DIVERSIFICATION DYNAMICS AND UNSUPERVISED DISCOVERY OF MICROBIAL UNITS IN THE EARTH MICROBIOME

BY

ALICE DOUCET BEAUPRÉ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Ecology, Evolution, and Conservation Biology
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Associate Professor James P. O'Dwyer, Chair and Director of Research
Professor Carla Cáceres
Professor Tandy Warnow
Professor Sergei Maslov

Abstract

Until recently, much of the microbial world was hidden from view. A global research effort has changed this, unveiling and quantifying microbial diversity across an enormous range of critically-important contexts, from the human microbiome, to plant-soil interactions, to marine life. Yet what has remained largely hidden is the interplay of ecological and evolutionary processes that led to the diversity we observe in the present day. In this thesis we introduce two theoretical frameworks, one at the macroevolutionary scale and the other at the mesoscopic scale where intricacies of abundances and environmental specificities begin to matter. At the macroscopic scale we identify an imbalance between gradual, ongoing diversification and rapid bursts across a vast range of microbial habitats and find universal quantitative similarities in the tempo and mode of diversification, independent of habitat type. This signature persists even when the quality and length of our sequence data and consequent resolution of the phylogeny is relatively low compared to the timescale of the processes. At the mesoscopic scale we discover a rich hierarchy of organization and niche signals in the pattern of abundances in the microbial diversity of the global ocean and in the process identify three putatively novel microbiomes.

To Lucia and Georgia.

Acknowledgments

This project would not have been possible without the support of many people. I owe my life to Lucia, my partner in crime, who endured me and stood by me and brought light, love, support, and growth during what was, and still is, a difficult transitional period of my life. Many thanks to my adviser, James P. O'Dwyer, for his eternal patience, understanding, encouragement, and trust. Thank you to PEEC Director Angela Kent for her support and for helping me secure the financial means to persist in my research during times of struggle. Many thanks also to Liz Barnabe for her indispensable and constant help in navigating the sinuous graduate process. And finally thanks to my parents, sisters, and friends for all their love and support throughout the years.

Table of Contents

Chapter 1

Introduction

Documenting universal patterns in ecological systems and evolution has a very long history stretching all the way back to the very early days of their respective fields. So has the drive to try and explain their origins using both simple and complicated statistical and dynamical models. Just to give a very short list of examples, we can think here of Preston's canonical log-normal relative species abundances distribution[1]–[3], large-scale species area curve[4]—considered one of the few quasi-fundamental law of ecology—Hubbell's unified neutral model of diversity in metacommunities[5], MacArthur and Wilson's theory of island biogeography[6], [7] which was two years later supported by the ingenious experiment of Simberloff and Wilson on mangrove island[8], the long-tailed nature of rank-abundance distributions[9], [10], allometric scaling of organismal organization[11]–[14], and patterns of interactions and the stability of complex communities[15], [16].

Various efforts are now striving to catalogue the most diverse systems on Earth; microbiomes. Current efforts in this direction challenge the imagination in scale and complexity. We are referring here to projects like the Human Microbiome Project[17], the International Census of Marine Microbes[18], the TARA Oceans expedition[19], [20], and the Earth Microbiome Project[21]. The tremendous wealth of information these efforts have collected over the years begs for new models and concepts to explain them. How fast and in what way do microbes evolve? How do they organize themselves across macro- and microenvironments? Is there a simple way to capture the incredible complexity they harbor? How do we bridge the explanatory gap between the microscopic, the mesoscopic, and the macroscopic? In short, are there simplified or complex laws and universalities underlying the patterns we see in microbial communities? Are we succumbing in this search to 'physics envy'[22] or can we hope to be rescued by the unreasonable effectiveness of mathematics[23]?

What we need here, we believe, is a conceptual microscope—or is it a telescope?—that allows us to smoothly change the focus of our explanations inwards and outwards, that imbricates our theories and models one scale below and one scale above. Given the features and processes we hoped to capture in our models, there are few things more satisfying than to find out it can explain other features at larger scales, namely that it explain the self-organization and the 'emergence' of qualitatively distinct phenomena in the limit of a large number of interacting constituents[24]. Emergent phenomena and self-organization have been widely recognized and documented in the field of ecology and evolution of complex systems[25], but the current theoretical picture remains unsatisfying given its disparate and case-by-case nature. What is missing, indeed, is the microscope, a way to transform concepts, models, or theories into others in a way that connects them across scales. One might be tempted to borrow directly from the realm of theoretical physics where one such tool exists, the 'renormalization group'[26], [27], yet the success of the renormalization group in particle

and statistical physics in explaining the emergence of macroscopic phases of matter in terms of dissimilar microscopic laws is only made possible because of the high degree of symmetry enjoyed by the systems and models at play. Unfortunately we do not have the luxury of such symmetries in the biological and ecological realms and thus we are still today left struggling with 'the problem of pattern and scale' in ecology and evolution[28]. Biology is heterogeneous, complicated, tangled, and its systems are out-of-equilibrium and complex.

In this thesis I do not pretend to even marginally approach the formulation of such a conceptual microscope, but I will attempt to reaffirm its importance and the question it raises by highlighting what I believe are interesting patterns at various scales of the diversification and organization of microbial systems. I will first try to explain at the phenomenological level how these patterns can emerge from simple processes containing but a few justified elements, and then I will teach the machine to automatically learn and dissect the complexity left behind by those processes in the organization of microbiomes. I will, in a way, teach it to read the leaves of microbial communities.

In Chapter 2 we situate ourselves at the macroevolutionary scale to study the tempo and mode of evolution in the microbial tree of life. In Section 2.2 we construct two new processes, called the 'innovation' and 'heterogeneous innovation' processes, to complement the more traditional birth-death processes, also called speciation-extinction processes, by introducing a second time-scale, the fast time-scale, at which diversification can occur. We combine these processes together into a novel model called the Birth-Death-Heterogeneous-Innovation model (BDH). Then in Section 2.3 we introduce a novel exact goodness-of-fit test to determine whether our model is sufficient, rather than simply better, at explaining the patterns of diversification observed in empirical phylogenies. In preparation to the application of our model to microbial diversification, in Section 2.4 we transform the enormous phylogenetic tree of the Earth Microbiome Project (EMP) into a timetree by means of phylogenetic placement using SEPP and constrained phylogenetic optimization using family-level calibration points given by the Timetree project. In Section 2.6 we apply our methodology the the calibrated EMP timetree and reveal two universal features present across microbiomes. Finally in Section 2.7 we discuss limits to our models and implications of our results to the understanding of microbial niche space and give a tentative answer to the question of the tempo and mode of microbial evolution.

Zooming in to the mesoscopic scale, in Chapter 3 we set out to find signals of the aforementioned microbial niches in the ocean microbiome. In Section 3.2 we first explain how to use the community phylogenetic tree as an organizational tool to helps us reframe traditional microbial ecology data into biologically-informed 'phylogenetic abundance tables'. Then we introduce two nonparametric probabilistic generative models inspired by computational linguistics, the flat sample-wise Dirichlet process and the path-limited nested hierarchical Dirichlet process (pl-nhDP), which we use to discover and describe patterns of abundance in the phylogenetic abundance table. In Section 3.3 we describe how we can discover coarse-grained ecological microbial units as a certain kind of minimal optimal set of clades that captures the full topology of the structure inferred by the pl-nHDP model. In Section 3.4 we apply our unsupervised learning method to two microbial datasets, namely the zebrafish gut microbiome dataset and the TARA Ocean expedition dataset, where we find abundant 'signals of niches' across both known and unknown environmental features. Finally in Section 3.5 we discuss the useful of our model and how our results open up the question of what we mean by niches in an unsupervised learning context.

Chapter 2

Burstiness in the Earth Microbiome

2.1 Introduction

Large scale microbiome sampling and sequencing, for example the efforts behind the Earth Microbiome project (EMP)[21], The Human Microbiome project[29], and the TARA Oceans expedition[19], have documented global microbial diversity with unprecedented scope and resolution. The tools currently applied to these data, in particular the Metagenomic RAST server[30], the Mothur software[31]–[34], open-reference OTU picking[35], the QIIME software[36], and the Deblur method[37], allow us to quantify the amount and type of diversity found in microbial communities and yet we know remarkably little about the the underlying community dynamics and tempo of diversification that generated the biodiversity we observe. This gap in our knowledge calls out for robust new ecological and evolutionary theories that will allow us to connect mechanisms to observed patterns through the processes of dispersal, diversification, environmental selection, and ecological drift. [38], [39].

To address this challenge, we introduce a new methodology to bridge the gap between biological process and observed microbial biodiversity. Our approach leverages the inference of dynamical processes from evolutionary trees, previously applied to understand large-scale evolutionary structures[40]–[46] and the phylodynamics of viral populations on shorter timescales[47]–[50]. We also incorporate the recent identification of bursts of diversification in microbial phylogenies which emerged following the multiscale analysis of phylogenetic diversity[51]. The result is a model which includes traditional, slow processes for gradual speciation (one lineage goes to two lineages, which we call the 'birth' of a lineage) and extinctions (one lineage disappears, which we call 'death'), together with a third set of mechanisms incorporating the process of ecological 'innovation' potentially followed by radiative diversification.

We apply our framework to the EMP dataset, spanning 27751 samples from 96 studies of 96 habitats and 40 biomes[21], finding a previously unidentified balance of fast and slow evolutionary processes in these data, and a tendency towards universal behavior in the quantitative description of bursty diversification. We cannot directly quantify the traits and their changes through time that may have led to a given combination of rapid and gradual processes, but our results are strongly suggestive of a centre-ground in the long-standing debate over phyletic gradualism versus punctuated equilibrium[52].

Our knowledge of the diversification of microbial organisms is characterized by the branching of their evolutionary lineages reconstructed using genetic sequence data sampled in the present day[53]–[57]. Phylogenetic trees represent evolutionary relationships in the form of a tree with branch lengths in units of the average

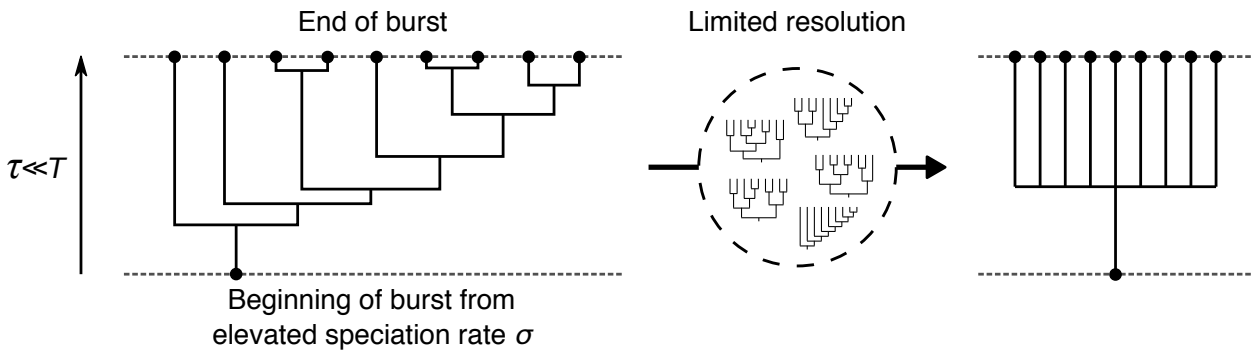


Figure 2.1: **Uncertainty generates polytomies.** Periods of fast diversification leave little to no signal in sequences with a limited number of base pairs, meaning that we cannot always distinguish between different possible orderings of diversification events using short sequences. These ambiguities often rightfully show up in bootstrap consensus trees or during calibration. Allowing for the presence of a process whereby one lineage effectively instantaneously diversifies into many over a very short timescale τ alleviates this issue and allows the inference of effective parameters associated with faster processes. Limited resolution in effect naturally lead to the coarse-graining of the class of all short periods of diversification events from one to n lineages by transforming them into a polytomy of some size k.

number of mutations between two points. To make the jump from molecular evolution of nucleotides to theoretical models of diversification through time we need to calibrate such phylogenetic trees and transform them into ultrametric trees, meaning where the path from every leaf to the root is equal, with branches in units of time, called chronograms[58] or more recently timetrees[59], [60]. This process assumes a molecular clock where evolutionary changes are taken to be mostly constant and is often calibrated e.g. using data from the fossil record or known, dated events in the evolution of life. Several software packages are available to generate them[61]–[65]. We can think of timetrees as the input data for macroevolutionary rate inference. Traditional macroevolutionary rate inference approaches allow for gradual speciation and extinction and various modifications have been introduced throughout the years on the way the rates of those processes are parametrized or vary in time[40]–[47]. More recently, the estimation of microevolutionary dynamical epidemiological SIR models of viral diversification using phylogenetic trees has been explored under what is an exciting new field called 'phylodynamics'[49], [50], [66].

Breaking away from seeking models and methods for the estimation of micro- and macroevolutionary rates of ever increasing complexity, we opt instead to create a more modest phenomenological model of diversification with a simple minimal addition; a third process called 'innovation'. In our innovation process, a lineage experiences a much faster rate of diversification, σ, for a very short time, τ. We can think of this event as representing the outcome of a key innovation that opens the opportunity for a rapid radiation, which is eventually saturated [67], [68]. The problem with inferring the parameters of the innovation process is that when σ is large, there will be parts of any reconstructed timetree where we may not have enough information in our sequence data to distinguish the true ordering of those branching events. Figure 2.1 shows a cartoon of how under those circumstances, instead of one lineage branching into two, one lineage will sometimes appear to 'instantaneously' branch into many. Even if we did have longer sequences, there is always a speed limit on what kinds of process we can accurately infer from these data.

Surprisingly, there is a way to bypass this speed limit by leveraging the distribution of sizes of these apparent bursts of branching. Even though we cannot resolve phylogenies down to the shortest timescales, the distribution of burst sizes still carries information about the parameters of the innovation process. The catch

is that we cannot distinguish between different values of σ and τ independently and instead the distribution of burst sizes only depends on the product of diversification rate and diversification time, $\sigma\tau$, as derived below. By looking at the evolutionary history through a blurred lens, we therefore collapse a multi-parameter family of models into a single parameter—reminiscent of the loss of information under coarse-graining in physics—so that at a sufficiently coarse resolution, many different fine-scale models map onto the same effective theory[26], [27], [69].

2.2 Methods

Coarse-Grained Timetrees

Consider a timetree $\mathcal{T}$, i.e. a rooted ultrametric phylogeny with branches in units of time. We will from now on call it a tree for short. We will denote its depth by T, namely the distance from any extant lineage in the present day to the root. We do not restrict ourselves to bifurcating trees and will allow internal nodes to have arbitrary outdegrees (number of immediate descendants) greater than or equal to 2. In other words we allow trees with polytomies. For computational purposes and without loss of generality, we rescale all branch lengths by T, which amounts to a change of units that normalizes the total tree depth to 1.

Our coarse-graining method is shown in Figure 2.2. In Figure 2.2 A we introduce K slices over the tree at equal distance $\Delta = T/K$ from each other. Each slice cuts the branches it intersects and effectively breaks the tree into small subtrees as shown in Figure 2.2 B. We will sometimes call those subtree 'observed past subtrees' when we want to highlight the fact that each lineage of a subtree from deeper in the tree must have all its descendant lineages with themselves observable descendants in the present day. Figure 2.2 C shows that for each subtree falling off of this slicing operation we associate a tuple (t, s, k), $t > s$, $k \geq 1$, where t is the time in the past of the single root lineage of the subtree, s the time in the past of the crown of the subtree, and k the number of leaves, or size, of the subtree. Note that t and s increases towards the past. For the slice closest to the present its subtrees are such that $s = 0$. Notice that for a given (t, s, k) there is a class of equivalence of subtrees of equal size, namely the set of all trees with k leaves. The probabilities of observing a subtree of size k, which we will derive in section 2.2.4, accounts for the marginalization over all those equivalent subtrees. This is a feature of the coarse-graining approach. We indicate this coarse-graining operation by $\mathcal{R}_K$ and the coarse-grained tree, following the slicing into K slices, by $\mathcal{T}_K = \mathcal{R}_K[\mathcal{T}]$. By an abuse of notion we identify the coarse-grained tree with the set of its tuples

$$\mathcal{T}_K = \{(t_i, s_i, k_i)\}_{i \in \mathcal{I}_{\mathcal{T}_K}}, \quad i = 1, \dots S \tag{2.1}$$

where $\mathcal{I}_{\mathcal{T}_K}$ is an index set over $\mathcal{T}_K$, S is the number of subtrees in $\mathcal{T}_K$ and thus of indexes in $\mathcal{I}_{\mathcal{T}_K}$. Given a uniform slicing we will alternatively use the notation which associate the index σ to a slice with boundaries $(t, s) = (\sigma\Delta, (\sigma - 1)\Delta)$ such that the set of tuples

$$\mathcal{T}_K = \{(t_{\sigma i}, s_{\sigma i}, k_{\sigma i})\}_{\substack{i \in \mathcal{I}_\sigma \\ \sigma = 1 \dots K}}. \tag{2.2}$$

where in particular $t_{Ki} = T$ and $s_{1i} = 0$. The usefulness of this notation resides in highlighting the redundancy

$$s_{\sigma,i} = t_{\sigma-1,i}. \tag{2.3}$$

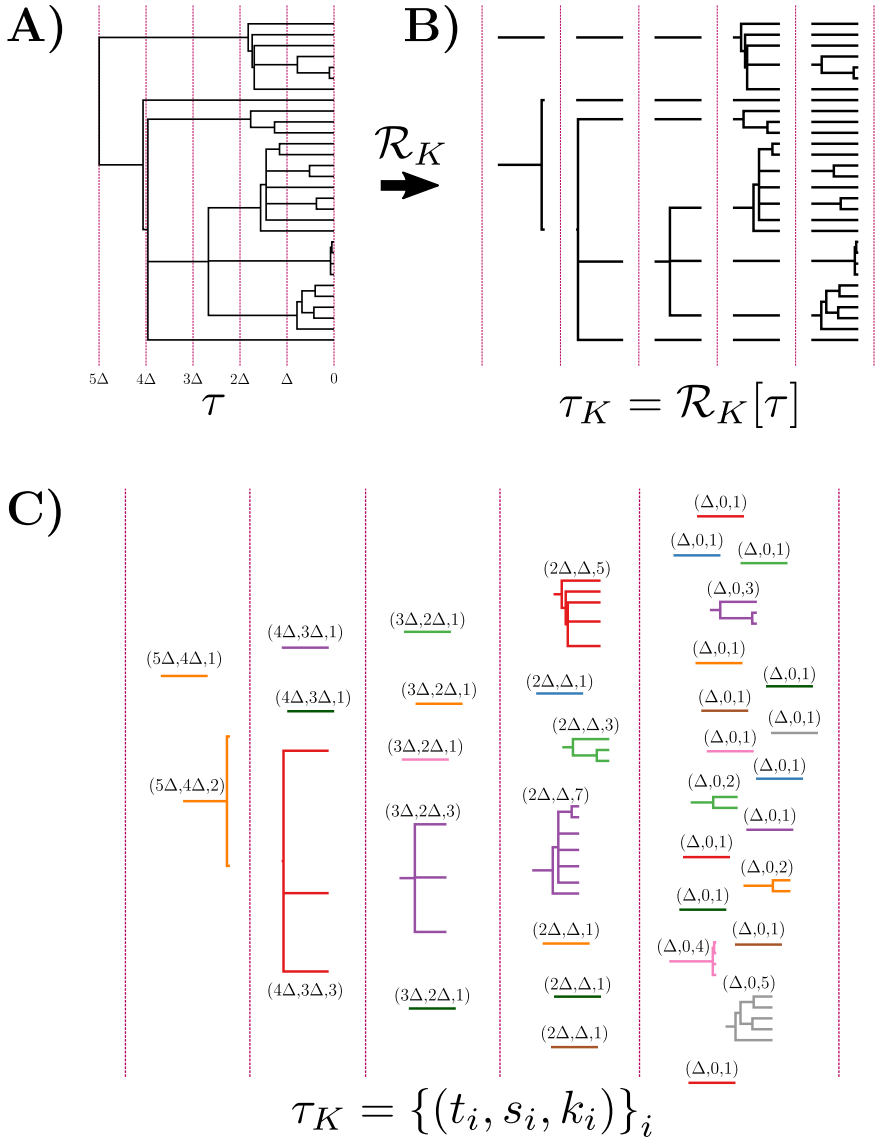


Figure 2.2: The coarse-graining process consists of three steps. **A** Introduce K slices of width $\Delta = T/K$. There is no loss of generality in using slices with equal width. **B** Slice the tree into observed past subtrees by cutting the tree at the intersection between branches and slice boundaries. **C** Associate a tuple (t, s, k) to each subtree where t is the time of the origin of the subtree, $s < t$ the time where descendant branches were cut by the neighboring slice closer to the present day, and k the number of leaves of the subtree.

2.2.1 Formalism

The equivalence between holomorphic/meromorphic generating functions and operational/Fock-space methods applied to classical objects has a long history in the field of nonequilibrium statistical mechanics of many-body systems, beginning with seminal papers from various authors[70]–[72]. See also Maslov's operational method [73] for a parallel development in the field of linear differential operators. Yet it is only relatively recently that those methods have been recognized as potent tools applicable to mathematical biology and ecology, (see e.g. [74], [75]). We will use this powerful mathematical framework in the following.

Generating Functions, Probabilities, and Observables

Let $p_n(t)$ denote the probability at time t of observing a system in state n. For our purposes, n will represent the number of lineages and thus take value in the non-negative integers. To any such $p_n(t)$ we associate the probability generating function (PGF)

$$\psi_t(z) = \sum_{n=0}^{\infty} p_n(t) z^n. \tag{2.4}$$

The normalization of $p_n(t)$ translates to

$$\psi_t(1) = \sum_{n\geq 0} p_n(t) = 1,$$

and we can recover individual state probabilities from the Taylor coefficients of the PGF around $z = 0$,

$$p_n(t) = \frac{1}{n!}\frac{\partial^n}{\partial z^n}\psi_t(z)\Big|_{z=0} := \frac{1}{n!}\psi_t^{(n)}(0). \tag{2.5}$$

Coefficients can also be recovered using Cauchy's integral formula

$$p_n(t) = \frac{1}{2\pi i}\oint_{|z|=r}\frac{\psi_t(z)}{z^{n+1}}dz = \frac{1}{2\pi r^n}\int_0^{2\pi} e^{-in\theta}\psi_t(re^{i\theta})d\theta, \tag{2.6}$$

with $0 < r < R$ the radius of a circular contour around $z = 0$ and R the radius of convergence of $\psi_t(z)$ about $z = 0$.

A wide class of Markovian master equations with transition matrix $\boldsymbol{W}$ specifying the time evolution of $p_n(t)$

$$\frac{dp_n(t)}{dt} = \sum_m W_{nm}p_m(t) - \sum_l W_{ln}p_n(t), \tag{2.7}$$

translate at the level of the PGF (using shorthand for partial derivatives $\partial_x := \frac{\partial}{\partial x}$) to

$$\partial_t\psi_t(z) = \mathcal{L}[z, \partial_z]\psi_t(z), \tag{2.8}$$

with $\mathcal{L}[z, \partial_z]$ the stochastic generator of forward time evolution. Details on how to obtain explicitly the mapping from $\boldsymbol{W}$ to $\mathcal{L}$ can be found for example in [71], [72]. We will give explicit expressions for $\mathcal{L}$ for all stochastic processes we are interested in in the following sections and will generally omit discussing $\boldsymbol{W}$.

Given an initial distribution $\psi_0(z)$, solutions to Equation 2.8 are formally given by

$$\psi_t(z) = e^{t\mathcal{L}[z,\partial z]}\psi_0(z). \tag{2.9}$$

This formalism, called holomorphic formalism, admits three important similarity transformations[76]: two shifts,

$$e^{x\partial_z}\psi(z,\partial_z)e^{-x\partial_z} = \psi(z+x,\partial_z) \tag{2.10}$$

and

$$e^{xz}\psi(z,\partial_z)e^{-xz} = \psi(z,\partial_z - x), \tag{2.11}$$

and a scaling transformation,

$$e^{xz\partial_z}\psi(z,\partial_z)e^{-xz\partial z} = \psi(ze^x, e^{-x}\partial_z). \tag{2.12}$$

Apart from probabilities given by Equation 2.5, and omitting all matters of convergence, with the knowledge of the PGF $\psi(z)$ for a random variable N we can obtain its factorial moments by taking repeated derivatives about $z = 1$, to wit

$$\langle N(N-1)...(N-k+1)\rangle = \frac{\partial^k}{\partial z^k}\psi(z)\bigg|_{z=1}.$$

Moments μ'_k can be obtained similarly following the change of variable $z \to e^t$, namely

$$\mu'_k = \langle N^k\rangle = \frac{\partial^k}{\partial t^k}\psi(e^t)\bigg|_{t=0},$$

and thus central moments μ_k, with $\mu := \mu'_1$ the mean,

$$\mu_k = \langle (N-\mu)^k\rangle = \frac{\partial^k}{\partial t^k}\left[e^{-\mu t}\psi(e^t)\right]\bigg|_{t=0}.$$

Finally we can extract cumulants

$$\kappa_k = \frac{\partial^k}{\partial t^k}\ln\psi(e^t)\bigg|_{t=0}$$

from the logarithm of the moment generating function, also called the cumulant generating function

2.2.2 Processes

Equations 2.10-2.12 together with the chain rule are sometimes sufficient for solving simple stochastic processes like the death (pure extinction) process, the Yule (pure birth) process, and the birth-death (speciation-extinction) process. For more complicated processes like the innovation and heterogeneous innovation processes introduced below we will need to resort to solving Equation 2.8 numerically. In later sections we will further show how to combine repeated numerical solving with and exponentially-converging approximation of Equation 2.6 into a numerically exact solver for the non-equilibrium probability distribution $p_n(t)$.

The Death (Extinction) Process

The death/extinction process with per capita death/per lineage extinction rate d consists in the instantaneous transition

$$A \xrightarrow{d} \emptyset,$$

where A represents any given lineage extant at time t. Each individual lineage goes to extinction with an exponentially distributed time to extinction $\sim e^{-dt}$. For a phylogeny with n extant lineages under the effect of this process, the waiting time between two extinction events anywhere in the phylogeny is thus exponentially

distributed with total rate dn, after each of which it changes state into a phylogeny with $n-1$ extant lineages. With no other process to replenish lineages the state $n=0$ is absorbing. In terms of combinatorial classes[77],

$$z^n \xrightarrow{dn} z^{n-1},$$

and therefore the generator

$$\mathcal{L}_{death} = d(1-z)\partial_z.$$

Notice how the partial derivative operator in the generator has the effect of enumerating the number of ways of pointing and removing any one of the n lineages in a state z^n of n lineages, i.e. it transforms $z^n \to nz^{n-1}$, and thus how the full stochastic generator transforms a state z^n into $dn(z^{n-1}-z^n)$. The formal solution of its master equation

$$\psi_t(z) = e^{dt(1-z)\partial_z}\psi_0(t),$$

where we can read off the characteristic timescale dt. Using Eqs. 2.10 and 2.12 to commute shift operators around we can write

$$\begin{aligned}\psi_t(z) &= e^{-\partial_z}e^{-dtz\partial_z}e^{\partial_z}\psi_0(z),\\ &= e^{-\partial_z}e^{-dtz\partial_z}\psi_0(z+1),\\ &= e^{-\partial_z}\psi_0(ze^{-dt}+1),\\ &= \psi_0(1-e^{-dt}+e^{-dt}z).\end{aligned} \tag{2.13}$$

For a initial state with a single lineage ($\psi_0(z)=z$) at some time 0 in the past, Equation 2.13 stipulates that at time t in the present the probability that the lineage remains extant (the survival probability) is e^{-dt} and the probability that it goes extinct somewhere between time 0 and time t (the extinction probability) is $1-e^{-dt}$. Notice that the extinction probability converges asymptotically to 1, once again indicating the presence of the absorbing state.

The Birth (Yule/Speciation) Process

The Yule process is the direct counterpart of the pure death process and thus, given a per capita birth/per lineage speciation rate b, consists in the instantaneous transition

$$A \xrightarrow{b} 2A.$$

In the absence of any other processes, each lineage undergoes binary speciation events (one lineage goes to two lineages) with inter-speciation waiting time exponentially distributed $\sim e^{-bt}$. In a phylogeny with n extant lineages the waiting time between speciation events happening anywhere in the phylogeny are exponentially distributed with total rate bn. After each such event the phylogeny transition into a state with $n+1$ extant lineages. Once again, in terms of combinatorial classes

$$z^n \xrightarrow{bn} z^{n+1},$$

therefore the generator is given by

$$\mathcal{L}_{birth} = b(z-1)z\partial_z,$$

and the formal solution of its master equation

$$\psi_t(z) = e^{bt(1-z)\partial_z}\psi_0(t).$$

Before proceeding with similarity transformations we introduce the change of variable $z \mapsto 1/(z+a)$. Let $y = 1/(z+a)$ and thus $z = (1-ay)/y$, so that

$$\begin{aligned} bt(z-1)z\partial_z &= bt\frac{1-(a+1)y}{y}\frac{1-ay}{y}\frac{dy}{dz}\partial_y, \\ &= bt\frac{1-(a+1)y}{y}\frac{1-ay}{y}(-y^2)\partial_y, \\ &= -bt(ay-1)((a+1)y-1)\partial_y. \end{aligned} \tag{2.14}$$

There are two obvious choices which will both put us back in a situation almost identical to the one encountered while solving for the death process. Indeed setting $a = -1$ or $a = 0$ eliminates the $y^2\partial_z$ monomial and leaves us with a differential operator solvable by quadrature. For no particular reason we proceed with $a = 0$. We now have

$$\begin{aligned} \psi_t(z) &= e^{bt(y-1)\partial_y}\psi_0(z(y)), \\ &= e^{-\partial_y}e^{bty\partial_y}e^{\partial_y}\psi_0(z(y)), \\ &= \psi_0(z((y-1)e^{bt}+1)), \\ &= \psi_0\left(\frac{1}{(y-1)e^{bt}+1}\right), \\ &= \psi_0\left(\frac{ze^{-bt}}{1-(1-e^{-bt})z}\right). \end{aligned} \tag{2.15}$$

For an initial condition with one lineage at $t = 0$, the state at time t is a random variable $N_t \sim \text{Geom}(1-e^{-bt})$, namely a geometric distribution with time-varying parameter $1 - e^{-bt}$.

The Birth-Death (Speciation-Extinction) Process

The birth-death process is the combination of the previous two processes with instantaneous transitions

$$\begin{aligned} A &\xrightarrow{b} 2A, \\ A &\xrightarrow{d} \emptyset, \end{aligned} \tag{2.16}$$

and the the generator

$$\mathcal{L}_{bd} = (bz-d)(z-1)\partial_z.$$

Using the same change of variable $y = 1/(z+a)$

$$\mathcal{L}_{bd} = -((1+a)y-1)(b(ay-1)+dy)\partial_y.$$

Two values for a once again cancel the $y^2\partial_y$ monomial, namely $a = -1$ and $a = -r := -d/b$. Setting $a = -r$ and defining $\Delta := b - d$,

$$\mathcal{L}_{bd} = \Delta\left(y - \frac{1}{1-r}\right)\partial_y.$$

Proceeding as before along our well threaded path,

$$\begin{aligned}\psi_t(z) &= e^{-\frac{1}{1-r}\partial_y} e^{\Delta t y \partial_y} e^{\frac{1}{1-r}\partial_y} \psi_0(z(y)), \\ &= \psi_0\left(r + \frac{1}{\frac{1}{1-r} + \left(\frac{1}{z-r} - \frac{1}{1-r}\right) e^{\Delta t}}\right).\end{aligned} \tag{2.17}$$

After a fair bit of algebra, during which we are naturally led to define $\omega(t) := e^{\Delta t}$ and $p_0(t) := r(\omega - 1)/(\omega - r)$ from isolating of the absorbing state contribution at $z = 0$,

$$\psi_t(z) = \psi_0\left(p_0(t) + (1 - p_0(t))\frac{\left(1 - \frac{p_0(t)}{r}\right) z}{1 - \frac{p_0(t)}{r} z}\right). \tag{2.18}$$

The expansion about $z = 0$ of the argument of $\psi_0(.)$ readily gives us the explicit time-dependent probabilities for a birth-death process with initial condition $p_n(0) = \delta_{1,n}$,

$$p_n(t) = \begin{cases} p_0(t), & n = 0, \\ (1 - p_0(t))\left(1 - \frac{p_0(t)}{r}\right)\left(\frac{p_0(t)}{r}\right)^{n-1}, & n \geq 1. \end{cases} \tag{2.19}$$

In other words, at time t we have a product random variable

$$N_t \sim \text{Bernoulli}(1 - p_0(t))\,\text{Geom}(1 - p_0(t)/r).$$

Innovation (Yule/Geometric Burst) Process

Consider a phylogeny of total depth T sporadically undergoing at rate ρ a process whereby we initiate a Yule process happening on a short timescale $\tau \ll T$ with a high speciation rate σ in such a a way that the product $\sigma\tau \sim \mathcal{O}(1)$. We can think of ρ as the rate at which innovations arise and open previously unexplored regions of niche space, allowing thus a period (over timescale τ) of rapid diversification (with diversification rate σ), followed by a saturation once ecological processes equilibrate anew. Looking at Equation 2.15 we see that the only important parameter that controls the characteristic size of bursts is $g := 1 - e^{-\sigma\tau}$, $0 < g < 1$ and is itself parametrized by fixed values of the product $\sigma\tau$. Indeed the mean burst size $\bar{k} = 1/(1 - g)$. Symbolically, each lineage

$$A \xrightarrow{\rho(1-g)g^{k-1}} kA, \quad k \geq 1, \tag{2.20}$$

and thus in terms of combinatorial classes

$$z^n \xrightarrow{\rho n(1-g)g^{k-1}} z^{n+k-1}.$$

The generator can now be constructed using the expansion for $1/(1 - xz)$ around $z = 0$. Alternatively, inspection of Equation 2.15 for the Yule process gives us an equivalent but more succinct combinatorial representation

$$z^n \xrightarrow{\rho n} \frac{(1 - g)}{1 - gz} z^n$$

and therefore its stochastic generator

$$\begin{aligned}\mathcal{L}_{geom} &= \rho\left(\frac{(1-g)}{1-gz}-1\right)z\partial_z,\\ &= \rho g\frac{(z-1)z}{1-gz}\partial_z.\end{aligned} \tag{2.21}$$

All models for which the product $\sigma\tau$ is equal are indistinguishable from one another. In other words g parametrizes a whole classes of innovation processes. The precise values of σ and τ can not be determined, only their product. If a geometric burst process happens over a timescale half as long and with a diversification rate twice as high as another diversification process it is then indistinguishable from the latter. Indeed $\sigma\tau = (2\sigma)(\tau/2) = (a\sigma)(\tau/a)$.

It is not obvious whether there exists a simple change of variable $z = f(y)$ that would transform the above expression into one with $y\partial_y$ as the highest degree monomials. If there were we could then solve the master equation exactly for the innovation process. We did attempt to do so by trying many different choice of change of variable but in the end seeking an analytic expression proved intractable. Indeed the factor $1/(1-gz)$ generates an infinite number of monomials with increasing degrees of z. Nonetheless in the next section we will describe a numerical method that allows us to extract exact time-varying probabilities for processes of this kind. Yet before doing so we would like to introduce one additional process at the core of this work.

Heterogeneous Innovation (Beta-Geometric Burst) Process

We have seen how a geometric burst process with a given value of g characterizes a whole class of geometric processes with rescaled rates and temporal extents. In phylogenies spanning billions of years of evolution across vast geographic regions there is no reason to expect that all bursts neatly fall within one given class, and therefore we need to introduce an additional degree of freedom to capture the heterogeneity across those classes. To do so we make g a random variable which follows a beta distribution $\text{Beta}(g|\alpha,\beta)$. The two parameters α and β control the weight given to values of g close to 0 and 1, respectively, and therefore the weight of burst processes with characteristically small and large mean burst sizes. Explicitly, and substituting η in place of ρ to distinguish between the rate of geometric innovation and the rate of heterogeneous innovation,

$$\begin{aligned}\int_0^1 \eta(1-g)g^{k-1}\,\text{Beta}(g|\alpha,\beta)dg &= \frac{\eta}{B(\alpha,\beta)}\int_0^1 g^{\alpha-1}(1-g)^{\beta-1}(1-g)g^{k-1}dg,\\ &= \frac{\eta}{B(\alpha,\beta)}\int_0^1 g^{(\alpha+k-1)+1}(1-g)^{(\beta+1)-1}dg,\\ &= \eta\frac{B(\alpha+k-1,\beta+1)}{B(\alpha,\beta)},\\ &= \eta\beta\frac{\Gamma(\alpha+k-1)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)\Gamma(\alpha)},\end{aligned} \tag{2.22}$$

where the beta function $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, and therefore symbolically

$$A \xrightarrow{\eta\beta\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\frac{\Gamma(\alpha+k-1)}{\Gamma(\alpha+\beta+k)}} kA.$$

Similarly for the generator we compound Equation 2.21 with a beta distribution as follows:

$$\begin{aligned}\mathcal{L}_{betageom} &= \frac{\eta(z-1)z}{B(\alpha,\beta)}\int_0^1 g^{\alpha-1}(1-g)^{\beta-1}\frac{g}{1-gz}dg\ \partial_z \\ &= \frac{\eta(z-1)z}{B(\alpha,\beta)}\int_0^1 \frac{g^{\alpha}(1-g)^{\beta-1}}{1-gz}dg\ \partial_z, \\ &= \eta\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}{}_2F_1(1,\alpha+1,\alpha+\beta+1;z)(z-1)z\partial_z, \\ &= \eta\frac{\alpha}{\alpha+\beta}{}_2F_1(1,\alpha+1,\alpha+\beta+1;z)(z-1)z\partial_z.\end{aligned} \tag{2.23}$$

where in going from the first to the second line we used the integral representation of Euler type for the ordinary hypergeometric function ${}_2F_1$. Omitting all details, the combinatorial class of the heterogeneous innovation process is represented by the tranformation

$$z^n \xrightarrow{\eta n} \frac{\beta}{\alpha+\beta}{}_2F_1(1,\alpha,\alpha+\beta+1;z)z^n. \tag{2.24}$$

Finally we want to mention that Equation 2.22 at large k,

$$A \xrightarrow{\overset{k\to\infty}{\sim} k^{-\beta-1}} kA,$$

which indicates that in this model the burst size distribution is a power-law with β controlling the exponent of the tail. This distribution has neither a finite mean nor variance. This is surprising given that it compounds geometric innovation models which themselves have finite mean and variance.

Incomplete Lineage Sampling

Except perhaps for large-scale datasets like The Earth Microbiome, we do not expect any given sample to contain all OTUs simply by virtue of limits on the experimenter's resources and sampling effort. This inherent incomplete lineage sampling can be approximated by a Bernouilli trial with success probability (or sampling

fraction) f. Given an unsampled PGF $\psi(z)$, the modified PGF following a sampling event is given by

$$\begin{aligned}
\psi(1-f+fz) &= \sum_{n=0}^{\infty} p_n(1-f+fz)^n, \\
&= \sum_{n=0}^{\infty}\sum_{k=0}^{n} p_n \binom{n}{k}(1-f)^k f^{n-k} z^{n-k}, \\
&= p_0\binom{0}{0}(1-f)^0(fz)^0, \\
&+ p_1\left[\binom{1}{0}(1-f)^0(fz)^1 + \binom{1}{1}(1-f)^1(fz)^0\right] \\
&+ p_2\left[\binom{2}{0}(1-f)^0(fz)^2 + \binom{2}{1}(1-f)^1(fz)^1 + \binom{2}{2}(1-f)^2(fz)^0\right] \\
&+ \ldots, \\
&= p_0\binom{0}{0}(1-f)^0(fz)^0 + p_1\binom{1}{1}(1-f)^1(fz)^0 + p_2\binom{2}{2}(1-f)^2(fz)^0 + \ldots \\
&+ p_1\binom{1}{0}(1-f)^0(fz)^1 + p_2\binom{2}{1}(1-f)^1(fz)^1 + p_3\binom{3}{2}(1-f)^2(fz)^1 + \ldots \\
&+ p_2\binom{2}{0}(1-f)^0(fz)^2 + p_3\binom{3}{1}(1-f)^1(fz)^2 + p_4\binom{4}{2}(1-f)^2(fz)^2 + \ldots \\
&+ \ldots \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} p_{n+m}\binom{n+m}{m}(1-f)^n f^m z^m.
\end{aligned} \tag{2.25}$$

To go from line three to line four and five we regroup terms that fall along the diagonals of line two.

This expression follows from the straightforward substitution $z \to 1-f+fz$ which stipulates that each lineage in the PGF is left intact if it is successfully sampled with probability f, and otherwise it is unsuccessfully sampled with probability $1-f$ and replaced by the null combinatorial object 1. Indeed one should see $1-f+fz$ as a stand-in for $(1-f)z^0 + (f)z^1$. The right-hand side of Equation 2.25 is also straightforward to interpret. In words, the probability of having n lineages at time t encapsulated in the coefficients of the power series of $\psi_t(z)$ are replaced after the shift $\psi(z) \to \psi(1-f+fz)$ by probabilities of observing n lineages after sampling with intensity f. All states with n or more lineages (hence the $n+k$, $k \geq 0$) now contribute to the probability of *observing* n lineages in a sample provided that exactly n of them are successfully sampled with joint success probability f^n and the k reminding ones all fail to be sampled with joint failure probability $(1-f)^k$. Finally the binomial coefficient $\binom{n}{k}$ accounts for the number of equivalent reorderings of n successes and k failures.

2.2.3 Numerically Exact Solution of the Master Equation

Method of Characteristics for First Order Quasi-Linear PDE

We mentioned above that master equations with generators containing a geometric or hypergeometric burst processes (Eqs. 2.21 and 2.23) are for all practical purposes intractable to solve analytically. Luckily all these processes and their combinations lead to quasi-linear master equations of first order of the form

$$\partial_t \mathcal{U}_t(z) = h(z)\partial_z \mathcal{U}_t(z). \tag{2.26}$$

which are in principal easily solved by the method of characteristics. We will moreover only have to care about solving such equations with initial conditions $\mathcal{U}_0(z) = z$. With those provisions, the method of characteristics transforms the partial differential equation Equation 2.26 into an ordinary autonomous differential equation where

$$\begin{aligned} \mathcal{U}_t(z) &= u_t, \\ \frac{du_t}{dt} &= h(z), \\ u_0 &= z. \end{aligned} \tag{2.27}$$

For the BD, BDI, and BDH models we have

$$\begin{aligned} h_{BD}(u) &= (bu - d)(u - 1), \\ h_{BDI}(u) &= \left(bu - d + \rho\frac{gu}{1 - gu}\right)(u - 1), \\ h_{BDH}(u) &= \left(bu - d + \eta\frac{\alpha}{\alpha + \beta}{}_2F_1(1, \alpha + 1, \alpha + \beta + 1, u)\right)(u - 1). \end{aligned} \tag{2.28}$$

Once again those equations are not solvable analytically (except for the BD model), but they are solvable numerically.

Hypergeometric Function ${}_2F_1$

As one can see the hypergeometric function ${}_2F_1$ makes an appearance in the equation of the BDH model. It is therefore imperative to have on hand a way to rapidly evaluate this function for various values of α, β, and z. Numerical functions are readily available through the `scipy` python package, but we have found their implementation to be unstable and incorrect in various region of the complex plane. We experimented with the `mppath` python package for multi-precision arithmetic but while the accuracy is excellent its slow speed made it completely impractical to use for our purposes. We therefore implemented our own method, `hyp2f1a1` to evaluate ${}_2F_1(a, b, c, z)$ for the special case $a = 1$. To do so we use Gauss' continuous fraction representation set at $a = 1$,

$${}_2F_1(1, b, c, z) = \cfrac{1}{1 + \cfrac{-bz}{c + \cfrac{(b - c)z}{c + 1 + \cfrac{-c(b + 1)z}{c + 2 + \cfrac{2(b - c - 1)z}{c + 3 + \cfrac{-(c + 1)(b + 2)z}{c + 4 + \ddots}}}}}}, \tag{2.29}$$

which we rewrite

$$
{}_2F_1(1,b,c,z) = \cfrac{1}{1-\cfrac{\frac{bz}{c}}{1-\cfrac{\frac{(c-b)z}{c(c+1)}}{1-\cfrac{\frac{c(b+1)z}{(c+1)(c+2)}}{1-\cfrac{\frac{2(c-b+1)z}{(c+2)(c+3)}}{1-\cfrac{\frac{(c+1)(b+2)z}{(c+3)(c+4)}}{1-\ddots}}}}}}. \tag{2.30}
$$

This form allows us to use the forward series recurrence algorithm[78] to evaluate the continued fraction in a stable way.

2.2.4 Observed Past Subtree Generating Function (OPSGF)

Solving the master equation gives us the probability generating function $\mathcal{U}_t(z) = \sum_{k=0}^{\infty} p_k(t) z^k$ of the probability distribution $p_k(t), k \geq 0$ of observing k lineages at time t given that we started at time 0 with only one lineage. Alternatively, if we take t to increase towards the past, if we start at time t in the past and observe at $t = 0$. Those are not yet the probabilities of interest. Each subtree within a phylogeny does start with one lineage at time t in the past and subtend k descendant lineages at some time s closer to the present, but only is each of those lineages at time s have at least one survive lineage at $t = 0$. To account for this we need must condition on survival of every one of the k lineages of the subtree, and on the survival of the initial lineage itself. Moreover we want to consider not only survival in terms of not going extinct, but of have been successfully sampled as well.

Generating Function Approach

We first describe the generating function approach to obtaining observed past subtree probabilities using the OPSGF. If this approach seems a bit obscure and too expedient we will show in the next subsection how to recover this conditional subtree generating function using a simpler but more involved probabilistic approach.

Consider first the probability $P(\text{extinct or extant but not sampled})$ of a lineage starting at time s to have its descendant either go extinct at some point between time s and 0, or to remain extant at time 0 yet not successfully sampled with probability $1 - f$. We can write this probability

$$
P(\text{extinct or extant but not sampled}) = p_0(s) + \sum_{k \geq 1} p_k(s)(1-f)^k. \tag{2.31}
$$

Within this sum the first term $p_0(s)$ is the probability that all descendants go extinct at some point between time s and the present, and all other terms $p_k(s)(1-f)^k$ are the probabilities that accounts for cases when k lineages remain extant but were unsuccessfully sampled under a sampling effort f. In term of the probability generating function this is simply

$$
P_s(\text{extinct or extant but not sampled}) = \mathcal{U}_s(1-f). \tag{2.32}
$$

It follows immediately that the probability of being both extant and successfully sampled

$$P_s(\text{extant and sampled}) = 1 - \mathcal{U}_s(1-f). \tag{2.33}$$

Consider now a lineage that starts at time t and evolves until time $s \leq t$ with probability generating function $\mathcal{U}_{t-s}(z)$. The atom z marks an 'extant' lineages present at time s, with $z^0 = 1$ the constant term associated with the probability of no lineages being present, i.e. being all extinct, at time s. It follows that the presence of n extant lineages will be associate with the monomial z^n times the probability of this event. We substitute each lineage z by individually splitting them into their descendant lineage which will go extinct between s and the present or are extant yet fail to be sampled, which we replace by the empty atom $z^0 = 1$ weighted by the probability $\mathcal{U}_s(1-f)$ of such an event, and those that are extant and sampled which we will mark with the atom y weighted by the associated probability of such an event, namely $y(1-\mathcal{U}_s(1-f))$. In other words we apply the substitution $z \mapsto \mathcal{U}_s(1-f) + y(1-\mathcal{U}_s(1-f))$. Therefore the unconditional PSGF

$$\begin{aligned}\tilde{\Phi}_f(y,t,s) &= \mathcal{U}_{t-s}\left(\mathcal{U}_f(1-f) + y(1-\mathcal{U}_s(1-f))\right), \\ &= \sum_{k\geq 0} \tilde{\phi}^{(k)}(t,s) y^k.\end{aligned} \tag{2.34}$$

We say unconditional because this generating function includes a constant term, i.e. the term with atom $y^0 = 1$, weighted by the probability $\tilde{\phi}^{(0)}(t,s)$ of not observing any lineage in the present, that is it does not remove the event whereby the whole subtree goes extinct. This probability

$$\begin{aligned}\tilde{\phi}_f^{(0)}(t,s) &= \tilde{\Phi}_f(0,t,s) \\ &= \mathcal{U}_{t-s}(\mathcal{U}_t(1-f)), \\ &= \mathcal{U}_t(1-f).\end{aligned} \tag{2.35}$$

Naturally this case can never be observed without fossil data and/or time-series. To condition on at least one lineage being extant and obtain the OPSGF, we subtract this probability and renormalize by the remaining total probability. The resulting generating function

$$\begin{aligned}\Phi_f(y,t,s) &= \frac{\tilde{\Phi}(y,t,s) - \tilde{\Phi}_f(0,t,s)}{1-\tilde{\Phi}_f(0,t,s)}, \\ &= \frac{\mathcal{U}_{t-s}(\mathcal{U}_s(1-f) + y(1-\mathcal{U}_s(1-f))) - \mathcal{U}_t(1-f)}{1-\mathcal{U}_t(1-f)}, \\ &= \sum_{k\geq 1} \phi_f^{(k)}(t,s) y^k.\end{aligned} \tag{2.36}$$

One can immediately see that $\phi_f^{(k)}(t,s),\ k \geq 1$ is a genuine normalized probability distribution by setting

$y = 1$. Indeed

$$\begin{aligned}
\Phi_f(1,t,s) &= \sum_{k\geq 1} \phi_f^{(k)}(t,s), \\
&= \frac{U_{t-s}(\mathcal{U}_s(1-f)+1-\mathcal{U}_s(1-f)) - U_t(1-f)}{1-\mathcal{U}_t(1-f)}, \\
&= \frac{\mathcal{U}_{t-s}(1) - \mathcal{U}_t(1-f)}{1-\mathcal{U}_t(1-f)}, \\
&= \frac{1-\mathcal{U}_t(1-f)}{1-\mathcal{U}_t(1-f)}, \\
&= 1.
\end{aligned} \tag{2.37}$$

Probabilistic Approach

The unconditional probability $\tilde{\phi}^{(k)}(t,s)$ of observing $k \geq 0$ lineages (technically we can not observe $k = 0$ unless we have a fossil record) at time s in the past descending from a unique lineage starting at time $t > s$ is given by the probability of going from 1 lineage at time t to $n + k$ lineages at time s times the probability that each of the k observed lineages have at least 1 observed extant and sampled lineage times the probability that every of the n lineages are extinct or if they are extant they were not sampled. We will make explicit the transition from m lineage to n lineages using the notation $m \to n$. The probability

$$\tilde{\phi}^{(k)}(t,s) = \sum_{n=0}^{\infty} p_{1\to n+k}(t-s)\binom{n+k}{k}\sum_{j=0}^{\infty} p_{n\to j}(s)(1-f)^j(1-\sum_{m=0}^{\infty} p_{1\to m}(s)(1-f)^m). \tag{2.38}$$

Using the independence between lineages we can write

$$\sum_{j=0}^{\infty} p_{n\to j}(t)z^j = \left(\sum_{j=0}^{\infty} p_{1\to j} z^j\right)^n$$

and therefore

$$\begin{aligned}
\tilde{\phi}^{(k)}(t,s) &= \sum_{n=0}^{\infty} p_{1\to n+k}(t-s)\binom{n+k}{k}\left(\mathcal{U}_s(1-f)\right)^n\left(1-\mathcal{U}_s(1-f)\right)^k, \\
\Rightarrow \sum_{k=0}^{\infty}\tilde{\phi}^{(k)}(t,s)y^k &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p_{1\to n+k}(t-s)\binom{n+k}{k}\left(\mathcal{U}_s(1-f)\right)^n\left(1-\mathcal{U}_s(1-f)\right)^k y^k, \\
&= \mathcal{U}_{t-s}(\mathcal{U}_s(1-f) + y(1-\mathcal{U}_s(1-f))), \\
&= \tilde{\Phi}_f(y,t,s),
\end{aligned} \tag{2.39}$$

where we used the identity Equation 2.25 to go from the second to the third line. The rest of the reasoning is the same; the observed past subtree distribution $\phi_f^{(k)}(t,s)$, $k \geq 1$ is simply obtained by omitting $\tilde{\phi}_f^{(0)}(t,s)$ from the distribution and renormalizing, namely

$$\phi_f^{(k)}(t,s) = \begin{cases} 0, & k = 0, \\ \frac{\tilde{\phi}_f^{(k)}(t,s)}{1-\tilde{\phi}_f^{(0)}(t,s)}, & k \geq 1, \end{cases} \tag{2.40}$$

for which it is easy to show that its probability generating function is nothing but the OPSGF Equation 2.36.

Coarse-Grain Property of the OPSGF

The OPSGF also satisfies the Chapman-Kolmogorov property

$$\Phi_f(\Phi(u,s,y),t,u) = \Phi_f(y,t,s).$$

This can be verified using Equation 2.36 and the Chapman-Kolmogorov property of $\mathcal{U}_t$ itself, to wit

$$\begin{aligned}
&\Phi_f(\Phi_f(y,r,s),t,r) = \\
&\frac{\mathcal{U}_{t-r}(\mathcal{U}_r(1-f) + \left[\frac{\mathcal{U}_{r-s}(\mathcal{U}_s(1-f)+y(1-\mathcal{U}_s(1-f)))-\mathcal{U}_r(1-f)}{1-\mathcal{U}_r(1-f)}\right](1-\mathcal{U}_r)) - \mathcal{U}_t(1-f)}{1-\mathcal{U}_t(1-f)}, \\
&= \frac{\mathcal{U}_{t-r}(\mathcal{U}_{r-s}(\mathcal{U}_s(1-f)+y(1-U_s(1-f))) - \mathcal{U}_t(1-f)}{1-\mathcal{U}_t(1-f)}, \\
&= \frac{\mathcal{U}_{t-s}(\mathcal{U}_s(1-f)+y(1-\mathcal{U}_s(1-f))) - \mathcal{U}_t(1-f)}{1-U_t(1-f)}, \\
&= \Phi_f(y,t,s). \qquad \square
\end{aligned} \tag{2.41}$$

Composing and collecting terms of the power series (and shortening the notation),

$$\begin{aligned}
\Phi_f(\Phi_f(y,r,s),t,r) = &\sum_{k\geq 1} \phi^1_{tr}\left(\phi^1_{rs}y + \phi^2_{rs}y^2 + \phi^3_{rs}y^3 + \phi^4_{rs}y^4 + \ldots\right) \\
&+ \sum_{k\geq 1} \phi^2_{tr}\left(\phi^1_{rs}y + \phi^2_{rs}y^2 + \phi^3_{rs}y^3 + \phi^4_{rs}y^4 + \ldots\right)^2 \\
&+ \sum_{k\geq 3} \phi^3_{tr}\left(\phi^1_{rs}y + \phi^2_{rs}y^2 + \phi^3_{rs}y^3 + \phi^4_{rs}y^4 + \ldots\right)^3 \\
&+ \sum_{k\geq 4} \phi^4_{tr}\left(\phi^1_{rs}y + \phi^2_{rs}y^2 + \phi^3_{rs}y^3 + \phi^4_{rs}y^4 + \ldots\right)^4 \\
&+ \ldots, \\
&= \phi^1_{rs}\phi^1_{rs}y \\
&+ \left(\phi^1_{tr}\phi^2_{rs} + \phi^2_{tr}\phi^1_{rs}\phi^1_{rs}\right)y^2 \\
&+ \left(\phi^1_{tr}\phi^3_{rs} + \phi^2_{tr}\left(\phi^1_{rs}\phi^2_{rs} + \phi^2_{rs}\phi^1_{rs}\right) + \phi^3_{tr}\phi^1_{rs}\phi^1_{rs}\phi^1_{rs}\right)y^3 \\
&+ \left(\phi^1_{tr}\phi^4_{rs} + \phi^2_{rs}\left(\phi^1_{rs}\phi^3_{rs} + \phi^3_{rs}\phi^1_{rs} + \phi^2_{rs}\phi^2_{rs}\right)\right. \\
&\left.+\phi^3_{ts}\left(\phi^1_{rs}\phi^1_{rs}\phi^2_{rs} + \phi^1_{rs}\phi^2_{rs}\phi^1_{rs} + \phi^2_{rs}\phi^1_{rs}\phi^1_{rs}\right) + \phi^4_{ts}\phi^1_{rs}\phi^1_{rs}\phi^1_{rs}\phi^1_{rs}\right)y^4 \\
&+ \ldots.
\end{aligned} \tag{2.42}$$

Induction gives us the decomposition

$$\begin{aligned}
\phi^{(k)}_f(t,s) &= \sum_{c\in\mathrm{Comp}(k)} \phi^{(|c|)}(t,u)\prod_{\lambda\in c}\phi^{(\lambda)}_f(u,s), \\
&= \sum_{\pi\in\mathrm{Part}(k)} |\mathrm{Comp}(\pi)|\phi^{(|c|)}_f(t,u)\prod_{\lambda\in\pi}\phi^{(\lambda)}_f(u,s).
\end{aligned} \tag{2.43}$$

in terms of partitions and compositions. Here $\mathrm{Comp}(k)$ is the set of compositions of the number k and $|c|$ the number of parts in a given composition. Similarly $\mathrm{Part}(k)$ is the set of partitions of k. By an abuse of notation $\mathrm{Comp}(\pi)$ is also the set of compositions equivalent to a partition π. Two compositions are said

to be equivalent to a partition π iff both multisets of their parts are equals, e.g. given the number 15, a composition $[5,4,1,5] \sim [1,5,5,4]$, but $[5,5,4,1] \not\sim [5,4,4,2]$. For a partition π of k, the size of the set of its equivalent compositions is given by the multinomial coefficient.

$$|\text{Comp}(\pi)| = \binom{|\pi|}{m_\pi(1), m_\pi(2), \ldots, m_\pi(k)} = \frac{|\pi|!}{\prod_{i=1}^{k} m_\pi(i)!} \tag{2.44}$$

where $m_\pi(i)$ is the multiplicity of parts of size i in k. For example for $k = 15$ and

$$\pi = [5,5,2,1,1,1] \Rightarrow m_\pi(i) = \begin{cases} 3, & i = 1 \\ 1, & i = 2, \\ 2, & i = 5, \\ 0, & i \in \{3,4,6,7,\ldots,15\}, \end{cases} \tag{2.45}$$

and thus $|\text{Comp}(\pi)| = 60$. In other words slicing a subtree of size k in two at some time u between t and s is equivalent to all ways of stitching subtrees from u to s with sizes $\sum \lambda = k$ at the crown of all subtrees from t to u of sizes $|c| \leq k$. Figure 2.3 shows the equivalent graphical decomposition of the above expansion. This property naturally gives us the Hastings ratio we will use in the Metropolis-Hastings MCMC algorithm described in Section 2.3 to sample coarse-grained tree using the space of integer multipartitions.

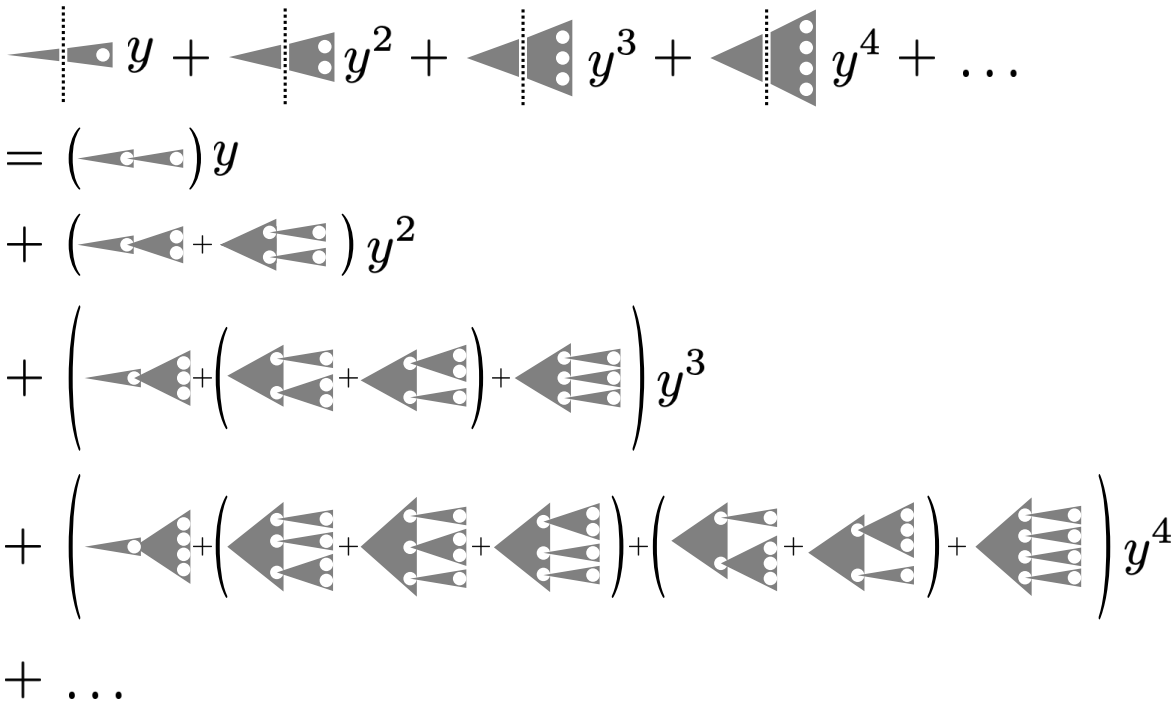


Figure 2.3: Graphical representation of the Chapman-Kolmogorov decomposition of observed past subtree probabilities. Terms in small parentheses give equal contributions.

2.2.5 Coarse-Grained Tree Likelihood

The likelihood L of observing a coarse-grained tree $\mathcal{T}_K$ represented by coarse-grain data $\{(t_i, s_i, k_i)\}_i$ given a sampling fraction f and a model with parameter values θ is the product of the observed past subtree probabilities of the coarse-grain data, namely

$$L[\mathcal{T}_K|f,\theta] = \prod_{i\in\mathcal{I}_{\tau_K}} \phi_f^{(k_i)}(t_i, s_i|\theta). \tag{2.46}$$

There are interesting cancellations in the above expression which are not explicitly mentioned in previous writings of the likelihood e.g. of the BD model applied to bifurcating tree[79]–[82]. Terms involved in these cancellations are essential in making $\phi^{(k)}(t,s)$ an actual probability distribution and writing e.g. partial likelihood that does not include all subtrees. Omitting subtrees is sometimes useful in speeding up the inference process. Using the notation of Equation 2.2 the coarse-grained likelihood

$$L[\mathcal{T}_K] = \prod_{\sigma=1\ldots K}\prod_{i\in\mathcal{I}_\sigma} \phi_f^{(k_{\sigma i})}(t_{\sigma i}, s_{\sigma i}|\theta). \tag{2.47}$$

We know that

$$\begin{aligned}\phi^{(k)}(t,s) &= \frac{1}{k!}\left.\frac{\partial^k \Phi(y,t,s)}{\partial y^k}\right|_{y=0},\\ &= \frac{(1-\mathcal{U}_s(1-f))^k}{1-\mathcal{U}_t(1-f)}\mathcal{U}_{t-s}^{(k)}(\mathcal{U}_s(1-f)).\end{aligned} \tag{2.48}$$

and therefore

$$L[\mathcal{T}_K] = \prod_{\sigma=1\ldots K}\prod_{i\in\mathcal{I}_\sigma} \frac{(1-\mathcal{U}_{s_{\sigma i}}(1-f))^{k_{\sigma i}}}{1-\mathcal{U}_{t_{\sigma i}}(1-f)}\mathcal{U}_{t_{\sigma i}-s_{\sigma i}}^{(k_{\sigma i})}(\mathcal{U}_{s_{\sigma i}}(1-f)). \tag{2.49}$$

Using the identity Equation 2.3 it follows that

$$L[\mathcal{T}_K] = \frac{f^{\sum_{i\in\mathcal{I}_1} k_i}\prod_{i\in\mathcal{I}_1} U_\Delta^{(k_{1i})}(1-f)}{1-\mathcal{U}_T(1-f)} \prod_{\sigma=2\ldots K}\prod_{i\in\mathcal{I}_\sigma} \mathcal{U}_\Delta^{(k_{\sigma i})}(\mathcal{U}_{(\sigma-1)\Delta}(1-f)). \tag{2.50}$$

where we obtain the f factor from the use of Equation 2.9, to wit

$$\begin{aligned}1-\mathcal{U}_{s_{1i}}(z)|_{z=1-f} &= 1-\mathcal{U}_0(z)|_{z=1-f},\\ &= 1-e^{t\mathcal{L}[z,\partial z]}z\Big|_{\substack{t=0\\ z=1-f}},\\ &\sim t\mathcal{L}[z,\partial z]z|_{\substack{t=0\\ z=1-f}},\\ &= f.\end{aligned} \tag{2.51}$$

.

Equivalence with Traditional Macroevolutionary Tree Likelihoods

In this subsection we want to show that in the limit of an infinitesimal coarse-graining $\Delta \to \delta t$, Equation 2.50 for the BD model recovers the traditional BD likelihoods found e.g. in[79]–[81]. Since those likelihoods

decompose the tree in terms of branches rather than subtrees we first need to show the identify

$$\phi_f^{(1)}(t,r)\phi_f^{(1)}(r,s) = \phi_f^{(1)}(t,s). \tag{2.52}$$

Equation 2.43 gives us this identity for free but here we will prove it explicitly using the chain rule. To begin with, the first derivative of the Chapman-Kolmogorov identity

$$\begin{aligned} &\frac{\partial}{\partial z}\mathcal{U}_{t-s}(z) = \frac{\partial}{\partial z}\mathcal{U}_{t-r}(\mathcal{U}_{r-s}(z)) \\ &\Rightarrow \mathcal{U}^{(1)}_{t-s}(z) = \mathcal{U}^{(1)}_{t-r}(\mathcal{U}_{r-s}(z))\mathcal{U}^{(1)}_{r-s}(z). \end{aligned} \tag{2.53}$$

Next we evaluate both sides at $z = \mathcal{U}_s(1-f)$ and obtain

$$\begin{aligned} \mathcal{U}^{(1)}_{t-s}(\mathcal{U}_s(1-f)) &= \mathcal{U}^{(1)}_{t-r}(\mathcal{U}_{r-s}(\mathcal{U}_s(1-f)))\mathcal{U}^{(1)}_{r-s}(\mathcal{U}_s(1-f)), \\ &= \mathcal{U}^{(1)}_{t-r}(\mathcal{U}_r(1-f))\mathcal{U}^{(1)}_{r-s}(\mathcal{U}_s(1-f)). \end{aligned} \tag{2.54}$$

Finally,

$$\begin{aligned} \phi_f^{(1)}(t,s) &= \mathcal{U}^{(1)}_{t-s}(\mathcal{U}_s(1-f))\frac{1-\mathcal{U}_s(1-f)}{1-\mathcal{U}_t(1-f)}, \\ &= \mathcal{U}^{(1)}_{t-r}(\mathcal{U}_r(1-f))\mathcal{U}^{(1)}_{r-s}(\mathcal{U}_s(1-f))\frac{1-\mathcal{U}_s(1-f)}{1-\mathcal{U}_t(1-f)}, \\ &= \frac{\mathcal{U}^{(1)}_{t-r}(\mathcal{U}_r(1-f)}{1-\mathcal{U}_t(1-f)}\left[\frac{1-\mathcal{U}_r(1-f)}{1-\mathcal{U}_r(1-f)}\right]\mathcal{U}^{(1)}_{r-s}(\mathcal{U}_s(1-f))(1-\mathcal{U}_s(1-f)), \\ &= \phi_f^{(1)}(t,r)\phi_f^{(1)}(r,s). \qquad \square \end{aligned} \tag{2.55}$$

The decomposition of a tree in [81] is constructed by associating a probability $P(\dashv, t, s)$ of a lineage to evolve from some time t until some time s at which point it undergoes a speciation event into two lineages. Under infinitesimal coarse-graining $\mathcal{R}_\infty$ a single lineage followed by such an event contributes to our likelihood

$$\begin{aligned} P(\dashv, t, s) &= \phi_f^{(2)}(s+\delta t, s) \prod_{i=0}^{N=(t-s-\delta t)/\delta t} \phi_f^{(1)}(t-i\delta t, t-(i+1)\delta t), \\ &= \phi_f^{(1)}(t-s-\delta t)\phi_f^{(2)}(s+\delta t, s). \end{aligned} \tag{2.56}$$

The contribution from the speciation event at the node

$$\phi_f^{(2)}(s+\delta t, s) = \frac{1}{2}\frac{(1-\mathcal{U}_s(1-f))^2}{1-\mathcal{U}_{s+\delta t}(1-f)}\mathcal{U}^{(2)}_{\delta t}(\mathcal{U}_s(1-f)). \tag{2.57}$$

Expanding $\mathcal{U}_{\delta t}$ to first order in δt for the BD model,

$$\begin{aligned} \mathcal{U}^{(2)}_{\delta t}(\mathcal{U}_s(1-f)) &\sim \frac{\partial^2}{\partial z^2}(1+\delta t(bz-d)(z-1)\partial_z)z\Big|_{z=\mathcal{U}_s(1-f)}, \\ &= 2b\delta t, \end{aligned} \tag{2.58}$$

and therefore

$$\phi_f^{(2)}(s+\delta t, s) = b\delta t\frac{(1-\mathcal{U}_s(1-f))^2}{1-U_{s+\delta t}(1-f)}. \tag{2.59}$$

Effecting all cancellations mentioned in the previous section and setting $s + \delta t \to s$ we find after a little algebra that the likelihood

$$\mathcal{L}[\mathcal{T}_\infty] = \frac{f^L b^{L-1} \delta t^{L-1} \prod_{i \in \text{branches}} \mathcal{U}^{(1)}_{t_i - s_i}(\mathcal{U}_{s_i}(1-f))}{1 - \mathcal{U}_T(1-f)}. \tag{2.60}$$

for a tree of size L and thus $L-1$ internal nodes. Changing units $\lambda = b = 1$, $r = \mu/\lambda = d/b$, and letting $\omega_t = e^{(1-r)t}$, we can write Equation 2 in [81] for constant birth and death rates

$$\Phi(t) = 1 - \frac{w_t}{\frac{1}{f} + \frac{w_t - 1}{1-r}}. \tag{2.61}$$

We claim now that for the BD model,

$$\mathcal{U}_T(1-f) = \Phi(T) \tag{2.62}$$

where we changed their t to our notation $t = T$. In this subsection Φ is their notation and not to be confused with the OPSGF. Our Equation 2.18 is written in term of the absorption probability

$$p_0(T) = r\frac{\omega_T - 1}{\omega_T - r}$$

and therefore

$$\omega_T = \frac{1 - p_0(T)}{1 - p_0(T)/r}.$$

Using this expression for ω_T and after some algebra

$$\begin{aligned}
\Phi(T) &= 1 - \frac{\frac{1-p_0(T)}{1-p_0(T)/r}}{\frac{1}{f} + \frac{p_0(T)/r}{1-p_0(T)/r}}, \\
&= p_0(T) + 1 - \frac{\frac{1-p_0(T)}{1-p_0(T)/r}}{\frac{1}{f} + \frac{p_0(T)/r}{1-p_0(T)/r}} + p_0(T), \\
&= p_0(T) + (1 - p_0(T)) \left[1 - \frac{\frac{1}{1-p_0(T)/r}}{\frac{1}{f} + \frac{p_0(T)/r}{1-p_0(T)/r}} \right], \\
&= p_0(T) + (1 - p_0(T)) \frac{\frac{1}{f} + \frac{p_0(T)/r}{1-p_0(T)/r} - \frac{1}{1-p_0(T)/r}}{\frac{1}{f} + \frac{p_0(T)/r}{1-p_0(T)/r}}, \\
&= p_0(T) + (1 - p_0(T)) \frac{\frac{1}{f} - 1}{\frac{1}{f} + \frac{p_0(T)/r}{1-p_0(T)/r}}. \\
&= p_0(T) + (1 - p_0(T)) \frac{1-f}{1 - \frac{p_0(T)/r}{1-p_0(T)/r} f}, \\
&= p_0(T) + (1 - p_0(T)) \frac{(1 - p_0(T)/r)(1-f)}{1 - p_0(T)/r(1-f)}, \\
&= \mathcal{U}_T(1-f). \quad \square
\end{aligned} \tag{2.63}$$

We further claim that

$$\mathcal{U}^{(1)}_{t-s}(\mathcal{U}_s(1-f)) = \Psi(s,t)$$

where the r.h.s is Equation 3 in [81]. We start by using Equation 2.62 to rewrite the l.h.s

$$\mathcal{U}^{(1)}_{t-s}(z)\Big|_{z=\mathcal{U}_s(1-f)}$$

so that we can first take the derivative w.r.t. z rather than w.r.t. $1-f$. Indeed

$$\begin{aligned}\mathcal{U}^{(1)}_{t-s}(z) &= \frac{\partial}{\partial z}\left(1-\frac{\omega_{t-s}}{\frac{1}{1-z}+\frac{\omega_{t-s}-1}{1-r}}\right),\\ &= \omega_{t-s}\left[1+\frac{\omega_{t-s}-1}{1-r}(1-z)\right]^{-2}\end{aligned} \tag{2.64}$$

such that

$$\begin{aligned}\mathcal{U}^{(1)}_{t-s}(z)\Big|_{z=\mathcal{U}_s(1-f)} &= \omega_{t-s}\left[1+\frac{\omega_{t-s}-1}{1-r}(1-z)\right]^{-2}\Bigg|_{z=\mathcal{U}_s(1-f)},\\ &= \omega_{t-s}\left[1+\frac{\omega_s\frac{\omega_{t-s}-1}{1-r}}{\frac{1}{1-z}+\frac{\omega_s-1}{1-r}}\right]^{-2}\Bigg|_{z=1-f},\\ &= \omega_{t-s}\left[1+\frac{\omega_s\frac{\omega_{t-s}-1}{1-r}}{\frac{1}{f}+\frac{\omega_s-1}{1-r}}\right]^{-2},\\ &= \Psi(s,t). \quad \square\end{aligned} \tag{2.65}$$

This is $\Psi(s,t)$ for constant birth and death rates. This complete the proof of the equivalence between Equation 2.60 and Equation 1 of [81]. The infinitesimal factor δt^{L-1} in Equation 2.60 has no effect other than indicating that our likelihood is a probability while their likelihood is a probability density.

Numerical Determination of Subtree Probabilities

Provided we can quickly, numerically, and repeatedly solve Equation 2.27 for any (most) values of z in the complex plane then we can use the method of complex derivation based on Cauchy's integral formula Equation 2.6 to extract probabilities $\phi^{(k)}_f(t,s)$ from $\Phi_f(z,t,s)$ up to very high values of k[83]–[85]. This is in stark contrast with finite difference methods to estimate derivatives and which would allow us to extract at best $\phi^{(1)}_f(t,s)$, $\phi^{(2)}_f(t,s)$, and perhaps $\phi^{(3)}_f(t,s)$ if we are lucky. It also contrasts with linear algebra methods based on matrix exponentiation algorithms. Those methods can access probabilities up to $k\sim\mathcal{O}(1,000)$ with rather bad truncation errors—especially in the BDH model—and have very large memory footprint $\mathcal{O}(k^3)$. Numerical Cauchy integration gives us access to extremely small probabilities $\phi^{(k)}_f(t,s)$ at order $k\sim\mathcal{O}(100,000)$ without issue and within floating point accuracy. If one is patient enough, derivatives at order $k\sim\mathcal{O}(1,000,000)$ are easily accessible as well. As explained in [85] this is possible if we are mindful about issues of floating-point under- and overflow which are in turn tamed by carefully choosing the radius r of the contour of Equation 2.6. This method uses the simple and exponentially convergent trapezoidal

approximation to Equation 2.6[86]

$$\begin{aligned}\phi_f^{(k)}(t,s) &= \frac{1}{k!}\frac{\partial^k}{\partial y^k}\Phi_f(y,t,s)\Big|_{y=0}, \\ &= \frac{1}{\pi r^k}\int_0^{2\pi} e^{ik\theta}\Phi_f(re^{i\theta},t,s), \\ &= \frac{1}{nr^k}\sum_{j=0}^{n-1} e^{-2\pi ijk/n}\Phi_f(re^{2\pi ij/n},t,s).\end{aligned} \tag{2.66}$$

Equation 2.66 is nothing but the Fast Fourier Transform (FFT) of Φ for which there are many readily available fast and accurate algorithms. For a slice with the largest subtree of size $k_{\max}$ we set $n = n_{\max} = 2^{\lceil \log_2 k_{\max}+1\rceil}$. The FFT then gives us all probabilities from $k = 0$, which vanishes, to $n_{\max} - 1$ from which we use only probabilities $k = 1$ to $k = k_{\max}$. For the BDH model the hypergeometric function introduces a branch point at $|y| = 1$ and we can simply set the integration radius $r = 1 - 1/n_{\max}$. For the BD and BDI model we find the integration radius automatically using the optimization criterion of [85] which we seed with an initial guess obtained from the stable pole-finding algorithm of [87]. Alternatively for the BD model we can also simply use the exact probability Equation 2.18.

To give a taste of the power of the approach using complex integration of generating functions, Figure **??** shows the match between subtree probabilities $\phi_f^{(k)}(t,s)$ obtained using Cauchy integration and matrix exponentiation. The matrix exponentiation approach uses Equation 2.38 where the transition probabilities $p_{1\to k}(t)$ is determined using the exact numerical exponentiation of the truncated transition matrix $\boldsymbol{W}$ of the BDH process. The complex integration approach uses the FFT to calculate Equation 2.66 for the same process. Because of the $\mathcal{O}(k^3)$ scaling of the algorithm behind numerical matrix exponentiation, we truncate at $k = 1024$ which roughly corresponds to the point beyond which it becomes numerically impractical, and eventually impossible, to proceed this way. For the complex integration approach we push the algorithm all the way up to $k = 2^{16}$ without issue. Notice the appearance of dramatic truncation artifacts which appears in the tail of the matrix exponentiation approach and the lack thereof for the complex integration approach.

Empirical Tree Likelihood

Figure 2.5 shows the 16 subtree size distributions obtained from coarse-graining the timetree of all samples of animal proximal gut microbiomes combined. While this set or distributions represents only one biome it is qualitatively very typical of trees across the EMP dataset. A few features are of note; the approximate power-law for $k \geq 2$ with a common exponent as shown by the green visual guide, in this case -1.65, the deviation from this power-law exponent in the first slice, and the deviation from the power-law at $k = 1$ across most slices. Figure 2.6 similarly shows the empirical subtree size distribution of the coarse-grained timetree of the whole Earth microbiome. The same features are present except the approximate power-law exponent here is -1.2.

The largest subtree in Figure 2.5 appears in the first slice and has size $k_{\max} = 5,556$. In the second slice $k_{\max} = 1,199$, and in the third slice $k_{\max} = 1,144$. For the whole Earth microbiome shown in Figure 2.6 the second and seventh slices contain subtrees of size $k_{\max} \sim 50,000$ and for several other slices $k_{\max} \sim 5,000 - 10,000$. This means that the number of sample points n_{max} taken along the contour in Equation 2.66 would range between $8,192$ and $16,384$ (half the points are needed in reality because of the positivity of spectrum of $\Phi_f(y,t,s)$ and thus the possibility of using the Hermitian FFT (hFFT)). This

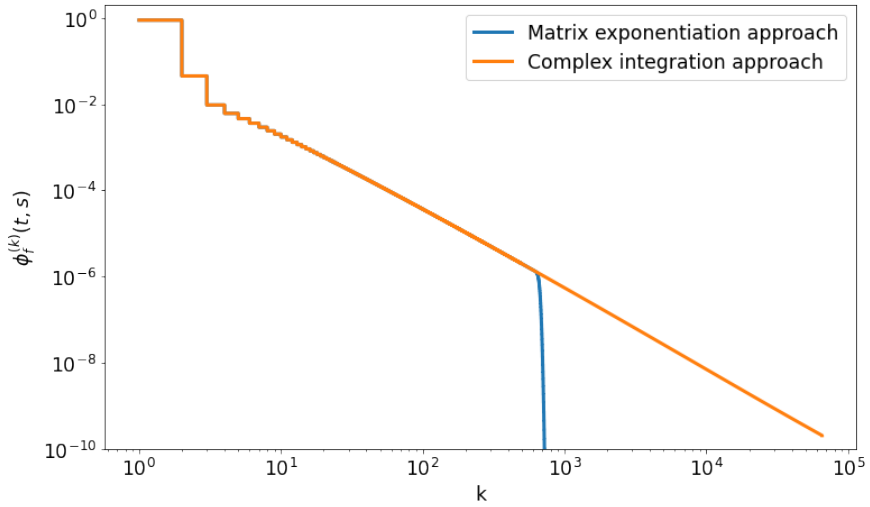


Figure 2.4: Comparison of the probabilistic approach using truncated matrix exponentiation and the generating function approach using complex integration. Parameters for the process are $\{f, b, d, \eta, \alpha, \beta\} = \{0.7, 1.0, 1.0, 1.0, 5.1, 1.1\}$ and times $t = 5/16$ and $s = 4/16$.

is no issue at all for the method presented in Subsection 2.2.5 except perhaps for speed. Indeed each evaluation of $\Phi_f(y, t, s)$ requires three evaluation of $\mathcal{U}$, namely $\mathcal{U}_t(z)$ and $\mathcal{U}_s(z)$ at $z = 1 - f$ and $\mathcal{U}_{t-s}(z)$ at $z = \mathcal{U}_s(1 - f) + y(1 - \mathcal{U}_s(1 - f))$. Combined and taking into account the use of the hFFT we must for some of the slices perform between $12,285$ and $24,573$ evaluations of the function $\mathcal{U}$ each of which requires the numerical integration of the differential equation Equation 2.26. Obviously this procedure can become slow for large timetrees.

Fortunately the power-law-like behavior already appears very early in the bulk of each subtree size distributions, i.e. for $k \geq 2$. This means that we can truncate the empirical distribution by imposing $k_{\max} \leq 256 - 4096$ as the maximum empirical subtree size that enters the computation of Equation 2.47. In other words neglecting outliers far in the tail induces a significant speed-up without sacrificing too much information from the empirical distribution. On the other hand this can sometimes lead to poor goodness of fit results.

On important point remains. Why is there an unmistakeable quantitative difference between the first slice and the others in the exponent of the tail of the empirical distribution as indicated by simple comparison with the guide. The first potential reason could be the fact that the original EMP phylogenies are obtained from OTU data. Using OTUs exposes us to potentially significant artifacts in the resolution of the arrangement of branching close to the present. In other words the quality of the data may be at fault. The second potential reason goes in the other direction. If we assume the OTU data to be valid, then the phylogenies will by design ignore their associated relative abundance. That is to say each OTU is represented as a single lineage even though it is surrounded by a cluster of closely related sequences, each of which could be valid sequences of microbes that have recently diverged away from the representative sequence. An OTU in effect coarse-grains its cloud of closely related sequences. The relative size of those clusters across OTUs is not reflected in the phylogeny. Adding back the relative sizes of those cluster as small polytomies sitting at the tip of each leaves may present a simple if approximative way to recover a more accurate empirical subtree size distribution in the first slice.

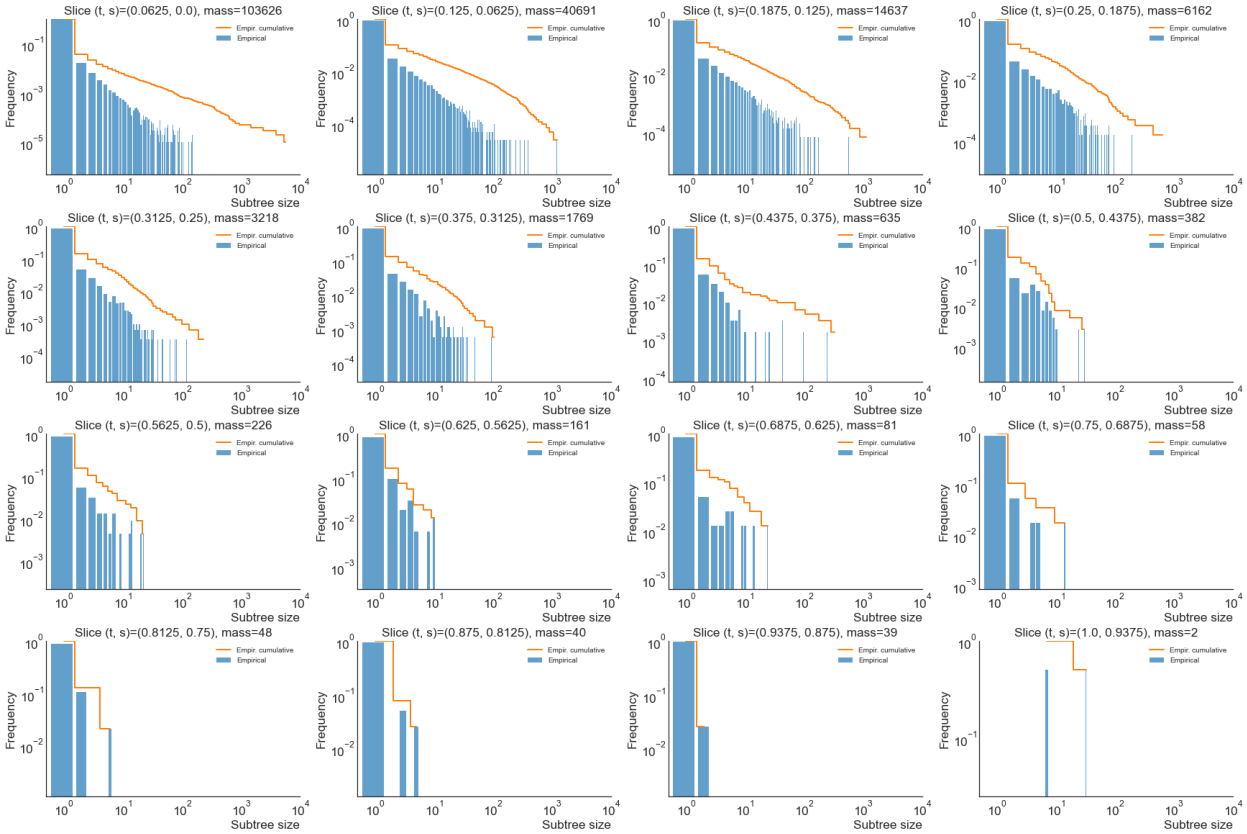


Figure 2.5: Shown in **blue** the empirical subtree size distributions of the animal proximal gut microbiome tree and in **orange** their associated cumulative distributions. The 16 subtree size distributions from the present to the past are shown from top-left to bottom-right. Notice how both the distributions and cumulative distributions suggest a heavy tail at least over some range of subtree sizes.

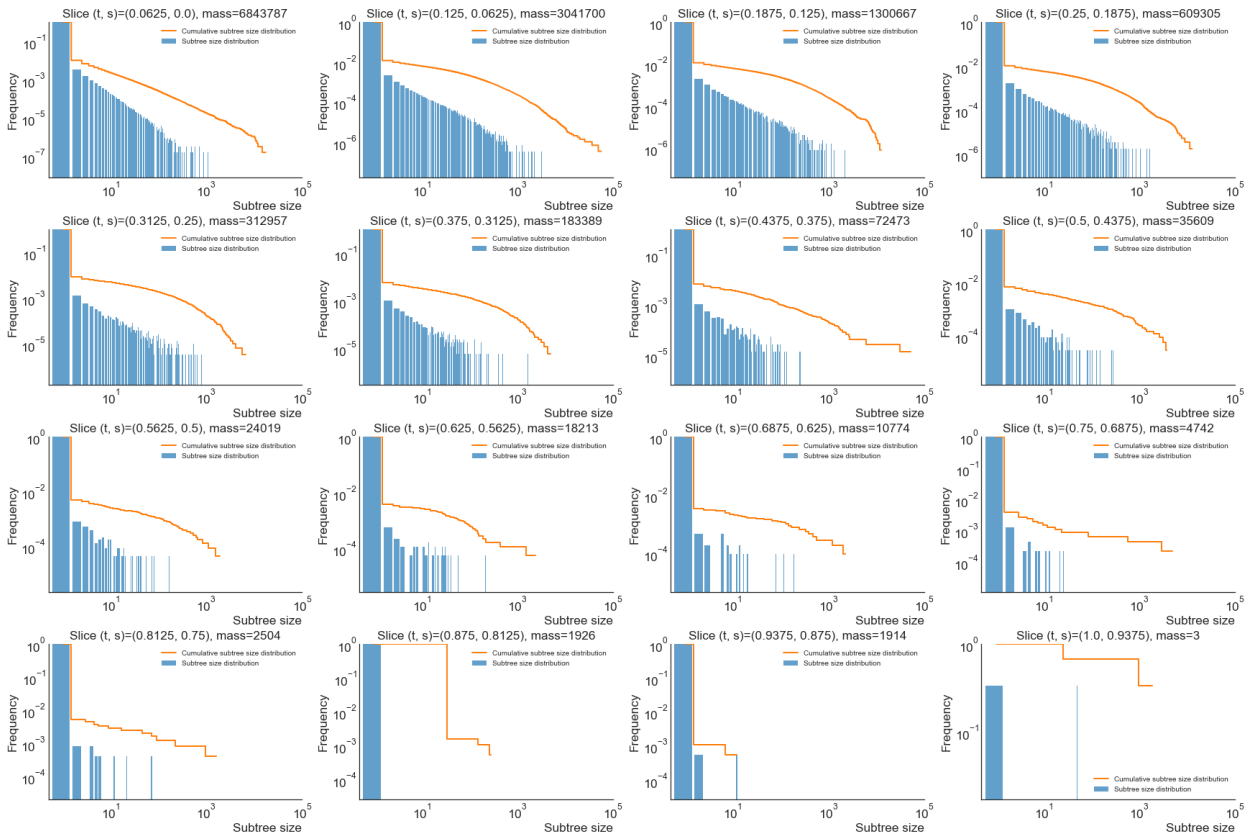


Figure 2.6: Shown in **blue** the empirical subtree size distributions of the whole Earth microbiome tree, and in **orange** their associated cumulative distributions. The 16 slice distributions from the present to the past are shown from top-left to bottom-right. Notice how both the distributions and cumulative distributions suggest a heavy tail at least over some range of subtree sizes.

Optimization and Parameter Identifiability

To find maximum likelihood estimates (MLE) for parameters $\{f^*, \theta^*\}$ we solve the optimization problem

$$f^*, \theta^* = \underset{f,\theta}{\operatorname{argmax}}\, \mathcal{L}[\mathcal{T}_K | f, \theta]. \tag{2.67}$$

Parameter sets for our models are given by

$$\begin{aligned} f, \theta \quad &\equiv bf, b-d, && \text{BD model} \\ &\equiv f, b, d, \rho, g, && \text{BDI model} \\ &\equiv f, b, d, \eta, \alpha, \beta, && \text{BDH model.} \end{aligned} \tag{2.68}$$

Notice that in the BD model b, d, and f are not completely independent and only two parameters are identifiable. The combination $bf, b-d$ is not unique and could be rewritten any other way but we chose this one because this is how it appears in the expression we will derive below. This has already been pointed out by [82]. There is a flat invariant direction in the likelihood along curves $bf = b'f'$ and $b-d = b'-d'$. In other words, if one finds MLEs f^*, b^*, and d^*, then one can choose a different $f^* \to f'$ and then sets $b^* \to b^* f^*/f'$ and $d^* \to b^*\,(f^*/f' - 1) - d^*$ without affecting the value of the ML. That is to say there are only two effective parameters rather than what appears to be three parameters in the BD model under incomplete lineage sampling. This flat direction may cause issues during optimization so one would simply set $f = 1$ to avoid them.

Proof of Non-Identifiability of the BD Model

All observables in the likelihood $L[\tau_K | f, b, d]$ are obtained from the OPSGF. Explicitly for the BD model

$$\Phi^{BD}(y, t, s) = \frac{bf(1 - e^{(b-d)s}) - (b-d)}{bf\left(1 - e^{(b-d)t} + y(e^{(b-d)t} - e^{(b-d)s})\right) - (b-d)}. \tag{2.69}$$

One can see by inspection only two combinations of parameters appear in , namely bf and $b-d$. There is thus a global unidentifiability in the BD model with incomplete lineage sampling.

The unconditional PSGF does not suffer the same problem because the extinction rate d would become independent, but one must remember they would be using it in a situation where subtrees going from 1 lineage to 0 lineages are observed, namely a situation where information about the fossil record is accessible.

Approximate Proof of Identifiability of the BDI and BDH Models

We do not have an explicit expression for the probability generating function of neither the BDI or BDH model and thus we cannot directly show the identifiability of their parameters. On the other hand we can expand Equation 2.9 up to some low order of ϵ with $t \to \epsilon t$ and$s \to \epsilon s$, inspect the various parameter combinations that emerge, and draw an approximate conclusion about identifiability or lack thereof. We do not show the derivation because of how unwieldy they become. Instead we use a computer algebra system and provide a short analysis without further details.

At first order of ϵ the OBSGF $\Phi^{BDI}(y, \epsilon t, \epsilon s)$ can be expressed in terms of parameters f, d, b, and ρg. At second order we also observe the same set of parameters plus g appearing by itself, which implies that unless improbable resummations happen we have identifiability for all 5 parameters of the model.

This result, provided it holds at higher orders of ϵ, indicates that in contrast to the BD model it is possible to infer the sampling fraction f in the BDI model. We claim that the same holds for the BDH model. We have two reason to believe this to be the case. First, the BDI model is nested within the BDH model and is recovered in the limit $\alpha+\beta \to \infty$ if the ratio $\alpha/(\alpha+\beta) = g$ is kept finite. Second, and perhaps obviously, the Gauss hypergeometric function ${}_2F_1(1, \alpha+1, \alpha+\beta+1, z)$ that appears in the generator of the BDH model is defined within $|z| < 1$ by a power series in z with coefficients $(\alpha+1)_n/(\alpha+\beta+1)_n$. Within those coefficients, for $n \geq 2$, α and β will always be independent and we have no reason to believe this would not transfer to Φ^{BDH}.

Numerical Proof of the Sloppiness and Potential Non-Identifiability of the BDH Model

Despite the argument made in the previous section, we find that numerical evidences point to the apparent and unfortunate non-identifiability of f and the sloppy, perhaps even flat direction[88] that appears to emerge between parameters f, b, d, and α.

Here's how we proceed. We generate a synthetic tree with 10,000 leaves poised at parameter values $\{f, b, d, \eta, \alpha, \beta\} = \{0.7, 2.0, 0.5, 1.2, 2.0, 1.4\}$. Then we proceed with the inference over b, d, η, α, and β while keeping $f = 0.7$ fixed. The resulting MLE for the parameters are $\{b^*, d^*, \eta^*, \alpha^*, \beta^*\} = \{2.0 \pm 0.1, 0.5 \pm 0.4, 0.9 \pm 0.2, 3.4 \pm 1.2, 1.6 \pm 0.2\}$ where the uncertainties are the 68% confidence intervals found along the the diagonal of the Fisher information, i.e. the diagonal of the inverse of the Hessian matrix, at the ML point. Those estimates are quite good and indicate that parameters can be inferred quite well when f is fixed.

Now let us run the inference again but this time including f. The optimization algorithm (Nelder-Mead) ends up drifting very slowing along f, b, d, and α with very minimal decrease in the negative log-likelihood. We nonetheless evaluate the Fisher information matrix again at a point along this very slow drift and find the approximate MLE $\{f^*, b^*, d^*, \eta^*, \alpha^*, \beta^*\} = \{0.4 \pm 0.6, 3.3 \pm 5.0, 2.1 \pm 5.7, 0.6 \pm 0.7, 5.1 \pm 6.6, 1.6 \pm 0.2\}$. The dramatic increase in the uncertainty along a direction combining f, b, d, α and to an extent η now indicates that the curvature of the negative log-likelihood along it is very low and therefore their identifiability questionable. It is even possible that some of these parameter uncertainties might eventually converge to infinity but for the fact that the optimization algorithm is incapable of finding the ridge along which the curvature vanishes simply because it is incapable of handling the case of singular models.

With this disappointing result in mind, we will resort to setting $f = 1$ during all model optimizations over empirical trees and accept the risk of missing the 'true", if it exists at all, MLE of f.

2.2.6 Simulating Synthetic Trees

To simulate synthetic trees we use a simple stochastic Gillespie algorithm[89], [90]. For the BDH model, three processes are possible; the birth process (B) with rate b per lineage, the death process (D) with rate d per lineage, and the heterogeneous innovation process (H) with rate η per lineage. The algorithm proceeds as follows:

1. Let T be the total depth of the tree, and n the current number of lineages. Let $N_{\text{cutoff}} > 0$ and $T_{\text{cutoff}} > 0$ be two stopping criteria. Start with $T = 0$ and a single lineage, the root, of length 0 and make it active,

2. Iterates:

(a) If $n \geq N_{\text{cutoff}}$ then sample leaves with intensity f and return the tree,

(b) Determine the total rate of events $\omega = (b + d + \eta)n$,

(c) Determine the waiting time until next event by drawing $\Delta t \sim \exp(\omega)$, or equivalently $\Delta t = -(\ln r)/\omega$ where $r \sim \text{Uniform}(0, 1)$,

(d) Add Δt to T and to the length of all n active lineages,

(e) If $T >= T_{\text{cutoff}}$ then sample leaves with intensity f and return the tree,

(f) Select a lineage l at random out of the n currently active lineages,

(g) Draw one of three processes B, D, or H with probabilities bn/ω, dn/ω, and $\eta n/\omega$, respectively,

   i. If B is drawn, then add two active child lineages l_1 and l_2 to lineage l and deactivate l,

   ii. If D is drawn, remove lineage l and contract l's parent if the parent is left with a single child, or return an empty tree if l is the root,

   iii. If H is drawn, draw $g \sim \text{Beta}(\alpha, \beta)$, then draw $k \sim \text{Geometric}(1 - g),\ k \geq 1$, if $k > 1$ then add k active child lineages $l_1, ..., l_k$ to l and deactivate l, and if $k = 1$ then simply leave l active,

Notice that it is possible that the tree returned by the algorithm may have more leaves than N_{cutoff} if the last process was the H process and k was drawn such that $n + k > N_{\text{cutoff}}$. Similarly for T if $T + \Delta t > T_{\text{cutoff}}$. This implies that we only obtain trees that satisfy the N_{cutoff} if the last process was a B process, and never one exactly satisfying $T = T_{\text{cutoff}}$. We nonetheless take the resulting tree to be sufficiently fine for all practical purposes.

2.3 Goodness of Fit

It is one thing to determine through model selection that the BDH model fits empirical coarse-grained trees better than, say, the BD or BDI models, but a better test must asked whether the BDH model stands on its own. In other words how likely it is that empirical coarse-grained trees are *typical* of coarse-grained trees generated by the BDH model poised at its ML parameter estimates. To do so we need a goodness-of-fit test (GOF) which captures the closeness between the set of empirical subtree size distributions and theoretical ML subtree size distributions generated by the BDH model.

2.3.1 Exact Goodness-of-Fit Test (eGOF)

We choose as our GOF the G-test[91]. The G-test is a generalization of Pearson's χ^2 test. Rather than using as a test statistic the square deviation between two distributions, it uses instead the test statistic

$$G = 2 \sum_i O_i \ln \frac{O_i}{E_i}, \tag{2.70}$$

where O_i is the expected observed (empirical) number of counts in category i and E_i is the expected (theoretical) number of counts. The test statistic G is directly related to the Kullback-Leibler divergence

D_{KL}. Indeed

$$\begin{aligned} G &= 2\sum_i O_i \ln \frac{O_i}{E_i}, \\ &= 2\left(\sum_i O_i\right) \sum_i P_i^{\text{empirical}} \ln \frac{O_i \sum_i E_i}{E_i \sum_i O_i}, \\ &= 2N \sum_i P_i^{\text{empirical}} \ln \frac{P_i^{\text{empirical}}}{P_i^{\text{theoretical}}}, \\ &= 2N \sum_i D_{\mathrm{KL}}\left(P_i^{\text{empirical}} \| P_i^{\text{theoretical}}\right), \end{aligned} \tag{2.71}$$

where $N = \sum_i O_i = \sum_i E_i$ is the total number of observations and all sums over i run over categories such that $O_i > 0$. Because of this relationship we call this test 'exact' in the sense that it captures a true information theoretic divergence between empirical and theoretical trees. Adapted to coarse-grained trees,

$$G_{\tau_K} = 2\sum_{\sigma=1}^{K} \sum_{i, k_i^\sigma > 0} O_i^\sigma \left(\ln P_i^{\sigma,\text{empirical}} - \ln \phi_f^{(k_i^\sigma)}(t_i^\sigma, s_i^\sigma)\right), \tag{2.72}$$

where σ runs over slices and $P_i^{\sigma,\text{empirical}} = O_i^\sigma / N^\sigma$.

2.3.2 Null Hypothesis for the eGOF

Usually for a goodness of fit test on simple distributions there is a known null distribution of the test statistic given for example in terms of a χ^2 distribution with a given number of degrees of freedom. Unfortunately we do know how to count degrees of freedoms for the distribution of G_{τ_K}. Intuitively the number of degrees of freedom might be taken to be the number of categories i with non-zero counts $O_i > 0$ minus the number of constraints, namely the total depth $T_{\mathcal{T}_K}^{\text{theoretical}} = T_{\mathcal{T}_K}^{\text{empirical}}$, that the sum of the size of subtrees in the first slice $N^{\text{theoretical}} = N^{\text{empirical}}$ is equal, the normalization of the subtree size distribution for each slice, and that at each interface between two slices the sum of the size of all subtrees must be equal to the number of subtrees in the previous slice. Unfortunately synthetic experiments we describe next do not agree with this intuition. We must therefore resort to generating this null distribution using simulations, more specifically using a Metropolis-Hastings MCMC algorithm.

Finding the null distribution of $G_{\mathcal{T}_K}^{\text{theoretical}}$ amounts to first finding a distribution of typical subtree size distributions satisfying all above constraints and calculating for each such typical subtree size distribution i its statistic $G_{\mathcal{T}_K}^i$. Then we compare $G_{\mathcal{T}_K}^{\text{empirical}}$ to the null distribution of $G_{\mathcal{T}_K}^i$ and determine how far in the tail $G_{\mathcal{T}_K}^{\text{empirical}}$ is. In an ideal world we would generate the set of all coarse-grained trees and calculate the test statistic for each of them, but unfortunately this task is computationally infeasible given the combinatorially explosive number of such coarse-grained trees. This is why we resort to approximating the null distribution using a MCMC algorithm.

The algorithm proceeds as follows:

1. Let the first coarse-grained tree in the chain be $\mathcal{T}_K$. Call this tree $\mathcal{T}$ for short. Calculate and return $G_{\mathcal{T}}$,

2. Iterate N_{iter} times:

(a) Propose a move $\mathcal{T}' = \mathcal{M}[\mathcal{T}]$,

(b) Accept move with probability $\alpha = \min\left[1,\ \frac{\Pr[\mathcal{T}']}{\Pr[\mathcal{T}]}\frac{g[\mathcal{T}|\mathcal{T}']}{g[\mathcal{T}'|\mathcal{T}]}\right]$,

(c) If the move is accepted set $\mathcal{T} = \mathcal{T}'$,

(d) Return $G_{\mathcal{T}}$.

We choose N_{iter} to be roughly 20 times the total number of subtrees $\sum_{\sigma=2}^{K}\sum_i O_i^\sigma$ such that each subtree will on average be resampled 20 times. Moves are proposed as follows:

1. Select a slice $\sigma = 2, ..., K$ at random where each slice is weighted by the number of subtrees it contains,
2. Select a subtree at random in slice σ,
3. For a selected subtree of size k, select at random and without replacement k subtrees $k_1, k_2, ...k_k$ in slice $\sigma - 1$. This is the current partition $\pi = \{k_1, k_2, ..., k_k\}$,
4. Let K $= \sum_{i=1}^{k} k_i$. Draw uniformly at random a partition π' of K using Algorithm 5 found in [92] and let $k' = |\pi'|$ the number of parts in π',
5. Propose $\mathcal{T}'$ where π is replaced with $\pi' = \{k_1', k_2', ...\}$.

With this proposal scheme, the Metropolis ratio

$$\frac{\Pr[\mathcal{T}']}{\Pr[\mathcal{T}]} = \frac{\phi_f^{(k')}(t_\sigma, s_\sigma)}{\phi_f^{(k')}(t_\sigma, s_\sigma)}\frac{\prod_i \phi_f^{(k_i')}(t_{\sigma-1}, s_{\sigma-1})}{\prod_i \phi_f^{(k_i)}(t_{\sigma-1}, s_{\sigma-1})}, \tag{2.73}$$

and Equation 2.44 gives us the Hastings ratio

$$\frac{g[\mathcal{T}|\mathcal{T}']}{g[\mathcal{T}'|\mathcal{T}]} = \frac{|\operatorname{Comp}(\pi)|^{-1}}{|\operatorname{Comp}(\pi')|^{-1}} = \frac{|\pi'|!}{\prod_{\lambda=1}^{K} m_{\pi'}(\lambda)!}\frac{\prod_{\lambda=1}^{K} m_\pi(\lambda)!}{|\pi|!}. \tag{2.74}$$

This completes the algorithm. It generates a chain of statistic values $G_{\mathcal{T}_1}, G_{\mathcal{T}_2}, ...G_{\mathcal{T}_{N_{\text{iter}}}}$. The stationary distribution of these values converges to the null distribution for G_{BDH}, namely the distribution of typical values of G of typical trees of the BDH model. The p-value for the test,

$$p = \frac{1}{N_{\text{iter}} - N_{\text{burnin}}}\sum_{i=N_{\text{burnin}}}^{N_{\text{iter}}} \mathbb{1}[G^{\mathcal{T}_i} < G_{\mathcal{T}}], \tag{2.75}$$

gives the probability that a realization of subtree size distributions generated by the BDH model diverges less from the theoretical subtree size distributions than the empirical subtree size distributions. For example, if $p = 0.95$ then it means that 95% of all coarse-grained trees generated by the BDH model are more typical than the empirical tree, or alternatively that 5% of trees generated by the BDH model are less typical than the empirical tree. In other words, the null hypothesis is that the model faithfully generates the empirical tree and its subtree size distributions. Therefore the test fails when the p-value is more than 0.95.

To verify whether this is a reasonable eGOF, we generate 1,000 BDH trees $\mathcal{T}_1, ...\mathcal{T}_{1000}$ of size approximately 10,000 with parameter values $\{b, d, \eta, \alpha, \beta\} = \{2.0, 0.7, 1.2, 3.0, 1.3\}$. For each tree $\mathcal{T}_j$ we then generate a chain of values $G^j_{\mathcal{T}_i}$ with $N_{\text{iter}} = 10^6$ and $N_{\text{burnin}} = 3 \times 10^5$ and calculate p_j. Figure 2.8 shows the distribution of those 1,000 p_j's. The distribution is almost uniform between 0 and 1 with a median of approximately 0.5 and therefore the eGOF is a reasonable one. Even though the mass is evenly split between $p < 0.5$ and $p > 0.5$,

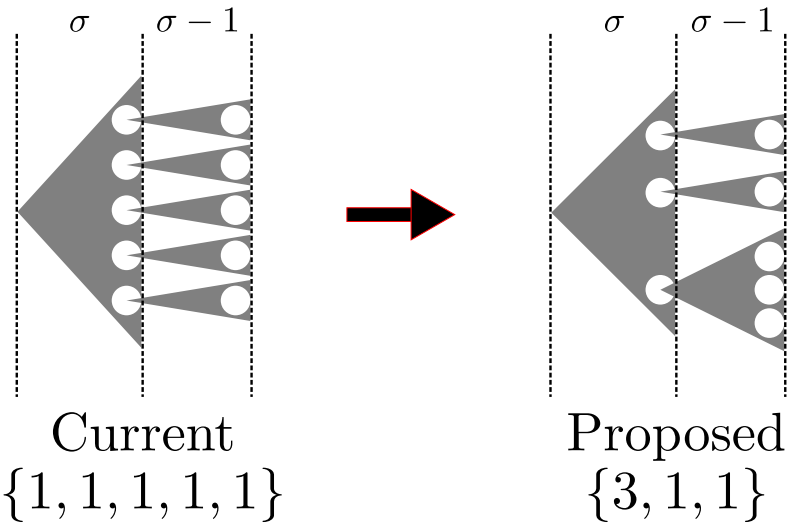


Figure 2.7: On the left we have a current partition and on the right a proposed partition. A subtree in slice σ represent the number of parts in a partition, and subtrees in slice $\sigma - 1$ represent the size of those parts. Proposing a move in the set of coarse-grained tree is equivalent to drawing at random a partition of the same number, in this case a partition of 5. In other word we move in the set of coarse-grained tree by reshuffling small parts of the trees. The Hastings ratio Equation 2.74 corrects for the number of ways to order partitions on the left and on the right.

the slight inflation of low p-values indicates an undiagnosed biased in the null distribution of trees we believe comes from the approximate truncation described in Section 2.2.6.

2.4 Data Preparation

2.4.1 Dataset

As the main application of our inference framework we are using the dataset provided by The Earth Microbiome Project[21], henceforth denoted EMP. More precisely we are using a subset of the dataset, namely the open-reference OTUs[35] placed on the greengenes 13.8[93], [94] core reference tree using SEPP[95]. Our method relies on the existence of a timetree where all leaves (representing OTUs) are at equal distance to the root and moreover in which the root has been identified. Fortunately the greengenes 13.8 reference tree is already rooted with Archaea as the outgroup. Both the open-reference placement tree given by the EMP and the greengenes 13.8 reference tree have branches lengths in units of substitutions per base pair and therefore need to be calibrated, and so we must recreate a suitable reference tree with calibration points out of which we can later construct a timetree.

Calibrating the Reference Tree

More details about this section can be found in Appendix A. To calibrate the greengenes reference tree we use the dataset behind the TimeTree of Life[60], [96], specifically for prokaryotes[97]. The prokaryote Timetree provides 101 calibration points at each node of a family-level tree topology. This topology together with the

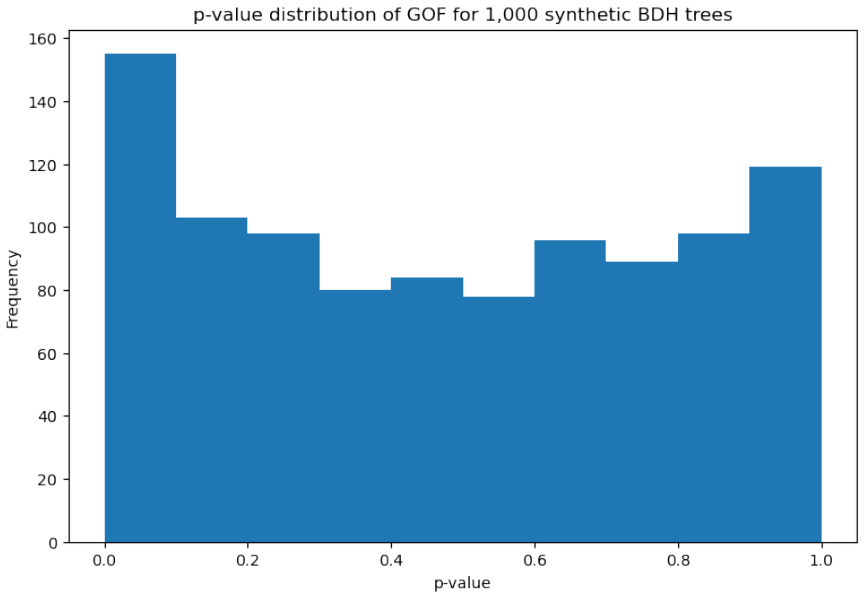


Figure 2.8: Distribution of p-values of the eGOF under the BDH model.

age of the calibration points can be found in Appendix B. Attached to those calibration points are polytomies of 16S sequences of representatives for each of those families. From the sequence data that was transmitted to us by Julie Marin we were able to keep 89 of the 101 calibration points after pruning families with no representatives. Conversely, many sequences did not find a taxonomic match in the family level topology and were dropped. In the end we have a tree with 6294 representative family level sequences decorating a family level topology with 89 calibration points.

Next we aligned and placed those 6294 representative timetree sequences into the greengenes 13.8 reference tree using SEPP. After this step, we found many conflictual placements, i.e. placements for which the taxonomic information between greengenes and timetree did not agree down to the level of families. To remediate this problem we built a constraint alignment for FastTree[98], [99] using all representative timetree sequences and their family level topology. Indeed each internal branch of the family level topology induces a split with groups of sequences left and right of that split. In other words removing an internal branch in an unrooted tree breaks it into two mutually exclusive subtree each containing their respective sequences. The constraint alignment is a tabulated representation of which sequences should indeed be constrained to fall left or right of given splits. FastTree expects that rows of the constraint alignment enumerate sequences and columns enumerate expected constrained splits. We then ran FastTree using this constraint alignment as input for a constrained topology search over the conflictual placement tree containing both greengenes and timetree sequences obtained in the previous step. Once FastTree terminated this search, we checked if there remained unsatisfied constraints. When that was the case, we used the output from FastTree and fed it back as its new input together with the same constraint alignment. We did this a few times until we did not observe any further decrease in the number of unsatisfied constraints. In our case this number stabilized at around 60 unsatisfied constraints. With this intermediate output tree we found which sequences are causing conflicts with the constraint alignment by intersecting the set of all sequences causing the largest mismatch (i.e. largest number of left-right crosses) between the observed splits and the expected splits in the constraint alignment. We identified 20 problematic sequences (out of 6294) as the worst minimal set

of sequences causing those remaining constraint violations. We pruned those problematic sequences from the intermediate tree and ran FastTree once more with the constraint alignment. Then we do a quick check with FastTree found that there were no more constraint violation. This is all to say that we restructured the greengenes reference tree by forcing it to adopt the family level topology provided by the prokaryote timetree and its representative sequences.

EMP Timetree

More details about this section can be found in Appendix A. Next we go back to the EMP dataset and using SEPP we place all 8,023,841 open-reference OTUs that were not part of the closed greengenes 13.8 reference set back onto the improved and calibrated greengenes+timetree reference tree. We then use MPL/PATHd8[63], [64] to ultrametricize the final EMP + greengenes + timetree placement tree. MPL/PATHd8 are the only methods we are aware of that can handle trees of size above $\mathcal{O}(10,000)$ OTUs. When correctly implemented they scale linearly with the number of leaves present in the tree. We want to mention that at first we attempted to improve slightly over the original authors' implementation of PATHd8 by introducing the possibility of weighting each path during calculation of the mean-path length (MPL). We refer to this modification we weighted MPL (wMPL). We used the relative abundance data provided by EMP as weights. This has the effect of taking into account the 'invisible' polytomies attached below each OTU, mimicking the unresolved cluster of sequences contained within an OTU and considering them in the calculation of mean-path lengths. This step is facultative and debatable, and ultimately we abandoned it because the goodness of fit tests from inferences ended up performing worst. In the end, and because of the calibration point at the LUCA, the EMP tree has total depth $T = 4.2$ Byr, which as mentioned in the beginning of Section 2.2 we rescale to 1.

2.5 Summary of Methods

Let us pause here and summaries the sequence of events that happen between inputing a phylogenetic tree from the EMP dataset and the eGOF:

1. Calibration of the phylogenetic tree from the EMP dataset into a timetree,
   (a) Place the sequences from the Timetree project onto the phylogenetic tree of the EMP dataset using SEPP,
   (b) Refine the reconstruction of the augmented EMP tree using FastTree using the known family- and higher level constraints given by the Timetree project sequences,
   (c) Root the resulting tree using Archaea as outgroup,
   (d) Ultrametricize the resulting tree using PATHd8 with the dated family- and higher level calibration points given by the Timetree project
2. Extract a tree for every level of the EMP ontologies out of the EMP timetree,
3. Coarse-grain each tree into distributions of OPS using the smallest number of slices which gives as many or more OPS as there are internal nodes,
4. Optimize the BD, BDI, and BDH models and find the MLE of parameters b, d, η, α, and β,
   (a) The log-likelihood Equation 2.46 is given by the sum of the logarithm of the OPS probabilities $\phi_f^{(k)}(t,s)$,

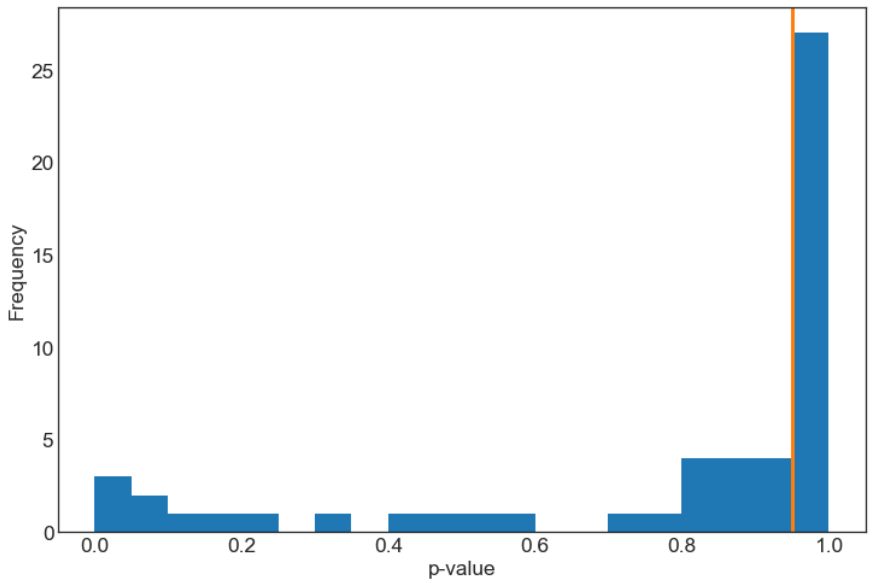


Figure 2.9: Distribution of eGOF p-values appearing in Table 2.1. Note that the BDH model passes the eGOF for 41% of all trees. The orange line indicates the threshold above which the eGOF fails with 95% confidence.

(b) The OPS probabilities are the Taylor coefficients of the OPSGF $\Phi_f(y,t,s)$ given by Equation 2.36 which we obtain by taking its FFT,

(c) To evaluate the FFT we need the value of $\Phi_f(y,t,s)$ at a large number of set points y in the complex plain, and by extension of the generating function $\mathcal{U}_t(z)$ of the stochastic process at a large number of points z,

(d) To evaluate $\mathcal{U}_t(z)$ at a single point z we solve Equation 2.27 numerically,

5. Do model comparison between BD, BDI, and BDH using the LRT,

6. Perform the eGOF on the best model.

2.6 Results

2.6.1 Inference on EMP

We performed ML estimation using the BDH model and report parameter estimates and eGOF for all trees up to level 4 of the EMP environmental ontologies (ENVO). The for the parameter estimates and eGOF p-values are shown in Table 2.1. We also report the results up to level 3 of the EMP ontologies (EMPO) and results are shown in Table 2.2. Of interest is that the BDH model passes the eGOF for roughly 50% of the ENVO trees, and 41% of the EMPO trees. Passing the eGOF suggests that those trees are typical of trees generated by the BDH model. The distribution of eGOF p-values is shown in Figure 2.9. Figure 2.10 shows the distribution of the ratio of MLE for the birth rate b over the innovation rate η appearing in Table 2.1. Most of the weight of the distribution lie left of the value $b^*/\eta^* = 0.1$ which indicates that slow gradual speciation occurs an order of magnitude less often than fast bursty speciation. In fact it is less than one of order of magnitude less in 85% of the cases. Figure 2.11 shows the distribution of the MLE for the death,

ENVO level 1	ENVO level 2	ENVO level 3	ENVO level 4	b^*	d^*	η^*	α^*	β^*	eGOF
aquatic				0.01	0.00	0.10	24.8	1.00	1.0
	freshwater			0.01	0.0	0.16	24.3	1.20	1.0
		lake		0.01	0.00	0.25	19.8	1.31	0.92*
			large lake	0.04	0.02	0.99	7.73	1.68	0.84*
			small lake	0.01	0.00	0.27	19.4	1.34	0.81*
		river		0.02	0.00	0.43	12.6	1.39	0.31*
			large river	0.03	0.0	0.83	7.77	1.46	1.0
			small river	0.04	0.0	0.61	10.8	1.52	1.0
		unspecified		0.03	1.88	0.26	17.5	1.23	0.86*
	marine			0.01	0.01	0.17	18.2	1.1	0.86*
		benthic		0.03	0.45	0.37	16.4	1.39	0.89*
			reef	0.00	5.03	5.6	1.77	1.9	0.94*
			unspecified	0.03	0.41	0.375	16.0	1.38	1.0
		estuarine		0.14	0.00	4.14	3.03	2.1	0.02*
		marginal sea		0.02	0.00	3.51	2.95	1.87	0.86*
		pelagic		0.03	5.38	0.92	10.6	1.72	0.0*
		unspecified		0.02	0.00	0.22	14.0	1.12	0.08*
	unspecified			0.02	0.00	0.84	4.58	1.35	1.0
terrestrial				0.01	0.00	0.10	23.9	0.98	1.0
	anthropogenic			0.0	0.0	0.14	14.8	0.945	0.97
		cropland		0.0134	0.0	0.292	23.1	1.46	0.82*
		dense settlement		0.01	0.01	0.225	12.0	1.06	1.0
			urban	0.0	0.0	0.251	8.02	0.96	1.0
			unspecified	0.05	0.00	1.89	4.08	1.68	1.0
		rangeland		0.05	0.0	0.75	6.19	1.37	1.0
		village		0.04	0.00	2.27	3.07	1.59	1.0
		unspecified		0.00	7.80	1.82	8.89	2.13	0.01*
	desert			0.08	0.00	1.42	5.65	1.58	1.0
		polar		0.00	3.14	2.7	3.2	1.64	0.98
		unspecified		0.09	0.00	1.69	5.24	1.65	0.9*
	forest			0.02	0.00	0.38	8.75	1.18	1.0
		broadleaf		0.02	0.0	0.52	5.6	1.19	1.0
		coniferous		0.19	2.78	1.4	7.56	1.78	1.0
			unspecified	0.03	1.74	2.37	4.25	1.73	0.21*
			temperate	0.00	3.23	2.39	4.48	1.78	0.47*
			tropical	0.176	7.98	1.79	7.74	1.95	0.09*
		mixed		0.06	0.00	1.4	4.2	1.52	1.0
			temperate	0.12	3.26	1.98	4.84	1.66	0.12*
			unspecified	0.04	0.00	0.76	4.59	1.44	0.74*
		unspecified		0.05	0.03	0.757	9.64	1.54	0.58*
	grassland			0.04	0.06	0.73	9.56	1.5	0.17*
		montane		0.11	2.22	0.96	8.08	1.51	1.0
		temperate		0.16	1.08	1.7	5.41	1.69	1.0
		tropical		0.18	2.04	2.87	4.91	2.07	0.41*
		unspecified		0.06	0.00	1.16	6.59	1.57	0.97
	mangrove			0.04	0.04	2.92	5.76	2.89	0.75*
	shrubland			0.03	0.00	0.46	7.91	1.24	1.0
		montane		0.18	4.57	1.97	6.43	1.88	0.05*
		subtropical		0.05	0.01	0.80	3.67	1.24	1.0
			mediterranean	0.05	0.00	0.79	4.07	1.29	1.0
		tropical		0.11	0.00	2.23	5.62	1.96	0.95*
		unspecified		0.02	0.00	0.61	6.43	1.26	1.0
	tundra			0.04	1.19	0.41	15.5	1.42	0.82*
	woodland	subtropical	mediterranean	0.0	1.53	0.31	6.62	1.16	1.0

Table 2.1: Parameter estimates and p-value of the eGOF for the BDH model for biomes up to the fourth level of ENVO. The BDH model passes the GOF for about 50% of trees.

EMPO level 1	EMPO level 2	EMPO level 3	b^*	d^*	η^*	α^*	β^*	eGOF
Control			0.05	0.00	0.85	3.45	1.26	1.0
	Negative	Sterile water blank	0.02	0.00	0.53	4.51	1.23	0.87*
	Positive	Mock community	0.00	0.00	9.87	1.03	1.84	0.83*
Free-living			0.01	0.00	0.09	13.0	0.54	1.0
	Non-saline		0.01	0.00	0.10	12.3	0.58	0.98
		Aerosol non-saline	0.04	0.00	0.35	5.17	1.16	1.0
		Sediment non-saline	0.02	0.00	0.37	20.2	1.6	1.0
		Soil non-saline	0.01	0.00	0.15	12.5	0.68	0.58*
		Surface non-saline	0.02	1.1	0.55	10.5	1.35	0.73*
		Water non-saline	0.02	1.5	0.15	21.9	1.04	0.90*
	Saline		0.01	0.0	0.24	12.9	0.99	0.68*
		Hypersaline saline	0.00	0.00	2.52	6.8	2.96	1.0
		Sediment saline	0.02	0.00	0.28	14.9	1.15	0.95*
		Surface saline	0.04	1.06	0.92	8.78	1.62	0.56*
		Water saline	0.10	2.46	0.58	13.0	1.56	0.88*
Host-associated			0.01	0.0	0.14	10.1	0.70	1.0
	Animal		0.01	0.0	0.21	6.22	0.87	1.0
		Animal corpus	0.22	3.59	2.4	3.68	1.69	0.89*
		Animal distal gut	0.02	0.00	0.35	5.31	0.98	1.0
		Animal proximal gut	0.29	3.33	1.72	5.61	1.62	1.0
		Animal secretion	0.03	0.00	0.78	4.29	1.26	1.0
		Animal surface	0.02	0.00	0.30	5.44	1.03	1.0
	Plant		0.02	0.60	0.28	21.8	1.31	1.0
		Plant corpus	0.04	0.00	3.53	2.99	1.93	0.96
		Plant rhizosphere	0.029	0.0	0.50	14.5	1.49	0.64*
		Plant surface	0.15	6.38	0.14	44.4	1.31	0.83*

Table 2.2: Parameter estimates and p-value of the eGOF for the BDH model for biomes up to the fourth level of EMPO. The BDH model passes the GOF for about 46% of trees.

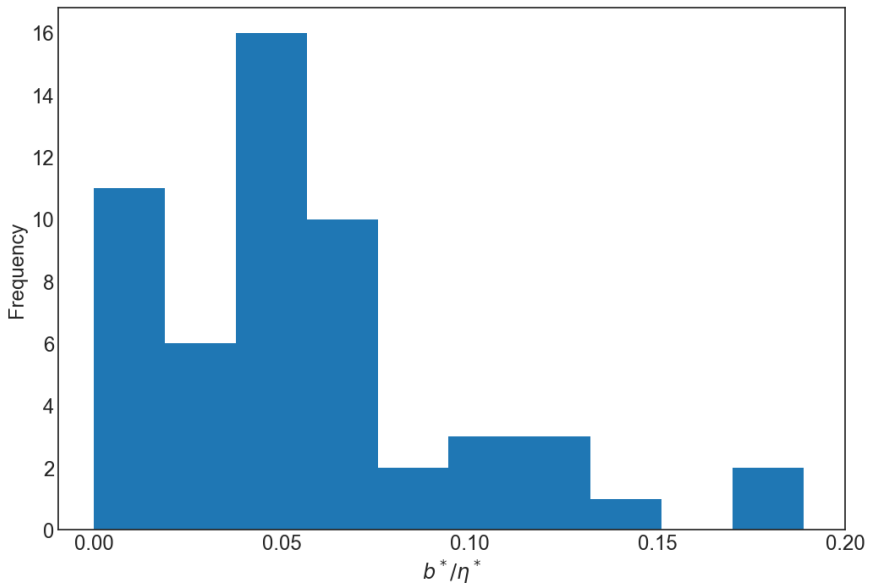


Figure 2.10: Distribution of ML estimates of the ratio b/η appearing in Table 2.1 between the death rate and the innovation rate for ENVO.

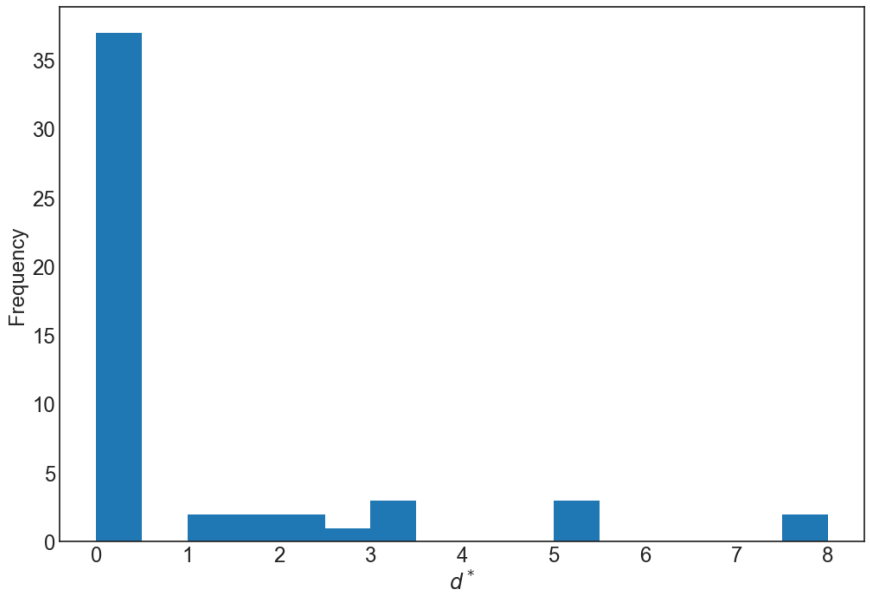


Figure 2.11: Distribution of ML estimates of the death rate d.

	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.90}$	$Q_{0.95}$	$Q_{0.99}$
mode ($\alpha = 4.6$, $\beta = 1.6$)	4	8	17	28	83
medians ($\alpha = 7.3$, $\beta = 1.5$)	6	13	29	49	156

Table 2.3: Quantiles of the beta-geometric distribution at the mode and medians of the MLE distribution appearing in Table 2.3

or extinction rate d appearing in Table 2.1. While the distribution shows a wide range of values for the estimates, a small majority of trees, about 57% of them, exhibit a death rate that is effectively 0 (less than 10^{-3}). Looking at Figure 2.12 which shows the MLE for d vs the eGOF p-value we see that there is a strong concentration of fits failing the eGOF when d^* is close to 0. We believe this to be partially artifactual and a result of bad fits and possibly related to the the phenomena of 0 inflated estimates for the extinction rate when inferring using extant phylogenies[100]. Finally Figure 2.15 shows the distribution of ML estimates for the shape parameter β vs α. The orange lines intersect in the middle of the modal bin of the distribution at $\{\alpha, \beta\} \approx \{4.6, 1.6\}$. The purple lines indicate the marginal medians of α^* and β^* where $\text{median}_{\alpha^*} = 7.3$, and $\text{median}_{\beta^*} = 1.5$. The 25% and 75% percentiles are $25\%_{\beta^*} = 1.3$, and $75\%_{\beta^*} = 1.7$ Parameters α and β control the shape of the burst size distribution of fast speciation events where a single lineage quickly speciate into k lineages, i.e. $A \to kA$ lineages. Their numerical value is hard to interpret as is. To give an idea of their meaning, let us consider the quantiles of the distribution they represent, namely the quantiles of the beta-geometric distribution. From Equation 2.22 we can write the quantile Q_p implicitly as

$$\sum_{k=1}^{Q_p} \frac{B(\alpha + k - 1, \beta + 1)}{B(\alpha, \beta)} = p. \tag{2.76}$$

Table 2.3 shows the quantiles of the beta-geometric distribution with α and β poised at the mode of the MLE distribution $\alpha = 3.7$ and $\beta = 1.4$ and the marginal medians $\alpha = 6.0$ and $\beta = 1.5$. They indicate that in 1% of fast bursty speciation events lineages can quickly diversify into around 100 or more new lineages. Now for

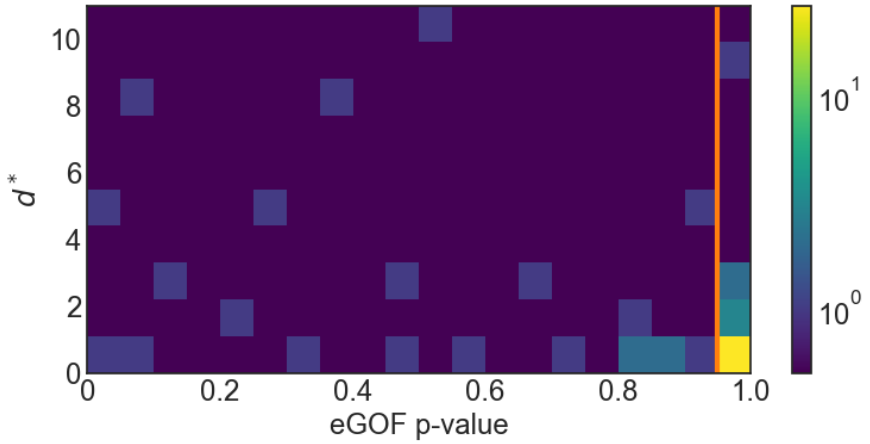


Figure 2.12: ML estimates for the death rate d as a function of the eGOF p-value. The **orange** line shows the 0.95 threshold.

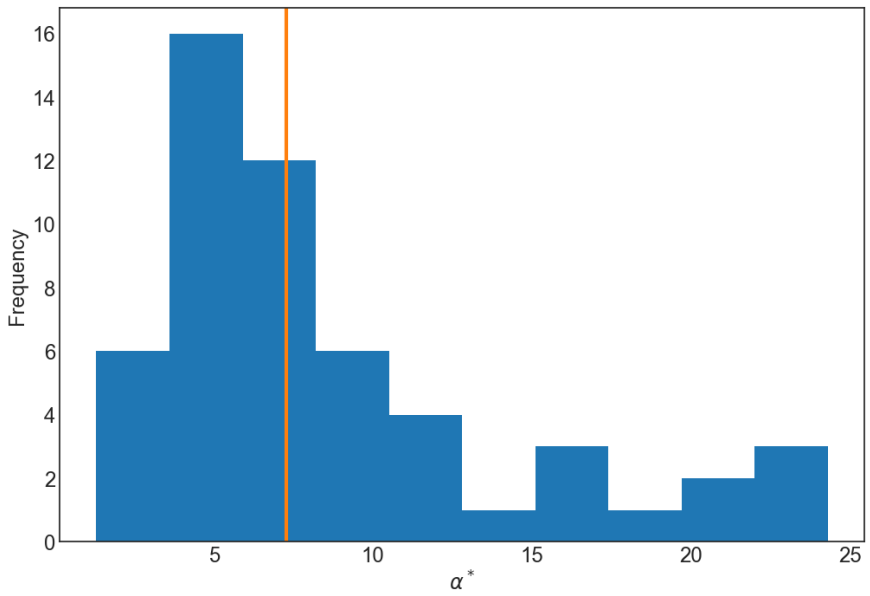


Figure 2.13: Distribution of ML estimates of the shape parameter α of the burst size distribution. The **orange** line is poised a $\text{median}_{\alpha^*} = 7.3$

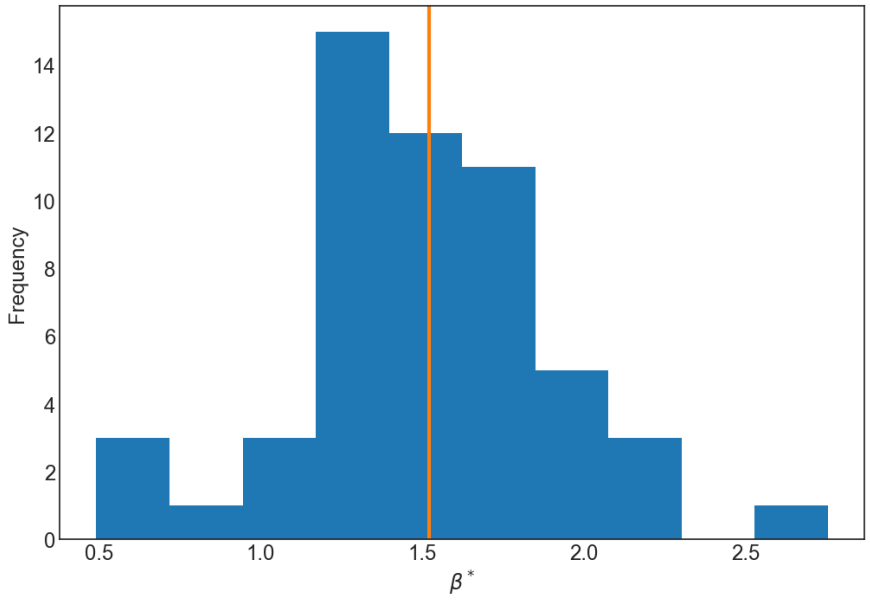


Figure 2.14: Distribution of ML estimates of the shape parameter β of the burst size distribution. The *orange* line is poised at $\text{median}_{\beta^*} = 1.5$

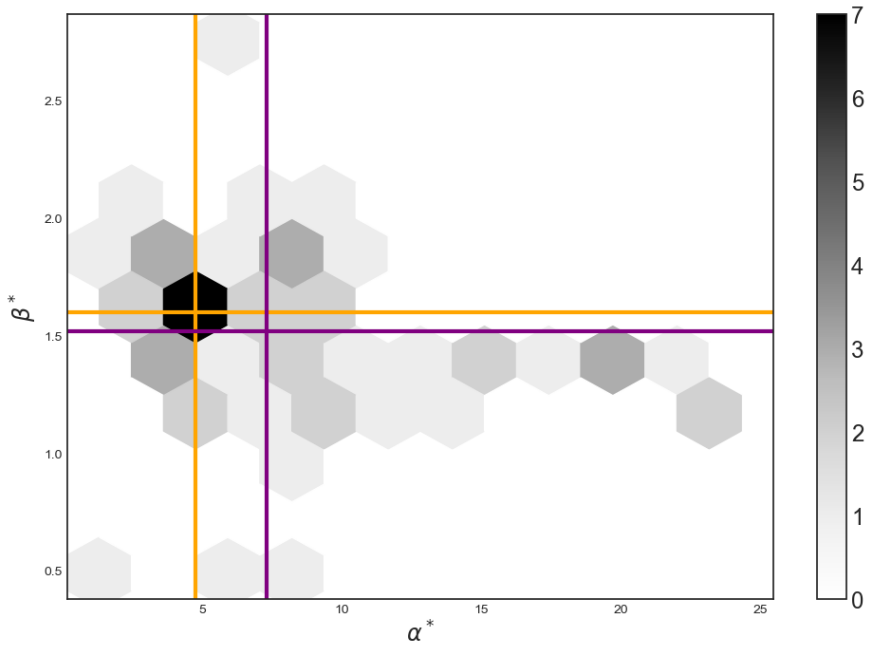


Figure 2.15: Distribution of ML estimates of the parameter β vs α appearing in Table 2.1. The orange lines intersect in the middle of the modal bin of the distribution at $\{\alpha, \beta\} \approx \{4.6, 1.6\}$. The purple lines indicate the marginal medians of α^* and β^* where $\text{median}_{\alpha^*} = 7.3$, and $\text{median}_{\beta^*} = 1.5$.

	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.90}$	$Q_{0.95}$	$Q_{0.99}$
aquatic ($\alpha = 7.2$, $\beta = 0.5$)	21	105	689	2775	69538
terrestrial ($\alpha = 6.9$, $\beta = 0.5$)	21	106	724	3000	80243

Table 2.4: Quantiles of the beta-geometric distribution at the mode and medians of the MLE distribution at ENVO level 1

something a bit more dramatic, let us look at the quantiles of the beta-geometric poised at the ML estimates of the parameters of the BDH model for the two trees at ENVO level 1, namely the aquatic and terrestrial trees. Those quantiles are reported in Table 2.4. We believe it is highly likely that those values, especially for the 95% and 99% quantiles, are spurious and do not reflect empirical reality. Even though subtrees of this size exist in a few of the subtree size distribution, meaning that bursts of this size within those subtrees are statistically allowed to occur, the frequency of those subtrees across those subtree size distributions is far from 1 in 20 and 1 in 100. This discrepancy is likely a manifestation of the fact that all models are wrong. Indeed the dramatic quantiles represent the weight of the ML distribution far in the tail. Yet we do not observe any subtree, and therefore no burst, that far in the empirical tail because it is likely that there is a cutoff imposed by nature which prevent such extreme events of fast diversification and the BDH model does not include any such mechanism, e.g. an exponential cutoff on the burst size distribution.

2.7 Discussion

We have introduced a new methodology to interpret what diversity in environmental sequence data can tell us about the ecological and evolutionary processes that shaped it. The key theoretical step in our new method is to recognize that faster diversification processes which appear intermittently and last only for a short time still leave a signature in imperfectly reconstructed phylogenies. This signature persists even when the quality and length of our sequence data and consequent resolution of the phylogeny is relatively low compared to the timescale of the processes.

Combining this realization with existing methods for inferring slower gradual processes from phylogenies we were able to quantify the balance of fast and slow processes, the so called *tempo* of evolution, and the parameter values that best describe the structure and distribution of burst sizes and therefore the *mode* of evolution.

Of note are also three technical advances without which we would be unable to fully leverage our theoretical approach. Firstly, the coarse-graining of trees introduces a new way of extracting empirical observables, the subtree size distribution across slices, which sidesteps the issue of poorly resolved and low resolution trees. Secondly, our generating function approach combined with the numerical complex integration step allows us to extract *exact* probabilities out of an infinite dimensional model. This is something methods which use perturbation theory or Markov matrix exponentiation are unable to achieve. Third and finally, our eGOF can confidently tell us whether empirical trees are typical or not of the model we use, in other words whether the model is sufficient to explain empirical features of coarse-grained trees. As far as we know current macroecological lineage based inference methods do not easily lend themselves to this kind of test because of the combinatorially difficult task of exploring the space of trees in an efficient manner.

Applying this framework to a large dataset encompassing heterogeneous habitats are stark: we almost always need these heterogeneous faster processes to complement gradual diversification in order to explain these data, and the parameters that best explain the observed fast heterogeneous structure of evolutionary

trees are surprisingly universal across studies and environmental context. Indeed the tempo is such that fast diversification happens almost always an order of magnitude more often than slow diversification, and the burst sizes distribution with parameter centered around $\alpha \sim 7$ and $\beta \sim 1.5$ indicates that the mode of diversification is exceedingly bursty.

In other words, our results indicate that one of the most salient feature of microbiomes is their remarkable and dramatic history of large and fast diversification events which happen again and again over several timescales all throughout the tree of life and across multiple environments. The sheer amount of fast diversification events and the number of new lineages they engender suggests that the number of niches, discovered or constructed through the ecological evolution of their microbes as they innovate, enter new environment, etc., must be just as dramatic. This loops back to the debate about the dimensionality of niche space[101] and whether it is bounded due to diversity dependent diversification processes[102], or unbounded[103] in line with a more ecologically informed understanding of diversification that goes beyond simple ideas of competitive pressure and towards openness of evolving species communities. Although we cannot yet explain why and the precise mechanisms that underly it, we suggests that the dimensionality of microbial niche space is marked and generated by a history of striking radiation events, e.g. following metabolic, ecological, and morphological innovations with putatively unbounded number of niche dimensions, and aligns with Gould and Eldredge's hypothesis of punctuated equilibrium rather than that of phyletic gradualism[52].

2.8 Conclusion

These results raise many questions, and open a number of doors for future investigation. Perhaps the primary open question is: what biological changes cause the bursts we observe in empirical trees? Are these genuinely due to innovations, where an adaptation opens the door to many further adaptations [67], [68], [104], [105]? They could also be the result of exploration of new habitats, disturbance opening up niche space to be invaded [106], [107], or something else entirely. Our current analysis cannot answer these questions clearly, but the evidence does clearly show that an explanation is necessary. Second, we have shown that a class of distinct fast processes all map on to the same observable phenomena at coarse temporal resolutions through a combination of their parameters. This is a quantitative example of a long-discussed idea in ecology that only a handful of parameters survive to describe phenomena at larger or longer scale. The assumption is inherent in neutral models, but also in other, simplified models of macroecological patterns [108]. Our approach can form the starting point of quantitatively understanding which parameters and processes 'upscale', and which do not[88]. Finally, why do we see such clearly convergent and universal patterns across divergent habitat types? The ecological and evolutionary constraints leading to the patterns we've seen deserve a fuller explanation.

Chapter 3

Unsupervised Discovery of Niche Signals and Microbial Units

3.1 Introduction

As we have seen in the previous chapter the pervasive burstiness of microbial phylogenies and the wealth of lineages it generates suggest the presence of a vast hierarchy of niches within which microbial diversify distributes itself. This hierarchy should leave a signature somewhere in the structure of a microbial community. To peak at this signature we need to zoom in to the mesoscopic scale and look at the way microbial abundance responses partition themselves across environmental gradients. By abundance responses we mean the numerical values by which the total abundance of a lineages split into the abundances of their daughter lineages. This response is therefore a proportion between 0 and 1 which changes or stays the same in different environmental conditions and for different lineages. If a lineage is completely filtered away, say, because of the presence of extreme conditions in a given environment, then you would expect the proportion to fall to 0. If it is well adapted, then you would expect that it will maintain a non-zero response.

Indeed, following repeated downstream events of diversification we would expect that there is a similar downstream hierarchical partitioning of abundance responses across the multiple environments spanned by such events together with potential reorganization of abundances within the standing diversity. One of our core assumption is therefore that if this scenario holds, then there must be signatures of it left in patterns of current day abundances.

This recursive partitioning of abundances gives us a natural way to define ecological microbial units. At one extreme, two sister clades for which one is present in only a handful of environments and absent in others, and vis-versa, serve as the prototypical ecological microbial unit. The response is an indicator that moves in tandem with its environment; those two sister clades 'discover' an environmental niche and the niche correlates with their presence or absence. At an intermediate level, two sister clades that respond in a coherent way across several environments, say by splitting their parent's abundance in three different ways, also acts as indicator not just of a single environment but of several. Maybe in the presence of high salinity the response is 0.3, in medium salinity the response is 0.5, i.e. they split evenly their parent's abundance, and at low salinity the response is 0.9. In other words their relative abundance 'discovers' an environmental gradient and thus the potential presence of two niches.

All this to say that we assume that the way the intensity of a response, namely its statistically inferred

value, clusters across various environments serve as an indicator of the presence of different ecological niches within different clades. Given that deeper clades can manifest this type of clustering and not just individual taxa and that we effectively set out to discover all of these clustered responses, we will in effect, if we are successful, find a representation of the repeated entry or construction of niches at all levels within a given phylogenetic tree.

Suppose now that we are given a set of microbiome samples and that we do not know *a priori* neither which set of, or how many niches are presented in each sample nor do we know how those niches break apart into smaller more specialized niches. We therefore find ourselves faced with a difficult hierarchical clustering problem. Indeed we do not and cannot assume that we know and have measured a list of environmental features like temperature, salinity, acidity, etc., which exhaustively covers all dimensions of potential niches across all microbiome samples. This problem is thus the purview of unsupervised machine learning methods which is, to make a reductive comparison, a generalization of traditional ordination methods like PCA or NMDS methods. One of the main advantage of unsupervised learning methods over ordination methods is that we will explicitly write and invert a probabilistic generative model rather than use *ad-hoc* clustering methods with arbitrarily chosen dissimilarity measures or distance metrics.

We take inspiration from the field of computational linguistic and more precisely probabilistic topic modeling, a machine learning approach which discovers a mixture of topics within a corpus of documents automatically[109]–[111]. Probabilistic topic modeling seek to find latent topics which may have generated the set of words in a document across a corpus of documents. Topics are latent and are represented by an unknown distribution over a dictionary of words, usually modeled by a multinomial distribution, and words are a single draw out of the multinomial distribution of their associated topic. This is what is called a 'bag-of-word' model whereby the order and syntax of the language is ignored and only the frequency of words matters. Topics are characterized by the set of probabilities that enter as parameters of their respective multinomial distribution and the distribution over those parameters which models the uncertainty in each topic is given by the conjugate prior of the multinomial distribution, namely a Dirichlet distribution. The unsupervised learning approach discovers automatically the number of topics, the distribution over words of a dictionary of each topics, and the topic associated with each word. In other words, we seek to find the distribution of topics, itself a distribution over distributions of words in a dictionary.

The correspondence goes as follow: each microbiome sample is a document constituted of a set of words. Each word is given by a couple of relative abundances between two sister clades. We therefore think of words as an abundance couple (k_l, k_r) taken to be a binomial draw with $k = k_1 + k_2$ where k_1 represent the relative abundance of one sister clade and k_2 the relative abundance of the other sister clade. Topics are binomial proportions ϕ, which we call the clade responses and the distribution characterizing the parameter of the response ϕ will be modeled using a beta distribution, namely the conjugate prior of the binomial distribution. A topic is therefore given by a binomial distribution over all binomial draws and the dictionary is the set of possible draws (k_1, k_2). To make a story short, each microbiome sample is a document generated by a mixture of topics, the proportion ϕ, and together they generate the set of words (k_1, k_1) across the clades of the phylogeny. In contrast to traditional topic modeling, each document has the same number of words, namely the number of clades in the phylogeny.

Now we go one step further. We do not assume that the mixture of topics in a microbiome sample is a flat collection of response ϕ. Instead we assume a hierarchy of topics in the sense that topics further separate into subtopics, themselves separating into sub-subtopic, and so on. Our probabilistic generative model will therefore seek to infer a hierarchy of topics. To do so we will need to use a further generalization of traditional

latent topic modeling called a nested Hierarchical Dirichlet process (nhDP). As hinted above and explained in the next section we will adapt this model to conventional microbial ecology datasets and in doing so settle on a slight specialization of the nhDP which we will call the path-limited nhDP, or pl-nhDP. Models of this kind and their generalizations have a long history within the field of computational linguistics, especially e.g. the work of Blei *et al.*[112]–[116]. Our method rely on a model in between the nested CRP [115] and the hierarchical nested Dirichlet processes [116]. Strong similarities also exist with the work of Ghahramani *et al.*[117] where ours use a much simpler transition kernel when navigating up the hierarchy of topics.

3.2 Methods

3.2.1 From Abundance Tables to Phylogenetically Informed Abundance Tables

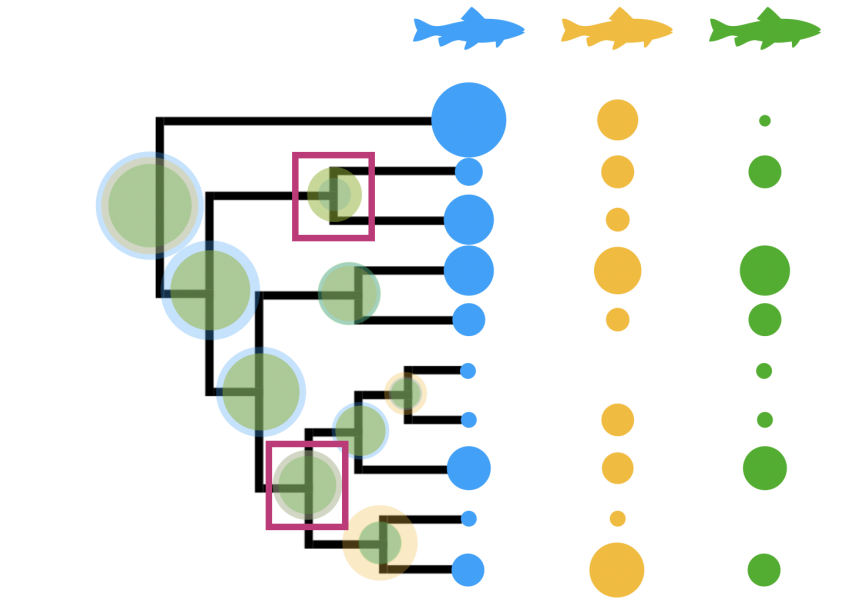
Microbial ecology datasets are traditionally comprised of two things. On one hand we have a phylogenetic tree which spans the set of all microbial 'species", usually imperfectly defined as OTUs or oligotypes which are cluster or component of a mixture of sequences in genetic space, found across all microbiome samples. On the other hand we have an abundance table for each of those species across samples (see Figure 3.1). Highlighted in pink on the figure are two clades with their respective sister clades. The proportion with which relative abundances of two sister clades split their parent's total abundance is what we called the clade response. Take for example the clade highlighted at the bottom and let's focus on sample Z0221. The top sister clade has abundance $2072 + 2 + 1289 = 3363$ while the bottom sister clade has abundance $788 + 2801 = 3589$. This constitute what we called a word, or abundance split, given by the couple $\{3363, 3589\}$. Together they split their parent's total abundance $3363 + 3589 = 6952$. One can thus say (in a ML context) that the response is approximately $3363/6952 \approx 0.48$. Indeed, rather than the direct abundances at the level of taxa as shown on the right in Figure 3.1), we can instead use the set of abundances of sister clades splitting their parent's abundance. We will call this the phylogenetically transformed abundance table, or phylogenetic abundance table for short. One can think of this table as decomposing taxa level abundances into phylogenetic contrasts.

3.2.2 Probabilistic Generative Modeling

We can use a product of binomials to write a very simplistic model which generate the phylogenetic abundance table across samples and microbial 'species'. This model takes the following form:

$$\begin{aligned} P(\{(k_{si_1}, k_{si_2})\}_{s\in\mathcal{S}, i\in\mathcal{I}} \mid \{\phi_{si}\}_{s\in\mathcal{S}, i\in\mathcal{I}}) &= \prod_{s\in\mathcal{S}}\prod_{i\in\mathcal{I}} P(k_{si_1}|n_{si}, \phi_{si}), \\ &= \prod_{s\in\mathcal{S}}\prod_{i\in\mathcal{I}} \binom{k_{si}}{k_{si_1}} \phi_{si}^{k_{si_1}} (1-\phi_{si})^{k_{si_2}}, \end{aligned} \tag{3.1}$$

where $\mathcal{S}$ is the set of microbial samples, $\mathcal{I}$ is the set of internal nodes of the phylogenetic tree, i_1 and i_2 are respectively the two sister clades of clade i, k_{si_1} and k_{si_2} their respective abundances, the abundance of the parent clade $k_{si} = k_{si_1} + k_{si_2}$, and ϕ_{si} is the response of clade i in sample s. This model is highly parametrized, i.e. #parameters= $|\mathcal{S}| \times |\mathcal{I}|$. One could simply set every ϕ_{si} at their MLE $\phi^*_{si} = k_{si_1}/k_{si}$ but this model would be the least parsimonious model possible, i.e. with as many parameters as there are entries in the phylogenetic abundance table. This would be completely uninformative as to which clades in which samples respond coherently or not. This is the equivalent of ascribing a single topic containing a unique



OTU.rep	Z0220	Z0221	Z0237	Z0244	Z0253	Z0254
	10	9	3	3	0	2
	552	1015	3960	64	3485	1519
	78	155	523	18	468	217
	1	2094	12	10	38	11
	176	1173	237	1	1062	142
	2644	2072	2030	47	146	7
	2017	2	127	10	19	0
	774	1289	2	1	40	10
	2248	788	415	684	9635	1601
	2403	2801	871	9779	709	8498

Figure 3.1: **Top** Cartoon of a traditional microbial ecology dataset where columns represent abundances in a sample across a species and rows represent abundances across samples of a given microbial species. **Bottom** More concretely, in actual datasets abundances are collected into a table. We highlighted in pink two parent clades and their respective sister clades which split the abundances of their parent. The proportion of this split is what we call the clade response.

word to every single word in a corpus of documents. There are two simple potential extensions to this model whereby we assume clustering of samples or clustering of clades. In the following we will only show the first one given that the second one is simply obtained by interchanging $\mathcal{S} \leftrightarrow \mathcal{I}$ everywhere.

Probabilistic Generative Model of Sample-Wise Clusterings

In this model we parametrize a given sample s by a vector of responses $\vec{\phi}_s = \{\phi_i\}_s$ and introduce clusters in the space of those (high-dimensional) vectors. This vector serves as the phylogenetic 'fingerprint' of a sample. Let $S = |\mathcal{S}|$ be the total number of samples and $\pi_{[S]}$ a partition over samples. Each cluster $\lambda \in \pi_{[S]}$ now has therefore an associated vector $\vec{\phi}_\lambda$ containing a response for each clade, i.e. $\vec{\phi}_\lambda = \{\phi_{\lambda i}\}_{i\in\mathcal{I}}$. Such a model can therefore be written

$$\begin{aligned} P(\{k_{si}\}_{s,i} \mid \pi_{[S]}, \{\vec{\phi}_\lambda\}_{\lambda\in\pi_{[S]}}) &= \prod_{\lambda\in\pi_{[S]}} \prod_{s\in\lambda} \prod_{i\in\mathcal{I}} \binom{k_{si}}{k_{si_1}} \phi_{\lambda i}^{k_{si_1}} (1-\phi_{\lambda i})^{k_{si_2}}, \\ &= \prod_{\lambda\in\pi_{[S]}} \left[\prod_{i\in\mathcal{I}} \phi_{\lambda i}^{\sum_{s\in\lambda} k_{si_1}} (1-\phi_{\lambda i})^{\sum_{s\in\lambda} k_{si_2}}\right] \left[\prod_{s\in\lambda} \prod_{i\in\mathcal{I}} \binom{k_{si}}{k_{si_1}}\right] \end{aligned} \tag{3.2}$$

Even though we reduced the number of parameter to $|\pi_{[S]}|$ vectors of $|\mathcal{I}|$ responses, we are left with finding the particular partition $\pi_{[S]}$ which best describes the set of abundances in the phylogenetic abundance table across samples. Even though we could in theory proceed with model selection, the number of partitions of a set $[S]$ is given by the Bell number B_S which has the unfortunate growth rate $B_S \sim \mathcal{O}(S^S)$, i.e. a super-exponential growth. To give an idea of this kind of growth, $B_4 = 15$, $B_{10} = 115975$, and $B_{40} = 1.6 \times 10^{35}$. This combinatorial explosion in the number of possible partition makes an approach using naive model selection completely impractical. The picture would be even more dire if we decided to partition clades rather than samples given that phylogenetic trees usually have many more clades than there are samples in the dataset.

This is where the use of nonparametric generative models can help. More specifically the use of the nonparametric prior over partition called the Dirichlet Process (DP)[118], and the closely related and simpler Chinese Restaurant Process (CRP)[119]. The DP is the infinite dimensional generalization of the Dirichlet distribution, namely a distribution over distributions of arbitrary dimensions, and the CRP, which is a marginalized version of the DP and in particular arises when using conjugate priors, provides a way to regularize the space over partitions. This regularization leads to a natural way to explore the space of partitions and lends itself to a simple sampling scheme using MCMC with Gibbs sampling. Given that the binomial distribution has a simple conjugate prior, the beta distribution, we can thus use the simpler CRP in the following. The CRP is a prior over partitions which can be written

$$P(\pi_{[S]}|\alpha) = \frac{\alpha^K}{\alpha^{(S)}} \prod_{\lambda\in\pi_{[S]}} (|\lambda|-1)!, \tag{3.3}$$

where the Pochhammer symbol, or rising factorial, $\alpha^{(S)} = \alpha(\alpha+1)\dots(\alpha+S-1) = \Gamma(\alpha+S)/\Gamma(\alpha)$. One important property of the CRP is that on average the number of clusters grows like $\alpha \log S$. This is why we say that it *regularizes* the space of partitions. Even though it can grow unboundedly with the number of data points, hence the label 'nonparametric', the number of parameters remains on average logarithmically bounded.

This expression can be understood simply through the metaphor of seat assignments used at certain Chinese restaurants whereby as customers pour in, the probability of sitting at a table that is already occupied

is proportional to the number of customers already sitting at that table, and the probability of sitting at an empty table is proportional to the concentration parameter $\alpha \geq 0$. The greater the concentration parameter α, the more likely new tables will be opened. In the opposite limit $\alpha \to 0$ everyone ends up sitting at the same table.

Concretely, when the first customer arrives, they are immediately seated at an empty table. When the second customer arrives, they are seated at the same table as the first customer with probability $1/(1+\alpha)$ and at a new empty table with probability $\alpha/(1+\alpha)$. There are thus two possible scenarios with two customers. Either both are seated at the same table, which we denote by the partition $\pi_{[2]} = \{\lambda_1\} = \{\{1,2\}\}$, or both are seated at their own table which we denote by the partition $\pi_{[2]} = \{\lambda_1, \lambda_2\} = \{\{1\},\{2\}\}$. Given a restaurant where S customers are seated at K tables with arrangement $\pi_{[S]} = \{\lambda_1, \ldots, \lambda_K\}$, then the next customer to arrive will be seater at table 1 with probability $|\lambda_1|/(S+\alpha)$, at table 2 with probability $|\lambda_2|/(S+\alpha)$ and so on and with probability $\alpha/(S+\alpha)$ they will be seated at an empty table. Notice that for a table λ with $|\lambda|$ customer, the numerator of the probability of reaching this number is always equal to $\alpha \times 1 \times 2 \times \ldots \times (|\lambda|-1) = \alpha(|\lambda|-1)!$ regardless of the order of customer arrival at the other tables. This is the case for every table $\lambda \in \pi_{[S]}$. This takes care of the numerator in Equation 3.3. Similarly, for the restaurant to reach a state with 1 customers, the assignment probability was made with proportionality factor $1/\alpha$. To reach a state with 2 customers from one with 1 customer, all possible assignments are made with a proportionality factor $1/(1+\alpha)$. To reach a state with S customers from one with $S-1$ customers all potential assignments were made with a proportionality factor $1/(S-1+\alpha)$. To reach a restaurant with S customers starting with no customer we therefore have that the probability of this state is proportional to $1/\alpha \times 1/(1+\alpha) \times \ldots \times 1/(S-1+\alpha)$. This takes care of the denominator in Equation 3.3. Notice that the final probability is therefore independent of the precise order with which seats were assigned at tables. There is no memory of the precise construction process of the ordering, but the probability of the next draw *is* correlated with the current state. In other words, draws out of a CRP are *not independent*, but the lake of memory of the CRP makes series of draws *exchangeable* sequences.

We are almost at the point where we can write a full generative model for the phylogenetic abundance table. We have the data likelihood which gives the probability of observing abundances given a set of responses, we have the probability of observing a given partitioning of responses across samples given the clustering parameter α, and now all that is missing is the probability of given response values in different sample clusters and across clades. To wit, the CRP gives us a nonparametric prior over over $\pi_{[S]}$, we simply need a prior over $\vec{\phi}_\lambda$. Given that our data likelihood is a product of binomial, it is natural to use the uninformative Jeffreys prior of the binomial distribution, namely the beta distribution with shape parameters identically $1/2$. To wit,

$$P(\vec{\phi}_\lambda) = \prod_{i \in \mathcal{I}} \text{Beta}(\phi_{\lambda i} \mid \beta, \beta) = \prod_{i \in \mathcal{I}} \frac{\phi_{\lambda i}^{\beta-1}(1-\phi_{\lambda i})^{\beta-1}}{B(\beta,\beta)} \tag{3.4}$$

where $\beta = 1/2$ and the beta function $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$. Using the CRP we can now write a probabilistic generative model with flat partitioning of samples

$$P(\{k_{si}\}_{\mathcal{S},\mathcal{I}}, \pi_{[S]}, \{\vec{\phi}_\lambda\}_{\lambda \in \pi_{[S]}} \mid \alpha, \beta) = \frac{\alpha^{|\pi_{[S]}|}}{\alpha^{(S)}} \prod_{\lambda \in \pi_{[S]}} (|\lambda|-1)! \prod_{i \in \mathcal{I}} \frac{\phi_{\lambda i}^{\beta-1}(1-\phi_{\lambda i})^{\beta-1}}{B(\beta,\beta)} \prod_{s \in \lambda} \binom{k_{si}}{k_{si_1}} \phi_{\lambda i}^{k_{si_1}} (1-\phi_{\lambda i})^{k_{si_2}}. \tag{3.5}$$

Given that the Jeffreys prior of the binomial distribution is also its conjugate prior we can integrate out $\vec{\phi}_\lambda$

and find the simpler generative model and its posterior distribution over $\pi_{[S]}$

$$P(\{k_{si}\}_{\mathcal{S},\mathcal{I}}, \pi_{[S]} \mid \alpha, \beta) = \frac{\alpha^{|\pi_{[S]}|}}{\alpha^{(S)}} \prod_{\lambda \in \pi_{[S]}} (|\lambda| - 1)! \left[\prod_{i \in \mathcal{I}} \frac{B(\beta + \sum_{s \in \lambda} k_{si_1}, \beta + \sum_{s \in \lambda} k_{si_2})}{B(\beta, \beta)} \right] \left[\prod_{s \in \lambda} \prod_{i \in \mathcal{I}} \binom{n_{si}}{k_{si_1}} \right],$$
$$\Rightarrow P(\pi_{[S]} \mid \{k_{si}\}_{\mathcal{S},\mathcal{I}}, \alpha, \beta) \propto \frac{\alpha^{|\pi_{[S]}|}}{\alpha^{(S)}} \prod_{\lambda \in \pi_{[S]}} (|\lambda| - 1)! \prod_{i \in \mathcal{I}} \frac{B(\beta + \sum_{s \in \lambda} k_{si_1}, \beta + \sum_{s \in \lambda} k_{si_2})}{B(\beta, \beta)}. \tag{3.6}$$

From this expression we can devise a simple Gibbs sampler which explores the posterior distribution over the space of partitions $\pi_{[S]}$. We start with a random partition π over samples $s = 1, \ldots, S$. Let $\pi_{-\sigma}$ be the current partition with sample σ removed. We want to reassign σ to any one of the cluster already present in $\pi_{-\sigma}$ or to its own individual cluster and in doing reach a new partition π' which can be the same as π or slightly different from it in that σ might belong to the same cluster or a different cluster. If we reassign σ to one of the current cluster $c \in \pi_{-\sigma}$ then the partition will change from π to $\pi' = \pi_{-\sigma} - c + c \cup \{\sigma\}$ and therefore $|\pi'| = |\pi_{-\sigma}| = K$. If we reassign σ to its own new cluster then the partition will change from π to $\pi' = \pi_{-\sigma} + \{\sigma\}$ and therefore $|\pi'| = |\pi_{-\sigma}| + 1 = K + 1$. Let

$$G_{\lambda,\mathcal{I}} = \prod_{i \in \mathcal{I}} \frac{B(\beta + \sum_{s \in \lambda} k_{si_1}, \beta + \sum_{s \in \lambda} k_{si_2})}{B(\beta, \beta)}.$$

The second line of Equation 3.6 therefore leads to the following probability for assignments,

$$P(\pi' | \pi_{-\sigma}, \{k_{\sigma i},\}_{\mathcal{I}}, \{k_{si},\}_{S_{-\sigma},\mathcal{I}}) \propto \begin{cases} \dfrac{\frac{\alpha^K}{\alpha^{(S-1)}} \left[\frac{|c-1|!|c|}{\alpha+S-1} G_{c+\{\sigma\},\mathcal{I}} \right] \left[\prod_{\lambda \in \pi_{-\sigma}-c} (|\lambda|-1)! G_{\lambda\mathcal{I}} \right]}{\frac{\alpha^K}{\alpha^{(S-1)}} \prod_{\lambda \in \pi_{-\sigma}} (|\lambda|-1)! G_{\lambda\mathcal{I}}}, & \pi' = \pi_{-\sigma} - c + c \cup \{\sigma\}, \\ \dfrac{\frac{\alpha^K}{\alpha^{(S-1)}} \left[\frac{\alpha}{\alpha+S-1} G_{\{\sigma\},\mathcal{I}} \right] \left[\prod_{\lambda \in \pi_{-\sigma}} (|\lambda|-1)! G_{\lambda\mathcal{I}} \right]}{\frac{\alpha^K}{\alpha^{(S-1)}} \prod_{\lambda \in \pi_{-\sigma}} (|\lambda|-1)! G_{\lambda\mathcal{I}}}, & \pi' = \pi_{-\sigma} + \{\sigma\}, \end{cases} \tag{3.7}$$

and after cancelling common factors

$$P(\pi' | \pi_{-\sigma}, \{k_{\sigma i},\}_{\mathcal{I}}, \{k_{si},\}_{S_{-\sigma},\mathcal{I}}) \propto \begin{cases} \frac{|c|}{\alpha+S-1} \frac{G_{c+\{\sigma\},\mathcal{I}}}{G_{c,\mathcal{I}}}, & \pi' = \pi_{-\sigma} - c + c \cup \{\sigma\}, \\ \frac{\alpha}{\alpha+S-1} G_{\{\sigma\},\mathcal{I}}, & \pi' = \pi_{-\sigma} + \{\sigma\} \end{cases} \tag{3.8}$$

Then the algorithm proceeds by repeatedly removing a randomly chosen sample σ from the current state π and reassigning it following Equation 3.8. Notice that this is nothing but the CRP described above augmented by the evidence provided by the phylogenetic abundance table. This is in essence the Gibbs algorithm found e.g. in [120] which at stationarity produces a chain over the *a posteriori* probability distribution of partitions $\pi_{[S]}$, namely over clusterings of samples.

We can summaries this probabilistic generative model of sample-wise clustering in a compressed distribu-

tional form

$$\begin{aligned} G &\sim \mathrm{DP}\left(\alpha \prod_{i \in \mathcal{I}} \mathrm{Beta}(\beta, \beta)\right), \\ \vec{\phi}_s \mid G &\sim G, \quad s \in \mathcal{S} \\ \{k_{si_1}, k_{si_2}\} \mid k_{si}, \phi_{si} &\sim \mathrm{Binomial}(k_{si_1}, k_{si_2} \mid k_{si}, \phi_{si}), \quad i \in \mathcal{I}, s \in \mathcal{S}. \end{aligned} \tag{3.9}$$

It says that first we draw a distribution over vectors $\vec{\phi}$ out of the DP. Then out of this distribution we draw a vector $\vec{\phi}_s$ for each sample $s \in \mathcal{S}$. Finally given each $\vec{\phi}_s$ we draw splits $\{k_{si_1}, k_{si_2}\}$ out of a binomial distribution with parameter $k_{si} = k_{si_1} + k_{si_2}$ and ϕ_{si} for each clade $i \in \mathcal{I}$ in the phylogenetic abundance table and across sample $s \in \mathcal{S}$. The Gibbs algorithm described above inverts this model to recover G and the assignments of $\vec{\phi}_s$ to atoms ϕ in G. This is possible because the *a posteriori* probability of the DP is an atomic distribution which is its own conjugate prior. The atoms in the posterior DP distribution are the clusters at some set of values $\vec{\phi}_\lambda$ together with a new atom at $\vec{\phi}_{\text{new}} \sim$ Beta.

At the beginning of this section we mentioned that there is a converse generative model of flat partitions, namely the one where we interchange $\mathcal{S} \leftrightarrow \mathcal{I}$ everywhere. This clade-wise model would find partitions over clades rather than samples and each cluster of this partition would be associated with a vector $\vec{\phi}_i = \{\phi_{is}\}_{s \in \mathcal{S}}$ of responses across samples (the rows of the phylogenetic abundance table) for clade i. This lead us to recognize the most cumbersome limitation of flat clustering models; there is nothing that guarantee that the clusters of the sample-wise model and those of the clade-wise model are compatible, or commensurable in that neither of them inform the other. Clustering of rows of the phylogenetic abundance table do not say anything about clustering of columns of the phylogenetic abundance table.

Probabilistic Generative Model of Nested Hierarchical Clusterings

In the context of our phylogenetic abundance table, the above model works only to cluster samples together. It informs us about which samples have the totality of their clades respond in similar ways within statistical certainty offered by binomial distribution. One would thus expect that we can only learn about the strongest niche signal; are there 2, 3, 4, or N niches which seem to lead to similar tree-wide responses and therefore similar relative abundances throughout the phylogenetic abundance table?

But what if we have the following scenario. Imagine that there is an overarching gradient, or niche dimension like temperature, along which there are two temperature niches t_1 and t_2. In niche t_1 a subset of clades x, y, and z respond with $\phi_x^{t_1}$, $\phi_y^{t_1}$ and $\phi_z^{t_1}$, and in niche t_2 they respond instead with $\phi_x^{t_2}$, $\phi_y^{t_2}$ and $\phi_z^{t_2}$. Then imagine there are sub-niches along a second dimension in niche space, say, salinity. At salinity sub-niche s_1 a particular clade α respond with $\phi_\alpha^{t_1,s_1}$ and in salinity sub-niche s_2 with $\phi_\alpha^{t_1,s_2}$. Imagine finally that this scenario continues recursively. One can quickly see that the previous flat sample-wise model (nor the clade-wise model for that matter) is completely incapable of capturing this more subtle hierarchy and we are left seeking a more complicated, certainly hierarchical model to handle it. Indeed we were inferring flat partitions of the form

$$\{\{A, B, D\}, \{C, F\}, \{E, G\}, \{H\}\},$$

but what we really need is to infer recursive partition, called partition refinements, of the form

$$\{\{\{A, D\}, B, C\}, \{\{F, G\}, \{H\}\}.$$

This recursive, hierarchical scheme we will represent as a tree, for example as shown in Figure 3.2.

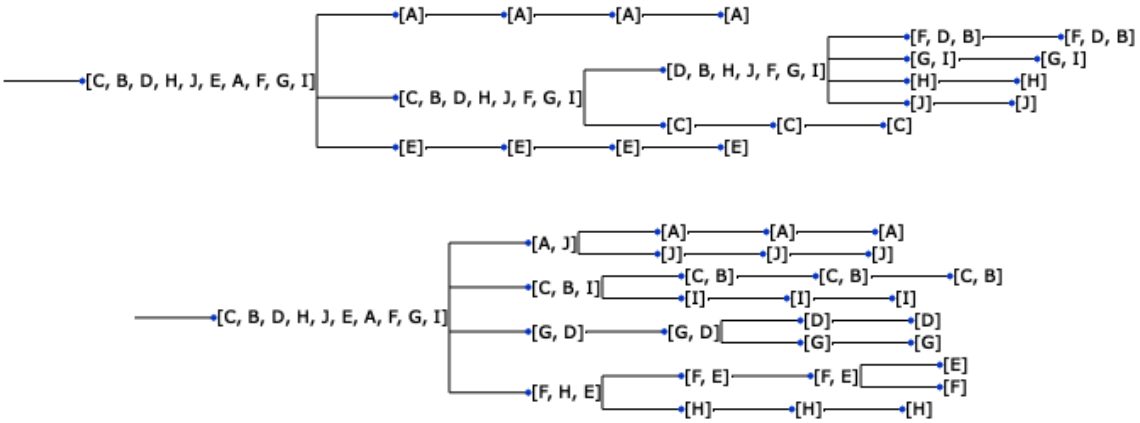


Figure 3.2: **Top** Tree representation of the refinement $\{\{A\}, \{\{\{F, D, B\}, \{G, I\}, \{H\}, \{J\}\}, \{C\}\}, \{E\}\}$ **Bottom** Tree representation of the refinement $\{\{\{A\}, \{J\}\}, \{\{C, B\}, \{I\}\}, \{\{G\}, \{D\}\}, \{\{\{F\}, \{E\}\}, \{H\}\}\}$. We collapsed degenerate refinements, i.e. given the top tree we should actually have $\{\{\{\{A\}\}\}\}$ instead of $\{A\}$ and so on for the other elongated branches.

Let us recapitulate the initial setting. Suppose we are sampling various microbiomes. Each microbiome sample represents a document with its words. The set of those words for a given sample/document s is the set of abundance splits $\{k_{si_1}, k_{si_2}\}_{s,\mathcal{I}}$ across the clades of the phylogenetic abundance table for that sample. Each word is understood to come from a topic which is characterized by a response ϕ that enters into the binomial distribution over realizations of abundance splits. Therefore documents are generated by a pool of latent topics, namely samples are generated by a pool of latent responses.

The purpose of the pl-nhDP model is to organize these topics, the responses, in a hierarchical way and to generate the abundance splits found in the phylogenetic abundance table from those responses. To help clarify the structure of our model consider Figure 3.3. In this structure there is one response sitting at each node. Samples are associated with paths from the leaf of the hierarchy to the root of the hierarchy. The nodes a sample path intersect creates the pool of responses that can generate the abundance splits down the column of the phylogenetic table. Clades are associated with slices across the hierarchy and 'cut' the hierarchy in disjoints subtree which joined together cover all leaves. Clade slices select the particular responses that will generate the abundance splits across rows of the phylogenetic table. If more than one clade slice intersect a path, then all those clades will respond coherently for that sample. If more than one sample path intersect a clade slice, then all those samples will respond coherently for that clade. If more than once clade slices intersect more than one paths, then all those clades across those samples will respond coherently. The pl-nhDP creates partitions that are commensurable across clades and across slices. This is the most important feature of the pl-nhDP.

Here's how we more precisely construct the pl-nhDP. First, we have a hierarchy of topics generated by a nested CRP. A simple way to understand the nested CRP is to extend the original metaphor of the CRP by the introduction of a recursive structure. Once a customer has been seated at a table and is done eating, they aredirected to a different restaurant where they begin the CRP anew. The restaurant they are directed to depends on which table they were seated at. This process produces a nested structure of restaurants where each table in each restaurant points to another restaurant and so on. At each node of the hierarchy, namely a

table in a given restaurant, is a particular dish that everyone at that table eats. This dish is the equivalent of a topic given by a particular response ϕ. Let us translate this in more precise terms. The hierarchical tree

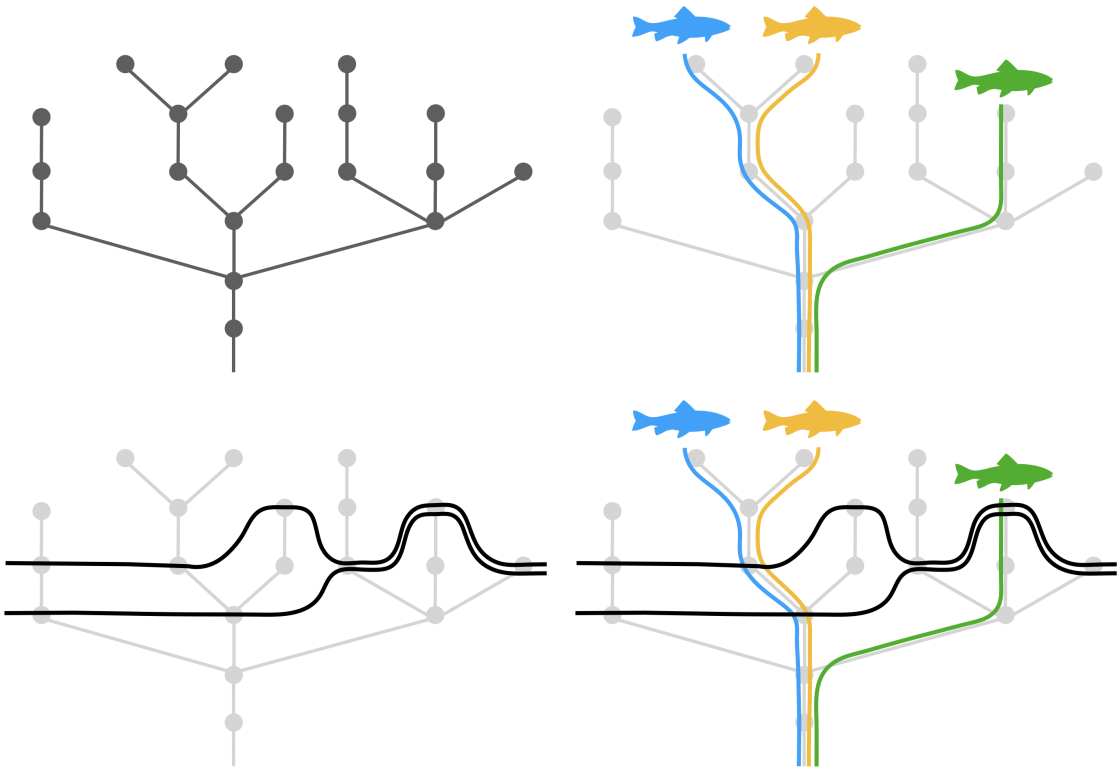


Figure 3.3: **Top-left** Hierarchy of topics. At each node of the hierarchy sits a response ϕ. **Top-right** Each sample, or document, is represented by a path (in blue, orange, and green), namely its sample path, from a leaf of the hierarchy to the root of the hierarchy and picks out a pool of potential responses which will be used to generated the abundance splits along its column of the phylogenetic abundance table. **Bottom-left** Each clade is represented by a slice (in black) across the hierarchy, namely its clade slice. The slice selects which responses in particular will generate the abundance splits within one row of the phylogenetic table. **Bottom-right** The pl-nhDP is the combination of the hierarchy, all sample paths, and all clade slices. For a given clade in a given sample, its abundance split is generated by the response found at the intersection of the clade slice with the sample path. If more than one clade slice intersect a path, then all those clades will respond coherently for that sample. If more than one sample path intersect a clade slice, then all those sample will respond coherently for that clade.

this process generates is an infinite multifurcating tree where at every node live individual Dirichlet processes. Fortunately in a context with finite data a given realization of the posterior hierarchical tree will be finite. We have thus

$$\begin{aligned} G_{\boldsymbol{\eta}_l} &\sim \mathrm{DP}\left(\alpha\,\mathrm{Beta}(\beta,\beta)\right), \quad \boldsymbol{\eta}_l \geq \mathbf{1}_l \text{ and } l \geq 0, \\ \phi_{(\boldsymbol{\eta}_l,\eta_{l+1})} \mid G_{\boldsymbol{\eta}_l} &\sim G_{\boldsymbol{\eta}_l}, \quad \eta_{l+1} \geq 1 \end{aligned} \tag{3.10}$$

where $\boldsymbol{\eta}_l$ is a tree coordinate up to depth l. For example in the top tree in Figure 3.2 the leaf of sample J would have coordinate $\eta_4 = (1, 2, 1, 4, 4)$. The symbol $\mathbf{1}_l$ represent the left-most (or top-most) coordinate vector $(1, 1, \ldots, 1)$ of length l. We use once again the Jeffreys prior ($\beta = 1/2$) as the base distribution from which to draw binomial parameters. The above specification stipulates that for each of the infinite number of nodes in the hierarchical tree we draw a random distribution over responses $0 < \phi < 1$, and out of this

distribution we draw a response $\phi_{(\eta_{l+1},\boldsymbol{\eta}_l)}$. Together, the set of all $G_{\boldsymbol{\eta}_l}$ exhausts all available responses that can generate a phylogenetic abundance table. This concludes the 'nested Dirichlet processes' part of our generative model.

Each sample needs a pool of responses to generate its abundance splits down its column of the phylogenetic abundance table. For each sample s we draw a path of Dirichlet processes sitting at nodes, the sample path, from the root to one of the tip of the nested DP,

$$G_l^s \mid G_{\boldsymbol{\eta}_l} \quad \sim \quad \lim_{\gamma\to 0} DP(\gamma G_{\boldsymbol{\eta}_l}) \quad l \geq 0, s \in \mathcal{S}. \tag{3.11}$$

The $\gamma \to 0$ limit of this hierarchical draw over all $G_{\boldsymbol{\eta}_l}$ insures that each level $l \geq 0$ we draw only one node. The node drawn at level l specifies which Dirichlet process to draw from at level $l+1$, e.g. if at level l we draw $G_{(\boldsymbol{\eta}_l,4)}$ out of $G_{\boldsymbol{\eta}_l}$, then at level $l+1$ we will draw the next Dirichlet process in the path out of $G_{(4,\boldsymbol{\eta}_l)}$. This process of stepping forth by selecting only one node at each level produces the said sample path. In other words at each level

$$G_l^s = \delta_{\phi_{\boldsymbol{\eta}_{l+1}^s}}.$$

Using this atomic limit we denote the random path drawn for sample s

$$\boldsymbol{\eta}^s = (\eta_1^s, \ldots \eta_l^s).$$

This path is the precise trajectory of restaurant a customer ended up going through. This concludes the 'path-limited' part of our generative model.

Finally, we add an additional hierarchical component to our model in order to draw a random slice for each individual clade, the clade slice. To do so we make use of stochastic switches sitting at each node $\boldsymbol{\eta}_l$ of the hierarchical tree. For each switch we draw a stopping probability $Q_{\boldsymbol{\eta}_l}$ that determines the distribution from which we will draw its state $U_{\boldsymbol{\eta}_l} = 0$ or 1. More precisely,

$$\begin{aligned} Q_{\boldsymbol{\eta}_l} \quad &\sim \quad \text{Beta}(\mu_1, \mu_2), \\ U^i_{\boldsymbol{\eta}_l} | Q_{\boldsymbol{\eta}_l} \quad &\sim \quad \text{Bernouilli}(Q_{\boldsymbol{\eta}_l}), \quad \boldsymbol{\eta}_l \geq i \in \mathcal{I}, \mathbf{1}_l, l \geq \end{aligned} \tag{3.12}$$

and the state of the root switch is always off so we always skip it. Then for each clade i within sample s we select a slice intersection H^{si} using the product of the states of switches from the root towards the leaves along the sample path, to wit

$$\begin{aligned} H^{si} \mid \boldsymbol{U}^i, \boldsymbol{G}^s \quad &\sim \quad \sum_{l\geq 1} G^s_{l-1} \; U^i_{\boldsymbol{\eta}_l^s} \prod_{m=1}^{l-1} (1 - U^i_{\boldsymbol{\eta}_l^s}), \\ &= \quad \sum_{l\geq 1} \delta_{\phi^s_{\boldsymbol{\eta}_l}} \; U^i_{\boldsymbol{\eta}_l^s} \prod_{m=1}^{l-1} (1 - U^i_{\boldsymbol{\eta}_l^s}), \quad s \in \mathcal{S}. \end{aligned} \tag{3.13}$$

Given that the state of the switch is either 1 or 0 notice how a clade selects only one node and its binomial proportion, the response, along each sample path. The above construction gives rise to an infinite tree-structured categorical distribution with probabilities along each sample path given by a Griffiths-Engen-McCloskey (GEM) process with parameters μ_1 and μ_2 which we set to 1/2. This concludes the 'hierarchical' part of our generative model.

Last but not least the data likelihood

$$\begin{aligned} \psi_{si} \mid H^{si} &\sim H^{si}, \\ k_{si_1}, k_{si_2} \mid k_{si}, \psi_{si} &\sim \text{Binomial}(n_{si}, \psi_{si}), \quad s \in \mathcal{S}, i \in \mathcal{I} \end{aligned} \tag{3.14}$$

Figure 3.4 shows an example of the GEM process overlaid on top of the nCRP for a real dataset of zebrafish gut microbiome samples. Figure 3.5 shows a subsample of clade slices inferred from the same dataset where one can see the remarkable heterogeneity that the pl-nhDP can extract from real data. The pl-nhDP

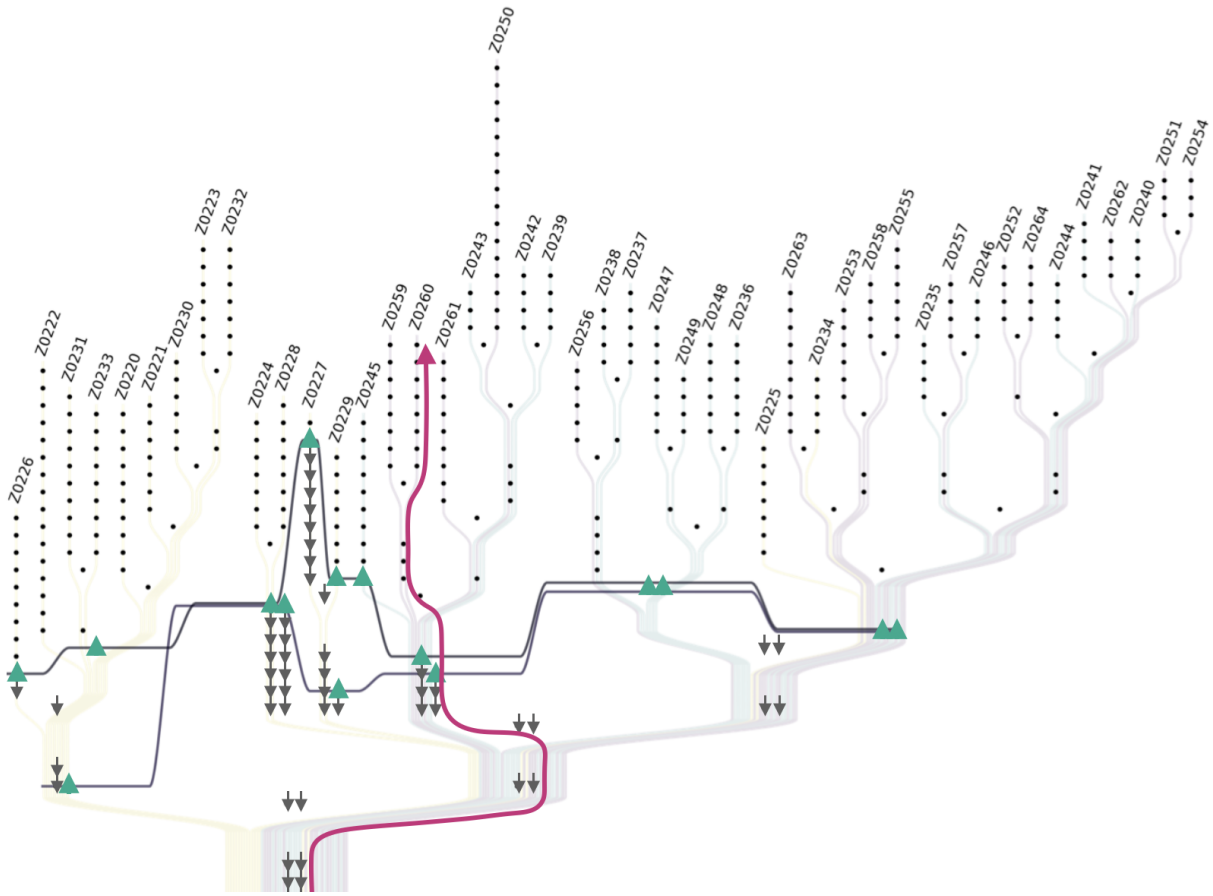


Figure 3.4: Example of a sample path and two clade slices on a real pl-nhDP. In **purple** a path from root to leaf which selects nodes throughout the hierarchical three generated using the nCRP. In **gray** and **green** the stochastic switches of the GEM process. Gray switches are in state 0 and green switches in state 1. Notice how a clade slice is made of a transversal set nodes in the state 1.

generative model gives for a given configuration the probability

$$P(\{k_{si_1}, k_{si_2}\}, \{\pi_{\boldsymbol{\eta}}\}, \{\boldsymbol{\eta}^s\}, \{\boldsymbol{U}^i\}) =$$
$$\prod_{\boldsymbol{\eta}} \left[\left[\frac{\alpha^{|\pi_{[P_{\boldsymbol{\eta}}]}|}}{\alpha^{(P_{\boldsymbol{\eta}})}} \prod_{\lambda \in \pi_{[P_{\boldsymbol{\eta}}]}} (|\lambda| - 1)! \right] \left[\prod_{si \in \boldsymbol{\eta}} \binom{k_{si}}{k_{si_1}} \right] \frac{B(\beta + \sum_{si \in \boldsymbol{\eta}} k_{si_1}, \beta + \sum_{si \in \boldsymbol{\eta}} k_{si_2})}{B(\beta, \beta)} \frac{B(\mu_1 + S_{\boldsymbol{\eta}}, \mu_2 + S_{>\boldsymbol{\eta}})}{B(\mu_1, \mu_2)} \right] \tag{3.15}$$

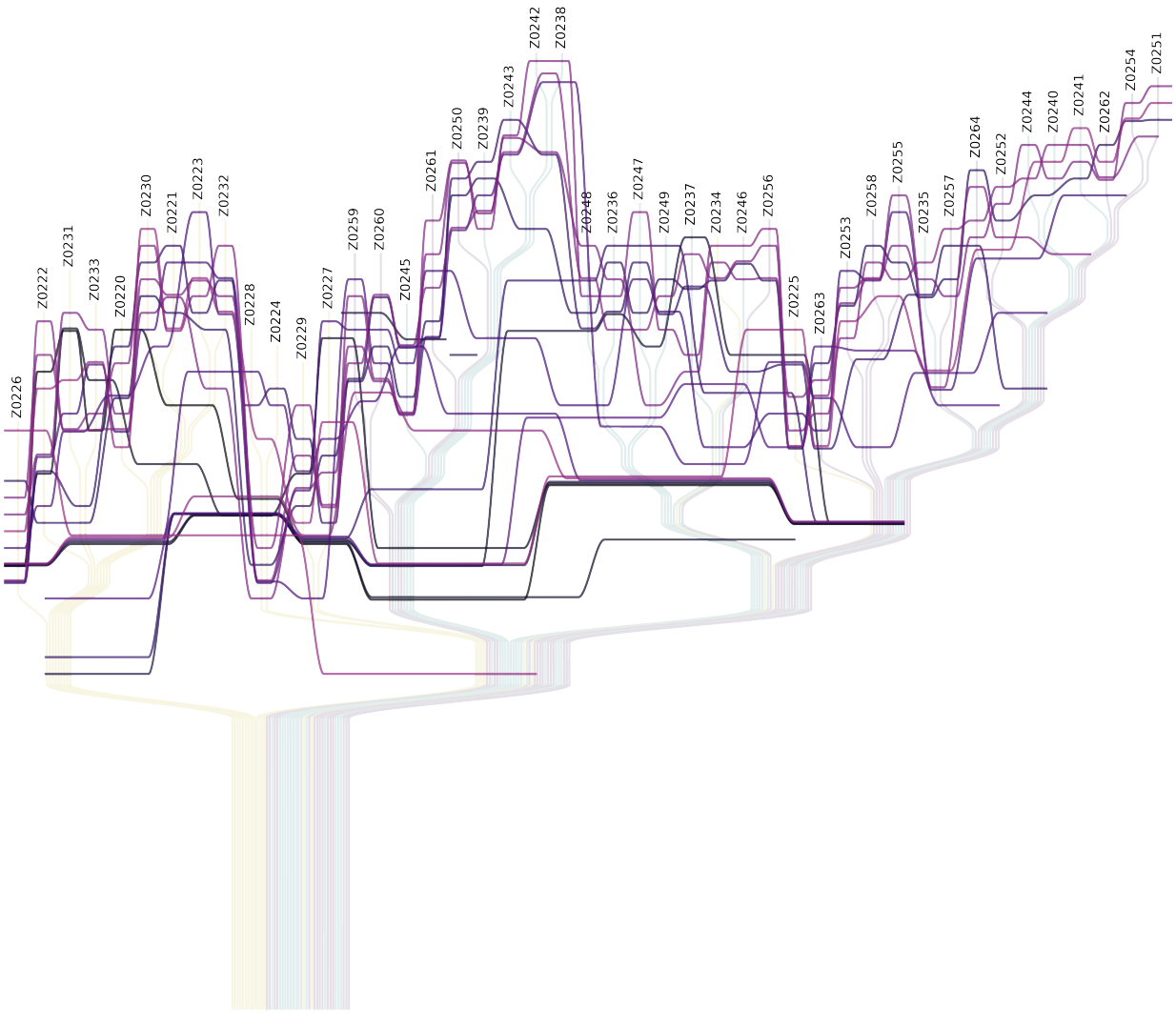


Figure 3.5: Example of the distribution of slices in a realization of the pl-nHDP from a real dataset of 45 zebrafish gut microbiome samples. Here only 6% (20) of the total number of slices (319) are shown. Already we can see the flexibility and complexity captured by the pl-nhDP. The darker the slice coloring, the closer its associated clade is to the root of the phylogenetic tree. This particular choice of slices in this figure is not arbitrary. It represents one candidate set of putative ecological microbial units which we discover as explained in the next section.

where $\boldsymbol{\eta}$ are the nodes in the hierarchical tree, $\{\boldsymbol{\eta}^s\}$ the set of sample paths, $\{\boldsymbol{U}^i\}$ the set of clade slices, $\{\pi_{[P_{\boldsymbol{\eta}}]}\}$ the set of partitions in the nCRP, $P_{\boldsymbol{\eta}}$ the number of sample paths crossing node $\boldsymbol{\eta}$, the partition over paths crossing node $\boldsymbol{\eta}$, $\{k_{si_1}, k_{si_2}\}$ the set of binomial draws for sample s and slice i in the phylogenetic abundance table, $si \in \boldsymbol{\eta}$ the sample path and clade slice indices intersecting at $\boldsymbol{\eta}$, and $S_{\boldsymbol{\eta}}$ and $S_{>\boldsymbol{\eta}}$ the number of slices at and above node $\boldsymbol{\eta}$.

Using this expression we can find the probability of various types of moves. The first type of move consists in first removing a clade slice $\boldsymbol{U}^j$ (namely the set of its on-off stochastic switches) and all the abundance splits $\{k_{sj_1}, k_{sj_2}\}_s$ at that clade across samples s and then reinserting it somewhere else in the hierarchy. Given a new position U^j,

$$P(\boldsymbol{U}^j|\{k_{si_1}, k_{si_2}\}, \{\pi_{\boldsymbol{\eta}}\}, \{\boldsymbol{\eta}^s\}, \{\boldsymbol{U}^i\}_{i\neq j}) = \left[\prod_{\boldsymbol{\eta}\in\boldsymbol{U}^j_+} \frac{\mu_1 + S^{-j}_{\boldsymbol{\eta}}}{\mu_1+\mu_2+S^{-j}_{\geq\eta}}\right]\left[\prod_{\boldsymbol{\eta}<\boldsymbol{U}^j_+} \frac{\mu_2 + S^{-j}_{>\boldsymbol{\eta}}}{\mu_1+\mu_2+S^{-j}_{\geq\boldsymbol{\eta}}}\right]$$
$$\times \left[\prod_s \binom{k_{sj}}{k_{sj_1}}\right]\left[\prod_{\boldsymbol{\eta}\in\boldsymbol{U}^j_+} \frac{B(\beta + \sum_{s\in\boldsymbol{\eta}} k_{sj_1} + \sum_{\substack{si\in\boldsymbol{\eta}\\ i\neq j}} k_{si_1}, \beta + \sum_{s\in\boldsymbol{\eta}} k_{sj_2} + \sum_{\substack{si\in\boldsymbol{\eta}\\ i\neq j}} k_{si_2})}{B(\beta + \sum_{\substack{si\in\boldsymbol{\eta}\\ i\neq j}} k_{si_1}, \beta + \sum_{\substack{si\in\boldsymbol{\eta}\\ i\neq j}} k_{si_2})}\right] \quad (3.16)$$

where $S^{-j}_{\boldsymbol{\eta}}$ means the number of slices at $\boldsymbol{\eta}$ when the slice j is removed from the hierarchy, $\boldsymbol{U}^j_+$ means the set of nodes at which the stochastic switches are on, namely the nodes along the clade slice, and $\boldsymbol{\eta} < U^j_+$ is the set of nodes at which the stochastic switches are off, namely those nodes that sit below the clade slice in the hierarchy. The second type of move consists in removing a sample path $\boldsymbol{\eta}^r$ and all abundance splits $\{k_{ri_1}, k_{ri_2}\}_i$ for this sample across all clades i from the hierarchy and then putting it back somewhere else. Given a new sample path $\boldsymbol{\eta}^r$,

$$P(\boldsymbol{\eta}^r|\{k_{si_1}, k_{si_2}\}, \{\pi_{\boldsymbol{\eta}}\}^{-r}, \{\boldsymbol{\eta}^s\}_{s\neq r}, \{\boldsymbol{U}^i\}) = \left[\prod_{l>0} \frac{\delta[P^{-r}_{\boldsymbol{\eta}^r_l}=0]\alpha + \delta[P^{-r}_{\boldsymbol{\eta}^r_l}>0]P^{-r}_{\boldsymbol{\eta}^r_l}}{\alpha + P^{-r}_{\boldsymbol{\eta}^r_{l-1}}}\right]$$
$$\times \left[\prod_i \binom{k_{ri}}{k_{ri_1}}\right]\left[\prod_{\boldsymbol{\eta}\in\boldsymbol{\eta}^s} \frac{B(\beta + \sum_{i\in\boldsymbol{\eta}} k_{ri_1} + \sum_{\substack{si\in\boldsymbol{\eta}\\ s\neq r}} k_{si_1}, \beta + \sum_{i\in\boldsymbol{\eta}} k_{ri_2} + \sum_{\substack{si\in\boldsymbol{\eta}\\ s\neq r}} k_{si_2})}{B(\beta + \sum_{\substack{si\in\boldsymbol{\eta}\\ s\neq r}} k_{si_1}, \beta + \sum_{\substack{si\in\boldsymbol{\eta}\\ s\neq r}} k_{si_2})}\right] \quad (3.17)$$

where $\boldsymbol{\eta}^r_l$ is the l-th node along the sample path $\boldsymbol{\eta}^r$ and $\{\pi_{\boldsymbol{\eta}}\}^{-r}$ and $P^{-r}_{\boldsymbol{\eta}}$ which here mean the set of partition in the hierarchy and the number of sample paths at node $\boldsymbol{\eta}$ when the sample path r is removed.

3.3 Discovery of Ecological Microbial Units

The discovery of the fundamental units of microbial diversity has changed substantially since the advent of the genomic era, especially in the way it is increasingly understood to be fundamentally different from the concept of species demarcation in eukaryotes[121], [122]. Our approach falls in line with previous work on the way to delineate microbial units in term of their ecology[123] rather than the way they cluster in genetic or genomic space. That is, namely, as 'ecotypes'[124], [125]. Ecotypes are defined as populations of organisms occupying the same ecological niche, and so we use what we call the 'signal of a niche' provided by the topology of the hierarchical tree as a baseline to discover what we call simply 'ecological microbial units'.

With a maximum *a posteriori* probability (MAP) hierarchy in hand we are now ready to proceed with the discovery of ecological microbial units. The set of nodes in the hierarchy parsimoniously exhausts

the heterogeneity of responses necessary to generate the phylogenetic abundance table. It organizes this heterogeneity in a hierarchical way where the closer one gets to the root of the hierarchical tree the stronger the data supports associated structures and similarities of responses. Moving towards the leaves, the model decomposes the dataset with increasing specificity and granularity. Many branches are spanned by multiple nodes, and this multiplicity captures the diversity of quantitative responses for which samples along that branch coherently respond the some factor.

Underlying the set of nodes in the hierarchy is the topology of the hierarchy itself, namely the set multifurcations and collapsed branches obtained after forgetting the multiplicity of nodes along them. This topology putatively corresponds to an exhaustive decomposition of the way grouping and subgrouping of samples are associated with particular, often unmeasured environmental factors. Indeed as we will see in the next section, clusters that emerge deep in the hierarchy can be sometimes associated with some important and known, controlled factors, for example an experimental treatment, but this is not so for smaller sub-clusters which might be associated with unspecified and uncontrolled factors. In other words, while the full hierarchy and its nodes exhausts the heterogeneity of responses, its underlying topology exhausts the heterogeneity of both known factors and unknown confounding factors. In this setting, the set of branches in the topology that a clade slice touches acts as stand in for some set of factors to which it responds in some arbitrary but coherent way.

If we accept the meaning of the topology and how clade slices interact with it, we can then ask what are the most 'important' clade slices which together completely cover it. We say that a clade slice *covers* a branch if one of its node sits along that branch. A clade slice can therefore cover several branches in the topology. We expect that the more clade slices we consider, the more they cover the set of all branches. After a certain number of clade is considered one would expect that the set of their slices completely covers the topology. We call this a complete covering. Finding a complete covering is interesting because its set of clades 'senses' the full set of known and confounding factors across all sampled microbiomes.

It is obvious that there is more than one set clades that can produce a complete coverings. If we are to attempt to find one such covering we must do so in a way that is as biologically sound as it is optimal. There are two ways to do so that we believe are more interesting than others. The first one is phylogenetically informed and the other topologically informed. In the phylogenetically informed approach, we rank all clades according to their distance to the root of the phylogenetic tree and therefore from the coarsest to the finest taxonomic level. In the topologically informed approach we are instead interested in prioritizing clades associated with slices that cut deeper in the topology of the hierarchical tree, and so we sort using the mean root distance in the topology of the branches cut by the clade slice. In other words from clades that respond more coherently on average to the underlying factors to those who respond with more granularity.

Then we proceed using a simple greedy approach. We add clade slices one by one to the covering following either the phylogenetic of topological sorting, each time eliminating the new branches the newly added clade slice covers. We also remain conservative whereby we skip clades slices that do not cover any new branches that were not already covered. Once we have a complete covering we stop. The set of daughter clades of the clades in the complete covering we take to be the 'ecological microbial units'. We must use the daughter clades because the heterogeneity of responses at one clade of the phylogenetic abundance table controls the way the abundance splits between its two daughter clades. In other words we need the daughter clades in order to see the response.

3.4 Results

3.4.1 Datasets

We use two datasets to showcase our method. The first one, the zebrafish gut dataset under diets with different levels of zinc, serves as a test-bench. The second one, the TARA Oceans expedition dataset, we hope will better highlight the power of our method.

Zebrafish Gut Microbiomes

The zebrafish dataset is constituted of two components: a completely bifurcated phylogenetic tree of 16S OTUs rooted at the {Archea, Bacteria} node, and an abundance table of those OTUs as obtained from the standard QIIME pipeline[36]. The dataset contains samples of gut microbiomes from 45 zebrafishes under three different treatments[126]: 15 fishes were exposed to a standard lab diet control (LDC), 15 were exposed to a defined diet with sufficient levels of zinc (DDC), and 15 where exposed to a defined diet with deficient levels of zinc (CZMD). The abundance table contains 1091 OTUs across 45 samples and therefore the phylogenetic abundance table contains 45×1090 entries as abundance splits $\{k_{si_1}, k_{si_2}\}$. To reduce the size of the dataset we dropped all records for which $k_{si} = k_{si_1} + k_{si_2} < 0.01 \times k_{s,\text{root}}$, namely for which the abundance of the parent clade drops below 1% of the total abundance in the sample s. This reduces the number of represented clades in the phylogenetic table from 1090 to 319. The final total number of abundance splits represented in the phylogenetic abundance table was 5373.

We now apply the pl-nhDP to the zebrafish dataset. The inferred MAP hierarchical tree is shown in Figure 3.6 with sample paths colored according to their associated treatment. One can see that almost all samples under the LDC treatment cluster apart from the two other treatments. The clusters for the DDC and CZMD treatments are a bit more mixed. The cluster where CZMD is mainly represented contains 10 of the 15 CZMD samples, while the DDC samples are dispersed between two sub-clusters together containing 11 of the 15 DDC samples.

Then using the phylogenetically informed discovery of ecological microbial units we find a complete covering of the topology of the hierarchical tree using only 20 clades out of the 319 clades present in the phylogenetic table. We use the phylogenetically informed discovery rather than the topological one simply because it needs fewer clades in the complete covering. We have already shown the set of clade slices with this particular complete covering in Figure 3.5. If our covering uses 20 clade slices then we get 40 ecological microbial units. This is an 8-fold reduction in the number of units compared to the use of collapsed OTUs. We say collapsed because as we mentioned above in preparing this dataset we reduced the number of lineages from 1090 to 319 by dropping abundance splits that fell below a 1% abundance threshold. Figure 3.7 shows the same set of ecological microbial units but this time overlaid on the phylogeny.

TARA Oceans Expedition

The TARA Oceans dataset[19], [20] consists of microbiome samples collected from three different depths at 68 stations across every oceans (except the Arctic Ocean) for a total of 243 samples. These three depths correspond to surface layer samples (SRF, n=113), deep chlorophyll maximum layer samples which for one of them sat at the marine oxygen minimum (DCM, n=74+1), mesopelagic zone samples which sometimes sat at the the marine oxygen minimum zone (MES, n=31 + 11), and finally marine epipelagic mixed layer samples (MIX, n=13). We used the accompanying abundance table of 16S rDNA fragments derived from

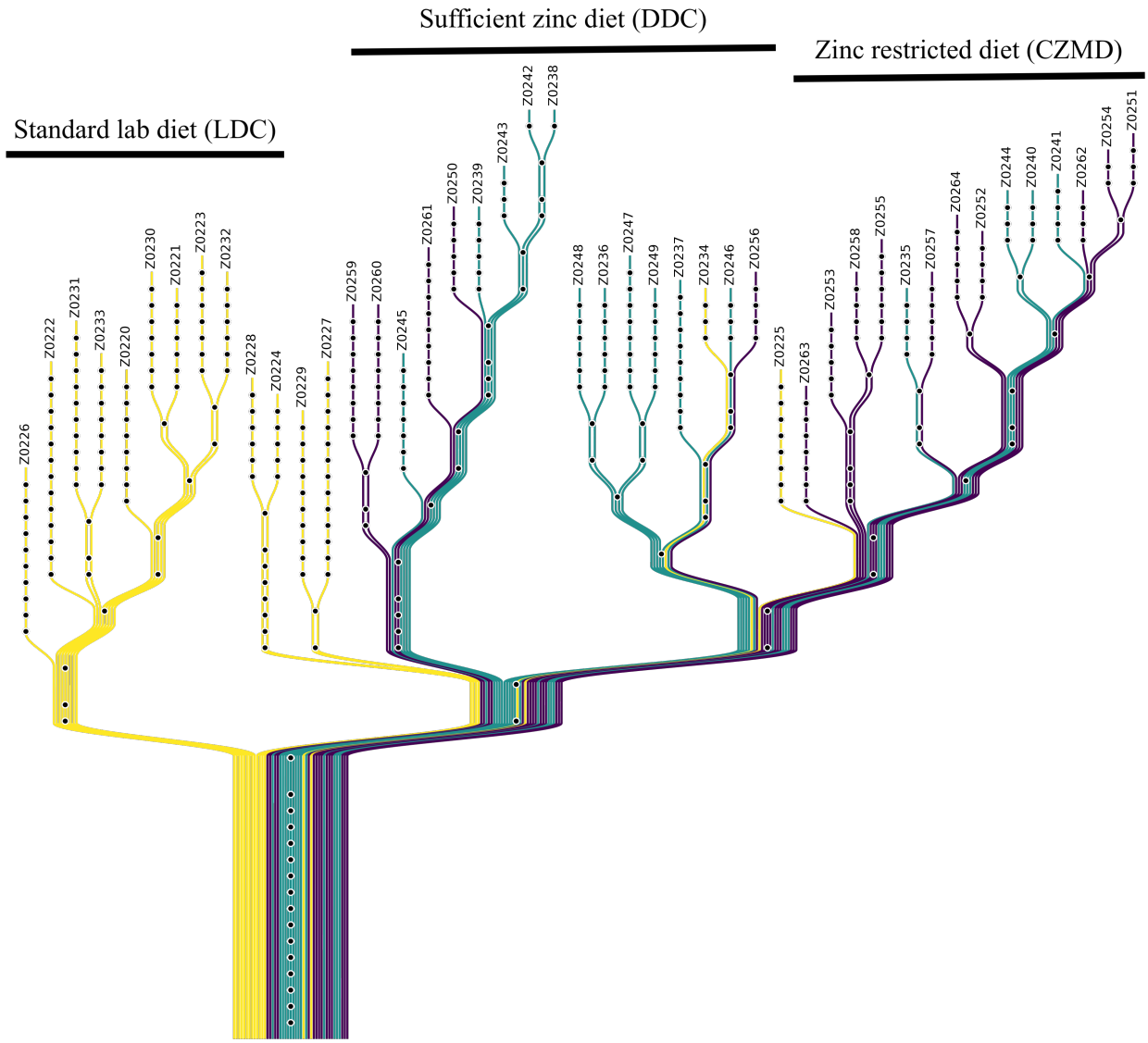


Figure 3.6: MAP hierarchy inferred from the pl-nhDP for the zebrafish dataset. Shown in **yellow** the samples paths under the LDC treatment, in **green** samples paths under the DDC treatment, and in **purple** the samples paths under the CZMD treatment. The **black** dots represent the nodes of the hierarchy at each of which sits a response.

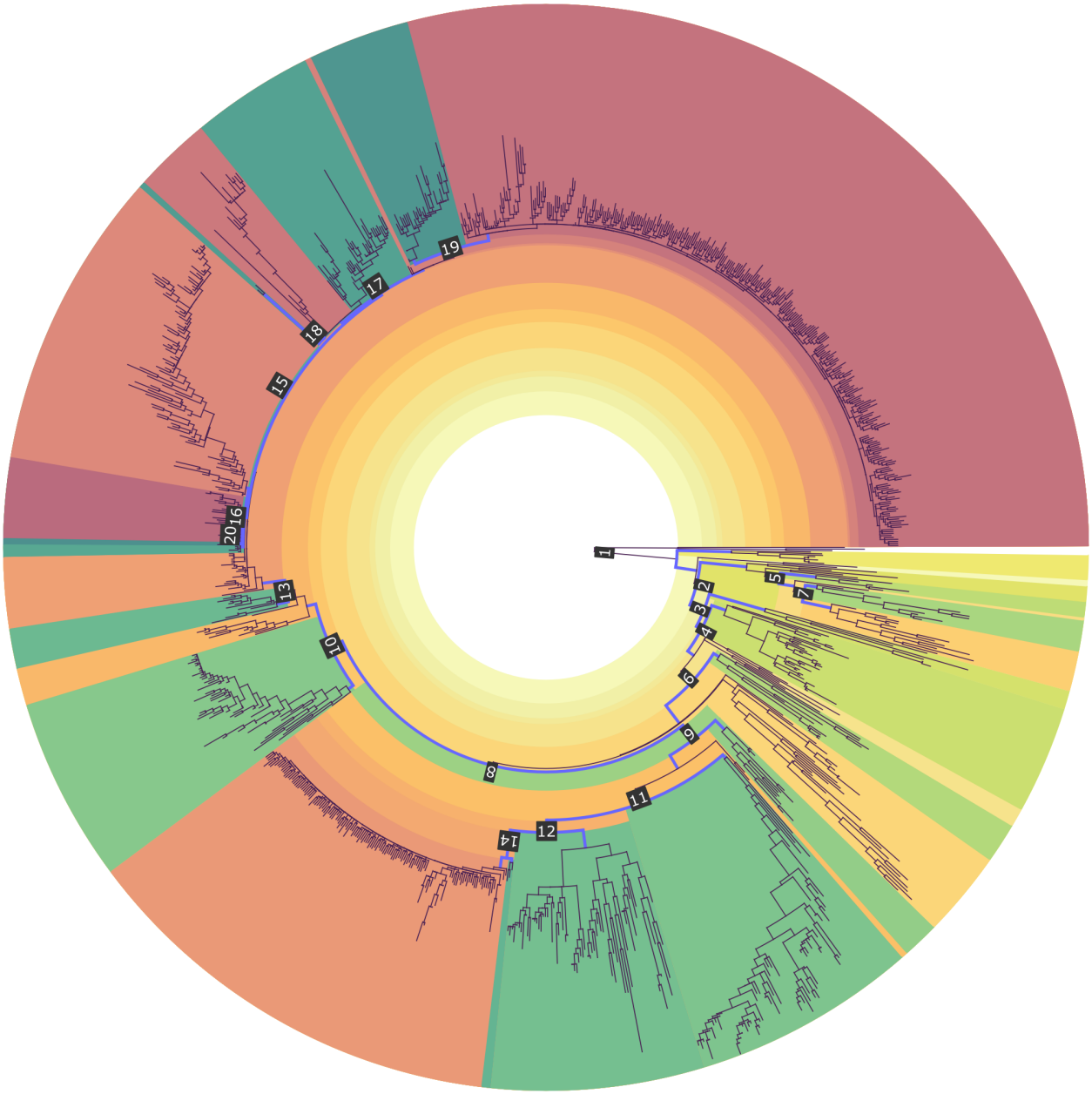


Figure 3.7: Set of 40 putative ecological microbial units displayed on top of the phylogenetic tree for the zebrafish gut microbiome dataset. Those ecological microbial units are the sister units, i.e. daughters of the clades in the complete covering, and are shown in randomly chosen red-green contrasts. The covering rank is indicated on the clades in the complete covering, i.e. at the parent node of the units.

Illumina sequenced metagenomes (mitags)[127] which offer a powerful alternative to 16S OTUs. This table includes 35649 mitags which themselves cover only 66 stations and 139 samples given that not all depths were represented at some stations. For the phylogenetic tree we placed those mitags sequences using SEPP on the 115 release of the SILVA SSU reference tree[128]. The final phylogenetic abundance table across all 139 samples comprises 1942479 abundance splits. To reduce the size of the dataset we drop all abundance splits where $k_{si} < 0.05 \times k_{s,\text{root}}$ and retain only 67007 abundance splits.

We first apply the flat sample-wise probabilistic generative model presented in Section 3.2.2 to the TARA Oceans dataset. Results are shown in Figure 3.8. The method infers five distinct clusters. The MES cluster includes all MES samples plus one DCM sample at station 137 off the coast of Mexico and one MIX sample at station 125 in the middle of the Pacific Ocean near the equator. The SRF cluster contains 47 out of the 63 SRF samples, and 13 out of the 42 DCM samples. The DCM cluster contains 25 out of the 42 DCM samples, 12 out of the 63 SRF sample, and 3 out of the 4 MIX samples. The CH cluster contains the SRF and DCM samples from station 93 off the coast of Chile. Finally the SO cluster contains all SRF and DCM samples in the Scotia sea and the South Ocean. While the stratification along depth of microbiome recapitulates known results[129], two novel clusters (CH and SO) are detected by the method which indicate the presence of two hypothetical microbiome that do not respond more strongly to some unknown factors than the depth stratification.

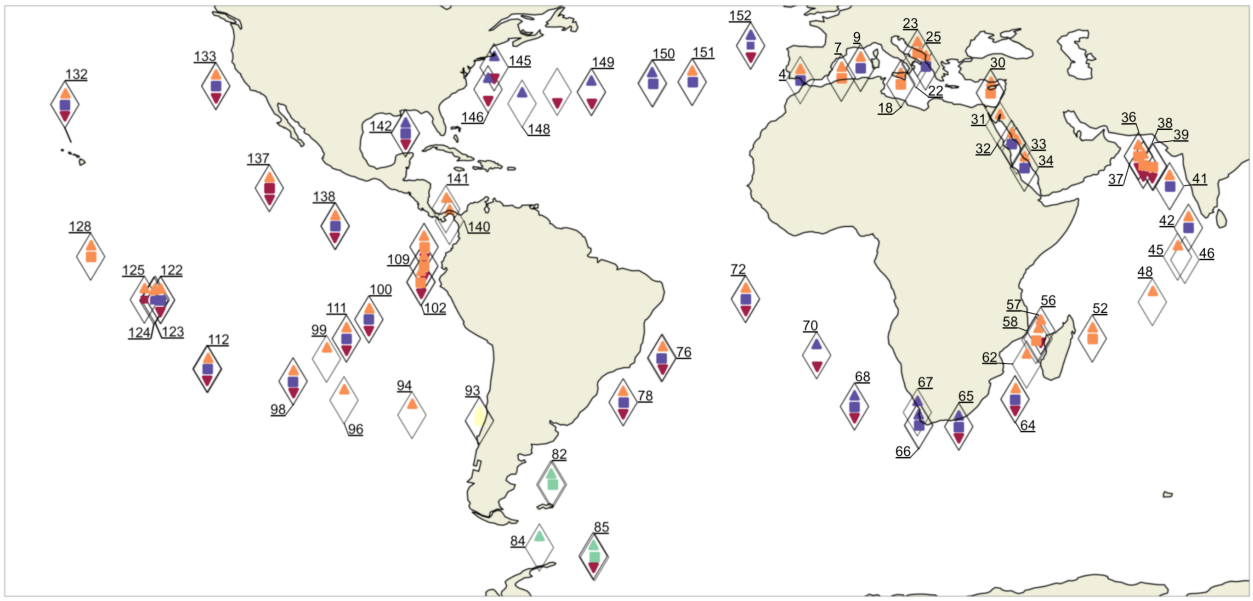


Figure 3.8: Results of the flat sample-wise clustering of the phylogenetic abundance table from the TARA Ocean dataset. Lozenges represent the stations from the the dataset. Two stations could not be correctly labelled because of mistakes in the metadata. In a given lozenge, the top triangle represent the SRF sample, the middle square the MIX/DCM sample, and the bottom triangle the MES sample. The five colors represent 5 different clusters inferred by the method. The MES cluster in **red** contains all MES samples, the SRF cluster in **orange** contains mostly SRF samples, the DCM cluster in **purple** contains mostly DCM samples, the CH cluster in **yellow** includes a unique station on the coast of Chile, and the SO cluster in **green** contains SRF and MIX/DCM samples in the Scotia Sea and the South Ocean.

We now apply the pl-nhDP to the TARA Oceans dataset. The resulting MAP hierarchy is shown in Figure 3.9. Perhaps encouragingly we see once again the emergence of the stratification along depth between SRF, DCM, and MES and the two geographically distinguished clusters CH and SO. The only difference

with Figure 3.8 at the coarsest level of the hierarchy is the appearance of a new cluster (AR/SWM) which combines two geographic locations, namely samples from stations 137 and 138 far off the South-West coast of Mexico, and samples from stations 37, 38, and 39 in the Arabian Sea.

Figure 3.10 show 10 of the 55 clade slices that are the (parents) of the discovered ecological microbial units using the phylogenetically informed method. We stopped at 10 slices because we can already guess how quickly we can get lost in the tangle. In bigger datasets this kind of visualization quickly becomes uninformative. Nonetheless we notice that most slices that cut far down in the hierarchy tend to spend the rest of their time at finer levels. They do not consistently cut deeply. To understand what this mean, Figure 3.11 shows the strength of the response at every node of the hierarchy. Notice how responses are consistently asymmetric deeper in the hierarchy and show more variability towards finer levels. This indicates that the stronger part of the signal picked up by ecological microbial units come from pattern of stark, coherent presence/absence in a select set of samples, and that across the rest of the samples the signal disintegrates into noise.

3.5 Discussion

In this chapter we took inspiration from computational linguistics and created models adapted to biological data to try to visualize and disentangle the heterogeneity inherent to large dataset of microbiome samples collected across multiple environments. The first step was to recognize the importance of the phylogenetic tree as an organizing principle through which we can reformulate abundance data into a set of responses, namely by decomposing abundance tables into phylogenetic abundance contrasts, the abundances split, from deep in the tree of life towards ever finer taxonomic levels.

Using this decomposition we first devised a simple flat sample-wise clustering method which uses the set of all abundance splits across a phylogenetic abundance table for a given sample as a unique, bulk fingerprint that distinguishes sample from each others. The associated probabilistic generative model uses a CRP to cluster these fingerprints into sets of samples with similar bulk responses. We were able to recover 5 clusters in the TARA Oceans dataset, 3 of which recapitulated the known stratification along depth, and 2 of which which seem to suggest the presence of hypothetically unique microbiomes; one in the coast of Chile near the city of Santiago, and the other in the Scotia sea and the South Ocean. While this result is encouraging in and of itself, the generative model does not allow us to look more precisely wherefrom those clusters emerge. It also, by virtue of its flat construction, cannot give us access to more than the strongest niche signal across samples. It is for example doubtful that there could be only 5 niches between which the global ocean microbiome distributes itself. Indeed this flat construction completely ignores the potential for the presence of sub-niches and the multi-dimensionality of niche space. Moreover the use of a bulk fingerprints (one vector of abundance splits across all clades for a given sample) does not lend itself to an analysis of which clades in particular are responsible for the emergence of those clusters. The complementary model is one where each clade is given a unique fingerprint, namely a vector of abundance splits across all samples for that clade. Clustering using this complementary approach cannot produce compatible clusters with those obtained from the sample-wise clustering. The absence of mutually compatible clusters prevents the joint analysis of samples and clade clusters when each of them were undertaken separately.

With this idea of the search for sub-niches in mind and of the need to produce a clustering whereby clades and samples inform each others we introduced a non-parametric hierarchical clustering model called the pl-nhDP. This model uses a double structures, the sample paths and the clade slices. Their points of

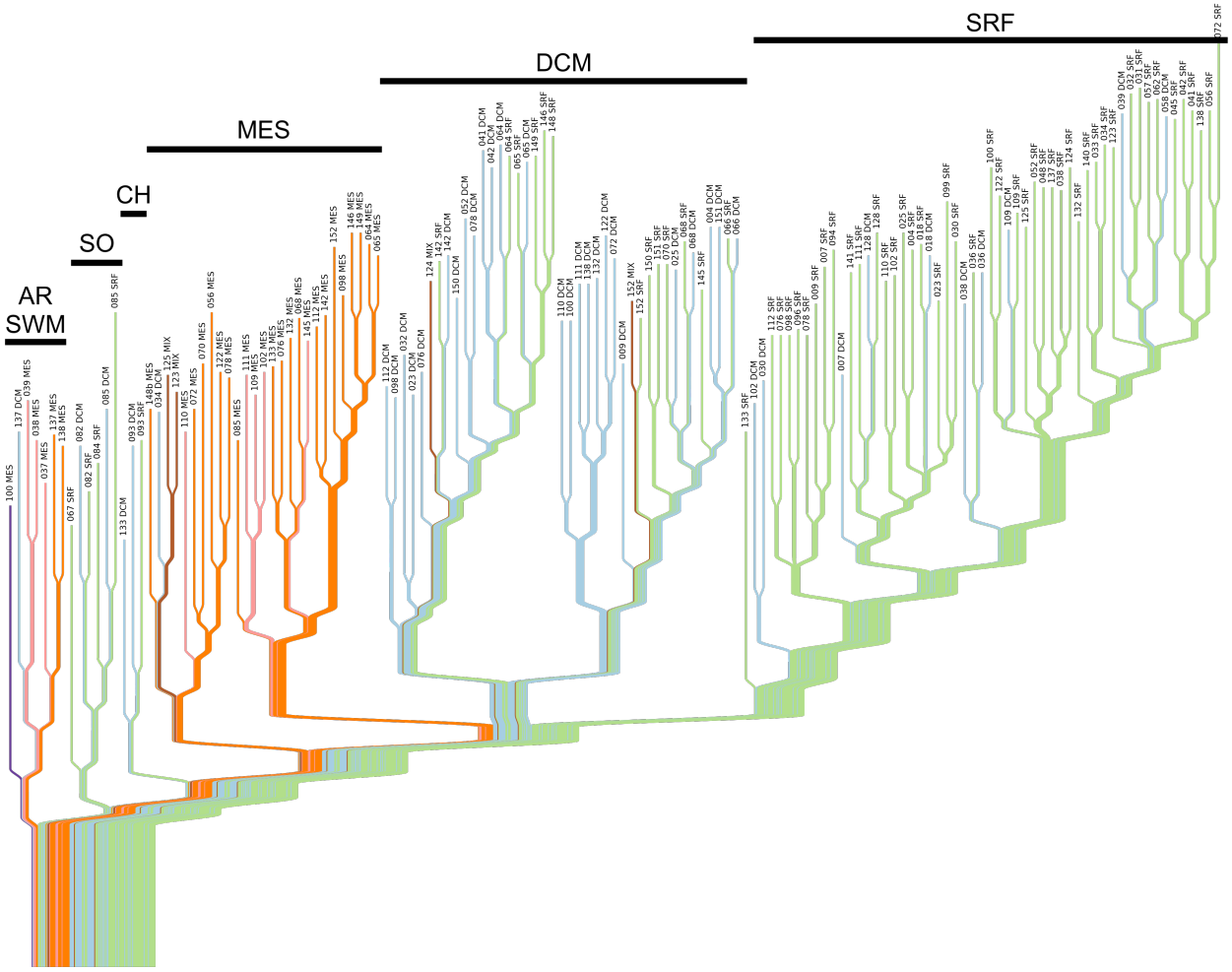


Figure 3.9: MAP hierarchy inferred from the pl-nhDP for the TARA Oceans dataset. Shown in **green** the SRF samples, in **blue** the DCM samples, in **orange** the MES samples, and in **brown** the MIX samples. There are two other types of samples, namely those shown in **pink**, which represented MES samples at the marine oxygen minimum zone, and in **purple** a DCM sample at the marine oxygen minimum zone. Six clusters are shows, three of them named after their main depth constituents (SRF, DCM, MES), and three of them after their location, CH for the coast of Chile, SO for the South Ocean, and AR/SWM for the Arabian sea and two stations far off the South-West coast of Mexico.

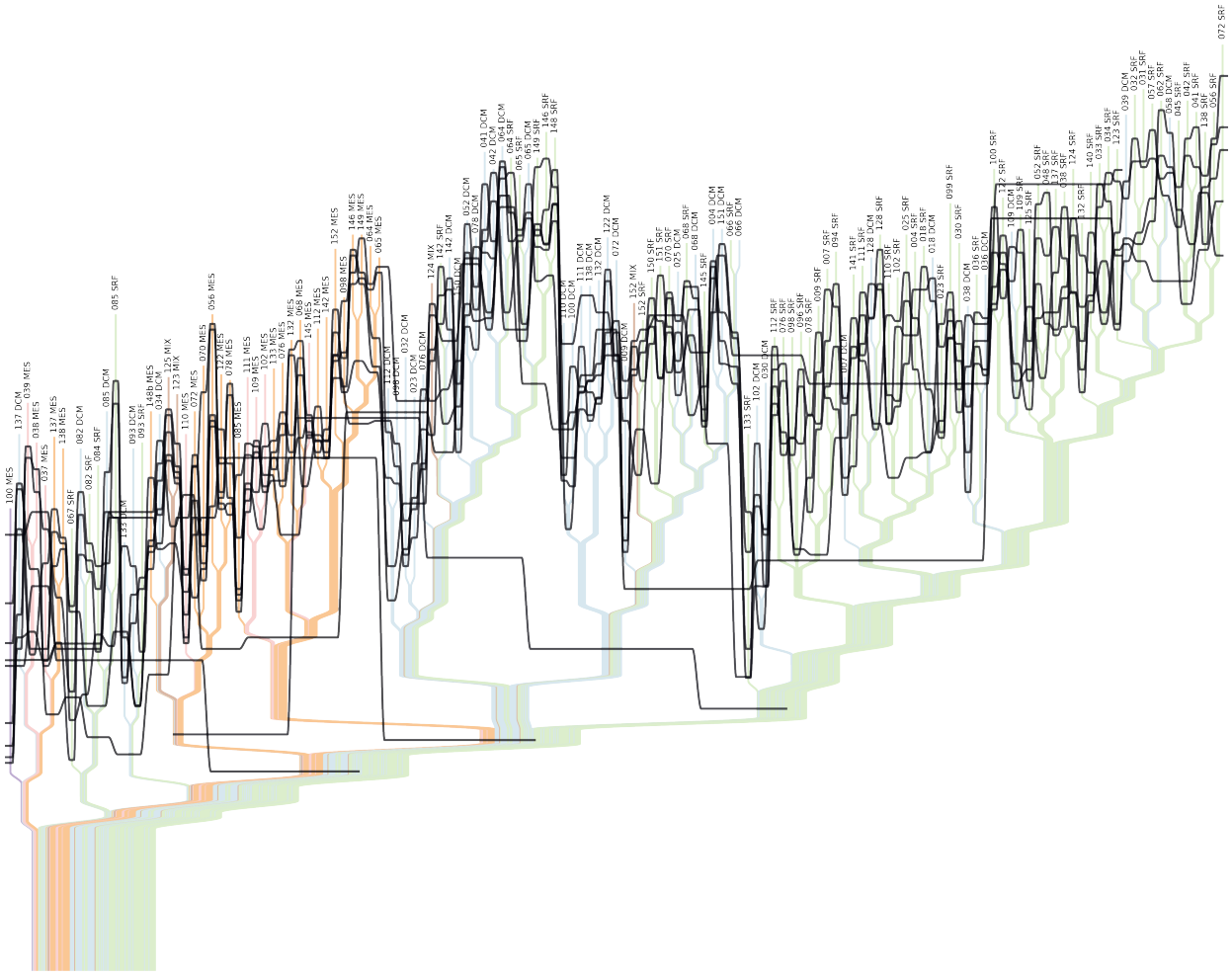


Figure 3.10: Clade slices associated with the (parents of) the first 10 microbial units discovered using the phylogenetically informed method.

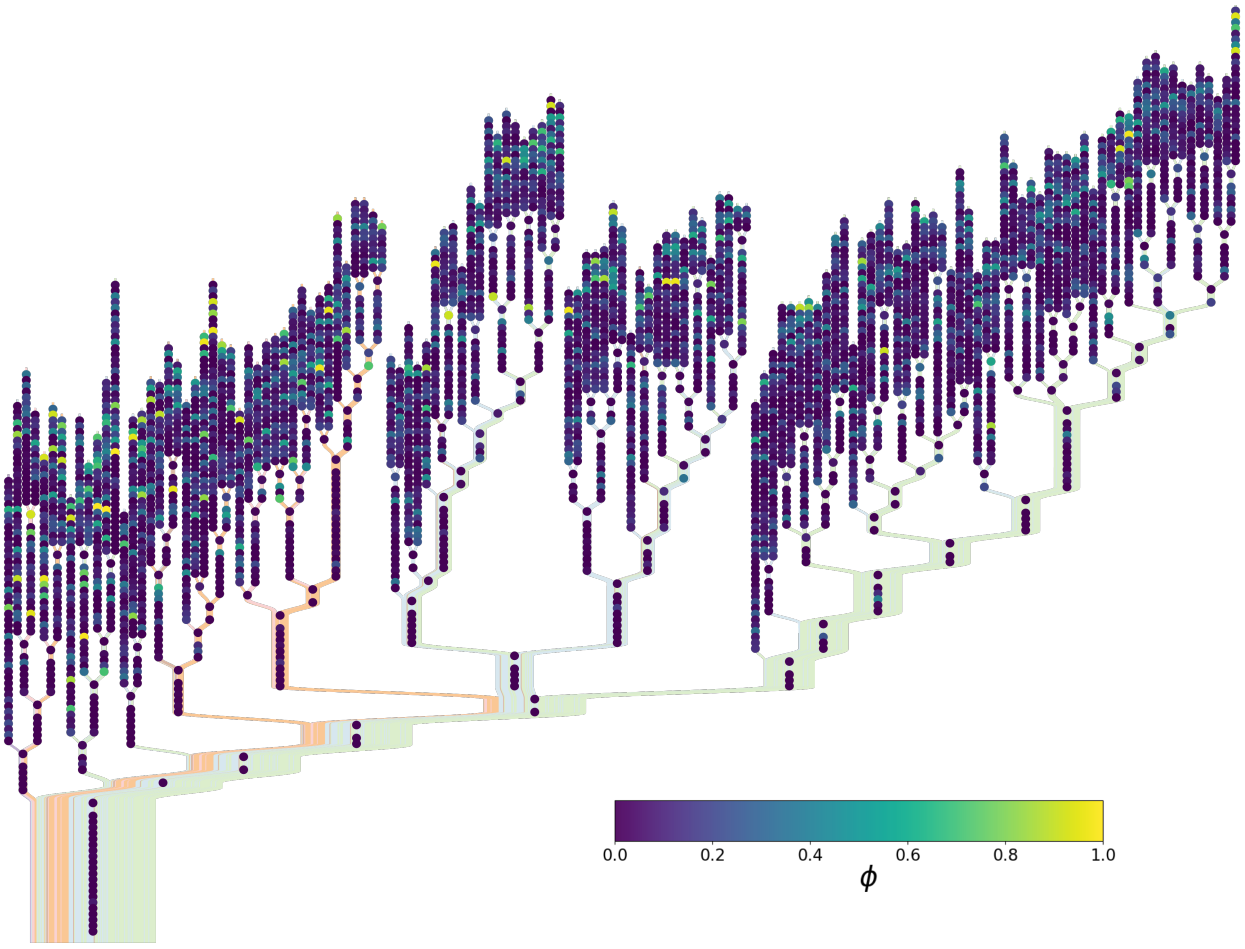


Figure 3.11: Strength of the pool of responses accessible to samples and clades.

intersection create compatible clusters; if many paths intersect many slices at one node of the hierarchy, then all those clades within those samples respond similarly. Indeed clusters are now groupings of sample *and* clades, and not simple grouping of samples *or* clades.

Applied to both the zebrafish gut dataset and the TARA Oceans we were able to recover coarse groupings associated with known factors. In the zebrafish data those were given by the diet treatment (LDC, DDC, and CZMD), and in the TARA Oceans dataset with depth stratification (SRF, DCM, MES) and geographic location (CH, SO, AR/SWM). Those last three clusters indicates the putative present of novel microbiomes driven most strongly by factors other than depth stratification. Whatever they may be they would have to be more informative than mere geography.

More interesting perhaps in the complex structure of branching within those clusters both in the zebrafish dataset and the TARA Oceans dataset. If we allow ourselves to speculate, the topology of this structure suggests what we set out to discover: a vast hierarchy of sub-niches. What we mean by niche here is quite minimal and must remain so if we are to define them in absence of any particular *a priori* knowledge about the environmental features associated with a sample. We define a niche, or rather the 'signal of a niche', as a pool of responses that are coherent for a certain number of samples across a certain number of clades. What exact environmental features correlate with each of those signals we have yet to analyze. We were able to associate a particularly strong signal in the environment to the coarser clusters (diet treatment in the zebrafish dataset and depth and geographic location in the TARA Oceans dataset) but we discovered this signal purely by inspecting the hierarchy after manually overlaying some of the known metadata associated with samples.

We finally used this construction to define what we called 'ecological microbial units'. These units we defined as a collection of sister clades which covers the totality of the topology of the hierarchical tree, or, in a way, that can 'sense' the various known and unknown factors that influence the way abundances split themselves throughout the phylogenetic abundance table. Those units are interesting because they are minimal and proceed from the coarsest towards the finest levels of the underlying phylogenetic tree, and are thus biologically informed. They are useful because if one were to focus on only them and to ignore the rest of the lineages present in a dataset, then the same topological complexity of the hierarchy would be recovered and with it the same niche signals. Ecological microbial units therefore offer a compressed and coarse-grained alternative to more arbitrarily defined OTUs and oligotypes and could help shed a simplifying light on the make-up of microbiomes and how they respond to their environment.

Unfortunately those units are not defined in a unique way. Even using the phylogenetically informed approach there is often a point where the next clade slice at the same depth in the phylogenetic tree does not sense any new information, i.e. any new branch in the hierarchy; the slice is redundant. In our conservative approach we skipped those slices even though there is no definite reason to do so. The less conservative approach keeps all the redundant slices but the flip-side is that the set of microbial units becomes quite large and therefore looses both its usefulness and minimalism. We haven't yet found the right criteria to cut through this conundrum.

3.6 Conclusion

The initial goal of this work was rather modest despite the level of technicality involved in conceptualizing and creating the pl-nhDP model. We wanted to know if there was a strong enough signal in traditional microbial ecology dataset to say something about the presence and heterogeneity of niches underlying an ensemble of

microbiomes without any *a priori* knowledge or information pertaining to them and of the way microbial diversity distributes itself across those hypothetical niches. Our results appear to answer this question in the affirmative. We found that the signal was strong enough to recapitulate known results about major niches in the global ocean microbiome, but also that there is enough signal to uncover a complex sub-structure of sub-niches of unknown origin. We are therefore forced to ask what are they and along which dimensions of environmental features do they organize?

By design the present method cannot answer this question. We say by design because at the core of the method, where the generative model meets the data, the parameters of interest are the (binomial) responses. The reason why a response sitting at a node of the hierarchical tree has its given value is simply that the data supports it. We do not know what it is about the samples with paths that intersect that node that give rise to this response. We only know that certain clades within those samples respond similarly. Tackling this question would require the introduction of substantial additional complexity to the model. We would need to proceed from the inference of responses to the inference of regression coefficients that give rise to those responses. This means that we would need to known how many such regression coefficients are needed and to which environmental factors those regression coefficients are associated. In the world of unsupervised nonparametric model this kind of challenge can be approached using infinite latent feature models like the Indian Buffet Process (IBP)[130]. The prospect of combining an IBP with the pl-nHDP is quite daunting to say the least.

This lead to another question: how deep would our knowledge of the environment itself need to be in order to capture most axes of niche organization? Given that our results suggest the alternative possibility of a hierarchical decomposition of niche space rather than a geometric, hypervolume concept, wouldn't the task become monumental toward ever increasing fine-grain levels as we proceed from the mesoscopic to the microscopic? Indeed we can, or rather we must imagine that niches do not simply align with macroscopic environmental features like temperature, light availability, pH, salinity and so on, but also along a more interacting concept in line with niches associated with the structure of food-webs, consumer-resource relationships, and all the way down to mutualistic, synthorphic, and allelopathic interactions. We want to believe that our method potentially detects signals from individual niches, but it is doubtful that we have yet the capacity, neither technical nor experimental, to fully explore them. The need or temptation to use more complex models, combining e.g. nonparametric clustering and infinite latent feature models like suggested above is in our opinion premature beyond one of technical showmanship.

Chapter 4

Conclusions

The goal of this thesis was to begin to bridge the gap between documenting microbial diversity and understanding the processes that shape it. Keeping in mind the importance of always thinking in terms of scales in ecology and evolution[28], [131], I developed a two pronged approach focusing in turn on the macroscopic and the mesoscopic scale. Doing so we were able to highlight some patterns that appear at those scales.

At the macroscopic level I developed a coarse-grained inference framework to infer the tempo and mode of the global microbial tree of life and used a novel dynamical diversification model called the birth, death, and heterogeneous innovation model which acknowledges and captures the empirical burstiness of microbial phylogenies. The coarse-graining step overcomes the unavoidable problem of incompletely resolved phylogenies common to microbial datasets. Using this approach, I identified two previously unknown universality pattern inherent to microbial diversification. First, we found that the tempo of evolution is such that there is always about one order of magnitude more fast bursty diversification events than slow diversification events. Second, the bursty mode of diversification seems universal across a vast range of different microbiome in that it manifests itself through the appearance of an exponent 1.5 ± 0.2 in the tail of the burst size distribution. This universal pattern suggests that microbes consistently enter new environments in a dramatic and punctuated fashion and therefore that the space of their niches is, at least phenomenologically speaking, rapidly expanded and unbounded[103].

At the mesoscopic level, I developed a generative model together with an unsupervised machine learning algorithm to automatically learn the complex hierarchical structure of factors (but not the factors themselves) that drive the variation of patterns of richness, diversity, and abundances seen across multiple microbiome samples. This model takes inspiration in the field of computational linguistic and topic modeling where topics and their associated "bag-of-words" are automatically discovered within corpuses of documents using hierarchical non-parametric Dirichlet process priors [115], [116]. I adapted these models for microbial data so that they simultaneously learns, in analogy with topic modeling, hierarchical clusters of clades and samples that respond in statistically coherent ways to the various unknown underlying environmental variables or treatments, in other words so that they can capture the 'signal of niches'. It moreover gives a natural way to define contextual microbial units as the minimal set of progressively lower phylogenetic clades that are sufficient to reconstruct its topology, or in other words to 'sense' the full set of known and unknown environmental factors. Applying the algorithm to microbiomes from the TARA Oceans expeditions I was able to recover the well-known stratification of ocean microbiomes along depth together with three potential novel

microbiomes present in the Scotia Sea and the South Ocean, near the coast of Chile, and in the Arabian see and far off the South-West coast of Mexico in the North Pacific ocean. Although we did find and were able visualize a wealth of potential niche signals the model nonetheless leave us in the dark as to what they are and what characterizes them. Finally, while we were indeed able to create custom visualizations of the complicated hierarchical structure, reminiscent of a tree covered with vines, we are yet to be able to easily understand the full scope of the microbial organization it captures. This deserves further work before our approach can serve as a tool for microbial ecologists.

Below the macroscopic and the mesoscopic lie the microscopic which we left addressed. To explore this realm we must look closely at which processes can give rise to the abundance patterns we observe in microbial datasets. Indeed we need to explain the distribution of responses across clades and samples rather than simply discover and organize it. Towards this goal I have started to develop an inference framework which seeks to recover the complex web of dynamical microbial interactions at play within ensemble of microbiomes once again using the simplifying lens of the evolutionary relationships given by the phylogenetic tree. Doing so will help to shed light on the many relationships of mutualism, syntrophy, and consumer-resource processes, their effects on timescales of hours and days, and how they might betray the signal and character of micro-niches which the previous two chapters hint at. Moreover we must consider that at this scale environmental stochasticity plays an important role. Indeed when we go out and look at microbiomes in the wild using environmental sampling we cannot take those microbiome samples as representation of fixed sets of abundances at equilibrium, but as true snapshots of dynamical systems out-of-equilibrium; nature plucks the cords of microbiomes and it is up to us to hear the music.

Perhaps we must first take a slight step back within the microscopic realm if we want to connect to the macroscopic scale. timescales of days and months are hard to square with those over which diversification events unfold. At the timescale of diversification one must consider that communities themselves evolve but the above approach assumes that the community interaction matrix is fixed yet those interactions must change in time. Instead of seeking fixed interactions to explain a phylogenetic abundance table, it makes more sense to seek to capture how a community starting with a few species end up after many generation with a many more species. This is necessary in order to go beyond the phenomenology of bursts. As it stands our bursts decorating the timetree of microbes are but Yule processes, namely phenomenological processes or pure and fast speciation happening on a very short timescale which leave as a signature a geometric burst size distribution characterized by a single parameter g. Moreover we then compounded this burst process with a beta distribution to accommodate for the possibility of heterogeneous g's and in doing so make the burst size distribution beta-geometrically distributed and characterized by two parameters, α and β. We do not know why those parameters have the value they have and as we mentioned already how they tie to more mechanistic models. This is where we want community-level modeling to make its mark. We need to allow members of the community to evolve and speciate through innovation, horizontal gene transfer, allopatry, etc., but we also want to stay away from trying to exactly capture these processes with intractably complex models. We want instead to capture the essence of their effects.

To be specific, what is the simplest way to capture the effect of evolving community interactions and their reorganization through time at at coarse-grained scale? No only that but what are the degrees of freedom and constraints underlying this reorganization? One of the best prototypical minimal model we can think of is the celebrated Bak-Sneppen model (BS)[132] in which species coevolve on a rugged fitness landscape and in which mutation moving one species towards a different place in the fitness landscape can propagate to its neighbor species with which it interacts in the community. It skips the explicitly consideration of the

precise underlying community dynamics, genetic, geographic, etc., that lead to these differences in fitness and to their correlated changes. It is, in other words, a coarse-grained model of evolution. This model displays coevolutionary avalanches with a size distribution power-law distributed with exponent -1. It is these exact kind of simplified models of ecologies with which we seek to replace our Yule/geometric model of bursts. The exponent of avalanche sizes in the BS model has been shown to vary with the dimensionality of the lattice in which the model unfolds (the original BS model happens on a 1-dimensional lattice). With constraints on the geometry of interactions, for example coming from the average number of accessible mutated metabolisms as dictated by a genetic, phenotypic, biochemical, and metabolic landscape, e.g. see [133]–[136], one can imagine variation on the theme of of the BS model—variations of which we must say are already abundant—that would seek to capture and explain as an emergent property the origin of the parameter we observe in the BDH model.

References

[1] F. W. Preston, "The Commonness, And Rarity, of Species," *Ecology*, vol. 29, no. 3, pp. 254–283, 1948, ISSN: 0012-9658. DOI: 10.2307/1930989.

[2] ——, "The Canonical Distribution of Commonness and Rarity: Part I," *Ecology*, vol. 43, no. 2, pp. 185–215, 1962, ISSN: 0012-9658. DOI: 10.2307/1931976.

[3] ——, "The Canonical Distribution of Commonness and Rarity: Part II," *Ecology*, vol. 43, no. 3, pp. 410–432, 1962, ISSN: 0012-9658. DOI: 10.2307/1933371.

[4] M. L. Rosenzweig *et al.*, *Species diversity in space and time*. Cambridge University Press, 1995.

[5] S. P. Hubbell, *The unified neutral theory of biodiversity and biogeography (MPB-32)*. Princeton University Press, 2011.

[6] R. H. MacArthur and E. O. Wilson, "An Equilibrium Theory of Insular Zoogeography," *Evolution*, vol. 17, no. 4, pp. 373–387, 1963, ISSN: 0014-3820. DOI: 10.2307/2407089. [Online]. Available: https://www.jstor.org/stable/2407089.

[7] E. O. Wilson and R. H. MacArthur, *The theory of island biogeography*. Princeton University Press, 1967.

[8] D. S. Simberloff and E. O. Wilson, "Experimental Zoogeography of Islands: The Colonization of Empty Islands," en, *Ecology*, vol. 50, no. 2, pp. 278–296, 1969, ISSN: 1939-9170. DOI: 10.2307/1934856.

[9] R. M. May, "Patterns of species abundance and diversity," *Ecology and evolution of communities*, pp. 81–120, 1975.

[10] E. C. Pielou, *Ecological diversity*, 574.524018 P5. 1975.

[11] O. Snell, "Die Abhängigkeit des Hirngewichtes von dem Körpergewicht und den geistigen Fähigkeiten," de, *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 23, no. 2, pp. 436–446, Jun. 1892, ISSN: 1433-8491. DOI: 10.1007/BF01843462.

[12] J. A. Thomson, "On Growth and Form," en, *Nature*, vol. 100, no. 2498, pp. 21–22, Sep. 1917, ISSN: 1476-4687. DOI: 10.1038/100021a0.

[13] J. Huxley, *Problems of relative growth*, ser. Dover books on the biological sciences. Dover, 1972, ISBN: 978-0-486-61114-3. [Online]. Available: https://books.google.com/books?id=fgApAQAAMAAJ.

[14] G. B. West, J. H. Brown, and B. J. Enquist, "A General Model for the Origin of Allometric Scaling Laws in Biology," *Science*, vol. 276, no. 5309, pp. 122–126, Apr. 1997. DOI: 10.1126/science.276.5309.122.

[15] R. M. May, "Will a Large Complex System be Stable?" en, *Nature*, vol. 238, no. 5364, pp. 413–414, Aug. 1972, ISSN: 1476-4687. DOI: 10.1038/238413a0.

[16] R. May, "Stability and complexity in model ecosystems. princeton, new jersey," *Princeton Univ. Press*, vol. 7, pp. 1415–1419, 1973.

[17] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The Human Microbiome Project," en, *Nature*, vol. 449, no. 7164, pp. 804–810, Oct. 2007, ISSN: 1476-4687. DOI: 10.1038/nature06244. [Online]. Available: https://www.nature.com/articles/nature06244.

[18] L. Amaral-Zettler, L. F. Artigas, J. Baross, L. Bharathi, A. Boetius, D. Chandramohan, G. Herndl, K. Kogure, P. Neal, C. Pedrós-Alió, A. Ramette, S. Schouten, L. Stal, A. Thessen, J. Leeuw, and M. Sogin, "A global census of marine microbes," in *Life in the World's Oceans: Diversity, Distribution and Abundance*, Blackwell Publishing Ltd, 2010, pp. 223–245.

[19] S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, C. Bowler, C. d. Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, and P. Bork, "Structure and function of the global ocean microbiome," en, *Science*, vol. 348, no. 6237, p. 1 261 359, May 2015, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1261359. [Online]. Available: http://science.sciencemag.org/content/348/6237/1261359.

[20] S. Pesant, F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, D. Iudicone, E. Karsenti, S. Speich, R. Troublé, C. Dimier, S. Searson, and Tara Oceans Consortium Coordinators, "Open science resources for the discovery and analysis of Tara Oceans data," eng, *Scientific Data*, vol. 2, p. 150 023, 2015, ISSN: 2052-4463. DOI: 10.1038/sdata.2015.23.

[21] L. R. Thompson, J. G. Sanders, D. McDonald, A. Amir, J. Ladau, K. J. Locey, R. J. Prill, A. Tripathi, S. M. Gibbons, G. Ackermann, J. A. Navas-Molina, S. Janssen, E. Kopylova, Y. Vázquez-Baeza, A. González, J. T. Morton, S. Mirarab, Z. Zech Xu, L. Jiang, M. F. Haroon, J. Kanbar, Q. Zhu, S. Jin Song, T. Kosciolek, N. A. Bokulich, J. Lefler, C. J. Brislawn, G. Humphrey, S. M. Owens, J. Hampton-Marcell, D. Berg-Lyons, V. McKenzie, N. Fierer, J. A. Fuhrman, A. Clauset, R. L. Stevens, A. Shade, K. S. Pollard, K. D. Goodwin, J. K. Jansson, J. A. Gilbert, and R. Knight, "A communal catalogue reveals Earth's multiscale microbial diversity," en, *Nature*, vol. 551, no. 7681, pp. 457–463, Nov. 2017, ISSN: 1476-4687. DOI: 10.1038/nature24621. [Online]. Available: https://www.nature.com/articles/nature24621.

[22] F. E. Egler, "" Physics Envy" in Ecology," *Bulletin of the Ecological Society of America*, vol. 67, no. 3, pp. 233–235, 1986.

[23] E. P. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences. Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959," en, *Communications on Pure and Applied Mathematics*, vol. 13, no. 1, pp. 1–14, 1960, ISSN: 1097-0312. DOI: 10.1002/cpa.3160130102.

[24] S. Kivelson and S. A. Kivelson, "Defining emergence in physics," en, *npj Quantum Materials*, vol. 1, no. 1, pp. 1–2, Nov. 2016, ISSN: 2397-4648. DOI: 10.1038/npjquantmats.2016.24.

[25] R. Solé and J. Bascompte, *Self-Organization in Complex Ecosystems. (MPB-42)*, en. Princeton University Press, Jan. 2012, ISBN: 978-1-4008-4293-3.

[26] K. G. Wilson, "Renormalization Group and Critical Phenomena. I. Renormalization Group and the Kadanoff Scaling Picture," *Physical Review B*, vol. 4, no. 9, pp. 3174–3183, Nov. 1971. DOI: 10.1103/PhysRevB.4.3174. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.4.3174.

[27] ——, "Renormalization Group and Critical Phenomena. II. Phase-Space Cell Analysis of Critical Behavior," *Physical Review B*, vol. 4, no. 9, pp. 3184–3205, Nov. 1971. DOI: 10.1103/PhysRevB.4.3184. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.4.3184.

[28] S. A. Levin, "The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture," en, *Ecology*, vol. 73, no. 6, pp. 1943–1967, 1992, ISSN: 1939-9170. DOI: https://doi.org/10.2307/1941447.

[29] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The Human Microbiome Project," *Nature*, vol. 449, no. 7164, pp. 804–810, Oct. 2007, ISSN: 0028-0836. DOI: 10.1038/nature06244. [Online]. Available: http://dx.doi.org/10.1038/nature06244.

[30] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. Edwards, "The metagenomics rast server—a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC Bioinformatics*, vol. 9, p. 386, 2008, ISSN: 1471-2105.

[31] P. D. Schloss and J. Handelsman, "Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness," *Applied and Environmental Microbiology*, vol. 71, no. 3, pp. 1501–1506, Mar. 2005. DOI: 10.1128/AEM.71.3.1501-1506.2005. [Online]. Available: https://journals.asm.org/doi/full/10.1128/AEM.71.3.1501-1506.2005.

[32] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. V. Horn, and C. F. Weber, "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities," en, *Applied and Environmental Microbiology*, vol. 75, no. 23, pp. 7537–7541, Dec. 2009, ISSN: 0099-2240, 1098-5336. DOI: 10.1128/AEM.01541-09. [Online]. Available: http://aem.asm.org/content/75/23/7537.

[33] S. M. Huse, D. M. Welch, H. G. Morrison, and M. L. Sogin, "Ironing out the wrinkles in the rare biosphere through improved OTU clustering," en, *Environmental Microbiology*, vol. 12, no. 7, pp. 1889–1898, 2010, ISSN: 1462-2920. DOI: 10.1111/j.1462-2920.2010.02193.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1462-2920.2010.02193.x.

[34] P. D. Schloss, "Reintroducing mothur: 10 Years Later," *Applied and Environmental Microbiology*, vol. 86, no. 2, e02343–19, DOI: 10.1128/AEM.02343-19. [Online]. Available: https://journals.asm.org/doi/full/10.1128/AEM.02343-19.

[35] J. R. Rideout, Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell, S. M. Gibbons, J. Chase, D. McDonald, A. Gonzalez, A. Robbins-Pianka, J. C. Clemente, J. A. Gilbert, S. M. Huse, H.-W. Zhou, R. Knight, and J. G. Caporaso, "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences," en, *PeerJ*, vol. 2, e545, Aug. 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.545. [Online]. Available: https://peerj.com/articles/545.

[36] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight, "QIIME allows analysis of high-throughput community sequencing data," en, *Nature Methods*, vol. 7, no. 5, pp. 335–336, May 2010, ISSN: 1548-7091.

[37] A. Amir, D. McDonald, J. A. Navas-Molina, E. Kopylova, J. T. Morton, Z. Z. Xu, E. P. Kightley, L. R. Thompson, E. R. Hyde, A. Gonzalez, and R. Knight, "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns," en, *mSystems*, vol. 2, no. 2, Apr. 2017, ISSN: 2379-5077. DOI: 10.1128/mSystems.00191-16. [Online]. Available: https://msystems.asm.org/content/2/2/e00191-16.

[38] M. Vellend, "Conceptual Synthesis in Community Ecology," *The Quarterly Review of Biology*, vol. 85, no. 2, pp. 183–206, Jun. 2010, ISSN: 0033-5770. DOI: 10.1086/652373. [Online]. Available: http://www.jstor.org/stable/10.1086/652373.

[39] E. K. Costello, K. Stagaman, L. Dethlefsen, B. J. Bohannan, and D. A. Relman, "The application of ecological theory toward an understanding of the human microbiome," *Science*, vol. 336, no. 6086, pp. 1255–1262, 2012. [Online]. Available: http://www.sciencemag.org/content/336/6086/1255.short.

[40] H. Morlon, T. Parsons, and J. Plotkin, "Reconciling molecular phylogenies with the fossil record," *Proc Natl Acad Sci*, vol. 109, pp. 327–332, 2011.

[41] M. R. May, S. Höhna, and B. R. Moore, "A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary," en, *Methods in Ecology and Evolution*, vol. 7, no. 8, pp. 947–959, Aug. 2016, ISSN: 2041-210X.

[42] S. Höhna, M. R. May, and B. R. Moore, "TESS: An R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates," *Bioinformatics*, vol. 32, no. 5, pp. 789–791, Mar. 2016, ISSN: 1367-4803.

[43] W. P. Maddison, P. E. Midford, S. P. Otto, and T. Oakley, "Estimating a Binary Character's Effect on Speciation and Extinction," *Systematic Biology*, vol. 56, no. 5, pp. 701–710, Oct. 2007, ISSN: 1063-5157.

[44] R. G. FitzJohn, W. P. Maddison, and S. P. Otto, "Estimating Trait-Dependent Speciation and Extinction Rates from Incompletely Resolved Phylogenies," *Systematic Biology*, vol. 58, no. 6, pp. 595–611, Dec. 2009, ISSN: 1063-5157.

[45] R. S. Etienne, B. Haegeman, T. Stadler, T. Aze, P. N. Pearson, A. Purvis, and A. B. Phillimore, "Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record," en, *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 279, no. 1732, pp. 1300–1309, Apr. 2012, ISSN: 0962-8452, 1471-2954.

[46] D. L. Rabosky, "Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees," *PLOS ONE*, vol. 9, no. 2, e89543, Feb. 2014, ISSN: 1932-6203.

[47] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes, "Unifying the epidemiological and evolutionary dynamics of pathogens," *Science*, vol. 303, no. 5656, pp. 327–332, 2004.

[48] R. Neher and O. Hallatschek, "Genealogies of rapidly adapting populations.," *Proc Natl Acad Sci*, vol. 110, pp. 437–442, 2013.

[49] D. Kühnert, T. Stadler, T. G. Vaughan, and A. J. Drummond, "Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model," en, *Journal of The Royal Society Interface*, vol. 11, no. 94, p. 20 131 106, May 2014, ISSN: 1742-5689, 1742-5662.

[50] ——, "Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data," en, *Molecular Biology and Evolution*, vol. 33, no. 8, pp. 2102–2116, Aug. 2016, ISSN: 0737-4038, 1537-1719.

[51] J. O'Dwyer, T. Sharpton, and S. Kembel, "Backbones of Evolutionary History Test Biodiversity Theory in Microbial Communities," *Proc Natl Acad Sci*, vol. 112, pp. 8356–8361, 2015.

[52] S. J. Gould and N. Eldredge, "Punctuated equilibria: The tempo and mode of evolution reconsidered," *Paleobiology*, pp. 115–151, 1977.

[53] J. Felsenstein, *PHYLIP (phylogeny inference package), version 3.5 c.* Joseph Felsenstein., 1993.

[54] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix," *Molecular Biology and Evolution*, vol. 26, no. 7, pp. 1641–1650, Jul. 2009, ISSN: 0737-4038. DOI: 10.1093/molbev/msp077. [Online]. Available: https://doi.org/10.1093/molbev/msp077.

[55] ——, "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments," en, *PLOS ONE*, vol. 5, no. 3, e9490, Mar. 2010, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0009490. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009490.

[56] A. Stamatakis, "RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, May 2014, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu033. [Online]. Available: https://doi.org/10.1093/bioinformatics/btu033.

[57] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, "RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference," *Bioinformatics*, vol. 35, no. 21, pp. 4453–4455, Nov. 2019, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz305. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz305.

[58] N. Wikström, V. Savolainen, and M. W. Chase, "Evolution of the angiosperms: Calibrating the family tree," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 268, no. 1482, pp. 2211–2220, Nov. 2001. DOI: 10.1098/rspb.2001.1782. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspb.2001.1782.

[59] S. Hedges, *Molecular evidence for the early history of living vertebrates. pe ahlberg, ed. major events in early vertebrate evolution: Paleontology, phylogeny, genetics and development 119*, 2001.

[60] S. B. Hedges and S. Kumar, *The Timetree of Life*, en. OUP Oxford, Apr. 2009, ISBN: 978-0-19-156015-6.

[61] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut, "Bayesian Phylogenetics with BEAUti and the BEAST 1.7," *Molecular Biology and Evolution*, vol. 29, no. 8, pp. 1969–1973, Aug. 2012, ISSN: 0737-4038. DOI: 10.1093/molbev/mss075. [Online]. Available: https://doi.org/10.1093/molbev/mss075.

[62] M. J. Sanderson, "R8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock," *Bioinformatics*, vol. 19, no. 2, pp. 301–302, Jan. 2003, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/19.2.301. [Online]. Available: https://doi.org/10.1093/bioinformatics/19.2.301.

[63] T. Britton, B. Oxelman, A. Vinnersten, and K. Bremer, "Phylogenetic dating with confidence intervals using mean path lengths," *Molecular Phylogenetics and Evolution*, vol. 24, no. 1, pp. 58–65, Jul. 2002, ISSN: 1055-7903. DOI: 10.1016/S1055-7903(02)00268-3. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1055790302002683.

[64] T. Britton, C. L. Anderson, D. Jacquet, S. Lundqvist, and K. Bremer, "Estimating Divergence Times in Large Phylogenetic Trees," en, *Systematic Biology*, vol. 56, no. 5, pp. 741–752, Oct. 2007, ISSN: 1063-5157, 1076-836X. DOI: 10.1080/10635150701613783. [Online]. Available: http://sysbio.oxfordjournals.org/content/56/5/741.

[65] S. A. Smith and B. C. O'Meara, "treePL: Divergence time estimation using penalized likelihood for large phylogenies," en, *Bioinformatics*, vol. 28, no. 20, pp. 2689–2690, Oct. 2012, ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bts492. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/28/20/2689.

[66] J. Barido-Sottani, T. G. Vaughan, and T. Stadler, "Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth–death model," en, *Journal of The Royal Society Interface*, vol. 15, no. 146, p. 20 180 512, Sep. 2018, ISSN: 1742-5689, 1742-5662. DOI: 10.1098/rsif.2018.0512. [Online]. Available: http://rsif.royalsocietypublishing.org/content/15/146/20180512.

[67] D. Schluter, *The ecology of adaptive radiation*. OUP Oxford, 2000.

[68] P. B. Rainey and M. Travisano, "Adaptive radiation in a heterogeneous environment," *Nature*, vol. 394, no. 6688, pp. 69–72, 1998. [Online]. Available: http://www.nature.com/nature/journal/v394/n6688/abs/394069a0.html.

[69] K. G. Wilson and M. E. Fisher, "Critical Exponents in 3.99 Dimensions," *Physical Review Letters*, vol. 28, no. 4, pp. 240–243, Jan. 1972. DOI: 10.1103/PhysRevLett.28.240. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.28.240.

[70] M. Doi, "Second quantization representation for classical many-particle system," *Journal of Physics A: Mathematical and General*, vol. 9, no. 9, pp. 1465–1477, 1976. [Online]. Available: http://dx.doi.org/10.1088/0305-4470/9/9/008.

[71] P. Grassberger and M. Scheunert, "Fock-Space Methods for Identical Classical Objects," en, *Fortschritte der Physik*, vol. 28, no. 10, pp. 547–578, 1980, ISSN: 1521-3979. DOI: 10.1002/prop.19800281004. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/prop.19800281004.

[72] L. Peliti, "Path integral approach to birth-death processes on a lattice," *Journal de Physique*, vol. 46, no. 9, p. 15, 1985. DOI: http://dx.doi.org/10.1051/jphys:019850046090146900.

[73] V. P. Maslov, *Operational methods*. Mir, 1976.

[74] M. Sasai and P. G. Wolynes, "Stochastic gene expression as a many-body problem," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2374–2379, Mar. 2003. DOI: 10.1073/pnas.2627987100. [Online]. Available: http://www.pnas.org/content/100/5/2374.abstract.

[75] P. J. Dodd and N. M. Ferguson, "A Many-Body Field Theory Approach to Stochastic Models in Population Biology," *PLoS ONE*, vol. 4, no. 9, e6855, 2009. DOI: 10.1371/journal.pone.0006855. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0006855.

[76] D. C. Mattis, "Exact Solution of the Diffusion-Limited Recombination Process $A + B \to \emptyset$," *Modern Physics Letters B*, vol. 11, no. 23, pp. 989–996, 1997. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0217984997001201.

[77] P. Flajolet and R. Sedgewick, *Analytic combinatorics*. cambridge University press, 2009.

[78] W. Gautschi, "A computational procedure for incomplete gamma functions," *ACM Trans. Math. Softw.*, vol. 5, no. 4, pp. 466–481, 1979.

[79] T. Gernhard, "The conditioned reconstructed process," *Journal of Theoretical Biology*, vol. 253, no. 4, pp. 769–778, Aug. 2008, ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2008.04.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022519308001811.

[80] T. Stadler, "On incomplete sampling under birth–death models and connections to the sampling-based coalescent," *Journal of Theoretical Biology*, vol. 261, no. 1, pp. 58–66, Nov. 2009, ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2009.07.018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022519309003300.

[81] H. Morlon, T. L. Parsons, and J. B. Plotkin, "Reconciling molecular phylogenies with the fossil record," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 16 327–16 332, Sep. 2011, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1102543108. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1102543108.

[82] T. Stadler, "How Can We Improve Accuracy of Macroevolutionary Rate Estimates?" en, *Systematic Biology*, vol. 62, no. 2, pp. 321–329, Mar. 2013, ISSN: 1063-5157, 1076-836X. DOI: 10.1093/sysbio/sys073. [Online]. Available: http://sysbio.oxfordjournals.org/content/62/2/321.

[83] J. Lyness and C. Moler, "Numerical Differentiation of Analytic Functions," *SIAM Journal on Numerical Analysis*, vol. 4, no. 2, pp. 202–210, Jun. 1967, ISSN: 0036-1429. DOI: 10.1137/0704019. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/0704019.

[84] J. N. Lyness and G. Sande, "Algorithm 413: ENTCAF and ENTCRE: Evaluation of Normalized Taylor Coefficients of an Analytic Function," *Commun. ACM*, vol. 14, no. 10, pp. 669–675, Oct. 1971, ISSN: 0001-0782. DOI: 10.1145/362759.362820. [Online]. Available: http://doi.acm.org/10.1145/362759.362820.

[85] F. Bornemann, "Accuracy and Stability of Computing High-order Derivatives of Analytic Functions by Cauchy Integrals," en, *Foundations of Computational Mathematics*, vol. 11, no. 1, pp. 1–63, Feb. 2011, ISSN: 1615-3383. DOI: 10.1007/s10208-010-9075-z. [Online]. Available: https://doi.org/10.1007/s10208-010-9075-z.

[86] L. N. Trefethen and J. A. C. Weideman, "The Exponentially Convergent Trapezoidal Rule," en, *SIAM Review*, vol. 56, no. 3, pp. 385–458, Jan. 2014, ISSN: 0036-1445, 1095-7200. DOI: 10.1137/130932132. [Online]. Available: http://epubs.siam.org/doi/10.1137/130932132.

[87] S. Ito and Y. Nakatsukasa, "Stable polefinding and rational least-squares fitting via eigenvalues," en, *Numerische Mathematik*, vol. 139, no. 3, pp. 633–682, Jul. 2018, ISSN: 0945-3245. DOI: 10.1007/s00211-018-0948-4. [Online]. Available: https://doi.org/10.1007/s00211-018-0948-4.

[88] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna, "Perspective: Sloppiness and emergent theories in physics, biology, and beyond," *The Journal of Chemical Physics*, vol. 143, no. 1, p. 010 901, Jul. 2015, ISSN: 0021-9606. DOI: 10.1063/1.4923066. [Online]. Available: https://aip.scitation.org/doi/full/10.1063/1.4923066.

[89] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," en, *Journal of Computational Physics*, vol. 22, no. 4, pp. 403–434, Dec. 1976, ISSN: 0021-9991. DOI: 10.1016/0021-9991(76)90041-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0021999176900413.

[90] ——, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, Dec. 1977, ISSN: 0022-3654. DOI: 10.1021/j100540a008. [Online]. Available: https://doi.org/10.1021/j100540a008.

[91] J. H. McDonald, "G-test of goodness-of-fit," *Handbook of biological statistics*, vol. 3, pp. 53–8, 2014.

[92] R. Arratia and S. DeSalvo, "Probabilistic divide-and-conquer: A new exact simulation method, with integer partitions as an example," *arXiv:1110.3856 [math]*, Nov. 2015. [Online]. Available: http://arxiv.org/abs/1110.3856.

[93] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB," EN, *Applied and Environmental Microbiology*, Jul. 2006. DOI: 10.1128/AEM.03006-05. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/AEM.03006-05.

[94] D. McDonald, M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, and P. Hugenholtz, "An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea," en, *The ISME Journal*, vol. 6, no. 3, pp. 610–618, Mar. 2012, ISSN: 1751-7370. DOI: 10.1038/ismej.2011.139. [Online]. Available: https://www.nature.com/articles/ismej2011139.

[95] S. Mirarab, N. Nguyen, and T. Warnow, "SEPP: SATé-enabled phylogenetic placement," eng, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 247–258, 2012, ISSN: 2335-6936.

[96] S. B. Hedges, J. Marin, M. Suleski, M. Paymer, and S. Kumar, "Tree of Life Reveals Clock-Like Speciation and Diversification," en, *Molecular Biology and Evolution*, vol. 32, no. 4, pp. 835–845, Apr. 2015, ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msv037. [Online]. Available: http://mbe.oxfordjournals.org/content/32/4/835.

[97] J. Marin, F. U. Battistuzzi, A. C. Brown, and S. B. Hedges, "The Timetree of Prokaryotes: New Insights into Their Evolution and Speciation," en, *Molecular Biology and Evolution*, msw245, Dec. 2016, ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msw245. [Online]. Available: http://mbe.oxfordjournals.org/content/early/2017/01/05/molbev.msw245.

[98] M. N. Price, P. S. Dehal, and A. P. Arkin, "Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix," *Molecular biology and evolution*, vol. 26, no. 7, pp. 1641–1650, 2009. [Online]. Available: http://mbe.oxfordjournals.org/content/26/7/1641.short.

[99] ——, "FastTree 2–approximately maximum-likelihood trees for large alignments," *PloS one*, vol. 5, no. 3, e9490, 2010. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009490.

[100] S. Louca and M. W. Pennell, "Why extinction estimates from extant phylogenies are so often zero," en, *Current Biology*, May 2021, ISSN: 0960-9822. DOI: 10.1016/j.cub.2021.04.066. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982221006138.

[101] G. E. Hutchinson, "Concluding Remarks," en, *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 22, pp. 415–427, Jan. 1957, ISSN: 0091-7451, 1943-4456. DOI: 10.1101/SQB.1957.022.01.039. [Online]. Available: http://symposium.cshlp.org/content/22/415.

[102] D. L. Rabosky, "Ecological limits and diversification rate: Alternative paradigms to explain the variation in species richness among clades and regions," en, *Ecology Letters*, vol. 12, no. 8, pp. 735–743, 2009, ISSN: 1461-0248. DOI: 10.1111/j.1461-0248.2009.01333.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1461-0248.2009.01333.x.

[103] L. J. Harmon and S. Harrison, "Species Diversity Is Dynamic and Unbounded at Local and Continental Scales," *The American Naturalist*, vol. 185, no. 5, pp. 584–593, May 2015, ISSN: 0003-0147. DOI: 10.1086/680859. [Online]. Available: https://www.journals.uchicago.edu/doi/full/10.1086/680859.

[104] C. H. Martin and P. C. Wainwright, "Multiple Fitness Peaks on the Adaptive Landscape Drive Adaptive Radiation in the Wild," *Science*, vol. 339, no. 6116, pp. 208–211, Jan. 2013. DOI: 10.1126/science.1227710. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1227710.

[105] Z. D. Blount, C. Z. Borland, and R. E. Lenski, "Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli," en, *Proceedings of the National Academy of Sciences*, vol. 105, no. 23, pp. 7899–7906, Jun. 2008, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0803151105. [Online]. Available: https://www.pnas.org/content/105/23/7899.

[106] K. Sneppen, P. Bak, H. Flyvbjerg, and M. H. Jensen, "Evolution as a self-organized critical phenomenon," en, *Proceedings of the National Academy of Sciences*, vol. 92, no. 11, pp. 5209–5213, May 1995, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.92.11.5209. [Online]. Available: https://www.pnas.org/content/92/11/5209.

[107] R. V. Solé and S. C. Manrubia, "Extinction and self-organized criticality in a model of large-scale evolution," *Physical Review E*, vol. 54, no. 1, R42–R45, Jul. 1996. DOI: 10.1103/PhysRevE.54.R42. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.54.R42.

[108] B. J. McGill, "Towards a unification of unified theories of biodiversity," *Ecology Letters*, vol. 13, no. 5, pp. 627–642, Apr. 2010, ISSN: 1461023X. DOI: 10.1111/j.1461-0248.2010.01449.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1461-0248.2010.01449.x.

[109] D. M. Blei and J. D. Lafferty, "Topic models," in *Text Mining*, Chapman and Hall/CRC, 2009, pp. 101–124.

[110] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012, ISSN: 0001-0782. DOI: 10.1145/2133806.2133826. [Online]. Available: https://doi.org/10.1145/2133806.2133826.

[111] A. Murakami, P. Thompson, S. Hunston, and D. Vajn, "What is this corpus about?: Using topic modelling to explore a specialised corpus," *Corpora*, vol. 12, no. 2, pp. 243–277, Aug. 2017, ISSN: 1749-5032. DOI: 10.3366/cor.2017.0118. [Online]. Available: https://www.euppublishing.com/doi/full/10.3366/cor.2017.0118.

[112] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical Topic Models and the Nested Chinese Restaurant Process," in *Advances in Neural Information Processing Systems*, vol. 16, 2004.

[113] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Advances in neural information processing systems*, 2005, pp. 1385–1392.

[114] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006, ISSN: 0162-1459. DOI: 10.1198/016214506000000302. [Online]. Available: https://doi.org/10.1198/016214506000000302.

[115] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies," *J. ACM*, vol. 57, no. 2, 7:1–7:30, Feb. 2010, ISSN: 0004-5411. DOI: 10.1145/1667053.1667056. [Online]. Available: http://doi.acm.org/10.1145/1667053.1667056.

[116] J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan, "Nested Hierarchical Dirichlet Processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 256–270, Feb. 2015, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2318728.

[117] Z. Ghahramani, M. I. Jordan, and R. P. Adams, "Tree-Structured Stick Breaking for Hierarchical Data," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds., Curran Associates, Inc., 2010, pp. 19–27. [Online]. Available: http://papers.nips.cc/paper/4108-tree-structured-stick-breaking-for-hierarchical-data.pdf.

[118] T. S. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, Mar. 1973, ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176342360.

[119] D. J. Aldous, "Exchangeability and related topics," en, in *École d'Été de Probabilités de Saint-Flour XIII — 1983*, D. J. Aldous, I. A. Ibragimov, J. Jacod, and P. L. Hennequin, Eds., ser. Lecture Notes in Mathematics, Berlin, Heidelberg: Springer, 1985, pp. 1–198, ISBN: 978-3-540-39316-0. DOI: 10.1007/BFb0099421.

[120] R. M. Neal, "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, Jun. 2000, ISSN: 10618600. [Online]. Available: http://www.jstor.org.proxy.uqar.qc.ca/stable/1390653.

[121] K. T. Konstantinidis, A. Ramette, and J. M. Tiedje, "The bacterial species definition in the genomic era," en, *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, no. 1475, pp. 1929–1940, Nov. 2006, ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2006.1920.

[122] F. M. Cohan and E. B. Perry, "A Systematics for Discovering the Fundamental Units of Bacterial Diversity," en, *Current Biology*, vol. 17, no. 10, R373–R386, May 2007, ISSN: 09609822. DOI: 10.1016/j.cub.2007.03.032.

[123] J. I. Prosser, B. J. M. Bohannan, T. P. Curtis, R. J. Ellis, M. K. Firestone, R. P. Freckleton, J. L. Green, L. E. Green, K. Killham, J. J. Lennon, A. M. Osborn, M. Solan, C. J. v. d. Gast, and J. P. W. Young, "The role of ecological theory in microbial ecology," en, *Nature Reviews Microbiology*, vol. 5, no. 5, pp. 384–392, May 2007, ISSN: 1740-1534. DOI: 10.1038/nrmicro1643. [Online]. Available: https://www.nature.com/articles/nrmicro1643.

[124] F. M. Cohan, "What are Bacterial Species?" *Annual Review of Microbiology*, vol. 56, no. 1, pp. 457–487, 2002. DOI: 10.1146/annurev.micro.56.012302.160634.

[125] A. Koeppel, E. B. Perry, J. Sikorski, D. Krizanc, A. Warner, D. M. Ward, A. P. Rooney, E. Brambilla, N. Connor, R. M. Ratcliff, E. Nevo, and F. M. Cohan, "Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics," en, *Proceedings of the National Academy of Sciences*, vol. 105, no. 7, pp. 2504–2509, Feb. 2008, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0712205105. [Online]. Available: http://www.pnas.org/content/105/7/2504.

[126] C. A. Gaulke, L. M. Beaver, C. R. Armour, I. R. Humphreys, C. L. Barton, R. L. Tanguay, E. Ho, and T. J. Sharpton, "An integrated gene catalog of the zebrafish gut microbiome reveals significant homology with mammalian microbiomes," en, Tech. Rep., Jun. 2020, p. 2020.06.15.153924. DOI: 10.1101/2020.06.15.153924. [Online]. Available: https://www.biorxiv.org/content/10.1101/2020.06.15.153924v1.

[127] R. Logares, S. Sunagawa, G. Salazar, F. M. Cornejo-Castillo, I. Ferrera, H. Sarmento, P. Hingamp, H. Ogata, C. de Vargas, G. Lima-Mendez, J. Raes, J. Poulain, O. Jaillon, P. Wincker, S. Kandels-Lewis, E. Karsenti, P. Bork, and S. G. Acinas, "Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities," en, *Environmental Microbiology*, vol. 16, no. 9, pp. 2659–2671, Sep. 2014, ISSN: 1462-2920. DOI: 10.1111/1462-2920.12250. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/1462-2920.12250.

[128] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools," *Nucleic Acids Research*, vol. 41, no. D1, pp. D590–D596, Jan. 2013, ISSN: 0305-1048. DOI: 10.1093/nar/gks1219. [Online]. Available: https://doi.org/10.1093/nar/gks1219.

[129] S. J. Giovannoni and U. Stingl, "Molecular diversity and ecology of microbial plankton," en, *Nature*, vol. 437, no. 7057, pp. 343–348, Sep. 2005, ISSN: 1476-4687. DOI: 10.1038/nature04158. [Online]. Available: https://www.nature.com/articles/nature04158.

[130] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *NIPS*, vol. 18, 2005, pp. 475–482.

[131] J. Chave and S. Levin, "Scale and scaling in ecological and economic systems," *Environmental and Resource Economics*, vol. 26, no. 4, pp. 527–557, 2003. [Online]. Available: http://link.springer.com/article/10.1023/B:EARE.0000007348.42742.49.

[132] P. Bak and K. Sneppen, "Punctuated equilibrium and criticality in a simple model of evolution," *Physical Review Letters*, vol. 71, no. 24, pp. 4083–4086, Dec. 1993. DOI: 10.1103/PhysRevLett.71.4083. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.71.4083.

[133] P. Schuster, "Landscapes and molecular evolution," en, *Physica D: Nonlinear Phenomena*, 16th Annual International Conference of the Center for Nonlinear Studies, vol. 107, no. 2, pp. 351–365, Sep. 1997, ISSN: 0167-2789. DOI: 10.1016/S0167-2789(97)00104-8.

[134] C. Reidys, P. F. Stadler, and P. Schuster, "Generic properties of combinatory maps: Neutral networks of RNA secondary structures," en, *Bulletin of Mathematical Biology*, vol. 59, no. 2, pp. 339–397, Mar. 1997, ISSN: 1522-9602. DOI: 10.1007/BF02462007.

[135] B. M. R. Stadler, P. F. Stadler, G. P. Wagner, and W. Fontana, "The Topology of the Possible: Formal Spaces Underlying Patterns of Evolutionary Change," en, *Journal of Theoretical Biology*, vol. 213, no. 2, pp. 241–274, Nov. 2001, ISSN: 0022-5193. DOI: 10.1006/jtbi.2001.2423.

[136] S. Maslov and K. Sneppen, "Specificity and Stability in Topology of Protein Networks," *Science*, vol. 296, no. 5569, pp. 910–913, May 2002. DOI: 10.1126/science.1065103.

Appendix A

Calibration of the EMP Tree

In this appendix we explain how we transform the EMP phylogenetic tree into a timetree. The short version goes as follows:

- Place the prokaryote sequences from the Timetree project on the greengenes 13.8 core reference tree using SEPP,
- Optimize the greengenes+timetree tree with FastTree constrained by the family-level tree from the Timetree project,
- Place the EMP OTU sequences on the constrained greengenes+timetree reference tree using SEPP,
- Use PATHd8 to ultrametrize the EMP+greengenes+timetree tree using the calibration point given by the Timetree project.

For the first step, we begin by running RAxML 8.2.12 using the following command to obtain the info RAxML info file which is needed for the `pplacer` step of SEPP:

```
raxmlHPC-PTHREADS -f e -m GTRCAT -H --no-bfgs \
                  -n gg_13_8_GTRCAT_for_SEPP
                  -s gg_13_8_99_otus.aligned.fasta \
                  -t gg_13_8_99_otus.tree \
                  -T 32 -p 424242
```

We then reformat it using the `reformat-info.py` helper from SEPP:

```
reformat-info.py RAxML_info.gg_13_8_GTRCAT_for_SEPP \
                 > RAxML_info.gg_13_8_GTRCAT_for_SEPP.reformatted
```

This file contains the MLE of the model parameters that best fit the greengenes reference tree, namely estimates of its branch lengths, GTR matrix, and CAT weights. Moving on, there are 11861 unique Timetree prokaryote sequences. The Timetree family-level calibrated tree contains 102 families (see Figure B.1 and Table B.1). We were unable to map all the prokaryote sequences onto one of those families using their taxonomy. Indeed only 90 of those families were represented in the set of sequences, which trimmed down the number of usable Timetree sequences to 6294 sequences. The families that were dropped are shown in Figure B.1. We do not know why we have this discrepancy. Now we place those 6294 sequences on the greengenes reference tree with SEPP 4.5.1 using the following command:

```
run_sepp.py --fragment bact_arch_tt_2009_remaining6K.unaligned.fasta
            --alignment gg_12_10_treecompatible_aligned.fasta \
            --tree RAxML_result.gg_12_10_GTRCAT_for_SEPP \
            --raxml RAxML_info.gg_12_10_GTRCAT_for_SEPP.reformatted \
            --output tt_on_gg_sepp --outdir sepp_tt_on_gg --tempdir sepp_tmp \
            -x 32 -seed 424242
```

The greengenes 13.8 reference tree is already rooted and therefore we can skip the rooting step. The filename for the placement tree is `tt_on_gg_sepp_placement.tog.tre`.

For the third step we begin by generating a FastTree constraint file file using the Timetree family-level tree decorated with the 6294 prokaryote sequences using the following command:

```
fasttree_constraints_alignment.py bact_arch_tt_2009_remaining6K.nwk \
                                  > timetree_6K_fasttree_constraints
```

This generates a constraint alignment of the following form:

```
Acetobacteraceti_Acetobacteraceae_[...]         01011101110111101111110111111 [...]
Acidianusbrierleyi_Sulfolobaceae_[...]          10111011111111111111111111111 [...]
Methanococcusaeolicus_Methanococcaceae_[...]    10110111011110111110110111111 [...]
```

This alignment has 6294 rows, and the block of 0's and 1's has 178 columns representing 178 constraints. Indeed, an unrooted tree of size N has $2(N-1)$ edges, and therefore the family-level constraint tree gives $2(90-1) = 178$ constraints, each representing a split induced by cutting an edge of the tree. Sequences that fall on one side of a split are given a 0 and those that fall on the other side of the split are given a 1. Then we run FastTree 2.1.11 in double precision using the following command:

```
FastTreeMP_DOUBLE  -constraints timetree_6K_fasttree_constraints \
                   -nt -gtr -fastest \
                   -intree tt_on_gg_sepp_placement.tog.tre \
                   < sepp_tt_on_gg/tt_on_gg_sepp_alignment_masked.fasta \
                   > tt_on_gg.1.nwk
```

We iterate a few times by feeding the tree from step `n` as `intree` for the step `n+1`, i.e

```
FastTreeMP_DOUBLE  -constraints timetree_6K_fasttree_constraints \
                   -nt -gtr -fastest \
                   -intree tt_on_gg.[n-1].nwk \
                   < sepp_tt_on_gg/tt_on_gg_sepp_alignment_masked.fasta \
                   > tt_on_gg.[n].nwk
```

After around 5 iterations, `tt_on_gg.5.nwk` contains a tree which breaks 62 out of the 178 constraints. Iterating once more gives a tree which breaks more constraints and therefore we stop. Then we run a heuristic to try and find the most likely sequences causing those constraints to be broken:

```
constraint_tree = ete3.Tree('bact_arch_tt_2009_remaining6K.nwk', format=1)
test_tree = ete3.Tree('tt_on_gg.5.nwk', format=1)
problematic_sequences = get_problematic_leaf_names(test_tree, constraint_tree)
```

The heuristic identifies 20 problematic sequences. Removing those sequences from both `constraint_tree` and `test_tree` resolves all broken constraints. We verify this using our own algorithm. Then we save a new reference tree `tt_on_gg.noprob.nwk`, alignment `tt_on_gg_alignment_masked.noprob.fasta`, and constraint tree `bact_arch_tt_2009_remaining6K.noprob.nwk` from which we eliminated the problematic sequences. Using FastTree we confirm once more that all constraints are satisfied. We mention here that following the removal of the problematic sequences we did not lose any of the 90 calibration points. Every family retained at least one representative sequence.

For the fourth step, we first recalculate a RAxML info file containing the MLE of model parameters because we changed the greengenes reference tree by placing Timetree sequences on it and by performing 5 iterations of constrained FastTree optimization. The command is as follows:

```
raxmlHPC-PTHREADS -f e -m GTRCAT -H --no-bfgs \
                  -n tt_on_gg_noprob_GTRCAT_for_SEPP
                  -s tt_on_gg_alignment_masked.noprob.fasta \
                  -t tt_on_gg.5.noprob.nwk \
                  -T 32 -p 434343
```

Then we reformat the info file using

```
reformat-info.py RAxML_info.tt_on_gg_noprob_GTRCAT_for_SEPP \
                 > RAxML_info.tt_on_gg_noprob_GTRCAT_for_SEPP.reformatted
```

For the fifth step we simply use SEPP again to place all 8,023,841 EMP sequences on our new reference tree using

```
run_sepp.py --fragment EMP_rep_set_fna
            --alignment tt_on_gg_sepp_alignment_masked.noprob.fasta \
            --tree RAxML_result.gg_13_8_GTRCAT_for_SEPP \
            --raxml RAxML_info.tt_on_gg_noprob_GTRCAT_for_SEPP.reformatted \
            --output emp_on_ttgg --outdir emp_on_ttgg --tempdir sepp_tmp \
            -x 32 -seed 434343
```

Finally, we generate a PATHd8 input file using

```
generate_pathd8_infile(calibration_tree='bact_arch_tt_2009_remaining6K.noprob.nwk',
                       node_age_dict_or_json='timetree_2009_node_ages.json',
                       tree_to_calibrate='emp_on_ttgg_placement.tog.tre',
                       outfile='emp_on_ttgg_pathd8_infile')
```

The file has the following format:

```
Sequence length = 3612;

(((((((200279:0.143462379,((33564:0.051103783,18294:0.107360841):0.09674 [...]

mrca: Thermoplasmaacidophilum_[...], Picrophilusoshimae_[...], fixage=992;
mrca: Anaplasmaphagocytophilum_[...], Aestuariispirainsulae_[...], fixage=2042;
mrca: Alysiellacrassa_[...], Cycloclasticuspugetii_[...], fixage=1993;
```

```
[...]
name of mrca: Thermoplasmaacidophilum_[...], Picrophilusoshimae_[...], name=12b;
name of mrca: Anaplasmaphagocytophilum_[...], Aestuariispirainsulae_[...], name=25;
name of mrca: Alysiellacrassa_[...], Cycloclasticuspugetii_[...], name=26;
[...]
```

The newick string is nothing but the content of `emp_on_ttgg_placement.tog.tre` obtained in the previous step, and then there is a list of pairs of sequence names that have as their MRCA a calibration point given by the calibrated Timetree. We only need one pair for each calibration point. We can now finally ultrametricize the EMP tree using

```
PATHd8 emp_on_ttgg_pathd8_infile emp_on_ttgg_pathd8_outfile
```

and from within `emp_on_ttgg_pathd8_outfile` we extract the EMP timetree.

Appendix B

Calibrated Family-Level Prokaryote TimeTree of Life

In Figure B.1 we reproduce the calibrated prokaryote TimeTree of Life from the Timetree project. The age of the nodes can be found in Table B.1.

	Node #	Age	Node #	Age	Node #	Age	Node #	Age	Node #	Age	Node #	Age
LUCA	0	4200										
Archaea	1b	4193	3b	3594	5b	3313	7b	3093	9b	2430	11b	1676
	2b	4187	4b	3468	6b	3160	8b	2799	10b	2216	12b	992
Bacteria	1	4189	16	2579	31	1834	46	1482	61	1121	76	751
	2	4179	17	2504	32	1806	47	1481	62	1104	77	751
	3	3306	18	2421	33	1775	48	1436	63	1069	78	744
	4	3134	19	2339	34	1753	49	1432	64	1055	79	668
	5	2979	20	2281	35	1753	50	1429	65	1042	80	662
	6	2908	21	2233	36	1747	51	1420	66	1037	81	634
	7	2897	22	2173	37	1673	52	1413	67	1030	82	621
	8	2874	23	2099	38	1653	53	1402	68	1028	83	616
	9	2849	24	2047	39	1621	54	1392	69	1027	84	594
	10	2762	25	2042	40	1620	55	1386	70	950	85	523
	11	2761	26	1993	41	1613	56	1326	71	937	86	509
	12	2739	27	1919	42	1612	57	1306	72	872	87	432
	13	2739	28	1899	43	1579	58	1224	73	871	88	380
	14	2687	29	1860	44	1554	59	1189	74	812		
	15	2607	30	1837	45	1498	60	1180	75	793		

Table B.1: Calibration point from the TimeTree of Life. Node numbers corresponds to those shown in Figure B.1

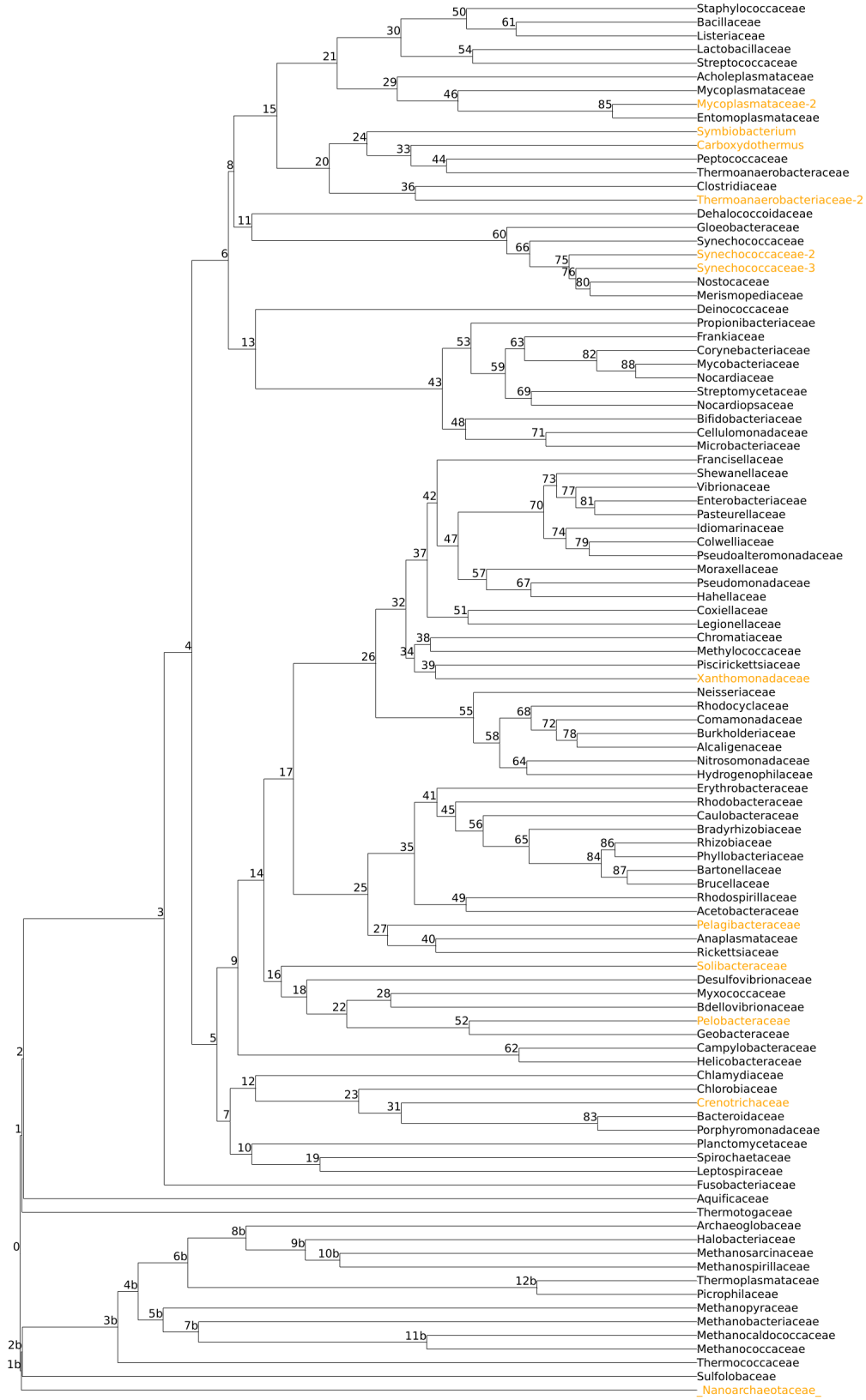


Figure B.1: Calibrated prokaryote TimeTree of Life. Node numbers correspond to calibrated nodes with ages found in Table B.1. In **orange** we highlight the families that are not represented in the set of sequences we obtained from the Timetree Project.

Appendix C

ML Fits for the BDH Model Across Environmental Ontologies

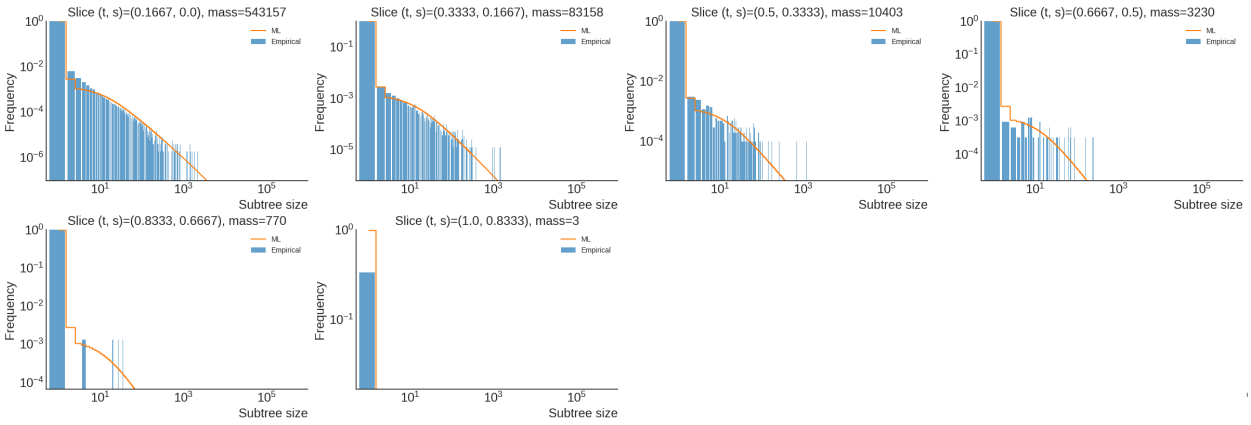


Figure C.1: Aquatic, freshwater biome (ENVO2)

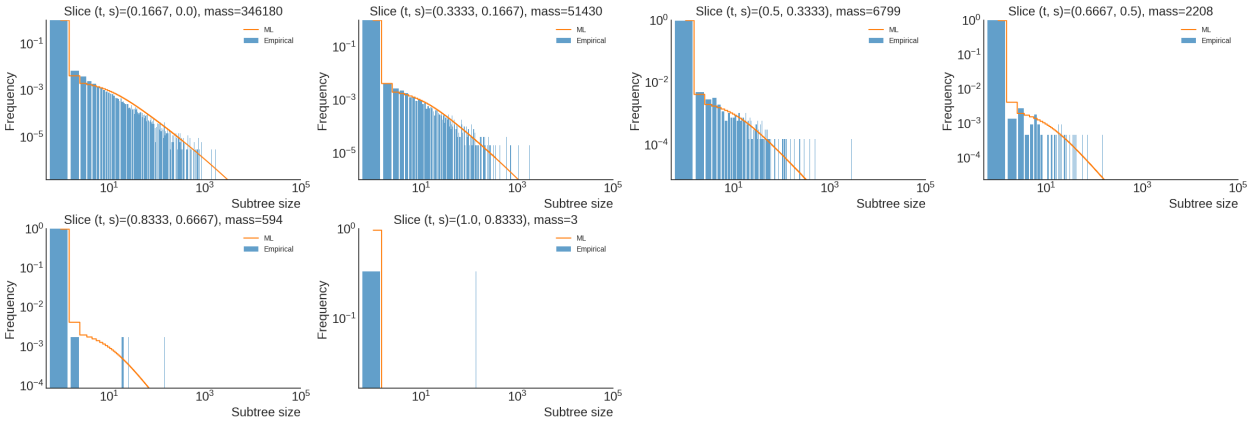


Figure C.2: Aquatic, freshwater lake biome (ENVO3)

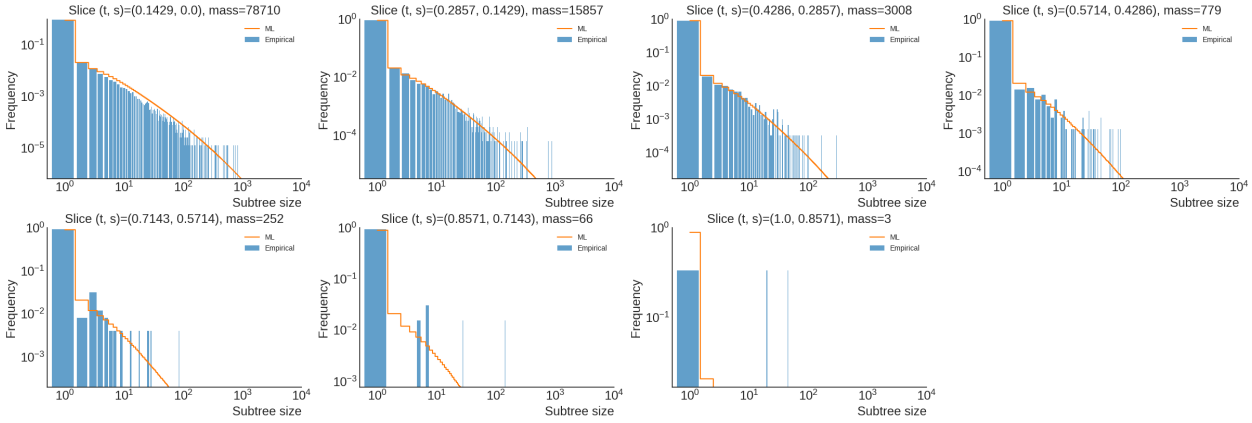


Figure C.3: Aquatic, large freshwater lake biome (ENVO4)

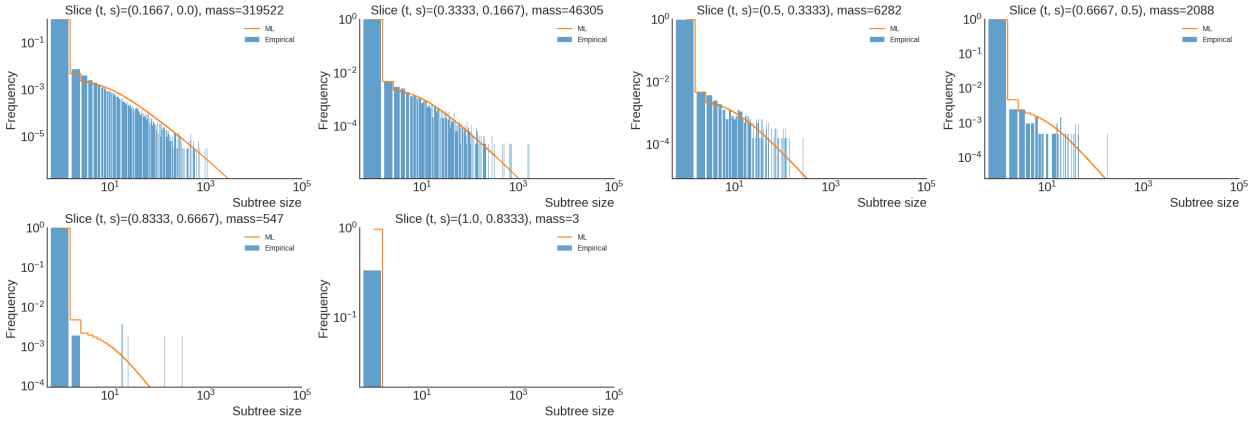


Figure C.4: Aquatic, small freshwater lake biome (ENVO4)

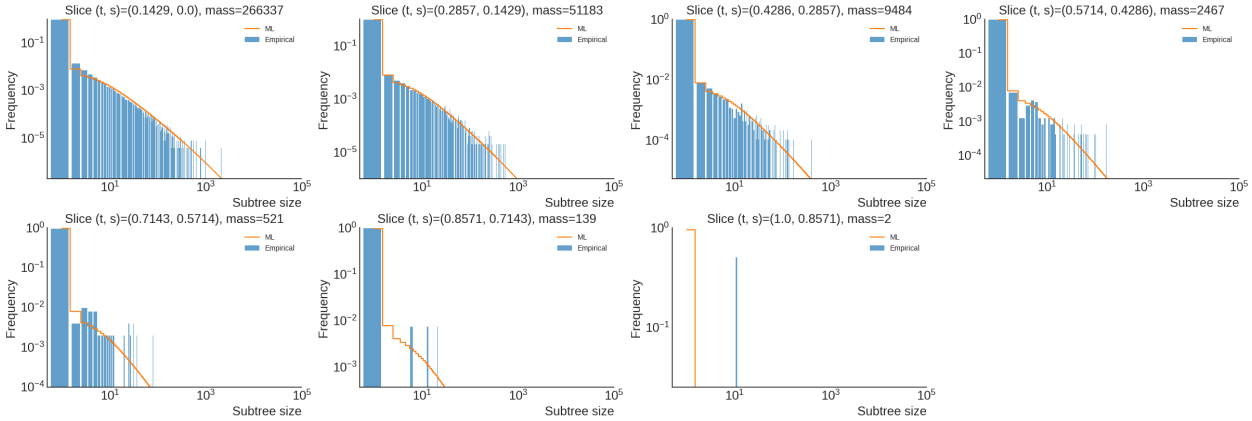


Figure C.5: Aquatic freshwater river biome (ENVO3)

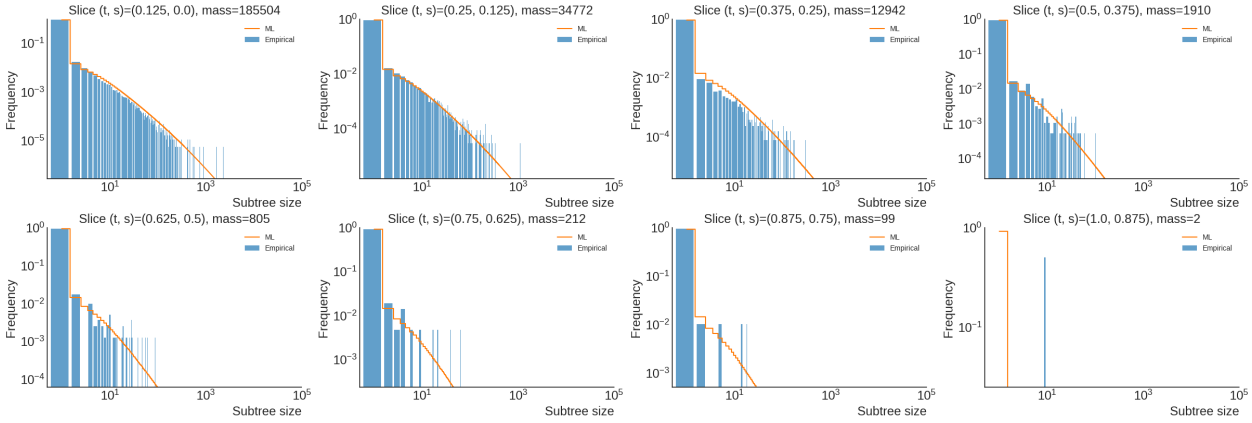


Figure C.6: Aquatic large freshwater river biome (ENVO4)

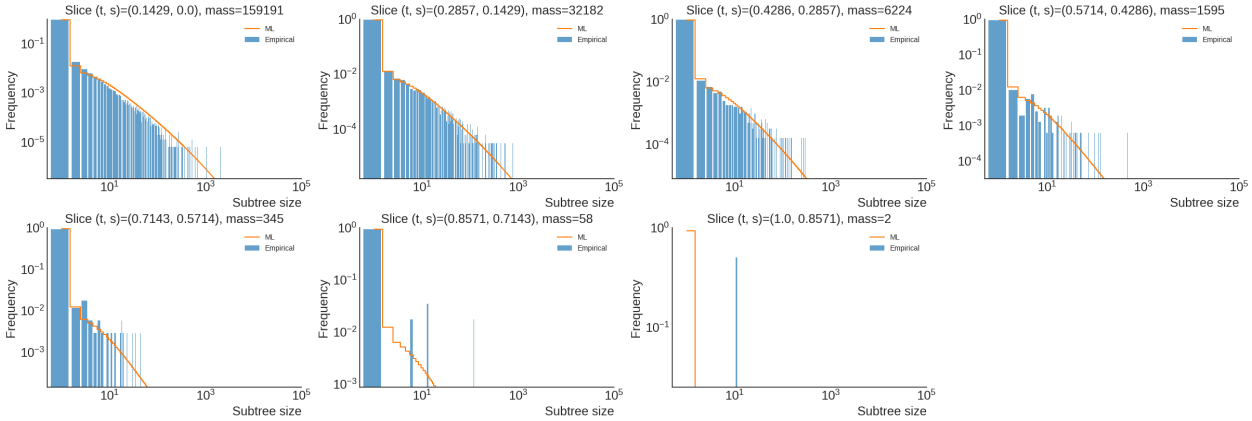


Figure C.7: Aquatic small freshwater river biome (ENVO4)

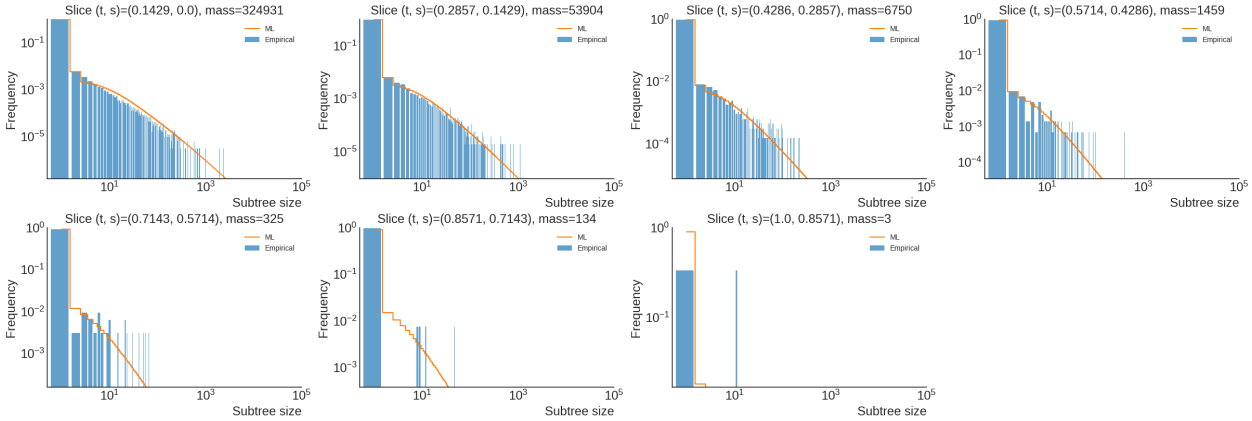


Figure C.8: Aquatic, unspecified freshwater biome (ENVO3)

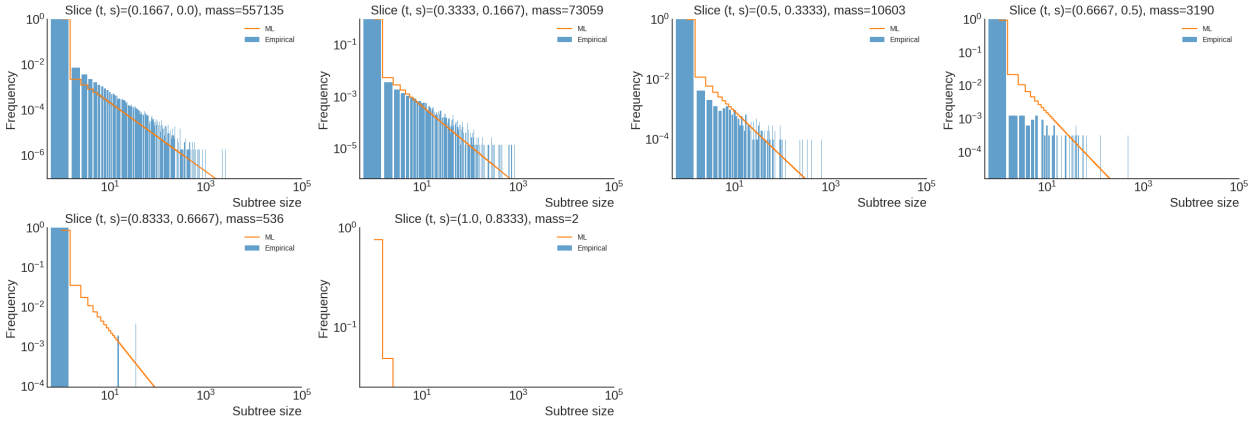


Figure C.9: Aquatic, marine biome (ENVO2)

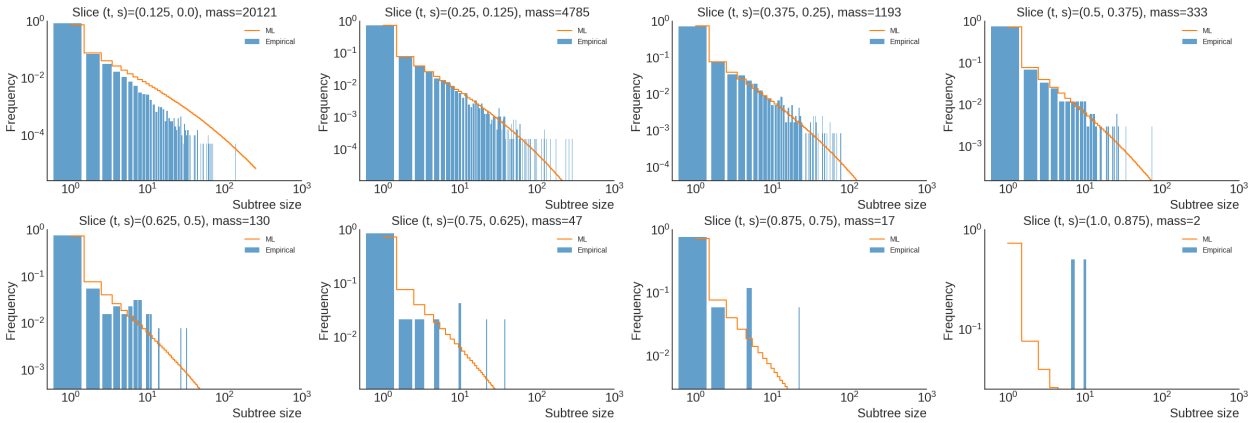


Figure C.10: Aquatic, estuarine marine biome (ENVO3)

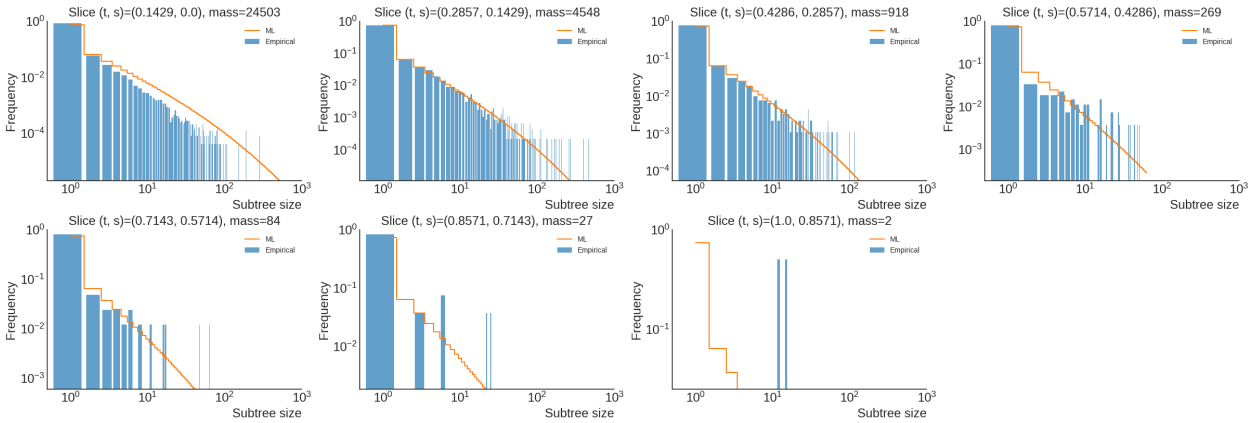


Figure C.11: Aquatic, marine marginal sea biome (ENVO3)

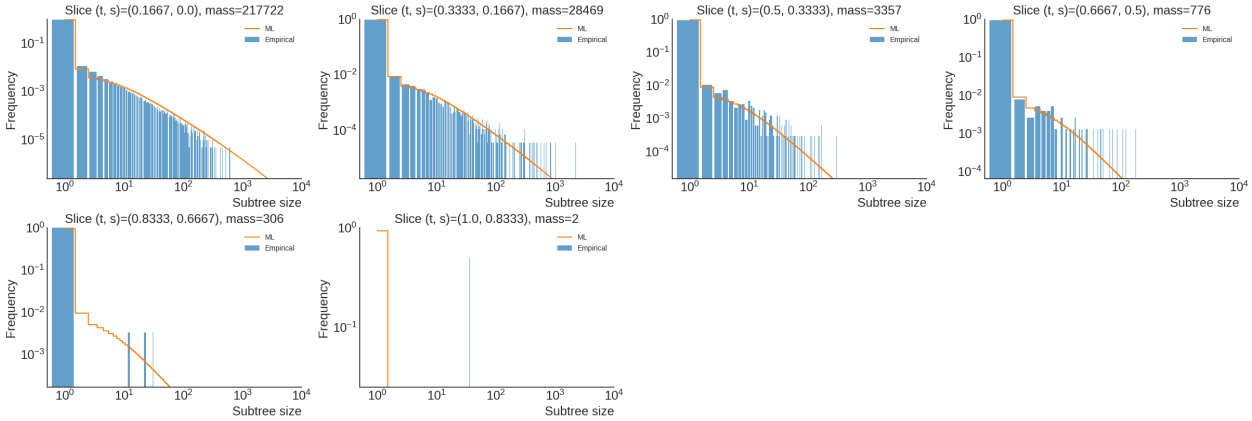


Figure C.12: Aquatic, marine benthic biome (ENVO3)

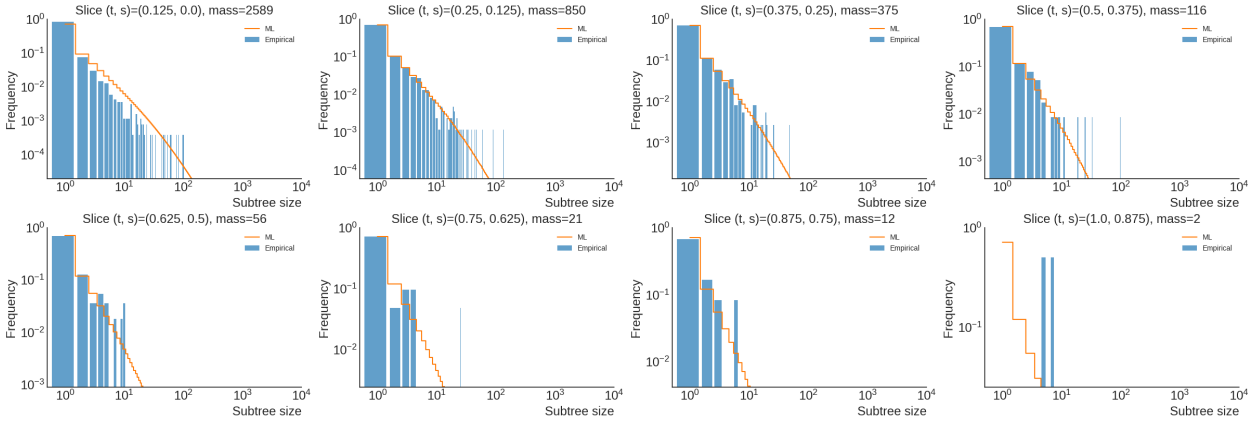


Figure C.13: Aquatic, marine benthic coral reef biome (ENVO4)

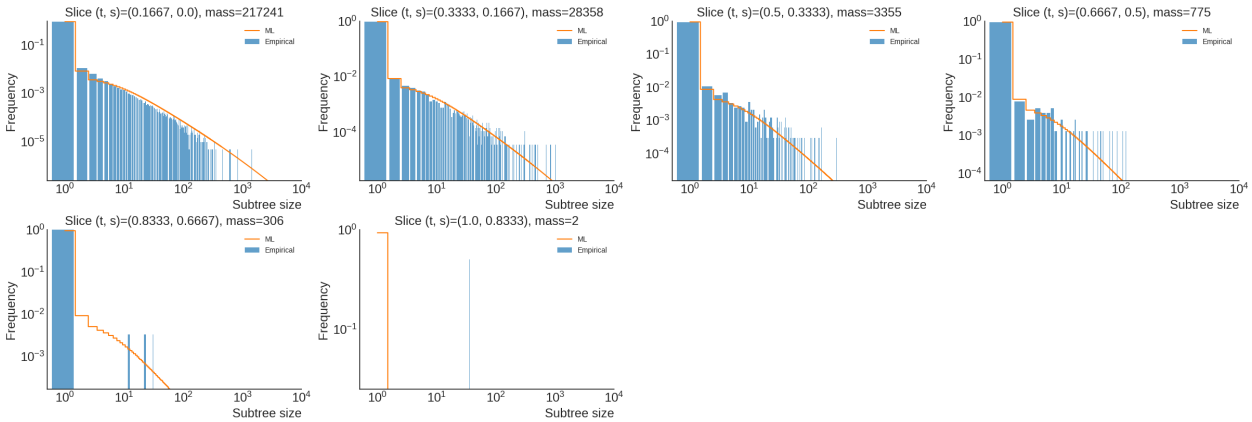


Figure C.14: Aquatic, unspecified marine benthic biome (ENVO4)

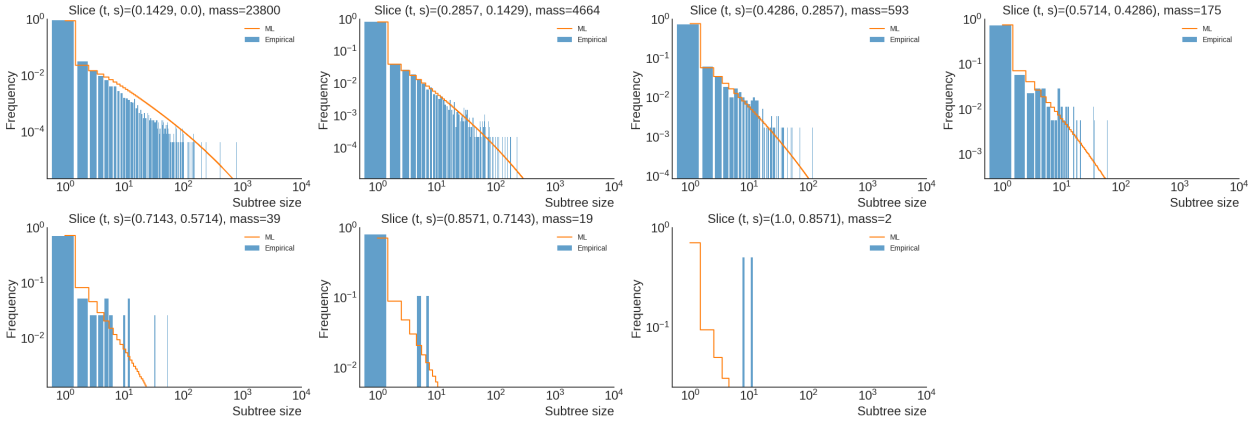


Figure C.15: Aquatic, marine pelagic biome (ENVO3)

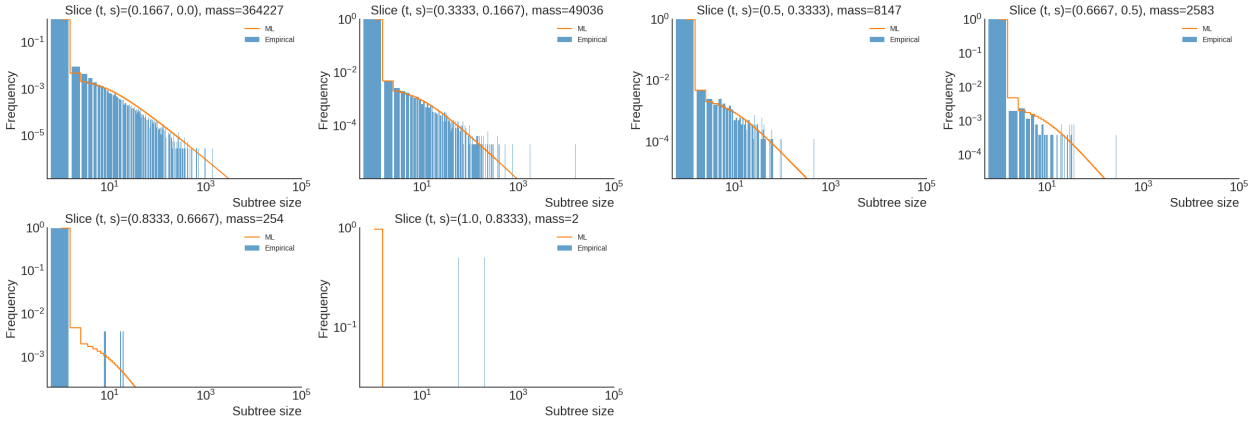


Figure C.16: Aquatic, unspecified marine biome (ENVO3)

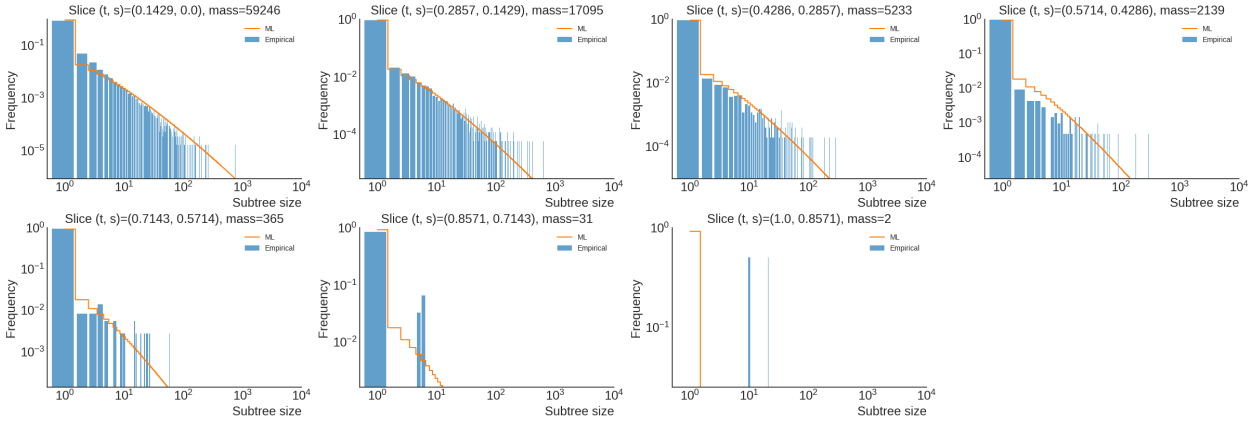


Figure C.17: Aquatic, unspecified (ENVO2)

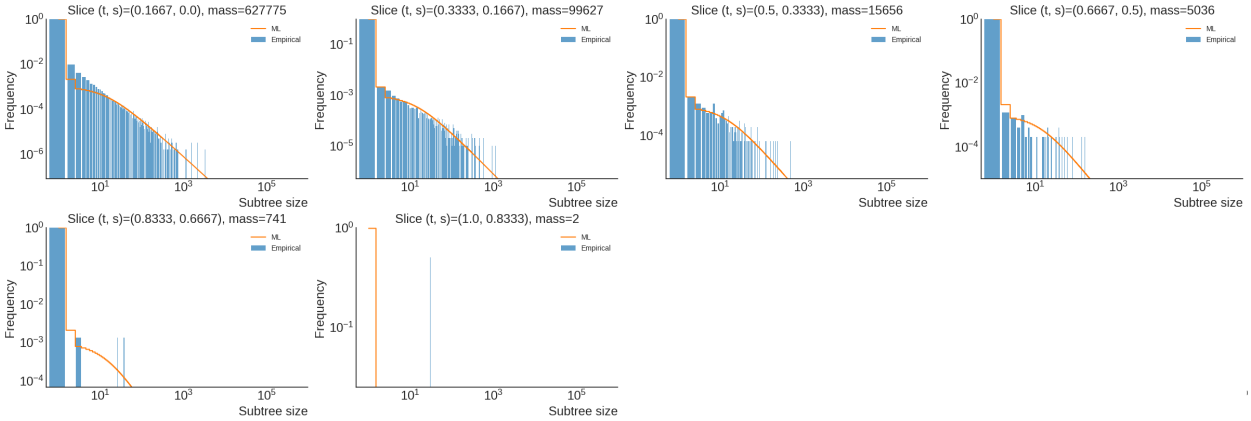


Figure C.18: Terrestrial, anthopogenic biome (ENVO2)

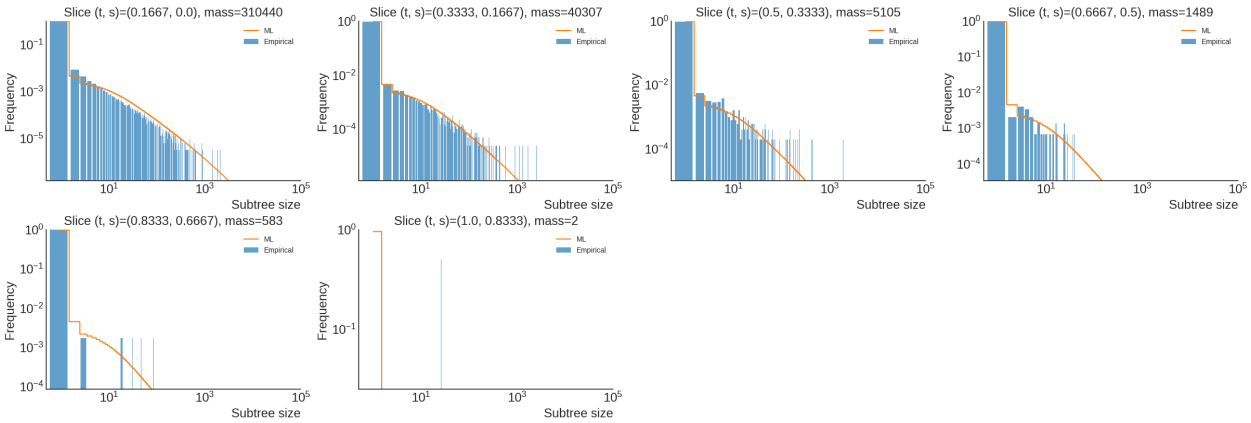


Figure C.19: Terrestrial, anthropogenic cropland biome (ENVO3)

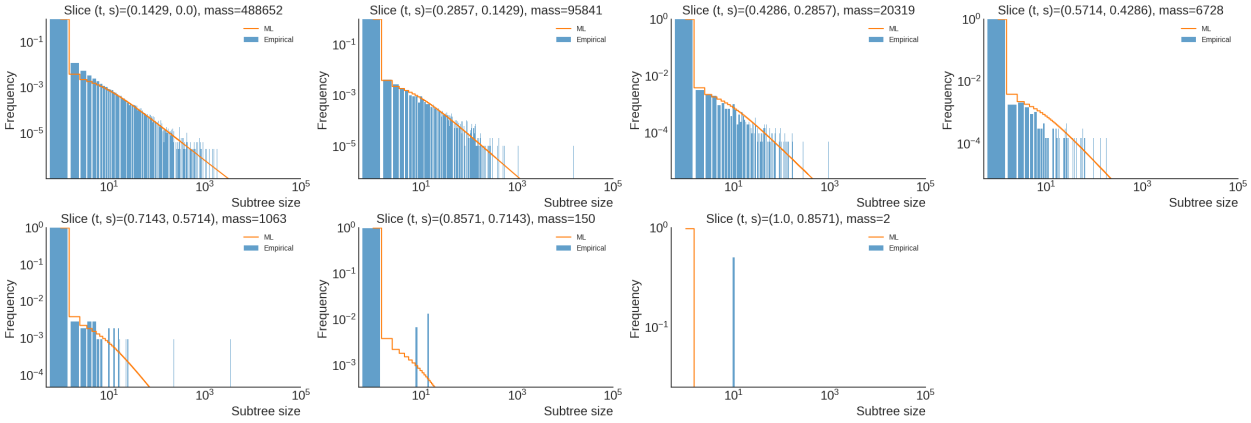


Figure C.20: Terrestrial, dense anthropogenic settlement biome (ENVO3)

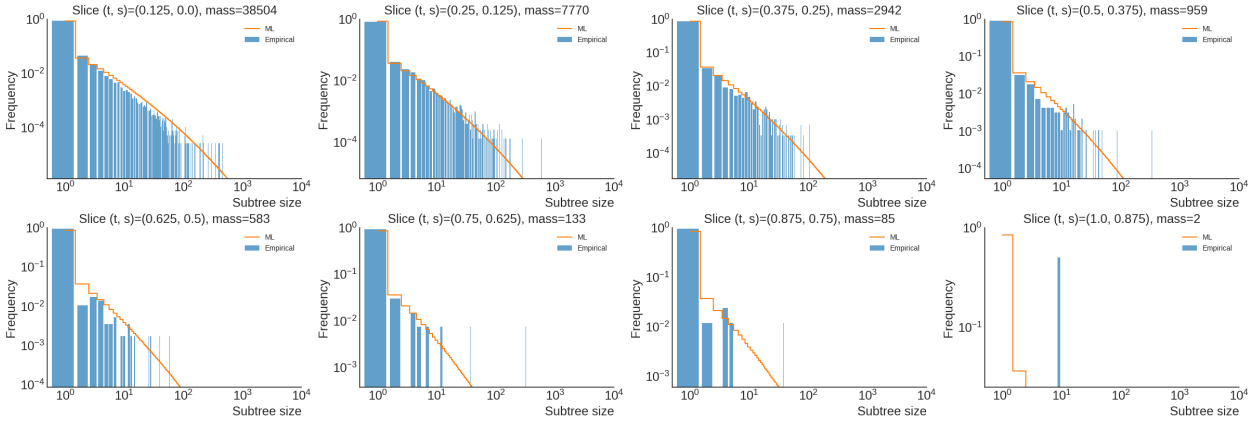


Figure C.21: Terrestrial, unspecified dense anthropogenic biome (ENVO4)

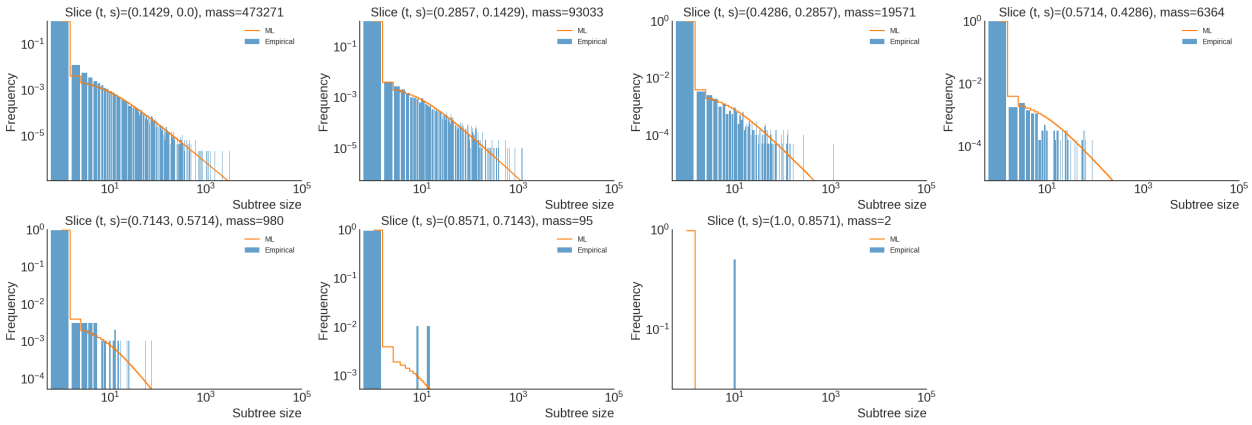


Figure C.22: Terrestrial, dense anthropogenic urban biome (ENVO4)

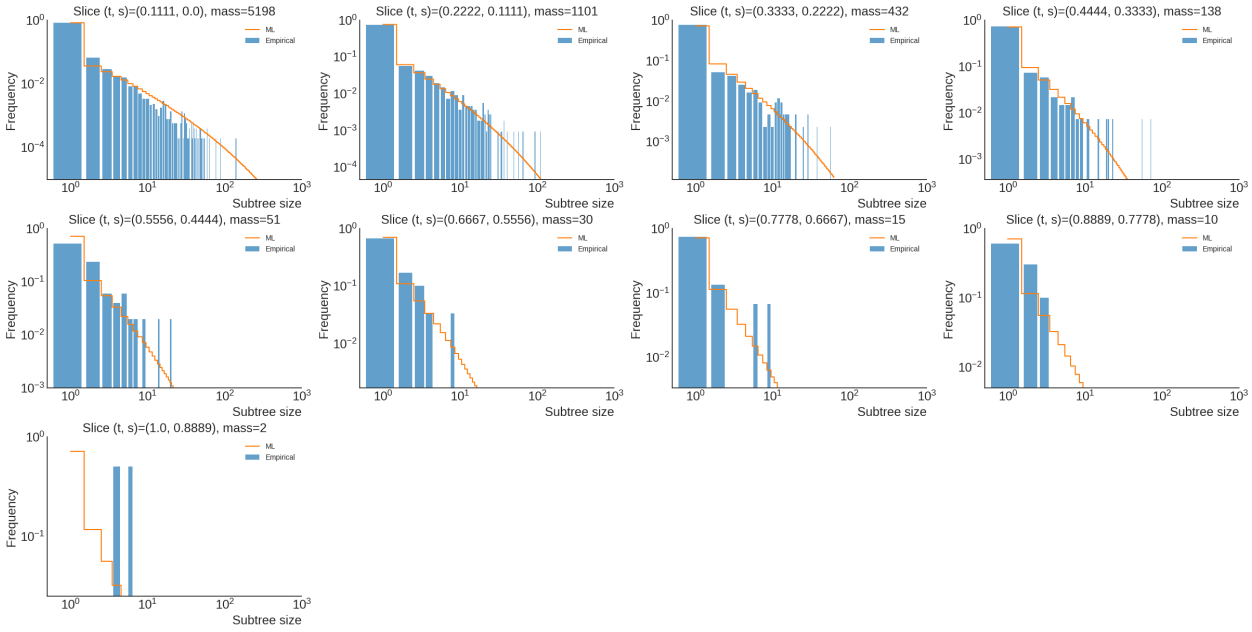


Figure C.23: Terrestrial, unspecified anthropogenic biome (ENVO3)

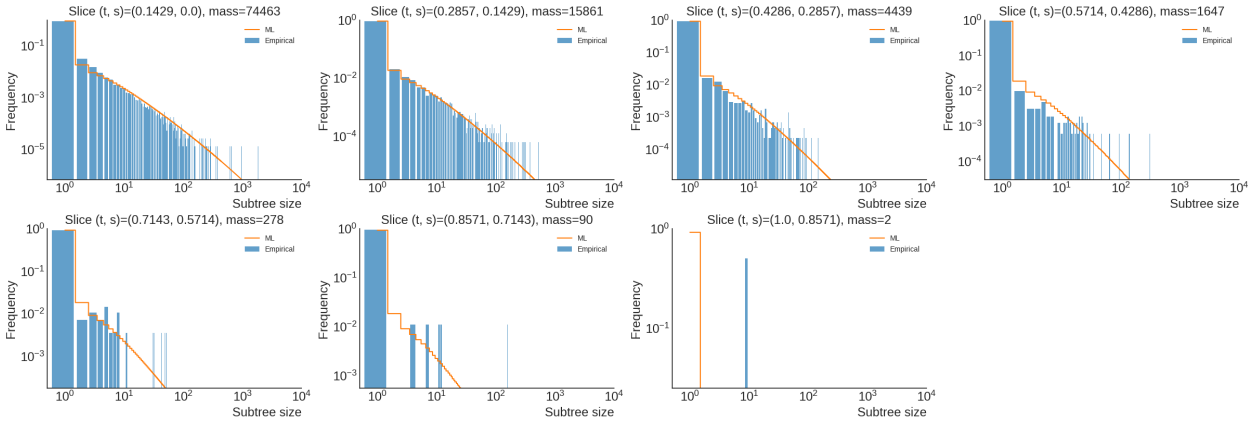


Figure C.24: Terrestrial, anthropogenic rangeland biome (ENVO3)

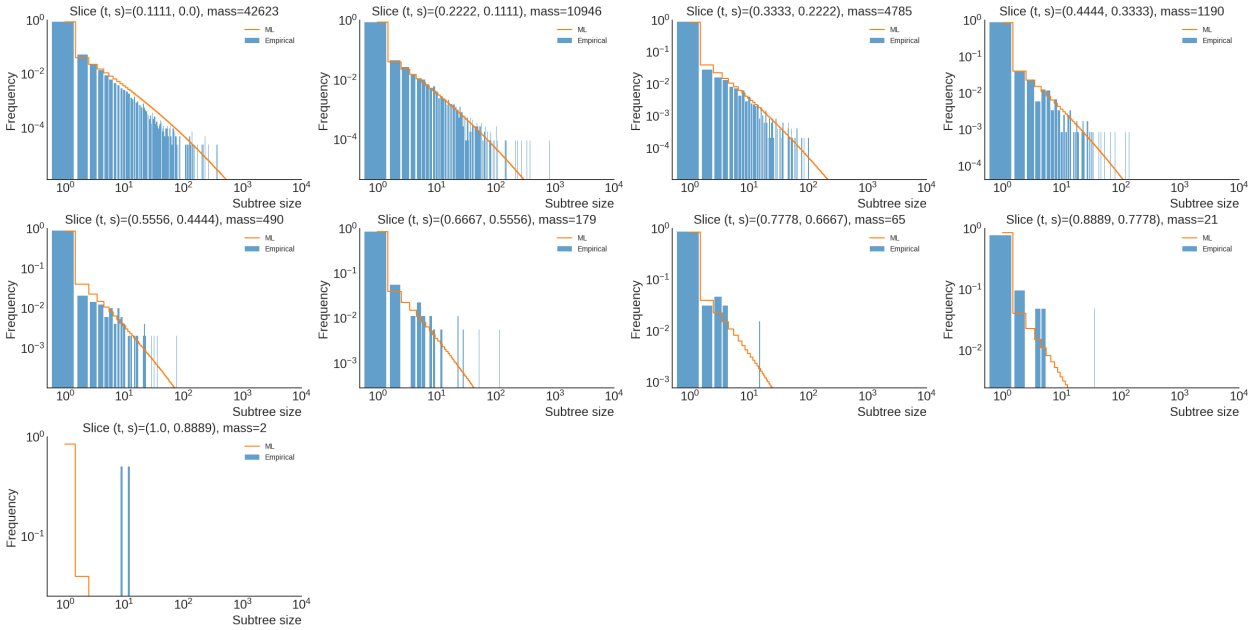


Figure C.25: Terrestrial, anthropogenic village biome (ENVO(3)

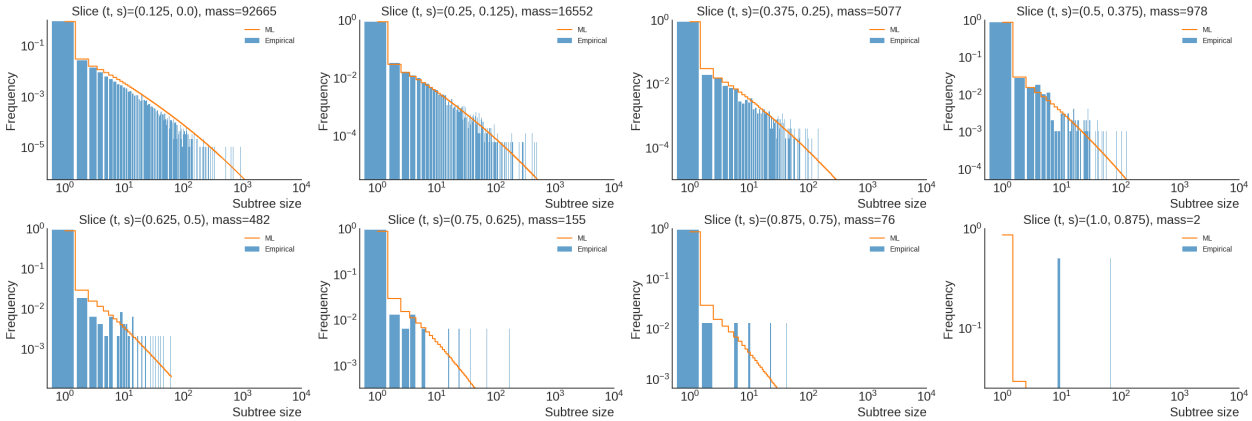


Figure C.26: Terrestrial, desert biome (ENVO2)

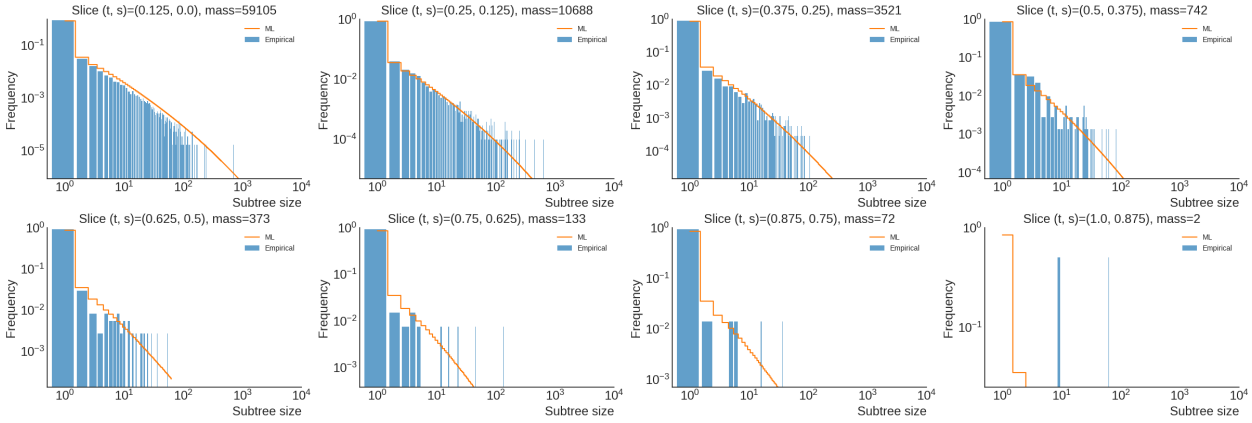


Figure C.27: Terrestrial, unspecified desert biome (ENVO3)

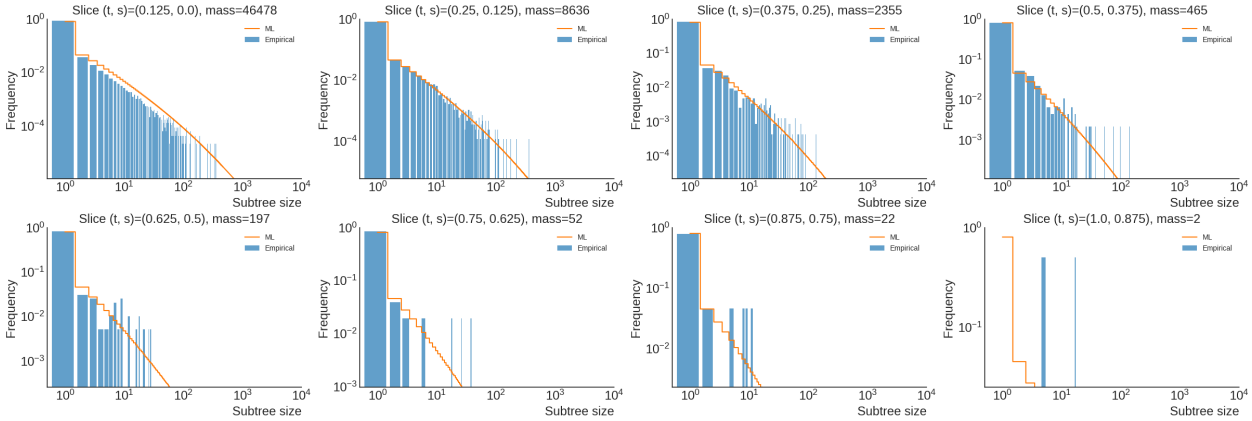


Figure C.28: Terrestrial, polar desert biome (ENVO3)

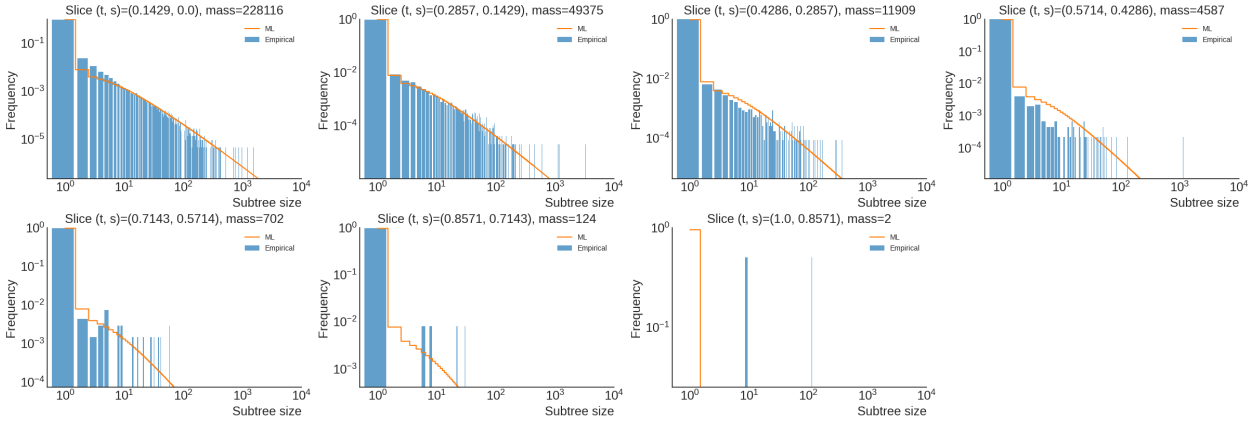


Figure C.29: Terrestrial, forest biome (ENVO2)

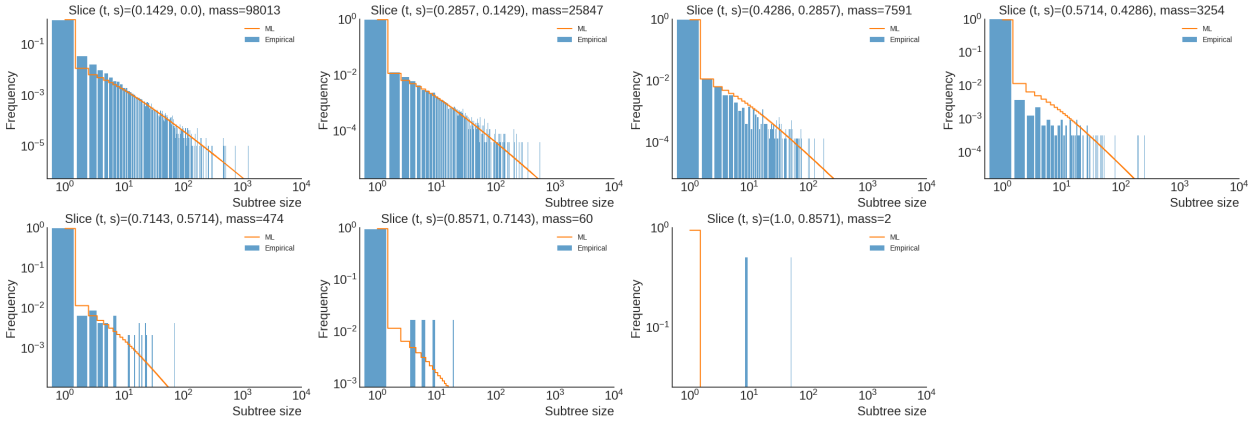


Figure C.30: Terrestrial, broadleaf forest biome (ENVO3)

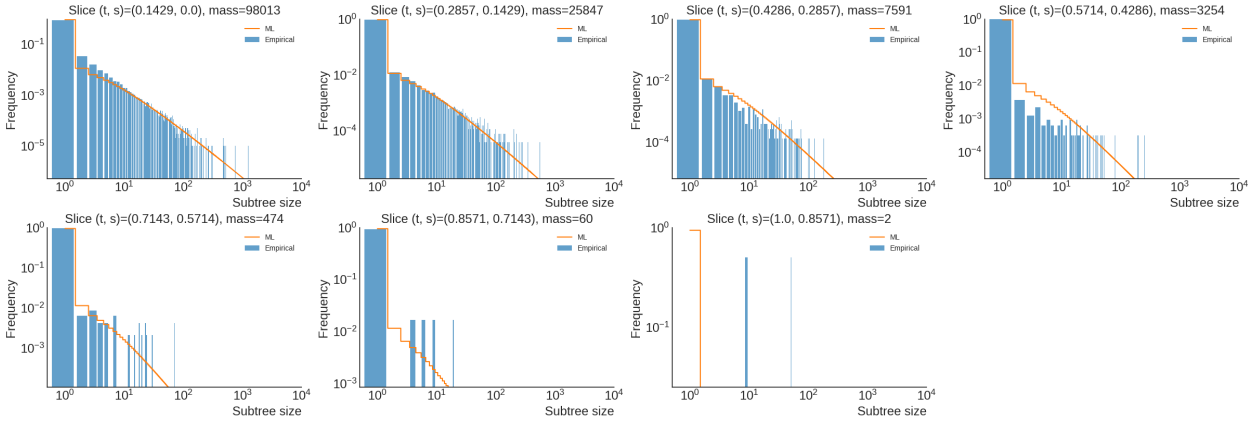


Figure C.31: Terrestrial, tropical broadleaf forest biome (ENVO3)

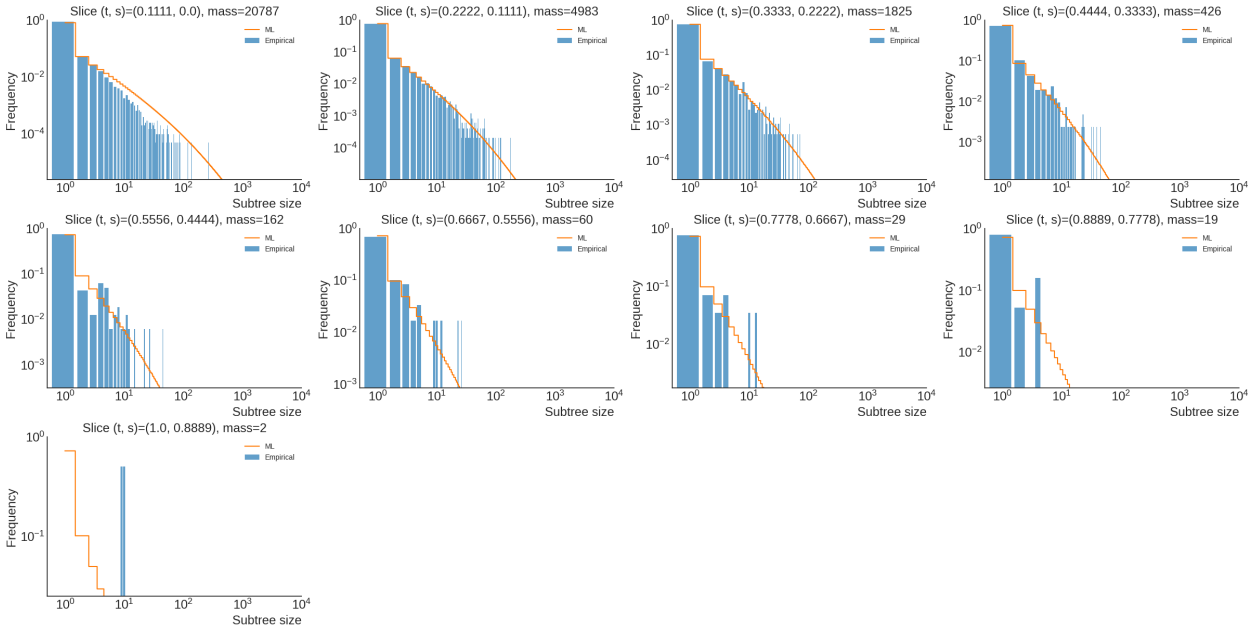


Figure C.32: Terrestrial, unspecified tropical broadleaf forest biome (ENVO4)

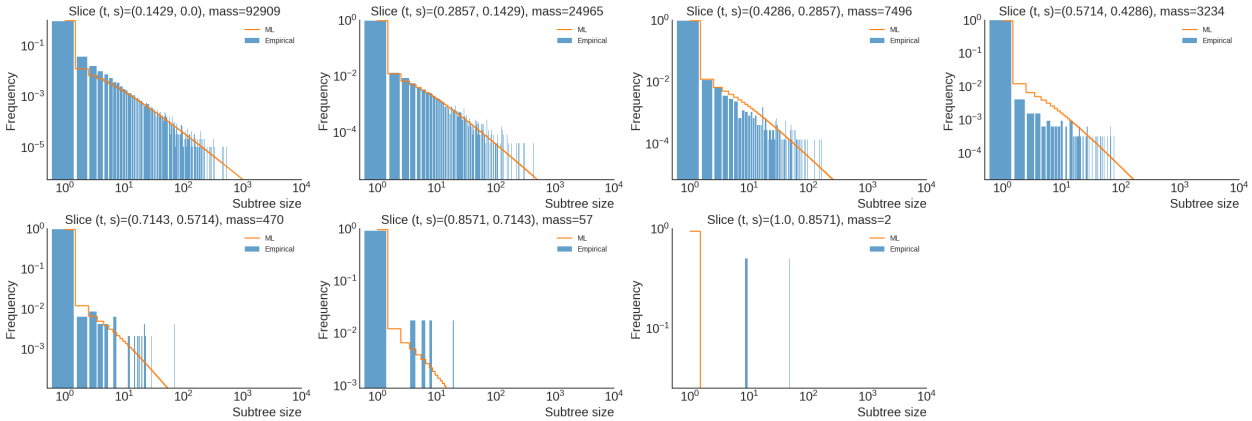


Figure C.33: Terrestrial, moist tropical broadleaf forest biome (ENVO4)

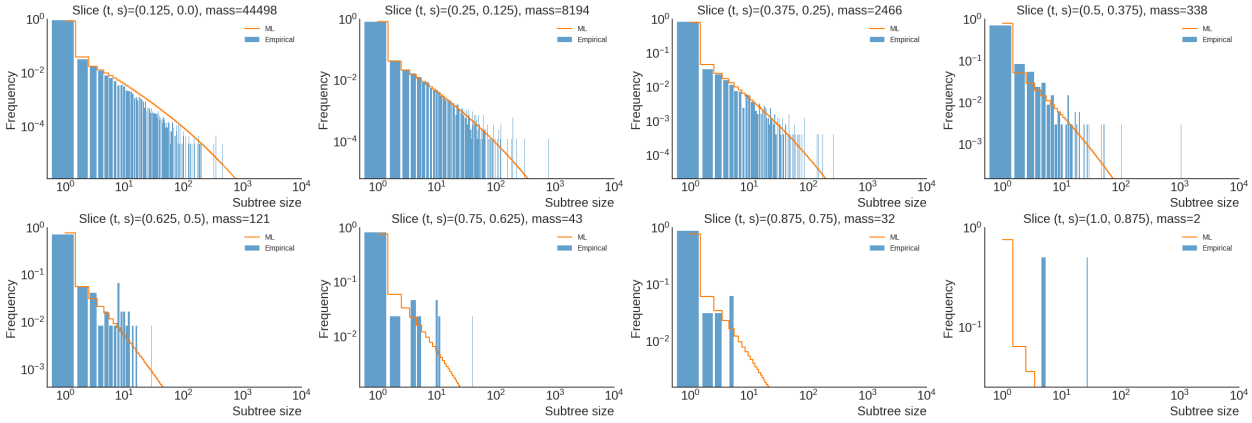


Figure C.34: Terrestrial biome, coniferous forest biome (ENVO3)

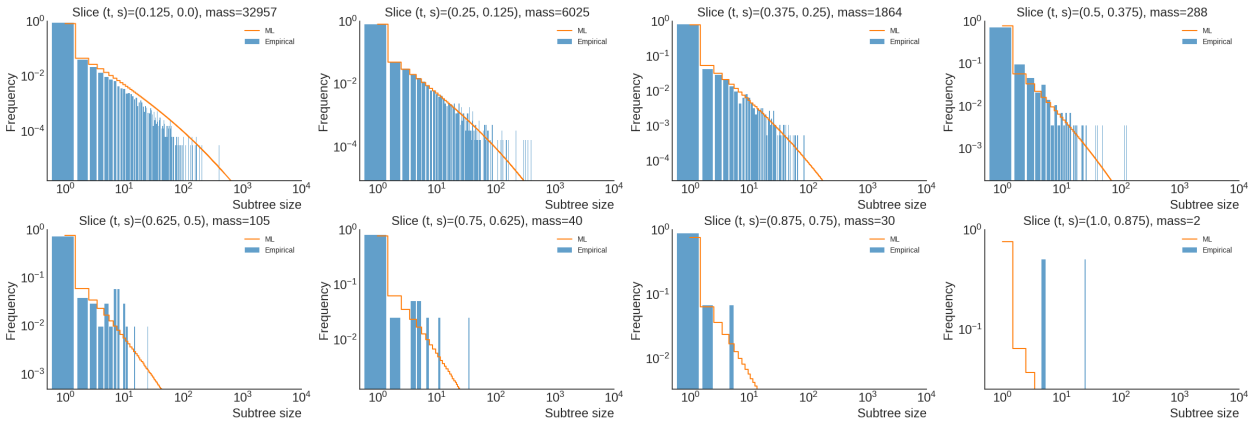


Figure C.35: Terrestrial, unspecified coniferous forest biome (ENVO4)

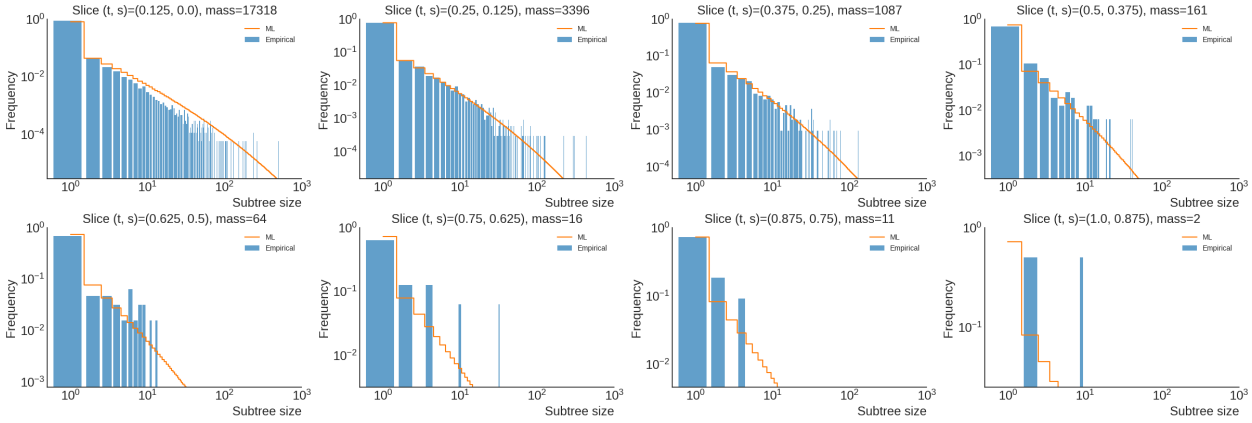


Figure C.36: Terrestrial, temperate coniferous forest biome (ENVO4)

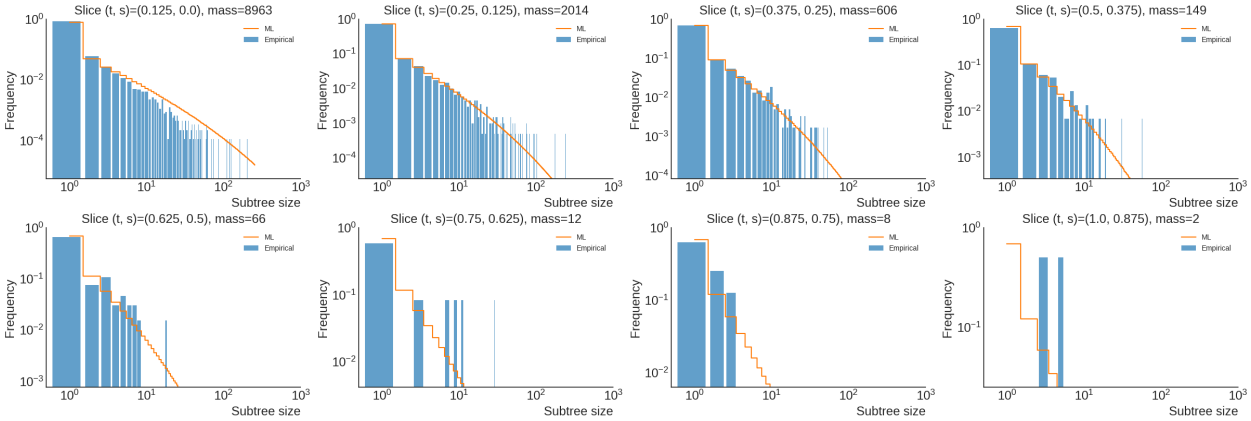


Figure C.37: Terrestrial, tropical coniferous forest biome (ENVO4)

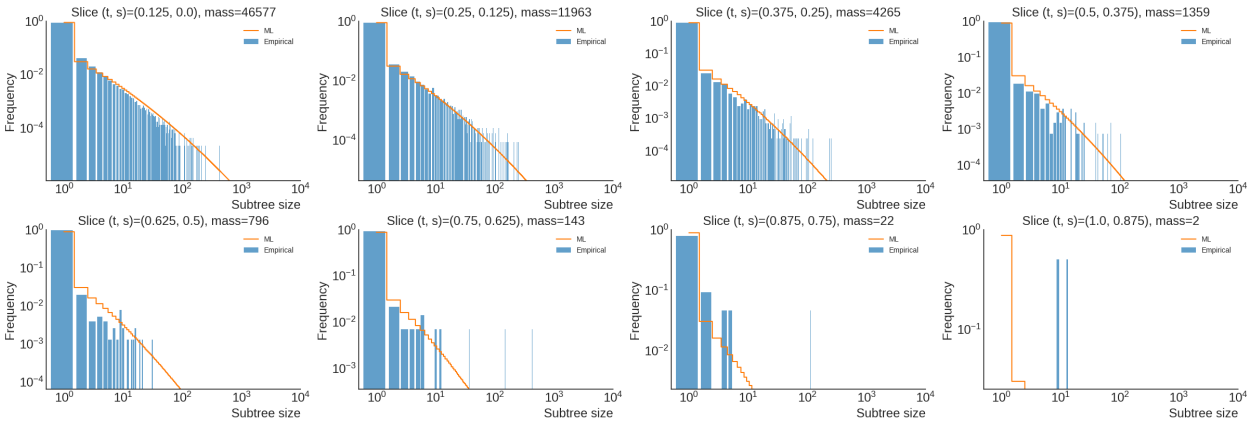


Figure C.38: Terrestrial, mixed forest biome (ENVO3)

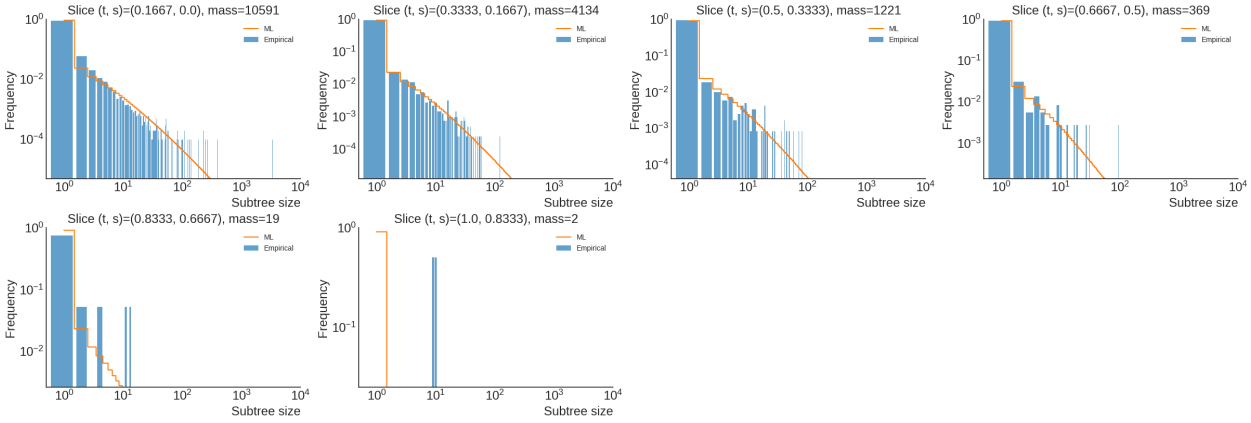


Figure C.39: Terrestrial, unspecified mixed forest biome (ENVO4)

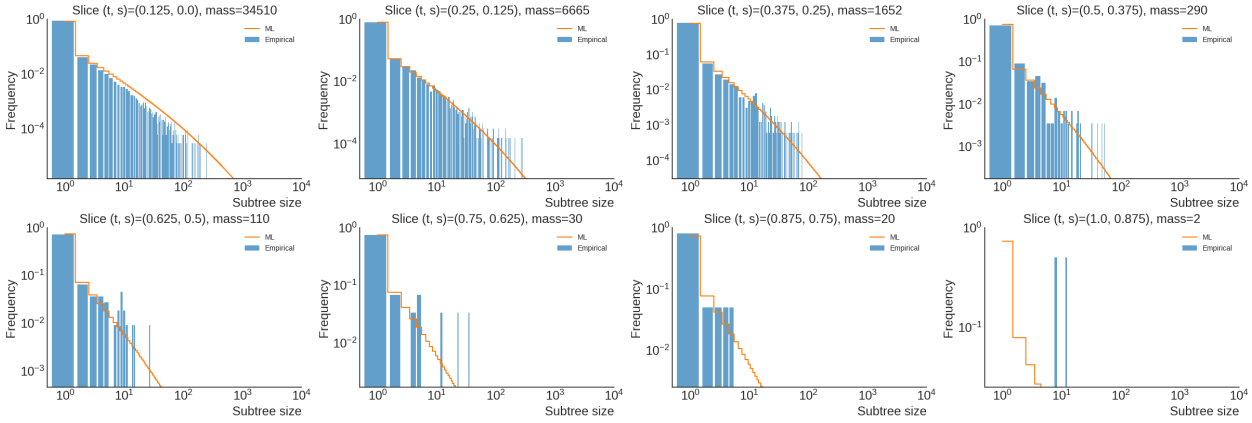


Figure C.40: Terrestrial, mixed temperate forest biome (ENVO4)

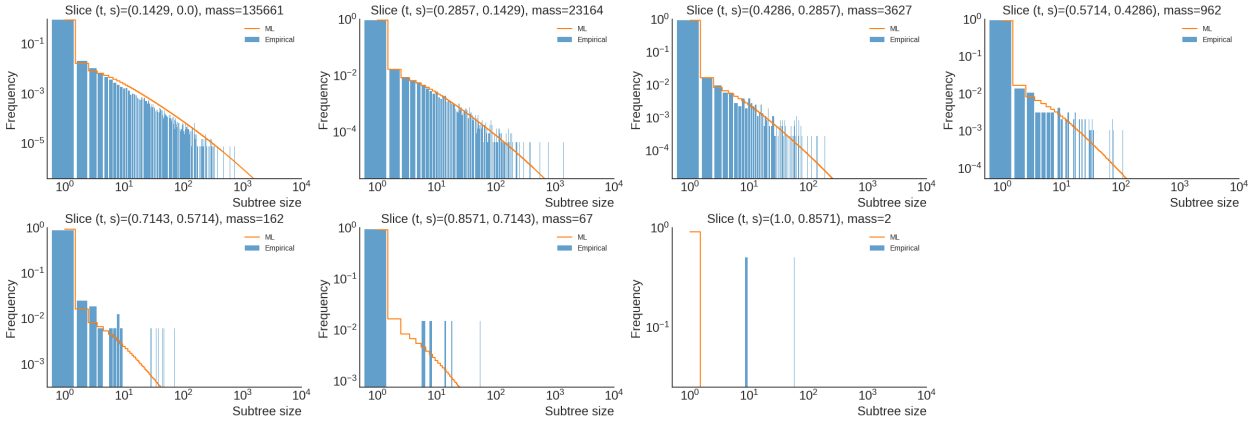


Figure C.41: Terrestrial, unspecified forest biome (ENVO3)

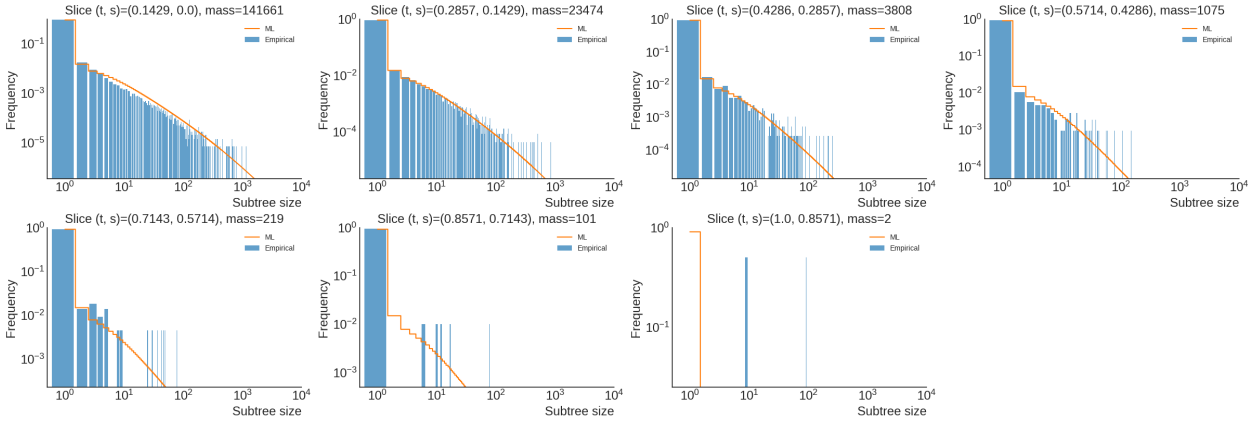


Figure C.42: Terrestrial, grassland biome (ENVO2)

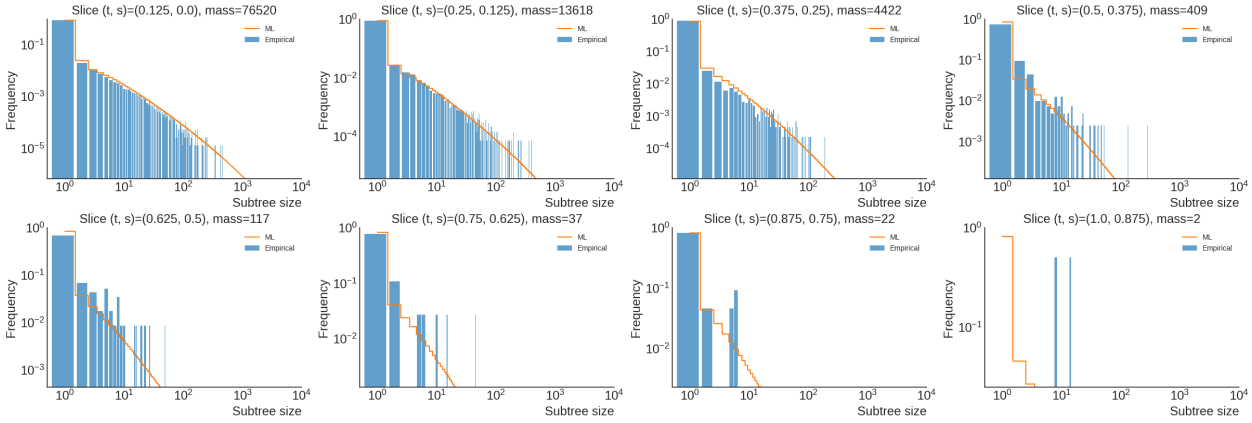


Figure C.43: Terrestrial, montane grassland biome (ENVO3)

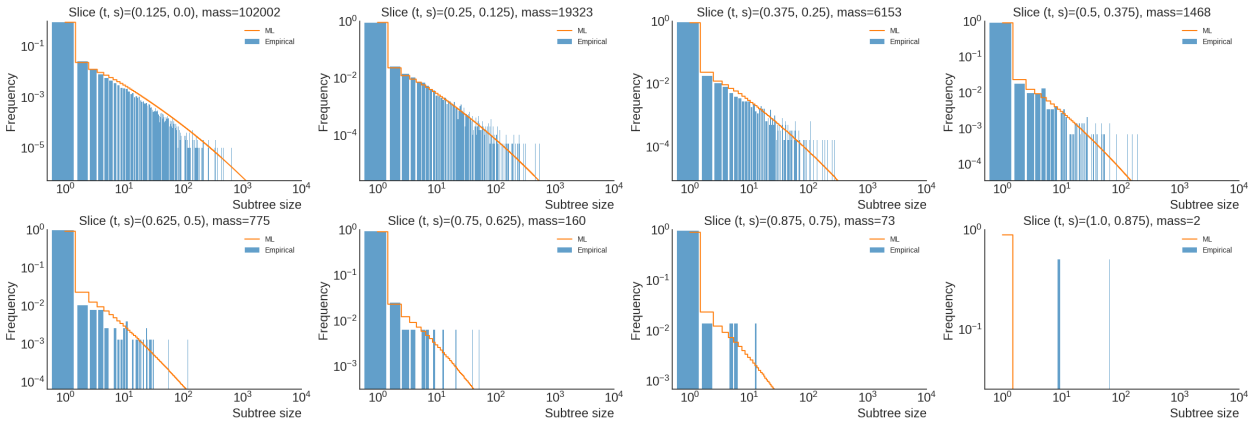


Figure C.44: Terrestrial, unspecified montane grassland biome (ENVO4)

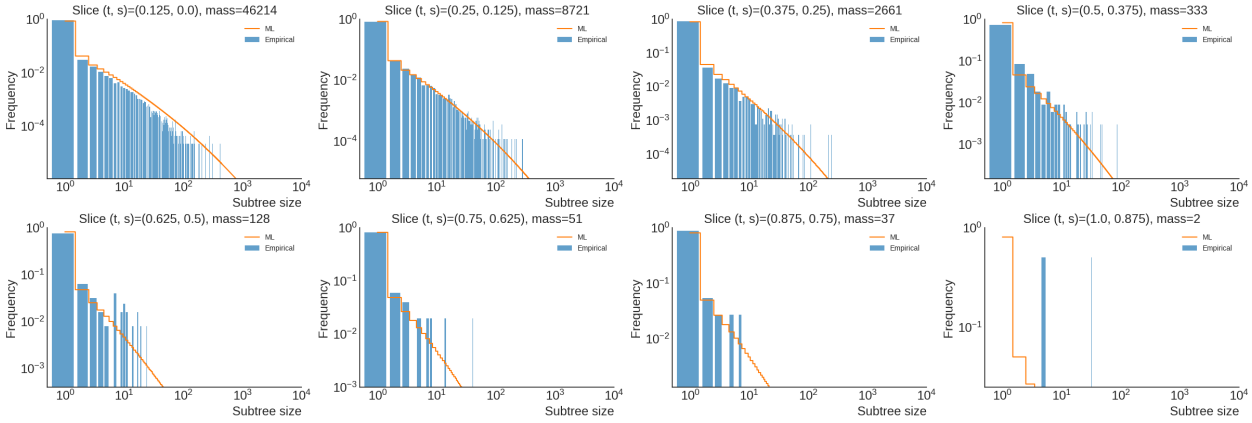


Figure C.45: Terrestrial, temperate grassland biome (ENVO3)

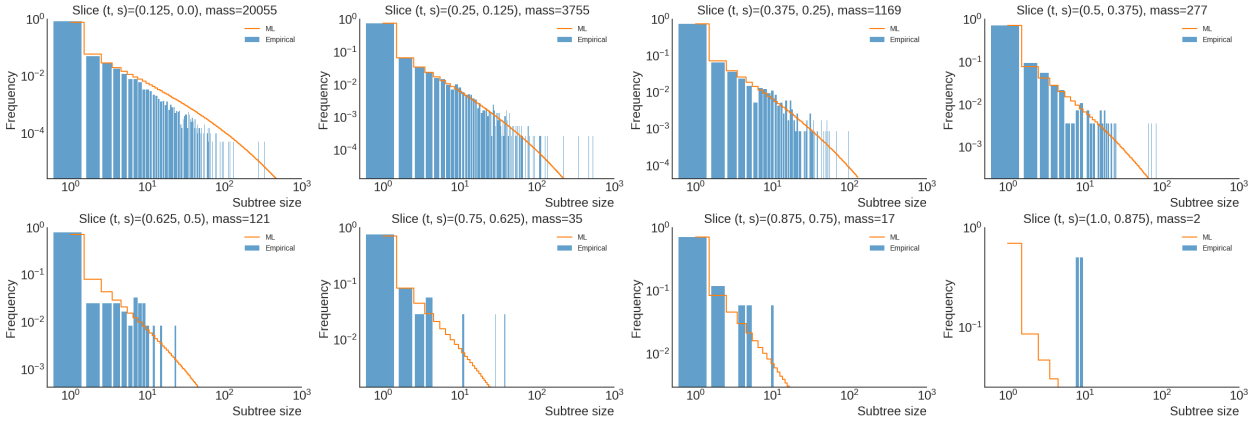


Figure C.46: Terrestrial, tropical grassland biome (ENVO3)

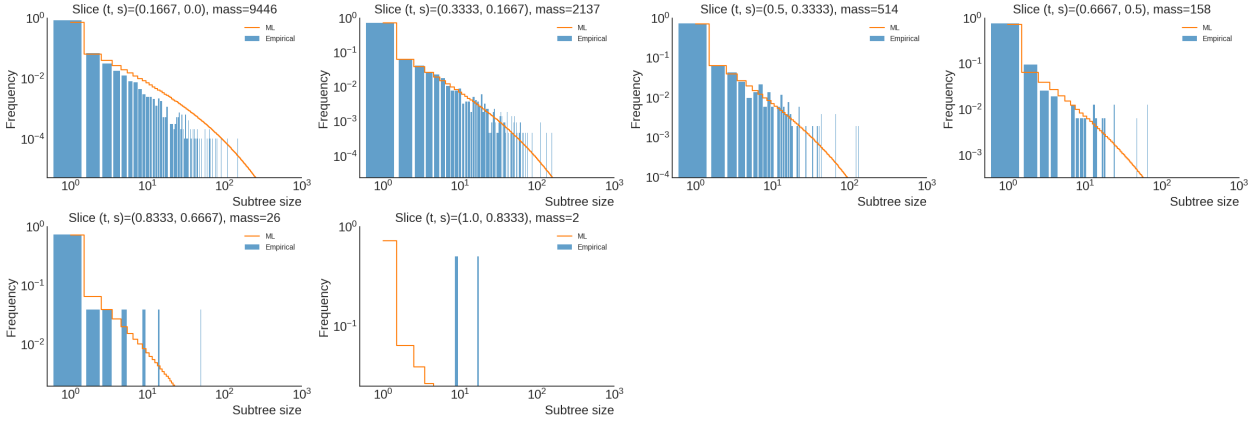


Figure C.47: Terrestrial, mangrove biome (ENVO2)

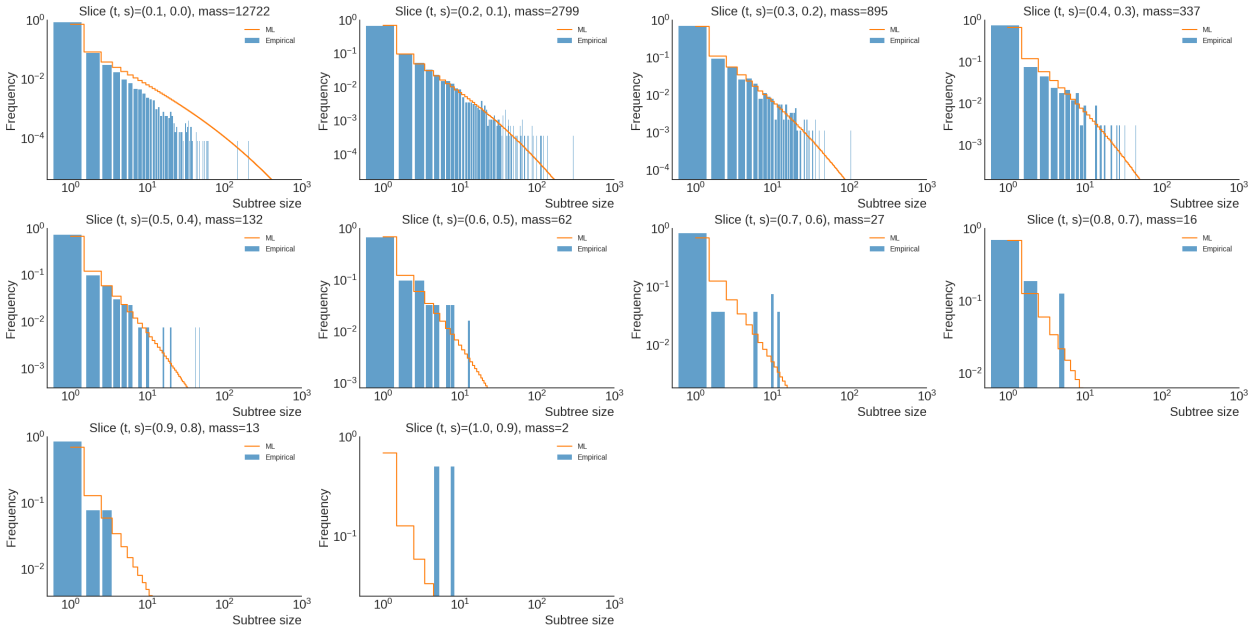


Figure C.48: Terrestrial, unspecified (ENVO2)

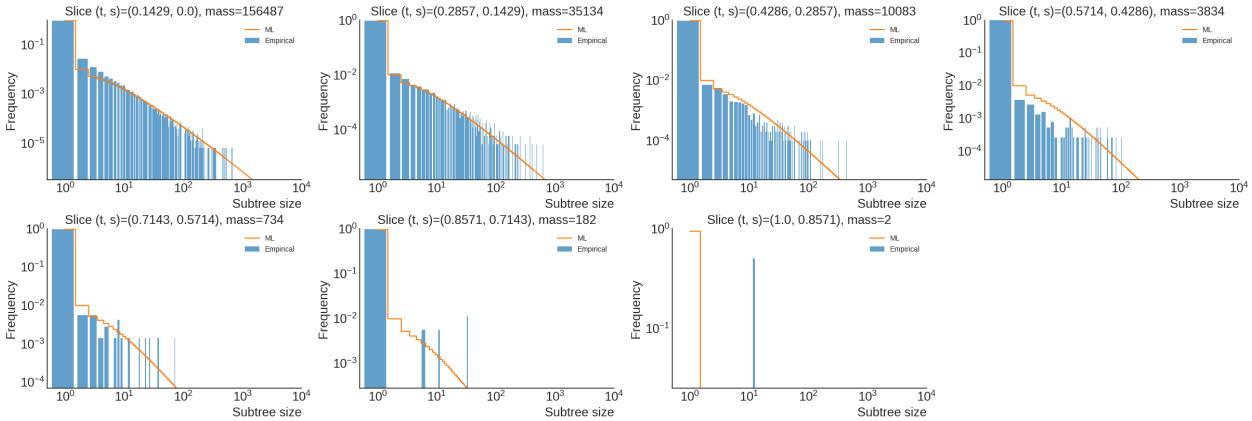


Figure C.49: Terrestrial, shrubland biome (ENVO2)

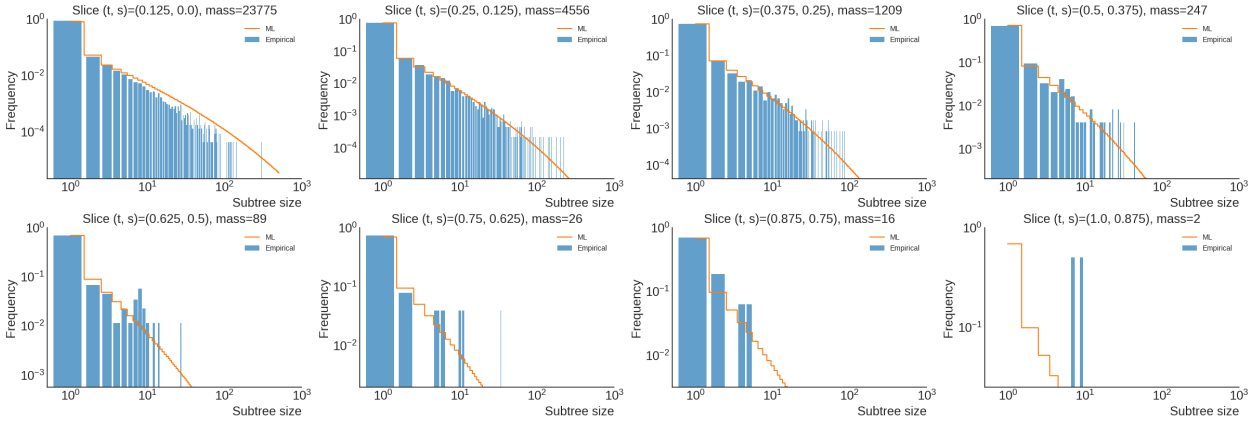


Figure C.50: Terrestrial, montane shrubland biome (ENVO3)

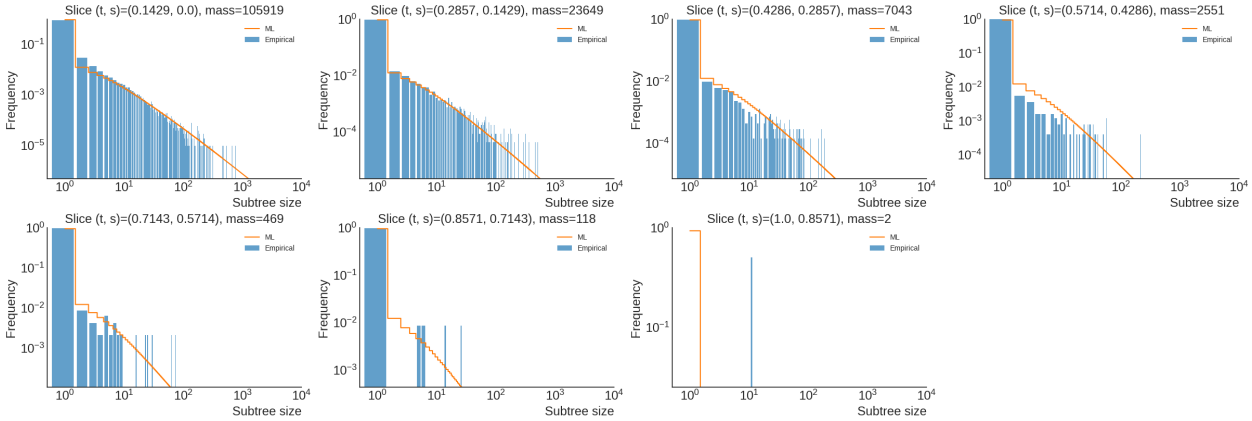


Figure C.51: Terrestrial, unspecified shrubland biome (ENVO3)

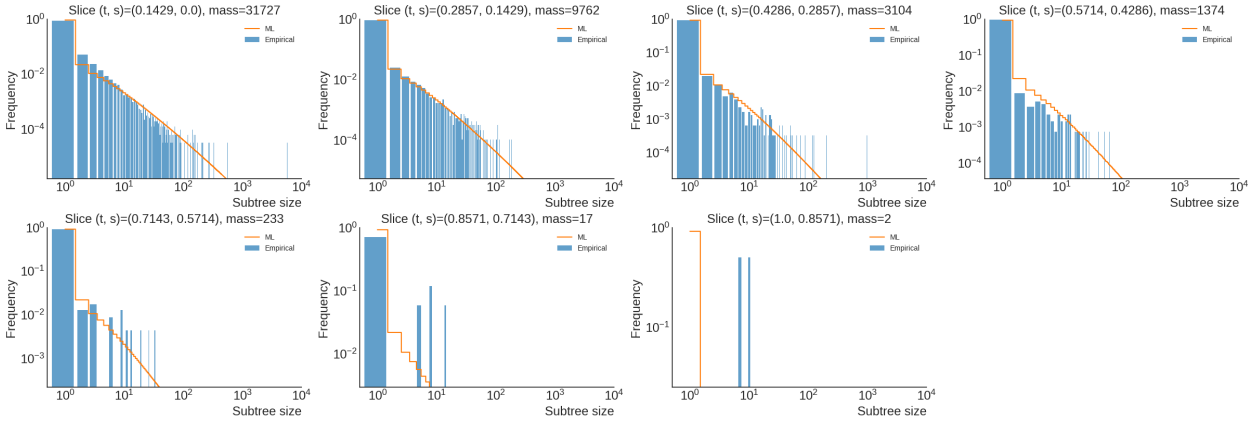


Figure C.52: Terrestrial, Mediterranen subtropical shrubland biome (ENVO4)

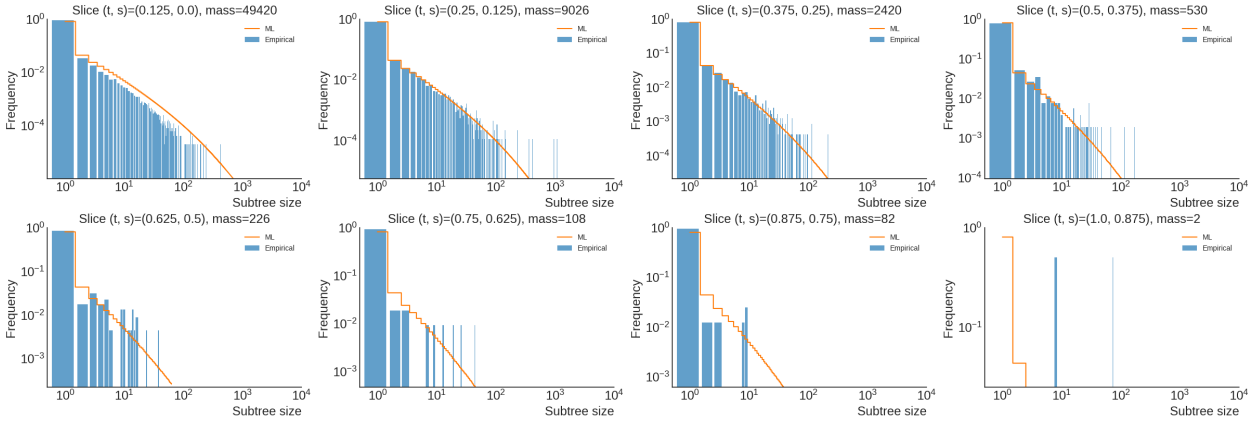


Figure C.53: Terrestrial, tropical shrubland biome (ENVO3)

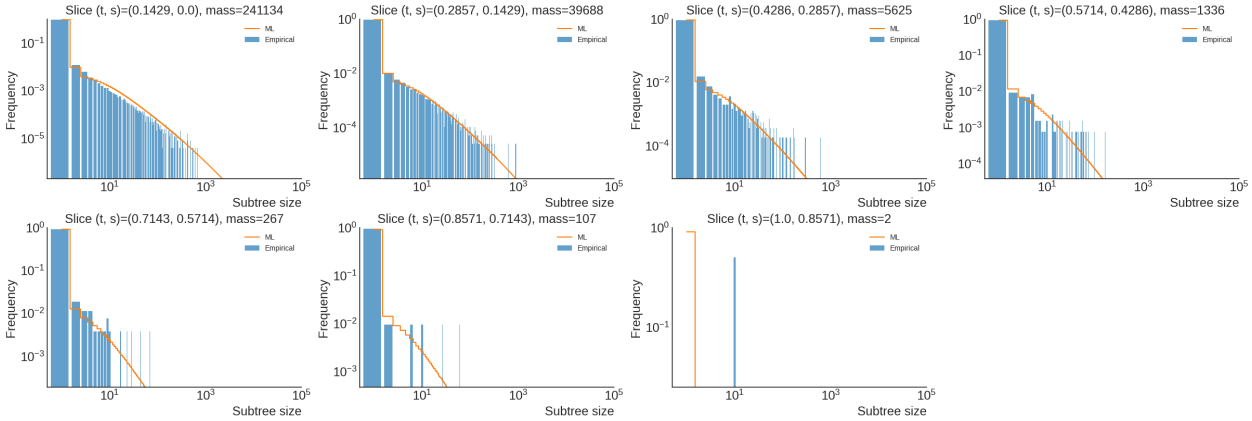


Figure C.54: Terrestrial, tundra biome (ENVO2)

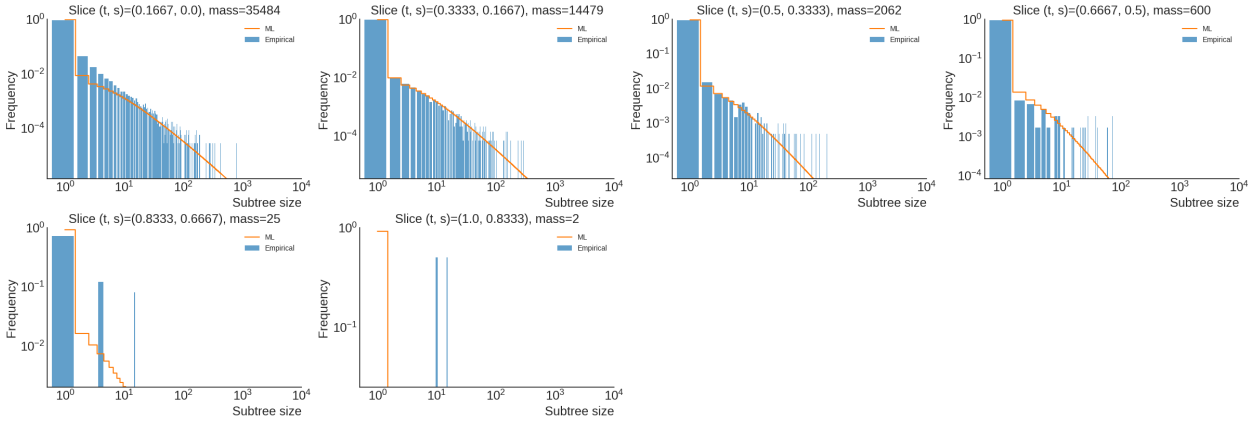


Figure C.55: Terrestrial, woodland biome (ENVO2)

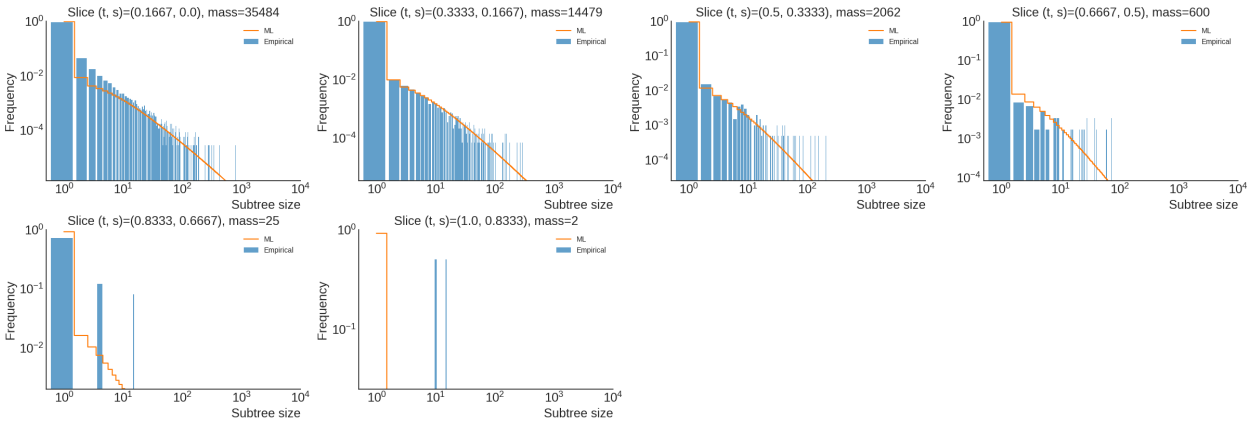


Figure C.56: Terrestrial, Mediterranean subtropical woodland biome (ENVO4)