UNIVERSITEIT GENT

Universiteit Gent
Faculteit Wetenschappen
Vakgroep Plantenbiotechnologie & Bioinformatica

Bayesian statistical methods in evolutionary genomics

Arthur Zwaenepoel

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor (PhD) of Science: Bioinformatics

Promotor: Prof. Dr. Yves Van de Peer
Academic year: 2022 – 2023

That number is the answer,
in the way that numbers answer.
That simple notion,
a coincidence among coincidences,
is all one needs to know.

Robert Ashley – Perfect Lives
Act 7: The Backyard (T'Be Continued)

Examination board

Prof. Dr. **Yves Van de Peer**
Promotor
VIB-UGent Center for Plant Systems Biology
Department of Plant Biotechnology & Bioinformatics
Ghent University/VIB, Belgium

Prof. Dr. **Klaas Vandepoele**
Chair of the examination committee
VIB-UGent Center for Plant Systems Biology
Department of Plant Biotechnology & Bioinformatics
Ghent University/VIB, Belgium

Prof. Dr. **Guy Baele**
Rega Institute for Medical Research
Department of Microbiology, Immunology & Transplantation
KU Leuven, Belgium

Dr. **Bastien Boussau**
CNRS, Laboratoire de Biométrie et Biologie Evolutive
Université Claude Bernard Lyon 1, France

Prof. Dr. Ir. **Lieven Clement**
Department of Applied Mathematics, Computer Science & Statistics
Ghent University, Belgium

Prof. Dr. Ir. **Steven Maere**
VIB-UGent Center for Plant Systems Biology
Department of Plant Biotechnology & Bioinformatics
Ghent University/VIB, Belgium

Preface

This dissertation is a product of about four and a half years of diligent study and incessant attempts at creating, finding, understanding and solving problems. Obviously, most of these attempts have failed mercilessly. Where they have failed, I invariably learned something. Where they have succeeded, the question remains whether the problems were in fact interesting. While writing this dissertation, I often stumbled on old abandoned ideas that I regret not having worked on with more perseverance. On the other hand, I found that for some of the things I did work on with perseverance, the latter was perhaps unjustified. I suppose this is the cruel fate of a scientist, which is somewhat exacerbated when working in relative isolation. At any rate, I had the opportunity to learn *a lot*, and because of that, it was well worth the effort.

In this text, I decided to collect most of my work in statistical evolutionary genomics, which certainly constitutes the most developed and unified part of the research I have done (which need not say much). The focus of the work is methodological and empirical. When I started on these projects a couple of years ago, I had, like many biologically trained aspiring researchers, a rather confused view on what statistics is and how one should do it. This dissertation reflects my gradual appreciation of what statistics is (or should be) and what it does (or can do) for the sciences, instigated by working on the problems I faced when trying to do research in evolutionary genomics. In that regard, I hope my thesis can communicate some of my enthusiasm about Bayesian statistical modeling in science, and in particular in those fields where this involves more than Gaussian distributions and linear regression models.

After this short preface (and the summary), I will switch to using the pronoun 'we'. This will be the so called *pluralis modestiae*, as all chapters were written by myself. I have been unable to adopt the 'I' pronoun throughout. Whether this is a sign of modesty or rather reluctance to bear individual responsability, I leave to the reader to judge. Virtually everything included in this thesis is

unpublished in the presented form, and where material is included that has appeared elsewhere before, this will be indicated accordingly. If, at times, the reader should wish to accuse me of *logorrhea*, I have little excuse but that I (still) do not find biology, nor statistics, easy topics to write on. I had the intent to write a short dissertation, but in part due to the urge to include many examples, this failed. If, moreover, the reader would find that I use too many footnotes for a scientific text, my only excuse shall be that I like footnotes. I assure the reader that they really are footnotes, and can be safely ignored.

While my PhD was largely a solo endeavor (and certainly too much so), there are of course a number of people who I wish to thank.

On the academic side, let me first thank my promotor, Yves, for trusting and supporting me, and for giving me the freedom to lay out my own research agenda. Whether granting me that freedom has beared its fruits, I again leave for the reader of this dissertation to judge. In addition, I warmly thank the members of the bioinformatics and evolutionary genomics group for their kindness and interest. I would also like to express my gratitude to the members of the examination committee for taking the effort to read my dissertation, as well as for their general appraisal and constructive feedback.

Next, I would like to thank professor Gertrudis Van de Vijver, for involving me in organizing her philosophy course on the life sciences and for infusing me with the Kantian spirit. I can only hope I have been reasonably attentive to the constitution of objectivity in the present work. At least I know I tried.

Among my friends, special thanks are due to Michiel, who has been, besides a friend, my closest colleague the last couple of years. Then, in approximate order of shared beers, I thank Brent, Torben, Jef and Siemen, [short pause for breath], Emiel, Lisa, Loïc, Janne, Elias, Brecht, Fien, Henri, Tom, Lukas, Stijn, Thomas, Ennio, Maarten, *etc.*

Infinite gratitude is due to my parents for their unconditional support and for making my life, up till now, (reasonably) easy. I would not wish to forget to thank my brother, in particular for tolerating me, which can be challenging, so I heard. My warm thanks also to my silly little feline friend, Rik.

Lastly, I am left searching for words to thank Katrien, without whom my life is literally unimaginable. I have lost track of where I end and you begin, and that is totally OK with me.

Arthur Zwaenepoel – September 21, 2022

Summary

Genome sequencing projects in the last two decades have resulted in a deluge of whole-genome data sets for a large number of species across the tree of life. These data reveal the staggering extent of extant genomic diversity, and call for close attention to the evolutionary processes which generate that diversity. The latter demand is what the field of *evolutionary genomics* seeks to address. Making sense of genomic diversity is however not straightforward, and requires the development of statistical models for the relevant evolutionary processes, in order to quantify rates of genome evolution by various causes, and enable the reconstruction of evolutionary history in a principled way. It is to this broader challenge that this dissertation is dedicated.

In the present work, I start from the conception of a genome as a 'bag of genes', evolving over long time scales through processes of speciation, polyploidization, hybridization, gene duplication and gene loss. Using probabilistic models of gene family evolution which account for these processes, I devise Bayesian hierarchical models to study evolutionary change and reconstruct evolutionary histories at the genomic scale. The work is firmly rooted in statistical phylogenetics, with a distinctively Bayesian commitment to assessing the adequacy of the assumed models in light of the observed data.

In the first half of this dissertation, I focus on models of gene *content* evolution along a known species tree. After briefly considering models for the paranome age and size distribution, I conduct an in-depth study of phylogenetic birth-death process models of gene family evolution, and devote special attention to the modeling and inference of ancient whole-genome duplications using statistical approaches. A new model for gene family evolution by gene duplication and loss based on a two-type branching process is proposed and studied in detail. In the second half of this dissertation, I switch to a more explicitly phylogenomic point of view, modeling the evolutionary processes which shape gene family phylogenetic trees at a genomic scale. I develop a

new approach for likelihood-free Bayesian inference of species trees under the multispecies coalescent model from observed empirical gene tree distributions. Lastly, I describe an approach for Bayesian gene tree reconciliation of multi-copy gene families under birth-death process models of gene family evolution and revisit the problem of statistically inferring ancient whole-genome duplications in a phylogenetic context. A brief conclusion and outlook on the future of statistical evolutionary genomics ensues.

Samenvatting

De zogeheten ‘genoomprojecten’ van de laatste kwarteeuw hebben een ware zondvloed aan grootschalige genomische data sets voortgebracht, zodat vandaag complete genoomsequenties beschikbaar zijn voor een aanzienlijk aantal taxa. Deze data onthullen een overweldigende diversiteit op het genomische niveau, die een nauwlettende aandacht eist naar de evolutionaire processen die deze diversiteit kunnen veroorzaken. Deze laatste eis is waar de *evolutionaire genomica* aan tracht tegemoet te komen. Het begrijpelijk maken van genomische diversiteit is echter allesbehalve vanzelfsprekend, en vereist de ontwikkeling van statistische modellen voor de relevante evolutionaire processen. Dergelijke modellen maken het mogelijk om, op basis van genomische data, de snelheden waarmee deze processen zich voltrekken te bepalen, alsook om de evolutionaire geschiedenis die met specifieke data geassocieerd is op een statistische basis te reconstrueren. Deze dissertatie is gewijd aan deze uitdagingen.

Ik vat mijn studie in het huidige werk aan met de conceptie van een genoom als een ‘zak met genen’, die evolueert over uitgestrekte tijdsintervallen aan de hand van speciatie, polyploidisatie, hybridisatie, genduplicatie en genverlies processen. Ik ontwikkel Bayesiaanse hiërarchische statistische modellen voor de grootschalige evolutie van genomen en de reconstructie van evolutionaire geschiedenissen, gebruikmakende van probabilistische modellen voor genfamilie evolutie die laatstgenoemde processen in rekening brengen. Het hele werk is sterk verankerd in statistische fylogenetica, met een uitdrukkelijk Bayesiaanse toewijding aan het kritisch evalueren van de adequaatheid van aangenomen theoretische modellen voor empirische data.

In de eerste helft van deze dissertatie leg ik de focus op modellen voor de evolutie van het aantal genen in een genfamilie in de context van een aangenomen evolutionaire geschiedenis voor de relevante genomen. Na een korte beschouwing van modellen voor genfamilie evolutie in een

enkel genoom, ga ik over op een diepgaande evaluatie van fylogenetische *birth-death* proces modellen voor de evolutie van genfamilies. Ik besteed bijzondere aandacht aan het modelleren en statistisch detecteren van historische genoomduplicaties. Daarna ontwikkel ik een nieuw probabilistisch model voor genfamilie evolutie gebaseerd op een tweesoortig vertakkingsproces (*two-type branching process*) om tegemoet te komen aan enkele van de problemen blootgelegd in de voorgaande studie. In de tweede helft van deze dissertatie verander ik van perspectief en beschouw ik meer expliciet de evolutionaire geschiedenissen van individuele genfamilies. Hier is het doel om de evolutionaire processen die variatie veroorzaken in fylogenetische bomen overheen het genoom te modelleren en in kaart te brengen. Ik ontwikkel een nieuwe *likelihood-free* Bayesiaanse methode om de evolutionaire geschiedenis op genoomniveau te reconstrueren onder de assumpties van het *multispecies coalescent* model, gebruikmakende van empirische distributies over genfamilie-specifieke fylogenetische bomen. In het laatste onderzoek in deze dissertatie beschrijf ik een methode voor Bayesiaanse *reconciliation* analyse met behulp van fylogenetische *birth-death* proces modellen. Ik beschouw opnieuw in detail het bepalen van historische genoomduplicaties in een fylogenetische context aan de hand van statistische methoden. Met een korte conclusie en blik op de uitdagingen en toekomst van de statistische evolutionaire genomica sluit ik deze dissertatie af.

Contents

List of abbreviations and symbols

Abbreviations

ABC	approximate Bayesian computation
AD	automatic differentiation
ALE	amalgamated likelihood estimation
ASDF	average standard deviation of split frequencies
BDP	birth-death process
CCD	conditional clade distribution
CTMC	continuous-time Markov chain
DL	duplication and loss
DLF	duplication, loss and x-functionalization
DLWGD	duplication loss and WGD
DP	dynamic programming
EP	expectation propagation
ESS	effective sample size
GO	Gene Ontology
GTR	general time-reversible [model]
HGT	horizontal gene transfer
HMC	Hamiltonian Monte Carlo
iid	independent and identically distributed
ILS	incomplete lineage sorting
IS	importance sampling
JC	Jukes & Cantor [model]
K80	Kimura's 1980 [model]
KL	Kullback-Leibler [divergence]
LRT	likelihood ratio test
MAP	maximum a posteriori
MBM	Markov branching model
MCMC	Markov chain Monte Carlo
MCSE	Monte Carlo standard error
ML	maximum likelihood
MLE	maximum likelihood estimate
MP	maximum parsimony
MRCA	most recent common ancestor
MSC	multispecies coalescent
NUTS	no U-turn sampler
pdf	probability density function

pgf	probability generating function
pmf	probability mass function
rjMCMC	reversible-jump Markov chain Monte Carlo
SBN	subsplit Bayesian network
SIS	sequential importance sampling
SSD	small-scale duplication
SSDL	small-scale duplication and loss
WGD	whole-genome duplication
WGM	whole-genome multiplication
WGT	whole-genome triplication

Symbols & notation

General

$[1..n]$	set of integers from 1 to n
$\mathcal{P}(A)$	powerset of set A
$\lvert A\rvert$	cardinality of set A
$\mathbb{E}$	expectation
$\mathbb{P}$	probability measure
$\mathbb{1}_A(\cdot)$	indicator function for set A, also $\mathbb{1}[\cdot \in A]$
$\delta_x(\cdot)$	Dirac delta function at x
$G(V, E)$	graph with vertex set V and edge set E
$p(\cdot)$	generic probability distribution (density or mass function)
$p(\cdot\lvert\cdot)$	generic conditional probability distribution (density or mass function)
y	generic data point or data set
θ	generic parameter
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
K_{S}	synonymous distance
My	million years
Gy	billion years

Chapter 2

λ	birth (duplication) rate
μ	death (loss) rate
$X(t)$	a continuous-time branching process
$p_{ij}(t)$	transition probability $\mathbb{P}\{X(t) = j \lvert X(0) = i\}$
$h_{i,t}(j)$	transient distribution of $X(t)$ for fixed t and $X(0) = i$
$\xi(k)$	pmf for the offspring distribution of a branching process
$g(s)$	pgf for the offspring distribution of a branching process
$f(s, t)$	pgf for a continuous-time branching process starting at $X(0) = 1$
$\epsilon(t)$	extinction probability at time t for a branching process starting at $X(0) = 1$

Chapter 3

S	species tree
$V(S)$	nodes (vertices) of S
$E(S)$	branches (edges) of S
$\mathcal{L}(S)$	leaves of S
$u, v, w, \ldots$	generic node labels
o	root node of a tree

$\rho(u)$	parent node of node u
$\langle u, v\rangle$	branch from node u to v
t_u	length (duration) of the branch leading to node u
S_u	subtree of S rooted in node u
X_u	random gene count at vertex u of S
$X_{[u]}$	random phylogenetic profile associated with S_u, i.e. $(X_v \colon v \in \mathcal{L}(S)$
X	shorthand for $X_{[o]}$
Y_u	number of genes at node u which leave observed descendants in $\mathcal{L}(S_u)$
$f_u(s)$	pgf for X_u conditional on $X_{\rho(u)} = 1$, i.e. $f(s, t_u)$
$g_u(s)$	pgf for X_u
ϵ_u	probability that a single lineage at u leaves no observed descendants
$Z_k(t)$	number of descendants at time t of gene k
$q_{ij}(t)$	transition probability $\mathbb{P}\{X(t) = j \mid X(0) = i, Z_1(t) > 0, \dots, Z_i(t) > 0\}$
π	prior distribution for X_o
E_u	the event of extinction below u, i.e. $X_{[u]} = (0, 0, \dots, 0)$
η	mean parameter for π
ζ	dispersion parameter for π when a beta-geometric distribution
q	generic WGD retention probability (also referred to as retention rate)[1]
K	Bayes factor in favor of $q = 0$ (Savage-Dickey density ratio)
μ_r	death rate for redundant genes in the DLF model
μ_{nr}	death rate for non-redundant genes in the DLF model
μ_1	death (loss) rate of a type 1 gene
μ_2	death (loss) rate of a type 2 gene
ν	x-functionalization rate (type 2 to type 1 transition)
$p_{ij}(k, l, t)$	transition probability $\mathbb{P}\{X(t) = (k, l) \mid X(0) = (i, j)\}$ for the two-type model
$f_{ij}(s_1, s_2, t)$	pgf for the transient distribution of the two-type model
$g_{ij}^{(u)}(\mathbf{s})$	joint pgf for the phylogenetic two-type model conditional on $X_{\rho(u)} = (i, j)$

Chapter 4

c_n	number of rooted trees with n leaves
$\mathcal{T}$	random tree topology (cladogram)
$C(\mathcal{T})$	clade set for $\mathcal{T}$
γ	generic clade (i.e. a generic element of $C(\mathcal{T})$ for some $\mathcal{T}$)
δ	generic split of a clade (i.e. an element of $\mathcal{P}(\gamma)$ for some γ)
$q_n(i)$	clade size distribution for a MBM
θ_γ	split distribution for clade γ in a CCD
$\#_\gamma$	number of possible splits of clade γ

Chapter 5

G	gene tree
ϕ	branch parameters
N	coalescent-effective population size (also N_e)
σ	gene-to-species map
$l_i(x)$	likelihood factor of site i in the context of EP
$q(x)$	EP global approximation
$q_i(x)$	EP site approximation for site i
$q_{-i}(x)$	EP cavity distribution for site i

[1] The letters q and p appear to be rather heavily overloaded throughout the present dissertation. However, the risk of confusion should be limited within any particular chapter.

$q_{\setminus i}(x)$	EP tilted distribution for site i
C_i	EP normalizing constant associated with site i
η_γ	natural parameter associated with clade γ in a CCD

Chapter 6

$\mathcal{G}_y$	set of gene trees compatible with sequence data or phylogenetic profile y
$\mathcal{R}$	reconciled gene tree
ρ	reconciliation map
$p_e(u,t)$	probabiity of lineage leading to gene tree node u passing through time point t along species tree branch e
$\phi_e(t, t+\Delta t)$	single-lineage propagation probability over a time slice of length Δt
$\psi_e(t, t+\Delta t)$	single-lineage represented duplication probability over a time slice of length Δt

1 Introduction

The present work aims to an understanding of genomic diversity from an evolutionary point of view, and this introductory chapter serves to delineate the scope of that endeavor, while situating the work presented in this dissertation within it. We will see that questions *about* evolutionary history, and our ability to learn about the evolutionary process *from* history, are (or at least should be) intimately related to *statistical* issues, and we will come to the thesis defended throughout this work; namely that the empirical study of evolution using genomic data, and the attempt to make sense of genomic data from an evolutionary perspective, are best conceived as statistical problems, where formal models of evolution are devised and confronted with data using Bayesian logic. We end by sketching, with broad brushstrokes, the kind of models we shall deal with in this dissertation.

1.1 Making evolutionary sense of genome data

1.1.1 Evolutionary biology and the genomic deluge

> It's a history book – a narrative of the journey of our species through time. It's a shop manual, with an incredibly detailed blueprint for building every human cell. And it's a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease.
>
> – Francis Collins[1]

After the human genome project was brought to (near) completion (Lander et al. 2001), it became clear that, despite ideological advertisements like the

[1] https://www.genome.gov/human-genome-project, last accessed May 22, 2022.

above, it did not solve *that* much. It did not yield a 'blueprint for building every human cell', nor cancer cure, and did not provide a 'narrative' of our evolutionary history. Doubtlessly, it has contributed to many new insights on these matters, but it should not be controversial to state that the net result was somewhat underwhelming. There seems to be no 'book of life' that we can read to solve all of biology. In fact, the main effect of the genomic revolution was to give birth to more questions[2], leading to the many '-omics' fields we have today. If something became crystal clear after the human genome project, it was that our ability to measure increasingly minute bits of matter vastly surpassed our ability to make biological sense of these measurements. We may sooth ourselves with the (questionable, to say the least) idea that all of biology is, somehow, *in there* (i.e., to use the book metaphor: there is a book, but it is written in a foreign language and unknown script), but that does not change the state of biology, only biologists.

Today, hundreds, if not thousands, of genomes have been sequenced to near completion from all over the tree of life. While these genome projects have contributed greatly to our appreciation of diversity at the genomic level, our inability to interpret all these data persists. One mode of interpretation, of making sense of these data, is making *evolutionary* sense of it, that is, to try to grasp how these data came to be, and why these (rather than different) data came to be, in the light of evolutionary theory. In other words, we can try to explain extant genomic diversity on the basis of evolutionary principles. This is a popular mode of interpretation, which has, since the so-called 'Modern Synthesis'[3], been increasingly considered as *the* mode of interpretation in biology – a tendency epitomized in the famous title of a somewhat less famous article of Dobzhansky: "Nothing in biology makes sense except in the light of evolution" (Dobzhansky 1973). This brand of evolutionism invites us to give an evolutionary explanation for every biological phenomenon. In Ernst Mayr's terms, to seek the 'ultimate' (as opposed to 'proximal') cause of biological phenomena (Mayr 1988). *Dually*, we may wish to use genomic data to make sense of evolution, that is, we may use genomic data to test theories and hypotheses about the evolutionary process. This is not the same as providing evolutionary explanations for observed genomic diversity, but is, or should be

[2] Which is of course a good thing, if, and only if, these happen to be the right questions.

[3] I am referring here to the 'second phase' of the Modern Synthesis, or the development of 'the synthetic theory of evolution', rather than its 'first phase', which corresponds to the synthesis of Mendelism and Darwinism (e.g. Provine 1971; Mayr and Provine 1980; Gould 2002). The former is associated with people like Huxley, Mayr, Dobzhansky, Stebbins and Simpson, whereas the latter is associated with the birth of theoretical population genetics and the illustrious names of Haldane, Wright and Fisher.

(I contend), closely related to it.

Influenced perhaps by its roots in natural history, such evolutionary explanations have often amounted to presenting a plausible historical narrative of adaptive divergence. In the polemic words of Michael Lynch:

> For the vast majority of biologists, evolution is nothing more than natural selection. This view reduces the study of evolution to the simple documentation of differences between species, proclamation of a belief in Darwin, and concoction of a superficially reasonable tale of adaptive divergence. (Lynch 2007)

In particular, much of the popular scientific literature on evolution takes this form. It is however not restricted to popular science. In evolutionary genomics this leads to questions of the form "Why has the genome of species *X* less genes for pathway *P* than species *Y*?", with answers in the form of a 'reasonable tale of adaptive divergence', e.g.: "Because species *X* evolved to live in habitat *A*, whereas species *Y* descends from a lineage that adapted to habitat *B*". Research papers with the generic title "Genome *X* provides insights in the evolution of *P*" often display these kinds of evolutionary tales. Here is a sample of recent literature:

> The chromosome-level genome assembly of the Japanese yellowtail jack *Seriola aureovittata* provides insights into genome evolution and efficient oxygen transport (Li et al. 2022)

> The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence (Huang et al. 2022)

> Chromosome-level pepino genome provides insights into genome evolution and anthocyanin biosynthesis (Song et al. 2022)

> The new *Haemaphysalis longicornis* genome provides insights into its requisite biological traits (Yu et al. 2022)

> Chromosome-level genome assembly of the dotted gizzard shad (*Konosirus punctatus*) provides insights into its adaptive evolution (B.-J. Liu et al. 2022)

As of May 2022, a search in the PubMed database reveals 351 hits for titles including 'genome', 'insights' and 'evolution', and almost every day a new article with similar-sounding title appears in our 'Recommended articles' list. Bibliometrics hence suggest that insights in (genome) evolution are accumulating at an unprecedented pace. We note that, whereas the generic title is

formed in a way suggestive of the second aspect of evolutionary genomics – i.e. using genomic data to make sense of evolution (supposedly, the genome provides insights in evolution, not the other way around) – it is at least equally, and probably more, the reverse direction of explanation that features in these papers. Indeed, first and foremost, these papers seek to fit genomic data in a historical evolutionary narrative.

Of course, this is a reasonable thing to do, and indeed, such evolutionary tales enable us to make sense of observed genomic diversity which remains unintelligible otherwise. However, as has been pointed out already a long time ago, for instance in the famous paper by Gould and Lewontin (1979) or Antonovics (1987)[4], and reiterated forcefully in recent times by people like Michael Lynch, there are considerable problems with the kind of adaptive storytelling that pervades evolutionary explanations for biological phenomena. The main issue lies not so much in these sort of speculative explanations themselves, which indeed *may* be true, but in the *scientific ideology*[5] that goes with them, which says that concocting a superficially plausible adaptive story finishes the job of the evolutionist. In that regard, as Gould and Lewontin (1979) noted, evolutionists often use *consistency* with evolution by natural selection as main criterion for assessing the scientific merit of an evolutionary explanation.

We would like, however, to bring evolutionary explanation in genomics, and dually, our study of the evolutionary processes affecting genomes, to the same standards as those adopted elsewhere in natural science. This means, constructing models on the basis of evolutionary theory, and testing these models by confronting them with data, introducing a *critical dimension* in evolutionary explanation. This, at the same time, brings the two aspects of evolutionary genomics closer together: to explain genomic diversity by means of models of evolution tells us something about the evolutionary process. In other words, coming up with a plausible adaptive tale should not be the goal of evolutionary biology, but rather the starting point, a hypothesis to be evaluated. To quote

[4] "The presentation of a confirmed theory of such broad scope led to a complacent acceptance and reduced evolutionary biology to everybody's toy and plaything. The ability to generate a simplistic speculation about some putative past selection process seemed to qualify anyone as an evolutionary biologist and, perhaps worse, led others to imagine that this is what professional evolutionary biologists do." – Antonovics (1987)

[5] Here we would like to refer to a splendid article by French philosopher Georges Canguilhem (Canguilhem 1977), from whom we learned to appreciate the concept of scientific ideology. We need to be careful here however, qualifying something prematurely as ideology is subscribing to another one. "La qualification comme idéologie d'un certain assemblage d'observations et de déductions, est postérieure à sa disqualification comme science par un discours qui délimite son champ de validité et qui fait ses preuves par la cohérence et l'intégration de ses résultats." (Canguilhem 1977, p41–42)

from Lynch (2007) again: "A strong belief in cells does not make one a cell biologist, and a strong belief in Darwin's principle of natural selection is not a sufficient condition for understanding evolution".

1.1.2 Evolutionary history, models and statistics

> Wat voor de meeste beoefenaren van exacte wetenschappen vanzelf spreekt, maar toch nog wel eens herhaald mag worden, omdat dit inzicht niet zeer verbreid is onder de velen die zulke vakken als sociologie, economie en geschiedschrijving beoefenen, is dit: de natuur is altijd compleet aanwezig, het verleden nooit.
>
> – W. F. Hermans (1981)[6]

The goal of providing evolutionary explanations for observed diversity is intimately related to the problem of reconstructing evolutionary history. Relatedly, to study the evolutionary process, we cannot, usually, evaluate our theories and models in the same way as we can, for instance, in mechanics, but have to rely on fragmentary observations of a long historical process. Indeed, evolution is in a way, like history, *never completely present*. Yet, the empirical study of evolution need not be relegated to the historian[7]. Indeed, the general principles and laws of evolutionary biology allow us to construct *models of evolution* which generate historical predictions[8], so that we can confront our models with empirical data. This brings us quickly into *statistics*, which we take to be the general attempt to formalize empirical methods in science, or more plainly, the science of confronting models with empirical data.

The reconstruction of evolutionary history is, roughly, the subject matter of the field known as *phylogenetics*, which has indeed gradually, but with consider-

[6] W. F. Hermans (1921 – 1995) was a Dutch scientist and (prolific) author of novels, poetry and essays. A pessimistic positivist with an academic career in physical geography, many of his essays reflect on the sciences. The quoted passage can be freely translated as: "What is evident for most researchers in the exact sciences, but might nonetheless be repeated from time to time, as this insight is not widespread among those who work in fields like sociology, economics or history, is the following: nature is always completely present, the past never is."

[7] We do not wish to suggest that the natural historian is superfluous, and has no claim to important contributions in (evolutionary) biology. That would be preposterous, recognizing that the whole field owes its very existence to the study of natural history. To recognize and accurately describe the diversity of living forms, and to conceive hypotheses on the evolutionary causes of that diversity was, and remains, needless to say, a vital component of the whole enterprise.

[8] This is of course an awkward term when what we are really talking about are *postdictions*. I will however stick to the common employment of the word prediction in the statistical sense – i.e. the use of a model to say something about unobserved quantities, whether or not these characterize past, present or future observables.

able pain (see the reminiscences in Felsenstein 2001), transformed into a statistical business. Not unlike how general physical principles inform historical inferences in geology, the principles of evolutionary genetics (Mendelism, common descent, mutation, recombination, population genetic constraints, *etc.*), together with information from previous observations, inform evolutionary models and inferences in phylogenetics. Using models of evolution, inference of the 'incompletely present' evolutionary history from observed data is no different from inference for unobserved quantities elsewhere in statistics.

If our goal is to study the evolutionary process *per se*, rather than to reconstruct particular evolutionary histories, we take a similar viewpoint. Indeed, to study evolution, in our view, amounts to devising formal models of the evolutionary process of interest and confronting them with empirical data[9]. Note, however, an important reversal that takes place here: while the phylogenetic concern is largely driven by the desire to make evolutionary sense of *given* data, the focus here is on evaluating models using empirical data, which suggests the data to come afterwards, somehow. Indeed, in theoretical matters, constituting what is considered *as data* is part of the scientist's job. It is, however, much more common to be confronted with empirical data for which a formal model is to be constructed. That is, the models we shall be considering and questions we shall be asking tend to be *conditioned by the data* from the outset, at least to an important extent. In particular, with regard to questions in genome evolution, the sorts of questions we succeed in asking are strongly conditional on the data available (essentially text files of nucleotide sequences). While this may sound as a platitude, or at least as inevitable, it is important to be aware of such constraints. Furthermore, it tends to bring us even more into statistics, that is, in its *practical* aspects, where not only the inferential task gets its distinctly statistical flavor, but also the devising and revising of the very models of interest becomes increasingly a statistical endeavor.

1.1.3 Probabilistic models of evolutionary processes

Our goal is hence to devise formal models of the evolutionary process, but what do such models look like? Clearly, there is no equation, like $y = gt^2/2$

[9]The sociologist Charles Tilly puts the reasons for preferring such an approach clearly: "*Post hoc* interpretation of data minimizes the opportunity to recognize contradictions between arguments and evidence, while adoption of formalisms increases that opportunity. Formalisms blindly followed induce blindness. Intelligently adopted, however, they improve vision. Being obliged to spell out the argument, check its logical implications, and examine whether the evidence conforms to the argument promotes both visual acuity and intellectual responsibility." (Tilly 2004) (quoted in Gelman and Shalizi 2013)

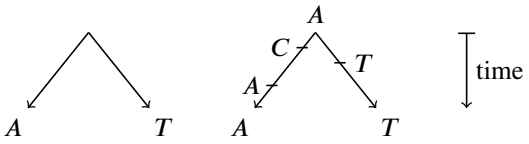


Figure 1.1: Evolutionary processes do not preserve all historical information in present-day data. On the left, the present-day (observed) pattern of states for a pair of homologous sites is shown, whereas on the right, the evolutionary history with all substitutions that did occur is shown. While under the assumption of common descent (homology), we know that at least one substitution must have occurred, we cannot tell how many did in fact occur in the evolutionary history from present-day observations. Note that even if we knew the ancestral state, we would still not be able to tell.

in classical mechanics, which can tell us what nucleotide sequence we are supposed to observe after one million years of evolution given some suitable initial conditions. The problems we meet are twofold. On the one hand, our models will always be unable to take into account all relevant factors which influence the evolutionary process. Clearly, the UV radiation strength in some place on Saturday the 21st of November *anno* 20 million years (My) B.C. will have an influence on whether or not a particular ancestral sequence at that time and place in history underwent an A to G mutation, but it is of course impossible to take this up in our model. Not only is it practically impossible (we do not have the relevant measurements, nor the capability, computationally and conceptually, to take all this information into account), it is also undesirable. Indeed, we do not consider these sorts of data relevant for evolutionary explanations at such time scales, and would prefer a theory or model which does not require such detailed information. On the other hand, we have the problem that the evolutionary processes erase the historical signals they leave in extant data over time. Indeed, when we observe a different nucleotide at a homologous site in two homologous sequences (fig. 1.1), how can we know whether one, two, or more mutations happened since their divergence?

Our concerns are hence analogous to those which have led physicists to adopt probabilistic models in the field known as statistical mechanics. It is not only the practical impossibility of tracking the dynamics of individual particles in large physical systems which motivates, for instance, the kinetic theory of gases, but also the more positive conviction that the detailed dynamics of these particles should not feature explicitly in a scientific explanation for the thermodynamic behavior of these systems. Hence, as a result of these considerations, our models will be stochastic, and if they are not, they will typically be conceived as approximations of stochastic models (chapter 2 provides some

examples of the latter situation). By deliberately ignoring certain aspects of the process and taking them up in the model as having some random structure, we hope to capture relevant aspects of the problem, while acknowledging variation due to causes whose effect is uncertain. We illustrate the sort of probabilistic models we shall be using by presenting an important class of models ubiquitously used in phylogenetics in the following example[10].

Example (CTMC models of sequence evolution). An important class of probabilistic models we shall be using often, but not deal with very explicitly, are continuous-time Markov chain (CTMC) models of sequence evolution. The most common such models describe the evolution of a *single site* in a sequence over time as a Markovian stochastic process $\{X(t)\}$ on a finite state space. Let $p_{ij}(t) = \mathbb{P}\{X(t) = j | X(0) = i\}$ be the transition probability of going from state i to state j over a time span t. The Markov assumption for a finite state space entails the Chapman-Kolmogorov identity

$$p_{ij}(t) = \sum_k p_{ik}(t - \Delta t) p_{kj}(\Delta t)$$

The process is defined by its infinitesimal transition rates

$$q_{ij} = \lim_{\Delta t \downarrow 0} \frac{p_{ij}(\Delta t) - p_{ij}(0)}{\Delta t}$$

for every pair of states i and j. A CTMC model for nucleotide substitution is hence defined by an infinitesimal rate matrix (generator) with, in general, the following form

$$Q = \begin{bmatrix} q_{AA} & q_{AT} & q_{AC} & q_{AG} \\ q_{TA} & q_{TT} & q_{TC} & q_{TG} \\ q_{CA} & q_{CT} & q_{CC} & q_{CG} \\ q_{GA} & q_{GT} & q_{GC} & q_{GG} \end{bmatrix}$$

The rate matrix Q is related to the transition probability matrix $P(t)$ intuitively in that, for small Δt, $P(\Delta t) \approx I + Q\Delta t$. The model entails that, when currently in state i, a substitution occurs after an exponentially distributed waiting time with mean $1/\sum_{j \neq i} q_{ij}$, upon which the chain (nucleotide site) takes a new state $j \neq i$, with probability proportional to q_{ij}. Different assumptions about the

[10] Throughout, we shall designate the start of an example by **Example (title)** and its end by a □ symbol. Hereby we apologize to those mathematicians who insist on reserving the □ for proofs. There will be no formal proofs in this dissertation, so there is no risk of confusion. The goal of our 'example' sections is to separate some of the technical content and concrete analyses from the main argument, while nevertheless including these detailed treatments at the appropriate place with respect to the latter.

molecular evolutionary process can be encoded in the structure of Q. For instance, the simplest possible model (the so-called Jukes & Cantor (JC) model (Jukes and Cantor 1969)) assumes that, conditional on a substitution occurring, all substitutions (from any given state) are equally likely, whereas Kimura's model (K80) (Kimura 1980) assumes that upon substitution, transitions (in the molecular biological sense, i.e. purine to purine or pyrimidine to pyrimidine substitutions) occur with a different probability than transversions.

We can derive a general solution for the transition probability matrix over any time span t by considering the Chapman-Kolmogorov equations

$$p_{ij}(t + \Delta t) = \sum_k p_{ik}(t)[\delta_{jk} + q_{kj}\Delta t + o(\Delta t)]$$

(where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise) or, in a single matrix expression,

$$P(t + \Delta t) = P(t)(I + Q\Delta t + o(\Delta t))$$

Subtracting $P(t)$ from both sides, dividing by Δt, and taking $\Delta t \downarrow 0$, we get the Kolmogorov backward differential equation

$$\frac{dP(t)}{dt} = P(t)Q$$

Which gives us the intuitive solution $P(t) = \exp(Qt) = I + Qt + (Qt)^2/2! + (Qt)^3/3! + \ldots$. Being able to compute transition probabilities for the process is a prerequisite for efficient statistical inference under the model. □

Clearly, this is a very crude model of sequence evolution, glossing over the many subtleties of the mutational process, but it is hard to imagine statistical phylogenetics without it. The model admits asking several basic questions, of the sort "Given evolutionary rates Q and initial state A, what is the expected number of substitutions over a time span Δt? What is the probability that the end state is T? What is the probability that more than five substitutions happened? What is the expected total amount of time spent in state G? …" Already, such quantities start to sound interesting, as we can imagine how they can feature in scientific inferences about the processes of nucleotide substitution from observed sequence data.

But note that the CTMC model of sequence evolution in itself cannot tell us much about observed sequence data. Indeed, given some sequence where we assume the sites to evolve independently in accordance with such a CTMC model, we cannot tell much about the sequence nor the likely parameters of

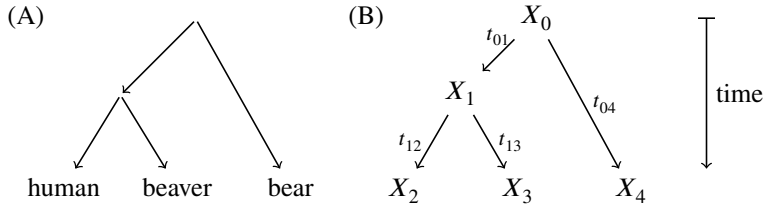


Figure 1.2: (A) An example three-taxon phylogenetic tree. (B) A simplified representation of the PGM for the phylogenetic CTMC model associated with the tree in (A).

the evolutionary model. To actually do something interesting with these models, we need multiple realizations of the process. However, in an evolutionary context, we do not, usually, have the typical kind of statistical sample consisting of some number of independent realizations of the process. Instead, we make observations of multiple non-independent realizations of the process in different taxa (individuals, populations, species, …), related by some evolutionary history. The simplest case, under the assumption of plain common descent, is that the observed data are related by a tree structure, a *phylogenetic tree* or *phylogeny*, which represents ancestor-descendant relationships. The usual strategy in phylogenetics is then to embed a model for an evolutionary process in a phylogenetic tree structure which describes the evolutionary relationships of the observed data. We illustrate this for the CTMC models of sequence evolution in the following example.

Example (phylogenetic CTMC model). Assume we sample nucleotide sequence data for three taxa, say, for definiteness, human, beaver and bear. Further assume we have succeeded in identifying the homology relationships among the different sites in the sequences, that is, those groups of sites which descend from a common ancestral site. Assuming a one-to-one correspondence between homologous sites in the different sequences, we obtain a *sequence alignment*, which can be represented as a matrix wherein each column represents a site and each row represents a taxon. We assume that the taxa are related by a phylogenetic tree, depicted in fig. 1.2 (A), and assume that, for each site, the CTMC model operates independently along different branches of the tree from past to present, where at each bifurcation in the tree, two independent CTMC processes are started with as initial state the end state of the parent branch. The resulting evolutionary model is called a *phylogenetic CTMC* model (e.g. Höhna et al. 2014). More explicitly, for a given node v in the tree, we denote by $X_v \in \{A, T, C, G\}$ the random variable which repre-

sents the state at that node. The phylogenetic tree together with the X_v define a probabilistic graphical model (PGM) (Jordan 2003; Höhna et al. 2014) shown in fig. 1.2 (B). When node v has parent node u, the phylogenetic CTMC model entails that the probability distribution for v is given by

$$\mathbb{P}\{X_v = j | X_u = i\} = p_{ij}^{(v)}(t_{uv}) = \left[\exp(Q^{(v)} t_{uv})\right]_{i,j}$$

where $p_{ij}^{(v)}(t)$ are the transition probabilities for the ordinary CTMC which is assumed to operate along the branch leading to node v (see above). Note that, in general, we may assume different Markov processes for the different branches of the phylogeny, so that in the three-taxon example we would have four different rate matrices defining the phylogenetic CTMC process. The basic assumption of the phylogenetic CTMC, captured by its PGM representation, entails the following conditional independence relationship for a node u with descendant nodes v and w

$$\mathbb{P}\{X_v = i, X_w = j | X_u = k\} = \mathbb{P}\{X_v = i | X_u = k\}\mathbb{P}\{X_w = j | X_u = k\}$$

This completely defines the probabilistic structure of the model and can be used to compute various expectations given a suitable parameterization. For instance, given some prior distribution on the root state and the rate matrices for the different branches, one can, by appropriately marginalizing over ancestral states (using the so-called pruning or variable elimination algorithm), compute the probability of an observed column in the sequence alignment under the model. □

The basic model outlined in the preceding example underlies that part of statistical phylogenetics that is most used in practice. In a statistical context, the model admits asking questions like: "What is the probability to observe *A* in human, *T* in beaver and *T* in bear for a putative homologous site assuming the phylogenetic tree in fig. 1.2?". The latter question shows how the model provides us with a probability yardstick to assess hypotheses as to which phylogeny is somehow most likely to represent the true branching pattern of the taxa under consideration. The latter is sometimes considered the core objective of phylogenetics in general, but it should be clear that the basic model admits addressing many other questions about the evolutionary process (e.g. are the rates of evolution similar across the branches of the tree? What is the likely ancestral state?).

Before moving on to statistics, a few more words about probability must be said at this stage. It is quite common for authors in evolutionary biology to as-

sert that evolution *is* a stochastic process[11], but this conflates the evolutionary process with our knowledge thereof. By adopting stochastic models, we do not necessarily subscribe to a view that evolution would somehow *be* stochastic, only that we are in a state of uncertainty when it comes to knowledge about these processes. Note that we are not ready to discount the statement that evolution would be inherently stochastic (that would depend strongly on just what *precisely* we understand 'evolution' to consist of and what we take probability to mean). We simply wish to note that this need not be relevant for our adoption of stochastic models, much in the same way as it does not matter whether or not the universe is deterministic (or whether that is a meaningful question) when we use the predictions of thermodynamics. Organic evolution is no more a problem for Laplace's demon than celestial mechanics. Probability is for us then, first and foremost, an epistemic concept, serving as a quantitative measure of a state of uncertainty about a logical proposition.[12] Our adoption of probabilistic models merely signals a willingness to reason quantitatively in the face of uncertainty, and involves no positive statement about any kind of non-determinism[13].

1.2 Bayesian statistics

Establishing models, probabilistic or otherwise, does not, in itself, have anything to do with empirical data. As we have already noted above, confronting models with empirical data is the subject domain of *statistics*. Again, we point out that, in practice, statistics also takes up a role in the reverse direction, which consists of devising (relatively generic) models to make sense of empirical data (think of linear regression, ANOVA *etc.*), which are widely adopted across the sciences. In other words, practical statistics is not only about inference and criticism for given scientific models, but also about the

[11] For a list of illustrative quotations, the reader may consult the appendices of Stoltzfus (2021). Here is one from Lynch (2007), from whom we have been quoting already quite a bit: "Evolution is an inherently stochastic process, starting from the chance events that produce single mutations in single individuals and proceeding through a series of fortuitous steps that gradually lead to the spread of those mutations to every member of the descendent population."

[12] For a compelling argument as to why probability as usually formalized in mathematics (i.e. in measure theoretic terms) corresponds to an adequate quantitative measure for the plausibility of propositions, we refer the reader to Jaynes (2003) and Cox (1961).

[13] Indeed, as Harold Jeffreys stressed, any theory of probability which would make positive claims about the world of this sort would be highly suspect: "If the rules of themselves say anything about the world, they will make empirical statements independently of observational evidence, and thereby limit the scope of what we can find out by observation. If there are such limits, they must be inferred from observation; we must not assert them." (Jeffreys 1961)

art of establishing certain types of models of nature itself.[14] In the present section we focus however on the former two aspects: inference and criticism.

Several 'schools' of statistics exist. These are often put in the two opposing camps of Bayesians and frequentists, with sometimes a third camp of Fisherian likelihoodists included, but even with the latter extension, the classification ignores many of the subtleties in the different positions practitioners hold (in particular, it ignores the numerous scientists with a pragmatic "will use whatever does the job" approach, and it ignores a perhaps even more important split, between those in favor of 'null hypothesis significance testing' (NHST) and those who believe NHST is fundamentally misguided). Our remarks on probability in the previous section, however, already appear to put us in the Bayesian camp, and indeed, we defend a Bayesian approach towards statistics. Both schools make use of probability and probabilistic models of course, but they diverge in their view on just what the role of probability is in making statistical inferences.

1.2.1 Frequentist and Bayesian statistical inference

Let us first briefly consider what has long been, and still is (but less so), the dominant school of statistics (while remaining aware of our overly polarized classification). In the frequentist context, a parametric probabilistic model $\mathcal{M}(\theta)$, with $\theta \in \Omega$ a generic parameter, is assumed as a model for some observed data set $y = (y_1, y_2, \ldots, y_n) \in \mathcal{Y}$. Typically the data are assumed to be iid (independent and identically distributed) draws from $\mathcal{M}$, and the frequentist statistician imagines y as a realization of a random vector $Y = (Y_1, Y_2, \ldots, Y_n) \sim_{\text{iid}} \mathcal{M}(\theta)$. Now we consider the situation where θ is unknown, and the goal is to come up with an estimate of θ (or some function thereof), call it $\hat{\theta}$, based on the observed data y. To come up with such an estimate, the frequentist considers some function $t : \mathcal{Y} \to \Omega$ so that $\hat{\theta} = t(y)$ and studies the behavior of $\hat{\Theta} = t(Y)$ as an estimator of θ. Writing the probability density function for observed data y as a function of the parameter $\ell(\theta, y) = p(y|\theta)$, an important example of an estimator is the maximum likelihood estimator (MLE) $\hat{\theta} = \text{argmax}_\theta \ell(\theta, y)$. The probabilistic properties of $\hat{\Theta}$ are then considered as informative about the statistical properties of the particular estimate $\hat{\theta}$ as an estimate of θ. Typical properties of interest, which

[14] Here we gloss over another aspect of statistics, that of (experimental) *design*. In the context of phylogenetics and much of evolutionary biology in general, we do not have the luxury that we can worry much about design (although the issues related to *taxon sampling* are arguably related to design questions).

one would like to keep as small as possible, are the *bias* and *variance* of the estimator (Lehmann and Casella 2006; Efron and Hastie 2021).

For instance, considering a univariate model, if we happen to know that $\mathbb{P}\{\hat{\Theta} > \theta + z\} = 0.1$, then we can say that, assuming y is indeed a realization of Y under the model $\mathcal{M}$, $\hat{\theta}$ will exceed the true parameter value by more than z in 10% of the cases if this experiment were repeated. The latter is the source of 'frequency' in frequentism: under the frequentist viewpoint we are invited to think of y as a realization of a random experiment which could, theoretically, be repeated indefinitely. We assess the statistical performance of some procedure, embodied by t, by considering how often it would lead us to make certain errors (such as reporting an estimator which exceeds the true value by more than z) if we were to repeat the experiment and statistical procedure. Note, however, that many of the probabilistic properties of an estimator that we might be interested in will depend on the true value of θ, as for instance in our example above, where we considered the property $\hat{\Theta} > \theta + z$. This is a major source of difficulty, and mathematical ingeniousness, in frequentist statistics.

Frequentist statistics is hence largely limited to making statements *before* having observed any data, answering questions like: If we were to use this estimator, what accuracy could we expect? If the hypothesis is true, how likely is it that a random data set will lead us to reject it? (Jaynes 2003). The reason for this is ultimately the reluctance to bring knowledge about parameters on the same (probabilistic) footing as knowledge about the natural processes we are seeking to model. There is, of course, much more to frequentist statistical practice. In particular we gloss over the many important differences within that practice. As far as *inference* is concerned, however, this short discussion will suffice for our purposes of comparing the two main approaches towards statistics.

The Bayesian approach, originated by Thomas Bayes and Pierre-Simon Laplace, is characterized by a more straightforward use of probability to make inferential statements (and decide on acts[15]). Under the Bayesian conception, probability serves as a quantitative measure of uncertainty about some proposition, such as the proposition '$\theta \in [0, 1]$'. The initial task of

[15]Many works on Bayesian statistics motivate the Bayesian approach explicitly in a decision theoretic context (e.g. Savage 1972; Robert 2007; Bernardo and Smith 2009). Bernardo and Smith (2009) show clearly how reporting a posterior probability distribution or uncertainty interval can be seen as a special case of deciding on an act in the face of uncertainty. We shall have no use for decision theory in the present work however, and will not make explicit the loss functions we use implicitly when reporting some Bayes estimator.

a Bayesian statistician is to come up with a joint probability model for the data y and parameter θ with distribution $p(y, \theta)$, which formally connects the observations to the unknowns. Elementary probability theory implies that $p(y, \theta) = p(y|\theta)p(\theta)$, so that the construction of a joint probability model decomposes in an encoding of our state of knowledge about the parameter θ, marginalized over possible data, i.e. the *prior distribution* $p(\theta)$, and a choice of *sampling distribution* $p(y|\theta)$.[16] By Bayes' theorem, we can quantify our state of uncertainty about the parameter conditional on having observed data y by evaluating the *posterior distribution*

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta)$$

Using the posterior distribution, we can compute the posterior probability of propositions of interest to make inferential statements. For instance, for $\theta \in \mathbb{R}$, we can compute the posterior probability for a proposition like '$\theta \in [0, 1]$' as $\mathbb{P}\{\theta \in [0, 1]|y\} = \int_0^1 p(\theta|y)d\theta$. Such a posterior probability expresses the post-data degree of belief we *should* have that the statement is in fact true, *if* we assume the joint probability model $p(y, \theta)$ to be adequate[17].

Clearly, what the Bayesian and frequentist approaches have in common is their assumption of a probabilistic model for the data-generating process. The latter will typically be the same either in a Bayesian or frequentist analysis. The main difference between the two approaches is really in the conception of just what the role of probability in statistics is. For the frequentist, probability measures something about a random process, typically a limiting frequency of some event in an infinite number of repeated trials. This is a measure for something which is supposed to *be* random independent of any state of knowledge. The data are modeled *as if* they were the product of some random process $\mathcal{M}(\theta)$ (e.g. a CTMC model of sequence evolution), and probabilities only 'exist' within the confines of such a random process. Clearly, the parameter θ is not random in any such sense, and it would make no sense to model it as such, since it is supposed to have a definite value which could not have turned out otherwise. Hence, it makes no sense, under a frequency conception of probability, to assign a probability distribution to it.

[16] We follow the abuse of notation adopted (for instance) in Gelman et al. (2013) where all distribution functions that appear in some Bayesian inference problem are symbolized by $p(\cdot)$ and $p(\cdot|\cdot)$. Different p symbols may therefore reflect different distribution functions in the same expression. We use the same notation for discrete and continuous random variables.

[17] This already points towards an approach for assessing whether our models are adequate: if our *actual* degree of belief for some statement does not correspond to what Bayes' theorem tells us it *should* be, this signals issues with the model (or a calculation error).

For many a Bayesian (including your present author), the idea of randomness independent of a state of knowledge fails to be intelligible, and would bear on philosophical and physical problems far beyond the usual scope we attribute to statistics.[18] However, if probability is indeed considered a quantitative measure of 'reasonable degree of belief', it makes perfect sense to encode our knowledge about θ in a probability distribution, and once we have done so, Bayes' theorem, or the principle of *inverse probability* as it is sometimes called, provides us with a principled approach to exhibit just how our knowledge about θ changes in light of the data. Note that, contrary to what is sometimes claimed, the Bayesian does not somehow assume that there is no true value of θ and that θ itself *is* somehow random. This conflates the frequentist conception of probability with the Bayesian, i.e. it conflates our state of knowledge about θ, expressed as a probability measure on a suitable space, with θ itself.

For us then, frequentist statistical approaches often appear reasonable, but we have no real use for them. Indeed, since our aim is to quantify our post-data uncertainty about some unknown quantity in the form of a probability distribution over the space of possible values for that quantity, the way to get there is by assuming a joint probability model for all the unknowns and the data, and rely on conditional probability to obtain the posterior for the quantity of interest. As we noted above, the joint probability model decomposes in a sampling distribution and prior distribution, the former of which is typically shared with a frequentist take on the same problem. Criticism of the Bayesian approach hence focuses on the latter, i.e. the prior distribution. In particular, an often repeated reproach is that the prior would somehow be 'subjective', and hence ruin our aspirations towards an 'objective' statistical analysis. We regard this as a rather misguided criticism, in that the whole statistical endeavor is about 'subjective' models which are used to probe the external world and make sense of empirical data. More concretely, in typical applications, the sampling distribution is at least as suspect as the prior regarding the perceived issues with 'objectivity.'[19] The choice of the prior distribution is no less a modeling question than the choice of sampling distribution, and it is part of the statistical scientist's modeling effort to specify just what, if anything, it is we are ready to assume concerning the unknown quantitaties that feature in the model. Furthermore, many non-Bayesian methods, can be interpreted as Bayesian methods with some implicitly assumed prior.

[18] In the words of de Finetti: "probability does not exist" (de Finetti 1974).

[19] "It is perhaps merely an accident of history that skeptics and subjectivists alike strain on the gnat of the prior distribution while swallowing the camel that is the likelihood." (Gelman and Robert 2013)

1.2.2 Bayesian inference in practice

The ultimate merit of Bayesianism is that it provides a simple method (in theory) for statistical inference, enabled by its endorsement of an epistemic concept of probability. Indeed, the statistical challenge, from a Bayesian point of view, lies not so much in the problem of inference, which is dealt with automatically and does not require *ad hoc* devices as in frequentism. By answering scientific questions directly (as opposed to giving answers involving hypothetical repeated experiments) and providing an automatic method to do so, Bayesianism gives statistics back to the scientist, allowing one to focus on the models and their adequacy to the data. In practice, however, the automatic inference method provided by Bayes' theorem often presents a considerable *computational* challenge, which, historically, presented a serious problem for successful application of Bayesian methods to all but a handful of trivial problems.

Today, powerful algorithms and computers enable practical applications of Bayesian inference to complicated problems (see Appendix A for an overview of relevant methods for the present work). The solution of an inference problem for a praticing Bayesian typically consists of a sample of N points from the posterior distribution, which can then be employed for approximating expectations with respect to the posterior distribution using Monte Carlo methods. For instance, let $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ be a sample from the posterior distribution $p(\theta|y)$, with $\theta^{(i)} \in \mathbb{R}$, the posterior mean value of θ can be approximated by

$$\mathbb{E}[\theta|y] = \int \theta p(\theta|y) d\theta \approx \frac{1}{N} \sum_{i=1}^{N} \theta^{(i)}$$

Probability is of course equivalent to the expectation of an indicator, so that posterior probabilities about certain inferential statements, for instance that θ lies in some set $A \subseteq \mathbb{R}$, can be derived in the same way, e.g.

$$\mathbb{P}\{\theta \in A|y\} = \mathbb{E}[\mathbb{1}_A(\theta)|y] \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_A(\theta^{(i)})$$

We illustrate the problem of statistical inference for a parametric model by reconsidering our CTMC model of sequence evolution.

Example (distance estimation). Consider two homologous DNA sequences of equal length $N = 50$, aligned in a data matrix y so that homologous char-

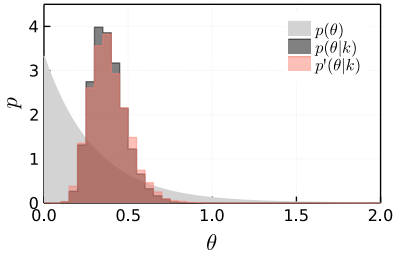


Figure 1.3: Prior and posterior distributions for the distance $\theta = 3\lambda t$ under the Jukes & Cantor model with $k = 15$ different sites for a pairwise sequence alignment of length $N = 50$. $p(\theta)$ is the exponential prior with mean 0.3, $p(\theta|k)$ shows a sample from the posterior for the model with the exponential prior and $p'(\theta|k)$ shows the posterior for a model with an improper prior for $\log \theta$. For references to color in figures, we refer the reader to the digital version of this dissertation.

acters appear in the same column of y:

$$y = \begin{bmatrix} \texttt{AGACTTGAATCATCTTGTGGTATAGGTCGTGGTGCCGAGTGGTCCCTAGC} \\ \texttt{AGACTTGGCGGGACTTGCACTATAGGTCGTGGTGCGCGATACTCCCTAGC} \end{bmatrix}$$

We assume the Jukes & Cantor model of sequence evolution (see above) and aim to estimate its parameter based on the observed sequence data. To do so, we shall need the relevant sampling distribution. One can show (see e.g. Yang (2006)) that the transition probability matrix has the following closed form

$$P(t) = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix} \quad \text{where} \quad \begin{aligned} p_0(t) &= \tfrac{1}{4} + \tfrac{3}{4}e^{-4\lambda t} \\ p_1(t) &= \tfrac{1}{4} - \tfrac{1}{4}e^{-4\lambda t} \end{aligned}$$

Where the non-diagonal elements of the rate matrix Q are all equal to λ. The expected number of substitutions at a site over a time t is equal to $\theta = 3\lambda t$ under this model. The latter quantity, which involves a product of a substitution rate and time interval, is, by physical analogy, usually called a *molecular distance*, and serves as an alternative parameter for the model. The reversibility of the process entails further that the probability of observing this pair of sequences is the same whether one considers them to have diverged from a common (unknown) ancestral sequence a time $t/2$ in the past or when one assumes the one to be the ancestor at a time t in the past of the other.

Let $k = 15$ be the number of observed differences between the two sequences. Our goal will be to estimate the expected number of substitutions per site $\theta = 3\lambda t$ given k observed differences. From the symmetry of the JC model, it is easy to see that k is a sufficient statistic for the model, that is, $p(y|k)$ is independent of θ, so that $p(y, k|\theta) = p(y|k)p(k|\theta) \propto p(k|\theta)$. To conduct inference, we hence only need the sampling distribution for k. The latter can be straightforwardly obtained from the transition probabilities of the JC model

$$p(k|\theta) \propto \left(\frac{1}{4} + \frac{3}{4}e^{-4\theta/3}\right)^{N-k} \left(\frac{1}{4} - \frac{1}{4}e^{-4\theta/3}\right)^{k}$$

The posterior distribution for the distance is then $p(\theta|k) \propto p(k|\theta)p(\theta)$. Let us assume an exponential prior density $p(\theta)$ with mean 0.3. A sample of size 10000 from the posterior distribution then provides a posterior mean estimate for the expected number of substitutions per site of $\hat{\theta} = 0.38$ (MCSE < 0.01) with 95% posterior uncertainty interval of $(0.21, 0.62)$. This means that, after having oberving the data, we assign 95% probability mass to the statement $\theta \in (0.21, 0.62)$, or, in other words, we are 'pretty sure' (or 'quite confident') that $\theta \in (0.21, 0.62)$, assuming the model.[20] When using an improper prior, $p(\log\theta) \propto 1$, the posterior is almost indistinguishable from the analysis with an exponential prior distribution (fig. 1.3). For comparison, the MLE is $\hat{\theta}_{\text{ML}} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{k}{N}\right) = 0.38$. □

Note that our choice for the prior distribution in the above problem was somewhat arbitrary. We know θ is positive, and that it should not be *too* far from the observed proportion of different sites k/N, but besides that, we have no compelling reasons to select a particular prior distribution. However, in the above example, choosing two different priors compatible with the little prior information we have does not affect our posterior inferences substantially, so it would be silly to concentrate our criticism on the prior assumptions while the sampling distribution (JC model) is a much more substantial, and potentially problematic, modeling assumption in this case.

[20] Note that the expected number of substitutions per site for a particular data set with a proportion k/N of different sites, which we could write as $\mathbb{E}[\theta_y]$, is necessarily larger than k/N. The θ parameter we estimate here, however, reflects the expected number of substitutions per site for a random sequence simulated from the CTMC model, of which our observed data is (granted the modeling assumption) but one example.

1.2.3 Statistical criticism

> As Rubin put it, "Bayesian" is an approach for making statistical inferences, "Frequentist" is an approach for evaluating statistical inferences."
>
> – Andrew Gelman

Statistics is about more than inference, which corresponds merely to its deductive aspect. Indeed, conditional on an assumed model, we use the logic of conditional probability to arrive at posterior inferences (and we can use the whole of mathematical probability theory, and many of the frequentist's results, to help us in this) – and that's it as far as inference is concerned. This procedure does not, however, tell us anything about whether our model assumptions are in fact reasonable. When we are analyzing time series data for a falling object using a linear model, there is nothing in Bayes' theorem that will tell us we are being silly. Bayesian inference alone cannot tell us anything about the model, but is restricted to answering questions conditional on the model. The model is the constraint which makes inference possible.

One can, of course, always try to embed the assumed model in a bigger model. For instance, in the sequence evolution example, we may consider two rivaling models: the JC model $\mathcal{M}_1$ and the K80 model $\mathcal{M}_2$. If we succeed in assigning prior probabilities to the two models, Bayesian inference for the posterior probability of either model is straightforward in principle

$$p(\mathcal{M}_1|y) = \frac{p(y|\mathcal{M}_1)p(\mathcal{M}_1)}{p(y|\mathcal{M}_1)p(\mathcal{M}_1) + p(y|\mathcal{M}_2)p(\mathcal{M}_2)}$$

If we regard the two models as forming a 'null hypothesis' – 'alternative hypothesis' pair, one can use the *Bayes factor* as a device for doing a sort of Bayesian hypothesis test. The posterior odds in favor of $\mathcal{M}_2$ is

$$\frac{p(\mathcal{M}_2|y)}{p(\mathcal{M}_1|y)} = \frac{p(y|\mathcal{M}_2)}{p(y|\mathcal{M}_1)} \frac{p(\mathcal{M}_2)}{p(\mathcal{M}_1)}$$

where $p(y|\mathcal{M}_2)/p(y|\mathcal{M}_1)$ is the Bayes factor in favor of $\mathcal{M}_2$. The Bayes factor measures the amount of evidence the data bears on a hypothesis , i.e. it is the factor which transforms the prior odds into the posterior odds (Jeffreys 1961; Kass and Raftery 1995; Jaynes 2003).

However, this bigger model hardly solves our problem. Indeed, by comparing two models, we have learned nothing about how adequate either is as a model for the observed data. By evaluating posterior model probabilities, we do not

assess the adequacy of either model for the data, but only their relative adequacy. In addition, it is not straightforward to interpret $p(\mathcal{M}_i)$ and $p(\mathcal{M}_i|y)$, the latter having an interpretation as the posterior probability that $\mathcal{M}_i$ is the true data-generating model conditional on the assumption that either one of the models is indeed the true model. Now, Box's famous aphorism in mind, we do not usually like to think of our models as true or false when assessing their adequacy, as they are usually deliberate simplifications of a (hypothetical) more complicated model for the natural system under study. The Bayes factor has similar issues, not measuring the evidence for a certain model in the data *as such*, but measuring the evidence for a model relative to another *conditional* on the bigger model. Of course, this is only to be expected, and to lament this would only highlight how vain we were in thinking it possible to assess the probability or evidence for a model without reference to an alternative. There is no such thing as an unconditioned inference[21]. A well-posed question is necessarily conditional on an assumed model.

However, the posterior distribution within a model provides other possibilities for garding us against misleading inferences for nonsense models. Essentially, what we want is to assess the fit of our models to the data in a more direct way, evaluating whether the data, or various aspects thereof, are actually plausible to be observed under the assumed model. The key to do so in the Bayesian framework is *prediction*. Specifically, we can use the *posterior predictive distribution*

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

which represents our uncertainty when predicting the value for new or unobserved (but potentially observed) data $\tilde{y}$ under the model after having observed data y. In particular, if the model adequately fits the data, we expect that data sets simulated from the posterior predictive distribution resemble our actually observed data y. That is, we ask a question of the form: Given our current state of knowledge about the model, represented by the posterior $p(\theta|y)$, how surprised would we be to see the data set y we did in fact observe? To express this in somewhat frequentist verbiage, we ask ourselves: If we were to observe an ensemble of data sets from the assumed data-generating process with values of θ in accordance with our current (Bayesian) state of knowledge about θ, how likely is it to see something which resembles y? This provides us with a powerful strategy to use Bayesian inference to signal issues with our models, indeed, we can use the predictive distribution under the model to

[21] While this is likely to be stressed in some way in many a statistical work, nowhere is this point articulated with more force than in Jaynes (2003) (see in particular chapter 5).

point towards problems with the model, without having to step outside of it and embed it in a bigger one.

To actually assess these questions, we need to be specific about what we wish to mean when we say a simulated data set $\tilde{y}$ 'resembles' y. Two main strategies are usually advocated, using either test quantities and posterior predictive p-values or graphical comparisons (Gelman et al. 2013). For the former, one identifies a certain statistic of the data, $T(y)$, and compares the value of $T(y)$ to the posterior predictive distribution of $T(\tilde{y})$, for instance by reporting an estimate of the posterior predictive p-value $\mathbb{P}\{T(\tilde{y}) > T(y)|y\}$. If we are able to simulate 'fake data sets' like y under the model, then, given a sample $\theta^{(1)}, \ldots, \theta^{(2)}$ from $p(\theta|y)$, we can generate a sample $\tilde{y}^{(1)}, \ldots, \tilde{y}^{(n)}$ from $p(\tilde{y}|y)$, and estimate the posterior predictive p-value using Monte Carlo. If T is chosen so that a larger $T(y)$ is less likely under the model (i.e. suggests a poor fit), than a small posterior predictive p-value indicates that the model provides a poor fit *to that aspect of the data captured by the statistic* T. The principle of graphical posterior predictive checks is of course really the same, but relies on a graphical display of certain features of y along replicate data sets simulated from the posterior predictive distribution. The latter may be more suggestive of *what*, if anything, is wrong with the model, rather than merely suggesting *that* something is wrong. We illustrate this important pillar of Bayesian statistical analysis by reconsidering the sequence evolution problem.

Example (distance estimation, continued) We investigate the fit of the JC model for the data set considered in the previous example using posterior predictive simulations. An obvious test quantity to assess the fit of the model to the data would be the observed number of different sites in pairwise sequence alignments simulated from the model. However, since this is a sufficient statistic for the model (see above) this will be unable to display potential discrepancies with the observed data set (Gelman et al. 2013, chap. 6). Indeed, we estimate a posterior predictive p-value for this statistic of 0.49, which suggests that this aspect of the data is perfectly captured by the model (the estimated posterior mean number of different sites is 14.7, and the 95% uncertainty interval for the number of observed differences is $(7, 24)$).

A more interesting test quantity would consider the site pattern frequencies. There are six different site patterns (`AC`, `AG`, `AT`, `CG`, `CT` and `GT`), each of which are equally probable under the JC model. The observed site pattern frequencies in the data are $T(y) = (0.07, 0.40, 0.07, 0.33, 0.07, 0.07)$ (in the lexicographical order also used above). This already suggests some discrepancy with the JC model, although we can not, simply by looking at these values,

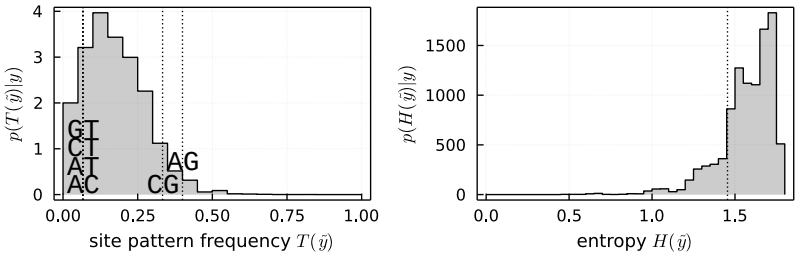


Figure 1.4: Posterior predictive distributions for site pattern frequencies and the entropy of the site pattern distribution for the JC model (based on 10000 data sets simulate from $p(\tilde{y}|y)$). Vertical dotted lines mark the values for the relevant quantities associated with the observed data y.

know whether this is *significantly* different from the JC prediction. A potentially interesting test statistic would be the entropy of the site pattern distribution ($-\sum_{i=1}^{6} f_i \log f_i$ where f_i is the frequency of pattern i), as the JC model induces the maximum entropy distribution (a uniform distribution) over site patterns and is thus a possibly significant modeling assumption. Posterior predictive simulations indicate that the entropy for the observed site patterns is somewhat low but not all that surprising (fig. 1.4), with a posterior predictive p-value of 0.16. Posterior predictive site pattern frequencies further suggest that the frequency of the AG pattern is somewhat higher than expected under the model ($p = 0.02$).

While the JC model appears to fit the data reasonably well from the perspective of the site pattern frequencies, we consider an alternative model which does not, in general, predict a uniform site pattern distribution. In particular, we assess the K80 model, which has two parameters: the distance d (expected number of substitutions per site) and the transition to transversion ratio (or bias) κ, where $\kappa = 1$ reduces the K80 model to the JC model. Using the same exponential prior for the distance d and an exponential prior distribution with mean 1 for κ, we estimate the posterior mean distance for the K80 model at $\hat{\theta} = 0.39$ with 95% uncertainty interval $(0.21, 0.62)$, which is the same as for the JC model. For κ, we estimate the marginal posterior mean at $\hat{\kappa} = 1.64$ $(0.41, 3.82)$. Clearly, there is considerable uncertainty about the value of the latter parameter. While suggestive of a transition bias $\kappa > 1$, the posterior under the K80 model is compatible with the JC model ($\kappa = 1$). The probability of the observed site pattern frequency for AG is, as expected, now higher (fig. 1.5 (A)). The marginal likelihood for the JC model

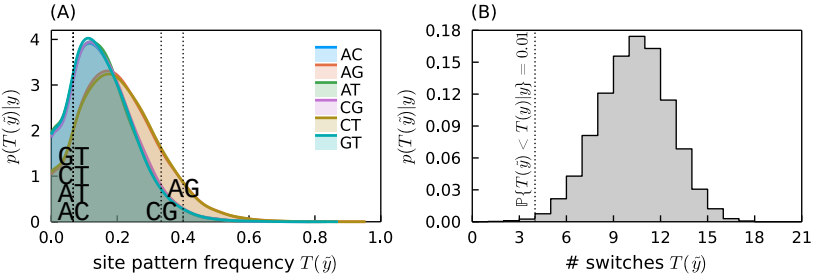


Figure 1.5: (A) Posterior predictive simulation of site pattern frequencies for the K80 (two-parameter) model (see fig. 1.4). Posterior predictive site pattern distributions are shown using a Gaussian kernel density estimate (KDE) for visual clarity. (B) Posterior predictive distribution for the number of switches between invariant sites and sites with an observed substitution for the JC model (the prediction for the K80 model is of course the same). The dotted vertical line marks the observed number of switches for data y.

is $p(y|\mathcal{M}_1) = \int p(y|\theta)p(\theta)d\theta$ which we can compute using numerical quadrature, giving $\log p(y|\mathcal{M}_1) = -48.4$. Similarly, the marginal likelihood for the K2P model is computed as $p(y|\mathcal{M}_2) = \iint p(y|\theta,\kappa)p(\theta)p(\kappa)d\theta d\kappa$, which on a log-scale amounts to -48.8. The Bayes factor 'in favor' of $\mathcal{M}_2$ is 0.7, and hence suggests there is no reason for favoring the two-parameter model.

An important assumption in both CTMC models we considered so far is independent evolution across sites. Indeed, we have been assuming that different columns in the pairwise alignment are iid realizations of the evolutionary model. A closer look at the data suggests that this assumption might be violated, consider the following display:

$$y = \begin{bmatrix} \texttt{.......AATCAT....TGG..............CGAG.GG........} \\ \texttt{.......GCGGGA....CAC..............GCGA.AC........} \end{bmatrix}$$

where we have represented invariant sites by a '.' character. We assess whether this is expected under the JC model by simulating alignments from the posterior predictive distribution and computing as a test quantity the number of times we switch from an invariant site to an observed difference moving along the alignment from left to right. We find that this is indeed a conspicuous pattern under the model, with the posterior predictive probability of observing a data set at least as extreme as y under the model about 0.01. This suggests that, if we wish to use a more realistic model, a reasonable avenue to explore would be taking into account some form of context-dependence in the substitution process (see e.g. Baele (2012)). A hidden Markov model would be an

obvious first choice for expanding the model in that direction for the present data set. □

In contrast to the frequentist question "What data would we expect under the assumed model?", Bayesian posterior predictive checks ask "What data would we expect under the assumed model, given the data we did in fact observe?". The latter question addresses much more adequately whether we have any reason to be *surprised* by the data under the assumed model and prior (Jaynes 2003). Clearly, whether or not we find problems with the model in this way depends crucially on which aspects of the data we care about. There are usually many potentially interesting test quantities for a given data set and model, and which ones we choose to evaluate should be motivated by which aspects of the data for which we think model fit is important. Devising good probes to critically assess model fit hence rests, like constructing models, ultimately on imagination and sound scientific judgement. This is in contrast with the 'automatic' nature of Bayesian inference for a given model.

Lastly, we note that in the above example, the Bayes factor actually *favors* $\mathcal{M}_1$, despite $\mathcal{M}_1$ being a special case of $\mathcal{M}_2$. This is because the simpler model makes more precise predictions, or conversely, the predictive distribution of the more complicated model tends to be more spread out over the space of possible data, so that the marginal likelihood $p(y|\mathcal{M}_2)$ is lower. This natural penalty for 'model complexity' in Bayesian inference has been called a 'Bayesian Occam's Razor' (MacKay 2003). Hence, when there are several competing models which all fit the data reasonably well, a Bayesian model selection procedure can be helpful to guard against issues of overfitting, when the latter is a concern. However, Bayesian model selection, as we illustrated above, is no substitute for checking model fit using the posterior predictive distribution. Whereas the above analysis favors the JC model, posterior predictive checks suggest there is no harm in assuming the K80 model instead[22].

1.2.4 Some concluding remarks about Bayesianism

We conclude this section by acknowledging our debt. The view on Bayesian statistical analysis expounded above has been most strongly influenced by Jaynes (2003) and Gelman et al. (2013); and to a lesser extent by MacKay

[22] […] which was, as a matter of fact, used as the true data generating process for simulating the data set in our toy example above (with $\kappa = 1.8$, and some additional autocorrelation in the substitution process along the sequence).

(2003) and Jeffreys (1961). In a nutshell, we are inclined to subscribe to the viewpoint expressed in Gelman and Shalizi (2013):

> […] the hypothesized model makes certain probabilistic assumptions, from which other probabilistic implications follow deductively. Simulation works out what those implications are, and tests check whether the data conform to them.

a viewpoint with which many a statistician, Bayesian or otherwise, will agree. Our view on probability, which paves the way for a proper appreciation of Bayesian statistical methods as the best way to bring this viewpoint into action, has been shaped by Jaynes (2003), Savage (1972), de Finetti (1974) and Jeffreys (1961). A core message in all of these authors is that probability is not simply a branch of mathematics that scientists can safely ignore, as they ignore topology or commutative algebra, but is really at the heart of the epistemology of science, quite independent of its mathematical formalization. We wish this were more widely taught. Of course, there is, inevitably, some ideology here too. We need ideas to guide practice, and the ideas coming from the Bayesian corner strike us as compelling in that regard.

Perhaps most importantly, what Bayesianism allowed us to see clearly, is that the main challenge in any statistical analysis lies not so much in its statistical aspects proper, but rather in the challenge of devising good models. Indeed, Bayesian statistics (with modern computational tools) is extremely powerful in practice, allowing us to conduct statistical inference for models never imagined by R. A. Fisher or other towering figures of the classical school. With the constraints on the types of models we can analyze gradually eroding, we are confronted (again) with the fact that it is really the science which is the hard thing, not the statistics. The creative endeavor which is the devising of theories and construction of formal models of nature is what limits our ability to make sense of empirical data, not our logical means to confront these two heterogeneous realms with each other.[23]

1.3 Genome evolution

Having established our broader goals and methodological commitments, we shall now seek to circumscribe our subject area, genomic diversity and the

[23] Although computational means remain limiting as well, despite the tremendous advances of the last decade. The creative challenge of devising *good approximations* hence remains of crucial importance as well.

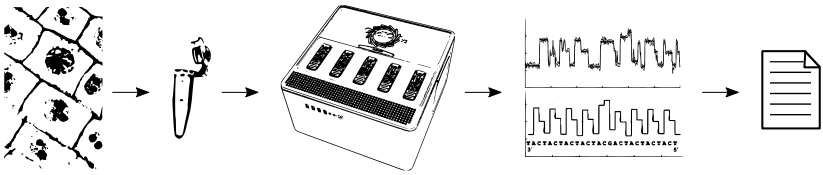


Figure 1.6: From tissue to genome sequence. Intricate chemical technology allows the determination of the sequence of base pairs in minute quantities of matter isolated from organic tissues. Here pictured is Oxford Nanopore's GridION.

evolutionary processes which shape it, in more detail. Of course, we are not the first to express a desire to study genomic data and genome evolution using statistical models of evolution, and in the present section we shall lay out the conceptual foundations due to previous workers in the field on which we build in this dissertation. As elsewhere in science, our object is not a simple given, but is constituted in a somewhat contingent way which may escape our criticism. In particular, as we have already noted above, the way in which we gather data conditions how we imagine our object of study, and *vice versa*. Clearly, how we conceive of a 'genome' in practice determines which aspects of genomic diversity we actually observe and shall seek to account for using our models. We shall hence first dedicate some attention to those concepts and abstractions we allow to structure our models.

1.3.1 Genomes as bags of genes

Of what consists the 'genomic data' and 'genomic diversity' we have been talking about in the first part of this chapter? Being bioinformaticians, when we refer to a 'genome' and 'genomic data', we are essentially referring to digital text files, which store the putative sequences of bases of a collection of DNA molecules isolated from some tissue material. The technological details of this process are staggering, but a rough sketch of the standard pipeline is shown in fig. 1.6. It should be held in mind that the end stage of these sophisticated manipulations is a rather long shot from the usual genetic definition of a genome as "all genetic information of an organism". Even if we grant the uncritical identification of "genetic information" with the sequence of base pairs in a collection of chromosomes, the usual bioinformatic conception of a genome, which we shall assume throughout the present work, ignores many potentially genetically relevant aspects thereof. Not only are modifications of the DNA and spatial features beyond the linear topology of eukaryotic chro-

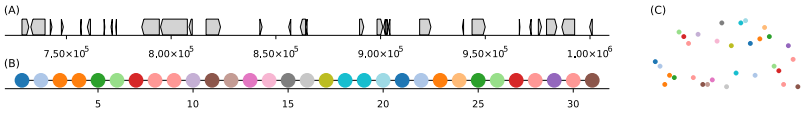


Figure 1.7: (A) The bioinformatic conception of an annotated genome. Protein-coding genes are shown as rectangles with a point showing whether the coding sequence is on the sense or antisense strand. (B) Associated gene list, ignoring both directionality and anchor points of the genes along the genomic sequence, retaining only gene order. Each dot represents a gene, whereas the colors represent homology relationships. (C) 'Bag-of-genes' associated with (A) and (B).

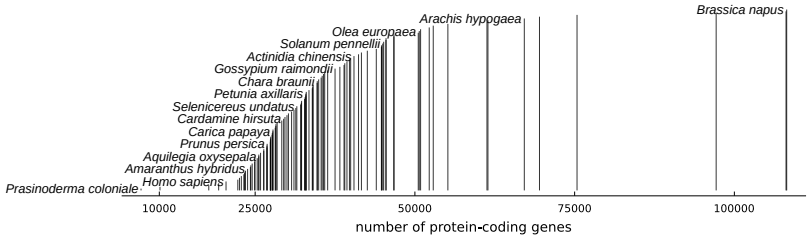


Figure 1.8: Number of protein-coding genes in the 100 genomes included in the PLAZA 5.0 database (Van Bel et al. 2022), and *Homo sapiens*. Species names are shown for several representative taxa.

mosomes ignored, we also ignore heterogeneity in the genetic material within an individual organism. In the present work, and most of evolutionary and comparative genomics, we go even further and shall mostly ignore the heterogeneity within a population (but see chapter 5), and speak without further restraint of *the* human genome, *the Arabidopsis* genome, whereas clearly, what we actually study in evolutionary genomics is, at best, a specimen with respect to the latter.

A next fateful step is the identification of 'genes' along these genome sequences. The concept of gene being a rather murky one, this is of course another source of conceptual difficulties and hard choices. In evolutionary genomics, the term 'gene' is usually employed to refer to a class of relatively stable and independent DNA segments that can be identified on the basis of certain structural properties (clearly, a rather long shot from the Mendelian factor), mainly DNA sequences which code for the usual biological macromolecules (proteins, tRNAs, rRNAs, miRNAs, and the rest of the RNA zoo). Protein-coding DNA sequences, which are relatively straightforward to identify, make up one such class, and they shall be the primary objects of study in

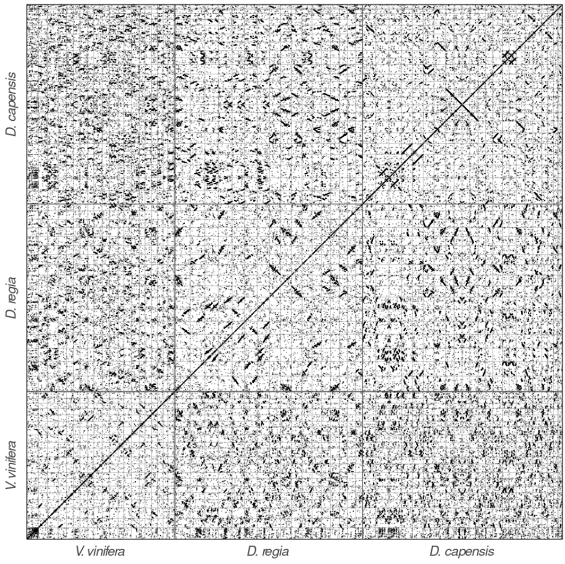


Figure 1.9: Scatter plot representation of the gene homology matrices (see main text) for all pairwise comparisons among the genomes of *Vitis* and two *Drosera* species. Gray and black lines separate chromosomes and genomes respectively.

this thesis, in that we shall base our study of genome evolution on what happens in this class. The bioinformatic identification of protein-coding genes is the result of a complicated pipeline, of which we shall be ignoring the specifics throughout the present work, assuming we have *the* collection of protein-coding sequences for a given genome sequence available.

What we end up with after applying our bioinformatic pipeline (i.e. the end result of a 'genome project'), and shall call a genome, is a collection of DNA sequences with a class of elements called genes anchored along them (fig. 1.7 A). *Comparing* these genomes reveals a rather dazzling amount of variation that demands evolutionary study. For one, the number of protein-coding genes varies considerably across genomes, as one can grasp from a glance at fig. 1.8. The absence of a broad correlation between the number of protein-coding genes and perceived 'organismal complexity' is generally considered as striking in this regard (referred to as the G-value paradox, see e.g. Hahn and Wray (2002)). What causes this variation? Which evolutionary processes can generate such patterns of diversity? At what rates must these processes operate in order to yield patterns that match the observed data? These types of questions

animate evolutionary genomics and the present work. If we further succeed in identifying *homologous genes*, we can conduct more detailed comparisons, as for instance displayed in fig. 1.9, where we show gene homology matrices for pairwise genome comparisons. In this representation, each genome is represented along the x- and y-axis as a collection of strings of genes, or gene lists (fig. 1.7 B), and a dot at coordinate (i, j) in the matrix represents a homology relationship between the ith gene along the x-axis and jth gene along the y-axis. Such a simple representation reveals a lot of variation whose structure is determined by evolutionary processes.

Clearly, the *gene list* view of fig. 1.7 (B) and fig. 1.9 is an abstraction which permits potentially insightful comparisons of genomes and opens up avenues for devising models of evolution which could enable us to unlock some of the information in genomic data about the evolutionary process and reconstruct plausible evolutionary histories. In most of our work, however, we will go even further in our abstraction of a genome and discard the information of where each gene is located along the genome, so that a genome is simply considered as a set of genes. This is sometimes called the *bag of genes* model of a genome (Huynen and Bork 1998) (fig. 1.7 C). To see what sort of evolutionarily relevant structure such a bag can have, we need to go a bit deeper into the various ways in which genes can be homologous to each other.

1.3.2 Gene families

> Biologists should realize that before long we shall have a subject which might be called 'protein taxonomy' – the study of the amino acid sequences of the proteins of an organism and the comparison of them between species. It can be argued that these sequences are the most delicate expression possible of the phenotype of an organism and that vast amounts of evolutionary information may be hidden away within them.
>
> – Francis Crick (1958)

Francis Crick, as astute and prescient as ever, was of course right in his prediction. Indeed, as the previous sections already suggest, much of this thesis is concerned with the development of statistical methods for unveiling some of the evolutionary information hidden away in protein sequences, in order to learn about the processes which shape genome evolution and reconstruct evolutionary histories. Crick's 'protein taxonomy' can be identified with what we nowadays would call the study of *gene families* and their evolution, which constitute the basic evolutionary units in the 'bag of genes' conception of a

genome, much in the same way as we have treated *sites* in our sequence evolution examples above.

A gene family, for us, is a set of homologous genes, that is, a set of genes which derive from some common ancestral gene.[24] Now, if by 'descend' we mean 'derived from a template-based copy of a stretch of DNA', then, presumably, most genes in extant genomes descend from a handful of ancient genes, and we are left with a rather useless concept. To appreciate the concept of gene family as it features in the present work, consider a collection of genomes $G = \{G_1, G_2, \ldots, G_n\}$, and assume their evolutionary relationships are adequately represented by a phylogenetic tree. To fix ideas, consider the three-taxon phylogeny depicted in fig. 1.10. This phylogenetic tree will be habitually referred to as the *species tree*, as a consequence of the common situation where the genome collection consists of genomes from distinct taxonomic species. In the latter case, bifurcations in the species tree are supposed to represent *speciation* 'events'[25]. In fig. 1.10, we see that the *most recent common ancestor* (MRCA) of the genomes in our collection is represented by an (unobserved) ancestral genome G_5. Concomitantly, if we now consider each genome as a 'bag of genes', so that G_j is identified with a set of genes $\{g_{j1}, g_{j2}, \ldots, g_{j,N_j}\}$, we see that, barring *de novo* origin of genes, each gene traces back to a common ancestral gene in G_5. A set of genes from G which trace back to a single common ancestral gene in the MRCA of G is referred to as an *orthogroup*, several examples of which are shown in fig. 1.10. In the present work, we shall identify the term 'gene family' with orthogroup.

A little more needs to be said about the term orthogroup, which is itself derived from *ortholog* (or sometimes *orthologue*), a concept due to Fitch (1970). Orthology is a particular kind of pairwise homology relationship. Two genes from *distinct genomes* are said to be orthologous if they derive from a common ancestral gene in the MRCA of their respective genomes. We can further broaden the concept by saying that two genes are orthologous *with respect to* ancestral genome $\mathcal{A}$ if they derive from a common ancestral gene in $\mathcal{A}$ (which need not be the MRCA of the relevant genomes). An orthogroup for a collection of genomes G is then a collection of genes where all between-genome

[24] Note that in molecular biology and more 'functionally' oriented fields in genomics, a gene family is sometimes defined using criteria such as having a similar biological function or sharing a particular molecular feature like a protein domain. Of course, due to evolution, such gene families will not be unrelated to ours.

[25] Speciation is hardly an 'event' but rather a, typically gradual, process. On long time scales however, it is convenient to think of speciation as an essentially instantaneous event, like a mutation. This collapsing of extended processes into discrete events is a common occurrence in phylogenetics.

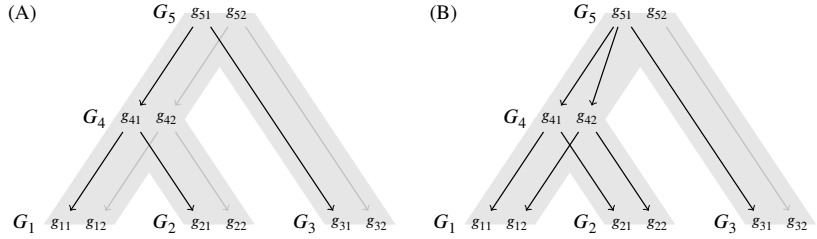


Figure 1.10: (A) Species tree (in gray) for three extant genomes G_1, G_2 and G_3. Orthogroups $\{g_{11}, g_{21}, g_{31}\}$ and $\{g_{12}, g_{22}, g_{32}\}$ are shown embedded in the tree with ancestor-descendant relationships between genes marked by a directed edge. (B) As in (A) but showing different orthology relationships among the set of genes (with extant orthogroups $\{g_{11}, g_{12}, g_{21}, g_{22}, g_{31}\}$ and $\{g_{32}\}$). Here g_{11} and g_{12}, for instance, are paralogous, whereas g_{11} and g_{22} are paralogous with respect to G_4, but orthologous with respect to G_5.

pairs of genes are orthologous with respect to the MRCA of G.

Now, homologous gene pairs do not only originate through divergence of their respective genomes (i.e. speciation if we are considering species trees), but also through *duplication* events, which create diverging gene pairs within a single genome. For instance, in fig. 1.10 (B), the ancestral gene g_{51} has two descendant genes in genome G_4 due to a gene duplication along the branch between G_4 and G_5. Pairs of genes which diverged through such within-genome gene duplication events are said to be *paralogous* (Fitch 1970). In addition, not all ancestral genes leave observed descendants in all genomes under study due to *gene loss*, whereby a gene is removed from the set of protein-coding genes. In fig. 1.10 (B) for instance, g_{52} does not leave any descendants in G_4, and consequentially, no descendants of this gene are observed in G_1 and G_2.

As a result of these processes (and others, see below), the evolutionary history of a gene family is not simply determined by the evolutionary history of the associated genomes (as in e.g. fig. 1.10 A), but is rather the result of an evolutionary process distinct, but not independent, from the one generating the genome-scale tree (Maddison 1997). The result is that there is considerable variation in both the size of gene families across the genome and their phylogenetic histories. By studying the collection of gene families in a genome, i.e. variation within and across bags of genes, we can learn about the key processes that generate genomic diversity at this level of abstraction. To do so, we shall need models of gene family evolution which explain how gene families

evolve within their evolving host genomes.[26]

1.3.3 Models of gene family evolution

We introduced the concept of gene family as the sensible evolutionary unit of analysis for studying genome evolution from the 'bag of genes' perspective, and we noted that multiple evolutionary processes lead to variation in gene family evolutionary histories across the genome. In this penultimate section, we provide a high-level overview of these sources of variation, what we can learn from them, and how previous workers have sought to model them. We will stick to an informal style, deferring detailed treatments to the more technical chapters of this dissertation. We refer the reader to Szöllősi et al. (2015) for an exceptionally exhaustive review on these matters. The recent edited volume by Scornavacca, Delsuc, and Galtier (2020) is also helpful in that regard.

1.3.3.1 Gene genealogies

The first source of variation in evolutionary histories across the genome is a consequence of elementary population genetics. Looking back at fig. 1.10, we realize that we have been grossly simplifying things, and that there is in fact no such thing as *the* ancestral G_5 genome, just like there is no such thing as *the* human genome. Genes are passed on generation by generation from parents to children, and not from species to descendant species. A set of genes in a gene family will therefore trace back a *genealogy* or *gene tree* inside the

[26] A short note on *practical* inference of gene families is due. Clearly, the concept of gene family is tied up with the processes which we seek to model, as what is or is not a proper gene family depends on its evolutionary history. Indeed, we do not *observe* gene families, so we have to infer them from the protein-coding sequences, ideally based on a model of evolution. The problem of gene family inference is essentially a *clustering* problem. In a Bayesian framework, it is straightforward to formulate such a clustering problem with an arbitrary within-cluster probabilistic model, so that in theory, one could infer gene families and their evolutionary histories jointly. This is however computationally prohibitive, and as far as we are aware, unexplored terrain. In practice then, we shall rely on a collection of gene families inferred by other, more or less heuristic, means, and assume these as data, i.e. as known without error. We use OrthoFinder in the present work (Emms and Kelly 2019), which uses the common strategy of clustering a sequence similarity graph into gene families using Markov clustering (Van Dongen 2000). As a compensation for our heuristically inferred orthogroups, we shall typically depart from the strict assumption that our 'observed' gene families are derived from a *single* ancestral gene in the MRCA of the species tree, but rather assume a small but unknown number of ancestral genes for each gene family, which we shall model by a parametric distribution.

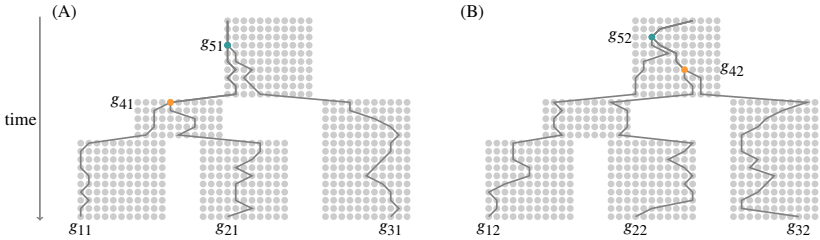


Figure 1.11: Gene genealogies for two gene families within a species tree. We show a model with a population of a constant size $N = 15$ evolving in discrete non-overlapping generations. Each dot represents an individual genome in a given generation in some population, reproductively isolated from the other contemporaneous populations. The lines trace back the genealogy for two observed gene families. (A) The genealogy for the first family is congruent with the species tree. (B) The genealogy for the second family is incongruent with the species tree due to ILS.

set of *populations* represented by the species tree (fig. 1.11). Moreover, if the latter involve sexually reproducing individuals, recombination and Mendelian segregation will cause the gene genealogies to be different in general for different samples of genes. Not only do the simple consequences of Mendelian inheritance lead to quantitative variation in the time since divergence for different orthologous genes across the genome, they can also lead to qualitatively different phylogenetic tree topologies, a phenomenon referred to as *deep coalescence* or *incomplete lineages sorting* (ILS), as illustrated in fig. 1.11 (B).

Clearly, in order to account for this source of variation one would have to model the underlying population genetic processes. The usual strategy to do so is to model the gene genealogies using *coalescent processes*, which describe the ancestry for a sample of extant genes in a single population backwards in time (Hein, Schierup, and Wiuf 2004). The key determinants of the extent of this type of variation are demographic parameters ϕ, in particular parameters which are related to the *population size*. If we succeed in establishing a reasonable model, one can use observed sequence data $y_1, \ldots, y_n$ for n gene families to learn about the species (population-level) tree S and ϕ using hierarchical models of the following form

$$\mathcal{T}_1, \ldots, \mathcal{T}_n | S, \phi \sim \text{CoalescentModel}(S, \phi)$$
$$y_i | \mathcal{T}_i, \psi_i \sim \text{PhyloCTMC}(\mathcal{T}_i, \psi_i)$$

Where $\text{PhyloCTMC}(\mathcal{T}, \phi)$ represents a phylogenetic CTMC model with tree $\mathcal{T}$ and parameter ψ as defined in sec. 1.1.3. As we will see repeatedly, the

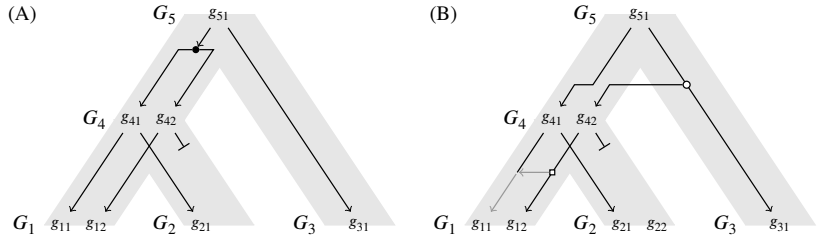


Figure 1.12: (A) Evolution at the locus level, or locus tree, (black lines) within a species tree (gray tree shape), with gene duplication and loss events. The black circle represents a gene duplication event, whereas the ⊣ arrowhead indicates a gene loss event. (B) Locus tree within a species tree with gene loss, horizontal gene transfer and gene conversion. The white circle indicates a transfer event, whereas the white square indicates a gene conversion event.

model combines a genome-level model with with a family-level model. The most common model along these lines is the *multispecies coalescent* (MSC) model (Tajima 1983; Hudson 1983; Pamilo and Nei 1988; Rannala et al. 2020), which will be the focus of our chapter 5.

1.3.3.2 Locus-level events

The gene tree as we described it in the previous paragraph traces the evolutionary history of a *single gene locus* in a collection of populations. However, as we already noted above, new loci are created and lost throughout evolution, leading to another source of evolutionary variation. Notably, these processes lead to variation in gene copy number across gene families, which is the rule rather than the exception.

Various events and processes have been considered in models of evolution at the locus level in the literature. *Gene duplication* events cause the duplication of a gene locus *within* a species tree branch, whereas *gene loss* events lead to the loss of a locus. These two processes are illustrated in fig. 1.12 (A). Gene family evolution by gene duplication and loss is often modeled using birth-death process (BDP) models which operate along the species tree, much like we defined the phylogenetic CTMC model as a CTMC model operating along a phylogenetic tree (some examples include Hahn et al. 2005; De Bie et al. 2006a; Csűrös and Miklós 2009; Arvestad, Lagergren, and Sennblad 2009; Sjöstrand et al. 2012; Szöllősi et al. 2012; Boussau et al. 2013; Szöllősi, Rosikiewicz, et al. 2013; Tasdighian et al. 2017; Zwaenepoel and Van de

Peer 2019a, 2020). The usual model for sequence data $y_1, \dots y_n$ has a similar form as above

$$\begin{aligned} \mathcal{G}_1, \dots, \mathcal{G}_n | S, \theta &\sim \text{PhyloBDP}(S, \theta) \\ y_i | \mathcal{G}_i, \psi_i &\sim \text{PhyloCTMC}(\mathcal{G}_i, \psi_i) \end{aligned}$$

Where it is assumed that the sequences y_i evolve along the *locus tree* $\mathcal{G}_i$ generated by the phylogenetic BDP model according to a phylogenetic CTMC model. Note, however, that because of the variation generated in gene copy number, we do not need the sequence data to learn about the parameters of the genome-scale process. Indeed, observed gene family *sizes* already provide relevant information for statistical inference (e.g. Hahn et al. 2005; Csűrös and Miklós 2009; Zwaenepoel and Van de Peer 2020). Phylogenetic BDP models of gene family evolution by gene duplication and loss form a major part of the present work and will be discussed in detail in the coming chapters.

Other processes which can be considered at the locus level are *horizontal gene transfer* (HGT) and *gene conversion*. The first involves copying a locus from a single species tree lineage into another, contemporaneous lineage, whereas the latter involves the replacement of the gene at one locus by a gene at a homologous locus within the same species. Both are illustrated in fig. 1.12 (B). *De novo* gene origin, or reactivation from a pseudogenized gene, is another locus-level event (McLysaght and Hurst 2016). We shall not consider these processes in the present dissertation. Importantly, the various evolutionary events thought to operate at the locus level are highly idealized. Indeed, there are no known *in vivo* molecular mechanisms which could accurately copy an arbitrary single gene, and actual gene duplicates are often partial or chimeral (e.g. Katju and Lynch 2003). Similarly, gene conversion often affects only parts of a gene sequence. Furthermore, a single HGT, duplication or loss event may involve multiple genes, violating the assumption of independence across families, *etc.*

In addition, models at the locus level usually do not take into account the population-level processes which generate actual genealogies. One can, however, consider the locus tree generated by a phylogenetic BDP as analogous to the species tree in our discussion of coalescent models, and generate a gene genealogy within a a locus tree. This is the so called three-tree model of Rasmussen and Kellis (2012; see also Mallo, Oliveira Martins, and Posada 2016)), which is of the form

$$\mathcal{G}_1, \dots, \mathcal{G}_n | S, \theta \sim \text{PhyloBDP}(S, \theta)$$

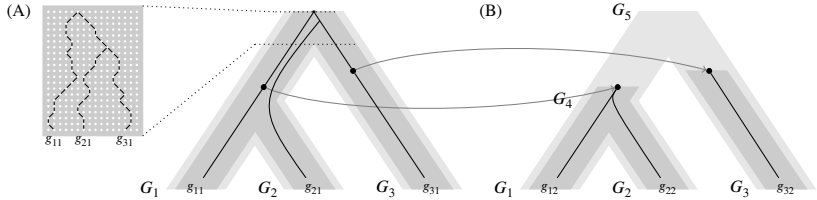


Figure 1.13: Example of a gene tree within a locus tree within a species tree. (A) Species tree (light gray) with locus tree (dark gray) for the first locus, and gene genealogy (black lines) within the locus tree. The inset on the left shows a close-up of the underlying population-level coalescence process (assuming a constant-size Wright-Fisher population model). (B) Locus trees for the second and third locus, which originate through gene duplication events along the branch from G_5 to G_4 and the branch form G_5 to G_3 respectively. It is assumed that gene duplications fix instantaneously.

$$\mathcal{T}_i|\mathcal{G}_i, \phi \sim \text{CoalescentModel}(\mathcal{G}_i, \phi)$$
$$y_i|\mathcal{T}_i, \psi_i \sim \text{PhyloCTMC}(\mathcal{T}_i, \psi_i)$$

generating a gene tree $\mathcal{T}$ in a locus tree $\mathcal{G}$ in a species tree. This is exemplified in fig. 1.13. This model has the two previous models as special cases, indeed, if we assume the locus tree to be identical to the species tree, we are generating a gene genealogy in a species tree according to a coalescent model. On the other hand, if we assume, in the three-tree model, that the gene genealogy is identical to the locus tree, we end up with a phylogenetic BDP as above. Usually, when either identification is made, the tree at the lowest level is referred to as *the gene tree*, irrespective of whether it represents a gene genealogy or a sequence of locus-level events.

Note however that this does not fully account for the interaction between population-level and locus-level processes, which is a somewhat artifical distinction anyhow. In particular, gene duplication (or loss) is a mutational event taking place in a single genome, and has to spread through the population in order to establish as a new locus at the population level (i.e. as a new locus of 'the' species' genome). There is, in other words, for each locus-level event a phase in which the population is *polymorphic* for the locus-level variant. If a speciation event were to occur during such a polymorphic phase, different daughter species may fix different locus-level variants, leading to ILS at the locus level. The length of such a polymorphic phase is determined by population genetic parameters, in particular the effective population size. Ignorance of locus-level polymorphism amounts, from a population genetics perspective, to the assumption that these mutations, when they fix, sweep through the popu-

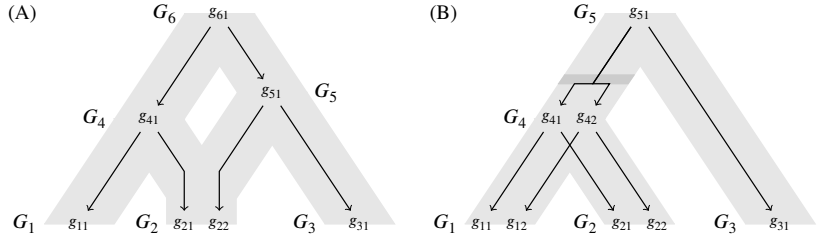


Figure 1.14: Locus trees in the presence of polyploidization. (A) Allopolyploid hybridization between a descendant of G_4 and a descendant of G_5 gave rise to allopolyploid lineage G_2. (B) Autopolyploidization, where the polysomic phase (if any) is suggested by the dark gray time span in the species tree.

lation instantaneously, and that when they do not fix, they are purged quickly from the population. A recent attempt at coherently defining an integrative model of gene family evolution which accounts for the latter issues may be found in Li et al. (2021).

1.3.3.3 Genome-scale events

We have been considering the evolution of gene families within the context provided by an assumed species tree S (where we stress that we are making rather liberal use of the term species, as we do for the term gene), which represents the evolution of a set of *populations* that descend from a common ancestral population through time. The species tree represents the top level of the evolutionary hierarchy, in the sense that we assume the whole genome (the entire bag) to evolve within a single species tree. Of course, we may also wish to model the evolutionary processes which generate species trees themselves and consider the effects of evolutionary processes at this level on the observed 'bags of genes'. Relevant 'events' at this level include *speciation*, *extinction*, *hybridization* and *polyploidization*.

We will not deal explicitly with models for macroevolutionary dynamics, which are concerned with the processes of speciation and extinction and the shapes of species trees (but see chapter 4). We deal quite extensively with polyploidization, or whole-genome duplication (WGD), which leads to a duplication of the entire bag of genes. We will see that additional aspects, besides the simple correlated duplication of loci across gene families, must be taken into account to adequately model WGDs. Note that hybridization

and allopolyploidization cause the species 'tree' to be not very tree-like after all, lending itself rather to a network representation (fig. 1.14 A). However, because the evolutionary histories at the lower levels of the hierarchy remain tree-like, we do not usually have to worry about phylogenetic networks (see for instance the locus tree in fig. 1.14 A, also our *Drosera* examples in chapter 5 and chapter 6).

Ancestral autopolyploidy on the other hand does not involve reticulation at the species tree level, and does not in itself create new loci, at least as long as there is polysomic inheritance. Only after a process referred to as *rediploidization* (Wolfe 2001) is disomic inheritance reinstated and can the duplicated genes be regarded as distinct loci evolving according to the locus-level processes considered above (fig. 1.14 B). As should be obvious from fig. 1.14 (B), when this rediploidization process is not completed before the occurrence of a speciation event, yet another source of ILS, referred to as 'lineage-specific ohnolog resolution' by Robertson et al. (2017) can lead to variation in gene trees across families. Finally, we note that the difference between allo- and autopolyploidy is not always clear cut. Many autopolyploids are formed through the hybridization of unreduced gametes from distinct parents, in which case the whole-genome 'duplication' is really a merger of two different genomes as in allopolyploidy. The difference in inheritance mode can however be significant for gene family evolution. Indeed, if there is no recombination among the merged genomes, homeologous genes will have diverged before the WGD event (as is clear from fig. 1.14 A), whereas if there is recombination, homeologous gene pairs are more likely to diverge after the WGD event (Roux and Pannell 2015). However, when considering long time scales (as we typically do in this dissertation), the difference may become of little relevance.

1.4 Aims and outline of this thesis

In this introductory chapter, we have sought to motivate our approach towards the challenge of making sense of genomic data and genome evolution. We have provided an overview of what we take to be most important aspects of the Bayesian paradigm in statistics. Lastly, we sketched a high-level picture of the evolutionary concepts and models that will be used to study genomic diversity in th ensuing chapters. The broad aims of the rest of this dissertation are then (1) to devise statistical *models* of genome-scale evolution from the 'bag of genes' point of view, (2) to implement efficient *methods* for inference under these models, (3) to *assess* to what extent these models are adequate,

and in which ways they need to be expanded in order to adequately fit the data, and (4) to *use* these models and methods to learn about the processes of genome evolution and evolutionary history in a principled way.

Specifically, in **chapter 2**, which is more of a prelude, we deal with gene family evolution within a single genome. This short chapter introduces the birth-death process models of gene family evolution by gene duplication and loss which take center stage in later chapters of the work. We take a brief look at age and family size distributions and simple models which can account for their characteristic form.

In **chapter 3** we consider gene family evolution by gene duplication and loss in a phylogenetic setting. We present the theory of phylogenetic birth-death process models and implement methods for genome-scale Bayesian inference of rates of gene family evolution from gene count data. We spend considerable effort on assessing how well simple birth-death process models fit observed patterns of genome evolution using a number of example empirical data sets. We study the inference of ancient whole-genome duplications from gene count data. Lastly we develop and evaluate a new model of gene family evolution based on multi-type branching processes to account for some of the failures of the simple birth-death model of gene family evolution.

In **chapter 4** we switch perspectives from gene counts to gene trees. This chapter serves as a prelude to the two chapters that succeed it, defining the concept of a conditional clade distribution that features heavily in the latter.

In **chapter 5** we implement a likelihood-free Bayesian inference method for species tree inference under the multispecies coalescent model. The method takes a somewhat intermediate position between two widely adopted strategies for the latter goal, taking advantages from both. Several case studies are presented.

In **chapter 6** we focus on model-based gene tree reconciliation for phylogenetic birth-death process models. We develop a Bayesian approach based on the amalgamation principle of Szöllősi, Rosikiewicz, et al. (2013) for joint genome-scale joint inference of gene trees and their reconciliation under models of gene family evolution which account for gene duplication, loss and whole-genome duplications. We study in detail the problem of phylogenomic inference of ancient whole-genome duplications.

We end with a brief conclusion assessing what we did and did not achieve in the body of the present work. We note that, while we describe previously published methods (mainly in chapter 3 and chapter 6), virtually all of the work

in this thesis is unpublished in the presented form. Where we have drawn from our previously published articles, this will be indicated accordingly.

2 Single-genome models of gene family evolution

We shall start our inquiry on what we can learn about genome evolution from the 'bag of genes' point of view, introduced in the preceding chapter, by taking a look at the gene content of individual genomes. Following the popular practice of using an '-ome' suffix, the collection of all protein-coding gene families in a *single* genome will be loosely referred to as the *paranome*. The paranome is one of the most elementary sources of information about genome evolution and has been studied since the first genome projects provided the necessary data (Huynen and Van Nimwegen 1998; Lynch and Conery 2000, 2003; Karev et al. 2002; Maere et al. 2005). Two features of the paranome are of special interest: (1) the gene family size distribution and (2) the age distribution. Both aspects will be dealt with in this still rather introductory chapter, not only because they are interesting for their own sake, but also because they are important for the inference of ancient whole-genome duplications (WGDs) and genome evolutionary rates. The main purpose of this chapter, however, is to present some simple results that appear (to us at least) not to be widely known, and to introduce some key ideas we shall work with later – in particular birth-death process models of gene family evolution.

2.1 The size and age distribution of gene families

Two aspects of the collection of gene families within a single genome can provide statistical insights in genome evolution. The (empirical) *size distribution* is simply the number of families of a certain size. The (empirical) *age distribution* records for each duplicate gene its time since duplication. Importantly however, neither the size distribution nor the age distribution are directly observable from our bag of genes. Before embarking on our attempts to devise models for the latter, we need to specify what it is exactly that we assume to observe.

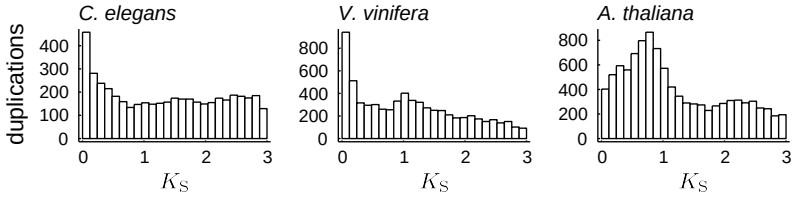


Figure 2.1: Examples of whole-paranome age distributions for *Caenorhabditis elegans*, *Vitis vinifera* and *Arabidopsis thaliana*, where age is measured by the pairwise synonymous divergence K_S, estimated using maximum-likelihood under the model of Goldman and Yang (1994) using the `codeml` program of Z. Yang (2007a).

Note first of all that the concept of *orthogroup* as defined in the previous chapter is not well-defined in the single-genome setting. Indeed, there is no MRCA with respect to which we could meaningfully define an orthogroup. In practice however, paralogous gene families within a single-genome can be inferred using standard orthogroup inference tools (which are approximate anyhow, as we noted in chapter 1), by clustering the sequence similarity graph for the gene set under consideration. The assumption is then that, since all gene families are defined with respect to the same sequence similarity threshold and clustering algorithm, their sizes are comparable, and we roughly obtain orthogroups with respect to some distant ancestor. Clearly this is *ad hoc*, but we shall have to live with it. We shall henceforth assume that we have meaningful gene families, and hence the associated size distribution.

The age distribution for a given paranome is however less straightforwardly obtained, even if we succeed in delimiting families. Obviously, extant gene sequences do not bear a tag which says when they were born. What we have at our disposal to estimate the age of a gene duplicate is its sequence, which in itself does not say anything about its age either. The *molecular divergence* between two sequences, however, provides information about their *divergence time*, which in the context of a strictly paralogous family corresponds to the time of a duplication event. Estimating divergence times from molecular data is a complicated problem, as molecular divergence provides information about the evolutionary *distance*, which is a product of the substitution rate and divergence time, but not the product of the two (see our examples in chapter 1). If, however, the substitution rate is approximately constant across the genome and over time, the divergence time between two genes will be a simple linear function of the molecular distance between them. The latter assumption is known as the *molecular clock*, and if it were to hold, one could use molecu-

lar distances to construct an approximation of the age distribution. While the molecular clock is usually a rather terrible assumption (see also later chapters), for putatively neutral sites (i.e. sites where substitutions are unlikely to have a fitness effect, or unlikely to have very different fitness effects at least) within a single genome over a not too long time scale, it may reasonably hold.

For this reason, the *synonymous distance* (i.e. the expected number of synonymous substitutions per synonymous site between two coding sequences), or K_S, has often been used to measure divergence between sequences for these purposes. Assuming a CTMC model of sequence evolution defined at the codon level, such as the one of Muse and Gaut (1994) or Goldman and Yang (1994), one can obtain an estimate of the synonymous distance using maximum likelihood or Bayesian inference (Yang 2006). One can thus estimate a K_S-based age distribution as a proxy for the true age distribution, and this was first taken to great advantage in Lynch and Conery (2000). As we will not make extensive use of age distributions beyond some simple considerations in the present chapter and chapter 6, we will not dwell on the details of inferring these distributions further here, and refer the reader to Vanneste, Van de Peer, and Maere (2013), Zwaenepoel and Van de Peer (2019b) and Sensalari, Maere, and Lohaus (2021) instead. It is however important to note that such K_S age distributions, as used in practice, are rather crude estimates of the true age distribution, usually relying on pairwise ML distance estimates and rather *ad hoc* strategies for estimating the K_S-scale age of duplication events deeper in the gene family phylogeny. The hope is of course that in the coarse-scale picture of the whole paranome these sources of error will not matter too much. Some examples of K_S-based whole-paranome age distributions are shown in fig. 2.1. Note the conspicuous 'bumps' in the distributions of *A. thaliana* and *V. vinifera*, which are widely believed to be the signatures of ancient WGD events (Blanc and Wolfe 2004b; Vanneste, Van de Peer, and Maere 2013).

2.2 Deterministic models for the paranome

Before delving into the main topic of interest in this chapter, we first introduce two historically rather important *deterministic* population models for the whole-paranome age distribution. Both use a difference (or differential) equation to model the number of gene duplication events of a certain age (on a K_S scale) retained in extantly observed paranomes.

2.2.1 The demographic model of Lynch & Conery

In one of the first studies that sought to quantify rates of gene family evolution from genome data, Lynch and Conery (2003) used a very simple demographic model to estimate the genome-wide gene duplication and loss rate from a histogram of the K_S distribution (see also Lynch and Conery (2000) for an even simpler model). Assuming a per-gene duplication rate λ and a loss rate μ for duplicated gene copies, they modeled the number of duplicated gene copies $N(t)$ per ancestral gene over discrete time steps of length Δt recursively as

$$N(t + \Delta t) = N(t) + \lambda \Delta t[1 + N(t)] - \mu \Delta t N(t) \tag{2.1}$$

It is however somewhat more convenient to consider the model in continuous time, using the associated ODE

$$\frac{dN(t)}{dt} = (\lambda - \mu)N(t) + \lambda$$

Assuming equilibrium $dN(t)/dt = 0$ and $\lambda < \mu$, we find that the number of duplicate genes per family equals $\lambda/(\mu - \lambda)$. At equilibrium, the size of a gene family is therefore $\mu/(\mu - \lambda)$. Denoting by $F(t)$ the number of duplicate genes of an age $< t$ in the extant duplicate gene pool, and assuming the population is (and has been) at equilibrium, we have under the stated assumptions

$$F(t + \Delta t) - F(t) = \frac{\mu}{\mu - \lambda} \lambda \Delta t e^{-\mu t}$$

Where the assumption of $\lambda \ll \mu$ entails that the probability of a gene undergoing additional duplications upon a duplication, before losing a duplicate, is negligible. Dividing both sides by Δt and taking the limit as $\Delta t \downarrow 0$ we obtain the age distribution in continuous time

$$D(t) = \frac{dF(t)}{dt} = \frac{\lambda \mu}{\mu - \lambda} e^{-\mu t} \approx \lambda e^{-\mu t}$$

Which corresponds to a simple exponential survival law, entailing a linear relationship between $\log D(t)$ and t.

Clearly the exponential model only fits the data reasonably in the low K_S region, with substantial deviations easily spotted by eye at higher K_S, and as a result parameters should only be estimated from the young age cohort of duplicates. Lynch and Conery (2003) fitted the discrete time model to a histogram of the K_S distribution of interest for the domain $K_S < 0.1$ in frequency

bins of $K_S = 0.01$ using a least squares regression of $\log D(t)$ on t, obtaining parameter estimates for λ and μ on a K_S 'time' scale. The slope of the regression b provides an estimator for the loss rate $\hat{\mu} = -b$, and the duplication rate can be estimated as $\hat{\lambda} = e^a \hat{\mu}/(\hat{\mu} + e^a) \approx e^a$, where a is the intercept (when the histogram is expressed on a density scale). While straightforward, this parameter estimation procedure is quite delicate and serves only to estimate short-term rates of genome evolution. We provide an illustration using the age distributions shown in fig. 2.1.

Example (Lynch and Conery (2003)). We fit log-linear regressions to the K_S age histogram of the number of retained duplication events per family for *C. elegans* and *V. vinifera*. We consider K_S intervals of length 0.01 over a range of 0.1 and 0.2 K_S. Estimates are shown in tbl. 2.1 and the regression lines and relevant data are plotted in fig. 2.2.

Table 2.1: Estimates of duplication rates (λ, number of duplication events per gene per K_S), loss rates (μ, number of gene loss events per duplicated gene per K_S) and half-lives of duplicate genes ($t_{1/2}$, on a K_S scale) for *C. elegans* and *V. vinifera* based on the regressions graphed in fig. 2.2. Estimates based on regressions on different K_S ranges are shown.

Species	K_S range	$\hat{\lambda}$	$\hat{\mu}$ (SE)	$t_{1/2}$
C. elegans	< 0.2	0.5	5.8 (1.0)	0.12
C. elegans	< 0.1	0.5	10.1 (2.8)	0.07
V. vinifera	< 0.2	1.3	6.7 (0.8)	0.11
V. vinifera	< 0.1	1.5	10.5 (1.4)	0.07

The K_S range chosen for the regression analysis can however have quite a strong effect on the estimated parameter values (tbl. 2.1). Furthermore, this approach uses only a tiny fraction of the available K_S distribution to estimate parameters of interest relevant for genome evolution over short evolutionary time scales. Nevertheless, the approach can be useful to get very crude estimates for the short term rates of gene duplication and loss. Using the substitution rate estimates of Lynch and Conery (2003), the duplication rate estimate would be about 1 duplication per 100 million years (My) for both *C. elegans* and *V. vinifera*, whereas the half life of a duplicated gene would be 2.2 and 5.4 My respectively. □

Note that we cannot easily apply this approach to the *A. thaliana* K_S distribution shown above, as here the young cohort of duplicate genes used for the regression approach ($K_S < 0.2$ say) seems to contain WGD-derived gene duplicates, violating model assumptions.

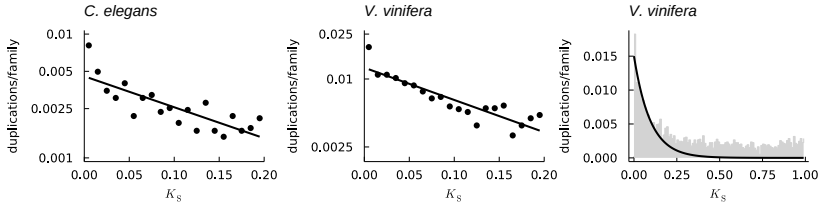


Figure 2.2: Log-linear regressions for the *C. elegans* and *V. vinifera* K_S distributions, yielding duplication and loss rate estimates under the model of Lynch and Conery (2003) displayed in tbl. 2.1. The rightmost plot shows the K_S distribution for *Vitis* over an interval of $1K_S$ with the fitted exponential distribution (black line).

2.2.2 The model of Maere *et al.*

Maere et al. (2005) and Vanneste, Van de Peer, and Maere (2013) considered a more sophisticated model akin to the one of Lynch and Conery (2003), but which accounts for whole-genome duplication (WGD) events and aims at modeling the full age distribution (as opposed to a small range where equilibrium assumptions are reasonable). Similar to the model of Lynch and Conery (2003), the model is specified as a discrete evolutionary model setting, modeling the paranome size in discrete time steps corresponding to K_S units. Notably, the model is specified directly in terms of the number of retained duplicates in each discrete age class (i.e. the age distribution D above).

Consider a genome evolving in discrete time steps $1, 2, \dots, T$ by small-scale duplication and loss (SSDL), with a finite number k of WGD events occurring at time points $(t_1, t_2, \dots, t_k)$. Maere et al. (2005) model the number of gene duplicates $D_i(x,t)$ of age x in time interval $t \in (1, 2, \dots, T)$ retained from 'duplication mode' i (with $i = 0$ corresponding to small-scale duplication, and $i > 0$ corresponding to WGD i) according to the following laws

$$
\begin{aligned}
D_0(1,t) &= \lambda \Delta t \Big(\sum_{x'=1}^{\infty} D(x', t-1) + G_0 \Big) \\
D_i(1,t) &= \delta(t, t_i) \Big(\sum_{x'=1}^{\infty} D(x', t-1) + G_0 \Big) && i > 0 \\
D_i(x,t) &= D_i(x-1, t-1) \big(x/(x-1) \big)^{-\alpha_i} && x > 1, i \geq 0 \qquad (2.2)
\end{aligned}
$$

with $D(x,t) = \sum_i D_i(x,t)$. Here G_0 is the initial number of genes in the genome, δ denotes the Dirac delta function, λ is the per-gene rate of small-

scale duplication and α_i is the power law decay constant for duplication mode i, determining the rate of gene loss. Note that under this model and for positive α_i, the decay rates decrease over time, so that the longer a duplicate has been preserved, the less likely it becomes to get lost. This is of course supposed to model the various processes by which gene duplicates become stably established in a genome (e.g. by sub- or neofunctionalization), and renders the model applicable to long-term evolutionary scales (in contrast to the model of Lynch and Conery (2003)). Maere et al. (2005) set $T = 50$ with each time step corresponding to $0.1 K_S$.

Two additional points should be noted about this model. The first is that, as in the model of Lynch and Conery (2003), it is assumed that there is a 'base' set of G_0 genes which is not subject to gene loss. The loss rate in this model (embodied by the α_i) is therefore associated exclusively with the loss of duplicate gene copies. Secondly, by modeling the age distribution and not the K_S distribution (albeit with age on K_S scale), the model results in a discrete peak of duplicated genes at WGD times $t_i, i \in \{1, \dots, k\}$. Obviously however, variation in the realized number of substitutions, as well as differences in substitution rates across genes, causes the WGD-associated peaks in the distribution to be distributed around some mean K_S age. The latter is addressed by Maere et al. (2005) by assuming a Poisson model for the number of substitutions since duplication conditional on the age of the duplication event. Maere et al. (2005) used this model to study gene family evolution in *Arabidopsis thaliana*. Specifically, they fitted the model (using simulated annealing and the χ^2 distance as objective function) to K_S age distributions for different functional classes of genes (based on Gene Ontology (GO) annotations), and used the fitted α_i to compare the long-term duplicate retention patterns of gene duplicates derived from the SSDL process with WGD-derived duplicates.

Probably due to the lack of a simple parametric form and estimation procedure, the above model has not been widely adopted in evolutionary genomics. In a statistically somewhat *ad hoc* way, one can however use the age distribution generated by eq. 2.2 as a probability density for the empirical age distribution, and use the latter as a likelihood function in a Bayesian model. This approach can be used to estimate duplication rates and decay parameters, as well as the timing of ancient WGDs on a K_S scale. We illustrate this approach briefly for the *Vitis* paranome. We shall not have the opportunity to use the model in the rest of our work however, and will consequently not go into much detail here.

Example (Maere et al. (2005)). To conduct Bayesian inference for the model of Maere et al. (2005), we generate an approximate likelihood function

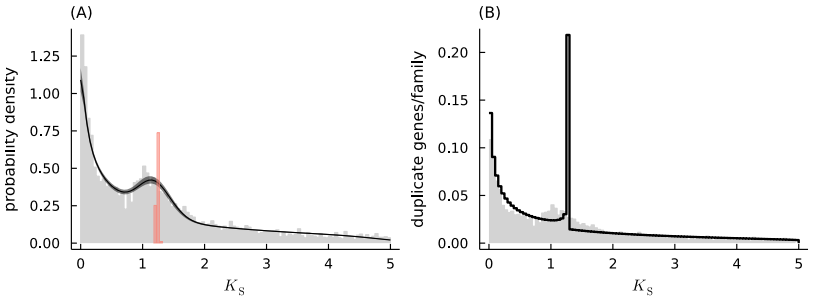


Figure 2.3: The model of Maere fitted to the *Vitis* age distribution. (A) Empirical K_S age distribution normalized as a probability distribution (histogram) and posterior predictive smoothed age distribution (posterior mean age distribution with 95% uncertainty intervals). The posterior distribution for the WGD time point (on a K_S scale) is shown in red. (B) Empirical age distribution for the observed number of duplicate genes per family (gray) and posterior mean unsmoothed age distribution (i.e. the predicted distribution for eq. 2.2 parameterized by the posterior mean parameter values).

$p(y|\lambda, \alpha, t)$ for the observed duplication ages on a K_S scale by simulating the deterministic model of eq. 2.2 (we use $\Delta t = 0.05 K_S$ and $T = 100$), applying Poisson smoothing and interpolating the density (using a quadratic B-spline). We use an exponential prior with mean 1 for both the duplication rate λ and the decay rates α_0 and α_1, and assign a discrete Uniform prior to the WGD age bin. We sample from the posterior using a simple adaptive Metropolis-Hastings algorithm (see Appendix A; for references to relevant code, see Appendix B). The posterior predictive empirical K_S distribution and K_S-scale age distribution are shown in fig. 2.3. The model appears to provide a very good fit, although with suspiciously narrow uncertainty intervals. We obtain a posterior mean duplication rate of 0.99 per gene per K_S with a (0.82, 1.10) 95% uncertainty interval, which is roughly similar to the short term rate estimated above using the young duplicate cohort. Similarly, applying the same model (without WGD) to *C. elegans*, we obtain a duplication rate estimate of 0.31 $(0.12, 0.45)$ duplications per gene per K_S. □

2.3 Probabilistic models of paranome evolution

While useful, the above models have important limitations. Both models consider the whole paranome as evolving in a deterministic manner, with a single constant rate of gene duplication for the entire population of genes. In the

deterministic setting, there is no principled way to relate either the predicted (equilibrium) number of genes in a family to observed family sizes or the predicted number of genes in an age interval with an observed age distribution, and inference amounts to fitting the model to the data assuming some optimization criterion (e.g. least-squares in the case of Lynch and Conery (2003)). Indeed, these models do not admit asking questions like: what is the probability that an extant family consists of n members? What is the probability that an extant duplicate pair derives from a duplication that happened a time t in the past? Similar to how we model long-term sequence evolution (see chapter 1), we would like to explicitly model the variation in outcomes of the assumed evolutionary process using stochastic models. It is arguably more reasonable to consider, for instance, any particular paranome and its associated age distribution a stochastic realization of some random process than to model the number of genes in a family using an ODE or difference equation. Furthermore, explicit probabilistic models allow using all the data directly for parameter inference using Bayesian methods or otherwise. It is therefore insightful to consider stochastic analogs of the simple models considered above. We will have extensive use for such models later when we consider gene family evolution along a species phylogeny, where the deterministic models are no longer applicable.

2.3.1 Branching processes and birth-death process models

Virtually all conceivable models of gene family evolution by small-scale duplication and loss, deterministic or probabilistic, assume birth-death like evolution of genes within families, where gene copies are treated as 'particles' which give rise to offspring particles with a certain rate. Conceived probabilistically, these models of evolution by SSDL naturally lead to continuous-time Markov processes which belong to the class of *branching processes*, *birth-death processes*, or both.

A continuous-time branching process is a model for a population of particles evolving according to some probabilistic law. Each particle in the system 'lives' for some (typically random) time interval, after which it gives rise to a random number of offspring particles distributed according to an offspring distribution with probability mass function (pmf) ξ and probability generating function (pgf) $g(s) = \sum_{k=0}^{\infty} \xi(k)s^k$. The number of particles in the population is assumed to evolve as a continuous-time Markov chain $\{X(t) \in \mathbb{N} \colon t \in \mathbb{R}^+\}$, and the transition probabilities $p_{ij}(t) = \mathbb{P}\{X(t) = j | X(0) = i\}$ satisfy the so-

called branching property (Athreya and Ney 1972)

$$\sum_{j=0}^{\infty} p_{ij}(t)s^j = \Big(\sum_{j=0}^{\infty} p_{1j}(t)\Big)^i$$

which is tantamount to requiring the evolution of different particles in the system to be statistically independent. More general continuous-time branching processes can be defined, with for instance age-dependence or multiple types of particles (see chapter 3), but they all satisfy the fundamental assumption that offspring particles behave independently conditional on the parent particle, so that the branching property holds. Because of the branching property, many interesting results can be derived for branching processes using manipulations of the associated generating functions.

Continuous-time birth-death processes (BDPs) are similarly defined as Markov processes on the state space of the nonnegative integers, but with the constraint that transitions occur only between neighboring states. The branching property need not hold for BDPs, but when it does, the BDP is obviously also a branching process. A rich body of theory for such processes has been developed in various fields, such as probability theory 'proper' (Kendall 1948; Karlin and McGregor 1957; Allen 2010), statistics (Crawford, Minin, and Suchard 2014; Tavaré 2018), operations research (in particular queueing theory, e.g. Kendall (1953)), epidemiology (Bailey 1990; Crawford and Suchard 2012; Stadler et al. 2012; Kühnert et al. 2014)[1], population genetics (Moran 1958; Crawford and Suchard 2012) and various parts of phylogenetics and evolutionary genomics, such as phylogenetic tree inference (Thompson et al. 1975; Rannala and Yang 1996), macroevolution (Yule 1925; Nee 2006; Lambert and Stadler 2013), phylodynamics (Stadler et al. 2012; Kühnert et al. 2014) and gene family evolution (Hahn et al. 2005; Csűrös and Miklós 2009). We will make extensive use of BDPs in the remainder of this thesis, and hence it will be worthwhile to work out some basic properties of them here.

More formally, a continuous-time Markov chain $\{X(t)\colon t \in \mathbb{R}^+\}$, where $X(t) \in \mathbb{N}$, is a time-homogeneous BDP if it has transition probabilities of the following form

$$p_{ij}(\Delta t) \stackrel{\text{def}}{=} \mathbb{P}\{X(t+\Delta t) = j \,|\, X(t) = i\} \tag{2.3}$$

[1]The by now all too familiar compartmental models for epidemic progression such as SIS, SIR and suchlike are instances of birth-death models.

$$= \begin{cases} \lambda_i \Delta t + o(\Delta t) & j = i + 1 \\ \mu_i \Delta t + o(\Delta t) & j = i - 1, i \geq 1 \\ 1 - (\lambda_i + \mu_i)\Delta t + o(\Delta t) & i = j \\ o(\Delta t) & |i - j| > 1 \end{cases} \tag{2.4}$$

Note that time-homogeneity implies that the transition probabilities over an interval $(t, t + \Delta t)$ only depend on Δt and not on t, justifying the above notation. The above specification, where the λ_i and μ_i are arbitrary functions of the state, defines the so-called *general* BDP. Note in particular that when $\lambda_i = 0$ for all $i > N$ for some N, the state space is effectively bounded by N. Karlin and McGregor (1957) were the first to study general BDPs *in extenso*, where they showed, among other things, the existence of transition probabilities, and developed a mathematically sophisticated theory for deriving transient distributions and other properties. However, analytical expressions or straightforward numerical methods for efficiently computing quantities such as transition probabilities associated with the general BDP are not available, and sophisticated numerical methods or approximations (for instance by truncating the state space and using the usual matrix-based methods for finite CTMCs) must be adopted. Crawford and Suchard (2012) for instance developed a numerical approach based on a continued fraction representation of the Laplace transform of the transient distribution to compute transition probabilities.

Tractable special cases of the BDP appear however for particular functions λ_i and μ_i. A very important special case is the *simple linear* BDP[2], obtained when $\lambda_i = i\lambda$ and $\mu_i = i\mu$ for all $i \in \mathbb{N}$. A suggestive way to characterize the linear BDP is to consider the evolution of a single particle in the population. Upon its birth, the particle lives for an exponentially distributed time with mean $1/(\lambda + \mu)$, after which the particle either is removed from the system or is replaced by two identical daughter particles, with probability $\mu/(\lambda + \mu)$ and $\lambda/(\lambda + \mu)$ respectively. In the latter case, two independent copies of the same process are started. Clearly then, the linear BDP is a continuous-time branching process with offspring pgf $g(s) = (\mu + \lambda s^2)/(\mu + \lambda)$. A solution for the pgf of the transition probabilities $p_{1j}(t)$ can be derived from the Kolmogorov forward equations (see e.g. Crawford and Suchard (2012) or Allen (2010)), yielding

$$f(s,t) = \sum_{j=0}^{\infty} p_{1j}(t)s^j = \frac{\mu(s-1) + (\lambda s - \mu)e^{(\mu-\lambda)t}}{\lambda(s-1) + (\lambda s - \mu)e^{(\mu-\lambda)t}}$$

[2]Sometimes also referred to as the Kendall process, after Kendall (1948).

and by the branching property we have $f_i(s,t) = \sum_{j=0}^{\infty} p_{ij}(t)s^j = f(s,t)^i$. In the *critical* case where $\lambda = \mu$, we have

$$f(s,t) = \frac{1-(\lambda t-1)(s-1)}{1-\lambda t(s-1)}$$

An important quantity is the *extinction probability* $\epsilon(t)$, i.e. the probability that a single particle does not have descendants after a time t. This can be obtained from the pgf by sending s to zero

$$\epsilon(t) = f(0,t) = \begin{cases} \frac{\mu(1-e^{(\mu-\lambda)t})}{\lambda-\mu e^{(\mu-\lambda)t}} & \lambda \neq \mu \\ \frac{\lambda t}{1+\lambda t} & \lambda = \mu \end{cases} \tag{2.5}$$

Bailey (1990) further derived an expression for the general transition probabilities using the pgf

$$p_{ij}(t) = \sum_{k=0}^{\min(i,j)} \binom{i}{k}\binom{i+j-k-1}{i-1}\alpha^{i-k}\beta^{j-k}(1-\alpha-\beta)^k \tag{2.6}$$

where $\alpha = \epsilon(t)$ and $\beta = (\lambda/\mu)\epsilon(t)$.

We denote by $h_{i,t}(j) = p_{ij}(t)$ the distribution of $X(t)$ for fixed t and $X(0) = i$ and refer to it as the *transient distribution*. The transient distribution $h_{1,t}(j)$ of the simple linear BDP has been referred to as a *shifted geometric distribution* by Csűrös and Miklós (2009), which has the following probability mass function

$$h_{1,t}(j) = \begin{cases} \alpha & j = 0 \\ (1-\alpha)(1-\beta)\beta^{j-1} & j > 0 \end{cases}$$

From this we see that the transient distribution of the number of descendants of a single particle *conditioned on non-extinction* is geometric with parameter $1-\beta$. Other properties such as the conditional mean and variance of $X(t)$ are readily derived from the pgf. Importantly, the linear BDP has no stationary distribution, with almost sure extinction for $\lambda < \mu$ and non-zero probability of so-called explosion for $\lambda > \mu$. The linear BDP has itself two other commonly encountered BDPs as special cases, being the *pure birth*[3] and *pure death* processes, obtained when $\mu = 0$ and $\lambda = 0$ respectively.

[3] Often referred to as the Yule process after Yule (1925), a pioneering work in the application of stochastic models to evolutionary problems, although Yule actually developed a more complicated model in that study (see below). Also sometimes referred to as the Yule-Furry process after Wendell Furry who applied it to problems related to radioactive decay.

2.3.2 A stochastic version of the model of Lynch & Conery

After this short detour introducing birth-death process models we return to the problems of gene family evolution from the single-genome perspective. The cognate stochastic variant of the model of Lynch and Conery (2003) would be a continuous-time BDP with the following transition probabilities

$$p_{ij}(\Delta t) = \begin{cases} (i+1)\lambda\Delta t + o(\Delta t) & j = i+1 \\ i\mu\Delta t + o(\Delta t) & j = i-1 \\ 1 - (i(\lambda+\mu)+\lambda)\Delta t + o(\Delta t) & i = j \\ o(\Delta t) & |i-j| > 1 \end{cases}$$

Where the random variable $X(t)$ denotes the number of duplicated genes per ancestral gene at time t in a gene family, so that the family size is $X(t) + 1$. This model is almost – but not quite – the simple linear BDP discussed above, the difference being that the duplication rate λ_i when in state i is equal to $(i+1)\lambda$ instead of $i\lambda$. In fact, the model is equivalent to a special case of the so-called linear *birth-death-immigration* process (BDIP) (Bailey 1990; Allen 2010) with immigration rate ν equal to the birth rate λ. Furthermore, it turns out that the same BDIP model is used for modeling the insertion and deletion dynamics in the famous TKF91 model used for evolutionary sequence alignment (Thorne, Kishino, and Felsenstein 1991; Holmes and Bruno 2001). In fig. 2.4 we show example realizations simulated from this model.

Letting $p_i(t) \stackrel{\text{def}}{=} \mathbb{P}\{X(t) = i\}$ we can derive the differential difference equations for the $p_i(t)$ using the Markov property, considering a time interval Δt we note that

$$\begin{aligned} p_0(t+\Delta t) &= (\mu\Delta t)p_1(t) + (1-\lambda\Delta t)p_0(t) + o(\Delta t) \\ p_i(t+\Delta t) &= i(\lambda\Delta t)p_{i-1}(t) + (i+1)(\mu\Delta t)p_{i+1}(t) \\ &\quad + [1 - \{i(\lambda+\mu)+\lambda\}\Delta t]p_i(t) + o(\Delta t) \end{aligned}$$

Subtracting $p_i(t)$ from both sides, dividing by Δt, and sending $\Delta t \to 0$, we obtain

$$\begin{aligned} \frac{dp_0(t)}{dt} &= \mu p_1(t) - \lambda p_0(t) \\ \frac{dp_i(t)}{dt} &= i\lambda p_{i-1}(t) + (i+1)\mu p_{i+1}(t) - [i(\lambda+\mu)+\lambda]p_i(t) \end{aligned}$$

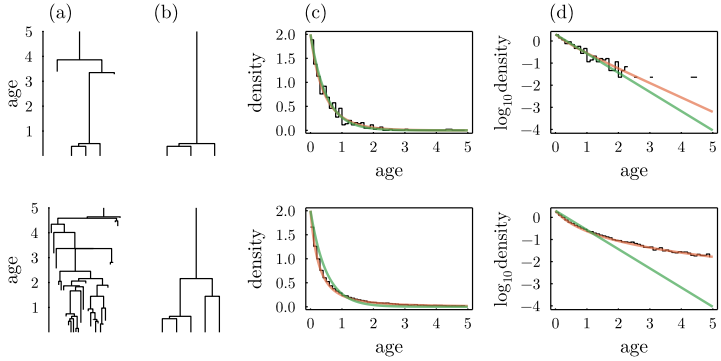


Figure 2.4: Simulations of the paranome under the stochastic BDIP model. The top row shows results from a model with $\lambda = 0.75$ and $\mu = 2$, whereas the bottom row shows results for $\lambda = 1.5$ and $\mu = 2$. (a) Representative random gene trees simulated from the BDIP models. (b) The reconstructed trees for the gene trees in (a), pruning the duplicates which are lost before reaching the present. (c) The whole-paranome age distributions based on a simulation of 5000 independently evolving gene families. The orange line shows the probability density function for the model (eq. 2.8), whereas the green line shows the exponential approximation $\mu e^{-\mu t}$ thereof. (d) as in (c) but on a $\log_{10}$ scale.

The pgf for the BDIP can be derived (Bailey 1990) but we will have little use for it here. It will be more interesting to consider the equilibrium situation studied by Lynch and Conery (2003) in the deterministic model, which corresponds in the stochastic setting to the *stationary distribution* π of the Markov process. Assuming $\lambda < \mu$, the stationary distribution can be derived from the differential-difference equations above by setting $dp_i/dt = 0$. We arrive at the following stationary pmf for $i \geq 1$

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \pi_0$$

where π_i of course denotes $\mathbb{P}\{X(t) = i\}$ at stationarity. Employing the constraint that $\sum_{i=0}^{\infty} \pi_i = 1$ and assuming $\lambda < \mu$, we further obtain

$$\pi_0 = \left(\sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i\right)^{-1} = 1 - \frac{\lambda}{\mu}$$

As a result the stationary probability distribution of the number of duplicate genes per ancestral gene is a geometric distribution with parameter $1 - \lambda/\mu$.

Assuming stationarity, we can obtain the age distribution under this model as follows. Let T denote the time before the present of a random duplication and let I_o be the indicator of whether the duplication is observed in an extant genome, we have

$$\begin{aligned} F(\tau) = \mathbb{P}\{T < \tau | I_o = 1\} &= \frac{\mathbb{P}\{T < \tau, I_o = 1\}}{\mathbb{P}\{I_o = 1\}} \\ &\propto \int_0^\tau \sum_{k=1}^{\infty} \left(\frac{1}{k}\bar{\epsilon}(t) + \left(1 - \frac{1}{k}\right)\bar{\epsilon}(t)^2 \right) k\lambda\pi_k dt \\ &= \lambda\left(1 - \frac{\lambda}{\mu}\right) \int_0^\tau \bar{\epsilon}(t)\left(\bar{\epsilon}(t)\frac{\mu^2}{(\lambda-\mu)^2} + \epsilon(t)\frac{\mu}{\mu-\lambda}\right) dt \end{aligned} \tag{2.7}$$

Where $\bar{\epsilon}(t) = (1 - \epsilon(t))$ is the probability that a lineage, extant at time t, survives until the present given that it evolves according to a linear birth-death process (see eq. 2.5). Some calculus shows that

$$F(t) = \frac{\mu\left(e^{\lambda t} - e^{\mu t}\right)}{\lambda e^{\lambda t} - \mu e^{\mu t}}$$

and the probability density function is

$$f(t) = \frac{\mu(\lambda-\mu)^2 e^{(\lambda+\mu)t}}{(\lambda e^{\lambda t} - \mu e^{\mu t})^2} \tag{2.8}$$

For $\lambda \ll \mu$, $f(t) \approx \mu e^{-\mu t}$, giving, as expected of course, an exponential age distribution as in the approximation used in the deterministic model of Lynch and Conery (2003). In fig. 2.4 we show simulated age distributions together with the predicted densities. We confront the equilibrium model with the *C. elegans* paranome in the next example.

Example (*C. elegans* BDIP). We infer parameters of the BDIP using the age and size distribution, assuming stationarity. We consider two models, one with the age distribution given by the plain BDIP prediction (eq. 2.8), and another where the age distribution is assumed to be a mixture of the BDIP equilibrium distribution and a uniform component (see below). Specifically, the former amounts to

$$\begin{aligned} \mu &\sim \text{Uniform}(0, 100) \\ \alpha &\sim \text{Beta}(1, 1) \\ t|\mu, \lambda = \alpha\mu &\sim_{\text{iid}} f \qquad \text{(see eq. 2.8)} \end{aligned}$$

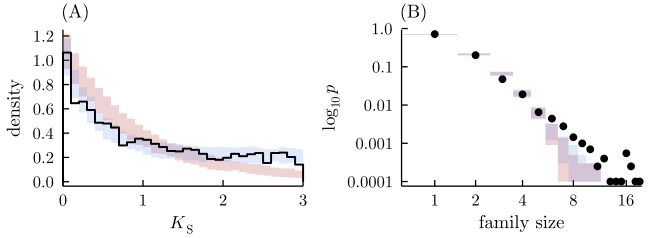


Figure 2.5: Posterior predictive distribution for the *C. elegans* (a) age distribution and (b) family size distribution for the linear BDIP model. The black lines/dots display the observed data. In red the 95% uncertainty interval is shown for the posterior predictive distribution for the linear BDIP model alone while in blue the same is shown for a mixture of the linear BDIP age distribution (eq. 2.8) and a Uniform$(0, 3)$ distribution as model, with a uniform prior on the mixture proportions.

$$x|\alpha \sim_{\text{iid}} \text{Geometric}(1-\alpha)$$

Where x is the vector of gene family sizes and t the observed duplication ages. The mixture model is a straightforward extension of this model. Note that this does not model the individual families explicitly, but rather the entire paranome under the said equilibrium assumptions. Posterior predictive distributions are shown in fig. 2.5. The basic model yields a loss rate of 1.1 $(1.0, 1.1)$ and duplication rate of 0.3 $(0.3, 0.4)$ per gene per K_S, whereas the mixture model yields a loss rate of 2.6 $(2.3, 3.0)$ and duplication rate of 0.8 $(0.7, 1.0)$ per gene per K_S. These estimates are different, but of the same order of magnitude, as the estimates reported above. □

Although mathematically convenient, this linear birth-death process is clearly an inappropriate model of long-term gene family evolution in general. Ignoring the distorting influence of whole-genome duplications, neither the geometric stationary distribution of gene family sizes, nor the exponential age distribution are observed in whole-genome data sets. For small families, with ≤ 5 gene copies say, the simple linear birth-death process may provide a reasonable model for the size distribution, perhaps salvaging the widespread usage of related models in phylogenetic analyses of gene family evolution (see our next chapter however for a detailed study). Nevertheless, the lack of fit of this simple model should already prompt us to search for more realistic models which could give us more insight in the evolutionary processes that govern the evolution of gene families.

Three main avenues for improvement are directly suggested by visual inspection of the age and family size distributions. Firstly, the *power law* like dis-

tribution of gene family sizes – witnessed by the quasi linear relationship between frequency and family size when plotted along double logarithmic axes (fig. 2.5) – suggests that duplication and loss in a gene family does not follow the assumption of constant rates within or across families assumed in the simple BDP model. Secondly, the observation that the K_S distribution does not decline to zero at higher K_S values suggests that duplicate genes may become essential to some extent, such that they are no longer susceptible to loss at the rate of newborn duplicates. In a model for the age distribution this can be accounted for in an *ad hoc* way by assuming a mixture model with a uniform component, as we did in our *C. elegans* example above (fig. 2.5), however such a device does not explicitly model the evolutionary causes of the phenomenon. Recall that we effectively assumed a 'base' set of ancestral genes (gene families) that are not subject to loss (we assumed only duplicate gene copies to be susceptible to gene loss). This is of course a simplification, as sometimes genes from the 'base' set do get lost, while duplicated gene copies could adopt functions that alleviate their redundancy that makes them susceptible to gene loss (by, for instance, subfunctionalization or neofunctionalization). Alternatively, some form of age-dependent loss, as in e.g. the model of Maere et al. (2005), could account for this, although this would entail a steady growth of the genome over time. Lastly, of course WGD and other large-scale events also affect both the family size and age distribution, and accounting for these events when relevant is of course a necessary condition for obtaining a decent model fit. We will consider the power law issue in the next section, while the other issues are addressed in the next chapter.

2.3.3 The power law size distribution

While the simple paranome model outlined above predicts a geometric family size distribution, it has since long been known that gene family sizes tend to show a power law tail (Huynen and Van Nimwegen 1998; Karev et al. 2002; Karev, Wolf, and Koonin 2003). A discrete random variable X follows a power law distribution if it has a pmf of the form

$$\mathbb{P}(X = k) = \alpha k^{-\beta}$$

Power laws have been observed in virtually any scientific domain, and they arise from a variety of mathematical models. Several authors have considered models which could give rise to power law size distributions in evolutionary genomics. In one of the earliest paper on the subject, Huynen and Van Nimwegen (1998) considered a very general but rather *ad hoc* model in which gene

family sizes fluctuate due to multiplicative noise. Specifically, the family size X_t at time t is assumed to be the product of X_{t-1} and α_t, where α_t is a random variable with some probability law, typically with mean 1. Such a model can be shown to lead to a power law distribution for large t. The essential feature of this model is that the per-gene duplication and loss rates are correlated *within* families, which in the latter model is a consequence of the magnitude of the fluctuations growing linearly with family size.

Contrary, however, to what Huynen and Van Nimwegen (1998) claimed, there are a number of ways by which relatively simple birth-death like processes can give rise to power law size distributions, some of which are relevant for genome evolutionary processes. We consider two of these, starting with Yule's model.

2.3.3.1 Yule's model

Consider for instance a *pure-birth* process $X(t)$ with $X \in \{1, 2, \dots\}$ and $t > 0$. The transition probabilities of the process are characterized by the following differential difference equations

$$\frac{dp_{ij}(t)}{dt} = (j-1)\lambda p_{i,j-1}(t) - j\lambda p_{ij}(t)$$

Which can be solved succesively to find the transient distribution at time t

$$p_i(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{i-1}$$

which is again a geometric distribution, this time with parameter $e^{-\lambda t}$. In particular, if we consider N particles independently evolving through time by a pure-birth process, and label all particles at time t that were derived from the same ancestral particle as members of the same *family*, the distribution of family sizes $X(t)$ at time t will be geometric.

Yule (1925) (!) considers such a pure-birth process, embedded in another pure-birth process. While we will not have much use for it here, we shall consider it in some more detail because of its historical significance. Specifically, he assumes that each particle (species in Yule's model) in the system does not only duplicate at rate λ, but also gives rise to new families (or genera in Yule's application) at rate γ. As a result, families in the system are of different random *ages*, with the age A of a family a random variable with an exponential law $f_A(t) = \gamma e^{-\gamma t}$. Yule considers the family size distribution under this model,

which is a compound distribution

$$\mathbb{P}\{X(t) = k\} = \mathbb{E}_A[\mathbb{P}X(\tau) = k|A = \tau] \tag{2.9}$$

$$= \int_0^t \mathbb{P}\{X(\tau) = k|A = \tau\} f_A(\tau) d\tau \tag{2.10}$$

$$= \int_0^t \left[e^{-\lambda\tau}(1 - e^{-\lambda\tau})^{k-1}\right]\left[\gamma e^{-\gamma\tau}\right] d\tau \tag{2.11}$$

The first factor in the integral is the pmf of the geometric distribution with parameter $e^{-\lambda\tau}$, where τ, the age of the family, itself has an exponential pdf.

For $k = 1$ and $t \to +\infty$ this can be computed as

$$\pi_1 = \mathbb{P}\{X(t) = 1\} = \int_0^\infty \gamma e^{-\gamma\tau} e^{-\lambda\tau} d\tau = (1 + \lambda/\gamma)^{-1}$$

The general solution can be recursively expressed as

$$\pi_k = \frac{(k-1)(\lambda/\gamma)}{1 + k(\lambda/\gamma)} p_{k-1}, \quad k > 1$$

Which is the main result of Yule's 1925 paper. Writing $\rho = \gamma/\lambda$, the resulting pmf can be expressed analytically as

$$\pi_k = \rho \frac{\Gamma(1+\rho)\Gamma(k)}{\Gamma(1+\rho+k)} = \rho \mathrm{B}(1+\rho, k)$$

where B is the Beta function. The resulting distribution is referred to as the *Yule-Simon* distribution with parameter ρ after Simon (1955) who generalized Yule's derivation of this probability law. For large k, we have that $\Gamma(k)/\Gamma(1+\rho+k) \sim k^{-(1+\rho)}$, so that $\pi_k \approx \rho\Gamma(1+\rho)k^{-(1+\rho)}$. The size distribution under this model therefore has an approximate power law tail.

While this pure-birth model does not appear to make sense as a model of gene family evolution by duplication and loss, note that when $\tau \sim$ Exponential(γ) we have that $\lambda\tau \sim$ Exponential(γ/λ) = Exponential(ρ). In general, we have shown that the compound distribution

$$W|\rho \sim \text{Exponential}(\rho)$$
$$X|W \sim \text{Geometric}(e^{-W})$$

implies that $X|\rho \sim$ YuleSimon(ρ). Note that ρ is the *rate* of the exponen-

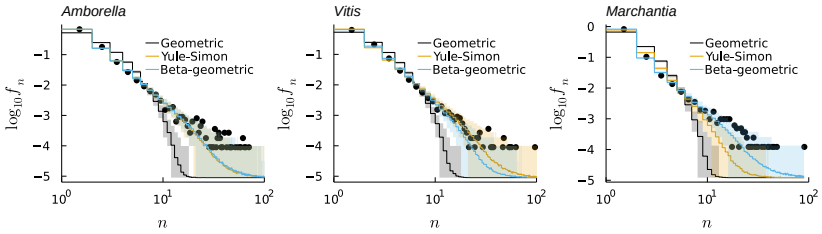


Figure 2.6: Geometric, Yule-Simon and beta-geometric stationary distributions fitted to observed gene family size distributions for three paranomes (for *Amborella*, *Vitis* and *Marchantia*. The black dots show the observed size distribution (across all taxa), whereas the lines show the mean frequencies and 95% posterior predictive intervals based on 1000 simulations from the posterior predictive distribution for each model.

tial distribution (not *scale*). In other words, when a random variable $X|p \sim \text{Geometric}(p)$, and p is distributed such that $-\log p \sim \text{Exponential}(\rho)$, the marginal pmf of X is a YuleSimon(ρ) distribution. Recall that under the simple linear BDIP model derived above, the stationary distribution for the family size X is Geometric($1 - \lambda/\mu$), so that when $\log(1 - \lambda/\mu) \sim \text{Exponential}(\rho)$, we would obtain a Yule-Simon distribution. The latter is equivalent to assuming that $\mu/(\mu - \lambda) \sim \text{Pareto}(\rho, 1)$, which may be easier to interpret.

This suggests that when duplication and loss rates are assumed to vary across gene families according to a fairly simple parametric model, a power law size distribution can be obtained, and we need not invoke more complicated BDP models to explain power law behavior (as several authors have done e.g. Karev et al. 2002). Similar conclusions were reached in a simulation study by Hughes and Liberles (2008), where the authors found that heterogeneity in pseudogenization rates was sufficient for generating a power-law tail. They did not however present the simple statistical arguments we consider here. As expected, we find that the Yule-Simon distribution provides a much better fit to the observed size distribution than the geometric model (fig. 2.6).

2.3.3.2 Beta-geometric distribution

A more flexible two-parameter model for rate variation across families in the linear birth-death process, which also gives rise to a power law size distribution, is obtained when we assume that $1 - \lambda/\mu \sim \text{Beta}(\alpha, \beta)$ (or equivalently, $\lambda/\mu \sim \text{Beta}(\beta, \alpha)$). The resulting compound stationary distribution becomes

a so-called beta-geometric BG(α, β) distribution with pmf

$$\begin{aligned}\pi_k &= \int_0^1 \eta(1-\eta)^{k-1} f(\eta;\alpha,\beta) d\eta \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 \eta^{\alpha}(1-\eta)^{\beta+k-2} d\eta \\ &= \frac{\mathrm{B}(\alpha+1,\beta+k-1)}{\mathrm{B}(\alpha,\beta)}\end{aligned}$$

for $k > 0$ and where $f(\eta;\alpha,\beta) = \mathrm{B}(\alpha,\beta)^{-1}\eta^{\alpha-1}(1-\eta)^{\beta-1}$ is the density of the beta distribution. The mean of this distribution is $(\alpha+\beta)/\alpha$, and the expected value of λ/μ across families under this model is $(\alpha+\beta)/\beta$. Note that, again, for large k, the BG distribution is approximated by a power law, in this case $\pi_k \sim \mathrm{B}(\alpha,\beta)^{-1}\Gamma(\alpha+1)k^{-(\alpha+1)}$.

An alternative parameterization which is often more useful defines the BG distribution using the mean $\eta = \alpha/(\alpha+\beta)$ and dispersion parameters $\zeta = \alpha+\beta$, so that $1/\eta$ is the mean family size, and the distribution approaches a geometric distribution as $\zeta \to +\infty$. The BG distribution provides an excellent fit to observed gene family size distributions (fig. 2.6). In addition to providing a reasonable model for the paranome size distribution, the BG distribution serves as an apt prior for the number of ancestral genes in a family in phylogenetic analyses (see later chapters). Furthermore, a fit of the BG distribution to observed data provides information on the ratio of λ/μ and the variation of this ratio across families.

2.4 Concluding remarks and segue

We have used this short and rather incomprehensive chapter to introduce some fundamental ideas that will be of use later, in particular birth-death process models of gene family evolution. We already see the important tension between the short and long term evolutionary processes that frustrate our attempts at modeling genome evolution. We will not dwell on the single-genome setting further in this thesis, but we note that there is still much room for improvement here. In particular, a tractable probabilistic variant of the model of Maere et al. (2005) would be of considerable interest for the statistical inference of ancient WGDs from K_S age distributions, a problem which at present remains addressed in a rather *ad hoc* way in the literature, using for instance Gaussian mixture models (see e.g. Zwaenepoel and Van de Peer 2019b;

Tiley, Barker, and Burleigh 2018; Sensalari, Maere, and Lohaus 2021). Age-dependent BDPs present however considerable mathematical challenges. We shall come back to this at the end of the next chapter.

A particularly interesting avenue for developing better statistical models of paranome evolution, possibly accounting for WGDs, would be to make use of the theory of coalescent point processes (CPPs) (Lambert and Stadler 2013) for modeling K_S-scale gene trees for paralogous families. CPPs are random processes formulated retrospectively (see chapter 5 for more on coalescent processes) which are associated with BDP models, the latter being forward-time models (i.e. describing evolution from an ancestral state to a present state). In particular, Lambert and Stadler (2013) show how some forms of age-dependence in BDP models lead to tractable CPPs, which may be used as a basis for deriving the likelihood function of observed ages for such processes. Adopting the CPP viewpoint could unlock likelihood-based inference for more realistic BDP models using empirical age distributions, an approach we have started to explore and hope to develop further in the future.

In the next chapter we take the BDP models to a comparative genomic setting, modeling the evolution of gene family content (but not the gene trees or age distributions explicitly) along a species phylogeny. We take up the challenge of statistical inference of ancient WGDs from comparative genomic data, building on the work of Rabier, Ta, and Ané (2014), and develop more adequate (or so we hope, at least) models of gene family evolution, using the foundations provided in the preceding paragraphs.

3 Phylogenetic birth-death process models

In this chapter[1] we shall take some of the ideas introduced in chapter 2 to the comparative genomic setting. Specifically, we will consider birth-death process (BDP) models of gene family evolution defined along a phylogeny which describes the evolutionary relationships among a collection of genomes. The goal is to learn about the processes of genome evolution by means of statistical phylogenetic analyses of gene family content across a set of species.

The statistical problems in the present chapter have the following general form. Consider a species tree S with leaves $\mathcal{L}(S)$ and with branch lengths on some suitable timescale. For some gene family, let X_u with $u \in V(S)$ be a discrete random variable denoting the number of genes in the family (the *gene count*) in the (ancestral) genome associated with node u of S. Let S_u be the subtree of S rooted in u, and let $X_{[u]}$ denote the random vector of gene counts at the leaves of S_u, i.e. $(X_v : v \in \mathcal{L}(S_u))$, assuming some suitable ordering on the leaf set (fig. 3.1). We reserve o as a symbol for the root node of S and define $X \stackrel{\text{def}}{=} X_{[o]}$. X will be referred to as a *phylogenetic profile*, and the $X_{[u]}$ for $u \neq o$ as *partial* phylogenetic profiles. A phylogenetic BDP model of gene family evolution is defined by the following recursive generative process:

$$
\begin{aligned}
X_o &\sim \pi(\cdot) \\
X_v | X_u = x &\sim h_{x,t_v}(\cdot) \qquad\qquad v \in V(S),\ v \neq o,\ u = \rho(v)
\end{aligned}
$$

Where π is some discrete distribution for the number of ancestral genes at the root of S, $\rho(u)$ denotes the parent node of u, and h_{x,t_v} is the transient distribution for a BDP along branch $\langle u, v \rangle$ of length t_v when started with x initial genes (see chapter 2). These interdependent random variables define a probabilistic graphical model with the same graph structure as S (see chapter 1).

[1] This chapter freely draws from our published work in Zwaenepoel and Van de Peer (2019a), Zwaenepoel and Van de Peer (2020), and Zwaenepoel and Van de Peer (2021) (preprint).

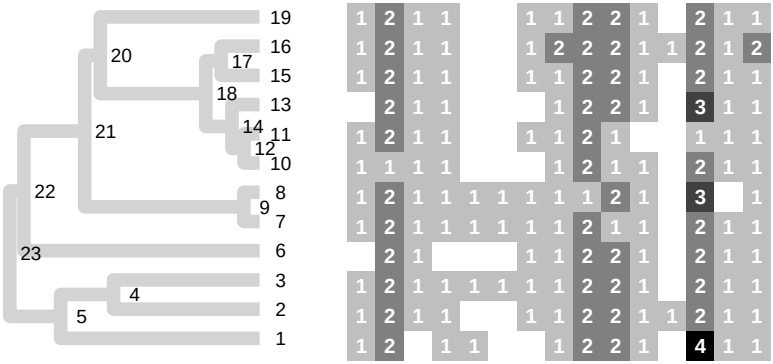


Figure 3.1: Example of a collection of phylogenetic profiles, or gene count matrix, for 15 gene families from a 12 taxon species tree. Each column of the matrix represents a phylogenetic profile, with each row corresponding to the species represented by the aligned tip in the tree on the left. Numbers in the tree label the associated nodes, so that in our notation, for instance, $X_{[14]} = (X_{10}, X_{11}, X_{13})$, which for the first family in this example is the observed partial profile $(1, 1, 0)$.

Assuming such a phylogenetic model of gene family evolution with parameter θ, our goal will be to approximate the posterior distribution

$$p(\theta|X, S) \propto \prod_{i=1}^{N} p(X^{(i)}|\theta, S)p(\theta|S)$$

conditional on a collection of observed phylogenetic profiles $(X^{(1)}, \ldots, X^{(N)})$, in order to learn about θ and the processes of genome evolution that it is supposed to model.

The rest of this chapter is organized in three parts: first we develop the statistical theory and computational tools for phylogenetic BDP models based on the linear BD(I)P. We describe in detail the techniques for computing the phylogenetic likelihood of continuous-time branching process models, due to Csűrös and Miklós (2009), on which we base our new implementation for Bayesian statistical inference of these models in a probabilistic programming context. We then consider the phylogenetic BDP model with whole-genome duplications of Rabier, Ta, and Ané (2014), embed it in our Bayesian framework for modeling gene content evolution and develop a new approach for model-based inference of ancient WGDs using reversible-jump MCMC (Zwaenepoel and

Van de Peer 2020). Lastly, in an attempt to overcome some of the issues associated with the linear BDP model of gene family evolution, we develop and study a model based on a multi-type branching process (Zwaenepoel and Van de Peer 2021).

3.1 Bayesian inference for the linear BDP and variants

3.1.1 The likelihood for phylogenetic BDPs

The likelihood function $p(X|\theta, S)$ presents significant challenges already for fairly simple models. Recall that a general BDP is a continuous-time Markov chain on the non-negative integers with infinitesimal rates as in eq. 2.4. We can hence express the process in terms of its (infinite dimensional) rate matrix (infinitesimal generator, see also chapter 1)

$$Q = \lim_{\Delta t \downarrow 0} \frac{P(\Delta t) - P(0)}{\Delta t} = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & \dots \\ \mu_1 & -(\mu_1 + \lambda_1) & \lambda_1 & \dots \\ 0 & \mu_2 & -(\mu_2 + \lambda_2) & \dots \\ 0 & 0 & \mu_3 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Where $P(t)$ is the transition probability matrix with entries $p_{ij}(t)$. Noting that $p_{ij}(t + \Delta t) = \sum_k p_{ik}(t)[\delta_{kj} + q_{kj}\Delta t + o(\Delta t)]$ we get the Kolmogorov backward differential equation

$$\frac{dP(t)}{dt} = P(t)Q$$

which describes the transition probabilities in terms of the infinitesimal generator. The parallel with CTMC models with a finite state space immediately suggests an approach for computing the likelihood function under a general BDP. If we put an upper bound on the state space, we can compute a transition probability matrix $P(t) = \exp(Qt) = I + Qt + \frac{Q^2}{2}t^2 + \dots$ for each branch of the species tree. This transition probability matrix can then be used to compute the phylogenetic likelihood in the ordinary way, using the pruning algorithm, i.e. by marginalizing over the unobserved random variables in the probabilistic graphical model (e.g. fig. 3.2). While in practical applications one should always be able to choose a reasonable bound on the state space, this approach is nevertheless rather inelegant and can be computationally quite demanding as well as numerically unstable (Crawford, Minin, and Suchard 2014).

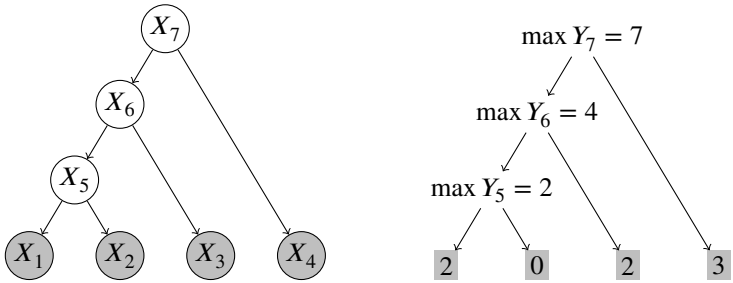


Figure 3.2: (Left) Probabilistic graphical model for a schematic four-taxon species tree with relevant random variables X_v, $v \in [1..7]$. The gray nodes are assumed to be observed while the white nodes are not, i.e. the associated phylogenetic profile is $X = X_{[7]} = (X_1, X_2, X_3, X_4)$. (Right) Example of an observed phylogenetic profile $(2, 0, 2, 3)$ with the upper bound for the random variables Y_v (the number of ancestral genes which leave observed descendants in $\mathcal{L}(S_v)$) indicated for the internal nodes of S.

A different approach was suggested in Csűrös and Miklós (2009; see also Csűrös 2022). The key idea there is to condition on a different random variable while marginalizing over the possible evolutionary trajectories in the pruning algorithm. Instead of conditioning on the ancestral gene count X_u at node u, we condition on the number of *surviving lineages* Y_u at node u, that is, the number of lineages at node u which leave observed descendants at the leaves of S_u.

$$\mathbb{P}\{X|\theta, S\} = \sum_{k=0}^{\infty} \mathbb{P}\{X|Y_o = k, \theta, S\}\mathbb{P}\{Y_o = k|\theta, S\} \tag{3.1}$$

The first factor, considered as a function of θ, is referred to as the *conditional survival likelihood*. We can recursively expand $\mathbb{P}\{X|Y_o = k, \theta, S\}$ along the tree, suggesting a pruning algorithm for computing the likelihood. For instance, writing out the marginalization explicitly for fig. 3.2, we have (dropping dependence on S and θ throughout)

$$\begin{aligned} &\mathbb{P}\{X|Y_7 = k\} \\ &= \mathbb{P}\{Y_1, Y_2, Y_3, Y_4|Y_7 = k\} \\ &= \sum_{i=0}^{\infty} \mathbb{P}\{Y_1, Y_2|Y_5 = i\}\mathbb{P}\{Y_5 = i, Y_3, Y_4|Y_7 = k\} \end{aligned} \tag{3.2}$$

$$= \sum_{i=0}^{\infty} \mathbb{P}\{Y_1, Y_2 | Y_5 = i\} \Big(\sum_{j=0}^{\infty} \mathbb{P}\{Y_5 = i, Y_3 | Y_6 = j\} \mathbb{P}\{Y_6 = j, Y_4 | Y_7 = k\} \Big)$$

Note however that we do not have conditional independence of child nodes given the parent node for the Y_u, e.g. $\mathbb{P}\{Y_1, Y_2 | Y_5\} \neq \mathbb{P}\{Y_1 | Y_5\}\mathbb{P}\{Y_2 | Y_5\}$ in general[2]. Importantly, the number of lineages Y_u at node u which survive until the present is bounded by the observed partial profile $X_{[u]}$, so that the infinite sums in eq. 3.1 and eq. 3.2 are actually finite and we no longer need to artificially truncate the state space (fig. 3.2, right). While, conceptually, conditioning on survival works for general phylogenetic BDPs, efficient computation of the required transition probabilities for the conditional process, as well as the probability $\mathbb{P}\{Y_o = k | \theta, S\}$ that k ancestral genes survive until the present, relies crucially on specific properties of the linear BDP.

3.1.2 Specializing for the linear BDP

The linear BDP (see chapter 2), as a model of gene family evolution, is characterized by a per-gene duplication rate λ and a per-gene loss rate μ, i.e. a general BDP with the additional constraints that $\lambda_i = i\lambda$ and $\mu_i = i\mu$ for $i \geq 0$. We have seen that the linear BDP has a closed form for the transition probability $p_{ij}(t)$ (eq. 2.6), so that instead of relying on matrix exponentiation, we could compute the entries of the transition probability matrix directly. This is the strategy adopted in the influential work of Hahn et al. (2005), among others. While evading numerical issues associated with computing $\exp(Qt)$, the method still relies on a user-defined upper bound and appears not to make optimal use of the special structure of the linear BDP, in particular the branching property which entails independent evolution in distinct lineages. Here we show how the branching property enables efficient computation of many important quantities associated with the process conditioned on survival.

3.1.2.1 Probability generating functions and extinction probabilities

Recall the probability generating function (pgf) for a single lineage in the linear BDP

$$f(s,t) = \sum_{j=0}^{\infty} p_{1j}(t) s^j = \frac{\mu(s-1) + (\lambda s - \mu)e^{(\mu-\lambda)t}}{\lambda(s-1) + (\lambda s - \mu)e^{(\mu-\lambda)t}}$$

[2]To see the dependence, note for instance that $Y_1 + Y_2 \geq Y_5$ always.

The branching property amounts to the fact that $\sum_{j=0}^{\infty} p_{ij}(t)s^j = f(s,t)^i$. Consider now a phylogenetic BDP for species tree S. Let $f_u(s) = f(s, t_u)$ be the pgf for the number of descendant genes Z_u at node $u \in V(S)$ starting from a single ancestral gene at the beginning of the branch leading to node u. Let $g_u(s)$ be the pgf for the total number of genes extant at node u. For a child node v of u, an argument well known from the theory of branching processes shows that

$$\begin{aligned} g_v(s) &= \sum_{k=0}^{\infty} \mathbb{P}\{X_v = k\}s^k \\ &= \sum_{k=0}^{\infty} s^k \sum_{m=1}^{\infty} \mathbb{P}\{X_v = k | X_u = m\}\mathbb{P}\{X_u = m\} \\ &= \sum_{m=1}^{\infty} \mathbb{P}\{X_u = m\} \sum_{k=0}^{\infty} \mathbb{P}\Big(\sum_{i=1}^{m} Z_{i,v} = k \Big) s^k \\ &= \sum_{m=1}^{\infty} \mathbb{P}\{X_u = m\}(f_v(s))^m \\ &= g_u(f_v(s)) \end{aligned}$$

Similarly, the *joint* pgf for the number of genes at child nodes v and w of u is given by $g_u(f_v(s)f_w(s))$. This generalizes neatly to the entire tree structure. For instance, notwithstanding a parenthetical labyrinth, we can represent the joint pgf for the phylogenetic profile in fig. 3.2 as

$$\begin{aligned} g_S(\mathbf{s}) = g_S(s_1, s_2, s_3, s_4) &= \sum_{k_1=0}^{\infty} \cdots \sum_{k_4=0}^{\infty} \mathbb{P}\{X_1 = k_1, \ldots, X_4 = k_4\} s_1^{k_1} \ldots s_4^{k_4} \\ &= g_7(f_6(f_5(f_1(s_1)f_2(s_2))f_3(s_3))f_4(s_4)) \end{aligned}$$

Here $g_7(s)$ will be the pgf of the prior distribution on the number of genes at the root of S (we write $g_o(s)$ for general trees). We write $f_S(\mathbf{s})$ for the associated joint pgf of the number of descendants at the leaves of S coming from a single ancestral gene at the root of S, so that $g_S(\mathbf{s}) = g_o(f_S(\mathbf{s}))$.

The joint pgf allows efficient computation of many important quantities, in particular various kinds of extinction probabilities. For instance, the probability of extinction everywhere (i.e. the observed profile is the zero profile $\mathbf{0} = (0, 0, \ldots, 0)$) corresponds to $g_S(\mathbf{0})$. The joint pgf for the phylogenetic process can hence be evaluated using the pgfs for the lineage-specific branching

processes by means of a postorder traversal, enabling the computation of extinction probabilities across the phylogeny. For instance, the probability that a single lineage observed at node u goes extinct in the subtree S_u is simply $f_{S_u}(\mathbf{0})$. The latter quantity plays a crucial role in phylogenetic BDPs, and we will denote it by ϵ_u. More complicated probabilities can be computed recursively from the generating functions, such as the probability of extinction at none of the leaves of S, or the marginal probability of extinction at certain leaves (e.g. the probability of extinction everywhere except in leaf node '1' is $g_S(1, 0, 0, \ldots, 0)$, *etc.*).

3.1.2.2 Transition probabilities conditional on survival

The branching property further allows us to derive the transition probabilities for the process conditional on non-extinction of a particular number of ancestral genes. To see this, recall that the linear BDP is characterized by the following transition probabilities (where we assume a fixed time interval and drop the dependence on t)

$$p_{1,k} = \begin{cases} \alpha & k = 0 \\ (1-\alpha)(1-\beta)\beta^{k-1} & k > 0 \end{cases} \tag{3.3}$$

With α and β as in eq. 2.6. Note that for $k > 1$ we have $p_{1,k} = \beta p_{1,k-1}$. Clearly, such relations can only hold when different genes evolve independently. Consider now a similar linear BDP, but where there is a fixed probability ϵ that a descendant gene is not actually observed. One can, again using generating function arguments which rely crucially on independence, show (Csűrös and Miklós 2009) that the probability that a single ancestral gene will leave k *observed* descendants is

$$\tilde{p}_{1,k} = \sum_{j=0}^{\infty} \binom{k+j}{k} p_{1,k+j} \epsilon^j (1-\epsilon)^k = \begin{cases} \alpha' & k = 0 \\ (1-\alpha')(1-\beta')\beta'^{k-1} & k > 0 \end{cases}$$

where

$$\alpha' = \frac{\alpha(1-\epsilon) + (1-\beta)\epsilon}{1-\beta\epsilon}$$
$$\beta' = \frac{\beta(1-\epsilon)}{1-\beta\epsilon}$$

In other words, the modified process, which is a linear BDP with a fixed probability (here ϵ) of not observing a random descendant, is again a linear BDP. We call it the ϵ-modified process.

From there we can derive the transition probabilities conditional on the number of surviving lineages. Let $X(t)$ be a linear BDP and define

$$q_{ij}(t) = \mathbb{P}\{X(t) = j, \forall k \in [1..i] : Z_k(t) > 0 | X(0) = i\} \tag{3.4}$$

Here $Z_k(t)$ is the random variable counting the number of descendants after a time t of the kth ancestral gene. In words then, $q_{ij}(t)$ is the probability that i ancestral genes all have descendants and give rise to j genes in total after a time t. Clearly, $q_{i,0} = 0$, $q_{1,j} = p_{1,j}$ and $q_{i,j} = 0$ for $j < i$. Csűrös and Miklós (2009) showed that, again because of independence, the $q_{ij}(t)$ can be related recursively as follows (dropping dependence on t)

$$\begin{aligned} q_{ij} &= \sum_{k=1}^{j-i+1} p_{1,k} q_{i-1,j-k} \\ &= p_{1,1} q_{i-1,j-1} + \sum_{k=2}^{j-i+1} p_{1,k} q_{i-1,j-k} \\ &= p_{1,1} q_{i-1,j-1} + \beta q_{i,j-1} \end{aligned}$$

Consider now k ancestral genes at node $u \in S$ evolving down the branch leading to node v of length t_v, and let ϵ_v be the extinction probability associated with node v. The probability that none of the k genes go extinct and give rise to $k+j$ lineages at node v which have observed descendants will be $q_{k,k+j}(t_v)$, using the α' and β' associated with the ϵ_v-modified process.

3.1.2.3 The phylogenetic likelihood

While these transition probabilities are not directly helpful to compute the likelihood, together with the extinction probabilities across the tree derived above they enable an efficient, if somewhat complicated, algorithm for computing the likelihood of a phylogenetic profile (Csűrös and Miklós 2009; Csűrös 2022). Firstly, note that (dropping dependence on θ and S as before)

$$\mathbb{P}\{X | Y_o = k\} = \mathbb{P}\{X_{[u]}, X_{[v]}, Y_o = k | X_o = k\} / \mathbb{P}\{Y_o = k | X_o = k\}$$

where the denominator is $(1-\epsilon_o)^k$. Let $\mathcal{A}_v$ be the set of lineages extant at the parent node of v which leave observed descendants in S_v, we can simply take this set to be $\{1, 2, \ldots, k\}$ when there are k surviving lineages. We can derive a recursive algorithm starting from the following decomposition

$$\begin{aligned}
&\mathbb{P}\{X_{[u]}, X_{[v]}, Y_o = k | X_o = k\} \\
&\quad = \sum_{B \subset [1..k]} \mathbb{P}\{X_{[u]}, X_{[v]}, \mathcal{A}_u = B, \mathcal{A}_v - \mathcal{A}_u = B^c | X_o = k\} \\
&\quad = \sum_{B \subset [1..k]} \mathbb{P}\{X_{[u]} | \mathcal{A}_u = B\} \mathbb{P}\{X_{[v]}, \mathcal{A}_v - \mathcal{A}_u = B^c | \mathcal{A}_u = B\} \\
&\qquad\qquad \times \mathbb{P}\{\mathcal{A}_u = B | X_o = k\} \\
&\quad = \sum_{j=0}^{k} \mathbb{P}\{X_{[u]} | Y'_u = j\} \mathbb{P}\{X_{[v]}, Y'_v \geq k - j | Y_o = k\} \binom{k}{j} (1-\epsilon'_u)^j \epsilon'^{k-j}_u
\end{aligned}$$

Here Y'_u is a random variable denoting the number of genes at the start of the branch leading to node u which have observed descendants down S_u, so that $Y'_u \leq Y_o$ and $Y'_u + Y'_v \geq Y_o$. Similarly ϵ'_u is the probability that a single lineage extant at the start of the branch leading to node u goes extinct down S_u, i.e. $\epsilon'_u = f_u(f_{S_u}(\mathbf{0}))$. Csűrös and Miklós (2009) further show how the first two factors can be computed recursively, using only the extinction probabilities and transition probabilities for the conditional process derived above.

3.1.2.4 The prior distribution on the root

The second factor in eq. 3.1, the probability that k ancestral genes at the root leave observed descendants, does not provide further challenges for the linear BDP. Recall that we have defined the generative process in terms of a prior distribution π on the number of lineages X_o at the root. We can express $\mathbb{P}\{Y_o = k | \theta, S\}$ in terms of this distribution

$$\mathbb{P}\{Y_o = k | \theta, S\} = \sum_{j=0}^{\infty} \mathbb{P}\{Y_o = k | X_o = k + j, \theta, S\} \pi(k + j)$$

This is valid for the general BDP. For the linear BDP however, applying the branching property allows us to express this as

$$\sum_{j=0}^{\infty} \binom{k+j}{j} \epsilon_o^j (1-\epsilon_o)^k \pi(k+j)$$

Depending on the choice of π, this series may or may not have a closed form.

Since the linear BDP *conditional on non-extinction* has a geometric quasi-stationary distribution, a natural choice for π would be a geometric distribution, and this has been adopted in several studies (e.g. Rabier, Ta, and Ané (2014), Zwaenepoel and Van de Peer (2019a), Zwaenepoel and Van de Peer (2020)). For a geometric distribution with parameter η (and mean $1/\eta$), the series simplifies to

$$\begin{aligned}
\mathbb{P}\{Y_o = k|\theta, S\} &= \sum_{j=0}^{\infty} \binom{k+j}{j} \epsilon_o^j (1-\epsilon_o)^k \eta(1-\eta)^{k+j-1} \\
&= (1-\epsilon_o)^k \frac{\eta(1-\eta)^{k-1}}{(1-(1-\eta)\epsilon_o)^{k+1}}
\end{aligned}$$

The size distribution of gene families is however universally overdispersed with respect to the geometric distribution, showing an approximate power law tail, as we discussed in chapter 2. For the beta-geometric distribution, which should be a more reasonable prior distribution (recall fig. 2.6), we did not find a closed form solution. However, using the property of the Beta function that $\mathrm{B}(\alpha, \beta+1) = \mathrm{B}(\alpha,\beta)\frac{\beta}{\alpha+\beta}$ we can find

$$\begin{aligned}
\mathbb{P}\{Y_o = k|\theta, S\} &= \sum_{j=0}^{\infty} \binom{k+j}{j} \epsilon_o^j (1-\epsilon_o)^k \frac{\mathrm{B}(\alpha+1, \beta+k+j-1)}{\mathrm{B}(\alpha,\beta)} \\
&= (1-\epsilon_o)^k \sum_{j=0}^{\infty} A_{k,j}
\end{aligned}$$

With the recursion relations

$$\begin{aligned}
A_{1,0} &= \frac{\alpha}{\alpha+\beta} \\
A_{k,0} &= A_{k-1,0} \frac{\beta+k-2}{\alpha+\beta+k-1}
\end{aligned}$$

$$A_{k,j} = A_{k,j-1} \frac{\epsilon_o(k+j)(\beta+k+j-2)}{j(\alpha+\beta+k+j-1)}$$

Which allow efficient approximation of the infinite series above by some partial sum sufficiently far in the series. The sum converges very rapidly, so that in practice no more than 10 terms are needed. For both the geometric and beta-geometric distributions, these results are easily modified if the relevant domain is $k \geq 0$ rather than $k \geq 1$ (as is the case in the BDIP setting, see below).

3.1.2.5 Conditioning on the sampling process

It is important to take into account sampling biases when conducting likelihood-based statistical inference. One source of bias is that the phylogenetic BDP model generates all-zero profiles with non-zero probability, whereas such profiles cannot be observed in typical comparative genomic data sets. Let E_u be the event of extinction of a family below node u and let $\bar{E}_u$ be its negation. For a phylogenetic profile y, the appropriate conditional likelihood is then

$$p(y|\theta, S, \bar{E}_o) = \frac{p(y|\theta, S)}{1 - \mathbb{P}\{E_o\}}$$

Where $\mathbb{P}\{E_o\}$ is easily obtained using the generating function techniques for computing extinction probabilities.

Furthermore, we often apply certain filtering steps to the data, and we should condition the likelihood accordingly. For instance, to rule out *de novo* gain of genes in arbitrary subtrees of the phylogeny, we can filter the data so that at least one gene is present in each clade stemming from the root of the species tree (as in e.g. Rabier, Ta, and Ané (2014) and Zwaenepoel and Van de Peer (2019a)). Labeling the daughter nodes of o by u and v, the likelihood of a profile y conditional on the event of non extinction in both clades is then

$$p(y|\theta, S, \bar{E}_u \cap \bar{E}_v) = \frac{p(y|\theta, S)}{\mathbb{P}\{\bar{E}_u \cap \bar{E}_v\}} = \frac{p(y|\theta, S)}{1 - \mathbb{P}\{E_u\} - \mathbb{P}\{E_v\} + \mathbb{P}\{E_o\}}$$

Where the relevant probabilities can again be computed from the probability generating functions.[3]

[3]Note that $\mathbb{P}\{E_u\} \neq \epsilon_u$, the latter being the probability of a *single lineage extant at node u* going extinct in S_u, whereas the former is the probability that a *family* has no observed descendants in S_u. The latter probability therefore involves also the prior on the number of lineages at

3.1.3 Extensions for other BDPs

In Csűrös and Miklós (2009) (see also Csűrös (2022)) it is further shown how the above results generalize readily to the birth-death immigration setting, already considered in chapter 2. In the linear BDIP, we assume a constant immigration rate κ, so that $p_{i,i+1}(\Delta t) = (i\lambda + \kappa)\Delta t + o(\Delta t)$. The transient distribution of a linear BDIP is a negative binomial distribution, and the process has a stationary distribution whenever $\lambda < \mu$. As we have shown in the previous chapter, this limiting distribution is geometric in the special case where $\kappa = \lambda$. The linear BDIP can be used as a model of gene family evolution by gene duplication, gene loss and horizontal gene transfer or *de novo* gene origin (together referred to as 'gene gain'). A different use for the linear BDIP was suggested in the previous chapter: if we assume $\kappa = \lambda$ and model the number of *excess* genes in a family (i.e. the number of duplicate genes *per* family), we recover a stochastic analog of the model of Lynch and Conery (2003), which corresponds to a general BDP with $\mu_i = \max\{0, (i-1)\mu\}$ and $\lambda_i = i\lambda$ for the total gene count. This model can be applied to putatively 'essential' gene families which cannot get lost without inflicting an insurmountable fitness cost.

Independence properties in probabilistic models of biological phenomena usually signal assumptions made for the sake of mathematical convenience rather than biological or epistemic relevance. In the case of gene family evolution, a particularly pressing issue derives from the functional roles of gene duplicates. If duplicates within a family are functionally redundant to some extent, independence and a constant per-gene loss rate do not seem very reasonable assumptions. We address these issues in much more detail below, but we note here that general BDPs do not necessarily have the branching property, so that a system of k particles will not have the same time evolution as the sum of k systems with a single particle. This is because in general BDPs, the per-gene duplication or loss rate can depend on the number of genes in the family. When the branching property does not hold, we cannot easily resort to algorithms of the sort proposed by Csűrös and Miklós (2009). However, importantly, we need not necessarily resort to the matrix exponentiation method either. Crawford and Suchard (2012) (see also Crawford, Minin, and Suchard (2014) and Crawford, Ho, and Suchard (2018)) showed how numerical inversion of the Laplace transform of the transient distribution is a feasible strategy for computing transition probabilities. Moreover, the same authors proposed expectation-maximization (EM) algorithms for inference of discretely observed BDPs. These ideas remain to be taken to the phylogenetic setting.

the root.

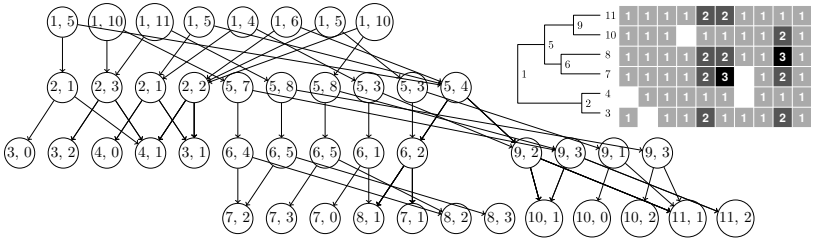


Figure 3.3: Example of the directed acyclic graph (DAG) data structure associated with a particular count matrix and phylogeny S, shown in the inset on the top right. Each node in the graph represents a unique partial phylogenetic profile and is labeled as $(u, \max Y_u)$ where the $u \in V(S)$ correspond to the node labels in S and $\max Y_u$ is the upper bound on the number of ancestral genes at u which have observed descendants in S_u. An edge between two nodes indicates that the bottom node is a partial profile with respect to the upper node. In this particular example, conducting dynamic programming along the DAG reduces the number of partial likelihood evaluations from 110 for a naive algorithm to 45.

However, as we shall argue below, the general BDP may not be the most adequate generalization if our goal is to develop more realistic and biologically relevant models of gene family evolution.

3.1.4 Implementation

We implemented algorithms for computing the phylogenetic likelihood in a flexible Julia (Bezanson et al. 2017) package called DeadBird. We implemented an efficient variant of the algorithm by Csűrös and Miklós (2009) outlined above which computes the conditional survival likelihood for a collection of gene families using dynamic programming (DP) across a directed acyclic graph (DAG) structure which summarizes the observed data (more specifically a so-called multitree, see fig. 3.3). This dynamic programming approach, which is akin to the approach of Pond and Muse (2004), will ensure that likelihood computations are only performed once for each unique partial phylogenetic profile in the data set. This strategy can lead to significant speedups over naive DP algorithms, although the gains are dependent on the variability in the data. For the complete data set of which the example in fig. 3.3 shows a tiny segment, an over ten-fold reduction in the number of partial likelihood evaluations can be achieved in this way.

Our implementation admits the specification of phylogenetic BDP models with branch-specific duplication, loss and gain rates; mixture models of rate

variation across families[4]; various prior distributions on the ancestral gene count and various likelihood corrections for filtering strategies (all of these are exemplified in the next section). We also implement algorithms for sampling ancestral states, enabling inference of ancestral gene family sizes and Bayesian analysis of gene family expansion and contraction. Furthermore, our implementation plays nicely with many other tools in the Julia scientific computing ecosystem, and is, for instance, fully compatible with (forward-mode) automatic differentiation (AD) and the Turing.jl probabilistic programming language (Ge, Xu, and Ghahramani 2018). This allows extremely flexible model specification, where we can embed phylogenetic BDPs in essentially arbitrary probabilistic programs (see the next section for examples). Compatibility with AD systems admits the application of highly efficient and accurate optimization and sampling techniques which use gradient information (gradient descent and cognate optimizers, Hamiltonian Monte Carlo and cognate samplers), enabling both maximum likelihood and Bayesian inference of model parameters. Not only is our implementation *much* more flexible than other implementations we know of, it is also quite fast, as illustrated in the following example.

Example (CAFE comparison). We compare our implementation for statistical inference of model parameters for phylogenetic BDPs against the popular CAFE software (De Bie et al. 2006b; Mendes et al. 2021). Assuming a model where the duplication and loss rates across the tree are determined by a single parameter λ (referred to as the *turnover rate*), we estimate λ by maximum likelihood for a data set consisting of 1000 gene families for a nine-taxon dicot species tree. We obtain a MLE of $\hat{\lambda} = 0.282$ events per 100 My using the BFGS[5] optimizer in about 2.7 seconds, whereas CAFE (v5, Mendes et al. (2021)), which uses a zeroth-order method instead (the Nelder-Mead downhill simplex method), takes 1 minute and 34 seconds to arrive at $\hat{\lambda} = 0.275$ events per 100 My. For a data set of 10000 gene families from 12 mammal species, available from the CAFE5 GitHub repository, we obtain a MLE of $\hat{\lambda} = 0.19$ events per 100 My in about 70 seconds using our implementation, while CAFE takes slightly less than 5 minutes to find the same estimate up two two significant digits. All analyses were conducted using a single CPU. Note that both CAFE and our implementation can make use of multiple cores if desired. □

[4]Note that when the data is modeled according to a discrete mixture of phylogenetic BDPs, the DAG structure can only be used when we marginalize over the different components.

[5]Broyden-Fletcher-Goldfarb-Shanno algorithm, for those who insist. This is a popular algorithm to find the minimum of some function $f(x)$, using the gradient $\nabla f(x)$ and an iteratively improved estimate of the Hessian matrix.

3.1.5 But does it fit?

The phylogenetic linear BDP is widely used as a model of gene family evolution in evolutionary genomics, where researchers employ popular ML-based packages such as CAFE (Hahn et al. 2005), BadiRate (Librado, Vieira, and Rozas 2012) or Count (Csűrös 2010) to estimate duplication and loss rates and seek to infer patterns of gene family expansion and contraction based on the inferred model. The adequacy of the model to empirical data is however rarely assessed. In this section, we address the question of how well the linear BDP model does in fact fit empirical gene count data, and we use posterior predictive simulations to address this question. The examples in this section motivate the developments presented in the rest of this chapter.

Example (rice). We conduct Bayesian inference of gene duplication and loss rates for a data set of six closely related rice (*Oryza*) species, taken from Stein et al. (2018).[6] Because these species are closely related, an assumption of constant duplication and loss rates across the species tree may be reasonable. After filtering out gene families which do not have at least one gene in both clades stemming from the root, the data set consists of 30638 gene families. As an illustration of how statistical phylogenetic BDP models can be defined using our DeadBird library, we include a code snippet for the specification of the phylogenetic linear BDP here:

```
# Load the required libraries
using DeadBird, NewickTree, Distributions, Turing
using CSV, DataFrames

# Load the gene count matrix and species tree
data = CSV.read("oryza-6taxa.csv", DataFrame)
tree = readnw(readline("oryza-6taxa.nw"))

# Construct the DAG object
dag, bound = CountDAG(data, tree)

# Define the probabilistic program
@model model1(dag, bound, tree) = begin
    η ~ Beta(1,1)
    ζ ~ Exponential(1)
    λ ~ Exponential(0.2)
```

[6] Data sets used in our examples throughout this dissertation are listed in Appendix B.

```
    μ ~ Exponential(0.2)
    θ = ConstantDLG(λ=λ, μ=μ, κ=zero(λ))
    p = ShiftedBetaGeometric(η, ζ+1)
    dag ~ PhyloBDP(θ, p, tree, bound)
end

# Sample from the posterior distribution using the NUTS sampler
chain = sample(model1(dag, bound, tree), NUTS(), 500)
```

The probabilistic program is defined using the Turing.jl library below `@model model1 [ ... ]`. The notation closely follows the usual mathematical notation for Bayesian hierarchical models (see chapter 1). We use Exponential priors for the rate parameters with a mean of 0.2, based on a preliminary analysis using maximum likelihood estimation with a fixed prior distribution for the ancestral gene count.[7] `ConstantDLG(λ=λ, μ=μ, κ=κ)` gathers the rate parameters of a linear phylogenetic BDP model where the duplication, loss and gain rates λ, μ and κ are constant across the phylogeny (i.e. all branches are assumed to share the same rates). `ShiftedBetaGeometric(η, ζ)` constructs a BG distribution with pmf $\pi(k)$ for $k \geq 1$. The `PhyloBDP` object gathers all the required components for the complete specification of a phylogenetic BDP, including the tree structure, appropriate conditioning strategy and ancestral gene count distribution (`p`). We here use the default conditioning strategy, which assumes the data is filtered so that there is at least one gene in each clade stemming from the root. Note that in this example, we do not fix the prior distribution on the number of ancestral genes in a family, but treat the prior parameters themselves as latent variables. We assume that $\zeta > 1$, so that the variance of the BG distribution is bounded. We obtain a sample from the posterior distribution using the NUTS[8] algorithm (Hoffman and Gelman 2014). In tbl. 3.1 we show relevant summary statistics for the sample of the posterior.

At a glance, posterior predictive simulations appear to indicate a reasonable fit, however some problems are immediately apparent (fig. 3.4). In particular, the

[7] We note that, in this analysis, we have a very large data set so that the posterior is dominated by the likelihood, and the parameterization of the prior distribution for the rates does not have an appreciable influence. We cannot, however, simply resort to an improper flat prior on the positive real line, since the posterior with a likelihood derived from the transient distribution of a linear BDP appears to be improper in that case. We thank prof. Clement for drawing attention to this.

[8] For 'No U-turn Sampler' a HMC sampling algorithm with some additional heuristics that make it the go-to sampler in many probabilistic programming environments. See also Appendix A.

Table 3.1: Posterior mean, standard deviation, Monte Carlo standard error (MCSE), effective sample size (ESS) and 2.5% and 97.5% quantiles of the marginal posterior distribution for the four parameters in the constant rates phylogenetic linear BDP analysis of the six-taxon rice data set. Duplication and loss rates are in units of events per gene per million years.

parameter	mean	std.	MCSE	ESS	2.5%	97.5%
η	0.91	0.002	0.0001	484	0.91	0.91
ζ	3.36	0.212	0.0097	278	3.01	3.77
λ	0.08	0.001	0.0001	467	0.08	0.09
μ	0.39	0.003	0.0001	758	0.39	0.40

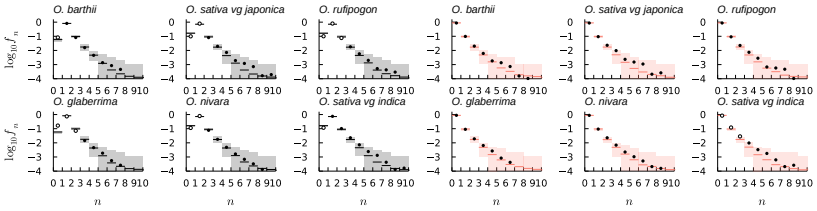


Figure 3.4: Posterior predictive simulations for the rice data set. In gray, posterior predictive family size distributions (log-frequency $\log_{10} f_n$ as a function of family size n) are shown for the linear phylogenetic BDP analysis with constant rates across the phylogeny. In red, the same is shown for the BDIP model, where along the x-axis the number of duplicate genes per family is shown. The dots mark the observed family size distribution in each respective genome, with a dot marked in white when it falls outside of the 95% posterior predictive interval.

frequencies for small families are not well fitted by the model. For each of the six genomes, the number of extinct gene families is not compatible with the posterior predictive distribution, with the observed frequency falling outside the 95% posterior predictive interval. For four out of six genomes, the same holds for the frequency of single-copy families. Furthermore, on average, 1.4 simulations are required to obtain a profile compatible with the filtering condition. Together with the predicted number of completely extinct families per observed family (which is 0.02), this implies that, roughly, some 11000 families should be filtered from the data, whereas the actual data suggests this to be more on the order of 3000 families. Together these observations suggest the linear BDP fits the data rather poorly. Note that predictions for the much rarer large families tend to be compatible with the observed data. This appears largely due to the prior distribution on the ancestral gene count being sufficiently flexible to accommodate the power law tail.

We conducted a similar analysis but using the linear BDIP with $\kappa = \lambda$ for the number of duplicate genes per family (i.e. the phylogenetic BDP analog of the Lynch and Conery (2003) model). Note that this model applies only to families which leave observed descendants at all leaves of S (or 'nowhere-extinct' families), which amounts to 19169 gene families for the six-taxon rice data set. The probabilistic program defining the model is:

```
@model model2(dag, bound, tree) = begin
    η ~ Beta(1,1)
    ζ ~ Exponential(1)
    μ ~ Exponential(1)
    λ = μ * (1 - η)
    θ = ConstantDLG(λ=λ, μ=μ, κ=λ)
    p = BetaGeometric(η, ζ+1)
    dag ~ PhyloBDP(θ, p, tree, bound, cond=:none)
end
```

Here the DAG object is based on the matrix of 'excess' genes, so that 0 indicates a single-copy state. We now use `cond=:none` to indicate that no likelihood correction is needed (there are no phylogenetic profiles with non-zero probability that are not observed because of our filtering steps). Recall that the stationary distribution of this model is geometric with parameter $1 - \lambda/\mu$, so that by setting $\lambda = \mu(1 - \eta)$, we constrain the model in such a way that the implied stationary distribution has the same mean as the prior distribution on the number of genes at the root. The results are summarized in tbl. 3.2.

Table 3.2: Posterior summary for the BDIP model applied to the rice data. See also tbl. 3.1. Note that $\lambda = (1 - \eta)\mu$ in this model.

parameter	mean	std.	MCSE	ESS	2.5%	97.5%
η	0.89	0.002	0.0001	322	0.88	0.89
ζ	2.01	0.138	0.0061	376	1.75	2.27
μ	0.74	0.012	0.0004	385	0.72	0.77
λ	0.08	0.001	0.0001	-	0.08	0.09

Although this analysis applies only to a subset of the data analyzed above under the linear BDP model, we consider a crude comparison of the estimated rates of gene family evolution between the two analyses, as these should be roughly similar if our modeling assumptions (in particular the assumptions of independent evolution of distinct families and independent evolution of gene copies within a family) are appropriate. The duplication rate estimate is indeed indistinguishable compared to the estimate for the ordinary linear BDP, and

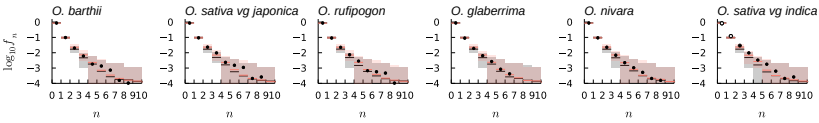


Figure 3.5: Posterior predictive simulations for the discrete mixture BDIP model applied to the six-taxon rice data set. In gray and red posterior predictive family size distributions are shown for the ordinary BDIP and the discrete mixture respectively. See also fig. 3.4.

the estimate for η also corresponds closely to its analog in the linear BDP model. Notably however, the loss rate per duplicate gene estimated for the linear BDIP is almost double the loss rate parameter for the BDP model, and the ζ parameter decreased considerably. The latter may be a result of the $\eta = 1 - \lambda/\mu$ constraint connecting the root prior and stationary distribution. Since the BDIP process implies a geometric stationary distribution and thus a gradual relaxation of the power law tail to a geometric tail, the ζ parameter can 'overcompensate' for this by increasing the overdispersion in the root family size distribution. Posterior predictive simulations further indicate a good fit to the observed family size distributions within individual genomes (fig. 3.4).

The results above already suggest it not to be necessary to invoke rate variation across branches or families to explain the observed data (except perhaps for *O. sativa vg. indica*). Nevertheless, for the sake of illustration, we shall now consider rate variation across families using the following discrete mixture model (variation across branches will be considered below):

```
@model model3(dag, bound, tree, K=4) = begin
    η ~ Beta(1, 1)
    ζ ~ Turing.FlatPos(1.)
    μ ~ Turing.FlatPos(0.)
    α = discretebeta(η, ζ, K)
    λ = μ .* (1 .- α)
    θ = [ConstantDLG(λ=λ[i], μ=μ, κ=λ[i]) for i=1:K]
    p = BetaGeometric(η, ζ)
    M = [PhyloBDP(θ[i], p, tree, bound, cond=:none) for i=1:K]
    dag ~ MixtureModel(M)
end
```

Note that there are countless strategies to account for rate heterogeneity across families, but here we again base ourselves on the beta-geometric hypothesis as outlined in chapter 2. Furthermore, in this model we make the 'unifor-

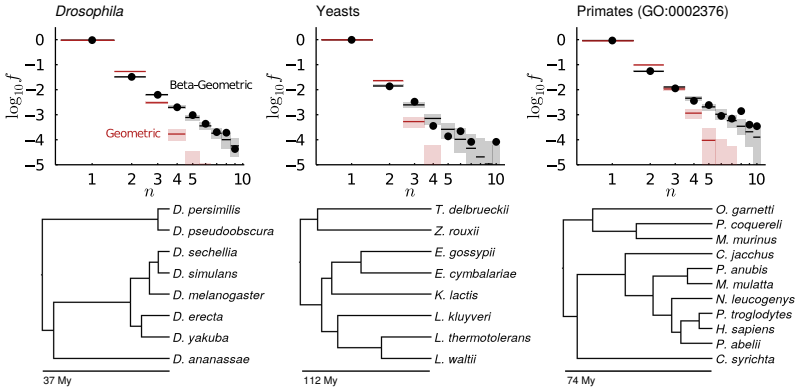


Figure 3.6: Top row: Geometric and beta-geometric stationary distributions fitted to observed gene family size distributions for three data sets. The black dots show the observed size distribution (across all taxa), whereas the lines show the mean frequencies and 95% posterior predictive intervals based on 1000 simulations from the posterior predictive distribution for both models. Bottom row: Phylogenies for the relevant data sets.

mitarian'[9] assumption that the patterns of rate heterogeneity that putatively determine the shape of the ancestral family size distribution are the same as the patterns of rate heterogeneity relevant for the evolutionary process within the species tree. Specifically, we assume a fixed μ across families, and that $\lambda_i = (1 - Z_i)\mu$ for family i, where Z_i is a beta-distributed random variable with mean η and dispersion parameter ζ, the same parameters as for the beta-geometric ditribution on the ancestral family size. We approximate this model by a discrete mixture, a popular approach for dealing with rate heterogeneity in phylogenetic models (e.g. Yang 2006), using four rate classes in this example. We find that accounting for rate heterogeneity across families does not appear to lead to a significantly better fit (fig. 3.5), although, as expected, this does lead to a slight increase in the ζ parameter ($\hat{\zeta} = 2.23$, $[2.07, 2.37]$ 95% interval). □

Example (*Drosophila*). We study the same simple models in the context of a

[9] Uniformitarianism is the doctrine that the natural laws which legislate today are the same as those which did in the past. In the context of genome evolution it refers more concretely to the assumption that the processes and rates of genome evolution in the past were, on average, the same as in the present (Lynch 2007). As such, the doctrine of uniformitarianism translates into a *principle* for the specification of priors and sampling distributions in the application of Bayesian statistics to evolutionary biology.

data set consisting of 12 *Drosophila* species (Drosophila 12 Genomes Consortium 2007), extensively studied in e.g. Hahn, Han, and Han (2007), Heger and Ponting (2007) and Stark et al. (2007). A dated species tree was downloaded from the TimeTree database (Kumar et al. 2017) and gene families were inferred using OrthoFinder (v2.4) (Emms and Kelly 2019). In these analyses, we fix the η and ζ parameter based on a fit of the beta-geometric distribution to the non-extinct gene family size distribution (not taking into account the phylogeny), resulting in $\hat{\eta} = 0.96$ and $\hat{\zeta} = 4.0$ (fig. 3.6, tbl. 3.3). When considering a data set of 12163 gene families across a subset of eight species (applying the usual filtering strategy of discarding gene families that have no observed descendants in one of the two clades stemming from the root) the standard linear BDP model yields an estimated loss rate of 0.24 expected loss events per gene per 100 My; whereas considering the set of gene families that are retained with at least one copy across all taxa, and assuming the linear BDIP model, we obtain a loss rate of 3.94 expected loss events per *duplicated* gene per 100 My.

Table 3.3: Marginal posterior rate parameter estimates for the *Drosophila*, yeast and primates data. All rates are on a scale of 'events per gene per 100 My'. *Critical linear BDP* refers to the constraint that $\lambda = \mu$ (leading to a critical branching process) whereas *linear BDIP+B4* refers to a model with λ/μ distributed across families according to a discretized Beta distribution with $K = 4$ classes. For all analyses, the prior on the number of lineages at the root was a beta-geometric distribution with η and ζ parameters fixed to the marginal posterior mean values obtained from the stationary distribution fit (fig. 3.6, *Drosophila*: $\eta = 0.96, \zeta = 4.01$, yeasts: $\eta = 0.98, \zeta = 4.06$, primates: $\eta = 0.93, \zeta = 3.12$).

Parameter	*Drosophila*	Yeasts	Primates (GO:0002376)
Critical linear BDP			
λ	0.19 (0.19, 0.20)	0.040 (0.038, 0.041)	0.15 (0.14, 0.15)
Linear BDP			
λ	0.17 (0.16, 0.17)	0.022 (0.020, 0.023)	0.13 (0.12, 0.14)
μ	0.24 (0.23, 0.25)	0.059 (0.056, 0.062)	0.17 (0.16, 0.18)
Linear BDIP			
λ	0.18 (0.17, 0.19)	0.016 (0.014, 0.019)	0.12 (0.11, 0.14)
μ	3.88 (3.66, 4.10)	1.65 (1.43, 1.88)	2.27 (2.07, 2.46)
Linear BDIP + B4			
λ	0.16 (0.14, 0.18)	0.010 (0.007, 0.014)	0.18 (0.15, 0.22)
μ	3.90 (3.64, 4.20)	1.47 (1.18, 1.14)	2.49 (2.18, 2.89)

Posterior predictive simulations show again that the linear BDP does not fit the data well for small gene families. In contrast with the rice data set, the basic lin-

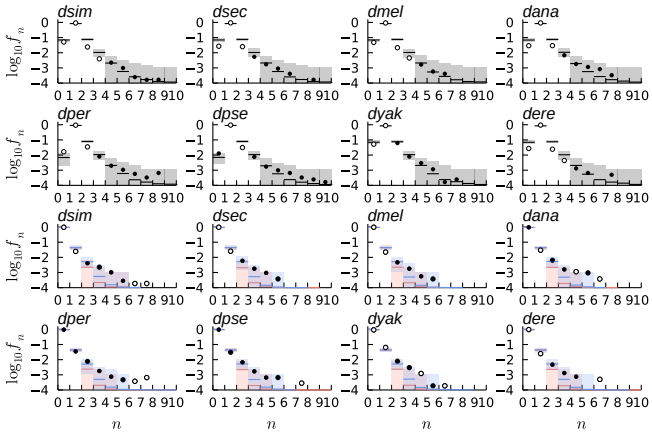


Figure 3.7: Posterior predictive gene family size distributions for the *Drosophila* data. In black the results for the linear BDP are shown while in blue and red the results for the linear BDIP with and without rate heterogeneity across families for the nowhere-extinct gene families are shown respectively. Taxon labels are derived from the first three letters of the species names (fig. 3.6). See also fig. 3.4.

ear BDIP for the nowhere-extinct families now also fails to provide a decent fit. The difference appears to be due to the timescales involved, with the estimated loss rates high enough to result in a more or less geometric size distribution at the leaves of the species tree despite the beta-geometric root count distribution. However, acknowledging rate heterogeneity under the beta-geometric assumption (assuming a fixed μ and using a $K = 4$ class discrete approximation to a beta distribution with unknown parameters for $1 - \lambda/\mu$) leads to a somewhat improved fit with similar duplication and loss rates. (fig. 3.7, tbl. 3.3). □

Example (yeasts). Exactly the same analysis was conducted for a data set of 4855 gene families from eight yeast species (the set of species included in the yeast gene order browser (YGOB, v7-Aug2012; Byrne and Wolfe (2005)) that did not undergo the *Saccharomyces cerivisiae* genome duplication) (fig. 3.6). For this data set we inferred a dated species tree using r8s (Sanderson 2003) under the molecular clock assumption, using the phylogeny inferred by OrthoFinder and a root calibration of 112 My (Beimforde et al. 2014). The estimated rates for this data set are markedly lower, yielding a duplication rate on the order of 0.01 – 0.04 duplication events per gene per 100 My (tbl. 3.3). Again a very strong difference is observed between the linear BDP and BDIP model. Model fit is similar to the *Drosophila* example, with little improve-

ment after accounting for heterogeneity across families. □

Example (primates). The same analyses were performed for a data set from 11 primate species, where we selected 1921 gene families which contain a human homolog annotated with the gene ontology (GO) term GO:0002376 (immune system process). In this analysis, we exclude the human lineage from the data so that this filtering strategy does not bias our estimates too much. Using a dated species tree downloaded from TimeTree, we obtain results roughly similar to the *Drosophila* data set (tbl. 3.3), where the linear BDP model does not fit the frequency spectrum for small families and the BDIP model predicts a geometric tail (and hence underestimates the number of large families). Accounting for rate heterogeneity across families again appears to accommodate latter issue, and reveals a residual lack of fit presumably due to rate heterogeneity across lineages. In particular, the predicted size distribution for the chimpanzee genome does not correspond well to the observed distribution, which is presumably due to our filtering strategy based on the presence of a homolog in the human genome. □

What these examples seem to imply is that, when dealing with 'core' gene families which do not go extinct, the 'Lynch & Conery'-like linear BDIP sometimes fits the data quite well (as for instance in the rice example); but that the more commonly employed linear BDP model fits the data poorly in virtually all cases. The discrepancies are dependent on the genomes and their phylogeny under consideration, as highlighted by the much greater difference in estimated loss rates for the *Drosophila*, yeast and primates data compared to the rice data, which may in part be due to the different timescales considered. The widely diverging estimates and overall lack of fit render the interpretations of the parameter estimates as rates of genome evolution vacuous.

A plausible explanation for the observed discrepancies is that gene family *extinction* is subject to different rates than gene loss in a multi-copy family. If the loss rate in a single-copy state is (much) lower than the loss rate in a multi-gene family, the μ parameter in the linear BDP will be pulled towards a value which may be unrepresentative for either rate. The lack of fit of the linear BDIP model for long timescales however suggests that the difference in rates between the single-copy state and multi-copy state may not provide the full picture. Biologically, of course, this is to be expected, since many duplicated genes, although stably established in the genome, exhibit some functional redundancy, so that the loss rate per gene should be different depending on the *distribution* of gene function over members of the family, and not directly on the number of genes in the family. We return to these problems in much more

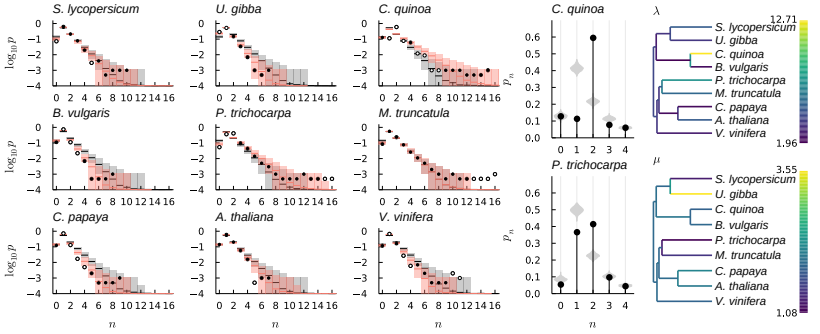


Figure 3.8: Bayesian inference of linear BDP models of gene family evolution for a nine-taxon dicot data set. The left (3×3) panel shows posterior predictive gene family size distributions for each genome for a constant rates phylogenetic linear BDP model in gray and a phylogenetic linear BDP with branch-specific rates (using an uncorrelated relaxed clock prior) in red. The middle pair of plots show a detail of the posterior predictive family size distribution (violin plots) and conserved size distribution (black dots) for small families for quinoa and poplar. The pair of trees show the phylogeny with branches colored according to the marginal posterior mean duplication rate λ and loss rate μ inferred with the branch-specific rates model. The color scale goes from blue over green to yellow.

detail later (see sec. 3.3) when we develop models of gene family evolution to account for this.

In our final example, we consider rate variation across lineages and the impact of ancient WGD events by studying a data set from dicotyledonous plants.

Example (dicots). We investigate a data set of 1000 gene families associated with a nine-taxon dicot phylogeny. The 1000 gene families were randomly sampled without replacement from a larger data set, filtered so that there is at least one observed gene in each clade stemming from the root. We estimate a prior distribution for the ancestral gene family size by fitting a beta-geometric distribution to the paranomes, and based on that choose $\eta = 0.92$ and $\zeta = 10$ for further analyses. For our phylogenetic BDP models we use a dated phylogenetic tree from the TimeTree database. Inspection of posterior predictive gene family size distributions under the constant rates linear BDP model show that there is likely considerable variation in duplication and loss rates across branches of the species tree (fig. 3.8). To accommodate this, we adopt an uncorrelated relaxed log-normal duplication/loss rate clock (DL clock) model

for rate variation across the phylogeny. Specifically we assume that

$$\begin{aligned} r_1 &\sim \mathcal{N}(\log(3), 2) \\ r_2 &\sim \mathcal{N}(\log(3), 2) \\ \log \lambda_i | r_1 &\sim \mathcal{N}(r_1, \tau) && i = 1, \ldots, 2n-2 \\ \log \mu_i | r_2 &\sim \mathcal{N}(r_2, \tau) && i = 1, \ldots, 2n-2 \end{aligned}$$

where the prior choice for r_1 and r_2 (the mean log rates) was based on the duplication and loss rate estimates for the constant rates model and where n is the number of taxa. Assigning a prior for τ (the variance of the clock model) tends to lead to very large τ estimates and numerical issues, likely due to model violations, so we constrain rate variation across branches by setting $\tau = 0.1$. Results for this analysis are shown in fig. 3.8 and we tabulate the rate estimates for the tip lineages in tbl. 3.4.

Table 3.4: Duplication and loss rate estimates with 95% uncertainty intervals (UI) under the uncorrelated relaxed clock model for the tip lineages of the nine dicot data set (see fig. 3.8). Rates are on a scale of expected number of events per gene per billion years (Gy).

Species	λ	95% UI	μ	95% UI
V. vinifera	2.3	(2.0, 2.5)	1.6	(1.4, 1.8)
A. thaliana	3.4	(3.0, 3.8)	1.8	(1.5, 2.0)
C. papaya	2.1	(1.8, 2.3)	1.9	(1.6, 2.2)
M. truncatula	3.7	(3.3, 4.0)	1.3	(1.1, 1.5)
P. trichocarpa	5.0	(4.7, 5.4)	1.1	(0.9, 1.2)
B. vulgaris	2.1	(1.8, 2.5)	1.6	(1.3, 2.0)
C. quinoa	12.7	(11.8, 13.6)	1.6	(1.3, 1.9)
U. gibba	2.7	(2.3, 3.0)	3.6	(3.1, 4.0)
S. lycopersicum	3.4	(3.0, 3.7)	1.2	(1.0, 1.4)

Clearly, rate variation across branches can account for some of the variation in the data, but the fit of the linear BDP for small families remains quite poor for a number of species (e.g. *Beta vulgaris*, *Carica papaya*, *Vitis vinifera*, *Populus trichocarpa* and *C. quinoa*). Several aspects of these results should be highlighted. Firstly, *Utricularia gibba* (bladderwort) is a representative of a lineage characterized by massive genome size reduction and high gene family turnover rates (Ibarra-Laclette et al. 2013; Carretero-Paulet et al. 2015), and we find, as expected, strongly increased gene loss rates in this lineage relative to other branches in S. Secondly, for poplar and quinoa, two species characterized by a very well-preserved WGD signature in their genomes, we obtain markedly increased duplication rate estimates. Despite the high duplication

rates, in all likelihood caused by the recent WGDs affecting these genomes, we observe a very strong lack of fit, with a much larger proportion of gene families of size 2 than expected under the posterior (fig. 3.8). We may conclude that, as expected, the linear BDP model cannot accomodate the effects of ancient WGDs. □

3.2 Modeling and inference of whole-genome duplications

> Complicated phenomena, in which several causes concurring, opposing, or quite independent of each other, operate at once, so as to produce a compound effect, may be simplified by subducting the effect of all the known causes, as well as the nature of the case permits, either by deductive reasoning or by appeal to experience, and thus leaving, as it were, a *residual phenomenon* to be explained.[10]
>
> John Herschel (1831)

As we showed in the last example of the previous section, ancient polyploidy, when unaccounted for, presents an obvious source of model violation for phylogenetic BDPs as models of gene content evolution. Not only is it important to account for ancient WGDs[11] if we aim to quantify gene duplication and loss, but also, more positively, the signatures left by ancient WGDs in comparative genomic data sets can be harnessed to infer ancient WGDs in a phylogenetic context by statistical means – a challenging problem of considerable interest in the study of genome evolution, especially in plants.

Rabier, Ta, and Ané (2014) were the first to consider the statistical inference of WGDs using a phylogenetic model of gene family evolution which accounts for ancient WGDs, and they applied their model to gene count data and gene trees for the Butler et al. (2009) yeast data set. The same method was used and evaluated for a land plant data set by Tiley, Ané, and Burleigh (2016). The model of Rabier, Ta, and Ané (2014), but with different methods, has been used by ourselves for modeling gene family evolution and statistical inference of WGDs from gene trees (Zwaenepoel and Van de Peer 2019a, see

[10] Also quoted in Stigler (2016), where the author takes the idea of *residual* (in a broad sense) to constitute the 'seventh pillar of statistical wisdom'.

[11] Somewhat inaccurately, we shall use 'WGD' in a colloquial sense throughout, referring to both whole-genome duplications proper and multiplications of a higher level. One could use 'whole-genome multiplication', however we have been unable to adopt the habit to do so. Furthermore, it is not so clear whether it is actually helpful to talk of multiplications of genomes. As there are no molecular mechanisms to, for instance, triplicate a genome, it seems safe to say that whatever event we talk about will be the result of one or more duplication or hybridization events.

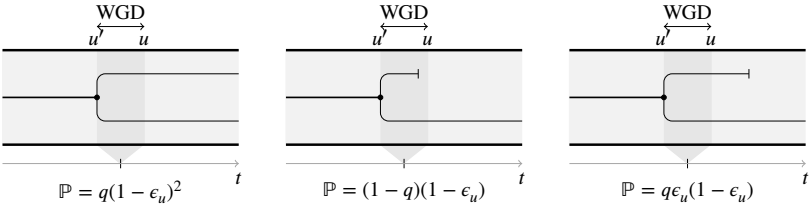


Figure 3.9: Diagram to illustrate the DLWGD model. Three different scenario's for the evolution of a single lineage within a species tree branch with a WGD node u are shown. Time t goes from left to right. The time interval in which the WGD occurs and rediploidization process completes is indicated in dark gray. The probability (under the model) that, conditional on a single lineage reaching WGD node u, the outcome at the end of the branch is observed is shown below each scenario, where ϵ_u is the extinction probability for a lineage right after u. Note that in the third scenario, a WGD-derived duplicate is lost due to the SSDL process, and not due to rediploidization.

chapter 6) and gene counts (Zwaenepoel and Van de Peer 2020). In this section we provide an overview of the model of Rabier, Ta, and Ané (2014), hereafter referred to as the DLWGD (duplication, loss and WGD) model, discuss the issues associated with statistical inference of WGDs using phylogenetic BDPs and outline our Bayesian approach for model-based detection of WGDs in a phylogenetic context, published in Zwaenepoel and Van de Peer (2020).

3.2.1 The DLWGD model of Rabier *et al.*

To bring WGDs in the phylogenetic BDP model, Rabier, Ta, and Ané (2014) propose a simple model for a WGD event. They insert for each hypothetical WGD an additional node in the species tree, which we shall refer to as a 'WGD node'. Each WGD node u is a vertex of S at distance t_u from its parent node $\varrho(u)$ and has outdegree 1. The DLWGD model assumes that a collection of gene families evolves independently according to a linear BDP process along the branches of S, but that at a WGD node, each extant lineage in each family duplicates instantaneously (fig. 3.9). Upon duplication at the WGD node, the resulting pair of daughter lineages is either retained in duplicate with probability q, or returns to a single lineage with probability $(1-q)$; and this happens independently for all lineages which pass through the WGD node. The latter (retention) process is supposed to model the rediploidization process[12], parameterized by a single parameter q, the retention probability (or

[12]We here use the term *rediploidization* in a somewhat loose sense, as in e.g. Wolfe (2001), as "[…] the evolutionary process by which a polyploid genome turns into a diploid one".

retention rate). Note that this model of a polyploidization/rediploidization cycle assumes that rediploidization occurs in a negligible time interval relative to the typical branch length of S. The model is easily extended to allow for higher level multiplications, although in that case an additional design choice needs to be addressed, namely whether each copy should be independently retained with probability q, or whether for a k-fold multiplication we would rather introduce $k-1$ retention parameters determining the probability of retaining $0, \dots, k$ copies.

The model as stated and illustrated in fig. 3.9 describes the evolution of a single lineage when passing a WGD node. In the probabilistic graphical model which defines the phylogenetic BDP, this model then takes the form

$$X_u | X_{u'} = n \sim n + \text{Binomial}(n, q)$$

where X_u is the number of genes extant after rediploidization and $X_{u'}$ is the number of genes extant *just before* the WGD associated with vertex u (we formally think of u' as an additional node in the tree at a distance 0 from u towards the root, fig. 3.9). This leads to a transition probability across the WGD node of the Binomial form

$$p^{(u)}_{n,n+k} = \begin{cases} \binom{n}{k} q^k (1-q)^{n-k} & \text{if } 0 \leq k \leq n \\ 0 & \text{else} \end{cases}$$

The transition probabilities for the process conditional on survival can be obtained using the following recursive relations

$$\begin{aligned} q^{(u)}_{i,j} &= \tilde{p}^{(u)}_{1,1} q^{(u)}_{i-1,j-1} + \tilde{p}^{(u)}_{1,2} q^{(u)}_{i-1,j-2} \\ q^{(u)}_{1,1} &= \tilde{p}^{(u)}_{1,1} = (1-q)(1-\epsilon_u) + 2q\epsilon_u(1-\epsilon_u) \\ q^{(u)}_{1,2} &= \tilde{p}^{(u)}_{1,2} = q(1-\epsilon_u)^2 \end{aligned}$$

Where the $q^{(u)}_{i,j}$ refer to transition probabilities conditional on survival (see eq. 3.4 along the branch leading to node u, whereas q (without subscript) refers to the retention probability. The change in model structure affected by introducing a WGD node of course affects the extinction probabilities as well. Note that the pgf for the number of descendant genes after rediploidization of a sin-

Rediploidization is sometimes understood alternatively as the re-establishment of disomic inheritance after autopolyploidy, a process which is of course related to rediploidization in the former sense.

gle lineage going through a WGD is

$$f_u(s) = \sum_{k=0}^{\infty} \mathbb{P}\{X_{u'} = k | X_u = 1\} s^k = (1-q)s + qs^2$$

So that we can still compute extinction probabilities using the pgfs along the phylogeny as outlined above. In particular, if v is the child node of the WGD node, the probability that a lineage extant right before the WGD node (i.e. extant at u') leaves observed descendants down the tree is $f_{S_{u'}}(\mathbf{0}) = f_u(f_{S_v}(\mathbf{0}))$. With these probabilities available, we can apply the algorithm of Csűrös and Miklós (2009) to compute the conditional survival likelihood.

3.2.2 Statistical inference of WGDs from gene count data

As should already be clear by now, statistical inference of WGDs using the DLWGD or related models boils down to a model selection problem, where we ask questions of the sort "Does a model with a WGD on branch $\langle v, w \rangle$ fit the data better than a model without?". Rabier, Ta, and Ané (2014) addressed the statistical issue using maximum likelihood inference and a likelihood ratio test (LRT) to compare models. To test whether or not a particular gene count matrix y provides evidence for an ancient WGD along some branch of the species tree, they assumed constant duplication and loss rates across the phylogeny, and consider two models: $\mathcal{M}_0$ without the WGD node of interest and $\mathcal{M}_1$ with the WGD node. They maximize the log-likelihood for both models with respect to the parameters

$$\ell_0 = \max_{\lambda,\mu} p_{\mathcal{M}_0}(y|\lambda,\mu)$$
$$\ell_1 = \max_{\lambda,\mu,q} p_{\mathcal{M}_1}(y|\lambda,\mu,q)$$

and compute the LRT test statistic $\Lambda = 2(\ell_1 - \ell_0)$. The authors compare this test statistic against the relevant asymptotic distribution (which is a mixture of a Dirac mass δ_0 and a χ^2_1 distribution, due to the null hypothesis being associated with q lying on the boundary of the parameter space). The approach generalizes readily to more complicated situations with multiple hypothetical WGDs marked along S. Note that we have assumed a WGD to be associated with a fixed time point along S, but of course, this is just another parameter which could be taken up in the estimation problem.

As we showed in Zwaenepoel and Van de Peer (2019a), there are several issues

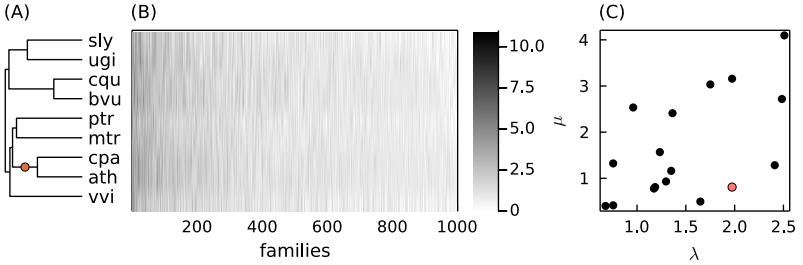


Figure 3.10: Example simulation showing the issue of testing the presence of WGDs with the method of Rabier, Ta, and Ané (2014) in the presence of deviations from a strict DL clock. (A) Species tree used for the simulation, with the test WGD marked by the red dot. (B) Simulated count matrix. (C) Scatter plot of duplication and loss rates, the dot marked in red marks the rates for the branch on which the hypothetical WGD is located.

with this approach, the most prominent one being the assumption of constant rates, or a strict duplication/loss (DL) clock, across the phylogeny. Clearly, as soon as there is variation in duplication and loss rates across the species tree – which should be the rule rather than the exception – this overly stringent assumption may cause q to actually capture an increased duplication rate or decreased loss rate along the branch of interest relative to the other branches in the tree. We illustrate this problem with a simulated example.

Example (LRT test for WGD in the presence of rate heterogeneity). We simulated a data set of 1000 gene families for the nine dicot species tree (see previous example) with branch-specific duplication and loss rates. We simulate branch-specific log-scale duplication rates λ from a $\mathcal{N}(\log(1.5), 0.5)$ distribution and obtain correlated loss rates by setting $\mu_i = \lambda_i + Z$ where $Z \sim \mathcal{N}(0, 0.5)$. We set the duplication rate for the branch leading to the subtree (ath, cpa) equal to the 75% percentile of the λ vector, and the loss rate to the 25% percentile of the μ vector. This creates a branch which has a somewhat higher duplication rate and a somewhat lower loss rate than most branches in S, but not in a particularly extreme way (see fig. 3.10 C). We then apply the LRT statistic of Rabier, Ta, and Ané (2014), assuming a model with constant duplication and loss rates across the phylogeny, to test whether a WGD has occurred along the branch leading to (ath, cpa). We find strong support for the WGD model, with the LRT for the hypothesis $q > 0$ equal to $\hat{\Lambda} = 29.2$, amounting to rejection at the 0.001 level. The associated ML rate estimates are $\hat{\lambda} = 1.3, \hat{\mu} = 1.7, \hat{q} = 0.06$. □

While this is of course but a single example and may seem somewhat contrived, the problem is clear enough, and we refer the reader for more extensive simulations to our work in Zwaenepoel and Van de Peer (2019a). Of course, a retention probability of 0.06 may not seem very relevant biologically, but note that for a well-supported WGD like the one in *Saccharomyces*, Rabier, Ta, and Ané (2014) do in fact report $\hat{q} = 0.07$. Clearly then, rate heterogeneity presents a challenge for this approach. If duplication and loss rates show a sufficiently strong correlation with substitution rates, which would be expected if both are largely driven by genetic drift, one could mitigate this issue by using a phylogenetic tree with molecular branch lengths instead of a timetree, and estimate duplication and loss rates on a molecular distance time scale (this strategy was employed by e.g. Tasdighian et al. (2017)). Alternatively, one could relax the assumption of constant λ and μ using independent branch rates, fixed rate classes, or more flexible local clock models (Yoder and Yang 2000), but this is a rather arbitrary approach which can be very sensitive to taxon sampling, and often leads to numerical issues (Zwaenepoel and Van de Peer 2019a). Furthermore, as the principle underlying the statistical inference of WGDs is signaling a deviation from a phylogenetic BDP model of the SSDL process, assuming models where rates can vary arbitrarily across branches can reduce the power to detect ancient WGDs.

A Bayesian approach proves more helpful. Assuming either a time-calibrated or molecular species tree S, we can assume relaxed clock models similar to those applied in molecular divergence time estimation for the evolution of the gene duplication and loss rate across S, as we first applied in the context of gene tree reconciliation in Zwaenepoel and Van de Peer (2019a) (see chapter 6). Such an approach also allows investigating critically the impact of assuming more or less rate variation across S on the inference of WGDs. One could then base inference of WGDs on an inspection of the retention probability posterior for the WGD of interest, or alternatively, mimicking the LRT approach, compute a Bayes factor to test the hypothesis of $q = 0$.

While we do not actually recommend the latter approach, as it takes the model too seriously (i.e. it only makes sense if the linear BDP is in fact a fully adequate model of the 'background' SSDL process, see also our discussion on model selection in chapter 1), we did use it in Zwaenepoel and Van de Peer (2019a) and briefly outline the approach here. Let $p(\theta, q)$ be the prior density for the retention probability q and all other parameters θ under the DLWGD model, and let $p_0(\theta)$ be the prior density for θ under the model without WGD. We will typically have $p(\theta|q = 0) = p_0(\theta)$, in which case a result from Dickey (1971) and Verdinelli and Wasserman (1995) shows that for observed data y

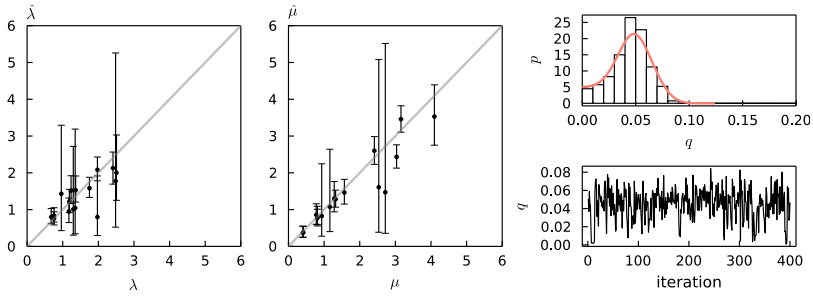


Figure 3.11: Posterior inference for the DLWGD model analyzed in fig. 3.10 but with an uncorrelated relaxed DL clock model. The left two plots show marginal posterior mean branch-specific duplication and loss rate estimates with 95% uncertainty intervals plotted against the true rates. On the right hand side a histogram and kernel density estimate for the retention probability posterior q (using a Gaussian kernel and bandwidth of $h = 0.01$, and using reflection at zero to correct for the boundary effect) are shown together with a trace plot of q during the sampling algorithm.

the Bayes factor in favor of the no-WGD model can be computed using the Savage-Dickey density ratio

$$K = \frac{p(q = 0|y)}{p(q = 0)} = \frac{\int p(q = 0, \theta|y)d\theta}{\int p(q = 0, \theta)d\theta}$$

For a uniform distribution on q, K is simply the marginal posterior density of q at 0. For a sample from the posterior, we can approximate $p(q = 0|y)$ by a kernel density estimate (KDE, with boundary correction) and provide an estimate for K. Values of $\log_{10} K < -1$ could then be considered as providing evidence for $q > 0$. We illustrate the Bayesian approach towards analyzing WGDs using gene count data and phylogenetic BDPs with relaxed clock models using the same simulated data set as above and the dicots data set of sec. 3.1.5.

Example (WGD and rate heterogeneity revisited). We revisit the last simulation example and analyze the same data set but now account for rate heterogeneity across S, assuming an uncorrelated log-normal relaxed DL clock. Specifically, we use the following Bayesian hierarchical model

$$\begin{aligned} r &\sim \mathcal{N}(\log(1.5), 1) \\ \tau &\sim \text{Exponential}(1) \\ \log \lambda_i | r, \tau &\sim \mathcal{N}(r, \tau) \qquad\qquad i = 1, \ldots, 2n-2 \end{aligned}$$

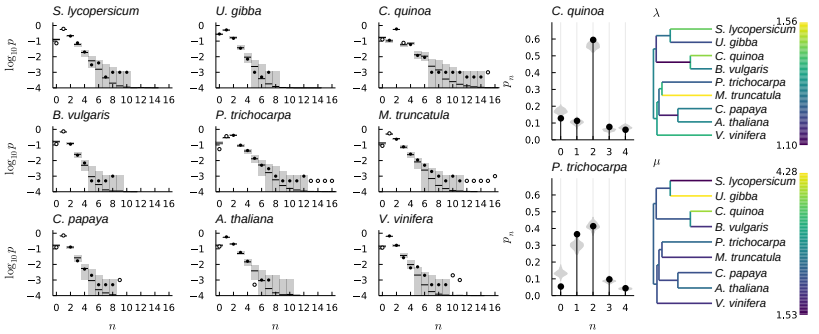


Figure 3.12: Posterior inference for the nine dicot data set using a DLWGD model with branch-specific duplication and loss rates. Compare with fig. 3.8.

$$\begin{aligned} \log \mu_i | r, \tau &\sim \mathcal{N}(r, \tau) && i = 1, \dots, 2n-2 \\ q &\sim \text{Beta}(1, 1) \\ y | \lambda, \mu, q &\sim \text{PhyloBDP}(S, \lambda, \mu, q) \end{aligned}$$

We obtain a sample from the posterior using the NUTS algorithm with Turing.jl. A graphical display of the posterior distribution is provided in fig. 3.11. Most branch-specific duplication and loss rates are accurately estimated, although some estimates have a large variance. The marginal posterior distribution for the retention probability q is clearly compatible with the hypothesis $q = 0$. The Savage-Dickey estimate of the Bayes factor is $\log_{10} K = 0.7$, suggesting the data favors the no-WGD model. □

Example (dicots revisited). We revisit the dicot example of sec. 3.1.5, now accounting for WGDs. Many of the tip branches of this phylogenetic tree are known to be associated with polyploid episodes (see e.g. Van de Peer, Mizrachi, and Marchal 2017). We construct a DLWGD model with a single WGD located at the midpoint of each tip branch of S and conduct inference using the following hierarchical model

$$\begin{aligned} r_1, r_2 &\sim_{\text{iid}} \mathcal{N}(\log 1.5, 2) \\ \log \lambda_i | r_1 &\sim \mathcal{N}(r_1, \tau) && i = 1, \dots, 2n-2 \\ \log \mu_i | r_2 &\sim \mathcal{N}(r_2, \tau) && i = 1, \dots, 2n-2 \\ q_j &\sim \text{Beta}(1, 1) && j = 1, \dots, 9 \\ y | \lambda, \mu, q &\sim \text{PhyloBDP}(S, \lambda, \mu, q) \end{aligned}$$

Where we again set $\tau = 0.1$. A graphical display of the posterior distribution is shown in fig. 3.12 and parameter estimates are tabulated in tbl. 3.5.

Clearly, the model provides a much better fit to the data, and much of the variation in duplication and loss rates in our WGD-unaware analysis is now explained by WGD, with the inferred variation in duplication rates now very limited. We note, in passing, that the long-term duplication rate estimates appear to be almost 10-fold lower than our crude estimates for *Vitis* from chapter 2. We find no evidence for WGD along the *Vitis*, *Beta* and *Carica* tip branches, which is completely in line with our expectations, as no genomic analyses have reported large-scale duplication events for these lineages. Remarkable, despite the massive gene loss and genome compaction in *Utricularia*, we do find a retention probability markedly different from zero. *Utricularia*, despite these notable features that may lead one to expect the absence of WGD, is also known to have undergone multiple WGD events in its relatively recent evolutionary history (Ibarra-Laclette et al. 2013). Our analysis suggest that this can still be detected statistically from gene counts despite massive gene loss (but note that we did not investigate multiple WGDs along this branch). We find retention probabilities significantly > 0 for *Arabidopsis*, *Medicago*, *Solanum*, *Populus* and *Chenopodium*, all in line with expectations from the evolutionary genomic literature (although, as in the case of *Utricularia*, some may in fact reflect multiple events). We note that although the model provides a much better fit to observed family size distributions for lineages with very strong WGD signatures, like quinoa and poplar, posterior predictive simulations show that the DLWGD model, as used here, can still not perfectly account for the observed data in these lineages, with for instance the number of single-copy and extinct families in poplar not predicted accurately. This may be due to the fixed timing of the WGD in our analysis, as well as due to fundamental limitations of the linear BDP as a model of gene family evolution. □

3.2.3 Studying retention patterns using the DLWGD model

A topic which has received a lot of attention in the literature on ancient WGD is the differential retention of WGD-derived duplicates across gene families (Blanc and Wolfe 2004a; Maere et al. 2005; De Smet et al. 2013; Li et al. 2015; Tasdighian et al. 2017). In a pioneering study, Maere et al. (2005) devised a birth-death like demographic model for the whole-paranome age distribution of duplicated genes (see chapter 2) and used this model to study the differential retention of different types of gene duplicates across putative functional classes of gene families in *A. thaliana*. They find a conspicuous as-

Table 3.5: Marginal posterior mean parameter estimates and 95% uncertainty intervals for the tip branches of the dicot phylogeny under the DLWGD model with an uncorrelated relaxed DL clock. Duplication and loss rates are in events per gene per Gy. The last column shows the $\log_{10}$ Bayes factor in favor of the no-WGD model, based on a KDE of the Savage-Dickey density ratio. See also fig. 3.12. Compare with tbl. 3.4.

Species	λ	95% UI	μ	95% UI	q	95% UI	$\log_{10} K$
V. vinifera	1.4	(1.1, 1.7)	1.8	(1.6, 2.0)	0.07	(0.02, 0.12)	-0.3
A. thaliana	1.3	(1.1, 1.6)	2.2	(1.9, 2.6)	0.20	(0.16, 0.24)	<-3
C. papaya	1.3	(1.1, 1.6)	2.0	(1.7, 2.3)	0.02	(0.00, 0.04)	1.5
M. truncatula	1.6	(1.3, 1.9)	1.8	(1.5, 2.0)	0.30	(0.24, 0.34)	<-3
P. trichocarpa	1.2	(1.0, 1.4)	2.2	(1.7, 2.7)	0.69	(0.61, 0.77)	<-3
B. vulgaris	1.3	(1.1, 1.5)	1.8	(1.5, 2.1)	0.01	(0.00, 0.03)	1.5
C. quinoa	1.5	(1.3, 1.7)	3.7	(3.1, 4.3)	0.98	(0.96, 1.00)	<-3
U. gibba	1.2	(1.0, 1.4)	4.3	(3.8, 4.9)	0.20	(0.15, 0.26)	<-3
S. lycopersicum	1.4	(1.2, 1.7)	1.5	(1.3, 1.8)	0.22	(0.17, 0.26)	<-3

sociation of certain functional classes of genes with distinct modes of duplication, suggesting that genes involved in, for instance, developmental processes and transcriptional regulation are more likely to be retained after WGD, but less likely to be retained after SSD, in comparison with the whole genome, a pattern that has been termed 'reciprocal retention' (Freeling and Thomas 2006; Freeling 2009; Tasdighian et al. 2017). Findings in accord with theirs have been reported in many later plant genome papers. Reciprocal retention of protein-coding genes is generally thought to be associated with dosage-balance constraints on the evolution of gene copy number, which arise when a distortion of the stoichiometric balance of certain pathways or protein complexes has a (strong) negative fitness effect (Birchler and Veitia 2012).

In this section, we consider two case studies, employing our Bayesian inference tools to study these types of patterns. In the first, we investigate whether Gene Ontology (GO) functional categories can serve as predictors for the retention probabilities of a gene family by using a regression approach. In the second case study, we consider a mixture model to study reciprocal retention patterns, in a somewhat similar vein as Tasdighian et al. (2017). Both analyses presented are highly tentative, and serve mainly to illustrate the flexibility of the Bayesian approach.

Example (regression on GO categories). We study a five-taxon subset of the dicots data set studied above, consisting of *Vitis*, *Arabidopsis*, *Carica*, *Populus* and *Medicago*. We obtained the relevant subset of gene families from the dicots data set and annotated the gene families with GO terms from the Plant GO Slim subset, using only the 'biological process' and 'molecular function'

ontologies. To do so, we use the GO annotations for *A. thaliana* available in PLAZA 4.5 (Van Bel et al. 2018), and assign a GO term to a gene family when at least 50% of the *Arabidopsis* genes in the family have that GO annotation. We further discard those GO terms that were assigned to fewer than 100 families of the total set. The final data set consists of 5824 gene families with 34 GO Slim terms. We record this data in a 5824×34 binary matrix.

Let x_i be the ith row of X (i.e. the 1×34 row vector of GO annotations for family i). We consider the following regression model

$$\begin{aligned}
\lambda &\sim \text{Exponential}(1.6) \\
\mu &\sim \text{Exponential}(2.0) \\
\bar{q}_1, \bar{q}_2, \bar{q}_3 &\sim_{\text{iid}} \text{Beta}(1, 1) \\
\beta &\sim \text{MVN}(\mathbf{0}, I_{34}) \\
q_{ij} &= \text{logistic}\big(\text{logit}(\bar{q}_j) + x_i\beta\big) \\
y_i|\lambda, \mu, q_i &\sim \text{PhyloBDP}\big(\lambda, \mu, (q_{i,1}, q_{i,2}, q_{i,3})\big)
\end{aligned}$$

where $q_{i,1}, q_{i,2}$ and $q_{i,3}$ are the retention probabilities in family i for the *Arabidopsis*, *Medicago* and *Populus* WGD nodes respectively, and where β is a 34-dimensional column vector of regression coefficients. Note that $\text{logit}(x) = \log(x/(1-x))$ and that $\text{logistic}(x) = \text{logit}^{-1}(x)$. We obtain a sample from the posterior using the NUTS algorithm and conduct posterior predictive simulations for the family size distributions as in the above examples (fig. 3.13).

Taken at face value, our results confirm at least in part those of Maere et al. (2005) and the many GO enrichment analyses reported in the plant genomics literature, with GO terms such as *DNA-binding transcription factor* and *signal transduction* having a positive effect on the WGD retention probability parameter, and terms like *photosynthesis*, *DNA metabolic process* and *cell cycle* among those with negative effect. Other results are less congruent, with developmental terms near the bottom of the ranking based on the regression coefficients, and certain metabolic processes near the top. An examination of model fit indicates that the regression model does not fit the genome-wide observed size distributions better than a simple DLWGD model with constant duplication and loss rates across the tree and a shared WGD retention probability across all families. Looking at the size distributions within functional categories, we see that regression model sometimes fits the data better, sometimes worse, and often equally well as the simpler model, with different patterns for different species and functional categories (fig. 3.13). Note that a lack of fit for *A. thaliana* is expected, as due to our annotation strategy, we filtered

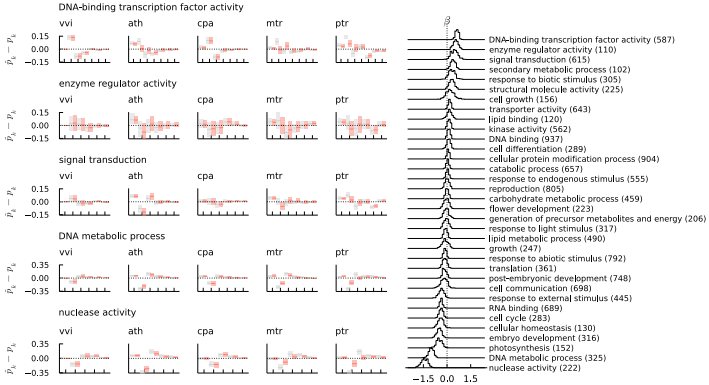


Figure 3.13: GO Slim regression analysis for the five-taxon dicot set using the DLWGD model. The left plots show the average deviation of the posterior predictive family size distribution from the observed family size distribution (with 95% posterior predictive intervals) for a selection of GO categories and families with a size $0 \leq k \leq 6$ (along the x-axes). Samples from the posterior distributions for the regression coefficients for each of the GO Slim categories are shown on the right, ordered by their respective marginal posterior mean. The number in brackets indicates the number of families with the relevant GO annotation.

out all families without *A. thaliana* representative from the data, but did not condition the likelihood on this filtering step in the presented analysis. □

While suggestive, it seems premature to make any strong biological conclusion from the regression analysis in the above example. In particular, the analysis is strongly dependent on the accuracy of the functional annotation and the aptness of the GO Slim set as an evolutionarily relevant partitioning of biological functions. Additionally, the analysis above did not account for heterogeneity in DL rates across families, so that, for instance, GO terms which tend to be associated with high duplication rates may end up with a positive regression coefficient, although they need not be especially strongly retained after WGD. Most importantly, even if the annotation were sound, the assumption that different functional annotations contribute linearly to the retention rate on a logit scale is of course very *ad hoc*. Clearly, it is one thing to look for enrichments of GO terms in certain gene sets, but quite another to use GO terms as linear predictors in a regression analysis. What the example does show clearly is that substantive hypotheses of the form 'property P has an effect on the retention probabilities (or DL rates)' can be assessed rather straightforwardly using Bayesian hierarchical models in a probabilistic programming environment, if one succeeds to clearly formulate such hypotheses.

Example (Reciprocal retention mixture model). In Tasdighian et al. (2017), the authors used a model similar to the DLWGD model outlined above to study reciprocal retention patterns in a phylogenetic context. Specifically, they assumed a critical linear BDP (i.e. $\lambda = \mu$), and a fixed retention rate of $q = 1$ for each WGD marked along the phylogeny. Using a species tree for 37 angiosperms with branch lengths on a molecular distance scale, they fitted the model by ML independently across families, assuming a single λ for each family. The rationale of the latter study was that the class of gene families which tend to be retained in duplicate after WGD but not SSD should be characterized by a small value of λ relative to the whole genome.

We consider the same five-taxon data set as in the previous example, but now seek to study the reciprocal retention patterns that are the focus of Tasdighian et al. (2017). To do so, we consider a rather crude model following the basic rationale of the latter authors, assuming there are three classes of gene families: (1) families with large retention probabilities and small DL rates, (2) families which show no particularly 'preferred' mode of duplication and (3) families with low retention probabilities and high DL rates. We note that the last class is no essential component of the reciprocal retention theory. We translate this into the following mixture model

$$
\begin{aligned}
\log \lambda_0 &\sim \mathcal{N}(\log 1.6, 1) \\
\log \mu_0 &\sim \mathcal{N}(\log 1.6, 1) \\
\bar{q}_1, \bar{q}_2, \bar{q}_3 &\sim_{\text{iid}} \text{Beta}(1, 1) \\
\alpha_1, \alpha_2 &\sim \text{Exponential}(1) && a = (-\alpha_1, 0, \alpha_2) \\
\beta_1, \beta_2 &\sim \text{Exponential}(1) && b = (\beta_1, 0, -\beta_2) \\
w &\sim \text{Dirichlet}(1, 1, 1) \\
z_i | w &\sim \text{Categorical}(w) \\
\lambda_j &= \exp(\log \lambda_0 + a_j) \\
\mu_j &= \exp(\log \mu_0 + a_j) \\
q_{jk} &= \text{logistic}(\text{logit}(\bar{q}_k) + b_j) \\
y_i | z_i, \lambda, \mu, q &\sim \text{PhyloBDP}(S, \lambda_{z_i}, \mu_{z_i}, (q_{z_i,1}, q_{z_i,2}, q_{z_i,3}))
\end{aligned}
$$

This is a mixture model with three components, where the duplication rates are $\lambda_0 e^{-\alpha_1}$, λ_0 and $\lambda_0 e^{\alpha_2}$ respectively, with $\alpha_j > 0$, so that $\lambda_1 < \lambda_2 < \lambda_3$ (the same holds for the loss rates). The retention probabilities are similarly, but oppositely determined by the β_j on a logit scale, so that $q_{1,k} > q_{2,k} > q_{3,k}$ for WGD node k. To sample from the posterior, we marginalize the like-

lihood over the component indicator z_i, i.e. we compute $p(y_i|\lambda, \mu, q, w) = \sum_j p(y_i|z_i = j, \lambda_j, \mu_j, q_j)w_j$. This ensures that all sampled parameters are real-valued, so that we can use a HMC sampler. After obtaining an MCMC sample, one can easily sample a component indicator for each MCMC iterate to obtain a posterior sample for the latter (see below).

Table 3.6: Posterior mean parameter estimates and 95% uncertainty intervals for the reciprocal retention mixture model.

component	w	λ	μ
1	0.14 (0.11, 0.18)	0.91 (0.85, 0.97)	1.01 (0.94, 1.08)
2	0.75 (0.71, 0.78)	0.92 (0.86, 0.98)	1.02 (0.94, 1.09)
3	0.11 (0.09, 0.13)	7.40 (6.57, 8.33)	8.23 (7.14, 9.36)
	q_1 (*Arabidopsis*)	q_2 (*Medicago*)	q_3 (*Populus*)
1	0.68 (0.59, 0.76)	0.79 (0.72, 0.86)	0.95 (0.93, 0.97)
2	0.10 (0.08, 0.12)	0.16 (0.14, 0.19)	0.50 (0.48, 0.53)
3	0.10 (0.08, 0.12)	0.16 (0.14, 0.19)	0.50 (0.47, 0.53)

In tbl. 3.6 we tabulate the posterior mean and 95% uncertainty intervals for the relevant parameters of the three mixture components, using a sample of 5000 gene families as data set. We find that the the first two components have the same DL rates, whereas the second and third component share the same retention rates. The first component is associated with markedly higher retention probabilities, whereas the last component is associated with very high DL rates. The three components have relative posterior weight of roughly 15%, 75% and 10% respectively. The model fits the observed family size distributions reasonably well (fig. 3.14), with the usual issues for small families (see above). Clearly, this mixture model suggests that there is a relatively large class of gene families for which a model with a (much) higher retention probability fits better, but which are otherwise associated with unexceptional DL rates, and a smaller class of gene families for which a high DL rate model fits the gene content patterns better, but with otherwise unexceptional retention probabilities. It appears therefore that the signal for *reciprocal* retention is rather weak in this data set, with the highly-retained gene families post-WGD not particularly associated with lower DL rates compared to the rest of the genome. On the other hand, a quite pronounced pattern of *differential* retention does appear from this analysis.

To take a closer look at the functional classes associated with the posterior mixture components, we computed the posterior component assignment probabilities for each gene family in the complete data set ($n = 11884$). Specifi-

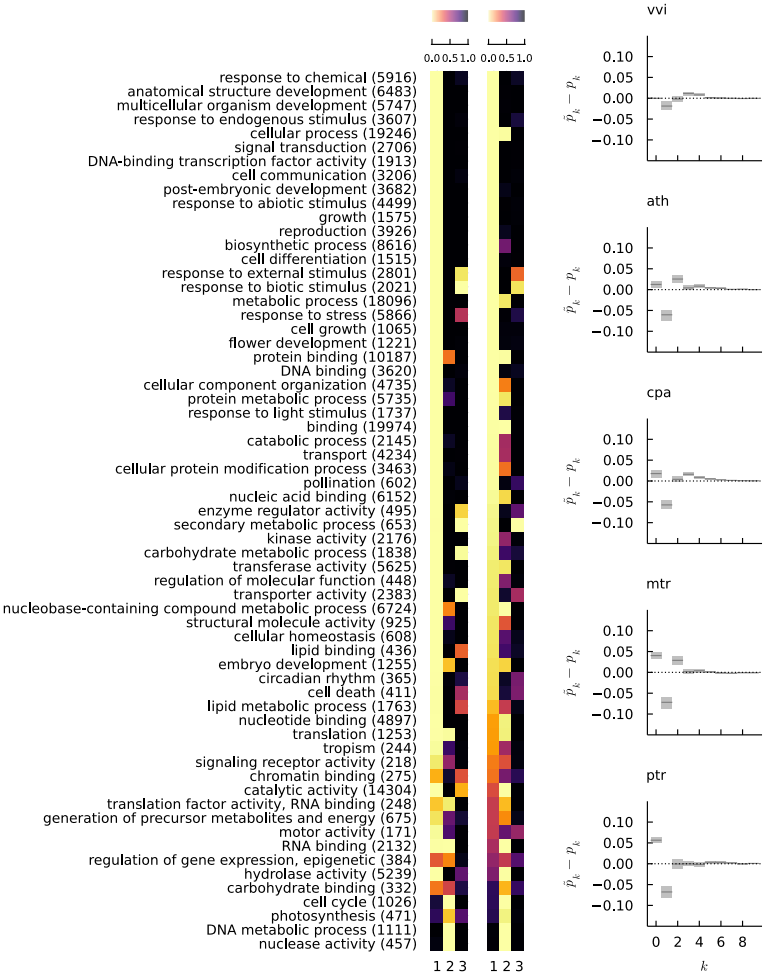


Figure 3.14: (Left) GO term representation in the posterior component assignments for the reciprocal retention mixture model. In the left heatmap we show, for each GO term and for each mixture component, the hypergeometric tail area for observing the posterior expected number of *genes* with that term in that mixture component. In the right heatmap we show for each term the hypergeometric tail area for observing the posterior expected number of *families* with that term in each mixture component, where a term was assigned to a family when at least 10% of the genes in the family were annotated with the term. The total number of genes in the data set annotated for each term is indicated in brackets. (Right) Average deviation of the posterior predictive size distribution from the observed size distribution, with 95% posterior predictive intervals (as in fig. 3.13).

cally, for each family i, we compute

$$p(z_i = j|y) = \int \frac{p(y_i|z_i = j)p(z_i = j|\theta)}{\sum_l p(y_i|z_i = l)p(z_i = l|\theta)} p(\theta|y)d\theta$$
$$\approx \frac{1}{N} \sum_{k=1}^{N} \frac{p(y_i|z_i = j)p(z_i = j|\theta^{(k)})}{\sum_l p(y_i|z_i = l)p(z_i = l|\theta^{(k)})}$$

where $\theta^{(1)}, \ldots, \theta^{(N)}$ is our sample from the posterior (for the 5000 families subset). We then estimated for each mixture component j and for each GO term r in the data set the posterior expected number of genes with the relevant GO term in the mixture component, that is

$$N_{j,r} = \sum_g \mathbb{1}[r \in \mathrm{GO}_g] p(z_{\phi(g)} = j|y_{\phi(g)})$$

where $\phi(g)$ is the gene family of which gene g is a member and GO_g denotes the set of GO terms associated with gene g. We use the $N_{j,r}$ values to compute the usual hypergeometric tail area as a measure for the relative representation of a GO term in a set of gene families. These values are shown in fig. 3.14. We find a large number of overrepresented terms in the highly-retained class (component 1), among which many of the usual suspects, such as *DNA-binding transcription factor*, *kinase activity*, and several development related terms. The second and largest component tends to be associated with the same GO terms we found to have negative regression coefficients in the regression analysis of the previous example, notably *cell cycle*, *photosynthesis* and *DNA metabolic process*. The third component, which involves the smallest number of families, shows fewer clear enrichments, although *secondary metabolic process*, *response to biotic stimulus* and *transporter activity* may be overrepresented in this class, but not more so than in the high-retention class (component 1). Note that some of these were found to have positive regression coefficients in our analysis above, which could indicate that in the regression analysis, the covariate-induced retention rate variation captures variation due to heterogeneity in DL rates, rather than heterogeneity in retention after WGD. □

Clearly, one could devise many types of mixture models along these lines to classify gene families in clusters with similar retention characteristics, and the one presented in the above example is but one of the many possible models that could capture potential reciprocal retention patterns. Indeed, we see again that once the flexibility of Bayesian modeling is properly unlocked, the challenge

lies mainly in devising good models, and less in inference.

3.2.4 Model-based detection of WGDs

An important limitation of the approach outlined above is the requirement of formulating a set of WGD hypotheses in the phylogeny of interest *a priori*, i.e. before conducting the statistical analysis. Ideally, we would like to discover putative WGDs 'automatically' under the DLWGD model, under some prior on the incidence of WGDs along S. Specifically, we seek to conduct joint inference of branch-specific duplication and loss rates, the number of WGDs and the locations and retention probabilities associated with these from a phylogenetic profile matrix y. This challenging problem was tackled in Zwaenepoel and Van de Peer (2020), where we devised an approach based on reversible-jump MCMC (rjMCMC) to sample from the complicated posterior distribution over what we termed *WGD configurations* on S.

Let $\theta = (\log \lambda, \log \mu)$ be the branch-specific duplication and loss parameters for a phylogenetic linear BDP on species tree S, so that the pair $\theta_e = (\log \lambda_e, \log \mu_e)$ denotes the BDP parameter for branch e. We denote the total tree length as T, and assume a mapping $[0, T] \to E(S) \times \mathbb{R}^+$ associating with each point in $[0, T]$ a location along the phylogeny (i.e. a pair consisting of an edge of S and a time point along that edge). We use $\phi = \{(t_1, q_1), \ldots (t_k, q_k)\}$ to denote a WGD configuration for k WGDs along S, where $t_i \in (0, T)$ marks the time point of WGD i and q_i is the associated retention probability. We will use ψ as a shorthand for (ϕ, θ). We will construct a MCMC sampler which samples from

$$p(\psi | y, S) \propto p(y | \theta, \phi, S) p(\theta | S) p(\phi | S) \tag{3.5}$$

To do so, we need two new ingredients: (1) a prior distribution for WGD configurations $p(\phi | S)$ and (2) a MCMC kernel to propose new WGD configurations ϕ' conditional on a current configuration ϕ'.

3.2.4.1 Prior for WGD configurations

We use a simple and intuitive construction for the prior of ϕ. Let $\varpi = |\phi|$ denote the number of WGDs in the configuration, then we have the following prior for WGD configurations:

$$\varpi \sim f_{\varpi}(\cdot)$$

$$t_i|\varpi \sim \text{Uniform}(0, T) \qquad i = 1, \ldots, \varpi$$
$$q_i|\varpi \sim \text{Beta}(\alpha, \beta) \qquad i = 1, \ldots, \varpi$$

where f_ϖ is some distribution on the non-negative integers, such as a geometric or Poisson distribution. Evaluating the prior density for a WGD configuration ϕ according to this model is straightforward.

In addition, in order to compute Bayes factors for WGD hypotheses for specific branches, we will need the prior probability of a WGD occurring on a given branch j. Let ϖ_j denote the random number of WGDs on branch j, the prior probability $\mathbb{P}\{\varpi_j = k\}$ can be obtained from the WGD configuration prior as

$$\begin{aligned}\mathbb{P}\{\varpi_j = k\} &= \sum_{n=k}^{\infty} \mathbb{P}\{\varpi_j = k|\varpi = n\} f_\varpi(n) \\ &= \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{t_j}{T}\right)^k \left(1 - \frac{t_j}{T}\right)^{n-k} f_\varpi(n)\end{aligned}$$

When f_ϖ has finite support this suffices to compute the relevant prior probabilities. For a geometric prior with parameter ξ on the total number of WGDs, this further simplifies to

$$\mathbb{P}\{\varpi_j = k\} = \frac{\xi\left(\frac{t_j}{T}(1-\xi)\right)^k}{\left[1 - \left(1 - \frac{t_j}{T}(1-\xi)\right)\right]^{k-1}}$$

whereas for a Poisson(ξ) prior on ϖ we get $\varpi_j \sim \text{Poisson}(\xi t/T)$.

3.2.4.2 Reversible-jump proposal kernels

More challenging is the task to devise suitable MCMC proposal kernels to sample from the posterior over WGD configurations. The fundamental issue is that different WGD configurations represent different DLWGD models parameterized by a parameter vector of varying dimension (i.e. a DLWGD model for n taxa with $\varpi = k$ WGDs and branch-specific rates has a $2k+2n-2$ dimensional parameter), so that constructing a suitable reversible proposal kernel which preserves the target distribution is not straightforward. We construct a

forward proposal kernel P_f which adds a WGD to the model

$$\psi = (\theta, \{(t_1, q_1), \dots (t_k, q_k)\}) \xrightarrow{P_f} (\theta', \{(t_1, q_1), \dots (t_k, q_k), (t_{k+1}, q_{k+1})\}) = \psi'$$

and a reverse kernel P_r which removes a WGD from the configuration. To ensure these trans-dimensional moves preserve detailed balance, we use the theory of Green (1995). For the forward move, we need a vector of random numbers u from some suitable joint density $g(u)$ such that $(\psi', u') = h(\psi, u)$, where h is a (non-random) invertible function and u' are the numbers necessary for the reverse move (with density g'), so that $(\psi, u) = h^{-1}(\psi', u')$. The acceptance probability for the forward move is $\min\{1, \alpha(\psi, \psi')\}$ where (dropping dependence on S)

$$\alpha(\psi, \psi') = \frac{p(y|\psi')p(\theta')p(\phi')g'(u')p_r(\psi')}{p(y|\psi)p(\theta)p(\phi)g(u)p_f(\psi)} \left| \frac{\partial(\psi', u')}{\partial(\psi, u)} \right|$$

where $p_f(x)$ and $p_r(x)$ are the probabilities of executing a forward, respectively reverse, move when in state x. Note that the last factor is the absolute value of the determinant of the Jacobian matrix for the transformation $(\psi, u) \to (\psi', u')$.

The simplest possible move adds a WGD at a uniformly random point along the phylogeny with a random retention rate q. Let $\varpi = k$, so that $\varpi' = k + 1$, in that case $u = (u_1, u_2) = (t_{k+1}, q_{k+1})$ and $u' = ()$ (the 0-dimensional vector), with $u_1 = t_{k+1} \sim \text{Uniform}(0, T)$ and $q_{k+1} \sim g_2(\cdot)$. We have $(\phi', u') = h(\phi, u) = (\phi \cup \{(t_{k+1}, q_{k+1})\}, ())$. The acceptance probability simplifies to

$$\alpha(\psi, \psi') = \alpha(\phi, \phi') = \Lambda \frac{f_\varpi(k+1)}{f_\varpi(k)} \frac{p(q)}{g_2(q)} \tag{3.6}$$

where Λ is the likelihood ratio. Since we expect a correlation between duplication and loss rates and retention probabilities in the posterior, we may obtain more efficient proposal kernels (i.e. with higher acceptance probabilities) by proposing changes to relevant duplication and loss rates when a WGD is proposed or removed. For instance, when a WGD is introduced on branch j, it may be beneficial to decrease the duplication rate λ_j concomitantly. To do so, still using the formalism of Green (1995), we simply generate $u = (u_1, u_2, u_3) \sim g$ and let $\psi' = (\theta', \phi')$ where $\theta' = (\theta_1, \dots, \theta'_j, \dots, \theta_{2n-2})$ with $\theta'_j = (\log \lambda_j - u_3, \log \mu_j)$, and where ϕ' is as before. The reverse move requires a single random variable u'_3 and we take this variable to have the same

density as u_3, so that the acceptance probability becomes

$$\alpha(\psi, \psi') = \alpha(\phi, \phi') \frac{p(\theta'_j | \theta_{-j})}{p(\theta_j | \theta_{-j})}$$

where $\alpha(\phi, \phi')$ is as in eq. 3.6 and θ_{-j} is the vector of branch parameters excluding branch j. A proposal kernel which also updates μ_j is easily derived using the same approach.

3.2.4.3 Implementation and application

An MCMC algorithm using the reversible jump proposals described above was implemented to sample from the posterior density in eq. 3.5. The statistical performance on simulated and empirical data of this approach was studied extensively in Zwaenepoel and Van de Peer (2020), and we refer the reader to that publication for details. The conclusion of our work there still stands: many challenges for this approach remain, both computational and statistical. Computationally, problems arise not only with regard to resources but also the complexity of the implementation of rjMCMC algorithms. For the method to be more flexible and useful, as well as less bug-prone, it would be desirable to implement the reversible-jump kernels in a way compatible with a probabilistic programming framework, so as to enable standard samplers for within-dimensional moves and more flexible specification of hierarchical models, but this is a considerable effort in itself. While yielding sensible rate estimates and locating many well-known WGDs in studied data sets, the combination of computational intensiveness – which together with the not-so flexible implementation precludes a smooth Bayesian workflow (which would involve extensive experimentation with different models, posterior predictive simulation, *etc.*) – and somewhat underwhelming statistical performance performance on simulated data demands renewed attention to this problem.

Example (dicots revisited). In Zwaenepoel and Van de Peer (2020), we studied the same data set for nine dicot species as in the previous examples. We find that our inferences of the number of WGDs are sensitive to prior assumptions on duplication and loss rate variation across S. Duplication and loss rate estimates are largely similar to those obtained from our analysis for a fixed WGD configuration (sec. 3.2.2). We find strong evidence for the *Arabidopsis*, poplar and quinoa WGD, but find that the results for *Medicago* and tomato are dependent on the assumed DL clock prior (fig. 3.15). Contrary to the analysis above, we now do not find support for WGD in *Utricularia*. It is unclear, at this

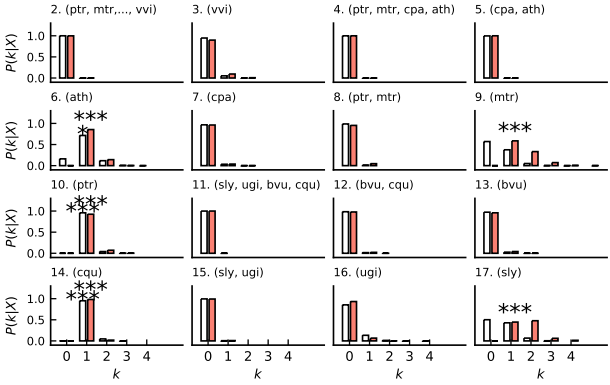


Figure 3.15: Posterior probabilities for the number of WGDs on each branch in the nine dicots phylogeny (figure 3 from Zwaenepoel and Van de Peer (2020)). The branches are identified by their respective clades, using three letter codes (***Genus species***) for the species names. In red and white, two different priors on duplication and loss rate heterogeneity across the tree are shown. Asterisks indicate the magnitude of the associated Bayes factor in favor of the model with k WGDs, where $(*)$ $0.5 < \log_{10} K < 1$, $(**)$ $1 < \log_{10} K < 2$, and $(***)$ $\log_{10} K > 2$. For details on the used priors and algorithm settings, we refer to Zwaenepoel and Van de Peer (2020).

point, what is causing the different results compared to the fixed WGD configuration analysis in sec. 3.2.2, but we note that the fixed-dimensional analysis performed in Zwaenepoel and Van de Peer (2020) yielded similar results as the rjMCMC approach. The latter fixed-dimensional analysis was conducted with a somewhat different DL clock model and a different prior distribution on the ancestral gene count[13] compared to our analysis above. We have to defer, however, a closer analysis to future work, where we hope to study the statistical issues underlying these discrepancies in more depth, while at the same time improving the implementation of the rjMCMC approach. □

3.2.5 Concluding remarks

While an interesting problem from a statistical perspective, WGD inference from gene count data may not be very 'data-efficient'. To have gene count data is to have genomic data, and having genomic data means that there is

[13]This study was conducted before our realization that the beta-geometric distribution model serves as a more realistic and flexible prior, and we used a geometric prior with parameter derived from the mean family size instead.

much more information available to investigate ancient WGDs than mere gene counts. Two such sources of information should be noted, and feature in future chapters of this thesis. Firstly, gene sequences provide information about the *gene trees* of multi-copy gene families, which provide, potentially, more fine-grained insights in gene family evolution than mere counts. Secondly, gene count data completely ignores the spatial structure of a genome, and with genomic data of sufficiently high quality, synteny and co-linearity information may be much more helpful to unveil ancient WGDs.

A potentially interesting application of the methods developed in this section which does take the latter into account would be the statistical modeling of gene counts in *microsyntenic* or *anchor* gene families (Zhao et al. 2021). An anchor gene family consists of homologs which share the same gene context, i.e. whose neighboring genes tend to be homologous as well, so that paralogs derived from small-scale duplications are not (ideally) within the same anchor gene family. Duplication within an anchor family can be assumed to proceed only through large-scale duplication events, such as WGD, single-chromosome duplications or large segmental duplications (whatever these may be). In Zhao et al. (2021), we used anchor families and associated phylogenetic profiles for the purpose of phylogenetic inference, but the potential of this representation of genomic data for evolutionary genomics more broadly remains nearly untapped. Assuming a model with a constant rate (independent of family size) of duplication *not* due to WGD, a loss rate linear in the family size, and WGD, we have started experimenting with the rjMCMC approach for inference of WGDs from this type of data.

Of course, the consideration that statistical inference from gene counts may not be a data-efficient way for the purposes of detecting ancient WGDs, does not invalidate our concern with modeling the effect of WGDs on phylogenetic patterns of gene family content. Statistical inference based on phylogenetic BDPs and gene count data remains the most used approach in evolutionary genomics for the study of gene family evolution. We have shown that WGDs present a major source of model violation for phylogenetic linear BDP models and that accounting for WGDs leads to a markedly improved fit and more biologically reasonable duplication rate estimates. Our implementation of phylogenetic BDPs in the DeadBird library provides efficient means to fit complicated (fixed-dimensional) Bayesian phylogenetic BDP models using the rich scientific computation ecosystem in Julia. Our trans-dimensional inference approach using rjMCMC, while promising, requires further work to live up to its expectations.

Importantly, as we noted in sec. 3.1.5 above, WGDs are not the only source of model violation, and more fundamental limitations of the linear BDP remain an important issue for modeling gene family evolution. We stress that these are indeed limitations of the linear BDP, and are not addressed by embedding the linear BDP in ever-more complicated hierarchical models including rate heterogeneity across families and lineages (which is evidently also highly important), as is shown through our posterior predictive simulations. Needless to say, better models of the SSDL process may enable more adequate modeling of WGD as well – as we are essentially modeling the deviation from the SSDL process as due to WGD. Furthermore, a more rich model for the SSDL process may also enable better models for the WGD events themselves compared to the single-parameter DLWGD model of Rabier, Ta, and Ané (2014). We deal with these challenging problems in the rest of the present chapter.

3.3 A two-type branching process model for duplication and loss[14]

While ubiquitously used, the standard phylogenetic BDP models considered above actually provide a rather awkward fit to comparative genomic data. Two related observations make this quite clear. The first is that posterior predictive simulations show that the linear BDP has a hard time providing a good fit to the size distribution for small gene families. Consider for instance the simulations for the linear BDP model in fig. 3.7. For each species in the analysis, the predicted number of single-copy families, which are the most common families in the data, underestimates the actual frequency of single-copy gene families, while the prediction for the frequency of gene families in zero- or two-copy state tends to overestimate the respective observed frequency. It seems, then, that in order for the linear BDP to explain the frequencies of the dominant families (those in zero-, single- or two-copy state), the duplication and loss rates are pulled towards values where they succeed at explaining these best on average, but happen to explain none well in particular. Similar observations hold for the rice data set in fig. 3.4.

The second, related observation, is that the Lynch & Conery-like linear BDIP model, which can be interpreted as a linear BDP conditioned on non-extinction, tends to provide a somewhat better fit while arriving at very different loss rate estimates. For instance, considering the full array of gene

[14]This section is a slight adaptation of our preprint article in Zwaenepoel and Van de Peer (2021).

families across a set of eight *Drosophila* species, the standard linear BDP model yields an estimated loss rate of 0.24 expected loss events per gene per 100 My; whereas considering the set of gene families that are retained with at least one copy across all taxa, and assuming the linear BDIP (in which these gene families cannot go extinct), a loss rate of 3.94 expected loss events per *duplicated* gene per 100 My is obtained. The same discrepancy appears when comparing loss rate estimates from the linear BDP model with crude estimates based on age distributions of duplicated genes (Lynch and Conery 2003, chapter 2).

One of the key assumptions underlying the standard linear BDP model that is generally violated, and which may lead to these observations, is the assumption of a single and constant loss rate per gene *within* a family. Many gene duplicates, although stably established in the genome, exhibit some functional redundancy (examples abound in the molecular biological literature). When a duplicated gene is fully or partially functionally redundant with its parental gene, it seems likely that both copies are subject to higher gene loss rates until either one gene of the duplicate pair is lost or has adopted a distinct function (neofunctionalization), or both genes underwent 'complete' subfunctionalization (hereafter, neo- and subfunctionalization are referred to jointly as 'x-functionalization'). In other words, when a set of genes is (partially) functionally redundant, we may expect an increased per-gene loss rate μ_r compared to a set of non-redundant genes with per-gene loss rate μ_{nr}, as in the former case there will be weaker purifying selection against pseudogenes (in the case of full redundancy for instance, a null-mutant will be effectively neutral (Walsh 2003)). This however induces non-independence among distinct genes in a family, because when one copy in a functionally redundant gene pair gets lost, the loss rate of the retained copy will drop back to μ_{nr}. This contrasts with the commonly employed BDP models which obey the branching property that distinct genes evolve independently. This non-independence represents a serious obstacle for developing more accurate stochastic models of gene family evolution.

In this section, we start by formalizing a model of gene family evolution which deals with functional redundancy in line with the informal model outlined above. We then develop an alternative model of gene family evolution based on a two-type continuous-time branching process to approximate the stochastic evolution of gene families under such a 'redundancy-aware' model, while retaining the independence assumption (branching property) that keeps our models tractable. We conduct statistical inference for our new model in a phylogenetic context, and discuss the implications of our results for long-term

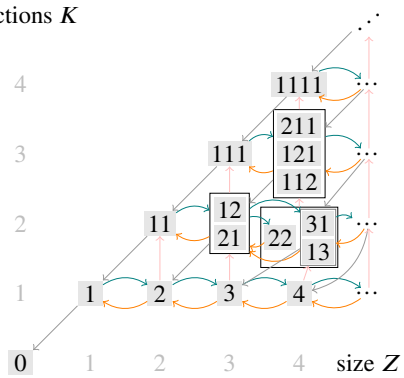


Figure 3.16: Schematic representation of the state space of the continuous-time Markov chain induced by the duplication-loss-functionalization model. A state is represented by a vector x with $K = k$ entries (the number of functions coded for by the family) with $Z = \sum_i x_i$ the total family size. Teal edges represent duplication events, orange edges loss events for redundant genes, gray edges loss of non-redundant genes and pink edges x-functionalization 'events'. States which are indistinguishable for particular transitions are grouped by a box with respect to these transitions (e.g. duplication from state $x = (1, 1, 1)$ yields $(2, 1, 1)$, $(1, 2, 1)$ and $(1, 1, 2)$ with equal probability).

evolutionary dynamics of gene content and the evolutionary genetics of gene duplication.

3.3.1 Redundancy-aware models of gene family evolution

3.3.1.1 Duplication-loss-functionalization

We first describe an idealized *duplication-loss-functionalization* (DLF) model with non-independent evolution of gene copies. We assume a gene family of n genes is partitioned in a random number $K \leq n$ of functional classes, and for $K = k$ we represent a gene family at time t as the length k array of gene counts in each functional class $X(t) = [X_1(t), X_2(t), \ldots, X_k(t)]$, $X_i(t) > 0$, $i \leq k$. An extinct gene family is represented by the empty array. The total gene family size is denoted by $Z(t) = \sum_{i=1}^{k} X_i(t)$. We assume birth-death dynamics, i.e. transitions increment or decrement $Z(t)$ by at most one unit, and we assume that all genes duplicate at rate λ. With respect to the loss dynamics, we assume that genes in a functional class i with $X_i > 1$ suffer loss at rate μ_r

per gene (loss of a redundant copy), whereas genes in a single-copy functional class suffer loss at rate μ_{nr} (loss of a non-redundant copy), an event which leads to a decrease in the number of functional classes in the family. Lastly, we assume that for each class i for which $X_i > 1$, genes shift to a new, not yet existing functional class at rate ν per gene, decrementing X_i and increasing the number of functional classes in the family (neo- or subfunctionalization). Note that the latter event does not affect $Z(t)$. The state space for the CTMC $X(t)$ is illustrated in fig. 3.16. While this model does not admit efficient statistical inference (as far as we are aware), it is straightforward to simulate from using standard techniques.

3.3.1.2 Quasi birth-death processes

Two approaches for symplifying the DLF model come quite naturally. Firstly, the scheme in fig. 3.16 brings to mind so-called (level-dependent) *quasi birth-death (QBD) processes*, which can be seen as bivariate Markov chains with two types of transitions – transitions between *levels* and transitions within a level (Latouche and Ramaswami 1999). The infinitesimal generator of a QBD can be written in block matrix form as

$$Q = \begin{bmatrix} L^{(0)} & B^{(0)} & 0 & 0 & \dots \\ D^{(1)} & L^{(1)} & B^{(1)} & 0 & \dots \\ 0 & D^{(2)} & L^{(2)} & B^{(2)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Where $L^{(i)}$ is the rate matrix for the birth-death process *within* level i, $D^{(i)}$ is the matrix of transition rates between the states of level i and level $i-1$ and $B^{(i)}$ is the matrix of transition rates between the states of level i and $i+1$. The QBD is called level-independent (sometimes 'homogeneous', e.g. Latouche and Ramaswami (1999)) when $L^{(i)}$, $D^{(i)}$ and $B^{(i)}$ are not dependent on i for $i > 0$, and level-dependent otherwise. QBD processes have been studied quite extensively in queueing theory, and a number of techniques have been developed for their stochastic analysis, although work related to statistical inference for such models is scarce. We are unaware of any applications in biology.

More concretely, as a model of gene family evolution, the state $(K(t), Z(t))$ under the DLF model evolves according to a QBD process, although not a particularly tractable one (it is level-dependent, with a rather complicated infinitesimal generator). Here, the number of functions K corresponds to the 'level' of the QBD and we have the usual linear BDP duplication/loss dynam-

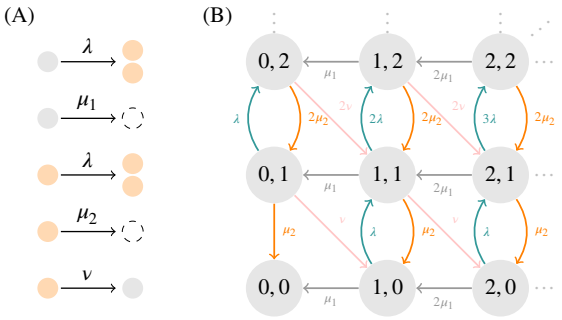


Figure 3.17: The two-type duplication-loss model. (A) Illustration of the possible events and their rates for type 1 (gray circles) and type 2 (yellow cirles) genes. (B) State space of the associated continous-time Markov chain. Each node represents a state (X_1, X_2), where X_1 and X_2 are the numbers of 'type 1' and 'type 2' genes respectively. Non-zero instantaneous transition rates are marked along the edges, with different colors for different evolutionary events.

ics within a level, i.e. for fixed K, whereas transitions between levels model gain and loss of function 'events', or, to be more specific, x-functionalization and loss of a non-redundant gene respectively. While the QBD process induced by the DLF model may be too complicated to work with, other more tractable QBD models derived from it could yield useful models for gene family evolution.

3.3.1.3 Multi-type branching processes

Alternatively, we may conceive of a gene family as consisting of multiple types of genes which evolve independently according to different stochastic laws. In particular, we consider a model with two types of genes, 'type 1' genes which are lost at a low rate, and 'type 2' genes which are associated with high loss rates. In this way, we obtain a two-type continuous-time branching process $X(t) = (X_1(t), X_2(t)) \in \mathbb{N} \times \mathbb{N}$, where $X_1(t)$ and $X_2(t)$ denote the number of 'type 1' and 'type 2' genes at time t respectively. We can then think of such a family as consisting of $Z(t) = X_1(t) + X_2(t)$ genes, coding for $X_1(t)$ functions. We allow for four types of events: gene duplication at rate λ per gene, loss of a type 1 gene at rate μ_1 per type 1 gene, loss of a type 2 gene at rate μ_2 per type 2 gene and a conversion of a type 2 to a type 1 gene (x-functionalization) at a rate ν per type 2 gene. The stochastic evolution of the process is thus

determined by the rate parameters $\theta = (\lambda, \mu_1, \mu_2, \nu)$ in the following way

$$p_{ij}(k,l,\Delta t) = \mathbb{P}\{X(t+\Delta t) = (k,l) | X(t) = (i,j)\}$$
$$= \begin{cases} (i+j)\lambda\Delta t + o(\Delta t) & k = i,\ l = j+1 \\ i\mu_1\Delta t + o(\Delta t) & k = i-1,\ l = j \\ j\mu_2\Delta t + o(\Delta t) & k = i,\ l = j-1 \\ j\nu\Delta t + o(\Delta t) & k = i+1,\ l = j-1 \\ 1 - ((i+j)\lambda + i\mu_1 + j(\mu_2+\nu))\Delta t + o(\Delta t) & k = i,\ l = j \\ o(\Delta t) & \text{else} \end{cases} \tag{3.7}$$

for $i, j, k, l \geq 0$. The process is a bivariate birth-death process and is Markovian (see fig. 3.17 for a graph representation of the state space). It is also a QBD, although this is not a particularly helpful characterization of the process.

As a model of gene family evolution by gene duplication and loss, $X_1(t)$ denotes the number of 'base' genes in the family, while $X_2(t)$ denotes the number of 'excess' genes in the family for which the eventual fate (non-, neo- or subfunctionalization) is yet to be determined at time t. The model defined above assumes that all genes duplicate at the same per-gene rate λ and that the fate of a type 2 gene is resolved after an exponentially distributed time with mean $\mu_2 + \nu$, with the probability of nonfunctionalization being $\mu_2/(\mu_2+\nu)$ and the probability of sub- or neofunctionalization $\nu/(\mu_2+\nu)$ (turning an excess gene into a base gene). We further assume type 1 genes are removed with rate μ_1. Throughout we assume $\{\mu_1, \lambda, \nu\} < \mu_2$. We refer to this model as the 'two-type duplication-loss (DL) model'. Note that for $\mu_1 = \nu = 0$, and $X_1(0) = 1$, we obtain the linear BDIP model for non-extinct families considered above.

Importantly, because of the independence assumption in the latter approach, μ_2 cannot be directly interpreted as the rate of gene loss of a functionally redundant gene μ_r. To see this, consider a state such as 'four genes coding for a single function'. In the DLF model, this state is represented as [4] while in the two-type branching process it would be represented as $(1, 3)$. All four genes are functionally redundant in this case, and under the DLF model we would have a total loss rate $4\mu_r$ where μ_r is the loss rate for a functionally redundant gene. Instead in the branching process model, the total rate of gene loss will be $\mu_1 + 3\mu_2$, so that when $\mu_1 \ll \mu_2$ the rate of loss per redundant gene is $\approx 3\mu_2/4$. Under the two-type model the rate of gene loss per *excess* gene μ_2 is constant for different family sizes, while the rate of gene loss per *redundant* gene increases with increasing number of excess genes per base

gene, approximately equaling $\mu_2(n-1)/n$ for a group of n redundant genes.

3.3.1.4 General BDPs?

We have already defined the general BDP as a birth-death process with arbitrary state-dependent duplication and loss rates λ_i and μ_i above (chapter 2). A natural question that emerges is whether such a general BDP, or at least a less restrictive BDP than the linear one, could yield a reasonable model of gene family evolution that can deal with loss rate heterogeneity within gene families. Although such a general BDP would evidently provide an improved fit to gene count data, it would not admit a straighforward interpretation as a model of gene family evolution, in contrast with the models considered above. The essential difference is of course that the QBD and multi-type models introduce a latent variable which models the evolution of gene function within a family, and on which loss rates depend, whereas a general BDP does no such thing. For a general BDP model, a gene family of n genes will always have the same exponential lifetime distribution, and the loss rate μ_n can only be interpreted as 'the rate of loss in a family of size n', which does not appear to be very meaningful, and would depend on the timescale and species tree with respect to which gene families are defined. On the other hand, in the case of a QBD process the lifetime distribution depends on the level random variable, while for the two-type process it depends on the partitioning of the n family members in type 1 and type 2 genes, and the rate parameters retain in both cases a biologically meaningful interpretation.

3.3.2 Inference for the two-type phylogenetic BDP

Likelihood-based inference of model parameters from comparative genomic data requires that we can efficiently compute transition probabilities under the model. Neither the QBD models nor multi-type branching processes admit straightforward calculation of transition probabilities, however, so we have to resort to slightly more involved techniques. While the QBD models are attractive and we could draw a lot of inspiration and ideas from their treatment in queueing theory, we have not made a lot of progress down that path. We leave the QBD models for future work and shall henceforth be concerned with statistical inference for the two-type branching process model in a phylogenetic context.

3.3.2.1 Transition probabilities for the two-type model

The transient distributions for multi-type branching processes like the two-type DL model are analytically intractable in general. Xu et al. (2015) however presented a numerical approach for computing transition probabilities for Markovian multi-type branching processes based on inversion of the associated probability generating functions. Let $f_{ij}(s_1, s_2, t)$ denote the pgf for the two-type DL model in eq. 1 when starting at $X(0) = (i, j)$, i.e.

$$f_{ij}(s_1, s_2, t) = \sum_{k}^{\infty} \sum_{l}^{\infty} p_{ij}(k, l, t) s_1^k s_2^l$$

Importantly, the branching property implies the following relationship among the probability generating functions for different $X(0)$

$$f_{ij}(s_1, s_2, t) = f_{10}(s_1, s_2, t)^i f_{01}(s_1, s_2, t)^j$$

(e.g. Athreya and Ney (1972)), so we may work with f_{10} and f_{01} and recover the desired pgfs easily.

In contrast with the linear BDP, we have no closed form solution for f_{10} and f_{01}. We can however, using techniques from Bailey (1990) and Xu et al. (2015), derive a system of ordinary differential equations of which the pgf constitutes a solution. Let $p_{ij}(k, l, dt) = r_{ij}(k, l)dt + o(dt)$, and define the auxilliary generating functions

$$u_{10}(s_1, s_2) = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} r_{10}(k, l) s_1^k s_2^l$$

and $u_{01}(s_1, s_2)$ analogously in terms of $r_{01}(k, l)$. Note that the r_{10} and r_{01} functions are determined by the definition of the process in terms of its infinitesimal rates in eq. 3.7. Filling in the relevant parameters we obtain

$$\begin{aligned} u_{10}(s_1, s_2) &= \mu_1 - (\lambda + \mu_1)s_1 + \lambda s_1 s_2 \\ u_{01}(s_1, s_2) &= \mu_2 + \nu s_1 - (\lambda + \nu + \mu_2)s_2 + \lambda s_2^2 \end{aligned}$$

The pgfs for the process are related to the auxilliary generating functions

$$f_{10}(s_1, s_2, t) = \sum_{k}^{\infty} \sum_{l}^{\infty} \Big(\mathbb{1}_{k=1,l=0} + r_{10}(k, l)t + o(t) \Big) s_1^k s_2^l$$

and similarly for f_{01}, so that we get

$$f_{10}(s_1, s_2, t) = s_1 + u_{10}(s_1, s_2)t + o(t)$$
$$f_{01}(s_1, s_2, t) = s_2 + u_{01}(s_1, s_2)t + o(t)$$

Note furthermore that $\partial f_{ij}(s_1, s_2, t)/\partial t = u_{ij}(s_1, s_2)$. We can expand f_{10} as a Taylor series around t

$$f_{10}(s_1, s_2, t + h) = f_{10}(s_1, s_2, t) + h \frac{\partial f_{10}(s_1, s_2, t + h)}{\partial h}\Big|_{h=0} + o(h)$$

Now exploiting the property of the branching process that

$$f_{10}(s_1, s_2, t + h) = f_{10}\{f_{10}(s_1, s_2, t), f_{01}(s_1, s_2, t), h\}$$

we can rewrite the Taylor expansion as

$$\begin{aligned} &f_{10}(s_1, s_2, t + h) \\ &\quad = f_{10}(s_1, s_2, t) + h \frac{\partial f_{10}\{f_{10}(s_1, s_2, t), f_{01}(s_1, s_2, t), h\}}{\partial h}\Big|_{h=0} + o(h) \\ &\quad = f_{10}(s_1, s_2, t) + h u_{10}\{f_{10}(s_1, s_2, t), f_{01}(s_1, s_2, t)\} + o(h) \end{aligned}$$

Which shows that

$$\frac{\partial f_{10}(s_1, s_2, t)}{\partial t} = u_{10}\{f_{10}(s_1, s_2, t), f_{01}(s_1, s_2, t)\}$$

and analogously for $\partial f_{01}/\partial t$. Combining this result with the generating functions u_{10} and u_{01}, we arrive at the following system of non-linear ordinary differential equations (ODEs)

$$\begin{aligned} f'_{10} &= \mu_1 - (\lambda + \mu_1) f_{10} + \lambda f_{10} f_{01} \\ f'_{01} &= \mu_2 + \nu f_{10} - (\lambda + \nu + \mu_2) f_{01} + \lambda f_{01}^2 \end{aligned} \tag{3.8}$$

Where the arguments s_1, s_2 and t are omitted for notational convenience and differentiation is with respect to t.

No closed form solution for $f_{10}(s_1, s_2, t)$ and $f_{01}(s_1, s_2, t)$ can be obtained for this system, and we solve these ODEs numerically using the `Tsit5` solver implemented in DifferentialEquations.jl (Rackauckas and Nie 2017; Tsitouras, Famelis, and Simos 2011). To obtain transition probabilities from the pgfs we use the numerical inversion method of Xu et al. (2015), which involves

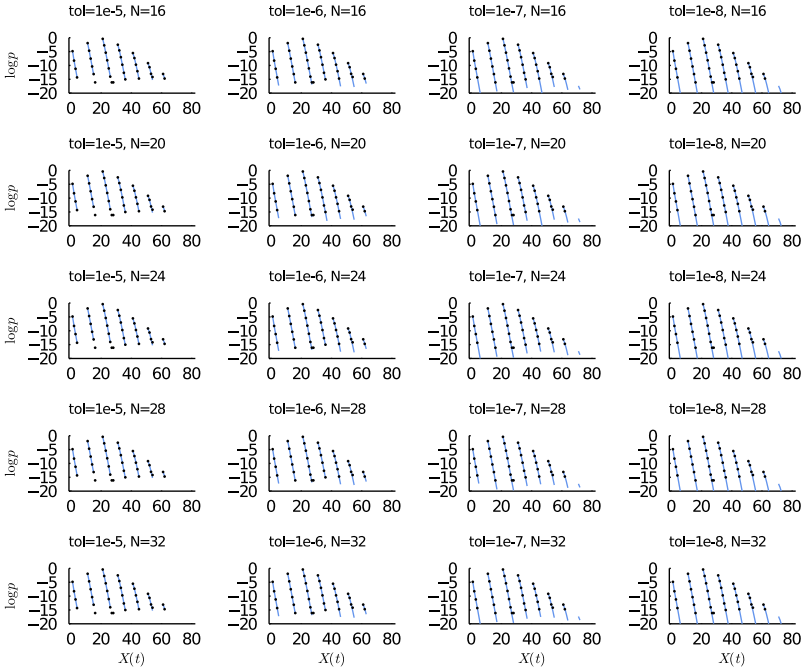


Figure 3.18: Comparison of log transition probabilities computed using the pgf method (blue lines) and estimated using Monte Carlo simulations (black dots) for different algorithm settings (tolerance settings in the ODE solver (`tol`) and FFT length N). Parameters were $\lambda = 0.2, \mu_1 = 0.1, \nu = 0.2$ and $\mu_2 = 5$. Transition probabilities are computed from the state $X(0) = (2,3)$ to states $X(t) = (i,j)$ where $0 \leq i < 8$, $0 \leq j < 10$ and $t = 1$. Target states along the x-axes are ordered $[(0,0),(0,1),\ldots,(0,9),(1,0),(1,1)\ldots]$. Monte Carlo estimates are based on 10 million independent simulations from the model.

re-expressing the pgf as a Fourier series $f_{jk}(e^{2\pi i w_1}, e^{2\pi i w_2}, t)$ so that the coefficients corresponding to the transition probabilities are given by the inverse Fourier transform, which can be computed numerically using the fast Fourier transform (FFT) along an $N \times N$ grid. Choice of a higher N should lead to more accurate transition probabilities, and determines the maximum state from and to which we can compute transition probabilities. In practice the numerical error is dominated by the tolerance settings in the ODE solver, so that increasing N beyond relatively small values (e.g. $N = 16$) does not lead to much gain in accuracy for transition probabilities among states with reasonable probability in our applications (fig. 3.18).

3.3.2.2 Count data likelihood along a phylogeny

With a reasonably efficient method for computing the transition probabilities available, likelihood-based statistical inference in a phylogenetic context is possible. Similar to our discussion of the linear phylogenetic BDP above, we denote by $X_{1,u}$ and $X_{2,u}$ the number of type 1, respectively type 2, genes at node $u \in V(S)$, with $X_u = (X_{1,u}, X_{2,u})$. We assume the same tree-structured generative process as for the phylogenetic BDPs considered above but now with the two-type DL process operating along the branches of S. In the present case we do not, however, actually observe the state of the process at the leaves of the species tree. Specifically, we cannot observe the number of type 1 and type 2 genes at the leaves of S, but assume that we can only observe $Z_u = X_{1,u} + X_{2,u}$ for $u \in \mathcal{L}(S)$. The resulting probabilistic graphical model (PGM) is depicted for a hypothetical three-taxon tree ($m = 3$) in fig. 3.19.

Assuming we can decide on a reasonable bound for the X_1 and X_2 variables, the likelihood of the observed data $p(y|\theta, \phi) = \prod_{i=1}^{n} p(y_i|\theta, \phi)$ conditional on the parameters of the branching process θ and the prior distribution for the root ϕ can be computed using variable elimination along the PGM, i.e. using Felsenstein's pruning algorithm, and integrating the marginal likelihood values at the root over a suitable prior distribution on the root state. For the latter, we use, as before, the beta-geometric distribution with mean η and dispersion ζ as a prior for the *total* number of genes Z_o at the root node o of the species tree. However, to conduct inference, a prior for $(X_{1,o}, X_{2,o})$ is required. Here we assume that there is at least one type 1 gene in each family, and that among the $Z_o - 1$ remaining genes, each gene is of type 1 with probability r and type 2 with probability $1 - r$, so that the number of type 2 genes is a Binomial$(Z_o - 1, 1 - r)$ random variable. We use the resulting distribution for $(X_{1,o}, X_{2,o})$ in our analyses for the two-type DL model and refer to it as the

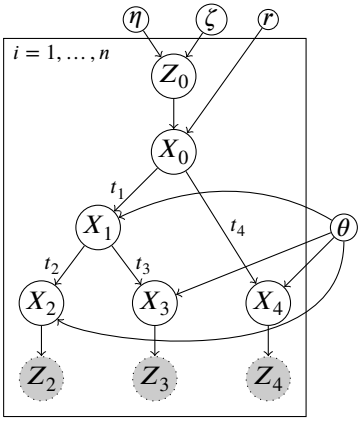


Figure 3.19: Probabilistic graphical model for the phylogenetic two-type DL model of gene content evolution. We use the notation of Höhna et al. (2014), where circular, unshaded nodes represent unobserved random variables, dotted circular nodes represent conditionally deterministic variables and shaded nodes represent observed variables. Note that $\theta = \{\lambda, \mu_1, \mu_2, \nu\}$. Priors for η, ζ and θ used in Bayesian analyses under the model are not shown.

BG-Binomial prior (fig. 3.19).

We note that, given that the two-type DL model has the branching property, it may be possible to devise an algorithm similar to the one of Csűrös and Miklós (2009) outlined above for the linear phylogenetic BDP, which relies on the conditional survival likelihood and does not demand an artificial bound on the natural state space of the BDP. However, the lack of a recursive or closed form expression for the transition probabilities (as far as we are aware) makes this quite challenging, and it is unclear at present whether any gains in terms of computational efficiency or numerical stability are to be expected from such an approach.

3.3.2.3 Extinction probabilities and condition factors

As for the single-type phylogenetic BDP models considered above, in order to compute the correct likelihood, we need to account for the biases induced by the sampling process. In particular, to rule out *de novo* gain of genes in arbitrary subtrees of the phylogeny, we usually filter the data so that at least one gene is present in each clade stemming from the root of the species tree.

Using the same notation as in sec. 3.1.2.5, we see that we require the probability $\mathbb{P}(\bar{E}_u \cap \bar{E}_v)$. We again can rely on the conditional independence of subtrees of S and properties of the pgfs to obtain

$$\begin{aligned}\mathbb{P}(\bar{E}_u \cap \bar{E}_v) &= \sum_{k=1}^{\infty}\sum_{l=1}^{k} \mathbb{P}\{\bar{E}_u \cap \bar{E}_v | X_o = (l, k-l)\}\mathbb{P}\{X_o = (l, k-l)\} \\ &= \sum_{k=1}^{\infty}\sum_{l=1}^{k} \left(1 - g^{(u)}_{l,k-l}(\mathbf{0})\right)\left(1 - g^{(v)}_{l,k-l}(\mathbf{0})\right)\mathbb{P}\{X_o = (l, k-l)\}\end{aligned} \tag{3.9}$$

Here the last factor is given by the prior distribution on the number of ancestral lineages in a gene family at the root of the species tree, and $g^{(u)}_{i,j}(\mathbf{s}) = g^{(u)}_{10}(\mathbf{s})^i g^{(u)}_{01}(\mathbf{s})^j$ is the joint pgf for the leaf observations in the subtree rooted in node u, conditional on the state at the parent of u, say u', being (i, j). Assuming there are m leaves below u and labeling the entries of $\mathbf{s}$ as $(s_{1,1}, s_{1,2}, s_{2,1}, s_{2,2}, \ldots, s_{m,1}, s_{m,2})$, where the first index refers to the leaf node in the subtree below u and the second to the gene type, this pgf is written more explicitly as

$$\begin{aligned}g^{(u)}_{10}(\mathbf{s}) = \sum_{(k_1,l_1)} \cdots \sum_{(k_m,l_m)} &\mathbb{P}\{X_1 = (k_1, l_1), \ldots, X_m = (k_m, l_m) | X_{u'} = (1, 0)\} \\ &\times s_{1,1}^{k_1} s_{1,2}^{l_1} \cdots s_{m,1}^{k_m} s_{m,2}^{l_m}\end{aligned}$$

and of course analogously for $g^{(u)}_{01}$. By the branching property and conditional independence of disjoint subtrees, this pgf can be evaluated efficiently using a postorder traversal and the single branch pgf f_{10}. Specifically, consider a node u, a distance t_u from its parent, with (if it is not a leaf) child nodes v and w. We have the following recursive relation:

$$\begin{aligned}g^{(u)}_{10}(\mathbf{s}) &= f_{10}\left[h^{(u)}_{10}(\mathbf{s}), h^{(u)}_{01}(\mathbf{s}), t_u\right] \\ h^{(u)}_{10}(\mathbf{s}) &= \begin{cases} s_{u,1} & \text{if } u \text{ is a leaf} \\ g^{(v)}_{10}(\mathbf{s}) g^{(w)}_{10}(\mathbf{s}) & \text{else} \end{cases}\end{aligned}$$

with analogous recursions holding for $g^{(u)}_{01}$ and $h^{(u)}_{01}$. With the extinction probabilities available for the child nodes of the root u and v, we can approximate the condition factor $\mathbb{P}\{\bar{E}_u \cap \bar{E}_v\}$ by the first couple of terms in eq. 3.9.

3.3.2.4 Bayesian inference

With an algorithm for the likelihood $p(y|\theta, \phi)$ available we can perform Bayesian inference for the two-type DL model using Markov Chain Monte Carlo (MCMC) given suitable prior distributions. Unless stated otherwise, we adopt the following priors in our analyses:

$$\begin{aligned} \mu_2 &\sim \text{Exponential}(5) \\ a_1, a_2, a_3 &\sim_{\text{iid}} \text{Beta}(1, 1) \\ (\lambda, \mu_1, \nu) &= (a_1\mu_2, a_2\mu_2, a_3\mu_2) \\ r &\sim \text{Beta}(1, 1) \end{aligned}$$

Where we assume λ, μ_1, ν are all $< \mu_2$. We fix η and ζ to the posterior mean values obtained from the fit of the beta-geometric distribution to the relevant data. We implemented a simple adaptive Metropolis-within-Gibbs (MWG) algorithm (Roberts and Rosenthal 2009, see also Appendix A) to sample from the posterior distribution. All methods are implemented in the Julia programming language (Bezanson et al. 2017) and the associated package is freely available online (see Appendix B).

3.3.3 Simulation experiments

3.3.3.1 Estimation for simulated data under the two-type DL model

We first assess our ability to recover true parameter values of the two-type DL model for simulated data from the same model. In particular, the fact that in empirical data sets we assume only to observe the total gene count Z_u for each family at each leaf node u (and not the counts for each type $X_u = (X_{1,u}, X_{2,u})$) is a potential source of identifiability issues. Indeed, it is clear that for a single branch, it would be impossible to identify model parameters of the two-type DL model based on nothing more than a single observation of Z for each family. By considering gene family counts along a phylogeny, a single family provides multiple correlated observations of the evolutionary process, providing information about gene content in ancestral branches of the tree. Observations of Z along the leaves of a phylogeny should therefore also provide information about ancestral gene content at the type level (X_1, X_2).

Simulations of data sets along the *Drosophila* phylogeny (see fig. 3.6) consisting of 1000 gene families across a range of randomly drawn parameter values

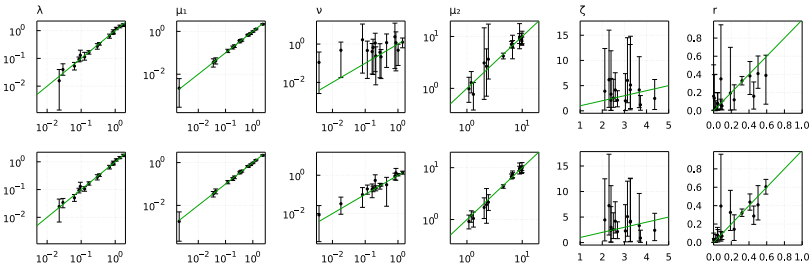


Figure 3.20: Posterior mean parameter estimates and 95% uncertainty intervals (y-axis) as a function of the true simulated value (x-axis) for simulations of data sets of 1000 gene families. μ_2 values were drawn uniformly from the interval (1, 10), while λ/μ_2 μ_1/μ_2 and ν/μ_2 were drawn from a Beta(9, 1) distribution. r was drawn from a Beta(3, 1) distribution and ζ from a lognormal distribution with mean 3 and variance 0.2. Plots are shown pairwise in six columns, with on top the results conditioning on the incompletely observed data $Z = X_1 + X_2$ while on bottom the results conditioning on the fully observed data $X = (X_1, X_2)$ are shown.

indicate that overall, parameters can be estimated accurately even for relatively small data sets (fig. 3.20). As expected, using the completely observed two-type data does lead to a considerably lower variance in the marginal posterior distribution for ν compared to the collapsed data. It may be that the marginal posterior mean for ν overestimates the true value when using the incompletely observed data, however all uncertainty intervals contain the true value, so this risk seems to be minor. Also noticeable, but less dramatic is the difference in posterior variance for μ_2 and r (the probability of excess genes at the root to be of type 1).

In addition, we simulated a data set of 10.000 gene families along the eight-taxon *Drosophila* phylogeny using model parameters that seemed reasonable based on exploratory analyses of subsets of the actual *Drosophila* data set ($\lambda = 0.2, \mu_1 = 0.1, \nu = 0.2, \mu_2 = 5, \eta = 0.95, \zeta = 4$ and $r = 0.5$). We obtain similar results as for our smaller simulations (fig. 3.21), in particular we find that the posterior variance for ν is much higher when using the incompletely observed data. Nevertheless, the associated posterior mean values seem to align rather well with each other and the true value, suggesting that inference for the two-type model in a phylogenetic context is indeed possible without observing a type-specific census. When we perform parameter inference for the single-type linear BDP model and the single-type model without extinction (linear BDIP) for the same simulated data set, we find that the duplication rate tends to be underestimated in both models (tbl. 3.7). The loss rate of the

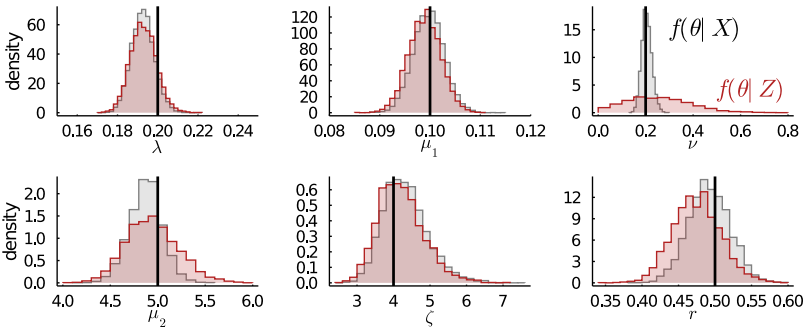


Figure 3.21: Marginal posterior distributions for a simulated data set of 10.000 gene families, simulated along the eight-taxon *Drosophila* phylogeny from the two-type DL process with parameters $\lambda = 0.2, \mu_1 = 0.1, \nu = 0.2, \mu_2 = 5$ and a bounded Beta-geometric prior on the number of lineages at the root with $\eta = 0.95$, dispersion $\zeta = 4$ and bound at $Z = 10$. When there is more than one lineage at the root, each additional lineage has a $r = 0.5$ probability of being a type 2 gene. In black the posteriors conditional on the fully observed data (i.e. $X^{(u)} = (X_1^{(u)}, X_2^{(u)})$ tuples for each leaf node u) are shown, while in red the posteriors conditional on the total gene count $Z^{(u)}$ are shown.

default single-type model only very slightly overestimates μ_1 while the loss rate in the single-type model without extinction strongly underestimates μ_2.

3.3.3.2 Estimation for simulated data under the DLF model

We next evaluate to what extent the two-type DL model can approximate the dynamics of the idealized DLF model, which entails non-independent evolution of gene copies within a family (see methods). We again simulated multiple data sets of 1000 gene families as well as a large data set of 10.000 gene families for the eight-taxon *Drosophila* phylogeny. Note that the total loss rate for a family consisting of one redundant pair will be $2\mu_r$ in the DLF model, whereas the total loss rate for such a family under the two-type DL model (corresponding to the state $X = (1, 1)$) will be $\mu_1 + \mu_2 \approx \mu_2$. Because such small families dominate the data, we expect that $\mu_2 \approx 2\mu_r$.

In line with this expectation, we find that the estimated value of μ_2 under the two-type DL model corresponds to twice the simulated loss rate per redundant gene (μ_r) under the DLF model (fig. 3.22). Similar observations hold for ν, although the large posterior variance blurs the expected relationship of ν in the

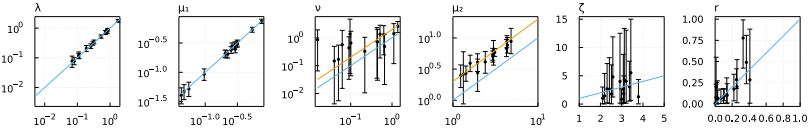


Figure 3.22: Posterior mean parameter estimates and 95% uncertainty intervals (y-axis) as a function of the true simulated value (x-axis) for simulations of data sets of 1000 gene families. Data sets were simulated under the DLF model, while inference was performed under the two-type DL model. μ_2 values were drawn uniformly from the interval (1, 5), while λ/μ_2 μ_1/μ_2 and ν/μ_2 were drawn from a Beta(9, 1) distribution. r was drawn from a Beta(3, 1) distribution and ζ from a lognormal distribution with mean 3 and variance 0.2. Inference is based on the total gene count (Z). The orange lines for μ_2 and ν mark the expected approximate relationship between the simulated parameter value in the DLF model and the corresponding parameter in the two-type DL model (i.e. $\mu_2 \approx 2\mu_r$ and $\nu \approx 2\nu_{\text{DLF}}$).

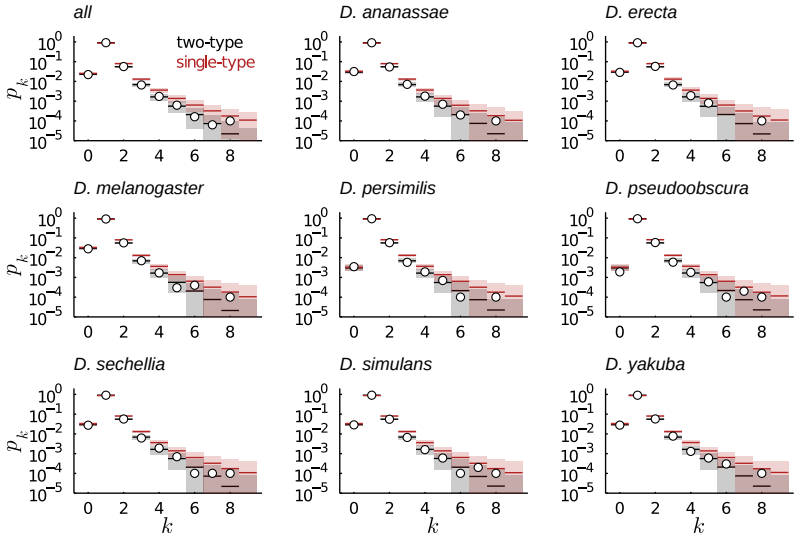


Figure 3.23: Posterior predictive simulations for the family size (k) distribution under the two-type DL model (gray) and the default single-type linear BDP model (red) applied to a data set of 10.000 families simulated under the DLF model. The dots mark the observed size distribution in the simulated data set.

Table 3.7: Posterior distribution summary showing marginal posterior means and 95% uncertainty intervals for various models for the simulation replicate associated with fig. 3.21. We show parameter estimates for the single type models in the same rows as the parameters with a somewhat similar interpretation in the two-type model for the sake of comparison.

Parameter	Two-type DL model	Linear BDP	Linear BDIP (non extinct)
λ	0.19 (0.18, 0.21)	0.14 (0.14, 0.15)	0.14 (0.13, 0.15)
μ_1	0.10 (0.09, 0.11)	0.12 (0.11, 0.13)	-
μ_2	4.94 (4.44, 5.51)	-	1.25 (1.11, 1.42)
ν	0.25 (0.03, 0.54)	-	-
ζ	4.20 (3.15, 5.60)	6.77 (4.62, 9.96)	3.69 (2.65, 5.11)
r	0.48 (0.41, 0.54)	-	-

DLF model to twice ν in the two-type DL model. Additionally, we note that the posterior mean μ_1 and λ values for the two-type DL model are very accurate estimators of the corresponding parameters under the DLF model. For the large data set simulated under the DLF model with $\lambda = 0.2$, $\mu_{nr} = 0.1$, $\mu_r = 3$ and $\nu = 0.2$, we obtain an estimate for the loss rate per excess gene μ_2 of 5.99 (5.35, 6.65), again coinciding with twice the true loss rate per redundant gene. Similarly, the posterior mean estimate for ν under the two-type DL model is obtained as 0.35 (0.10, 0.62), also approximately corresponding to twice the true value of the underlying DLF model. The much higher posterior variance for this parameter makes the correspondence again however less clear. These simulations therefore suggest it is reasonable to assume that twice the loss rate per *excess* gene, estimated under the two-type DL model, may serve as an approximation to the more intuitive loss rate per *redundant* gene in the DLF model. Posterior predictive simulations for the fitted two-type DL model further show that the posterior predictive distribution is fully compatible with the data simulated under the DLF model (fig. 3.23), indicating that the two-type DL model indeed can serve as an approximation to the DLF model.

3.3.4 Analysis of *Drosophila*, yeast and primate data

We performed Bayesian inference for the two-type DL model using gene count data from *Drosophila*, yeast and primates already analyzed in sec. 3.1.5, and compare the estimated model parameters (tbl. 3.8) to parameter estimates for different single-type BDP models (see tbl. 3.3). We note that for the yeast and primates data sets, initial tests indicated that numerical inaccuracies in the likelihood for the two-type model could lead to a failure of the MCMC

algorithm to converge in some runs. Increasing the FFT length from $N = 16$ (as employed in our simulations) to $N = 32$ and decreasing tolerance settings in the ODE solver alleviated these issues and resulted in proper convergence.

Interestingly, the estimated rates for slow processes in the *Drosophila* and primates data sets are roughly on the same scale, corresponding to about 0.1 to 0.3 events per gene per 100 My, whereas the yeast data set yields parameter estimates that are considerably lower. Of course, this is not wholly unexpected, given the vastly different genomic organization of yeast species, with small genomes typically consisting of 5000 to 7000 predicted genes. Clearly, a failure to model the different loss rates within a family, supposedly caused by functional redundancy, leads to a loss rate estimate μ that is dominated by the large number of non-redundant genes that only rarely get lost. The loss rate in the default single-type BDP model therefore more closely resembles μ_1 than μ_2 in the two-type model. The opposite holds for the single-type model for the number of *excess* genes in non-extinct families (the linear BDIP). Indeed, we find that for the yeast and primates data sets, the posterior mean parameter estimates appear to correspond to a combination of the corresponding estimates for the linear BDP and BDIP models (tbl. 3.3), with μ_1 in the two-type model roughly the same as μ in the linear BDP model and μ_2 roughly corresponding to μ in the linear BDIP model fit. This is especially the case for the primates data, where the rate of type 2 to type 1 transitions appears to be negligible compared to the other rates, and we end up with a two-type model that is essentially like a linear BDP but with a distinct rate for $1 \rightarrow 0$ transitions. In all data sets, the loss rate for an excess (type 2) gene μ_2 is an order of magnitude higher than other rate parameters, with an implied half-life ($t_{1/2}$) of a type 2 gene of about 11, 39 and 30 My for the three data sets respectively (tbl. 3.8). If we assume $\mu_2 \approx 2\mu_r$ (see above), these can be interpreted as the half-lives of duplicate *pairs* under the DLF model. We further note that the ratios of λ/μ_2 of about 0.04, 0.02 and 0.06 are in agreement with the beta-geometric stationary distribution fit of the single-type model. The marginal posterior mean estimate of the probability that a new duplicate gets established eventually (i.e. becomes a type 1 gene, rather than suffering loss, p_x) is a mere 3% for the *Drosophila* data, about 2% for the yeast data and 0.5% for the primates data. Again, under the DLF model, this value can be interpreted as the probability that a duplicate *pair* undergoes successful sub- or neofunctionalization so that it is stably established in the genome.

For the *Drosophila* data, and to a lesser extent the yeast data, we find that the duplication rate estimated for the two-type DL model is higher than for the single-type models, which is in line with our simulations above. However

the loss rates for the linear BDIP model (see tbl. 3.3) tend to be somewhat lower than the estimates for μ_2 for the two-type model. This is consistent with the dynamics the two-type process is supposed to model, as duplicated genes that sub- or neofunctionalize should pull back the loss rate towards the loss rate of non-redundant genes in the corresponding single-type model. In other words, when assuming all duplicated genes to have the same loss rate (as in the single-type non-extinct model), the presence of duplicate genes that have become essential leads to a downwardly biased loss rate when interpreted as the rate of pseudogenization of redundant duplicate genes. We note that an additional explanatory factor may be that the subset of families which do not go extinct has a lower average loss rate per excess gene than the full data set.

Table 3.8: Marginal posterior parameter estimates for the *Drosophila*, yeast and primates data sets (see sec. 3.1.5). For all analyses, the prior on the number of lineages at the root was a beta-geometric distribution with η and ζ parameters fixed to the marginal posterior mean values obtained from the stationary distribution fit (fig. 3.6, tbl. 3.3). We use an exponential prior for μ_2 with mean equal to the posterior mean value of μ under the single-type model with no extinction. $t_{1/2}$ and p_x denote the half-life (in My) and probability of x-functionalization (in %) of a type 2 gene respectively.

Parameter	*Drosophila*	Yeasts	Primates (GO:0002376)
λ	0.26 (0.24, 0.28)	0.03 (0.02, 0.03)	0.13 (0.12, 0.14)
μ_1	0.19 (0.18, 0.20)	0.06 (0.05, 0.06)	0.13 (0.12, 0.13)
μ_2	6.37 (5.98, 6.93)	1.79 (1.64, 1.97)	2.32 (2.15, 2.50)
ν	0.20 (0.08, 0.35)	0.05 (0.01, 0.10)	0.01 (0.00, 0.05)
r	0.12 (0.10, 0.15)	0.01 (0.00, 0.03)	0.01 (0.00, 0.03)
$t_{1/2}$ (My)	11 (10, 12)	39 (35, 42)	30 (28, 32)
p_x (%)	3.1 (1.3, 5.0)	2.3 (0.4, 5.2)	0.6 (0.0, 2.0)

Posterior predictive simulations indicate that the two-type model provides a better fit to the size distribution than the single-type model for all data sets (see fig. 3.24 for the *Drosophila* data). For all three data sets we find that the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(p, \tilde{p})$ from the posterior predictive frequency distribution $\tilde{p}$ to the observed frequency distribution p for the two-type model is less than half of the same KL divergence obtained under the default single-type model. We note that the different degrees of correspondence of the posterior predictive distributions to the observed size frequency distributions for different taxa suggests rate heterogeneity *across branches* of the species tree, a complication we ignore here. As expected, the posterior predictive size distribution is underdispersed with respect to the true distribution further in the tail, which is a consequence of ignoring rate heterogeneity *across families*. This is less so for the single-type model, where the tail of the

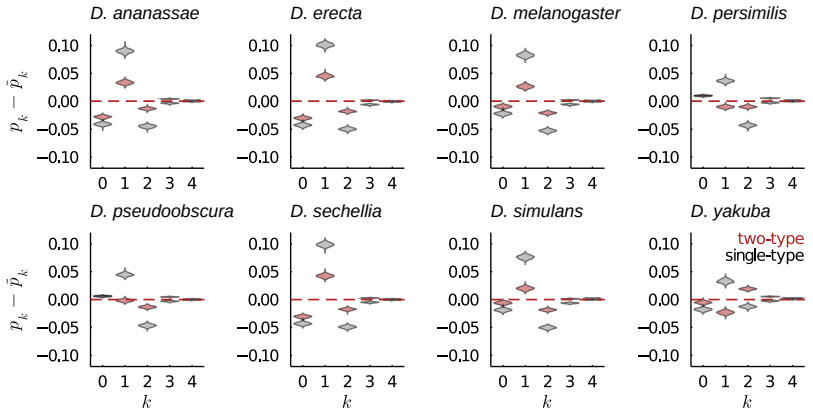


Figure 3.24: Posterior predictive densities of the gene family size distributions for each of the eight leaves of the *Drosophila* phylogeny. The distribution of the differences of the observed gene family size frequencies (p_k) with the frequencies observed in 1000 posterior predictive simulations ($\tilde{p}_k$) is plotted. In red and gray posterior predictive distributions for the two-type and single-type (linear BDP) model are shown respectively.

posterior predictive distribution is closer to the observed distribution. This is simply a result of the distribution on the number of lineages at the root (which, we remind the reader, is derived from a phylogeny-unaware beta-geometric fit of the observed data) being better preserved under the process, which has a lower overall event rate. We note that the posterior predictive distribution for the non-zero counts under the two-type model is nearly indistinguishable from the posterior predictive distribution for the single-type model for non-extinct families.

While rate heterogeneity across families could be accounted for by using a mixture modeling approach, we refrain from doing so in the present study for reasons of computational feasibility. We may however apply an *ad hoc* procedure to apply rate heterogeneity across families in our posterior predictive simulations by considering the fit of the BG distribution to the data. As we discussed in chapter 2, the BG distribution is the stationary distribution of a special case of the linear BDIP model with rate heterogeneity under the model where the ratio λ/μ_2 is iid distributed according to a beta distribution across families. The parameters of this stationary distribution can be easily estimated from the data, and we estimated the dispersion parameter for the *Drosophila* data for instance at $\zeta = 4.01$ (fig. 3.6). The linear BDIP model can however be seen as a special case of the two-type DL model where $\mu_1 = \nu = 0$, and

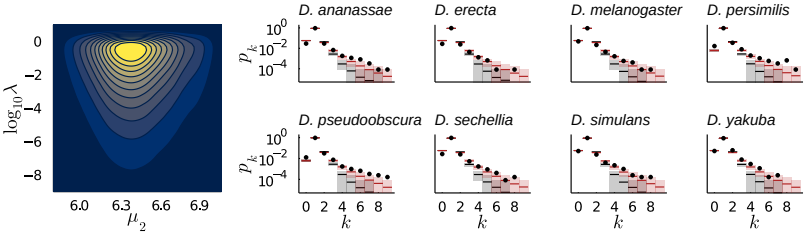


Figure 3.25: Posterior predictive simulations assuming λ/μ_2 is Beta distributed across gene families with dispersion $\zeta = 4.01$. On the left the joint distribution of family-specific λ and posterior mean μ_2 values is shown. The eight panels on the right show the posterior predictive family size distributions with (red) and without (black) the *post hoc* procedure to account for rate heterogeneity across families based on the Beta-Geometric assumption.

where the initial state is $X = (1, X_2)$. If we assume λ/μ_2 to be beta-distributed across families, and take for each replicate simulation of the posterior distribution the value of λ/μ_2 as the mean of this beta distribution, we may simulate from the posterior distribution under this assumption as follows: For each of the $i \in (1, \dots, 1000)$ replicate simulations we sample a $(\lambda/\mu_2, \mu_2)_i$ pair from the posterior and compute $\alpha_i = \zeta(\lambda_i/\mu_{2,i})$ and $\beta_i = \zeta(1-\lambda_i/\mu_{2,i})$. For each of the $j \in (1, \dots, n)$ families in the ith simulation replicate, we sample a random value for $\xi_{ij} = \lambda_{ij}/\mu_{2,ij}$ by sampling from $\text{Beta}(\alpha_i, \beta_i)$. We then obtain λ_{ij} and $\mu_{2,ij}$ by assuming $\mu_{2,ij} = \mu_{2,i}$ and $\lambda_{ij} = \xi_{ij}\mu_{2,i}$. Note that this procedure translates the assumed variation in λ/μ_2 across families to heterogeneity in λ alone across families. The joint posterior distribution of the family-specific rates obtained following this procedure is shown in fig. 3.25. The resulting posterior predictive distribution still fits the small family sizes well (as in fig. 3.24), but now no longer underestimates the proportions of larger families, and yields predictions compatible with the observed power law tail (fig. 3.25). This again clearly shows that the power-law tail of the gene family size distribution can be explained by rate heterogeneity across families.

3.3.5 Whole-genome duplications in the two-type model

In sec. 3.2, we described the DLWGD model of Rabier, Ta, and Ané (2014), which provides one way to account for WGDs in the context of the linear phylogenetic BDP. The essence of the DLWGD model is that it seeks to account for the massive gene loss that is thought to occur after whole-genome duplication using a single parameter, the retention probability q. With probability q

the duplicate gene survives rediploidization and becomes just another gene independently evolving in the phylogenetic BDP, while with probability $1-q$ the gene is lost before rediploidization completes. Clearly, one can hardly think of a more crude model of post-polyploidy genome evolution (except perhaps a no-parameter model, where all gene families simply duplicate, and gene loss after polyploidy is simply modeled by the linear BDP, as in Tasdighian et al. (2017)).

While this model makes sense in the context of the linear BDP model, this is only because the loss rate in the latter mostly accounts for the rate of non-redundant gene loss, or gene family extinction. The rate of loss of (partly) redundant duplicated genes is likely much higher than the loss rate inferred under the linear BDP model, as exemplified by our loss rate estimates under the linear BDIP (for non-extinct families) model and μ_2 in the two-type DL model. From the perspective of the latter models, the problem presented by WGDs takes a rather different form, in that we now need not so much model the massive loss of redundant duplicates created by WGD, but rather the *excess* retention after WGD.

Making the connection with the population genetics of duplicate gene retention will clarify this point. If WGD-derived duplicate genes are completely redundant, as we presume many will be, then a double-recessive model in which recessive mutations at the two loci occur with some mutation rate, and where only the double recessive genotype has reduced fitness, may be a relevant model for the evolution of these loci.[15] Predictions of the time until silencing under such a double-recessive model of duplicate gene evolution however indicate that the observed degree of gene loss after polyploidy is actually much lower than expected under such a model (Watterson 1983; Walsh 2003). In other words, long-term retention of gene duplicates after WGD tends to be more common than expected under a model of strict gene redundancy, suggesting that some portion of the genome violates the double recessive (or a cognate) model. Biological explanations for this phenomenon include subfunctionalization through degenerative mutations (as for instance in the duplication-degeneration-complementation (DDC) model of Force et al. (1999)) as well as gene dosage effects and haploinsufficiency (Kondrashov and Koonin 2004; Birchler and Veitia 2010; Makino and McLysaght 2010) –

[15]Note that in such a model, we typically assume that we can identify two different unlinked diploid loci, i.e. the parental locus and the duplicated locus are two independent disomically inherited loci (e.g. Watterson 1983). This would be a reasonable model for allotetraploids, but less so for autotetraploids with tetrasomic inheritance (either through multivalent formation or random bivalent formation).

all of which lead to a violation of the double recessive model.

The two-type model admits a different approach for modeling retention after WGD which may appear more relevant than the simple DLWGD model in light of the observations noted in the previous paragraph. We describe the model for general k-level multiplications (k-WGMs). We start from the same simple model as the one of Rabier, Ta, and Ané (2014), supposing that WGD events are marked along the species tree S by nodes with in- and outdegree one, and that each WGM event is associated with a single parameter q. At a WGM event, we assume that *all* genes (both of type 1 and type 2) duplicate k times. We assume that the $k-1$ duplicates of a type 2 gene are all of type 2, whereas each of the $k-1$ duplicates of a type 1 gene is of type 1 with probability q or of type 2 with probability $1-q$ independently. Clearly, if the total number of genes in the family before the k-WGM is z, the number of genes after the event will be kz. In contrast with DLWGD model considered above, gene loss after WGD is not modeled by a process distinct from the SSDL process. Instead, we assume that some type 1 genes duplicate to give rise to excess (type 2) genes, which are prone to rapid loss, whereas others give rise to new base (type 1) genes, which get lost only rarely. The latter is of course supposed to model, for instance, dosage sensitive genes, where duplicated copies are supposed to be essential in a similar way as their ancestor was before the WGM event.

The transition probability of the two-type process at a WGM node u under this model is easily obtained. Using the same notation as in sec. 3.2, we have

$$\mathbb{P}\{X_u = (a,b) | X_{u'} = (c,d)\} = \begin{cases} 0 & \text{if } (a+b \neq k(c+d)) \text{ or } \neg(c \leq a \leq kc) \\ \binom{(k-1)c}{a-c} q^{a-c}(1-q)^{kc-a} & \text{else} \end{cases}$$

Similarly, the pgfs required for computing extinction probabilities along the phylogeny can be obtained as

$$f_{10}(s_1, s_2) = \sum_{j=0}^{k-1} \binom{k-1}{j} q^j (1-q)^{k-1-j} s_1^{j+1} s_2^{k-1-j}$$
$$f_{01}(s_1, s_2) = s_2^k$$

With the transition probabilities and pgfs available, we can use the pruning algorithm and recursions for extinction probabilities as before to compute the phylogenetic likelihood under this two-type DLWGD model.

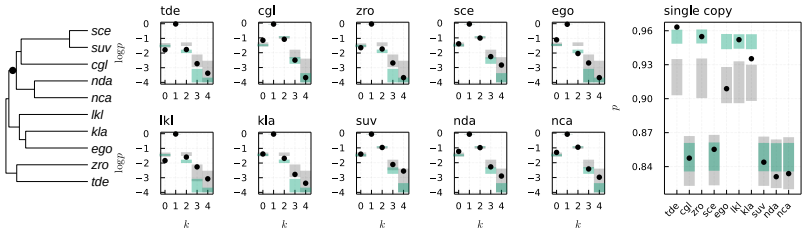


Figure 3.26: Posterior predictive simulations for the single-type and two-type DLWGD model applied to the 10-taxon yeast data set. The phylogeny is shown on the left with the WGD marked by the black circle. The middle set of 10 panels shows the posterior predictive family size distribution for small families for the 10 taxa with the 95% posterior predictive interval for the single-type model shown in gray and for the two-type model in green, with the observed frequencies marked by the black dots. A detail of the proportion of single-copy families is shown on the right.

Example (yeast WGD). We applied the two-type DLWGD model to the well-known WGD event in yeast, using a 10-taxon data set obtained from YGOB (fig. 3.26). We use the beta-geometric distribution with parameters $\eta = 0.98$ and $\zeta = 4$ for the prior on the family size at the root and condition on non-extinction in both clades stemming from the root, as in the yeast data analyses above. We verified the identifiability of the model by conducting inference for simulated data (not shown). We find a posterior mean retention probability estimate of $q = 0.08$ $(0.07, 0.09)$ for the single-type DLWGD model, whereas the estimate for the q parameter for the two-type model was 0.04 $(0.04, 0.05)$, suggesting that about 5% of duplicated genes were retained after WGD. We note that we find a somewhat high type 2 loss rate μ_2 compared to our yeast analyses above, with estimated posterior mean 4.9 $(4.6, 5.3)$, as well as a high x-functionalization rate of 0.35 $(0.30, 0.41)$. While it is hard to assess whether these rates are reasonable and what this could signal, a possible and biologically interesting explanation is differential retention patterns in the different clades which share the WGD. Since in the two-type model, all genes duplicate after the WGD, a low q combined with a high ν and μ_2 allows WGD-derived duplicate genes of type 2 to shift to type 1 genes differentially in lineages that diverge after the WGD. We note that the two-type model may provide a slightly better fit than the single-type one as judged from the posterior predictive family size distributions (fig. 3.26), although the difference is marginal and neither model provides a particularly good fit, possibly in part due to variation in rates across lineages. □

We note that *inference* of WGDs is less straightforward under this model, since

$q = 0$ is no longer representative for a no-WGD model, indeed $q = 0$ would entail that all genes that are born due to the WGD event start their lives as type 2 genes. To infer WGDs under the two-type DLWGD model one would therefore have to resort to more general model selection techniques, which tend to be computationally intensive (as they usually rely on an estimate of the marginal likelihood). We do not consider the two-type DLWGD model further here, noting, as we did above, that a more detailed study of WGD would benefit from considering more data than mere gene counts. We shall briefly come back to this in chapter 6.

3.3.6 Discussion on the two-type model

We have described a two-type continuous-time branching process model of gene family evolution, showed the feasibility of estimating its parameters from incomplete data in a phylogenetic context using simulations, and performed Bayesian inference of model parameters for comparative genomic data sets. Comparisons with closely related single-type phylogenetic BDP models highlight the shortcomings of these models and indicate how a two-type process may provide a first step towards more realistic stochastic models of gene family evolution, providing more detailed and biologically meaningful quantitative insights in the associated evolutionary processes.

The main virtue of the two-type DL model is the discrimination between non-redundant functional genes that are only rarely lost, and duplicates which may be largely redundant and therefore more prone to loss by pseudogenization. We noted that the model proposed here can be viewed as an approximation to the gene family evolution dynamics where duplicated genes are (partially) functionally redundant and evolve in a non-independent manner (as in the DLF model). Compared to the DLF model, the loss rate μ_2 in the two-type process does *not* correspond to the loss rate per redundant gene, but rather the loss rate per *excess* gene in a family. We explicitly examined the connection between the two models using simulated data and find that parameter estimates under the two-type DL model can be interpreted in the context of the more intuitive DLF model.

An alternative approach to account for differences in loss rates within a family which may be natural to adopt is to consider an age-dependent process (more precisely a 'budding' age-dependent branching process (Greenman and Chou 2016)), where the loss rate of a duplicated gene decreases through time. Such a model has been considered by Zhao et al. (2015), albeit not in the context

of gene content evolution along a phylogeny. Inference of model parameters relies however on knowledge of the gene family trees and ages of duplication events (dated reconciled gene trees in the comparative genomic setting), the estimation of which is of course an extremely challenging statistical problem in its own right. Without access to such high-quality gene trees, inference from gene counts alone seems highly challenging. Indeed, statistical inference of age-dependent branching processes more generally is an active research topic of considerable mathematical sophistication (e.g. Fok and Chou 2013; Greenman and Chou 2016).

It is important to stress what is *not* modelled by the phylogenetic two-type DL process. Most importantly, we do not explicitly model the population genetics of fixation of copy number variants (CNVs). Clearly, most duplication events will either be lost due to drift if neutral, or lost by purifying selection when deleterious, and as a result leave no trace in extant genomes. The duplication rate λ in the model should therefore be interpreted as a proxy for the rate at which gene duplications that rise to high frequency occur. We have however no guarantee that the observed copy numbers are monomorphic in their respective populations, so the presence of low-frequency CNVs in the data may cause λ to overestimate the rate at which duplications that fix in the population occur. Similar considerations hold for the loss rate parameters μ_1 and μ_2, which should be interpreted as rates of gene deletion events, not the rates of loss of unfixed duplicate genes by genetic drift or purifying selection. Of course, exactly the same issues hold for single-type models as well. We note that, if gene duplications were predominantly neutral (which is of course extremely unlikely), the estimated λ should roughly correspond to the per-gene duplicative mutation rate. For *Drosophila*, the duplicative mutation rate has been estimated at 1.25×10^{-7} duplications per gene per generation (Schrider et al. 2013). Taking this as a crude estimate of the order of magnitude of the duplicative mutation rate, and considering a long-term average generation time between 7 to 20 days, the expected duplication rate λ under neutrality would be *at least* a 100 times larger than the rates we estimated under the two-type DL model, suggesting that the vast majority of duplications is deleterious. A more detailed picture of the distribution of fitness effects of new duplications remains however elusive.

The two-type DL model may provide new quantitative insights in the long-term evolutionary dynamics of duplicated genes. In our analyses of comparative genomic data sets, we find that the loss rate per excess gene is much higher than the loss rate per base gene, suggesting that even when gene duplicates establish in a population, the selective pressures ensuring their mainte-

nance differ strongly from those maintaining typical single-copy genes. For instance, for the *Drosophila* genus, we estimated that on average half of the duplicated genes are maintained over a period of approximately 11 My and that about 3% of the duplicated genes may eventually become stably established (complete x-functionalization) so that their loss rates reflect those of base genes. The half-life of a typical base gene on the other hand is estimated at 371 My. Concomitantly, our results indicate that estimates of long-term gene loss rates based on simple BDP models (as in e.g. Hahn, Han, and Han (2007) for *Drosophila*) likely underestimate the actual loss rates of duplicated genes significantly, while estimates of duplication rates tend to be in rough agreement with those presented here.

What then causes gene duplicates that establish in a population to remain prone to higher loss rates? Or conversely, how are gene duplicates that are non-essential and prone to loss at relatively high rates established in the first place? Of course, completely redundant duplicates may drift to fixation, and it is unsurprising that such genes would be prone to higher loss rates than their essential counterparts (Walsh 2003). Many gene duplicates that rise to fixation may however be adaptive (Han et al. 2009; Innan and Kondrashov 2010; Kondrashov 2012), demanding an answer to the first question. As Kondrashov (2012) stressed, it may be misleading to think of gene duplicates as either completely redundant or non-redundant, as many adaptive duplications may establish as a consequence of positive selection for increased dosage in a stressful environment, despite being *qualitatively* redundant. In such cases, a changing environment or the emergence of other genetic variants may alter selection pressures over time, and while some duplicated gene may promote adaptation during some environmental challenge, it may return to a state of complete redundancy or even come at a fitness cost later (Kondrashov 2012). Our phylogenetic analyses may go some way supporting this view of long-term gene family evolution, where many of the duplicated genes in multi-copy gene families reside for some time in the genome, but eventually suffer loss before undergoing complete sub- or neofunctionalization.

4 Tree distributions and phylogenomic forests

In the previous chapters, we have modeled gene family evolution across the genome without explicit reference to the locus and gene trees introduced in chapter 1. We have conditioned our statistical analyses of the content of our 'bags of genes' on a known species tree, which provided the backbone along which our models of evolution were supposed to operate. In the following two chapters, trees will feature much more prominently, and in particular probability distributions over a set of trees will be our bread and butter. Indeed, the latter chapters share a *phylogenomic* theme, where the subfield of *phylogenomics* roughly refers to the study of genome evolution from the perspective of a genome-scale collection of phylogenetic trees. In the present interlude, we shall hence be a bit more formal in our treatment of phylogenetic trees and introduce some of the concepts which are central to our work in the rest of this dissertation.

4.1 Probability distributions on trees

A lot of phylogenetics concerns the inference of *evolutionary trees*, which come in various kinds (fig. 4.1). As before, we use *phylogeny* as a general term for any graph representation of evolutionary relationships among taxonomic units, which may be species, populations, (sub)genomes, loci, alleles or even human languages (we refer to all these, somewhat inaccurately, as *taxa* for the sake of brevity). By a *cladogram*, or rooted tree topology, we understand a leaf-labeled directed graph, with a designated root node which has indegree zero. A well known counting argument (see e.g. Felsenstein 2004) shows that there are

$$c_n = \prod_{\substack{1 \leq i \leq 2n-3 \\ i \text{ odd}}} i = \frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{4.1}$$

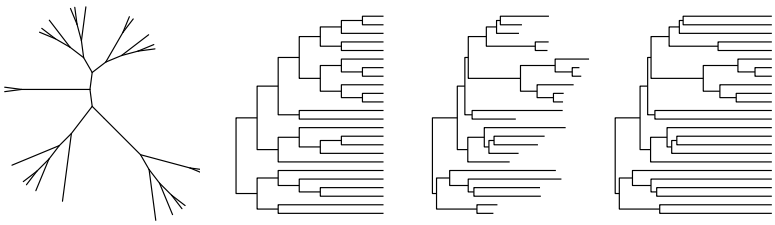


Figure 4.1: Different kinds of phylogenetic trees. From left to right: unrooted tree topology (displayed branch lengths are meaningless), rooted tree topology or cladogram (displayed branch lengths are meaningless), rooted phylogram, timetree.

distinct rooted tree topologies on n leaves. This number grows rather dramatically with n, with, for instance, the number of rooted tree topologies for 20 taxa already exceeding 8×10^{21}. An *unrooted* tree topology is an undirected graph with no cycles. The number of unrooted n-taxon trees can be easily obtained by noting that after designating one arbitrary taxon as root one obtains a rooted topology, and each of the rooted topologies on the $n - 1$ remaining taxa corresponds to a distinct unrooted tree on the original n-taxon set. A *phylogram* is a cladogram with a positive real number assigned to each edge, referred to as the branch's length. A *timetree* is a phylogram where branch lengths measure calendar time.

Probability distributions on phylogenetic trees are of interest for various reasons and arise in different ways. We may distinguish three main situations:

1. Uniform distributions on certain combinatorial classes of trees (e.g. unrooted trees, cladograms, ranked trees, *etc.* for a fixed set of leaves).
2. Tree distributions induced by stochastic processes such as branching processes, birth-death processes, coalescent processes or fragmentation/coagulation processes.
3. Distributions on trees arising in statistical inference settings, such as empirical measures derived from MCMC samplers or bootstrapping algorithms.

This is no clear-cut classification, but serves as an aid for further discussion. The first class is rather self-explanatory: having specified some combinatorial class, such as "all unrooted tree topologies on m leaves", we can assign an arbitrary probability measure on this (discrete) sample space.

In the second class of tree distributions, the stochastic process which induces the distribution typically serves as a parametric model of an evolutionary process of interest, as in the case of population genetic models like Kingman's n-

coalescent, macroevolutionary birth-death processes, or the duplication-loss models of gene family evolution we have dealt with in the previous chapter. Many of these models induce distributions on timetrees for some associated time scale. Clearly, any distribution on timetrees induces *a fortiori* a distribution on cladograms and unrooted trees. An important further distinction in these types of models arises from whether the model generates trees on a specified number of leaves, or not. The former typically correspond to evolutionary models which are specified *backwards* in time (e.g. coalescent processes) whereas the latter correspond to *forward-time* models (e.g. birth-death processes), although the two are not unrelated (see e.g. coalescent point processes; Lambert and Stadler 2013). Note that in the latter case a distribution on m-taxon trees can usually be obtained by *conditioning* on the event that m taxa are in fact generated by the forward-time process, however, this conditioning need not be analytically or computationally tractable.

The third class of tree distributions arises in a rather different setting. For instance, in classical Bayesian phylogenetic inference for continuous-time Markov chain (CTMC) models of sequence evolution, the goal is to determine the joint posterior distribution over tree topologies $\mathcal{T}$, branch lengths ϕ (expected number of substitutions per site) and substitution model parameters θ conditional on observed sequence data y. Virtually all Bayesian phylogenetic inference relies on MCMC methods to simulate approximately from the joint posterior distribution and to estimate posterior probabilities of phylogenies. Such an MCMC sample will constitute an *empirical tree distribution* (in the sense of an empirical (random) measure), which serves as an approximation to the actual posterior distribution over phylogenetic trees relating the different sequences in y. Bootstrap or jackknife methods in phylogenetic inference also give rise to tree distributions of this sort. Often these distributions arise in inference settings where one believes there is a true underlying tree topology 'around' which the tree sample is distributed.

Considering such empirical distributions over tree topologies on a fixed leaf set $\mathcal{L}$ of size m, we note that these can in general be represented by a categorical distribution with parameter $\{p_j : j = 1, \ldots, c_m\}$ (assuming a suitable indexing of the space of tree topologies has been fixed). Assuming we have an iid sample $(\mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(N)})$ from this distribution, the parameter of the categorical distribution over tree space can be estimated by the empirical measure, which uses the observed tree frequencies in the sample

$$\hat{p}_j = \frac{1}{N} \sum_{i=1}^{N} [\mathcal{T}^{(i)} = T_j] \qquad j = 1, \ldots, c_m \tag{4.2}$$

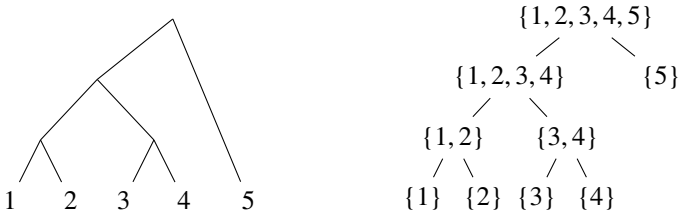


Figure 4.2: The correspondence between a rooted cladogram $\mathcal{T}$ (left) and the associated clade collection $C(\mathcal{T})$ (right), depicted as a Hasse diagram with respect to the inclusion ($\subseteq$) order.

However, this estimator is not always desirable as it may have large variance for small N and will likely assign probability zero[1] to a large portion of tree space. The issue is of course that for m sufficiently large, we are trying to estimate c_m parameters using a sample $N \ll c_m$. For continuous probability spaces, there are two commonly used strategies for estimating densities from samples: (1) approximation by simpler parametric densities (e.g. a Gaussian distribution) and (2) nonparametric density estimation, or *smoothing*, methods, such as histograms or kernel densities. A natural question is whether one can devise reasonable analogous strategies for distributions over tree topologies.

4.2 Markov branching models

Let $\mathcal{L}(v)$ denote the set of leaves in the subtree rooted in v for some internal node v of a rooted tree $\mathcal{T}$, and let $\mathcal{L} \stackrel{\text{def}}{=} \mathcal{L}(\text{root of } \mathcal{T})$. We define the *clade* associated with node v to be the set $\mathcal{L}(v)$. The rooted tree topology $\mathcal{T}$ on a set of leaves $\mathcal{L}$ of size m can be represented as a collection of clades $C(\mathcal{T}) = \{\mathcal{L}(v) \colon v \in V(\mathcal{T})\}$. The tree topology coincides with the usual Hasse diagram on $\mathcal{P}(\mathcal{L})$ (the powerset of $\mathcal{L}$) ordered by inclusion and restricted to $C(\mathcal{T})$ (fig. 4.2). We define the *split* of a clade γ to be the pair of daughter clades associated with γ, e.g. $(\{1, 2, 3, 4\}, \{5\})$ is the split of the clade $\mathcal{L}$ in fig. 4.2. Following the usual notation for sets, we let $|\gamma|$ refer to the size of clade γ, i.e. the number of leaves contained in it.

[1] From a Bayesian perspective, any estimator which assigns probability zero to some parts of the space under consideration is highly suspect. One should only assign probability zero to some event when one is absolutely certain that the event will not obtain, which usually means that the event entails some logical contradiction.

Algorithm 1 MBMRecursion(T, γ, q)

if $1 < |\gamma| \leq m$ **then**
  $X \sim q_{|\gamma|}(\cdot)$
  $\gamma_1 \sim \text{Uniform}(\{\delta \subset \gamma : |\delta| = X\})$
  $\gamma_2 \leftarrow \gamma - \gamma_1$
  $T \leftarrow T \cup \{\gamma_1, \gamma_2\}$
  $T \leftarrow \text{MBMRecursion}(T, \gamma_1, q)$
  $T \leftarrow \text{MBMRecursion}(T, \gamma_2, q)$
end if
return T

An interesting and useful parametric family of distributions over rooted tree topologies based on the correspondence between trees and clade collections has appeared repeatedly in the literature (e.g. Maddison and Slatkin 1991; Aldous 1996; Blum and François 2006; Jones 2011). Aldous (1996) refers to these as *Markov branching models* (MBMs). A MBM $(\mathcal{L}, \{q_n : 2 \leq n \leq m\})$ consists of a leaf set $\mathcal{L}$ where $|\mathcal{L}| = m$ and a collection of probability distributions $q_n(i)$ on $i \in [1..n-1]$ where $q_n(i) = q_n(n-i)$. A MBM generates a random cladogram by initiating the recursion in algorithm 1 with $T = \{\mathcal{L}\}$ and $\gamma = \mathcal{L}$. When the algorithm terminates, the clade collection T is isomorphic to a rooted cladogram in the sense of fig. 4.2.

Clearly the splitting process in a MBM is Markovian, as the split of a clade γ does not depend on its parent clade. Depending on the symmetric probability distribution we choose we will get different distributions on cladograms. In its most general form, the split size distributions q_n are arbitrary symmetric categorical distributions on $[1..n-1]$, so that a MBM on m leaves is specified by at most

$$\sum_{k=1}^{m/2} 2(k-1) = \frac{m(m-2)}{4} \qquad m \text{ even}$$

$$\sum_{k=1}^{(m-1)/2} 2(k-1) + \frac{m-1}{2} = \frac{(m-1)^2}{4} \qquad m \text{ odd}$$

parameters (i.e. the number of parameters in an arbitrary MBM grows as $O(m^2)$). Several well known distributions on cladograms can be formulated as MBMs with additional constraints on the split size distributions q_n. For instance:

1. The constant rates birth-death model (CRBD), where $q_n(i) = 1/(n-1)$ for all n. This is the distribution on cladograms with m leaves induced by birth-death processes with constant birth and death rates *conditioned* on generating m extant leaves. The so conditioned pure-birth process (Yule-Furry process) is a special case. Kingman's n-coalescent (see next chapter) induces the same distribution on cladograms.
2. The proportional to distinguishable arrangements (PDA) model, characterized by
$$q_n(i) = \frac{1}{2}\binom{n}{i}\frac{c_i c_{n-i}}{c_n}$$
where c_i is as in eq. 4.1. This distribution is notable because it gives rise to a uniform distribution on rooted cladograms.
3. The random partition tree model (RPM) (Maddison and Slatkin 1991), where
$$q_n(i) = \begin{cases} \binom{n}{i}(2^{n-1}-1)^{-1} \text{ if } n \text{ odd} \\ \binom{n}{i}(2^n-2)^{-1} \text{ else} \end{cases}$$
where the probability of a split of size i for clade γ of size n is proportional to the number of possible splits of size i of clade γ.

Note that while linear birth-death processes and Kingman's coalescent induce different distributions on timetrees, they both induce the CRBD distribution on cladograms. These models have been studied extensively in the context of empirical tree data sets and phylogenetic tree (im)balance. Further discussion and analysis in that context can be found in Aldous (1996), Blum and François (2006) and Jones (2011), among others.

The split distributions for the above MBMs are motivated by combinatorial considerations. Interestingly however, Aldous (1996) showed that all these models appear as special cases in a one-parameter family of MBMs, which he called the β-splitting model. Under the β-splitting model, the MBM split distribution is defined as

$$q_n(i) = \frac{1}{s(\beta, n)}\frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(m-i+1)} \tag{4.3}$$

Where $\beta \in [-2, \infty)$ and $s(\beta, n)$ is a normalizing constant. The β parameter determines the degree of balance[2] of a typical tree, with larger β resulting

[2] A tree is balanced when the subclades of any particular split tend to have similar sizes. We will not go into formal details on how to measure tree balance here and trust that the reader has an intuitive feel for this notion. We refer to Felsenstein (2004) and Jones (2011) for detailed treatments.

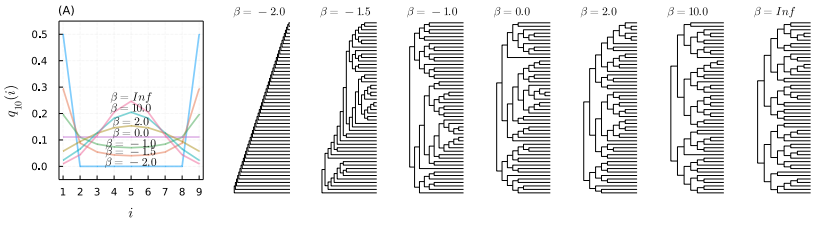


Figure 4.3: Aldous' β-splitting model. The subclade size distribution $q_{10}(i)$ for different values of β is shown in the leftmost plot (A). The trees show random realizations from the β-splitting model on 20 taxa for different values of β.

in more balanced trees (fig. 4.3). At the extreme end of $\beta = -2$, we obtain the uniform distribution on "comb" trees (i.e. a uniformly randomly labeled comb tree shape). The PDA model appears for $\beta = -3/2$, while the CRBD model appears for $\beta = 0$. At the other extreme we obtain the RPM model of Maddison and Slatkin (1991) as $\beta \to \infty$. Large-scale analyses of empirical trees (Blum and François 2006; Jones 2011) have shown that $\beta \approx -1$ provides a good fit, indicating that empirical trees are typically more balanced than the ubiquitously used uniform (PDA) model predicts, but less balanced than predictions for the similarly widespread CRBD model – a finding which has some repercussions for the choice of prior distribution in Bayesian phylogenetic analyses (Jones 2011). The special case $\beta = -1$ is known as Aldous' branching (AB) model.

4.3 Conditional clade distributions

There is a close connection between MBMs such as the β-splitting model and a smoothing method for empirical tree distributions proposed by Larget (2013). The latter method is based on what Larget termed the *conditional clade distribution* (CCD). Here we will take the liberty to generalize the definition of a CCD such that both MBMs and Larget's CCD appear as a special case. We return to Larget's construction, which we term the *empirical* CCD, below.

Instead of defining a MBM on a leaf set $\mathcal{L}$ using the set of symmetric probability distributions for *subclade sizes* $\{q_n : 1 < n \leq m\}$, we define a CCD using a collection of arbitrary categorical distributions $\theta = \{\theta_\gamma : \gamma \subseteq \mathcal{L},\ |\gamma| > 1\}$ on the *subclades* themselves. A CCD generates a cladogram similarly as a MBM, but with $\gamma_1 \sim \theta_\gamma$ in algorithm 1. In more detail, consider such a collection of *conditional split distributions* θ, and assume we have fixed a total order rela-

tion $<$ on $\mathcal{P}(\mathcal{L})$. Any non-leaf clade $\gamma \subseteq \mathcal{L}$ in a tree T will have a split $(\delta, \gamma - \delta)$ where $\delta \subset \gamma$ and $\delta < \gamma - \delta$. We formally identify a split of a given clade γ with its smallest subclade, in this case δ. Now let $\mathcal{T}$ be a random tree and $Y \subset \gamma$ be the random split of γ in $\mathcal{T}$. In the CCD family of distributions, we assume that the *conditional clade probability* $\mathbb{P}(Y = \delta|\gamma) = \theta_{\gamma,\delta}$, so that the distribution over splits of γ is categorical with parameter $\theta_\gamma = (\theta_{\gamma,1}, \ldots, \theta_{\gamma,\#_\gamma})$, where $\#_\gamma = 2^{|\gamma|-1} - 1$ is the number of possible splits of γ. Note that this definition of a CCD subsumes the MBM family *sensu* Aldous. Indeed, in the special case where θ_γ is a function which only depends on γ through $|\gamma|$, we will obtain a MBM as defined above. For instance, a β-splitting CCD will be defined by

$$\theta_{\gamma,\delta} = q_{|\gamma|}(|\delta|)\binom{|\gamma|}{|\delta|}^{-1} \qquad \text{where } \delta \subset \gamma,\ \gamma \subseteq \mathcal{L} \tag{4.4}$$

with $q_n(i)$ defined as a function of β as in eq. 4.3. Given a parameterized CCD $\mathcal{M}$, computing the likelihood $p_{\mathcal{M}}(\mathcal{T}|\theta)$ of a rooted tree topology $\mathcal{T}$ under the model is straightforward and can be done in $O(|\mathcal{L}|)$ time.

Note that the number of possible splits $\#_\gamma$ of a clade γ grows exponentially in the size of the clade, so that the number of parameters associated with an arbitrary CCD on m leaves will be

$$\sum_{k=2}^{m}\binom{m}{k}(2^{k-1}-2) = 2\sum_{k=0}^{m}\binom{m}{k+2}(2^k-1) = \frac{3^m - 2^{m+2} + 2m + 3}{2}$$

which is $O(3^m)$, a rather dramatic increase compared to a general MBM, but still a considerable reduction compared to an arbitrary categorical distribution on the space of possible tree topologies on m leaves (compare for instance 1.7×10^9 with the 8×10^{21} mentioned in the introduction for $m = 20$). This reduction is a consequence of the Markovian assumption: the split distribution of a clade γ does not depend on its parent or sister clades. In other words, the topology of the subtree on the set γ only depends on γ and not on the topology induced on $\mathcal{L} - \gamma$. The CCD family of distributions is hence characterized by the *conditional independence of disjoint subtrees*.

4.3.1 The empirical CCD

The smoothing method proposed by Larget (2013), and used in an elegant way by Szöllősi, Rosikiewicz, et al. (2013) in the context of phylogenomic forestry (see below), is based on this principle of conditional independence of

disjoint subtrees and leads to a particular CCD which we call the *empirical* CCD. Recall that in the smoothing problem, the goal is to, given some sample of tree topologies $(\mathcal{T}_1, \dots \mathcal{T}_N)$, construct a more adequate approximation of the distribution on tree topologies than the naive empirical measure eq. 4.2. Larget proposed to approximate the distribution of $\mathcal{T}$ by the CCD with split distributions estimated from the sample. Specifically, let n_γ be the number of times the clade γ is observed in the sample, and let $n_{\gamma,\delta}$ be the number of times the split δ is observed as a subclade of γ in the sample. The conditional clade probability $\theta_{\gamma,\delta}$ in the empirical CCD is then defined as

$$\theta_{\gamma,\delta} = \frac{n_{\gamma,\delta}}{n_\gamma} = \frac{n_{\gamma,\delta}}{\sum_{i=1}^{\#_\gamma} n_{\gamma,\delta_i}} =: \hat{p}_{\gamma,\delta}$$

Given a sample of N rooted tree topologies on a leaf set $\mathcal{L}$ of size m, a simple postorder traversal for each observed tree in the sample suffices to compute the split count $n_{\gamma,\delta}$, and hence the associated empirical CCD can be constructed in $O(mN)$ time. Clearly, if the counts are stored (and not merely the $\theta_{\gamma,\delta}$) the empirical CCD can be updated in a sequential fashion as more data comes in by simply updating the split counts. If the sample $(\mathcal{T}_1, \dots, \mathcal{T}_N)$ would be weighted in any way, so that $(w_1, \dots, w_N)$ is a vector of associated weights, the empirical CCD can obviously be generalized to take this into account simply by multiplying the contribution of $\mathcal{T}_i$ to the CCD by its respective weight w_i.

Szöllősi, Rosikiewicz, et al. (2013) proved that the empirical CCD *sensu* Larget (2013) corresponds to the maximum entropy probability distribution $p(\mathcal{T})$ over the space of cladograms for a fixed leaf set subject to the constraint that the marginal split frequencies induced by p correspond to the observed marginal split frequencies $n_{\gamma,\delta}/n_\gamma$ in the sample, which provides some motivation for its use as a smoothing distribution. That the empirical CCD satisfies this constraint is obvious of course, but that it should maximize the entropy over all discrete probability distributions over tree space is less so. We note that since the CCD forms an exponential family (see also below) with sufficient statistic $n_{\gamma,\delta}$, it must follow from the theory of exponential families that the empirical CCD is the maximum entropy CCD distribution subject to a constraint on marginal split frequencies. However, this does not entail that the empirical CCD is the maximum entropy distribution over all possible tree distributions (which are not necessarily in the CCD family) subject to this constraint. It is however not hard to show that the maximum entropy distribution subject to the constraint on observed marginal split frequencies is indeed a CCD, and hence the empirical CCD.

Consider a leaf set $\mathcal{L}$ of size m and fix any index set $i \in \{1, \ldots, c_m\}$ on the space of possible rooted tree topologies so that T_i is a well-defined tree topology and assume we have some sample of rooted tree topologies $(\mathcal{T}_1, \ldots, \mathcal{T}_N)$. Let $\{p_1, \ldots, p_{c_m}\}$ be the probability distribution we wish to construct based on the sample, where $p_i = \mathbb{P}\{\mathcal{T} = T_i\}$. The entropy of this probability distribution on tree topologies is

$$H = -\sum_{i=1}^{c_m} p_i \log p_i$$

Let $\hat{p}_{\gamma,\delta}$ be the observed frequency of the split (γ, δ) in the sample, and let $p_{\gamma,\delta}$ be the probability of this split induced by the tree distribution $\{p_i\}$. We now wish to find the probability distribution $\{p_i\}$ which maximizes the entropy subject to the constraint that $\sum_{i=1}^{c_m} p_i = 1$ and that

$$p_{\gamma,\delta} = \frac{\sum_{i=1}^{c_m} \mathbb{1}_{\gamma,\delta}(T_i) p_i}{\sum_{i=1}^{c_m} \mathbb{1}_{\gamma}(T_i) p_i} = \hat{p}_{\gamma,\delta} \quad \forall \gamma \subseteq \mathcal{L}, \delta \subset \gamma \tag{4.5}$$

The usual approach to solve such a constrained maximization problem is to construct the Lagrangian

$$\mathscr{L} = -\sum_i p_i \log p_i - \alpha\Big(\sum_i p_i - 1\Big) - \sum_{\gamma,\delta} \lambda_{\gamma,\delta}\Big(\sum_i \mathbb{1}_{\gamma,\delta}(T_i) p_i - \hat{p}_{\gamma,\delta}\Big)$$

where α and $\lambda_{\gamma,\delta}$ are Lagrange multipliers. We then solve $\partial\mathscr{L}/\partial p_i = 0$ for p_i to find that

$$p_i \propto \exp\Big(-\sum_{\gamma,\delta} \mathbb{1}_{\gamma,\delta}(T_i)\lambda_{\gamma,\delta}\Big) = \exp\Big(-\sum_{(\gamma,\delta)\in T_i} \lambda_{\gamma,\delta}\Big) = \prod_{(\gamma,\delta)\in T_i} \Phi_{\gamma,\delta}$$

So that the probability $p_i = \mathbb{P}\{\mathcal{T} = T_i\}$ under the maximum entropy distribution is the product of split-specific factors. Hence, the maximum entropy probability distribution must be a CCD, and the theory of exponential families implies then that it is the empirical CCD, i.e. that $\Phi_{\gamma,\delta} = p_{\gamma,\delta}$. The conclusion is that, if one is willing to derive a probability distribution over trees from a collection of observed split frequencies, then the obvious thing to do, i.e. constructing a CCD by matching the moments of the split distributions to the empirical split distributions, is indeed the right thing to do, in the sense that it represents the minimally informative distribution over trees subject to the stated constraints (see e.g. Jaynes 2003).

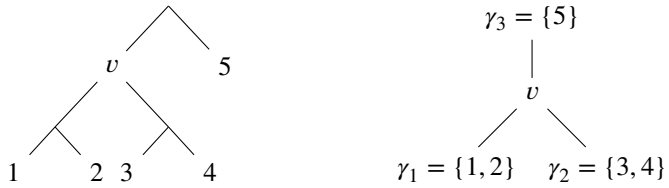


Figure 4.4: (Left) Unrooted tree topology $\mathcal{T}$, drawn as a pseudo-rooted tree. (Right) Clade neighborhood of internal node v in $\mathcal{T}$. Each clade is the outgroup clade for the pair consisting of the two remaining clades.

4.3.2 Empirical CCD for unrooted trees

The empirical CCD as discussed above is easily generalized to unrooted tree topologies by marginalizing over the different possible root positions. An m-taxon tree has $2m - 3$ possible root positions. A naive algorithm to construct an empirical CCD from a collection of unrooted tree topologies would iterate, for each unrooted tree $\mathcal{T}$, over the $i = 1, \ldots, 2m - 3$ possible rooted tree topologies $\tilde{\mathcal{T}}_i$, updating the observed splits with $\tilde{\mathcal{T}}_i$ as in the rooted case but with weight $1/(2m-3)$ (or any other weight if we have some informative prior probability distribution over the possible root positions). For large trees and large samples this induces however a considerable computational overhead. A more efficient algorithm is possible with only a single pass over the unrooted tree. Note that an unrooted tree is not typically represented as an undirected graph data structure but rather as a *pseudo-rooted tree*, i.e. an arbitrarily rooted tree (fig. 4.4). Algorithm 2 will correctly update the (weighted) marginal split counts $\{n_{\gamma,\delta}\}$ of an empirical CCD given some observed pseudo-rooted tree topology T. The weight w_{ij} added at line 8 is the proportion of possible rootings of T in which the split (γ_i, γ_j) appears in the associated rooted tree. The weight $1/(2m - 3)$ added to $n_{\mathcal{L},\gamma_k}$ at line 9 accounts for the fact that any clade in the unrooted tree is a possible subclade of the root clade $\mathcal{L}$ (i.e. the leaf set).

4.3.3 Problems with the empirical CCD

Two potential problems with the empirical CCD should be noted: (1) the assumption of conditional independence of disjoint subtrees may be too stringent and may not provide a reasonable fit to empirical tree distributions, and (2) while covering a much larger portion of tree space than the empirical mea-

Algorithm 2 AddUnrootedTree(T, $\{n_{\gamma,\delta}\}$)

for each non-root internal node v in $V(T)$ **do**
  Identify the clade neighborhood $(\gamma_1, \gamma_2, \gamma_3)$ at v (see fig. 4.4).
  for each clade pair (γ_i, γ_j) in the clade neighborhood **do**
    $\gamma \leftarrow \gamma_i \cup \gamma_j$
    $\delta \leftarrow \min(\gamma_i, \gamma_j)$
    $\gamma_k \leftarrow \mathcal{L} - \gamma$
    $w_{ij} \leftarrow (2|\gamma_k| - 1)/(2m - 3)$
    $n_{\gamma,\delta} \leftarrow n_{\gamma,\delta} + w_{ij}$
    $n_{\mathcal{L},\gamma_k} \leftarrow n_{\mathcal{L},\gamma_k} + 1/(2m - 3)$
  end for
end for

sure, the empirical CCD still assigns zero probability mass on a large part of tree space. The first issue is of course an issue of the CCD family in general, not restricted to the empirical CCD, whereas the second is concerned specifically with the empirical CCD.

The first issue is the main motivation for the work on *subsplit Bayesian networks* (SBNs) of Cheng Zhang and Matsen IV (2018a) (also Cheng Zhang and Matsen IV (2018b)). SBNs are similar to CCDs but allow for more complicated conditional dependence structures, resulting in a more flexible parametric approximation of an arbitrary distribution on cladograms. In Cheng Zhang and Matsen IV (2018a), the authors showed that the SBN better approximates empirical tree distributions than the CCD by comparing the CCD and SBN approximations based on relatively small MCMC samples of trees against the empirical measure obtained from large 'gold standard' MCMC samples from the same target distribution. The conclusion is that empirical tree distributions arising for instance in Bayesian phylogenetic inference using MCMC tend to violate the principle of conditional independence at least to some degree, so that SBNs provide an improvement over the empirical CCD. SBNs were used by Cheng Zhang and Matsen IV (2018b) as variational distributions on cladograms in the context of variational Bayesian (VB) phylogenetic inference, a promising alternative for the computationally intensive MCMC methods that dominate Bayesian phylogenetic inference. We will return to the use of parametric distributions on cladograms for variational inference in the next chapter, albeit in a rather different context than the latter authors.

The second issue pertains to the undesirable situation already mentioned above, namely that, at least from a Bayesian standpoint, it is undesirable to

assign probability zero to outcomes which are not considered impossible (when the prior assigns non-zero probability to some set in the parameter or model space, and the data does not logically contradict this set, the posterior probability of that set should be positive). If one is willing to make some parametric assumptions, treating the tree smoothing problem in a more explicitly Bayesian way provides a solution for these issues.

4.3.4 Bayesian estimation of CCDs

The CCD, as a collection of categorical distributions, forms an exponential family, and hence has many nice mathematical properties, such as the existence of a conjugate family. If we consider the θ_γ unknown, and model them as Dirichlet distributed random vectors, we obtain a conjugate prior distribution for a CCD which can be characterized as a collection of Dirichlet densities $\{a_\gamma : \gamma \subseteq \mathcal{L}\, |\gamma| > 1\}$ with parameters $a_\gamma = (a_1, a_2, \ldots, a_{\#_\gamma})$. We refer to this distribution as a *Dirichlet-CCD*. Note that a realization from a Dirichlet-CCD on the leaf set $\mathcal{L}$ is a CCD on $\mathcal{L}$. A particular class of Dirichlet-CCD distributions we will be using is the one for which the mean is a β-splitting CCD. Specifically, we define for $\beta \geq -2$ and $\alpha \geq 0$ a β-Dirichlet-CCD so that for each clade γ the Dirichlet density over split distributions has the parameter $a_\gamma = (\alpha\theta_{\gamma,1}, \ldots, \alpha\theta_{\gamma,\#_\gamma})$ where $\theta_{\gamma,\delta}$ is defined as in eq. 4.4. When used as a prior in a Bayesian context, α can be interpreted as a pseudocount, the information embodied by the prior being equivalent to α observations from the associated β-splitting model.

Consider now again the problem of learning a CCD from a sample of tree topologies. The existence of a conjugate family allows replacing the empirical CCD by a Bayes estimator if we are willing to make a parametric prior assumption. Assuming a β-Dirichlet-CCD prior, entailing the prior mean is a β-splitting model, the posterior distribution will again be a Dirichlet-CCD, now with parameters $a_\gamma = (\alpha\theta_{\gamma,1} + n_{1,\gamma}, \ldots, \alpha\theta_{\gamma,\#_\gamma} + n_{\gamma,\#_\gamma})$. The marginal posterior probability of the split Y_γ of clade γ in a random tree being δ is then

$$\theta'_{\gamma,\delta} = \mathbb{P}(Y = \delta|\gamma, \alpha, \beta, y) = \frac{\alpha\theta_{\gamma,\delta} + n_{\gamma,\delta}}{\sum_{i=1}^{\#_\gamma} \alpha\theta_{\gamma,i} + n_{\gamma,i}} \tag{4.6}$$

A Bayes estimator for the CCD after observing the data set y can be easily derived from the posterior as the CCD with conditional split distributions given by eq. 4.6. Note that the empirical CCD is obtained as the special case where

$\alpha = 0$, and that whereas the support of the empirical CCD is restricted by the observed clades in y, the support in the case where $\alpha > 0$ will be the full tree space. Crucially however, for a finite sample the $n_\gamma = (n_{\gamma,1}, \ldots n_{\gamma,\#_\gamma})$ will typically be sparse for most γ, so that for computations we need not represent the huge number of parameters that define such a CCD in general.

4.4 Phylogenomic forests

Before embarking on specific applications of the CCD in Bayesian statistical modeling of genome evolution, we discuss briefly the general role these distributions could play in phylogenomic modeling. In discrete mathematics and machine learning (and ecology, of course), researchers sometimes refer to a collection of trees as a *forest*, a notable example being the random forest in machine learning. What we would propose to call a *phylogenomic forest* can be defined in the following way

$$\Psi_1, \ldots, \Psi_n | \theta \sim_{\text{iid}} f_1(\cdot|\theta)$$
$$\mathcal{T}_i^{(1)}, \ldots, \mathcal{T}_i^{(N)} | \Psi_i, \psi_i \sim_{\text{iid}} f_2(\cdot|\Psi_i, \psi_i)$$

Where f_1 will be a tree distribution of the second class above (i.e. a tree distribution induced by some evolutionary model) and f_2 a distribution of the third class (an empirical tree distribution).

This is perhaps best illustrated by an example, studied in detail in a later chapter. Consider the multispecies coalescent (MSC) model (Hudson 1983; Pamilo and Nei 1988) and let $\theta = (S, \phi)$ where S is a species tree topology and ϕ a vector of effective population sizes for the branches of S, so that $f_1(\cdot|\theta)$ is the MSC distribution over gene tree topologies (Degnan and Salter 2005). The collection $(\Psi_1, \ldots, \Psi_n)$ is then the set of gene trees for n loci. Now suppose we have inferred gene tree topologies from sequence data for each locus using a standard Bayesian phylogenetics program, obtaining a sample of tree topologies $(\mathcal{T}_i^{(1)}, \ldots, \mathcal{T}_i^{(N)})$ for locus i. We assume this sample to be an iid sample from *some* distribution over tree topologies which depends in some way on the true tree topology Ψ_i and some other parameter ψ_i. Note that we are not specifying a full hierarchical Bayesian model for the tree topologies and sequence data here, but rather assume (1) a model for the gene tree topology associated with each locus and (2) that we are given a sample of tree topologies for each locus *obtained by other means*, which may even be unknown to us.

The reason for introducing this idea formally is to stress that this is *one point of view* one could adopt when seeking to learn about genome evolution from genome sequence data using Bayesian methods. It is however not a very commonly adopted point of view, with Bayesian approaches in phylogenomic modeling usually explicitly including the sequence data in a hierarchical model of the form

$$\Psi_1, \ldots, \Psi_n | \theta \sim_{\text{iid}} f_1(\cdot|\theta)$$
$$X_i^{(1)}, \ldots, X_i^{(M)} | \Psi_i, \eta_i \sim_{\text{iid}} f_2(\cdot|\Psi_i, \eta_i)$$

Where $X_i^{(j)}$ denotes column j of the multiple sequence alignment (of length M) for gene family i, and f_2 is a typical phylogenetic CTMC (Höhna et al. 2016). In the MSC example this is the approach taken by, for instance, Rannala and Yang (2003) and Heled and Drummond (2009), as we shall discuss in more detail in the next chapter. Of course, the reason this point of view is adopted is that the X_i are supposed to be the 'observed' data, and the only source of information about the Ψ_i. This is a rather strong assumption, since it is by no means clear how the X_i could be observed, with each $X_i^{(j)}$ in fact being the result of a complicated bioinformatic pipeline assumed to unveil homology among individual residues. So even if the process of sequence evolution would be adequately modeled by an iid CTMC across individual sites (itself a strong assumption), it remains to be seen if one can actually ascertain the supposedly observed sites. It is only with a great deal of willful ignorance that we can call a multiple sequence alignment 'observed data'.

One can take the hierarchical model one step further by including sequence alignment in the probabilistic model (Redelings and Suchard 2005), but besides the tremendous computational challenges associated with such an approach, many conceptual difficulties remain. In particular, the inference of which sequences to align (i.e. gene-level homology inference) remains out of the model's scope, and we are still required to assume that the atomization of an often biologically functional DNA sequence into individual sites makes evolutionary sense. While the more integrative hierarchical models that are called for are of course highly welcome (Szöllősi et al. 2015), it appears nevertheless to be a kind of "*turtles all the way down*" situation. If taken seriously, the Bayesian maxim that we should take into account all the available information (and theory) in devising a joint probability model for the data and unknowns would never allow us to take off, let alone carry through Bayes' theorem and actually compute a posterior in reasonable time.

Nevertheless, in phylogenomic modeling, pragmatic considerations force us to start somewhere, and we would contend that, for certain purposes, it *may* be as reasonable to start a Bayesian analysis for a genome-level evolutionary model with empirical distributions of locus-specific tree topologies (computed using standard phylogenetic tools for tree inference under CTMC models) as to start from the associated multiple sequence alignments, although the latter analysis could subsume the former. More specifically, inference of θ in a phylogenomic forest model, while conceptually an approximation of the sequence alignment model, could be very useful, even if only for the 'division of labor' it implies computationally. In the next two chapters, we will consider Bayesian inference for models of genome evolution from this perspective.

5 Likelihood-free Bayesian inference for the multispecies coalescent

An important hierarchical phylogenomic model that has received considerable attention is the multispecies coalescent (MSC). The MSC model is a *retrospective* model for the genealogy G (*gene tree*) of a *single* locus within a species tree or population tree S (see chapter 1). The main reason for sustained interest in the MSC model is that it provides a fairly tractable model for some of the population-level processes that cause variation in gene genealogies (but not family sizes) across the genome, more specifically the phenomenon of *incomplete lineage sorting* (ILS), or *deep coalescence* (chapter 1), and allows statistical inference of species trees from molecular sequence data while accounting for this variation.

The basic hierarchical model for sequence data $(X_1, \ldots, X_n)$ from n loci, conditional on a species tree S with branch parameters ϕ (see below) can be written as

$$\begin{aligned} G_i|\mathcal{L}_i, S, \phi &\sim_{\text{iid}} \text{MSC}_{\mathcal{L}_i}(S, \phi) \\ X_i|G_i, \theta_i &\sim \text{PhyloCTMC}(G_i, \theta_i) \end{aligned} \tag{5.1}$$

for $i = 1, \ldots, n$. Here $\mathcal{L}_i$ denotes the leaf set for locus i, and we assume the existence of a map $\sigma\colon \mathcal{L}_i \to \mathcal{L}(S)$ associating with each gene the species or population from which it was sampled. The hierarchy involves a genome and gene level model, assuming the locus tree to be identical to the species tree, with the pair (S, ϕ) constituting the genome-level parameter. The MSC model is more formally defined below, but we note already that a crucial, and rather controversial, assumption underlying this hierarchical model is that there is no recombination within a locus (so that there exists a single genealogy for each locus), and free recombination among loci (which is embodied by the assumption of independence in eq. 5.1). Importantly, as stated above, the model

is a retrospective one, and conditions on having observed a particular sample of n loci with leaf sets $\mathcal{L}_i$ (see below). It is a generative model for a collection of *genealogies* relating a fixed set of extant genes, and not a generative model for a collection of genomes in the sense that for instance a birth-death process model of gene family evolution is.

Although a wealth of methods has been developed for statistical inference of species trees (and associated demographic parameters) under the MSC model, considerable statistical and computational challenges persist and continue to invite new approaches. Existing popular methods range from full-Bayesian joint inference of species trees and gene genealogies (Heled and Drummond 2009; Rannala and Yang 2017) to fast heuristic methods based on summary statistics of the (assumed known) empirical gene tree distribution (Liu, Yu, and Edwards 2010; Chao Zhang et al. 2018; Zhang et al. 2020) (reviewed in e.g. Xu and Yang (2016), Rannala et al. (2020), Mirarab, Nakhleh, and Warnow (2021)). In this chapter we employ the theory of conditional clade distributions (CCDs) developed in chapter 4 to devise likelihood-free Bayesian inference methods for the MSC model. Before motivating our interest in such methods, a brief introduction to the MSC model will be presented.

5.1 Gene genealogies and the MSC

5.1.1 The Wright-Fisher model and Kingman's n-coalescent

The MSC model is derived from the widely used coalescent models of molecular population genetics, in particular Kingman's n-coalescent model. Kingman's n-coalescent arises in turn from considering the forward-in-time neutral Wright-Fisher (WF) population model retrospectively.

The WF model, named after Wright (1931) and Fisher (1930) is arguably the simplest nontrivial population genetic model for finite populations. Under the WF model, we assume (1) a haploid population of constant size N, (2) discrete non-overlapping generations, (3) no selection, (4) panmyxis, (5) no recombination. In the strictly neutral model, a diploid WF population of N individuals may be modeled by a haploid one of $2N$ individuals, reflecting the fact that under the assumptions stated, Mendelian inheritance ensures that we may treat a diploid population as if it were obtained by random sampling from a *gene pool* without loss of essential information. Note that this device fails for higher ploidy levels, as biparental inheritance in, for instance, a tetraploid population evolving under WF assumptions does not amount to an independent

sampling of four gene copies from a gene pool (but see Arnold, Bomblies, and Wakeley (2012)).

With these assumptions, the WF model further postulates that the number of descendants of each (haploid) individual is distributed according to a multinomial distribution with N cells and cell probabilities $1/N$. If we track the evolution of a particular allele which is currently at frequency $X_0/N = i/N$ in the WF population, we obtain a Markov chain X_t on the state space $[0..N]$, where $X_t = N$ indicates fixation of the allele by generation t and $X_t = 0$ indicates loss from the population by generation t. Under the WF model, the transition probability for X_t is given by a Binomial law:

$$\mathbb{P}\{X_{t+1} = j | X_t = i\} = \binom{N}{i} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j} \qquad 0 \leq j, i \leq N$$

The process X_t is a discrete time martingale, and it can be seen by considering the symmetry of the model that the probability that a gene at frequency p will eventually fix in the population is exactly p. Clearly, this is the ultimate 'bean-bag genetics' model (Mayr 1959; Haldane 1964), the genetic composition of the offspring generation being determined by randomly sampling genes with replacement from the current generation. Note that it is customary in population genetics to define the *effective population size* N_e of some population of interest as the population size of a WF population with similar statistical properties as the population under consideration[1]. We will use N throughout and understand it to refer to the effective population size in this sense.

Considering this process now backwards in time, we see that the probability that two genes in the extant population find their most recent common ancestor (MRCA) t generations in the past has a geometric distribution

$$\mathbb{P}\{T_c = t\} = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}$$

We say the two lineages *coalesce*, or merge into one lineage, at T_c. From this we make the important observation that the expected coalescence time of two lineages is N (or $2N$ in a diploid population). If we consider now a sample of k genes, the probability that there is no coalescence event among any of the k genes in the previous generation is (Hein, Schierup, and Wiuf 2004; Kingman

[1] Here we gloss over the many subtleties associated with the concept of effective population size and the various ways of defining it. Walsh and Lynch (2018) (chapter 3) provide an informative discussion and many relevant references.

1982a)

$$\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{k+1}{N}\right)=1-\binom{k}{2}\frac{1}{N}+\mathcal{O}(N^{-2})$$

Assuming that $k \ll N$, so that $\mathcal{O}(N^{-2})$ is negligible and we may ignore the possibility of multiple coalescence events in a single generation, we see that the probability for k lineages to coalesce into $k-1$ lineages t generations in the past is approximately

$$\mathbb{P}\{T_k = t\} \approx \left(1-\binom{k}{2}\frac{1}{N}\right)^{t-1}\binom{k}{2}\frac{1}{N}$$

Assuming an initial sample of $n \ll N$ genes, this model describes a discrete-time ancestral process which consists of a series of $n-1$ coalescence events with expected waiting times $N/\binom{n}{2}, N/\binom{n-1}{2}, \ldots, N$.

One way to derive Kingman's n-coalescent, after Kingman (1982a) and Kingman (1982b), is to rescale time in the discrete-time ancestral process just described to a continuous time scale so that one time unit is equivalent to N (or $2N$ in the diploid case) generations. The continuous-time ancestral process thus obtained is characterized similarly by a series of $n-1$ coalescence events, where the waiting time distributions are given by

$$\mathbb{P}\{T_k \leq t\} = 1 - e^{-\frac{k(k-1)}{2}t}$$

i.e. $T_k \sim \text{Exponential}(k(k-1)/2)$. This characterizes Kingman's n-coalescent in terms of waiting times between consecutive coalescence events: we start with n lineages, after an exponentially distributed waiting time T_n, two lineages out of the n, sampled uniformly without replacement, coalesce and the process starts anew with $n-1$ lineages. This proceeds until all lineages have coalesced into the MRCA of the sample under consideration. It should be noted that Kingman's n-coalescent arises from many different models, notably the Cannings model (which includes the WF and Moran models as special cases) where the distribution over the number of offspring of each of the N individuals is only required to be exchangeable (Kingman 1982a; Etheridge 2011).

Note that Kingman's n-coalescent defines a continuous-time Markov chain Y_t on the space of partitions of $[1..n]$: we start with the maximally fine partition $Y_0 = \{\{1\},\{2\},\ldots,\{n\}\}$ with n elements and after an exponentially distributed time t a new, coarser, partition is obtained by replacing two uni-

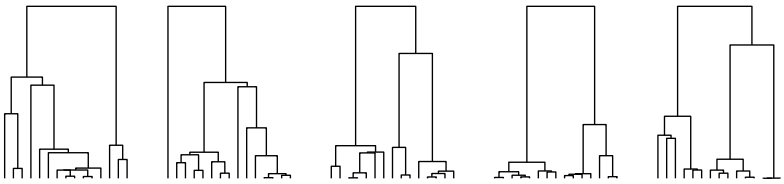


Figure 5.1: Genealogies for five independent random realizations of Kingman's coalescent on a sample of 15 genes. Note the expected behavior where many coalescence events occur in short succession near the leaves of the tree while much longer waiting times tend to be observed towards the root. Note also the rather balanced tree shapes that arise from the standard coalescent (the distribution over cladograms coincides with the β-splitting model with $\beta = 0$, see chapter 4).

formly randomly drawn elements by their union. The process can thus be decomposed in a *jump process* and *waiting time process* (Hein, Schierup, and Wiuf 2004), and the two processes conceived together induce a distribution on timetrees (on a timescale of N generations) with n leaves, parameterized by a single parameter: the effective population size N or coalescence *rate* N^{-1}. From this description it is clear that the simulation of genealogies and associated statistics under Kingman's coalescent can be done in an extremely efficient manner, and fig. 5.1 shows the obligate example of a bunch of gene genealogies simulated under the coalescent model. The relative efficiency of simulation, compared to evaluation, of the sampling distribution under Kingman's coalescent has lead to considerable interest in simulation-based statistical inference in molecular population genetics and is associated with the origin of the so-called approximate Bayesian computation methods (Tavaré et al. (1997); Sisson, Fan, and Beaumont (2018); see also Appendix A).

5.1.2 The multispecies coalescent model

The MSC model, as commonly understood, arises from a *structured* coalescent process and some additional assumptions about recombination between and within loci. With 'structured coalescent', we refer loosely to a situation where a coalescent process is defined for a set of WF populations (generally of different sizes), derived from some common ancestral population. The different populations are assumed to be geographically or reproductively isolated from each other so that two genes can coalesce only if they are in the same population. Of course, one can generalize this to allow for more complicated demographic scenarios, involving population expansion and contraction, dif-

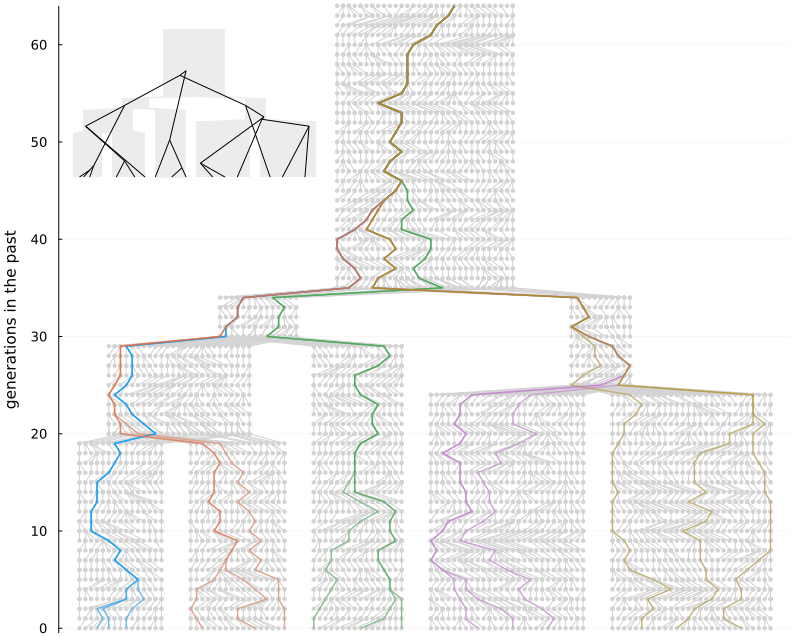


Figure 5.2: Example simulation for a structured coalescent and associated Wright-Fisher population model on which the standard multispecies coalescent model is based. Every circle represents an individual gene while a line connecting two genes represents an ancestor – descendant relation. Different reproductively isolated populations (with different population sizes) are shown separated from each other. Note that this is a simulation for the discrete-time ancestral process of which the standard MSC model is the continuous-time variant. The colored lines trace back the ancestry of three sampled genes per extant population. The inset figure in the upper left corner highlights the induced gene tree topology for the sample, which clearly manifests incomplete lineage sorting (ILS) with respect to the population/species tree.

Algorithm 3 MSCGeneTree($S, \phi, \mathcal{L}_G, \sigma$)

Require:
species tree S,
branch parameters $\phi = \{\phi_u : u \in V(S)\}$,
sampled genes $\mathcal{L}_G$,
gene-to-species map $\sigma : \mathcal{L}_G \to \mathcal{L}(S)$
Ensure:
for each node u in a postorder traversal of S **do**
    if u is a leaf node **then**
        $g(u) \leftarrow \{\{x\} \in \mathcal{L}_G : \sigma(x) = u\}$
    else
        $g(u) \leftarrow g'(v) \cup g'(w)$, where v and w are the child nodes of u.
    end if
    $g'(u) \leftarrow$ CensoredCoalescent$(g(u), e^{\phi_u})$
end for

ferent patterns of migration between different populations, hybridization, introgression, polyploidization *etc.*, and inference for such models is an active topic in population genetics (Beerli and Felsenstein 2001; Gutenkunst et al. 2009; Hey 2010; Roux and Pannell 2015; Excoffier et al. 2021), phylogeography (Nielsen and Beaumont 2009; Drummond et al. 2012) and phylogenetics (Wen, Yu, and Nakhleh 2016; Flouri et al. 2020).

In the present chapter we focus on the standard MSC model, denoted MSC(S, ϕ), where we assume a collection of WF populations related by a strictly bifurcating *species tree* S with leaf set $\mathcal{L}(S)$. A simulation of such a collection of WF populations evolving down a species tree is displayed in fig. 5.2. The standard MSC model is obtained as the continuous-time ancestral process associated with such a collection of WF populations. We express the branch lengths ϕ of the species tree under the MSC model in log-coalescent time units (i.e. the length of the time interval spanned by the branch in units of N generations, where N is the effective population size of that particular branch). Under the standard MSC model, every branch of the species tree is associated with a *censored* Kingman n-coalescent (Rannala and Yang 2003), i.e. a coalescent process which starts at the target vertex of a given branch and is terminated at the source vertex after a time given by the associated branch length (algorithm 4). The k uncoalesced lineages then enter the population associated with the parent branch together with l uncoalesced lineages of the sister population and the process starts anew,

Algorithm 4 CENSOREDCOALESCENT(A, t),

Require:
sample of genes $A = \{A_1, \ldots A_n\}$,
branch length t in N_e time units
Ensure:
$r \sim \text{Exponential}(n(n-1)/2)$
$t \leftarrow t - r$
while $t > 0$ **do**
    Draw two random elements, A_i and A_j, without replacement from A
    $A \leftarrow (A - \{A_i\} - \{A_j\}) \cup \{A_i \cup A_j\}$
    $n \leftarrow n - 1$
    $r \sim \text{Exponential}(n(n-1)/2)$
    $t \leftarrow t - r$
end while
return A

now as a $k + l$ coalescent (algorithm 3). The branch leading to the root is taken to be of infinite length, so that the genealogical process associated with it is the standard Kingman coalescent and terminates at the MRCA of the multispecies sample.

In the present study, we shall not be interested in branch lengths of gene trees under the MSC, focusing only on gene tree topologies G, the sampling distribution of which has been studied in detail in e.g. Degnan and Salter (2005). Assuming a species tree S with branch lengths ϕ and a set of extant genes $\mathcal{L}_G$ sampled from species of S, with the map $\sigma\colon \mathcal{L}_G \to \mathcal{L}(S)$ assigning to each gene the species from which it was sampled, the generative model for a gene tree topology G on $\mathcal{L}_G$ under a MSC(S, ϕ) process is defined according to algorithms 3 and 4. The process induces a sequence of partitions of the gene set which is isomorphic to the gene tree topology, which can be obtained by some additional bookkeeping in algorithms 3 and 4. If we want to refer to the MSC induced distribution on gene trees for a specific set of sampled leaves $\mathcal{L}_G$, we also write $\text{MSC}_{\mathcal{L}_G}(S, \phi)$ (see also eq. 5.1). Finally, note that when only a single gene is sampled from each species (as is often the case in phylogenetic studies), no coalescence events can occur in the terminal branches of S, and concomitantly, gene trees for such samples do not provide information about the branch length of the terminal branches.

5.2 Inference for the MSC model

5.2.1 Overview of existing approaches

As already alluded to above, inference of species trees under the MSC has been one of the most active areas of methodological development in phylogenetics throughout the past decade, and many different methods exist. The available and actively used approaches fall roughly in three camps: (1) joint Bayesian inference of gene trees and species trees, (2) summary methods which assume gene trees as input data and (3) summary methods based on site patterns (Xu and Yang 2016; Bryant and Hahn 2020; Mirarab, Nakhleh, and Warnow 2021).[2]

The first type of methods are theoretically straightforward. After completing the hierarchical evolutionary model in eq. 5.1 by specifying a prior distribution for S, ϕ and the θ_i (note that this may itself involve a hierarchical model), statistical inference of both the species tree and gene trees amounts, in theory, to no more than a straightforward application of Bayesian logic. Specifically, we arrive at a marginal posterior distribution for S, ϕ which has the form

$$p(S, \phi|X) \propto p(S, \phi) \prod_{i=1}^{n} \int_{G_i} \left(\int_{\theta_i} p(X_i|G_i, \theta_i) p(\theta_i) d\theta_i \right) p(G_i|S, \phi) d\theta_i$$

Where G_i is here taken to represent the gene genealogy (i.e. a timetree) for locus i. Computationally however, the problem is extremely challenging due to the awkward geometry of the posterior distribution. Engineering efficient MCMC proposal kernels, which take into account the correlation between the species tree and gene tree topologies among other issues, appears to be very challenging (Rannala and Yang 2017), and MCMC chains are known to show very poor mixing for these problems. Implementations of the joint Bayesian inference approach include BPP (Rannala and Yang 2003; Flouri et al. 2018), StarBeast (Heled and Drummond 2009; Ogilvie, Bouckaert, and Drummond 2017) and RevBayes (Höhna et al. 2016).

The main advantage of these approaches is that they are statistically efficient,

[2] We are here glossing over several approaches which do not fall readily in any of these classes. In particular, SNAPP (Bryant et al. 2012) and SNAPPER (Stoltz et al. 2021) enable Bayesian inference of species trees from unlinked biallelic markers without explicitly dealing with gene trees. We also note that PoMo (De Maio, Schrempf, and Kosiol 2015; Borges et al. 2021), although technically not assuming the MSC model, is another approach for conducting inference of species trees from multi-locus data sets in the presence of ILS.

using most of the available information in a single inferential procedure, carrying uncertainty through the complete hierarchical model under the guarantees of Bayesian coherence[3]. In addition, while many of the summary methods (see below) focus only on species tree inference, the full Bayesian approach provides estimates of gene tree topologies, substitution rates, ancestral effective population sizes, and all other parameters one manages to include in the hierarchical evolutionary model. In particular with respect to gene trees this is interesting, as in the joint estimation setting the species tree provides information about the gene trees (and vice versa), and the latter are often challenging to reliably estimate (Szöllősi et al. 2015). Note that this is a double-edged sword, since to the extent that the assumptions of the MSC model are violated, the inferred gene trees will be more or less biased. A flipside of the comprehensive nature of joint inference methods, besides the computational issues, is that we *have* to take everything on board: we have to devise a prior for a dated species tree, for the molecular clock, for substitution rate heterogeneity across and within loci, *etc.* Of course, this is not a bug but a feature, as it forces us to be explicit about the manifold assumptions at stake. However, in combination with the computational demands of these models, the complexity of the whole approach may become a serious impediment to a smooth 'Bayesian workflow' (Gelman et al. 2020).

The second class of methods, referred to as summary methods (or sometimes 'shortcut methods'; Springer and Gatesy (2016)), rely on standard phylogenetic tools to infer gene tree topologies, and assume these topologies *as data* for inference under the MSC. The likelihood of a gene tree *topology* (i.e. an *undated* cladogram) under the MSC model is however intractable, requiring the enumeration of all 'coalescent histories' compatible with a given gene tree (Degnan and Salter 2005; Wu 2012, 2016), so that these approaches rely on approximate inference methods based on summary statistics such as rooted triple (Liu, Yu, and Edwards 2010) or quartet frequencies (Mirarab et al. 2014) (but see Wu (2012) and Wu (2016) for an implementation of maximum likelihood

[3]It should be noted here that while the joint Bayesian inference approach indeed allows for a much more adequate quantification of uncertainty in parameter and model estimates than other methods considered here, the extent to which uncertainty is taken into account should not be exaggerated. As we have noted before, many data processing steps separate the 'raw' data from the input data for these methods (a set of aligned sequences) which are not taken up in the Bayesian hierarchical model (think of base calling, sequence assembly, gene annotation, orthogroup inference, multiple sequence alignment, *etc.*). Also, Bayesian theoretical guarantees are only as nice as our posterior approximations are accurate, so computational issues can not entirely be separated from these considerations. "[…] what's the motivation for modeling everything probabilistically? Sure, it's coherent – but so is some mental patient who thinks he's Napoleon and acts daily according to that belief." – Andrew Gelman

species tree inference from gene tree topologies using the exact MSC likelihood). The unifying idea is to compare the observed frequencies of quartets or triplets with the expected frequencies under the MSC model and search for the species tree which fits the observed frequencies best. While statistically less efficient than a joint Bayesian inference approach, many of these methods can be proven to give statistically consistent results under the MSC model, provided they are applied using accurate gene trees as input data. ASTRAL (Mirarab et al. 2014; Chao Zhang et al. 2018; Zhang et al. 2020) is arguably the most popular summary method to date, and has been used in many recent large-scale phylogenomics studies (e.g. Leebens-Mack et al. 2019; Jarvis et al. 2014; Prum et al. 2015). ASTRAL uses a dynamic programming algorithm with some additional heuristics to identify the species tree which maximizes the number of quartets shared by an input set of unrooted gene trees, a strategy which, provided the gene trees are correct, results in a statistically consistent estimator of the species tree under the MSC. Unreliable input gene tree data presents however a considerable challenge to these summary methods (Springer and Gatesy 2016; Bryant and Hahn 2020). Recall in particular that under the MSC model, we assume no recombination within a locus, so that it is preferable to analyze relatively short sequences. Gene tree estimators for short sequence alignments will however have a high variance, causing problems for summary methods. We note that heuristic methods to account for gene tree uncertainty have recently been proposed for ASTRAL (Zhang and Mirarab 2022).

Both of the above classes of methods focus on modeling gene trees as the unit of analysis, either assumed to be given as input data or modeled as latent variables. Hence, the loci to which the MSC model is applied are *genes*, where a 'gene' is typically taken to be a protein-coding DNA sequence or sometimes an exon in this context, so that the assumption of no recombination becomes potentially problematic. The third class of methods, of which we name only SVDQuartets (Chifman and Kubatko 2014), adopts a rather different approach which circumvents this issue, using a collection of (putatively) unlinked sites as loci assumed to be evolving under the MSC. Quartet trees are determined based on site pattern counts for these loci, and the species tree is inferred using a method based on phylogenetic invariants (Felsenstein 2004). We will not consider these methods here further.

5.2.2 Motivation for a likelihood-free Bayesian approach

The intractability of the MSC likelihood for gene tree topologies $p(G|S, \phi)$ forces methods which use gene trees as input data to rely on summary statistics, which are not sufficient statistics for the MSC model. While the likelihood is intractable to evaluate, it is, however, extremely straightforward to simulate gene trees under the MSC model (algorithm 3), suggesting that likelihood-free approximate Bayesian computation (ABC) approaches may provide an interesting alternative (see Appendix A). ABC methods relying on the efficient simulation of genealogies under the coalescent have been the main approach for demographic inference under structured coalescent models in population genetics[4], and it is somewhat remarkable that their use in phylogenetic applications has not been considered extensively so far. Depending on the ABC approach taken, such methods have the potential to be more computationally efficient than joint Bayesian inference methods, while at the same time they may (but need not) be able to take into account uncertainty in the input gene tree data. In addition, likelihood-free methods may generalize more easily to more complicated models, such as the multi-species network coalescent (MSNC) (Wen, Yu, and Nakhleh 2016), the multilocus multispecies coalescent (MLMSC) (Rasmussen and Kellis 2012; Li et al. 2021) or MSC models with allopolyploid hybridization (see below).

There are multiple potentially fruitful avenues for likelihood-free phylogenetic inference for the MSC and cognate models. Fan and Kubatko (2011) explored a basic ABC approach where they simulated species trees and enumerate the distribution over gene trees, using the χ^2 distance between the expected and observed gene tree distribution as an ABC kernel. Note that this method does not make use of the fact that gene trees can be efficiently simulated, relying on a computationally intensive enumeration of the MSC-induced gene tree distribution which is intractable for even moderately sized species trees[5]. An obvious solution would be to approximate the expected gene tree distribution using simulation, however, a naive application of this idea will not work very well. As the simulated gene trees will usually not span the full tree space, the χ^2 distance or Kullback-Leibler (KL) divergence will often diverge, and the algorithm will not be stable unless a huge number of gene trees are simulated.

[4]For a nice overview of how the history of ABC approaches in Bayesian statistics is intimately related to challenging statistical problems that arose in population genetics, see Tavaré (2018).

[5]The program `hybrid-coal` (Zhu and Degnan 2017) takes about 4.2 seconds to enumerate the probabilities for the 135135 gene trees on eight taxa. The ABC method of Fan and Kubatko (2011) hence becomes quite useless already for only a few taxa.

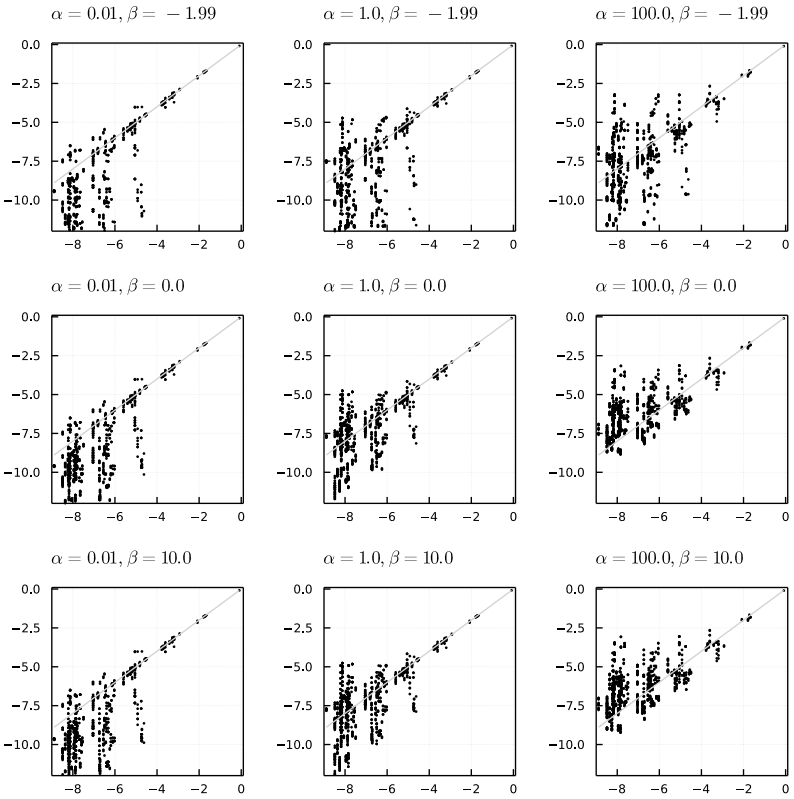


Figure 5.3: CCD approximation to the MSC-induced gene tree distribution for an eight-taxon species tree, using different prior distributions for the CCD estimator. Every dot represents a gene tree with probability $> 10^{-9}$, with gene tree probabilities computed using `hybcoal` (Zhu and Degnan 2017). The x-coordinate shows the true gene tree probability on a $\log_{10}$ scale, while the y-coordinate corresponds to the gene tree probability approximated by a CCD posterior derived from a sample of 10000 gene trees assuming a β-splitting Dirichlet-CCD prior with prior weight α. For this particular example, we found that $\alpha = 0.38$ and $\beta = 2.8$ minimizes the KL divergence between the CCD approximation and the true gene tree distribution.

By using the tree smoothing methods described in the previous chapter we could however salvage this approach, using a CCD with full support to approximate the MSC-induced gene tree distribution. The same idea could be used in a Bayesian *synthetic likelihood* approach (Wood 2010; Drovandi et al. 2018), where we substitute the intractable MSC likelihood by a CCD likelihood (the 'synthetic likelihood'), with the CCD estimated from simulated gene trees under the model. This synthetic likelihood can then be used in a standard algorithm such as MCMC for sampling from the posterior distribution over species trees. In fig. 5.3, we show to what extent such a CCD could approximate the MSC likelihood for different parameterizations of the CCD prior, by comparing the synthetic against the analytic likelihood for gene tree topologies on an eight-taxon species tree. There are several drawbacks to Bayesian synthetic likelihood or related ABC approaches. Besides the general computational and statistical problems with using noisy likelihood estimators, the approach also does not readily generalize to a situation where the taxonomic composition across loci differs, due to missing data or gene loss for instance, which is usually the case. While we have experimented with these approaches, we will not explore these in this chapter, and defer a detailed study to future work.

The flexibility of ABC approaches for likelihood-free inference leads to many other potentially viable strategies to unlock simulation-based inference under the MSC. In the remainder of this chapter, we describe a new approach for joint species tree inference and gene tree reconciliation under the MSC, based on likelihood-free expectation propagation (EP), also known as EP-ABC (Barthelmé and Chopin 2014). EP approximates a posterior distribution in a data-partitioned way, updating a variational approximation of the posterior data point by data point. This avoids the need to simulate the entire phylogenomic forest and the need to devise an ABC kernel for comparing the simulated and observed data in a genome-wide fashion. Our new approach makes use of the properties of the conditional clade distribution (CCD) described in chapter 4, both to take into account gene tree uncertainty and to approximate the posterior distribution over species trees.

5.3 Likelihood-free expectation propagation for the MSC

5.3.1 Overview of the likelihood-free EP approach

Let $y = (y_1, \dots y_n)$ denote the data set consisting of n loci, where y_i is the gene tree topology for locus i, assumed to be known without error (we will

deal with gene tree uncertainty later). Assuming the sampling distribution for the true gene tree topologies to be given by an MSC model, our central goal is to infer the species tree topology S and log-scale branch lengths ϕ. We assume the posterior distribution for (S, ϕ) can be written as

$$p(S, \phi|y) = \frac{1}{Z} p(y|S, \phi)p(S, \phi) = \frac{1}{Z} p(S)p(\phi) \prod_{i=1}^{n} p(y_i|S, \phi) \tag{5.2}$$

where Z (the marginal likelihood) is a normalizing constant. The problem which motivates our approach is the intractability of the likelihood term $p(y_i|S, \phi)$. We can alternatively express the posterior as

$$\begin{aligned} p(S, \phi|y) &\propto p(S)p(\phi) \int \delta_y(y^*)p(y^*|S, \phi)dy^* \\ &= p(S)p(\phi) \prod_{i=1}^{n} \int \delta_{y_i}(y_i^*)p(y_i^*|S, \phi)dy_i^* \end{aligned} \tag{5.3}$$

where δ is a Dirac measure. The most common approach for likelihood-free Bayesian inference would then proceed by substituting some summary statistic $s(y)$ and $s(y^*)$ for y and y^* respectively, replacing δ_y by some kernel function and approximating the resulting integral using some Monte Carlo technique (e.g. rejection sampling), yielding the usual form of an ABC posterior (Sisson, Fan, and Beaumont (2018); see Appendix A). In the context of species tree inference, such an approach was taken by Fan and Kubatko (2011), as was described above.

We will however adopt an approach due to Barthelmé and Chopin (2014) based on expectation propagation (EP) (Minka 2001; Seeger 2005; Vehtari et al. 2020), where we construct a variational approximation to the factorized posterior in eq. 5.3. EP assumes that the density to be approximated, $\pi(x)$ say, can be factorized as

$$\pi(x) = \frac{1}{Z} \prod_{i=0}^{n} l_i(x) \tag{5.4}$$

which is clearly the case when π is a posterior distribution for a model with n iid data points (in which case we have one of the l_i, say l_0, representing the prior). EP will construct an approximation $q(x)$ of π which admits the same

factorization

$$q(x) \propto \prod_{i=0}^{n} q_i(x)$$

where q, the global approximation, is a distribution from an exponential family and the *site approximations* q_i are members of the associated unnormalized exponential family (Seeger 2005). The EP algorithm then proceeds by iteratively updating the individual site approximations in such a way that q progressively improves as an approximation of π. The EP site update algorithm for site i has the following general form, given the current approximation q:

1. Construct the *cavity distribution* $q_{-i}(x) = q(x)/q_i(x)$
2. Construct the *tilted distribution* $q_{\setminus i}(x) = l_i(x)q_{-i}(x)$
3. Update q_i so that $q(x) = q_i(x)q_{-i}(x)$ approximates $q_{\setminus i}(x)$

(Vehtari et al. 2020). Usually in step (3) the updated q_i is obtained through moment matching, which for q in an exponential family amounts to minimizing the KL divergence between the tilted distribution and the global approximation with respect to the natural parameter of q_i (Minka 2001; Seeger 2005).

As Vehtari et al. (2020) stress, there is no compelling reason why in step (3) one should use moment matching – any numerical or Monte Carlo technique to approximate $q_{\setminus i}$ may yield a viable EP algorithm. In the likelihood-free algorithm of Barthelmé and Chopin (2014), the site approximation is also updated by moment matching, but the moments of the tilted distribution are now estimated using simulation. Applied specifically to the MSC inference problem: if we cast eq. 5.3 in the form of eq. 5.4 we will have

$$l_i(S, \phi) = \int \delta_{y_i}(y_i^*) p(y_i^* | S, \phi) dy_i^*$$

and we can express the tilted distribution as

$$q_{\setminus i}(S, \phi) \propto q_{-i}(S, \phi) \int \delta_{y_i}(y_i^*) p(y_i^* | S, \phi) dy_i^* \tag{5.5}$$

Given that we can simulate (S, ϕ) from the cavity distribution q_{-i}, that the MSC model admits efficient simulation of $y_i^* \sim p(\cdot | S, \phi)$ and that $\mathbb{P}(y_i^* = y_i) > 0$, we can approximate the tilted distribution using simulation, and use moment matching (or another strategy) to update q_i based on the sample from the tilted distribution. The challenge now remains to devise a suitable exponential family of distributions q which can be used to approximate eq. 5.2.

The EP approach also provides an approximation for the marginal likelihood (also known as the evidence), provided we can evaluate the marginal log-likelihood $\Phi(S, \phi)$ (also called log-partition function) for a member of the variational family. This is of course very useful for model selection purposes, allowing one to compute Bayes factors for competing species tree topologies for instance. To see this, we write the EP approximation with site-specific normalizing constants (here for a generic argument x)

$$q(x) = p(x) \prod_{i=1}^{n} \frac{q_i(x)}{C_i}$$

Seeger (2005) showed that one can update the approximation for C_i using the following relation

$$\log C_i = \log Z_{\setminus i} - \Phi(x) + \Phi_{-i}(x)$$

where Φ_{-i} is the log-partition function for the cavity distribution, and $Z_{\setminus i}$ is the normalizing constant of the tilted distribution. Provided we can approximate the latter using Monte Carlo simulation, we can update the site-wise contributions to the marginal likelihood and approximate the marginal likelihood under the model, the latter being given by

$$\log Z \approx \sum_{i=1}^{n} \log C_i + \Phi(x) - \Phi_0(x)$$

where Φ_0 is the log-partition function for the prior (Seeger 2005).

5.3.2 Use of the CCD as variational approximation

We now discuss how the conditional clade distribution (CCD) introduced in the previous chapter gives rise to a suitable family of distributions over cladograms which we can use in a likelihood-free EP algorithm. Crucially for our EP approach, the CCD model for a set of m taxa $\mathcal{L}$ constitutes an exponential family, as is easily seen by considering that its distribution is equivalent to a product of $2^m - m - \binom{m}{2} - 1$ categorical distributions (one for each clade γ for which $|\gamma| > 2$, resulting in a total of $\mathcal{O}(3^m)$ parameters). This means its density (or probability mass function) can be expressed in the form

$$f(x|\eta) = \exp(\eta^T t(x) - \Phi(\eta))$$

where η is called the natural parameter of the family f, $t(x)$ is the sufficient statistic and Φ is the log-normalizer. The natural parameter is associated with the moment parameter θ by an invertible transformation. In the CCD case, this means that for a clade γ with conditional split distribution given by the moment parameter θ_γ, the associated natural parameter $\eta \in \mathbb{R}^{\#_\gamma - 1}$ will be

$$\eta_\gamma = (\eta_{\gamma,1}, \ldots, \eta_{\gamma,\#_\gamma - 1}) = \left(\log \frac{\theta_{\gamma,1}}{\theta_{\gamma,\#_\gamma}}, \ldots, \log \frac{\theta_{\gamma,\#_\gamma - 1}}{\theta_{\gamma,\#_\gamma}} \right)$$

Note that the operations required for performing a site update in the EP algorithm are particularly straightforward in natural parameter space, the natural parameter η_{-i} of the cavity distribution being related to the natural parameter of the global approximation η and site approximation η_i as $\eta - \eta_i$.

To obtain an exponential family which can approximate the species tree posterior, we assume the posterior approximation q factors as

$$q(S, \phi) = q_\eta(S) q_\zeta(\phi) \propto \prod_{i=0}^{n} q_{\eta,i}(S) q_{\zeta,i}(\phi) \tag{5.6}$$

Here $q_\eta(S)$ will be a CCD with natural parameter η as discussed above, whereas we use for $q_\zeta(\phi)$ a multivariate Gaussian distribution with mean $\mu = (\mu_{\gamma,\delta} \colon \gamma \subset \mathcal{L}, \delta \subset \gamma)$ and covariance matrix $\sigma^2 I$ with $\sigma^2 = (\sigma^2_{\gamma,\delta} \colon \gamma \subset \mathcal{L}, \delta \subset \gamma)$. We denote the natural parameter of the multivariate Gaussian factor of q by ζ. It is important to stress that the family defined by eq. 5.6 consists of joint distributions over the entire species tree space, which are not defined in terms of a more intuitive conditional structure (e.g. as in Zhang et al. 2020), where we would have a similar CCD model for the m-taxon species tree S, and conditional on a particular S a multivariate Gaussian for the $2m - 2$ branches in S. To see that our model does not have such a structure, note that a random draw from a distribution in the family defined by eq. 5.6 *formally* consists of a species tree topology and a vector of $\mathcal{O}(3^m)$ real numbers (two for each possible split). Of course we need only explicitly represent the $2m - 2$ relevant branch lengths when performing computations, however, it is important to take into account the full space when estimating a new distribution in the family (for instance through moment matching).

5.3.3 Implementation of the likelihood-free EP algorithm

With our variational family in hand we return to step (3) of the EP algorithm. Following Barthelmé and Chopin (2014), we can use a rejection sampling approach as in algorithm 5 to approximate the tilted distribution for each site i, corresponding to data point y_i. The accepted simulations will constitute a sample from the tilted distribution $q_{\setminus i}(S, \phi)$ (eq. 5.5). We then approximate the tilted distribution within the exponential family q by estimating a CCD from the accepted $S^{(j)}$ to obtain an estimate of the natural parameter of the tilted distribution $\hat{\eta}_{\setminus i}$ and using moment matching to estimate $\hat{\xi}_{\setminus i}$.

We use the Bayes CCD estimator discussed in chapter 4 under a β-splitting Dirichlet-CCD prior with parameters α and β to estimate the CCD from the accepted species tree simulations, ensuring the approximation has the complete tree space as support. For β we use the same value as for the CCD prior for the species tree, whereas α is a user-defined parameter. We then update q using a convex combination of the approximation to the tilted distribution and the previous global approximation, i.e. $\eta' = (1 - \lambda)\eta + \lambda\hat{\eta}_{\setminus i}$ and similarly for ζ. The site approximation is then updated accordingly. Note that both α and λ are parameters which affect the 'learning rate' of the EP algorithm: choosing larger α will cause the tilted distribution approximations to be more strongly shrunk towards the prior CCD, whereas a smaller λ will lead to a less drastic update of the global approximation for each site iteration. Unless stated otherwise, we use $\lambda = 0.1$ and $\alpha = 10^{-4}$ in our experiments below. We also implemented an EP algorithm for fixed species tree topologies with a multivariate Gaussian on $\mathbb{R}^{2m-2}$ as variational family for the branch parameters, using the same algorithmic approaches for approximating the tilted distribution and performing EP updates.

Algorithm 5 Rejection sampling for approximation of $q_{\setminus i}$

1: **for** $j = 1, \ldots, M$ **do**
2: Simulate a draw from the cavity distribution $(S^{(j)}, \phi^{(j)}) \sim q_{-i}$
3: Simulate a gene tree on the leaf set of locus i, $y_i^* \sim p(\cdot | S^{(j)}, \phi^{(j)})$
4: Accept $(S^{(j)}, \phi^{(j)})$ with probability $\delta_{y_i}(y_i^*)$
5: **end for**

5.3.4 Accounting for gene tree uncertainty

In the above, we have assumed that the data consists of gene tree topologies for n loci, known without error. This is an unrealistic situation, since in practice gene trees are usually inferred from molecular sequence data using statistical phylogenetic methods. This is of course the main problem shared by all summary methods, however, in the EP approach described above this can be accounted for in a rather natural way. To alleviate the issue, we assume instead that the true tree topology Y_i for locus i is unknown, and that we have a probability distribution ξ_i over gene tree topologies at our disposal which adequately describes our uncertainty about Y_i, so that eq. 5.3 becomes

$$p(S, \phi|y) = p(S)p(\phi) \prod_{i=1}^{n} \int \xi_i(y_i^*) p(y_i^*|S, \phi) d\, y_i^* \tag{5.7}$$

Where we assume $\xi_i(y_i^*) \approx \mathbb{P}\{Y_i = y_i^*\}$. The same substitution of ξ_i for δ_{y_i} applies to eq. 5.5 and line 4 in algorithm 5 of course. Here ξ_i could be derived from a sample from the posterior distribution over gene trees under some standard phylogenetic CTMC model of evolution, obtained using for instance MrBayes or RevBayes (Ronquist et al. 2012a; Höhna et al. 2016). Alternatively, a sample of trees could be obtained using nonparametric bootstrapping methods (Hoang et al. 2018; Chao Zhang et al. 2018). The sample could be used as such (i.e. using the sample frequencies as estimates of the tree probabilities) or a smoothed distribution such as a CCD or a subsplit Bayesian network (SBN; Cheng Zhang and Matsen IV (2018a)) could be used. Note that ξ_i assumes the role of the kernel function the ABC context (Appendix A).

5.3.5 Improving the tilted approximation

The efficiency of the whole strategy will depend crucially on how efficiently we manage to approximate the tilted distribution. For trees of appreciable size and weakly informative or non-informative priors for S, the acceptance probability in the rejection sampling algorithm may be very low, in which case it may take a very long time before we get to a reasonable posterior approximation q. The simulation-based approximation of the tilted distribution is however a fairly standard ABC problem, so in principle we can use the whole battery of ABC approaches that have been developed (see Sisson, Fan, and Beaumont 2018, chap. 4, for an overview).

Approaches based on importance sampling (IS) may alleviate efficiency issues

Algorithm 6 SIS with rejection control for approximation of $q_{\setminus i}$

Require:
gene tree probability distribution ξ_i
cavity distribution q_{-i}
number of particles M
rejection threshold c
stopping rule
Ensure:
$g \leftarrow q_{-i}$
while stopping rule not satisfied **do**
  for $j = 1, \ldots, M$ **do**
    Simulate species tree $(S_j, \phi_j) \sim g$
    Simulate gene tree $y_i^* \sim p_i(\cdot|S_j, \phi_j)$
    Compute importance weight $w_j \leftarrow \xi_i(y_i^*)q_{-i}(S_j, \phi_j)/g(S_j, \phi_j)$
    Accept (S_j, ϕ_j) with probability $r_j = \min(1, w_j/c)$
    Update importance weight $w_j \leftarrow w_j/r_j$
  end for
  Construct a new g from the accepted (w, S, ϕ)
end while
return the final accepted (w, S, ϕ) as a weighted sample from $q_{\setminus i}$

to some extent, but require careful monitoring of importance weights in order to obtain a stable approximation. A basic IS approach uses $\xi_i(y_i^*)$ as an importance weight for each simulated $(S^{(j)}, \phi^{(j)})$ pair. This importance sampling step integrates nicely with the recycling technique of Barthelmé and Chopin (2014), which leads to additional computational gains. More sophisticated IS approaches such as sequential importance sampling (SIS) rejection-control IS (Liu, Chen, and Wong 1998; Liu 2001; Peters, Fan, and Sisson 2012) or sequential Monte Carlo (SMC) can further improve the approximation of the tilted distribution for a given computational budget. In our experiments below, we use a SIS algorithm with rejection control, along the lines of algorithm 6, where we select a rejection threshold dynamically based on some quantile of the importance weight distribution. We refer to this algorithm as EP-ABC-SIS. As a stopping rule we use a threshold on the effective sample size (ESS) as defined by Kong (1992)

$$\mathrm{ESS} = \Big(\sum_{j=1}^{M} w_j^2 \Big)^{-1}$$

which is widely used in importance sampling applications, together with a maximum number of SIS steps. When the final set of importance weights does not reach a specified minimum ESS threshold, the site update has failed, and we let the EP algorithm proceed to the next site. If a particular site consistently fails to be updated across subsequent EP passes, the locus under consideration will not contribute to the posterior approximation. Depending on one's point of view, this could be considered a bug or a feature, as this results in a kind of automatic detection, and discarding, of outlier loci. Another strategy which avoids this would be to continue sampling until some minimum number of accepted simulations is reached, without setting a hard upper bound on the number of simulation replicates. In our specific application, such a strategy risks however to spend a lot of computational time on outlier gene families which strongly violate the MSC model assumptions – leading to the less than desirable situation where most time is spend making our inferences more biased due to model violations.

Another strategy for approximating the tilted distribution would be to use an MCMC sampler. Starting from an initial sample from the cavity, we could sample from the tilted distribution by applying an MCMC transition kernel for a specified number of iterations. While we will not pursue this strategy here, we note that this may further alleviate issues that arise from applying IS-based algorithms to high-dimensional target distributions, and would definitely be a worthwhile avenue for further algorithmic improvements.

5.4 Applications of the EP approach

5.4.1 Simulation experiments

For small trees of three or four taxa, it is straightforward to approximate the posterior distribution by numerical quadrature, enabling a detailed verification of the correctness of the implementation and quality of the EP approximation. fig. 5.4 shows the EP approximation for a simulated data set consisting of 200 gene families from an unbalanced four-taxon phylogeny, obtained using the SIS algorithm with $M = 10^4$ and as stopping rule a minimum ESS of 1000 after at most five iterations of the SIS inner loop. We used a Uniform prior ($\beta = -1.5$) for the species tree, and a Gaussian prior with mean 0 and variance 5 for the log-scale branch parameters. The posterior approximation has probability mass ≈ 1 on the true species tree, and the gaussian approximation to the posterior density for the branch parameters matches the true posterior den-

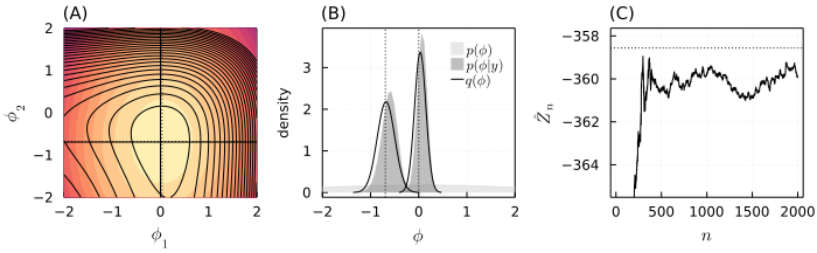


Figure 5.4: Simulation experiment for a four-taxon species tree. (A) Comparison of the joint posterior for the two internal branch lengths computed by numerical quadrature (black lines) with the posterior approximation obtained using the EP-ABC-SIS algorithm (colored contour plot). The vertical and horizontal lines mark the posterior mean for each parameter for both the posterior computed by quadrature and the EP posterior. (B) Marginal posterior densities for the two internal branch lengths for the posterior computed by quadrature $p(\phi|y)$ and the EP posterior approximation $q(\phi)$. The dotted lines mark the true (simulated) values. The prior density $p(\phi)$ is shown as well. (C) Trace plot of the marginal likelihood estimator throughout 10 passes of the EP algorithm. The marginal likelihood computed by quadrature is indicated by the dotted line.

sity extremely well (fig. 5.4 A,B). We find that the algorithm converges quite rapidly, with the approximation stabilizing after about two full EP passes over the data (fig. 5.4 C). Furthermore, we find that the EP algorithm provides a fair approximation to the marginal likelihood $\sum_S \int_\phi p(y, S, \phi)$ in the three and four-taxon problems where we can compute the latter by numerical quadrature.

We next assessed the performance of the EP algorithm with the SIS sampler on a 10-taxon species tree, simulated from a β-splitting distribution with $\beta = -1$ and with log-scale internal branch lengths sampled from a mixture of two normal distributions, one with mean $\log 0.5$ and weight 0.2 and another with mean $\log 3$ and weight 0.8, with both components having variance 0.5. We use the same prior and algorithm settings as before, except for using $M = 10^5$. The EP posterior is displayed in fig. 5.5, together with the simulated species tree. We find that the algorithm retrieves the correct species tree and provides a posterior distribution for the branch lengths compatible with the simulated branch parameters.

5.4.2 Yeast data set

We analyzed the rather famous data set of Rokas et al. (2003) using the proposed likelihood-free EP algorithm. This data set consists of 106 protein-

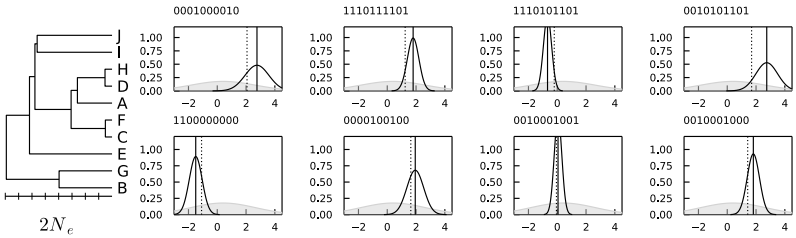


Figure 5.5: EP posterior approximation for 100 loci simulated from a 10 taxon species tree. The species tree is shown on the left, with branch lengths in N_e generations. Note that the terminal branch lengths are not meaningful. The Gaussian approximations to the marginal posterior distribution of the log-scale branch length for each of the eight internal branches is shown, with the true (simulated) branch length indicated by the dotted gray line. The posterior probability of the true species tree was ≈ 1.

coding genes from seven budding yeast (*Saccharomyces*) species and one outgroup (*Candida albicans*) and was one of the first empirical data sets to draw broad attention to the issues of heterogeneity in gene trees across the genome in phylogenetic inference.

We investigate the effect of taking into account gene tree uncertainty by conducting the analysis using tree topologies inferred by maximum likelihood (using IQ-TREE (Minh et al. 2020), with the GTR + Γ4 model) on the one hand, and using empirical CCDs estimated from MCMC samples on the other. To estimate the CCDs, we sampled 10000 tree topologies under the GTR+Γ4 model and default prior settings in MrBayes (Ronquist et al. 2012b) (sampling a tree every 50 iterations for a total of 550000 iterations, discarding 1000 samples as burn-in). We use a uniform prior for S and a Gaussian distribution with mean $\log 2$ and variance 5 for the log-scale branch parameters, and use the SIS algorithm with $M = 10000$ and default settings for conducting EP.

MAP species trees inferred using the EP algorithm for both data sets are shown in fig. 5.6 (A) and (C). Notably, both analyses result consistently in a posterior approximation which concentrates on a single tree topology, but which differs for the two data sets. We refer to the MAP tree for the CCD data and ML gene tree data as topology 1 and 2 respectively. We also ran the EP algorithm with a fixed species tree topology, analyzing both data sets assuming the MAP species tree inferred for the other data set (fig. 5.6 B & D), which confirms that indeed both data sets are significantly better fit by different species tree topologies. Note that both topologies have appeared in previous studies (e.g. Edwards, Liu, and Pearl 2007; Fan and Kubatko 2011; Flouri et al. 2020),

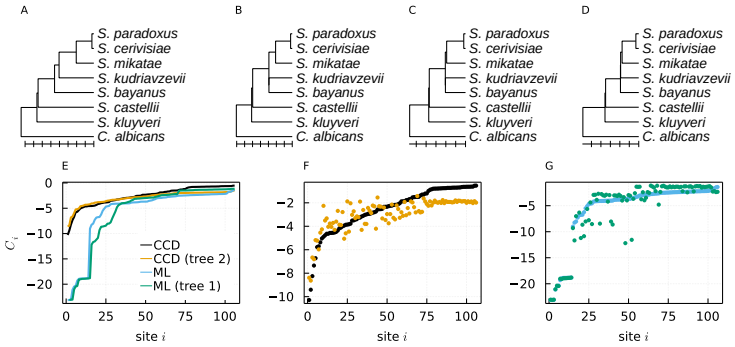


Figure 5.6: Analysis of the Rokas et al. (2003) yeast data set using the EP algorithm. (A) MAP species tree using MCMC-derived CCDs as input data. (B) Species tree with MAP branch lengths based on the CCD input data for the tree (tree 2) in (C). (C) MAP species tree using ML gene trees as input data. (D) Species tree with MAP branch lengths for the tree in (A) (tree 1) using ML gene trees as input data. The estimated marginal log-likelihoods are -242, -283, -549 and < -606 respectively. One unit along the x-axis represents $10N_e$ generations. (E) Ordered site-wise contributions C_i to the marginal likelihood for the four different analyses associated with (A-D). The ordering is different for each analysis. (F) C_i for the CCD based analyses, with sites in the same order. (G) C_i for the ML based analyses, with sites in the same order. All analyses were performed using the EP-ABC-SIS algorithm with a uniform species tree prior ($\beta = -1.5$) and a $\mathcal{N}(\log(2), 5)$ prior for the log-scale coalescent branch lengths, performing five passes over the data.

one of which (Flouri et al. 2020) suggesting that some of the observed gene tree heterogeneity is due to introgression between the *S. kudriavzevii* and *S. bayanus* lineages.

Unsurprisingly, an inspection of the site-wise contribution to the marginal likelihood indicates a markedly better fit of the MSC model when taking into account gene tree uncertainty (fig. 5.6 E), and shows that for the CCD-based analysis the MAP tree does seem to provide an improved fit across much of the data (fig. 5.6 F). Note that, while for the analyses based on the ML tree data the marginal likelihood cannot be reliably estimated, the site-wise contributions are still informative to estimate an upper bound to the marginal likelihood and visually compare model fit. For the ML tree input data, we see that many loci are actually better explained by the species tree topology 1, and that the posterior concentration on the topology 2 appears to be driven by a handful of genes which are highly unlikely under topology 1 but not so under topology 2. Also telling is the posterior expectation for the sum of the internal branch lengths, which we estimate at 119 with 95% uncertainty interval (47, 262)

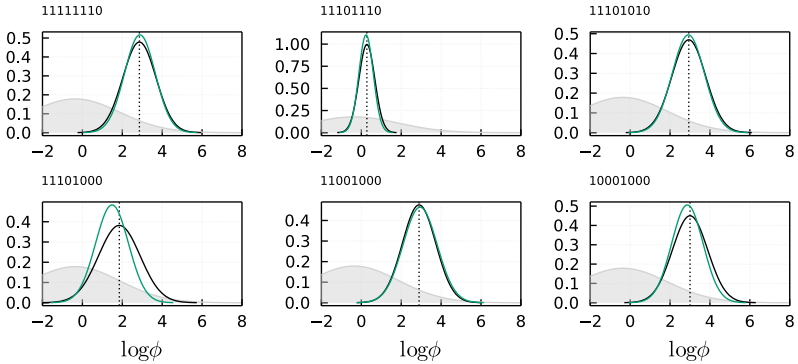


Figure 5.7: Marginal posterior approximation for the log-scale coalescent branch lengths for the Rokas et al. (2003) data set, using MCMC-derived CCDs as input data (analysis associated with MAP tree (A) in fig. 5.6). The Gaussian approximation obtained with the standard EP-ABC-SIS algorithm is shown for each of the six internal branches in black. The Gaussian densities in green show the marginal posterior approximations obtained with the full multivariate Gaussian approximating family (with the species tree held fixed in the EP algorithm). The gray density shows the prior distribution.

for the CCD-based analysis, compared to 64 (29, 126) for the analysis based on ML trees (both in N_e units). This clearly shows the strong bias towards higher inferred degrees of discordance when not taking into account gene tree uncertainty.

In fig. 5.7 we show the Gaussian posterior approximations for the branch parameters for the CCD-based analysis. From these plots we can conclude that for long species tree branches, as expected, there is little information in the data to estimate the branch length with high precision. Finally, we also implemented the same EP-ABC-SIS algorithm for fixed species trees with a full multivariate Gaussian distribution as approximating family for the branch length posterior (i.e. with non-zero covariance terms), and compare the resulting EP posterior to our simpler model with independent Gaussian components (fig. 5.7). The marginal posterior approximations for both analysis agree very well, suggesting not much is to be gained by using a more complicated variational family for the branch parameters

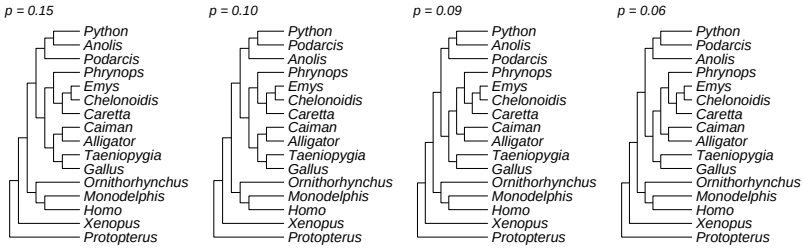


Figure 5.8: The four most probable trees and their respective probabilities for the prior distribution constructed from ASTRAL species trees. The ASTRAL trees were inferred from bootstrap samples of gene trees from the Chiari et al. (2012) data. We used a β-splitting Dirichlet-CCD with $\beta = -1$ and $\alpha = 10$. Most of the variation seems to involve the relationships within lepidosaurs and the position of turtles relative to birds and crocodiles.

5.4.3 The problem with turtles

The position of turtles within the amniote phylogeny has been a longstanding topic of controversy in systematics. Traditional debates centered around amniote skull morphology, where the anapsid skulls of turtles were assumed to be the ancestral morphology of diapsid reptilians, an assumption which became increasingly contested over time. The situation was however not resolved immediately with the advent of molecular data sets, with different phylogenetic hypotheses receiving decisive statistical support in different studies (see Brown and Thomson 2017 for an overview of recent phylogenetic studies). Based on molecular data, turtles have been considered sister to all other reptilians, sister to lepidosaurs (lizards, snakes and tuataras), sister to archosaurs (birds and crocodolians) or sister to crocodilians (Chiari et al. 2012), although most studies point to one of the latter two hypotheses. Brown and Thomson (2017) conducted a detailed study comparing phylogenetic information for various hypotheses concerning the placement of turtles within the amniote phylogeny using different published data sets, showing that different outcomes are often driven by a handful of loci. In this section, we investigate the data set of Chiari et al. (2012) (also studied in Brown and Thomson (2017)) using the EP-ABC-SIS method. This data set consists of 248 loci from 16 taxa spanning the amniote phylogeny. As input data, we use CCDs estimated from MCMC samples of gene tree topologies obtained using MrBayes (with the GTR + Γ4 model and default priors) as for the yeast data set discussed above.

Importantly, when the number of taxa increases, IS-based algorithms may

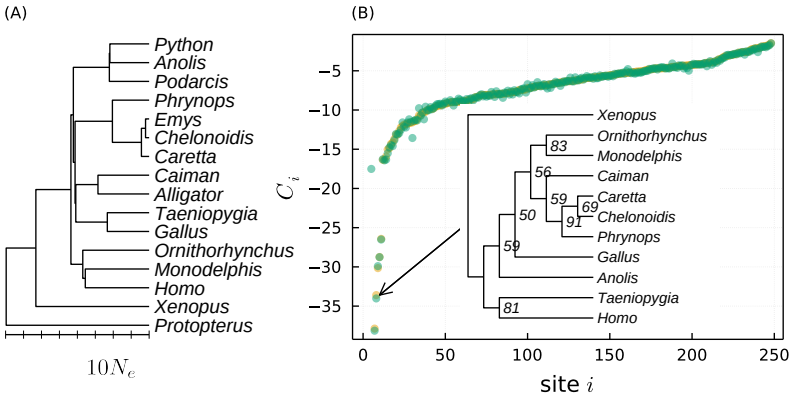


Figure 5.9: (A) MAP species tree for the Chiari et al. (2012) data set. (B) Site-wise contribution C_i to the marginal likelihood for two independent runs (in yellow and green) of the EP-ABC-SIS algorithm. Sites are ordered according to the C_i for the first run. The tree in the inset of (B) shows the consensus tree for outlier gene `ENSGALG00000002969` with clade credibilities (from the MrBayes MCMC sample) on a % scale.

have difficulties obtaining reasonable approximations for the tilted distribution of S. Besides being inefficient, the algorithm may be sensitive to the number of particles (M) used, resulting in unstable approximations for insufficiently large M. This can be alleviated to some extent by using more informative prior distributions, if such information is available – as is usually the case in phylogenetic problems. Moreover, the CCD prior for the species tree provides quite some flexibility for constructing prior distributions, without assigning zero probability to large parts of tree space. One approach would be to estimate a CCD from the observed gene trees with a strong β-splitting Dirichlet-CCD prior (i.e. with large α). Another, arguably better, approach would be to use a fast heuristic algorithm such as ASTRAL to infer a collection of species trees for bootstrap samples of gene trees, and to use the inferred species trees to estimate the CCD prior. We use the latter approach here. Specifically, we constructed 100 data sets where we sampled 1000 loci with replacement, and for each locus one gene tree from the posterior distribution under the GTR + Γ4 model. We then ran ASTRAL (v5.5.3) on each bootstrap sample and constructed a CCD from the resulting collection of species trees. In fig. 5.8 we show the four most probable species trees under a prior obtained using this strategy.

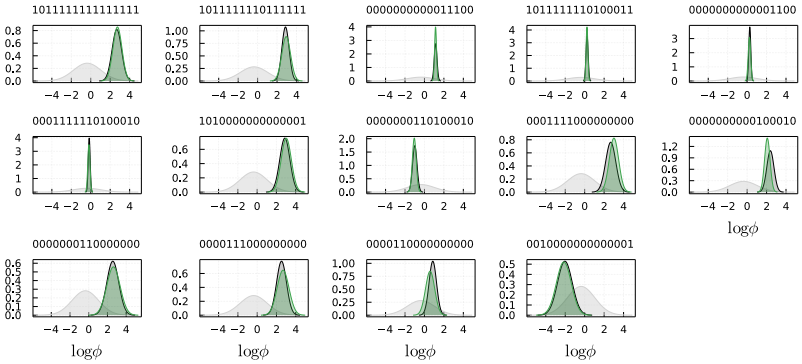


Figure 5.10: EP posterior approximation for log-scale coalescent branch lengths of the MAP species tree inferred for the Chiari et al. (2012) data set. The prior density (a $\mathcal{N}(\log 2, 5)$ distribution) is shown in light gray, while the two colored densities show the marginal Gaussian posterior approximations obtained from two independent runs of the EP-ABC-SIS algorithm (with $M = 50000$, performing five passes over the full data set). The bitstrings above each plot show the clade codes associated with each branch.

The EP posterior concentrates on the species tree with turtles sister to birds and crocodilians (Archosauria), and supports the grouping of *Anolis* and *Python* (Toxicofera) with *Podarcis* sister to this clade (fig. 5.9). The EP posterior approximation does not vary significantly across independent runs (fig. 5.10), suggesting the chosen algorithm settings do lead to a stable variational approximation. Note that in the MSC-based analyses of Chiari et al. (2012) (using MP-EST (Liu, Yu, and Edwards 2010)), the authors found a sister relationship of turtles to crocodiles using the same loci, but using ML trees without taking into account gene tree uncertainty. They attributed their results to substitution saturation, as an analysis using gene trees inferred at the amino acid level resulted in turtles sister to archosaurs, agreeing with their concatenation-based Bayesian phylogenetic analysis.

Brown and Thomson (2017) however showed that some of the gene trees which show extremely strong support for a crocodilian sister relationship likely include paralogous genes, providing an alternative explanation for these conflicting results. Using the EP results, we also found a number of outlier loci whose gene trees are highly unlikely under the MSC model (fig. 5.11). All of these gene families show evolutionarily highly implausible clades (for instance with paraphyletic mammals or birds), suggestive of false orthology assignments. Note that none of the loci reported by Brown and

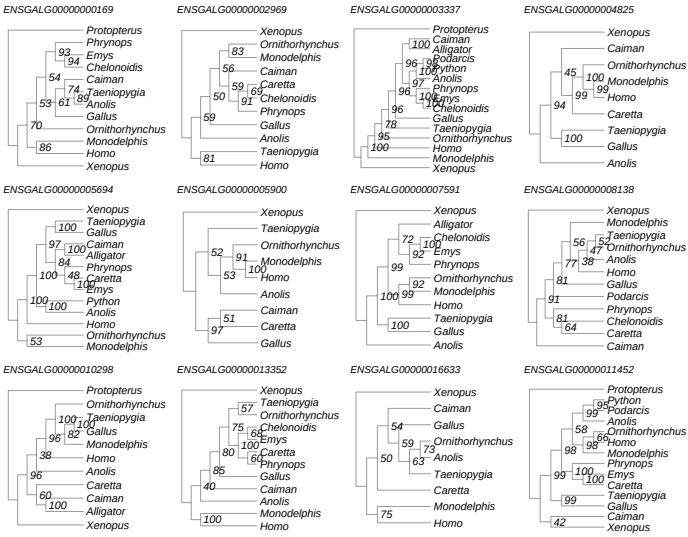


Figure 5.11: Majority rule consensus trees with clade credibilities from a MrBayes MCMC sample for the twelve outlier gene families (having $C_i < -20$ in either one of the two EP runs shown in fig. 5.9).

Thomson (2017) to include paralogous clades are signaled as outliers by the EP results, and they even have a fairly high likelihood under the posterior MSC model (with C values of -13.7 and -8.9), showing that discordance due to other sources may be wrongly modeled as due to ILS. While the approach thus clearly signals wildly implausible gene trees under the MSC, when there is some discordance at some species tree node due to ILS or variance in the gene tree estimates, it is not possible to flag putative model violations due to, for instance, paralogy without external information. Clearly, while the short branch lengths in fig. 5.9 may signal ILS to some extent, they are highly likely to be significantly biased downwards due to other sources of discordance, hampering biologically meaningful interpretations of inferences under the MSC model. This is a recurring (and somewhat sad) theme in phylogenetic inference, stressing once again the need for more realistic models of gene family evolution that can deal with the major sources of evolutionary variation and enable us to learn about the processes of genome evolution, without simultaneously incurring statistical biases due to restrictive model assumptions.

5.4.4 *Drosera* allopolyploid hybridization

The sundews (*Drosera*) constitute a large cosmopolitan genus of carnivorous plants, notorious for their flypaper-like leaf morphology which they use to trap insects. The evolutionary history of the genus is complicated due to extensive hybridization, likely both allopolyploid and homoploid in nature (see for instance Brittnacher (2010)). In the context of an ongoing genomic study of the sundew lineage, the genomes of *Drosera capensis* and *Drosera regia* were sequenced and high-quality genome assemblies were obtained (Renner et al., *in preparation*). Comparative genomic analyses revealed that the *D. capensis* genome exhibits a six-fold multiplication with respect to homologous regions in, for instance, *Beta vulgaris* (beet), while *D. regia* shows a clear triplicate structure (fig. 5.12). *D. capensis* can further be shown to be derived from a recent allopolyploid by comparing its genome against, for instance, *D. spatulata* (see chapter 6), so that its six-fold multiplication level likely derives from an ancient hexaploid history followed by rediploidization and a more recent allotetraploid phase. The *D. regia* triplicated structure also likely derives from an ancient hexaploid phase, however, differences in rates of molecular evolution and genome rearrangement render it difficult to determine whether the two lineages share a diploid, tetraploid or hexaploid ancestor (or whatever combination of those), despite having complete genome assemblies available.

To trace the evolutionary relationships of the nine subgenomes of *D. capensis* and *D. regia*, we inferred colinear alignments within the macrosyntenic clusters shown in fig. 5.12. We identified those colinear blocks for which six *D. capensis* homologous regions could be identified and three *D. regia* regions (6:3 colinear regions). For each position in the colinear alignment with more than three anchors, we identified an outgroup gene in *Beta vulgaris* and inferred a multiple sequence alignment using MAFFT (Katoh and Standley 2013). We then inferred gene trees for each individual set of anchors. Additionally, we constructed for each of the identified 6:3 colinear regions a concatenated alignment of all anchors within a colinear block and inferred a phylogeny for each such 6:3 block. All trees were inferred using ML with IQ-TREE (Minh et al. 2020), using default settings and the GTR + Γ4 model of sequence evolution. Note that without phasing the different colinear stretches into their respective subgenomes, we cannot construct a reasonable concatenated alignment for the *whole* data set (for instance, without additional information, it is unclear which of the six colinear stretches for *D. capensis* in one block should be concatenated to which of the six in another).

The results of this exploratory analysis are displayed in fig. 5.13. With 12

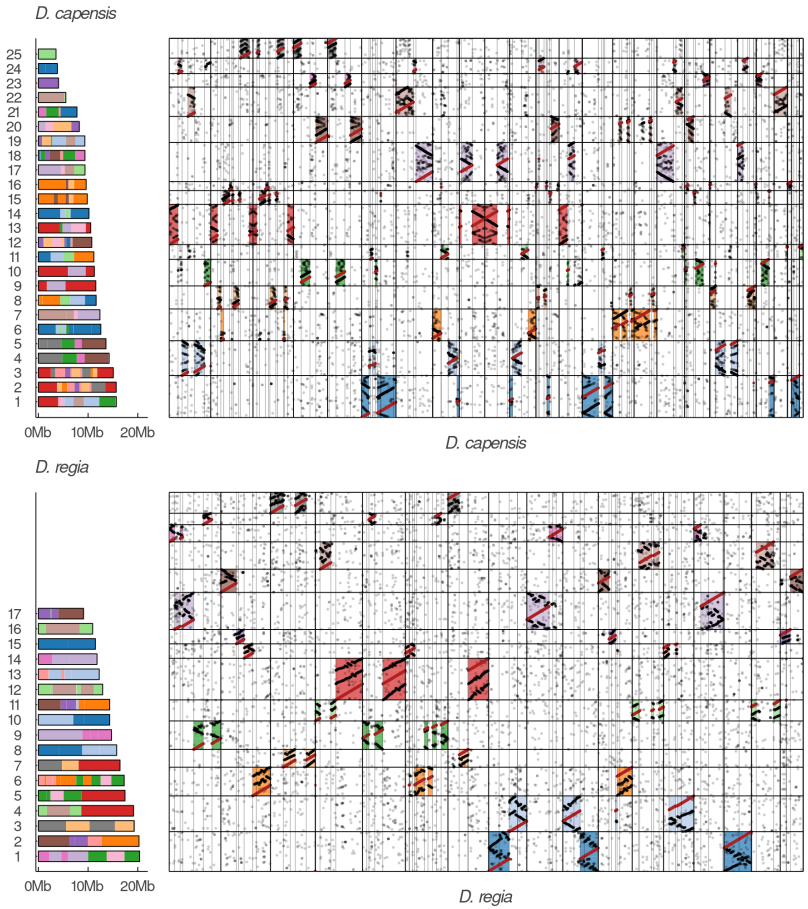


Figure 5.12: Macrosyntenic analysis of the *Drosera capensis* and *Drosera regia* genomes. We developed a Bayesian segmentation and clustering algorithm for the inference of genome-scale patterns of homology and putative ancestral blocks of genes (*macrosynteny clusters*). The method is a generalization of Nakatani and McLysaght (2017) and will not be presented in this dissertation. The bar graphs on the left show the chromosomal organization of the two species with homologous regions painted by macrosyntenic cluster. The scatter plots on the right show the within-genome gene homology matrix where the relevant genome is represented by the x-axis, with each x-coordinate representing a gene and the black and gray lines representing chromosome and syntenic segment boundaries respectively. The same genome is represented along the y-axis, but with the segments reordered by macrosyntenic cluster (indicated by the colored regions). Each dot represents a homologous gene.

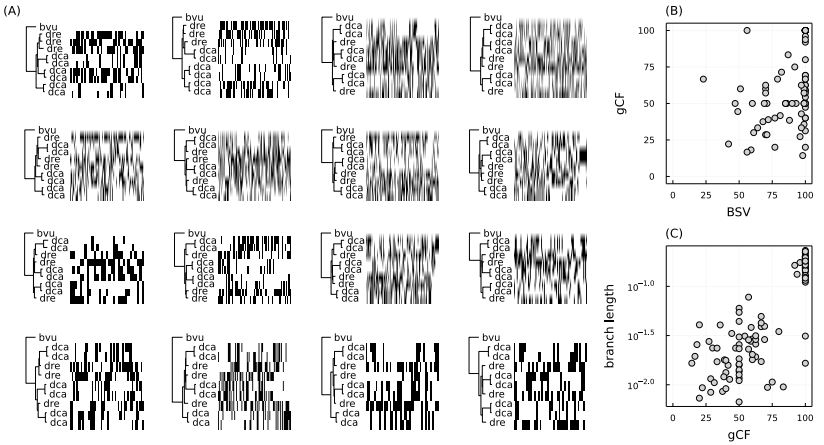


Figure 5.13: Phylogenetic analysis of 6:3 colinear regions in *Drosera capensis* and *Drosera regia*. (A) Maximum likelihood tree topologies for concatenated alignments of 6:3 colinear regions, with associated taxon occupancy matrices for the concatenated alignments (heatmaps, black/white is presence/absence of an anchor in the taxon label in the tree aligned with the relevant row). (B) Relationship between gene concordance factors (gCF) and bootstrap support value (BSV) and (C) gCF and branch length.

distinct topologies for 16 alignments, the situation clearly appears complicated. What we can conclude from these results is that, indeed, the *D. capensis* subgenomes come in pairs due to the recent allotetraploidy event in this lineage. However, not much can be said with respect to the events deeper in the tree, with many different relationships among the three pre-WGD *D. capensis* lineages and three *D. regia* subgenomes having some support. What is clear however is that the molecular distances for branches emanating from the crown *Drosera* group are all very small, which suggests (1) that ILS could be relevant (and concatenation problematic) and (2) that there may be considerable uncertainty in gene tree estimates due to a limited number of substitutions being informative for the branches of interest. Furthermore, note that the assumption behind the concatenation approach that the genes in a colinear stretch share the same history may be severely violated here if recombination between distinct subgenomes after polyploidy is prevalent (but note that this would also affect the assumption of a single tree being representative for each individual gene to a lesser extent).

The absence of a reliable phasing of the individual genes into subgenomes prevents a straightforward analysis of this data set under the MSC model as-

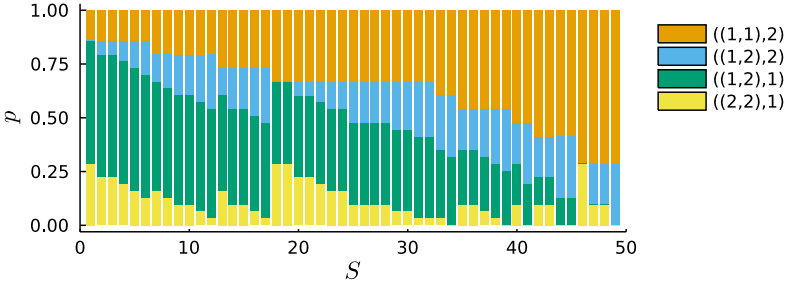


Figure 5.14: The probability distribution over rooted triples in MUL gene trees for each of the 49 six-taxon MUL species trees consisting of 2×3 subgenomes (labeled '1' and '2') assuming no ILS. There are four kinds of triples in this setting, shown in Newick format on the right. Note that there are distinct species trees which induce the same distribution over rooted triples, hence S is not identifiable on the basis of the distribution of rooted triples in observed gene trees.

sumption. An equivalent way to express the problem is that we have no unambiguous gene-to-subgenome map σ, as we do in our usual applications of the MSC model. What we need is therefore an approach for inference under the MSC based on multi-labeled gene trees (MUL) trees, where the same leaf label appears multiple times in a single gene tree (in our case, six leaves are labeled as *D. capensis*, whereas three leaves are labeled as *D. regia* genes). Joint Bayesian inference of gene trees and species trees would in principle admit this, using the subgenome phase-swap Markov kernel implemented by Freyman, Johnson, and Rothfels (2020) in RevBayes (Höhna et al. (2016)). However, we did not manage to get decently mixing MCMC chains in reasonable time, neither in a basic concatenation analysis nor in a hierarchical model under the MSC (even after making further modifications to the Markov kernel motivated by this specific problem, allowing for instance joint swap moves for multiple genes), presumably in part due to the nested nature of multiple ancient genome duplications in this problem. Summary methods for MUL trees are not available, and will likely not work in general due to identifiability issues, as even in the absence of ILS, the species tree appears not to be identifiable based on the distribution of rooted triples (fig. 5.14).

We adapted the EP approach to enable inference of MUL species trees from MUL gene trees by making the assumption that all assignments of genes to subgenomes of their respective species are equally likely *a priori*, and simulating gene trees accordingly. This is a rather naive approach, which does not learn the subgenome phasing explicitly as part of the EP

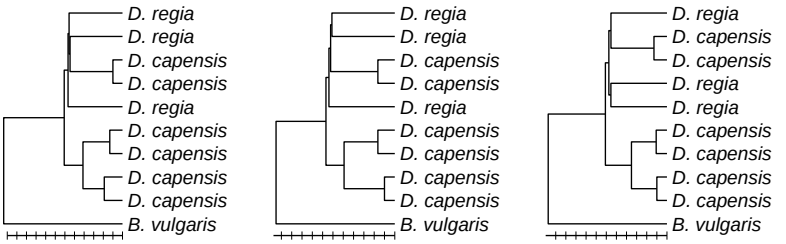


Figure 5.15: EP-based inference of a MUL species tree under the MSC for the *Drosera* 6:3 colinear block data set. The three species tree topologies with near-identical marginal likelihood estimate are shown. The scalebar is displayed with units of $2N_e$. Terminal branch lengths are meaningless.

algorithm, nor considers the possibility of unequal sampling probabilities for different subgenomes (which would be expected when gene loss during the complicated rediploidization process is biased across subgenomes). As often the case in these kind of Bayesian mixture models, the approach also leads to a strongly multimodal posterior distribution due to the permutation invariance of the likelihood (i.e. permuting the labels of the subgenomes *within* a species does not change the likelihood). Nevertheless, the approach does seem to provide reasonable results across multiple runs of the EP algorithm, converging however to a single mode of the posterior.

Applying this approach to the 229 loci with at least six taxa in the 6:3 colinear block data set (using CCDs derived from MrBayes MCMC samples as before, a uniform prior on species tree topologies and a $\mathcal{N}(\log 2, 3)$ prior for the log-scale branch lengths), we find that the posterior concentrates on three different species tree topologies across multiple runs of the EP algorithm (fig. 5.15). We do not find that the CCD approximation to the posterior distribution of S has significant probability mass on all three trees at once, but rather that it tends to collapse on one of these, signaling potential issues in the EP approach for joint inference of (S, ϕ) in this use case. Increasing the number of particles in the SIS algorithm partly alleviates this problem, but in combination with the degeneracy of the posterior distribution the problem is hard to mitigate completely. This problem does not appear of course when we fix the species tree, so that we can verify suggested species trees by conducting model comparisons on the basis of estimated marginal likelihood values. We find that the three species tree topologies have near identical marginal log-likelihood estimates ($\hat{Z} \approx -1044$), suggesting equal posterior probabilities

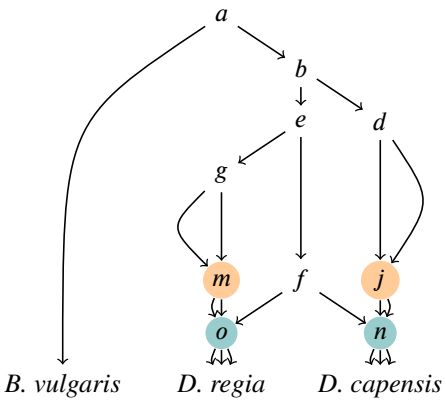


Figure 5.16: Example of a possible allopolyploid hybridization network compatible with the MUL species trees of high posterior probability under the MSC. Speciation nodes of putative iploid ancestors which can be represented in observed gene trees are labeled and unshaded, whereas allotetraploidization nodes are displayed in orange and allohexaploidization nodes in teal. Note that this is but one of the many phylogenetic networks compatible with the likely MUL species trees. The recent WGD event in *D. capensis* is not shown.

for the different species tree topologies. Moreover, we find, as expected, that the *D. capensis* subgenomes come in pairs, as a result of the recent allotetraploidy in this lineage. These results suggest a fairly distinct subclade of four *D. capensis* subgenomes, with the other subgenomes involved in a complicated knot. We note that this does seem to agree with our syntenic analyses, where we often find for a given *D. regia* segment two *D. capensis* segments with somewhat stronger conserved synteny and colinearity, although this is hard to quantify accurately without dedicated models. A graphical example of one of the many possible allopolyploid hybridization networks compatible with these MUL trees is displayed in fig. 5.16

Of course, even if we ignore the problems with multimodal posteriors, the statistical analysis presented in the preceding paragraphs is highly tentative, and should be interpreted cautiously. For one, it remains to be shown that the discordance in this data set is actually due to ILS and can be adequately modeled by the MSC. Moreover, we make the assumption that, while the species level phylogeny is a network, the subgenome phylogeny remains a tree. If allopolyploid hybridization among distinct lineages is common however, we may expect ongoing introgression to be common as well, which would violate this assumption. Another important, and also rather likely, model violation

would be gene conversion among homeologs (homologous genes on distinct subgenomes) within a species. Lastly, we have to mention the fact that the identification of homologs from distinct subgenomes is itself fairly complicated, and we should leave open the possibility that even in the completely unrealistic case where all modeling assumptions would hold for *true* homeologs, our sampled genes do not in fact conform to such an idealized set of loci. Clearly, if we want to adequately model these processes, we stand in need of more integrative models of genome evolution that are currently wanting. Such complicated models, if one is able to formally state them, are however not likely to have tractable likelihood functions (a simple model like the MSC even has no such thing!). What we hoped to illustrate here, then, is that for such complicated models, likelihood-free Bayesian inference (using EP or otherwise) may be a viable approach towards statistical inference.

5.5 Discussion

We have seen how Bayesian inference of phylogenomic models from gene tree distributions can be performed using efficient simulation algorithms, thereby bypassing the evaluation of the often intractable likelihood function of tree topologies associated with models of gene family evolution. In particular, we have explored a new approach based on expectation propagation which makes use of an exponential family over tree topologies that admits a sparse representation, allowing variational inference of species tree topologies under the MSC in a data-partitioned way. Although more work is needed for improving the stability of the proposed algorithms, in particular for applications to larger taxon sets, we believe the EP approach presents an interesting and viable strategy for learning about genome evolution from phylogenomic forests.

Note that, while we focused on the problem of species tree inference, the method presented above can also be used to infer *reconciled gene trees* (see the next chapter for more details). Indeed, the (weighted) sample of gene tree topologies in each individual site update provides an approximation of the posterior distribution over reconciled gene trees under the MSC model, given some empirical tree distribution. We relied on this feature in the present chapter indirectly when we sought to identify outlier loci which present likely model violations. We see that the proposed method is really somewhere in between the two main classes of commonly used methods outlined above, conducting joint Bayesian inference of species trees and gene trees, not using a full hierarchical model for the sequence data, but modeling only the variation

in gene tree topologies instead.

Importantly, relying on simulation, the proposed approach is rather flexible with regard to models of gene family evolution, which often do not have a tractable likelihood function for observed gene tree topologies. However, we have only considered MSC-like models of gene family evolution, which describe the genealogy of a *given* sample retrospectively, and hence can efficiently be simulated. In the case of forward-time evolutionary models, such as birth-death like models of gene family evolution within a species tree, we do not have simple methods for simulating gene trees *conditional on the sampled genes*, and we have to resort to less efficient means such as rejection sampling to simulate trees under the model which have non-zero probability to be compatible with the observed data. This makes the suggested methods unlikely to be useful for species trees and gene families of even moderate size for forward-time models. Designing new types of ABC kernels for trees that enable a meaningful quantification of the distance between trees of different sizes could be part of a solution to such issues, although it is certainly not obvious how this should be done. Nevertheless, despite these issues, likelihood-free methods like the one developed here should be among the prime candidates to consider for conducting statistical analyses of genome evolution under complicated models, such as the MLMSC in the sense of Li et al. (2021).

A major source of information which is not taken up in the approaches we have suggested here are the molecular distances associated with the branches of the gene trees. The only way by which this source of information presently enters the analysis is through the side-effect that short distances tend to imply gene tree uncertainty, and hence discordance of inferred gene trees with species trees. It remains an open question whether, and how, we can meaningfully incorporate distance estimates from phylogenetic CTMC models in the 'phylogenomic forest' point of view. We will take up this question again, but will not resolve it, in the next chapter, where we take the phylogenomic forests to the multi-copy gene family setting.

6 Bayesian gene tree reconciliation for multi-copy gene families

We continue our explorations in modeling genome evolution from the phylogenomic forest point of view introduced in chapter 4 and 5, but return to the multi-copy setting of chapter 3[1]. In contrast with the previous chapter, our focus will here be on the problem of *gene tree - species tree reconciliation*, or simply gene tree reconciliation for the sake of brevity, where we assume a *known* species tree S and seek to 'fit' the gene trees 'inside' S, assuming some genome-scale model of gene family evolution.

Gene tree reconciliation for multi-copy families has been a major approach for inferring homology relationships (e.g. Huerta-Cepas et al. 2007; Van Bel et al. 2018; Emms and Kelly 2019), for the analysis of genome evolution across deep time scales (Blomme et al. 2006; Hahn, Han, and Han 2007; Hahn 2007; De Smet et al. 2013; Li et al. 2016) and for the inference of ancient polyploidy (Jiao et al. 2011; Li et al. 2015; Marcet-Houben and Gabaldón 2015; McKain et al. 2016; Thomas, Ather, and Hahn 2017; Yang et al. 2018; Z. Li et al. 2018; Zwaenepoel and Van de Peer 2019a; Roelofs et al. 2020; Leebens-Mack et al. 2019). Most commonly used methods for reconciling gene trees with species trees do not rely on probabilistic models of gene family evolution nor statistical methods for inference, and are based on the principle of maximum parsimony under some scoring scheme or other, more or less *ad hoc*, methods. In addition, most methods for gene tree reconciliation assume the gene tree to be known without error (see also our discussion in the previous chapter), a likely problematic assumption which may lead to considerable biases in the inferred rates of gene family evolution or number of inferred evolutionary events

[1] This chapter draws freely from our published work in Zwaenepoel and Van de Peer (2019a). The method developed there was further used by us in Roelofs et al. (2020) and Chen et al. (2022), as well as by a number of other authors (two examples we know of are Wickell et al. (2021) and Liu et al. (2021)).

(e.g. gene duplication, loss and horizontal gene transfer events) (Hahn 2007). The pervasive combination of disacknowledging uncertainty in inferred gene trees (or doing so in *ad hoc* ways) and relying on naive reconciliation algorithms in phylogenomic studies has lead to a literature scattered with claims and inferences of which the reliability is hard to assess, with a number of controversies ensuing, in particular in the context of phylogenomic inference of ancient WGDs (e.g. Jiao et al. 2011; Li et al. 2015; Ruprecht et al. 2017; Ren et al. 2018; Zwaenepoel et al. 2019; Zwaenepoel and Van de Peer 2019a; Wang et al. 2019; Z. Li et al. 2018; Nakatani and McLysaght 2019; Roelofs et al. 2020; Leebens-Mack et al. 2019; Huang et al. 2020; Chen et al. 2022).

It seems wise then to take a step back and search for a statistically more adequate approach for gene tree reconciliation. In this chapter we develop methods for Bayesian gene tree reconciliation for multi-copy gene families using phylogenetic BDP models. Our approach is based on the amalgamated likelihood approximation due to Szöllősi, Rosikiewicz, et al. (2013), using a two-step method which makes use of the empirical conditional clade distribution of Larget (2013), defined in chapter 4. We will see that in this approach, the branching property is even more vital for efficient statistical inference, and in terms of the underlying models of gene family evolution we will hence stick to the phylogenetic birth-death processes of chapter 3. We will again devote special attention to the problem of modeling and inference of ancient WGDs in a phylogenetic context, describing in detail the method proposed in Zwaenepoel and Van de Peer (2019a). We conclude with a discussion of the pitfalls of the proposed methods and the challenges that remain.

6.1 Statistical gene tree reconciliation

6.1.1 From gene counts to gene trees

Recall that we have conceived of a genome, or a set of genomes, as a collection of genes partitioned into evolutionarily relevant subsets (the 'bag of genes' model), where the latter we have termed *gene families*. Genes within a gene family are assumed to be homologous, or more specifically, they are assumed to be descendants of an ancestral gene in some suitably chosen ancestor (see chapter 1). Moreover, we assume that a gene family contains *all* descendants of such an ancestral gene, including gene duplicates. Considering this collection of gene families as given data, we have, in chapter 2 and 3, considered evolutionary models that could generate those data, and attempted

to confront the data with these models to estimate parameters which inform us about genome evolutionary processes. In the latter step, we have previously, however, tacitly assumed that we possess no more information than the gene content of each family. When we label each gene by its respective genome $k \in [1..m]$, the observed data consisted of nothing more than n multisets with elements from $[1..m]$, which we have usually summarized in an $n \times m$ count matrix recording the multiplicity for each species in each family.

Obviously, such an approach wilfully ignores a tremendous amount of information that genomic data provides about the evolutionary models we are considering. Specifically, all stochastic models we considered generate *locus trees* or *gene trees* (see chapter 1), i.e. phylogenetic trees with as leaves the relevant gene family members.[2] If we are prepared to assume some model of sequence evolution, extant observed gene sequences provide information about the gene tree (this is of course the fundamental idea of statistical molecular phylogenetics), and hence, if molecular sequence evolution occurs on similar time scales as gene family evolution (which appears to be the case, see e.g. chapter 2), sequence data provides information about gene family evolution. Of course, the reverse is also true, and considering the evolution of gene families within a species tree may greatly inform the inference of gene trees from sequence data (Szöllősi, Rosikiewicz, et al. 2013; Szöllősi et al. 2015).

When modeling gene content alone using phylogenetic BDP models, we have implicitly marginalized the likelihood over all tree topologies and node ages that could have generated the observed phylogenetic profile x. In other words, for a species tree S and parameters of a model of gene family evolution θ, the likelihood function used in statistical inference from gene counts can be expressed as

$$p(x|S,\theta) = \sum_{G \in \mathcal{G}_x} \int p(G,t|S,\theta)dt$$

Where $\mathcal{G}_x$ denotes the set of gene tree topologies with extant leaves compatible with the profile x, and t the vector of branch lengths of the gene tree. Note that the algorithms described for inference for phylogenetic BDPs from gene count data do not in fact perform this marginalization, but instead make use

[2]Note that virtually all models used in practice generate *either* a locus tree *or* a gene tree, assuming some sort of constraint on the one which is not explicitly modeled. For instance, the MSC model considered in the previous chapter generates a gene tree (genealogy), and it is assumed that the locus tree is identical to the species tree. On the other hand, the phylogenetic BDP models considered in the present chapter generate locus trees, and it will be assumed that a gene tree is determined by its associated locus tree. In the latter case, we refer to the locus tree as a gene tree, following a common abuse of terminology.

of the transient distributions of the BDP process to marginalize over ancestral states along the internal nodes of S.

Taking into account the associated sequence data y admits a more detailed modeling approach if we are prepared to make assumptions about the molecular evolutionary process. The likelihood function for aligned sequence data y can be expressed as

$$p(y|\phi, \theta, S) = \sum_{G \in \mathcal{G}_y} \int p(y|G, t, \phi) p(G, t|S, \theta) dt \tag{6.1}$$

where ϕ represents the (typically vector-valued) parameter for the model of sequence evolution. Here $p(y|G, t, \phi)$ is the usual likelihood for phylogenetic CTMC models, while $p(G, t|S, \theta)$ is the likelihood of the gene tree under the model of gene family evolution. Importantly, for sufficiently rich models of multi-copy gene family evolution, statistical methods based on eq. 6.1 not only provide a means for inference of S and θ, but also allow the statistical inference of gene tree topologies and implied homology relations, and in principle allow doing all this jointly (barring computational limitations, of course). Note that most models of gene family evolution generate time-calibrated gene trees, so that one cannot bypass the specification of a molecular clock model in order to compute $p(y|G, t, \phi)$ (i.e. ϕ includes substitution rate parameters). Likelihood-based methods (Bayesian or otherwise) for inference of phylogenetic BDPs and related models of gene family evolution explicitly based on eq. 6.1 have been scarce, but Prime-DLRS and related methods (Åkerborg et al. 2009; Sjöstrand et al. 2012; Mahmudi et al. 2013; Ullah et al. 2015) provide a notable example.

In the remainder of this chapter we will deal with this problem from the phylogenomic forest point of view, that is, we will not deal with the model of sequence evolution explicitly, but will assume that we have a probability distribution over gene trees at our disposal which adequately describes the probability of a gene tree given the observed sequence data under some phylogenetic CTMC model of sequence evolution. Roughly, the idea, due to Szöllősi, Rosikiewicz, et al. (2013), is to replace eq. 6.1 by

$$p(y|\theta, S) = \sum_{G \in \mathcal{G}_y} p(y|G) p(G|\theta, S) \approx \sum_{G \in \mathcal{G}_y} \xi_y(G) p(G|\theta, S) \tag{6.2}$$

Where $\xi_y(G)$ is an estimate of the posterior probability of G given the sequence data y under a phylogenetic CTMC model of sequence evolution. In this set-

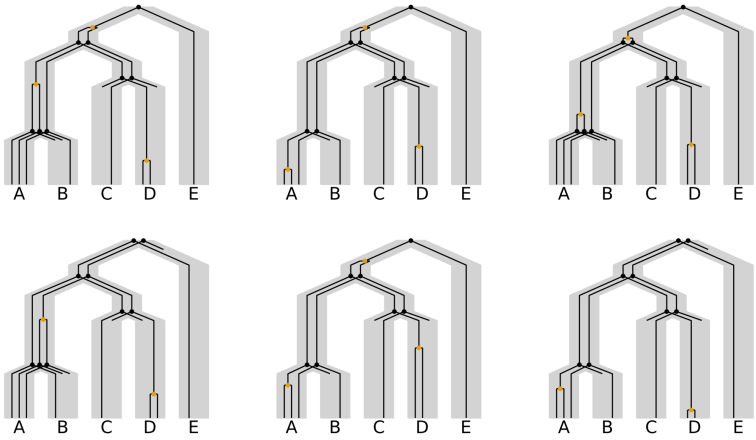


Figure 6.1: Example of reconciled gene trees. We show six different reconciliations for a single gene tree topology (sampled iid from a phylogenetic BDP model using methods developed below). The species tree is shown with broad gray branches, whereas the gene tree is displayed in black within the species tree. In this particular reconciliation, each node of the gene tree is either associated with a duplication event (orange), a speciation event (black) or a loss event. Note that four out of six sampled reconciled trees are different.

ting however, similar challenges as in the MSC problem appear, namely, computing $p(G|S,\theta)$ efficiently so that the sum in eq. 6.2 can be completed in reasonable computational time. Computing $p(G|S,\theta)$, the probability of observing a particular gene tree topology for a given species tree and model of gene family evolution, constitutes the core problem of *probabilistic gene tree reconciliation*. Before delving into this, we will need to be more precise about what we understand by a reconciled gene tree.

6.1.2 Reconciled gene trees

6.1.2.1 Informal description

The concept of gene tree - species tree reconciliation, together with parsimony-based strategies for inferring reconciled gene trees given a known species tree, stems from Goodman et al. (1979), and was elaborated on considerably in Page (1994) and Page and Charleston (1997), before becoming a core concept in comparative genomics in the post-genomic era. The title of the pa-

per by Goodman et al., "Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences", is as good as any informal description of the goal of gene tree reconciliation. As individual gene trees differ from the species tree in general (and necessarily so for multi-copy gene families), the challenge to explain the evolutionary history of the gene family, as represented by the gene tree, with respect to the evolutionary history of the associated populations, as represented by the species tree, comes up naturally. We can think of providing such an explanation as 'fitting' the gene tree within the species tree in some way (see fig. 6.1 for an illustration). From the perspective of the three-tree model introduced in chapter 1 (Rasmussen and Kellis 2012), this essentially boils down to inferring the locus tree conditional on the gene tree and species tree.

Of course, to do this in a systematic fashion, we need some model of gene family evolution which postulates a set of possible locus-level *evolutionary events* in terms of which we are allowed to explain the gene tree, such as for instance gene duplication, gene loss, horizontal gene transfer, *de novo* gene origin and deep coalescence[3]. Given such a model, we then need an approach to infer a plausible evolutionary history for the gene family in terms of these evolutionary events in a way compatible with both the gene tree and species tree. If our model is probabilistic in nature, such as a phylogenetic BDP model, we can use likelihood-based statistical inference methods to do so, if not, the typical approach is to devise some *ad hoc* scoring scheme and use maximum parsimony with (possibly heuristic) optimization algorithms (as in e.g. the pioneering work of Goodman et al. (1979)). Despite considerable advances in statistical phylogenetics, the latter approach is still far more common in the context of gene tree reconciliation. Particularly common is the parsimony approach which seeks to minimize the number of gene duplication and loss events required to fit the gene tree inside the species tree. As is clear from our example in fig. 6.1, where we show reconciliations sampled under a probabilistic phylogenetic BDP model, there are many possible reconciliations for a given gene tree, of which the most parsimonious one is but one example, a fact that is not always well-appreciated.

[3]In the previous chapter, where we did not study multi-copy gene families, the only event available to explain the (possibly discordant) relation between the gene tree and species tree was deep coalescence. We admit that it is slightly awkward to refer to deep coalescence as an evolutionary event distinct from speciation, but we trust that the reader is prepared to grant us this common abuse of terminology.

6.1.2.2 Definitions and notation

It will be helpful to make some definitions and introduce some suitable notation, some of which repeated from chapter 4 and 5. Note that as before a gene family will always correspond to an orthogroup. Let $G(V, E)$ be the graph representation of a tree topology $\mathcal{T}$, with $V = V(\mathcal{T})$ the set of nodes (or vertices) and $E = E(\mathcal{T})$ the set of branches (or edges) of the tree. We shall again denote the set of leaves of a tree $\mathcal{T}$ by $\mathcal{L}(\mathcal{T})$. In a rooted tree each branch $\langle u, v\rangle \in E(\mathcal{T})$ can be identified with its target node v. For a rooted tree $\mathcal{T}$, we denote by $\mathcal{T}_v$ the subtree of $\mathcal{T}$ rooted in the node v. A tree may or may not be associated with *branch lengths* which formally correspond to a mapping $l\colon E(\mathcal{T}) \to \mathbb{R}^+$. We write l_v for the length of the branch $\langle u, v\rangle$ leading to node v of a rooted tree. We recall that a timetree is a phylogenetic tree where branch lengths are interpreted to measure time intervals on some suitable scale. As before, by a *species tree* S, we will usually refer to a timetree where the leaves represent extant species and internal nodes speciation events. We define a gene tree in the same way as we did in the previous chapter, that is:

Definition. A *gene tree* (G, σ) is a binary rooted tree associated with some species tree S together with a gene-to-species map $\sigma\colon \mathcal{L}(G) \to \mathcal{L}(S)$, associating with each gene u the species $\sigma(u)$ from which it derives.

For our purposes in the present chapter, we employ the following formal definition of a gene tree reconciliation:

Definition. A *reconciled gene tree* $\mathcal{R}$ is a tuple (G, ρ) where G is a gene tree associated with species tree S and $\rho\colon V(G) \to E(S)\times\mathbb{R}^+$, is the *reconciliation map*, where $\forall u \in \mathcal{L}(G)\colon \rho(u) = (\sigma(u), 0)$.

In words, ρ maps a gene tree node to a branch of the species tree and a time point along that branch, where we associate time point 0 with the target node of the branch. Note that we could alternatively use a definition where we assign to each time point along S a unique number in $[0, T]$ where T is the total tree length, and have a reconciliation map $\rho\colon V(G) \to [0, T]$, however, for our purposes it will be more convenient to take up the species tree branches explicitly in ρ.

With the above definition of a reconciliation, we have that, when $\rho(u) = (e, 0)$ for a gene tree node u, the node is mapped to the species tree *node* e and represents divergence through speciation (fig. 6.1). When $\rho(u) = (e, t)$ with $t \neq 0$, u is mapped to a *branch* of S, in which case the gene tree node reflects a dupli-

cation event. The gene tree as we have defined it is essentially an incomplete reconciliation, where ρ is only defined for the leaf nodes of the tree. Note that there are more general ways to formally define a reconciliation, but all of them of course consider some mapping from gene tree to species tree nodes or *vice versa*. In particular, when the model admits more evolutionary events than duplication and loss (e.g. horizontal gene transfer or deep coalescence), the above definition will not suffice. We will not, however, deal with such models in the present chapter, so that the definition formulated here, similar to the definition in Arvestad, Lagergren, and Sennblad (2009), will suffice for our needs.

6.2 Gene tree reconciliation for the phylogenetic linear BDP

At this point, it may be unclear what purpose the gene tree reconciliation has when our immediate goal is the calculation of $p(G|\theta, S)$, i.e. the probability of observing a gene tree *topology* given a phylogenetic BDP with parameters θ. However, note that for a gene tree topology G, we can write

$$p(G|\theta, S) = \sum_{\rho \in S^{V(G)}} p(G, \rho|\theta, S) \tag{6.3}$$

where $S^{V(G)}$ is a slight abuse of notation for denoting the set of all possible reconciliation maps. For a phylogenetic linear BDP, $p(G, \rho|\theta, S)$ can be computed relatively easily. In words, then, eq. 6.3 suggests that to compute the likelihood of a gene tree topology G under some model of gene family evolution, we can marginalize over all possible ways to fit G in the species tree. We now proceed by describing an algorithm to perform this marginalization for an observed gene tree G under the linear phylogenetic BDP model. We shall develop the theory for rooted gene trees, but will relax this assumption later when we deal with gene tree uncertainty[4].

[4]Indeed, an unrooted tree can be thought of as a probability distribution over rooted trees, where we assume each possible rooting equally probable *a priori*. An algorithm for dealing with unrooted gene trees will hence be a simple corollary of our treatment of reconciliation with gene tree uncertainty below.

6.2.1 The gene tree likelihood

6.2.1.1 Differential equations for the phylogenetic BDP

For a linear BDP model of gene family evolution by gene duplication and loss, with duplication rate λ and loss rate μ, we can compute $p(G|S, \lambda, \mu)$ by solving a system of ODEs recursively. To see this, let $p_e(u,t)$ with $t \in [0, l_e]$ be the probability that the lineage (gene tree branch) leading to gene tree node $u \in V(G)$ passes through branch e at a distance t from the endpoint of e, and note that

$$\begin{aligned} p_e(u, t+\Delta t) &= (1-\mu\Delta t)\times \\ &\Big[\underbrace{(1-\lambda\Delta t)p_e(u,t)}_{\text{propagation}} + \underbrace{\lambda\Delta t p_e(v,t)p_e(w,t)}_{\text{represented duplication}} + \underbrace{2\lambda\Delta t p_e(u,t)\epsilon_e(t)}_{\text{duplication and loss}}\Big] + o(\Delta t) \\ &= \big(1-(\mu+\lambda)\Delta t\big)p_e(u,t) + \lambda\Delta t\big(p_e(v,t)p_e(w,t) + 2p_e(u,t)\epsilon_e(t)\big) + o(\Delta t) \end{aligned} \tag{6.4}$$

Where v and w are the daughter nodes of u, and $\epsilon_e(t)$ is the probability that a lineage present at time t in branch $e \in E(S)$ leaves no observed descendants (the *extinction probability*). If u is a leaf node, we assume the terms involving v and w to be zero. The extinction probabilities can again be computed straightforwardly using the probability generating function for the linear BDP and a postorder traversal of S (as in chapter 3). The rationale for the different terms in eq. 6.4 is illustrated in fig. 6.2.

We can of course derive an ODE from the above, which will define the $p_e(u,t)$ for a given species tree branch e recursively. Subtracting $p_e(u,t)$ from both sides of eq. 6.4, dividing by Δt, and taking the limit as the latter goes to 0, we obtain

$$\frac{dp_e(u,t)}{dt} = -(\lambda+\mu)p_e(u,t) + \lambda\big(p_e(v,t)p_e(w,t) + 2p_e(u,t)\epsilon_e(t)\big) \tag{6.5}$$

Unsurprisingly, these ODEs are very similar to those used in macroevolutionary studies, where one seeks to fit birth-death process models to dated species trees (e.g. Nee et al. 1994; Maddison, Midford, and Otto 2007; Rabosky 2014). The system of ODEs in eq. 6.5 can be solved for the complete species tree by specifying a suitable set of boundary conditions. Specifically, we have that

for nodes $e \in \mathcal{L}(S)$

$$p_e(u,0) = \begin{cases} 1 & \text{if } u \in \mathcal{L}(G) \wedge \sigma(u) = e \\ 0 & \text{else} \end{cases}$$

and for internal nodes e of S

$$p_e(u,0) = \underbrace{p_f(v,l_f)p_g(w,l_g) + p_f(w,l_f)p_g(v,l_g)}_{\text{represented speciation}} + \underbrace{p_f(u,l_f)\epsilon_g + p_g(u,l_g)\epsilon_f}_{\text{speciation and loss}} \tag{6.6}$$

Where in the latter equation we again assume that the terms involving v and w are 0 when $u \in \mathcal{L}(G)$. Here, $\epsilon_e = \epsilon_e(l_e)$ is the probability that a single lineage present at the start of species tree branch e has no observed descendants (the extinction probability). The ϵ_e values can be computed recursively using the probability generating function of the process in a postorder traversal of the species tree (see chapter 3). Using this recursive set of boundary conditions, one can solve the system of ODEs numerically along the species tree using dynamic programming over the gene tree nodes in a post-order traversal of S. In this way we can compute for each gene tree node $u \in V(G)$ the probability that the lineage leading to u 'passes through' branch e of the species tree at time t.

6.2.1.2 Recursions for the root

Special care is needed for the probabilities at the root node o of S. Assuming a single gene at the root, we see that $p(G|S,\lambda,\mu) = p_o(x,0)$, where x is the root node of the gene tree. Of course, we are usually in no position to make such a stringent assumption (chapter 1). A straightforward way to account for this is to add a 'virtual branch' of a certain length l_o leading to the root of S and extend the phylogenetic BDP model accordingly (as in e.g. Szöllősi et al. (2012); Szöllősi, Rosikiewicz, et al. (2013)). Assuming a single lineage at the beginning of the virtual root branch, the intra-branch system of ODEs can be solved again to obtain $p_o(x,l_o)$ as the likelihood of the gene tree topology under the phylogenetic BDP model. Given that the transient distribution of a linear BDP conditioned on non-extinction is a geometric distribution, such an approach is equivalent to assuming a geometric prior on the number of lineages at the root. Choosing the length of the virtual branch (or the duplication and loss rate of the associated BDP) is however a rather awkward way of

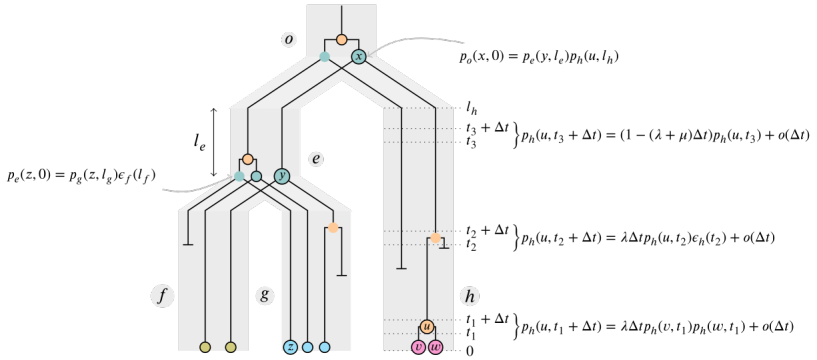


Figure 6.2: Illustration of the rationale behind the recursive algorithm for computing the likelihood of a reconciled tree under a linear phylogenetic BDP model. The dark gray rectangles represent the species tree branches, with branch labels shown in the associated gray circles. The reconciled tree, with duplication nodes in orange and speciation nodes in green, is drawn inside the species tree. Nodes with a black border represent nodes in the observed (unreconciled) gene tree G. We show an example for each of the evolutionary events that can occur in a short time slice Δt with their associated probabilities under the phylogenetic BDP model.

specifying a prior distribution for the number of lineages at the root.

It would be desirable to be able to specify the prior on the number of genes at the root in a more flexible way. Let ϵ_o denote the probability that a single gene at o does not leave observed descendants at the leaves of S, and let Y_o be the number of lineages at o which do leave observed descendants, assuming a geometric prior on the number of genes at the root, we already noted in chapter 3 that for a geometric prior with parameter η we have

$$f(k) = \mathbb{P}\{Y_o = k\} = (1-\epsilon_o)^k \frac{\eta(1-\eta)^{k-1}}{(1-(1-\eta)\epsilon_o)^{k+1}}$$

Note that this probability can be expressed recursively using

$$f(1) = \frac{\eta(1-\epsilon_o)}{\tilde{\eta}^2}$$
$$f(k) = \frac{1-\eta}{\tilde{\eta}} f(s-1)$$

and where $\tilde{\eta} = (1 - (1 - \eta)\epsilon_o)$. Note furthermore that

$$f(i + j) = \frac{(1 - \eta)\tilde{\eta}}{\eta} f(i) f(j)$$

Using this bit of algebra, we can devise a recursion for computing the probability $p_o(u)$ of the gene tree rooted in node u under the phylogenetic BDP model with a geometric prior on the ancestral gene family size. Specifically, let f and g be the child nodes of the root o in S, for internal gene tree node u with child nodes v and w

$$p_o(u) = \underbrace{\frac{(1 - \eta)\tilde{\eta}}{\eta} p_o(v) p_o(w)}_{\text{duplication before the root}} + \frac{\eta(1 - \epsilon_o)}{\tilde{\eta}^2} \Big(\underbrace{p_f(v, l_f) p_g(w, l_g) + p_f(w, l_f) p_g(v, l_g)}_{\text{root speciation}} + \underbrace{p_f(u, l_f)\epsilon_g(l_g) + p_g(u, l_g)\epsilon_f(l_f)}_{\text{root speciation and loss}} \Big)$$

and similarly for a leaf node u of G

$$p_o(u) = \frac{\eta(1 - \epsilon_o)}{\tilde{\eta}^2} \big(p_f(u, l_f)\epsilon_g(l_g) + p_g(u, l_g)\epsilon_f(l_f) \big)$$

Given the $p_e(u, t)$, which we obtain by solving the system eq. 6.5 along S, we can hence compute the $p_o(u)$ using a postorder traversal over G. The marginal likelihood of the gene tree topology rooted in node u under the assumed phylogenetic BDP with geometric prior on the root family size is then given by $p_o(u)$.

6.2.1.3 Whole-genome duplications

Unsurprisingly, adapting the above recursions to the DLWGD model of Rabier, Ta, and Ané (2014) (see chapter 3) is straightforward. As in our previous treatment of the phylogenetic DLWGD model, we again introduce an additional node in the species tree to mark a WGD event. Given the pgf for the transition across the WGD node, $f(s) = (1-q)s+qs^2$, extinction probabilities

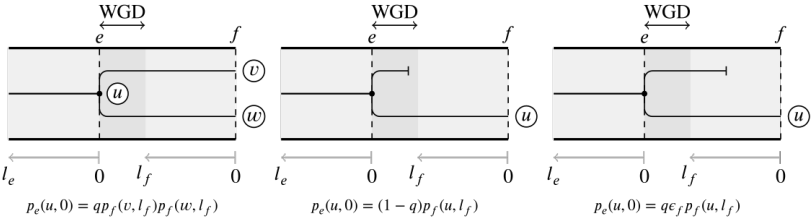


Figure 6.3: Illustration for the reconciliation likelihood recursion at a WGD node under the DLWGD model of Rabier, Ta, and Ané (2014). The species tree branch is shown in gray, e is the WGD node, and f is its daughter node, both marked by vertical dashed lines. The u, v and w labels mark gene tree nodes. Each of the three scenarios corresponds to one of the terms in eq. 6.7.

are still straightforwardly calculated in a single postorder traversal of S. If e is a WGD node and f is its single daughter node, we will have for each gene tree node u with children v and w

$$p_e(u,0) = \underbrace{qp_f(v,l_f)p_f(w,l_f)}_{\text{represented retention}} + \underbrace{(1-q)p_f(u,l_f)}_{\text{non-retention}} + \underbrace{2q\epsilon_f p_f(u,l_f)}_{\text{retention and loss}} \tag{6.7}$$

The evolutionary events corresponding to the three terms in this expression are illustrated in fig. 6.3. Again we use the same expression for a leaf node u, assuming in that case the terms involving v and w to equal zero. Note that dealing with higher multiplication levels in the case of gene trees would require a more detailed model of the polyploidization history. Indeed, a simplistic model which assumes that a WGM induces a multifurcation in the gene tree will not work, since observed gene trees will, usually, be strictly bifurcating. Furthermore, as we have noted before in chapter 3, such a model is biologically awkward anyway. Also note that, as before, this is in the first place a model for autotetraploidy, but can serve as a reasonable model for allopolyploidy as well. However, in the latter case, we necessarily assume that the divergence between the involved subgenomes postdates the closest speciation event above the WGD node in S. In addition, we assume that the time interval between the divergence of the subgenomes and their merger through allopolyploidy is short, so that the probability of a duplication or loss event occurring in this interval is negligible.

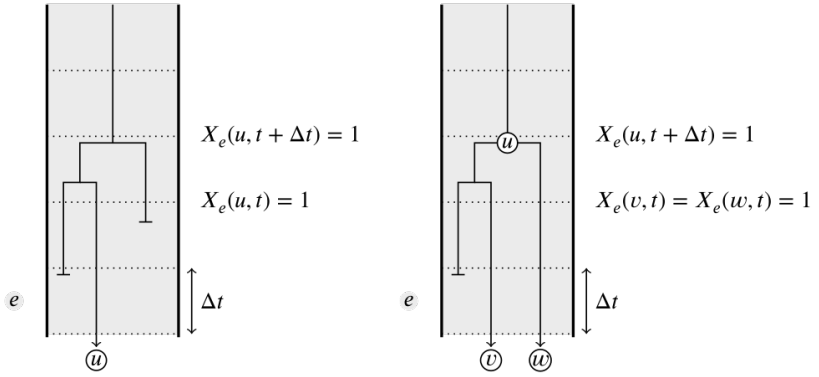


Figure 6.4: Illustration of the discretized calculation of the reconciliation likelihood for linear BDP models. In the discretized model, we assume that, for each branch e of the species tree (here depicted as a gray rectangle), at most one *represented* duplication event occurs within a time slice of length Δt. Under this assumption, we can calculate the probability of observing the lineage leading to gene tree node u for the boundaries of each time slice recursively from the tipward slice to the rootward slice. The left diagram illustrates the propagation of a single lineage through a time slice, despite duplication events occurring within the slice. The right diagram shows a similar scenario for a represented duplication. See the main text for further details.

6.2.1.4 Discretization

Numerically solving the recursive set of ODEs with high accuracy could become computationally prohibitive for phylogenomic data sets. We can however use theoretical results for the linear BDP model to approximate the likelihood using a coarse discretization without sacrificing too much numerical accuracy.

Consider a species tree branch e of length l_e, and divide the branch in n_e time slices of length Δt (fig. 6.4). We will take another viewpoint on the random process associated with the gene tree reconciliation, already implicit in our derivation of the ODEs for the marginal likelihood above. Let $X_e(u, t)$ be the binary random variable[5] which takes the value one when the lineage leading to gene tree node u passes through branch e of S at time t, and zero otherwise. For a discretized species tree branch with time slices of length Δt, the probability

[5]Given that the tuple (e, t) marks the time point along the phylogeny S, it may seem more natural to adopt a notation like $X_u(e, t)$. We however retain the suffix referring to the species tree branch, to stress the independence (and compositionality) of the evolutionary processes in distinct species tree branches. Perhaps $X_{e,u}(t)$ would be even more natural, but also typographically uglier.

of propagating a single gene lineage through the slice (see fig. 6.4, on the left) is

$$\begin{aligned}\phi_e(t, t+\Delta t) &= \mathbb{P}\{X_e(u, t+\Delta t) = 1 | X_e(u,t) = 1\} \\ &= \sum_{k=0}^{\infty} \binom{k+1}{k} p_{1,k+1}(\Delta t)\epsilon_e(t)^k \\ &= \frac{(1-\alpha(\Delta t))(1-\beta(\Delta t))}{(1-\beta(\Delta t)\epsilon_e(t))^2}\end{aligned}$$

where α and β were defined in chapter 2. Similarly[6], in the case of a duplication, for child nodes v and w of u, we have

$$\begin{aligned}\psi_e(t, t+\Delta t) &= \mathbb{P}\{X_e(u, t+\Delta t) = 1 | X_e(v,t) = 1, X_e(w,t) = 1\} \\ &= \sum_{k=0}^{\infty} \binom{k+2}{k} p_{1,k+2}(\Delta t)\epsilon_e(t)^k \\ &= \phi_e(t, t+\Delta t)\frac{\beta(\Delta t)}{(1-\beta(\Delta t)\epsilon_e(t))}\end{aligned}$$

The latter scenario is also illustrated in fig. 6.4 (right hand side). We note that these results can also be derived using a recursive argument, as we did in chapter 3.

Using these results, we can approximate the likelihood by replacing the within branch ODE of eq. 6.5 by the following recursion

$$p_e(u, t+\Delta t) = \underbrace{\phi(t, t+\Delta t)p_e(u,t)}_{\text{propagation}} + \underbrace{\psi(t, t+\Delta t)p_e(v,t)p_e(w,t)}_{\text{represented duplication}}$$

With the inter-branch recursions given by eq. 6.6 as before. It is important to stress that this does not merely provide a coarse discrete approximation to the ODEs of the continuous-time linear BDP model, but rather an exact calculation for an approximate *model*, in which we assume that at most a single *represented* duplication occurs within a time slice of length Δt (fig. 6.4). In fig. 6.5 we compare the discretized model likelihood against the ODE solution. As ex-

[6] As an aside, we note that while these kind of manipulations of sums may be utterly trivial for mathematicians, computer scientists and physicists, they can appear somewhat magical to the uninitiated (at least that used to be the case for your humble author). This is a great opportunity, however, to mention the wonderful book by Graham et al. (1989), where one can learn the tricks of the trade.

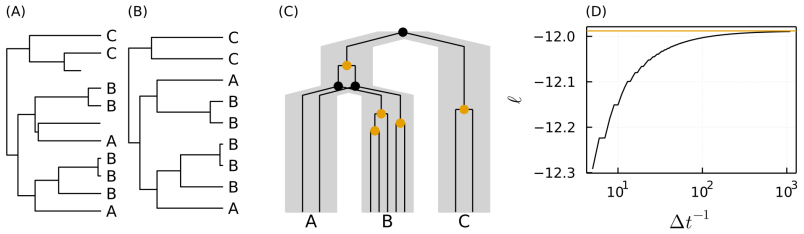


Figure 6.5: Comparison of the discretized model likelihood to the solution of the ODE system for the phylogenetic linear BDP marginal likelihood for a simulated gene tree. (A) Simulated gene tree for a phylogenetic linear BDP model. (B) Reconstructed gene tree (with extinct lineages pruned). (C) Example reconciliation for the reconstructed gene tree, sampled from the linear phylogenetic BDP model. (D) In black, the discretized marginal likelihood ℓ for decreasing time slice length Δt is shown, whereas the likelihood obtained by solving the ODE system of sec. 6.2.1.1 is indicated by the orange line.

pected, the ODE solution provides an upper bound for the discretized model, since the latter assumes the exact same probabilistic model, but restricts the number of admissible scenarios for fitting the gene tree in the species tree to those with at most a single represented duplication in one time slice.

6.2.1.5 Conditioning on the sampling process

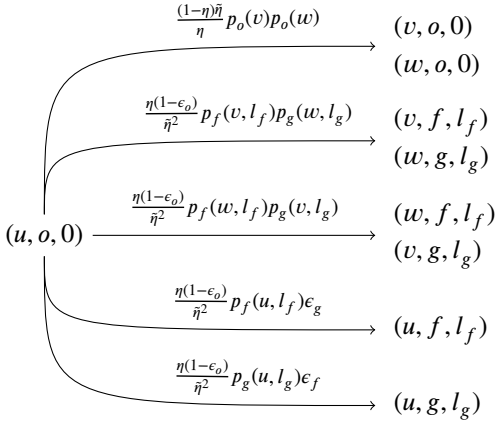
As was already discussed in chapter 3, it is important to correct the marginal likelihood in accordance with the sampling process when the latter is biased. We shall for instance, as in the mentioned chapter, usually deal with families which have at least one observed descendant in both clades stemming from the root of the species tree. The necessary conditioning factor can however be computed exactly as before, and we refer the reader to chapter 3 for the relevant details.

6.2.2 Sampling reconciled trees

The above theory allows the calculation of the marginal likelihood of a gene tree topology G given a species tree and a phylogenetic BDP model of gene family evolution, and hence likelihood-based inference for the parameters of the phylogenetic BDP (and S, if desired) when we assume G to be observed. However, at this point it is not yet clear how to conduct inference of reconciled

gene trees using the framework we are developing. In the discretized model, where the algorithm to compute the gene tree likelihood has essentially been cast in a common dynamic programming form, we can easily sample reconciliations for a given gene tree topology from the probability distribution induced by a phylogenetic BDP model by means of a stochastic backtracking algorithm. That is, conditional on a phylogenetic BDP model with species tree S and parameters θ, we can sample from $p(\rho|G, \theta, S)$. Using such samples, we can estimate quantities of interest like the expected number of represented duplications on a branch of the species tree, the probability that two genes are paralogs rather than orthologs with respect to some ancestor, the probability that a gene pair of interest derives from a duplication event in a particular branch of the species tree, *etc.*

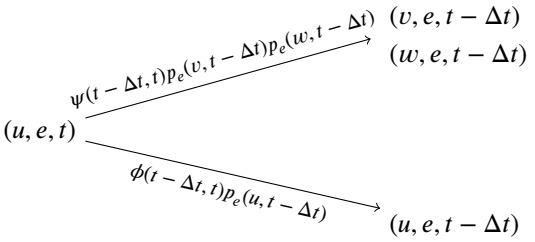
Assume that we have computed the $p_e(u, t_i)$ for each $u \in V(G)$, $e \in E(S)$ and $i \in [1..n_e]$. The stochastic backtracking algorithm will grow a reconciled tree from the root to the tips in a preorder traversal of S. Let $Y = (Y_G, Y_S, Y_t)$ denote the state of the backtracking algorithm. First, the part of the gene tree reconciled within the root of S is sampled in accordance with the geometric model outlined above. Given that gene tree node u is reconciled to the root node o of S, i.e. the state of the backtrace is $Y = (u, o, 0)$, there are three classes of admissible events (duplication before the root, represented speciation at the root and speciation followed by loss at the root), which are each sampled with a certain probability, leading to an update of the state Y according to the following diagram



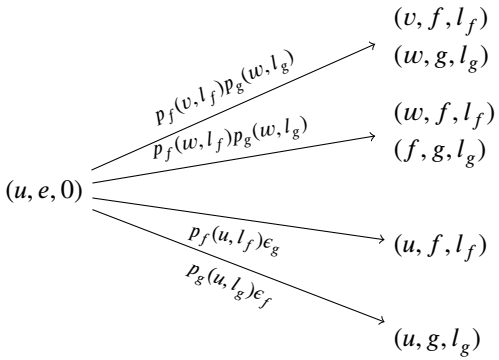
Where the expressions on top of the arrows mark the (unnormalized) probability of the associated transition. Here, the first three cases imply a bifurcation

in the gene tree, while the last two imply an unobserved speciation node in the gene tree. The backtracking algorithm is then recursively applied to the new state(s).

When $Y_t \neq 0$, the next state, involving either duplication or propagation of a single lineage, is sampled according to the following diagram



At speciation (internal) nodes in S, i.e. when $Y_t = 0$, $Y_S \neq o$ and $Y_S \notin \mathcal{L}(S)$, a represented speciation or speciation with subsequent loss occurs, which amounts to sampling one of the following transitions



The recursion terminates at leaf nodes of the species tree. Note that under the assumption that there is at least one observed descendant in each clade stemming from the root, the gene tree root x is reconciled with probability 1 to the root node o of the species tree, i.e. $\mathbb{P}\{X_o(x, 0) = 1\} = 1$, so that we can initiate the recursive sampling algorithm from the state $(x, o, 0)$. The reconciled gene trees shown in fig. 6.1 were sampled using this algorithm. Clearly, the backtracking algorithm is readily extended to the DLWGD model, or any other model with similar structure.

6.2.3 Taking into account gene tree uncertainty

In the previous section we outlined an approach to compute $p(G|S,\theta)$, the probability of observing a gene tree topology G given a species tree S and a phylogenetic linear BDP model (with or without WGD events) with parameters θ. This is of course still a far cry from computing $p(y|\theta,S)$ for sequence data y (eq. 6.2). Szöllősi, Rosikiewicz, et al. (2013) however noted how the framework for computing the marginal reconciliation likelihood outlined above can be adapted straightforwardly to deal with uncertainty in the gene tree when the latter can be adequately captured by a conditional clade distribution (CCD, see chapter 4). That is, they describe an algorithm, called *amalgamated likelihood estimation* (ALE), that allows computing

$$p(y|\theta,S) \approx \sum_{G\in\mathcal{G}_y} \xi_y(G)p(G|\theta,S)$$

whenever the gene tree distribution ξ_y is a CCD. When this CCD has restricted support, as in the case of a (typical) empirical CCD (which is used in the paper by Szöllősi, Rosikiewicz, et al. (2013)), the computational complexity of the algorithm is only slightly increased with respect to the fixed G case, and the increase is proportional to the degree of gene tree uncertainty.

We present the adaptation of the system of ODEs for the exact reconciliation likelihood under the phylogenetic BDP (eq. 6.5). The adaptation for the discretized model is analogous. Let, similar to eq. 6.5, $p_e(\gamma,t)$ be the probability that the lineage leading to *clade* γ passes through time point t along branch e of S, and let, following chapter 4, $\theta_{\gamma,\delta}$ be the probability of observing the split $(\delta,\gamma-\delta)$ of clade γ under the CCD. By the conditional independence property for disjoint subtrees under the CCD model, it is easy to see that

$$\frac{dp_e(\gamma,t)}{dt} = -(\lambda+\mu)p_e(\gamma,t) + 2\lambda p_e(\gamma,t)\epsilon_e(t) + \lambda\sum_{\delta\subset\gamma}\theta_{\gamma,\delta}p_e(\delta,t)p_e(\gamma-\delta,t)$$

and that for an internal branch e of S with daughter branches f and g, we have, analogous to eq. 6.6,

$$\begin{aligned} p_e(\gamma,0) = \sum_{\delta\subset\gamma}\theta_{\gamma,\delta}\big[&p_f(\delta,l_f)p_g(\gamma-\delta,l_g) + p_f(\gamma-\delta,l_f)p_g(\delta,l_g)\big] \\ &+ p_f(\gamma,l_f)\epsilon_g + p_g(\gamma,l_g)\epsilon_f \end{aligned}$$

with again, for a tip branch e of S, $p_e(\gamma,0)$ equal to 1 when γ is a leaf node of

G with $\sigma(\gamma) = e$, and 0 otherwise. In the DLWGD model, we will of course have an analogous counterpart to eq. 6.7

$$p_e(\gamma, 0) = q \sum_{\delta \subset \gamma} \theta_{\gamma,\delta} p_f(\delta, l_f) p_f(\gamma - \delta, l_f) + (1 - q) p_f(\gamma, l_f) + 2q\epsilon_f p_f(\gamma, l_f)$$

Szöllősi, Rosikiewicz, et al. (2013) further showed how to extend the recursions for the discretized model to a phylogenetic BDP model with horizontal gene transfer (their ODT and exODT models; Szöllősi et al. 2012; Szöllősi, Tannier, et al. 2013). We will not deal with horizontal gene transfer in the present work however.

Clearly, to sample reconciled trees, we can use a similar backtracking algorithm for the discretized model as before, now not only determining the reconciliation map ρ through the backtrace, but also resolving the gene tree topology G. Specifically, conditional on the phylogenetic BDP model, we can sample approximately from $p(\mathcal{R}|y, \theta, S)$, and hence conduct joint inference of gene trees and their reconciliations under a phylogenetic BDP model for a known species tree. Recall furthermore that a CCD can represent a distribution over *unrooted* tree topologies (chapter 4), so that we can conduct reconciliation of unrooted trees using this approach. Hence, the ALE approach allows joint inference of the gene tree topology, the location of the root, and the homology relationships among gene family members under a phylogenetic BDP model of gene family evolution, based on a sample of tree topologies obtained using standard tools for phylogenetic inference under CTMC models of sequence evolution (e.g. MrBayes, Ronquist et al. (2012a)).

6.2.4 ALE with a two-type branching process model

The DP algorithm at the core of the ALE approach can be adapted to deal with more sophisticated models of gene family evolution. As noted before, the assumption that the loss rate *per* gene in a family is independent of the number of genes in the family is particularly problematic, with, for instance, the per-gene loss rate in a single-copy gene family typically being lower compared to the per-gene loss rate in the same family right after a duplication event. Furthermore, such more sophisticated models may admit more biologically reasonable modeling of WGDs (chapter 3).

By an analogous argument as for the phylogenetic linear BDP model, it is easy to see that the following system of differential equations describes the probability $p_e(u, m, t)$ of a gene tree lineage leading to node u passing through

time t along branch e as a type $m \in \{1, 2\}$ gene for the two-type DL model described in chapter 3:

$$\begin{aligned}
\frac{dp_e(u,1,t)}{dt} &= -(\lambda+\mu_1)p_e(u,1,t) + \nu p_e(u,2,t) \\
&\quad + \lambda\big(p_e(v,1,t)p_e(w,2,t) + p_e(v,2,t)p_e(w,1,t) \\
&\quad\quad + p_e(u,1,t)\epsilon_e(2,t) + p_e(u,2,t)\epsilon_e(1,t)\big) \\
\frac{dp_e(u,2,t)}{dt} &= -(\lambda+\mu_2+\nu)p_e(u,2,t) \\
&\quad + \lambda\big(p_e(v,2,t)p_e(w,2,t) + 2p_e(u,2,t)\epsilon_e(2,t)\big)
\end{aligned} \tag{6.8}$$

Recall that in chapter 3 we derived the following ODEs for the probability generating functions of the two-type branching process model

$$\begin{aligned}
f'_{10} &= \mu_1 - (\lambda+\mu_1)f_{10} + \lambda f_{10}f_{01} \\
f'_{01} &= \mu_2 + \nu f_{10} - (\lambda+\nu+\mu_2)f_{01} + \lambda f_{01}^2
\end{aligned} \tag{6.9}$$

which we can solve to obtain the extinction probabilities $\epsilon_e(m,t)$ for $e \in E(S)$. The moral is of course that we expect that for any model which satisfies the branching property, i.e. where distinct lineages evolve independently, it should be possible to derive a suitable recursion along these lines. We shall not consider the two-type model in the context of statistical gene tree reconciliation further here, deferring this to future work.

6.2.5 Implementation

We implemented the ALE algorithm for phylogenetic BDP models (with possible WGDs) in a Julia (Bezanson et al. 2017) package called WHALE[7], first used in Zwaenepoel and Van de Peer (2019a). Again (see chapter 3), we ensure differentiability by means of forward-mode automatic differentiation[8], and implement our methods such that they can be employed within a probabilistic programming framework, enabling the flexible specification of hierarchical Bayesian models as well as the usage of sophisticated sampling algorithms

[7] WHALE is not exactly an acronym, but rather a kind of portmanteau of whole-genome duplication and ALE. It is of course suggestive of a large cetacean, but this has little to do with the present topic (note however that all vertebrates have ancient polyploid ancestors).

[8] We note that the version of WHALE associated with our study in Zwaenepoel and Van de Peer (2019a) did not allow for AD, where we instead relied on gradient-free (and hence less efficient) methods for inference.

and optimizers for Bayesian inference and maximum likelihood estimation. The package is available at https://github.com/arzwa/Whale.jl. In the following example we illustrate our implementation using a simulated data set.

Example (simulated data). We use WHALE to simulate (reconciled) 1000 gene trees for a five-taxon phylogeny with a single WGD event, where we assume log-scale duplication and loss rates to vary across branches according to a Normal distribution with mean $\log 0.1$ and standard deviation of 0.5. We assume a geometric prior distribution on the number of ancestral lineages with parameter $\eta = 0.7$ and a retention probability for the single WGD event of $q = 0.1$. The following snippet of Julia code will conduct the simulation:

```
# Load required libraries
using Whale, NewickTree, Distributions, Turing

# Set up the model, `W` marks the WGD node
S = nw"(((((A:0.3,B:0.3):0.2)W:0.3,(C:0.6,D:0.6):0.2):0.2,E:1);"
n = length(postwalk(t))-2
λ = rand(Normal(log(0.1), .5), n)
μ = rand(Normal(log(0.1), .5), n)
θ = DLWGD(λ=λ, μ=μ, q=[0.1], η=0.7)
M = WhaleModel(M, S, 0.01)

# Simulate reconciled trees
trees, _ = simulate(M, 1000)
```

We next conduct Bayesian inference using WHALE for this simulated data set using the Turing.jl probabilistic programming environment, assuming uniform priors for η and q, and an uncorrelated relaxed clock prior for the branch rates, with a Normal $\mathcal{N}(\log 0.2, 1)$ prior for the the log-scale mean rate and an Exponential prior for the standard deviation of the branch rates (τ). The following snippet shows the model specification:

```
# Construct the CCDs from the simulated trees (input data)
y = read_ale(aleobserve(trees))

# Specify the Bayesian hierarchical model
@model model(M, y, n) = begin
    η ~ Beta()
    q ~ Beta()
    r ~ Normal(log(0.2), 1)
    τ ~ Exponential()
```

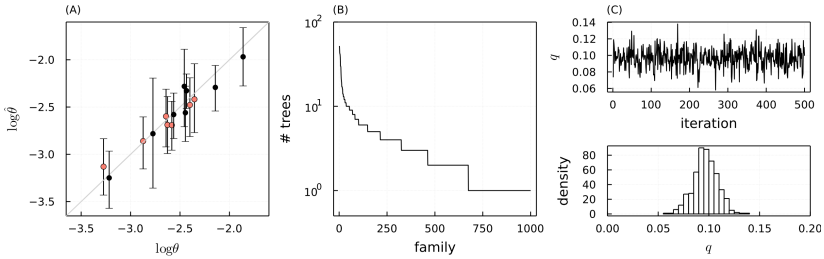


Figure 6.6: Bayesian inference using WHALE and Turing.jl for simulated gene trees on a five-taxon species tree. (A) Marginal posterior mean rate estimates and 95% posterior density intervals for the duplication (black) and loss (red) rates. (B) Number of distinct reconciled trees in the sample from the posterior distribution for each gene family. (C) Trace plot and marginal posterior histogram for the retention rate parameter q for the single WGD in the model.

```
    λ ~ MvNormal(fill(r, n), τ)
    μ ~ MvNormal(fill(r, n), τ)
    y ~ M((λ=λ, μ=μ, η=η, q=[q]))
end

# Sample from the posterior using the NUTS algorithm
chain = sample(model(M, y, n), NUTS(), 500)
```

In fig. 6.6 we show the marginal posterior means and estimated 95% posterior density intervals for the duplication and loss rates, where we find that we nicely recover the simulated rates. The other parameters are also recovered as expected, with $\hat{q} = 0.1\ (0.07,\ 0.12)$, $\hat{\eta} = 0.69\ (0.66, 0.71)$ and $\hat{\tau} = 0.46\ (0.29, 0.73)$. In addition, we used the backtracking algorithm to sample reconciled gene trees from the posterior distribution. We note that although the gene trees are assumed known in this example, there are often many possible reconciliations for the same fixed (unrooted) gene tree topology. □

6.3 Phylogenomic inference of whole-genome duplications

While the above discussion points to the general relevance of statistical reconciliation for making sense of gene family evolution from genomic data, our original interest in model-based gene tree reconciliation stems from the problem of inferring the occurrence of WGD events in the evolutionary history of

a group of taxa of interest by means of phylogenomic approaches. We have already alluded to this particular application of gene tree reconciliation and its problems in the introduction of the present chapter and provide a more detailed treatment in the present section.

6.3.1 Unveiling ancient WGDs from genomic data

Methods for unveiling ancient WGDs from genomic data can be crudely classified into three main approaches. In the first approach, we take advantage of the expectation that a WGD event leaves a distinctive signature in the distribution of divergence times between duplicated gene pairs. To do so, one commonly estimates the synonymous distance (K_S), or some other putatively neutral molecular distance, for the whole paranome, and visualizes the resulting distribution. In such a K_S distribution, ancient WGDs will be visible as distinctive components against the approximately exponentially distributed background coming from the SSDL process (see chapter 2). There are a couple of pitfalls with this approach, discussed in detail in Vanneste, Van de Peer, and Maere (2013) (see also Tiley, Barker, and Burleigh (2018); Zwaenepoel et al. (2019)). Most importantly, this approach will fail for very ancient events due to saturation of the estimated molecular distance. Furthermore, it can be challenging to locate putative ancient WGD events in a phylogeny based on K_S distributions due to differences in substitution rates across different lineages (see for instance Sensalari, Maere, and Lohaus (2021) or Chen et al. (2022) for a recent appreciation of this issue). Although statistically well-motivated model-based efforts for inference of WGDs from K_S distributions have not been considered in much detail, it is clear, at least in principle, how we could employ models like the DLWGD model of Rabier, Ta, and Ané (2014) or the model of Maere et al. (2005) for inferential purposes in this regard.

The second approach is based on the expectation that WGD events should lead to large colinear blocks in the genome(s) of interest. Such colinearity- or synteny-based information has often been considered the strongest evidence for ancient WGDs. In particular, the combination of syntenic and information from molecular divergence estimates has been vital for the discrimination of WGD-derived and SSD-derived paralogs. A major drawback is, however, that high-quality genome assemblies are required, and while considerable progress continues to be made in that regard, these are still nontrivial to obtain, especially for large plant genomes. Nevertheless, even with high-quality assemblies, interpretation of syntenic signal for very ancient putative WGDs is not always unequivocal. In particular, the temporal (either relative or absolute)

framing of a WGD event based on syntenic data is complicated and requires high-quality genomes of multiple related lineages. Here we note that, although there exists a rich literature in computational biology and combinatorics on genome rearrangement and ancestral genome reconstruction, statistical models of genome evolution at this level are barely existent. The usage of structural information residing in the linear organization of genomes for the inference of WGDs (and rates of genome rearrangement *etc.*) remains largely in a pre-statistical state.

The last set of methods, which constitute our present topic, are united by their usage of phylogenetic information in individual gene families. We have already discussed the methods based on gene counts and phylogenetic BDP models in chapter 3, but these have not been widely used and have limited power for unveiling ancient WGD events deep in a phylogeny (although this may in part be due to inadequate models of gene family evolution – see our discussion in chapter 3). Approaches based on analyzing gene trees of multi-copy gene families have been far more common in practice (e.g. Jiao et al. 2011; Li et al. 2015; McKain et al. 2016; Thomas, Ather, and Hahn 2017; Z. Li et al. 2018; Yang et al. 2018; Leebens-Mack et al. 2019, and too many plant genome papers to list here). Virtually all of these methods employ a two-step approach[9] where (1) gene tree topologies are inferred from multiple sequence alignments using ML inference assuming some phylogenetic CTMC model of sequence evolution, and (2) subsequently reconciled with an assumed species tree topology by means of some form of most parsimonious reconciliation[10].

[9] Of course, we stress once more (see also chapter 1 and chapter 5) that to label these methods as 'two-step methods' glosses over many important details. Most bioinformatics analyses involve *n*-step methods, where *n* is large (however much they are of the 'joint Bayesian model for everything' type). Indeed, to arrive at the input data for these two-step methods we have to sequence genomes (DNA extraction, purification, modification, sequencing, base-calling, *etc.*), assemble them, annotate protein-coding genes, infer orthogroups and align homologous genes. Thinking *too* much about the details of all these other steps when trying to conduct inference in evolutionary biology can however lead to a form of statistical paralysis that is not very useful.

[10] Usually, the number of duplication and loss events to fit the gene tree in the species tree (DL score) is minimized by means of least (or lowest, or last) common ancestor (LCA) reconciliation, as e.g. defined in Zmasek and Eddy (2001); or some algorithm which bears resemblance to LCA reconciliation. In the latter case, often an *ad hoc* approach is used without a stated formal principle (such as minimizing the DL score) but which is implicitly motivated by some vague parsimony argument. Here is an example from the methods section of a recent paper: "[…] we applied two basic requirements for the determination of a reliable duplication event: (1) at least one common species' genes are present in two child branches; and (2) the bootstrap values of the parental node and one of the child nodes are both ≥50%. After scoring gene duplications in a large-scale analysis on gene families, we were able to confidently identify the nodes with concentrated gene duplications across the phylogeny, which possibly support the WGD events." (Y. Liu et al. 2022). Note, in passing, that it is unclear how to conduct this procedure for unrooted trees.

In these approaches, a larger than expected number of duplication events inferred for a particular branch of the species tree is regarded as indicative for an ancient WGD. These methods have been applied both to genomic and transcriptomic data sets, and many large-scale phylotranscriptomic studies report results obtained by these means (Ren et al. 2018; Z. Li et al. 2018; Leebens-Mack et al. 2019 are notable examples). So far, most of the support for very ancient WGD events, and in particular their phylogenetic position, has been obtained using such gene tree reconciliation approaches.

6.3.2 Issues with inference of WGDs from gene trees

Clearly, there are many serious pitfalls with the commonly used two-step approaches. Firstly, how many gene duplicates would be indicative of WGD? Without some model of 'background' gene family evolution by small-scale gene duplication and loss, one would have to adopt a threshold value which would be arbitrary; with such a model, it is quickly realized that the parsimony-based duplication count is not the relevant measure to consider. Furthermore, the number of duplication events inferred for specific branches can be very sensitive to taxon sampling, and some signal for a putative WGD event on a particular branch may be absent or weakened when the branch is subdivided by adding more taxa to the analysis.

These are however not the most serious issues. A more fundamental problem in the standard two-step methods is the ignorance of gene tree uncertainty in the reconciliation step. In particular, reconciliation (model-based or otherwise) of incorrect or uncertain ML trees may lead to grossly overestimated numbers of gene duplication events deep in the tree (Hahn 2007). Not only can an inferred ML (or MAP) tree fail to be the correct ML tree, but, more importantly, there may be many trees for which the sequence data provides near-equal evidence under the assumed model of sequence evolution (Salter 2001). Obviously, if this is the case, relying on the most parsimonious reconciliation of the single ML tree is a perilous strategy to reconstruct the evolutionary history of a gene family. Furthermore, many of the methods used in practice rely on rooted gene trees, whereas standard phylogenetic tools infer unrooted trees[11]. The root location is itself however uncertain, and even when outgroups are available, rooting trees for multi-copy gene families is often not

[11]Note however that parsimony criteria can also be adopted for rooting the gene tree, and that many parsimony-based reconciliation programs (e.g. NOTUNG; Chen, Durand, and Farach-Colton 2000) can be used to infer the location of the root as part of the reconciliation problem.

straightforward. We illustrate the problem of ignoring gene tree uncertainty in two-step approaches in the following example[12]:

Example (yeast). A random sample of 100 gene families for a twelve-taxon yeast data set was obtained (see Appendix B). Maximum likelihood (ML) phylogenies were inferred using IQ-TREE, with 1000 bootstrap replicates (using IQ-TREE's ultrafast bootstrap approximation (Hoang et al. 2018)). We marked the yeast WGD event shared by the clade below the MRCA of *Saccharomyces cerivisiae* (sce) and *Tetrapisispora phaffii* (tph) along the phylogeny, and used WHALE with a simple constant rates model and flat non-informative priors for λ, μ and q, and setting $\eta = 0.98$ based on the average non-extinct gene family size in yeast species that are not derived from the ancient polyploid. We applied the same model to the ML trees and the empirical CCDs derived from the bootstrap replicates (note that an ML tree, i.e. a single topology, can be represented as a trivial empirical CCD). Posterior mean parameter estimates and 95% posterior uncertainty intervals are displayed in tbl. 6.1

Table 6.1: Marginal posterior mean estimates for the various model parameters of the DLWGD model with constant rates across S applied to maximum likelihood gene tree topologies (ML) or CCDs derived from maximum likelihood tree topologies of 1000 bootstrap replicates for 100 random gene families from the twelve-taxon yeast data set.

data set	λ	μ	q
ML trees	0.14 (0.11, 0.18)	1.25 (1.20, 1.31)	0.19 (0.13, 0.26)
CCDs	0.04 (0.02, 0.05)	0.35 (0.32, 0.39)	0.27 (0.19, 0.35)

Clearly, ignoring gene tree uncertainty leads to considerably higher parameter estimates, with almost a fourfold difference in the estimated duplication and loss rates. This is expected: noisy ML gene tree topologies which are incompatible with the species tree will lead to larger estimated evolutionary rates, as the latter result in a larger variance in the sampling distribution. Notably, the WGD retention rate estimate also differs, the direction of the difference however not being predictable in the same way. A related effect is that ignoring uncertainty in the gene tree topologies leads to *more* uncertainty in the gene tree reconciliation (fig. 6.7). The estimated posterior probability of the MAP reconciled tree is much larger on average when inference is based on the CCD than on the ML tree, and concomitantly, the entropy of the posterior sample is much lower. There is little correlation between the entropy of the

[12]Note that we illustrate the issue using the statistical reconciliation approach developed in the first part of the present chapter. The issue of gene tree uncertainty is pertinent for any two-step approach, independent of whether maximum parsimony or model-based statistical inference is conducted.

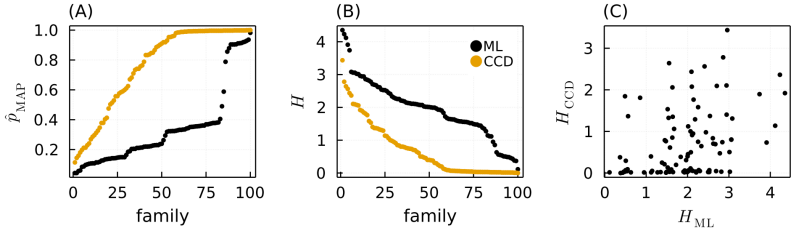


Figure 6.7: Whale reconciliation results for 100 gene families of the 12 yeast data set, using either ML trees as input data or CCDs derived from 1000 bootstrap replicates. (A) Estimated posterior probability $\hat{p}_{\mathrm{MAP}}$ of the MAP reconciled tree and (B) entropy H of the posterior sample over reconciled trees. Families are sorted independently for each data set. (C) Scatter plot of the entropy of the posterior sample of reconciled trees for each family for both data sets (i.e. each dot represents the entropy of the sample from the posterior distribution over reconciled gene trees for a single family for the ML tree as input data H_{ML} versus the empirical CCD as input data H_{CCD}).

posterior distributions for a single family for both data sets. In line with all this, reconciled trees based on the CCD show much fewer duplication and loss events compared to reconciled trees based on ML trees (fig. 6.8). The biases first brought to attention by Hahn (2007) are clearly visible here, where reconciliations based on ML trees tend to show a lot of duplications near the root followed by losses down the tree (fig. 6.9). Uncertainty-aware reconciliations based on the CCDs clearly do not show this bias to the same extent. □

Many authors are aware of some of the above issues, and mitigate these by resorting to intuitive but *ad hoc* filtering criteria and combining evidence from gene trees with other sources, such as K_{S} distributions and synteny (e.g. Yang et al. 2018; Li and Barker 2020). Here is an illustrative example from Yang et al. (2018)

> "To map polyploidy events in each subclade, we extracted orthogroups from each subclade homolog tree, requiring no more than two missing ingroup taxa. When two or more taxa overlapped between the two daughter clades, a gene duplication event was recorded to the most recent common ancestor (MRCA) on the subclade species tree (Yang et al. 2015). In this procedure, each node on a species tree can be counted at most once per orthogroup to avoid nested gene duplications inflating the number of duplications scored. Two alternative filters were applied for comparison. The first filter required an average bootstrap percentage of each orthogroup to be at least 50. Alternatively, we also tested a local topology filter that only mapped a gene duplication event when the sister clade of the gene duplication node in the orthogroup contained a subset of the

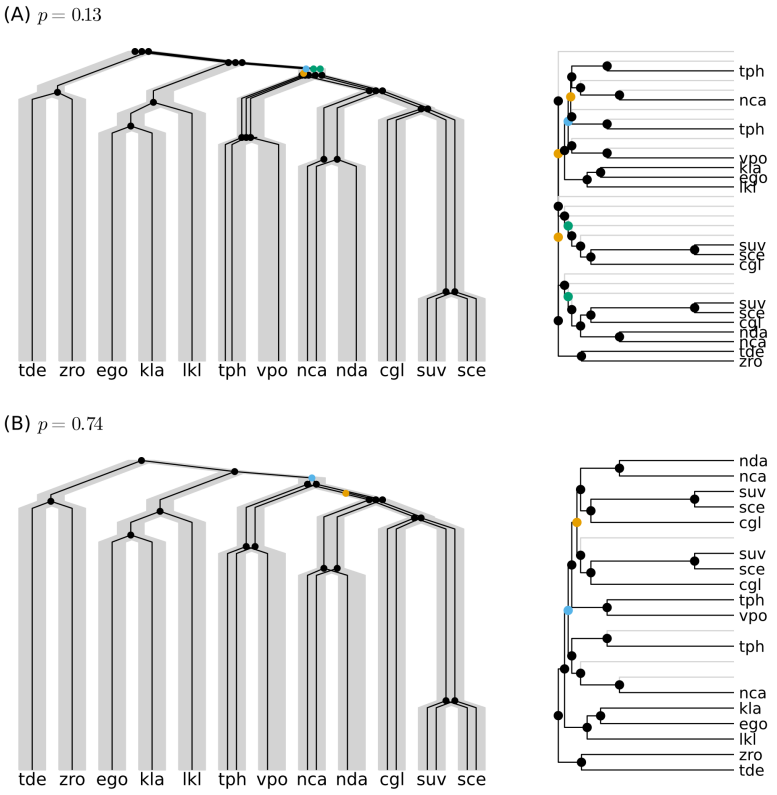


Figure 6.8: (A) MAP reconciled tree for a random family assuming a fixed (ML) tree topology. (B) MAP reconciled tree for the same family using the CCD as data. On the left we show the 'fit' inside the species tree (see fig. 6.1), while on the right we show the gene tree topology. The estimated posterior probability (p) is indicated above each reconciled tree. Black nodes mark speciation events, orange nodes mark duplication events, blue nodes mark retention after WGD and green nodes mark non-retention after WGD. Gray branches indicate subtrees where the gene lineage present at the source vertex of the branch did not leave observed descendants (i.e. went extinct or otherwise unsampled).

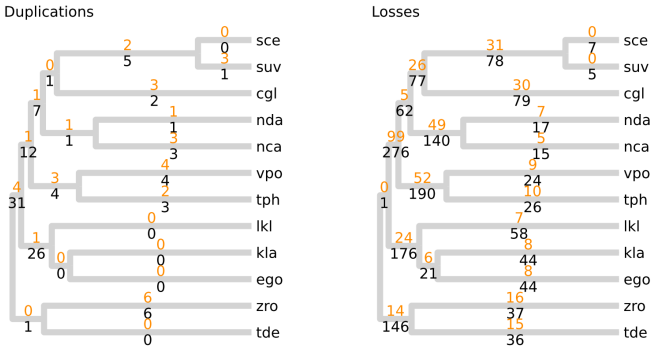


Figure 6.9: Expected number of represented duplication and loss events for the data set of 100 gene families from 12 yeast species. The numbers correspond to the average number of duplication/loss nodes in the gene trees for each branch of the species tree across 1000 samples of reconciled trees for each of the 100 families. In black (below each branch) the numbers for the ML tree input data are shown, whereas in orange we show the estimates for the CCD-based analysis which takes into account uncertainty in the gene trees. Note that the loss event count for the branch leading to the MRCA of *sce* and *tph* includes losses after the hypothetical WGD (i.e. a non-retention 'event' in the DLWGD model).

taxa in the corresponding sister clade in the species tree."

Similar recipes can be found in papers from other authors. While of course reasonable and motivated by prudence, this is a far cry from a direct tackling of finding out what evidence the data bears for our questions of interest. Instead, what we get is a further complexification of the already bewildering garden of forking paths that is characteristic of many bioinformatic analyses.

The methods developed in the present chapter should allow for better approaches towards the problem of WGD inference from gene trees. In particular, the statistical framework admits to address two types of questions in a scientifically more satisfactory manner: (1) "What is the evidence for a gene pair to be derived from a duplication event along branch e of the species tree given the sequence data" and (2) "What is the evidence for a WGD event occurring along a particular branch of the species tree given the sequence data?". The central assumptions which shall condition our answers to these questions are of course (1) birth-death like evolution of gene families and (2) standard Markovian sequence evolution. Our answers will be approximate, in the sense that we use the ALE approximation to the joint likelihood, adopting the phylogenomic forest point of view instead of

full joint (Bayesian) inference. In the next section we illustrate, by means of several examples, the use of statistical reconciliation under phylogenetic BDP models for the study of ancient WGDs.

6.3.3 Statistical inference under the DLWGD model

> *Freedom:* It's not user friendly.
> *Disbelief:* Remains an option.
>
> Statistical lexicon – Andrew Gelman

The DLWGD model of Rabier, Ta, and Ané (2014) (see chapter 3 and sec. 6.2.1.3) is a model of genome-wide gene family evolution that admits statistical assessment of hypothetical WGDs in a phylogenetic setting. We already illustrated the approach of Rabier, Ta, and Ané (2014), and discussed its shortcomings, in chapter 3. Evidently, the ML-based model selection approach (using the likelihood ratio test statistic) of these authors carries over without complications to the reconciliation setting[13]. Consider the following running example:

Example (land plants). It is widely believed that an ancient WGD event preceded the diversification of seed plants and another one the diversification of angiosperms. More recently, a WGD has been reported to have preceded the diversification of gymnosperms as well (Y. Liu et al. 2022). The actual support for many of these claims is however rather unclear. The first study to report a seed plant and angiosperm WGD (Jiao et al. 2011), which relied on gene tree reconciliation and phylogenomic dating, is likely plagued by methodological issues (Ruprecht et al. 2017), and so far, synteny information has not provided conclusive evidence for more than one WGD preceding the angiosperm crown (Albert et al. 2013). The claim of a gymnosperm-specific WGD in Y. Liu et al. (2022) is based on a two-step reconciliation approach which accounts for uncertainty in the gene trees using *ad hoc* filtering criteria and does not assume a model of gene family evolution. Liu et al. (2021) used WHALE in a Bayesian analysis for the *Ginkgo biloba* genome paper and show support for an ancient WGD in the seed plant, but not gymnosperm,

[13]Note that in Rabier, Ta, and Ané (2014), the authors did in fact also study a gene tree reconciliation-based method for the inference of ancient WGDs using the ML/LRT approach. However, they reported less accurate estimates and lower power to detect WGDs compared to their gene count method, likely because of the various simplifying assumptions and coarse optimization approach adopted for the sake of computational feasibility. The ALE approach of Szöllősi, Rosikiewicz, et al. (2013) adopted by us in WHALE circumvents many of the computational issues in Rabier, Ta, and Ané (2014).

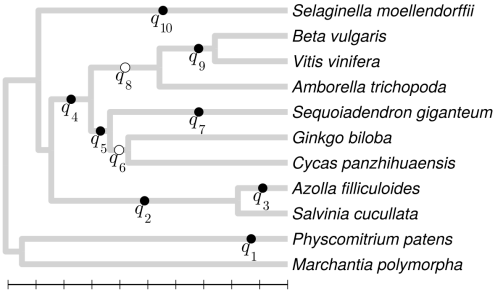


Figure 6.10: Species tree for the 11-taxon land plant data set. The time scale along the horizontal axis is 50 My. Dots mark WGD hypotheses assessed in our analyses with WHALE . The time point of the hypothetical WGD is set to the midpoint of the relevant branch for all but q_1 and q_9, where we used WGD date estimates reported in Vanneste, Van de Peer, and Maere (2013) and Tiley, Ané, and Burleigh (2016) respectively. The white dots mark hypothetical WGDs where we coudl not reject the null hypothesis ($q = 0$) using the LRT under the constant rates (strict DL clock) DLWGD model.

stem branch. However, the latter study did not report the details of the employed statistical model, which renders their analysis moot (and suspect)[14]. We note that both the *Ginkgo* and *Cycas* genomes show a clear sign of WGD in their respective K_S distributions, and that this signature likely reflects a single WGD event which precedes the divergence of the two (Roodt et al. 2017). In Zwaenepoel and Van de Peer (2019a), we showed that, when taking into account gene tree uncertainty, it is not straightforward to conclusively associate the putative WGD-derived duplicates in *Ginkgo* with either the seed plant or gymnosperm stem branch.

Clearly then, we consider the problem unsolved. To study the problem using the methods developed in the present chapter we obtained a data set consisting of the complete set of protein-coding sequences from 11 taxa sampled across the phylogeny of land plants. Specifically, we include *Physcomitrium patens*, *Marchantia polymorpha*, *Selaginella moellendorffii*, *Azolla filiculoides*, *Salvinia cucullata*, *Ginkgo biloba*, *Sequoiadendron*

[14]This is all the more upsetting as in the *Reporting Summary* that comes with the paper of Liu et al. (2021) (the *Nature* group introduced these in 2017 to combat irreproducibility (Nature 2017)), the authors mark the 'not applicable' checkbox next to the item with the following description: "For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings". We also take the opportunity to note the rather silly identification on Nature's part of Bayesian analysis with MCMC.

giganteum, *Cycas panzhihuaensis*, *Amborella trichopoda*, *Vitis vinifera* and *Beta vulgaris* (see Appendix B for details). We used Orthofinder (Emms and Kelly 2019) to delineate orthogroups, filtered out those families which did not contain at least one gene in each clade stemming from the root of the species tree, and obtained a random subset of 1000 gene families for the sake of computational efficiency. We inferred codon-level multiple sequence alignments with MAFFT (Katoh and Standley 2013), trimmed the alignments using TrimAL (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009, using the `-automated1` strategy) and obtain an empirical CCD for each gene family based on a sample of 10000 gene trees from the posterior distribution under the GTR + Γ4 model with default priors using MrBayes (sampling every 50 iterations for a total of 550000 iterations, discarding 1000 samples as burn-in, running two chains in parallel). We restrict our analyses further to the 949 families for which we obtained a minimum ESS of 100 for the total tree length and for which the average standard deviation of split frequencies (ASDF) was lower than 0.05. We use a dated species tree from the study of Morris et al. (2018) in our analyses.

A ML-based statistical analysis of the data is presented in tbl. 6.2. We fix the parameter of the geometric prior on the number of ancestral lineages to $\eta = 1/1.5$ (based on the average non-extinct family size in the data) and estimate a single duplication and loss rate for the entire species tree S, as well as retention probabilities for 10 WGD hypotheses specified along S (see tbl. 6.2 and fig. 6.10). We test each WGD hypothesis against the model including all other WGDs but the focal one. The LRT approach of Rabier, Ta, and Ané (2014) would result in the rejection of the null hypothesis ($q = 0$, which we identify with the absence of a WGD) for each of the marked WGD hypotheses, except for the event associated with the *Ginkgo-Cycas* ancestor and the hypothetical angiosperm-specific event. This includes both the hypothetical seed plant and gymnosperm WGDs, with the latter having a fairly high MLE for the retention probability ($\hat{q} = 0.19$). □

While attractive, the statistical analysis presented in the above example can be positively misleading. Any statistical inference is of course conditional on the assumed model, which may or may not be reasonable. In this case, as was already noted in chapter 3, our model assumptions include a strict DL clock, i.e. constant rates of duplication and loss across the phylogeny, an assumption which is very likely to be problematic. In Zwaenepoel and Van de Peer (2019a) we showed that, as could be expected, violations of the DL clock assumption in the LRT based method of Rabier, Ta, and Ané (2014) can lead to high false positive rates, as well as limited power. Indeed, intuitively, if

Table 6.2: ML-based analysis for the land plant data set. The table records MLEs for the 10 retention probability parameters corresponding to the marked WGD events in fig. 6.10. Each row displays the MLEs for an analysis holding none (first row) or a single retention rate parameter fixed to 0 (indicated by the dash). ℓ is the maximum log-likelihood value. Those log-likelihoods which do not differ significantly from the full model log-likelihood are marked in bold.

q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	ℓ
0.36	0.08	0.12	0.08	0.19	0.01	0.08	0.00	0.09	0.07	**-24633.44**
-	0.02	0.11	0.07	0.18	0.00	0.04	0.00	0.07	0.00	-24786.39
0.36	-	0.12	0.08	0.19	0.01	0.07	0.00	0.09	0.06	-24642.72
0.35	0.08	-	0.08	0.19	0.00	0.06	0.00	0.09	0.05	-24714.60
0.35	0.07	0.12	-	0.20	0.00	0.07	0.00	0.09	0.05	-24675.97
0.35	0.06	0.12	0.10	-	0.03	0.08	0.00	0.09	0.04	-24839.20
0.36	0.08	0.12	0.08	0.19	-	0.08	0.00	0.09	0.07	**-24633.95**
0.35	0.07	0.12	0.08	0.19	0.00	-	0.00	0.09	0.05	-24647.21
0.36	0.08	0.12	0.08	0.19	0.01	0.08	-	0.09	0.07	**-24633.44**
0.35	0.06	0.12	0.08	0.19	0.00	0.07	0.00	-	0.05	-24670.41
0.36	0.07	0.12	0.08	0.19	0.01	0.07	0.00	0.09	-	-24639.21

some branch in the phylogeny which underwent a WGD happens to have a lower rate of small-scale duplication and/or higher rate of gene loss, we may expect the power to detect WGD to be diminished with respect to strict clock situation, whereas a branch with an increased (small-scale) duplication rate and/or decreased loss rate relative to the rest of the species tree could lead to false positive WGD inferences under the DLWGD model as formulated by Rabier, Ta, and Ané (2014). As we noted in Zwaenepoel and Van de Peer (2019a), this likely explains to some extent the observations of Tiley, Ané, and Burleigh (2016), where the authors report that the power to detect certain ancient WGDs depends on the sampled taxa.

While the ML-based approach does not readily provide us with a means to investigate the extent of (unavoidable) model violations, we can investigate model fit for the constant rates model by using a Bayesian analysis and conducting posterior predictive simulations. We illustrate this continuing our analysis of the land plant data set.

Example (land plants, continued). We reconsider the land plant data set analyzed above, now using a Bayesian model with a Beta$(4, 2)$ prior for the parameter of the Geometric distribution on the number of lineages at the root, an exponential prior with mean 1 for both the duplication and loss rate and uniform prior for the retention probabilities $q_1, q_2, \ldots, q_{10}$. The marginal posterior mean parameter estimates and 95% uncertainty intervals are shown in

tbl. 6.3. We note that we estimate $\eta = 0.86\ (0.84, 0.89)$[15], which is quite a bit higher than the estimate based on extant family sizes (0.66). This would indicate that the ancestral gene family size was, on average, 1.16, instead of the presently observed 1.5. As expected, the marginal posterior means correspond almost exactly to the MLEs. The Bayes factors (which constitute a Bayesian analog of the LRTs of Rabier, Ta, and Ané (2014), see chapter 3) lead to similar conclusions, with the data not providing evidence for non-zero retention probabilities for the hypothetical *Cycas-Ginkgo* WGD nor the seed plant WGD under the model.

Table 6.3: Posterior mean and 95% uncertainty intervals for the various parameters in the constant rates DLWGD model. The last column shows the $\log_{10}$ bayes factor against the WGD hypothesis, computed using the Savage-Dickey density ratio (see chapter 3). Duplication and loss rates are on a scale of expected number of events per lineage per 100 My.

	mean	2.5%	97.5%	$\log_{10} K$
η	0.86	0.84	0.89	
λ	0.06	0.06	0.07	
μ	0.09	0.09	0.10	
q_1	0.36	0.32	0.41	< -3
q_2	0.08	0.04	0.12	-1.0
q_3	0.12	0.10	0.15	< -3
q_4	0.08	0.06	0.11	< -3
q_5	0.19	0.16	0.22	< -3
q_6	0.01	0.00	0.02	1.7
q_7	0.08	0.05	0.11	< -3
q_8	0.01	0.00	0.02	1.8
q_9	0.09	0.07	0.12	< -3
q_{10}	0.07	0.03	0.11	-0.9

We conducted posterior predictive simulations to assess the fit of the constant rates DL model to the data. Specifically, we sample N parameter vectors $(\theta_1, \ldots, \theta_N)$ from the posterior distribution, and for each θ_i, $1 \leq i \leq N$, we obtain a set Ψ_i of reconciled trees for all 949 families using the stochastic backtracking algorithm, as well as a set $\tilde{\Psi}_i$ of reconciled trees for 949 simulated gene families conditional on θ_i. In our notation:

$$\Psi_i = (\mathcal{R}_1, \ldots, \mathcal{R}_n) \text{ where } \mathcal{R}_j \sim p(\mathcal{R}_j|y) = \int p(\mathcal{R}_j|y_j, \theta)p(\theta|y)d\theta$$

[15] As elsewhere, we report point estimates (marginal posterior means) with 95% posterior density intervals, which are both estimated from a MCMC sample from the posterior. The MCSE of the estimators is virtually always negligible and not reported.

$$\tilde{\Psi}_i = (\tilde{\mathcal{R}}_1, \ldots, \tilde{\mathcal{R}}_n) \text{ where } \tilde{\mathcal{R}}_j \sim p(\tilde{\mathcal{R}}|y) = \int p(\mathcal{R}|\theta)p(\theta|y)d\theta$$

where $n = 949$ for this specific example. Note that each $\tilde{\Psi}_i$ constitutes a sample from the posterior predictive distribution of the same size as the observed data. We can use these simulations to assess the discrepancies between the observed data and predictions under the model by comparing statistics T of interest between Ψ and $\tilde{\Psi}$, such as, for instance, the number of duplication or loss events on each branch of the species tree.[16] If the model provides a reasonable fit, we would expect $T(\Psi) \approx T(\tilde{\Psi})$, whereas any systematic discrepancy between the two suggests that T points at an aspect of the data which is not adequately captured by the model. In fig. 6.11, we show a graphical model check (black scatter plots), comparing the posterior number of duplications on each branch of the species tree against posterior predictive simulations thereof for $N = 100$ draws from the posterior. These simulations indicate that the model provides a reasonable fit in some parts of the species tree, whereas it has more issues in other parts. Clearly, the number of represented gene duplication events along the branch leading to *Marchantia* (mpo) is much smaller than expected under the model, and similar observations hold for *Cycas*, *Ginkgo* and the moss stem branch ((ppa,mpo), to a lesser degree). The opposite appears to hold for *Salvinia*, *Vitis*, *Beta* and the euphyllophyte stem (scu,cpa). The lack of fit seems to be more extreme when we consider related test statistics (figs. 6.12, 6.13, 6.14). The most obvious explanation for the observed discrepancies appears to be rate heterogeneity across lineages, where the globally (phylogeny-wide) estimated λ and μ do not appear to reasonably predict the number of duplication events for individual branches of S. □

As in chapter 3, we can combat these issues in the Bayesian framework by adopting a hierarchical model to describe the variation of gene duplication and loss rates across the species tree S. The approach is entirely analogous to what we described for our gene count analyses under the phylogenetic BDP model of gene family evolution, so we refer the reader to the relevant chapter. We pick up the land plant example again to illustrate the Bayesian inference approach first used in Zwaenepoel and Van de Peer (2019a).

Example (land plants, continued). To account for rate heterogeneity across branches of S, we specify the following hierarchical model (where $n = 11$ is

[16] For a discussion of posterior predictive test quantities which depend not only on the data y, but also on the parameters θ, see Gelman et al. (2013) chapter 6 (p.148).

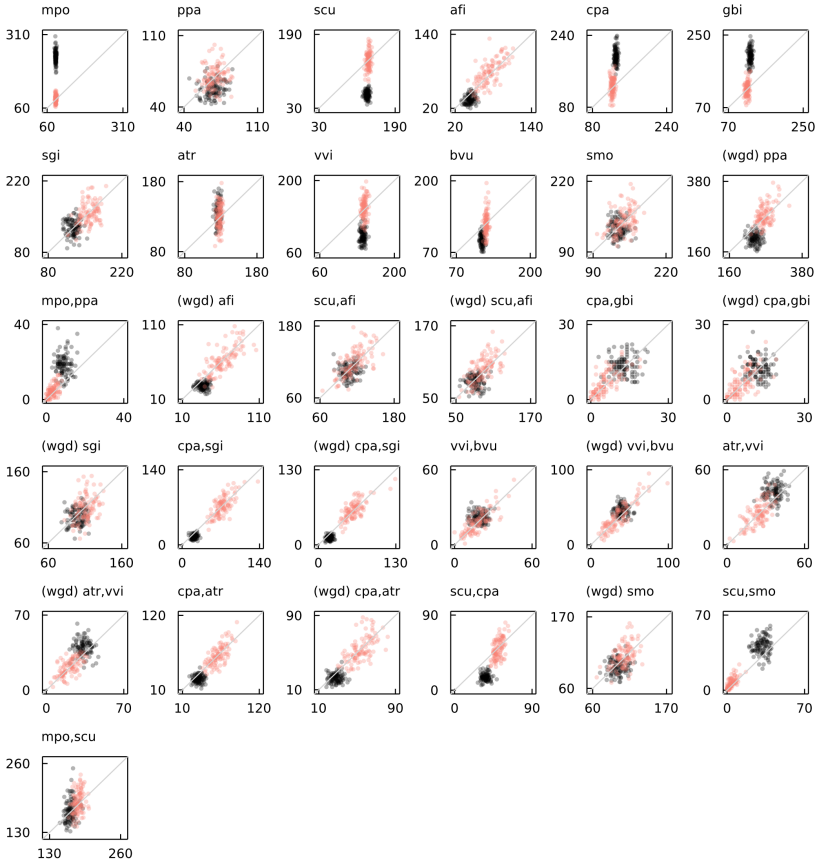


Figure 6.11: Posterior predictive model checks for the constant rates DL model (black) and the relaxed DL clock model (red) fitted to the land plant data set. For each branch of the species tree we show a scatter plot where each dot is associated with a single parameter vector θ drawn from the posterior distribution $p(y|\theta)$. For each dot the x-coordinate marks the number of duplication events in the sample of reconciled trees *for the observed data* conditional on θ for the relevant branch of $\boldsymbol{S}$, whereas the y-coordinate marks the number of duplication events in a data set of 949 families drawn from the predictive distribution conditional on θ. The titles indicate the target vertex of the relevant branch, where, for instance, *mpo,ppa* refers to the branch leading to the MRCA of *M. polymorpha* and *P. patens*, and *(wgd) ppa* refers to the branch leading to the WGD node parental to *P. patens* (see fig. 6.10 for the relevant WGD nodes).

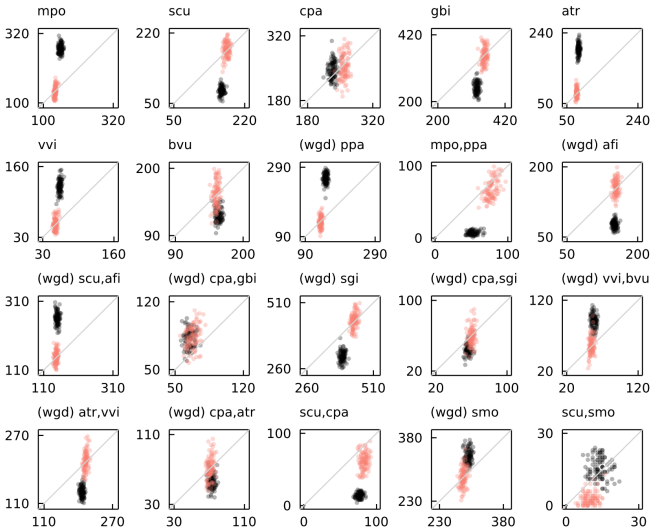


Figure 6.12: As in fig. 6.11, but showing, for each branch of the species tree, the number of represented speciation events which are followed by extinction down the subtree rooted in the relevant branch. For instance in the *vvi* panel, we show the number of *B. vulgaris*, *V. vinifera* nodes in the simulated reconciled trees which are followed by extinction along the branch leading to *vvi*.

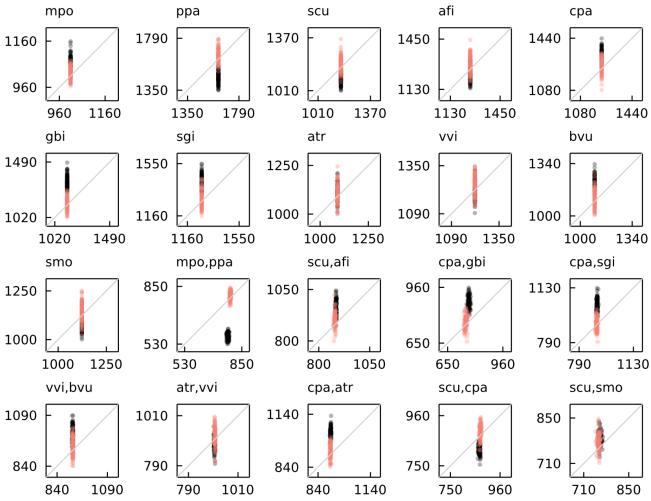


Figure 6.13: As in fig. 6.11, but showing the number of gene tree nodes reconciled to the speciation nodes or leaf nodes associated with the indicated branch. For leaf nodes (the first 11 plots), this amounts to a comparison of the observed family sizes (the x-coordinate is the total number of genes of that species in the 949 gene families) against posterior predictive simulations thereof.

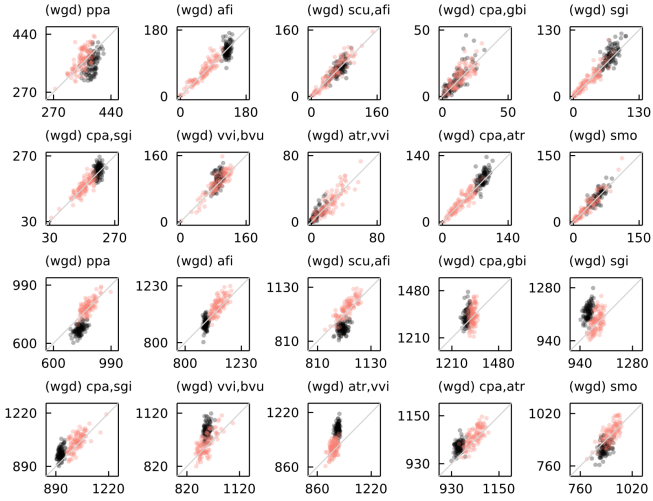


Figure 6.14: As in fig. 6.11, but showing the number of WGD retention events (top two rows) and non-retention events (bottom two rows) for each node of the species tree.

the number of taxa, and $m = 10$ is the number of hypothetical WGDs)

$$\begin{aligned}
\eta &\sim \text{Beta}(4, 2) \\
\tau &\sim \text{Exponential}(0.5) \\
r_1 &\sim \mathcal{N}(\log(0.07), 1) \\
r_2 &\sim \mathcal{N}(\log(0.07), 1) \\
\log \lambda_i | r_1, \tau &\sim \mathcal{N}(r_1, \tau) && i = 1, \dots, 2n-2 \\
\log \mu_i | r_2, \tau &\sim \mathcal{N}(r_2, \tau) && i = 1, \dots, 2n-2 \\
q_j &\sim \text{Uniform}(0, 1) && j = 1, \dots, m
\end{aligned}$$

In molecular evolutionary terms, this amounts to an uncorrelated log-normal relaxed clock model (actually, two independent ones). Of course, many different models can be conceived to account for rate variation, among which we note models where the rates are themselves considered as quantitative traits evolving along S according to some stochastic process, such as, for instance, a (geometric) Brownian motion or a multivariate process modeling the joint evolution of duplication and loss rates. We explored this in more detail in Zwaenepoel and Van de Peer (2019a) and Zwaenepoel and Van de Peer (2020).

A sample from the posterior distribution for the 10 retention probability parameters is shown in fig. 6.15. Bayes factors computed using the Savage-Dickey density ratio under a KDE approximation yield substantial or strong evidence for non-zero q only for the gymnosperm (q_5), eudicot (q_9) and *Physcomitrium* (q_1) events. We find that the standard deviation of the clock model τ has a marginal posterior mean of 1.0 (0.76, 1.33), and that there is, concomitantly, substantial rate variation. Some branch rates appear suspiciously large, such as the loss rate of 0.88 (0.60, 1.19) events per gene per 100 My for the moss stem branch. The latter would entail that a single lineage at the root has a probability of 0.23 (0.16, 0.29) to leave no descendants at the moss crown group. Note, however, that we condition the likelihood on there being at least one observed descendant in each clade stemming from the root. In other words, the estimated rates may yield a good fit conditional on non-extinction, while making a possibly unrealistic prediction with regard to the number of unobserved families (i.e. families which are extinct or do not leave observed descendants in either clade stemming from the root). Indeed, we find that to obtain a simulated data set of 947 families with at least one observed descendant in both clades stemming from the root, we have to reject 317 (234, 424) simulated gene families as unobservable under our filtering strategies, whereas for the constant rates model analyzed above, this was 183 (159, 209). Nevertheless,

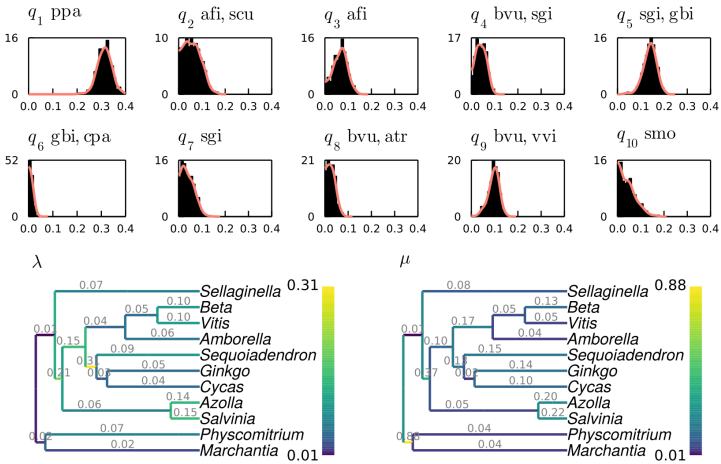


Figure 6.15: Marginal posterior distributions for the retention probabilities q and species trees colored by branch-specific marginal posterior mean duplication rates (λ, left) and loss rates (μ, right). The indices for the retention probability parameters correspond to those indicated in fig. 6.10. The orange lines show the KDEs fitted to the marginal posterior distributions on which we base the calculation of the Savage-Dickey ratio. The number on top of each branch is the marginal posterior mean rate estimate for that particular branch.

posterior predictive model checks suggest that the model fits the data rather well, in the sense that random gene families simulated under the posterior (conditional on non-extinction, as noted above) agree very well with reconstructed reconciled trees for the observed data in terms of duplication, loss, speciation and WGD event counts as well as observed family sizes at the leaves of S (figs. 6.11, 6.12, 6.13, 6.14, red scatter plots). For the loss-related posterior predictive test quantity (fig. 6.12), we still observe some problematic branches, especially deep in the phylogeny, such as the branches leading to the euphyllophytes and the tracheophytes, as well as, to a lesser extent, the *Cycas*-*Ginkgo* stem branch.

Clearly, under this model, evidence for many putative ancient WGDs is considerably weaker than under the (generally more problematic) strict DL clock assumption. Consider for instance the putative core leptosparangiate shared WGD (see also Chen et al. 2022), with retention probability parameter q_2. In the constant rates analysis, we find strong support for non-zero q_2, whereas the relaxed clock analysis shows that, while the data is compatible with q_2 at

least as high as 10%, the data is also compatible with $q_2 = 0$. Similar observations hold for the other WGD hypotheses that would be accepted when naively applying LRT-based model selection or Bayes factors computed under the constant rates DL model. We note that while a WGD in the *Azolla* lineage is well supported from other sources (Chen et al. 2022), our analysis is rather undecisive with regard to this event. While the marginal posterior histogram and mean of $q_3 = 0.07$ $(0.01, 0.13)$ are suggestive of WGD, the posterior is compatible with $q_3 = 0$. Similarly, the marginal posterior distribution for q_4 may suggest some signal for the hypothetical seed plant WGD event, however the posterior shows that under the relaxed clock model, the data are perfectly compatible with $q_4 = 0$. □

What can we learn from the land plant example? It seems that gene tree topologies *alone* cannot provide conclusive evidence for many WGD events when they are analyzed under flexible models which provide a reasonable fit to the data. Stated in other words, if we admit rate variation across lineages of the species tree, we are lead to the conclusion that observed gene tree *topologies* are largely compatible with a linear phylogenetic BDP model, even when WGDs are actually known to have occurred (as for *Azolla* in the land plants analysis). On the other hand, for some known WGDs, such as the *Physcomitrium* and eudicot WGD[17] in the example above, we do find consistent support even if we admit rate variation, suggesting that a flexible phylogenetic BDP cannot always adequately capture the genome-wide effect of a WGD on gene family evolutionary histories. Similar observations were made for Bayesian analyses based on gene counts in chapter 3. It is, in that regard, rather unsurprising that evidence for WGD events deep in the species tree is scarce when gene tree topologies are properly analyzed. Uncertainty in the gene trees for these deep splits will result in many reconciliations being reasonably plausible under the model, rendering any strong signal for a deviation from the linear BDP model unlikely. Do note however that the analysis presented above is based on a reduced data set of 1000 families and a fairly small phylogeny for the sake of computational efficiency. In the era of extensive genome sequencing, one can always gather more data to address these kind of evolutionary questions, but this comes at the cost of increased computational demands and a more complicated (and sometimes precarious) statistical workflow.

Importantly, notwithstanding the observation that the model provides a rea-

[17] This is in fact thought to be a triplication, i.e. involving an ancestral hexaploid phase, a detail we gloss over here. We note that this renders the associated retention probability q hard to interpret.

sonable fit to the data, care needs to be taken when interpreting results like those presented in the above example. The interpretation of non-zero q as the probability of duplicate retention after WGD, or as signal for WGD in a weaker sense, rests on a number of assumptions which are almost surely not met. Firstly, we reiterate that what the WGDs in the DLWGD model capture are genome-wide deviations from a linear BDP, the latter being assumed as an adequate model of small-scale duplication and loss within a branch. When phrased this way, it is clear that these genome-wide deviations from a phylogenetic linear BDP need not in fact correspond to WGDs, even if the model was designed with this in mind. In particular, ILS, (possibly allopolyploid) hybridization, delayed rediploidization (Robertson et al. 2017) and introgression may cause genome-wide deviations from a BDP model which could, potentially, lead to non-zero q. Hence, it would be unwise to claim, for instance, strong evidence for a shared gymnosperm WGD based on the above results, without assessing whether those gene pairs that supposedly diverged after a duplication along the gymnosperm stem branch are in fact compatible with the hypothesis of a WGD event. We return to this problem shortly.

Besides the basic flaws and caveats associated with the DLWGD model, four serious, and related, weaknesses of the approach exemplified above should be noted. The first is that we are specifying WGD hypotheses along S *a priori*, and are not discovering them from the data in an 'automatic' way. The reversible-jump MCMC algorithm for inference across model space, outlined in chapter 3 and Zwaenepoel and Van de Peer (2020), could be of use here, but would lead to considerable computational challenges. The second is an implementation detail[18] which is not a fundamental limitation of the approach, namely that in the above, we have fixed the time point of each WGD along the associated branch of S, whereas we usually have no clue about this *a priori* (in the land plant example, we used substantive prior information for this only for *Physcomitrium*). The third is our reliance on an assumed known dated species tree. If the topology of the latter is incorrect, this will render the whole analysis problematic for obvious reasons, whereas systematic errors in relative branch lengths could lead to biases in the rate estimates and estimates of rate heterogeneity across S. The last weakness is that we do not make use of branch length information. Clearly, this would provide additional information for distinguishing small-scale duplication from WGD-derived duplication events, and hence further improve duplication and loss rate estimates

[18] By implementation detail, we mean to suggest something that presents no theoretical issues, but does require a considerable programming effort. In this case, allowing the WGD timings to vary along a designated branch should indeed present no theoretical difficulties, but would require a reimplementation of a nontrivial part of WHALE, which will be done in due time.

while increasing the power to detect WGD events. This could be particularly impactful for relatively recent WGD events, where the molecular distances between WGD-derived duplicates should be strongly correlated and the associated node ages concentrated to a rather narrow peak, whereas the node ages for small-scale duplication events should, under neutral evolution, be approximately exponentially distributed under the linear BDP model (see chapter 2).

6.3.4 Probabilistic homology inference

The statistical reconciliation approach developed (and advocated) in the present chapter gives us another powerful tool for studying WGD, quite independent of the statistical analysis under the DLWGD model first proposed by Rabier, Ta, and Ané (2014). This tool is the statistical assessment of homology relationships, which allows us to address questions like "what is the probability that a given gene pair derives from a duplication event on branch e of S?". This allows us to provide a statistically more adequate means to quantify the number of duplication and loss events along each branch of the species tree compared to the standard two-step approach, by acknowledging both gene tree and reconciliation uncertainty. To clarify, we return once more to the land plants data set.

Example (land plants, continued). The analysis above indicated strong evidence for a gymnosperm-specific WGD conditional on the model assumptions, while the results are not conclusive with respect to a seed plant nor angiosperm specific WGD, although the former appears more probable than the latter based on the analyzed data. As noted above, while suggestive, it would be unwise to base any strong claims on this kind of analysis, since there exist plenty of biological processes which may lead to systematic model violations that could be mistaken for WGD under the DLWGD model. To further assess these results, we leave the genome-wide view and zoom in on individual gene pairs. For each family in the full data set (which consists of gene tree distributions for 7526 families, obtained using the same methods as the set of 949 families above), we sample 100 reconciled trees under the posterior distribution from the analysis above based on the 949 families set. That is, denoting the small data set by $y_{[949]}$, we sample for each family $i = 1, \dots, 7526$ from

$$p(\mathcal{R}_i|y) = \int p(\mathcal{R}_i|y_i, \theta)p(\theta|y_{[949]})d\theta$$

Based on these samples, we can estimate for each homologous gene pair in

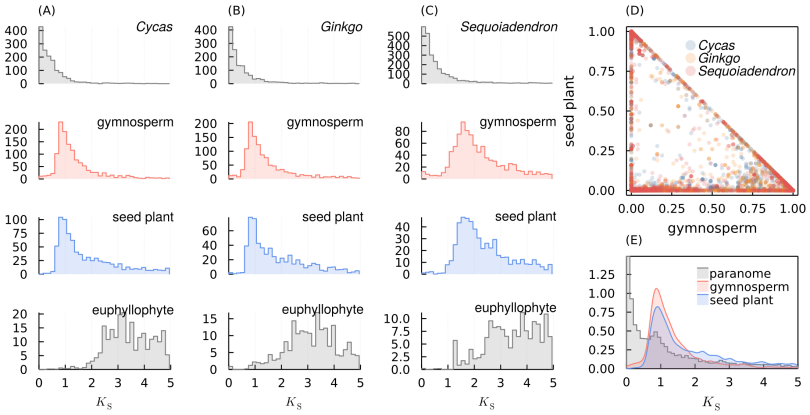


Figure 6.16: Statistical homology assessment and K_S distributions for duplication events along different branches of the species tree for the three gymnosperm taxa in the land plants data set. (A) K_S distribution for paralogous *Cycas* gene pairs weighted by the posterior probability of being reconciled as a duplication to the *Cycas* tip branch, gymnosperm stem branch, seed plant stem branch and euphyllophyte stem branch. (B) and (C) show the same but for *Ginkgo* and *Sequoiadendron* respectively. (D) Scatter plot of the posterior probability of being reconciled to the gymnosperm vs. seed plant stem branch for each duplicate gene pair. (E) Whole-paranome K_S distribution for *Cycas* with the putative gymnosperm and seed plant signatures overlaid (we show KDEs for the same distributions as displayed in the middle two plots of panel (A)).

the data set the posterior probability distribution over possible reconciliations. In addition, we estimated for each homologous gene pair in the data set the synonymous distance by maximum likelihood using `codeml` (Z. Yang 2007b, using `runmode=-2, CodonFreq=2` and default settings otherwise).

In fig. 6.16, we show, for duplicate gene pairs in *Cycas*, *Ginkgo* and *Sequoiadendron*, a histogram of pairwise K_S estimates where each estimate is weighted by the estimated posterior probability for the associated pair of being reconciled as a duplication event along (1) the relevant tip branch, (2) the gymnosperm stem branch, (3) the seed plant stem branch and (4) the euphyllophyte stem branch. The K_S distributions for the duplicate pairs associated with the tip branches clearly show the expected exponential shape, whereas we find that the duplicates associated with the putative gymnosperm WGD show a clear peak around $K_S \approx 1$, coinciding with the signature in the whole-paranome K_S distribution. This provides additional evidence that the signature in the whole-paranome K_S distribution is associated with a gymnosperm-specific event. However, we find that the K_S distribution for

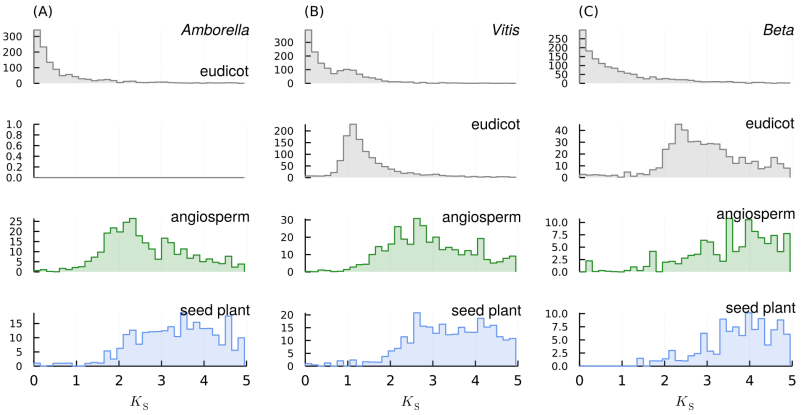


Figure 6.17: As in fig. 6.16, but focusing on duplicate genes within angiosperm genomes, i.e. (A) *Amborella trichopoda*, (B) *Vitis vinifera* and (C) *Beta vulgaris*.

the putative gymnosperm-specific duplicates overlaps almost completely with the distribution for gene pairs that are likely (under the model) to derive from duplication events, possibly from a WGD, associated with the seed plant stem branch. Therefore, if both a gymnosperm and seed plant WGD did occur, it appears not to be possible to distinguish between the two based on molecular divergence. These results may however also suggest that there is only a single event, with gene tree uncertainty 'diluting' the signal over the two branches. The latter is to some extent contradicted by the posterior reconciliation probabilities (fig. 6.16 D), where we can see that most duplicates are either reconciled with high probability to the seed plant branch or the gymnosperm branch, but rarely have high posterior probability for both. For the angiosperms in the analysis, we similarly find no clear distinction between putative angiosperm-specific duplicates and duplicates associated with the seed plant stem in terms of molecular divergence (fig. 6.17). Notably, the higher substitution rates in angiosperms compared to gymnosperms lead to a higher variance in both the K_S estimates and putative WGD signatures, which renders ancient WGD hypotheses even harder to assess in this part of the phylogeny. □

Our results in the land plants example confirm our previous analyses reported in Zwaenepoel and Van de Peer (2019a), which were conducted with a different sample of genomes. Taking everything together, a gymnosperm-specific WGD event appears plausible, while we must remain rather inconclusive concerning the hypothesized seed plant or angiosperm WGD events. Note

that the actual history is almost surely more complicated. It is, for instance, conceivable that the seed plant ancestor was a polyploid which did not undergo complete rediploidization before the divergence of angiosperms and gymnosperms. Many more evolutionary scenarios can be conceived, limited only by our imagination. Comprehensive models which take every possible evolutionary phenomenon into account will remain a fantasm. If we are to figure out these sorts of details, however, it will involve finding discrepancies between the data and predictions under some simpler model, a task for which Bayesian statistical inference followed by model checking is *the Right Thing*[19].

The above case study further shows how we can use the results from a statistical reconciliation analysis in combination with external information to check our models further, beyond the posterior predictive simulations of event counts we used earlier. Specifically, we showed that the temporal information in molecular sequence data agrees well with the fitted model of gene family evolution. Of course, the reason we did this is because there is actual cogent information which we did not yet take into account in our analysis, i.e. branch length information. Ideally, this is taken up directly in the analysis, but while awaiting the necessary methodological developments, the weighted K_S distributions shown in the above example provide an informative way to take this information up in the model checking phase.

In the above example analysis our inferences were still conducted under the DLWGD model with a relaxed clock model of rate variation along a time-calibrated species tree. We can however use statistical reconciliation to assess the evidence for WGD in a way more akin to the standard two-step approach, without having to assume a dated species tree and explicit model of duplication and loss rate variation across the tree, nor having to specify a set of WGD hypotheses *a priori*. Indeed, irrespective of whether the DLWGD model actually provides a good model of the evolutionary process, we may use the ALE approach developed in the present chapter to conduct gene tree reconciliation in a way which avoids the pitfalls of the standard two-step approach. Our assumptions on the evolutionary process will be less precarious, while our inferences will be restricted to simpler questions. All this should become more clear in the following example.

Example (Droseraceae). As discussed in chapter 5, the Droseraceae family

[19] "That which is compellingly the correct or appropriate thing to use, do, say, *etc.* Often capitalized, always emphasized in speech as though capitalized. Use of this term often implies that in fact reasonable people may disagree." (The Jargon file, v4.4.7).

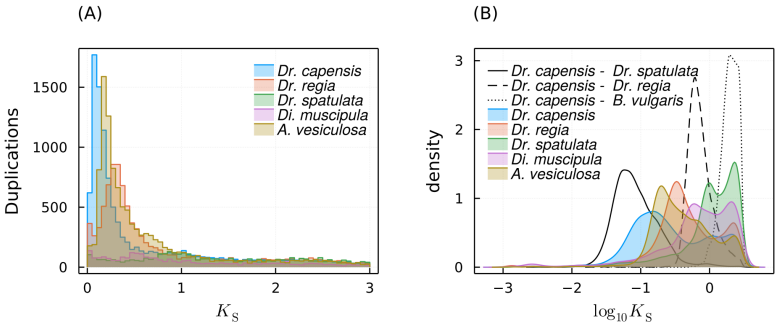


Figure 6.18: Whole-paranome and reciprocal best hit (RBH) ortholog K_S distributions for the Droseraceae data set. (A) Histograms of node-averaged K_S estimates for the paranomes of the five Droseraceae species. (B) Kernel density estimates for whole-paranome and RBH-ortholog K_S distributions on a $\log_{10}$ scale.

of carnivorous plants (sundews and relatives) presents a particularly complicated phylogenetic situation, with an evolutionary history characterized by rampant polyploidization and hybridization. Analyses of genome structure in chromosome-scale assemblies of *Drosera capensis* and *Drosera regia* reveal that the former underwent a recent WGD, whereas the latter shows a clear triplicate structure (see fig. 5.12 in chapter 5). *D. capensis* and *D. spatulata* further seem to show evidence for a more ancient event, likely a triplication (see again fig. 5.12). Whole-paranome K_S distributions are consistent with these observations (fig. 6.18). We note that an assessment of relative substitution rates suggests that the *D. capensis* synonymous substitution rate is about three times that of *D. regia*, so that the *D. regia* WGT is in fact considerably more ancient than the *D. capensis* WGD event, despite their overlapping K_S distributions. The relationship (if any) between the *D. regia* WGT and putative ancient WGT in *D. capensis* was investigated in chapter 5. Here we focus on the recent event in *D. capensis*. The whole-paranome K_S distribution for *D. spatulata* suggests there to be no trace of the recent *D. capensis* WGD in its close relative *D. spatulata*, despite the synonymous divergence of *D. capensis* - *D. spatulata* orthologs being on average smaller than the synonymous divergence associated with the *D. capensis* WGD (fig. 6.18). This indicates either remarkably strong differences in synonymous substitution rates among these closely related lineages, or, more likely, that the recent WGD in *D. capensis* is associated with an allopolyploid hybridization event, wherein the *D. capensis* lineage derives from an allopolyploid hybrid between two lineages that diverged at distinct time points from the *D. spatulata* lineage.

We conduct statistical reconciliation using WHALE to perform a more detailed analysis at the gene family level. We consider a five taxon data set (consisting of *D. capensis*, *D. spatulata*, *D. regia*, *Chenopodium quinoa* and *Beta vulgaris*) (see Appendix B). We inferred *anchor families* (or microsynteny clusters, see e.g. Zhao et al. (2021)) for the whole data set by inferring orthogroups with OrthoFinder (Emms and Kelly 2019) (using default settings) and anchor pairs using I-ADHoRe 3.0 (Proost et al. 2012; Fostier et al. 2011). Anchor families are then obtained as the connected components in the graph where genes are nodes and edges represent anchor pair relationships (i.e. genes which are both homologous and colinear). For each family, we inferred an amino acid MSA using MAFFT (Katoh and Standley 2013) and derived a CCD based on a sample of 10000 trees from the posterior distribution of gene tree topologies obtained using MrBayes, assuming the GTR + Γ4 model and default priors for model parameters, branch lengths and topologies (see also the land plants example above). We restrict the analysis further to those families for which the MrBayes sample had a minimum ESS > 200 for the total tree length and ASDF < 0.025. The final set consists of 6066 anchor families. Note that, since gene duplicates derived from small-scale gene duplication events are not expected to be co-linear, variation in gene copy number within and across anchor families should be largely due to duplications of large chromosomal regions, as for instance caused by WGD or single-chromosome duplications.

We assume an undated species tree phylogeny (see fig. 6.19) and estimate the expected number of duplication and loss events per lineage for each branch under the phylogenetic linear BDP model of gene family evolution, instead of the associated rates per time unit. To do so, we set all species tree branches to one, and assume an iid Exponential prior with mean 0.5 for the branch-specific duplication and loss rates. We assume a Beta$(5, 1)$ prior on the root, roughly motivated by the observed anchor family size for *B. vulgaris*, which is representative of a lineage which has not undergone any WGDs after the eudicot (γ) triplication event (which should be shared by all genomes in the sample). We sample, for each family, 1000 reconciled trees from the posterior distribution and estimate the posterior probability distribution over possible reconciliations for each duplicate gene pair in the data set. These probability distributions are displayed in fig. 6.19. Clearly, most anchor pairs in *D. capensis* appear to derive from duplication events along the *D. capensis* - *D. spatulata* stem branch, and these appear to be associated with the most recent WGD signature in the *D. capensis* K_S distribution (around $K_S \approx 0.2$). On the other hand, the (much less numerous) anchor pairs from *D. spatulata* that

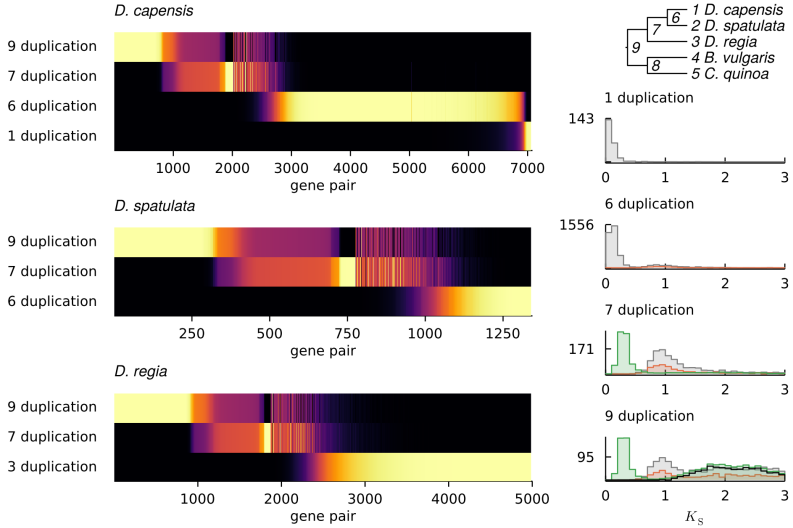


Figure 6.19: Statistical reconciliation analysis for the Droseraceae data set. The heatmaps show the posterior reconciliation probability distributions for each duplicate gene pair in *D. capensis*, *D. spatulata* and *D. regia*, i.e. showing for each pair (column) the estimated posterior probability of being reconciled as a duplication to specific species tree branches (rows) (the color scale is linear with black corresponding to probability 0 and yellow to probability 1. The numeric labels for each branch are displayed in the phylogeny in the upper right corner. The K_S distributions for gene pairs in *D. capensis* (gray), *D. spatulata* (orange) and *D. regia* (green) are shown on the right, where we show for each relevant branch of the species tree weighted histograms, with weights derived from the estimated posterior probabilities (as in fig. 6.16).

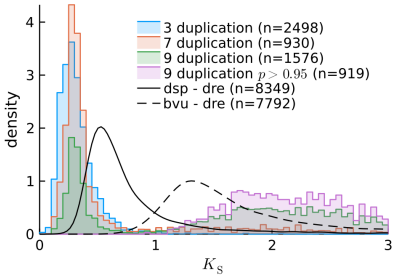


Figure 6.20: Weighted K_S distributions for *D. regia* anchor pairs (weighted by posterior reconciliation probability), as well as anchor pairs between *D. regia* and *D. spatulata* and *B. vulgaris*. We show the sum of the relevant weights *n* as an indicator of the effective number of gene pairs in each component.

have high probability to be derived from duplications along this branch are associated with the older WGD signature shared by (at least) *D. spatulata* and *D. capensis* (at $K_S \approx 1$). This supports the allopolyploidy hypothesis outlined above.

The K_S distribution for the *D. regia* tip branch is nearly indistinguishable from the K_S distribution associated with anchor pairs that putatively derive from the Droseraceae stem branch (fig. 6.20). We note that there appears to be considerable uncertainty in the reconciliations for the deeper events, with many anchor pairs having high posterior probability to reconcile to both the root and the Droseraceae stem. The K_S distributions clearly suggest that those which are reconciled with $p > 0.95$ to the root are really ancient duplicates (probably from the eudicot WGT), whereas the others seem to be associated with the Droseraceae-specific events (fig. 6.20). Note that the low substitution rate in *D. regia* relative to the other *Drosera* species renders the interpretation of K_S distributions difficult. The analysis sheds little additional light on the evolutionary sequence of polyploidization events associated with the anicent WGTs in *D. regia* and *D. capensis*. □

In this example we see clearly how taking into account molecular divergence (branch length information) in gene tree reconciliation analyses has the potential to substantially improve inferences. Indeed, information from K_S distributions suggests that reconciliation of (unrooted) gene tree topologies alone cannot clearly discriminate between putative Droseraceae-specific duplication events and eudicot-shared duplications, assigning appreciable probability for the former to be reconciled to the root of the species tree. We found additional evidence for the hypothesis that *D. capensis* derives from an allopolyploid ancestor, but we have not, however, explicitly accounted for this in our analysis. In our final example we show how to account for allopolyploidization in the statistical reconciliation approach.

Example (*Drosera capensis* allopolyploidy). To further assess the allopolyploidy hypothesis, we extended the WHALE algorithm to deal with so-called multi-labeled (MUL) trees (see also chapter 5). Let $\varrho(e)$ be the species of which node e in the species tree represents a subgenome. If we assume that any assignment of a gene from an (ancient) allo-k-ploid to any of its k subgenomes is equally likely *a priori*, we can modify the ALE recursion for leaf nodes so that

$$p_e(\gamma, 0) = 1/k$$

whenever γ is a leaf clade from $\varrho(e)$ and species $\varrho(e)$ is an allo-k-ploid species with k represented subgenomes in the species tree. Using this slight modi-

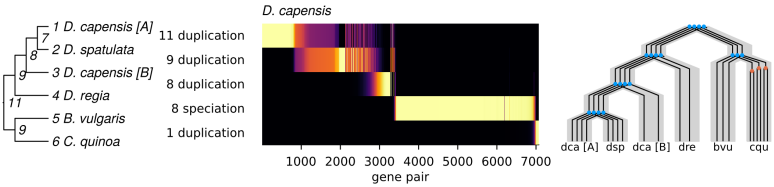


Figure 6.21: Posterior reconciliation probability distributions for the *D. capensis* anchor pairs in the MUL tree reconciliation analysis (see fig. 6.19). The MUL tree with associated branch labels is shown on the left. Example reconciled MUL tree, showing two retained *D. capensis* gene pairs from the allopolyploidy event.

fication of the ALE algorithm, we can conduct statistical reconciliation for MUL trees. In fig. 6.21 we show the posterior reconciliation probability distributions for each *D. capensis* anchor pair. Clearly, these results suggest that the majority of *D. capensis* anchor pairs derive from the putative allopolyploidization event. We again find that deeper events in the gene trees tend to be associated with more uncertain reconciliations.

Note that not only does this allow us to assess allopolyploidy hypotheses, it also admits a statistical approach towards assigning homeologous genes to their respective subgenomes. Indeed, we can sample reconciled trees from the posterior distribution, and for each homeolog simply record the frequency of seeing it reconciled to each of the eligible subgenomes (MUL species tree leaves). However, for the present phylogenetic situation, one can not assign an individual gene from *D. capensis* to a subgenome when it has no retained duplicate using topological information alone. That is, one can not discriminate between '((dca,dsp),loss)' and '((loss,dsp),dca)'. Indeed, the posterior distribution suggests that virtually all gene loss after the allopolyploidy event occurred along the B subgenome lineage, with about 60% of the lineages passing through the subgenome divergence suffering loss along the B subgenome lineage (tbl. 6.4). This is however an artifact, and should not be interpreted as the amount of gene loss along the B lineage, but rather as the amount of gene loss after allopolyploidy across both the A and B lineages. Consequentially, we can only conduct subgenome assignment for retained homeologous pairs. Doing so, we find that about 60% of the *D. capensis* anchor pairs in our data set which reconcile with posterior probability > 0.95 to the node representing the divergence of the subgenomes can be unambiguously phased with respect to the two subgenomes (posterior probability > 0.95 for the subgenome assignment of each gene in the pair). On the other hand, for a substantial portion of the anchor pairs (the other 40%), uncertainty in the gene trees precludes a

definite assignment. □

Table 6.4: Expected number of gene tree nodes per anchor family for each possible species tree node/event combination. Note that the number in the 'speciation' column for leaf nodes (first six rows) gives the average number of genes per family in the data set reconciled to that leaf node.

clade	duplication	loss	speciation
D. capensis (A)	0.02	0.07	1.21
D. capensis (B)	0.00	0.58	0.69
D. spatulata	0.00	0.04	1.22
D. regia	0.39	0.10	1.66
B. vulgaris	0.00	0.04	1.09
C. quinoa	0.90	0.02	2.01
D. capensis (A), *D. spatulata*	0.00	0.01	1.15
D. capensis (A), *D. capensis* (B)	0.04	0.13	0.68
D. capensis, *D. regia*	0.19	0.00	1.14
B. vulgaris, *C. quinoa*	0.05	0.10	1.07
D. capensis, *B. vulgaris*	0.18	.	1.08

We note that the MUL tree approach for dealing with allopolyploidy is qualitatively different from the DLWGD model, in that it does model the divergence of the genomes which come together in a polyploidization event, but does not model the rediploidization process. Indeed, the WGD event is not explicitly represented in the model, and we cannot, therefore, model retention and non-retention of duplicates after WGD as a process separate from the linear BDP model. To do so, we would need a model which explicitly acknowledges the reticulation in the species tree[20].

6.4 Concluding remarks

In this chapter, we have outlined an approach for statistical gene tree reconciliation with phylogenetic BDP models of gene family evolution. Using the principle of amalgamation and the ALE algorithm of Szöllősi, Rosikiewicz, et al. (2013) for approximating the marginal likelihood of the sequence data under a model of sequence and gene family evolution, we have shown how we can jointly reconstruct a gene tree and its reconciliation, or, in other words, conduct model-based gene tree reconciliation while acknowledging uncertainty

[20] Here, we should mention the work of Jones, Sagitov, and Oxelman (2013), where the authors assess more closely the difference between modeling allopolyploid hybridization using MUL trees and phylogenetic networks. The setting there is however quite different, in that the authors focus on species tree inference under the MSC model and do not consider multi-copy gene families.

in the gene tree topology. We implemented the methods in a flexible probabilistic programming framework allowing Bayesian statistical inference for essentially arbitrary hierarchical models with a phylogenetic BDP sampling distribution. This includes, but is not restricted to, relaxed DL clock models to account for rate heterogeneity across lineages of the species tree. We showed, using example analyses, how the statistical approach to gene tree reconciliation can be used to study and infer ancient WGDs in a phylogenetic context, avoiding thereby many of the pitfalls associated with standard two-step approaches used for these purposes.

As we noted in the introduction of the present chapter, approaches for joint Bayesian inference under integrative models for sequence and (multi-copy) gene family evolution have been developed, so an important question is why we should be interested in the ALE approach towards gene tree reconciliation at all. Clearly, an important reason is computational efficiency, which bears not only on issues of time, resources and CO_2, but also of statistical workflow and implementation effort. Indeed, the statistical analyses are hard enough already, and computational bottlenecks render the cycle of designing, fitting, checking and expanding statistical models a painful process, which is likely to affect the quality of the analysis – at least in our experience. By separating the inference of the gene tree topologies from the reconciliation, we can delegate the former to highly optimized MCMC samplers for phylogenetic inference under CTMC models of sequence evolution (such as MrBayes), while focusing our modeling (and implementation) efforts on the models of gene family and genome evolution that we are interested in. The flipside of this division of labor is of course that we lose statistical efficiency. By not being able to take up sequence divergence information beyond the gene tree topologies, we are not fully exploiting all available information in the inference of reconciled gene trees and model parameters. Note that the ALE approach not only leads to computational gains by splitting the problem in two independent steps, but also that because of the efficient marginalization over the gene tree distribution in the amalgamation algorithm we do not need to explicitly deal with reconciled trees during posterior inference. This enables the use of efficient gradient-based MCMC samplers developed for continuous parameter spaces when fitting the model of gene family evolution.

The use of an empirical CCD as an approximation of the posterior distribution over gene tree topologies may be subjected to further scrutiny. Clearly, the ALE algorithm relies crucially on the conditional independence assumption that underpins the CCD family of tree distributions. Furthermore, in order for the ALE algorithm to be feasible, it is necessary for the CCD to

have a relatively sparse support, which further seems to single out empirical CCDs derived from samples of tree topologies. Nevertheless, it would definitely be interesting to investigate whether efficient algorithms for computing the marginal likelihood under a phylogenetic BDP model can be devised for other approximating families which should be more realistic, such as subsplit Bayesian networks (Cheng Zhang and Matsen IV 2018a, 2018b).

The presented approach is intimately tied to the phylogenetic linear BDP model. We have noted how the ALE approach would generalize to more complicated models of gene family evolution, such as the two-type model studied in chapter 3, however, the computational feasibility and statistical performance of this remains to be assessed. We note that Szöllősi, Rosikiewicz, et al. (2013) develop the ALE approach for a model with horizontal gene transfer, a phenomenon which we ignored in the present chapter. A potentially more serious issue is that the linear phylogenetic BDP model does not account for the underlying population-level processes, and assumes, for instance, that divergence of two lineages in the species tree coincides with the divergence of two gene tree lineages. That is, we have been assuming congruence between the locus and gene tree in the sense of Rasmussen and Kellis (2012) (chapter 1). Hence we are not acknowledging the possibility of ILS, neither at the gene nor locus level. Important steps in the direction of integrating MSC-like models for gene genealogies with models of gene family evolution by duplication and loss were taken in Rasmussen and Kellis (2012) and more recently Li et al. (2021), but statistical inference for these models remains challenging. Lastly, we have been assuming a fixed species tree topology in our analyses, so that our parameter space can be brought in bijection with $\mathbb{R}^d$, enabling the application of efficient samplers (as already noted above). It would of course be interesting to use the ALE approach for joint inference of species trees and reconciled gene trees under phylogenetic BDP models of gene family evolution (as for instance PHYLDOG does (Boussau et al. 2013)). This would however again entail a departure from the comfortable setting where we have a real parameter space, and would reintroduce the problem of sampling from a distribution over tree space. Nevertheless, the marginalization over the gene tree distribution in the ALE algorithm entails that we would not have to sample gene tree topologies jointly with species trees, which is the main computational bottleneck in joint inference methods under the MSC (Rannala and Yang 2013). The potential for species tree inference using the ALE approach remains untapped to date.

Important applications of phylogenomic reconciliation analyses are the quantification of rates and events of genome evolutionary processes and the infer-

ence and analysis of ancient WGDs. We have seen how standard two-step approaches based on ML gene tree reconstruction and naive, parsimony-guided reconciliation can be problematic for these purposes, and we have shown through examples how statistical reconciliation can tackle these problems in a much more adequate way. Once again, we noted how ALE-like methods which can take into account branch length information would be desirable, as we currently cannot make use of the distinctively correlated molecular distances of WGD-derived duplicates in the inference of model parameters and reconciled trees. Such an approach could also lead to improved methods for dating ancient WGD events, a challenging problem of considerable interest, given that many questions about the long-term evolutionary importance of polyploidy hinge on our ability to place ancient WGDs in geological time (Fawcett, Maere, and Van De Peer 2009; Vanneste et al. 2014; Clark and Donoghue 2018). Note that the approach presented above could already prove to be helpful in this regard. Indeed, we could use gene families for which there is a reconciled gene tree with distinctively high posterior probability to conduct divergence time estimation with fixed gene tree topologies, using fossil (node) calibrations for certain speciation nodes in the gene trees. Such an approach was in fact adopted in Clark, Puttick, and Donoghue (2019), although, as in Vanneste et al. (2014), the authors conducted molecular dating for each selected gene family independently. One could however share information across families, as is typically done in divergence time inference for species trees from single-copy gene families. To our knowledge, however, no implementation exists at present for estimating divergence times in multi-copy gene families on a genome-wide scale, but we note that, conceptually, when we condition on the *reconciled* gene tree topology, this should be relatively straightforward.

7 Conclusion and future perspectives

En résumé, sans vouloir instruire le lecteur, nous serions payé de nos peines, si nous pouvions le convaincre de pratiquer un exercise où nous sommes maître: se moquer de soi-même. Aucun progrès n'est possible dans la connaissance objective sans cette ironie autocritique.

Gaston Bachelard (1949)

Time has come to look back and face the bold *desiderata* set out in chapter 1. The reader will have noticed that we are still rather far removed from supplanting the plausible adaptive stories that pervade evolutionary genomics with model-based statistical inferences. We believe we did contribute to that challenge, or at least provided some tools to do so. In this final chapter, we shall consider what we did and did not achieve in this regard, and what we believe to be promising avenues for future research.

7.1 Modeling the bag

In order to start making sense of the pile of genomic data, we adopted the bag of genes conception of a genome. The bag was partitioned in evolutionarily relevant units, gene families, and we sought to devise statistical models that describe the variation of these units within and across genomes. We have approached this general problem in two distinct ways. In the first part of this work we considered models for gene content evolution, modeling variation of gene family sizes in the phylogenetic context provided by the species tree. In the second part we adopted a less coarse point of view where we considered gene trees and distributions thereof, modeling variation in evolutionary histories across the genome at a finer scale.

7.1.1 Gene content evolution

In chapters 2 and 3, we focused on modeling variation in gene family sizes using (phylogenetic) birth-death process models. These models are intuitive, in that they appear as reasonable models of the mutational processes that drive small-scale gene duplication and loss. In particular, the linear BDP, which obeys the branching property, appears as a reasonable yet simple model in that regard, assuming that the rate of a duplicative mutation is a property of an individual gene, not a gene family. However, already in the single-genome setting, we found the BDP models to be somewhat less than ideal as models of gene family evolution. Furthermore, in chapter 3 we used Bayesian hierarchical models which account for many of the plausible additional sources of variation (specifically, variation in the number of ancestral genes, rate variation across lineages and rate variation across families), and showed that the linear phylogenetic BDP model is not usually an adequate model of gene family evolution. One source of model violation that we could account for quite succesfully was whole-genome duplication, and we harnessed this source of variation to conduct inference of ancient WGDs in a phylogenetic context and to study differential retention of gene duplicates after WGD and SSD across the genome. However, the fundamental inadequacy of the linear BDP remains, and plagues the interpretation of parameter estimates as rates of long-term genome evolution.

The lack of fit of the linear BDP model has some repercussions for common practices in evolutionary genomics which were not the focus of the present dissertation. Indeed, the linear BDP model is ubiquitously used in model-based approaches to study gene family *expansion* and *contraction* across phylogenies, forming the basis of popular ML inference tools such as CAFE (De Bie et al. 2006a), Count (Csűrös 2010) and BadiRate (Librado, Vieira, and Rozas 2012). Such analyses are often taken as providing a statistical foundation to claims of adaptive evolution at the genomic level. Many studies have reported 'significantly' expanding or contracting gene families by means of an approach which boils down to signaling gene families for which transitions between inferred (ML) ancestral states are unlikely under the fitted linear BDP model, usually without taking into account lineage or family-specific variation in evolutionary rates. In other words, the most common use of these tools is to signal lack of fit of a simple linear BDP model, but the causes of this lack of fit are rarely, if ever, investigated. We showed that the linear BDP almost never provides an adequate fit to the data, so it would appear that these approaches are really rejecting a strawman model. While this commonly adopted approach

can signal interesting families, the claim that it provides 'a statistical foundation for evolutionary inferences'[1] appears moot. Indeed, the mere fact that some approach allows one to compute a p-value does not make that approach statistically more sound than providing, for instance, a ranking of gene families based on some summary statistic of the phylogenetic profile matrix (an approach which has also appeared from time to time in the literature). We believe a lack of fit should prompt one to look for better models which can accommodate the observed variation, in a similar way as we have for instance accounted for WGD events.

Luckily, the approach is somewhat less precarious than the above may suggest, because the ML ancestral states are typically reasonably robust to model violations, although the computed p-values remain meaningless if the 'null' model does not fit a single family. A worthwhile avenue for further research would hence be to assess the robustness of evolutionary claims of 'significant' expansion or contraction based on the linear BDP in the face of the pervasive model violations we highlighted. We note that the approach for detecting gene family expansion and contraction under the linear BDP really amounts to an approximate form of posterior predictive model checking, relying on point estimates (MLEs) instead of a posterior distribution over parameters and ancestral states. In our view, the whole approach would benefit from a more explicit Bayesian treatment, and in this work we provided the building blocks to do so. An in depth study of how one could conduct a more meaningful statistical analysis of gene family expansion and contraction using the linear phylogenetic BDP in Bayesian hierarchical models is however deferred to future work.

But the problem is really with the linear BDP itself, and merely embedding the linear BDP model in ever more complicated hierarchical models and sophisticated Bayesian analyses will not solve the issue. The fundamental issue is that, while the linear BDP (or any other branching process) may appear a reasonable model of the mutational process, it is not a suitable model for the *population genetic* processes that determine the fate of gene duplication and loss mutations, and hence the long-term evolution of the 'bag of genes'. In particular, the effect on fitness of degenerative mutations which lead to gene loss will differ systematically depending on the number of genes in a family, violating the independence assumptions that underpin the ubiquitously used BDP models. We explored models which can account for this in sec. 3.3, where we focused in particular on a two-type branching process model. While this model is interesting both from a probabilistic and evolutionary biological

[1] See e.g. the CAFE 5 GitHub page: https://github.com/hahnlab/CAFE5 (last accessed: May 2022).

perspective (as we argued in the relevant chapter), the gain in model fit was not as great as we would have hoped, while the computational cost increased markedly.

Given that we had to restrict analyses with the two-type model to relatively small gene families and models which do not take into account variation across lineages, it might be worthwhile to adopt a general BDP model instead. In particular, a simple model with $\lambda_i = i\lambda$ for $i > 0$, $\mu_i = i\mu$ for $i > 1$ and $\mu_1 < \mu$ a third parameter to be estimated from the data may already accommodate the most serious violation of the linear BDP model. While this may seem like only a minor modification of the linear BDP, the mathematical treatment is vastly different (as far as we can tell) because of the lack of the branching property. Calculation of transition probabilities would hence have to be performed approximately using matrix exponentiation (as for the usual CTMC models of sequence evolution) or the numerical techniques of Crawford and Suchard (2012). While this would only address part of the issue (as we have noted in sec. 3.3), if the main goal is to get rid of the model violations for the linear BDP, this may be the most efficient strategy.

However, a more ambitious goal is to work out substantive models of evolution, which are not merely tailored to yield good predictive performance, but which involve stochastic processes that we consider adequate models of the evolutionary processes of interest, and where we can ascribe meaningful interpretations to key parameters in the context of a scientific theory. This kind of desire has been expressed repeatedly in statistical evolutionary genomics (e.g. Szöllősi et al. 2015; Scornavacca, Delsuc, and Galtier 2020), and demands, in particular, close attention to how we can link the microevolutionary population genetic processes with the long-term phylogenetic time scales. Important work in this regard is performed for models of sequence evolution, we think for instance of mutation-selection models of codon evolution (Rodrigue, Philippe, and Lartillot 2010; Rodrigue and Lartillot 2017; Latrille, Lanore, and Lartillot 2021) or polymorphism-aware models of sequence evolution (Wilson et al. 2011; De Maio, Schrempf, and Kosiol 2015; Borges, Szöllősi, and Kosiol 2019), and indeed, much of the literature surrounding the multispecies coalescent is somewhere between population genetics and more long term evolutionary genomics (e.g. Hobolth et al. 2011). Strikingly, virtually no such models have been developed for the processes of gene duplication and loss, which are clearly among the most important determinants of long-term genome evolution. So far, little, if any, of the rather extensive population genetics theory on gene duplication and loss (e.g. Kimura and King 1979; Watterson 1983; Ohta 1987; Force et al. 1999; Walsh 2003), has found

its way into evolutionary genomics. Our two-type branching process model goes somewhat in that direction, but not nearly as far as we would like it to.

Additionally, when considering the population genetic processes that affect genomic diversity, we find that the different sources of variation all arise in the same 'population genetic environment' (Lynch 2007). This leads to all sorts of correlations which could be taken up in statistical models. In particular, the effects of population size on genetic drift affect newborn point mutations and gene duplications in similar ways, so that rates of long-term sequence and gene family evolution should be correlated. Devising models which can account for and harness evolutionary correlations across these different phenotypes, esentially bringing in more explicitly *quantitative genetics* in our study of long-term genome evolution, appears to be another promising avenue for further research. Approaches like the one of Lartillot and Poujol (2011) or Lartillot (2012) may serve as an inspiration in that regard.

7.1.2 Phylogenomic forestry

In the second part of this thesis we dipped a bit deeper in the bag, not only considering gene family sizes, but also the evolutionary relationships among gene family members, as represented by their *gene tree*. Many researchers have sought to learn about genome evolution from gene trees, either for reconstructing species trees or for inferring reconciled gene trees, which anchor the evolution of a gene family within the phylogenetic context provided by the species tree.

Modeling the collection of gene trees across the genome admits more detailed inferences about the evolutionary processes that cause genomic variation, but comes at a substantially increased computational cost. Indeed, while joint Bayesian hierarchical models roughly of the form

$$S \sim \text{GenomeLevelModel} \tag{7.1}$$

$$\mathcal{G}_i | S \sim \text{GeneFamilyLevelModel} \qquad 1 \leq i \leq n \tag{7.2}$$

$$y_i | \mathcal{G}_i \sim \text{SequenceLevelModel} \qquad 1 \leq i \leq n \tag{7.3}$$

(where $S, \mathcal{G}$ and y represent the species tree, gene tree and sequence data respectively) may be relatively easy to specify in theory, conducting inference in reasonable time with reasonable resources turns out to be rather challenging. With data sets of continuously increasing size, these approaches are infeasible and unsustainable. Hence many people have sought to 'shortcut' inference for

these models, essentially splitting the model in two independent components, where gene trees are inferred from sequence data, and the gene trees are then considered as data for studying the processes at the genome level.

However, learning from inferred trees asumed as *data* is, although widespread, a statistically tricky business. Most approaches have relied on gene trees estimated from sequence data using ML, treating these as if observed. Not only are these estimators noisy due to errors in the alignment, model violations and the use of heuristic optimization algorithms for inference, even if there were no such sources of error, the inferred gene tree remains a statistical inference with associated uncertainty. Ignoring this uncertainty is likely a major source of issues for phylogenomic inference of species trees under the multispecies coalecent model and phylogenomic reconciliation analyses. The former have become a mainstay in large-scale phylogenetic studies, whereas the latter aim to quantify the extent of gene duplication, gene loss and horizontal gene transfer over long evolutionary time scales and are frequently used as a basis for, again, claims of adaptive genome evolution and inference of ancient WGD events.

In our work in chapters 4, 5 and 6 we adopted an intermediary viewpoint, relying on a two step approach which breaks up of the hierarchical model of eq. 7.3, while taking into account uncertainty in the gene trees. Our approach hinged crucially on a class of distributions over tree topologies, the CCD, first studied by Larget (2013) and here considered in some detail in chapter 4. In chapter 5, we show how the CCD can be used for inference under the MSC from uncertain gene tree topologies, using an approach based on likelihood-free expectation propagation. Here, the CCD finds dual use: as an ABC kernel on the one hand and as a variational family for the species tree posterior on the other. Our usage of EP appears to be unique in the phylogenetic literature, and while our method is still in a rather experimental stage, we believe it shows the merits of this somewhat underappreciated (at least in genomics) approach to Bayesian inference. As we noted in the discussion of the relevant chapter, likelihood-free approaches are typically much more flexible when it comes to inference, bypassing the need to implement efficient algorithms for evaluating the typically computationally intensive or intractable likelihood function, focusing instead on the simulation of 'fake data' under the model. In that regard we hope that our approach could enable inference for more adequate models of genome evolution which have so far remained elusive. In addition, the shift of focus from evaluating the likelihood, which is in itself a meaningless number, to simulation of the data-generating process is interesting from a Bayesian perspective in that the assessment of model fit is to some extent taken up already

during inference. Indeed, in contrast with typical likelihood-based methods, which 'fail silently' as far as model fit is concerned, ABC methods can fail rather loudly when the model does not fit the data. We made use of this fact when we considered outlier loci in our analysis of the Chiari et al. (2012) data set in chapter 5.

Our work in chapter 6 was mostly motivated by the problem of phylogenomic inference of ancient WGDs, an important problem in plant evolutionary genomics which has received a lot of attention. The literature on the subject, as we have noted, is characterized by an application of rather *ad hoc* methods, and building on the work of Szöllősi, Rosikiewicz, et al. (2013) and Rabier, Ta, and Ané (2014), we sought to develop a more principled approach based on statistical reconciliation using phylogenetic BDP models. Of course, this approach suffers from the same issues of model fit as our analyses of gene counts under the phylogenetic linear BDP. However, as we showed using posterior predictive simulations, this appears not to be an issue for inference of the number of evolutionary events along different branches of the species tree (this is similar to the robustness of inferred ancestral gene counts to model violations noted above). Again, however, better substantive models of the evolutionary process would be welcome and would likely benefit the inference of ancient WGDs. We noted, briefly, how the two-type process could be implemented for the ALE-based statistical reconciliation approach, and this appears to us as another fruitful avenue for further research. We note that this need not be computationally prohibitive if one is willing to make certain approximations in the discretized ALE algorithm, for instance by approximating the continuous-time two-type process by a discrete-time branching process.

Developing aproaches to take up branch length information in the methods we presented in chapter 5 and chapter 6 should be a likely rewarding effort. Indeed, it appears to us that the main disadvantage of our phylogenomic forest methods compared to the 'full-Bayes' approaches based on eq. 7.3 is their inability to take up information from molecular divergence beyond the effect the latter has on topological uncertainty. Our analysis of the *Drosera* data in chapter 6 highlighted this. If we would be able to take up such information in the analysis, this might further enable considerable improvements for molecular divergence estimation from multi-copy gene families. Indeed, it appears plausible to us that, if one could take into account molecular divergence of the sequences in a gene family in genome-wide reconciliation analyses along the lines of those we conducted in chapter 6, one could further push the Bayesian hierarchical model to include unknown node ages in the species tree. This seems, to us, an exciting area of further research, which would not

only provide an interesting approach for inferring time-calibrated species trees, but also would yield a statistically well-founded approach for dating ancient WGDs. The latter is in turn vital for our assessment of the macroevolutionary importance of ancient WGDs (Van de Peer, Mizrachi, and Marchal 2017; Clark and Donoghue 2018).

7.2 Beyond the bag

But really, our preoccupation with the 'bag of genes' model of a genome is becoming increasingly unjustified. With the quality of genome assemblies for non-model organisms now routinely becoming on a par with what used to be the exceptional level of assembly contiguity of the human, mouse or *Arabidopsis* genomes; models of evolution which ignore the wealth of evolutionary information residing in genome *structure* appear increasingly ill-motivated. We are of course not the first to lament this state of affairs, and similar sentiments are expressed at regular intervals in the literature (e.g. Yang 2006; Szöllősi et al. 2015; Scornavacca, Delsuc, and Galtier 2020). Devising tractable statistical models at this level of organization is however notoriously complicated. The key question here is: how can we conceptualize a genome in a way that retains structural information, yet yields feasible modeling approaches?

The 'gene list' conception of a genome, already introduced in chapter 1, is one strategy that is commonly taken as a starting point. At this level of abstraction, the problem is usually framed as the modeling of *genome rearrangement* processes, which essentially consider models of evolution which can turn one gene list into another by means of a set of admissible evolutionary events (inversions, translocations, fusions, fissions, *etc.*) (Hannenhalli and Pevzner 1995; Bafna and Pevzner 1996; Pevzner and Tesler 2003; Fertin et al. 2009). The combinatorial challenges associated with problems of this kind have spawned an entire industry of computer scientists and mathematicians working on algorithmic problems which are often rather remote from evolutionary concerns. *Statistical* modeling of genome rearrangement processes has not really taken off yet, with not much progress after several promising early attempts (e.g. Larget, Kadane, and Simon 2002; York, Durrett, and Nielsen 2002; Durrett, Nielsen, and York 2004; Miklós and Tannier 2010). We here note the work of Nakatani and McLysaght (2017) and Nakatani et al. (2021) which describes a Bayesian approach towards ancestral genome reconstruction that may open up new opportunities for modeling genome evolution at a coarse scale. Such *macrosynteny* approaches, on which we are presently

working, currently serve mainly to delineate genomic homology on a large scale (fig. 5.12 provides an illustration of our results in that regard). However, the foundation of these methods in relatively simple parametric models may enable their embedding in a more explicit phylogenetic context and the incorporation of more substantive modeling assumptions regarding large-scale genome evolution. Whether this is indeed the case remains an open question at this point.

However, we have not, usually, modeled a collection of sequences or a bag of genes *an sich*, and have additionally relied on rather strong *assumptions of homology* to do so. Indeed, nothing remains of our approaches if we do not have orthogroups and multiple sequence alignments, although both are rather questionable constructs. Even if we would have methods that succeed with high accuracy in uncovering such homology relationships among sites and genes, many aspects of actual evolutionary processes are bracketed when making the assumption that we can in fact atomize genes and sequences in the way we do. For instance, a typical protein-coding gene consists of multiple protein domains, which may have quite different evolutionary histories, something we have glossed over completely. Sequences evolve not only by point mutations, but also by insertions, deletions, gene conversion *etc.*, all of which trouble the assertion of homology relationships in the form of an alignment matrix and are ignored in typical phylogenetic analyses using CTMC models of sequence evolution. Nevertheless, a rich literature and wealth of methods has emerged from these seemingly precarious foundations, and they continue to help us in making sense of (molecular) evolution at the sequence and genome scale, as we have had ample opportunity to show in this thesis.

By analogy, then, another question appears: how can we 'atomize' a genome, conceived as a gene list (or collection of such lists), in a way which retains structural information, but does allow us to devise tractable models? The relevant homology concepts that can be defined for this representation of a genome are *synteny* (roughly, conserved gene content) and *colinearity* (conserved gene order). We have ourselves used the latter to peak a little bit beyond the bag in our work, using information from synteny and colinearity analyses in both chapter 5 and 6 to select interesting subsets of gene families. However, we have not taken up this information in any way in the probabilistic models of evolution that are at the core of our statistical inferences. Several potentially rewarding avenues for further research appear to us in this regard. We noted, for instance, in chapter 3 that modeling gene counts may not be a very data-efficient way for inferring ancient WGDs in a phylogenetic context, especially with increasing availability of high-quality chromosome-scale genome

assemblies. However, nothing prevents us from applying similar models to more rich data, such as for instance phylogenetic profiles of anchor families (e.g. Zhao et al. 2021, see also chapter 6). This could, finally, provide us a robust yet reasonable simple statistical method for the inference of ancient WGDs in a phylogenetic context. Indeed, we conjecture that an approach based on a BDP-like model for anchor families, together with something like the rjMCMC strategy to sample across the space of possible WGD hypotheses (like the one we developed in Zwaenepoel and Van de Peer (2020) and chapter 3) could solve many of the issues of statistical WGD inference. Similarly, incorporating synteny or colinearity information in gene tree reconciliation, for instance by modeling microsynteny networks in the sense of Zhao and Schranz (2019), could yield interesting methods for species tree inference, WGD inference and studying rates of genome structure evolution. Duchemin et al. (2017) is a noteworthy example of an attempt to do so, although not using probabilistic models of evolution nor statistical inference.

7.3 Epilogue

So what good are all these models and methods? Do they help us to make evolutionary sense of genomic data? Do they enable a better understanding of genome evolution? Do they actually help us to prevent falling for the *just-so* stories of Gould (1978)? For one thing, the naturalist's eye is lost in the face of large genomic data sets, and we need models at least as *probes* to interpret these data in an evolutionary context. Indeed a model is like an arrow which we can throw at empirical data to get something out of the latter. Hence, what we see is what the model allows us to see. Without models, there is not much to be seen, and without *statistical* models, what we see is challenging to interpret. In this regard, statistical evolutionary genomics is already a highly succesful endeavor, and essential to our current approaches for understanding genomic diversity. However, at present, we remain stuck in a rather high-level discourse, talking about genome-wide and long-term rates of evolution, patterns of lineage and family variation, inferring phylogenies at various scales *etc.*, but rarely come to the point of what many a biologist would recognize as an evolutionary explanation of some biological phenomenon. Without much knowledge about what all these genes and genomes are supposed to mean for the development and maintainance of a living system, however, evolutionary inferences at this level of detail remain out of scope. Unraveling all these functional questions is however, in our view, hardly the task of evolutionary biology. Nevertheless, our high-level evolutionary inferences (and particu-

lar evolutionary reconstructions) may be helpful in constructing hypotheses concerning functional aspects and the evolutionary constraints they engender, which may in turn lead to more substantive and detailed evolutionary hypotheses. Whenever such evolutionary hypotheses are championed, one should seek to test them, and this *is* the task of evolutionary biology. Devising models and confronting them with data is then, once more, the 'right thing' to do. We hope to have made some contribution to our means to do so.

Give any one thought a push:
it falls down easily;
but the pusher and the pushed produce
that entertainment called a discussion.
Shall we have one later?

John Cage – Lecture on Nothing

A Bayesian computation

Many applications of Bayesian statistical methods lead to serious computational challenges. Once a suitable probability model has been determined, the objective of a Bayesian analysis is usually to arrive at some Bayes estimator $\hat{\theta}$ for an unknown quantity θ of interest, having the following general form

$$\hat{\theta} = \text{argmin}_{\hat{\theta}} \mathbb{E}[L(\theta, \hat{\theta})|y] \tag{A.1}$$

Where L is a suitable loss function and y is the observed data (Robert 2007; Bernardo and Smith 2009). For general L, we will not have a closed form for (A.1), nor for the expectation in it, and simple algorithms like numerical quadrature are only feasible for low-dimensional θ. Numerical approximation of challenging high-dimensional integrals is therefore a routine task in Bayesian inference, and a whole battery of algorithms has been developed to this end. It is no exaggeration to say that we owe the revival of Bayesian statistics in the present century largely to the availability of efficient general purpose algorithms and increased computational power. Problems in phylogenetics and evolutionary genomics are among those presenting serious computational challenges, and we use a number of different methods for approximating posterior distributions in this thesis. This chapter provides a brief overview of some of the relevant methods.

A.1 Monte Carlo integration and sampling

Let $X : \Omega \to \mathbb{R}$ be a random variable with distribution function μ and let h be a measurable function, then we can approximate the integral (provided it

exists)

$$\mathbb{E}_\mu[h] = \int_{-\infty}^{\infty} h(x)\mu(dx) \approx \frac{1}{N}\sum_{i=1}^{N} h(X_i) \qquad \text{where } X_i \sim \mu$$

Here $(X_1, X_2, \dots, X_n)$ constitutes a sample from the density μ. By the strong law of large numbers the Monte Carlo approximation converges almost surely to the expectation. If we can therefore *simulate* random draws from the density μ, we can use these simulations to estimate intractable integrals. Note that the same idea can be used to approximate the probability density function (pdf) of μ using simulation. Specifically, using an iid sample $(X_1, \dots, X_n)$ from μ, we can approximate the pdf by the (random) *empirical measure*

$$\hat{\mu}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}(x)$$

where $\delta_x(\cdot)$ is the Dirac mass at x. Seen in this light, the Monte Carlo approximation of $\mathbb{E}_\mu[h]$ is equivalent to $\mathbb{E}_\mu[h] \approx \mathbb{E}_{\hat{\mu}}[h]$. The conclusion is the same: the problem of integration has been recast as a problem of simulation.

For general probability measures μ, simulating random draws is itself however a challenging task, and we have to take recourse to *sampling algorithms* which can be used to simulate from arbitrary densities. The goal of a sampling algorithm is to generate random realizations from a prescribed probability distribution $\mu(x)$ (the *target*) using random numbers[1] from probability distributions which admit direct simulation (i.e. the uniform distribution and those distributions that can be related to it in a relatively simple way, using for instance the inverse transform method, see e.g. Devroye (2006)). The rest of this chapter will be a brief exposition of the main sampling algorithms that are used in the context of this thesis. We will specialize the discussion to the Bayesian setting, where the aim is to sample from a target distribution $p(\theta|y) \propto p(y|\theta)p(\theta)$, where $\theta \in \Theta$. Note that the marginal likelihood $p(y)$ is in general an unknown normalizing constant, so that the goal is to simulate from $p(\theta|y)$ assuming only the possibility to evaluate the *unnormalized* target distribution $\pi(\theta) = p(y|\theta)p(\theta)$. We will also consider the so-called *likelihood-*

[1] One might object that we only have *pseudo-random number* generators at our disposal, such as the Mersenne twister algorithm. These algorithms generate a deterministic sequence of bit-strings with certain statistical properties that render it indistinguishable from a sequence of random points from Lebesgue measure on $[0, 1]$. I am however reluctant to ascribe meaning to the notion of *pseudo*-randomness, for it is unclear to me what *true* randomness would be. Pseudo-uniform I may find acceptable.

free[2] setting where we seek to sample from a posterior density $\pi(\theta)$ without having to evaluate $p(y|\theta)$. Many good resources on the topic of simulation and computational methods for Bayesian inference exist, of which we name Robert and Casella (1999), MacKay (2003), Gelman et al. (2013) and Sisson, Fan, and Beaumont (2018).

A.2 Rejection sampling

Two relatively straightforward methods for sampling from an arbitrary target are *rejection sampling* and *importance sampling*. While these are typically not sufficiently efficient for high-dimensional spaces, they assume an important role in the context of likelihood-free approximate Bayesian computation (ABC) algorithms, among which those ABC algorithms developed in this thesis. We shall briefly describe both in this and the following section.

The basic idea of *rejection sampling*, also called the 'accept-reject' algorithm, is that when $X \sim f(x)$ and $U \sim \text{Uniform}(0,1)$, the joint distribution of the pair $(X, Uf(X)) \sim \text{Uniform}(A)$, where

$$A = \{(x,u) : 0 < u < f(x),\ x \in \text{dom} f\}$$

This can be seen by considering

$$\begin{aligned} X &\sim f_X(x) \\ U|X &\sim \text{Uniform}(0, f_X(X)) \end{aligned}$$

So that the joint density $f_{X,U}(x,u) = f_{U|X}(u|x)f_X(x) = \frac{1}{f(x)}f(x) = 1$ for $0 < u < f(x)$. Hence, if we can simulate pairs from $\text{Uniform}(A)$ and marginalize over U, we obtain a sample from f. Now this appears quite useless, since we cannot easily simulate uniformly from the set A without being able to simulate from the density f of interest in the first place. Nevertheless, this result has been considered important enough to be called the 'fundamental theorem of simulation' (Robert and Casella 1999).

The reason this does lead to a viable strategy is that we can usually simulate iid pairs $Y = \{Y_1, \ldots, Y_n\}$ uniformly from some set A' such that $A \subseteq A'$,

[2] This is somewhat of a misnomer, since in no way can we actually bypass the likelihood or the likelihood principle that is at the core of Bayesian statistics. We merely assume the likelihood function to be intractable or computationally expensive to evaluate.

and that $Y \cap A$ will then be an iid sample uniformly distributed on A (Devroye 2006, theorem 3.2). To obtain an iid sample from $f(x)$ using this idea, the usual strategy is to select a sampling density $g(x)$ which admits simulation by direct methods (the instrumental density) and a constant M such that $Mg(x) > f(x)$. With such a $g(x)$ and M available, it is straightforward to sample pairs uniformly from a set A' for which $A \subseteq A'$ by using the following sampling scheme:

$$\begin{aligned} X &\sim g(x) \\ U|X &\sim \text{Uniform}(0, Mg(X)) \end{aligned}$$

Upon obtaining a size m iid sample of (X_i, U_i) pairs, the set $\{X_i \colon U_i < f(X_i), 1 \leq i \leq m\}$ will constitute an iid sample of random size from f. Note that this approach will work equally well when f is an unnormalized density as long as the condition $Mg(x) > f(x)$ is satified. One can show that the number of (X, U) pairs needed to simulate one realization of $Z^{-1}f(x)$, where Z is a normalizing constant, has a geometric distribution with mean Z/M, so the smaller M the more efficient the algorithm. Specializing to the Bayesian setting, we see that to simulate a single draw from $p(\theta|y)$ using rejection sampling, we simulate a realization θ from another density g which dominates $p(\theta|y)$ and accept θ as a draw from the target with probability

$$p_a(\theta) = \frac{\pi(\theta)}{Mg(\theta)}$$

where M is an upper bound on the density ratio $\pi(\theta)/g(\theta)$. If we consider using the prior as sampling density, we see that $p_a \propto p(y|\theta)$, and that $M \geq \max_\theta p(y|\theta)$.

A.3 Importance sampling

The key observation leading to the idea of *importance sampling* is that the same integral can be expressed with respect to different measures. Specifically, if g is absolutely continuous with respect to f, we have the following identity

$$\mathbb{E}[h(X)] = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx$$

While apparently trivial as a statement about integrals, from the probabilistic perspective this is an important realization, since a random variable with den-

sity g may have very different properties compared to a random variable distributed according to f (for instance, g may admit efficient simulation, when f does not). The quantity $w(x) = f(x)/g(x)$ is called the (normalized) *importance weight*. The importance sampling Monte Carlo approximation is then

$$\mathbb{E}[h] \approx \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{g(X_i)} h(X_i) \qquad \text{where } X_i \sim g$$

We can use importance sampling in the setting where the target density is known up to a normalizing constant. Consider again the Bayesian setting where we want to approximate $\mathbb{E}[h(\theta)|y]$, we have

$$\begin{aligned} \mathbb{E}[h(\theta)|y] &= \int_{\Theta} h(\theta) p(\theta|y) d\theta = \frac{1}{p(y)} \int_{\Theta} h(\theta) \frac{p(y,\theta)}{g(\theta)} g(\theta) d\theta \\ &\approx \frac{1}{N p(y)} \sum_{i=1}^{N} h(\theta_i) w(\theta_i) \end{aligned} \tag{A.2}$$

where $(\theta_1, \theta_2, \ldots, \theta_N)$ is an iid sample from density $g(\cdot)$. Now, from the above, it follows that

$$p(y) = \int_{\Theta} p(y,\theta) d\theta \approx \frac{1}{N} \sum_{i=1}^{N} w(\theta_i)$$

Plugging this in eq. A.2, we get

$$\mathbb{E}[h(\theta)|y] \approx \sum_{i=1}^{N} h(\theta_i) \frac{w(\theta_i)}{\sum_{i=1}^{N} w(\theta_i)} := \sum_{i=1}^{N} h(\theta_i) W(\theta_i)$$

Where $W(\theta_i)$ are the *self-normalized* importance weights. Clearly, to compute the $W(\theta)$, we only need to evaluate the target up to a constant. We can obtain expectations with respect to a posterior distribution *and* get an estimate of the marginal likelihood $p(y)$ using this strategy, the latter being very useful in the context of Bayesian model selection.

The efficiency of an importance sampling algorithm is determined by the variation of the importance weights. If the $W(\theta_i)$ vary wildly, with many having very small values ≈ 0 and some $W(\theta_i)$ rather large, the resulting importance sampling estimate will be dominated by the small number of samples with large weights, and the resulting estimator will have a high variance. When $g(\theta) \propto p(\theta|y)$, we see that $W(\theta_i) \propto 1$, so that each sample contributes equally to the computed expected value. To measure the deviation from this ideal situ-

ation, one can use the so-called *effective sample size*[3] (ESS) (Kong 1992; Liu 2001)

$$\text{ESS} = \Big(\sum_{i=1}^{N} W(\theta_i)^2 \Big)^{-1}$$

which provides a crude estimate of the number of equally weighted samples to which our actual sample is equivalent. A number of methods are available for reducing the variance of the importance weights and stabilizing importance sampling estimators. An interesting example, which also provides other diagnostic tools than the ESS, is the Pareto-smoothed importance sampling method of Vehtari et al. (2015).

Importance sampling constitutes the basis of a number of more involved methods for sampling from complicated target distributions, such as sequential importance sampling (SIS) (Liu, Chen, and Wong 1998) and sequential Monte Carlo (SMC) (Doucet, Freitas, and Gordon 2001) methods. The latter are sometimes referred to as *particle filters*, as we can view these as propagating a set of N 'particles', $\{(\theta_1, w_1), \ldots (\theta_N, w_N)\}$, which are propagated through a series of reweighting, resampling and perturbation steps so that they eventually provide a high-quality sample from the target. The essential idea behind these methods is that, while it is often difficult to design a good importance sampling density which is reasonably well aligned with the target, it is often feasible to learn such a sampling density algorithmically. We shall not discuss these methods here in detail, referring to e.g. Doucet, Johansen, and others (2009) for a well-known treatment of such methods.

A.4 Markov chain Monte Carlo

In 1953, at the (in)famous Los Alamos national labaratory, Metropolis et al. invented an exceptionally powerful method to simulate the equilibrium dynamics of a complicated system arising in statistical mechanics, without having to simulate the actual many-particle dynamical system (Metropolis et al. 1953). Instead, the authors simulated a much simpler ergodic Markov chain, with a stationary distribution provably equivalent to the equilibrium distribution of their system of interest. The method, and its generalization by Hast-

[3] Note that the ESS is a rather general concept, and much like other 'effective sizes' (e.g. effective population size in population genetics, effective number of loci in quantitative genetics, *etc.*) can be defined in multiple ways. In particular, the ESS in the context of Markov chain Monte Carlo has more or less the same meaning, but a very different mathematical definition.

ings (1970), was heavily used in statistical physics and chemistry, but was only gradually realized to be of much wider significance in statistics in general (Gelfand and Smith 1990; Geyer 2011; Robert and Casella 2011). The general method of sampling from an arbitrary target distribution by simulating an ergodic Markov chain with the desired target as stationary distribution came to be known as Markov chain Monte Carlo (MCMC), and is the single most important methodological advance at the root of the revival of Bayesian methods in statistics. Indeed, one author goes as far as to claim that:

> "Whatever the philosophical analysis that leads you to conclude that a particular statistical procedure is 'the Right Thing'[4], that is what you must do, because some form of MCMC will enable you to do it." (Geyer 1998)

The Metropolis-Hastings (MH) algorithm has been further generalized by Green (1995), and forms the basis for the extremely powerful Hamiltonian Monte Carlo (HMC) algorithms (such as the "no U-turn sampling" (NUTS) algorithm of Hoffman and Gelman (2014)). In this section we briefly recapitulate the principle behind MCMC methods and provide a quick sketch of the main methods used in this thesis.

A.4.1 The Metropolis-Hastings algorithm

Assume we wish to sample from a target distribution π, on a probability space $(\mathcal{X}, \mathcal{F}, \mathbb{P})$. As already indicated, the basic idea of MCMC is to simulate a Markov chain $(X_1, X_2, \dots)$ which has π as stationary distribution. To do so, we shall devise a *Markov proposal kernel* P, defined as

$$P(x, A) = \mathbb{P}\{X_n \in A | X_{n-1} = x\} \qquad A \in \mathcal{F}$$

which preserves π, that is, for which

$$\pi(A) = \int P(x, A)\pi(dx) \tag{A.3}$$

The strength of MCMC methods lies in the fact that there exist generic strategies to design such proposal kernels. One such strategy is the Metropolis-Hastings update, which works for any distribution π for which we can eval-

[4] "That which is compellingly the correct or appropriate thing to use, do, say, *etc.* Often capitalized, always emphasized in speech as though capitalized. Use of this term often implies that in fact reasonable people may disagree." (The Jargon file, v4.4.7).

uate its density up to a normalizing constant. Suppose the Markov chain is at iteration n in state $X_n = x$. The Metropolis-Hastings update makes use of two steps to provide a Markovian transition which generates X_{n+1} while preserving the target density: (1) a proposal step and (2) an accept-reject step. In the proposal step, we use a proposal distribution q, which may or may not be dependent on the present state x, to obtain a new value of the state x', i.e. we sample

$$x' \sim q(\cdot|x)$$

Next, we compute the acceptance probability for the proposal

$$a(x', x) = \min\left(1, \frac{\pi(x')q(x|x')}{\pi(x')q(x'|x)}\right)$$

and we set the state of the chain at the next iteration as

$$X_{n+1} = \begin{cases} x' & \text{with probability } a(x', x) \\ x & \text{else} \end{cases}$$

If we run such a simulation long enough, starting from a suitable point X_0 in the state space, we will generate a sample $(X_1, X_2, \ldots, X_N)$ which is distributed according to the target density π. The probability of an accepted proposal is

$$p_a(x) = \int a(x', x)q(x'|x)dx'$$

So that we can express the Markov kernel as

$$P(x, A) = (1 - p_a(x))\mathbb{1}_A(x) + \int_A a(x', x)q(x'|x)dx'$$

for which it is not hard to show that it satisfies eq. A.3, (Geyer 1998).

To make this more concrete, consider the very simple one-dimensional problem we have dealt with in chapter 1, i.e. the estimation of the distance parameter θ for the Jukes & Cantor (JC) model. For a simple symmetric proposal kernel q (e.g. a Gaussian or uniform distribution with mean 0), the following bit of code implements the MH transition kernel for an unnormalized target $\pi(\theta)$:

```
function mhstep(θ, p, q, π)
    θ_ = θ + rand(q) # propose change to θ
    p_ = π(θ_)       # evaluate target at θ_
```

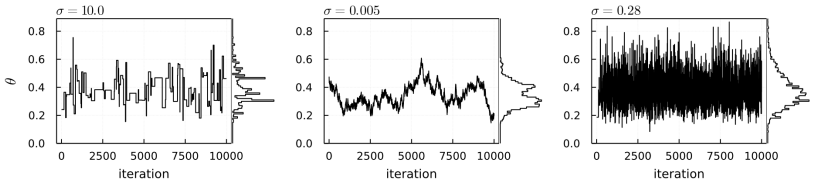


Figure A.1: Trace plots and posterior histograms for the random walk MH algorithm with different standard deviations σ of the proposal kernel, applied to the problem of estimating the distance θ under the Jukes-Cantor model for a pairwise alignment of $N = 50$ sites with $k = 15$ observed differences (see chapter 1). The rightmost plot is the result of automated tuning of the proposal standard deviation using the adaptive MCMC method of Roberts and Rosenthal (2009).

```
    # accept/reject
    return ifelse(p_ - p > log(rand()), (θ_, p_), (θ, p))
end
```

Given a suitable initial state $(\theta, \pi(\theta))$, iterating this simple function will simulate a Markov chain with as stationary distribution the target π. In fig. A.1 we show such simulated Markov chains for Gaussian proposal kernels with different settings for the standard deviation σ of the proposal kernel. Clearly, the choice of the proposal kernel has a considerable influence on the efficieny of the sampler. Choosing too large a σ will result in a very low acceptance probability in the accept/reject step of the algorithm, and hence an inefficient sampling algorithm. Choosing too small a σ on the other hand leads to a very high acceptance probability, but the resulting chain will explore the high probability region of the target very slowly.

Theoretical results show that in a one-dimensional context, a proposal kernel with an acceptance probability of about 0.44 often will lead to an optimal exploration of the space. In order to select a proposal kernel for MH updates with such an acceptance probability, one can automatically tune the scale of the kernel during the sampling algorithm, if one ensures that the resulting Markov chain still has the desired stationary distribution. Roberts and Rosenthal (2009) describe several strategies for constructing adaptive MCMC samplers using such variants of the MH algorithm. The rightmost plot in fig. A.1 shows the trace of a Markov chain simulated using the MH algorithm with automatic tuning of the standard deviation σ of the Gaussian kernel. Methods for automatic tuning of multivariate proposal kernels have also been developed.

The MH algorithm, in particular its adaptive variant, is quite efficient for sam-

pling from low-dimensional target distributions. For high-dimensional target distributions, one can use such simple MH proposal kernels compositionally by sampling from the conditional distributions. Specifically, consider for instance a target posterior density $p(\theta|y)$, where $\theta = (\theta_1, \theta_2, \dots, \theta_n) \in \mathbb{R}^n$, one can construct MH proposal kernels for each conditional posterior distribution, and sample, for $i = 1, \dots, n$, θ_i using the MH kernel according to

$$\theta_i|\theta_{-i} \sim p(\theta_i|\theta_{-i}, y)$$

where $\theta_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n)$. This is sometimes called a Metropolis-within-Gibbs (MWG) algorithm. Formally, however, it is a special case of the MH algorithm, but one with a rather complicated Markov kernel (Geyer 1998).

A.4.2 Hamiltonian Monte Carlo

When the target distribution is differentiable, samplers which make use of the geometry of the target distribution can be used. Such samplers use information from the gradient of the target to enable extremely efficient exploration of the regions of high density (the 'typical set' in information theoretic jargon). The Hamiltonian Monte Carlo (HMC) algorithm is the prime example of this class of methods. We will very briefly discuss the principle behind HMC here, and refer the reader to MacKay (2003), Neal (2011) and the pedagogical exposition in Betancourt (2017) for more information.

HMC, also originally a simulation algorithm for particle systems in physics, can be used to sample from target densities $\pi(\theta)$ where $\theta \in \mathbb{R}^n$. The algorithm is essentially a MH algorithm which uses a particularly efficient proposal kernel that reduces the random walk behavior of standard MH algorithms. To construct the HMC proposal kernel, we associate with each θ_i an auxiliary *momentum parameter* ϕ_i, so that we transform our n dimensional parameter space in a $2n$ dimensional *phase space*. We then choose a joint probability distribution on the latter which factors as

$$\pi(\theta, \phi) = \pi(\phi|\theta)\pi(\theta)$$

so that $\int \pi(\theta, \phi) d\phi = \pi(\theta)$ is the target from which we wish to sample. Let us now write

$$\pi(\theta, \phi) \propto \exp(-H(\theta, \phi))$$

with

$$H(\theta, \phi) = -\log \pi(\phi|\theta) - \log \pi(\theta)$$

which we can rewrite suggestively as

$$H(\theta, \phi) = K(\phi, \theta) + V(\theta)$$

which a physicist will recognize as a Hamiltonian for a system with potential energy $V(\theta)$ and kinetic energy $K(\phi, \theta)$. Having identified the correspondence of our joint density on phase space with a classical physical system, we can use theory form classical mechanics to help us devise an efficient proposal kernel. In particular, we can make use of conservative Hamiltonian dynamics to simulate trajectories through phase space which conserve the total energy. That is, we can simulate trajectories according to Hamilton's equations

$$\frac{d\phi}{dt} = \frac{\partial K(\phi, \theta)}{\partial \phi}$$
$$\frac{d\theta}{dt} = -\frac{\partial K(\phi, \theta)}{\partial \theta} - \frac{\partial V(\theta)}{\partial \theta}$$

Where we see the gradient of the target distribution appearing in the potential energy term. The basic idea is then the following: if we manage to start from a point sampled from the joint density over momentum and parameter variables, we can simulate the conservative Hamiltonian dynamics for some time, ending up in a point with the same density under the joint distribution on phase space. Because of the way the joint density factors, we can then marginalize out the momentum variable to obtain a sample from our target of interest.

The MCMC algorithm will then iterate the following three steps:

1. sample $\phi|\theta$
2. simulate Hamiltonian dynamics to arrive at (ϕ', θ')
3. accept/reject (ϕ', θ') using the MH acceptance probability

The technical details of HMC involve the choice of the kinetic energy term $\pi(\phi|\theta)$ and the specifics of how we simulate the Hamiltonian trajectories. We will not dwell on these here and point the reader to the already suggested references. Importantly, HMC does not rely (exclusively) on a random walk to sample from the target density, and instead can make, if suitably parameterized, big steps in the target space, reducing the autocorrelation in the simulated Markov chain considerably. An important limitation of HMC methods is that they are only defined for target densities with as domain $\mathbb{R}^n$, so that we cannot use HMC to sample from discrete probability spaces, such as, for instance,

the space of phylogenetic trees on a given leaf set. Note that we did not have such a restriction on the domain of the target density in the MH algorithm discussed above, which can equally well be applied to discrete spaces. Another restriction is of course that we need to be able to compute the gradient of the target density. Powerful automatic differentiation methods make differentiability however less of an issue than it used to be.

A.5 Approximate Bayesian computation

Approximate Bayesian computation (ABC) is an umbrella term for a number of methods for approximating posterior distributions and expectations without evaluating the likelihood function $p(y|\theta)$. For this reason, these methods are referred to as *likelihood-free* methods for Bayesian inference. Historically, the development of these approaches was largely motivated by the complicated problems which arise in evolutionary genetics, where evaluating the sampling distribution is often computationally intractable (Tavaré 2018), but simulation from the sampling distribution is very efficient.

The basic, and very intuitive, idea behind likelihood-free methods for Bayesian inference is that, to infer the likely value for a parameter θ given some data set y, we can (1) simulate parameter values $\theta_1, \dots, \theta_m$ from some density g (for instance the prior), (2) simulate *pseudo-data* $\tilde{y}$ from $p(y|\theta_i)$, and (3) compare the pseudo-data to the actually observed data. Those θ_i for which the pseudo-data is sufficiently similar to the actually observed data should then provide an idea of the posterior $p(\theta|y)$.

It is this simple idea for the *approximation* of the posterior which is formalized in ABC methods. Consider first the above scheme, but where we sample θ_i from the prior, and accept the simulated pseudo-data $\tilde{y} \in \mathcal{Y}$ if it matches the observed data *exactly*. Clearly, this approach simulates samples from $p(\theta|y)$ exactly, using the following characterization of the latter

$$p(\theta|y) \propto p(y|\theta)p(\theta) = \int \delta_y(\tilde{y})p(\tilde{y}|\theta)p(\theta)d\tilde{y}$$

Of course, this is quite useless if the space of possible data $\mathcal{Y}$ is continuous or some large finite set, as we will never simulate the observed y exactly, or do so with very small probability. In general, an ABC approach addresses this

issue by sampling instead from an approximation of the posterior

$$p(\theta|y) \propto p(y|\theta)p(\theta) \approx \int \xi(y,\tilde{y})p(\tilde{y}|\theta)p(\theta)d\tilde{y}$$

Where $0 \leq \xi(y,\tilde{y}) \leq 1$, and $\xi(y,\tilde{y}) = 1$ whenever $y = \tilde{y}$. Here the *kernel function* ξ introduces a certain tolerance window: instead of accepting only simulations which match the observed data, we shall accept simulations which are sufficiently similar. Concretely, the basic ABC rejection algorithm, using the prior as sampling density, reduces to the following scheme: for $i = 1, \ldots, m$

1. simulate $\theta_i \sim p(\theta)$
2. simulate $\tilde{y}_i \sim p(\tilde{y}|\theta)$
3. accept θ_i with probability $\xi(y,\tilde{y}_i)$, if not accepted return to (1)

The resulting sample $(\theta_1, \ldots \theta_m)$ will be sampled *approximately* from $p(\theta|y)$, with the quality of the approximation determined by ξ. Note that if ξ is interpreted as a density for a measurement error model, the ABC approach can be viewed as an exact approach under the joint model which combines the prior, sampling distribution and error distribution (Wilkinson 2013). Lastly, note that we can use other sampling densities g for θ instead of the prior in step (1), in which case we need to correct the acceptance probability by an importance weight factor $p(\theta_i)/g(\theta_i)$.

For $y \in \mathbb{R}^n$, one typically uses some *smoothing kernel function* $K_h(\|y - \tilde{y}\|)$, i.e. a symmetric probability density function with mean 0 and finite variance. A common example in ABC is the *uniform kernel*, where $K_h(\|y - \tilde{y}\|) = \mathbb{1}\left[\|y - \tilde{y}\| \leq h\right]$. When using such a kernel function, it is clear that as $h \to 0$, $K_h(\|y - \tilde{y}\|) \to \delta_{\tilde{y}}(y)$, so that the ABC approximation to the posterior becomes more and more accurate, at the expense of requiring more simulations. The kernel is hence a first source of approximation in ABC. Another source of approximation stems from the use of (insufficient) summary statistics as substitutes for the data y. Indeed, because the actual data y is typically of rather high-dimension, devising an efficient kernel function can be very hard, and as a result, one often relies on a low-dimensional representation of the data, replacing y by a set of suitably chosen summary statistics $s(y)$. This introduces a second level of approximation in ABC applications, in that the approximation quality is dependent on the degree to which the chosen statistics are sufficient for the model. Our applications in the present work are however rather atypical ABC approaches in that we do not rely on summary statistics nor on smoothing kernel function on a real-valued domain. We will hence not discuss the various complications that arise when choosing suitable kernels and summary

statistics, referring the reader to the recent volume edited by Sisson, Fan, and Beaumont (2018) for a detailed treatment of many of these issues.

The above sketches the ABC posterior approximation, and a simple rejection-based algorithm to sample from it. Of course, having identified the ABC posterior as just another target density, we can use the whole ensemble of Monte Carlo sampling techniques to sample from the resulting approximate posterior (under the constraint of course that any evaluation of the intractable likelihood is to be avoided). For instance, in the algorithm sketched above, where we sampled θ_i from the prior, we can simply use $\xi(y, \tilde{y}_i)$ as an importance weight for θ_i, instead of doing rejection sampling. More sophisticated samplers, using MCMC and particle-based methods such as SIS and SMC have been used succesfully in the context of ABC (see also chapter 5). For a comprehensive overview of the many sampling approaches for ABC problems, we refer again to Sisson, Fan, and Beaumont (2018) (chapter 4).

B Data sets and software

B.1 Data sets

Throughout this dissertation, we uses various data sets in our examples. Here we specify the specific taxa, abbreviations and sources for the various data sets (unless this was already done in the main text).

B.1.1 Rice

id	taxon	source	accession
oba	*Oryza barthii*	Stein et al. (2018)	oba
osj	*Oryza sativa* vg. *japonica*	.	osj
oru	*Oryza rufipogon*	.	oru
ogl	*Oryza glaberrima*	.	ogl
oni	*Oryza nivara*	.	oni
osi	*Oryza sativa* vg. *indica*	.	osi

B.1.2 *Drosophila*

id	taxon	source	accession
dere	*Drosophila erecta*	NCBI	GCF_003286155.1_DereRS2
dana	*Drosophila ananassae*	.	GCF_003285975.2_DanaRS2.1
dsim	*Drosophila simulans*	.	GCF_000754195.2_ASM75419v2
dper	*Drosophila persimilis*	.	GCF_003286085.1_DperRS2
dpse	*Drosophila pseudoobscura*	.	GCF_009870125.1_UCI_Dpse_MV25
dmel	*Drosophila melanogaster*	.	GCF_000001215.4_Release_6_plus_ISO
dyak	*Drosophila yakuba*	.	GCF_000005975.2_dyak_caf1
dsec	*Drosophila sechellia*	.	GCF_004382195.1_ASM438219v1

B.1.3 Yeasts

id	taxon	source	accession
ecy	*Eremothecium cymbalariae*	YGOB, v7-Aug2012	Ecymbalariae
ego	*Eremothecium gossypii*	.	Egossypii
kla	*Kluyveromyces lactis*	.	Klactis
lkl	*Lachancea kluyveri*	.	Lkluyveri
lth	*Lachancea thermotolerans*	.	Lthermotolerans
lwa	*Lachancea waltii*	.	Lwaltii
tde	*Torulaspora delbrueckii*	.	Tdelbrueckii
zro	*Zygosaccharomyces rouxii*	.	Zrouxii
sce	*Saccharomyces cerivisae*	.	Scerevisiae
suv	*Saccharomyces uvarum*	.	Suvarum
tph	*Tetrapisispora blattae*	.	Tblattae
vpo	*Vanderwaltozyma polyspora*	.	Vpolyspora
nda	*Naumovozyma dairenensis*	.	Ndairenensis
nca	*Naumovozyma castellii*	.	Ncastellii
cgl	*Candida glabratra*	.	Cglabrata

B.1.4 Primates

id	taxon	source	accession
Hsapi	*Homo sapiens*	Ensembl (v102)	Homo_sapiens.GRCh38
Ptrog	*Pan troglodytes*	.	Pan_troglodytes.Pan_tro_3.0
Pabel	*Pongo abelii*	.	Pongo_abelii.PPYG2
Nleuc	*Nomascus leucogenys*	.	Nomascus_leucogenys.Nleu_3.0
Mmula	*Macaca mulatta*	.	Macaca_mulatta.Mmul_10
Panub	*Papio anubis*	.	Papio_anubis.Panu_3.0
Csaba	*Chlorocebus sabaeus*	.	Chlorocebus_sabaeus.ChlSab1.1
Cjacc	*Callithrix jacchus*	.	Callithrix_jacchus.ASM275486v1
Csyri	*Carlito syrichta*	.	Carlito_syrichta.Tarsius_syrichta
Mmuri	*Microcebus murinus*	.	Microcebus_murinus.Mmur_3.0
Pcoqu	*Propithecus coquereli*	.	Propithecus_coquereli.Pcoq_1.0
Ogarn	*Otolemur garnettii*	.	Otolemur_garnettii.OtoGar3

B.1.5 Land plants

Note that taxa in the table below appear in multiple data sets (referred to as the dicots and land plants data sets in the main text). *Carica papaya* and *Cycas panzhihuaensis* do not occur together in a single analysis, hence no confusion is possible.

id	taxon	source	accession
mpo	*Marchantia polymorpha*	PLAZA 5.0	mpo
ppa	*Physcomitrium patens*	.	ppa
atr	*Amborella trichopoda*	.	atr
sgi	*Sequoiadendron giganteum*	.	sgi
vvi	*Vitis vinifera*	.	vvi
bvu	*Beta vulgaris*	.	bvu
cqu	*Chenopodium quinoa*	.	cqu
mtr	*Medicago truncatula*	.	mtr
ptr	*Populus trichocarpa*	.	ptr
sly	*Solanum lycopersicum*	.	sly
ath	*Arabidopsis thaliana*	.	ath
cpa (1)	*Carica papaya*	.	cpa
afi	*Azolla filliculoides*	F.-W. Li et al. (2018)	afi
scu	*Salvinia cucullata*	.	scu
gbi	*Ginkgo biloba*	Liu et al. (2021)	gbi
cpa (2)	*Cycas panzhihuaensis*	Y. Liu et al. (2022)	cpa

B.1.6 *Drosera*

id	taxon	source	accession
dca	*Drosera capensis*	unpublished (Renner et al. in prep.)	dca
dre	*Drosera regia*	.	dre
dsp	*Drosera spatulata*	.	dsp
dmu	*Dionaeae muscipula*	.	dmu
ave	*Aldrovanda vesiculosa*	.	ave
vvi	*Vitis vinifera*	PLAZA 5.0	vvi
bvu	*Beta vulgaris*	.	bvu
cqu	*Chenopodium quinoa*	.	cqu

B.2 Software

Much research effort is hidden away in codebases. The main methodological developments presented in this dissertation are implemented in free and open source software libraries, implemented in the Julia programming language (Bezanson et al. 2017). They are available at the following URLs (last accessed, May 2022):

1. DeadBird (https://github.com/arzwa/DeadBird.jl): Package for Bayesian and ML statistical inference for phylogenetic BDP models of gene content evolution.

2. Beluga (https://github.com/arzwa/Beluga.jl): Package for inference of ancient WGDs using reversible-jump MCMC and phylogenetic BDP models of gene content evolution.
3. TwoTypeDLModel (https://github.com/arzwa/TowTypeDLModel): Implements algorithms to compute the phylogenetic likelihood for the two-type branching process model developed in chapter 3, as well as a simple MCMC sampler to conduct Bayesian inference.
4. SmoothTree (https://github.com/arzwa/SmoothTree.jl): Package for working with CCDs and conducting species tree inference under the MSC using likelihood-free expectation propagation.
5. Whale (https://github.com/arzwa/Whale.jl): Package for Bayesian and ML inference of reconciled gene trees and ancient WGDs using amalgamated likelihood estimation for phylogenetic BDP models of gene family evolution.

Additional software developed by myself and used in this thesis include

1. NewickTree (https://github.com/arzwa/NewickTree.jl): A Julia package for working with phylogenetic trees, centered around the Newick representation.
2. MacroSynteny (https://github.com/arzwa/MacroSynteny): A julia package for conducting Bayesian analyses of macrosynteny (in development, used in the *Drosera* analyses in chapter 5).
3. MaereModel (https://github.com/arzwa/MaereModel): Some julia code for conducting approximate Bayesian inference for the model of Maere et al. (2005).
4. AdaptiveMCMC (https://github.com/arzwa/AdaptiveMCMC.jl): Implements various adaptive proposal kernels to use as building blocks in MH-like MCMC algorithms.
5. wgd (https://github.com/arzwa/wgd): A Python pipeline for constructing K_S distributions and conducting colinearity analyses using genomic sequence data.

In addition, we have relied on a lot of free and open source software in our research, developed by many devoted programmers across the world. Relevant bioinformatics and phylogenetics software used for evolutionary genomics research has been credited throughout the body of this work, but here we would like to thank the developers of pandoc, vim, the Julia programming language, the many Julia packages we have relied on (in particular, Distributions.jl, Turing.jl, Optim.jl, DifferentialEquations.jl, DataFrames.jl and Plots.jl), LaTeX, tikz, git, and many more, for making our lives easier.

Bibliography

Albert, Victor A, W Bradley Barbazuk, Claude W dePamphilis, Joshua P Der, James Leebens-Mack, Hong Ma, Jeffrey D Palmer, Steve Rounsley, David Sankoff, and others. 2013. "The Amborella Genome and the Evolution of Flowering Plants." *Science* 342 (6165): 1241089.

Aldous, David. 1996. "Probability Distributions on Cladograms." In *Random Discrete Structures*, 1–18. Springer.

Allen, Linda JS. 2010. *An Introduction to Stochastic Processes with Applications to Biology*. CRC Press.

Antonovics, Janis. 1987. "The Evolutionary Dys-Synthesis: Which Bottles for Which Wine?" *The American Naturalist* 129 (3): 321–31.

Arnold, Brian, Kirsten Bomblies, and John Wakeley. 2012. "Extending Coalescent Theory to Autotetraploids." *Genetics* 192 (1): 195–204.

Arvestad, Lars, Jens Lagergren, and Bengt Sennblad. 2009. "The Gene Evolution Model and Computing Its Associated Probabilities." *Journal of the ACM (JACM)* 56 (2): 1–44.

Athreya, Krishna B., and Peter E. Ney. 1972. *Branching Processes*. Grundlehren Der Mathematischen Wissenschaften. Berlin Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-65371-1.

Åkerborg, Örjan, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. 2009. "Simultaneous Bayesian Gene Tree Reconstruction and Reconciliation Analysis." *Proceedings of the National Academy of Sciences* 106 (14): 5714–9.

Baele, Guy. 2012. "Context-Dependent Evolutionary Models for Non-Coding Sequences: An Overview of Several Decades of Research and an Analysis of Laurasiatheria and Primate Evolution." *Evolutionary Biology* 39 (1): 61–82.

Bafna, Vineet, and Pavel A Pevzner. 1996. "Genome Rearrangements and Sorting by Reversals." *SIAM Journal on Computing* 25 (2): 272–89.

Bailey, Norman T. J. 1990. *The Elements of Stochastic Processes with Applications to the Natural Sciences*. John Wiley & Sons.

Barthelmé, Simon, and Nicolas Chopin. 2014. "Expectation Propagation for Likelihood-Free Inference." *Journal of the American Statistical Association* 109 (505): 315–33.

Beerli, Peter, and Joseph Felsenstein. 2001. "Maximum Likelihood Estimation of a Migration Matrix and Effective Population Sizes in N Subpopulations by Using a Coalescent Approach." *Proceedings of the National Academy of Sciences* 98 (8): 4563–8.

Beimforde, Christina, Kathrin Feldberg, Stephan Nylinder, Jouko Rikkinen, Hanna Tuovila, Heinrich Dörfelt, Matthias Gube, et al. 2014. "Estimating the Phanerozoic History of the Ascomycota Lineages: Combining Fossil and Molecular Data." *Molecular Phylogenetics and*

Evolution 78: 386–98.

Bernardo, José M, and Adrian FM Smith. 2009. *Bayesian Theory*. Vol. 405. John Wiley & Sons.

Betancourt, Michael. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv Preprint arXiv:1701.02434*.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B Shah. 2017. "Julia: A Fresh Approach to Numerical Computing." *SIAM Review* 59 (1): 65–98.

Birchler, James A, and Reiner A Veitia. 2010. "The Gene Balance Hypothesis: Implications for Gene Regulation, Quantitative Traits and Evolution." *New Phytologist* 186 (1): 54–62.

———. 2012. "Gene Balance Hypothesis: Connecting Issues of Dosage Sensitivity Across Biological Disciplines." *Proceedings of the National Academy of Sciences* 109 (37): 14746–53.

Blanc, Guillaume, and Kenneth H Wolfe. 2004a. "Functional Divergence of Duplicated Genes Formed by Polyploidy During Arabidopsis Evolution." *The Plant Cell* 16 (7): 1679–91.

———. 2004b. "Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes." *The Plant Cell* 16 (7): 1667–78.

Blomme, Tine, Klaas Vandepoele, Stefanie De Bodt, Cedric Simillion, Steven Maere, and Yves Van de Peer. 2006. "The Gain and Loss of Genes During 600 Million Years of Vertebrate Evolution." *Genome Biology* 7 (5): 1–12.

Blum, Michael GB, and Olivier François. 2006. "Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance." *Systematic Biology* 55 (4): 685–91.

Borges, Rui, Bastien Boussau, Sebastian Höhna, Ricardo J Pereira, and Carolin Kosiol. 2021. "Polymorphism-Aware Estimation of Species Trees and Evolutionary Forces from Genomic Sequences with Revbayes." *bioRxiv*.

Borges, Rui, Gergely J Szöllősi, and Carolin Kosiol. 2019. "Quantifying Gc-Biased Gene Conversion in Great Ape Genomes Using Polymorphism-Aware Models." *Genetics* 212 (4): 1321–36.

Boussau, Bastien, Gergely J Szöllősi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. 2013. "Genome-Scale Coestimation of Species and Gene Trees." *Genome Research* 23 (2): 323–30.

Brittnacher, John. 2010. "Drosera Hybrids." *Drosera Hybrids | ICPS*. https://www.carnivorousplants.org/cp/evolution/DroseraHybrids.

Brown, Jeremy M, and Robert C Thomson. 2017. "Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses." *Systematic Biology* 66 (4): 517–30.

Bryant, David, Remco Bouckaert, Joseph Felsenstein, Noah A Rosenberg, and Arindam RoyChoudhury. 2012. "Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis." *Molecular Biology and Evolution* 29 (8): 1917–32.

Bryant, David, and Matthew W Hahn. 2020. "The Concatenation Question." No commercial publisher| Authors open access book.

Butler, Geraldine, Matthew D Rasmussen, Michael F Lin, Manuel AS Santos, Sharadha Sakthikumar, Carol A Munro, Esther Rheinbay, et al. 2009. "Evolution of Pathogenicity and Sexual Reproduction in Eight Candida Genomes." *Nature* 459 (7247): 657–62.

Byrne, Kevin P, and Kenneth H Wolfe. 2005. "The Yeast Gene Order Browser: Combining Curated Homology and Syntenic Context Reveals Gene Fate in Polyploid Species." *Genome*

Research 15 (10): 1456–61.

Canguilhem, Georges. 1977. “Qu’est-Ce Qu’une Idéologie Scientifique?” *Idéologie et Rationalité Dans L’histoire Des Sciences de La Vie* 33: 45.

Capella-Gutiérrez, Salvador, José M Silla-Martínez, and Toni Gabaldón. 2009. “TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–3.

Carretero-Paulet, Lorenzo, Pablo Librado, Tien-Hao Chang, Enrique Ibarra-Laclette, Luis Herrera-Estrella, Julio Rozas, and Victor A Albert. 2015. “High Gene Family Turnover Rates and Gene Space Adaptation in the Compact Genome of the Carnivorous Plant Utricularia Gibba.” *Molecular Biology and Evolution* 32 (5): 1284–95.

Chen, Hengchi, Yuhan Fang, Arthur Zwaenepoel, Sanwen Huang, Yves Van de Peer, and Zhen Li. 2022. “Revisiting Ancient Polyploidy in Leptosporangiate Ferns.” *bioRxiv*.

Chen, Kevin, Dannie Durand, and Martin Farach-Colton. 2000. “NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees.” *Journal of Computational Biology* 7 (3-4): 429–47.

Chiari, Ylenia, Vincent Cahais, Nicolas Galtier, and Frédéric Delsuc. 2012. “Phylogenomic Analyses Support the Position of Turtles as the Sister Group of Birds and Crocodiles (Archosauria).” *Bmc Biology* 10 (1): 1–15.

Chifman, Julia, and Laura Kubatko. 2014. “Quartet Inference from Snp Data Under the Coalescent Model.” *Bioinformatics* 30 (23): 3317–24.

Clark, James W, and Philip CJ Donoghue. 2018. “Whole-Genome Duplication and Plant Macroevolution.” *Trends in Plant Science* 23 (10): 933–45.

Clark, James W, Mark N Puttick, and Philip CJ Donoghue. 2019. “Origin of Horsetails and the Role of Whole-Genome Duplication in Plant Macroevolution.” *Proceedings of the Royal Society B* 286 (1914): 20191662.

Cox, Richard T. 1961. *The Algebra of Probable Inference, 1961*. Johns Hopkins University Press.

Crawford, Forrest W, Lam Si Tung Ho, and Marc A Suchard. 2018. “Computational Methods for Birth-Death Processes.” *Wiley Interdisciplinary Reviews: Computational Statistics* 10 (2): e1423.

Crawford, Forrest W, Vladimir N Minin, and Marc A Suchard. 2014. “Estimation for General Birth-Death Processes.” *Journal of the American Statistical Association* 109 (506): 730–47.

Crawford, Forrest W, and Marc A Suchard. 2012. “Transition Probabilities for General Birth–Death Processes with Applications in Ecology, Genetics, and Evolution.” *Journal of Mathematical Biology* 65 (3): 553–80.

Crick, Francis HC. 1958. “On Protein Synthesis.” In *Symp Soc Exp Biol*, 12:8. 138-63.

Csűrös, Miklós. 2010. “Count: Evolutionary Analysis of Phylogenetic Profiles with Parsimony and Likelihood.” *Bioinformatics* 26 (15): 1910–2.

———. 2022. “Gain-Loss-Duplication Models for Copy Number Evolution on a Phylogeny: Exact Algorithms for Computing the Likelihood and Its Gradient.” *Theoretical Population Biology*.

Csűrös, Miklós, and István Miklós. 2009. “Streamlining and Large Ancestral Genomes in Archaea Inferred with a Phylogenetic Birth-and-Death Model.” *Molecular Biology and Evolution* 26 (9): 2087–95.

De Bie, Tijl, Nello Cristianini, Jeffery P Demuth, and Matthew W Hahn. 2006a. “CAFE: A Computational Tool for the Study of Gene Family Evolution.” *Bioinformatics* 22 (10): 1269–71.

———. 2006b. “CAFE: A Computational Tool for the Study of Gene Family Evolution.” *Bioin-*

formatics 22 (10): 1269–71.

de Finetti, Bruno. 1974. *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons.

Degnan, James H, and Laura A Salter. 2005. "Gene Tree Distributions Under the Coalescent Process." *Evolution* 59 (1): 24–37.

De Maio, Nicola, Dominik Schrempf, and Carolin Kosiol. 2015. "PoMo: An Allele Frequency-Based Approach for Species Tree Estimation." *Systematic Biology* 64 (6): 1018–31.

De Smet, Riet, Keith L Adams, Klaas Vandepoele, Marc CE Van Montagu, Steven Maere, and Yves Van de Peer. 2013. "Convergent Gene Loss Following Gene and Genome Duplications Creates Single-Copy Families in Flowering Plants." *Proceedings of the National Academy of Sciences* 110 (8): 2898–2903.

Devroye, Luc. 2006. "Nonuniform Random Variate Generation." *Handbooks in Operations Research and Management Science* 13: 83–121.

Dickey, James M. 1971. "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters." *The Annals of Mathematical Statistics*, 204–23.

Dobzhansky, Theodosius. 1973. "Nothing in Biology Makes Sense Except in the Light of Evolution." *The American Biology Teacher* 35 (3): 125–29.

Doucet, Arnaud, Nando de Freitas, and Neil Gordon. 2001. "An Introduction to Sequential Monte Carlo Methods." In *Sequential Monte Carlo Methods in Practice*, 3–14. Springer.

Doucet, Arnaud, Adam M Johansen, and others. 2009. "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later." *Handbook of Nonlinear Filtering* 12 (656-704): 3.

Drosophila 12 Genomes Consortium. 2007. "Evolution of Genes and Genomes on the Drosophila Phylogeny." *Nature* 450 (7167): 203.

Drovandi, Christopher C, Clara Grazian, Kerrie Mengersen, and Christian Robert. 2018. "Approximating the Likelihood in Abc." *Handbook of Approximate Bayesian Computation*, 321–68.

Drummond, Alexei J, Marc A Suchard, Dong Xie, and Andrew Rambaut. 2012. "Bayesian Phylogenetics with Beauti and the Beast 1.7." *Molecular Biology and Evolution* 29 (8): 1969–73.

Duchemin, Wandrille, Yoann Anselmetti, Murray Patterson, Yann Ponty, Sèverine Bérard, Cedric Chauve, Celine Scornavacca, Vincent Daubin, and Eric Tannier. 2017. "DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies." *Genome Biology and Evolution* 9 (5): 1312–9.

Durrett, Richard, Rasmus Nielsen, and Thomas L York. 2004. "Bayesian Estimation of Genomic Distance." *Genetics* 166 (1): 621–29.

Edwards, Scott V, Liang Liu, and Dennis K Pearl. 2007. "High-Resolution Species Trees Without Concatenation." *Proceedings of the National Academy of Sciences* 104 (14): 5936–41.

Efron, Bradley, and Trevor Hastie. 2021. *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*. Vol. 6. Cambridge University Press.

Emms, David M, and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 1–14.

Etheridge, Alison. 2011. *Some Mathematical Models from Population Genetics: École d'Été de Probabilités de Saint-Flour Xxxix-2009*. Vol. 2012. Springer Science & Business Media.

Excoffier, Laurent, Nina Marchi, David Alexander Marques, Remi Matthey-Doret, Alexandre Gouy, and Vitor C Sousa. 2021. "Fastsimcoal2: Demographic Inference Under Complex Evolutionary Scenarios." *Bioinformatics* 37 (24): 4882–5.

Fan, Helen Hang, and Laura S Kubatko. 2011. “Estimating Species Trees Using Approximate Bayesian Computation.” *Molecular Phylogenetics and Evolution* 59 (2): 354–63.

Fawcett, Jeffrey A, Steven Maere, and Yves Van De Peer. 2009. “Plants with Double Genomes Might Have Had a Better Chance to Survive the Cretaceous–Tertiary Extinction Event.” *Proceedings of the National Academy of Sciences* 106 (14): 5737–42.

Felsenstein, Joseph. 2001. “The Troubled Growth of Statistical Phylogenetics.” *Systematic Biology*, 465–67.

———. 2004. *Inferring Phylogenies*. Vol. 2. Sinauer associates Sunderland, MA.

Fertin, Guillaume, Anthony Labarre, Irena Rusu, Stéphane Vialette, and Eric Tannier. 2009. *Combinatorics of Genome Rearrangements*. MIT press.

Fisher, RA. 1930. “The Genetical Theory of Natural Selection.”

Fitch, Walter M. 1970. “Distinguishing Homologous from Analogous Proteins.” *Systematic Zoology* 19 (2): 99–113.

Flouri, Tomáš, Xiyun Jiao, Bruce Rannala, and Ziheng Yang. 2018. “Species Tree Inference with Bpp Using Genomic Sequences and the Multispecies Coalescent.” *Molecular Biology and Evolution* 35 (10): 2585–93.

———. 2020. “A Bayesian Implementation of the Multispecies Coalescent Model with Introgression for Phylogenomic Analysis.” *Molecular Biology and Evolution* 37 (4): 1211–23.

Fok, Pak-Wing, and Tom Chou. 2013. “Identifiability of Age-Dependent Branching Processes from Extinction Probabilities and Number Distributions.” *Journal of Statistical Physics* 152 (4): 769–86.

Force, Allan, Michael Lynch, F Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. 1999. “Preservation of Duplicate Genes by Complementary, Degenerative Mutations.” *Genetics* 151 (4): 1531–45.

Fostier, Jan, Sebastian Proost, Bart Dhoedt, Yvan Saeys, Piet Demeester, Yves Van de Peer, and Klaas Vandepoele. 2011. “A Greedy, Graph-Based Algorithm for the Alignment of Multiple Homologous Gene Lists.” *Bioinformatics* 27 (6): 749–56.

Freeling, Michael. 2009. “Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition.” *Annual Review of Plant Biology* 60: 433–53.

Freeling, Michael, and Brian C Thomas. 2006. “Gene-Balanced Duplications, Like Tetraploidy, Provide Predictable Drive to Increase Morphological Complexity.” *Genome Research* 16 (7): 805–14.

Freyman, William A, Matthew G Johnson, and Carl J Rothfels. 2020. “Homologizer: Phylogenetic Phasing of Gene Copies into Polyploid Subgenomes.” *bioRxiv*.

Ge, Hong, Kai Xu, and Zoubin Ghahramani. 2018. “Turing: A Language for Flexible Probabilistic Inference.”

Gelfand, Alan E, and Adrian FM Smith. 1990. “Sampling-Based Approaches to Calculating Marginal Densities.” *Journal of the American Statistical Association* 85 (410): 398–409.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. CRC Press.

Gelman, Andrew, and Christian P Robert. 2013. “ ‘Not Only Defended but Also Applied’: The Perceived Absurdity of Bayesian Inference.” *The American Statistician* 67 (1): 1–5.

Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. “Philosophy and the Practice of Bayesian Statistics.” *British Journal of Mathematical and Statistical Psychology* 66 (1): 8–38.

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020.

"Bayesian Workflow." *arXiv Preprint arXiv:2011.01808*.

Geyer, Charles J. 1998. "Markov Chain Monte Carlo Lecture Notes." *Course Notes, Spring Quarter* 80.

———. 2011. "Introduction to Markov Chain Monte Carlo." *Handbook of Markov Chain Monte Carlo* 20116022: 45.

Goldman, Nick, and Ziheng Yang. 1994. "A Codon-Based Model of Nucleotide Substitution for Protein-Coding Dna Sequences." *Molecular Biology and Evolution* 11 (5): 725–36.

Goodman, Morris, John Czelusniak, G William Moore, Alejo E Romero-Herrera, and Genji Matsuda. 1979. "Fitting the Gene Lineage into Its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences." *Systematic Biology* 28 (2): 132–63.

Gould, Stephen Jay. 1978. "Sociobiology: The Art of Storytelling." *New Scientist* 80 (1129): 530–33.

———. 2002. *The Structure of Evolutionary Theory*. Harvard University Press.

Gould, Stephen Jay, and Richard C Lewontin. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme." *Proceedings of the Royal Society of London. Series B. Biological Sciences* 205 (1161): 581–98.

Graham, Ronald L, Donald E Knuth, Oren Patashnik, and Stanley Liu. 1989. "Concrete Mathematics: A Foundation for Computer Science." *Computers in Physics* 3 (5): 106–7.

Green, Peter J. 1995. "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika* 82 (4): 711–32.

Greenman, Chris D, and Tom Chou. 2016. "Kinetic Theory of Age-Structured Stochastic Birth-Death Processes." *Physical Review E* 93 (1): 012112.

Gutenkunst, Ryan N, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. 2009. "Inferring the Joint Demographic History of Multiple Populations from Multidimensional Snp Frequency Data." *PLoS Genetics* 5 (10): e1000695.

Hahn, Matthew W. 2007. "Bias in Phylogenetic Tree Reconciliation Methods: Implications for Vertebrate Genome Evolution." *Genome Biology* 8 (7): 1–9.

Hahn, Matthew W, Tijl De Bie, Jason E Stajich, Chi Nguyen, and Nello Cristianini. 2005. "Estimating the Tempo and Mode of Gene Family Evolution from Comparative Genomic Data." *Genome Research* 15 (8): 1153–60.

Hahn, Matthew W, Mira V Han, and Sang-Gook Han. 2007. "Gene Family Evolution Across 12 Drosophila Genomes." *PLoS Genet* 3 (11): e197.

Hahn, Matthew W, and Gregory A Wray. 2002. "The G-Value Paradox." *Evolution and Development* 4 (2): 73–75.

Haldane, John Burdon Sanderson. 1964. "A Defense of Beanbag Genetics." *Perspectives in Biology and Medicine* 7 (3): 343–60.

Han, Mira V, Jeffery P Demuth, Casey L McGrath, Claudio Casola, and Matthew W Hahn. 2009. "Adaptive Evolution of Young Gene Duplicates in Mammals." *Genome Research* 19 (5): 859–67.

Hannenhalli, Sridhar, and Pavel A Pevzner. 1995. "Towards a Computational Theory of Genome Rearrangements." *Computer Science Today*, 184–202.

Hastings, W Keith. 1970. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications."

Heger, Andreas, and Chris P Ponting. 2007. "Evolutionary Rate Analyses of Orthologs and Paralogs from 12 Drosophila Genomes." *Genome Research* 17 (12): 1837–49.

Hein, Jotun, Mikkel Schierup, and Carsten Wiuf. 2004. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, USA.

Heled, Joseph, and Alexei J Drummond. 2009. "Bayesian Inference of Species Trees from Multilocus Data." *Molecular Biology and Evolution* 27 (3): 570–80.

Hermans, Willem Frederik. 1981. "Over Popper." *Hollands Maandblad*, 398–409.

Herschel, JFW. 1831. "A Preliminary Discourse on the Study of Natural Philosophy."

Hey, Jody. 2010. "Isolation with Migration Models for More Than Two Populations." *Molecular Biology and Evolution* 27 (4): 905–20.

Hoang, Diep Thi, Olga Chernomor, Arndt Von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2018. "UFBoot2: Improving the Ultrafast Bootstrap Approximation." *Molecular Biology and Evolution* 35 (2): 518–22.

Hobolth, Asger, Julien Y Dutheil, John Hawks, Mikkel H Schierup, and Thomas Mailund. 2011. "Incomplete Lineage Sorting Patterns Among Human, Chimpanzee, and Orangutan Suggest Recent Orangutan Speciation and Widespread Selection." *Genome Research* 21 (3): 349–56.

Hoffman, Matthew D, and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.* 15 (1): 1593–1623.

Holmes, Ian, and William J Bruno. 2001. "Evolutionary Hmms: A Bayesian Approach to Multiple Alignment." *Bioinformatics* 17 (9): 803–20.

Höhna, Sebastian, Tracy A Heath, Bastien Boussau, Michael J Landis, Fredrik Ronquist, and John P Huelsenbeck. 2014. "Probabilistic Graphical Model Representation in Phylogenetics." *Systematic Biology* 63 (5): 753–71.

Höhna, Sebastian, Michael J Landis, Tracy A Heath, Bastien Boussau, Nicolas Lartillot, Brian R Moore, John P Huelsenbeck, and Fredrik Ronquist. 2016. "RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language." *Systematic Biology* 65 (4): 726–36.

Huang, Chien-Hsun, Xinping Qi, Duoyuan Chen, Ji Qi, and Hong Ma. 2020. "Recurrent Genome Duplication Events Likely Contributed to Both the Ancient and Recent Rise of Ferns." *Journal of Integrative Plant Biology* 62 (4): 433–55.

Huang, Xiong, Wenling Wang, Ting Gong, David Wickell, Li-Yaung Kuo, Xingtan Zhang, Jialong Wen, et al. 2022. "The Flying Spider-Monkey Tree Fern Genome Provides Insights into Fern Evolution and Arborescence." *Nature Plants*, 1–12.

Hudson, Richard R. 1983. "Testing the Constant-Rate Neutral Allele Model with Protein Sequence Data." *Evolution*, 203–17.

Huerta-Cepas, Jaime, Anibal Bueno, Joaquín Dopazo, and Toni Gabaldón. 2007. "PhylomeDB: A Database for Genome-Wide Collections of Gene Phylogenies." *Nucleic Acids Research* 36 (suppl_1): D491–D496.

Hughes, Timothy, and David A Liberles. 2008. "The Power-Law Distribution of Gene Family Size Is Driven by the Pseudogenisation Rate's Heterogeneity Between Gene Families." *Gene* 414 (1-2): 85–94.

Huynen, Martijn A, and Peer Bork. 1998. "Measuring Genome Evolution." *Proceedings of the National Academy of Sciences* 95 (11): 5849–56.

Huynen, Martijn A, and Erik Van Nimwegen. 1998. "The Frequency Distribution of Gene Family Sizes in Complete Genomes." *Molecular Biology and Evolution* 15 (5): 583–89.

Ibarra-Laclette, Enrique, Eric Lyons, Gustavo Hernández-Guzmán, Claudia Anahí Pérez-Torres, Lorenzo Carretero-Paulet, Tien-Hao Chang, Tianying Lan, et al. 2013. "Architecture and Evolution of a Minute Plant Genome." *Nature* 498 (7452): 94–98.

Innan, Hideki, and Fyodor Kondrashov. 2010. "The Evolution of Gene Duplications: Classifying

and Distinguishing Between Models." *Nature Reviews Genetics* 11 (2): 97–108.

Jarvis, Erich D, Siavash Mirarab, Andre J Aberer, Bo Li, Peter Houde, Cai Li, Simon YW Ho, et al. 2014. "Whole-Genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds." *Science* 346 (6215): 1320–31.

Jaynes, Edwin T. 2003. *Probability Theory: The Logic of Science*. Cambridge university press.

Jeffreys, Harold. 1961. *Theory of Probability*. Oxford University Press.

Jiao, Yuannian, Norman J Wickett, Saravanaraj Ayyampalayam, André S Chanderbali, Lena Landherr, Paula E Ralph, Lynn P Tomsho, et al. 2011. "Ancestral Polyploidy in Seed Plants and Angiosperms." *Nature* 473 (7345): 97–100.

Jones, Graham R. 2011. "Tree Models for Macroevolution and Phylogenetic Analysis." *Systematic Biology* 60 (6): 735–46.

Jones, Graham, Serik Sagitov, and Bengt Oxelman. 2013. "Statistical Inference of Allopolyploid Species Networks in the Presence of Incomplete Lineage Sorting." *Systematic Biology* 62 (3): 467–78.

Jordan, Michael I. 2003. "An Introduction to Probabilistic Graphical Models." unpublised.

Jukes, Thomas H, and Charles R Cantor. 1969. "Evolution of Protein Molecules." *Mammalian Protein Metabolism* 3: 21–132.

Karev, Georgy P, Yuri I Wolf, and Eugene V Koonin. 2003. "Simple Stochastic Birth and Death Models of Genome Evolution: Was There Enough Time for Us to Evolve?" *Bioinformatics* 19 (15): 1889–1900.

Karev, Georgy P, Yuri I Wolf, Andrey Y Rzhetsky, Faina S Berezovskaya, and Eugene V Koonin. 2002. "Birth and Death of Protein Domains: A Simple Model of Evolution Explains Power Law Behavior." *BMC Evolutionary Biology* 2 (1): 18.

Karlin, Samuel, and James McGregor. 1957. "The Classification of Birth and Death Processes." *Transactions of the American Mathematical Society* 86 (2): 366–400.

Kass, Robert E, and Adrian E Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (430): 773–95.

Katju, Vaishali, and Michael Lynch. 2003. "The Structure and Early Evolution of Recently Arisen Gene Duplicates in the Caenorhabditis Elegans Genome." *Genetics* 165 (4): 1793–1803.

Katoh, Kazutaka, and Daron M Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80.

Kendall, David G. 1948. "On the Generalized" Birth-and-Death" Process." *The Annals of Mathematical Statistics* 19 (1): 1–15.

———. 1953. "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain." *The Annals of Mathematical Statistics*, 338–54.

Kimura, Motoo. 1980. "A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences." *Journal of Molecular Evolution* 16 (2): 111–20.

Kimura, Motoo, and Jack L King. 1979. "Fixation of a Deleterious Allele at One of Two" Duplicate" Loci by Mutation Pressure and Random Drift." *Proceedings of the National Academy of Sciences* 76 (6): 2858–61.

Kingman, John Frank Charles. 1982a. "On the Genealogy of Large Populations." *Journal of Applied Probability* 19 (A): 27–43.

———. 1982b. "The Coalescent." *Stochastic Processes and Their Applications* 13 (3): 235–48.

Kondrashov, Fyodor A. 2012. "Gene Duplication as a Mechanism of Genomic Adaptation to

a Changing Environment." *Proceedings of the Royal Society B: Biological Sciences* 279 (1749): 5048–57.

Kondrashov, Fyodor A, and Eugene V Koonin. 2004. "A Common Framework for Understanding the Origin of Genetic Dominance and Evolutionary Fates of Gene Duplications." *Trends in Genetics* 20 (7): 287–90.

Kong, Augustine. 1992. "A Note on Importance Sampling Using Standardized Weights." *University of Chicago, Dept. Of Statistics, Tech. Rep* 348.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S Blair Hedges. 2017. "TimeTree: A Resource for Timelines, Timetrees, and Divergence Times." *Molecular Biology and Evolution* 34 (7): 1812–9.

Kühnert, Denise, Tanja Stadler, Timothy G Vaughan, and Alexei J Drummond. 2014. "Simultaneous Reconstruction of Evolutionary History and Epidemiological Dynamics from Viral Sequences with the Birth–Death Sir Model." *Journal of the Royal Society Interface* 11 (94): 20131106.

Lambert, Amaury, and Tanja Stadler. 2013. "Birth–Death Models and Coalescent Point Processes: The Shape and Probability of Reconstructed Phylogenies." *Theoretical Population Biology* 90: 113–28.

Lander, Eric S, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome."

Larget, Bret. 2013. "The Estimation of Tree Posterior Probabilities Using Conditional Clade Probability Distributions." *Systematic Biology* 62 (4): 501–11.

Larget, Bret, J Kadane, and D Simon. 2002. "A Markov Chain Monte Carlo Approach to Reconstructing Ancestral Genome Rearrangements." *Mol. Biol. Evol* 22: 486–89.

Lartillot, Nicolas. 2012. "Interaction Between Selection and Biased Gene Conversion in Mammalian Protein-Coding Sequence Evolution Revealed by a Phylogenetic Covariance Analysis." *Molecular Biology and Evolution* 30 (2): 356–68.

Lartillot, Nicolas, and Raphaël Poujol. 2011. "A Phylogenetic Model for Investigating Correlated Evolution of Substitution Rates and Continuous Phenotypic Characters." *Molecular Biology and Evolution* 28 (1): 729–44.

Latouche, Guy, and Vaidyanathan Ramaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM.

Latrille, Thibault, Vincent Lanore, and Nicolas Lartillot. 2021. "Inferring Long-Term Effective Population Size with Mutation–Selection Models." *Molecular Biology and Evolution* 38 (10): 4573–87.

Leebens-Mack, James H, Michael S Barker, Eric J Carpenter, Michael K Deyholos, Matthew A Gitzendanner, Sean W Graham, Ivo Grosse, et al. 2019. "One Thousand Plant Transcriptomes and the Phylogenomics of Green Plants."

Lehmann, Erich L, and George Casella. 2006. *Theory of Point Estimation*. Springer Science & Business Media.

Li, Fay-Wei, Paul Brouwer, Lorenzo Carretero-Paulet, Shifeng Cheng, Jan De Vries, Pierre-Marc Delaux, Ariana Eily, et al. 2018. "Fern Genomes Elucidate Land Plant Evolution and Cyanobacterial Symbioses." *Nature Plants* 4 (7): 460–72.

Li, Ming, Xiwen Xu, Shanshan Liu, Guangyi Fan, Qian Zhou, and Songlin Chen. 2022. "The Chromosome-Level Genome Assembly of the Japanese Yellowtail Jack Seriola Aureovittata Provides Insights into Genome Evolution and Efficient Oxygen Transport." *Molecular Ecology Resources*.

Li, Qiuyi, Celine Scornavacca, Nicolas Galtier, and Yao-Ban Chan. 2021. "The Multilocus Multispecies Coalescent: A Flexible New Model of Gene Family Evolution." *Systematic*

Biology 70 (4): 822–37.

Li, Zhen, Jonas Defoort, Setareh Tasdighian, Steven Maere, Yves Van de Peer, and Riet De Smet. 2016. "Gene Duplicability of Core Genes Is Highly Consistent Across All Angiosperms." *The Plant Cell* 28 (2): 326–44.

Li, Zheng, Anthony E Baniaga, Emily B Sessa, Moira Scascitelli, Sean W Graham, Loren H Rieseberg, and Michael S Barker. 2015. "Early Genome Duplications in Conifers and Other Seed Plants." *Science Advances* 1 (10): e1501084.

Li, Zheng, and Michael S Barker. 2020. "Inferring Putative Ancient Whole-Genome Duplications in the 1000 Plants (1KP) Initiative: Access to Gene Family Phylogenies and Age Distributions." *GigaScience* 9 (2): giaa004.

Li, Zheng, George P Tiley, Sally R Galuska, Chris R Reardon, Thomas I Kidder, Rebecca J Rundell, and Michael S Barker. 2018. "Multiple Large-Scale Gene and Genome Duplications During the Evolution of Hexapods." *Proceedings of the National Academy of Sciences* 115 (18): 4713–8.

Librado, Pablo, Filipe G Vieira, and Julio Rozas. 2012. "BadiRate: Estimating Family Turnover Rates by Likelihood-Based Methods." *Bioinformatics* 28 (2): 279–81.

Liu, Bing-Jian, Kun Zhang, Shu-Fei Zhang, Yi-Fan Liu, Jia-Sheng Li, Ying Peng, Xun Jin, et al. 2022. "Chromosome-Level Genome Assembly of the Dotted Gizzard Shad (Konosirus Punctatus) Provides Insights into Its Adaptive Evolution." *Zoological Research* 43 (2): 217.

Liu, Hailin, Xiaobo Wang, Guibin Wang, Peng Cui, Shigang Wu, Cheng Ai, Nan Hu, et al. 2021. "The Nearly Complete Genome of Ginkgo Biloba Illuminates Gymnosperm Evolution." *Nature Plants* 7 (6): 748–56.

Liu, Jun S. 2001. *Monte Carlo Strategies in Scientific Computing*. Vol. 10. Springer.

Liu, Jun S, Rong Chen, and Wing Hung Wong. 1998. "Rejection Control and Sequential Importance Sampling." *Journal of the American Statistical Association* 93 (443): 1022–31.

Liu, Liang, Lili Yu, and Scott V Edwards. 2010. "A Maximum Pseudo-Likelihood Approach for Estimating Species Trees Under the Coalescent Model." *BMC Evolutionary Biology* 10 (1): 1–18.

Liu, Yang, Sibo Wang, Linzhou Li, Ting Yang, Shanshan Dong, Tong Wei, Shengdan Wu, et al. 2022. "The Cycas Genome and the Early Evolution of Seed Plants." *Nature Plants*, 1–13.

Lynch, Michael. 2007. *The Origins of Genome Architecture*. Vol. 98. Sinauer Associates Sunderland, MA.

Lynch, Michael, and John S Conery. 2000. "The Evolutionary Fate and Consequences of Duplicate Genes." *Science* 290 (5494): 1151–5.

———. 2003. "The Evolutionary Demography of Duplicate Genes." In *Genome Evolution*, 35–44. Springer.

MacKay, David JC. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge university press.

Maddison, Wayne P. 1997. "Gene Trees in Species Trees." *Systematic Biology* 46 (3): 523–36.

Maddison, Wayne P, Peter E Midford, and Sarah P Otto. 2007. "Estimating a Binary Character's Effect on Speciation and Extinction." *Systematic Biology* 56 (5): 701–10.

Maddison, Wayne P, and Montgomery Slatkin. 1991. "Null Models for the Number of Evolutionary Steps in a Character on a Phylogenetic Tree." *Evolution* 45 (5): 1184–97.

Maere, Steven, Stefanie De Bodt, Jeroen Raes, Tineke Casneuf, Marc Van Montagu, Martin Kuiper, and Yves Van de Peer. 2005. "Modeling Gene and Genome Duplications in Eukaryotes." *Proceedings of the National Academy of Sciences* 102 (15): 5454–9.

Mahmudi, Owais, Joel Sjöstrand, Bengt Sennblad, and Jens Lagergren. 2013. "Genome-Wide

Probabilistic Reconciliation Analysis Across Vertebrates." In *BMC Bioinformatics*, 14:1–11. 15. Springer.

Makino, Takashi, and Aoife McLysaght. 2010. "Ohnologs in the Human Genome Are Dosage Balanced and Frequently Associated with Disease." *Proceedings of the National Academy of Sciences* 107 (20): 9270–4.

Mallo, Diego, Leonardo de Oliveira Martins, and David Posada. 2016. "SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees." *Systematic Biology* 65 (2): 334–44.

Marcet-Houben, Marina, and Toni Gabaldón. 2015. "Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage." *PLoS Biology* 13 (8): e1002220.

Mayr, E. 1959. "Where Are We? Cold Spring Harbor Symp." *Quant. Biol.* 24: 1–14.

Mayr, Ernst. 1988. *Toward a New Philosophy of Biology: Observations of an Evolutionist*. 211. Harvard University Press.

Mayr, Ernst, and William B Provine. 1980. *The Evolutionary Synthesis*. Vol. 231. Cambridge, MA: Harvard University Press.

McKain, Michael R, Haibao Tang, Joel R McNeal, Saravanaraj Ayyampalayam, Jerrold I Davis, Claude W Depamphilis, Thomas J Givnish, J Chris Pires, Dennis Wm Stevenson, and James H Leebens-Mack. 2016. "A Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution Across the Poales." *Genome Biology and Evolution* 8 (4): 1150–64.

McLysaght, Aoife, and Laurence D Hurst. 2016. "Open Questions in the Study of de Novo Genes: What, How and Why." *Nature Reviews Genetics* 17 (9): 567–78.

Mendes, Fábio K, Dan Vanderpool, Ben Fulton, and Matthew W Hahn. 2021. "CAFE 5 Models Variation in Evolutionary Rates Among Gene Families." *Bioinformatics* 36 (22-23): 5516–8.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92.

Miklós, István, and Eric Tannier. 2010. "Bayesian Sampling of Genomic Rearrangement Scenarios via Double Cut and Join." *Bioinformatics* 26 (24): 3012–9.

Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. 2020. "IQ-Tree 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." *Molecular Biology and Evolution* 37 (5): 1530–4.

Minka, Thomas Peter. 2001. "A Family of Algorithms for Approximate Bayesian Inference." PhD thesis, Massachusetts Institute of Technology.

Mirarab, Siavash, Luay Nakhleh, and Tandy Warnow. 2021. "Multispecies Coalescent: Theory and Applications in Phylogenetics." *Annual Review of Ecology, Evolution, and Systematics* 52: 247–68.

Mirarab, Siavash, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. 2014. "ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation." *Bioinformatics* 30 (17): i541–i548.

Moran, Patrick Alfred Pierce. 1958. "Random Processes in Genetics." In *Mathematical Proceedings of the Cambridge Philosophical Society*, 54:60–71. 1. Cambridge University Press.

Morris, Jennifer L, Mark N Puttick, James W Clark, Dianne Edwards, Paul Kenrick, Silvia Pressel, Charles H Wellman, Ziheng Yang, Harald Schneider, and Philip CJ Donoghue. 2018. "The Timescale of Early Land Plant Evolution." *Proceedings of the National Academy of Sciences* 115 (10): E2274–E2283.

Muse, Spencer V, and Brandon S Gaut. 1994. “A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome.” *Molecular Biology and Evolution* 11 (5): 715–24.

Nakatani, Yoichiro, and Aoife McLysaght. 2019. “Macrosynteny Analysis Shows the Absence of Ancient Whole-Genome Duplication in Lepidopteran Insects.” *Proceedings of the National Academy of Sciences* 116 (6): 1816–8.

———. 2017. “Genomes as Documents of Evolutionary History: A Probabilistic Macrosynteny Model for the Reconstruction of Ancestral Genomes.” *Bioinformatics* 33 (14): i369–i378. https://doi.org/10.1093/bioinformatics/btx259.

Nakatani, Yoichiro, Prashant Shingate, Vydianathan Ravi, Nisha E Pillai, Aravind Prasad, Aoife McLysaght, and Byrappa Venkatesh. 2021. “Reconstruction of Proto-Vertebrate, Proto-Cyclostome and Proto-Gnathostome Genomes Provides New Insights into Early Vertebrate Evolution.” *Nature Communications* 12 (1): 1–14.

Nature. 2017. “Announcement: Towards Greater Reproducibility for Life-Sciences Research.” *Nature* 546: 8–8.

Neal, Radford M. 2011. “MCMC Using Hamiltonian Dynamics.” *Handbook of Markov Chain Monte Carlo* 2 (11): 2.

Nee, Sean. 2006. “Birth-Death Models in Macroevolution.” *Annu. Rev. Ecol. Evol. Syst.* 37: 1–17.

Nee, Sean, Edward C Holmes, Robert Mccredie May, and Paul H Harvey. 1994. “Extinction Rates Can Be Estimated from Molecular Phylogenies.” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 344 (1307): 77–82.

Nielsen, Rasmus, and Mark A Beaumont. 2009. “Statistical Inferences in Phylogeography.” *Molecular Ecology* 18 (6): 1034–47.

Ogilvie, Huw A, Remco R Bouckaert, and Alexei J Drummond. 2017. “StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates.” *Molecular Biology and Evolution* 34 (8): 2101–14.

Ohta, Tomoko. 1987. “Simulating Evolution by Gene Duplication.” *Genetics* 115 (1): 207–13.

Page, Roderic DM. 1994. “Maps Between Trees and Cladistic Analysis of Historical Associations Among Genes, Organisms, and Areas.” *Systematic Biology* 43 (1): 58–77.

Page, Roderic DM, and Michael A Charleston. 1997. “From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem.” *Molecular Phylogenetics and Evolution* 7 (2): 231–40.

Pamilo, Pekka, and Masatoshi Nei. 1988. “Relationships Between Gene Trees and Species Trees.” *Molecular Biology and Evolution* 5 (5): 568–83.

Peters, Gareth W, Yanan Fan, and Scott A Sisson. 2012. “On Sequential Monte Carlo, Partial Rejection Control and Approximate Bayesian Computation.” *Statistics and Computing* 22 (6): 1209–22.

Pevzner, Pavel, and Glenn Tesler. 2003. “Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes.” *Genome Research* 13 (1): 37–45.

Pond, Sergei L Kosakovsky, and Spencer V Muse. 2004. “Column Sorting: Rapid Calculation of the Phylogenetic Likelihood Function.” *Systematic Biology* 53 (5): 685–92.

Proost, Sebastian, Jan Fostier, Dieter De Witte, Bart Dhoedt, Piet Demeester, Yves Van de Peer, and Klaas Vandepoele. 2012. “I-Adhore 3.0—Fast and Sensitive Detection of Genomic Homology in Extremely Large Data Sets.” *Nucleic Acids Research* 40 (2): e11–e11.

Provine, William B. 1971. “The Origins of Theoretical Population Genetics.”

Prum, Richard O, Jacob S Berv, Alex Dornburg, Daniel J Field, Jeffrey P Townsend, Emily Mo-

riarty Lemmon, and Alan R Lemmon. 2015. “A Comprehensive Phylogeny of Birds (Aves) Using Targeted Next-Generation Dna Sequencing.” *Nature* 526 (7574): 569–73.

Rabier, Charles-Elie, Tram Ta, and Cécile Ané. 2014. “Detecting and Locating Whole Genome Duplications on a Phylogeny: A Probabilistic Approach.” *Molecular Biology and Evolution* 31 (3): 750–62.

Rabosky, Daniel L. 2014. “Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees.” *PloS One* 9 (2): e89543.

Rackauckas, Christopher, and Qing Nie. 2017. “DifferentialEquations.Jl – A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia.” *Journal of Open Research Software* 5 (1): 15. https://doi.org/10.5334/jors.151.

Rannala, Bruce, Scott V Edwards, Adam Leaché, and Ziheng Yang. 2020. “The Multi-Species Coalescent Model and Species Tree Inference.” No commercial publisher| Authors open access book.

Rannala, Bruce, and Ziheng Yang. 1996. “Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference.” *Journal of Molecular Evolution* 43 (3): 304–11.

———. 2003. “Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using Dna Sequences from Multiple Loci.” *Genetics* 164 (4): 1645–56.

———. 2013. “Improved Reversible Jump Algorithms for Bayesian Species Delimitation.” *Genetics* 194 (1): 245–53.

———. 2017. “Efficient Bayesian Species Tree Inference Under the Multispecies Coalescent.” *Systematic Biology* 66 (5): 823–42.

Rasmussen, Matthew D, and Manolis Kellis. 2012. “Unified Modeling of Gene Duplication, Loss, and Coalescence Using a Locus Tree.” *Genome Research* 22 (4): 755–65.

Redelings, Benjamin D, and Marc A Suchard. 2005. “Joint Bayesian Estimation of Alignment and Phylogeny.” *Systematic Biology* 54 (3): 401–18.

Ren, Ren, Haifeng Wang, Chunce Guo, Ning Zhang, Liping Zeng, Yamao Chen, Hong Ma, and Ji Qi. 2018. “Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms.” *Molecular Plant* 11 (3): 414–28.

Robert, Christian, and George Casella. 2011. “A Short History of Mcmc: Subjective Recollections from Incomplete Data.” *Handbook of Markov Chain Monte Carlo* 49.

Robert, Christian P. 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Vol. 2. Springer.

Robert, Christian P, and George Casella. 1999. *Monte Carlo Statistical Methods*. Vol. 2. Springer.

Roberts, Gareth O, and Jeffrey S Rosenthal. 2009. “Examples of Adaptive Mcmc.” *Journal of Computational and Graphical Statistics* 18 (2): 349–67.

Robertson, Fiona M, Manu Kumar Gundappa, Fabian Grammes, Torgeir R Hvidsten, Anthony K Redmond, Sigbjørn Lien, Samuel AM Martin, Peter WH Holland, Simen R Sandve, and Daniel J Macqueen. 2017. “Lineage-Specific Rediploidization Is a Mechanism to Explain Time-Lags Between Genome Duplication and Evolutionary Diversification.” *Genome Biology* 18 (1): 1–14.

Rodrigue, Nicolas, and Nicolas Lartillot. 2017. “Detecting Adaptation in Protein-Coding Genes Using a Bayesian Site-Heterogeneous Mutation-Selection Codon Substitution Model.” *Molecular Biology and Evolution* 34 (1): 204–14.

Rodrigue, Nicolas, Hervé Philippe, and Nicolas Lartillot. 2010. “Mutation-Selection Models of Coding Sequence Evolution with Site-Heterogeneous Amino Acid Fitness Profiles.” *Pro-*

ceedings of the National Academy of Sciences 107 (10): 4629–34.

Roelofs, Dick, Arthur Zwaenepoel, Tom Sistermans, Joey Nap, Andries A Kampfraath, Yves Van de Peer, Jacintha Ellers, and Ken Kraaijeveld. 2020. "Multi-Faceted Analysis Provides Little Evidence for Recurrent Whole-Genome Duplications During Hexapod Evolution." *BMC Biology* 18 (1): 1–13.

Rokas, Antonis, Barry L Williams, Nicole King, and Sean B Carroll. 2003. "Genome-Scale Approaches to Resolving Incongruence in Molecular Phylogenies." *Nature* 425 (6960): 798–804.

Ronquist, Fredrik, Maxim Teslenko, Paul Van Der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. 2012a. "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space." *Systematic Biology* 61 (3): 539–42.

———. 2012b. "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space." *Systematic Biology* 61 (3): 539–42.

Roodt, Danielle, Rolf Lohaus, Lieven Sterck, Riaan L Swanepoel, Yves Van de Peer, and Eshchar Mizrachi. 2017. "Evidence for an Ancient Whole Genome Duplication in the Cycad Lineage." *PLoS One* 12 (9): e0184454.

Roux, Camille, and John R Pannell. 2015. "Inferring the Mode of Origin of Polyploid Species from Next-Generation Sequence Data." *Molecular Ecology* 24 (5): 1047–59.

Ruprecht, Colin, Rolf Lohaus, Kevin Vanneste, Marek Mutwil, Zoran Nikoloski, Yves Van de Peer, and Staffan Persson. 2017. "Revisiting Ancestral Polyploidy in Plants." *Science Advances* 3 (7): e1603195.

Salter, Laura A. 2001. "Complexity of the Likelihood Surface for a Large Dna Dataset." *Systematic Biology* 50 (6): 970–78.

Sanderson, Michael J. 2003. "R8s: Inferring Absolute Rates of Molecular Evolution and Divergence Times in the Absence of a Molecular Clock." *Bioinformatics* 19 (2): 301–2.

Savage, Leonard J. 1972. *The Foundations of Statistics*. Courier Corporation.

Schrider, Daniel R, David Houle, Michael Lynch, and Matthew W Hahn. 2013. "Rates and Genomic Consequences of Spontaneous Mutational Events in Drosophila Melanogaster." *Genetics* 194 (4): 937–54.

Scornavacca, Celine, Frédéric Delsuc, and Nicolas Galtier. 2020. "Phylogenetics in the Genomic Era." No commercial publisher.

Seeger, Matthias. 2005. "Expectation Propagation for Exponential Families."

Sensalari, Cecilia, Steven Maere, and Rolf Lohaus. 2021. "Ksrates: Positioning Whole-Genome Duplications Relative to Speciation Events Using Rate-Adjusted Mixed Paralog–Ortholog Ks Distributions." *bioRxiv*.

Simon, Herbert A. 1955. "On a Class of Skew Distribution Functions." *Biometrika* 42 (3/4): 425–40.

Sisson, Scott A, Yanan Fan, and Mark Beaumont. 2018. *Handbook of Approximate Bayesian Computation*. CRC Press.

Sjöstrand, Joel, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. 2012. "DLRS: Gene Tree Evolution in Light of a Species Tree." *Bioinformatics* 28 (22): 2994–5.

Song, Xiaoming, Haibin Liu, Shaoqin Shen, Zhinan Huang, Tong Yu, Zhuo Liu, Qihang Yang, et al. 2022. "Chromosome-Level Pepino Genome Provides Insights into Genome Evolution and Anthocyanin Biosynthesis in Solanaceae." *The Plant Journal*.

Springer, Mark S, and John Gatesy. 2016. "The Gene Tree Delusion." *Molecular Phylogenetics and Evolution* 94: 1–33.

Stadler, Tanja, Roger Kouyos, Viktor von Wyl, Sabine Yerly, Jürg Böni, Philippe Bürgisser, Thomas Klimkait, et al. 2012. "Estimating the Basic Reproductive Number from Viral Sequence Data." *Molecular Biology and Evolution* 29 (1): 347–57.

Stark, Alexander, Michael F Lin, Pouya Kheradpour, Jakob S Pedersen, Leopold Parts, Joseph W Carlson, Madeline A Crosby, et al. 2007. "Discovery of Functional Elements in 12 Drosophila Genomes Using Evolutionary Signatures." *Nature* 450 (7167): 219–32.

Stein, Joshua C, Yeisoo Yu, Dario Copetti, Derrick J Zwickl, Li Zhang, Chengjun Zhang, Kapeel Chougule, et al. 2018. "Genomes of 13 Domesticated and Wild Rice Relatives Highlight Genetic Conservation, Turnover and Innovation Across the Genus Oryza." *Nature Genetics* 50 (2): 285–96.

Stigler, Stephen M. 2016. "The Seven Pillars of Statistical Wisdom." In *The Seven Pillars of Statistical Wisdom*. Harvard University Press.

Stoltz, Marnus, Boris Baeumer, Remco Bouckaert, Colin Fox, Gordon Hiscott, and David Bryant. 2021. "Bayesian Inference of Species Trees Using Diffusion Models." *Systematic Biology* 70 (1): 145–61.

Stoltzfus, Arlin. 2021. *Mutation, Randomness, and Evolution*. Oxford University Press.

Szöllősi, Gergely J, Bastien Boussau, Sophie S Abby, Eric Tannier, and Vincent Daubin. 2012. "Phylogenetic Modeling of Lateral Gene Transfer Reconstructs the Pattern and Relative Timing of Speciations." *Proceedings of the National Academy of Sciences* 109 (43): 17513–8.

Szöllősi, Gergely J, Wojciech Rosikiewicz, Bastien Boussau, Eric Tannier, and Vincent Daubin. 2013. "Efficient Exploration of the Space of Reconciled Gene Trees." *Systematic Biology* 62 (6): 901–12.

Szöllősi, Gergely J, Eric Tannier, Vincent Daubin, and Bastien Boussau. 2015. "The Inference of Gene Trees with Species Trees." *Systematic Biology* 64 (1): e42–e62.

Szöllősi, Gergely J, Eric Tannier, Nicolas Lartillot, and Vincent Daubin. 2013. "Lateral Gene Transfer from the Dead." *Systematic Biology* 62 (3): 386–97.

Tajima, Fumio. 1983. "Evolutionary Relationship of Dna Sequences in Finite Populations." *Genetics* 105 (2): 437–60.

Tasdighian, Setareh, Michiel Van Bel, Zhen Li, Yves Van de Peer, Lorenzo Carretero-Paulet, and Steven Maere. 2017. "Reciprocally Retained Genes in the Angiosperm Lineage Show the Hallmarks of Dosage Balance Sensitivity." *The Plant Cell* 29 (11): 2766–85.

Tavaré, Simon. 2018. "On the History of Abc." In *Handbook of Approximate Bayesian Computation*, 55–69. Chapman; Hall/CRC.

Tavaré, Simon, David J Balding, Robert C Griffiths, and Peter Donnelly. 1997. "Inferring Coalescence Times from Dna Sequence Data." *Genetics* 145 (2): 505–18.

Thomas, Gregg WC, S Hussain Ather, and Matthew W Hahn. 2017. "Gene-Tree Reconciliation with Mul-Trees to Resolve Polyploidy Events." *Systematic Biology* 66 (6): 1007–18.

Thompson, Elizabeth Alison, Kroonm Thompson, EA Thompson, and others. 1975. *Human Evolutionary Trees*. CUP Archive.

Thorne, Jeffrey L, Hirohisa Kishino, and Joseph Felsenstein. 1991. "An Evolutionary Model for Maximum Likelihood Alignment of Dna Sequences." *Journal of Molecular Evolution* 33 (2): 114–24.

Tiley, George P, Cécile Ané, and J Gordon Burleigh. 2016. "Evaluating and Characterizing Ancient Whole-Genome Duplications in Plants with Gene Count Data." *Genome Biology and Evolution* 8 (4): 1023–37.

Tiley, George P, Michael S Barker, and J Gordon Burleigh. 2018. "Assessing the Performance of Ks Plots for Detecting Ancient Whole Genome Duplications." *Genome Biology and Evo-*

lution 10 (11): 2882–98.

Tilly, Charles. 2004. "Observations of Social Processes and Their Formal Representations." *Sociological Theory* 22 (4): 595–602.

Tsitouras, Ch., I. Th. Famelis, and T. E. Simos. 2011. "On Modified Runge–Kutta Trees and Methods." *Computers & Mathematics with Applications* 62 (4): 2101–11. https://doi.org/10.1016/j.camwa.2011.06.058.

Ullah, Ikram, Joel Sjöstrand, Peter Andersson, Bengt Sennblad, and Jens Lagergren. 2015. "Integrating Sequence Evolution into Probabilistic Orthology Analysis." *Systematic Biology* 64 (6): 969–82.

Van Bel, Michiel, Tim Diels, Emmelien Vancaester, Lukasz Kreft, Alexander Botzki, Yves Van de Peer, Frederik Coppens, and Klaas Vandepoele. 2018. "PLAZA 4.0: An Integrative Resource for Functional, Evolutionary and Comparative Plant Genomics." *Nucleic Acids Research* 46 (D1): D1190–D1196.

Van Bel, Michiel, Francesca Silvestri, Eric M Weitz, Lukasz Kreft, Alexander Botzki, Frederik Coppens, and Klaas Vandepoele. 2022. "PLAZA 5.0: Extending the Scope and Power of Comparative and Functional Genomics in Plants." *Nucleic Acids Research* 50 (D1): D1468–D1474.

Van de Peer, Yves, Eshchar Mizrachi, and Kathleen Marchal. 2017. "The Evolutionary Significance of Polyploidy." *Nature Reviews Genetics* 18 (7): 411.

Van Dongen, Stijn Marinus. 2000. "Graph Clustering by Flow Simulation." PhD thesis.

Vanneste, Kevin, Guy Baele, Steven Maere, and Yves Van de Peer. 2014. "Analysis of 41 Plant Genomes Supports a Wave of Successful Genome Duplications in Association with the Cretaceous–Paleogene Boundary." *Genome Research* 24 (8): 1334–47.

Vanneste, Kevin, Yves Van de Peer, and Steven Maere. 2013. "Inference of Genome Duplications from Age Distributions Revisited." *Molecular Biology and Evolution* 30 (1): 177–90.

Vehtari, Aki, Andrew Gelman, Tuomas Sivula, Pasi Jylänki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P Cunningham, David Schiminovich, and Christian P Robert. 2020. "Expectation Propagation as a Way of Life: A Framework for Bayesian Inference on Partitioned Data." *J. Mach. Learn. Res.* 21: 17–11.

Vehtari, Aki, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. 2015. "Pareto Smoothed Importance Sampling." *arXiv Preprint arXiv:1507.02646*.

Verdinelli, Isabella, and Larry Wasserman. 1995. "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio." *Journal of the American Statistical Association* 90 (430): 614–18.

Walsh, Bruce. 2003. "Population-Genetic Models of the Fates of Duplicate Genes." In *Origin and Evolution of New Gene Functions*, 279–94. Springer.

Walsh, Bruce, and Michael Lynch. 2018. *Evolution and Selection of Quantitative Traits*. Oxford University Press.

Wang, Haifeng, Chunce Guo, Hong Ma, and Ji Qi. 2019. "Reply to Zwaenepoel et Al.: Meeting the Challenges of Detecting Polyploidy Events from Transcriptomic Data." *Molecular Plant* 12 (2): 137–40.

Watterson, GA. 1983. "On the Time for Gene Silencing at Duplicate Loci." *Genetics* 105 (3): 745–66.

Wen, Dingqiao, Yun Yu, and Luay Nakhleh. 2016. "Bayesian Inference of Reticulate Phylogenies Under the Multispecies Network Coalescent." *PLoS Genetics* 12 (5): e1006006.

Wickell, David, Li-Yaung Kuo, Hsiao-Pei Yang, Amra Dhabalia Ashok, Iker Irisarri, Armin Dadras, Sophie de Vries, et al. 2021. "Underwater Cam Photosynthesis Elucidated by Isoetes

Genome." *Nature Communications* 12 (1): 1–13.

Wilkinson, Richard David. 2013. "Approximate Bayesian Computation (Abc) Gives Exact Results Under the Assumption of Model Error." *Statistical Applications in Genetics and Molecular Biology* 12 (2): 129–41.

Wilson, Daniel J, Ryan D Hernandez, Peter Andolfatto, and Molly Przeworski. 2011. "A Population Genetics-Phylogenetics Approach to Inferring Natural Selection in Coding Sequences." *PLoS Genetics* 7 (12): e1002395.

Wolfe, Kenneth H. 2001. "Yesterday's Polyploids and the Mystery of Diploidization." *Nature Reviews Genetics* 2 (5): 333–41.

Wood, Simon N. 2010. "Statistical Inference for Noisy Nonlinear Ecological Dynamic Systems." *Nature* 466 (7310): 1102–4.

Wright, Sewall. 1931. "Evolution in Mendelian Populations." *Genetics* 16 (2): 97.

Wu, Yufeng. 2012. "Coalescent-Based Species Tree Inference from Gene Tree Topologies Under Incomplete Lineage Sorting by Maximum Likelihood." *Evolution: International Journal of Organic Evolution* 66 (3): 763–75.

———. 2016. "An Algorithm for Computing the Gene Tree Probability Under the Multispecies Coalescent and Its Application in the Inference of Population Tree." *Bioinformatics* 32 (12): i225–i233.

Xu, Bo, and Ziheng Yang. 2016. "Challenges in Species Tree Estimation Under the Multispecies Coalescent Model." *Genetics* 204 (4): 1353–68.

Xu, Jason, Peter Guttorp, Midori Kato-Maeda, and Vladimir N. Minin. 2015. "Likelihood-Based Inference for Discretely Observed Birth–Death-Shift Processes, with Applications to Evolution of Mobile Genetic Elements." *Biometrics* 71 (4): 1009–21. https://doi.org/10.1111/biom.12352.

Yang, Ya, Michael J Moore, Samuel F Brockington, Jessica Mikenas, Julia Olivieri, Joseph F Walker, and Stephen A Smith. 2018. "Improved Transcriptome Sampling Pinpoints 26 Ancient and More Recent Polyploidy Events in Caryophyllales, Including Two Allopolyploidy Events." *New Phytologist* 217 (2): 855–70.

Yang, Ya, Michael J Moore, Samuel F Brockington, Douglas E Soltis, Gane Ka-Shu Wong, Eric J Carpenter, Yong Zhang, et al. 2015. "Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing." *Molecular Biology and Evolution* 32 (8): 2001–14.

Yang, Ziheng. 2006. *Computational Molecular Evolution*. Oxford University Press.

———. 2007a. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8): 1586–91.

———. 2007b. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8): 1586–91.

Yoder, Anne D, and Ziheng Yang. 2000. "Estimation of Primate Speciation Dates Using Local Molecular Clocks." *Molecular Biology and Evolution* 17 (7): 1081–90.

York, Thomas L, Richard Durrett, and Rasmus Nielsen. 2002. "Bayesian Estimation of the Number of Inversions in the History of Two Chromosomes." *Journal of Computational Biology* 9 (6): 805–18.

Yu, Zhijun, Biao He, Zhen Gong, Yan Liu, Qianghui Wang, Xiaomin Yan, Tiantian Zhang, et al. 2022. "The New Haemaphysalis Longicornis Genome Provides Insights into Its Requisite Biological Traits." *Genomics* 114 (2): 110317.

Yule, George Udny. 1925. "II.—A Mathematical Theory of Evolution, Based on the Conclusions of Dr. JC Willis, Fr S." *Philosophical Transactions of the Royal Society of London. Series*

B, Containing Papers of a Biological Character 213 (402-410): 21–87.

Zhang, Chao, and Siavash Mirarab. 2022. "Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-Based Species Trees." *bioRxiv*.

Zhang, Chao, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. 2018. "ASTRAL-Iii: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees." *BMC Bioinformatics* 19 (6): 15–30.

Zhang, Chao, Celine Scornavacca, Erin K Molloy, and Siavash Mirarab. 2020. "ASTRAL-Pro: Quartet-Based Species-Tree Inference Despite Paralogy." *Molecular Biology and Evolution* 37 (11): 3292–3307.

Zhang, Cheng, and Frederick A Matsen IV. 2018a. "Generalizing Tree Probability Estimation via Bayesian Networks." In *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, 1449–58. Curran Associates, Inc.

———. 2018b. "Variational Bayesian Phylogenetic Inference." In *International Conference on Learning Representations*.

Zhao, Jing, Ashley I Teufel, David A Liberles, and Liang Liu. 2015. "A Generalized Birth and Death Process for Modeling the Fates of Gene Duplication." *BMC Evolutionary Biology* 15 (1): 1–11.

Zhao, Tao, and M Eric Schranz. 2019. "Network-Based Microsynteny Analysis Identifies Major Differences and Genomic Outliers in Mammalian and Angiosperm Genomes." *Proceedings of the National Academy of Sciences* 116 (6): 2165–74.

Zhao, Tao, Arthur Zwaenepoel, Jia-Yu Xue, Shu-Min Kao, Zhen Li, M Eric Schranz, and Yves Van de Peer. 2021. "Whole-Genome Microsynteny-Based Phylogeny of Angiosperms." *Nature Communications* 12 (1): 1–14.

Zhu, Sha, and James H Degnan. 2017. "Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent." *Systematic Biology* 66 (2): 283–98.

Zmasek, Christian M, and Sean R Eddy. 2001. "A Simple Algorithm to Infer Gene Duplication and Speciation Events on a Gene Tree." *Bioinformatics* 17 (9): 821–28.

Zwaenepoel, Arthur, Zhen Li, Rolf Lohaus, and Yves Van de Peer. 2019. "Finding Evidence for Whole Genome Duplications: A Reappraisal." *Molecular Plant* 12 (2): 133–36.

Zwaenepoel, Arthur, and Yves Van de Peer. 2019a. "Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates." *Molecular Biology and Evolution* 36 (7): 1384–1404.

———. 2019b. "Wgd—Simple Command Line Tools for the Analysis of Ancient Whole-Genome Duplications." *Bioinformatics* 35 (12): 2153–5.

———. 2020. "Model-Based Detection of Whole-Genome Duplications in a Phylogeny." *Molecular Biology and Evolution* 37 (9): 2734–46.

———. 2021. "A Two-Type Branching Process Model of Gene Family Evolution." *bioRxiv*.

Curriculum vitae

Personal information

Name	Arthur Zwaenepoel
Date of Birth	20 February 1995
Place of Birth	Ghent, Belgium
Address	Zilverenberg 14E, 9000 Ghent (Belgium)
Phone	+32 473852569
email	arthur.zwaenepoel@gmail.com
Github	https://github.com/arzwa

Education

2017 - 2022	**PhD candidate in Bioinformatics** Ghent University, Belgium VIB/UGent department of plant systems biology Fellow of the Flanders Fund for Scientific Research (FWO)
2015 - 2017	**Master of science in Bioinformatics (systems biology)** Ghent University, Belgium *summa cum laude*
2013 - 2015	**Honours programme 'Quetelet Colleges'** Ghent University, Belgium
2012 - 2015	**Bachelor of science in Biochemistry & Biotechnology** Ghent University, Belgium *summa cum laude*

Publications

An asterisk (*) next to my name indicates (co)first authorship.

1. T. Zhao, **A. Zwaenepoel**, J.-Y. Xue, S.-M. Kao, Z. Li, M. E. Schranz, and Y. Van de Peer, "Whole-genome microsynteny-based phylogeny of angiosperms," *Nature Communications*, vol. 12, no. 1, pp. 1–14, 2020.
2. X. Wang, S. Chen, X. Ma, A. E. Yssel, S. R. Chaluvadi, M. S. Johnson, P. Gangashetty, F. Hamidou, M. D. Sanogo, **A. Zwaenepoel**, *et al.*, "Genome sequence and genetic diversity analysis of an under-domesticated orphan crop, white fonio (digitaria exilis)," *GigaScience*, vol. 10, no. 3, p. giab013, 2021.
3. **A. Zwaenepoel*** and Y. Van de Peer, "Model-based detection of whole-genome duplications in a phylogeny," *Molecular biology and evolution*, vol. 37, no. 9, pp. 2734–2746, 2020.
4. D. Roelofs, **A. Zwaenepoel***, T. Sistermans, J. Nap, A. A. Kampfraath, Y. Van de Peer, J. Ellers, and K. Kraaijeveld, "Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution," *BMC biology*, vol. 18, pp. 1–13, 2020.
5. J. Zhang, X.-X. Fu, R.-Q. Li, X. Zhao, Y. Liu, M.-H. Li, **A. Zwaenepoel**, H. Ma, B. Goffinet, Y.-L. Guan, *et al.*, "The hornwort genome and early land plant evolution," *Nature plants*, vol. 6, no. 2, pp. 107–118, 2020.
6. **A. Zwaenepoel*** and Y. Van de Peer, "Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates," *Molecular biology and evolution*, vol. 36, no. 7, pp. 1384–1404, 2019.
7. **A. Zwaenepoel*** and Y. Van de Peer, "wgd—simple command line tools for the analysis of ancient whole-genome duplications," *Bioinformatics*, vol. 35, no. 12, pp. 2153–2155, 2019.
8. **A. Zwaenepoel***, Z. Li, R. Lohaus, and Y. Van de Peer, "Finding evidence for whole genome duplications: a reappraisal," *Molecular plant*, vol. 12, no. 2, pp. 133–136, 2019.
9. **A. Zwaenepoel***, T. Diels, D. Amar, T. Van Parys, R. Shamir, Y. Van de Peer, and O. Tzfadia, "Morphdb: prioritizing genes for specialized metabolism pathways and gene ontology categories in plants," *Frontiers in plant science*, vol. 9, p. 352, 2018.

Preprints

1. H. Chen, Y. Fang, **A. Zwaenepoel**, S. Huang, Y. Van de Peer, Z. Li "Revisiting Ancient Polyploidy in Leptosporangiate Ferns," *bioRxiv*, 2022.
2. **A. Zwaenepoel***, & Y. Van de Peer. "A two-type branching process model of gene family evolution," *bioRxiv*, 2021.

Contributed presentations

1. Probabilistic Modeling in Genomics, Aussois, France, (2020), "Bayesian statistical inference of ancient whole-genome duplications".
2. International Conference on Polyploidy, Ghent, Belgium, (2019), "Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates".

3. ALPHY (Alignment & Phylogeny) meeting, Paris, France, (2019), "Inference of ancient whole-genome duplications using statistical gene tree reconciliation".

Invited presentations

1. Polyploidy webinar (April 2021) – https://www.barkerlab.net/polyweb, "Statistical inference of whole-genome duplications in a phylogenetic context".

Courses and workshops

1. EMBO 10-day practical course on Computational Molecular Evolution (2018) – Heraklion (Crete, Greece).
2. ForBio workshop on polyploid population genetics (2019) – Drøbak (Norway)

Teaching experience

Teaching assistant / co-lecturer for the following courses:

- *Bioinformatics II* (MSc Biochemistry & Biotechnology, Prof. Yves Van de Peer) (2018-2022): I developed and organized practical sessions on statistical phylogenetics for this course.
- *Evolutionary biology* (MSc Bioinformatis, Prof. Yves Van de Peer) (2018-2022): I have taught theory and organized practical sessions on statistical phylogenetics for this course.
- *Philosophical issues in the life sciences* (MA philosophy, Prof. Gertrudis Van de Vijver) (2020-2022): I co-organized this course with professor Van de Vijver, guiding discussions on key texts in the history and philosophy of biology. I have also lectured on 20th century evolutionary biology and the history of genetics in this course.

Project supervisor for Master 1 group projects (Design Project)

- In the 2018 – 2019 academic year I supervised a group of four students working on a project related to bioinformatic inference of ancient genome duplications.

- In the 2019 – 2020 academic year I supervised a group of five students working on a phylogenomics project, solving hard phylogenetics problems using advanced statistical tools.

Supervisor for Master thesis students (Msc. Bioinformatics):

- Sam Vanmassenhove (2019-2020): Variational inference for Bayesian models of gene family evolution
- Michael Vandevoorde (2020-2021): PolyStab: Individual-based modeling to study polypoid establishment from an eco-evo perspective

Other academic

Co-organizer of the International Conference on Poyploidy in Ghent, Belgium (2019) (I did the scheduling of the talks, designed the conference programme booklet and various other tasks).

Co-editor for the book "Methods in Molecular Biology: Polyploidy", with Yves Van de Peer and Zhen Li.

I have reviewed for: Molecular Biology & Evolution, PLOS Genetics, Journal of Evolutionary Biology, Communications Biology, Proceedings of the Royal Society B.

Languages, skills and other interests

Languages: Dutch (native), English (fluent), French (good)

Programming languages: Julia (preferred), Python (proficient), R (ok), Perl (rusty). I have hacked on existing C++ codebases.

Other relevant skills: Bayesian statistics, probabilistic modeling, probabilistic programming, statistical phylogenetics, population genetics, evolutionary theory, Bioinformatics

Other interests: Philosophy, mathematics, natural science, computer science, literature, music, running, cycling, skateboarding