of a rooted binary phylogenetic tree T, we obtain an unrooted phylogenetic tree, which we will denote as $T^{-\rho}$. Let $Y(n)$ be the set of unrooted trees with the leaf set X, $|X| = n$. Then the distribution on $T(n)$ induces the distribution on $Y(n)$ as:

$$\mathbb{P}(\mathcal{Y} = Y) := \sum_{\substack{T \in T(n) \\ s.t. T^{-\rho} = Y}} \mathbb{P}(\mathcal{T} = T).$$

In this section, we are interested in the distribution of the clade probabilities for an unrooted tree that has been induced by the rooted tree under the PDA model. Note that for a finite set with cardinality n, there are exactly $\varphi(n-1)$ unrooted trees in $Y(n)$.

We say that a subset A of X is a *clan* of an unrooted tree $Y \in Y(n)$ if and only if $A|(X-A)$ is a split of Y, which means that removing one (necessarily unique) edge in Y decomposes Y into two rooted subtrees on A and $X - A$.

5.4.1 Clan probability under the PDA model (I)

Given a subset A of X, let $q'_n(A)$ be the probability that A is a clan of a random unrooted PDA tree. By the EP property, this depends only on $a = |A|$ and $n = |X|$, so we will

write it as $q'_n(a)$.

Recall Lemma 14:

> *Given a rooted binary X-tree, T, a set A is a clan of $T^{-\rho}$ if and only if either A is a clade of T or $X - A$ is a clade of T.*

This suggests the following lemma.

Lemma 17. *For a proper subset A of X, we have:*

$$q'_n(A) = p'_n(A) + p'_n(X - A) - \hat{p}'_n(A, X - A).$$

As a corollary, we have the following results on $p'_n(a)$ and $q'_n(a)$.

Theorem 13. *For $1 \leq a < n$, we have:*

$$q'_n(a) = p'_{n-1}(a). \tag{5.15}$$

Proof. Since we have:

$$\begin{aligned} q'_n(a) &= \frac{\varphi(a)\varphi(n-a+1) + \varphi(n-a)\varphi(a+1) - \varphi(a)\varphi(n-a)}{\varphi(n)} \\ &= \frac{\varphi(a)\varphi(n-a)\left[2(n-a+1) - 3 + 2(a+1) - 3 - 1\right]}{\varphi(n)}, \end{aligned}$$

we have $q'_n(a) = \dfrac{\varphi(a)\varphi(n-a)}{\varphi(n-1)} = p'_{n-1}(a)$ as required. □

Thus by combining the results of Theorem 13 and Corollary 4, one can show that that $q'_n(a)$ is 'unimodal'

5.4.2 Clan probability under the PDA model (II)

Theorem 14. *Let $A_1, \ldots, A_k$ be k disjoint (non-empty) subsets of X, with $a_i = |A_i|$, $|X| = n$ and $m = \sum_{i=1}^{k} a_i$. Then the probability that a random unrooted PDA tree $\mathcal{Y}$ has $A_1, \ldots, A_k$ as clans is $q'_{(m,n)}(A_1, \ldots, A_k)$. By the EP property, we have:*

$$q'_{(m,n)}(a_1, \ldots, a_k) = \frac{\varphi(n-m+k-1)\prod_{i=1}^{k}\varphi(a_i)}{\varphi(n-1)}.$$

Proof. Since each tree in $\mathcal{Y}$ has a probability of $1/\varphi(n-1)$, it remains to compute the number of trees that have $A_1, \ldots, A_k$ as clans. Such a tree can be constructed in two steps:

1. Build an unrooted tree on $\left(X \setminus \bigcup_{i=1}^{k} A_i\right) \cup \{x_1, \ldots, x_k\}$, where $x_1, \ldots, x_k$ are not leaves in X but serve as "placeholders" in the second step.

2. Replace each x_i with a tree in $\mathcal{T}_{A_i}$.

There are $\varphi(n-m+k-1)$ different choices of tree in the first step, and there are $\prod_{i=1}^{k}\varphi(a_i)$ different ways to replace $x_1,\ldots,x_k$ with trees in $\mathcal{T}_{A_1},\ldots,\mathcal{T}_{A_k}$. The claim then follows. □

Theorem 15. *Let A and B be two compatible subsets of X. We use $q'_n(A,B)$ to denote the probability that A and B are clans. From the EP property, we have:*

$$q'_n(A,B)=\begin{cases} p'_n(a) & \text{if } A=B \quad \text{[Case 1]}\ ;\\ \dfrac{\varphi(b)\varphi(n-b)\varphi(a)\varphi(b-a)\varphi(n-b)}{\varphi(n-1)\varphi(b-1)}, & \text{if } A\subsetneq B \quad \text{[Case 2]}\ ;\\ \dfrac{\varphi(a)\varphi(n-a)\varphi(b)\varphi(a-b)\varphi(n-a)}{\varphi(n-1)\varphi(a-1)}, & \text{if } B\subsetneq A \quad \text{[Case 3]}\ ;\\ \dfrac{\varphi(a)\varphi(b)\varphi(n-a-b+1)}{\varphi(n-1)}, & \text{if } A\cap B=\emptyset, A\cup B\subseteq X_n \ \text{[Case 4]}\ ;\\ 0, & \text{otherwise} \quad \text{[Case 5]}\ ; \end{cases}$$

Cases 2 and 3 follow by applying Theorem 13 and then Theorem 9 (i) twice. The second claim follows from Theorem 14.

5.4.3 Correlation results under the PDA model (II)

Let A and B be two compatible subsets of X, i.e. $A,B\subsetneq X$, with $|A|,|B|\geq 2$.

Theorem 16. *The events 'A is a clan' and 'B is a clan' are positively correlated under the PDA process.*

Proof. Similar to the proof of Theorem 12, which shows that two clans are positively correlated, it is sufficient to show that: $q'_n(A,B)\geq q'_n(A)q'_n(B)$. Recalling the first four cases of Theorem 15, here, we discuss the cases where one set is contained in the other or that the two sets are disjoint.

By symmetry, we can assume that $A\subseteq B$. To show that:

$$\frac{\varphi(b)\varphi(n-b)\varphi(a)\varphi(b-a)\varphi(n-b)}{\varphi(n-1)\varphi(b-1)}\geq\frac{\varphi(b)\varphi(n-b)}{\varphi(n-1)}\frac{\varphi(a)\varphi(n-a)}{\varphi(n-1)},$$

it is sufficient to show that $\varphi(n-1)\varphi(b-a)\geq\varphi(b-1)\varphi(n-a)$, which certainly holds by Lemma 16.

If $A\cap B=\emptyset$, we show that:

$$\frac{\varphi(a)\varphi(b)\varphi(n-a-b+1)}{\varphi(n-1)}\geq\frac{\varphi(b)\varphi(n-b)}{\varphi(n-1)}\frac{\varphi(a)\varphi(n-a)}{\varphi(n-1)},$$

which follows from Lemma 16.

□

5.5 Further discussion

5.5.1 Size of a randomly selected internal node

One particular function in `apTreeshape` attracted our attention, namely `cladesize`, which randomly samples one of the interior nodes of a Yule tree or a PDA tree, then returns the number of descendants of the sampled interior node. Let K be a discrete random variable denoting the size of a clade in a phylogenetic tree with n leaves. The manual of `apTreeshape` (Bortolussi *et al.*, 2009) suggest that according to the YHK model, the probability that the number of descendants of an internal node of the tree equals a is:

$$\mathbb{P}_{\mathrm{YHK}}(K = a) = \frac{2n}{(n-1)a(a+1)}, \tag{5.16}$$

for $a = 2, 3, ..., n-1$, and $\mathbb{P}_{\mathrm{YHK}}(K = a) = \dfrac{1}{(n-1)}$ for $a = n$. This can easily be shown by using the result from Lemma 10. Since there are $n-1$ interior nodes in a rooted binary tree, $a = n$ is equivalent to the trivial case of selecting the root, which has the probability $1/(n-1)$. For $a = 2, 3, ..., n-1$, let $X_n(a)$ be the number of interior nodes for which the number of descendants is a in an n-taxon tree. The expected value of $X_n(a)$ under the YHK model is given by Equation (4.11) as:

$$\mathbb{E}_{\mathrm{YHK}}[X_n(a)] = \frac{2n}{a(a+1)}, \quad 2 \leq a \leq n-1,$$

which is also equal to $(n-1) \times \mathbb{P}_{\mathrm{YHK}}(K = a)$. We can now rearrange the equation below:

$$\mathbb{E}_{\mathrm{YHK}}[X_n(a)] = (n-1) \cdot \mathbb{P}_{\mathrm{YHK}}(K = a), \tag{5.17}$$

Equation (5.16) then follows.

The manual of `apTreeshape` also suggests another formula that calculates the probability that the number of descendants of a random internal node of a PDA tree is equal to a, when the number of leaves tends to infinity. Here, we show the exact probability for any size of leaf set by using a similar approach to that used for deriving Equation (5.16): by dividing Equation (5.10) by $1/(n-1)$, we have the probability that the number of descendants of a random internal node of a PDA tree is equal to a, because:

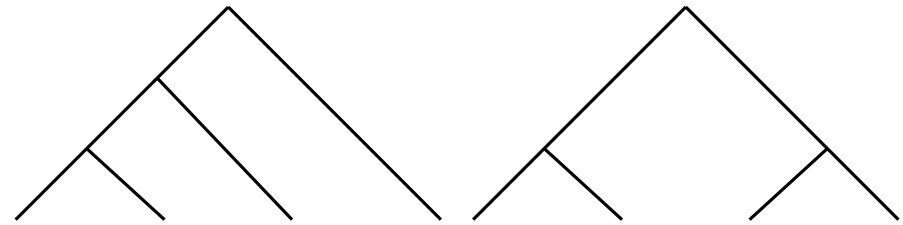
$$\mathbb{P}_{\mathrm{PDA}}(K = a) = \frac{1}{(n-1)} \binom{n}{a} \binom{n-1}{a-1} \binom{2n-2}{2a-2}^{-1}. \tag{5.18}$$

5.5.2 Expected value of the Sackin index

The Sackin index (Sackin, 1972) is used for measuring the balance of a tree. It is defined as the sum of the number of interior nodes (include the root node) d_i on the path of every leaf node i to the root:

$$S_n = \sum_{i=1}^{n} d_i.$$

For two binary trees with the same leaf set, the one which has the smaller Sackin index is more balanced than the other. For example, for unlabelled four-taxon trees, there are only two possible topologies: a completely imbalanced tree (Figure 5.4a) and a completely balanced tree (Figure 5.4b). The Sackin indices of the completely imbalanced tree and the completely balanced tree are 9 and 8 respectively.



(a) Four-taxon completely imbalanced tree (b) Four-taxon completely balanced tree

Figure 5.4: An imbalanced four-taxon tree and a balanced four-taxon tree.

Suppose that A_i is a clade of an X-tree T. Note that, the Sackin index can be also defined as the sum of the non-trivial clade (non-leaf) sizes at all interior nodes of the tree:

$$S_n = \sum_{1<|A_i|\leq n} |A_i|, \tag{5.19}$$

where $A_i \in \mathcal{C}(T)$.

We now use this definition to show an alternative derivation for the expected Sackin index under the YHK model (Blum *et al.*, 2006; Kirkpatrick and Slatkin, 1993)

$$\mathbb{E}_{\text{YHK}}(S_n) = 2n \sum_{j=2}^{n} \frac{1}{j}.$$

From Equation (5.19), we take the expected value of the Sackin index and then we can obtain:

$$\mathbb{E}(S_n) = \mathbb{E}\left(\sum_{1<|A_i|\leq n} |A_i|\right) = \sum_{i=1}^{n-1} \mathbb{E}(|A_i|) = (n-1)\mathbb{E}(|A_i|).$$

Let K be a discrete random variable denoting the size of a clade in a phylogenetic tree with n leaves. We then have $\mathbb{E}(|A_i|) = \sum_{a=2}^{n} a\mathbb{P}(K = a)$. Similar to Equation (5.17), we have $\mathbb{P}(K = a) = \frac{1}{n-1}\mathbb{E}(X_n(a))$. Therefore:

$$\mathbb{E}(S_n) = (n-1)\sum_{a=2}^{n} a\mathbb{P}(K = a) = (n-1)\sum_{a=2}^{n} \frac{a}{n-1}\mathbb{E}(X_n(a)) = \sum_{a=2}^{n} a \cdot \mathbb{E}(X_n(a)). \tag{5.20}$$

Note that for YHK trees, $a \cdot \mathbb{E}(X_n(a)) = n$ if $a = n$. Therefore as required:

$$\begin{aligned}\mathbb{E}_{\mathrm{YHK}}(S_n) &= \sum_{a=2}^{n-1} \frac{2n}{a+1} + n = 2n \sum_{a=1}^{n-1} \frac{1}{a+1} \\ &= 2n \sum_{j=2}^{n} \frac{1}{j} \ .\end{aligned}$$

Similarly, by combining Equations (5.10) and (5.20), we show that the expected value of the Sackin index under the PDA model is:

$$\mathbb{E}_{\mathrm{PDA}}(S_n) = \sum_{a=2}^{n} a \cdot \binom{n}{a} \binom{n-1}{a-1} \binom{2n-2}{2a-2}^{-1}.$$

Chapter 6

Probabilities of gene trees of a given species network

Abstract

This chapter introduces a novel method to calculate gene tree probabilities of a given species network, which works by conditioning on coalescent events occurring below hybridization nodes to remove hybridization events recursively from the network. This results in gene tree probabilities being linear combinations of the gene tree probabilities in given species trees, weighted by the probabilities of events below hybridization nodes.

6.1 Introduction

We consider more generic species networks than those of Meng and Kubatko (2009), Kubatko (2009) and Yu *et al.* (2011) (for example, see Figure 6.1), by allowing speciation events to occur before or after hybridization events, and are flexible with the number of hybridization events and the number of sampled lineages per species.

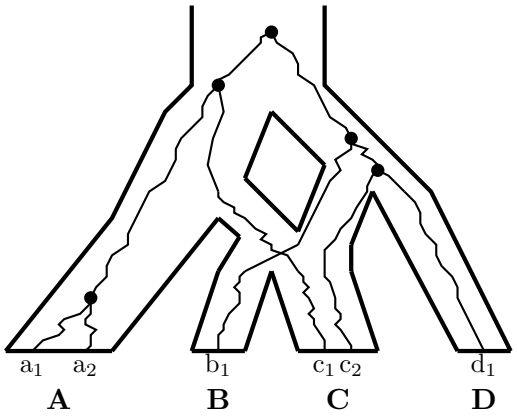


Figure 6.1: Example of a binary gene tree with the topology $(((a_1,a_2),c_1),(b_1,c_2,d_1))$ in a network with the topology `((((B,C)s1)h1#H1,A)s2,(h1#H1,D)s3)r`.

Yu *et al.* (2011) computed the gene tree probabilities in a specific example: A population was separated into two groups in the past. Each of these subgroups was then separated again, but two of these sub groups were merged instantly, and this was followed again by a separation. This network allows speciation to occur after hybridization. Yu *et al.* (2012) express a species network as its equivalent multi-labelled species tree, and the gene tree probabilities are then computed using a modification of the method in Degnan and Salter (2005).

6.2 Methodology

In order to correctly refer to the branches of a species network, we first introduce a systematic way of labelling all the internal branches and nodes, namely the *modified post-order traversal method*, followed by two operations of simplifying a complex rooted network to simpler structures, and how to operate the simplification procedure recursively. As a result, we compute gene tree probabilities of a given species network by taking the sum of gene probabilities of given species trees.

6.2.1 Modified post-order traversal method

The post-order tree traversal method described by Felsenstein (2004) and Valiente (2002) enumerates all the nodes of a rooted binary tree recursively. This method starts from the bottom of the tree, and labels each node after visiting its descendants. Because of the unique mapping between a node and its parental node, this method ensures that all branches are uniquely labelled from bottom of the tree to the root in ascending order.

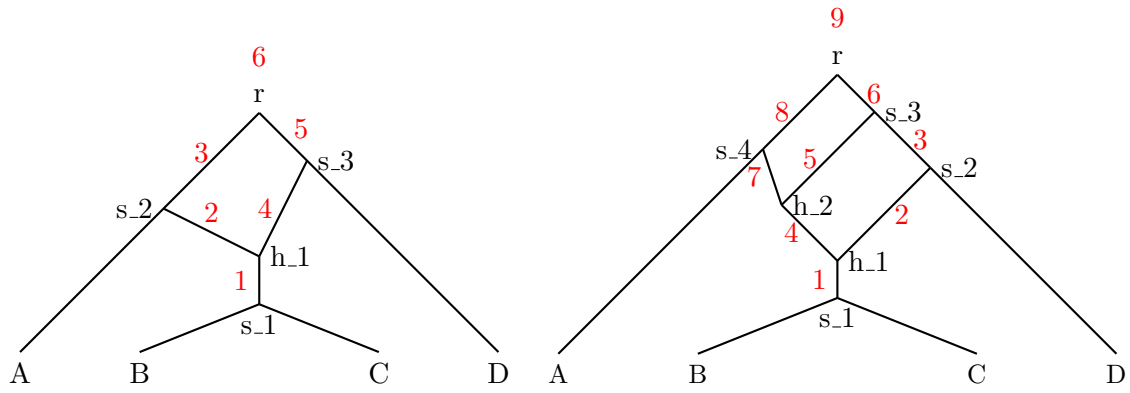
The modified post-order traversal generalises to a rooted network W, which can be treated as a directed graph $G = (V, E)$ and labels the intererior branches of G. This method labels the internal branches of a tree in a manner consistent with the post-order tree traversal method.

The new method guarantees that all branches are visited twice and are labelled while exiting. The traversal process starts from the root, and travels to every child node. Unless the child node is a tip node or has been labelled, the traversal process continues towards to the bottom of the network. A branch is labelled if either the bottom of the network is reached or all the descendant nodes have been visited.

This algorithm is feasible for all rooted networks (see Figures 6.2a and 6.2b) or trees. It labels all interior edges of a directed graph $G = (V, E)$ as follows.

For $v \in V$, let $C(v)$ be the set of child nodes of node v. We use $o_1(v)$ and $o_2(v)$ to denote the label on the branch from v to its parents. $o_2(v)$ is defined only if v is a hybrid node. Suppose that $\mathcal{I}(v)$ is a indicator function that shows whether all the branches under v have been labelled. During the traversal process,

$$\mathcal{I}(v) = \begin{cases} \texttt{FALSE} & \text{all branches below } v \text{ are labelled,} \\ \texttt{TRUE} & \text{otherwise.} \end{cases}.$$



(a) Level-1 network (W^1_{IV}) (b) Level-2 network (W^2_{IV})

Figure 6.2: Examples of using post-order traversal to label the branches of a level-1 and a level-2 network. All internal branches are enumerated and marked in red. Note that the branch lengths of the networks shown in (a) and (b) are not to scale.

All branches are labelled by the function $f(v, o_{in})$ for all $v \in V$. We use o_{in} to denote the number that labelled the previous branch. The value returned by the function $f(v, o_{in})$ is the current label while exiting node v. To ensure that all branches are visited twice, the traversal process starts from the root ρ, and calls $f(\rho, 0)$. Initially, for all $v \in V$, let $\mathcal{I}(v)$ be `FALSE`.

Algorithm 6.1 Recursive algorithm for modified post-order traversal.

if $|C(v)|$ is 0 **then**
  return o_{in}
else
  if $\mathcal{I}(v)$ is `TRUE` **then**
    if $o_2(v)$ is nonempty **then**
      return $o_2(v) \leftarrow o_{in} + 1$
    end if
  else
    for $v' \in C(v)$ **do**
      $o_{in} \leftarrow f(v', o_{in})$
    end for
    $\mathcal{I}(v) \leftarrow$ `TRUE`
    return $o_1(v) \leftarrow o_{in} + 1$
  end if
end if

6.2.2 *Decomposition operations*

In this section, we propose two operations to simplify a complex phylogeny structure into simpler structures with fewer hybridization events (Zhu *et al.*, 2011b). Then the coalescent model of (Degnan and Salter, 2005) is extended to obtain the distribution of gene trees

in a given network. To demonstrate this procedure, we first consider simple cases of that one individual is sampled from each population at the present. We first make several restrictions and assumptions for the gene tree T and the network W in this section:

- The gene tree T and the network W are rooted.
- Gene tree T and network W have the same number of external edges (not necessary for section 6.2.5).
- Gene tree T is binary.
- An interior node of W can only have at most two parent nodes; a hybrid node refers to an internal node of W which has two parent nodes.
- We do not consider the case that a hybrid node is also a leaf node (not necessary for section 6.2.5).

The network W is initially reduced to a set of simpler networks ($SG(W)$) in a single step in the reduction process. Let $P(T|W)$ be the probability of gene tree T given a species network W, by the *Law of total probability*, we have the following:

$$\begin{aligned} P(T|W) &= \sum_{w^* \in SG(W)} P(T|W^* = w^*, W) P(W^* = w^*|W) \\ &= \sum_{w^* \in SG(W)} P(T|W^* = w^*) P(W^* = w^*|W), \end{aligned} \tag{6.1}$$

where W^* is a random variable that depends on W.

For any $w^* \in SG(W)$, w* implies either a particular parental branch that some lineages have followed at a hybrid node or some specific coalescences that have occurred beneath a hybridization node.

Note that, prior to decomposing a network, we need to rank each node from the bottom of the network to the top: Tip nodes have rank one; an interior node's rank is one plus the highest rank of its child nodes. This ensures that we perform the simplification operations in a correct order — always operate on the node with lower rank first, since the coalescent process starts from the bottom of a network, then gradually move towards to the root node.

The key of simplifying a network is to remove the interior nodes of the network in a specific order, along with the branches that are connected to the node. Here we define several functions to assist us identifying which nodes should firstly be removed. Suppose $G = (V, E)$ is a directed graph. We index all the nodes in V; for $v \in V$, let $r(v)$ be the rank of v, and $p(v)$ be the number parent node of v. We use indicator function $h(v)$ to identify if a node v is a hybrid node:

$$h(v) = \begin{cases} 1, & \text{if } p(v) = 2; \\ 0, & \text{otherwise.} \end{cases}$$

Let $hd(v)$ and $t(v)$ be the indicator function that take values

$$hd(v) = \begin{cases} 1, & \text{if } v \text{ is a descendant node of a hybrid node;} \\ 0 & \text{otherwise;} \end{cases}$$

and

$$t(v) = \begin{cases} 1, & \text{if } v \text{ is a leaf node;} \\ 0 & \text{otherwise,} \end{cases}$$

respectively.

Algorithm 6.2 Algorithm to choose the *index* of the node to be removed in order to simplify a network.

```
1: index = |V| − 1; I = 1;
2: for I < |V| do
3:    if (h(v_I) + hd(v_I)) ∗ (1 − t(v_I)) ≥ 1 and r(v_I) < r(v_index) then
4:       index = I;
5:    end if
6:    I = I + 1;
7: end for
8: if I = |V| − 1 then
9:    return index = −1
10: else
11:    return index
12: end if
```

Thus, we can apply Algorithm 6.2 to find which node should be removed from the network: If the algorithm returns value -1, it means that W is already tree-like, and do not need to be simplified; otherwise, it returns the index of the node that we need to perform the following operations.

Decomposition operation 1

If the chosen node is an interior descendant node s of a hybrid node, then this implies that s has a single parent node (otherwise s is a hybrid node), and child nodes of s are the leaf nodes of W (since s has the lowest rank beside the tips). The first step of operation 1 is to remove s from W, along with all the edges that are connected to s.

Let set D denote all the leaf nodes descendant from s. We now enumerate all possible ways to partition D. For example, if $D = \{\alpha_1, \alpha_2, \alpha_3\}$, let D' be one of the possible partitions of D. D' could be $\{\{\alpha_1\}, \{\alpha_2\}, \{\alpha_3\}\}$, $\{\{\alpha_1, \alpha_2\}, \{\alpha_3\}\}$, $\{\{\alpha_1\}, \{\alpha_2, \alpha_3\}\}$, $\{\{\alpha_1, \alpha_3\}, \{\alpha_2\}\}$ or $\{\{\alpha_1, \alpha_2, \alpha_3\}\}$. We treat every element of any D' as a new leaf node. In the second part of operation 1, we create a new graph w^*, by connecting the elements of D' to the parent node of s. Notice, if the element of D' contains more than 1 leaf node, this implies that by changing from graph W to w^*, we need to coalesce these leaves on

the branch that connects s and its parent node.

To calculate the probability of these events, we let $u = |D|$, and $v = |D'|$ and t be the branch length from s to its parent node. Then the probability of u lineages coalesce into v lineages within time t is (Degnan and Salter, 2005; Rosenberg, 2002; Saunders *et al.*, 1984; Tajima, 1983; Takahata and Nei, 1985):

$$p_{uv}(t) = \sum_{k=v}^{u} e^{-k(k-1)t/2} \frac{(2k-1)(-1)^{k-v}}{v!(k-v)!(v+k-1)} \times \prod_{y=0}^{k-1} \frac{(v+y)(u-y)}{(u+y)}. \tag{6.2}$$

Therefore, we have:

$$P(W^* = w^*|W) = \frac{w}{c} p_{uv}(t) \mathcal{I}_{w^*}(T), \text{ for } w^* \in SG(W), \tag{6.3}$$

where c is the number of ways for u lineages coalesce into v lineages (see Equation (4.3)), which is equal to $\prod_{i=v}^{u} \binom{i}{2}$, and w is the number of repeated topologies with u lineages coalescing into v lineages (see Equation (4.4)). This is equal to $w = (u-v)! \prod_{j=1}^{u-v} \frac{1}{1+a_j}$, where a_j is the number of interior nodes that are below the coalesced nodes, and

$$\mathcal{I}_{w^*}(T) = \begin{cases} 1, & \text{if the lineages in } w^* \text{ can lead to tree topology } T; \\ 0, & \text{otherwise.} \end{cases}$$

For instance, if the gene tree is $(((a_1, d_1), c_1), b_1)$ and $w^* = W_1$ in Figure 6.3, then $\mathcal{I}_{w^*}(T) = 0$.

Remark 1. Operation 1 removes an internal node of network W. Thefore, any reduced network w^*, $w^* \in SG(W)$, has one less interior node than network W.

Decomposition operation 2

Before applying operation 2 on a hybrid node h of W, we need to make sure that operation 1 have been applied to all the interior descendant from h. This implies that all the child nodes of h are the leaf nodes of W. Let p_L and p_R be the two parent nodes of h. We use H to denote the set of child nodes of h and $\mathbb{C}_H$ to denote the collection of all the subsets of H. The first step of operation 2, is to remove h from W, and all the edges connected to h.

We then introduce two new nodes, h_L and h_R. For any $L \in \mathbb{C}_H$, we have a new graph w^*, connect $l \in L$ to h_L, then connect h_L to p_L, and connect $r \in H \setminus L$ to h_R, then connect h_R to p_R. Let $m_L = |L|$, $m_R = |H \setminus L|$, and $m = |H|$. The parameter γ is the probability that one lineage is attached to p_L. Thus, we obtain the set of simpler networks $SG(W)$ and the probabilities $P(W^* = w^*|W)$ for any $w^* \in SG(W)$:

$$P(W^* = w^*|W) = \gamma^{m_L}(1-\gamma)^{m_R}, \qquad \text{where } m_L + m_R = m. \tag{6.4}$$

Remark 2. Operation 2 removes an internal node of network W. The newly added nodes h_L and h_R are effectively external nodes: as all the nodes below h_L and h_R are leaf nodes, we can treat h_L and h_R as leaf nodes, but sampling multiple lineages from each of them (details see Section 6.2.5). Thefore, any reduced network w^*, $w^* \in SG(W)$, has one less interior node than network W.

6.2.3 *Simplifying a network recursively*

Apply operations 1 and 2 recursively on any networks in $SG(W)$ until all the simplified network structures are tree-like. The problem of obtaining gene tree probabilities from species trees has already been solved (Degnan and Salter, 2005). The approach outlined in this chapter will therefore reduce the probability of a gene tree, given a species network, to a linear combination of gene tree probabilities of given species trees.

Let $AG_T(W)$ be an ordered list of directed graphs (trees or networks), and $|AG_T(W)|$ is the number of elements in the list. Here we borrow the concepts of set operations "$\bigcup$" and "\" for our use. Let $AG_T(W) \bigcup SG(W)$ denote gradually appending the elements of $SG(W)$ to the end of the list $AG_T(W)$, then indexing the new elements of $AG_T(W)$ from $|AG_T(W)|+1$ to $|AG_T(W)|+|SG(W)|$. For an element G of $|AG_T(W)|$, we define operation $AG_T(W) \setminus \{G\}$, as removing the element G from $AG_T(W)$, the index of any element behind G is now one less. Then we apply Algorithm 6.3 to simplify a network W, and then compute the probability for gene tree T.

Algorithm 6.3 Recursive algorithm for simplify a network.

Initialize $AG_T(W) = \{W\}$ and $I = 1$;
while $I \leq |AG_T(W)|$ **do**
  Apply Algorithm 6.2 to G_I, $G_I \in AG_T(W)$ to choose the *index* of the node needs to be removed;
  if *index* is positive **then**
    if $p(v_{index})$ is 1 **then**
      Perform decomposition operation 1 on v_{index}, obtain $SG(G_I)$.
    else
      Perform decomposition operation 2 on v_{index}, obtain $SG(G_I)$.
    end if
    $AG_T(W) \leftarrow AG_T(W) \setminus \{G_I\}$.
    $AG_T(W) \leftarrow AG_T(W) \bigcup SG(G_I)$.
  else
    $I = I + 1$;
  end if
end while

During the decomposition process, different sequences of removing the hybrid nodes may lead to the same sub-species trees W'. For $W' \in AG_T(W)$, we use $C(W, W')$ to denote the collection of ways to decompose W into W'. Each sequence of decomposition

corresponds to a unique weight ω_c. Thus by simplifying Equation (6.1), we have:

$$P(T|W) = \sum_{W' \in AG_T(W)} P(T|W^* = W', W) \sum_{c \in C(W,W')} \omega_c. \tag{6.5}$$

Figure 6.3 demonstrates the decomposition of a species network on four taxa with two hybridization nodes (Figure 6.2b). Notice that even though W_1 and W_{14} have the same topology, the branch lengths of these two trees differ. We consider them to be different species trees. For different gene trees, according to coalescent events, $AG_T(W^2_{\text{IV}})$ may differ.
For example, if the gene tree is $(((a_1,d_1),c_1), b_1)$, $AG_T(W^2_{\text{IV}}) = \{W'_4, W'_5, \ldots, W'_{12}\}$, but when the gene tree is $(((a_1,b_1),c_1),d_1)$, $AG_T(W^2_{\text{IV}}) = \{W'_1, W'_2, \ldots, W'_{12}\}$.

Once a network has been decomposed into species trees, it is straight forward to compute the probability of the gene tree given each species tree using the coalescent history approach of Degnan and Salter (2005) and Degnan (2010) (if there is more than one allele sampled per species), or using a recursive approach with ancestral configurations (Wu, 2012).

Ancestral configurations list the lineages occurring in each ancestral population, and this approach enumerates the possible ancestral configurations consistent with the number of lineages in each ancestral population. The recursion is done by computing the probability of an ancestral configuration conditional on the configuration in more ancestral populations.

6.2.4 Example

This section demonstrates an example of calculating gene tree probability of (((a,b),c),d) in W^1_{IV} (details see Table 6.1).Since lineages of b and c can not coalesce in branch 1 of W^1_{IV}, they both have to travel through population 1. At node h_1 of W^1_{IV}, there are four possibilities between b and c:

Scenario (1) lineages b and c both enter branch 2 of W^1_{IV}; no coalescent occurs in branch 2; then b and c both enter branch 3.

Scenario (2) lineage b enter branch 2, then enter branch 3; lineage c enter branch 4, then enter branch 5.

Scenario (3) lineage c enter branch 2, then enter branch 3; lineage b enter branch 4, then enter branch 5.

Scenario (4) lineages b and c both enter branch 4 of W^1_{IV}; no coalescent occurs in branch 4; then b and c both enter branch 5.

Let λ_1 be the branch length of branch 1. For all four scenarios, there is no coalescent event between lineages b and c within branch 1. Therefore, we calculate the probability of

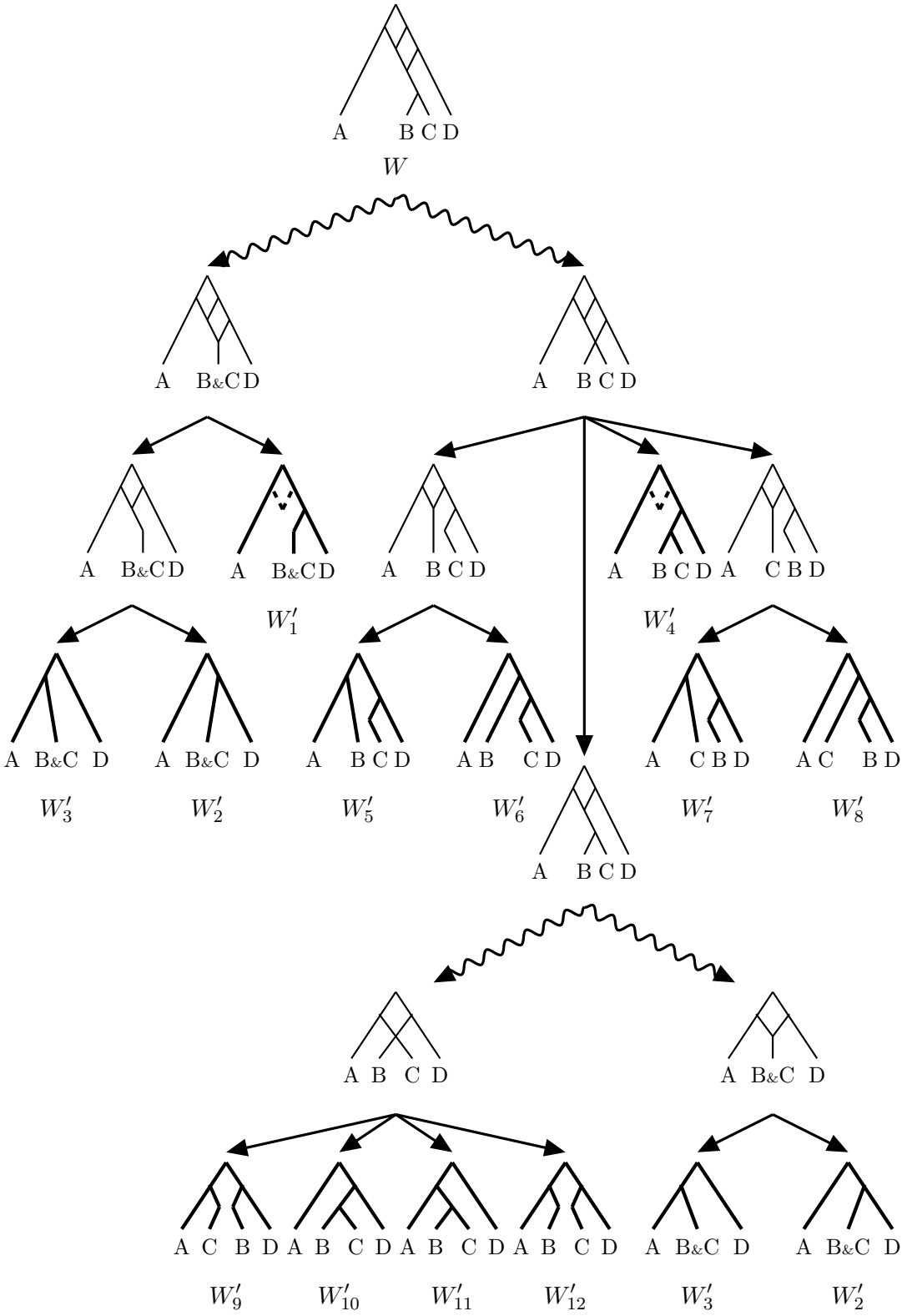


Figure 6.3: Illastration of decomposing the network in Figure 6.2b. Snake shape and straight line arrows represent the decomposition operation 1 and 2 respectively. The trees with thicker branches are the final trees after decomposing the network.

2 lineages entering and exiting a population with time λ_1, which is $p_{22}(\lambda_1)$. Suppose that γ is the probability of a lineage travel to the left at the hybrid node (h_1), then enter branch 2. Therefore, we consider lineages b and c independently, and calculate the probability for each scenario as shown by Table 6.1.

Scenario		(1)	(2)	(3)	(4)
Probability of scenario $(*)$		$\gamma^2 p_{22}(\lambda_1)$	$\gamma(1-\gamma)p_{22}(\lambda_1)$	$\gamma(1-\gamma)p_{22}(\lambda_1)$	$(1-\gamma)^2 p_{22}(\lambda_1)$
(((a,b),c),d)	$(3,3,6)$	$p_{22}(\lambda_2)\frac{1}{3}p_{31}(\lambda_3)$	–	–	–
	$(3,6,6)$	$p_{22}(\lambda_2)\frac{1}{3}p_{32}(\lambda_3)\frac{1}{3}$	$p_{21}(\lambda_3)p_{22}(\lambda_5)\frac{1}{3}$	–	–
	$(6,6,6)$	$p_{22}(\lambda_2)p_{33}(\lambda_3)\frac{1}{18}$	$p_{22}(\lambda_3)p_{22}(\lambda_5)\frac{1}{18}$	$p_{22}(\lambda_3)p_{22}(\lambda_5)\frac{1}{18}$	$p_{22}(\lambda_4)p_{33}(\lambda_5)\frac{1}{18}$

Table 6.1: Gene tree probability of (((a,b),c),d) given the species network W^1_{IV} shown in Figure 6.2. The entries in each column are gene tree probabilities conditional on each scenario. Each row is the probability of a specific sequence which indicates the branch that the internal nodes of tree (((a,b),c),d) coalesce in. Thus, the gene tree probability is equal to the sum of the weighted (by the probabilities of the scenarios) column sums (conditional probabilities under each scenario).

The gene tree probabilities under scenarios 1, 2, 3 and 4 are consistent with computing the gene tree probability of (((a,b),c),d) in given species trees (((B,C):2,A):3,D):6, ((A,B):3,(C,D):5):6, ((A,C):3,(B,D):5):6 and (((B,C):4,D):5,A):6, where the numbers after colons denote the branch labels. These trees can be obtained by performing operation 1 on s_1, and followed by performing operation 2 on h_1. We can then calculate gene tree probability by conditioning on species trees first, then weight these probabilities according to the scenario probabilities.

Notice that, in Table 6.1, the row (3,3,6) has one probability entry, but rows (3,6,6) and (6,6,6) have more than one. Degnan and Salter (2005) refers the sequences (3,3,6), (3,6,6) and (6,6,6) as the coalescent histories, which indicate the species tree branches, of which the internal nodes of gene tree coalesce in, in a specific order. If coalescent histories are defined as the list of populations in which nodes of the gene tree occur, as in Degnan and Salter (2005), then Table 6.1 suggests determining the probability of a coalescent history by summing over probabilities of lineages taking different paths through the network at hybridization nodes. The approach taken in the recursion, however, does not directly enumerate coalescent histories on networks, but rather first reduces the network to trees, and then enumerates coalescent histories on each of these trees.

6.2.5 Multiple lineages sampled from each population in the present

In the previous section, we have discussed how to compute gene tree probabilities for a given network when only one lineage is sampled per species. In which case, coalescent events would not happen in the external branches of the phylogeny structures. However, one would sample more than one individual from each population in practice. Consider the example shown in Figure 6.4(a): two samples have been take from each population A, B and C. Lineages in populations A and C do not travel through the hybrid node.

Therefore, the coalescent events in populations A and C can be treated as coalescent events in internal branches. However, for the two lineages in the population B, we need to consider the similar four scenarios mentioned in the previous example given that the two lineages in the hybrid species do not coalesce, in additional to two cases that after the two lineages in B have coalesced, the new lineage goes to the left or the right.

In this section, we introduce a technique to manipulate the species network in order to compute the gene tree probabilities when multiple individuals are sampled from a species. Suppose that L is the label of a descendant leaf node from a hybrid node in a directed graph $G = (V, E)$. When n individuals are sampled from L, we attach n leaf nodes to graph G below the node "L", namely from "L_1" to "L_n". Then the L node becomes a internal node of G. This problem then becomes as to compute gene tree probabilities when one individual is sampled from each (hybrid) species. Since these external branches do not actually exist, we assign them length 0. For instance, we can rewrite the species network (Figure 6.4(a)) string `(((B:1)h_1#.5:1, A:2)s_1:1, (h_1#.5:1, C:2)s_2:1)r` as `((((B1:0, B2:0)B:1)h_1#.5:1, A:2)s_1:1, (h_1#.5:1, C:2)s_2:1)r`, which is illustrated by Figure 6.4(b), and then apply Algorithm 6.3 to simplify the network.

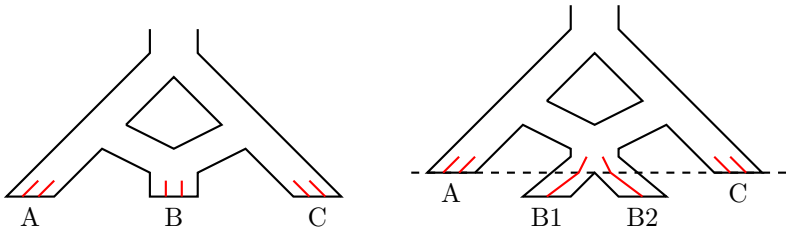


Figure 6.4: Illustration of manipulating the species network in order to compute the gene tree probabilities of gene trees for a given species network when more than one lineages are sampled from the descendant species of a hybrid species. (a) A three-taxon species network with one hybrid node; B is the hybrid species; and two individuals are sampled from each species. (b) Adding nodes "$B1$" and "$B2$" below node B. The three-taxon network becomes a "four"-taxon network.

Theorem 17. *The probabilities of gene trees given species networks are given by Equation (6.1) and Algorithm 6.3 correspond to the correct probabilities.*

Proof. Let k_s and k_h be the numbers of speciation nodes and hybridization nodes respectively, and the total number of interior nodes is $k = k_s + k_h$. In a directed graph $G = (V, E)$, k_s is always at least one (the root node is considered as a speciation node). Therefore, for the trivial case $k = 2$, we need to consider the two cases of $k_h = 0$ and $k_h = 1$. If $k_h = 0$, then G is tree-like, and Algorithm 6.3 returns G. In this case, Equation (6.1) returns the probability of the gene tree given the species tree, which can be computed correctly (Degnan and Salter, 2005; Wu, 2012); this is true for all cases when $k = k_s$.

If $k_h = 1$ and $k_s = 1$, G is simply a loop below the root node, and all leaf nodes are connected to the hybrid node. Apply Algorithm 6.3, the hybrid node is removed by

operation 2 in a single step, any $w^* \in SG(W)$ is a tree, the probability computed by Equations (6.1) and (6.4) is the correct probability.

Suppose that our method works for some $k \geq 2$. In a network W with $k+1$ total internal nodes, apply Algorithm 6.3 to simplify W to $SG(W)$. If a hybrid descendant internal node is removed by operation 1, any $w^* \in SG(W)$ has at most k internal nodes, and $P(T|W^* = w^*)$ is correct. Thus the probability computed by Equations (6.1) and (6.3) is the correct probability.

If a hybrid node is removed by operation 2, any $w^* \in SG(W)$ has one less internal node than W (see Remark 2). Thus, $P(T|W^* = w^*)$ is valid for any $w^* \in SG(W)$, and the probability computed by Equations (6.1) and (6.4) is the correct probability.

□

6.3 `hybrid_coal`

Applying the algorithm described in the previous section, we have developed the program `hybrid_coal` to calculate gene tree probabilities given species networks using C++. Details and examples of `hybrid_coal` can be found in Appendix C.

`hybrid_coal` currently implements the coalescent history approach (Degnan and Salter, 2005), but the algorithm presented in this chapter could be used by calling STELLS (Wu, 2012) to compute the probabilities of gene trees of given species trees. The recursive approach is implemented in the program STELLS is more efficient than the coalescent history approach for moderate to larger trees (for instance, 12 or more taxa), while the coalescent history approach implemented in COAL (Degnan and Salter, 2005) can be more efficient for smaller trees.

Considering that the linear combination of species trees used to compute gene tree probabilities involves species trees of various sizes, the speed of the algorithm could potentially be optimised by using either the coalescent history approach or the ancestral approach, depending on the species tree being used.

`hybrid_coal` can also produce `Maple` script for deriving the theoretical probabilities of gene trees of given species networks. In the next section, we will use this function and discuss some problems in identifying networks.

6.4 Discussion and future work

6.4.1 Identifiability (I)

We also explore the possibilities of differentiating level-k species networks via the number of distinct gene tree probabilities. This approach shows that it is theoretically possible to determine whether a collection of gene trees comes from a species network rather than a species tree, or from a level-2 network rather than a level-1 network.

An approach used by Allman *et al.* (2011) for determining whether the probabilities of gene tree topologies can be used to identify species trees is to use the number of distinct

gene tree probabilities. For example, for a four-taxon unbalanced species tree with one individual sampled per species, there are seven distinct gene tree probabilities, while for a four-taxon balanced species tree, there are only five distinct gene tree probabilities (see Table 6.5).

Knowing the number of distinct gene tree probabilities therefore determines whether or not the species tree is balanced. Although, in practice, determining the number of distinct gene tree probabilities is difficult, since many probabilities will be close to 0 and therefore some topologies may not be observed from the data, this is still a useful theoretical tool for addressing identifiability issues and shows that there is information in the gene tree probabilities about the network topology. We use the approach of counting distinct gene tree probabilities on several species trees and level-k networks (see Tables 6.4 and 6.5).

Our results suggest that by sampling one individual per species, one can often determine whether or not a hybridization event occurred in the past because the number of distinct gene tree probabilities tends to be larger than would occur if we were given a species tree with no hybridization events. Sampling only one individual per species tends to be insufficient for determining how many hybridization events have occurred, at least for the three-, four-, five- and six-taxon examples that we tried.

6.4.2 Identifiability (II)

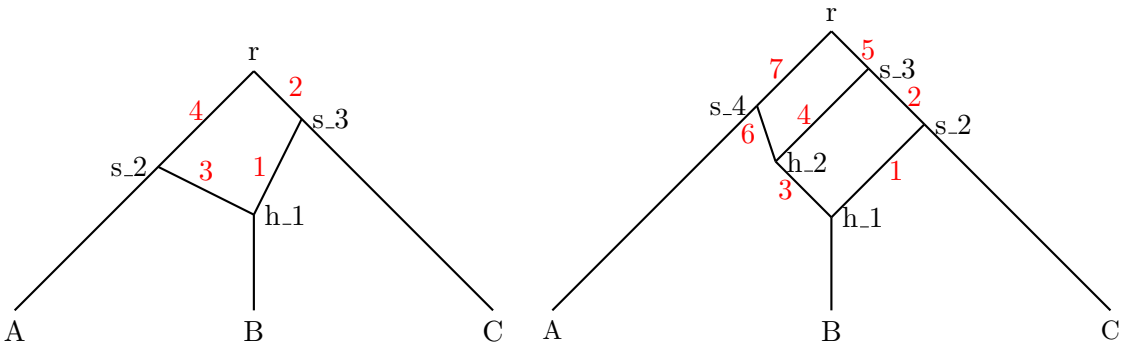
T^1_{III}	$((a_1,c_1),b_1);$	T^2_{III}	$(a_1,(b_1,c_1));$	T^3_{III}	$((a_1,b_1),c_1);$

Table 6.2: The three gene trees on three species

Consider the network W^1_{III} (see Figure 6.5a) and the gene trees in Table 6.2. The branch lengths and the hybridization parameters are denoted as λ'_i and γ'_j respectively. The gene tree probabilities are:

$$
\begin{aligned}
P(T^1_{\text{III}}|W^1_{\text{III}}) &= \frac{1}{3}\gamma'_1 \exp(-\lambda'_2) + \frac{1}{3}(1-\gamma'_1)\exp(-\lambda'_4),\\
P(T^2_{\text{III}}|W^1_{\text{III}}) &= \gamma'_1(1-\frac{2}{3}\exp(-\lambda'_2)) + \frac{1}{3}(1-\gamma'_1)\exp(-\lambda'_4),\\
P(T^3_{\text{III}}|W^1_{\text{III}}) &= \frac{1}{3}\gamma'_1 \exp(-\lambda'_2) + (1-\gamma'_1)(1-\frac{2}{3}\exp(-\lambda'_4)).
\end{aligned}
$$

Now consider the network W^2_{III} (see Figure 6.5b). Let the branch lengths and hy-



(a) Level-1 network (W^1_{III})

(b) Level-2 network (W^2_{III})

Figure 6.5: A level-1 network and a level-2 network of size three.

bridization parameters be denoted as λ_i'' and γ_i'' respectively. Let us assume the following:

$$\begin{aligned}\lambda_2' = &-\log((\gamma_1' \exp(\lambda_5'') - \gamma_1''\gamma_2'' + \gamma_1''\gamma_2'' \exp(\lambda_5'') + \gamma_1'' \exp(-\lambda_2'' - \lambda_5'') \exp(\lambda_5'') + \gamma_2'' \\ &- \gamma_1'' \exp(\lambda_5'') - \gamma_2'' \exp(\lambda_5''))/\gamma_1') + \lambda_5'', \\ \lambda_4' = &-\log((-1 + \gamma_2'' + \gamma_1'' - \gamma_1''\gamma_2'' + \gamma_1' \exp(\lambda_7'') - \gamma_1'' \exp(\lambda_7'') - \gamma_2'' \exp(\lambda_7'') \\ &+ \gamma_1''\gamma_2'' \exp(\lambda_7''))/(-1 + \gamma_1')) + \lambda_7''.\end{aligned}$$

We then have:

$$\begin{aligned}P(T^1_{\text{III}}|W^1_{\text{III}}) &= P(T^1_{\text{III}}|W^2_{\text{III}}), \\ P(T^2_{\text{III}}|W^1_{\text{III}}) &= P(T^2_{\text{III}}|W^2_{\text{III}}), \\ P(T^3_{\text{III}}|W^1_{\text{III}}) &= P(T^3_{\text{III}}|W^2_{\text{III}}).\end{aligned}$$

Similarly, for W^3_{III} and W^4_{III}, there exists expressions for γ_1' and branch lengths and hybridization parameters which make the gene tree probabilities equal. Thus, sampling one individual per taxon cannot identify species networks.

6.4.3 *Identifiability (III)*

T^1_{IV}	(((a$_1$,d$_1$),c$_1$),b$_1$);	T^2_{IV}	((a$_1$,(c$_1$,d$_1$)),b$_1$);	T^3_{IV}	(((a$_1$,c$_1$),d$_1$),b$_1$);
T^4_{IV}	((a$_1$,c$_1$),(b$_1$,d$_1$));	T^5_{IV}	(((a$_1$,c$_1$),b$_1$),d$_1$);	T^6_{IV}	((a$_1$,d$_1$),(b$_1$,c$_1$));
T^7_{IV}	(a$_1$,((b$_1$,d$_1$),c$_1$));	T^8_{IV}	(a$_1$,(b$_1$,(c$_1$,d$_1$)));	T^9_{IV}	(a$_1$,((b$_1$,c$_1$),d$_1$));
T^{10}_{IV}	((a$_1$,(b$_1$,c$_1$)),d$_1$);	T^{11}_{IV}	(((a$_1$,d$_1$),b$_1$),c$_1$);	T^{12}_{IV}	((a$_1$,(b$_1$,d$_1$)),c$_1$);
T^{13}_{IV}	(((a$_1$,b$_1$),d$_1$),c$_1$);	T^{14}_{IV}	((a$_1$,b$_1$),(c$_1$,d$_1$));	T^{15}_{IV}	(((a$_1$,b$_1$),c$_1$),d$_1$);

Table 6.3: The 15 gene trees on 4 species

Consider the level-1 network W^1_{IV} (Figure 6.2a). Branch lengths and the hybridization parameters were denoted as λ_i and γ_i respectively. We have derived the theoretical

probabilities for all 15 gene trees (see Table 6.3):

$$P(T^1_{\text{IV}}|W^1_{\text{IV}}) = P(T^{11}_{\text{IV}}|W^1_{\text{IV}}) = q_1 = X_1; \tag{6.6}$$
$$P(T^2_{\text{IV}}|W^1_{\text{IV}}) = P(T^{12}_{\text{IV}}|W^1_{\text{IV}}) = q_2 = X_1 + X_2; \tag{6.7}$$
$$P(T^3_{\text{IV}}|W^1_{\text{IV}}) = P(T^{13}_{\text{IV}}|W^1_{\text{IV}}) = q_3 = X_1 + X_3; \tag{6.8}$$
$$P(T^4_{\text{IV}}|W^1_{\text{IV}}) = P(T^{14}_{\text{IV}}|W^1_{\text{IV}}) = q_4 = 2X_1 + X_2 + X_3 + X_4; \tag{6.9}$$
$$P(T^5_{\text{IV}}|W^1_{\text{IV}}) = P(T^{15}_{\text{IV}}|W^1_{\text{IV}}) = q_5 = X_1 + X_3 + X_5; \tag{6.10}$$
$$P(T^7_{\text{IV}}|W^1_{\text{IV}}) = P(T^8_{\text{IV}}|W^1_{\text{IV}}) = q_6 = X_1 + X_2 + X_6; \tag{6.11}$$
$$P(T^6_{\text{IV}}|W^1_{\text{IV}}) = q_7 = 2X_1 + X_7; \tag{6.12}$$
$$P(T^9_{\text{IV}}|W^1_{\text{IV}}) = q_8 = X_1 + X_6 + X_7 + X_8;; \tag{6.13}$$
$$P(T^{10}_{\text{IV}}|W^1_{\text{IV}}) = q_9 = X_1 + X_5 + X_7 + X_9, \tag{6.14}$$

where:

$$\begin{aligned}
X_1 =& a^2\frac{1}{18}p_{22}(\lambda_1)p_{22}(\lambda_2)p_{33}(\lambda_3) + b^2\frac{1}{18}p_{22}(\lambda_1)p_{22}(\lambda_4)p_{33}(\lambda_5) + c\frac{1}{9}p_{22}(\lambda_1)p_{22}(\lambda_3)p_{22}(\lambda_5);\\
X_2 =& b^2\frac{1}{9}p_{22}(\lambda_1)p_{22}(\lambda_4)p_{32}(\lambda_5) + c\frac{1}{3}p_{22}(\lambda_1)p_{22}(\lambda_3)p_{21}(\lambda_5);\\
X_3 =& a^2\frac{1}{9}p_{22}(\lambda_1)p_{22}(\lambda_2)p_{32}(\lambda_3) + c\frac{1}{3}p_{22}(\lambda_1)p_{21}(\lambda_3)p_{22}(\lambda_5);\\
X_4 =& cp_{22}(\lambda_1)p_{21}(\lambda_3)p_{21}(\lambda_5);\\
X_5 =& a^2\frac{1}{3}p_{22}(\lambda_1)p_{22}(\lambda_2)p_{31}(\lambda_3);\\
X_6 =& b^2\frac{1}{3}p_{22}(\lambda_1)p_{22}(\lambda_4)p_{31}(\lambda_5);\\
X_7 =& a\frac{1}{3}p_{21}(\lambda_1)p_{22}(\lambda_3) + b\frac{1}{3}p_{21}(\lambda_1)p_{22}(\lambda_5) + a^2\frac{1}{3}p_{22}(\lambda_1)p_{21}(\lambda_2)p_{22}(\lambda_3)\\
&+ a^2\frac{1}{9}p_{22}(\lambda_1)p_{22}(\lambda_2)p_{32}(\lambda_3) + b^2\frac{1}{3}p_{22}(\lambda_1)p_{22}(\lambda_4)p_{31}(\lambda_5) + b^2\frac{1}{9}p_{22}(\lambda_1)p_{22}(\lambda_4)p_{32}(\lambda_5);\\
X_8 =& bp_{21}(\lambda_1)p_{21}(\lambda_5) + b^2p_{22}(\lambda_1)p_{21}(\lambda_4)p_{21}(\lambda_5);\\
X_9 =& ap_{21}(\lambda_1)p_{21}(\lambda_3) + a^2p_{22}(\lambda_1)p_{21}(\lambda_2)p_{21}(\lambda_3).
\end{aligned}$$

Here, $a = \gamma$, $b = \gamma$ and $c = \gamma(1-\gamma)$.

When branch lengths and hybridization parameters are non-zero, we make pair-wise comparisons among equations (6.9) and (6.7), (6.9) and (6.8), (6.11) and (6.7), (6.10) and (6.8), we have the following inequalities:

$$q_4 > q_2,\ q_4 > q_3,\ q_5 > q_3,\ \text{and}\ q_6 > q_2.$$

These inequalities also hold for network W^2_{IV}, as a result of being unable to differentiate the networks when only one locus is considered. However, from these inequalities, and Equations (6.6) to (6.14), we can determine that species B and C form a clade in the

species networks, rather than other pairs of species.

The identifiability of the complex networks improves if we sample more individuals per species. For example, for the level-1, level-2 and level-3 networks shown in Table 6.5, there are nine different probability classes out of the 15 four-taxon gene tree topologies. However, we observe that when two individuals are sampled from one hybrid species and two individuals are sampled from a non-hybrid species, or when we sample three individuals from one hybrid species, the number of distinct gene tree probabilities is different between the level-1 four-taxon network and the level-2 four-taxon network. However, there is still a lack of information to separate level-3 networks from level-2 networks.

6.4.4 Future work

This research is still ongoing . `hybrid_coal` is not yet the best implementation of the algorithm we have described in the previous sections. Note that in Equation (6.5), since for all values of w' in $AG_T(W)$ are independent from each other, this allows us to calculate the values of $P(T|W' = w')$ separately. This enables the possibility of using parallel computing to calculate the gene tree probabilities. In this case, `hybrid_coal` can potentially be written in CUDA and perform parallel computations on graphics cards.

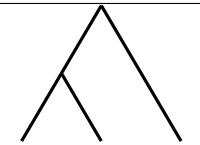
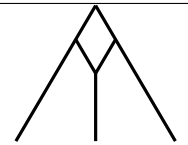
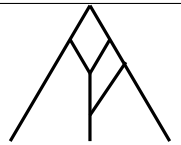
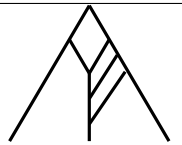
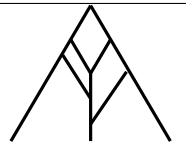
	 W_{III}^0	 W_{III}^1	 W_{III}^2	 W_{III}^3	 W_{III}^4
a_1, b_1, c_1	2	3	3	3	3
$a_1, a_2, b_1, b_2, c_1, c_2$	81	123	123	123	123
$a_1, a_2, b_1, b_2, b_3, c_1$	76	104	104	104	104

Table 6.4: Number of distinguishable gene tree probabilities, given three-taxon species networks.

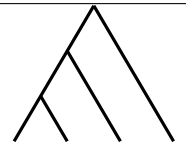
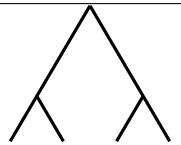
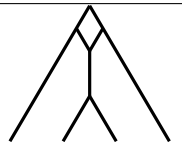
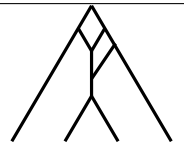
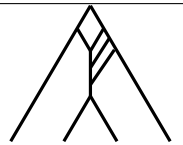
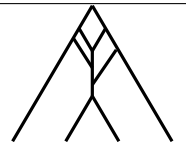
	 W_{IV}^{01}	 W_{IV}^{02}	 W_{IV}^1	 W_{IV}^2	 W_{IV}^3	 W_{IV}^4
a_1, b_1, c_1, d_1	7	5	9	9	9	9
a_1, b_1, b_2, c_1, d_1	31	20	52	52	52	52
$a_1, b_1, b_2, c_1, d_1, d_2$	122	94	264	269	269	269
$a_1, b_1, b_2, b_3, c_1, d_1$	106	63	259	318	318	318

Table 6.5: Number of distinguishable gene tree probabilities, given four-taxon species networks.

Chapter 7

Simulating genealogies

Abstract

In phylogenetic and population genetic studies, the assumption of bifurcating trees may not apply, particularly, for the cases where relationships among species are described by a network rather than a tree or where the relationships within the a single population may contain polytomy nodes. Simulation studies are interested in investigating the genealogies of hybrid species or marine organisms where one individual may have a massive number of offspring. One of the outputs of this thesis is developing the program `hybrid-Lambda`, which can simulate bifurcating or multifurcating genealogies within a species tree or network, under appropriate assumptions. This chapter reviews the methodology of the simulating procedure.

7.1 Introduction

Species trees are used to describe species relationships. Gene trees are used to describe the mutation history of alleles. Frequently, these trees are assumed to be bifurcating, for simulation studies (Degnan and Salter, 2005; Hudson, 1990). In these studies, gene trees are simulated under a particular coalescent process called the Kingman coalescent, which produces binary trees.

Let us consider the relationships among the species that are described by a network. Under the probabilistic model of gene trees introduced in the previous chapter, one can use the package `ms` (Hudson, 2002) to simulate genealogies within a general species network. However the input of `ms` is very tedious when the network is sophisticated. Other than `ms`, simulation packages are mostly designed for specific studies, for example `phylonet` (Yu *et al.*, 2011), which is not feasible for general simulation use. Therefore, simulating gene trees from species network with a simple expression becomes one motivation for this study.

For organisms where one individual can produce a large number of offspring, such as oysters and other marine organisms (Sargsyan and Wakeley, 2008), the Kingman coalescent is not appropriate, as it only allows one parent node to have two child nodes. Thus we

consider models that allow more than two lineages to coalesce simultaneously, i.e. multiple merger coalescent, also known as the Λ-coalescent (Eldon and Wakeley, 2006; Pitman, 1999).

However, the probability of a multiple merger gene tree becomes messy when either or both of the number of populations and the number of sampled individuals per population increase. Thus, studies of multiple merger coalescence for many individuals can only be undertaken by simulations. The program `simcoal` (Excoffier *et al.*, 2000) can simulate multiple merger coalescent trees. However, `simcoal` assumes that coalescence occurs generation by generation, which is different from continuous time approximation. Therefore, the program `simcoal` is not feasible for simulations under the Λ-coalescent.

The program `hybrid-Lambda` has been developed to simulate gene trees of a given species network allowing multiple merger coalescence (see Figure 7.1). The simulation procedure is explained in detail in this chapter.

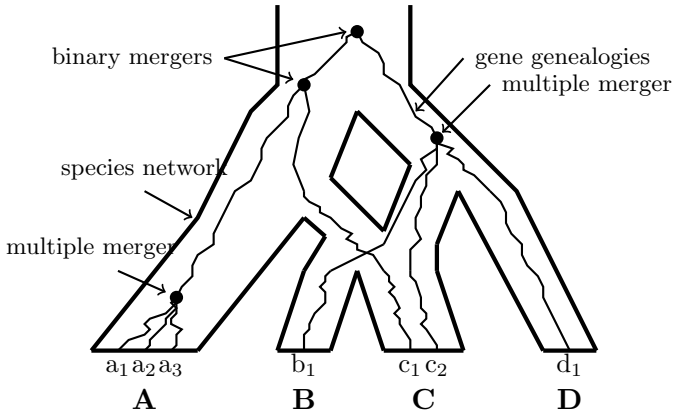


Figure 7.1: Example of a multiple merger gene tree simulated in a network with the topology `((((B,C)s1)h1#H1,A)s2,(h1#H1,D)s3)r`.

7.2 Method

In this section, we will discuss gene tree simulation in species network under the coalescent process. Apart from hybridization of two populations, we assume there is no migration and recombination among the sequences. We will discuss both the Kingman coalescent and the multiple merger coalescent in diploid species, in which each individual carries two copies of the genome. Hence if the effective population size of i is $N_i/2$, there are N_i gene copies.

For τ_i generations of species i, we assume that species i has a constant population size. We rescale the time by the number of gene copies N_i to make it continuous, as $\frac{\tau_i}{N_i}$.

Firstly, let us consider the simplest model, the Kingman coalescent model, which suggests that the waiting time of two lineages to coalesce is an exponential random variable with a rate of 1. Suppose there are n' lineages entering a population (also known as the

number of living lineages). There are $\binom{n'}{2}$ ways to choose two lineages from n'. Therefore, the time X in which n' lineages coalesce to $n'-1$ lineages is an exponential random variable with the rate $\binom{n'}{2}$, i.e. $X \sim \text{Exp}\left(\binom{n'}{2}\right)$.

For the Λ-coalescent, if the coalescent parameter is between zero and one, the rate parameter of k mergers between or among b lineages (λ_{bk}) is determined by Equation (7.1) (Pitman, 1999; Eldon, 2011); if the coalescent parameter is between one and two, then rate parameter is obtained from Equation (7.2) (Eldon and Wakeley, 2006):

$$\lambda_{bk} = \binom{b}{k} \frac{B(k-\alpha, b-k+\alpha)}{B(2-\alpha, \alpha)}, \tag{7.1}$$

where $B(2-\alpha, \alpha)$ is a beta function.

$$\lambda_{bk} = \binom{b}{k} \psi^k (1-\psi)^{b-k}. \tag{7.2}$$

If the coalescent event is Λ-coalescent, we need to consider cases that two to n' lineages will coalesce simultaneously. We use X to denote the waiting time for the next Λ-coalescent event. Here we introduce two approaches to simulate X.

1. Let $X_2, X_3, \ldots, X_{n'}$ be the time that two, three, ..., n' lineages coalesce into one respectively. Therefore, $X = \min\{X_2, X_3, \ldots, X_{n'}\}$.

2. Since each X_i is a exponential random variable with the rate of $\lambda_{n'i}$, for $i = \{2, \ldots n'\}$. Therefore, $X \sim \text{Exp}\left(\sum_{i=2}^{n'} \lambda_{n'i}\right)$.

These two methods are described by Algorithm 7.1a and Algorithm 7.1b respectively. Both algorithms are easy to implement. However, the time complexity of Algorithm 7.1b is worse in comparison. Thus we apply Algorithm 7.1a in practice. By considering the Kingman coalescent, we establish Algorithm 7.2 to simulate more general cases of n_c lineages to coalesce in time l.

1: **for** i in (2 to n') **do**
2: Propose $l_i \sim \text{Exp}(\lambda_{n'i})$.
3: **end for**
4: For $n_c \in \{2, 3, \ldots, n'\}$, we have $l_{n_c} = \min\{l_2, \ldots, l_{n'}\}$, and $l \leftarrow l_{n_c}$

(a)

1: Propose $l \sim \text{Exp}\left(\sum_{i=2}^{n'} \lambda_{n'i}\right)$.
2: Propose $n_c = i$, with probability $\frac{\lambda_{n'i}}{\sum_{i=2}^{n'} \lambda_{n'i}}$.

(b)

Algorithm 7.1: Propose the waiting time l of n_c lineages to coalesce under the Λ-coalescent.

Algorithm 7.2 Propose the waiting time l of n_c lineages to coalesce.

1: **if** Λ-coalescent **then**
2: Apply Algorithm 7.1a.
3: **else**
4: $l \sim \text{Exp}\left(\binom{n'}{2}\right)$, $n_c \leftarrow 2$.
5: **end if**

7.3 Simulating coalescent time in coalescent units

Recall Equation (6.2), the probability of the coalescent events $p_{uv}(t)$ within a time interval t only depends on the numbers of lineages entering (u) and exiting (v) the population. Therefore, the key to simulating the genealogy is to keep track the number of living lineages in a population.

7.3.1 Simulating lineage coalescent events in a time interval

Let t_0, t_1 be the time at the two ends of one particular population, where t_0 is more recent than t_1. Let L_n be the set of lineages entering this population from the present at t_0, and $n = |L_n|$. Let L'_n be the set of living lineages within the population between time t_0 and t_1, and $n' = |L'_n|$. We refer the absolute time of a lineage as the time from the bottom of a lineage to the present, denoted as a. We use b to denote the branch length of a lineage, i.e. the time difference between the two ends of a lineage. Let l_0 be the remaining time from the coalescent event to the time at the top of the population t_1. We use the algorithm described in Algorithm 7.3 to update the branch length of genealogy lineages within a time interval.

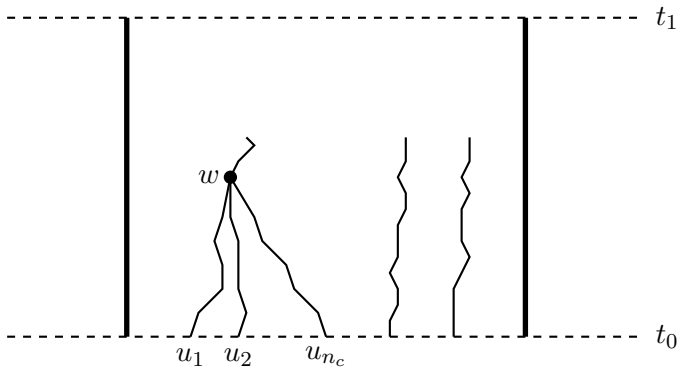


Figure 7.2: Lineages coalescing in an internal branch.

7.3.2 Simulating lineage coalescent events above the root

All lineages will eventually all coalesce above the root, which implies that the branch lengths of lineages can extend to infinity above the root of the phylogeny. Thus, it is

Algorithm 7.3 Algorithm for updating genealogy branch lengths within a time interval.

1: **if** $n' > 1$ **then**
2: Apply Algorithm 7.2 to propose a new branch length l and the number of lineages n' to coalesce.
3: **end if**
4: *initialize:* $n' \leftarrow n$, $L'_n \leftarrow L_n$, $l_0 \leftarrow t_1 - t_0$.
5: **while** $n' > 1, l_0 > l$ **do**
6: **for** i in (1 to n_c) **do**
7: Choose u uniformly from L'_n.
8: $b_u \leftarrow b_u + l$.
9: $L'_n \leftarrow L'_n \setminus \{u\}$,
$n' \leftarrow n' - 1$.
10: **end for**
11: Introduce a new lineage w.
12: $a_w \leftarrow a_u + b_u$.
13: $\forall i \in L'_n$, $b_i \leftarrow b_i + l$.
14: $L'_n \leftarrow L'_n \cup \{w\}$,
$n' \leftarrow n' + 1$.
15: Update l_0, $l_0 \leftarrow l_0 - l$.
16: Apply Algorithm 7.2 to propose a new branch length l and the number of lineages n' to coalesce.
17: **end while**
18: $\forall u \in L'_n, b_u \leftarrow t_1 - a_u$.

not necessary to consider the remaining time in a population (l_0 in Algorithm 7.3). By applying the similar approach mentioned in the previous section, we make the following modifications to Algorithm 7.3 to update the branch lengths of the genealogy at the root of the phylogeny:

1. Remove the initialisation of l_0, $l_0 \leftarrow t_1 - t_0$ at line 4.

2. Remove $l_0 > l$ of the while loop condition at line 5.

3. Remove line 15 and line 18.

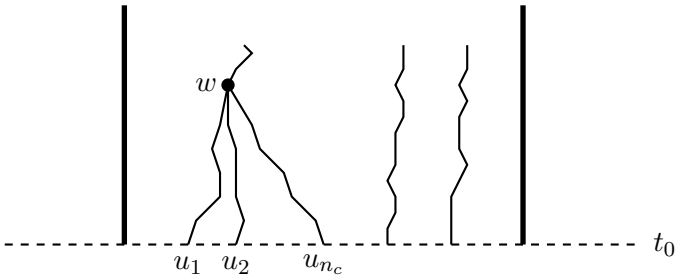


Figure 7.3: Demonstration of lineages coalescing above the root.

7.3.3 Simulating lineage coalescent events at a hybrid node

Since there are two populations at the hybrid node of the phylogeny, and the coalescent events are independent between the two populations, we can apply Algorithm 7.3 in each population. However, the genealogys' branches need to be assigned randomly to the two populations (see Figure 7.4).

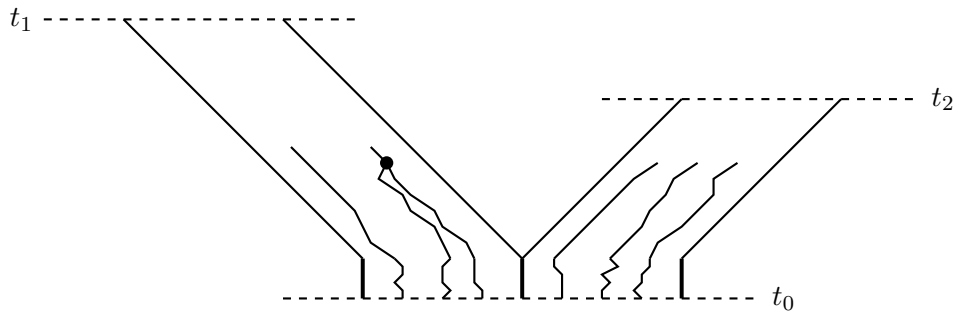


Figure 7.4: Lineages coalescing at a hybrid node.

Let L_n denote the set of lineages entering the hybrid node at t_0, let γ be the probability of a lineage entering the left-hand population. Then, L_n is divided into L_L and L_R, where L_L and L_R are the sets of lineages entering the left-hand and right-hand populations respectively. Let t_1 and t_2 be the times at the top of the left-hand and right-hand populations respectively. We can then apply Algorithm 7.3 twice to update the genealogy branch lengths. This procedure is described by Algorithm 7.4.

Algorithm 7.4 Algorithm to updating genealogy branch lengths at a hybrid node.

$\forall i \in L_n$, propose $\delta \sim Unif(0,1)$
if $\delta < \gamma$ **then**
  $i \in L_L$
else
  $i \in L_R$
end if
$n_L \leftarrow |L_L|$, and $n_R \leftarrow |L_R|$
Apply Algorithm 7.3 and initialise with $n' \leftarrow n_L$, $L'_n \leftarrow L_L$, $l_0 \leftarrow t_1 - t_0$ at line 4.
Apply Algorithm 7.3 and initialise with $n' \leftarrow n_R$, $L'_n \leftarrow L_R$, $l_0 \leftarrow t_2 - t_0$ at line 4.

7.4 Simulating coalescent time in number of generations

We use B to denote the genealogy branch lengths in term of the number of generations. Let N be the number of gene copies in a population. To convert the branch lengths from the coalescent time to the number of generations, we make the following modifications to Algorithm 7.3:

1. At line 8, add $B_u \leftarrow B_u + l \times N$.

2. At line 13, add $B_i \leftarrow B_i + l \times N$.

3. At line 18, before updating b_i, add the following:

if $a_i > t_0$ **then**
  $B_i \leftarrow (t_1 - a_i) \times N$
else
  $B_i \leftarrow (t_1 - a_i - b_i) \times N + B_i$
end if.

7.5 Segregating site data

In this chapter, we do not consider the possibility of sequence recombination. Mutation is the only source of DNA sequence changes. Under the infinite site model, we assume that mutations always occur on the new site; the number of mutations is a Poisson random variable with the mean of $\frac{\theta}{2}t$ (Wakeley, 2008), where θ is the mutation rate per generation. In this section, we will discuss how to simulate segregating site data from the simulated genealogies.

7.5.1 Expected number of mutations

Let μ be the mutation rate per locus per generation. Therefore, for N gene copies, the mutation rate per generation is $2N\mu$. After we simulate the genealogy branch lengths t in coalescent units, we convert it to the number of generations τ, which is equal to Nt. Let M be the number of mutations. The expected value of M is now $\frac{\theta}{2}t$. Since $\theta = 2N\mu$ and $t = \frac{\tau}{N}$, we have:

$$\mathbb{E}(M) = \mu\tau.$$

We first calculate the total branch lengths in a generation and simulate the number of mutations. We then assign each mutation to the gene tree branch i with the probability of $\frac{\tau_i}{\sum \tau_i}$, and write it as a Newick formatted string (see the example in Figure 7.5).

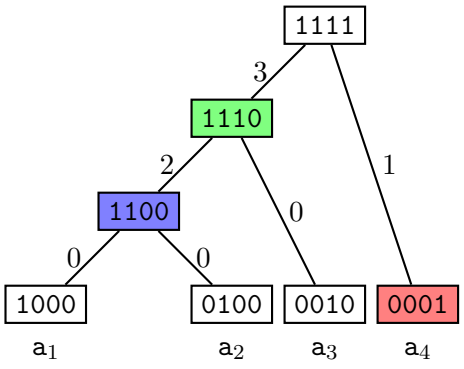


Figure 7.5: Simulated geneology `(((a1:0,a2:0):2,a3:0):3,a4:1)`, where the branch lengths indicate the number of mutations.

7.5.2 *Simulating segregating site data*

Note that the number of segregating sites is the total number of mutations in the history of a sample. This enables us to simulate segregating site data from the simulated number of mutations.

At each node of the gene tree, we can use the numbers 1 and 0 to indicate whether or not a particular sample is a descendant from this node. Therefore, for a gene tree of k samples, node i can be expressed as a $0, 1$ vector v_i of length k. We build a matrix S of k rows, whose columns are the $0, 1$ vectors v_i repeated M_i times, which is the number of mutations on the edge that is connected to node i towards the root. Therefore, the rows of the matrix S are the simulated segregating site data. An example of this procedure is demonstrated in Figures 7.5 and 7.6.

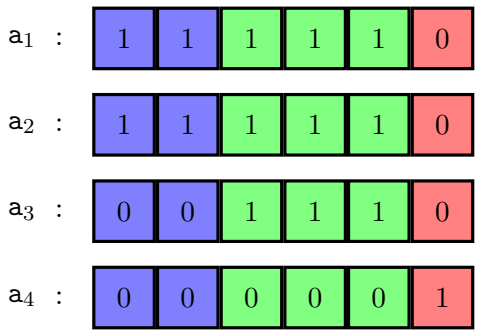


Figure 7.6: Segregating site data associated with the simulated genealogy `(((a1:0,a2:0):2,a3:0):3,a4:1)`. Taxa $\mathtt{a}_1$ and $\mathtt{a}_2$ both have the same sequence 111110; taxon $\mathtt{a}_3$ has the sequence 001110; and taxon $\mathtt{a}_4$ has the sequence 000001.

7.6 `hybrid-Lambda`

Applying the algorithm described in the previous section, we have developed the program `hybrid-Lambda` to simulate genealogies of a given species tree or network, allowing multiple merger coalescence.

`hybrid-Lambda` can sample multiple lineages from each species, then simulate lineage coalescent time under a Markov process. New coalescent events are conditional on prior coalescent events, as well as on population branch lengths. By default, the Kingman coalescent is used. For the Λ-coalescent, the current version of `hybrid-Lambda` uses Algorithm 7.1a to propose the coalescent time.

The program input file is a character string that describes the relationships among species. Standard Newick format (Olsen, 1990) is used for inputting species trees and outputting gene trees, the interior nodes of which are not labelled. An extended Newick formatted string (Cardona *et al.*, 2008; Huson *et al.*, 2010) labels all internal nodes, and is used for inputting species networks. (See Figure 7.1).

As the population branch specifies, `hybrid-Lambda` requires the input network (tree) branch lengths to be in coalescent units. However, this is not essential. Coalescent units can be converted through an alternative input file with numbers of generations as the branch lengths, divided by the corresponding effective population sizes. By default, effective population sizes of all branches are assumed to be equal and unchanged. Users can change this parameter using the command line or by using a(n) (extended) Newick formatted string to specify population sizes on all branches though another input file.

According to the molecular clock, the simulation requires species structures to be *ultrametric*, i.e. lengths of all paths from tip to root are equal. `hybrid-Lambda` checks the distances in coalescent units between the root and all tip nodes, and prints out warning messages if the ultrametric assumption is violated.

7.6.1 Features

`hybrid-Lambda` outputs simulated gene trees in three different files: one contains gene trees with branch lengths in coalescent units; one converts branch lengths from coalescent units to the number of generations; one uses the number of expected mutations as branch lengths. Moreover, `hybrid-Lambda` has an option that allows segregating data from the infinite site model to be simulated for each gene tree generated.

Besides outputting gene tree files, `hybrid-Lambda` also provides several functions for analytical purposes:

- user-defined random seeds for simulation,
- a frequency table of gene tree topologies,
- a figure of the species network or tree (this function only works when LaTeX or `dot` is installed), and
- when gene trees are simulated from two populations, `hybrid-Lambda` can generate a table of relative frequencies of the cases where gene trees are reciprocally monophyletic and polyphyletic, as well as cases where each population is monophyletic or paraphyletic.

A detailed description and examples can be found in Appendix D

Gene tree topologies	Probabilities	Expected counts	Actual counts
(((A,D),C),B)	0.00157014	157.014	143
((A,(C,D)),B)	0.0104929	1049.29	1019
(((A,C),D),B)	0.0104929	1049.29	1054
((A,C),(B,D))	0.0577347	5773.47	5730
(((A,C),B),D)	0.0158282	1582.82	1571
((A,D),(B,C))	0.0985005	9850.05	9704
(A,((B,D),C))	0.0158282	1582.82	1597
(A,(B,(C,D)))	0.0158282	1582.82	1523
(A,((B,C),D))	0.338803	33880.3	34028
((A,(B,C)),D)	0.338803	33880.3	33812
(((A,D),B),C)	0.00157014	157.014	157
((A,(B,D)),C)	0.0104929	1049.29	1071
(((A,B),D),C)	0.0104929	1049.29	1062
((A,B),(C,D))	0.0577347	5773.47	5907
(((A,B),C),D)	0.0158282	1582.82	1622

Table 7.1: Expected and acutal counts of simulating 100,000 gene trees given the network `'((((B:1,C:1)s1:1)h1#.5:1,A:3)s2:1,(h1#.5:1,D:3)s3:1)r;'`. The gene tree probabilities were calculated by the program `hybrid_coal`; the simulated gene trees and frequency table were generated by `hybrid-Lambda`. A Chi-square test was conducted on this data set, with the null hypothesis of the expected count and actual were from the same distribution. The test statistics retured a p-value of 0.562, and we failed to reject the H null at 0.05 level. This data set provided insufficient evidence to suggest that the gene trees were simulated from an alternative distribution.

Chapter 8

Concluding comments

To sum up, in this thesis, we studied several modern techniques and models that are used in macro evolution and micro evolution, such as the YH and PDA models, the Tree-Puzzle process, the Kingman and multiple merger coalescent models. We investigated the properties of the evolutionary models in depth, particularly for the YHK model and the PDA model.

In the PDA model, new leaf nodes are uniformly added to an edge of the existing tree, whereas the Yule tree selects a pendant edge randomly and adds a new node to this pendant edge. During the construction process, PDA, RTP and RTP$'$ can all attach new leaves to interior edges. For the PDA process, this has probability of almost 1/2 (and much less for RTP), as the number of leaves increases. In the case of RTP$'$, beyond seven leaves, all further leaves are attached to a pendant edge, just as in the YH model. However, we found evidence that the variational distance between YH and RTP$'$ appeared to remain bounded away from zero even when n tends to infinity, which suggested that they are still two distinct tree construction methods. Thus, we conjectured that the RTP process was different from the YH process. Even so, we have verified that the RTP process would eventually not add new leaves onto interior edges after some point, which would make the RTP process become more like the YH process.

By comparing the YH model and the Kingman model, we found that these two different processes had the same probability distribution on tree topologies. Therefore, we generalised these two models as the YHK model. We applied the three properties that both YHK and PDA shared: the EP, GE and SC properties to derive the probabilities that a set is a clade and that two compatible sets are clades in rooted trees. These clade probabilities were extended to the cases of $k > 2$ compatible clades under both the YHK and PDA models.

By using the clade probabilities in rooted trees, we derived the clan probabilities in the unrooted trees. Under both the YHK and PDA models, we found that two clades were compatible, they were positively correlated. However, the correlation appeared to be weak. Similarly, two clans under the PDA model were also positively correlated. We also related the clade probabilities to the balance of a rooted tree, and showed an alternative

proof and a new expression for computing the expected value of the Sackin index under the YHK model and the PDA model respectively

In this thesis, we also considered the scenarios that phylogenies and genealogies were not binary trees, in which the speciation and coalescent process could not be modelled by the YH process and the Kingman process respectively.

For closely related species, interbreeding is likely to occur, which makes it very difficult to infer the phylogeny from genealogies, because of the incongruent gene tree shapes. In this research, we focused on calculating the probabilities of gene trees for a given set of taxa, with the presence of both lineage sorting and hybridization events. I developed a program for this project that can decompose a species network into a sequence of networks, according to the gene trees' coalescent histories, and can then compute the gene tree probability, given the species network, as a linear combination of the gene tree probabilities, given a set of species trees.

We are still working on the identification of networks from a set of gene trees and their frequencies. Our prelimneary results have made us believe that using the different number of gene tree probabilities can help us to identify certain types and numbers of hybridization patterns.

Finally, I developed another program to simulate gene trees within a species network or tree. This program is different from 'ms' and 'simcoal'. It allows the user to simulate genealogies from the Kingman coalescent process or the Λ-coalescent process. This program can undertake some simple analysis of the simulated gene tree topologies and simulate a segregating data-set from the infinite site model.

Bibliography

Aldous, D. (1995). *Random Discrete Structures*. New York: Springer.

Aldous, D. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science 16*(1), 23–34.

Allman, E. S., J. H. Degnan, and J. A. Rhodes (2011). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology 62*, 833–862.

Árnason, E. (2004). Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics 166*, 1871–1885.

Bagchi, A. and A. K. Pal (1985). Asymptotic normality in the generalized Polya–Eggenberger urn model, with an application to computer data structures. *SIAM Journal on Algebraic and Discrete Methods 6*(3), 394–405.

Barthélemy, J.-P. and A. Guenoche (1991). *Trees and Proximity Representations*. London: Wiley.

Beckenbach, A. (1994). Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models. In B. Golding (Ed.), *Non-neutral Evolution*, pp. 188–198. New York: Chapman and Hall.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distribution via Pólya Urn schemes. *The Annals of Statistics 1*(2), 353–355.

Blum, M. G. B., O. Francois, and S. Janson (2006). The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *The Annals of Applied Probability 16*(4), 2195–2214.

Bortolussi, N., E. Durand, M. Blum, and O. Francois (2006). aptreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics 22*(3), 363–364.

Bortolussi, N., E. Durand, M. Blum, and O. Francois (2009). Analyses of phylogenetic treeshape. http://cran.r-project.org/web/packages/apTreeshape/apTreeshape.pdf, Feb 22, 2013.

Brown, J. K. M. (1994). Probabilities of evolutionary trees. *Systematic Biology 43*(1), 78–91.

Brumfield, R. T. and M. D. Carling (2010). The influence of hybrid zones on species tree inference in manakins. In L. L. Knowles and L. S. Kubatko (Eds.), *Estimating Species Trees, Practical and Theoretical Aspects*, pp. 115–127. Hoboken, NJ: Wiley-Blackwell.

Bruno, W. J., N. D. Socci, and A. L. Halpern (2000). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution 17*, 189–197.

Buckley, T. R., M. Cordeiro, D. C. Marshall, and C. Simon (2006). Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada Dugdale*). *Systematic Biology 55*(3), 411–425.

Cardona, G., F. Rossell, and G. Valiente (2008). Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics 9*(532-540).

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

Colless, D. H. (1982). Review of "Phyogenetics: The theory and practice of phylogenetic systematics". *Systematic Zoology 31*, 100–104.

Cummings, M. P., M. C. Neel, and K. L. Shaw (2008). A genealogical approach to quantifying lineage divergence. *Evolution 62*(9), 2411–2422.

Darlu, P. and G. Lecointre (2002). When does the incongruence length difference test fail? *Molecular Biology and Evolution 19*, 432–437.

Daubin, V. and H. Ochman (2004). Quartet mapping and the extent of lateral transfer in bacterial genomes. *Molecular Biology and Evolution 1*, 86–89.

de Quieroz, K. (2007). Species concepts and species delimitation. *Systematic Biology 56*, 879–886.

de Quieroz, K. and J. Gauthier (1990). Phylogeny as a central principle in taxonomy: phylogenetic definitions of taxon names. *Systematic Zoology 39*(4), 307–322.

Degnan, J. H. (2010). Probabilities of gene tree topologies with intraspecific sampling given a species tree. In L. L. Knowles and L. S. Kubatko (Eds.), *Estimating Species Trees, Practical and Theoretical Aspects*, pp. 53–77. Hoboken, NJ: Wiley-Blackwell.

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

Donnelly, P. and T. Kurtz (1999). Particle representations for measure-valued population models. *The Annals of Probability 27*(1), 166–205.

Dress, A., A. von Haeseler, and M. Krueger (1986). Reconstructing phylogenetic trees using variants of the "four-point condition". *Studien zur Klassifikation 17*, 299–305.

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

Eldon, B. (2011). Estimation of parameters in large offspring number models and ratios of coalescence times. *Theoretical Population Biology 80*, 16–28.

Eldon, B. and J. H. Degnan (2012). Multiple merger gene genealogies in two species: monophyly, paraphyly, and polyphyly for two examples of Lambda coalescents. *Theoretical Population Biology 82*, 117–130.

Eldon, B. and J. Wakeley (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics 172*, 2621–2633.

Excoffier, L., J. Novembre, and S. Schneider (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.

Felsenstein, J. (1978). The number of evolutionary trees. *Systematic Zoology 27*, 27–33.

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

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

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology 20*(4), 406–416.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution 14*(7), 685–695.

Guigo, R., I. Muchnik, and T. F. Smith (1996). Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution 6*(2), 189 – 213.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology 59*(3), 307–321.

Guindon, S. and O. Gascuel (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology 52*(5), 696–704.

Harding, E. F. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability 3*(1), 44–77.

Hedgecock, D. (1994). Does variance in reproductive success limit effective population sizes of marine organisms? In A. Beaumont (Ed.), *Genetics and Evolution of Aquatic Organisms*, pp. 1222–1344. London: Chapman and Hall.

Hedgecock, D., M. Tracey, and K. Nelson (1982). Genetics. In L. Abele (Ed.), *The Biology of Crustacea*, Volume 2, pp. 297–403. New York: Academic Press.

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

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*(301), 13–30.

Holland, B. and V. Moulton (2003). Consensus networks: A method for visualising incompatibilities in collections of trees. In G. Benson and R. Page (Eds.), *Algorithms in Bioinformatics*, Volume 2812 of *Lecture Notes in Computer Science*, pp. 165–176. Springer Berlin Heidelberg.

Holland, B. R., S. Benthin, P. J. Lockhart, V. Moulton, and K. T. Huber (2008). Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology 8*, 202–213.

Holland, B. R., K. T. Huber, V. Moulton, and P. J. Lockhart (2004). Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution 21*(7), 1459–1461.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys Evolution Biology 7*, 1–44.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model. *Bioinformatics 18*, 337–338.

Hudson, R. R. and J. A. Coyne (2002). Mathematical consequences of the genealogical species concept. *Evolution 56*(8), 1557–1565.

Huelsenbeck, J. P. and F. Ronquist (2001). MrBayes: Bayesian inference of phylogeny. *Bioinformatics 17*, 754–755.

Huson, D., R. Rupp, and C. Scornavacca (2010). *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge, UK: Cambridge University Press.

Huynh, T. N. D., J. Jansson, N. B. Nguyen, and W.-K. Sung (2005). Constructing a smallest refining galled phylogenetic network. *Annual International Conference on Research in Computational Molecular Biology 2005, LNBI 3500*, 265–280.

Joly, S., P. A. McLenachan, and P. J. Lockhart (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist 174*, 54–70.

Jordan, C. (1869). Sur les assemblages des lignes. *Journal für die Reine und Angewandte Mathematik 70*, 185–190.

Kang, A. N. C. and D. A. Ault (1975). Some properties of a centroid of a free tree. *Information Processing Letters 4*(1), 18–20.

Karr, A. F. (1993). *Probability*. New York: Springer-Verlag.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability 19*, 27–43.

Kirkpatrick, M. and M. Slatkin (1993). Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution 47*(4), 1171–1181.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2000). *Continuous Multivariate Distributions, Volume 1, Models and Applications* (2 ed.). New York: Wiley.

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

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

Lapointe, F.-J., P. Lopez, Y. Boucher, J. Koenig, and E. Bapteste (2010). Clanistics: a multi-level perspective for harvesting unrooted gene trees. *Trends in Microbiology 18*, 341–347.

Leigh, J. W., E. Susko, M. Baumgartner, and A. J. Roger (2008). Testing congruence in phylogenomic analysis. *Systematic Biology 57*(1), 104–115.

Li, S., D. K. Pearl, and H. Doss (2000). Phylogenetic tree construction using Markov Chain Monte Carlo. *Journal of the American Statistical Association 95*(450), 493–508.

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

Liu, L., L. Yu, and D. K. Pearl (2010). Maximum tree: a consistent estimator of the species tree. *Journal of Mathematical Biology 60*, 95–106.

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

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

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

Mahmoud, H. (2008). *Pólya Urn Models*. Boca Raton: Chapman and Hall/CRC.

Mallet, J., M. Beltrán, W. Neukirchen, and M. Linares (2007). Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology 7*, 28.

Matsumoto, M. and T. Nishimura (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform Pseudo-Random number generator. *ACM Transactions on Modeling and Computer Simulation 8*(1), 3–30.

McKenzie, A. (2000). *Stochastic Speciation Models for Evolutionary Trees*. Ph. D. thesis, University of Canterbury, Christchurch.

McKenzie, A. and M. A. Steel (2000). Distributions of cherries for two models of trees. *Mathematical Biosciences 164*, 81–92.

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

Mitchell, S. L. (1978). Another characterization of the centroid of a tree. *Discrete Mathematics 24*, 277–280.

Nieselt-Struwe, K. and A. von Haeseler (2001). Quartet-mapping, a generalization of the likelihood-mapping procedure. *Molecular Biology and Evolution 7*(18), 1204–1219.

Olsen, G. (1990). Gary Olsen's interpretation of the "Newick's 8:45" tree format standard. http://evolution.genetics.washington.edu/phylip/newick_doc.html. Feb 21, 2013.

Paradis, E., J. Claude, and K. Strimmer (2004). Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics 20*, 289–290.

Pinelis, I. (2003). Evolutionary models of phylogenetic trees. *Proceedings of the Royal Society of London. Series B: Biological Sciences 270*(1522), 1425–1431.

Pitman, J. (1999). Coalescents with multiple collisions. *The Annals of Probability 27*(4), 1870–1902.

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

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

Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology 61*, 225–247.

Rosenberg, N. A. (2003). The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly and polyphyly in a coalescent model. *Evolution 57*(7), 1465–1477.

Rosenberg, N. A. (2007). Statistical tests for taxonomic distinctiveness from observations of monophyly. *Evolution 61*(2), 317–323.

Sackin, M. J. (1972). "Good" and "bad" phenograms. *Systematic Zoology 21*(2), 225–226.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability 36*, 1116–1125.

Saitou, N. and M. Nei (1987). The Neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution 4*(4), 406–425.

Sang, T. and Y. Zhong (2000). Testing hybridization hypotheses based on incongruent gene trees. *Systematic Biology 49*, 422–434.

Sankoff, D. D. (1975). Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics 28*, 35–42.

Sankoff, D. D. and P. Rousseau (1975). Locating the vertices of a Steiner tree in arbitrary space. *Mathematical Programming 9*, 240–246.

Sargsyan, O. and J. Wakeley (2008). A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoretical Population Biology 74*, 104–114.

Saunders, I. W., S. Tavaré, and G. A. Watterson (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability 16*, 471–491.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics 18*(3), 502–504.

Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications 106*, 107–139.

Semple, C. and M. A. Steel (2003). *Phylogenetics*. Oxford, UK: Oxford University Press.

Shaw, K. L. (1998). Species and the diversity of natural groups. In D. J. Howard and S. H. Berlocher (Eds.), *Endless Forms: Species and Speciation.*, pp. 44–56. Oxford, UK: Oxford University Press.

Slowinski, J. B. (1990). Probabilities of n-trees under two models: a demonstration that asymmetrical interior nodes are not improbable. *Systematic Zoology 39*(1), 89–94.

Smythe, R. T. (1996). Central limit theorems for urn models. *Stochastic Processes and their Applications 65*, 115–137.

Sokal, R. and C. Michener (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin 38*(22), 1409–1438.

Steel, M. A. (1993). Distributions on bicoloured binary trees arising from the principle of parsimony. *Discrete Applied Mathematics 43*, 245–261.

Steel, M. A. (2012). Root location in random trees: A polarity property of all sampling consistent phylogenetic models except one. *Molecular Phylogenetics and Evolution 65*(1), 345 – 348.

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

Strimmer, K., N. Goldman, and A. von Haeseler (1997). Bayesian probabilities and quartet puzzling. *Molecular Biology and Evolution 2*(14), 210–211.

Strimmer, K. and A. von Haeseler (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution 13*(7), 964–969.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics 105*, 437–460.

Takahata, N. and M. Nei (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics 110*, 325–344.

Tavaré, S. (1984). Line-of-descent and genealogical processes and their applications in population genetics models. *Theoretical Population Biology 26*(2), 119–164.

Than, C. and L. Nakhleh (2009). Species tree inference by minimizing deep coalescences. *PLoS Computational Biology 5*(9), 12.

Than, C., D. Ruths, and L. Nakhleh (2008). PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics 9*, 322.

Valiente, G. (2002). *Algorithms on Trees and Graphs*. Berlin, Germany: Springer.

van Iersel, L., J. Keijsper, S. Kelk, L. Stougie, F. Hagen, and T. Boekhout (2008). Constructing level-2 phylogenetic networks from triplets. *Annual International Conference on Research in Computational Molecular Biology 2008. LNCS 4955*, 450–462.

Vinh, L. S., A. Fuehrer, and A. von Haeseler (2011). Random tree-puzzle leads to the Yule–Harding distribution. *Molecular Biology and Evolution 28*(2), 873–877.

Wakeley, J. (2008). *Coalescent theory: An Introduction*. Greenwood Village, CO: Roberts and Co.

Watterson, G. A. (1984). Lines of descent and the coalescent. *Theoretical Population Biology* (26), 77–92.

Wilkinson, M., J. O. McInerney, R. P. Hirt, P. G. Foster, and T. M. Embley (2007). Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends in Ecology and Evolution 22*(3), 114–115.

Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution 66*(3), 763–775.

Yu, Y., J. H. Degnan, and L. Nakhleh (2012). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet. 8*, e1002660.

Yu, Y., C. Than, J. H. Degnan, and L. Nakhleh (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology 60*(2), 138–149.

Yule, G. U. (1925). A mathematical theory of evolution. based on the conclusions of Dr. J.C. Willis, F.R.S. In *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, Volume 213, pp. 21–87.

Zhu, S., J. H. Degnan, and M. A. Steel (2011a). Clades, clans and reciprocal monophyly under neutral evolutionary models. *Theoretical Population Biology 79*, 220–227.

Zhu, S., J. H. Degnan, and M. A. Steel (2011b). Probabilistic modeling of gene trees given species networks. Poster, http://www.newton.ac.uk/programmes/PLG/Zhu.pdf. Feb 21, 2013.

Zhu, S. and M. A. Steel (2013). Does Random Tree Puzzle produce Yule–Harding trees in the many-taxon limit? *Mathematical Biosciences*. In press.

Appendix A

Symbols used

Chapter 2

$\mathbb{P}(\mathcal{E})$	probability of the event $\mathcal{E}$
$\mathbb{P}(\mathcal{E}_1 \vert \mathcal{E}_2)$	conditional probability of the event $\mathcal{E}_1$ given the event $\mathcal{E}_2$
$A \vert A^c$	split of A and A^c
$\mathbb{E}[X]$	expected value of the random variable X
$\mathrm{var}(X)$	variance of the random variable X
$f_X(x)$	probability mass/density function of the random variable X
$F_X(x)$	probability distribution function of the random variable X

Chapter 3

T_n	labelled and unrooted binary tree with n leaves
t_n	unlabelled and unrooted binary tree with n leaves
$\mathcal{T}(n)$	set of labelled and unrooted binary trees with n leaves
$\mathcal{S}(n)$	set of unlabelled and unrooted binary trees with n leaves
$\mathbb{P}_{\mathrm{YH}}(\mathcal{E})$	probability of the event $\mathcal{E}$ under the YH process
$\mathbb{P}_{\mathrm{RTP}}(\mathcal{E})$	probability of the event $\mathcal{E}$ under the RTP process
$\mathbb{P}_{\mathrm{RTP}'}(\mathcal{E})$	probability of the event $\mathcal{E}$ under the modified RTP process
C_n	number of cherries in a YH tree
C_n^*	number of cherries in a RTP tree

Chapters 4 and 5

T_X or T	rooted tree on X
T_n	rooted tree with n leaves
Y_X or Y	unrooted tree on X
$T^{-\rho}$	unrooted tree induced from the rooted tree T
$\varphi(n)$	number of rooted binary trees with n taxa
$\mathcal{C}(\mathcal{T})$	collection of clades of all X-trees

$\mathbb{P}_{\mathrm{YH}}(\mathcal{E})$	probability of event $\mathcal{E}$ under the YH process
$\mathbb{P}_{\mathrm{K}}(\mathcal{E})$	probability of event $\mathcal{E}$ under the Kingman process
$\mathbb{P}_{\mathrm{YHK}}(\mathcal{E})$	probability of event $\mathcal{E}$ under the YHK process
$\mathbb{P}_{\mathrm{PDA}}(\mathcal{E})$	probability of event $\mathcal{E}$ under the PDA model
$p_n(A)$	probability that a subset A of X, $n = \|X\|$, is a clade of a YHK tree
$p_n(A, B)$	probability that the subsets A and B of X, $n = \|X\|$, are clades of a YHK tree
$q_n(A)$	probability that a subset A of X, $n = \|X\|$, is a clan of an unrooted YHK tree
$q_n(A, B)$	probability that subsets A and B of X, $n = \|X\|$, are clans of an unrooted YHK tree
$p'_n(A)$	probability that a subset A of X, $n = \|X\|$, is a clade of a PDA tree
$p'_n(A, B)$	probability that the subsets A and B of X, $n = \|X\|$, are clades of a PDA tree
$q'_n(A)$	probability that a subset A of X, $n = \|X\|$, is a clan of an unrooted PDA tree
$q'_n(A, B)$	probability that the subsets A and B of X, $n = \|X\|$, are clans of an unrooted PDA tree

Chapter 6

T or T'	gene tree
W	species network
W'	species tree
λ_i	branch length in coalescent units
γ_j	probability that a lineage at a hybrid node is from the first parent node

Chapter 7

τ_i	branch length of branch i in number of generations
t_i	branch length of branch i in coalescent units
$N_i/2$	population size of branch i, assuming that the population size does not change within a branch
N_i	number of gene copies in branch i
μ	constant mutation rate per locus per generation
θ_i	mutation rate per generation of population i

Appendix B

Technical details

Justification of Lemma 4

Proof. At edge e, suppose that A and B partition X_n, where $n-1 \geq k \geq 1$, $|A| = k$ and $|B| = n-k$. Let $\{a, b, c\}$ be a subset of X_n of size three. Suppose that a new leaf x is to be attached to e. Let q be a split of $\{x, a, b, c\}$, such that $q = xc|ab$, $xa|bc$, $xb|ac$ with equal probabilities. Suppose that a and b are always on one side of e. We consider the following four cases:

$$\begin{cases} \text{Case I} & c \in B \text{ and } \{a, b\} \subseteq A; \\ \text{Case II} & \{a, b, c\} \subseteq B; \\ \text{Case III} & c \in A \text{ and } \{a, b\} \subseteq B; \\ \text{Case IV} & \{a, b, c\} \subseteq A. \end{cases}$$

We use Q_{I}, Q_{II}, Q_{III} and Q_{IV} to denote the set of quartet trees on the leaf set $\{x, a, b, c\}$ in Cases I, II, III and IV respectively, and let Q be the entire set of quartet trees for the leaf set $\{x, a, b, c\}$. Since the four cases are mutually exclusive, each Q_i partitions Q, $i \in \{\text{I}, \text{II}, \text{III}, \text{IV}\}$, and the size of each Q_i is $|Q_{\text{I}}| = \binom{k}{2} \times \binom{n-k}{1}$, $|Q_{\text{II}}| = \binom{n-k}{3}$, $|Q_{\text{III}}| = \binom{n-k}{2} \times \binom{k}{1}$ and $|Q_{\text{IV}}| = \binom{k}{3}$.

Let $w(e)$ be a random variable for the weight that is added to e for a quartet tree from $\{x, a, b, c\}$. Consider $w(e)$ for each case $\{\text{I}, \text{II}, \text{III}, \text{IV}\}$. Then we have:

- Cases I and III: $w(e) = \begin{cases} 1, & w.p.\ \frac{2}{3}; \\ 0 & w.p.\ \frac{1}{3}, \end{cases}$
- Cases II and IV: $w(e) = 0$.

Let $W_i(e)$, $i \in \{\text{I}, \text{II}, \text{III}, \text{IV}\}$, be the sum of all the weights added to the edge e. $W_{\text{I}}(e)$ is a binomial random variable with the parameters $\binom{k}{2}\binom{n-k}{1}$ and $\frac{2}{3}$; $W_{\text{III}}(e)$ is a binomial random variable with the parameters $\binom{n-k}{2}\binom{k}{1}$ and $\frac{2}{3}$; $W_{\text{II}} = W_{\text{IV}} = 0$. Let

$W_n(e)$ be the sum of $W_i(e)$, so that $W_n(e) = W_{\mathrm{I}}(e) + W_{\mathrm{III}}(e)$. Let $n_1 = \binom{k}{2}\binom{n-k}{1}$, and $n_2 = \binom{n-k}{2}\binom{k}{1}$. Then we have:

$$n_1 + n_2 = \frac{k(n-k)(n-2)}{2},$$

and so $W_n(e)$ consists of this many independent trials with a probability of success on each trial of $\frac{2}{3}$. That is, $W_n(e)$ is a binomial random variable with the parameters $\frac{k(n-k)(n-2)}{2}$ and $\frac{2}{3}$.

□

Justification of Inequality (3.1)

Let E_n^{P} denote the set of pendant edges of the current X_n-tree T_n, and let E_n^{I} be the set of interior edges.

Lemma 18. *For any $e'' \in E_n^{\mathrm{P}}$ and any $e' \in E_n^{\mathrm{I}}$, the expected pendant edge total weight $W_n(e'')$ and the expected interior edge total weight $W_n(e')$ satisfy the inequality:*

$$\mathbb{E}\left[W_n(e')\right] - \mathbb{E}\left[W_n(e'')\right] \geq \frac{1}{3}\left[n^2 - 5n + 6\right] > 0. \tag{B.1}$$

Proof. $W_n(e'')$ and $W_n(e')$ are binomial random variables with the same probability of success, $\frac{2}{3}$, but a different number of trials $\binom{n-1}{2}$ and $\frac{k(n-k)(n-2)}{2}$, where $k \in \{2, \ldots, n-2\}$. Thus we have:

$$\mathbb{E}[W_n(e'')] = \frac{2}{3}\binom{n-1}{2}, \qquad \mathbb{E}[W_n(e')] = \frac{2}{3}\frac{k(n-k)(n-2)}{2}.$$

For a fixed n, $\mathbb{E}\left[W_n(e')\right] - \mathbb{E}\left[W_n(e'')\right]$ is a function of k. Therefore, to find the minimum difference between these two expected values, we need to find the value(s) of k for which $\mathbb{E}\left[W_n(e')\right] - \mathbb{E}\left[W_n(e'')\right]$ is minimal.

Let $y = (n-2)(n-k)k - (n^2 - 3n + 2)$, then $\frac{dy}{dk} = (n-2)(n-2k)$. When $k = \frac{n}{2}$, $\frac{dy}{dk} = 0$, $\frac{d^2y}{dk^2} < 0$. Thus, there is a maximum at $k = \frac{n}{2}$, and minimum values occur at $k = 2$ or $k = n-2$. Therefore, when $k = 2$ or $k = n-2$:

$$\frac{1}{3}\left[n^2 - 5n + 6\right] \leq \mathbb{E}\left[W_n(e')\right] - \mathbb{E}\left[W_n(e'')\right].$$

Moreover, it is easily shown that for $n > 3$, $\frac{1}{3}\left[n^2 - 5n + 6\right] > 0$. Therefore

$$\mathbb{E}\left[W_n(e')\right] - \mathbb{E}\left[W_n(e'')\right] \geq \frac{1}{3}\left[n^2 - 5n + 6\right] > 0.$$

□

Theorem 18. *For any $e'' \in E_n^{\mathrm{P}}$ and any $e' \in E_n^{\mathrm{I}}$, we have:*

$$\mathbb{P}\left(W_n(e'') \geq W_n(e')\right) \leq 2\exp(-\frac{1}{576}n).$$

Proof. Let $W_n'' = W_n(e'') - \mathbb{E}\left[W_n(e'')\right]$,
$W_n' = W_n(e') - \mathbb{E}\left[W_n(e')\right]$, and $\beta = \mathbb{E}\left[W_n(e')\right] - \mathbb{E}\left[W_n(e'')\right]$. By Lemma 18, for $n \geq 4$, $\beta \geq 2dn^2$, where $d = \frac{1}{48}$.

Now, we have:

$$\begin{aligned}
\mathbb{P}\left(W_n(e'') \geq W_n(e')\right) =& \mathbb{P}\left(W_n'' - W_n' \geq \beta\right) \\
\leq& \mathbb{P}\left(W_n'' \geq \frac{\beta}{2} \text{ or } -W_n' \geq \frac{\beta}{2}\right) \\
\leq& \mathbb{P}\left(W_n'' \geq \frac{\beta}{2}\right) + \mathbb{P}\left(-W_n' \geq \frac{\beta}{2}\right) \\
\leq& \mathbb{P}\left(W_n'' \geq dn^2\right) + \mathbb{P}\left(-W_n' \geq dn^2\right).
\end{aligned}$$

We now apply Hoeffding's inequality to the two terms on the right. Suppose that $\{Y_i, i = 1, 2, 3, ..., N\}$ are independent Bernoulli random variables, and let $Y = \sum_{i=1}^{N} Y_i$. By Hoeffding's inequality (Hoeffding, 1963), we have:

$$\mathbb{P}\left(Y - \mathbb{E}(Y) \geq t\right) \leq \exp\left(-2t^2/N\right),$$

$$\mathbb{P}\left(-(Y - \mathbb{E}(Y)) \geq t\right) \leq \exp\left(-2t^2/N\right).$$

Taking $Y = W_n'$ (and W_n''), $t = dn^2$, and $N = \frac{k(n-k)(n-2)}{2}$ in the previous string of inequalities, gives:

$$\begin{aligned}
\mathbb{P}\left(W_n(e'') \geq W_n(e')\right) &\leq 2\exp(-\frac{1}{576\frac{k}{n}(1-\frac{k}{n})(1-\frac{2}{n})}n), \\
&\leq 2\exp(-\frac{1}{576}n).
\end{aligned}$$

□

Justification of Inequality (3.2)

Proof. We will use Theorem 18 to establish Inequality (3.2). For $e'' \in E_n^{\mathrm{P}}$ and $e' \in E_n^{\mathrm{I}}$, let D be the event that $\min_{e'' \in E_n^{\mathrm{P}}}\{W_n(e'')\} < \min_{e' \in E_n^{\mathrm{I}}}\{W_n(e')\}$.

Consider the complement of the event D,

$$D^c = \left(\min_{e \in E_n^{\mathrm{P}}} \{W_n(e'')\} < \min_{e' \in E_n^{\mathrm{I}}} \{W_n(e')\} \right)^c,$$

such that: $W_n(e') < \min_{e'' \in E_n^{\mathrm{P}}} \{W_n(e'')\}$, $W_n(e') \leq W_n(e'')$, for all $e'' \in E_n^{\mathrm{P}}$. Let $A_{e'',e'}$ be the event that $W_n(e'') > W_n(e')$. Then we have $D^c \subseteq \bigcup_{(e'',e') \in P \times I} A_{e'',e'}$, and so

$$\mathbb{P}\left(D^c\right) \leq \mathbb{P}\left(\bigcup_{(e'',e') \in P \times I} A_{e'',e'} \right).$$

According to Boole's inequality, we have:

$$\mathbb{P}\left(\bigcup_{(e'',e') \in P \times I} A_{e'',e'} \right) \leq \sum_{(e'',e') \in P \times I} \mathbb{P}(A_{e'',e'}). \tag{B.2}$$

Now, the number of pendant edge is n, i.e. $|P| = n$, and the number of interior edges is $n-3$, i.e. $|I| = n-3$. Thus, $|P \times I| = n(n-3)$, and so, by Theorem 18, $\mathbb{P}(A_{e'',e'}) = \mathbb{P}\left(W_n(e'') \geq W_n(e')\right) \leq 2\exp(-\frac{1}{576}n)$. Thus we have:

$$\sum_{(e'',e') \in P \times I} \mathbb{P}(A_{e'',e'}) \leq n(n-3)2\exp(-\frac{1}{576}n) \leq 2n^2 \exp(-\frac{1}{576}n). \tag{B.3}$$

Therefore:

$$\mathbb{P}\left(\min_{e'' \in E_n^{\mathrm{P}}} \{W_n(e'')\} \leq \min_{e' \in E_n^{\mathrm{I}}} \{W_n(e')\} \right) \geq 1 - 2n^2 \exp(-\frac{1}{576}n).$$

□

Justification of Inequality (3.3)

Proof. Since $\dfrac{k^2 \exp(-ck)}{\exp(-ck/2)} = k^2 \exp(-ck/2)$, and $k^2 \exp(-ck/2) \leq 1$ for $c \geq \dfrac{4\log k}{k}$ and $k > 1$, we have:

$$k^2 \exp(-ck) \leq \exp(-\frac{c}{2}k).$$

Thus $\sum_{k=m}^{\infty} k^2 \exp(-ck) \leq \sum_{k=m}^{\infty} \exp(-\frac{c}{2}k)$, where $c \geq \dfrac{4\log k}{k}$ and $k > 1$. Since $\sum_{k=m}^{\infty} \exp(-\frac{c}{2}k)$ is the sum of a geometric series:

$$\sum_{k=m}^{\infty} \exp(-\frac{c}{2}k) = \frac{\exp(-cm/2)}{1 - \exp(-c/2)}.$$

For $m \geq m_0$, $\exp(-cm/2) \leq \exp(-cm_0/2)$. Therefore, $\sum_{k=m}^{\infty} k^2 \exp(-ck) \leq \frac{\exp(-cm_0/2)}{1-\exp(-c/2)}$.

□

Justification of Lemma 7

Proof. We use induction to show Lemma 7.

For the trivial case $n = 2$, the maximal clades of T_2 are the leaves, which have size 1, i.e. $1 \leq u \leq 1$, $u = 1$, which leads the probability in Equation (4.6) to be equal to 1.

Suppose that for $n \leq k$, $\mathbb{P}_n(U = u) = \begin{cases} \frac{1}{n-1}, & u \in \{1, 2, \ldots, n-1\}, \\ 0, & \text{otherwise.} \end{cases}$ is true.

For $u = 1$, since $\mathbb{P}_k(U = 1) = \frac{1}{k-1}$, $\mathbb{P}_{k+1}(U = 1) = \mathbb{P}_k(U = 1)\frac{k-1}{k} = \frac{1}{k}$.

For $1 < u \leq n-1$:

$$\begin{aligned} \mathbb{P}_{k+1}(U = u) &= \mathbb{P}_k(U = u)\frac{k-u}{k} + \mathbb{P}_k(U = u-1)\frac{u-1}{k} \\ &= \frac{1}{k}. \end{aligned}$$

□

Justification of Lemma 8

Proof. Let U be a discrete uniform random variable on 1 to $n-1$, that is:

$$\mathbb{P}(U = i) = \begin{cases} \frac{1}{n-1}, & i \in \{1, 2, \ldots, n-1\}; \\ 0, & \text{otherwise.} \end{cases}$$

Let V be a random variable, such that $V = \min\{U, n-U\}$. For $j \in \{1, \ldots, \frac{n}{2}\}$:

$$\begin{aligned} \mathbb{P}(V = j) &= \mathbb{P}(U = j \text{ or } n - U = j) \\ &= \mathbb{P}(U = j) + \mathbb{P}(U = n-j) - \mathbb{P}(U = n - U = j) \\ &= \mathbb{P}(U = j) + \mathbb{P}(U = n-j) - \mathbb{P}(U = j = \frac{n}{2}) \\ &= \frac{1}{n-1} + \frac{1}{n-1} - \begin{cases} 0, & n \text{ is odd,} \\ \frac{1}{n-1}, & n \text{ is even,} \end{cases} \\ &= \begin{cases} \frac{2}{n-1}, & 1 \leq j < \frac{n}{2}; \\ \frac{1}{n-1}, & j = \frac{n}{2}. \end{cases} \end{aligned}$$

□

Justification of Lemma 10

Proof. $X_n(a)$ is the number of proper clades of size a in a random YHK tree. Let Z_1 and Z_2 be random variables such that: $Z_1(a) = \begin{cases} 1, & a = r, \\ 0, & \text{otherwise} \end{cases}$; and $Z_2(a) = \begin{cases} 1, & a = n - r, \\ 0, & \text{otherwise.} \end{cases}$ We now have $X_n(a) = X_r(a) + X_{n-r}(a) + Z_1(a) + Z_2(a)$. We can combine $Z_1(a)$ and $Z_2(a)$ as $X_n(a) = X_r(a) + X_{n-r}(a) + Z(a)$, where

$$Z(a) = \begin{cases} 1, & a = r \text{ or } a = n - r; \\ 2, & a = r = \dfrac{n}{2}; \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathcal{E}_r$ be the event that the maximal clades are of size $r, n - r$, where $r \leq \dfrac{n}{2}$. By probability theory, we have:

$$\mathbb{E}[\omega] = \sum_{\alpha} \mathbb{E}[\omega|\mathcal{E}_\alpha]\mathbb{P}[\mathcal{E}_\alpha].$$

Let $\omega = X_n(a)$, and let $\mathcal{E}_\alpha$ be the event $\mathcal{E}_r$. We have:

$$\mathbb{E}[X_n(a)] = \sum_{r \geq 1}^{\lfloor \frac{n}{2} \rfloor} \mathbb{E}[X_n(a)|\mathcal{E}_r]\mathbb{P}[\mathcal{E}_r],$$

and:

$$\mathbb{E}[X_n(a)] = \mathbb{E}[X_r(a) + X_{n-r}(a) + Z(a)].$$

By the linearity of the expected value, the expected values of the sum of random variables is the sum of the expected value of these random variables:

$$\begin{aligned} \mathbb{E}[X_n(a)|\mathcal{E}_r] &= \mathbb{E}[(X_r(a) + X_{n-r}(a) + Z(a))|\mathcal{E}_r] \\ &= \mathbb{E}[X_r(a)|\mathcal{E}_r] + \mathbb{E}[X_{n-r}(a)|\mathcal{E}_r] + \mathbb{E}[Z(a)|\mathcal{E}_r], \end{aligned}$$

where $E[Z(a)|\mathcal{E}_r] = \begin{cases} 1, & a = r \text{ or } a = n - r; \\ 2, & a = r = \dfrac{n}{2}; \\ 0, & \text{otherwise.} \end{cases}$

When $n = 1$, the Yule tree T_n is just a single point and there are no strict clades other than the root. Therefore, $X_1(a)$ is always 0 and the expected value is zero.

When $n = 2$, the strict clade size can only be one, since $1 \leq a \leq 2 - 1$. There exist two maximal clades of size one. Previously, we found $X_1(1) = 0$, and $\mathbb{E}[X_1(1)] = 0$, so $\mathbb{E}[X_2(1)] = 2$.

When $n = 3$, there is a maximal clade of size one, and the other has size two. There is only one possible topology for an unlabelled T_3, so $\mathbb{E}[X_3(1)] = 3$, and $\mathbb{E}[X_3(2)] = 1$.

Suppose that when $n = k - 1$, we have:

$$\mathbb{E}[X_n(a)] = \begin{cases} \dfrac{2(k-1)}{a(a+1)}, & 1 \le a \le k-1; \\ 0, & a > k-1. \end{cases}$$

For $n = k$, since $1 \le r \le n - r \le n - 1$, we have:

$$\mathbb{E}[X_r(a)] = \begin{cases} \dfrac{2(r)}{a(a+1)}, & 1 \le a \le r-1, \\ 0, & a > r-1 \end{cases};$$

and:

$$\mathbb{E}[X_{n-r}(a)] = \begin{cases} \dfrac{2(n-r)}{a(a+1)}, & 1 \le a \le n-r-1, \\ 0, & a > n-r-1. \end{cases}$$

When $a < \dfrac{n}{2}$ and n is odd, we have:

$$\begin{aligned} \mathbb{E}[X_n(a)] &= \frac{2}{n-1}\left[\sum_{r=1}^{\frac{n-1}{2}} \mathbb{E}[X_{n-r}(a)] + \sum_{r=a+1}^{\frac{n-1}{2}} \mathbb{E}[X_r(a)] + 1\right] \\ &= \frac{2}{n-1}\left[\frac{2n(\frac{n-1}{2}) - 2\sum_{r=1}^{\frac{n-1}{2}} r}{a(a+1)} + \frac{2\sum_{r=a+1}^{\frac{n-1}{2}} r}{a(a+1)} + 1\right] \\ &= \frac{2n}{a(a+1)}. \end{aligned}$$

When $a < \dfrac{n}{2}$, and n is even, we have:

$$\begin{aligned} \mathbb{E}[X_n(a)] &= \frac{2}{n-1}\left[\sum_{r=1}^{\frac{n}{2}-1} \mathbb{E}[X_{n-r}(a)] + \sum_{r=a+1}^{\frac{n}{2}-1} \mathbb{E}[X_r(a)] + 1\right] + \frac{1}{n-1}\left[\frac{2(n-\frac{n}{2})}{a(a+1)} + \frac{2\frac{n}{2}}{a(a+1)}\right] \\ &= \frac{2n}{a(a+1)}. \end{aligned}$$

When $a = \frac{n}{2}$ and n is even, we have:

$$\begin{aligned}\mathbb{E}[X_n(a)] &= \frac{2}{n-1}\left[\sum_{r=1}^{\frac{n}{2}-1}\mathbb{E}[X_{n-r}(a)]\right] + \frac{1}{n-1}\times 2 \\ &= \frac{2}{n-1}\left[\sum_{r=1}^{\frac{n}{2}-1}\mathbb{E}[X_{n-r}(a)] + 1\right] \\ &= \frac{2n}{\frac{n}{2}(\frac{n}{2}+1)} \\ &= \frac{2n}{a(a+1)}.\end{aligned}$$

When $a > \frac{n}{2}$ and n is even or odd, we have:

$$\begin{aligned}\mathbb{E}[X_n(a)] &= \frac{2}{n-1}\left[\sum_{r=1}^{\lfloor\frac{n}{2}\rfloor}\mathbb{E}[X_{n-r}(a)] + 0 + 1\right] \\ &= \frac{2}{n-1}\left[\sum_{r=1}^{n-a-1}\mathbb{E}[X_{n-r}(a)] + 0 + 0 + 1\right] \\ &= \frac{2n}{a(a+1)}.\end{aligned}$$

□

Further justification of Corollary 3

Proof.

Case 1. It is clear that $p_n(A,B) = p_n(A) = p_n(B) = p$, and for $0 < p < 1$, we have $p > p^2$. Thus $\rho_n(A,B) > 1$.

Cases 2 and 3:

$$\frac{p_n(A,B)}{p_n(A)p_n(B)} = \frac{\frac{4n}{a(a+1)(b+1)}\binom{n}{b}^{-1}\binom{b}{a}^{-1}}{\frac{2n\times 2n}{(a(a+1)b(b+1)}\binom{n}{a}^{-1}\binom{n}{b}^{-1}} = \frac{bn(n-1)\cdots(n-a+1)}{nb(b-1)\cdots(b-a+1)},$$

where, $\frac{n-1}{b-1} > 1, \cdots, \frac{n-a+1}{b-a+1} > 1$. Thus $\rho_n(A,B) > 1$.

Case 4:

$$\begin{aligned}\frac{p_n(A,B)}{p_n(A)p_n(B)} &= \frac{\frac{2}{n-1}\binom{n}{a}^{-1}}{\frac{2n}{a(a+1)}\binom{n}{a}^{-1}\frac{2n}{(n-a)(n-a+1)}\binom{n}{n-a}^{-1}} \\ &= \frac{1}{n-1}\binom{n}{a}\left(\frac{2n\times n}{(n-a)(n-a+1)a(a+1)}\right)^{-1} \\ &= \frac{n(n-1)(n-2)\cdots(n-a+1)(a+1)a(n-a)(n-a+1)}{[a(a-1)!]\times 2n\times n\times(n-1)} \\ &= \binom{n-2}{a-1}\frac{(a+1)(n-a+1)}{2n}\end{aligned}$$

For $a \geq 2$ and $n-a \geq 2$, we have $n \geq 4$. Thus $\binom{n-2}{a-1} \geq 2$.
Since $(a+1)(n-a+1) = an - a^2 + a + n - a + 1 = (an - a^2 + 1) + n > n$, this implies that $\dfrac{(a+1)(n-a+1)}{n} > 1$, and thus $\rho_n(A,B) > 1$.

Case 5:

$$\begin{aligned}\frac{p_n(A,B)}{p_n(A)p_n(B)} &= \frac{4n\binom{n-b}{a}^{-1}\binom{n}{b}^{-1}\frac{n(a+b)(a+b+1)(a+b-1)-(a+b-1)(a+b)[a(a+1)+b(b+1)+ab]+ab(a+1)(b+1)}{(a+b)(a+b+1)(a+b-1)ab(a+1)(b+1)}}{\frac{2n\times 2n}{ab(a+1)(b+1)}\binom{n}{a}^{-1}\binom{n}{b}^{-1}} \\ &= \frac{1}{n}\binom{n-b}{a}^{-1}\binom{n}{a}\frac{n(a+b)(a+b+1)(a+b-1)+G(a,b)}{(a+b)(a+b+1)(a+b-1)} \\ &= \frac{n(n-1)\cdots(n-a+1)a!}{(n-b)(n-b-1)\cdots(n-b-a+1)a!}\frac{n(a+b)(a+b+1)(a+b-1)+G(a,b)}{n(a+b)(a+b+1)(a+b-1)} \\ &= \frac{n(n-1)\cdots(n-a+1)}{(n-b)(n-b-1)\cdots(n-b-a+1)}\frac{n(a+b)(a+b+1)(a+b-1)+G(a,b)}{n(a+b)(a+b+1)(a+b-1)}.\end{aligned}$$

where $G(a,b) = -(a+b-1)(a+b)[a(a+1)+b(b+1)+ab] + ab(a+1)(b+1)$
For $a \geq 2$, $b \geq 2$ and $n \geq (a+b+1)$, $\rho_n(A,B) > 1$.

□

Appendix C

hybrid_coal user manual

C.1 Introduction

In phylogenetic studies, trees are used for describing evolutionary histories. In particular, a species tree presents population divergences, and a gene tree indicates the times when genes started to differentiate within populations. Even though speciation is driven by gene mutations, using a single gene tree to infer the species tree is not ideal. Often, the inconsistency among gene trees and species trees makes describing the relationships between and among species very difficult. Common causes of the conflict include gene duplication, horizontal gene transfer, incomplete lineage sorting, and hybridization (Holland *et al.*, 2008; Meng and Kubatko, 2009). If speciation events occur close together, it is likely that some gene copies remain the same after species divergence. This inconsistency between gene trees and species trees is referred as incomplete lineage sorting.

Hybridization refers to interbreeding between species. Offspring who carry genes from both parental species then reproduce and form a new species. For closely related species, however, both lineage sorting and hybridization are likely to occur, e.g. an avian genus *Manacus* (Brumfield and Carling, 2010) or the New Zealand alpine cicadas (Buckley *et al.*, 2006). Here, I present research on the probabilistic modelling of coalescence with lineage sorting in hybridized species. In these models, the relationships among species are represented by a network rather than a tree, while relationships at the gene level are still represented by trees.

`hybrid_coal` has been developed to calculate the gene tree probabilities within a species network.

C.2 Download and installation

`hybrid_coal` can be downloaded from https://code.google.com/p/hybrid-coal/. Extract the source code by executing the following command:

```
tar -xf hybrid_coal-VERSION.tar.gz.
```

It is fairly standard to compile `hybrid-Lambda` on UNIX-like systems. In the directory `hybrid_coal-VERSION`, execute the following command:

```
$./bootstrap
$make
```

Note: The command `make doxygen-run` will generate `HTML` documentation of the source code.

C.3 Notation

C.3.1 Coalescent parameters

Under the coalescent process, the waiting time for lineages to coalesce is an exponential random variable. The Kingman coalescent process only allows two lineages to coalesce at a time. Thus, the mean waiting time for b lineages to coalesce into $b-1$ lineages is $\binom{b}{2}$ per unit of time.

C.3.2 Input/output formats

The input file for `hybrid_coal` is a character string that describes the relationships among species. Standard Newick format (Olsen, 1990) is used for inputting species trees and outputting gene trees, e.g.:

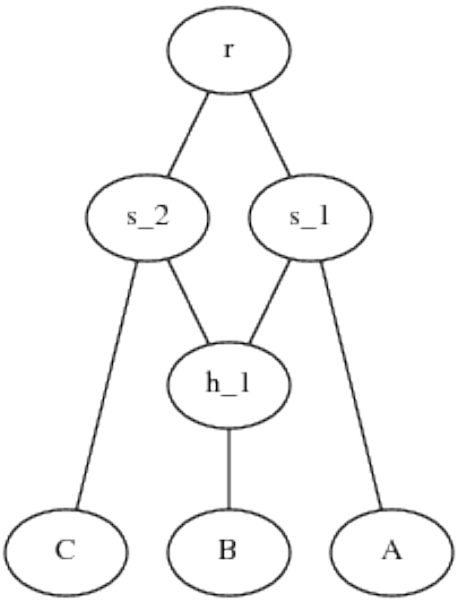
$$((A:t_A, B:t_B):t_{AB}, C:t_C), \tag{C.1}$$

where t_i denotes the branch length from i to its parent node in coalescent units.

Extended Newick formatted strings (Cardona *et al.*, 2008; Huson *et al.*, 2010) label all internal nodes, and are used for inputting species networks. In the network string, the descendants of a hybrid node are recorded before the hybrid node the first time the hybrid node occurs; otherwise, it is written as a tip node. For example:

$$(((B:t_B)h\#\gamma:t_h^{s1}, A:t_A)s1:t_{s1}, (h\#\gamma:t_h^{s2}, C:t_C)s2:t_{s2})r, \tag{C.2}$$

where # identifies the hybrid node.



At a hybrid node, lineages travel to either parent node with given probabilities. The parameter γ denotes the probability of that the lineage goes to the first parent node. Since the hybrid node has two parent nodes, the branch length needs to be specific, i.e. t_i^j denotes the branch length from i to j in coalescent units.

C.3.3 Method

The coalescent model of Degnan and Salter (2005) is extended to obtain the distribution of gene trees (T) in a given network (W). The network W is initially reduced to a set of simpler networks ($SG(W)$) in a single step of the reduction process. By standard probability theory, we have the following:

$$P(T|W) = \sum_{w^* \in SG(W)} P(T|W^* = w^*)P(W^* = w^*|W)$$

When each network w^* is obtained from W by removing an edge or a node, we will apply the recursion to reduce the networks on W^* until all the simpler network structures are tree-like. The problem of obtaining gene tree probabilities from species trees has already been solved by Degnan and Salter (2005). The approach we have outlined will therefore reduce the probability of a gene tree, given a species network, into a linear combination of gene tree probabilities, given species trees.

C.4 Commands:

C.4.1 Generating a list of all gene tree topologies in a taxa set

```
hybrid_coal -sp INPUT1 -gtopo
```

```
hybrid_coal -sp INPUT1 -gtopoF OUTPUT
```

`INPUT1` is a Newick formatted string (see Section C.3.2), which does not have to be a binary tree. For example, to generate a gene tree topology for the taxon set $\{A, B, C\}$:

```
hybrid_coal -sp '(A,B,C)r;'-gtopo.
```

The default setting will save the gene tree topologies to the file `GENE_topo`. The file name can be specified via the option `-gtopoF`.

```
(A,C),B);
(A,(B,C));
((A,B),C);
```

To generate gene tree topologies for multiple lineages for the same species, e.g. for the taxon set $\{A, B, C\}$ with two lineages of species A, use the command:

```
hybrid_coal -sp '(A_1,A_2,B,C)r;' -gtopoF A2BC.
```

```
(((A_1,C),B),A_2);
((A_1,(B,C)),A_2);
(((A_1,B),C),A_2);
((A_1,B),(A_2,C));
(((A_1,B),A_2),C);
((A_1,C),(A_2,B));
(A_1,((A_2,C),B));
(A_1,(A_2,(B,C)));
(A_1,((A_2,B),C));
((A_1,(A_2,B)),C);
(((A_1,C),A_2),B);
```

```
((A_1,(A_2,C)),B);
(((A_1,A_2),C),B);
((A_1,A_2),(B,C));
(((A_1,A_2),B),C);
```

C.4.2 Calculating gene tree probabilities of a given species network

```
hybrid_coal -sp INPUT1 [-gt INPUT2] [-out OUTPUT]
```

INPUT1 is a(n) (extended) Newick formatted string (see Section C.3.2), which can be entered through the command line or a text file.

The flags -gt and INPUT2 are optional. INPUT2 is a Newick formatted string of a gene tree topology, which can be entered through the command line or from a text file, where users can specify several gene trees. If gene trees are not specified, hybrid_coal will generate all possible gene tree topologies and then compute the probabilities. For example:

```
hybrid_coal -sp '((A:1,B:1):1,C:2)r;'.
```

will print the following message:

```
   1  ((A,C),B)   0.122626
   2  (A,(B,C))   0.122626
   3  ((A,B),C)   0.754747
       Total         1
Species Input: ((A:1,B:1):1,C:2)r;
Species structure: ((A:1,B:1):1,C:2)r;
Total probability: 1
Gene tree probabilities produced in file: out_coal.
```

The gene tree probabilities are saved in the file out_coal by default. Users can specify the filename via the option -out.

C.4.3 Generating Maple script for the gene tree probabilities

hybrid_coal can also generate Maple script to calculate the gene tree probabilities. The option -symb enables users to calculate the symbolic probabilities of the gene trees for analytic work. By default, the Maple script is saved in the file maple_prob.mw. Users can specify the filename via the option -mapleF.

```
hybrid_coal -sp INPUT1 [-gt INPUT2] -maple [-symb].

hybrid_coal -sp INPUT1 [-gt INPUT2] -mapleF OUTPUT.
```

C.4.4 Generating coalescent histories for the gene tree probabilities

hybrid_coal can also generate extensive LaTeX code for users to study the coalescent history of a gene tree within a network.

```
hybrid_coal -sp INPUT1 [-gt INPUT2] -latex.

hybrid_coal -sp INPUT1 [-gt INPUT2] -latexF OUTPUT.
```

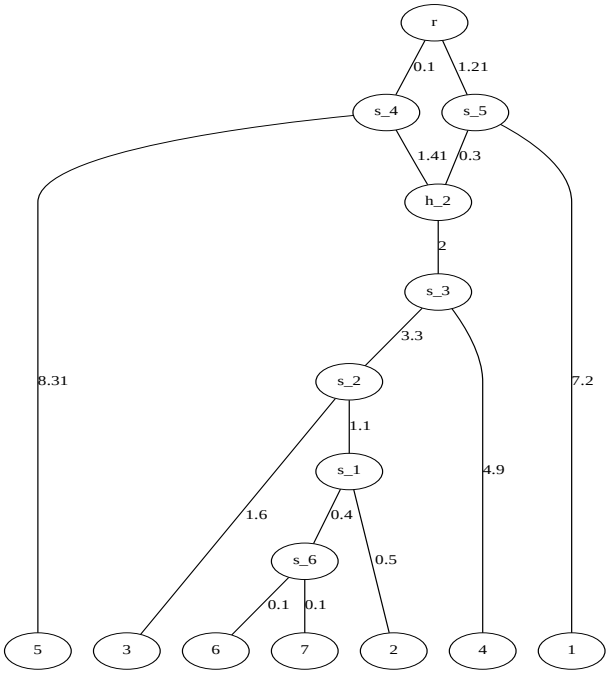
C.4.5 Commands for other features:

Plot

```
hybrid_coal -sp INPUT -dotF OUTPUT [-branch].
```

`hybrid_coal` uses the program `dot` to generate figures. The option `[-branch]` will label the branch lengths in the figure, e.g.:

```
hybrid_coal -sp trees/7_tax_sp_nt1_para -dot -branch.
```



If the option `-dot` is used instead of `-dotF OUTPUT`, the figure will be saved in the file `figure.pdf` by default.

Alternatively, by replacing `-dotF` with `plotF`, `hybrid_coal` can generate LaTeX code for plotting a network/tree. If `-plot` is used instead of `-plotF OUTPUT`, LaTeX code will be saved in the file `texfigure.tex` by default.

C.5 Summary of command line options

`-h` or `-help`	Help. List the following content.
`-sp INPUT`	Input the species network/tree string through the command line or from a file. Branch lengths of the `INPUT` are in coalescent units.
`-gt INPUT`	Input the gene tree string through the command line or from a file.
`-latex` / `-latexF`	Generate the coalescent history of a gene tree within a species network.
`-maple`/ `-mapleF`	Generate a `Maple` executable script file to calculate the gene tree probabilities of given species networks.

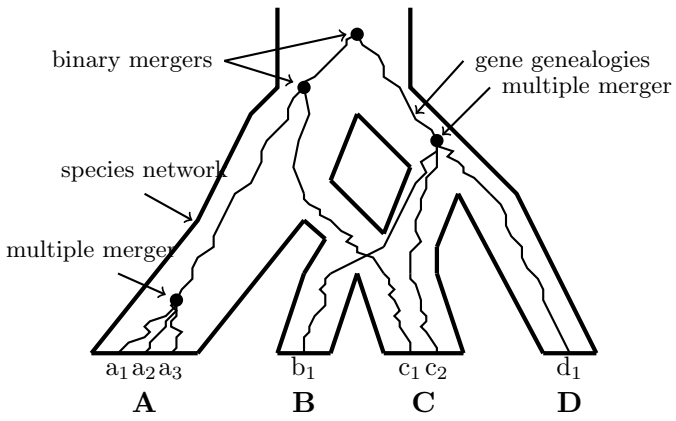
`-symb`	Enable the `Maple` script to calculate the symbolic gene tree probabilities.
`-gtopo` / `-gtopoF`	Generate the gene tree topologies of a given set of taxa.
`-plot/-dot [option]`	Use LaTeX(`-plot`) or `Dot` (`-dot`) to draw the input (defined by `-sp`) network/tree.
`-plotF/-dotF FILE`	Generated figure will be saved in `FILE`.

Appendix D

hybrid-Lambda user manual

D.1 Introduction

`hybrid-Lambda` is a software package that can simulate gene trees within a rooted species network or a rooted species tree under the coalescent process. The main feature of this program is that users can choose to use the standard Kingman coalescent process, which produces bifurcating genealogies, or two other Λ-coalescent processes, which produce multifurcating genealogies. The other feature is that `hybrid-Lambda` uses extended Newick formatted strings to make it easier to represent hybridization events between species.



D.2 Download and installation

`hybrid-Lambda` can be downloaded from https://code.google.com/p/hybrid-lambda/. Extract the source code by executing the following command:

```
tar -xf hybrid-Lambda-VERSION.tar.gz.
```

It is fairly standard to compile `hybrid-Lambda` on UNIX-like systems. In the directory `hybrid-Lambda-VERSION`, execute the following command:

```
$./bootstrap
$make
```

Note: The command `make doxygen-run` will generate `HTML` documentation of the source code.

D.3 Notation

D.3.1 Coalescent parameters

Under the coalescent process, the waiting time for lineages to coalesce is an exponential random variable. The Kingman coalescent process allows only two lineages to coalesce at a time. Thus the mean of the waiting time for b lineages to coalesce into $b-1$ lineages is $\binom{b}{2}$ per unit of time. However, for the Λ-coalescent, if the coalescent parameter ψ is between 0 and 1 (Eldon and Wakeley, 2006), the rate λ_{bk} at which k out of b active ancestral lineages merge is:

$$\lambda_{bk} = \binom{b}{k} \psi^k (1-\psi)^{b-k}, \quad \psi \in (0,1). \tag{D.1}$$

If the coalescent parameter α is between 1 and 2, the rate is:

$$\lambda_{bk} = \binom{b}{k} \frac{B(k-\alpha, b-k+\alpha)}{B(2-\alpha, \alpha)}, \quad \alpha \in (1,2), \tag{D.2}$$

where $B(\cdot,\cdot)$ is the beta function (Schweinsberg, 2003).

D.3.2 Input/output formats

The input file for `hybrid-Lambda` is a character string that describes the relationships among species. Standard Newick format (Olsen, 1990) is used for inputting species trees and outputting gene trees, e.g.:

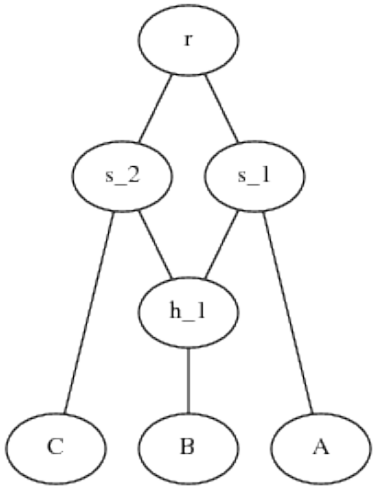
$$((A:t_A, B:t_B):t_{AB}, C:t_C), \tag{D.3}$$

where t_i denotes the branch length from i to its parent node. `hybrid-Lambda` uses the values of t_i to assign parameters for different inputs. In Expression (D.3), the interior nodes are not labelled. However, this is not essential. `hybrid-Lambda` can also recognise input Newick string whose nodes are all labelled.

Extended Newick formatted strings (Cardona *et al.*, 2008; Huson *et al.*, 2010) label all internal nodes, and are used for inputting species networks. In the network string, the descendants of a hybrid node are recorded before the hybrid node the first time the hybrid node appears, otherwise is written as a tip node. For example,

$$(((B:t_B)h\#\gamma:t_h^{s1}, A:t_A)s1:t_{s1}, (h\#\gamma:t_h^{s2}, C:t_C)s2:t_{s2})r, \tag{D.4}$$

where # identifies the hybrid node.



At a hybrid node, lineages travel to either parent nodes with given probabilities. The parameter γ denotes the probability of that the lineage goes to the first parent node. Since the hybrid node has two parent nodes, the branch length needs to be specific, i.e. t_i^j denotes the branch length from i to j.

Normally, standard Newick formatted and extended Newick formatted strings do not include branch lengths at the root node. However, in this program, this is required, as the input strings assign population size or the coalescent parameter at the root. Thus, Expressions (D.3) and (D.4) are associated with Expressions (D.5) and (D.6) respectively:

$$((A : t_A, B : t_B) : t_{AB}, C : t_C) : t_{root}, \tag{D.5}$$

$$(((B : t_B)h\#\gamma : t_h^{s1}, A : t_A)s1 : t_{s1}, (h\#\gamma : t_h^{s2}, C : t_C)s2 : t_{s2})r : t_r. \tag{D.6}$$

D.4 Commands for simulation:

D.4.1 Simulating gene trees

```
hybrid-Lambda -spcu INPUT [-num N] [-seed SEED] [-gF OUTPUT-FILE].
```

`INPUT` is a(n) (extended) Newick formatted string (see Section D.3.2), which can be entered through the command line or from a text file. If the input is followed by the flag `-spcu`, its branch lengths must be in coalescent units. The value `N` following the flag `-num` is the number of gene trees simulated. Users can specify a random seed for simulation by declaring it after `-seed`. By default, the branch lengths of the output trees are in coalescent units. They are saved in the file `GENE_TREE_coal_unit`. The flag `-gF` enables the users to name the output files. For example:

```
hybrid-Lambda -spcu '((1:1,2:1):1,3:2);' -num 3 -seed 2 -gF example1 -log
```

will print the following message:

```
Default Kingman coalescent on all branches
Random Seed  2  used
Produced gene tree files:
example1_coal_unit
3 trees simulated.
```

The following gene trees are saved in the file `example1_coal_unit`:

```
(1_1:2.98119,(3_1:2.55301,2_1:2.55301):0.428181);
(3_1:6.66739,(2_1:1.06869,1_1:1.06869):5.5987);
(3_1:2.38966,(2_1:1.0722,1_1:1.0722):1.31746);
```

D.4.2 Gene tree output options and user-defined mutation rate

```
hybrid-Lambda -spcu INPUT [-gF OUTPUT-FILE] [-mu MU] [option]
```

By default, the mutation rate $\mu = 0.00005$ is used. The flag `-mu` makes it possible for users to define a constant mutation rate. Moreover, the options in `[ ]` enable more manipulations of the output gene trees. These options include:

`-sim_mut_unit`	Convert the simulated gene tree branch lengths to mutation units.
`-sim_num_gener`	Convert the simulated gene tree branch lengths to number of generations.
`-sim_num_mut`	Simulate the number of mutations on each branch of the simulated gene trees.
`-sim_Si_num`	Generate the file `out_table`, which includes the number of segregating sites and the total branch length of the gene tree in coalescent units.

For example, suppose the input network string in the file `4_tax_sp_nt1_para` is

```
((((B:1,C:1)s1:1)h1#.5:1,A:3)s2:1,(h1#.5:1,D:3)s3:1)r.
```

```
hybrid-Lambda -spcu trees/4_tax_sp_nt1_para -gF example2 -num 2 -mu 0.00003
                        -sim_mut_unit -sim_num_mut
```

will generate the following files:

```
$ cat example2_coal_unit
((B_1:1.9099,C_1:1.9099):2.82957,(A_1:4.05317,D_1:4.05317):0.686292);
((D_1:3.77974,(C_1:1.2291,B_1:1.2291):2.55064):0.369812,A_1:4.14956);
$ cat example2_mut_unit
((B_1:0.57297,C_1:0.57297):0.848871,(A_1:1.21595,D_1:1.21595):0.205888);
((D_1:1.13392,(C_1:0.36873,B_1:0.36873):0.765192):0.110944,A_1:1.24487);
$ cat example2_num_mut
((B_1:1,C_1:1):2,(A_1:1,D_1:1):0);
((D_1:0,(C_1:1,B_1:1):0):0,A_1:3); .
```

D.4.3 User-defined population sizes

```
hybrid-Lambda -spcu INPUT-1 -pop INPUT-2
```

By the default setting, the population sizes for each species are assumed to be equal and unchanged at any time, which is 10,000. This can be reassigned to other constant values followed by `-pop`. As a result, the branch lengths of the gene trees in number of generations will change. This can be observed though the option `-sim_num_gener`. For example, to simulate gene trees within a species network/tree with a population size of 25,000, we use the following:

```
hybrid-Lambda -spcu INPUT -num N -pop 25000 -sim_num_gener.
```

This command will also produce gene trees for which the branch lengths are in number of generations, saved in the file `GENE_TREE_num_gener`.
Note: The population size refers to the number of gene copies, not the number of individuals.

Instead of inputting a species network with branch lengths in coalescent units, input strings can have branch lengths representing the number of generations. Moreover, in the following example, we demonstrate that if the population sizes are assumed to vary, the input strings in Expression (D.5) can specify the population sizes on all branches and the root.

```
hybrid-Lambda -spng '(A:50000,B:50000)r;' -pop '(A:50000,B:50000)r:40000;'.
```

D.4.4 Simulating multiple samples per species

```
hybrid-Lambda -spcu INPUT -S n1 n2 ...
```

`hybrid-Lambda` sorts the taxa names in a particular order. At each character of a taxon name, it sorts:

- numerics in ascending order,
- letters in alphabetical order,
- numerics then letters,
- upper-case letters then lower-case letters.

To sample multiple individuals, the order of the sample sizes needs to follow the order of the taxa names.

For example:

```
hybrid-Lambda -spcu '((((A:1.1,B:1.1):2.1,a:2.2):1.1,13D:.2):.3,4:.3);'
                                -S 2 4 3 6 5 .
```

The order of the taxon names is `13D`, `4`, `A`, `B` and `a`. Thus, the program will sample 2 individuals in taxon `13D`, four samples from taxon `4`, three samples from taxon `A`, six samples from taxon `B` and five samples from taxon `a`.

D.4.5 Simulating gene trees with multiple merger coalescents

```
hybrid-Lambda -spcu INPUT-1 -mm INPUT-2
```

The Kingman coalescent is assumed by default. To use the Λ-coalescent, the coalescent parameter needs to be specified after `-mm`. For details, see Equation (D.1) and (D.2). Moreover, similar to assigning particular population sizes on branches (see the examples in Section D.4.3), coalescent parameters can be specified as well. In this case, to assume the Kingman coalescent within some population, the multiple merger parameter needs to be set to 2. For example:

```
hybrid-Lambda -spcu '(A:1,B:1)r;' -mm '(A:1.9,B:.2)r:2;' -S 3 4.
```

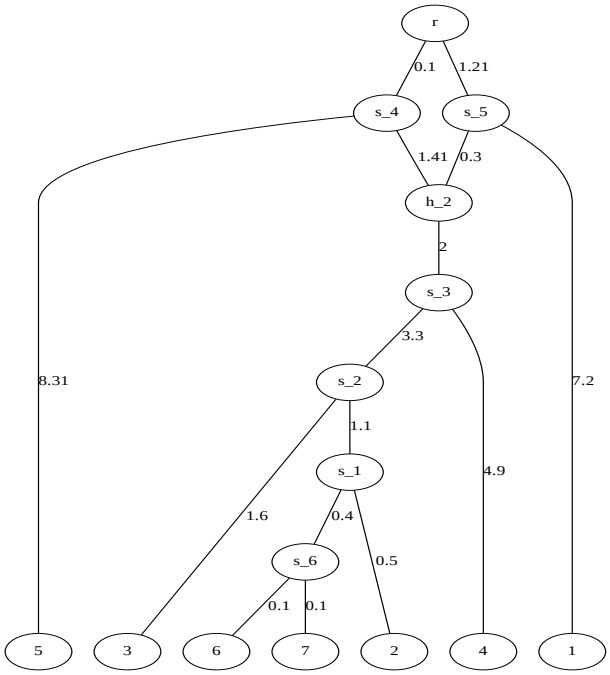
D.4.6 Commands for other features:

Plot

```
hybrid-Lambda -spcu INPUT -dotF OUTPUT [-branch]
```

`hybrid-Lambda` uses the program `dot` to generate figures. The option `[-branch]` will label the branch lengths in the figure, e.g.:

```
hybrid-Lambda -spcu trees/7_tax_sp_nt1_para -dot -branch
```



If the option `-dot` is used instead of `-dotF OUTPUT`, the figure will be saved in the file `figure.pdf` by default.

Alternatively, by replacing `-dotF` with `plotF`, `hybrid-Lambda` can generate LaTeX code for plotting a network or tree. If `-plot` is used instead of `-plotF OUTPUT`, the LaTeX code will be saved in the file `texfigure.tex` by default.

Analysing the frequencies of gene trees

`hybrid-Lambda` can generate a topology frequency table for the simulated gene trees by:

```
hybrid-Lambda -spcu INPUT -num N -fF OUTPUT .
```

```
hybrid-Lambda -gt INPUT -fF OUTPUT
```

reads trees in the file `INPUT`, and generates a topology frequency table in the file `OUTPUT`. If the option `-f` is used instead of `-fF OUTPUT`, the analysed frequency table will be saved in the file `freq_out` by default.

Simulating and analysing the monophyly topology of the gene trees

```
hybrid-Lambda -spcu INPUT-1 -S n1 n2 -mono [-mm INPUT-2]
```

Recent studies on the shape of genealogies between two species investigated the probabilities of monophyletic taxa (Eldon and Degnan, 2012; Rosenberg, 2003). The option `-mm` generates a frequency table of the gene trees whose taxa are monophyletic in one population, both populations, or are paraphyletic and polyphyletic. For example:

```
hybrid-Lambda -spcu '(A:5,B:5)r;' -mono -num 100 -mm .1 -S 4 4 -log
```

will print the following message:

```
   A mono     B mono Recip mono     A para     B para  Polyphyly
     0.02       0.01           0       0.02       0.01       0.97
Random Seed  1342238826  used
Produced gene tree files:
GENE_TREE_coal_unit
100 trees simulated.
```

D.5 Summary of command line options

`-h` or `-help`	Help. List the following content.
`-spcu INPUT`	Input the species network/tree string through the command line or from a file. Branch lengths of the `INPUT` are in coalescent units.
`-spng INPUT`	Input the species network/tree string through the command line or from a file. Branch lengths of the `INPUT` are in number of generations.
`-pop INPUT`	Population sizes are defined by a single numerical constant, or a string which specifies the population size on each branch. The string can be inputted through the command line or from a file. By default, the population size 10,000 is used.
`-mm INPUT`	Multiple merger parameters are defined by a single numerical constant, or a string which specifies the parameter on each branch. The string can be inputted through the command line or from a file. By default, the Kingman coalescent is used.
`-S n1 n2 ...`	Specify the number of samples for each taxon.
`-num N`	The number of gene trees to be simulated.
`-seed SEED`	User defined random `SEED`.
`-mu MU`	User defined constant mutation rate μ. By default, a mutation rate of 0.00005 is used.
`-gF FILE [option]`	Specify the filename for the simulated gene trees.
`-sim_mut_unit`	Convert the simulated gene tree branch lengths to mutation units.
`-sim_num_gener`	Convert the simulated gene tree branch lengths to number of generations.
`-sim_num_mut`	Simulate the number of mutations on each branch of the simulated gene trees.

`-f`	Generate a topology frequency table for a set of input trees or simulated gene trees. The frequency table is saved in the file `freq_out` by default.
`-fF FILE`	The topology frequency table will be saved in `FILE`.
`-gt FILE`	Specify the file to analyse tree topology frequencies.
`-tmrca [FILE]`	Return the TMRCA of the gene trees. TMRCA file is saved in "tmrcaFILE" by default, user can define the TMRCA file name by option FILE.
`-log [FILE]`	Enable the log function. Log file is saved in "LOG" by default, user can define the log file name by option FILE.
`-mono`	Generate a frequency table of monophyletic, paraphyletic and polyphyletic trees.
`-plot/-dot [option]`	Use LaTeX(`-plot`) or `Dot` (`-dot`) to draw the input (defined by `-spcu`) network (tree).
`-branch`	Branch lengths will be labelled in the figure.
`-plotF/-dotF FILE`	The generated figure will be saved in `FILE`.
`-plotF/-dotF FILE`	The generated figure will be saved in `FILE`.

Acknowledgements

Funding: New Zealand Marsden Fund (Sha Zhu and James Degnan), Engineering and Physical Sciences Research Council (Bjarki Eldon). This work was partly conducted while JD was a Sabbatical Fellow at the National Institute for Mathematical and Biological Synthesis, an institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Award #EF-0832858, with additional support from The University of Tennessee, Knoxville.

Included files: C++ Mersenne twister pseudo-random number generator (Matsumoto and Nishimura, 1998).