Combinatorial and probabilistic methods in biodiversity theory

A thesis submitted in
partial fulfilment
of the requirements for
the Degree of
Doctor of Philosophy in Mathematics
by
Beáta Faller

Supervisors:
Professor Mike Steel and
Associate Professor Charles Semple

University of Canterbury
Department of Mathematics and Statistics
2010

To my father, Ferenc Faller.
24 October 1954 - 20 February 2010

Abstract

Phylogenetic diversity (PD) is a measure of species biodiversity quantified by how much of an evolutionary tree is spanned by a subset of species. In this thesis, we study optimization problems that aim to find species sets with maximum PD in different scenarios, and examine random extinction models under various assumptions to predict the PD of species that will still be present in the future.

Optimizing PD with Dependencies is a combinatorial optimization problem in which species form an ecological network. Here, we are interested in selecting species sets of a given size that are ecologically viable and that maximize PD. The NP-hardness of this problem is proved and it is established which special cases of the problem are computationally easy and which are computationally hard. It is also shown that it is NP-complete to decide whether the feasible solution obtained by the greedy algorithm is optimal. We formulate the optimization problem as an integer linear program and find exact solutions to the largest food web currently in the empirical literature. In addition, we give a generalization of PD that can be used for example when we do not know the true evolutionary history. Based on this measure, an optimization problem is formulated. We discuss the complexity and the approximability properties of this problem.

In the generalized field of bullets model (g-FOB), species are assumed to become extinct with possibly different probabilities, and extinction events are independent. We show that under this model the distribution of future phylogenetic diversity converges to a normal distribution as the number of species grows. When extinction probabilities are influenced by some binary character on the tree, the state-based field of bullets model (s-FOB) represents a more realistic picture. We compare the expected loss of PD under this model to that under the associated g-FOB model and find that the former is always greater than or equal to the latter. It is natural to further generalize the s-FOB model to allow more than one binary character to affect the extinction probabilities. The expected future PD obtained for the resulting trait-dependent field of bullets model (t-FOB) is compared to that for the associated g-FOB model and our previous result is generalized.

Acknowledgements

First and foremost, I would like to thank Mike Steel for dropping out of the train in Mannheim to meet me and for inviting me to visit his group at the University of Canterbury in New Zealand. I express my deepest gratitude to my supervisors Charles Semple and Mike Steel for offering me a PhD position with an exciting research project, for their ongoing support and guidance throughout the different stages of the PhD process, and for their motivation, inspiration, and valuable contributions to my work.

I am grateful to my further collaborators Travis Ingram, Fabio Pardi, and Dominic Welsh, all of whom have been great to interact with and to learn from. I would like to thank Magnus Bordewich, Stefan Grünewald, Peter Lockhart, Karen Magnusson-Ford, Arne Mooers, Elchanan Mossel, Sebastien Roch, Raazesh Sainudiin, Michael Snook, and Alexander Zelikovsky for useful discussions relating to the work presented in the thesis.

I express my appreciation to my examiners Mike Atkinson, Vincent Moulton, and Mike Steel for spending time on my thesis and for their valuable comments, and to Peter Humphries for proof reading various parts of this thesis. I am grateful to my Masters supervisor Tamás Móri for his continuous encouragement and support.

I would like to thank the New Zealand Marsden Fund, the Allan Wilson Centre for Molecular Ecology and Evolution, and the Mathematics and Statistics Department at the University of Canterbury for their generous funding, and thank Charles and Mike once again for finding me this funding.

Many thanks to the postgraduate students and all the people and visitors at the Mathematics and Statistics Department for their assistance and for making my time at work so enjoyable. It has been great to be part of such a supportive and friendly community. Special thanks to Simone Linz, Dietrich Radel, and Meghan Williams Tripp.

Finally, I extend my heartfelt thanks to my family and friends, who have graciously coped with all of my absences and have helped and encouraged me throughout my studies. I am especially grateful to Mária and Axel Graf, Ferenc Faller, Éva Kricskovics (Oma), Barbara Faller and Péter Szepesi, Szabolcs Pivarcsi, Rozália and Mátyás Rutterschmidt, and István Sibalin (Steve).

Contents

Chapter 1

Introduction

The current rapid rate of extinction of many diverse species has focused attention on predicting and maximizing future biodiversity. There are numerous ways to measure the biodiversity of a group of species, and one which recognizes the evolutionary linkages between species is phylogenetic diversity (PD). It was introduced by Dan Faith in 1992 with the aim of measuring and protecting biodiversity [20]. Namely, the high rates of recent species extinctions and the limited resources available for conservation urge the need for placing conservation priorities on different species. Faith proposed to use phylogenetic diversity as a selection criteria for preserving biodiversity (for example, feature diversity) based on phylogenetic information. Briefly, given a subset of taxa, the phylogenetic diversity of that subset is the sum of the evolutionary distances of the edges of the phylogenetic tree that connects this subset. Here, the distance assigned to an edge may refer to the amount of genetic change on that edge, its temporal duration, or perhaps other features such as morphological diversity.

Under various possible interpretations of the edge lengths, PD has been widely used for quantifying and maximizing present and expected future biodiversity [9, 16, 21, 47, 48, 49, 54, 63, 70, 80]. For example, the Noah's Ark problem [34, 55, 78] attempts to maximize expected future phylogenetic diversity by allocating resources that increase the survival probabilities in a constrained way.

This thesis develops new combinatorial and computational insights, algorithms, and stochastic models that can be applied to predict and maximize the PD score of future species. The problems studied can be grouped into two classes: combinatorial optimization problems and probabilistic models. Accordingly, the thesis has been divided into two parts.

The organization of the thesis is as follows. We first give some preliminaries in Chapter 2, where we define fundamental notions that are used in both parts of the thesis. Definitions that are used only in one of the two parts are introduced either in the introductory chapter of the appropriate part or where they are needed.

After describing the most fundamental concepts of the thesis in the Preliminaries, we study PD maximization problems in Part I. This part starts with Chapter 3, which introduces the terminology used only in Part I. In Chapter 4, the computational complexity of the problem Optimizing PD with Dependencies and its special cases is analyzed. We show that the problem is computationally hard in general and that polynomial-time instances of it have a very special structure. It is also established in this chapter that it is NP-complete to decide whether the feasible solution obtained by the greedy algorithm can be improved. This is followed by Chapter 5, which uses the approach of integer linear programming to find exact solutions to real instances of this problem. The last chapter of Part I deals with a maximization problem that is based on the more general biodiversity measure called phylogenetic diversity for cluster systems. We show that for this problem a polynomial-time greedy algorithm always produces a solution whose value is at least $1 - e^{-1}$ times the value of an optimal solution. It is also proved that no polynomial-time algorithm can achieve an approximation ratio higher than this value unless P=NP. Our proofs are based on the fact that the new biodiversity measure is a submodular set function and on the notion of an approximation preserving reduction called L-reduction.

In Part II, probabilistic methods are used to develop new mathematical results concerning three species extinction models. We start this part with a chapter in which we introduce the discussed models and define the terminology used only in Part II. This is followed by Chapters 8, 9, and 10, which present our main results on the three models. In Chapter 8 we prove that, under the generalized field of bullets model (g-FOB), the distribution of future PD is asymptotically normal. We also describe an algorithm to derive the exact distribution. Chapter 9 introduces the more realistic state-based field of bullets model (s-FOB) and compares the expected biodiversity loss under this model to that under the simpler g-FOB model. In the proof of this result, we first use the classical four functions theorem to establish a generic inequality that applies to two-state Markov processes on trees. This new inequality, combined with the FKG inequality, is used to prove the above relationship. Interestingly, our new Markov inequality also allows us to derive a purely combinatorial result concerning the parsimony score of binary characters on trees. In the last chapter, our most realistic extinction model is described and the expected value and variance of the resulting future PD are compared with the corresponding values of the g-FOB model. In order to show a relationship between the first moments similar to the one above, we generalize our two-state inequality using a form of the four functions theorem for finite distributive lattices. We then apply the new Markov inequality, alongside a general version of the FKG inequality, to prove the required relationship.

From this thesis, three publications have been published or accepted [25, 26, 71], another publication has been submitted [27], and one further paper is in preparation [24].

Moreover, there is a chapter in the thesis presenting unpublished research. This is Chapter 6, in which we prove three main results. The first two describe the approximability properties of the problem Optimizing PD for Cluster Systems. As there is a strong relationship between this problem and the problem called Max k-Cover, results on the approximability of one of these problems immediately carry over to the other, and vice versa. Therefore, the result stated in Theorem 6.1(i) was first established in [15], while Theorem 6.1(ii) follows immediately from a result of Feige [28]. In Chapter 6 of this thesis, we describe our proof for part (i) and also show how [28] can be used to derive the statement in part (ii).

Finally, it should be noted that most of the results in [71] were established by Mike Steel and have been reproduced for completeness.

Chapter 2

Preliminaries

In biology, phylogenetic trees and their rooted counterparts are used to describe the evolutionary relationships between taxa (for example, species). Briefly, a phylogenetic tree is a tree whose leaves are labelled by the elements of a finite set, which represents a set of present-day taxa. If such a tree is rooted, then it is regarded as describing the evolution of the species that label the leaves of the tree from a common hypothetical ancestral species at the root. The other interior vertices of the tree correspond to further hypothetical ancestral species or, alternatively, to past speciation events. In biology, rooted phylogenetic trees are also called evolutionary trees or cladograms. We now formally define these concepts.

2.1 Phylogenetic trees

Unless otherwise stated, X denotes a non-empty finite set in this thesis. Most of the notation and terminology of the thesis follows [65].

Definition 2.1. *A* phylogenetic X-tree $\mathcal{T}$ *is an ordered pair* (T, ϕ)*, where* $T = (V_T, E_T)$ *is a tree with no degree-two vertices, and* ϕ *is a bijection from* X *into the leaf set of* T*.*

Definition 2.2. *A* rooted phylogenetic X-tree $\mathcal{T}$ *is an ordered pair* (T, ϕ)*, where* $T = (V_T, E_T)$ *is a rooted tree in which the root has degree at least two and all the other interior vertices have degree at least three, and* ϕ *is a bijection from* X *into the set of leaves of* T*.*

Figure 2.1 (a) illustrates a phylogenetic X-tree, while Figure 2.1 (b) shows an example of a rooted phylogenetic X-tree.

A (rooted) phylogenetic X-tree is also called a *(rooted) phylogenetic tree on* X or, if there is no ambiguity, a *(rooted) phylogenetic tree.*

Two phylogenetic X-trees $\mathcal{T}_1 = (T_1, \phi_1)$ and $\mathcal{T}_2 = (T_2, \phi_2)$, with $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$, are *isomorphic* if there exists a bijection $\psi\colon V_1 \to V_2$ that induces a bijection from E_1 to E_2 and satisfies $\phi_2 = \psi \circ \phi_1$. The isomorphism of rooted phylogenetic X-trees is defined analogously, with the additional requirement that ψ takes the root of T_1 to the root of T_2. We regard isomorphic phylogenetic trees as being equivalent.

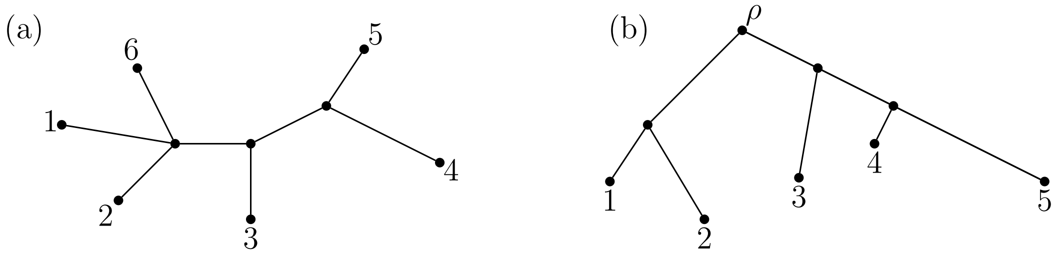


Figure 2.1: (a) A phylogenetic X-tree with $X = \{1, 2, 3, 4, 5, 6\}$. (b) A rooted phylogenetic X-tree with root ρ and with $X = \{1, 2, 3, 4, 5\}$.

Let $\mathcal{T}$ be a (rooted) phylogenetic X-tree (T, ϕ). The tree T is called the *underlying tree of* $\mathcal{T}$ and ϕ is called the *labelling map of* $\mathcal{T}$. The set X is called the *label set of* $\mathcal{T}$. For convenience, we often refer to the vertices and edges of T as the vertices and edges of $\mathcal{T}$ respectively. Accordingly, we often write $V_{\mathcal{T}}$ for V_T and $E_{\mathcal{T}}$ for E_T. If $e \in E_{\mathcal{T}}$ and e is incident with a leaf, e is called a *pendant edge* of $\mathcal{T}$. A non-pendant edge in $E_{\mathcal{T}}$ is called an *interior edge* of $\mathcal{T}$. Similarly, a non-leaf vertex $v \in V_{\mathcal{T}}$ is called an *interior vertex* of $\mathcal{T}$. Furthermore, we can view X as the set of leaves of the underlying tree T, and so we denote the leaves of T by the elements of X since ϕ is implicitly determined.

In some of the chapters, we regard a rooted phylogenetic tree as a directed tree by viewing each edge as an arc directed away from the root. In this case, the orientation of the edges corresponds to a temporal ordering.

For a detailed description of the properties of phylogenetic trees, we refer the reader to [65].

2.2 Phylogenetic diversity

In phylogenetics, the edges of a phylogenetic tree are frequently regarded just as pairs of vertices, without any lengths assigned to them. However, in several problems, we do assign a length or weight to each edge. This length refers to the amount of the genetic change on that edge, its temporal duration, or other features such as morphological diversity. Edge lengths play an important role when we aim to measure biological diversity based on phylogenetic information. Phylogenetic diversity (PD) is based on these evolutionary distances: the PD of a subset of the leaf set of a phylogenetic tree measures how much total genetic or evolutionary diversity in the tree is spanned by only the species in the subset [16, 20, 21, 47, 80].

Definition 2.3. *Let $\mathcal{T}$ be a phylogenetic X-tree and suppose that λ is a map that assigns* *a non-negative real-valued length* $\lambda(e)$ *to each edge* e *of* $\mathcal{T}$*. The* phylogenetic diversity *of a subset* Y *of* X*, denoted* $\mathrm{PD}_{(\mathcal{T},\lambda)}(Y)$*, is defined by*

$$\mathrm{PD}_{(\mathcal{T},\lambda)}(Y) = \sum_{e \in E_{\mathcal{T}(Y)}} \lambda(e),$$

where $E_{\mathcal{T}(Y)}$ *is the edge set of the minimal subtree of* $\mathcal{T}$ *connecting the leaves in* Y*.*

An example to represent this concept is shown in Figure 2.2 (a).

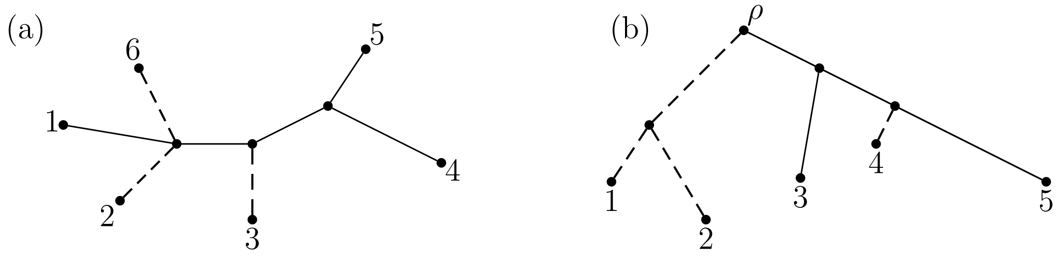


Figure 2.2: (a) The phylogenetic diversity of the subset $\{1, 4, 5\}$ of the leaf set of the tree on the left is the sum of the lengths of the solid edges. (b) The phylogenetic diversity of the set $\{3, 5\}$ in the rooted tree on the right is the sum of the lengths of the solid edges.

For subsets of the leaf set of a rooted phylogenetic tree, PD is defined as the sum of the edge lengths of the minimal subtree connecting the species in the subset and the root vertex [20]. We define this formally next and illustrate it in Figure 2.2 (b).

Definition 2.4. *Let* $\mathcal{T}$ *be a rooted phylogenetic* X*-tree with root vertex* ρ *and suppose that* λ *is a map that assigns a non-negative real-valued length* $\lambda(e)$ *to each edge* e *of* $\mathcal{T}$*. The* phylogenetic diversity *of a subset* Y *of* X *is*

$$\mathrm{PD}_{(\mathcal{T},\lambda)}(Y) = \sum_{e \in E_{\mathcal{T}(Y,\rho)}} \lambda(e),$$

where $E_{\mathcal{T}(Y,\rho)}$ *is the edge set of the minimal subtree of* $\mathcal{T}$ *connecting the leaves in* Y *and the root vertex* ρ*.*

Since it is clear in each section whether the trees under consideration are rooted or not, it is also clear which definition of phylogenetic diversity is meant by $\mathrm{PD}_{(\mathcal{T},\lambda)}$. Furthermore, we often denote $\lambda(e)$ by λ_e and, if there is no ambiguity, $\mathrm{PD}_{(\mathcal{T},\lambda)}(Y)$ by $\mathrm{PD}_{\mathcal{T}}(Y)$ or, more briefly, by $\mathrm{PD}(Y)$.

Both notions of phylogenetic diversity have a number of nice combinatorial properties. A summary of these can be found in [33] and [35].

Part I

Optimization problems

Chapter 3

Introduction to PD maximization problems

A basic question in conservation biology is how to maximize future biodiversity as species face extinction. When conservation decisions require prioritizing some species over others, one solution is to select a set of species with maximum phylogenetic diversity. The basic PD optimization problem aims to find a k-element subset of a given species set that has maximum PD among all such subsets. In Chapters 4 and 5, we consider the extension of this problem where we are only interested in selecting subsets of the taxa that are ecologically viable. Chapter 4 presents complexity theoretic results, while Chapter 5 deals with applications of the problem. In some situations, evolution is not tree-like; even if it is, we do not always know the true tree. The approximability properties of an optimization problem that can be used in these cases are discussed in Chapter 6.

Intractability is a central mathematical notion of the following chapters. We continue the present chapter by introducing some informal concepts from this area. For formal definitions see, for example, [32] and [41].

3.1 Introductory thoughts on intractability

We are often interested in finding efficient algorithms for solving a problem. In general, efficiency covers all the various computing resources needed to execute an algorithm. However, by efficient algorithms one frequently means fast algorithms, where time requirements are expressed in terms of the instance size.

The *time complexity function* of an algorithm expresses its time requirements by giving, for each possible input length, the largest amount of time needed by the algorithm to solve a problem instance of that size. A simple distinction that gives considerable insight into time complexity is between polynomial time algorithms and exponential time algorithms. We say that a function $f(n)$ is $O(g(n))$ whenever there exists a constant c such that $|f(n)| \leq c|g(n)|$ for all values of $n \geq 0$. A *polynomial-time algorithm* is one whose time complexity function is $O(p(n))$ for some polynomial function p, where n denotes the input length. Any algorithm that is not a polynomial-time algorithm is called an *exponential time algorithm*, even if the time complexity function is not an exponential function. The distinction between the two types of algorithms was first discussed in [13] and [18]. This distinction is central to our notion of intractability and to the theory of NP-completeness: we refer to a problem as *intractable* if there is no polynomial-time algorithm for solving it.

The earliest intractability results are the undecidability results of Alan Turing [74, 75], presented in 1936. In these papers, Turing demonstrated that there are problems that are so hard that no algorithm at all can be given for solving them. Such problems are called *undecidable*. A variety of other problems are today known to be undecidable. These problems are intractable in a particularly strong sense.

The first decidable intractable problems were found in the 1960s. These problems, and all the decidable intractable problems found since then, cannot be solved in polynomial time using even a non-deterministic computer. A non-deterministic computer model is able to pursue an unbounded number of independent computation sequences in parallel.

All the provably intractable problems known to date fall into these two categories. However, most of the apparently intractable problems are decidable and also solvable in polynomial time by non-deterministic computers. To prove intractability results for these problems, a new technique needs to be introduced.

Besides proving intractability, finding relationships between the difficulties of different problems is also one of the major goals of theoretical computer science. Giving a transformation that maps any instance of a problem into an equivalent instance of a second problem is used to reduce one problem to another. Such a reduction allows us to convert any algorithm that solves the second problem into an algorithm that solves the original one. Many examples of reductions were found in the 1960s, foreshadowing the results that were established later in the theory of NP-completeness.

The foundations of NP-completeness were laid in a paper by Stephen Cook, in 1971 [14]. In his paper, Cook emphasized the significance of polynomial-time reducibility, and focused attention on decision problems that can be solved in polynomial time by a non-deterministic computer model. These problems form the set NP. He proved that one problem in NP, the satisfiability problem, has the property that every other problem in NP can be reduced to it in polynomial time. Therefore, if the satisfiability problem can be solved in polynomial time, then so can every problem in NP, and if any problem in NP is intractable, then the satisfiability problem must be intractable too. This means that the satisfiability problem is one of the hardest problems in the class NP.

Following these findings, Richard Karp [39] presented a number of results proving that the decision problem version of many well-known combinatorial problems are as hard as the satisfiability problem. Since then, a large collection of other problems have been proved equivalent in difficulty to these problems. This equivalence class, consisting of the hardest problems in NP, is called the class of *NP-complete problems*. (We do not define the notion of this class formally here. However, as we use it throughout the thesis, we assume that the reader is familiar with the definition and with the techniques used to prove that a problem belongs to the class.)

With Cook's powerful ideas, the question of whether or not the NP-complete problems are intractable arose and is today considered to be one of the most important open questions of mathematics and computer science. The knowledge that a problem is NP-complete does not imply that it is intractable, but it definitely suggests that a major breakthrough will be needed to solve it in polynomial time.

The theory of NP-completeness is designed to be applied to decision problems only. However, it is possible to extend the notion of polynomial transformation to prove that other types of problems are at least as hard as NP-complete problems. The class of hard problems in this more general sense is called the class of NP-hard problems.

Finally, we note that today, numerous complexity classes are known, and a wide range of techniques has been established to prove hardness results and to cope with hard problems. However, there is still a large number of open questions waiting for resolution.

3.2 Computational problems

As mentioned above, *decision problems* are central in the theory of NP-completeness. Such problems have two possible solutions, either the answer 'yes' or the answer 'no'. The format we use for specifying decision problems consists of two parts: the first part specifies a *generic instance* of the problem in terms of sets, graphs, functions, numbers, and so on, and the second part states a *yes-no question* asked in terms of the generic instance.

A more general class of computational problems, called *search problems*, can be viewed as a collection of instances with a set of solutions for every instance. In the following chapters, we deal with optimization problems, which are special search problems. Informally, an *optimization problem* consists of a set of instances, a set of feasible solutions for each instance, a measure for each instance/feasible solution pair (also called the value of the feasible solution), and a goal function, which is either the function MIN or the function MAX. The set of *optimal solutions* for an instance is the set of feasible solutions whose value is optimal (minimum or maximum depending on whether the goal function is MIN or MAX).

For any optimization problem, there is a corresponding decision problem, each instance of which includes an additional numerical bound B. If the optimization problem aims to minimize the measure of the feasible solutions, then the question of the decision problem asks whether there is a feasible solution whose measure is no more than B. If the goal is to maximize the measure of the feasible solutions, then the question is whether there is a feasible solution whose value is at least B. It is important to note that an optimization problem is at least as hard as the corresponding decision problem. This fact is used several times in the first part of the thesis.

Chapter 4

Hardness of Optimizing PD with Dependencies

In the context of conservation biology, maximizing phylogenetic diversity is a prominent selection criteria for deciding which species to conserve (see, for example, [9, 20, 23, 48, 54, 63, 62, 70]). In its most direct application to conservation, one selects a k-element subset of species that maximizes PD over all k-element subsets. While PD makes a comparison between species to capture the notion of diversity, the conservation of individual species are considered in isolation. In real ecosystems this can be problematic as species frequently depend on other species for their survival—there is no point conserving a species if all the species it depends on go extinct [76, 81]. In this chapter, we consider an extension of the PD selection criteria, where only subsets that are ecologically viable are considered for conservation.

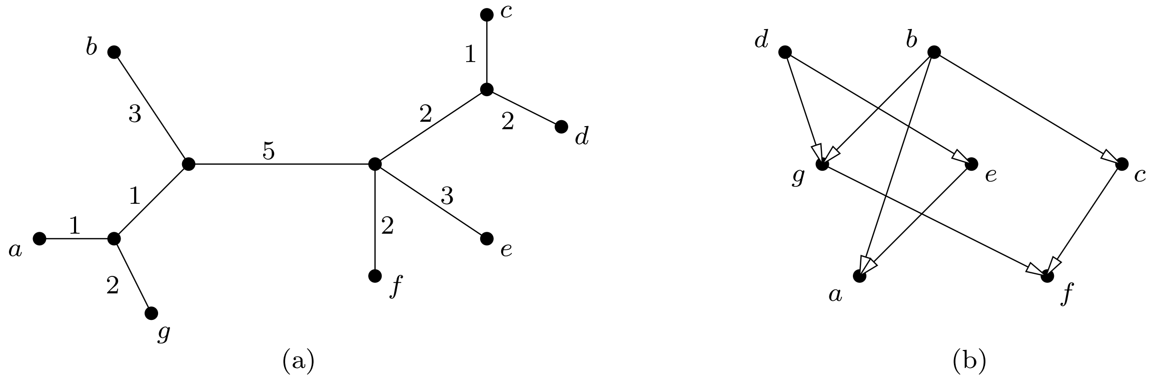
4.1 Problem definition

Given a phylogenetic X-tree $\mathcal{T}$ with non-negative real-valued weights on its edges and a fixed integer k, the *PD optimization problem* is to find

$$\max\{\mathrm{PD}(S) : S \text{ is a } k\text{-element subset of } X\}.$$

Pardi and Goldmann [54] and Steel [70] independently showed that a solution to this problem can be found in polynomial time using a greedy algorithm. An edge-weighted phylogenetic X-tree $\mathcal{T}$ with $X = \{a, b, c, d, e, f, g\}$ is shown in Figure 4.1(a). For this tree and for $k = 3$, the PD optimization problem has three optimal solutions: $\mathrm{PD}(\{b, d, g\}) = \mathrm{PD}(\{b, d, e\}) = \mathrm{PD}(\{g, d, e\}) = 15$.

To allow for ecological dependencies in the conserving of species, we extend the PD optimization problem to additionally include an acyclic directed graph $D = (X, A)$. Here, D could be an ecological network, for example a food web, where $(u, v) \in A$, for $u, v \in X$, precisely if taxon u feeds or preys on taxon v. We say that a subset S of X is *viable* if, for each $s \in S$, there is a directed path in D from s to a vertex with out-degree zero in which every vertex in the path is in S. The optimal solutions $\{b, d, g\}$, $\{b, d, e\}$, and $\{g, d, e\}$, obtained for the tree in Figure 4.1(a) and for $k = 3$, are not viable in the digraph D represented in Figure 4.1(b). Three-element viable subsets of the vertex set of D are,



(a)

(b)

Figure 4.1: (a) A phylogenetic X-tree $\mathcal{T}$ and (b) a food web D on X.

for example, $\{a, b, f\}$ and $\{a, e, d\}$. Under the food-web interpretation, a set S is viable if, for each taxon in S that is not at the bottom of the food chain, there is a taxon in S that it feeds or preys on. Formally, the problem we are interested in is the following:

Decision problem: Optimizing PD with Dependencies
Instance: A phylogenetic X-tree $\mathcal{T}$, a non-negative real-valued weighting λ on the edges of $\mathcal{T}$, an acyclic digraph $D = (X, A)$, a positive integer k, and a non-negative real number d.
Question: Is there a viable subset S of X of size at most k with $\mathrm{PD}(S) \geq d$?

As stated, this problem has been considered by Moulton *et al.* [48] and Spillner *et al.* [68]. The first paper was interested in the problem in the context of greedoids and greedy algorithms, while the second paper noted without proof that the problem was NP-complete. The purpose of the present study is to establish which variations of Optimizing PD with Dependencies are computationally easy and which variations of it are computationally hard. In addition to the immediate significance of knowing the complexity of the restricted problems, these results increase our knowledge of the essential elements which made the original problem NP-complete.

The next section contains some preliminaries that are used throughout the present chapter. A *star tree* is a (phylogenetic) tree with exactly one interior vertex. In Section 4.3, we show that Optimizing PD with Dependencies is NP-complete even if $\mathcal{T}$ is a star tree. Section 4.4 considers polynomial-time instances of Optimizing PD with Dependencies when $\mathcal{T}$ is a star tree. Such instances rely on the underlying graph of D containing no (undirected) cycles. An opposite extreme to consider is when $\mathcal{T}$ is arbitrary, but the underlying graph of D is a rooted tree. However, as we show in Section 4.5, this particular possibility is also NP-complete. For both intrinsic and practical reasons, greedy algorithms have been frequently considered in the context of phylogenetic diversity. A curious feature of the problem Optimizing PD with Dependencies which gives some additional indication of its hardness is highlighted in Section 4.6, where we show that it is NP-complete to decide if the feasible solution obtained by the greedy algorithm can be bettered. Throughout most of this chapter, we restrict ourselves to unrooted phylogenetic trees. However, in the last section, we consider the extension of our earlier results to rooted phylogenetic trees, including such trees satisfying the 'molecular clock hypothesis'.

4.2 Vertex Cover and the star tree problem

Vertex Cover is a classical NP-complete problem and is frequently used for completeness reductions. As we use Vertex Cover several times in this chapter, we give a formal definition of it here. Furthermore, we also describe a problem equivalent to Optimizing PD with Dependencies in case $\mathcal{T}$ is a star tree.

For a graph $G = (V, E)$, a *vertex cover* of G is a subset V' of V such that, for each edge $\{u, v\} \in E$, at least one of u and v belongs to V'. The NP-complete decision problem Vertex Cover [39] is the following:

Decision problem: Vertex Cover
Instance: A graph $G = (V, E)$ and a positive integer $m \leq |V|$.
Question: Is there a vertex cover of G of size at most m?

A special instance of Optimizing PD with Dependencies is when $\mathcal{T}$ is a star tree. Because a star tree contains no non-pendant edges, this special instance can be reformulated as a problem on acyclic digraphs. In particular, let w be the weighting on the vertices of D defined by setting $w(v) = \lambda(\{u, v\})$ for each $v \in X$, where u denotes the interior vertex of $\mathcal{T}$. In this setting, for any subset S of X, set $\mathrm{PD}(S) = \sum_{v \in S} w(v)$. It is now easily checked that the following decision problem is equivalent to Optimizing PD with Dependencies when $\mathcal{T}$ is a star tree and $k \geq 2$.

Decision problem: Optimizing PD in Vertex-Weighted Food Webs
Instance: An acyclic digraph $D = (X, A)$, a non-negative real-valued weighting w on the vertices of D, a positive integer k, and a non-negative real number d.
Question: Is there a viable subset S of X of size at most k with $\mathrm{PD}(S) \geq d$?

The above equivalence will be freely used several times in this chapter. Although the typical model of evolution is a bifurcating tree, there are instances for which it appears that the underlying model is more star-like than bifurcating (for example, see [79] and the references therein). Thus, restricting Optimizing PD with Dependencies to when $\mathcal{T}$ is a star tree is also of practical importance.

4.3 NP-completeness of Optimizing PD with Dependencies

In this section, we show that the decision problem Optimizing PD with Dependencies is NP-complete even if $\mathcal{T}$ is a star tree. In particular, recalling the equivalence in Section 4.2, we prove the following theorem.

Theorem 4.1. Optimizing PD in Vertex-Weighted Food Webs *is* NP-*complete.*

Proof. Evidently, Optimizing PD in Vertex-Weighted Food Webs is in NP since, given a subset S of X of size at most k, one can easily check in polynomial time if S is viable and $\mathrm{PD}(S) \geq d$. To complete the proof of the theorem, we show that there is a polynomial-time reduction from Vertex Cover to Optimizing PD in Vertex-Weighted Food Webs.

Given a graph $G = (V, E)$ and a positive integer m, we construct an instance of Optimizing PD in Vertex-Weighted Food Webs as follows. Let D be the acyclic digraph whose vertex set X is the (disjoint) union of V and E, and whose arc set A is defined to be

$$A = \{(e, v) : e \in E,\ v \in V,\ v \text{ is an end-vertex of } e \text{ in } G\}.$$

Let w be the weight function $w : X \to \mathbb{R}^{\geq 0}$ specified by assigning weight 1 to each vertex in $X \cap E$ and weight 0 to each vertex in $X \cap V$. Clearly, this construction can be accomplished in polynomial time.

We now show that there is a vertex cover of G of size at most m if and only if there is a viable subset S of X of size at most $|E| + m$ with $\mathrm{PD}(S) \geq |E|$. First, suppose that $V' \subseteq V$ is a vertex cover for G with $|V'| \leq m$. Then, the construction of D implies that $V' \cup E$ forms a viable subset of X. Since $|V' \cup E| = |V'| + |E| \leq m + |E|$ and $w(V' \cup E) = |E|$, it follows that $V' \cup E$ is a viable subset of X of size at most $|E| + m$ and with weight at least $|E|$. Conversely, suppose that there is a viable subset S of X of size at most $|E| + m$ and with weight at least $|E|$. Since the vertices in $X \cap V$ have weight 0, the subset S must contain all $|E|$ vertices with weight 1; that is, it must contain

$X \cap E$. Therefore, $S = V' \cup E$ for some $V' \subseteq V$. Since S is viable, there is an arc from each $e \in E$ to some vertex $v \in V'$. In terms of G, this implies that V' is a vertex cover of G. As $|S| \leq |E| + m$, it follows that $|V'| \leq m$, completing the proof of the theorem. □

Remark. Theorem 4.1 tells us that Optimizing PD with Dependencies is NP-complete even if $\mathcal{T}$ is a star tree. However, the proof of this theorem says the problem remains NP-complete if D is a bipartite digraph with vertex partition $V_1 \cup V_2$, where each vertex in V_1 has in-degree zero and out-degree two and each vertex in V_2 has out-degree zero. Moreover, it is also interesting to note that we could have used any restricted version of Vertex Cover for the reduction provided the version is NP-complete. For example, it has been shown that Vertex Cover remains NP-complete if G is cubic and planar [46]. A graph is *cubic* if each vertex has degree three. Thus, Optimizing PD with Dependencies remains NP-complete if $\mathcal{T}$ is a star tree and D is a bipartite graph as described above with the additional properties that each vertex in V_2 has in-degree three and D is planar. To see planarity, observe that D can be obtained by taking a planar drawing of the planar graph G, subdivide each edge of G, and, for each resulting vertex u, direct the incident edges away from u.

4.4 Star tree and food tree

In contrast to the NP-completeness results of this chapter, we have the following theorem.

Theorem 4.2. Optimizing PD in Vertex-Weighted Food Webs *can be solved in polynomial time if D is either*

(i) *a rooted tree with all arcs directed away from the root or*

(ii) *a rooted tree with all arcs directed towards the root.*

Theorem 4.2 is an immediate consequence of what appears to be a well-known dynamic programming algorithm for solving the following problem (see, for example, [43]). Let T be a rooted tree with root r and let k be a positive integer. Suppose that the vertices of T are assigned real-valued weights. The problem is to find a maximum-weighted subtree of T with root r and at most k vertices.

We briefly outline the dynamic programming algorithm here in the language of our problem. For convenience, we view Optimizing PD in Vertex-Weighted Food Webs as an optimization problem. First consider (ii). Let D be a digraph satisfying (ii) in the statement of Theorem 4.2 and let r be the root of D. Thus, D contains exactly one vertex of out-degree zero, namely r. Let v be a vertex of D. We denote the subset of vertices u of D for which (u, v) is an arc in D by $I(v)$. Furthermore, we denote the rooted subtree of D with root v whose vertex set is precisely the subset of vertices x of D for which there is a directed path from x to v by $D(v)$.

For a vertex v of T and a non-negative integer $q \leq k$, let $S(v, q)$ denote the value of an optimal solution of Optimizing PD in Vertex-Weighted Food Webs when D is chosen to be $D(v)$ and the size of the viable subset is at most q. Note that $S(r, k)$ denotes the value of an optimal solution for the original problem. Clearly, for any vertex v of T, we have $S(v, 0) = 0$ and, for each vertex u of in-degree zero, $S(u, q) = w(u)$ for $1 \leq q \leq k$. The dynamic programming algorithm starts at vertices of in-degree zero and works itself towards r using the recursion

$$S(v, q) = w(v) + \max_{\{q_u : \sum_{u \in I(v)} q_u \leq q-1\}} \sum_{u \in I(v)} S(u, q_u)$$

for $1 \leq q \leq k$. It is shown in [43] that this approach leads to a quadratic-time algorithm for finding $S(r, k)$.

If D is a digraph satisfying (i) in the statement of Theorem 4.2, then we simply modify the above algorithm in the obvious way to find a minimum-weight subtree of D rooted at r with at least $n - k$ vertices. The complement of the value of the resulting solution; that is, $\sum_{u \in V(D)} w(u)$ minus this value, gives the desired optimal value.

Despite the above positive results, we end this section with the following conjecture, where no constraints are placed on the direction of the arcs.

Conjecture 4.3. Optimizing PD in Vertex-Weighted Food Webs *when the underlying graph of D is a tree is NP-complete.*

4.5 Arbitrary phylogenetic tree and food tree

In this section, we show that Optimizing PD with Dependencies is still NP-complete if $\mathcal{T}$ is an arbitrary phylogenetic tree while D is a rooted tree. In particular, we establish the following theorem.

Theorem 4.4. Optimizing PD with Dependencies *when $\mathcal{T}$ is an arbitrary phylogenetic tree and D is either*

(i) *a rooted tree with all arcs directed away from the root or*

(ii) *a rooted tree with all arcs directed towards the root* *is NP-complete.*

Proof. We prove (i). The proof of (ii) is similar and omitted. Since Optimizing PD with Dependencies is in NP, this particular instance of the problem is also in NP. Like the NP-completeness proof for Theorem 4.1, the reduction is from Vertex Cover. However, for this proof, we use the restricted version of Vertex Cover in which the graph is cubic and planar. It is shown in [46] that Vertex Cover remains NP-complete under these restrictions.

Let $G = (V, E)$ be a cubic, planar graph. We construct an instance of the restricted version of Optimizing PD with Dependencies described by (i) as follows. Colour the edges of G with three colours $\{1, 2, 3\}$ such that no two edges incident with the same vertex receive the same colour. Due to a classic construction of Tait [73], this is equivalent to four-colouring the faces of a planar drawing of G, which can be done in quadratic time [61]. For each colour $c \in \{1, 2, 3\}$, let

$$V_c = \{u_c : u \in V\},$$

and let T_c be the tree with leaf set V_c that consists of a (central) vertex z_c of degree $|V|/2$, where the $|V|/2$ neighbours of z_c each have degree three, and the $|V|$ leaves are arranged so that, for each edge $\{u, v\}$ of G coloured c, the vertices u_c and v_c are adjacent to the same degree-three vertex. As G is a cubic graph, T_c is well-defined for each $c \in \{1, 2, 3\}$. Let $\mathcal{T}$ be the phylogenetic X-tree that is constructed by starting with components T_1, T_2, and T_3 and two new (isolated) vertices x and y, and then connecting these components with new edges $\{x, y\}$, $\{y, z_1\}$, $\{y, z_2\}$, and $\{y, z_3\}$. Observe that the leaf set of $\mathcal{T}$ is $V_1 \cup V_2 \cup V_3 \cup \{x\}$. We specify the weighting function λ by setting

$$\lambda(e) = \begin{cases} 0 & \text{if } e \text{ is a pendant edge incident with a vertex in } V_1 \text{ or } V_2; \\ N & \text{if } e \text{ is a pendant edge incident with a vertex in } V_3; \\ 0 & \text{if } e = \{x, y\} \text{ or } e = \{y, z_c\} \text{ for some } c \in \{1, 2, 3\}; \\ 1 & \text{otherwise,} \end{cases}$$

where N is sufficiently large, say $N > |E|$. With this construction and weighting, our phylogenetic tree and corresponding edge weighting is complete. Now let D be the associated rooted tree with vertex set $V_1 \cup V_2 \cup V_3 \cup \{x\}$ and arc set

$$\bigcup_{u \in V} \{(x, u_3), (u_3, u_2), (u_2, u_1)\}.$$

Note that x is the root of D. Clearly, both $\mathcal{T}$ and D can be constructed in polynomial time.

We complete the proof by showing that G has a vertex cover of size at most m if and only if there is a viable subset S of X of size at most $3m$ such that $\mathrm{PD}(S) \geq |E| + mN$. Suppose first that there is a vertex cover $V' \subseteq V$ for G with $|V'| = m$. By selecting S to be the set $\{u_c : c \in \{1, 2, 3\}$ and $u \in V'\}$, we have a viable subset of X of size $3m$. Moreover, observing that there are exactly $|E|$ edges in $\mathcal{T}$ with weight 1 (each corresponding to a distinct edge of G), $\mathrm{PD}(S) = |E| + mN$.

Conversely, suppose that there is a viable subset S of X of size at most $3m$ that has PD score at least $|E| + mN$. Since $N > |E|$ and $\mathrm{PD}(S) \geq |E| + mN$, it follows that S must contain at least m leaves of T_3 so that the minimal subtree of $\mathcal{T}$ connecting the elements of S includes m edges with weight N. But then, as S is viable, for each such leave u_3 in S, the set S also includes u_1 and u_2. Therefore, $|S| = 3m$ and S consists of exactly these vertices. As $\mathrm{PD}(S) \geq |E| + mN$, it now follows that the minimal subtree of $\mathcal{T}$ connecting the elements in S must contain all $|E|$ edges with weight 1. In turn, this implies that $V' = \{u \in V : u_3 \in S\}$ is a vertex cover of G. As $|V'| = m$, we have completed the proof of the theorem. □

4.6 Improving greedy solutions is hard

Greedy algorithms have been regularly considered as approaches for solving problems that optimize some measure of diversity (see, for example, [8, 9, 54, 37, 48, 70]). There are a variety of reasons for this consideration. First, they are fast, simple to use and implement, and, more importantly, solve the original PD problem exactly [54, 70] and provide sharp approximation algorithms for other PD-related problems [8, 9]. Indeed, the fact that the original PD problem can be solved in this way motivated Moulton *et al.* [48] to consider PD and the greedy algorithm in detail. Second, in the context of conservation biology, they underlie the desirable property of stability [8]. In particular, one would like the set of species to be targeted for conservation to be stable as budgets vary. For example, if, given some initial budget, one selects a set of species to conserve resulting from a diversity-based method, one would like most of that set to remain if the budget was to be adjusted up or down at a later date and the chosen set of species to conserve was reselected under the new budget.

In this section, we consider the following greedy approach to solving Optimizing PD in Vertex-Weighted Food Webs.

Algorithm: Greedy(D, w, k)
Input: An acyclic digraph $D = (X, A)$, a non-negative real-valued weighting w on the vertices of D, and a positive integer k.
Output: A viable subset of X of size k.

```
Step 1 Let $S$ be the empty set and set counter $c = 0$.
Step 2 If $c = k$, STOP; otherwise, select an element $z$ of $X - S$ so that $S \cup \{z\}$ is viable and maximizes PD$(S \cup \{z\})$ − PD$(S)$.
Step 3 Set $S = S \cup \{z\}$ and $c = c + 1$, and return to Step 2.
```

It is not difficult to construct a counterexample to show that Greedy does not necessarily find an optimal solution to Optimizing PD in Vertex-Weighted Food Webs. Of course, since Greedy is trying to solve an NP-hard problem, this is not surprising. However, what is perhaps unexpected is that deciding if there is a feasible solution better than that returned by Greedy is NP-complete as we show next. It would be interesting to know of other situations where improving greedy solutions was a provably hard problem.

Decision problem: Greedy Optimality
Instance: An acyclic digraph $D = (X, A)$, a non-negative real-valued weighting w on the nodes of D, a positive integer k, and the PD score g of the solution returned by Greedy applied to (D, w, k).
Question: Is there a viable subset S of X of size at most k such that PD$(S) > g$?

Theorem 4.5. Greedy Optimality *is* NP-*complete.*

Proof. Greedy Optimality is clearly in NP since, given a subset S of X, one can easily verify in polynomial time whether S is viable and PD$(S) > g$. To complete the proof of the theorem, we show that there is a polynomial-time reduction from Vertex Cover to Greedy Optimality.

Let $G = (V, E)$ and m be a given instance of Vertex Cover. Let D be the acyclic digraph whose vertex set is the union of $V \cup E$ and $U = \{u_1, u_2, \ldots, u_{|E|+m-1}\}$, and whose arc set is the union of

$$\{(e, v) : e \in E,\ v \in V,\ v \text{ is an end-vertex of } e \text{ in } G\}$$

and

$$\{(u_i, u_{i-1}) : i \in \{2, 3, \ldots, |E| + m - 1\}\}.$$

Now let w be any function from the vertex set X of D to $\mathbb{R}^{\geq 0}$ that is defined, for all $x \in X$, by setting

$$w(x) = \begin{cases} 0 & \text{if } x \in V; \\ 1 & \text{if } x \in E; \\ \delta & \text{if } x \in \{u_1, u_2, \ldots, u_{|E|+m-2}\}; \\ \alpha & \text{if } x = u_{|E|+m-1}, \end{cases}$$

where $(|E| + m - 2)\delta + \alpha = |E| - \epsilon$ for some $\epsilon > 0$ and $0 < \delta < \frac{1-\epsilon}{|E|+m-2}$. Clearly, such a function exists.

Let $k = m + |E|$ and let g_k be the solution of Greedy applied to (D, w, k). Observe that $g_k = |E| - \epsilon$ and that any set corresponding to this solution must contain all of the elements in U and exactly one element in V. We next show that there is a vertex cover for G of size at most m if and only if there is a viable subset S of X of size at most $m + |E|$ such that $\mathrm{PD}(S) > g_k$.

Suppose first that there is a vertex cover V' of G of size at most m. Then, by taking the subset $V' \cup E$ of the vertex set of D, we have a viable subset of size at most $m + |E|$ whose weight is $|E|$. In particular, $\mathrm{PD}(V' \cup E) > g_k$.

For the converse, suppose that there is a viable subset S of X of size at most $m + |E|$ such that $\mathrm{PD}(S) > g_k$. If $E \subset S$, then we have a vertex cover for G of size at most m by choosing the set $V \cap S$. Therefore, we may assume that E is not a subset of S. Furthermore, if $u_{|E|+m-1} \in S$, then, as S is viable, $U \subseteq S$. In this case, as $|S| \leq m + |E|$, we have $S = U$ or $S = U \cup \{v\}$ for some $v \in V$. But then $\mathrm{PD}(S) = |E| - \epsilon = g_k$, which is a contradiction. Thus, we may also assume that $u_{|E|+m-1} \notin S$. Since E is not a subset of S, it follows that

$$\mathrm{PD}(S) \leq (|E| + m - 2)\delta + |E| - 1. \tag{4.1}$$

But δ was chosen so that $\delta < \frac{1-\epsilon}{|E|+m-2}$; that is,

$$(|E| + m - 2)\delta < 1 - \epsilon.$$

Combining this with (4.1), we get

$$\mathrm{PD}(S) \leq (|E| + m - 2)\delta + |E| - 1 < 1 - \epsilon + |E| - 1 = |E| - \epsilon,$$

contradicting the fact that $\mathrm{PD}(S) > g_k = |E| - \epsilon$. It follows that E must be a subset of S and so there is a vertex cover of G of size at most m. Since the reduction can be done in polynomial time, this completes the proof of the theorem. □

As noted earlier, Moulton *et al.* considered Optimizing PD with Dependencies in the context of greedy algorithms. They observed that in the trivial case $\mathcal{T}$ is a star tree in which all edge weights are equal the problem is solvable via a greedy algorithm. One can extend this observation further by showing that Optimizing PD in Vertex-Weighted Food Webs is solvable via Greedy if D has the property that, whenever P is a directed path in D, then $w(u) \leq w(v)$ for each $(u, v) \in P$. An interesting problem would be to determine precisely when Optimizing PD in Vertex-Weighted Food Webs is solvable via Greedy. However, Theorem 4.5 shows that any such characterization will not be validated in polynomial time unless P=NP.

4.7 Rooted and clock-like phylogenetic trees

In practice, one frequently works with rooted phylogenetic trees and, therefore, the rooted analogue of PD. Recall the definitions of these in Section 2.2. In the present short section, we review the implications of our earlier results in this setting.

The rooted analogue of Optimizing PD with Dependencies, called Optimizing Rooted PD with Dependencies, is the same as that in the unrooted setting but with the rooted phylogenetic tree replacing the (unrooted) phylogenetic tree and using the appropriate definition of PD (see Definition 2.4). A *rooted star tree* is a rooted phylogenetic tree in which the only interior vertex is the root. As in the unrooted setting, when $\mathcal{T}$ is a rooted star tree, Optimizing Rooted PD with Dependencies is equivalent to Optimizing PD in Vertex-Weighted Food Webs. A minor point to note is that, unlike the unrooted setting where $k \geq 2$ for this equivalence to work, there is no restriction on k in the rooted equivalence. It is now easily seen that Theorems 4.1, 4.2, and 4.5 apply to Optimizing Rooted PD with Dependencies too. Furthermore, the rooted analogue of Theorem 4.4 also holds. This can be easily checked by making minor changes to the proof of Theorem 4.4. In particular, distinguishing the interior vertex y as the root in the constructed tree and using PD as defined in Definition 2.4 in the course of the reduction.

In biology, it is sometimes reasonable to assume that mutations in evolution occur at a constant rate. This assumption is called the *molecular clock* assumption. Mathematically speaking, this assumption implies that, in a rooted phylogenetic tree, the sum of the lengths of the edges from the root to each leaf is the same. The notion of the existence of a molecular clock first appeared in [82] followed by [64]. Now consider Optimizing Rooted PD with Dependencies under the assumption that the edge-weights of $\mathcal{T}$

satisfy the molecular clock; $\mathcal{T}$ is also called *clock-like* in this case. If $\mathcal{T}$ is a star tree, then Optimizing Rooted PD with Dependencies is trivially solvable in polynomial time [48]. However, if $\mathcal{T}$ is arbitrary and D is a food tree, then Optimizing Rooted PD with Dependencies is NP-complete.

Theorem 4.6. Optimizing Rooted PD with Dependencies *when $\mathcal{T}$ is a rooted phylogenetic tree with the molecular clock assumption and D is either*

(i) *a rooted tree with all arcs directed away from the root or*

(ii) *a rooted tree with all arcs directed towards the root*

is NP-complete.

Proof. We just outline the proof of (i). The proof of (ii) is similar. We use a reduction from the restricted version of Vertex Cover in which G is cubic and planar. The proof is essentially the same as the proof of Theorem 4.4, and so we just highlight the necessary changes.

Distinguish the interior vertex y of the phylogenetic tree constructed in the proof of Theorem 4.4 to obtain a rooted phylogenetic tree $\mathcal{T}_y$ with root y. Using the original weighting function λ , we make $\mathcal{T}_y$ clock-like with the following weighting function λ_y:

$$\lambda_y(e) = \begin{cases} N+1 & \text{if } e = \{x, y\}; \\ N & \text{if } e \in \{\{y, z_1\}, \{y, z_2\}\}; \\ \lambda(e) & \text{otherwise.} \end{cases}$$

Setting $k = 3m + 1$ and $d = |E| + (m + 3)N + 1$ completes the necessary changes. □

The paper entitled ‘Optimizing phylogenetic diversity with ecological constraints’ [26] is a result of the work presented in this chapter. I would like to thank my co-authors Charles Semple and Dominic Welsh for their collaboration, and Magnus Bordewich, Peter Lockhart, Michael Snook, Mike Steel, and Alexander Zelikovsky for useful discussions relating to this work.

Chapter 5

Using linear programming to find solutions to a real instance

We have seen that Optimizing (Rooted) PD with Dependencies is an NP-hard optimization problem, except in the simple case when the phylogenetic tree is a star tree and the food web is either a rooted directed tree with all arcs directed away from the root or a rooted directed tree with all arcs directed towards the root. However, real phylogenetic trees are typically not star-like, and real food webs contain undirected and even directed cycles. To find exact solutions to real and possibly large instances, we generalize our optimization problem to allow directed cycles and formulate this more general and more realistic problem as an integer linear programming problem. We then solve this for an empirical food web with 249 vertices with a MATLAB solver and are able to find solutions for all possible values of k in seconds. In this chapter, we consider only rooted phylogenetic trees.

5.1 The real problem

As real ecological networks may contain directed cycles, we let $D = (X, A)$ be any directed graph. Just as in the previous chapter, there is an arc from a vertex $u \in X$ to a vertex $v \in X$ in A precisely if taxon u feeds or preys on taxon v. In this chapter, we often refer to vertices in D with out-degree zero as *source vertices*. Recall that viability was defined only for acyclic digraphs in Section 4.1. The same definition is suitable for arbitrary digraphs: a subset S of X is *viable* if, for each $s \in S$, there is a directed path in D from s to a source vertex such that every vertex in this path is in S. The biological meaning of a viable species set in an arbitrary digraph is that each non-source species has at least one prey and is connected to a source (a set of species in a directed cycle cannot persist if disconnected from the sources). Note that if the food web has no source species, that is, $D = (X, A)$ has no vertices of out-degree zero, then X has no viable subsets. However, the real instances of the problem that we are interested in contain food webs with at least one source species. Moreover, in real food webs $D = (X, A)$ under consideration, X is typically viable. We are now in the position to formally define the optimization problem we are interested in:

Optimization problem: Optimizing PD in Real Ecological Networks
Instance: A rooted phylogenetic X-tree $\mathcal{T}$, a non-negative real-valued weighting λ on the edges of $\mathcal{T}$, an arbitrary digraph $D = (X, A)$, and a positive integer k.
Goal: Find a viable subset S of X of size k that maximizes PD among all such subsets.

While it is obvious that the existence of a vertex of out-degree zero in D is a necessary (and sufficient) condition for X to have viable subsets, it is not straightforward to see under what conditions does X have a viable subset of size k? To answer this question, let us recall the definition of a greedoid. Let X be a finite set and let $\mathscr{F}$ be a collection of subsets of X. The pair $(X, \mathscr{F})$ is a *greedoid* if it satisfies the following two conditions:

(G1) If $F \in \mathscr{F}$ and $F \neq \emptyset$, then there is an element x in F such that $F - \{x\} \in \mathscr{F}$.
(G2) If $F_1, F_2 \in \mathscr{F}$ and $|F_2| = |F_1| + 1$, then there is an element x in $F_2 - F_1$ such that $F_1 \cup \{x\} \in \mathscr{F}$.

Moulton *et al.* [48] showed that if $D = (X, A)$ is a digraph and $\mathscr{F}$ is the collection of viable subsets F of X, then $(X, \mathscr{F})$ is a greedoid. An immediate consequence of this result is that X has a viable subset of size k precisely if a maximal viable subset of X has cardinality at least k. This follows from property (G1). As real food webs for which we aim to solve Optimizing PD in Real Ecological Networks have a viable vertex set X, they also have a viable subset of any size between 1 and $|X|$.

5.2 ILP formulation

To formulate Optimizing PD in Real Ecological Networks as an integer linear programming problem, we require further definitions and notations. Let λ_e denote the value that λ assigns to an edge e of $\mathcal{T}$. Let $E_{\mathcal{T}}$ denote the set of edges of $\mathcal{T}$, and let E_i and E_p denote the set of interior and the set of pendant edges of $\mathcal{T}$ respectively. For an interior edge $e \in E_i$, let X_e denote the set of leaves that are separated from the root of $\mathcal{T}$ by e. A digraph D is *strongly connected* if there is a directed path from each vertex in D to every other vertex in D. We denote the collection of strongly-connected subgraphs of $D = (X, A)$ none of whose vertices are source vertices by $\mathscr{S}_D$. If $C = (V_C, A_C)$ is a subgraph in $\mathscr{S}_D$, then $\delta^{-}(C)$ denotes the set of vertices v such that there is an arc (u, v) in D with $u \in V_C$ and $v \notin V_C$. Now we formulate our problem as a 0-1 integer linear programming problem.

$$
\begin{array}{lll}
\text{maximize} & \displaystyle\sum_{\substack{e=\{u,v\}\in E_p\\ v\in X}} \lambda_e x_v + \sum_{e\in E_i} \lambda_e x_e & \\
\text{subject to} & x_v, x_e \in \{0,1\} & \text{for all } v\in X, e\in E_i \\
& \displaystyle\sum_{v\in X} x_v = k & \\
& \displaystyle\sum_{v\in X_e} x_v \geq x_e & \text{for all } e\in E_i \\
& \displaystyle\sum_{u\in V_C} x_u - \sum_{v\in\delta^-(C)} x_v \leq |V_C| - 1 & \text{for all } C\in\mathscr{S}_D
\end{array}
$$

Here, the variables x_v correspond to the species in X and the variables x_e correspond to the interior edges of $\mathcal{T}$. They take value 1 if the species (or interior edge) is in the solution and 0 if not. Exactly k species must be chosen, and an interior edge may be chosen only if at least one species below that edge is selected. We do not require an interior edge to be selected if there are selected species below that edge. However, in an optimal solution, such an edge is selected if its λ-value is positive. Finally, whenever we select each vertex from a strongly connected subgraph C in $\mathscr{S}_D$, there has to be an arc in D from a vertex of C to a selected vertex outside of C. This condition is efficient to check for real food webs, and, as the next proposition shows, it is equivalent to the viability of the selected set of vertices.

Proposition 5.1. *A subset Y of the vertex set X of a digraph $D = (X, A)$ is viable if and only if it satisfies the following condition:*

(C) *For any subset Z of Y that induces a strongly connected subgraph in $\mathscr{S}_D$, there is a vertex q in $\delta^-(Z) \cap Y$.*

Proof. Assume first that Y is viable. Consider an arbitrary subset Z of Y that induces a strongly connected subgraph in $\mathscr{S}_D$. Since Y is viable, there is, for each $s \in Z$, a directed path in D from s to a vertex with out-degree zero in which every vertex is in Y. Let s_0 be any vertex in Z and let p be a directed path from s_0 to a source that is completely contained in Y. Since Z does not contain source vertices, p has to leave Z at some point, and so the first vertex of p that is outside of Z has to be in $\delta^-(Z) \cap Y$, as required.

Assume now that Y satisfies condition (C). We prove that Y is viable. Consider an arbitrary vertex, say s_0, in Y. We give a directed path from s_0 to a source vertex in which every vertex is in Y. If s_0 is a source vertex, then it forms a suitable path by itself.

Therefore, we assume that s_0 is not a source, and consider the vertex set M_0 of the strongly connected subgraph of D that satisfies $s_0 \in M_0 \subseteq Y$ and is maximal in Y. By (C), at least one of the elements of $\delta^-(M_0)$ is in Y. Let s_1 be an element in $\delta^-(M_0) \cap Y$. Let p_1 be a directed path from s_0 to s_1 whose vertices are all in $M_0 \cup \{s_1\}$. If s_1 is a source, we have found a directed path from s_0 to a source and the proof is complete. Otherwise, consider the vertex set M_1 of the strongly connected subgraph of D that satisfies $s_1 \in M_1 \subseteq Y$ and is maximal in Y. By (C), at least one of the elements of $\delta^-(M_1)$ is in Y. Let s_2 be in $\delta^-(M_1) \cap Y$. If s_2 is in M_0, then $M_0 \cup M_1$ induces a strongly connected subgraph in D, contradicting the maximality of M_0. Thus, $s_2 \in Y - (M_0 \cup M_1)$. Consider a directed path from s_1 to s_2 whose vertices are all in $M_1 \cup \{s_2\}$. This path, together with p_1, forms a directed path p_2 from s_0 to s_2. If s_2 is a source, we have found a suitable directed path and the proof is complete. Otherwise, we continue the above process. As there are finitely many vertices in Y, the process will terminate after visiting a finite number of vertices. If it terminates when finding a directed path from s_0 to a source, the proof is complete. So assume that the process terminates when arriving to a non-source vertex, say s_t, of Y along the directed path p_t whose vertices are all in $(\cup_{i=0}^{t-1} M_i) \cup \{s_t\}$. Since s_t is in $Y - (\cup_{i=0}^{t-1} M_i)$ (otherwise, M_{t-1} would not be maximal), there has to be a strongly connected subgraph of D whose vertex set M_t satisfies $s_t \in M_t \subseteq Y - (\cup_{i=0}^{t-1} M_i)$ and is maximal in Y. Since s_t is not a source, by (C), there has to be a vertex, say s_{t+1}, in $\delta^-(M_t) \cap Y$, and there has to be a directed path from s_t to s_{t+1} using vertices only in $M_t \cup \{s_{t+1}\}$. This path, together with p_t, forms a directed path from s_0 to s_{t+1}. This contradicts the fact that the process terminates at the non-source vertex s_t. Thus, termination occurs only when arriving to a source, giving a directed path from s_0 to a source in which every vertex is in Y. This completes the proof. □

Note that we assume that the set of instances for the problem Optimizing PD in Real Ecological Networks equals the set of instances for the above integer linear programming problem. The next theorem states that the two sets of optimal solutions corresponding to the two formulations are also equivalent.

Theorem 5.2. *Let I be an arbitrary instance of the integer linear programming problem, and consider the set $O_I = \{\mathbf{x} \colon \mathbf{x}$ is an optimal solution for $I\}$. For each $\mathbf{x} \in O_I$, let $Y_{\mathbf{x}}$ be the subset of X that contains a vertex $v \in X$ precisely if $x_v = 1$ in $\mathbf{x}$. Then, $\{Y_{\mathbf{x}} \colon \mathbf{x} \in O_I\}$ is the set of optimal solutions of* Optimizing PD in Real Ecological Networks *for the instance I.*

The proof of Theorem 5.2 is straightforward and is omitted.

5.3 Test Case

The constraints described above can be written in the form $\mathbf{Ax} \leq \mathbf{b}$ and input to any integer linear program solver. We use MATLAB to build $\mathbf{A}$ and $\mathbf{b}$ from a matrix that describes a rooted phylogenetic tree with lengths on its edges and from a matrix that describes a food web.

To evaluate whether realistically complex problems can be solved in practice, we used the largest food web currently in the empirical literature: a 249-node network with 3315 arcs based on feeding interactions in a large area of Caribbean shelf (details can be found in [51] and [6]). The only prior phylogenetic appraisal of this dataset [60] covered only 116 nodes, so we used a variety of sources to hand-construct a plausible clock-like rooted phylogenetic tree that includes all the species and trophospecies from the food web. This phylogeny was constructed by our biologist collaborator Travis Ingram. The food web features extensive cycling, with 553 strongly connected subgraphs. The large size and complexity of this food web means that it probably represents the most difficult real instance of the problem likely to be encountered. We ran the solver called `bintprog`, which is part of the Optimization Toolbox of MATLAB 7 and later, for all values of k between 1 and 249 on a server with an Intel Xeon processor (8 core, 3.2 GHz) and 32GB RAM. We found that in all cases an optimal solution was returned within 5 seconds.

5.4 Conclusions

By applying integer linear programming, we are able to effectively select maximum PD subsets of a given size of a species set while meeting the viability constraints imposed by a food web. Currently, we are working on the construction of a more realistic phylogenetic tree, for which Optimizing PD in Real Ecological Networks will be solved and optimal solutions will be interpreted. We also plan to study the structure of the set of all optimal solutions for a given instance. As there are other solvers that perform better than MATLAB's `bintprog`, we may use an alternative solver as well.

Related approaches have recently been proposed for other conservation and ecological problems, such as selecting a set of habitats to maximize PD [63], optimizing split diversity [45], or detecting community structures in networks [12]. Our application of integer linear programming to food web matrices might also be extended to investigate other problems involving ecological networks. For example, we might wish to select a set of species in a food web that maximizes total biomass or energy flow.

The work that has been presented in this chapter [24] is a result of joint work with Travis Ingram and Charles Semple. I would like to thank Travis and Charles for their collaboration, Karen Magnusson-Ford, Arne Mooers, and Mike Steel for useful discussions, and Carlos J. Melián for sharing food web data.

Chapter 6

On approximation of Optimizing PD for Cluster Systems

Given the phylogenetic X-tree of a species set X with lengths on its edges, in the basic PD optimization problem, one selects a k-element subset of X that maximizes PD over all k-element subsets [54, 70]. Since this optimization problem assumes that the evolutionary history of the species in X is known, it cannot be used in situations where we do not know the true tree or where evolution is not tree-like. In these cases, a more general biodiversity measure needs to be defined and, based on it, a more general optimization problem has to be formulated. Spillner et al. [68] introduced the measure 'phylogenetic diversity for split systems' ($\mathrm{PD}_{\mathscr{S}}$) and the problem Optimizing PD for Split Systems, which can be used when considering species whose evolution is better represented by an unrooted network rather than an unrooted tree. In this chapter, we give a different generalization of PD, which we call 'phylogenetic diversity for cluster systems' ($\mathrm{PD}_{\mathscr{C}}$). This measure is useful when the evolutionary history is best described by a rooted network [5], or when we do not know the true history, but we have a set of rooted phylogenetic trees (perhaps with different probabilities) and we want to maximize the expected PD. We consider the problem of finding a k-element subset of a given species set that maximizes $\mathrm{PD}_{\mathscr{C}}$ over all such subsets. We find that a greedy algorithm gives a $(1-e^{-1})$-approximation to this problem, and that there is no polynomial-time algorithm that achieves a better approximation ratio unless P=NP. We prove that, as a consequence, the problem Optimizing PD for Split Systems has a polynomial-time approximation algorithm with ratio $1-e^{-1}$.

In the next section, we formally define the measure $\mathrm{PD}_{\mathscr{C}}$. We then introduce the problem Optimizing PD for Cluster Systems and give an example of its applications. The results on the approximability properties of this optimization problem are stated in Section 6.2. I would like to note that these two results have already essentially appeared in the literature: as there is a strong relationship between our problem and the problem called Max k-Cover, the known approximability properties of this problem, described in [15] and [28], carry over to our problem. In Section 6.2, we describe our proofs to these results, the second of which also relies on the strong relationship between our problem and Max k-Cover. The consequence of these results for the approximability properties of Optimizing PD for Cluster Systems is stated and proved in the last section of this chapter.

6.1 Optimizing PD for Cluster Systems

Recall the definition of phylogenetic diversity for rooted phylogenetic trees (see Definition 2.4 in Section 2.2). The following definition generalizes this notion of PD.

Let X be a finite set, and let $\mathscr{C}$ be a collection of subsets of X. Furthermore, let w be a weighting function on $\mathscr{C}$ that assigns a non-negative real-valued weight to each member of $\mathscr{C}$. For a subset Y of X, we define the *phylogenetic diversity of Y relative to $\mathscr{C}$*, denoted by $\mathrm{PD}_{\mathscr{C}}(Y)$, as the sum of the weights of the members of $\mathscr{C}$ whose intersection with Y is non-empty. That is, we set

$$\mathrm{PD}_{\mathscr{C}}(Y) = \sum_{C \in \mathscr{C}, C \cap Y \neq \emptyset} w(C). \tag{6.1}$$

To see that $\mathrm{PD}_{\mathscr{C}}$ generalizes the notion of PD for rooted phylogenetic trees, consider the special case when X is the leaf set of a rooted phylogenetic X-tree $\mathcal{T}$. A subset C of X is a *cluster of $\mathcal{T}$* if there is an edge that has precisely C as its set of descendant leaves. Suppose that the edges of $\mathcal{T}$ have non-negative real-valued weights and let $\mathscr{C}$ be the set of all clusters of $\mathcal{T}$. For a cluster $C \in \mathscr{C}$, let $w(C)$ be the weight of the unique edge of $\mathcal{T}$ whose associated cluster is C. It is easy to see that in this setting, the phylogenetic diversity of a subset Y of X equals the phylogenetic diversity of Y relative to $\mathscr{C}$. That is, for any $Y \subseteq X$, we have $\mathrm{PD}_{\mathcal{T}}(Y) = \mathrm{PD}_{\mathscr{C}}(Y)$.

In this chapter, we consider the general case when $\mathscr{C}$ is an arbitrary collection of subsets of a finite set. In the following, we define and feature an optimization problem that is based on $\mathrm{PD}_{\mathscr{C}}$.

Optimization problem: Optimizing PD for Cluster Systems
Instance: A finite set X, a collection $\mathscr{C}$ of subsets of X, a non-negative real-valued weighting w on $\mathscr{C}$, and a positive integer k.
Goal: Find a subset Y of X of size k that maximizes $\mathrm{PD}_{\mathscr{C}}$ among all such subsets.

In the case when $\mathscr{C}$ is the collection of clusters of a rooted phylogenetic X-tree, Optimizing PD for Cluster Systems is just the basic PD optimization problem and is solvable in polynomial time using a greedy algorithm [54, 70]. However, as discussed later in this section, Optimizing PD for Cluster Systems is NP-hard in general. (This also follows immediately from the NP-hardness of Max k-Cover.)

One of the reasons why we are interested in solving the above problem is highlighted in the following example.

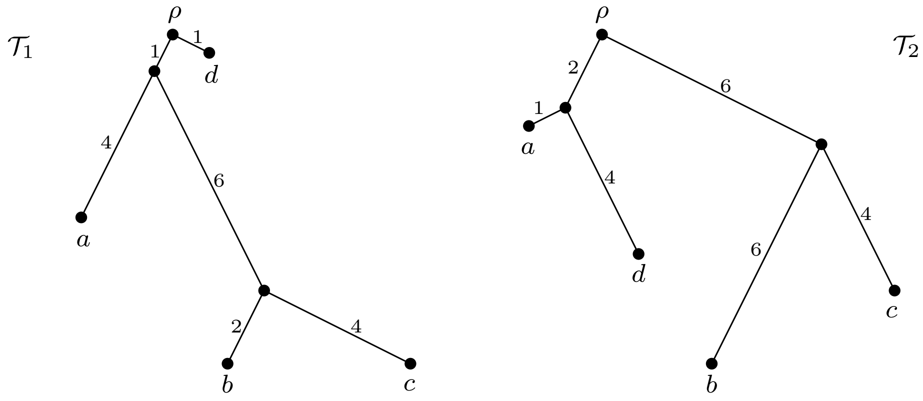


Figure 6.1: Two edge-weighted rooted phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$, both rooted at ρ.

Example. Let $X = \{a, b, c, d\}$ and consider the edge-weighted rooted phylogenetic X-trees shown in Figure 6.1. Assume that we do not know the evolutionary history of the species in X, but we know that either $\mathcal{T}_1$ or $\mathcal{T}_2$ represents it each with probability $\frac{1}{2}$, say. Consider the basic PD optimization problem with $k = 2$; that is, the problem of finding a two-element subset of X that has maximum PD among all two-element subsets. If $\mathcal{T}_1$ was the true tree, $\{a, c\}$ would be the optimal solution. However, if $\mathcal{T}_2$ was the true tree, the best two-element subset would be $\{b, d\}$. In such a situation, it may be safer to choose a subset of size two that maximizes the expected PD with respect to the probability distribution on the two trees. Let $W \subseteq X$ be of cardinality two and let $\mathbb{E}[\mathrm{PD}_{\mathcal{T}_1,\mathcal{T}_2}(W)]$ denote the expected PD of W with respect to the probability distribution on the two trees. Then, we have

$$\mathbb{E}[\mathrm{PD}_{\mathcal{T}_1,\mathcal{T}_2}(W)] = \frac{1}{2}\,\mathrm{PD}_{\mathcal{T}_1}(W) + \frac{1}{2}\,\mathrm{PD}_{\mathcal{T}_2}(W).$$

Consider now the cluster sets of $\mathcal{T}_1$ and $\mathcal{T}_2$. These are

$$\mathscr{C}_1 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{b, c\}, \{a, b, c\}\}$$

and

$$\mathscr{C}_2 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, d\}, \{b, c\}\},$$

respectively. For $i \in \{1,2\}$, let w_i assign to a cluster in $\mathscr{C}_i$ the weight of the edge corresponding to that cluster in $\mathcal{T}_i$. For example, $w_1(\{d\}) = 1$, $w_2(\{d\}) = 4$, and $w_2(\{a,d\}) = 2$. It is easy to see that $\mathbb{E}[\mathrm{PD}_{\mathcal{T}_1,\mathcal{T}_2}(W)]$ can be written as

$$\mathbb{E}[\mathrm{PD}_{\mathcal{T}_1,\mathcal{T}_2}(W)] = \sum_{C\in\mathscr{C}, C\cap W\neq\emptyset} w(C), \tag{6.2}$$

where $\mathscr{C} = \mathscr{C}_1 \cup \mathscr{C}_2$, $w(C) = \frac{1}{2}w_1(C)I_{C\in\mathscr{C}_1} + \frac{1}{2}w_2(C)I_{C\in\mathscr{C}_2}$, and $I_{C\in\mathscr{C}_i}$ takes the value 1 if $C \in \mathscr{C}_i$ and 0 otherwise. Since the right hand side of (6.2) is the $\mathrm{PD}_{\mathscr{C}}$ score of W under the above specified X, $\mathscr{C}$, and w, the problem of maximizing the expected PD is equivalent to solving Optimizing PD for Cluster Systems for our particular instance. A quick check of all two-element subsets of X shows that the unique optimal solution is $\{b,c\}$. This subset has the highest expected PD among all two-element subsets of X or, equivalently, it maximizes $\mathrm{PD}_{\mathscr{C}}$ over all such subsets.

The process described in the example also works in general when we are given a finite set of rooted phylogenetic trees with an arbitrary probability distribution on them; maximizing expected PD always leads to an instance of Optimizing PD for Cluster Systems. Furthermore, maximizing expected PD is equivalent to the problem Weighted Average PD on t (Rooted) Trees. For $t = 2$, this problem is solvable in polynomial time [10]. For $t \geq 3$, it is NP-hard [68], and so Optimizing PD for Cluster Systems is also NP-hard. However, Theorem 6.1 shows that there is a sharp approximation algorithm for it.

6.2 Approximability properties

Theorem 6.1. Optimizing PD for Cluster Systems *is an* NP-*hard optimization problem. However,*

(i) *there is a polynomial-time greedy algorithm with approximation ratio* $1 - e^{-1}$ *for this problem; and*

(ii) *there is no polynomial-time approximation algorithm for it with a ratio higher than* $1 - e^{-1}$ *unless* P=NP.

Theorem 6.1 does not present new results. As there is a strong relationship between Optimizing PD for Cluster Systems and the problem called Max k-Cover, results on the approximability of one of these problems immediately carry over to the other, and vice versa. Therefore, the result stated in Theorem 6.1(i) was established first in [15], while Theorem 6.1(ii) follows immediately from a result of Feige [28]. We state our proof for part (i) and show how [28] can be used to derive the statement in part (ii) in the next section.

Our proof of (i) uses the fact that $\mathrm{PD}_{\mathscr{C}}$ is a submodular set function. The greedy algorithm that actually gives the above-named approximation ratio is described in [50] in a more general setting. We briefly outline the algorithm here in the language of this chapter.

Algorithm: Greedy$(X, \mathscr{C}, w, k)$
Input: A finite set X, a collection $\mathscr{C}$ of subsets of X, a non-negative real-valued weighting w on $\mathscr{C}$, and a positive integer k.
Output: A subset of X of size k.

```
Step 1 Let $S$ be the empty set and set counter $c = 0$.
Step 2 If $c = k$, STOP and return $S$; otherwise, select an element $z$ of $X - S$ that maximizes $\mathrm{PD}_{\mathscr{C}}(S \cup \{z\}) - \mathrm{PD}_{\mathscr{C}}(S)$ among all elements of $X - S$ (with ties settled arbitrarily).
Step 3 Set $S = S \cup \{z\}$ and $c = c + 1$, and return to Step 2.
```

Greedy always produces a solution whose value is at least $1 - (1 - k^{-1})^k$ times the value of an optimal solution. This bound can be achieved for each k and has a limiting value of $1 - e^{-1}$ [50].

Remark. It would be interesting for future work to explore extensions to Theorem 6.1 (i) that allow costs to be assigned to the taxa. More precisely, suppose that each taxon has an associated positive real-valued cost associated with its conservation, and there is total budget B available to allocate. Then an extension to Optimizing PD for Cluster Systems is to select a subset of taxa to conserve that maximizes the PD score subject to the constraint that the sum of the costs of the taxa conserved does not exceed the budget B (Optimizing PD for Cluster Systems corresponds to the special case where all costs take the value 1). Recently, variations on the PD optimization problem on trees that allow taxon costs have allowed pseudo-polynomial-time exact algorithms and polynomial-time approximation algorithms [9, 55].

Proof of Theorem 6.1

We noted prior to the statement of Theorem 6.1 that Optimizing PD for Cluster Systems is NP-hard. Thus, the rest of this subsection establishes parts (i) and (ii). To prove Theorem 6.1 (i), we first verify the following lemma.

Lemma 6.2. *Let $\mathscr{C}$ be a collection of subsets of a finite set X and let w be a non-negative real-valued weighting on the elements of $\mathscr{C}$. Then, $PD_{\mathscr{C}}$ is a submodular set function. That is, for any two subsets A and B of X, we have*

$$PD_{\mathscr{C}}(A \cup B) + PD_{\mathscr{C}}(A \cap B) \leq PD_{\mathscr{C}}(A) + PD_{\mathscr{C}}(B). \tag{6.3}$$

Proof. Let A and B be arbitrary subsets of X. Apply Equation (6.1) to $A, B, A \cup B$ and $A \cap B$, and partition $\mathscr{C}$ into three sets as follows. For $i \in \{0, 1, 2\}$, let $\mathscr{C}_i$ consist of subsets in $\mathscr{C}$ whose intersection is non-empty with exactly i sets in $\{A, B\}$. Consider now the following cases. For a subset $C \in \mathscr{C}_0$, the weight $w(C)$ affects neither side of (6.3). For $C \in \mathscr{C}_1$, $w(C)$ appears exactly once on both sides of (6.3). Finally, for $C \in \mathscr{C}_2$, $w(C)$ appears exactly twice on the right hand side and at most twice on the left hand side. Noting that w is non-negative completes the proof. □

Proof of Theorem 6.1 (i). It is shown in [50] that a greedy heuristic can be used to approximate the following problem with approximation ratio $1 - e^{-1}$. Let S be a finite set and z be a real-valued function defined on the power set of S. Assume that z is submodular and non-decreasing and that $z(\emptyset) = 0$. The problem is to find a subset of S of size at most k that maximizes z among all such subsets. We complete the proof by showing that Optimizing PD for Cluster Systems is a special case of this problem. Take X as the finite set and $\mathrm{PD}_{\mathscr{C}}$ as the real-valued function on the power set of X. That is, set $S = X$ and $z = \mathrm{PD}_{\mathscr{C}}$. By Lemma 6.2, $\mathrm{PD}_{\mathscr{C}}$ is submodular. It is easy to see that $\mathrm{PD}_{\mathscr{C}}$ is also non-decreasing: for any subset A of X and for any element a in $X - A$, we have $\mathrm{PD}_{\mathscr{C}}(A \cup \{a\}) - \mathrm{PD}_{\mathscr{C}}(A) \geq 0$. Finally, $\mathrm{PD}_{\mathscr{C}}(\emptyset) = 0$. Theorem 6.1 (i) now follows. □

Before proving Theorem 6.1 (ii), we formally state the problem Max k-Cover and the definition of a type of approximability preserving reduction, called L-reduction.

Optimization problem: Max k-Cover
Instance: A finite set $S = \{s_1, \ldots, s_n\}$, a collection $\mathscr{F}$ of subsets of S, and a positive integer k.
Goal: Find a subset $\mathscr{F}' = \{F_1, \ldots, F_k\}$ of $\mathscr{F}$ of size k that maximizes the size of the set $\cup_{i=1}^{k} F_i$.

Feige [28] showed that no polynomial-time approximation algorithm for Max k-Cover can have an approximation ratio better than $1 - e^{-1}$ unless P=NP.

Let Π_1 and Π_2 be two arbitrary optimization problems. An *L-reduction* [4, 53] from Π_1 to Π_2 is a pair of polynomial-time computable functions f and g, and a pair of positive constants α and β that satisfy the following properties:

(I) If I is an instance of Π_1, then $f(I)$ is an instance of Π_2 with

$$\mathrm{opt}(f(I)) \leq \alpha\, \mathrm{opt}(I),$$

where $\mathrm{opt}(I)$ and $\mathrm{opt}(f(I))$ denote the size of an optimal solution to I and $f(I)$, respectively.

(II) If S is a feasible solution to $f(I)$, then $g(S)$ is a feasible solution to I with

$$|\mathrm{opt}(I) - c(g(S))| \leq \beta\, |\mathrm{opt}(f(I)) - c(S)|,$$

where $c(g(S))$ and $c(S)$ is the size of $g(S)$ and S, respectively.

It follows from the definition that if Π_1 L-reduces to Π_2, and there is a polynomial-time approximation algorithm for Π_2 with approximation ratio ϵ, then there is a polynomial-time approximation algorithm for Π_1 with approximation ratio $\alpha\beta\epsilon$ [53].

Proof of Theorem 6.1 (ii). We prove (ii) by giving an L-reduction with $\alpha = \beta = 1$ from Max k-Cover to Optimizing PD for Cluster Systems. By the previous remarks on Max k-Cover and L-reduction, this will imply that Optimizing PD for Cluster Systems cannot be approximated in polynomial time with an approximation ratio better than $1 - e^{-1}$ unless P=NP, as required.

Let I be an instance of Max k-Cover, and let R be an equivalence relation on S defined as follows. Two elements s_i and s_j of S are *equivalent* if and only if they are elements of precisely the same subsets in $\mathscr{F}$; that is, they satisfy $s_i \in F \Leftrightarrow s_j \in F$, for all F in $\mathscr{F}$. Let $[s_i]$ denote the equivalence class of $s_i \in S$ under R. We now give a function f that constructs from I an instance $f(I)$ of Optimizing PD for Cluster Systems; that is, it specifies a set, a collection of subsets of this set, a non-negative real-valued weight assigned to each subset in the collection, and a positive integer. Let $\mathscr{F}$ be the set and let $\mathscr{C}$ be the collection of subsets of $\mathscr{F}$ be defined as follows. For each equivalence class $[s_i]$ under R, there is a unique member $C_{[s_i]} = \{F \in \mathscr{F} : s_i \in F\}$ of $\mathscr{C}$. Let the weight of $C_{[s_i]} \in \mathscr{C}$ be the cardinality of the equivalence class $[s_i]$. Furthermore, let the positive integer in instance $f(I)$ equal k. Clearly, this construction can be accomplished in polynomial time.

To prove (I), we show that $\mathrm{opt}(I) = \mathrm{opt}(f(I))$, and so $\alpha = 1$. Suppose that $\mathscr{F}' = \{F_1, \ldots, F_k\}$ is an optimal solution to I. Then $\mathrm{opt}(I) = |\cup_{j=1}^{k} F_j|$. Trivially, $\mathscr{F}'$ is a feasible solution to $f(I)$. Moreover, $\mathscr{F}' \cap C_{[s_i]}$ is non-empty precisely if $\cup_{j=1}^{k} F_j$ contains s_i, in which case $[s_i] \subseteq \cup_{j=1}^{k} F_j$. By the choice of weighting, it now follows that $\mathrm{PD}_{\mathscr{C}}(\mathscr{F}') = |\cup_{j=1}^{k} F_j|$, and so $\mathrm{opt}(I) \leq \mathrm{opt}(f(I))$. By choosing an optimal solution to $f(I)$ and reversing this argument, it is also straightforward to show that $\mathrm{opt}(I) \geq \mathrm{opt}(f(I))$, as required.

For (II), let $\mathscr{F}'' = \{F^1, \ldots, F^k\}$ be a feasible solution to $f(I)$. Setting $g(\mathscr{F}'') = \mathscr{F}''$ gives a feasible solution to I with $c(g(\mathscr{F}'')) = c(\mathscr{F}'') = | \cup_{j=1}^{k} F^j|$. This can be seen by arguments similar to those used in the proof of (I). Trivially, g is computable in polynomial time. Thus, (II) is satisfied with $\beta = 1$. This completes the proof of Theorem 6.1 (ii). □

6.3 Optimizing PD for Split Systems

In this section, we first recall the definition of 'phylogenetic diversity for split systems' ($\mathrm{PD}_{\mathscr{S}}$) and the problem Optimizing PD for Split Systems from [68]. We then demonstrate how Theorem 6.1 (i) can be used to prove a result concerning the approximability properties of this problem.

Let X be a finite set. A bipartition $\{A, B\}$ of X is called a *split of* X, and a collection of splits of X is called a *split system* $\mathscr{S}$ *on* X. Consider such a split system $\mathscr{S}$ on X together with a weighting function w that assigns a non-negative real-valued weight to each split in $\mathscr{S}$. For a subset Y of X, the *phylogenetic diversity of* Y *relative to* $\mathscr{S}$, denoted by $\mathrm{PD}_{\mathscr{S}}(Y)$, is the sum of the weights of the splits $\{A, B\}$ in $\mathscr{S}$ for which both A and B have some common elements with Y [68]. That is, $\mathrm{PD}_{\mathscr{S}}(Y)$ is defined by

$$\mathrm{PD}_{\mathscr{S}}(Y) = \sum_{\substack{\{A,B\} \in \mathscr{S} \\ A \cap Y \neq \emptyset, B \cap Y \neq \emptyset}} w(\{A, B\}). \tag{6.4}$$

Spillner et al. [68] noted that if the split system $\mathscr{S}$ corresponds to a unique unrooted phylogenetic X-tree $\mathcal{T}$, then $\mathrm{PD}_{\mathscr{S}}$ corresponds to $\mathrm{PD}_{\mathcal{T}}$. Furthermore, they introduced and studied the following problem, which we consider in this section.

Optimization problem: Optimizing PD for Split Systems
Instance: A finite set X, a collection $\mathscr{S}$ of splits of X, a non-negative real-valued weighting w on $\mathscr{S}$, and a positive integer k.
Goal: Find a subset Y of X of size k that maximizes $\mathrm{PD}_{\mathscr{S}}$ among all such subsets.

It was shown in [68] that Optimizing PD for Split Systems is NP-hard, and it was noted without proof that for this problem, there is a polynomial-time greedy-type approximation algorithm with ratio $1 - e^{-1}$. We state this result and prove it using Theorem 6.1 (i). It should be noted that the algorithm suggested in [68] is more efficient than the algorithm used in our proof.

Theorem 6.3. *There is a polynomial-time approximation algorithm with approximation ratio* $1 - e^{-1}$ *for* Optimizing PD for Split Systems.

Proof. Given an arbitrary instance I of Optimizing PD for Split Systems, we describe how to produce a solution to I whose value is at least $1-e^{-1}$ times the value $\mathrm{opt}(I)$ of an optimal solution.

Let I consist of the finite set $X = \{x_1, \ldots, x_n\}$, the collection $\mathscr{S}$ of splits of X, the weighting function w on $\mathscr{S}$, and the positive integer k. For $i = 1, \ldots, n$, let I_i be the instance of Optimizing PD for Cluster Systems that is defined as follows. Let the base set be X. Let the collection $\mathscr{C}_i$ of subsets of X be the set that contains, for each bipartition $\{A, B\} \in \mathscr{S}$, precisely one of A and B, namely, $A \in \mathscr{C}_i$ if $x_i \in B$ and $B \in \mathscr{C}_i$ otherwise. If, for $\{A, B\} \in \mathscr{S}$, we have $A \in \mathscr{C}_i$, then let the weighting function on $\mathscr{C}_i$ assign weight $w(\{A, B\})$ to A; that is, for $A \in \mathscr{C}_i$ we set $w(A) = w(\{A, B\})$. Finally, let the positive integer pertaining to instance I_i be $k - 1$.

For $i = 1, \ldots, n$, let the set O_i be an optimal solution to I_i, and let $\mathrm{opt}(I_i)$ denote the value of O_i. That is, $\mathrm{opt}(I_i) = \mathrm{PD}_{\mathscr{C}_i}(O_i)$. Assume that O_i does not contain x_i. (If an optimal solution to I_i does contain x_i, taking out x_i from this solution and adding any element originally not in it, gives another optimal solution to I_i. This follows easily from the construction of $\mathscr{C}_i$.) Consider the k-element subsets $O_1 \cup \{x_1\}, \ldots, O_n \cup \{x_n\}$ of X and their $\mathrm{PD}_{\mathscr{S}}$ scores.

We first show that the set that has maximum $\mathrm{PD}_{\mathscr{S}}$ score among these sets, denoted by $M = O_m \cup \{x_m\}$, is an optimal solution to I. That is, we prove that $\mathrm{PD}_{\mathscr{S}}(M) = \mathrm{opt}(I)$. Assume the contrary that there exists a k-element subset Z of X the $\mathrm{PD}_{\mathscr{S}}$ score of which is greater than the $\mathrm{PD}_{\mathscr{S}}$ score of $O_i \cup \{x_i\}$, for $i = 1, \ldots, n$. That is, we have:

$$\mathrm{PD}_{\mathscr{S}}(Z) > \mathrm{PD}_{\mathscr{S}}(O_i \cup \{x_i\}), \text{ for } i = 1, \ldots, n. \tag{6.5}$$

Let x_j be an arbitrary element of Z. Denoting $Z - \{x_j\}$ by Z_j, we get $Z = Z_j \cup \{x_j\}$. Applying our assumption (6.5) to $i = j$ gives $\mathrm{PD}_{\mathscr{S}}(Z_j \cup \{x_j\}) > \mathrm{PD}_{\mathscr{S}}(O_j \cup \{x_j\})$. Using the definition of $\mathrm{PD}_{\mathscr{S}}$, this can be rewritten as follows:

$$\sum_{\substack{\{A,B\} \in \mathscr{S} \\ A \cap (Z_j \cup \{x_j\}) \neq \emptyset, B \cap (Z_j \cup \{x_j\}) \neq \emptyset}} w(\{A, B\}) > \sum_{\substack{\{A,B\} \in \mathscr{S} \\ A \cap (O_j \cup \{x_j\}) \neq \emptyset, B \cap (O_j \cup \{x_j\}) \neq \emptyset}} w(\{A, B\}). \tag{6.6}$$

If $\{A, B\}$ is a bipartition in $\mathscr{S}$, precisely one of A and B contains x_j. The other set (A or B), by definition, is a set in $\mathscr{C}_j$. If, for a bipartition $\{A, B\} \in \mathscr{S}$, x_j is, say, in B, then $B \cap (Z_j \cup \{x_j\}) \neq \emptyset$ on the left-hand side and $B \cap (O_j \cup \{x_j\}) \neq \emptyset$ on the right-hand side are satisfied, while $A \cap (Z_j \cup \{x_j\}) \neq \emptyset$ and $A \cap (O_j \cup \{x_j\}) \neq \emptyset$ are equivalent to

$A \cap Z_j \neq \emptyset$ and $A \cap O_j \neq \emptyset$, respectively, where $A \in \mathscr{C}_j$. Since this is true for each bipartition in $\mathscr{S}$, (6.6) simplifies to:

$$\sum_{A \in \mathscr{C}_j, A \cap Z_j \neq \emptyset} w(A) > \sum_{A \in \mathscr{C}_j, A \cap O_j \neq \emptyset} w(A).$$

But this is equivalent to $\mathrm{PD}_{\mathscr{C}_j}(Z_j) > \mathrm{PD}_{\mathscr{C}_j}(O_j)$, which contradicts the fact that O_j is optimal to I_j. Thus, M is an optimal solution to I.

Now we continue on finding an approximate solution to I. For $i = 1, \ldots, n$, apply the approximation algorithm that we described in Section 6.2 to I_i. If x_i appears in the obtained approximate solution $\hat{O}_i$ to I_i, then take out x_i from $\hat{O}_i$ and add any (but only one) element originally not voted in. Denote the resulting set by $\tilde{O}_i$. Note that, by the definition of $\mathscr{C}_i$, we get $\mathrm{PD}_{\mathscr{C}_i}(\tilde{O}_i) \geq \mathrm{PD}_{\mathscr{C}_i}(\hat{O}_i)$. Furthermore, since $\hat{O}_i$ is the output set of the approximation algorithm applied to I_i, we have $\mathrm{PD}_{\mathscr{C}_i}(\hat{O}_i) \geq [1 - (\frac{k-2}{k-1})^{k-1}]\mathrm{opt}(I_i)$. Thus, $\tilde{O}_i$ is an approximate solution to I_i, that does not contain x_i and that satisfies:

$$\mathrm{PD}_{\mathscr{C}_i}(\tilde{O}_i) \geq \left[1 - \left(\frac{k-2}{k-1}\right)^{k-1}\right] \mathrm{opt}(I_i). \tag{6.7}$$

Applying the above procedure to I_i for $i = 1, \ldots, n$ yields such an approximate solution to each instance: $\tilde{O}_1, \ldots, \tilde{O}_n$. Consider now the sets $\tilde{O}_1 \cup \{x_1\}, \ldots, \tilde{O}_n \cup \{x_n\}$ together with their $\mathrm{PD}_{\mathscr{S}}$ scores. We prove that the set that has maximum $\mathrm{PD}_{\mathscr{S}}$ score among these sets, denoted $\tilde{M} = \tilde{O}_m \cup \{x_m\}$, is an approximate solution to I satisfying:

$$\mathrm{PD}_{\mathscr{S}}(\tilde{M}) \geq \left[1 - \left(\frac{k-2}{k-1}\right)^{k-1}\right] \mathrm{opt}(I). \tag{6.8}$$

Since $1 - (\frac{k-2}{k-1})^{k-1} \geq 1 - e^{-1}$, once (6.8) has been established, the proof is complete.

By the definition of $\tilde{M}$, we have:

$$\mathrm{PD}_{\mathscr{S}}(\tilde{M}) \geq \mathrm{PD}_{\mathscr{S}}(\tilde{O}_i \cup \{x_i\}), \text{ for } i = 1, \ldots, n. \tag{6.9}$$

Using the definitions of $\mathrm{PD}_{\mathscr{S}}$ and $\mathrm{PD}_{\mathscr{C}_i}$ and reasonings similar to those used to simplify (6.6), it can be seen that $\mathrm{PD}_{\mathscr{S}}(\tilde{O}_i \cup \{x_i\}) = \mathrm{PD}_{\mathscr{C}_i}(\tilde{O}_i)$. This, combined with (6.9) and (6.7), gives:

$$\mathrm{PD}_{\mathscr{S}}(\tilde{M}) \geq \left[1 - \left(\frac{k-2}{k-1}\right)^{k-1}\right] \mathrm{opt}(I_i), \text{ for } i = 1, \ldots, n. \tag{6.10}$$

Now recall the definition of the set $M = O_m \cup \{x_m\}$, for which we proved that $\mathrm{PD}_{\mathscr{S}}(M) = \mathrm{opt}(I)$. Note that $\mathrm{PD}_{\mathscr{S}}(M) = \mathrm{PD}_{\mathscr{S}}(O_m \cup \{x_m\}) = \mathrm{PD}_{\mathscr{C}_m}(O_m) = \mathrm{opt}(I_m)$ and therefore, $\mathrm{opt}(I) = \mathrm{opt}(I_m)$. Applying (6.10) to $i = m$ gives (6.8), as required. □

It was noted in [68] that there exists a constant $\alpha > 0$ such that the problem of computing a k-element subset Y of X such that $\mathrm{PD}_{\mathscr{S}}(Y)$ is at least $1-\alpha$ times the maximum possible $\mathrm{PD}_{\mathscr{S}}$-value of a k-element subset is NP-hard. Theorem 6.3 implies that every α that satisfies this statement is less than e^{-1}. We end this chapter with a conjecture that states that the hardness of approximation is actually true for every $\alpha < e^{-1}$. That is, we conjecture that the approximation ratio $1-e^{-1}$ is best possible for Optimizing PD for Split Systems.

Conjecture 6.4. *There is no polynomial-time approximation algorithm for* Optimizing PD for Split Systems *with a ratio higher than* $1-e^{-1}$ *unless* P=NP.

The work that has been presented in this chapter is a result of joint work with Charles Semple and Mike Steel. I would like to thank Charles and Mike for their collaboration.

Part II

Probabilistic models

Chapter 7

Introduction to species extinction models

If one considers the phylogenetic diversity of the unknown subset of current species that will still be present at some future time, then this future phylogenetic diversity provides a measure of the impact of various extinction scenarios in biodiversity conservation. Under the simplest models of speciation, each taxon has the same probability of being extinct at some future time, and the extinction of taxa are treated as independent events; this is a simple type of 'field of bullets model'. A more realistic extension allows each species to have its own survival probability—this is the 'generalized field of bullets model' (g-FOB), which we study in Chapter 8. The g-FOB model also assumes that extinction events are independent. However, in some situations, extinction risks may be influenced by species characters, which may evolve according to a Markov process on the underlying tree. The 'state-based field of bullets model' and the 'trait-dependent field of bullets model' are based on this assumption, introducing dependencies between extinction events. In order to be able to study these models in Chapters 9 and 10, we first need to give some definitions.

7.1 Characters

In biology, characters describe different attributes of species, including morphological, behavioural, and genetic attributes. Mathematically, characters are functions.

Definition 7.1. *A* character on X *is a function* χ *from a non-empty subset* Y *of* X *into a set* C *of* character states. C *is also referred to as the* state set of χ. *If* $Y = X$, *we say that* χ *is a* full character, *and if* $|\{\chi(y) : y \in Y\}| = r$, *we say that* χ *is an* r-state character. *A character* χ *on* X *is a* binary character *if* χ *is a two-state full character.*

In this thesis, we only consider the way in which characters may evolve on some underlying phylogenetic tree. This is introduced in Sections 7.2 and 7.3. For other aspects of the theory and applications of characters, see [65].

7.2 Markov processes on trees

The concept of a Markov process on a tree has been extensively studied in physics, information theory, and evolutionary biology. In particular, finite-state Markov processes on trees are used to model the way in which characters of present-day species have evolved from the state present in some common ancestor [11, 29, 65, 69]. This concept of a Markov process on a tree generalizes the familiar notion of a Markov chain.

Definition 7.2. *Let T be a rooted tree with vertex set V. Viewing the edges of T as arcs directed away from the root, let $\leq$ be any total order on V such that, whenever (u, v) is an arc of T, we have $u < v$. A* Markov process on T *with state set C is a family $\{\xi_v : v \in V\}$ of random variables such that, whenever (u, v) is an arc of T,*

$$\mathbb{P}\left(\xi_v = \alpha_v | \bigcap_{w<v} \{\xi_w = \alpha_w\}\right) = \mathbb{P}\left(\xi_v = \alpha_v | \xi_u = \alpha_u\right), \tag{7.1}$$

where α_v, α_u, and the α_w-values are elements of C.

The *Markov property* (7.1) states that, for each arc (u, v) of T, the value of ξ_v, conditional on ξ_u, is independent of the ξ-values at all other earlier vertices.

For each arc $e = (u, v)$ of T, a Markov process on T with state set C induces an associated transition matrix, denoted $P(e) = [P(e)_{\alpha\beta}]$, with rows and columns indexed over C, and defined by

$$P(e)_{\alpha\beta} = \mathbb{P}(\xi_v = \beta | \xi_u = \alpha),$$

for all $\alpha, \beta \in C$. The matrix $P(e)$ is called the *transition matrix for edge e*. Let ρ denote the root of T and let $\pi_\alpha = \mathbb{P}(\xi_\rho = \alpha)$, for each $\alpha \in C$. Specifying π_α for every $\alpha \in C$ together with the transition matrices $P(e)$ for every arc e of T uniquely defines the Markov process on T.

As mentioned above, an example of a Markov process on a rooted tree is a finite Markov chain. This is a sequence $\xi_0, \xi_1, \xi_2, \ldots, \xi_n, \ldots$, of random variables taking values in some finite set C and that satisfies

$$\mathbb{P}\left(\xi_n = \alpha_n | \bigcap_{j<n} \{\xi_j = \alpha_j\}\right) = \mathbb{P}\left(\xi_n = \alpha_n | \xi_{n-1} = \alpha_{n-1}\right),$$

for all $n > 0$ and all α-values in C. A Markov chain can be regarded as a Markov process on the rooted tree that has vertex set $0, 1, 2, \ldots, n, \ldots$, with vertex 0 as the root and with an arc from i to $i + 1$ for each $i \geq 0$. It is also worth mentioning that Markov processes on trees form a special class of the more general *Markov random fields on graphs* [40, 65].

7.3 Character evolution on phylogenetic trees

Here, we link Markov processes on trees with phylogenetic trees and characters. We focus only on the basics that are essential to understand the following chapters.

Suppose that $\mathcal{T} = (T, \phi)$ is a rooted phylogenetic X-tree. A *Markov process on $\mathcal{T}$* is a Markov process on the rooted tree T.

For a Markov process on a rooted phylogenetic X-tree $\mathcal{T} = (T, \phi)$ and a full character $\chi \colon X \to C$, let

$$p(\chi) = \mathbb{P}\left(\bigcap_{x \in X} \{\xi_{\phi(x)} = \chi(x)\} \right).$$

This is the probability that, for all $x \in X$, the leaf of T labelled $\phi(x)$ takes the state specified by the character χ. Let $T = (V, E)$ and let ρ denote the root of T. Then, $p(\chi)$ can be expressed as

$$p(\chi) = \sum_{\bar{\chi} \colon V \to C, \bar{\chi} \circ \phi = \chi} \pi_{\bar{\chi}(\rho)} \prod_{e=(u,v) \in E} P(e)_{\bar{\chi}(u)\bar{\chi}(v)}, \tag{7.2}$$

where the sum is taken over all extensions $\bar{\chi}$ of χ to V [65]. Equation (7.2) expresses the fact that $p(\chi)$ is a marginal distribution, obtained by summing the probability of all the possible extensions $\bar{\chi}$ of χ to all the vertices of T.

For further definitions and results on general and specific Markov models on phylogenetic trees, we refer the reader to [65].

Chapter 8

Distribution of future PD under the g-FOB model

Under the field of bullets model, each taxon has the same constant probability of being extinct at some time in the future, and extinction events are independent (see, for example, [49, 58, 77]). This model is quite restrictive [57]; the more realistic generalized field of bullets model (g-FOB) allows each species to have its own survival probability. In the present chapter, we consider the g-FOB model and make predictions of the PD score of the set of taxa that survive. This future PD is a random variable with a well-defined distribution, but to date, most attention has focused on its mean, that is, the expected PD score of the species that survive. For example, the Noah's Ark problem [34, 78, 55] attempts to maximize expected future PD by allocating resources that increase the survival probabilities in a constrained way.

Clearly, one could consider other properties of the distribution of future PD—for example, the probability, let us call it the PL_0 value, that future PD is less than some critical lower limit L_0. Given different conservation strategies, we may wish to maximize expected PD or minimize the PL_0 value. A natural question is how are these two quantities related? Minimizing PL_0 is in line with a min-max approach to PD-based biodiversity conservation [22]. This is a familiar strategy in other fields, such as economics, where one wishes to minimize the risk of worst-case scenarios.

To address these sorts of questions, we need to know the full distribution of future PD. In this chapter, we show that future PD is asymptotically normally distributed. Our work was also motivated by the increasing trend in biology of constructing and analyzing phylogenetic trees that contain large numbers of species ($10^2 - 10^3$), and the suggestive form of distributions obtained by simulating future PD by sampling 12-leaf subtrees randomly from 64-leaf trees [49] (see also [77]).

To formally prove the normal limit law requires some care as future PD is not a sum of independent random variables, even though the survival events for the taxa at the leaves are treated independently. Consequently, the usual central limit theory does not immediately apply. The style of our proof has some similarities to the approach in [72], in which the authors established an asymptotic normal distribution for the parsimony score of a random assignment of character states to the leaves of a phylogenetic tree. However, the properties of parsimony score are quite different to phylogenetic diversity, requiring a somewhat different type of tree decomposition in the proof and other modifications. We also note that an asymptotic normal distribution of a quantity related to phylogenetic diversity was described in [52], however, in that paper, the tree is random rather than fixed.

This limit law has some useful consequences for applications. For example, it means that for a large tree, the PL_0 value can be estimated by the area under a normal curve to the left of $\frac{L_0-\mathbb{E}[\mathrm{PD}]}{\sqrt{\mathrm{Var}[\mathrm{PD}]}}$, where $\mathbb{E}[\mathrm{PD}]$ denotes the expected value of future PD. In particular, we see that the relation between the PL_0 value and expected future PD involves scaling by the standard deviation of future PD, so strategies that aim to maximize expected future PD may not necessarily minimize the PL_0 value.

Our normal distribution result is asymptotic; that is, it holds for large trees. However, it is also useful to have techniques for calculating the exact PD distribution on any given tree. In the following, we also show how this may be achieved by a pseudo-polynomial-time algorithm under the mild assumption that each edge length is (approximated by) an integer multiple of some fixed length.

This chapter is organized as follows. First, we consider rooted phylogenetic trees and phylogenetic diversity as defined in Definition 2.4 for rooted phylogenetic trees. Section 8.1 gives all the definitions and preliminary results used in Section 8.2. In particular, it shows how to determine the main parameters—mean and variance—of the distribution we set out to study. Section 8.2 contains the main result of this chapter—the asymptotic normality of this distribution—together with its formal proof and the conditions under which this result holds. Section 8.3 describes the algorithm to derive the exact distribution of future PD. In Section 8.4, we show how our results can be easily modified to handle PD for unrooted phylogenetic trees. Finally, Section 8.6 summarizes and discusses the main results of this chapter.

8.1 Mean and variance of future PD

Suppose we have a rooted phylogenetic X-tree $\mathcal{T}$ and a map λ that assigns a non-negative real-valued length λ_e to each edge e of $\mathcal{T}$. In the *generalized field of bullets model* (g-FOB), we have a triple $(\mathcal{T}, \lambda, p)$, where $\mathcal{T}$ is a rooted phylogenetic X-tree, λ is an edge length assignment map, and p is a map that assigns to each leaf $x \in X$ a probability p_x. Construct a random set X' by assigning each element x of X to X' independently with probability p_x. In biodiversity conservation, we regard X' as the set of taxa that will still exist (that is, not be extinct) at some time t in the future. Accordingly, we call p_x the *survival probability*

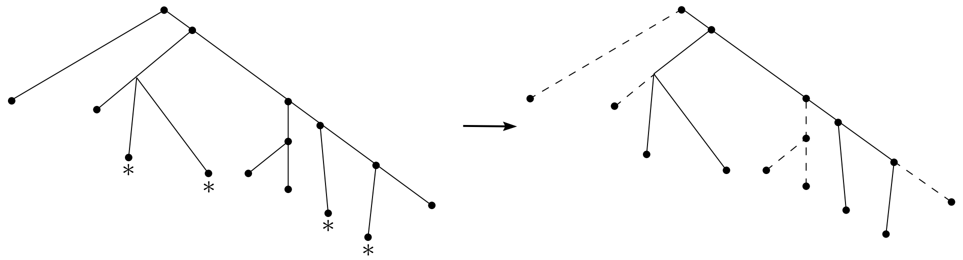


Figure 8.1: If only the taxa marked * in the rooted tree on the left survive, then the future phylogenetic diversity is the sum of the lengths of the solid edges in the tree on the right.

of x. The value p_x depends on t, and in a monotone decreasing fashion [35]. In this thesis, we consider t to be some fixed candidate time in the future—for example, 5 years or 100 years from now—rather than a continuous variable. The assignment of survival probabilities should ideally be based on population viability analysis [7], as discussed in [59]. This latter paper also describes an alternative assignment procedure based on IUCN Red List guidelines.

Considering the random variable $\varphi = \varphi_{\mathcal{T}} = \mathrm{PD}_{(\mathcal{T},\lambda)}(X')$, which is the phylogenetic diversity of the random subset X' of X consisting of those taxa that survive according to the process just described, we call φ *future phylogenetic diversity.* An example of this process is shown in Figure 8.1. Note that we are considering extinction here in the short-term (for example, tens or hundreds of years) rather than on evolutionary timescales, so the negligible increase in branch lengths of surviving species is ignored in calculating φ.

Note that in the g-FOB model, we can write

$$\varphi = \sum_e \lambda_e Y_e, \tag{8.1}$$

where Y_e is the binary random variable that takes the value 1 if e lies on a path between an element of X' and the root of $\mathcal{T}$ and that is 0 otherwise. Moreover,

$$\mathbb{P}[Y_e = 1] = 1 - \prod_{x \in C_e} (1 - p_x), \tag{8.2}$$

where C_e is the set of elements of X that are separated from the root of $\mathcal{T}$ by e. Consequently, if we let $P_e := \mathbb{P}[Y_e = 1]$, then

$$\mathbb{E}[\varphi] = \sum_e \lambda_e P_e. \tag{8.3}$$

Equation (8.1) suggests that for large trees, φ might be normally distributed, as it is a sum of many random variables. A normal distribution is also suggested by simulations described in [49, 77], though in that setting random samples of fixed size were drawn rather than selecting each taxon independently with a given probability, which leads to variable size samples. Nevertheless, one can relate these two approaches by setting the taxon selection probability in the g-FOB model equal to the proportion of taxa sampled in a fixed-sample-size setting. The proportion of taxa sampled in this setting of the g-FOB model has a mean that matches the fixed-sample-size setting, and a variance that tends to zero as the sample size grows.

Although φ is a sum of the random variables $(\lambda_e Y_e)$, these are not identically distributed and, more importantly, they are not independent. Therefore, a straightforward application of the usual central limit theorem seems problematic. We show that under two mild restrictions, a normal law can be established for large trees. Moreover, neither of these two mild restrictions can be lifted. We exhibit a counter-example to a normal law in both cases.

Since a normal distribution is determined once we know both its mean and variance, it is useful to have equations for calculating both these quantities. Equation (8.3) provides a simple expression for the mean, and we now present an expression for the variance that is also easy to compute. Given two distinct edges e, f of $\mathcal{T}$, we write $e <_{\mathcal{T}} f$ if the path from the root of $\mathcal{T}$ to f includes edge e (or, equivalently, if $C_f \subset C_e$).

Lemma 8.1.

$$Var[\varphi] = \sum_e \lambda_e^2 P_e(1 - P_e) + 2 \sum_{(e,f):e<_{\mathcal{T}} f} \lambda_e \lambda_f P_f(1 - P_e).$$

Proof. From Equation (8.1), we have:

$$\mathrm{Var}[\varphi] = \mathrm{Cov}[\varphi, \varphi] = \sum_{e,f} \lambda_e \lambda_f \,\mathrm{Cov}[Y_e, Y_f].$$

The covariance of Y_e and Y_f is:

$$\mathrm{Cov}[Y_e, Y_f] = \mathbb{E}[Y_e Y_f] - \mathbb{E}[Y_e]\mathbb{E}[Y_f] = \mathbb{P}[Y_e = 1, Y_f = 1] - \mathbb{P}[Y_e = 1]\mathbb{P}[Y_f = 1].$$

Now, we have the following cases:

(i) $e \neq f$ and neither $e <_{\mathcal{T}} f$ nor $f <_{\mathcal{T}} e$. In this case, the subtree of $\mathcal{T}$ with root edge e and the subtree of $\mathcal{T}$ with root edge f do not have any leaves in common, and so Y_e and Y_f are independent. Thus, $\mathrm{Cov}[Y_e, Y_f] = 0$.

(ii) $e <_{\mathcal{T}} f$. In this case, $C_f \subset C_e$ and so the survival of any taxon in C_f implies the survival of a taxon in C_e; that is, $Y_f = 1$ implies $Y_e = 1$ and we have $\mathrm{Cov}[Y_e, Y_f] = \mathbb{P}[Y_f = 1] - \mathbb{P}[Y_e = 1]\mathbb{P}[Y_f = 1] = P_f(1 - P_e)$.

(iii) $f <_{\mathcal{T}} e$. This is analogous to case (ii) (and, together with case (i), explains the factor of 2 in the expression on the right-hand side of our formula for $\mathrm{Var}[\varphi]$).

(iv) $e = f$. This case gives $\mathrm{Cov}[Y_e, Y_f] = \mathbb{P}[Y_e = 1](1 - \mathbb{P}[Y_e = 1]) = P_e(1 - P_e)$ (and corresponds to the first term on the right-hand side of our formula for $\mathrm{Var}[\varphi]$).

By considering these cases for $\mathrm{Cov}[Y_e, Y_f]$, we obtain the result claimed. □

A consequence of this lemma is the following lower bound on the variance of future PD, which will be useful later.

Corollary 8.2. *Consider the g-FOB model on $(\mathcal{T}, \lambda, p)$. If $E_p(\mathcal{T})$ denotes the set of pendant edges of $\mathcal{T}$, then*

$$Var[\varphi] \geq \sum_{e \in E_p(\mathcal{T})} \lambda_e^2 P_e(1 - P_e).$$

Proof. Notice that all the terms in the summation expression for $\mathrm{Var}[\varphi]$ in Lemma 8.1 are non-negative, and so a lower bound on $\mathrm{Var}[\varphi]$ is obtained by summing over those pairs (e, f) for which $e = f$ is a pendant edge of $\mathcal{T}$. This gives the claimed bound. □

8.2 Asymptotic normality of future PD

Consider a sequence of rooted phylogenetic trees:

$$\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n, \ldots,$$

where for each $n \geq 1$, $\mathcal{T}_n$ has a leaf label set $X = \{1, \ldots, n\}$. Furthermore, suppose that for each tree, we have an associated edge length function $\lambda = \lambda^{(n)}$ and a survival probability function $p = p^{(n)}$. Let $E(\mathcal{T}_n)$ and $E_p(\mathcal{T}_n)$ denote the set of edges and the set of pendant edges of $\mathcal{T}_n$ respectively. For the sequence of g-FOB models $(\mathcal{T}_n, \lambda^{(n)}, p^{(n)})$, we impose the following conditions:

(**C1**) For some $\epsilon > 0$ and for each n, we have:

$$\epsilon \leq p_x^{(n)} \leq 1 - \epsilon,$$

for all $x \in \{1, \ldots, n\}$ except for at most An^{α} values of x, where $A, \alpha \geq 0$ are constants, with $\alpha < \frac{1}{2}$.

(**C2**) Let $L(n) = \max\{\lambda_e^{(n)} : e \in E(\mathcal{T}_n)\}$. Then, for each n, we have:

$$\sum_{e \in E_p(\mathcal{T}_n)} \left(\lambda_e^{(n)}\right)^2 \geq Bn^{\beta} L(n)^2,$$

for some constants $B > 0, \beta > 2\alpha$.

Remarks concerning conditions (C1) and (C2). Condition (C1) simply says that the survival of most taxa is neither arbitrarily close to certain nor impossible. The term An^{α} provides the flexibility to allow for some of the taxa to have a survival probability that is very close to, or even equal to, 0 or 1.

Condition (C2) says, roughly speaking, that the pendant edges are, on average, not too short in relation to the longest edge in the tree. This is relevant for evolutionary biology, as it follows that for trees generated by a constant speciation rate pure birth model (see, for example, [19]) condition (C2) holds in expectation (for any $\alpha \in (0, \frac{1}{2})$). A more formal statement of this claim, and its proof, is given in [25].

Note that if condition (C2) holds for a value $\beta > 0$, then β is at most 1, since each term in the summation expression in (C2) is at most $L(n)^2$ and there are $O(n)$ of them.

□

Next, we state our main theorem, which describes the asymptotic normality of future phylogenetic diversity $\varphi_n = \varphi_{\mathcal{T}_n}$. Since phylogenetic trees often contain a large number of taxa, the result allows one to approximate the distribution of future phylogenetic diversity with a normal distribution.

Theorem 8.3. *Under conditions (C1) and (C2),* $(\varphi_n - \mathbb{E}[\varphi_n])/\sqrt{Var[\varphi_n]}$ *converges in distribution to* $N(0,1)$ *as* $n \to \infty$*, where* $N(0,1)$ *denotes a standard normally distributed random variable.*

We pause to note that one cannot drop either condition (C1) or (C2) in Theorem 8.3. It is clear that dropping (C1) is problematic. For example, set $p_x^{(n)} \in \{0,1\}$ for all x which leads to a degenerate distribution. As for (C2) the following example shows that we require β to be strictly positive.

Example: Condition (C2) cannot be removed. Consider a rooted tree $\mathcal{T}_n$ with n leaves. Leaves $1, \ldots, n-1$ have incident edges that each have length $\frac{1}{\sqrt{n-1}}$ and all these edges are incident with a vertex that is adjacent to the root by an edge of length 1. Leaf n has edge length 1 (see Figure 8.2). Consider a sequence of g-FOB models with $p_x^{(n)} = s$ for all x, n, where s is any number strictly between 0 and 1. Then, $\varphi_n = \frac{1}{\sqrt{n-1}}A_n + B_n + C_n$, where $\frac{1}{\sqrt{n-1}}A_n$ is the contribution to φ_n of the $n-1$ edges that are incident with leaves $1, \ldots, n-1$, term B_n is the contribution to φ_n of the edge that connects these $n-1$ edges to the root of $\mathcal{T}_n$, and C_n is the contribution to φ_n of the edge incident with leaf n. Notice that A_n is a sum of $n-1$ independent and identically-distributed binary $(0, 1)$ random variables, each of which takes the value 1 with probability s, and C_n is a binary random variable which takes the value 1 with probability s. Consequently, the variance of $\frac{1}{\sqrt{n-1}}A_n$ equals $s(1-s)$, the same as the variance of C_n. Moreover, B_n converges in probability to 1, and C_n is independent of A_n and B_n. Consequently, $\text{Var}[\varphi_n] \to 2s(1-s)$ as $n \to \infty$. Furthermore, by the standard central limit theorem, $\frac{\frac{1}{\sqrt{n-1}}A_n - \mathbb{E}[\frac{1}{\sqrt{n-1}}A_n]}{\sqrt{2s(1-s)}}$ converges in distribution to $N(0, \frac{1}{2})$ (a normal random variable with mean 0 and variance $\frac{1}{2}$). Thus, $(\varphi_n - \mathbb{E}[\varphi_n])/\sqrt{\text{Var}[\varphi_n]}$ converges to the random variable $N(0, \frac{1}{2}) + W$, where $W = (C_n - \mathbb{E}[C_n])/\sqrt{2s(1-s)}$ is independent of $N(0, \frac{1}{2})$ and takes the value $\frac{1-s}{\sqrt{2s(1-s)}}$ with probability s and takes the value $\frac{-s}{\sqrt{2s(1-s)}}$ with probability $1-s$. In particular, $(\varphi_n - \mathbb{E}[\varphi_n])/\sqrt{\text{Var}[\varphi_n]}$ does not converge in distribution to $N(0, 1)$. Notice that in this example, (C1) is satisfied, but (C2) fails since $\sum_{e \in E_p(\mathcal{T}_n)} (\lambda_e^{(n)})^2 = 2L(n)^2$.

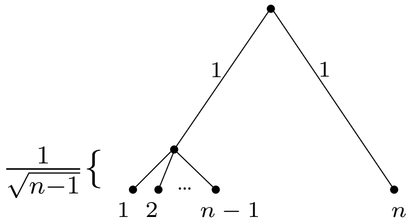


Figure 8.2: A rooted tree for which future phylogenetic diversity does not become normally distributed as n grows.

□

We now provide a brief, informal outline of the approach we use to prove Theorem 8.3. The main idea is to decompose $\mathcal{T}_n$ into a central core and a large number of moderately small pendant subtrees. Each edge in the central core separates the root from enough leaves so that we can be almost certain that at least one of these leaves survives—consequently the combined PD-contribution of this central core converges in probability to a fixed (non-random) function of n. Regarding the pendant subtrees, their contributions to the PD score are independent and although they are not identically distributed random variables, their combined variance grows sufficiently quickly that we can establish a normal law for their sum by a standard central limit theorem.

Proof of Theorem 8.3. We first note that it is sufficient to establish Theorem 8.3 under (C1) and (C2*), which can be viewed as a normalization of (C2):

(C2*) $L(n) = 1$, and $\sum_{e \in E_p(\mathcal{T}_n)} (\lambda_e^{(n)})^2 \geq Bn^\beta$ for constants $B > 0, \beta > 2\alpha$.

To see why this condition suffices, suppose we have established Theorem 8.3 under (C1) and (C2*). For a sequence $\mathcal{T}_n$ (with associated maps $\lambda^{(n)}$, $p^{(n)}$) satisfying (C1) and (C2), let $\mu_e^{(n)} = L(n)^{-1}\lambda_e^{(n)}$ for each edge e of $\mathcal{T}_n$ and each n. Note that, by Equation (8.1), the normalized φ score (namely $(\varphi_n - \mathbb{E}[\varphi_n])/\sqrt{\mathrm{Var}[\varphi_n]})$) for $(\mathcal{T}_n, \mu^{(n)}, p^{(n)})$ equals the normalized φ score for $(\mathcal{T}_n, \lambda^{(n)}, p^{(n)})$ and that $(\mathcal{T}_n, \mu^{(n)}, p^{(n)})$ satisfies (C2*). Thus, we will henceforth assume conditions (C1) and (C2*).

Next, we make a notational simplification: for the remainder of the proof, we will write $\lambda_e^{(n)}$ as λ_e and $p_x^{(n)}$ as p_x, but we will respect in the proof that these quantities depend on n. Also, for a sequence of random variables (Y_n), we write $Y_n \xrightarrow{P} a$ to denote that Y_n converges in probability to a constant a, and $Y_n \xrightarrow{D} Y$ to denote that Y_n converges in distribution to a random variable Y.

Since $\beta > 2\alpha$, we may select a value γ with $\alpha < \gamma < \beta/2$, and set $f(n) := n^\gamma$. We partition the edges of $\mathcal{T}_n$ into two classes E_1^n and E_2^n and we define a third class $E_{12}^n \subseteq E_1^n$ as follows: Let n_e denote the number of leaves of $\mathcal{T}_n$ that are separated from the root by e. Then set:

- E_1^n: edges e of $\mathcal{T}_n$ with $n_e \leq f(n)$;
- E_2^n: edges e of $\mathcal{T}_n$ with $n_e > f(n)$;
- E_{12}^n: edges $e \in E_1^n$ such that e is adjacent to an edge $f \in E_2^n$.

For an edge $e \in E_{12}^n$ of $\mathcal{T}_n$, we make the following definitions:

- t_e denotes the subtree of $\mathcal{T}_n$ consisting of edge e and all other edges of $\mathcal{T}_n$ that are separated from the root by e.
- φ_e^n denotes the future phylogenetic diversity of t_e, under the probabilistic model described above.

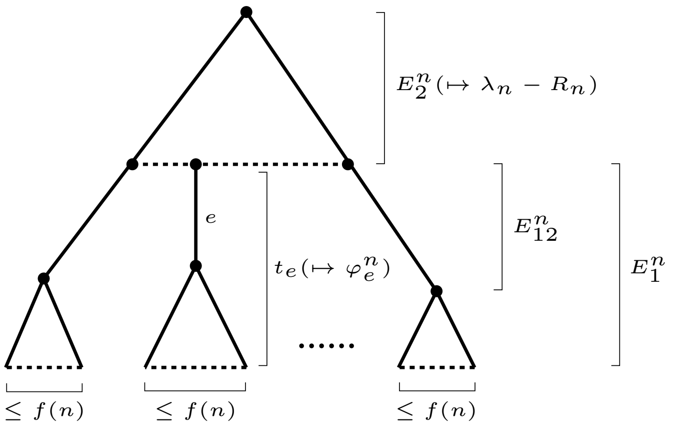

$E_2^n (\mapsto \lambda_n - R_n)$
e
$t_e (\mapsto \varphi_e^n)$
E_{12}^n
E_1^n
$\leq f(n)$
$\leq f(n)$
$\leq f(n)$

Figure 8.3: A representation of the decomposition of $\mathcal{T}_n$ in the proof of Theorem 8.3.

See Figure 8.3 for a schematic summary of these concepts.

For φ_n, Equation (8.1) gives:

$$\varphi_n = \sum_{e \in E_1^n} \lambda_e Y_e + \sum_{e \in E_2^n} \lambda_e Y_e = \sum_{e \in E_{12}^n} \varphi_e^n + \sum_{e \in E_2^n} \lambda_e Y_e. \tag{8.4}$$

Let

$$\lambda_n = \sum_{e \in E_2^n} \lambda_e, \; Z_n = \sum_{e \in E_{12}^n} \varphi_e^n, \text{ and } R_n = \sum_{e \in E_2^n} \lambda_e (1 - Y_e).$$

With this notation, we can rewrite (8.4) as

$$\varphi_n = \lambda_n + Z_n - R_n. \tag{8.5}$$

The next lemma states that the last term in this equation makes a vanishing contribution to φ_n.

Lemma 8.4. $R_n \xrightarrow{P} 0$.

Proof. Since $\mathrm{Var}[R_n] = \mathbb{E}[R_n^2] - \mathbb{E}[R_n]^2$ and $\mathbb{E}[R_n^2] \geq \mathbb{E}[R_n]^2$, it is sufficient to show that $\mathbb{E}[R_n^2] \to 0$ (the claim that $R_n \xrightarrow{P} 0$ then follows by Chebyshev's inequality). We have $R_n = \sum_{e \in E_2^n} \lambda_e (1 - Y_e)$, and so:

$$R_n^2 = \sum_{e,f \in E_2^n} \lambda_e \lambda_f (1 - Y_e)(1 - Y_f) \leq |E_2^n| \sum_{e \in E_2^n} (1 - Y_e),$$

since $\lambda_e, \lambda_f \leq 1$ by (C2*), and $(1 - Y_f) \leq 1$ for all $f \in E_2^n$. Thus,

$$\mathbb{E}[R_n^2] \leq |E_2^n|^2 \cdot \max\{\mathbb{P}[Y_e = 0] : e \in E_2^n\}. \tag{8.6}$$

Now, for any edge $e \in E_2^n$ there are at least $n^\gamma - An^\alpha$ elements x of C_e for which $p_x \geq \epsilon$ (by (C1)), and thus:

$$\mathbb{P}[Y_e = 0] \leq (1-\epsilon)^{n^\gamma - An^\alpha}.$$

Since $|E_2^n| < 2n$, Equation (8.6) and the inequality $\alpha < \gamma$ gives

$$\mathbb{E}[R_n^2] \leq 4n^2 \cdot (1-\epsilon)^{n^\gamma - An^\alpha} \to 0 \text{ as } n \to \infty,$$

as required. □

The next result is used in the proof of Lemma 8.6.

Lemma 8.5. *Under conditions (C1) and (C2*), we have*

$$\sum_{e \in E_p(\mathcal{T}_n)} \lambda_e^2 P_e(1-P_e) \geq B\epsilon^2(1+o(1))n^\beta,$$

where $o(1)$ denotes a term that tends to 0 as $n \to \infty$.

Proof. Let U_n be the set of those pendant edges e of $\mathcal{T}_n$ for which the leaf incident with e has its survival probability in the interval $[\epsilon, 1-\epsilon]$, and let V_n denote the set of the remaining pendant edges of $\mathcal{T}_n$. Clearly,

$$\sum_{e \in E_p(\mathcal{T}_n)} \lambda_e^2 P_e(1-P_e) \geq \epsilon^2 \sum_{e \in U_n} \lambda_e^2, \tag{8.7}$$

and by (C2*) we have

$$Bn^\beta \leq \sum_{e \in E_p(\mathcal{T}_n)} \lambda_e^2 \leq \sum_{e \in U_n} \lambda_e^2 + |V_n| \tag{8.8}$$

where the last term ($|V_n|$) is an upper bound on $\sum_{e \in V_n} \lambda_e^2$ by virtue of the bound $|\lambda_e| \leq 1$ (by (C2*)). Since $|V_n| \leq An^\alpha$, Equations (8.7) and (8.8) give

$$\sum_{e \in E_p(\mathcal{T}_n)} \lambda_e^2 P_e(1-P_e) \geq \epsilon^2(Bn^\beta - An^\alpha) = B\epsilon^2(1+o(1))n^\beta.$$

□

Let $\psi_n = (Z_n - \mathbb{E}[Z_n])/\sqrt{\mathrm{Var}[Z_n]}$. A key step in establishing Theorem 8.3 is the following lemma, which states that ψ_n is asymptotically normal.

Lemma 8.6. $\psi_n \xrightarrow{D} N(0,1)$.

Proof. We can apply a version of the central limit theorem for double arrays of random variables. The required theorem can be found in [66] (Corollary in Section 1.9.3, pp. 31-32) and states the following. For each n, let $X_{n1}, \dots, X_{nr}$ be $r = r(n)$ independent random variables with finite pth moments for some $p > 2$. Let

$$A_n = \sum_j \mathbb{E}[X_{nj}]; \quad B_n = \sum_j \mathrm{Var}[X_{nj}].$$

If

$$B_n^{-p/2} \sum_j \mathbb{E}[|X_{nj} - \mathbb{E}[X_{nj}]|^p] \to 0 \text{ as } n \to \infty, \tag{8.9}$$

then $W_n = (\sum_j X_{nj} - A_n)/\sqrt{B_n} \xrightarrow{D} N(0,1)$. We apply this by taking $\{X_{n1}, \dots, X_{nr}\} = \{\varphi_e^n : e \in E_{12}^n\}$, since the random variables $\{\varphi_e^n : e \in E_{12}^n\}$ are clearly independent. With our notation $Z_n = \sum_{e \in E_{12}^n} \varphi_e^n$, we have $A_n = \mathbb{E}[Z_n]$, $B_n = \mathrm{Var}[Z_n]$ and $W_n = \psi_n$. Thus, we only need to verify condition (8.9) in order to establish Lemma 8.6.

By Corollary 8.2, we have:

$$\mathrm{Var}[\varphi_e^n] \geq \sum_{f \in E_p(t_e)} \lambda_f^2 P_f (1 - P_f).$$

This lower bound and the independence of $\{\varphi_e^n : e \in E_{12}^n\}$, implies:

$$B_n = \mathrm{Var}[Z_n] = \sum_{e \in E_{12}^n} \mathrm{Var}[\varphi_e^n] \geq \sum_{e \in E_{12}^n} \sum_{f \in E_p(t_e)} \lambda_f^2 P_f (1 - P_f)$$

Consequently, by Lemma 8.5, and the fact that every pendant edge occurs in $E_p(t_e)$ for some $e \in E_{12}^n$ we obtain,

$$B_n \geq B\epsilon^2 (1 + o(1)) n^\beta. \tag{8.10}$$

Consider now the absolute central moments in (8.9). We have

$$\mathbb{E}[|X_{nj} - \mathbb{E}[X_{nj}]|^p] = \mathbb{E}[|\varphi_e^n - \mathbb{E}[\varphi_e^n]|^p] \leq L_e^p,$$

where L_e is the sum of the lengths of the edges of t_e. Since t_e has less than $2n_e$ edges, and the edge lengths are bounded from above by 1 (under (C2*)) and $e \in E_{12}^n$ implies $n_e \leq f(n)$, we obtain $L_e \leq 2n_e \leq 2f(n)$. Now we have

$$\mathbb{E}[|\varphi_e^n - \mathbb{E}[\varphi_e^n]|^p] \leq 2^p f(n)^p. \tag{8.11}$$

Combining the bounds (8.10) and (8.11), and noting that $|E_{12}^n| \leq 2n$ and $f(n) = n^\gamma$ we obtain:

$$\begin{aligned} B_n^{-p/2} \sum_{e \in E_{12}^n} \mathbb{E}[|\varphi_e^n - \mathbb{E}[\varphi_e^n]|^p] &\leq \frac{|E_{12}^n| 2^p f(n)^p}{(B\epsilon^2(1+o(1)))^{p/2} n^{\beta p/2}} \\ &\leq C(p) n^{1+p(\gamma-\beta/2)}, \end{aligned}$$

for some constant $C(p) > 0$ independent of n. Now, since $\gamma < \beta/2$, the exponent of n in the obtained upper bound is negative for any $p > (\beta/2 - \gamma)^{-1}$. Since there are some $p > 2$ satisfying this inequality and consequently satisfying condition (8.9), the proof of the lemma is complete. □

We return to the proof of Theorem 8.3. Using Equation (8.5) and the definition of ψ_n, we get

$$\begin{aligned} \frac{\varphi_n - \mathbb{E}[\varphi_n]}{\sqrt{\mathrm{Var}[\varphi_n]}} &= \frac{\lambda_n + Z_n - R_n - (\lambda_n + \mathbb{E}[Z_n] - \mathbb{E}[R_n])}{\sqrt{\mathrm{Var}[\varphi_n]}} \\ &= C_n \psi_n + D_n \end{aligned}$$

where

$$C_n = \frac{\sqrt{\mathrm{Var}[Z_n]}}{\sqrt{\mathrm{Var}[\varphi_n]}} \text{ and } D_n = -\frac{R_n - \mathbb{E}[R_n]}{\sqrt{\mathrm{Var}[\varphi_n]}}.$$

By Lemma 8.4 and the fact that $\mathrm{Var}[\varphi_n]$ does not converge to 0 (by Corollary 8.2, Lemma 8.5, and condition (C2*)), we have:

$$D_n \xrightarrow{P} 0. \tag{8.12}$$

Moreover, by (8.5), $\mathrm{Var}[\varphi_n] = \mathrm{Var}[Z_n] + \mathrm{Var}[R_n] - 2\,\mathrm{Cov}[Z_n, R_n]$, so that

$$C_n^{-2} - 1 = \frac{\mathrm{Var}[R_n]}{\mathrm{Var}[Z_n]} - 2\rho \frac{\sqrt{\mathrm{Var}[R_n]}}{\sqrt{\mathrm{Var}[Z_n]}},$$

where ρ is the correlation coefficient of R_n and Z_n. Now, by Lemma 8.4 we have $\lim_{n\to\infty} \mathrm{Var}[R_n] = 0$. Thus, since $\mathrm{Var}[Z_n]$ is bounded away from 0 (by (8.10)), and $\rho \in [-1, 1]$, we have:

$$\lim_{n\to\infty} C_n = 1. \tag{8.13}$$

To complete the proof of Theorem 8.3 we apply Slutsky's Theorem [17], which states that if X_n, Y_n, W_n are sequences of random variables, and $X_n \xrightarrow{P} a$, $Y_n \xrightarrow{P} b$, (where a, b are constants) and $W_n \xrightarrow{D} W$ (for some random variable W) then $X_n W_n + Y_n \xrightarrow{D} aW + b$. In our setting, we will take $X_n = C_n, Y_n = D_n, W_n = \psi_n$, and $W = N(0, 1)$ (the standard normal random variable). The condition that $\psi_n \xrightarrow{D} N(0,1)$ was established in Lemma 8.6, and the conditions $C_n \xrightarrow{P} 1$, $D_n \xrightarrow{P} 0$ were established in (8.13) and (8.12) (note that the convergence of a sequence of real numbers in (8.13) is just a special case of convergence in probability). Thus,

$$(\varphi_n - \mathbb{E}[\varphi_n])/\sqrt{\mathrm{Var}[\varphi_n]} = C_n\psi_n + D_n \xrightarrow{D} N(0,1),$$

which completes the proof of Theorem 8.3.

□

8.3 Computing the PD distribution

In this section, we describe an algorithm to calculate the distribution of $\varphi_{\mathcal{T}}$ efficiently under the g-FOB model. The approach we present here allows us to derive the exact distribution of $\varphi_{\mathcal{T}}$. Note that we do not require conditions (C1) or (C2) in this section. We make the simplifying assumption that the edge lengths are non-negative integer-valued, which implies that $\varphi_{\mathcal{T}}$ can only have values in the set $\{0, 1, \ldots, L\}$, where $L = \mathrm{PD}(X) = \sum_e \lambda_e$. This assumption is not problematic in practice, as we can rescale all the edge lengths so that they are arbitrarily close to integer multiples of some small value (in doing so we can approximate the correct distribution within any desired precision, as detailed at the end of this section).

We also assume that the input tree is such that the root has one incident edge, and all other non-leaf vertices have exactly three incident edges. This assumption does not affect the generality of our method as any tree can be modified to satisfy it, without changing the distribution for $\varphi_{\mathcal{T}}$. One can resolve multifurcations (interior vertices of degree greater than three) arbitrarily and possibly insert an edge below the root, always assigning length zero to the newly introduced edges.

Consistent with the notation used before, φ_e denotes the contribution to $\varphi_{\mathcal{T}}$ that comes from e and the edges separated from the root by e. Then, for any edge e and integer m, define

$$f_e(m) := \mathbb{P}[\varphi_e = m,\, Y_e = 1].$$

Also recall that $P_e = \mathbb{P}[Y_e = 1]$.

Clearly, if e is the only edge attached to the root of $\mathcal{T}$, then f_e and P_e are all that is needed to derive the distribution of $\varphi_{\mathcal{T}}$: simply observe that

$$\mathbb{P}[\varphi_{\mathcal{T}} = m] = \mathbb{P}[\varphi_e = m,\, Y_e = 1] + \mathbb{P}[\varphi_e = m,\, Y_e = 0] = f_e(m) + (1 - P_e) \cdot I_{m=0},$$

where I_p equals 0 or 1 depending on proposition p being false or true, respectively.

The algorithm then consists in doing a depth-first (bottom-up) traversal of all the edges, so that each time an edge e is visited, the values of P_e and $f_e(m)$, for all $m \in \{\lambda_e, \lambda_e + 1, \dots, L\}$, are calculated using the following recursions. We may then use the P_e and $f_e(m)$ values of the root edge to calculate the distribution of $\varphi_{\mathcal{T}}$.

Recursion for $f_e(m)$

- If e leads into leaf x, then

$$f_e(m) \;=\; \mathbb{P}[\varphi_e = \lambda_e,\, Y_e = 1] \cdot I_{m=\lambda_e} \;=\; p_x \cdot I_{m=\lambda_e}.$$

- If e leads into the tail of edges c and d, then

$$f_e(m) = \sum_{i=\lambda_c}^{m-\lambda_e-\lambda_d} f_c(i) \cdot f_d(-\lambda_e - i) + (1-P_d) \cdot f_c(m-\lambda_e) + (1-P_c) \cdot f_d(m-\lambda_e). \quad (8.14)$$

Note that whenever the term $f_c(m-\lambda_e)$ with $m - \lambda_e < \lambda_c$ or the term $f_d(m-\lambda_e)$ with $m - \lambda_e < \lambda_d$ is used in Equation (8.14), the algorithm will assume that its value is 0 and that therefore, there is no need to calculate and store $f_e(m)$ for m outside the range $\{\lambda_e, \lambda_e + 1, \dots, L\}$.

Equation (8.14) is easily proved. We have

$$\begin{aligned}
f_e(m) &= \mathbb{P}[\varphi_e = m,\, Y_c = 1,\, Y_d = 1] + \mathbb{P}[\varphi_e = m,\, Y_c = 1,\, Y_d = 0] \\
&\quad + \mathbb{P}[\varphi_e = m,\, Y_c = 0,\, Y_d = 1] \\
&= \mathbb{P}[\varphi_c + \varphi_d = m - \lambda_e,\, Y_c = 1,\, Y_d = 1] + \mathbb{P}[\varphi_c = m - \lambda_e,\, Y_c = 1,\, Y_d = 0] \\
&\quad + \mathbb{P}[\varphi_d = m - \lambda_e,\, Y_c = 0,\, Y_d = 1]
\end{aligned}$$

where the second equality is obtained by restating event $\varphi_e = m$ in terms of φ_c and φ_d, which is possible once we make assumptions on Y_c and Y_d. Thus,

$$\begin{aligned}
f_e(m) &= \sum_{i=0}^{m-\lambda_e} \mathbb{P}\left[\varphi_c = i,\, Y_c = 1\right] \cdot \mathbb{P}\left[\varphi_d = m - \lambda_e - i,\, Y_d = 1\right] + \\
&\quad \mathbb{P}[\varphi_c = m - \lambda_e,\, Y_c = 1] \cdot \mathbb{P}[Y_d = 0] + \mathbb{P}[\varphi_d = m - \lambda_e,\, Y_d = 1] \cdot \mathbb{P}[Y_c = 0] \\
&= \sum_{i=\lambda_c}^{m-\lambda_e-\lambda_d} f_c(i) \cdot f_d(m - \lambda_e - i) + (1 - P_d) \cdot f_c(m - \lambda_e) + (1 - P_c) \cdot f_d(m - \lambda_e).
\end{aligned}$$

where the first equality is obtained by using the independence between the survival events in C_c and C_d. Note that in the first expression in the second equality, the range of the sum has been reduced, as $f_c(i) = 0$ for $i < \lambda_c$ and $f_d(m - \lambda_e - i) = 0$ for $m - \lambda_e - i < \lambda_d$.

Recursion for P_e

- If e leads into leaf x, then $P_e = p_x$.
- If e leads into the tail of edges c and d, then $P_e = P_c + P_d - P_c P_d$.

Efficiency considerations

For any given e, the calculation of P_e is done in $O(1)$ time, whereas that of each of the $f_e(m)$ values requires $O(m) = O(L)$ time (see recursion (8.14)), giving a total of $O(L^2)$. Calling n the number of leaves in $\mathcal{T}$, there are $2n-1$ edges in $\mathcal{T}$ and the entire procedure takes $O(nL^2)$ time. Note that this means that the algorithm runs in pseudo-polynomial time. The reason for this is that each integer λ_e is described in the input by a string of length only $O(\log \lambda_e)$. Therefore, the length of the entire input is only $O(n \log L)$, and $O(nL^2)$ is not bounded by any polynomial function of this quantity.

A more efficient (but still pseudo-polynomial-time) version of the algorithm can be obtained by restricting the calculation of $f_e(m)$ to the values of $m \in \{\lambda_e, \lambda_e + 1, \dots, L_e\}$, where L_e is the maximum value that φ_e can attain (namely the sum of the lengths of all the edges separated from the root by e, including e itself). Note that the sum in (8.14) can then be further restricted to the values of i such that $i \leq L_c$ and $m - \lambda_e - i \leq L_d$. Using this more efficient algorithm, it is easy to see that the calculation of all the $f_e(m)$ values for a given internal edge e takes $O(L_c L_d + L_e)$ time, where c and d are the edges that e leads into. Noting that the sum of all the $L_c L_d$ terms, for all sister edges c and d, is bounded above by L^2, this shows that the running time of the entire procedure is $O(L^2 + nL)$. Since typically every pair of taxa in the tree is separated by at least one edge of positive length, we have that $n = O(L)$ and therefore, the running time above is equivalent to $O(L^2)$.

Regarding memory requirements, note that each time we calculate the information relative to e (namely P_e and $f_e(m)$), the information relative to the edges it leads to (if any) can be deleted, as it will never be used again. So, at any given moment the information of at most n active edges needs to be stored. (In practice the maximum number of active edges can be brought down to $O(\log n)$ by organising the depth-first traversal so that for each edge its larger subtree is always traversed first.) If we use the range restriction just described, the sizes of the $f_e(m)$ vectors for all the active edges sum to a number bounded above by $n + L$, and therefore, the algorithm requires $O(n + L)$ space, equivalent to $O(L)$ if $n = O(L)$.

Regarding the assumption that the edge lengths be expressed as integer multiples of some fixed unit, note that in some cases this unit may need to be very small, thus potentially causing L to be very large and our algorithm quite inefficient. In these cases it is possible to produce instead an arbitrarily precise approximation of the distribution of the future PD: if we round each edge length to the nearest integral multiple of ϵ/n, then

$$|\tilde{\varphi} - \varphi| = \sum_e |\lambda'_e - \lambda_e| Y_e \leq \sum_e \frac{\epsilon}{2n} Y_e < \epsilon,$$

where λ'_e is the rounded length of edge e and $\tilde{\varphi}$ is the rounded future PD. In other words, we can achieve any desired precision ϵ by re-expressing each edge length as a multiple of ϵ/n. However, precision is usually measured by number of decimal or binary places of accuracy. To get d decimal places of accuracy we need to take $\epsilon = 10^{-d}$. If L grows linearly with n/ϵ, the running time of our algorithm is exponential in the precision d. This can be problematic if we aim for a large number of decimal places of accuracy.

8.4 Asymptotic normality for unrooted trees

Let $\mathcal{T}$ be an unrooted phylogenetic X-tree and recall the definition of PD for unrooted phylogenetic trees in Definition 2.3. As the g-FOB model is also defined naturally on unrooted trees, it makes sense to consider the distribution of future unrooted PD under the g-FOB model in this setting. A natural question is whether Theorem 8.3 is still valid, (that is, is the future unrooted PD of unrooted trees also asymptotically normal under conditions (C1) and (C2)?). We now answer this question affirmatively and also show how to extend the computation of the exact future PD distribution to unrooted trees.

Let the random variable $\varphi' = \varphi'_{\mathcal{T}}$ denote the PD score of the random subset X' of X (consisting of those taxa that will still exist at some time t in the future). We call φ' the *future unrooted phylogenetic diversity.* In this model, we have

$$\varphi' = \sum_e \lambda_e Y'_e, \tag{8.15}$$

where Y'_e is the binary random variable which takes the value 1 if e lies on a path between some pair of taxa in X', and which is 0 otherwise. Moreover,

$$\mathbb{P}[Y'_e = 1] = (1 - \prod_{x \in X_1(e)} (1 - p_x))(1 - \prod_{x \in X_2(e)} (1 - p_x)), \tag{8.16}$$

where $X_1(e)$ and $X_2(e)$ are the two parts of the bipartition of X consisting of the two subsets of X that are separated by edge e. Thus, if we let $P_i(e)$ denote the probability that at least one taxon in $X_i(e)$ survives (for $i \in \{1, 2\}$), then the expected value of φ' (analogous to (8.3)) is

$$\mathbb{E}[\varphi'] = \sum_e \lambda_e P_1(e)P_2(e). \tag{8.17}$$

Regarding $\mathrm{Var}[\varphi']$, there is an analogous formula to that given in Lemma 8.1.

Consider now a sequence $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n, \ldots$ of unrooted phylogenetic trees where $\mathcal{T}_n$ has n leaves, and assume that this sequence satisfies conditions (C1) and (C2) when each $\mathcal{T}_n$ has associated edge length and leaf survival probability functions. As we shall see, Theorem 8.3 is still valid for unrooted PD; that is, under the same conditions, $(\varphi'_n - \mathbb{E}[\varphi'_n])/\sqrt{\mathrm{Var}[\varphi'_n]}$ converges in distribution to $N(0, 1)$ as $n \to \infty$.

To establish this asymptotic normality of φ'_n under conditions (C1) and (C2*) (and thereby (C1) and (C2)) requires slight modifications to the proof of Theorem 8.3, and we now provide an outline of the argument. The main difference is that now each edge e induces a bipartition $X = X_1(e) \cup X_2(e)$ of the taxon set and so we decompose $\mathcal{T}_n$ in a slightly different way. For simplicity, assume that $|X_1(e)| \leq |X_2(e)|$ and consider the following edge sets (the definition of the function $f(n)$ is as in the rooted case):

- E_1^n: edges e of $\mathcal{T}_n$ with $|X_1(e)| \leq f(n)$;
- E_2^n: edges e of $\mathcal{T}_n$ with $|X_1(e)| > f(n)$;
- E_{12}^n: edges $e \in E_1^n$ such that e is adjacent to an edge $f \in E_2^n$.

For φ'_n, we obtain the following equation:

$$\varphi'_n = \sum_{e \in E_1^n} \lambda_e Y'_e + \sum_{e \in E_2^n} \lambda_e Y'_e = \sum_{e \in E_1^n} \lambda_e Y'_e + \lambda_n - R'_n, \tag{8.18}$$

where $\lambda_n = \sum_{e \in E_2^n} \lambda_e$ and $R'_n = \sum_{e \in E_2^n} \lambda_e(1 - Y'_e)$. For an edge $e \in E_{12}^n$, let t_e denote the subtree with root edge e and with leaf set $X_1(e)$. Let $(\varphi_e^n)'$ denote the contribution to φ'_n by the edges in t_e. Furthermore, let φ_e^n be the rooted future phylogenetic diversity of t_e, $Z_n = \sum_{e \in E_{12}^n} \varphi_e^n$ as in the rooted case, $W_e = \varphi_e^n - (\varphi_e^n)'$, and $V_n = \sum_{e \in E_{12}^n} W_e$. With this notation, we get

$$\varphi'_n = \sum_{e \in E_{12}^n} (\varphi_e^n)' + \lambda_n - R'_n = \sum_{e \in E_{12}^n} \varphi_e^n - \sum_{e \in E_{12}^n} W_e + \lambda_n - R'_n = Z_n - V_n + \lambda_n - R'_n. \tag{8.19}$$

Now we can apply Lemma 8.6 and Slutsky's Theorem to complete the proof.

8.5 Computing the unrooted PD distribution

Finally, we show how the algorithm described in Section 8.3 for computing the PD distribution can be modified to calculate the distribution of unrooted PD. As before, we assume the edge lengths are non-negative integers and we preprocess $\mathcal{T}$ (possibly rooting it in an arbitrary vertex) so that the number of outgoing edges is 1 for the root and 2 for all the other non-leaf vertices. Since $\mathcal{T}$ is now rooted, C_e and the random variables Y_e are well defined. We also define φ'_e as the unrooted PD of the surviving taxa in C_e. Then, for any integer m, define

$$f'_e(m) := \mathbb{P}[\varphi'_e = m,\, Y_e = 1].$$

As before, if e is the root edge of $\mathcal{T}$, then f'_e and P_e are sufficient to derive the distribution of $\varphi'_{\mathcal{T}}$:

$$\mathbb{P}[\varphi'_{\mathcal{T}} = m] = f'_e(m) + (1 - P_e) \cdot I_{m=0}.$$

An algorithm to calculate the distribution of $\varphi'_{\mathcal{T}}$ can be obtained with a simple modification of the algorithm for $\varphi_{\mathcal{T}}$: for each edge e, in addition to calculating P_e and $f_e(m)$, also calculate $f'_e(m)$, for all $m \in \{0, 1, \ldots, L\}$. For this purpose, the following recursion is used (note that the f'_e values may depend on f_c and f_d as well as on f'_c and f'_d, which is why we retain the calculation of the f_e values even though they are not directly implicated in determining $\mathbb{P}[\varphi'_{\mathcal{T}} = m]$).

Recursion for $f'_e(m)$

- If e leads into leaf x, then
$$f'_e(m) \;=\; p_x \cdot I_{m=0}.$$
- If e leads into the tail of edges c and d, then
$$f'_e(m) = \sum_{i=\lambda_c}^{m-\lambda_d} f_c(i) \cdot f_d(m-i) + (1 - P_d) \cdot f'_c(m) + (1 - P_c) \cdot f'_d(m), \tag{8.20}$$

which is proved in a way similar to (8.14):

$$\begin{aligned} f'_e(m) &= \mathbb{P}[\varphi'_e = m,\, Y_c = 1,\, Y_d = 1] + \mathbb{P}[\varphi'_e = m,\, Y_c = 1,\, Y_d = 0] \\ &\quad + \mathbb{P}[\varphi'_e = m,\, Y_c = 0,\, Y_d = 1] \\ &= \mathbb{P}[\varphi_c + \varphi_d = m,\, Y_c = 1,\, Y_d = 1] + \mathbb{P}[\varphi'_c = m,\, Y_c = 1,\, Y_d = 0] \\ &\quad + \mathbb{P}[\varphi'_d = m,\, Y_c = 0,\, Y_d = 1] \\ &= \sum_{i=\lambda_c}^{m-\lambda_d} f_c(i) \cdot f_d(m-i) + (1 - P_d) \cdot f'_c(m) + (1 - P_c) \cdot f'_d(m). \end{aligned}$$

8.6 Concluding remarks

The main result of this chapter (Theorem 8.3) has been to establish a limiting normal distribution for future PD on large phylogenetic trees. This theorem assumes an underlying generalized field of bullets model, and imposes two further mild conditions (conditions (C1) and (C2)). In this setting Theorem 8.3 reduces the problem of computing the distribution of future PD to that of determining just two parameters—its mean and variance—and these can be readily computed by Equation (8.3) and Lemma 8.1. Using the resulting normal distribution one can easily compute the probability that future PD will fall below any given critical value. This may also be helpful in designing strategies to minimize this probability, analogous to the Noah's Ark problem, which tries to maximize expected future PD.

In practice, the use of a normal distribution based on Theorem 8.3 requires that the number of taxa is moderate (> 50), that the survival probabilities are not too extreme (condition (C1)), and that the length of the pendant edges on average are not too small in relation to the largest edge length in the tree (condition (C2)). If these conditions are violated, it would be prudent to use the exact algorithm we have described, as this requires neither a large number of taxa nor condition (C1) or (C2). To apply this algorithm may involve some small adjustment to the edge lengths to make them integral multiples of some common value.

We also showed that neither (C1) nor (C2) can be dropped completely from the statement of Theorem 8.3. However, it is likely that both conditions could be weakened somewhat, though at the risk of complicating their description and the proof of the theorem.

The paper entitled 'Distribution of phylogenetic diversity under random extinction' [25] is a result of the work presented in this chapter. I would like to thank my co-authors Fabio Pardi and Mike Steel for their educative collaboration on my first paper.

Chapter 9

Comparing expected PD loss under g-FOB and s-FOB

In this chapter, we investigate a generic inequality that applies to two-state Markov processes on trees, and provide two applications. In the first application, we consider the expected loss of phylogenetic diversity under a model in which extinction risk is associated with an underlying state that evolves on the tree. We are interested in comparing this expected loss to the expected loss under the g-FOB model, in which extinction events are treated independently. We find that when extinction events reflect phylogenetic history, then the expected loss of phylogenetic diversity is always greater than or equal to that predicted by an independent extinction scenario. In a second application, we derive a new, purely combinatorial result concerning the parsimony score of a binary character on a tree.

9.1 Two-state Markov processes on trees

Consider a Markov random field on a tree T with state space $\{0,1\}$, and for each vertex v of T, let $\xi(v)$ be the random state (0 or 1) that v is assigned. This process is usually described as follows. We have a root vertex ρ for which we specify a probability, say π_i, that $\xi(\rho)=i$, for $i \in \{0,1\}$. Direct all the edges of T away from ρ and for any arc (r,s) of the resulting directed tree $T=(V_T,A_T)$, let $P^{(r,s)}$ denote the 2×2 transition matrix for which the ij-entry (for $i,j \in \{0,1\}$) is the conditional probability that $\xi(s)=j$ given that $\xi(r)=i$. Specifying $\pi=[\pi_0,\pi_1]$ together with the transition matrices $P^{(r,s)}$ for all the arcs (r,s) of T uniquely defines the Markov random field on T (see Section 7.2 or [11, 65, 69]); an explicit formula appears below (Equation (9.1)). We will assume throughout that π is strictly positive and that $\det P^{(r,s)} \geq 0$ holds for each transition matrix. Notice that this determinant condition automatically holds if one views the transition matrix for an arc as describing the net effect of a continuous-time Markov process operating for some duration for that arc. Note however that we are not assuming that any such process is the same between the arcs of T.

For $U \subseteq V_T$, let $P(U)$ denote the probability that U is precisely the set of vertices of

T in state 0; that is: $P(U) = \mathbb{P}(\{v \in V_T : \xi(v) = 0\} = U)$. To express $P(U)$ in terms of the transition matrices and π, let $\delta(U, v) = 0$ if $v \in U$ and let $\delta(U, v) = 1$ if $v \in V_T - U$. Then, the Markov property gives:

$$P(U) = \pi_{\delta(U,\rho)} \cdot \prod_{(r,s)\in A_T} P^{(r,s)}_{\delta(U,r)\delta(U,s)}. \tag{9.1}$$

For any subset W of the leaf set X of T, let p_W denote the probability that W is precisely the set of leaves of T that are in state 0. This marginal probability is:

$$p_W = \sum_{U \in \mathscr{A}_W} P(U), \text{ where } \mathscr{A}_W := \{U \subseteq V_T : U \cap X = W\}. \tag{9.2}$$

A number of authors have noticed that certain inequalities hold for quadratic functions of the p_W values. For example, for any $x, y \in X$ with $x \neq y$, it is well known that $p_{\{x\}} \cdot p_{\{y\}} \leq p_{\{x,y\}} \cdot p_\emptyset$. Moreover, in [56] the following inequality was described: for subsets $\{x, y\}$ and $\{x, z\}$ of X where x, y, z are distinct, we have $p_{\{x,y\}}p_{\{x,z\}} \leq p_{\{x,y,z\}}p_{\{x\}}$. We now provide a much more general inequality.

Proposition 9.1. *For any two-state Markov process on a tree with leaf set X, and any two subsets Y, Z of X, we have $p_Y \cdot p_Z \leq p_{Y\cup Z} \cdot p_{Y\cap Z}$.*

Proof. Let A, B be arbitrary subsets of V_T. We first establish the following:

$$P(A) \cdot P(B) \leq P(A \cup B) \cdot P(A \cap B). \tag{9.3}$$

Applying Equation (9.1) to $U \in \{A, B, A \cup B, A \cap B\}$, the product $P(A) \cdot P(B)$ and the product $P(A\cup B)\cdot P(A\cap B)$ can each be written as a product of two entries of π multiplied by a product over the arcs (r, s) of T of two entries of $P^{(r,s)}$. Moreover, regardless of where r and s lie in relation to the sets A, B, the product of the two π terms agree in $P(A)\cdot P(B)$ and $P(A\cup B)\cdot P(A\cap B)$ (i.e., we have $\pi_{\delta(A,\rho)}\pi_{\delta(B,\rho)} = \pi_{\delta(A\cup B,\rho)}\pi_{\delta(A\cap B,\rho)}$), while the product of the two $P^{(r,s)}$ terms agree in $P(A) \cdot P(B)$ and $P(A \cup B) \cdot P(A \cap B)$, except for the cases in which either (i) $r \in A - B$ and $s \in B - A$ or (ii) $r \in B - A$ and $s \in A - B$. However, in both case (i) and (ii), the product $P^{(r,s)}_{01}P^{(r,s)}_{10}$ appears in the term for $P(A) \cdot P(B)$ while $P^{(r,s)}_{00}P^{(r,s)}_{11}$ appears in the term for $P(A \cup B) \cdot P(A \cap B)$, and the former term is less or equal to the second since $P^{(r,s)}_{00}P^{(r,s)}_{11} - P^{(r,s)}_{01}P^{(r,s)}_{10} = \det P^{(r,s)}$ and $\det P^{(r,s)} \geq 0$ by assumption. Consequently, all the terms in $P(A) \cdot P(B)$ are either less than or equal to (in cases (i) and (ii)) or equal to (in all remaining cases) the corresponding terms in $P(A \cup B) \cdot P(A \cap B)$. This establishes (9.3).

We now invoke a classical result of Ahlswede and Daykin [2] from 1978, sometimes called the 'four functions theorem'. A particular form of this theorem that suffices for our purposes is the following (we follow [3]). Suppose we have a finite set S and a function α that assigns a non-negative real number to each subset of S. Suppose that α satisfies the property that for all subsets A, B of S:

$$\alpha(A)\alpha(B) \leq \alpha(A \cup B)\alpha(A \cap B).$$

For a collection $\mathscr{C}$ of subsets of S, let $\alpha(\mathscr{C}) := \sum_{C \in \mathscr{C}} \alpha(C)$. Then for any two collection of subsets of S, $\mathscr{A}$ and $\mathscr{B}$, say, we have:

$$\alpha(\mathscr{A})\alpha(\mathscr{B}) \leq \alpha(\mathscr{A} \vee \mathscr{B})\alpha(\mathscr{A} \wedge \mathscr{B}), \tag{9.4}$$

where $\mathscr{A} \vee \mathscr{B} := \{E \subseteq S : E = A \cup B : A \in \mathscr{A}, B \in \mathscr{B}\}$, and where $\mathscr{A} \wedge \mathscr{B} := \{E \subseteq S : E = A \cap B : A \in \mathscr{A}, B \in \mathscr{B}\}$. We will apply this to our problem by taking $S = V_T, \alpha(U) = P(U)$ and noting that α satisfies the required hypothesis by (9.3). By the definition of $\mathscr{A}_W$ in (9.2), $\mathscr{A}_Y \vee \mathscr{A}_Z = \mathscr{A}_{Y \cup Z}$ and $\mathscr{A}_Y \wedge \mathscr{A}_Z = \mathscr{A}_{Y \cap Z}$. Thus, taking $\mathscr{A} = \mathscr{A}_Y$ and $\mathscr{B} = \mathscr{A}_Z$ in (9.4) we have $\alpha(\mathscr{A}_Y)\alpha(\mathscr{A}_Z) \leq \alpha(\mathscr{A}_{Y \cup Z})\alpha(\mathscr{A}_{Y \cap Z})$. The proposition now follows by observing that $p_W = \alpha(\mathscr{A}_W)$ for all subsets W of X, in particular the subsets $Y, Z, Y \cup Z$ and $Y \cap Z$. □

9.2 Expected PD loss under g-FOB and s-FOB

We first show how Proposition 9.1, together with another inequality, provides a general inequality concerning the loss of expected future biodiversity under species extinction models.

Suppose that $\mathcal{T}$ is a rooted phylogenetic X-tree, and with each arc $e = (u, v)$ of $\mathcal{T}$ there is an associated length λ_e. Recall the measure phylogenetic diversity as defined in Definition 2.4.

For each species $x \in X$, let E_x denote the event that species x is extinct at some future time t. Then, the expected phylogenetic diversity of the species that are extant at time t, referred to as *expected future PD* and denoted $\mathbb{E}[\varphi]$, is given by:

$$\mathbb{E}[\varphi] = \sum_{e=(u,v) \in A_{\mathcal{T}}} \lambda_e \cdot (1 - \mathbb{P}(\bigcap_{x \in C_v} E_x)) = \varphi_X - \sum_{e=(u,v) \in A_{\mathcal{T}}} \lambda_e \cdot \mathbb{P}(\bigcap_{x \in C_v} E_x), \tag{9.5}$$

where C_v denotes the subset of X which is separated from the root by v and which equals $\{v\}$ if v is a leaf vertex. Recall that the generalized field of bullets model (g-FOB) (generalizing an earlier model from [49]) assumes that the events E_x are independent.

Then, if we let $p_x = \mathbb{P}(E_x)$, the value of $\mathbb{P}(\bigcap_{x \in C_v} E_x)$ in (9.5) (the probability of the extinction of all the species descended from v) is given by:

$$\mathbb{P}(\bigcap_{x \in C_v} E_x) = \prod_{x \in C_v} p_x. \tag{9.6}$$

The assumption that the events E_x are independent is likely to be unrealistic in most settings (see, for example, [36, 67]). For example, species close together in $\mathcal{T}$ are more likely to share attributes that may put them at risk in a hostile future environment. As one topical scenario, consider extinction risk due to climate change. Suppose that the extinction risk of each species in X is partially influenced by some associated binary state (0 or 1) where state 0 confers an elevated risk of extinction under climate change. We suppose that these states are not known in advance for the species in X, and that this state has evolved under some Markovian model on $\mathcal{T}$. Once the states are determined at the leaves, then extinction proceeds according to the g-FOB model, where species x is extinct at time t with probability p_x^i if it is in state $i \in \{0, 1\}$. We call this a *state-based field of bullets* model (s-FOB). Note that this includes the g-FOB model as a special case where $p_x^0 = p_x^1$ for all x. Moreover, once we condition on the state for each leaf, an s-FOB model is just a g-FOB model with modified extinction probabilities, but we are assuming that these states are unknown (in line with the uncertainty over what features may be helpful for an organism in a future climate).

With any s-FOB model we also have an associated g-FOB model in which the extinction probability of each species x is the same as in the s-FOB model. That is, in the g-FOB model we set:

$$p_x = p_x^0 \mathbb{P}(\xi(x) = 0) + p_x^1 \mathbb{P}(\xi(x) = 1), \tag{9.7}$$

where ξ describes the Markov process for the binary character. A natural question arises: how does the future expected PD of an s-FOB model compare with that of its associated g-FOB model? The following result provides a general inequality.

Theorem 9.2. *Consider a fixed rooted phylogenetic X-tree with branch lengths. Consider an s-FOB model, in which state 1 is advantageous for each species; that is, $p_x^1 \leq p_x^0$ for all $x \in X$. Then, the expected future PD of this model is less than or equal to the expected future PD of the associated g-FOB model.*

Proof. In view of (9.5) and (9.6), it suffices to show that:

$$\prod_{x \in C_v} p_x \leq \mathbb{P}(\bigcap_{x \in C_v} E_x), \tag{9.8}$$

where p_x is defined by Equation (9.7). For each subset W of C_v, let p_W denote the probability that the set of elements of C_v in state 0 is precisely W. Then, $\mathbb{P}(\bigcap_{x\in C_v} E_x) = \sum_{W\subseteq C_v} p_W \prod_{x\in W} p_x^0 \prod_{x\in C_v - W} p_x^1$. Thus, if we let $f_x(W) = p_x^0$ if $x \in W$, and $f_x(W) = p_x^1$ if $x \in C_v - W$, then $\mathbb{P}(\bigcap_{x\in C_v} E_x) = \sum_{W\subseteq C_v} p_W \prod_{x\in C_v} f_x(W)$. Moreover, $p_x = p_x^0\mathbb{P}(\xi(x) = 0) + p_x^1\mathbb{P}(\xi(x) = 1) = \sum_{W\subseteq C_v} p_W f_x(W)$, where the second equality arises by considering in the summation those W containing x and those not containing x. Consequently, (9.8) is equivalent to the requirement that:

$$\prod_{x\in C_v}\left(\sum_{W\subseteq C_v} p_W f_x(W)\right) \leq \sum_{W\subseteq C_v} p_W \prod_{x\in C_v} f_x(W). \tag{9.9}$$

The proof of (9.9) involves combining Proposition 9.1 with the FKG inequality of Fortuin, Kasteleyn and Ginibre (1971) [31], a particular (and multivariate) form of which we now recall. Given a finite set S, suppose that $f_1, f_2, \ldots, f_n$ are functions from the power set of S into the non-negative real numbers, and that these satisfy the condition:

$$A \subseteq B \Rightarrow f_i(A) \leq f_i(B). \tag{9.10}$$

Furthermore, suppose that μ is a probability measure on the subsets of S which satisfies the log-supermodularity condition:

$$\mu(A)\mu(B) \leq \mu(A\cup B)\mu(A\cap B). \tag{9.11}$$

Then:

$$\prod_{i=1}^{n}\left(\sum_A \mu(A) f_i(A)\right) \leq \sum_A \mu(A) \prod_{i=1}^{n} f_i(A), \tag{9.12}$$

where the summations are over all subsets of S.

We apply this form of the FKG inequality by taking $S = \{1, \ldots, n\} = C_v$, $\mu(W) = p_W$, and f_x as defined above. Then, f_x satisfies (9.10) by the hypothesis that $p_x^1 \leq p_x^0$ for all x, while μ satisfies (9.11) by Proposition 2.1. Thus, inequality (9.12) provides the required inequality (9.9). This completes the proof. □

9.3 Combinatorics of parsimony

We now provide a second application of Proposition 9.1 to phylogenetics. Given a function $f : X \to \{0, 1\}$, the *parsimony score* of f on a tree T with leaf set X, denoted $l(f, T)$, is the minimum number of edges that have different states assigned to their endpoints, across all extensions $F : V_T \to \{0, 1\}$ of f (for further details see [65]). For $W \subseteq X$, let function f_W assign state 0 to the elements of W, and state 1 to the elements of $X - W$. We show that the parsimony score function for a given tree is submodular.

Theorem 9.3. *For any tree T with leaf set X and subsets Y, Z of X, we have:*

$$l(f_Y, T) + l(f_Z, T) \geq l(f_{Y \cup Z}, T) + l(f_{Y \cap Z}, T).$$

Proof. Consider the two-state Markov random field on T with $\pi_0 = \pi_1 = 0.5$, and set each transition matrix $P^{(r,s)}$ to be the symmetric 2×2 matrix with off-diagonal entry $\epsilon > 0$. Then, for any $W \subseteq X$, a straightforward calculation shows that:

$$p_W = C_W \epsilon^{l(f_W,T)}(1 + o(\epsilon)), \tag{9.13}$$

for a constant C_W that depends only on W and T and not ϵ (specifically, C_W is the number of minimal extensions of f_W to the vertices of T multiplied by $\frac{1}{2}$). Now Proposition 9.1, expressed using logarithms, states that:

$$-\log(p_Y) - \log(p_Z) \geq -\log(p_{Y \cup Z}) - \log(p_{Y \cap Z}). \tag{9.14}$$

Applying (9.13) (and noting that $\log(1 + o(\epsilon)) = o(\epsilon)$), the left-hand side of (9.14) is:

$$(l(f_Y, T) + l(f_Z, T)) \log\left(\frac{1}{\epsilon}\right) - \log(C_Y C_Z) + o(\epsilon),$$

while the right-hand side of (9.14) is:

$$(l(f_{Y \cup Z}, T) + l(f_{Y \cap Z}, T)) \log\left(\frac{1}{\epsilon}\right) - \log(C_{Y \cup Z} C_{Y \cap Z}) + o(\epsilon).$$

Theorem 9.3 now follows by letting ϵ tend to zero. □

Concluding remark. As a further phylogenetic application, we note that Proposition 9.1 provides a collection of polynomial inequalities on the p_W values, which have recently been studied for a particular class of Markov two-state models in [44]. These polynomial inequalities complement the much-studied phylogenetic invariants (polynomial identities in the p_W values), which hold under various restrictions on the Markov model.

Most of the results that have been presented in this chapter were proved by Mike Steel. These results led to the publication 'Markovian log-supermodularity, and its applications in phylogenetics' [71]. I would like to thank Mike for his educative and motivating collaboration.

Chapter 10

Mean and variance of future PD under g-FOB and t-FOB

In this chapter, we establish a generic Markov inequality for multivariate Markov processes that consist of k independent but not necessarily identical two-state Markov processes on a tree. The inequality has been specifically designed for the purpose of comparing a new species extinction model with existing ones in conservation biology. This new model is the generalized version of the s-FOB model from the previous chapter, in which the extinction risk of a species is associated with an underlying state that evolves on an evolutionary tree. In the more general setting, extinction is influenced by k independently evolved traits rather than only one, giving a more realistic model.

We compare the expected loss and the variance of phylogenetic diversity under this model to the corresponding values of the g-FOB model. We show that when extinction events reflect the evolutionary history of many characteristics, the expected loss of phylogenetic diversity is greater than or equal to that predicted under a model with independent extinction events. This generalizes the result presented in Section 9.2, and suggests that simple models that treat species extinctions independently may systematically underestimate the loss of phylogenetic diversity.

Given this inequality between the expected future phylogenetic diversity under these two models, we might expect a similar inequality to apply for the variance. However, we show that there is no similar relationship between the variances corresponding to the two models. There are examples for which the variance of future phylogenetic diversity under an independent extinction scenario (g-FOB) can be either smaller or greater than the variance under the model in which extinction events are influenced by k characteristics, even for $k = 1$.

In the next section, we define the multivariate Markov processes under scrutiny and then state and prove the Markov inequality. To demonstrate the phylogenetic application, Section 10.2 presents the inequality between the expected loss of phylogenetic diversity and Section 10.3 summarizes our findings concerning the variance of future phylogenetic diversity.

10.1 An inequality for Markov processes on trees

Let T be a rooted tree with root vertex ρ and with leaf set X. Consider k independent, non-identical two-state Markov processes on T, each of which with the state space $\{0,1\}$ (for a formal definition, see Section 7.2 or [11, 29, 65, 69]). For each vertex v of T and for $j=1,\ldots,k$, let $\xi_j(v)$ denote the random state that v is assigned in the jth Markov process. Furthermore, for $j=1,\ldots,k$ and for $i\in\{0,1\}$, let $\pi_i^{(j)}$ be the probability that $\xi_j(\rho)=i$. Viewing the edges of T as arcs directed away from the root, let $P^{(j)}(r,s)$ be the transition matrix assigned to arc (r,s) in the jth process. The il-entry $P^{(j)}(r,s)_{il}$ of this 2×2 matrix is, by definition, the conditional probability that $\xi_j(s)=l$ given that $\xi_j(r)=i$. For each j, having specified the probabilities $\pi_i^{(j)}$ and the transition matrices $P^{(j)}(r,s)$, $i\in\{0,1\}, (r,s)\in A_T$ (the arc set of T), the jth Markov process on T is uniquely defined.

We now combine these k Markov processes into a vector (having jth coordinate ξ_j) to provide a multivariate Markov process on T with state space $\{0,1\}^k$. In this process, each vertex v of T is assigned state $\boldsymbol{\xi}(v)=(\xi_1(v),\ldots,\xi_k(v))$. Let $\mathbf{i}=(i_1,\ldots,i_k)\in\{0,1\}^k$ and let $\pi_{\mathbf{i}}$ be the probability that $\boldsymbol{\xi}(\rho)=\mathbf{i}$. Then, by the independence of the k processes, we get $\pi_{\mathbf{i}}=\prod_{j=1}^k \pi_{i_j}^{(j)}$. Similarly, for the transition matrix $P(r,s)$ corresponding to arc (r,s) in the multivariate process, the entry $P(r,s)_{\mathbf{il}}$ in 'row $\mathbf{i}$' and 'column $\mathbf{l}$' (for $\mathbf{i}=(i_1,\ldots,i_k), \mathbf{l}=(l_1,\ldots,l_k)\in\{0,1\}^k$) becomes $\prod_{j=1}^k P^{(j)}(r,s)_{i_jl_j}$. This is the conditional probability that $\boldsymbol{\xi}(s)=\mathbf{l}$ given that $\boldsymbol{\xi}(r)=\mathbf{i}$. With these, the multivariate Markov process is uniquely defined.

We will assume throughout that all the π values are strictly positive and that the determinant of $P^{(j)}(r,s)$ is non-negative for each arc (r,s) and for each j. Note that this implies that $\det P(r,s)\geq 0$. Namely, it can be seen that $P(r,s)$ is the *Kronecker product* of the k matrices $P^{(j)}(r,s)$, and so $\det P(r,s)=(\det P^{(1)}(r,s)\times\ldots\times\det P^{(k)}(r,s))^{2^{k-1}}$ (see [38] for the definition and properties of the Kronecker product). However, we are neither assuming that any of the k processes are identical nor that within any of them, the arcs are assigned the same transition matrix.

Consider now a realization $\mathbf{U}=(U_1,\ldots,U_k)$ of $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_k)$. Note that $\mathbf{U}$ is a function from $V=V_T$ (the vertex set of T) into the set $\{0,1\}^k$ of character states. Let $P(\mathbf{U})$ denote the probability that $\boldsymbol{\xi}=\mathbf{U}$; that is, the probability that for each $v\in V$, v is assigned $\mathbf{U}(v)$. For $j=1,\ldots,k$, let $\delta_j(\mathbf{U},v)=0$ if the jth coordinate $U_j(v)$ of $\mathbf{U}(v)$ is 0 and let $\delta_j(\mathbf{U},v)=1$ if $U_j(v)=1$. Also, let $\delta(\mathbf{U},v)$ denote the state that v is assigned in $\mathbf{U}$. Now we are able to express $P(\mathbf{U})$ in terms of the transition matrices and the π values of the multivariate process, using the Markov property (we follow [65]). We have:

$$P(\mathbf{U}) \;=\; \pi_{\delta(\mathbf{U},\rho)} \cdot \prod_{(r,s)\in A_T} P(r,s)_{\delta(\mathbf{U},r)\delta(\mathbf{U},s)},$$

which, by the independence of the k two-state processes, gives:

$$\begin{aligned} P(\mathbf{U}) &= \prod_{j=1}^{k} \pi^{(j)}_{\delta_j(\mathbf{U},\rho)} \cdot \prod_{(r,s)\in A_T} \prod_{j=1}^{k} P^{(j)}(r,s)_{\delta_j(\mathbf{U},r)\delta_j(\mathbf{U},s)} \qquad (10.1) \\ &= \prod_{j=1}^{k} \left(\pi^{(j)}_{\delta_j(\mathbf{U},\rho)} \prod_{(r,s)\in A_T} P^{(j)}\,(r,s)_{\delta_j(\mathbf{U},r)\delta_j(\mathbf{U},s)} \right). \end{aligned}$$

Recall that a *lattice* $\mathscr{L}$ is a partially ordered set in which any two elements $a, b \in \mathscr{L}$ have a unique least upper bound $a \vee b$, called their *join*, and a unique greatest lower bound $a \wedge b$, which is their *meet*. A lattice is *distributive* if $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$ for all $a, b, c \in \mathscr{L}$ or equivalently $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$ for all $a, b, c \in \mathscr{L}$.

Let $\mathscr{L}_V$ be the set of all possible realizations of $\boldsymbol{\xi}$. Let $\mathbf{Y}, \mathbf{Z} \in \mathscr{L}_V$, and let $\leq$ be the partial order over $\mathscr{L}_V$ in which $\mathbf{Y} \leq \mathbf{Z}$ whenever $Y_j(v) \leq Z_j(v)$ for each vertex $v \in V$ and for each $j = 1, \ldots, k$, and in which $\mathbf{Y}$ and $\mathbf{Z}$ are incomparable otherwise. Clearly, any two elements $\mathbf{Y}$ and $\mathbf{Z}$ of the partially ordered set $(\mathscr{L}_V, \leq)$ have a join $\mathbf{Y} \vee \mathbf{Z}$ and a meet $\mathbf{Y} \wedge \mathbf{Z}$. The join $\mathbf{Y} \vee \mathbf{Z}$ is the realization of $\boldsymbol{\xi}$ that assigns to each vertex $v \in V$ state $(\max\{Y_1(v), Z_1(v)\}, \ldots, \max\{Y_k(v), Z_k(v)\})$, while the meet $\mathbf{Y} \wedge \mathbf{Z}$ is the realization of $\boldsymbol{\xi}$ that assigns to each vertex $v \in V$ state $(\min\{Y_1(v), Z_1(v)\}, \ldots, \min\{Y_k(v), Z_k(v)\})$. It follows that $(\mathscr{L}_V, \vee, \wedge)$ is a lattice on $\mathscr{L}_V$. It is easy to see that this lattice is distributive.

Recall that X denotes the leaf set of T and fix a non-empty subset W of X. For each function $\mathbf{U}$ in $\mathscr{L}_V$, define $\mathbf{u} = (u_1, \ldots, u_k)$ to be the restriction of $\mathbf{U}$ to W; that is, $\mathbf{u} = \mathbf{U}|_W$. With this we have $\mathbf{u}(v) = \mathbf{U}(v)$ for each leaf v in W. Since $\mathbf{u}$ is a function from the non-empty subset W of X into a set of character states, it is a character on X (see Definition 7.1). Let $\mathscr{L}_W$ be the set that contains, for each $\mathbf{U} \in \mathscr{L}_V$, the restricted function $\mathbf{u} = \mathbf{U}|_W$. Let $\mathbf{y}, \mathbf{z} \in \mathscr{L}_W$, and let $\leq$ be the partial order over $\mathscr{L}_W$ such that if $y_j(v) \leq z_j(v)$ for each $v \in W$ and for each $j = 1, \ldots, k$, we have $\mathbf{y} \leq \mathbf{z}$; otherwise $\mathbf{y}$ and $\mathbf{z}$ are incomparable. The join $\mathbf{y} \vee \mathbf{z}$ and the meet $\mathbf{y} \wedge \mathbf{z}$ can be obtained for any two elements $\mathbf{y}, \mathbf{z}$ of $\mathscr{L}_W$ analogously to the case of $\mathscr{L}_V$, defining the finite distributive lattice $(\mathscr{L}_W, \vee, \wedge)$. Now let $p(\mathbf{u})$ be the probability that for each leaf v in W, v is assigned $\mathbf{u}(v)$.

This marginal probability is given by:

$$p(\mathbf{u}) = \sum_{\mathbf{U}\in\mathscr{A}_{\mathbf{u}}} P(\mathbf{U}), \text{ where } \mathscr{A}_{\mathbf{u}} := \{\mathbf{U} \in \mathscr{L}_V : \mathbf{U}|_W = \mathbf{u}\}. \qquad (10.2)$$

An example to illustrate this concept is provided in Figure 10.1. Let $k = 1$ and let $\mathbf{u}$ be denoted by u. In this example, if $W = \{a, b, c, d\}$, $u(a) = u(c) = 0$, and $u(b) = u(d) = 1$, then:

$$\begin{aligned} p(u) &= \pi_0 P(\rho, a)_{00} P(\rho, b)_{01} P(\rho, s)_{00} P(s, c)_{00} P(s, d)_{01} \\ &+ \pi_0 P(\rho, a)_{00} P(\rho, b)_{01} P(\rho, s)_{01} P(s, c)_{10} P(s, d)_{11} \\ &+ \pi_1 P(\rho, a)_{10} P(\rho, b)_{11} P(\rho, s)_{10} P(s, c)_{00} P(s, d)_{01} \\ &+ \pi_1 P(\rho, a)_{10} P(\rho, b)_{11} P(\rho, s)_{11} P(s, c)_{10} P(s, d)_{11}. \end{aligned}$$

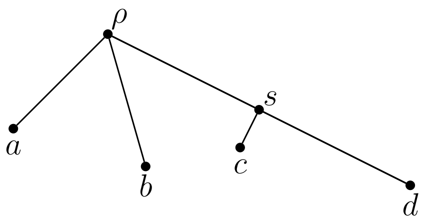


Figure 10.1: A rooted tree with leaf set $\{a, b, c, d\}$.

The following proposition extends Proposition 9.1, which dealt with the special case $k = 1$.

Proposition 10.1. *Consider k independent two-state Markov processes on a tree with leaf set X. Assume that for each of them, all the determinants of the transition matrices are non-negative. Then, for the corresponding multivariate process and for any two characters $\mathbf{y}, \mathbf{z} \colon W \to \{0,1\}^k$ on X from a fixed non-empty subset W of X, we have:*

$$p(\mathbf{y}) \cdot p(\mathbf{z}) \leq p(\mathbf{y} \vee \mathbf{z}) \cdot p(\mathbf{y} \wedge \mathbf{z}).$$

Proof. Consider any two elements $\mathbf{Y}$ and $\mathbf{Z}$ of $\mathscr{L}_V$. We first prove the following:

$$P(\mathbf{Y}) \cdot P(\mathbf{Z}) \leq P(\mathbf{Y} \vee \mathbf{Z}) \cdot P(\mathbf{Y} \wedge \mathbf{Z}). \tag{10.3}$$

Denote the term in the brackets of Equation (10.1) by $P_j(\mathbf{U})$ to get $P(\mathbf{U}) = \prod_{j=1}^k P_j(\mathbf{U})$. Applying this to $\mathbf{U} \in \{\mathbf{Y}, \mathbf{Z}, \mathbf{Y} \vee \mathbf{Z}, \mathbf{Y} \wedge \mathbf{Z}\}$, inequality (10.3) can be written in the form $\prod_{j=1}^k P_j(\mathbf{Y}) \prod_{j=1}^k P_j(\mathbf{Z}) \leq \prod_{j=1}^k P_j(\mathbf{Y} \vee \mathbf{Z}) \prod_{j=1}^k P_j(\mathbf{Y} \wedge \mathbf{Z})$. It is clear that proving $P_j(\mathbf{Y}) P_j(\mathbf{Z}) \leq P_j(\mathbf{Y} \vee \mathbf{Z}) P_j(\mathbf{Y} \wedge \mathbf{Z})$ for each j establishes (10.3).

So let j be an arbitrary index in $\{1, \ldots, k\}$ and consider the products $P_j(\mathbf{Y}) P_j(\mathbf{Z})$ and $P_j(\mathbf{Y} \vee \mathbf{Z}) P_j(\mathbf{Y} \wedge \mathbf{Z})$. These can each be written as a product of two $\pi^{(j)}$ values multiplied by a product over the arcs (r, s) of T of two entries of $P^{(j)}(r, s)$. The products of the two $\pi^{(j)}$ terms agree in $P_j(\mathbf{Y})P_j(\mathbf{Z})$ and $P_j(\mathbf{Y} \vee \mathbf{Z})P_j(\mathbf{Y} \wedge \mathbf{Z})$; that is, $\pi^{(j)}_{\delta_j(\mathbf{Y},\rho)}\pi^{(j)}_{\delta_j(\mathbf{Z},\rho)} = \pi^{(j)}_{\delta_j(\mathbf{Y}\vee\mathbf{Z},\rho)}\pi^{(j)}_{\delta_j(\mathbf{Y}\wedge\mathbf{Z},\rho)}$. The products of the two $P^{(j)}(r, s)$ entries agree in $P_j(\mathbf{Y})P_j(\mathbf{Z})$ and $P_j(\mathbf{Y} \vee \mathbf{Z})P_j(\mathbf{Y} \wedge \mathbf{Z})$, except for the cases in which either (i) $\delta_j(\mathbf{Y}, r) = 0$, $\delta_j(\mathbf{Y}, s) = 1$, $\delta_j(\mathbf{Z}, r) = 1$, and $\delta_j(\mathbf{Z}, s) = 0$ or (ii) $\delta_j(\mathbf{Y}, r) = 1$, $\delta_j(\mathbf{Y}, s) = 0$, $\delta_j(\mathbf{Z}, r) = 0$, and $\delta_j(\mathbf{Z}, s) = 1$. However, in both cases (i) and (ii), the product $P^{(j)}(r, s)_{01}P^{(j)}(r, s)_{10}$ appears in the term for $P_j(\mathbf{Y})P_j(\mathbf{Z})$ while $P^{(j)}(r, s)_{00}P^{(j)}(r, s)_{11}$ appears in the term for $P_j(\mathbf{Y} \vee \mathbf{Z})P_j(\mathbf{Y} \wedge \mathbf{Z})$. The former term is less than or equal to the second since $P^{(j)}(r, s)_{00}P^{(j)}(r, s)_{11} - P^{(j)}(r, s)_{01}P^{(j)}(r, s)_{10} = \det P^{(j)}(r, s)$, which is non-negative by our assumption. Consequently, all the terms in $P_j(\mathbf{Y})P_j(\mathbf{Z})$ are less than or equal to the corresponding terms in $P_j(\mathbf{Y} \vee \mathbf{Z})P_j(\mathbf{Y} \wedge \mathbf{Z})$. This establishes (10.3).

We now recall a form of the four functions theorem, a classical result of Ahlswede and Daykin [1]. Let $(\mathscr{L}, \vee, \wedge)$ be a finite distributive lattice and let α be a function that assigns a non-negative real number to each element of $\mathscr{L}$. For a subset $\mathscr{A} \subseteq \mathscr{L}$, set $\alpha(\mathscr{A}) = \sum_{A\in\mathscr{A}} \alpha(A)$. If α satisfies the property that for any two elements A, B of $\mathscr{L}$, $\alpha(A)\alpha(B) \leq \alpha(A \vee B)\alpha(A \wedge B)$, then

$$\alpha(\mathscr{A})\alpha(\mathscr{B}) \leq \alpha(\mathscr{A} \vee \mathscr{B})\alpha(\mathscr{A} \wedge \mathscr{B}), \tag{10.4}$$

where $\mathscr{A} \vee \mathscr{B} = \{A \vee B : A \in \mathscr{A}, B \in \mathscr{B}\}$ and $\mathscr{A} \wedge \mathscr{B} = \{A \wedge B : A \in \mathscr{A}, B \in \mathscr{B}\}$.

We apply this theorem by taking $\mathscr{L} = \mathscr{L}_V, \alpha = P$ and noting that α satisfies the required hypothesis by (10.3). Consider any fixed non-empty subset W of X and recall the definition (for $\mathbf{u} \in \mathscr{L}_W$) of $\mathscr{A}_\mathbf{u}$ in (10.2). Note that:

$$\mathscr{A}_\mathbf{y} \vee \mathscr{A}_\mathbf{z} = \mathscr{A}_{\mathbf{y}\vee\mathbf{z}}, \text{ and } \mathscr{A}_\mathbf{y} \wedge \mathscr{A}_\mathbf{z} = \mathscr{A}_{\mathbf{y}\wedge\mathbf{z}}.$$

Thus, taking $\mathscr{A} = \mathscr{A}_\mathbf{y}$ and $\mathscr{B} = \mathscr{A}_\mathbf{z}$ in (10.4) we deduce that:

$$\alpha(\mathscr{A}_\mathbf{y})\alpha(\mathscr{A}_\mathbf{z}) \leq \alpha(\mathscr{A}_{\mathbf{y}\vee\mathbf{z}})\alpha(\mathscr{A}_{\mathbf{y}\wedge\mathbf{z}}),$$

which is, by $\alpha = P$ and (10.2), equivalent to $p(\mathbf{y})p(\mathbf{z}) \leq p(\mathbf{y} \vee \mathbf{z})p(\mathbf{y} \wedge \mathbf{z})$. □

10.2 Expected future PD

In this section, we use Proposition 10.1 to obtain an inequality concerning the expected loss of biodiversity under the two species extinction models g-FOB and t-FOB. Consider a rooted phylogenetic X-tree $\mathcal{T}$. View the edges of $\mathcal{T}$ as arcs directed away from the root, and denote the vertex set and the arc set of $\mathcal{T}$ by $V_\mathcal{T}$ and $A_\mathcal{T}$, respectively. Let each arc

a in $A_{\mathcal{T}}$ be assigned a non-negative length λ_a. Recall the notion of phylogenetic diversity for rooted phylogenetic X-trees in Definition 2.4.

Assume that species in X undergo random extinction and let E_x denote the event that a species $x \in X$ is extinct at some fixed future time t. Consider the phylogenetic diversity φ of the group of species that are still extant at time t. Recall that this random variable is referred to as *future PD*.

The expected value of φ is:

$$\mathbb{E}[\varphi] = \sum_{a=(u,v)\in A_{\mathcal{T}}} \lambda_a \cdot (1 - \mathbb{P}(\bigcap_{x\in C_v} E_x)) = \varphi_X - \sum_{a=(u,v)\in A_{\mathcal{T}}} \lambda_a \cdot \mathbb{P}(\bigcap_{x\in C_v} E_x), \tag{10.5}$$

where C_v denotes the subset of X which is separated from the root by v and which equals $\{v\}$ if v is a leaf vertex. Recall that $\mathbb{E}[\varphi]$ is referred to as *expected future PD*.

In the *generalized field of bullets model* (g-FOB) [25], the events $E_x^{(g)} := E_x$ are independent, and so the probability $\mathbb{P}(\bigcap_{x\in C_v} E_x^{(g)})$ that all the species descended from v become extinct can be written as:

$$\mathbb{P}(\bigcap_{x\in C_v} E_x^{(g)}) = \prod_{x\in C_v} p_x, \tag{10.6}$$

where p_x denotes the probability $\mathbb{P}(E_x^{(g)})$.

We noted in the previous chapter that the assumption that the events E_x are independent is likely to be unrealistic in most settings. In particular, rates at which lineages become extinct may be influenced by some species traits [42, 30]. The state-based field of bullets model (s-FOB), studied in the previous chapter, is based on the idea that closely related species in $\mathcal{T}$ are more likely to share attributes that may put them at risk in a hostile future environment. It assumes that the extinction risk of each species is influenced by some associated binary state with values 0 and 1, where state 0 confers an elevated risk of extinction for example under climate change.

Here, we generalize this model and suppose that the extinction risk of each species x is influenced by k binary states, each of which takes values in $\{0, 1\}$, where state 1 is always advantageous over state 0 for x. We suppose that it is not known what features will help species survive and so the states are not known for the species in X. However, we assume that the k states have evolved under k independent Markovian models on $\mathcal{T}$ assigning a state in $\{0, 1\}^k$ to each species.

We assume further that if the states were determined at the leaves, then extinction would proceed according to the g-FOB model in which species x is extinct at time t with probability $p_x^{\mathbf{i}}$ if it is in state $\mathbf{i} \in \{0, 1\}^k$. Finally, we suppose that for each species $x \in X$

and any two states $\mathbf{i} = (i_1, \dots, i_k)$ and $\mathbf{l} = (l_1, \dots, l_k)$:

$$p_x^{\mathbf{i}} \leq p_x^{\mathbf{l}} \text{ whenever } l_j \leq i_j \text{ for each } j = 1, \dots, k. \tag{10.7}$$

This condition says that state $\mathbf{l}$ confers at least as high an extinction risk on a species x as state $\mathbf{i}$ if all the binary states in $\mathbf{i}$ are at least as advantageous for x as the binary states in $\mathbf{l}$. Note, however, that if condition $l_j \leq i_j$ is not satisfied for every j, there is no prescribed relationship between $p_x^{\mathbf{i}}$ and $p_x^{\mathbf{l}}$. We have the freedom to specify these relationships according to the needs of the model being studied, or leave them unspecified. For example, we may assume that the k binary states are ordered in a decreasing manner by their importance for survival and that $p_x^{\mathbf{i}} \leq p_x^{\mathbf{l}}$, whenever $l_j \leq i_j$ for the smallest coordinate $j \in \{1, \dots, k\}$ for which $i_j \neq l_j$. Alternatively, we may assume that all the states are equally important for survival and that $p_x^{\mathbf{i}} \leq p_x^{\mathbf{l}}$, whenever $\sum_{j=1}^k l_j \leq \sum_{j=1}^k i_j$; that is, the more coordinates of the state assigned to x are 1 the smaller is the extinction probability of x. In the following, we only assume the relationships described in (10.7).

We call the model described above the *trait-dependent field of bullets* model (t-FOB). In the case when $k = 1$, this model is the s-FOB model, whereas the case where for each x, we have $p_x^{\mathbf{i}} = p_x^{\mathbf{l}}$ for any two states $\mathbf{i}, \mathbf{l} \in \{0, 1\}^k$ gives the g-FOB model.

Given a t-FOB model, consider the g-FOB model in which the extinction probability of each species x is the same as in the t-FOB model. That is, if $\boldsymbol{\xi}$ describes the multivariate Markov process and the values $p_x^{\mathbf{i}}$ are the conditional extinction probabilities in the t-FOB model, then, in the associated g-FOB model, each species $x \in X$ goes extinct with probability

$$p_x = \mathbb{P}[E_x^{(g)}] = \mathbb{P}[E_x^{(t)}] = \sum_{\mathbf{i} \in \{0,1\}^k} p_x^{\mathbf{i}} \mathbb{P}(\boldsymbol{\xi}(\mathrm{x}) = \mathbf{i}), \tag{10.8}$$

where $E_x^{(t)}$ denotes the event E_x under t-FOB. Theorem 10.2 compares the loss of PD under a t-FOB model with the PD loss under the associated g-FOB model.

Theorem 10.2. *Consider a t-FOB model on a rooted phylogenetic X-tree $\mathcal{T}$ with non-negative arc lengths. The expected future PD of this model is less than or equal to the expected future PD of the associated g-FOB model.*

Proof. Let $\boldsymbol{\xi}$ and $p_x^{\mathbf{i}}$ denote the Markov process and the extinction probabilities of the t-FOB model, respectively. In view of (10.5) and (10.6), it suffices to show that:

$$\prod_{x \in C_v} p_x \leq \mathbb{P}(\bigcap_{x \in C_v} E_x^{(t)}), \tag{10.9}$$

where p_x is given in (10.8). Recall how we defined the lattice $(\mathscr{L}_W, \vee, \wedge)$ for a Markov process on a tree and for a non-empty subset W of the leaf set of the tree in the previous section, and consider $(\mathscr{L}_{C_v}, \vee, \wedge)$. Since, for $\mathbf{u} \in \mathscr{L}_{C_v}$, $p(\mathbf{u})$ denotes the probability that for each $x \in C_v$, x is assigned $\mathbf{u}(x) \in \{0,1\}^k$, we get:

$$\mathbb{P}(\bigcap_{x \in C_v} E_x^{(t)}) = \sum_{\mathbf{u} \in \mathscr{L}_{C_v}} p(\mathbf{u}) \prod_{x \in C_v} f_x(\mathbf{u}),$$

where $f_x(\mathbf{u})$ is the probability that x becomes extinct given that it is in state $\mathbf{u}(x)$; that is, $f_x(\mathbf{u}) = p_x^{\mathbf{u}(x)}$. Moreover, for each $x \in C_v$, we have:

$$p_x = \sum_{\mathbf{i} \in \{0,1\}^k} p_x^{\mathbf{i}} \mathbb{P}(\boldsymbol{\xi}(x) = \mathbf{i}) = \sum_{\mathbf{i} \in \{0,1\}^k} p_x^{\mathbf{i}} \left(\sum_{\mathbf{u} \in \mathscr{L}_{C_v} : \mathbf{u}(x) = \mathbf{i}} p(\mathbf{u}) \right) = \sum_{\mathbf{u} \in \mathscr{L}_{C_v}} p(\mathbf{u}) f_x(\mathbf{u}).$$

Now we can rewrite (10.9) as

$$\prod_{x \in C_v} \left(\sum_{\mathbf{u} \in \mathscr{L}_{C_v}} p(\mathbf{u}) f_x(\mathbf{u}) \right) \leq \sum_{\mathbf{u} \in \mathscr{L}_{C_v}} p(\mathbf{u}) \prod_{x \in C_v} f_x(\mathbf{u}). \tag{10.10}$$

The proof of (10.10) makes use of Proposition 10.1 as well as the following multivariate form of the FKG inequality of Fortuin, Kasteleyn and Ginibre (1971) [31]. Given a finite distributive lattice $(\mathscr{L}, \vee, \wedge)$, suppose that $f_1, f_2, \ldots, f_n$ are functions from $\mathscr{L}$ into the non-negative real numbers that satisfy, for any two elements A, B of $\mathscr{L}$, the condition that:

$$A \leq B \Rightarrow f_i(A) \geq f_i(B). \tag{10.11}$$

Furthermore, suppose that μ is a probability measure on the elements of $\mathscr{L}$ which satisfies the condition that

$$\mu(A)\mu(B) \leq \mu(A \vee B)\mu(A \wedge B) \text{ for any pair } A, B \in \mathscr{L}. \tag{10.12}$$

Then:

$$\prod_{i=1}^{n} \left(\sum_{A \in \mathscr{L}} \mu(A) f_i(A) \right) \leq \sum_{A \in \mathscr{L}} \mu(A) \prod_{i=1}^{n} f_i(A). \tag{10.13}$$

We apply this inequality by setting $\mathscr{L} = \mathscr{L}_{C_v}$, $\mu = p$ and $f_x(\mathbf{u}) = p_x^{\mathbf{u}(x)}$ for $\mathbf{u} \in \mathscr{L}_{C_v}$, $x \in C_v$. Note that f_x satisfies (10.11). Namely, $\mathbf{u} \leq \mathbf{y}$ (for $\mathbf{u}, \mathbf{y} \in \mathscr{L}_{C_v}$) means that $u_j(x) \leq y_j(x)$ for each coordinate j, which, by (10.7), implies $p_x^{\mathbf{u}(x)} \geq p_x^{\mathbf{y}(x)}$. Note also that μ satisfies (10.12) by Proposition 2.1. In view of these, (10.13) provides inequality (10.10), and the proof is complete. □

10.3 Variance of future PD

Consider now the variance of φ:

$$\mathrm{Var}[\varphi] = \mathrm{Cov}[\varphi, \varphi] = \sum_{a,b \in A_{\mathcal{T}}} \lambda_a \lambda_b \, \mathrm{Cov}[Y_a, Y_b], \tag{10.14}$$

where Y_a is the random variable that takes value 1 if arc a is part of the subtree connecting the survival species and the root and takes value 0 otherwise. Our goal is to compare the variance under a t-FOB model to the variance under the associated g-FOB model. It is easy to find examples in which the former variance is greater than the latter and so we will only show that the variance for a t-FOB model can be less than that of the associated g-FOB model. To this end, let $\mathcal{T}$ be the tree with leaf set $\{x, y\}$ in which the arcs b and c pointing to x and y, respectively, are incident with the single interior vertex of the tree, which is adjacent to the root by arc a. Consider $\mathrm{Cov}[Y_a, Y_a] = (1 - \mathbb{P}[E_x \cap E_y])\mathbb{P}[E_x \cap E_y]$, which is written as $(1 - \mathbb{P}[E_x^{(t)} \cap E_y^{(t)}])\mathbb{P}[E_x^{(t)} \cap E_y^{(t)}]$ in t-FOB and which becomes $(1 - p_x p_y)p_x p_y$ under g-FOB. Note that $\mathrm{Cov}[Y_a, Y_a]$ is less under a t-FOB than under the associated g-FOB if and only if $\mathbb{P}[E_x^{(t)} \cap E_y^{(t)}] > p_x p_y$ and $\mathbb{P}[E_x^{(t)} \cap E_y^{(t)}] + p_x p_y > 1$ hold. It is easy to see that these conditions can be satisfied by some t-FOB model (together with its g-FOB) on $\mathcal{T}$. Additionally, for any such t-FOB model, a value of λ_a can be chosen that is large enough in relation to λ_b and λ_c so that $\lambda_a^2 \, \mathrm{Cov}[Y_a, Y_a]$ is the dominant term in (10.14), resulting in a greater total variance for the corresponding g-FOB.

The following example describes an s-FOB (that is, a t-FOB with $k = 1$) under which the variance is less than the variance under the associated g-FOB.

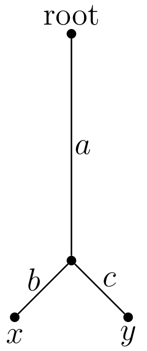


Figure 10.2: A rooted phylogenetic tree on $\{x, y\}$.

Example. Let $\mathcal{T}$ be the tree shown in Figure 10.2 with arc lengths $\lambda_a = 4$ and $\lambda_b = \lambda_c = 1$ and consider the following s-FOB model on $\mathcal{T}$. Let ξ be a two-state Markov process on

$\mathcal{T}$ with the state space $\{0,1\}$ so that $\pi_0 = \pi_1 = \frac{1}{2}$ and each arc is assigned the transition matrix $\left(\begin{smallmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{smallmatrix}\right)$. Let $p_x^0 = p_y^0 = \frac{7}{8}$ and $p_x^1 = p_y^1 = \frac{6}{8}$. A careful check shows that the variance under this model is less than the variance under the associated g-FOB model (in which $p_x = p_y = \frac{13}{16}$).

This chapter presents the results from the paper 'Trait-dependent extinction leads to greater expected biodiversity loss' [27]. I would like to thank my co-author Mike Steel for his collaboration.

Bibliography

[1] R. Ahlswede and V. Blinovsky, *Lectures on advances in combinatorics*, Springer, 2008.

[2] R. Ahlswede and D. E. Daykin, *An inequality for the weights of two families of sets, their unions and intersections*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **43** (1978), 183–185.

[3] I. Anderson, *Combinatorics of finite sets*, Dover Publications, New York, 1987.

[4] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi, *Complexity and approximation: Combinatorial optimization problems and their approximability properties*, Springer, Berlin, 1999.

[5] M. Baroni, C. Semple, and M. Steel, *A framework for representing reticulate evolution*, Annals of Combinatorics **8** (2004), 391–408.

[6] J. Bascompte, C. J. Melian, and E. Sala, *Interaction strength combinations and the overfishing of a marine food web*, Proceedings of the National Academy of Sciences of the USA **102** (2005), 5443–5447.

[7] S. R. Beissinger and D. R. McCullough, *Population viability analysis*, University of Chicago Press, Chicago, 2002.

[8] M. Bordewich, A. G. Rodrigo, and C. Semple, *Selecting taxa to save or sequence: desirable criteria and a greedy solution*, Systematic Biology **57** (2008), 825–834.

[9] M. Bordewich and C. Semple, *Nature reserve selection problem: a tight approximation algorithm*, IEEE/ACM Transactions on Computational Biology and Bioinformatics **5** (2008), 275–280.

[10] M. Bordewich, C. Semple, and A. Spillner, *Optimizing phylogenetic diversity across two trees*, Applied Mathematics Letters **22** (2009), 638–641.

[11] J. T. Chang, *Full reconstruction of Markov models on evolutionary trees: identifiability and consistency*, Mathematical Biosciences **137** (1996), 51–73.

[12] W. Y. C. Chen, A. W. M. Dress, and W. Q. Yu, *Checking the reliability of a linear-programming based approach towards detecting community structures in networks*, IET Systems Biology **1** (2007), 286–291.

[13] A. Cobham, *The intrinsic computational difficulty of functions*, Proceedings of the 1964 International Congress for Logic Methodology and Philosophy of Science (Y. Bar-Hiller, ed.), North Holland, Amsterdam, 1964, pp. 24–30.

[14] S. Cook, *The complexity of theorem-proving procedures*, Proceedings of the Third Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, New York (1971), 151–158.

[15] G. Cornuejols, M. L. Fisher, and G. L. Nemhauser, *Location of bank accounts to optimize float: an analytic study of exact and approximate algorithms*, Management Science **23** (1977), 789–810.

[16] R. H. Crozier, *Genetic diversity and the agony of choice*, Biological Conservation **61** (1992), 11–15.

[17] R. Durrett, *Probability: Theory and Examples*, Wadsworth and Brooks/Cole, Belmont, California, 1991.

[18] J. Edmonds, *Paths, trees, and flowers*, Canadian Journal of Mathematics **17** (1965), 449–467.

[19] A. W. F. Edwards, *Estimation of the branch points of a branching diffusion process*, Journal of the Royal Statistical Society: Series B **32** (1970), 155–174.

[20] D. P. Faith, *Conservation evaluation and phylogenetic diversity*, Biological Conservation **61** (1992), 1–10.

[21] ———, *The role of the phylogenetic diversity measure, PD, in bio-informatics: getting the definition right*, Evolutionary Bioinformatics Online **2** (2006), 301–307.

[22] ———, *Phylogenetic Diversity and Conservation*, Conservation Biology: Evolution in Action (S. P. Carroll and C. Fox, eds.), Oxford University Press, New York, 2008, pp. 99–115.

[23] D. P. Faith and A. M. Baker, *Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges*, Evolutionary Bioinformatics Online **2** (2006), 70–77.

[24] B. Faller, T. Ingram, and C. Semple, *Preserving phylogenetic diversity in ecological networks*, in preparation.

[25] B. Faller, F. Pardi, and M. Steel, *Distribution of phylogenetic diversity under random extinction*, Journal of Theoretical Biology **251** (2008), 286–296.

[26] B. Faller, C. Semple, and D. Welsh, *Optimizing phylogenetic diversity with ecological constraints*, Annals of Combinatorics, in press.

[27] B. Faller and M. Steel, *Trait-dependent extinction leads to greater expected biodiversity loss*, SIAM Journal on Discrete Mathematics, submitted.

[28] U. Feige, *A threshold of* $\ln n$ *for approximating Set Cover*, Journal of the Association for Computing Machinery **45** (1998), 634–652.

[29] J. Felsenstein, *Inferring phylogenies*, Sinauer Associates, Sunderland, Massachusetts, 2004.

[30] R. G. FitzJohn, W. P. Maddison, and S. P. Otto, *Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies*, Systematic Biology **58** (2009), 595–611.

[31] C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre, *Correlation inequalities on some partially ordered sets*, Communications in Mathematical Physics **22** (1971), 98–103.

[32] M. R. Garey and D. S. Johnson, *Computers and intractability: A guide to the theory of NP-Completeness*, W. H. Freeman and Company, San Francisco, 1979.

[33] K. Hartmann, *Biodiversity conservation and evolutionary models*, Ph.D. thesis, University of Canterbury, New Zealand, 2008.

[34] K. Hartmann and M. Steel, *Maximimizing phylogenetic diversity in biodiverstity conservation: greedy solutions to the Noah's Ark problem*, Systematic Biology **55** (2006), 644–651.

[35] ______, *Phylogenetic diversity: from combinatorics to ecology*, Reconstructing Evolution: New Mathematical and Computational Advances (O. Gascuel and M. Steel, eds.), Oxford University Press, 2007, pp. 171–196.

[36] S. B. Heard and A. O. Mooers, *Phylogenetically patterned speciation rates and extinction risks change the loss of evolutionary history during extinctions*, Proceedings of the Royal Society of London: Series B **267** (2000), 613–620.

[37] B. R. Holland, *Evolutionary analyses of large data sets: trees and beyond*, Ph.D. thesis, Massey University, New Zealand, 2001.

[38] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*, Cambridge University Press, 1994, corrected reprint of the 1991 original.

[39] R. M. Karp, *Reducibility among combinatorial problems*, Complexity of Computer Computations (R. E. Miller and J. W. Thatcher, eds.), Plenum Press, New York, 1972, pp. 85–103.

[40] R. Kindermann and J. L. Snell, *Markov random fields and their applications*, Dover Publications, American Mathematical Society, 1980.

[41] L. Lovász, *Complexity of algorithms*, Lecture notes, http://www.cs.elte.hu/ lovasz/complex.ps, translated version of the Hungarian original from 1992.

[42] W. P. Maddison, P. E. Midford, and S. P. Otto, *Estimating a binary character's effect on speciation and extinction*, Systematic Biology **56** (2007), 701–710.

[43] T. L. Magnanti and L. A. Wolsey, *Optimal trees*, Network Models, Handbook in Operations Research and Management Science (M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser, eds.), Amsterdam, North-Holland, 1995, pp. 503–615.

[44] F. A. Matsen, *Fourier transform inequalities for phylogenetic trees*, IEEE/ACM Transactions on Computational Biology and Bioinformatics **6** (2009), 89–95.

[45] B. Q. Minh, S. Klaere, F. Forest, and A. von Haeseler, *SDA-LP: A simple and unifying framework for optimizing split diversity and its application to the Cape flora of South Africa*, in preparation.

[46] B. Mohar, *Face covers and the genus problem for apex graphs*, Journal of Combinatorial Theory, Series B **82** (2001), 102–117.

[47] A. O. Mooers, S. B. Heard, and E. Chrostowski, *Evolutionary heritage as a metric for conservation*, Phylogeny and Conservation (A. Purvis, J. L. Gittleman, and T. M. Brooks, eds.), Cambridge University Press, 2005, pp. 120–138.

[48] V. Moulton, C. Semple, and M. Steel, *Optimizing phylogenetic diversity under constraints*, Journal of Theoretical Biology **246** (2007), 186–194.

[49] S. Nee and R. M. May, *Extinction and the loss of evolutionary history*, Science **278** (1997), 692–694.

[50] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, *An analysis of approximations for maximizing submodular set functions i*, Mathematical Programming **14** (1978), 265–294.

[51] S. Opitz, *Trophic interactions in Caribbean coral reefs*, Tech. Report 43, International Center for Living Aquatic Resources Management, Makaty City, Philippines, 1996.

[52] A. Panholzer, *The distribution of the size of the ancestor-tree and of the induced spanning subtree for random trees*, Random Structures and Algorithms **25** (2004), 179–207.

[53] C. H. Papadimitriou and M. Yannakakis, *Optimization, approximation, and complexity classes*, Journal of Computer and System Sciences **43** (1991), 425–440.

[54] F. Pardi and N. Goldman, *Species Choice for Comparative Genomics: Being Greedy Works*, PLoS Genetics **1** (2005), e71.

[55] ———, *Resource-aware taxon selection for maximizing phylogenetic diversity*, Systematic Biology **56** (2007), 431–444.

[56] J. Pearl and M. Tarsi, *Structuring causal trees*, Journal of Complexity **2** (1986), 60–77.

[57] A. Purvis, P.-M. Agapow, J. L. Gittleman, and G. M. Mace, *Nonrandom extinction and the loss of evolutionary history*, Science **288** (2000), 328–330.

[58] D. M. Raup, *Extinction: Bad Genes or Bad Luck?*, Oxford University Press, 1993.

[59] D. W. Redding and A. O. Mooers, *Incorporating evolutionary measures into conservation priorities*, Conservation Biology **20** (2006), 1670–1678.

[60] E. L. Rezende, E. M. Albert, M. A. Fortuna, and J. Bascompte, *Compartments in a marine food web associated with phylogeny, body mass, and habitat structure*, Ecology Letters **12** (2009), 779–788.

[61] N. Robertson, D. Sanders, P. Seymour, and R. Thomas, *A new proof of the four-colour theorem*, Electronic Research Announcements of the American Mathematical Society **2** (1996), 17–25.

[62] A. S. L. Rodrigues, T. M. Brooks, and K. J. Gaston, *Integrating phylogenetic diversity in the selection of priority areas for conservation: does it make a difference?*, Phylogeny and Conservation (A. Purvis, J. L. Gittleman, and T. M. Brooks, eds.), Cambridge University Press, 2005, pp. 101–119.

[63] A. S. L. Rodrigues and K. J. Gaston, *Maximising phylogenetic diversity in the selection of networks of conservation areas*, Biological Conservation **105** (2002), 103–111.

[64] V. M. Sarich and A. C. Wilson, *Immunological time scale for hominid evolution*, Science **158** (1967), 1200–1203.

[65] C. Semple and M. Steel, *Phlyogenetics*, Oxford University Press, 2003.

[66] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, Wiley, New York, 1980.

[67] H. Simianer, *Accounting for non-independence of extinction probabilities in the derivation of conservation priorities based on Weitzman's diversity concept*, Conservation Genetics **9** (2008), 171–179.

[68] A. Spillner, B. Nguyen, and V. Moulton, *Computing phylogenetic diversity for split systems*, IEEE/ACM Computational Biology and Bioinformatics **5** (2008), 235–244.

[69] M. Steel, *Recovering a tree from the leaf colourations it generates under Markov model*, Applied Mathematics Letters **7** (1994), 19–23.

[70] ———, *Phylogenetic diversity and the greedy algorithm*, Systematic Biology **54** (2005), 527–529.

[71] M. Steel and B. Faller, *Markovian log-supermodularity, and its applications in phylogenetics*, Applied Mathematics Letters **22** (2009), 1141–1144.

[72] M. Steel, L. Goldstein, and M. S. Waterman, *A central limit theorem for the parsimony length of trees*, Advances in Applied Probability **28** (1996), 1051–1071.

[73] P. G. Tait, *On the colouring of maps*, Proceedings of the Royal Society of Edinburgh: Section A **10** (1980), 501–503.

[74] A. Turing, *On computable numbers, with an application to the entscheidungsproblem*, Proceedings of the London Mathematical Society: Series 2 **42** (1936), 230–265.

[75] ———, *On computable numbers, with an application to the entscheidungsproblem: a correction*, Proceedings of the London Mathematical Society: Series 2 **43** (1936), 544–546.

[76] C. J. van der Heide, C. von den Bergh, and E. C. van Ierland, *Extending Weitzman's economic ranking of biodiversity protection: combining ecological and genetic considerations*, Ecological Economics **55** (2005), 218–223.

[77] D. P. Vazquez and J. L. Gittleman, *Biodiversity conservation: Does phylogeny matter?*, Current Biology **8** (1998), 379–381.

[78] M. L. Weitzman, *The Noah's Ark problem*, Econometrica **66** (1998), 1279–1298.

[79] J. B. Whitfield and P. J. Lockhart, *Deciphering ancient rapid radiations*, Trends in Ecology and Evolution **22** (2007), 258–265.

[80] L. Witting and V. Loeschcke, *The optimization of biodiversity conservation*, Biological Conservation **71** (1995), 205–207.

[81] L. Witting, J. Tomiuk, and V. Loeschcke, *Modelling the optimal conservation of interacting species*, Ecological Modelling **125** (2000), 123–143.

[82] E. Zuckerkandl and L. B. Pauling, *Molecular disease, evolution, and genetic heterogeneity*, Horizons in Biochemistry (M. Kasha and B. Pullman, eds.), Academic Press, New York, 1962, pp. 189–225.