
Phylogeny



CBMS-NSF REGIONAL CONFERENCE SERIES IN APPLIED MATHEMATICS

A series of lectures on topics of current research interest in applied mathematics under the direction of the Conference Board of the Mathematical Sciences, supported by the National Science Foundation and published by SIAM.

- GARRETT BIRKHOFF, *The Numerical Solution of Elliptic Equations*
D. V. LINDLEY, *Bayesian Statistics, A Review*
R. S. VARGA, *Functional Analysis and Approximation Theory in Numerical Analysis*
R. R. BAHADUR, *Some Limit Theorems in Statistics*
PATRICK BILLINGSLEY, *Weak Convergence of Measures: Applications in Probability*
J. L. LIONS, *Some Aspects of the Optimal Control of Distributed Parameter Systems*
ROGER PENROSE, *Techniques of Differential Topology in Relativity*
HERMAN CHERNOFF, *Sequential Analysis and Optimal Design*
J. DURBIN, *Distribution Theory for Tests Based on the Sample Distribution Function*
SOL I. RUBINOW, *Mathematical Problems in the Biological Sciences*
P. D. LAX, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*
I. J. SCHOENBERG, *Cardinal Spline Interpolation*
IVAN SINGER, *The Theory of Best Approximation and Functional Analysis*
WERNER C. RHEINBOLDT, *Methods of Solving Systems of Nonlinear Equations*
HANS F. WEINBERGER, *Variational Methods for Eigenvalue Approximation*
R. TYRRELL ROCKAFELLAR, *Conjugate Duality and Optimization*
SIR JAMES Lighthill, *Mathematical Biofluidynamics*
GERARD SALTON, *Theory of Indexing*
CATHLEEN S. MORAWETZ, *Notes on Time Decay and Scattering for Some Hyperbolic Problems*
F. HOPPENSTEADT, *Mathematical Theories of Populations: Demographics, Genetics and Epidemics*
RICHARD ASKEY, *Orthogonal Polynomials and Special Functions*
L. E. PAYNE, *Improperly Posed Problems in Partial Differential Equations*
S. ROSEN, *Lectures on the Measurement and Evaluation of the Performance of Computing Systems*
HERBERT B. KELLER, *Numerical Solution of Two Point Boundary Value Problems*
J. P. LASALLE, *The Stability of Dynamical Systems*
D. GOTTLIEB AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*
PETER J. HUBER, *Robust Statistical Procedures*
HERBERT SOLOMON, *Geometric Probability*
FRED S. ROBERTS, *Graph Theory and Its Applications to Problems of Society*
JURIS HARTMANIS, *Feasible Computations and Provable Complexity Properties*
ZOHAR MANNA, *Lectures on the Logic of Computer Programming*
ELLIS L. JOHNSON, *Integer Programming: Facets, Subadditivity, and Duality for Group and Semi-Group Problems*
SHMUEL WINOGRAD, *Arithmetic Complexity of Computations*
J. F. C. KINGMAN, *Mathematics of Genetic Diversity*
MORTON E. GURTIN, *Topics in Finite Elasticity*
THOMAS G. KURTZ, *Approximation of Population Processes*
JERROLD E. MARSDEN, *Lectures on Geometric Methods in Mathematical Physics*
BRADLEY EFRON, *The Jackknife, the Bootstrap, and Other Resampling Plans*
M. WOODROOFE, *Nonlinear Renewal Theory in Sequential Analysis*
D. H. SATTINGER, *Branching in the Presence of Symmetry*
R. TEMAM, *Navier-Stokes Equations and Nonlinear Functional Analysis*
MIKLÓS CSÖRGÖ, *Quantile Processes with Statistical Applications*
J. D. BUCKMASTER AND G. S. S. LUDFORD, *Lectures on Mathematical Combustion*
R. E. TARJAN, *Data Structures and Network Algorithms*

- PAUL WALTMAN, *Competition Models in Population Biology*
S. R. S. VARADHAN, *Large Deviations and Applications*
KIYOSI ITÔ, *Foundations of Stochastic Differential Equations in Infinite Dimensional Spaces*
ALAN C. NEWELL, *Solitons in Mathematics and Physics*
PRANAB KUMAR SEN, *Theory and Applications of Sequential Nonparametrics*
LÁSZLÓ LOVÁSZ, *An Algorithmic Theory of Numbers, Graphs and Convexity*
E. W. CHENEY, *Multivariate Approximation Theory: Selected Topics*
JOEL SPENCER, *Ten Lectures on the Probabilistic Method*
PAUL C. FIFE, *Dynamics of Internal Layers and Diffusive Interfaces*
CHARLES K. CHUI, *Multivariate Splines*
HERBERT S. WILF, *Combinatorial Algorithms: An Update*
HENRY C. TUCKWELL, *Stochastic Processes in the Neurosciences*
FRANK H. CLARKE, *Methods of Dynamic and Nonsmooth Optimization*
ROBERT B. GARDNER, *The Method of Equivalence and Its Applications*
GRACE WAHBA, *Spline Models for Observational Data*
RICHARD S. VARGA, *Scientific Computation on Mathematical Problems and Conjectures*
INGRID DAUBECHIES, *Ten Lectures on Wavelets*
STEPHEN F. MCCORMICK, *Multilevel Projection Methods for Partial Differential Equations*
HARALD NIEDERREITER, *Random Number Generation and Quasi-Monte Carlo Methods*
JOEL SPENCER, *Ten Lectures on the Probabilistic Method, Second Edition*
CHARLES A. MICCHELLI, *Mathematical Aspects of Geometric Modeling*
ROGER TEMAM, *Navier-Stokes Equations and Nonlinear Functional Analysis, Second Edition*
GLENN SHAFER, *Probabilistic Expert Systems*
PETER J. HUBER, *Robust Statistical Procedures, Second Edition*
J. MICHAEL STEELE, *Probability Theory and Combinatorial Optimization*
WERNER C. RHEINBOLDT, *Methods for Solving Systems of Nonlinear Equations, Second Edition*
J. M. CUSHING, *An Introduction to Structured Population Dynamics*
TAI-PING LIU, *Hyperbolic and Viscous Conservation Laws*
MICHAEL RENARDY, *Mathematical Analysis of Viscoelastic Flows*
GÉRARD CORNUÉJOLS, *Combinatorial Optimization: Packing and Covering*
IRENA LASIECKA, *Mathematical Control Theory of Coupled PDEs*
J. K. SHAW, *Mathematical Principles of Optical Fiber Communications*
ZHANGXIN CHEN, *Reservoir Simulation: Mathematical Techniques in Oil Recovery*
ATHANASSIOS S. FOKAS, *A Unified Approach to Boundary Value Problems*
MARGARET CHENEY AND BRETT BORDEN, *Fundamentals of Radar Imaging*
FIORALBA CAKONI, DAVID COLTON, AND PETER MONK, *The Linear Sampling Method in Inverse Electromagnetic Scattering*
ADRIAN CONSTANTIN, *Nonlinear Water Waves with Applications to Wave-Current Interactions and Tsunamis*
WEI-MING NI, *The Mathematics of Diffusion*
ARNULF JENTZEN AND PETER E. KLOEDEN, *Taylor Approximations for Stochastic Partial Differential Equations*
FRED BRAUER AND CARLOS CASTILLO-CHAVEZ, *Mathematical Models for Communicable Diseases*
PETER KUCHMENT, *The Radon Transform and Medical Imaging*
ROLAND GLOWINSKI, *Variational Methods for the Numerical Solution of Nonlinear Elliptic Problems*
BENGT FORNBERG AND NATASHA FLYER, *A Primer on Radial Basis Functions with Applications to the Geosciences*
MIKE STEEL, *Phylogeny: Discrete and Random Processes in Evolution*

Mike Steel

University of Canterbury
Christchurch, New Zealand

Phylogeny

Discrete and Random Processes in Evolution



SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS
PHILADELPHIA

Copyright © 2016 by the Society for Industrial and Applied Mathematics.

10 9 8 7 6 5 4 3 2 1

All rights reserved. Printed in the United States of America. No part of this book may be reproduced, stored, or transmitted in any manner without the written permission of the publisher. For information, write to the Society for Industrial and Applied Mathematics, 3600 Market Street, 6th Floor, Philadelphia, PA 19104-2688 USA.

Trademarked names may be used in this book without the inclusion of a trademark symbol. These names are used in an editorial context only; no infringement of trademark is intended.

Publisher	<i>David Marshall</i>
Acquisitions Editor	<i>Elizabeth Greenspan</i>
Developmental Editor	<i>Gina Rinelli Harris</i>
Managing Editor	<i>Kelly Thomas</i>
Production Editor	<i>Lisa Briggeman</i>
Copy Editor	<i>Matthew Bernard</i>
Production Manager	<i>Donna Witzleben</i>
Production Coordinator	<i>Cally Shrader</i>
Compositor	<i>Techsetters, Inc.</i>
Graphic Designer	<i>Lois Sellers</i>

Library of Congress Cataloging-in-Publication Data

Names: Steel, M. A. | Society for Industrial and Applied Mathematics.

Title: Phylogeny : discrete and random processes in evolution / Michael Steel, University of Canterbury, Christchurch, New Zealand.

Description: Philadelphia : Society for Industrial and Applied Mathematics, [2016] | Series: CBMS-NSF regional conference series in applied mathematics ; 89 | Includes bibliographical references and index.

Identifiers: LCCN 2016019513 | ISBN 9781611974478

Subjects: LCSH: Phylogeny. | Evolution (Biology)--Statistical methods. | Mathematical statistics. | Probabilities.

Classification: LCC QH367.5 .S74 2016 | DDC 576.8/8--dc23 LC record available at <https://lccn.loc.gov/2016019513>

siam is a registered trademark.

Contents

Preface	ix
Acknowledgments	xi
Commonly Used Symbols	xiii
1 Phylogeny	1
1.1 What is phylogenetics?	1
1.2 Preliminaries	2
1.3 Phylogenetic trees	9
2 Basic combinatorics of discrete phylogenies	15
2.1 Counting trees	15
2.2 Rooted trees as nested sets of clusters	18
2.3 Refinement, compatibility, and encoding	21
2.4 Unrooted trees as systems of splits	23
2.5 Tree rearrangement metrics	29
2.6 Consensus functions	36
3 Tree shape and random discrete phylogenies	41
3.1 Tree shapes	41
3.2 The shape of evolving trees	43
3.3 Measuring and modeling tree shape	53
3.4 Cherries and extended Pólya urn models	58
4 Pulling trees apart and putting trees together	63
4.1 Restriction and display	63
4.2 When is a collection of trees compatible?	67
4.3 Sets of trees that “define” and “identify” a phylogeny	72
4.4 Agreement subtrees	79
4.5 Phylogenetic decisiveness and terraces	82
5 Phylogenies based on discrete characters	87
5.1 Characters, homoplasy, and perfect phylogeny	87
5.2 Minimal evolution (maximum parsimony (MP))	100
5.3 Minimal evolution trees for a sequence of characters	107
6 Continuous phylogenies and distance-based tree reconstruction	111
6.1 Metrics from trees with edge lengths	111

6.2	Distance-based tree reconstruction methods	121
6.3	Generalizations and geometry	129
6.4	Phylogenetic diversity	133
7	Evolution on a tree: Part one	147
7.1	Nonhomogeneous Markov chains	148
7.2	From Markov chains to processes on trees	153
7.3	Classes and properties of models	157
7.4	The Hadamard story	164
7.5	Phylogenetic mixture models	171
8	Evolution on a tree: Part two	177
8.1	Preliminaries	177
8.2	Phylogeny reconstruction methods and properties	180
8.3	Algebraic analysis of Markov models	191
8.4	The infinite-state random cluster model	197
8.5	Additional topics	203
9	Evolution of trees	205
9.1	Yule pure-birth trees: The simplest model	206
9.2	Birth-death models	211
9.3	Gene trees and species trees	224
10	Introduction to phylogenetic networks	237
10.1	To tree or not to tree: Why networks?	237
10.2	Implicit (unrooted) networks	238
10.3	Explicit (directed) networks	245
10.4	Trees displayed by networks	253
10.5	Reconstructing networks	261
10.6	Additional topics	267
Bibliography		269
Index		291

Preface

The idea that all life on earth traces back to a common origin dates back at least to Charles Darwin's *Origin of Species*. Ever since, biologists have tried to piece together parts of this “tree of life” based on what we can observe today: fossils, and the evolutionary signal that is present in the genomes and phenotypes of different organisms. Mathematics has played a key role in helping transform genetic data into phylogenetic (evolutionary) trees and networks. In this book, I will explain some of the central concepts and basic results in phylogenetics, which benefit from several branches of mathematics, including combinatorics, probability, and algebra.

As well as providing an overview of this field, I have also tried to highlight many of the advances in this field that have taken place since my earlier book with Charles Semple (*Phylogenetics*, Oxford University Press, 2003). It is quite amazing how much this field has developed in the past dozen years. It soon became clear that one of the hardest tasks would be deciding what to leave out in order to complete this book on schedule. Rather than attempting a comprehensive survey of the current state of the art, this book provides an updated summary of the main theory, supplemented by a selection of topics that have mathematical appeal and either proven or potential biological relevance. In confining the scope to mathematical topics, there is comparatively little detail about the many impressive advances that have been made by the theoretical computer science community in developing more efficient algorithms or resolving outstanding computational complexity questions.

The chapters roughly follow the outline of the series of 10 lectures I gave at the NSF/CBMS Conference on Mathematical Phylogeny (Winthrop University) in June 2014, and is based around my November 2014 survey paper on phylogenetics in a special issue of the *American Mathematical Monthly* devoted to mathematical biology.

Where some material overlaps with the earlier book, *Phylogenetics*, I have tried to ensure that the presentation here is briefer, presented differently, or updated. For example, there is more emphasis here on stochastic models, particularly in the new material in Chapters 3, 8, and 9. Topics such as phylogenetic diversity are now described, and there is a more detailed treatment in Chapter 9 of how species trees and gene trees evolve. Chapter 10 provides an introduction to the very active research area of phylogenetic networks (which complements the excellent 2010 book [203] on this topic). Several of the specialized topics in the earlier book are also omitted here.

In this book, phylogenies (rather than the more abstract X -tree notion) are central, hierarchies rather than split systems take precedence, and rooted trees are drawn growing upwards like biological trees (however, in the final chapter, I bow to convention and orient phylogenetic networks downwards). Keeping the mathematical requirements modest has precluded descriptions of some of the more technical results in detail. However, I hope this makes the text a bit more accessible for readers from other disciplines, particularly biologists, who wish to better understand the mathematical foundations of phylogenetics.

Regarding the structure of this book, most chapters rely on earlier ones in some way, though some chapters can be read without having covered all the earlier material; Fig. 1 illustrates the main flow of concepts between the chapters.

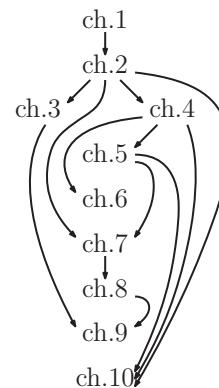


Figure 1. *The main flow of concepts between the chapters of this book.*

A number of exercises (and some examples) are highlighted in boxes throughout the chapters. The exercises marked with ⁺ are slightly more challenging than the others.

Acknowledgments

Funding for this work was made possible by the NZ Marsden Fund, the Allan Wilson Centre, and SIAM. I would like to thank Dietrich Radel for technical assistance with typesetting and figure preparation, Megan Foster for proofreading, Jotun Hein and the Oxford science book discussion group for helpful feedback, and Joseph Felsenstein for permission to reproduce Fig. 2.5.

I also thank numerous others who have provided helpful feedback, references, and critical comments on particular sections. These include Tanja Stadler, Erick Matsen, David Bryant, Marta Casanellas, Jotun Hein, Arne Mooers, Katharina Huber, Fred (Buck) McMorris, Jeremy Sumner, Barbara Holland, Elizabeth Allman, Elliott Sober, Noah Rosenberg, Fabio Pardi, Celine Scornavacca, Charles Semple, Paola Bonizzoni, Andrew Francis, Giulio Dalla Riva, and Olivier Gascuel.

Finally, I would like to thank Joseph Rusinko and Trent Kull (Winthrop University) for organizing the phylogenetic lecture series in 2014, on which this book is based, and Elizabeth Greenspan, Lisa Briggeman, John Rogosich and Gina Rinelli Harris at SIAM for their ongoing support and encouragement.

Commonly Used Symbols

Chapter 1

$\Delta[G]$	diameter of graph G , 3
$P(T; u, v)$	path in T connecting u and v , 4
$\text{cy}(G)$	cyclomatic number of graph G , 5
$\text{tw}(G)$	treewidth of G , 6
\preceq_T	partial order on vertices of a rooted tree T , 6
$\text{lca}_T(x, y)$	least common ancestor of x, y in T , 6
$S(T)$	symmetry group of T , 7
$c(T)$	number of cherries of T , 11
$P(X), P(n)$	phylogenetic trees on X (on $[n]$), 12
$RP(X), RP(n)$	rooted phylogenetic trees on X (on $[n]$), 12
$B(X), B(n)$	binary phylogenetic trees on X (on $[n]$), 12
$RB(X), RB(n)$	rooted binary phylogenetic trees on X (on $[n]$), 12
$b(n), r b(n)$	size of $B(n)$ and $RB(n)$, 12

Chapter 2

$c_T(v)$	cluster associated with v , 18
$\mathcal{C}(T)$	clusters associated with T , 18
$\mathcal{H}[\mathcal{S}]$	hierarchy associated with \mathcal{S} , 20
\preceq	tree refinement partial order, 21
$A B$	X -split $\{A, B\}$, 23
$\Sigma(T)$	splits of T , 23
$\overset{\circ}{\Sigma}(T)$	nontrivial splits of T , 23
$d_{\text{RF}}(T, T')$	Robinson–Foulds and distance between T and T' , 25
$B(\Sigma)$	Buneman graph for Σ , 26
$o(T)$	circular orderings of T , 29
$G_\theta(n)$	θ -tree space, 32
$A \perp B$ ($A \perp T$)	incompatibility of A with split B (with tree T), 36
φ_{AD}	Adams consensus function, 38

Chapter 3

$\text{Stab}(T)$	symmetry group of T ($= S(T)$), 41
λ_v	number of interior vertices descended from v (including v), 46
$c(\tau)$	number of cherries of tree shape τ , 49
$p_\theta(T)$	probability that $\mathcal{T} = T$ under model θ , 52
C_T	Colless index of T , 53
S_T	Sackin index of T , 53
χ_n	number of cherries in a YH model, 59
χ'_n	number of cherries in a uniform model, 60

Chapter 4

$T Y$	restriction of T to Y , 63
$\mathcal{L}(\mathcal{P})$	leaves present in trees from \mathcal{P} , 65
$\langle \mathcal{P} \rangle$	span of \mathcal{P} , 65
$\langle \mathcal{P} \rangle_B$	binary span of \mathcal{P} , 65
$\mathcal{Q}(T)$	quartet tree displayed by T , 65
$q(\mathcal{P})$	quartet trees displayed by at least one tree in \mathcal{P} , 66
$T_{\mathcal{Q}}$	the \mathcal{Q}^* -tree, 66
$\mathcal{R}(T)$	rooted triples displayed by T , 67
$r(\mathcal{R})$	rooted triples displayed by at least one tree in \mathcal{R} , 67
$\mathcal{A}_{\mathcal{R}}$	the BUILD tree, 67
$[\mathcal{R}, S]$	graph for finding $\mathcal{A}_{\mathcal{R}}$, 68
$\text{exc}(\mathcal{P})$	excess of \mathcal{P} , 70
$G(\mathcal{P})$	display graph of \mathcal{P} , 71
$\sigma(T)$	blocks in splits of T , 77
$d_{\mathcal{Q}}$	quartet metric, 80
$T \chi$	T restricted to the subsets of X in χ , 82

Chapter 5

$\Pi(f)$	partition of X induced by f , 88
$T[B]$	minimal subtree of T connecting B , 88
$q(\mathcal{C})$	quartet trees associated with \mathcal{C} , 89
$\text{int}(\mathcal{C})$	partition intersection graph for \mathcal{C} , 90
$\text{int}_T(\mathcal{C})$	$\text{int}(\mathcal{C})$ with additional edges induced by T , 90
$f_{(A B)}$	partial binary character corresponding to the split $A B$, 99
$\text{ps}(f, T)$	parsimony score of f on T , 100
$h(f, T)$	homoplasy score of f on T , 102
$\Sigma_\theta(T, r)$	splits in the θ -neighborhood of T , 103

Chapter 6

$d_{(T,l)}$	tree metric represented on T , 111
$\mathcal{A}[\delta]$	set of Apresjan clusters of δ , 114
δ_U	ultrametric from the tree of Apresjan clusters, 115
δ_{SD}	subdominant ultrametric, 115
$\tilde{d}_r(x,y)$	Gromov product, 117
$d_r(x,y)$	Gromov–Farris transform, 117
$\partial(x,y)$	symbolic ultrametric, 118
$\delta_{\mathcal{C}}$	normalized Hamming distance from \mathcal{C} , 119
L	total length of a tree, 125
$\lambda_T(x,y)$	coefficients in the generalized Paulin formula, 126
\mathcal{L}	a subset of $\binom{X}{2}$, 127
\mathbb{T}_X	continuous tree space, 131
$\tilde{\mathbb{T}}_X$	link of the origin, 131
$PD_{(T,l)}$	phylogenetic diversity, 133
$c_T(e)$	leaves descended from e , 137
ψ_T	phylogenetic diversity index, 138

Chapter 7

$\text{diag}(\pi)$	diagonal matrix with entries of π on the diagonal, 150
$P^{(e)}$	transition matrix for edge e , 153
$p(f T,\theta)$	probability of character f from (T,θ) , 154
J^{xy}	joint probability of states at x,y , 155
p_A	probability of pattern associates with A , 165
$\mathcal{E}(n)$	the subsets of $[n]$ of even size, 166
$P(T;B)$	path set induced by $B \in \mathcal{E}(n)$, 166
ω_i	$1 - 2p_i$, 166
H	Hadamard matrix, 167
\tilde{H}	symmetric Hadamard matrix, 169

Chapter 8

d_1	l_1 distance, 177
d_H	Hellinger distance, 177
d_{KL}	Kullback–Leibler divergence, 178
$I(X,Y)$	mutual information for X,Y , 178
\mathbb{O}_T	phylogenetic orange space for T , 190
\mathbb{O}_n	space of phylogenetic oranges, 190
$\text{flat}_{A B}(p)$	the flattening of p for the split $A B$, 195
RC_∞	infinite-state random cluster model, 197

Chapter 9

$N(t)$	number of leaves in birth-death tree at time t , 206
$\mathcal{T}_{(n)}$	Yule pure-birth tree with n leaves, 208
L_n	sum of the lengths of the edges, 208
$r (= \lambda - \mu)$	diversification rate, 212
$p_i(t)$	the probability that $N(t) = i$, 212
$\mathcal{T}, \mathcal{T}_t$	complete tree, 213
$\tilde{\mathcal{T}}, \tilde{\mathcal{T}}_t$	reconstructed tree, 213
$N_\tau(t)$	number of lineages in \mathcal{T}_t that have at least one lineage at time τ , 214
\mathcal{E}_τ	the event that $N(\tau) > 0$, 215
$\mathcal{H}_n^{\text{KC}}$	total height of the Kingman coalescent tree, 220
$\mathcal{L}_n^{\text{KC}}$	total sum of edge lengths of the Kingman coalescent tree, 220
γ	gamma statistic, 220
T_s, T_g	species tree, gene tree, 224
\mathcal{T}_g	gene tree under ILS (or LGT), 225
$\text{xl}(T, e)$	the number of extra lineages in e , 230
$\text{dc}(T, S)$	the deep coalescence cost for reconciling tree T within tree S , 231

Chapter 10

v, t, r	number of vertices, tree vertices, reticulate vertices in a network, 246
\preceq_N	partial order on the vertices of a directed network, 246
\tilde{N}	the collapsed network associated with N , 251
$c_N(v)$	the (hardwired) cluster induced by v , 252
$\mathcal{T}(N)$	phylogenetic trees displayed by N , 254
N_X	a binary network that displays each tree in $RB(X)$, 255
$\text{lca}_N(Y)$	lowest stable ancestor of Y in N , 263
$r(N)$	reticulation number of N , 265
$r(\mathcal{P})$	reticulation number of a set \mathcal{P} of trees, 266
$r(T_1, T_2)$	reticulation number of the pair of trees T_1, T_2 , 266

Chapter 1

Phylogeny

“I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics, for men thus endowed seem to have an extra sense” – Charles Darwin [95]

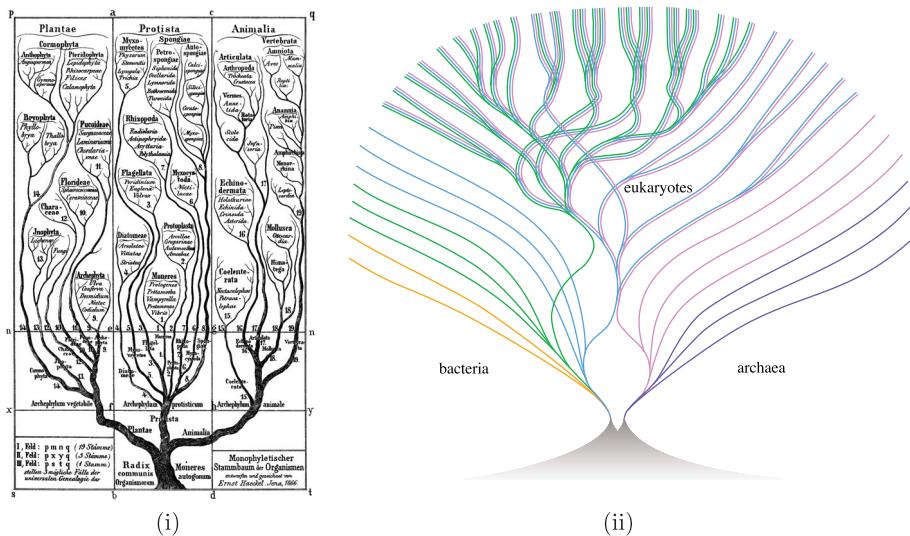


Figure 1.1. (i) Early depiction of a “tree of life” by Ernst Haeckel, 1866, in which plants and animals dominated two of the three main branches. (ii) A modern depiction of the network of life in which plants and animals are relegated to a small corner of one of the three domains of life (eukaryotes), and with some reticulate transfer events that complicate a single tree like history for life. Reprinted with permission from William Martin and John Wiley and Sons [240].

1.1 • What is phylogenetics?

All living organisms on earth harbor a signature of their evolutionary heritage within their DNA. By studying patterns and differences between the genetic makeup of different species, molecular biologists are able to piece together parts of the story of how life

today traces back to a common origin (as illustrated in Fig. 1.1). In this way, many basic questions can be answered. When did animals and plants diverge? Are fungi more closely related to plants or animals? How and when did photosynthesis arise? What is the closest living relative to the whales? Does speciation occur in bursts or at a steady rate? Other topics are proving more difficult to resolve—for example, deciphering the earliest history of life on earth.

Similar questions arise for evolutionary processes in other fields such as epidemiology (e.g., the relationship between different strains of influenza or HIV), linguistics (e.g., how languages diverged from one another over time), stemmatology (e.g., how ancient manuscripts are related to each other), and medicine (e.g., studying the “tree of cells” in the evolution of a tumor). In all these fields, the analysis relies on an underlying mathematical theory, grounded in combinatorics, algebra, and stochastic processes, with the concept of an evolutionary tree as a unifying object.

In this book, I describe a cross section of some of the key concepts in “phylogenetics,” which is the theory of reconstructing and analyzing trees or more complex networks from data observed in the present. In Chapter 2, we will learn about some combinatorial features of phylogenetic trees, namely their encoding by set systems, their enumeration, properties of tree shape, and metrics on trees. In Chapter 3, we will study the discrete properties of phylogenetic trees that arise under random models of evolution. In Chapter 4, we explore how trees can be built from small building blocks (subtrees), and in Chapter 5, we study the way in which trees can display discrete data “perfectly” and focus on tree reconstruction from data (discrete or distance-based), which may not perfectly fit a tree. Chapter 6 allows edges of a tree to have lengths, a seemingly minor embellishment which, in fact, leads to enhanced theory for tree reconstruction, geometric modeling, and biodiversity conservation.

In Chapter 7, we turn to stochastic models that allow data to evolve along the branches of the tree. We will see how tree reconstruction is possible from this evolved data for certain models (but not for others) and how simple reconstruction methods can be misled in various “zones” of statistical inconsistency. Chapter 8 provides a more in-depth look into the world of Markov processes on trees, through the eyes of information theory and algebra.

In Chapter 9, we study speciation and extinction models for phylogenies, and the predictions these make regarding the “shape” of evolutionary trees, as well as the expected loss of biodiversity under simple extinction scenarios. We also investigate random processes under which gene trees and species trees can disagree (and how the latter can be inferred from the former).

Finally, in Chapter 10, we turn to phylogenetic networks, which are able to represent data in a way that is not strictly tree like, due to reticulate processes such as the formation of hybrid species, lateral gene transfer, and endosymbiosis events.

We start with some basic notation and conventions that are used throughout the book, along with some background on graphs and trees. This is followed by an introduction to phylogenetic trees, first results, and some definitions of the main classes of phylogenies.

1.2 ■ Preliminaries

1.2.1 ■ Generic notation

Throughout, X will denote a finite set of size n , $[n]$ is shorthand for the set $\{1, \dots, n\}$, 2^S denotes the power set of S , and $\binom{S}{k}$ refers to the collection of subsets of S of size k . For a subset A of X , we let \overline{A} ($= X - A$) denote the complement of A .

We will denote the size of a set S by either $|S|$ or $\#S$, sometimes even writing $\#x : f(x)$ in place of $|\{x : f(x)\}|$. Occasionally, we use \coloneqq to emphasize that we are defining a quantity.

A *partition* Π of a set S is a collection of disjoint subsets of S (called *blocks*) whose union is S . A partition Π of S *refines* another partition Π' of S if every block of Π' is a union of blocks of Π (equivalently, every block of Π is contained in a block of Π').

For an event A we will let \mathbb{I}_A be the indicator variable that takes the value 1 when A occurs and 0 otherwise.

We will assume that the reader is familiar with the following notions: a partially ordered set (poset); a metric space; equivalence relations on a set; the order notation O , Θ , and Ω ; and asymptotic equivalence \sim . We also assume familiarity with basic notions of probability theory, algebra (groups and linear algebra), elementary differential equations, and graph theory. For a few sections of Chapters 6 and 7 some familiarity with notions from topology (e.g., homotopy, simplicial complexes) and algebraic geometry (ideals, varieties) is helpful but not essential.

Since graph theory will play a particularly prominent role, we will now review some of the main concepts and definitions required.

1.2.2 • Graphs and trees

An (undirected) *graph* $G = (V, E)$ consists of a set V of vertices and a set $E \subseteq \binom{V}{2}$ of edges. We will assume that all graphs in this book are finite (i.e., where $|V|$ and thus $|E|$ are finite). If $\{v, v'\} \in E$, then v and v' are said to be *adjacent*, while if $v \in e \in E$, then v and e are *incident*. Under this definition, a graph has no “parallel edges” (two or more edges containing the same pair of vertices) or “loops” (an edge from a vertex to itself). Any nonempty subset U of V defines an (induced) subgraph, namely $G[U] = (U, E')$, where $E' = \{\{u, v\} \in E : u, v \in U\}$.

The *degree* of a vertex v is the number of edges that are incident with v and is denoted $d(v)$. The “handshake lemma” states that

$$\sum_{v \in V} d(v) = 2|E|. \quad (1.1)$$

A *clique* in G is a subset of vertices for which each pair of vertices is adjacent. A *path* in G is a sequence v_1, \dots, v_k of distinct vertices, where each vertex is adjacent to the next one in the series. If, in addition, $k \geq 3$ and v_1 and v_k are adjacent, then the path forms a *cycle* of length k (a “ k -cycle”). The *diameter* of a graph $G = (V, E)$ is

$$\Delta[G] = \max\{d(u, v) : u, v \in V\},$$

where $d(u, v)$ is the length (number of edges) of the shortest path in G connecting u and v , in other words, the longest path needed (in a shortest path) to connect any two vertices of G . A graph G is *bipartite* if its vertex set can be partitioned into two nonempty subsets U and W so that all edges of G consist of a vertex from each subset. It is a classical result that a graph is bipartite if and only if it does not contain any cycles of odd length. This, in turn, is equivalent to the graph having a proper 2-coloring of its vertices (i.e., an assignment of one of two colors to each vertex so that no two adjacent vertices receive the same color). Since the existence of such a coloring can be determined in polynomial time, deciding whether or not a graph is bipartite is also easy. More generally, a graph is *multipartite* (or k -partite) if its vertex set can be partitioned into k subsets so that each edge of G consists of a pair of vertices from different sets.

Two graphs $G = (V, E)$ and $G' = (V', E')$ are regarded as equivalent if there is a *graph isomorphism* between them; in other words, a bijection $\varphi : V \rightarrow V'$ that satisfies the condition $\{u, v\} \in E \Leftrightarrow \{\varphi(u), \varphi(v)\} \in E'$.

Trees. A *tree* is a graph that is connected and has no cycles. There are many alternative characterizations of trees; for example, the following are equivalent:

- $T = (V, E)$ is a tree;
- $T = (V, E)$ is connected and it has $|V| - 1$ edges;
- for any two vertices u and v of T , there is a unique path from u to v .

We will denote this unique path connecting u and v in a tree T by $P(T; u, v)$. Note that we often treat $P(T; u, v)$ as the set of edges in the path (for example, by writing $e \in P(T; u, v)$). Trees will play a central and important role in this book, so it is helpful to define some basic concepts and discuss how they relate to graphs further.

Every tree T that has more than one vertex always has at least two vertices of degree one. Such vertices are called *leaves*, whereas the remaining vertices are said to be *interior* vertices. Similarly, an edge that is incident with a leaf is said to be a *pendant* edge; otherwise, it is an *interior* edge. Every tree with three or more vertices has an interior vertex. A *star tree* is a tree with a single interior vertex that is adjacent to all the leaves.

Lemma 1.1. *Any tree T is either a star tree or it has two interior vertices v_1 and v_2 , each of which is adjacent to exactly one nonleaf vertex.*

Proof: If T is not a star tree, then consider the longest path P in T consisting of interior vertices, and take v_1 and v_2 to be the endpoints of this path. Then v_1 and v_2 satisfy the property claimed, for otherwise P could be made longer.

A *cherry* in a tree is a pair of leaves that are adjacent to a common interior vertex (this vertex may also be adjacent to other leaves).

Exercise: Using Lemma 1.1, show that if a tree T has five or more vertices and no vertex of degree 2, then T has at least two disjoint pairs of leaves that form cherries of T .

Lemma 1.1 provides the basis for several induction arguments in phylogenetics, which typically proceed as follows. To prove a result about trees with n leaves (and which holds for star trees), select a vertex v_i of the type described, and delete the leaves that are adjacent to v_i to obtain a tree with fewer leaves than T , to which the induction hypothesis can be applied and from which (hopefully) this hypothesis can be extended to hold for T also.

A collection of trees is said to form a *forest*. If v is an interior vertex of T , we will let $T - v$ denote the graph obtained from T by deleting v and its incident edges. Similarly, if e is an edge of T , then $T - e$ is the graph obtained from T by deleting e . Notice that $T - v$ is a forest with at least two components, while $T - e$ is a forest with exactly two components. Often, we will talk about *subdividing an edge* $e = \{u, v\}$ of a tree; this means replacing $\{u, v\}$ by two or more edges $\{u, w_1\}, \{w_1, w_2\}, \dots, \{w_k, v\}$, where w_1, \dots, w_k ($k \geq 1$) are new vertices, and therefore have degree 2 in the resulting tree. The reverse of this process (replacing any path of vertices of degree 2 by a single edge) is called *suppressing vertices of degree 2*.

Trees enjoy certain combinatorial properties. One is the following *Helly property*: If $V_1, V_2, \dots, V_k \subseteq V$ comprise the vertex sets of a family of subtrees of a tree T , and these

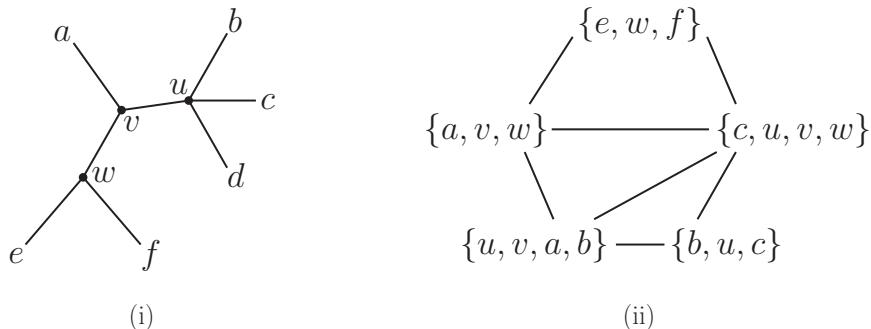


Figure 1.2. (i) A tree T with six leaves and three interior vertices. Vertex v is the median vertex of six triples of leaves, for example, $v = \text{med}_T(a, c, e)$. Two examples of cherries are $\{b, c\}$ and $\{e, f\}$. (ii) The intersection graph of a collection of subtrees of T (indicated by the vertices in the subtree) is necessarily a chordal graph.

sets satisfy the property that $V_i \cap V_j \neq \emptyset$, then $\bigcap_{i=1}^k V_i$ is also nonempty. To see this, first notice that if $V_i \cap V_j \neq \emptyset$, then $V_i \cap V_j$ is the set of vertices of a subtree of T . One can then show that the Helly property holds for $k = 3$ and then extend to larger values of k by induction. One consequence of the Helly property is that, given a tree T , if one takes any three vertices u, v , and w and considers the intersection in T of the three paths $P(T; u, v)$, $P(T; u, w)$, and $P(T; v, w)$, then this intersection must be nonempty. Moreover, in this case the intersection consists of just a single vertex, called the *median* of u, v , and w , denoted $\text{med}_T(u, v, w)$. It is easy to see that each interior vertex of T is the median of some triple of leaf vertices (cf. Fig. 1.2(i)).

Exercise: Show that if $\mathcal{P} = \{T_1, T_2, \dots, T_k\}$ is a collection of subtrees of T , and each tree in \mathcal{P} contains strictly more than half the leaves of T , then there is a vertex common to all trees in \mathcal{P} .

Given a finite graph $G = (V, E)$ consisting of c_G connected components, the *cyclo-matic number* of G , denoted $\text{cy}(G)$, is defined by

$$\text{cy}(G) = |E| - |V| + c_G.$$

Notice that $\text{cy}(G) = 0$ if and only if G is a forest (i.e., a tree or a disjoint union of two or more trees), while, in general, $\text{cy}(G)$ is the minimum number of edges that need to be removed from G in order to make it acyclic, so it is sometimes called the circuit rank of G (if one views G as a one-dimensional simplicial complex, then $\text{cy}(G)$ is the rank of the first homology group of this complex).

Chordal graphs. A graph is *chordal* if every cycle of length four or more has a chord (an edge between nonconsecutive vertices of the cycle). Thus a chordal graph either has no cycles (i.e., it is a forest) or its cycles can be broken up into 3-cycles. Chordal graphs are relevant to phylogenetics by virtue of an alternative (and at first sight, slightly surprising) characterization of the class of chordal graphs: they are precisely the graphs that are the intersection graphs of a family of subtrees of a tree. In other words, G is chordal if and only if for some tree T , G is isomorphic to a graph (\mathcal{T}, E) , where \mathcal{T} is a collection of

subtrees of T , and the edges in E are between any two trees in \mathcal{T} that share at least one vertex. An example is shown in Fig. 1.2(ii).¹

Chordal graphs provide an easy way to define an important integer associated with any finite graph G , namely its “treewidth.” This can be defined in terms of a certain way of describing G by a tree of subsets of V ; however, an equivalent direct definition is the following. For a graph G' , let $\omega(G')$ be the number of vertices in the largest clique in G' . Then the *treewidth* of G $\text{tw}(G)$ is the minimal value of $\omega(G') - 1$ over all chordal graphs G' obtained from G by adding zero or more edges. For example, since a tree is (trivially) a chordal graph, it follows that a connected graph G is a tree if and only if $\text{tw}(G) = 1$. A main importance of treewidth is that many computationally intractable (NP-hard) problems on graphs become solvable in polynomial (often linear) time on trees that have bounded treewidth.

1.2.3 • Directed graphs and rooted trees

A *directed graph* or, more briefly, *digraph* $G = (V, A)$, consists of a set V of vertices and a set of directed edges $A \subseteq V \times V$. A directed edge (u, v) is said to start at u and end at v (u is the “tail” and v the “head” of e). The *out-degree* and *in-degree* of any vertex v is the number of edges that start at v and end at v , respectively, and these values are denoted $d^+(v)$ and $d^-(v)$, respectively. The analogue of the handshaking lemma (eqn. (1.1)) for directed graphs is

$$\sum_{v \in V} d^-(v) = \sum_{v \in V} d^+(v) = |A|.$$

A *directed acyclic graph* (or DAG) is a directed graph that is *acyclic*; in other words, there is no directed path (a sequence of edges leading from one vertex to another) from any vertex to itself. In any DAG, a *proper descendant* of a vertex v is any vertex v' that can be reached by following a directed path from v , in which case, v is said to be an *ancestor* of v' .

The following lemma summarizes two fundamental properties of digraphs.

Lemma 1.2. *Let G be a finite digraph.*

- (i) *If G is acyclic, then it has at least one vertex of out-degree zero and at least one vertex of in-degree zero.*
- (ii) *G is acyclic if and only if there exists a total ordering on the vertices of G that satisfies the following condition: if v' is a proper descendant of v , then v is strictly less than v' in the ordering.*

A *rooted tree* is a tree $T = (V, E)$, where a vertex $v \in V$ is distinguished as a root vertex and all edges are directed away from this vertex. If, in addition, each nonleaf vertex of T has out-degree 2, we say that T is a *binary tree*.

Notice that for any rooted tree T , each edge is directed, and there is also a natural partial order \preceq_T on the vertices of T , defined by $u \preceq_T v$ if $u = v$ or if there is a directed path from u to v in T . Given a rooted tree $T = (V, E)$ and a nonempty subset A of V , there is a (unique) vertex v for which $v \preceq_T a$ for all $a \in A$ and which is maximal (under \preceq_T) with respect to this property;² this vertex v is called the *least common ancestor* (LCA) of A in T and is denoted $\text{lca}_T(A)$. For two vertices x, y , we will write $\text{lca}_T(x, y)$ in place of $\text{lca}_T(\{x, y\})$.

¹Chordal graphs have a further quite different characterization as graphs G that have a “perfect elimination ordering,” meaning that the vertices can be ordered v_1, v_2, \dots, v_n so that $\{v_1, v_2, \dots, v_i\}$ is a clique of G for all i .

²In other words, if $w \preceq_T a$ for all $a \in A$, then $w \preceq_T v$.

1.2.4 • Symmetries and “centers” of trees

Let $T = (V, E)$ be a rooted tree. An *automorphism* of T is a graph isomorphism from T to itself (i.e., a bijective map $\phi : V \rightarrow V$ for which $(u, v) \in E \Rightarrow (\phi(u), \phi(v)) \in E$). It is a classical result that an automorphism of a tree is fully determined by its action on the leaves of the tree (i.e., if the automorphisms ϕ and ϕ' agree on the leaf set of T , then $\phi = \phi'$). This can be seen by noting that every nonleaf vertex v of a tree T can be written as $v = \text{lca}_T(\{x, x'\})$ for some pair of leaves x, x' , and so for any automorphism ϕ of T , $\phi(v) = \text{lca}_T(\{\phi(x), \phi(x')\})$. Let $S(T)$ be the group of automorphisms of a rooted tree T under composition; we will call this the *symmetry group* of T . For any rooted tree, there is a procedure for calculating the order of its symmetry group. Mostly, we will deal with the case where T is a rooted binary tree, in which case $|S(T)| = 2^{s(T)}$, where $s(T)$ is the number of vertices of T where the two descendant subtrees are isomorphic.

What about the symmetry group of unrooted trees? First, the notion of automorphism carries over directly to unrooted trees. As before, any automorphism is determined by its action on the leaves of the tree (since each interior vertex v can be written as a median of three leaves). For a rooted tree, there is always a vertex that is fixed by every automorphism, namely the root vertex. For an unrooted tree, this is no longer the case (consider, for example, the tree on two vertices u and v , for which the nonidentity automorphism interchanges u and v). Despite this, something almost as strong holds: every tree has either a vertex or an edge that is fixed by every automorphism. One way to see this is based on a classical concept, the “centroid” of a tree, which we now explain.

Given an unrooted tree $T = (V, E)$ and a vertex $v \in V$, consider the graph obtained from T by deleting v and its incident edges, and let $\omega(v)$ be the maximum number of vertices in any component of this graph. A vertex v of T is called a *centroid* of T if $\omega(v)$ is minimal across all vertices of T . A result due to C. Jordan from 1869 states that any tree either has a unique centroid or two adjacent centroids [40]. Moreover, it is clear that if T has a unique centroid, then this is fixed by all automorphisms of T , whereas if T has two adjacent centroids, these are either fixed or interchanged by any given automorphism of T .

The symmetry group of T is now easily described: it is simply the symmetry group of the rooted tree obtained from T by selecting the centroid vertex (as the root) when this is unique or, if there are two centroids, subdividing the edge connecting them and making this new root of degree 2 the root. For example, the trees in Figs. 1.3 (a), (b), and (c) have symmetry groups of orders $7! \times 2!$, $2 \times (2!)^2$, and $3! \times (2!)^3$, respectively.

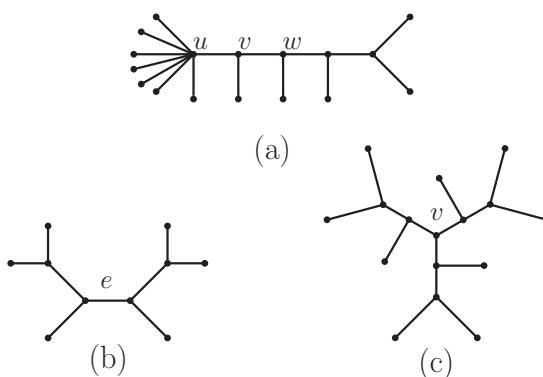


Figure 1.3. (a) A tree with the leaf-centroid (u), centroid (v), and center (w) marked; (b) a tree with a central symmetry edge e ; (c) a tree with a central symmetry vertex v .

Instead of the centroid, there is an alternative notion we could have used in describing $S(T)$. The *center* of a tree T is the set of vertices v in T for which the maximum distance (number of edges) from v to any other vertex of T is as small as possible. The center shares the same two properties as the centroid: it consists either of a single vertex (and this is fixed by all automorphisms of T) or two adjacent vertices (and these are either fixed or interchanged by any given automorphism of T) [40].

The center and centroid of a tree may differ; an example of this is shown in Fig. 1.3(a). If T has a *central symmetry vertex* v (i.e., the rooted tree components of the graph $T - v$ are isomorphic), then v is the unique center and centroid vertex of T . If T has a *central symmetry edge* $e = \{u, v\}$ (i.e., the two rooted tree components of the graph $T - e$ are isomorphic), then $\{u, v\}$ is both the center and centroid of T . It follows that rooting T using the center in the same way as we did for the centroid would have also described the same symmetry group of each unrooted tree T . These notions are also illustrated in Fig. 1.3.

Both the center and the centroid of a tree T can be easily computed in linear time in the number of vertices of the tree. The center is the midpoint(s) of any longest path in T , whereas the centroid of any tree has alternative characterizations. For example, a vertex is a centroid vertex of a tree T if and only if each component of the graph $T - v$ has at most half of the vertices of T [208] (this paper also showed that the centroid vertices of a tree are precisely the vertices that have a minimal average distance to all the other vertices of the tree). A further characterization of the centroid of a tree was established in [259].

There is yet one more notion of “center” that applies whenever a tree has no vertices of degree 2 (most of the trees we talk about in this book have this property). The *leaf-centroid* of T is the set of vertices v of T with the property that each of the components of $T - v$ contains at most half of the leaves of T . The leaf-centroid can be different from the (ordinary) centroid, even for a tree with no vertices of degree 2 (an example is shown in Fig. 1.3(a)). For trees in which all interior vertices are of degree 3, the leaf-centroid is the same as the centroid, by the first of the alternative characterizations of the centroid described above. For trees without vertices of degree 2, the leaf-centroid has a similar property to the ordinary centroid (and coincides with it on this class of trees when the tree has a central symmetry vertex or edge).

Proposition 1.3. *Let T be a tree with at least two leaves and no vertices of degree 2. The leaf-centroid of T consists of either a single vertex or two adjacent vertices.*

Proof. We can use Lemma 1.2. Convert any edge $e = \{u, v\}$ into a directed edge (u, v) if there are more leaves of T on the u side of this edge than there are on the v side of this edge (i.e., in the disconnected graph $T - e = (V, E - \{e\})$, the component containing u has more leaves of T than the component containing v). It is possible that some edge e of T is not given a direction; however, this can only occur when exactly half of the leaves of T are on either side of e , so, at most, one edge of T can remain undirected, and in that case the leaf-centroid consists of the endpoints of that edge. In the other case, all edges become directed and T becomes an acyclic directed graph. Therefore, by Lemma 1.2, there is a vertex v of out-degree 0 (which cannot be a leaf). This means that there is an interior vertex v of T for which each of the subtrees adjacent to v has less than $n/2$ leaves of T . Moreover, it is easily seen that there can be, at most, one such vertex. ■

The leaf-centroid is applied in Proposition 1.5 of the next section and at one point in Chapter 2.

1.3 ■ Phylogenetic trees

We first define a *rooted phylogenetic X-tree* to be a tree T with directed edges in which

- X is the set of leaves (vertices of out-degree 0),
- all the edges are directed away from some root vertex ρ ,
- every nonleaf (interior) vertex has out-degree at least 2.

The set X here may be a set of biological “species” (or, more generally “taxa”); however, in other applications of phylogenetics, X might be a set of languages (for which T describes their evolutionary history), strains of a virus (HIV, influenza virus, etc.), or a set of extant cells in a tumor (for which T describes how these developed from a single mutant cell).

Figure 1.4(a) shows a simple biological example of a rooted phylogenetic X -tree for a set X of five species. This reveals one relationship that is perhaps surprising to most non-biologists: the genetic data indicate that fungi are more closely related genealogically to animals than to plants. The interior vertices of a phylogenetic tree represent hypothetical ancestral species, with the root ρ being the “most recent common ancestor” of the species at the leaves.

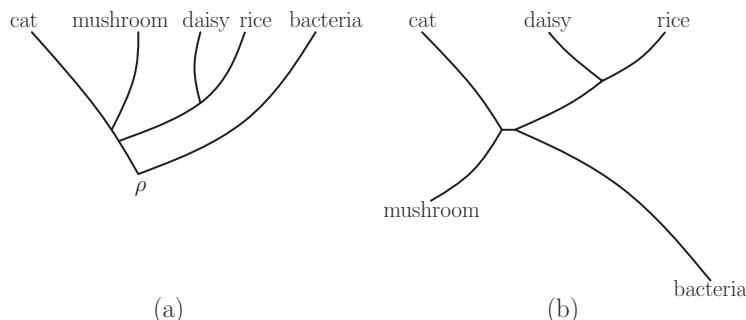


Figure 1.4. (a) A rooted phylogenetic tree with root ρ . (b) The associated unrooted phylogenetic tree obtained by ignoring (suppressing) the root vertex of the tree in (a).

We will think of two rooted phylogenetic X -trees as being *equivalent* if they are isomorphic as rooted trees by an isomorphism that is the identity map on X (i.e., the trees are equivalent up to relabeling of the nonleaf vertices). When T and T' are equivalent, we will denote this by writing $T \cong T'$. To the biologist, equivalent rooted trees tell the same story of how lineages split into the species we observe today, even though the trees may have different timing events for these splittings or be drawn in the plane in different ways. The most informative rooted binary trees (i.e., the trees that display the largest number of branching events) are the ones for which every interior vertex has out-degree exactly 2 (these are the *rooted binary phylogenetic X-trees*).

A convenient way to represent rooted phylogenetic trees is by representing the branching structure using nested parentheses and commas (referred to as “Newick format”). For example, the tree in Fig. 1.4(a) can be written as (((cat, mushroom), (daisy, rice)), bacteria). In the case of three species, we will often write $xy|z$ in place of $((x, y), z)$. Notice that any rooted phylogenetic tree always contains a cherry (two leaves adjacent to a common vertex).

Rooted trees appeal to biologists, since they show evolution happening through time, from the past to the present. However, it is often more convenient to consider unrooted

trees. One reason is that most methods for building trees from data can usually do so only up to the placement of the root, and thus produce unrooted trees (figuring out where the root goes usually comes later).³ The choice to work with either rooted or unrooted trees is somewhat analogous to the distinction in classical geometry between the affine and projective settings (respectively), where one viewpoint may have advantages over the other, depending on the questions at hand.

Definition: An (*unrooted*) *phylogenetic X-tree* is a tree T with leaf set X and where every interior (i.e., nonleaf) vertex is of degree, at least, 3. If the degree of every nonleaf vertex is exactly three, we say that T is a *binary phylogenetic X-tree*. Some mathematical authors have used the word “trivalent” or “ternary” rather than binary; however, “binary” is more standard in biology, and we will see that unrooted binary phylogenetic trees correspond naturally to rooted binary trees (a further term for binary trees, popular in biology, is “fully resolved” trees). Figure 1.4(b) shows the unrooted binary phylogenetic tree obtained from the rooted tree in part (a), by ignoring the root vertex.

Analogous to the rooted case, we will think of two (unrooted) phylogenetic X -trees T and T' as being equivalent, denoted $T \cong T'$, if they are isomorphic as (unrooted) trees by an isomorphism that is the identity map on X (i.e., trees are equivalent up to relabeling of the nonleaf vertices). The concepts of pendant and exterior edges carry over into the unrooted setting directly.

Here we describe some basic properties of binary phylogenetic trees, which will be useful in future chapters.

Lemma 1.4. *Let T be an unrooted binary phylogenetic tree with $n \geq 2$ leaves.*

- (i) *If T is binary, it has $2n - 3$ edges and $n - 2$ interior vertices.*
- (ii) *If T is not binary, it has fewer than $2n - 3$ edges and fewer than $n - 2$ interior edges.*

For rooted phylogenetic trees, the same results apply, except with $2n - 3$ and $n - 2$ being replaced by $2n - 2$ and $n - 1$, respectively.

Proof: As noted earlier, a connected graph is a tree if and only if the number $|V|$ of vertices exceeds the number $|E|$ of edges by 1. Therefore, if a tree T has i interior vertices, we have $|V| = i + n$ and so

$$|E| = i + n - 1. \quad (1.2)$$

Also, for any graph, the handshaking lemma tells us that the sum of the degrees of the vertices of any finite graph equals $2|E|$, since each edge is counted twice in this sum. Now, for part (i), where T is binary, the sum of the degrees is $[1 + 1 + \dots + 1(n \text{ times})] + [3 + 3 + \dots + 3(i \text{ times})]$, so

$$2|E| = n + 3i. \quad (1.3)$$

Combining eqns. (1.2) and (1.3), we see that $i = n - 2$, and so $|V| = i + n = 2n - 2$, which implies that $|E| = |V| - 1 = 2n - 3$ and $i = n - 2$. This completes the proof of Lemma 1.4, part (i). For part (ii), where T has at least one interior vertex of degree greater than 3, the sum of the degrees of the vertices is strictly greater than $n + 3i$, so $2|E| > n + 2i$, from which we obtain $i < n - 2$ and $|E| < 2n - 3$. The analogous results for rooted phylogenies follow by similar arguments. ■

The following result is particular to (unrooted) binary phylogenies.

³In addition, from a mathematical perspective, unrooted trees are often the more natural object to consider.

Proposition 1.5. *Any unrooted binary phylogenetic tree T with two or more leaves has an edge e for which each of the two components of $T - e$ contains at least one-third of the leaves of T .*

Proof: Select a vertex v that is a leaf-centroid for T and then select the edge of T that is incident with v for which the component of $T - e$ that does not contain v has the largest number (say, k) of the n leaves of T . Then $k \geq \frac{1}{3}n$; otherwise, the total number of leaves of T would be less than n . Moreover, since $k \leq \frac{1}{2}n$ (since v is a leaf-centroid vertex), the number of leaves in the component of $T - e$ that contains v is at least $\frac{1}{2}n$. ■

A binary phylogenetic tree (rooted or unrooted) is called a *caterpillar tree* if the interior vertices form a path (in the rooted case, the root is required to be at one end of this path). Examples are shown in Fig. 1.6 ((c) and (c)').

Exercise: For $n \geq 4$ show that, up to equivalence, the number of binary phylogenetic trees on $[n]$ that are caterpillars is $n!/8$ (for unrooted trees) and $n!/2$ (for rooted trees).

A further class of binary phylogenies is the set of *perfect trees*. In the rooted setting, these are rooted binary trees for which each leaf is the same number of edges h distant from the root, so the number of leaves is 2^h . In the unrooted setting, a perfect tree is a binary phylogenetic tree in which either (i) each leaf is the same number of edges from some fixed vertex v , or (ii) each leaf is the same number of edges from some fixed edge e . In case (i), $n = 3 \cdot 2^h$ for some h , and T consists of three disjoint perfect rooted phylogenies with 2^h leaves, the roots of which are adjacent to v . In case (ii), $n = 2^h$ for some h , and T consists of two disjoint perfect rooted phylogenies, the roots of which are joined by an edge. Perfect trees are sometimes also called “(complete) balanced” trees in the literature. These concepts are illustrated in Fig. 1.5. A perfect rooted phylogeny on four leaves is called a *fork tree*. Note that a binary tree (rooted or unrooted) is perfect if and only if its symmetry group is transitive on the leaves (i.e., for any two leaves x and y of the tree, there is an automorphism of the tree that sends x to y).

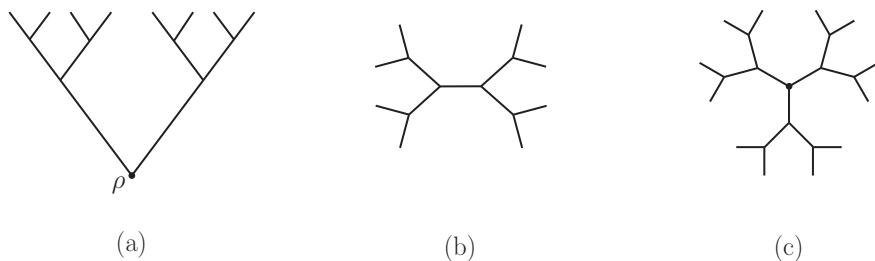


Figure 1.5. (a) A perfect rooted binary tree on eight leaves; perfect unrooted binary trees on (b) eight leaves and (c) on 12 leaves. The trees in (a), (b) and (c) have four, four, and six cherries, respectively. Leaf labels are not shown.

For a (rooted or unrooted) phylogenetic tree, let $c(T)$ denote the number of cherries in T (recall that a cherry is a pair of leaves that are adjacent to a common interior vertex). Caterpillar trees are precisely the binary trees with the fewest cherries ($c(T) = 1$ in the case of rooted trees and $c(T) = 2$ in the case of unrooted trees), whereas the maximal

value of $c(T)$ among n -leaf phylogenies is $\lfloor n/2 \rfloor$ and is attained by a variety of tree shapes, including (for n a power of 2) the perfect rooted trees.

X -trees. A slightly more general notion than a phylogenetic X -tree is that of an X -*tree*. This consists of a triple (V, E, ϕ) , where (V, E) is a tree and $\phi : X \rightarrow V$ has the property that for every vertex $v \in V$ of degree 1 or 2, there is some $x \in X$ with $\phi(x) = v$.⁴ Any phylogenetic X -tree T can be viewed as an X -tree for which ϕ maps X bijectively onto the set of leaves of the tree. More generally, X -trees allow several elements of X to be mapped to the same vertex, and T to have one or more vertices of degree 2, provided that each such vertex is in the range of ϕ .

Exercise: Show that for every phylogeny $T \in P(n)$ and every interior vertex v of T , there is a leaf x of T that lies within $\lfloor \log_2(\frac{n}{3}) \rfloor + 1$ edges from v .

1.3.1 • Key phylogenetic notation

Following [315], we will let $P(X)$ and $RP(X)$ denote the sets of unrooted and rooted phylogenetic X -trees (up to equivalence), while $B(X)$ and $RB(X)$ will denote the sets of unrooted and rooted binary phylogenetic X -trees. Thus when X has just four elements, $B(X)$ consists of the three *quartet trees* (for $X = \{i, j, k, l\}$ shown in Fig. 1.6(a)', and written $ij|kl$, and the other two are $ik|jl$ and $il|jk$), and $P(X)$ has one additional “star” tree that has a single nonleaf vertex of degree 4. When $X = [n] = \{1, \dots, n\}$, we will write $RP(n), P(n), RB(n)$, and $B(n)$ for $RP(X), P(X), RB(X)$, and $B(X)$, respectively, and we will let $rb(n) = |RB(n)|$ and $b(n) = |B(n)|$.

We will also often refer to a phylogenetic X -tree T as a *phylogeny* on X . When the leaf set of T is clear we will also refer to T as simply a phylogenetic tree or a phylogeny. For technical reasons, when $|X| = 1$ we take $RB(\{x\})$ and $B(\{x\})$ to consist of just a single isolated vertex labeled x . We denote the set of vertices and edges of T by $V(T)$ and $E(T)$ respectively, and the set of interior vertices and edges by $\overset{\circ}{V}(T)$ and $\overset{\circ}{E}(T)$, respectively.

We will use T for a (given) phylogenetic tree, \mathcal{T} for a randomly generated phylogenetic tree, \mathcal{P} for a sequence or set of trees, and \mathcal{C} for a sequence or set of subsets of X or functions on X .

Figure 1.6 displays some of the tree types that will appear in various places throughout the book.

⁴Two X -trees (V, E, ϕ) and (V', E', ϕ') are regarded as equivalent if there is a graph isomorphism φ from (V, E) to (V', E') for which $\phi'(x) = \varphi(\phi(x))$ for all $x \in X$.

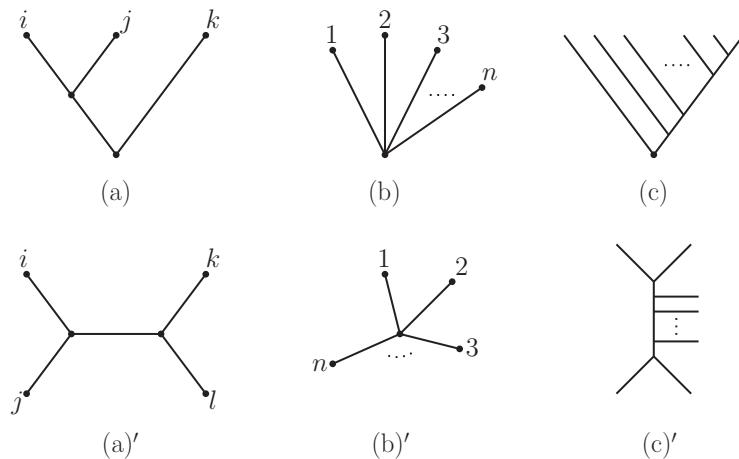


Figure 1.6. A selection of trees: a rooted triplet (a) and a quartet tree (a)' can be viewed as the basic building blocks of (rooted and unrooted) phylogenies; star trees (rooted (b) and unrooted (b)'; caterpillar trees (rooted (c), and unrooted (c)'), without the leaf labels being shown.

Chapter 2

Basic combinatorics of discrete phylogenies

2.1 • Counting trees

Counting trees has a long tradition in mathematics, with Cayley's n^{n-2} formula from 1889 for the total number of trees on n labeled vertices being the most famous example. Counting binary phylogenetic trees turns out to be easier, and it has a history that dates back to even earlier mathematical work, contemporary with Darwin [309]. To explain this, we first describe a close connection between rooted and unrooted phylogenetic trees. There are two natural ways to associate an unrooted phylogenetic X -tree with a rooted tree.

Adding an outgroup: Take a rooted phylogenetic tree T on $X - \{x\}$ and attach x to the root of T by a new edge to produce an unrooted tree. Species x is called an “outgroup” species.

Suppressing the root: Simply ignore the root vertex ρ of T . If the root has degree 2, then delete it and identify its two incident edges; if the root has degree ≥ 3 , then just treat this vertex as an interior vertex with no special root status. In either case, we let $T^{-\rho}$ denote this unrooted phylogeny.

Notice that the operation “adding an outgroup” provides a bijection

$$o_x : RP(X - \{x\}) \rightarrow P(X)$$

which restricts to a bijection from $RB(X - \{x\})$ to $B(X)$. On the other hand, “suppressing the root” results in a surjective map,

$$s : RP(X) \rightarrow P(X), T \mapsto T^{-\rho}$$

which restricts to a surjective map from $RB(X)$ to $B(X)$. These concepts are illustrated in Fig. 2.1 ((i), (ii), and (iii)).

Notice that the number of elements of $RB(X)$ which map to the same tree in $T \in B(X)$ is the number of edges in T , which is $2n - 3$ (where $n = |X|$), by Lemma 1.4. These observations show us that

$$|RB(X)| = (2n - 3)|B(X)| = (2n - 3) \cdot |RB(X - \{x\})|.$$

In particular, the number $rb(n)$ of rooted binary phylogenetic trees on a leaf set of size n satisfies $rb(n) = (2n - 3)rb(n - 1)$, which, together with $rb(2) = 1$, leads to the following

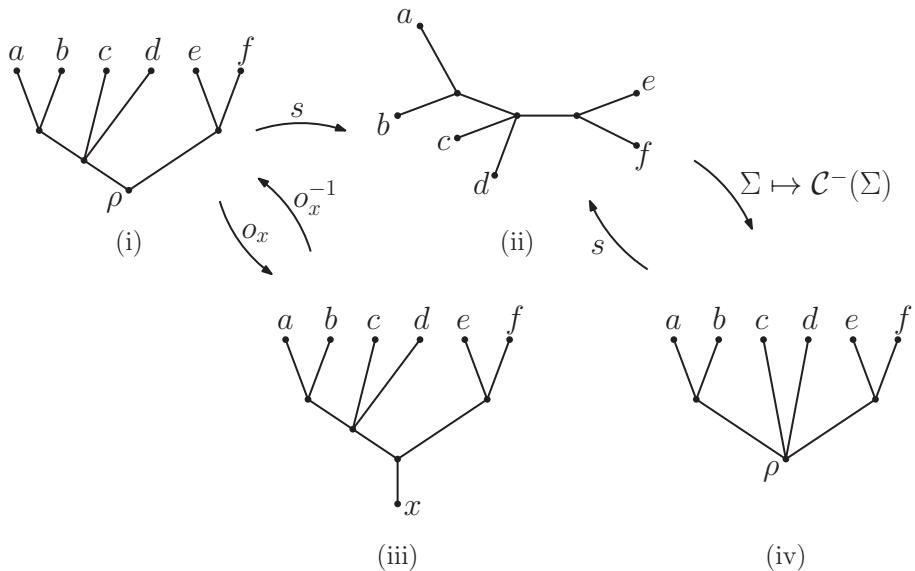


Figure 2.1. (i) A rooted phylogenetic tree on the leaf set $\{a, b, c, d, e, f\}$; (ii) the unrooted phylogenetic trees obtained by suppressing the root vertex of the phylogeny in (i); (iii) the unrooted phylogenetic tree obtained by adding an outgroup leaf x to the phylogeny in (i); (iv) the rooted phylogenetic tree obtained by rooting the phylogeny on (ii) on its leaf-centroid vertex (Proposition 2.5).

expression for all $n \geq 2$:

$$rb(n) = 1 \times 3 \times 5 \times \cdots \times (2n-3). \quad (2.1)$$

This product of the odd numbers is often written as the double factorial (in this case, $(2n-3)!!$). This product can also be expressed in terms of ordinary factorials and powers of 2 as follows:

$$rb(n) = \frac{(2n-2)!}{(n-1)!2^{n-1}}. \quad (2.2)$$

Graph theorists may recognize this quantity as the number of perfect matchings of a complete graph on $2n-2$ vertices. In other words, if there are $2n-2$ people in a room, eqn. (2.2) counts the number of scenarios in which each person shakes hands with precisely one other person. The bijection between this set of scenarios and the set $RB(n)$ is an interesting but nontrivial exercise [127]. For the number $b(n)$ of unrooted binary trees on a leaf set of size n , the bijection o_x described above gives $b(n) = rb(n-1) = (2n-5)!!$.

Applying Stirling's approximation $n! \sim \sqrt{2\pi \cdot n^{n+\frac{1}{2}}} e^{-n}$ to eqn. (2.2) reveals that $rb(n)$ grows very rapidly. For example, $rb(10)$ is around 34 million, while $rb(30)$ is more than 10^{38} . Biologists often want to build trees for hundreds (or even thousands) of species, so it's no surprise that mathematics has an important role to play in this task, as it would be impossible to check each tree to see how well it might "fit the data."

Two asymptotic expressions for $rb(n)$ that follow from Stirling's approximation are

$$rb(n) \sim \frac{1}{\sqrt{2}} \left(\frac{2}{e}\right)^n n^{n-1} \text{ and } \frac{rb(n)}{n!} \sim \frac{1}{2\sqrt{\pi}} 2^n n^{-3/2}. \quad (2.3)$$

The first expression in (2.3) can also be written more loosely as

$$rb(n) = 2^{n \log_2 n + O(n)}.$$

Exercise⁺: Show that the proportion of trees in $B(n)$ that have a centroid consisting of two vertices (i.e., there is an edge that separates $[n]$ into two equally sized sets), is $\frac{1}{2} \binom{n}{n/2} \frac{rb(n/2)^2}{b(n)}$ for n even. Show that this is asymptotically equivalent to $\frac{4}{\sqrt{\pi}} n^{-1/2}$ (which tends to zero, albeit somewhat slowly) by the second part of eqn. (2.3).

There is another way to arrive at eqn. (2.2), namely by using generating functions. Let $\varphi(x) = \sum_{n \geq 1} rb(n) \frac{x^n}{n!} = x + \frac{x^2}{2!} + 3 \frac{x^3}{3!} + 15 \frac{x^4}{4!} + \dots$. Then

$$\varphi(x) = \frac{1}{2} \varphi(x)^2 + x, \quad (2.4)$$

since deleting the root of a tree $T \in RB(n)$ for $n \geq 2$ results in a rooted binary tree (or an isolated leaf) on each of two leaf sets, Y_1 and Y_2 , that partition $[n]$. Solving this quadratic equation gives $\varphi(x) = 1 - \sqrt{1 - 2x}$, from which $rb(n)$ drops out as $n!$ times the coefficient of x^n in the Taylor expansion of $1 - \sqrt{1 - 2x}$. While this is a more complicated derivation, generating functions turn out to be very useful in other applications, such as in deriving exact explicit formulae for the number of “forests” of rooted binary trees on a given leaf set (we give another application in Chapter 10).

For the wider class $RP(n)$ of all rooted phylogenies on $[n]$, there is no known elegant simple expression for $r(n) = |RP(n)|$. However, generating functions again provide a way forward for developing recursive and asymptotic formulae for this quantity, as well as occasional exact results.⁵ For the exponential generating function $R(x) = \sum_{n \geq 1} \frac{r(n)}{n!} x^n = x + \frac{x^2}{2!} + 4 \frac{x^3}{3!} + \dots$, the analogue of eqn. (2.4) follows by a similar argument to give

$$R(x) = x + \sum_{k \geq 2} \frac{R(x)^k}{k!} = x + \exp(R(x)) - 1 - R(x), \quad (2.5)$$

and so $R(x) = \frac{1}{2}(\exp(R(x)) - 1 + x)$. A slight twist also allows us to count the number $r(n, k)$ of trees in $RP(n)$ that have k edges, by letting

$$R(x, y) = \sum_{n \geq 1, k \geq 0} \frac{r(n, k)}{n!} x^n y^k = x + \frac{1}{2!} x^2 y^2 + \frac{1}{3!} (3x^3 y^4 + x^3 y^3) + \dots$$

If we remove the root of $T \in RP(n)$ and this produces $r \geq 2$ rooted subtrees, then the number of edges is reduced by exactly r . Thus, the extension of eqn. (2.5) is

$$R(x, y) = x + \exp(yR(x, y)) - 1 - yR(x, y). \quad (2.6)$$

Putting $y = -1$ in this last equation leads to a curious combinatorial fact. Observe that $n!$ times the coefficient of x^n in $R(x, -1)$ counts the number of rooted phylogenies on $[n]$ with an even number of edges minus the number with an odd number of edges. Moreover, one copy of $R(x, -1)$ conveniently cancels from each side of eqn. (2.6) to give

$$R(x, -1) = -\ln(1 - x) = x + x^2/2 + x^3/3 + \dots$$

This means that there are precisely $n! \times \frac{1}{n} = (n-1)!$ more trees in $RP(n)$ that have an even number of edges than have an odd number. More results on the application of generating

⁵Further background on generating function techniques is provided, for example, in [325].

function techniques to enumerate phylogenies can be found in [143] and the references therein.

We end this section with a further enumerative result. Let $\mathcal{P} = \{T_1, T_2, \dots, T_k\}$ be a set of unrooted binary phylogenetic trees on the leaf sets Y_1, Y_2, \dots, Y_k that partition X , referred to as an *(unrooted) binary phylogenetic forest* on X . It will be helpful later (Chapter 5) to be able to count the number $N(\mathcal{P})$ of trees $T \in B(X)$ that we can form by connecting these trees together (more precisely, $N(\mathcal{P})$ is the number of trees $T \in B(X)$ that contain subdivisions of T_1, \dots, T_k as vertex disjoint subtrees). At first, this number seems difficult to calculate, but there is a remarkably compact formula for it. If $n_i = |Y_i| \geq 2$ and $n = |X|$, then an inductive argument (e.g., [315, Theorem 2.8.3]) shows that

$$N(\mathcal{P}) = \frac{b(n)}{b(n-k+2)} \prod_{i=1}^k (2n_i - 3). \quad (2.7)$$

Exercise: A *rooted phylogenetic forest* on $[n]$ is a set of rooted phylogenetic trees, whose leaf sets partition $[n]$ into one or more blocks. Explain why the number of rooted phylogenetic forests on $[n]$ is exactly twice $|P(n)|$ when $n \geq 2$.

2.2 • Rooted trees as nested sets of clusters

2.2.1 • Hierarchies

The 18th century Swedish taxonomist Carl Linnaeus noticed that much of the living world can be nicely organized into a “hierarchy” in which groups of living organisms are either disjoint or nested [232]. For example, cats and dogs comprise disjoint classes of organisms, but both are subsets of the class of mammals (the usual Linnean classification in biology from the highest level to the lowest comprises the sequence “kingdom, phylum, class, order, family, genus, species,” for which the mnemonic phrase *King Philip came over for green soup* can be helpful).

Formally, a *hierarchy* \mathcal{H} on a finite set X is a collection of nonempty subsets of X with the property that any two elements of \mathcal{H} are either nested (one is contained in the other) or disjoint. It will also be convenient here to require that any hierarchy on X contain the set X and all its singleton subsets. Thus \mathcal{H} forms a hierarchy if it does not contain the empty set and it satisfies the following two properties:

H1: For any two sets $A, B \in \mathcal{H}$, we have $A \cap B \in \{A, B, \emptyset\}$; and

H2: \mathcal{H} contains the entire set X and each singleton set $\{x\}$ for all $x \in X$.

The second condition is harmless: if \mathcal{H} is any collection of sets that satisfies **H1**, we can always add the extra elements mentioned by **H2** without violating **H1**.

There is a fundamental equivalence between hierarchies and rooted phylogenetic trees, which we now explain. Given a rooted phylogeny $T \in RP(X)$, and a vertex v of T , the *cluster* associated with v , denoted $c_T(v)$, is the subset of X that becomes separated from the root upon deletion of v . We will let

$$\mathcal{C}(T) = \{c_T(v) : v \in V(T)\}$$

denote the set of clusters associated with the vertices of T . For the tree in Fig. 2.1(i), the clusters consist of the sets $\{a, b\}$, $\{e, f\}$, and $\{a, b, c, d\}$, together with the *trivial clusters* consisting of X and the singleton sets (in this case $\{a, b, c, d, e, f\}, \{a\}, \{b\}, \{c\}, \{d\}, \{e\}$, and $\{f\}$), which are always present in any rooted phylogenetic X -tree. In biology, clusters are sometimes referred to as “clades” or “monophyletic groups.”

Any collection \mathcal{C} of subsets of X forms a directed graph, sometimes called the *cover digraph* of \mathcal{C} . The vertices of this graph are the elements of \mathcal{C} , and we place a directed edge from $B \in \mathcal{C}$ to $A \in \mathcal{C}$ precisely if B covers A (i.e., $A \subset B$ and there is no set $C \in \mathcal{C}$ with $A \subset C \subset B$). The clusters of any rooted phylogeny $T \in RP(X)$ form a hierarchy on X , and T is isomorphic to the cover digraph of this hierarchy under the map $v \mapsto c_T(v)$. Moreover, every hierarchy can be realized in this way, and nonequivalent phylogenetic trees give rise to different hierarchies. In other words, we have the following bijective correspondence between hierarchies and rooted phylogenetic X -trees (up to equivalence).

Proposition 2.1. *A collection of subsets \mathcal{C} of X is a hierarchy if and only if \mathcal{C} is the set of clusters of some rooted phylogenetic X -tree T . Moreover, T and T' have the same set of clusters if and only if $T \cong T'$.*

Proof: For a tree $T \in RP(X)$ with vertex set V , $\mathcal{C}(T)$ is a hierarchy; otherwise, there would be vertices v and v' and leaves x, y, z with $x \in c_T(v) - c_T(v')$, $y \in c_T(v) \cap c_T(v')$, and $z \in c_T(v') - c_T(v)$. This, in turn, would require two different directed paths in T from the root of T to y , which is impossible, since T is a tree and, any tree has a unique path (directed or undirected) connecting any two given vertices.

Conversely, suppose that \mathcal{C} is a hierarchy and let $T = T(\mathcal{C})$ be the cover digraph of \mathcal{C} , with the minor adjustment that we replace vertex $\{x\}$ by the element x for each $x \in X$. Since $X \in \mathcal{C}$, T has a root vertex of in-degree 0, and there is a directed path from this vertex to every vertex of T , and so T is connected. Also, since $\{x\} \in \mathcal{C}$ for all $x \in X$ (by H2), the set of vertices of T of out-degree 0 is precisely X . Also, if $C_1, C_2 \in \mathcal{C}$ with C_1 strictly contained in C_2 , then, by selecting any element $x \in C_2 - C_1$, there is a path in T from C_2 to x that does not go via C_1 . Thus, each vertex of T that is not in X has out-degree at least 2. It remains to show that T is a tree, which is equivalent to showing that T has no vertex of in-degree greater than 1. Suppose to the contrary that such a vertex $C \in \mathcal{C}$ exists in T . Then if (C', C) and (C'', C) are two directed edges of T that end at C , we can select leaves $x \in C'' - C'$, $y \in C' - C''$, and $z \in C \subseteq C' \cap C''$, from which it would follow that $C' \cap C''$ is neither C' nor C'' nor the empty set, violating condition H1 of a hierarchy. Thus $T = T(\mathcal{C})$ is indeed a rooted phylogenetic X -tree with cluster set \mathcal{C} .

The final part of Proposition 2.1 follows from the fact that the composition of the map $\mathcal{C} \mapsto T(\mathcal{C})$ followed by $T \mapsto \{c_T(v) : v \in V(T)\}$ is the identity map on the set of hierarchies on X , while the composition of these maps in reverse order provides an isomorphism from each tree in $RP(X)$ to itself (which is the identity map on the leaf set of each tree). ■

It is an easy exercise to show that the maximal hierarchies on X correspond to the set $RB(X)$ of binary phylogenies on X . By Lemma 1.4, these trees have $2n - 1$ vertices, where $n = |X|$. Therefore, by Lemma 2.1, this is the size of the largest hierarchy on a set of size n . Biologists typically regard binary trees as more informative than nonbinary ones, since the former show just one lineage splitting off at a time; by contrast, a vertex of out-degree 3 or more represents what biologists call a “polytomy” (usually interpreted

as uncertainty about the order of speciation events, rather than certainty about a sudden speciation event into multiple lineages).

Proposition 2.1 can also be extended from hierarchies to set systems that just satisfy the nesting condition **H1** (such systems are sometimes called “laminar families”). In these cases such set systems can be represented in a natural and bijective way with “rooted X -forests” (i.e., a collection of rooted trees with vertex set V , alongside a function $\phi : X \rightarrow V$ with the property that every unlabeled vertex has degree at least 3). For details, the reader is referred to [124].

Notice also that *any* collection \mathcal{S} of nonempty subsets of X contains a canonical subset

$$\mathcal{H}[\mathcal{S}] := \{A \in \mathcal{S} : A \cap B \in \{\emptyset, A, B\} \text{ for all } B \in \mathcal{S}\}, \quad (2.8)$$

which satisfies the hierarchy property **H1**. This may not be a maximal subset of \mathcal{S} satisfying **H1**, but $\mathcal{H}[\mathcal{S}] = \mathcal{S}$ if and only if \mathcal{S} satisfies **H1**.

2.2.2 ■ First applications

The utility of viewing a rooted phylogenetic tree as a set system (a hierarchy) is illustrated by two questions biologists often face. Suppose we have a collection of different trees that estimate the evolutionary history of the same set of taxa. These trees might have been constructed by comparing genetic data across these species, and different choices of which genetic data to use (e.g., different genes) could have resulted in different tree estimates. In other words, while there might be one underlying and unknown “true” species tree that we wish to infer, the phylogenetic trees constructed from the data will typically be merely imperfect estimates of this tree, since the data evolve randomly, a topic we will discuss in Chapter 6. Other reasons for trees to disagree is that different inference methods applied to the same data may estimate different phylogenies; also, in statistical analysis, trees are often estimated by resampling from the original data many times. So two problems arise:

- How can we compare different phylogenetic X -trees?
- Can we combine different phylogenetic X -trees into some “consensus” tree?

The hierarchy link provides a very simple solution to both questions. First, observe that we can define a distance d between any two rooted phylogenetic X -trees T and T' by taking $d(T, T')$ to be the number of clusters that are present in one but not both of the trees T and T' . This distance is the (rooted version) of the so-called “Robinson–Foulds (RF) metric” and since we will be considering other metrics on trees, we will write it as d_{RF} . It satisfies the triangle inequality and can be computed quickly (in linear time in $n = |X|$).

Turning to the consensus question, given a sequence of rooted phylogenetic X -trees T_1, T_2, \dots, T_k , let \mathcal{H}^* be the collection of subsets of X that are present as clusters in more than half of the corresponding hierarchies. The following lemma shows that \mathcal{H}^* forms the set of clusters of a tree, the so-called *majority rule consensus tree*.

Lemma 2.2. \mathcal{H}^* forms a hierarchy, and so corresponds to a rooted phylogenetic X -tree.

Proof. Let \mathcal{H}_j denote the hierarchy on X corresponding to T_j . Suppose that $C, C' \in \mathcal{H}^*$. By the “pigeonhole principle,” there must be some hierarchy \mathcal{H}_j that contains both C and C' . Consequently, either C and C' are disjoint or one is nested in the other. As this holds for all $C, C' \in \mathcal{H}^*$, condition **H1** holds. Moreover, **H2** holds also, since $\{x\}$ and X are elements of \mathcal{H}_j for every j and every $x \in X$. Thus \mathcal{H}^* forms a hierarchy and

so (by Proposition 2.1) corresponds to a rooted phylogenetic X -tree that is unique up to equivalence. ■

The majority rule consensus tree has the nice combinatorial property that it comes as “close as possible,” on average, to the input trees T_1, T_2, \dots, T_k under the RF metric; more precisely, it is a “median” tree T that minimizes $\sum_{i=1}^k d_{\text{RF}}(T, T_i)$ [99]. The proof of this is straightforward from the definitions. Moreover, this reveals that when $n = |X|$ is odd, the majority rule is the unique median tree. While computing the median tree for T_1, T_2, \dots, T_k can be achieved in polynomial time (in n and k), the problem of finding a *binary* median tree (i.e., a tree $T \in RB(X)$ that minimizes $\sum_{i=1}^k d_{\text{RF}}(T, T_i)$) turns out to be NP-hard [250]. In the last section of this chapter, we will consider other ways of constructing a consensus of trees.

In summary, clusters allow us to easily compare and combine phylogenetic trees. Once a biologist has a tree, it can help answer questions of interest, such as “how long ago did two given species have a common ancestor?” or “how did some given characteristic that varies between species (e.g., brain size) evolve?” First, however, we need a tree. A fundamental problem in phylogenetics is how to reconstruct—or infer—a tree from data present in the species today. Biologists also want to know how accurate such a reconstructed tree is likely to be. We will address these types of questions in later chapters. In this and the next two chapters, we will further explore the combinatorial and stochastic properties of phylogenies.

2.3 • Refinement, compatibility, and encoding

The collection of all hierarchies on X is partially ordered under set inclusion. This translates into a partial order on $RP(X)$. For trees T and T' in $RP(X)$ we say that T' *refines* T , written $T \preceq T'$, if every cluster of T is a cluster of T' . This ordering has a natural biological interpretation. If the hierarchy associated with T is a subset of the hierarchy associated with T' , then either T is equivalent to T' (when the hierarchies are equal) or T is obtained from T' by collapsing one or more edges. Thus $(RP(X), \preceq)$ is a poset (partially ordered set) in which every nonempty subset \mathcal{R} of $RP(X)$ has a greatest lower bound, namely the phylogeny whose associated hierarchy is the intersection of the hierarchies associated with the trees in \mathcal{R} (the *strict consensus tree*).

Compatibility. Suppose that we are now given a collection \mathcal{P} of one or more input phylogenies $T_1, \dots, T_k \in RP(X)$. Then \mathcal{P} is said to be *compatible* if there is a tree T in $RP(X)$ that is a common refinement of all the trees in \mathcal{P} . This holds if and only if the union of the hierarchies associated with the set of trees in \mathcal{P} is a hierarchy, and in this case we can take as our choice of T the tree that corresponds to the union of these hierarchies (T is also the unique minimal refinement of the set of input trees). The equivalence between hierarchies and rooted phylogenies allows a one-line proof of another classic early result in phylogenetics that might otherwise require a more tedious graph-theoretic argument:

A set \mathcal{P} of phylogenies on X is compatible if and only if each pair of phylogenies in \mathcal{P} is compatible.

2.3.1 • Encoding phylogenies by ternary relations

Given any collection \mathcal{C} of nonempty subsets of X , let us write $aa'|_{\mathcal{C}} b$ if there is a set $A \in \mathcal{C}$ with $a, a' \in A$ and $b \in X - A$. Similarly, if T is a rooted phylogeny on X , let us

write $aa'|_T b$ if

$$\text{lca}_T(a, b) = \text{lca}_T(a', b) \neq \text{lca}_T(a, a'),$$

where lca_T refers to the least common ancestor (LCA) relationship from Chapter 1. Informally, $aa'|_T b$ says that the leaves a and a' are more closely related to each other than either is to b .

These two concepts $|_{\mathcal{C}}$ and $|_T$ coincide when \mathcal{C} is a hierarchy on X , and when the tree $T \in RP(X)$ is the corresponding rooted phylogeny (i.e., when $\mathcal{C} = \mathcal{C}(T)$). In other words, for all $a, a', b \in X$,

$$aa'|_{\mathcal{C}} b \iff aa'|_T b.$$

Moreover, in this case, $\mathcal{C} = \{A \subseteq X : aa|_T b \text{ for all } a, a' \in A, b \in X - A\}$, and so the ternary relation $|_T$ determines the hierarchy that corresponds to T (and thereby \mathcal{C}) up to equivalence. In other words, the ternary relation $|_T$ encodes the phylogeny T , and $|_{\mathcal{C}}$ encodes the hierarchy \mathcal{C} . The ternary relation $|_{\mathcal{C}}$ also allows us to characterize when \mathcal{C} is a hierarchy, as follows.

Lemma 2.3. *A collection \mathcal{C} of subsets of X containing X is a hierarchy if and only if the following two conditions both hold:*

- (i) *There are no three elements a, b, c in X with $ab|_{\mathcal{C}} c$ and $ac|_{\mathcal{C}} b$; and*
- (ii) *for each distinct pair $a, b \in X$, we have $aa|_{\mathcal{C}} b$.*

Proof: Condition (i) is equivalent to the hierarchy condition H1. To see why, first suppose that \mathcal{C} satisfies H1. Then one cannot simultaneously have $ab|_{\mathcal{C}} c$ and $ac|_{\mathcal{C}} b$; otherwise, \mathcal{C} would contain sets A and B for which $a, b \in A$ and $c \in X - A$, and $a, c \in B$ and $b \in X - B$, which implies that $a \in A \cap B$, $b \in A - B$, $c \in B - A$, and so A and B are neither disjoint nor nested. Conversely, if \mathcal{C} is not a hierarchy, there are sets $A, B \in \mathcal{C}$ with $A \cap B$ neither empty nor equal to A or B . In that case there exist elements $a \in A \cap B$, $b \in A - B$, $c \in B - A$, and since $A \in \mathcal{C}$ we have $ab|_{\mathcal{C}} c$, and since $B \in \mathcal{C}$ we have $ac|_{\mathcal{C}} b$. Condition (ii) is equivalent to the condition that \mathcal{C} contains all the singleton sets (i.e., H2). ■

Now suppose that λ on X is an arbitrary ternary relation on X (i.e., a subset of $X \times X \times X$ where we write $ab \lambda c$ if $(a, b, c) \in \lambda$). In this more general setting, conditions (i) and (ii) from Lemma 2.3 (with $|_{\mathcal{C}}$ being replaced by λ) are necessary but not sufficient for the existence of a hierarchy \mathcal{H} on T with $\lambda = |_{\mathcal{H}}$. The extra conditions needed were formalized in Theorem 1 of [37], which showed that $\lambda = |_{\mathcal{H}}$ for a (unique) hierarchy \mathcal{H} on X if and only if λ satisfies the following three-point and four-point properties:

- (i) $ab \lambda c \implies ba \lambda c$, and $aa \lambda b \iff a \neq b$;
- (ii) If $bc \lambda d$ and either $ab \lambda d$ or $ab \lambda c$, then $ac \lambda d$.

Notice that these two properties imply that no three elements $a, b, c \in X$ exist with $ab \lambda c$ and $ac \lambda b$; however, this three-point condition alone is not sufficient for λ to coincide with $|_{\mathcal{H}}$ for some hierarchy \mathcal{H} .

Weak hierarchies. While the intersection of two or more hierarchies is always a hierarchy, their union need not be. Nevertheless, for any two hierarchies \mathcal{H} and \mathcal{H}' on X , the set $\mathcal{C} = \mathcal{H} \cup \mathcal{H}'$ satisfies the following (weak) Helly property:

$$A, B, C \in \mathcal{C} \implies A \cap B \cap C \in \{A \cap B, A \cap C, B \cap C\}.$$

A collection \mathcal{C} of nonempty subsets of X that satisfy this property is called a *weak hierarchy* on X . This class of set systems is much larger than the set systems that are merely the unions of two hierarchies. Moreover, just as a hierarchy on $[n]$ has size at most $2n - 1$, a weak hierarchy is also bounded in size by a polynomial in n , in this case $\binom{n+1}{2}$ (we will see why shortly). The cover digraph of a weak hierarchy thus provides a directed graph that need not be a tree—it is an example of a “phylogenetic network,” a topic we will explore further in Chapter 10.

Note that a collection of nonempty sets \mathcal{C} forms a weak hierarchy on X precisely if, for every three elements a, b, c of X , at most two of the three possible ternary relations holds: $ab|_{\mathcal{C}} c, ac|_{\mathcal{C}} b, bc|_{\mathcal{C}} a$. This implies that no subset of X of size 3 can be “shattered” by $\mathcal{C}' = \mathcal{C} \cup \{\emptyset\}$ (i.e., the intersection of any three-element set A with the sets in \mathcal{C}' is never all of 2^A). In this way, we can obtain the quadratic bound on the size of \mathcal{C} mentioned earlier, by applying the Sauer–Shelah lemma (this states that if a collection \mathcal{C}^* of subsets of an n -element set X is larger than $\sum_{i=0}^{k-1} \binom{n}{i}$, then \mathcal{C}^* shatters a subset of X of size k). It follows that \mathcal{C}' can have cardinality at most $1 + \binom{n}{1} + \binom{n}{2} = 1 + \binom{n+1}{2}$ and so $|\mathcal{C}| \leq \binom{n+1}{2}$. Moreover, there are weak hierarchies that realize this bound.

The quadratic bound on the size of a weak hierarchy can also be derived from another property they possess: If \mathcal{C} is a weak hierarchy, then for any $A \in \mathcal{C}$ there exists a pair of (possibly equal) elements $x, y \in A$ for which A is the intersection of all elements of \mathcal{C} that contain both x and y .

2.4 ■ Unrooted trees as systems of splits

In Section 2.1, we described a bijective correspondence between unrooted phylogenetic X -trees and rooted phylogenetic X -trees on $X - \{x\}$ (for any $x \in X$), and thereby to hierarchies on $X - \{x\}$. However, the choice of a particular element $x \in X$ is completely arbitrary, so we seek a more satisfactory way to describe an unrooted phylogenetic X -tree. This is based on the notion of an X -split, which is a bipartition of X into two nonempty parts (A and B , say), usually written $A|B$ as shorthand for $\{A, B\}$. Such a notion has clear biological meaning: for example, we can divide the set of all animal species into the vertebrates and the invertebrates.

Given any unrooted phylogenetic X -tree T , if we delete any particular edge e of T and consider the leaf sets of the two connected components of the resulting disconnected graph, we obtain a corresponding X -split, which we will refer to as a *split of T* that corresponds to e . For example, for each $x \in X$, every phylogenetic X -tree has the *trivial split* $\{x\}|(X - \{x\})$, corresponding to the edge incident with leaf x . We let $\Sigma(X)$ denote the set of all X -splits and let $\Sigma(T)$ (respectively, $\dot{\Sigma}(T)$) denote the X -splits that correspond to edges (respectively, interior edges) of T .

Notice that any two splits $A|\overline{A}$ and $B|\overline{B}$ of the same phylogenetic X -tree have the property that one of the four intersections

$$A \cap B, A \cap \overline{B}, \overline{A} \cap B, \text{ and } \overline{A} \cap \overline{B}$$

must be empty. If a collection Σ of X -splits has this property (for every pair of splits), we say that Σ is *pairwise compatible*. For example, the tree in Fig. 2.1(ii) has the splits

$$A|\overline{A} = \{a, b\}|\{c, d, e, f\} \text{ and } B|\overline{B} = \{a, b, c, d\}|\{e, f\},$$

so in this case $A \cap \overline{B} = \emptyset$. The pairwise compatibility of splits is the unrooted analogue of the hierarchy property **H1**, so it is not surprising that Proposition 2.1 has an equivalent formulation for unrooted trees.

Proposition 2.4. *A collection Σ of X -splits is the set of splits of some unrooted phylogenetic X -tree T if and only if Σ is pairwise compatible and contains the trivial splits. Moreover, two unrooted phylogenetic X -trees T and T' have the same set of X -splits if and only if $T \cong T'$.*

This result says that any pairwise compatible collection Σ of X -splits can be represented by a unique unrooted phylogenetic X -tree, namely the tree $T \in P(X)$ whose nontrivial splits are the nontrivial splits of Σ , just as every hierarchy on X can be represented by a rooted phylogenetic X -tree. The notion of refinement extends naturally from rooted phylogenies to unrooted ones: A tree T' in $P(X)$ refines a tree T in $P(X)$, written $T \preceq T'$, if and only if $\Sigma(T) \subseteq \Sigma(T')$. Thus $T \preceq T'$ means that $T \cong T'$ or that T is obtained from T' by collapsing certain interior edges (the ones corresponding to the splits in $\Sigma(T') - \Sigma(T)$).

If we drop the condition that Σ contains the trivial splits in Proposition 2.4, we obtain instead a corresponding bijection between X -trees (cf. Section 1.3) and set systems that are merely pairwise compatible. It follows that the number of X -trees (up to equivalence) is $2^n \cdot |P(n)|$, for $n = |X|$.

Any collection \mathcal{C} of nonempty subsets of X that satisfies the hierarchy condition H1 gives rise to a set Σ of pairwise compatible splits on X by the map $\tilde{s} : A \mapsto A|\bar{A}$ (where \bar{A} is the complement of A in X). This corresponds exactly to the operation s of suppressing the root of a tree for associating an unrooted phylogeny with a rooted one, as described at the start of this chapter (Section 2.1). The other operation we described there, namely o_x (adding an outgroup), provided a bijection from $RP(X - \{x\})$ to $P(X)$; at the level of hierarchies and split systems the inverse of this map is easily described; namely, associate each split $A|B$ in Σ with the set (A or B) that does not contain x . However, this bijection between hierarchies and the splits of an unrooted phylogeny is a little unsatisfying, as it requires an arbitrary choice of x and involves trees with different leaf sets (namely X and $X - \{x\}$).

This raises the question of whether there is a canonical way to associate a hierarchy on X with any pairwise compatible split system Σ containing the trivial splits for which the map \tilde{s} recovers Σ . In other words, is there a canonical way to root an unrooted phylogeny? There are natural ways to do this based on any one of the notions of center, centroid, and leaf-centroid introduced in Chapter 1. For any one of these different concepts of “centrality,” an unrooted phylogeny T has either (i) a single vertex or (ii) a pair of adjacent vertices that are central, and so we could root T on the single vertex (in case (i)) or the midpoint of edge connecting them (in case (ii)) to obtain a rooted phylogeny. An example of rooting a phylogeny on its leaf-centroid is illustrated in Fig. 2.1 ((ii) and (iv)).

Of the three notions, the leaf-centroid has the most simple description as a map from hierarchies to pairwise compatible split systems, as we now explain. Given a collection Σ of X -splits, let $\tilde{\mathcal{C}}(\Sigma)$ be the collection of subsets A of X with size at most $\frac{1}{2}|X|$ for which $A|\bar{A} \in \Sigma$ (i.e., the smaller block of each split; in case of a tie, both blocks). Part (i) of the following result dates back to 1977 [248].

Proposition 2.5. *Suppose Σ is a collection of X -splits that contains the trivial splits. Then*

- (i) Σ is a pairwise compatible if and only if $\tilde{\mathcal{C}}(\Sigma)$ is a hierarchy on X ; and
- (ii) when (i) holds and $T \in P(X)$ represents Σ , the tree $T' \in RP(X)$ that represents $\tilde{\mathcal{C}}(\Sigma)$ is obtained from T by rooting this tree on its leaf-centroid (vertex or edge).

Proof. Part (i) follows by a direct set-theoretic arguments and is left as an exercise (the “only if” direction also follows from part (ii)). For part (ii), if the leaf-centroid of T consists of a

single vertex v , let $T' \in RP(X)$ be obtained from T by rooting T on v (for this case, there is no split $A|B$ of T with $|A| = |B| = \frac{1}{2}|X|$). In the second case, where the leaf-centroid of T consists of two adjacent vertices v and v' , let $T' \in RP(X)$ be the tree obtained by rooting T on the midpoint of the edge $\{v, v'\}$. In either case, $\tilde{\mathcal{C}}(\Sigma) = \mathcal{C}(T')$, and the map $A \mapsto A|\bar{A}$ is a bijection from this set of clusters of T' to the splits of T . ■

2.4.1 • Unrooted analogues of hierarchy results

Most of the results described for rooted phylogenies (viewed as hierarchies) have direct analogues for unrooted phylogenies (viewed as pairwise compatible split systems). For example, the majority rule consensus tree of a collection \mathcal{P} of unrooted trees is defined in a directly analogous way (the tree that corresponds to splits that appear in more than half of the trees in \mathcal{P}). Compatibility of \mathcal{P} is also defined analogously, namely the existence of a common (unrooted) tree that refines each of the trees in \mathcal{P} .

The Robinson–Foulds distance between two trees $T, T' \in P(X)$, denoted $d_{RF}(T, T')$, is similarly defined as the number of splits in one tree but not the other, and can be written

$$d_{RF}(T, T') = |\Sigma(T) \Delta \Sigma(T')|,$$

where Δ is the symmetric difference operator (i.e., $A \Delta B = A \cup B - (A \cap B)$) [296]. Notice that $|A \Delta B| = |A| + |B| - 2|A \cap B|$. Since $|E(T)| = |\Sigma(T)|$ and the set $\overset{\circ}{E}(T)$ of interior edges of T is in one-to-one correspondence with the set $\overset{\circ}{\Sigma}(T)$ of nontrivial splits of T , there are various equivalent ways to write $d_{RF}(T, T')$. For example,

$$d_{RF}(T, T') = |\overset{\circ}{E}(T)| + |\overset{\circ}{E}(T')| - 2|\overset{\circ}{\Sigma}(T) \cap \overset{\circ}{\Sigma}(T')|. \quad (2.9)$$

The set-theoretic definitions of $d_{RF}(T, T')$ are computationally helpful, but they obscure the fact that $d_{RF}(T, T')$ has a simple and biologically relevant interpretation: it is the smallest number of edges in total that need to be collapsed in both T and T' to arrive at the same unrooted phylogeny. The RF metric is widely used by biologists, because there is a fast (polynomial-time) way to compute it, its interpretation is clear, and some of its basic properties are readily established. For instance, the maximal distance between any two trees in $P(n)$ is $2n - 6$, which is achieved precisely for pairs of binary phylogenies that do not share any nontrivial split. To see why, notice that for any two trees in $B(n)$, eqn. (2.9) simplifies to

$$d_{RF}(T, T') = 2n - 6 - 2|\overset{\circ}{\Sigma}(T) \cap \overset{\circ}{\Sigma}(T')|.$$

This equality also reveals that the d_{RF} distance between two binary trees is always an even integer, so the smallest nonzero distance is 2. For a given tree T in $B(n)$, there are exactly $2(n-3)$ trees T' with $d_{RF}(T, T') = 2$. These are the trees obtained from T by collapsing one of the $n-3$ interior edges, and then “re-expanding” the edge in one of two different ways. The number of trees $T' \in B(n)$ with $d_{RF}(T, T') = 4$ is quadratic with n , and its exact value depends on the number $c(T)$ of cherries of T . More precisely, for any tree T in $B(n)$, one has

$$|\{T' \in B(n) : d_{RF}(T, T') = 4\}| = 2n^2 - 8n - 12 + 6c(T).$$

Despite the dependence on $c(T)$, notice that this expression can be written asymptotically as $2n^2(1 + O(n^{-1}))$. This is part of a more general result from [104]: for any $T \in B(n)$

and any fixed k , the number of trees $T' \in B(n) : d_{\text{RF}}(T, T') = 2k$ is equal to

$$\frac{2^k n^k}{k!} (1 + O(n^{-1})).$$

Notice that we can also write

$$d_{\text{RF}}(T, T') = |\Sigma(T) - \Sigma(T')| + |\Sigma(T') - \Sigma(T)|. \quad (2.10)$$

In applications, T might represent some known “true” tree, while T' is some estimate of it, perhaps in a simulation study. In this case, the two terms on the right in (2.10) are sometimes reported separately as the “false negative” and “false positive” values, respectively. The former are splits that should have been present but were not found; the latter are splits in the estimated tree that were not present in the “true” tree.

One feature of d_{RF} that is sometimes annoying in biological data is that it can be very sensitive to small perturbations of the tree. For example, if a leaf of a caterpillar tree is moved from one end of the tree to the other, then the resulting tree is at the maximal RF distance from the original tree, even though the trees are otherwise identical. This feature is not usually a problem if one is comparing very similar trees; however, for other applications this sensitivity renders the RF metric too coarse. We will look at some alternative metrics on trees later in this chapter (based on tree rearrangements) and in Chapter 4 (based on subtrees), but there are also ways to modify the RF metric to make it much less sensitive, yet still computable efficiently. This is based on replacing the all-or-nothing scoring of two splits by whether they are identical or not with a more sensitive measure of the extent to which they differ. Two groups of authors ([39] and [231]) independently arrived at similar proposals for such modifications in papers published in the same journal issue in 2012.

The Buneman graph. We end this section by describing how to associate a connected graph to any set of X -splits, which returns a phylogenetic tree, precisely when the splits are pairwise compatible and contain the trivial splits. This is the *Buneman graph*, which belongs to the class of “median graphs.” This graph was introduced by Peter Buneman in 1971, and modifications of it have been extensively developed and applied more recently by the mathematician Hans-Jürgen Bandelt and colleagues, for building “haplotype networks.” Given a collection $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_k\}$ of X -splits, the associated Buneman graph $B(\Sigma) = (V, E)$ and its labeling function $f : X \rightarrow V$ is defined as follows: The vertex set V consists of all sets $v = \{S_1, S_2, \dots, S_k\}$, where $S_i \in \sigma_i$ and $S_i \cap S_j \neq \emptyset$ for all $i \neq j$. Two vertices v and v' are the ends of an edge in E precisely if $|v \cap v'| = k - 1$. For $v \in V$, with $v = \{S_1, \dots, S_k\}$, say, let $I_v = \bigcap_{i=1}^k S_i$. It can then be checked that $\{I_v : v \in V, I_v \neq \emptyset\}$ partitions X . For each $x \in X$, let $S_i(x)$ denote the block of the split σ_i that contains x . Then the function $f : X \rightarrow V$, defined by $f(x) = \{S_1(x), \dots, S_k(x)\}$, is well defined and its range contains all $v \in V$ of degree at most 2.

The graph $B(\Sigma)$ is connected, and it is a tree if and only if Σ is pairwise compatible. Moreover, in this case, the tree $B(\Sigma)$ is a phylogenetic X -tree in which the leaves are precisely the vertices that receive a (unique) label according to the labeling function f , provided that Σ contains the trivial splits. When Σ is not pairwise compatible, $B(\Sigma)$ contains cycles and can have a large number (exponential in $|X|$) of vertices and edges. There are concise expressions for these numbers. For example, the number of vertices of $B(\Sigma)$ is 1 plus the number of cliques in the graph that has vertex set Σ and an edge between each pair of splits that are incompatible. For example, if Σ is pairwise compatible,

then $|B(\Sigma)| = 1 + |\Sigma|$, since the only cliques in the incompatibility graph are the singleton vertices from Σ .

The Buneman graph has a number of other special properties,⁶ and we will return to it, and a generalization called the *quasi-median graph*, in Chapter 10.

Exercise⁺: Show that the Buneman graph is a bipartite graph. [Hint: Select $x \in X$ and consider the parity of the number of sets S_i in v that contain x .]

2.4.2 • Three more ways to encode an unrooted phylogeny

For unrooted trees, the analogue of the ternary relation $ab|_T c$ described earlier is a *quaternary relation* on X . Given a phylogenetic X -tree T write $aa'|_T bb'$ if there is some edge of T that separates the leaves a and a' from the leaves b and b' . In other words, $aa'|_T bb'$ holds if T has an X -split $A|B$ with $a, a' \in A$ and $b, b' \in B$. More generally, given an arbitrary collection Σ of X -splits, we can write $aa'|_\Sigma bb'$ if and only if there is a split $A|B \in \Sigma$ with $a, a' \in A$ and $b, b' \in B$ (thus $aa'|_T bb' \iff aa'|_{\Sigma(T)} bb'$). In this more general setting, Σ is the set of splits of a (unique) phylogenetic X -tree if and only if $|\Sigma$ satisfies the properties that (i) there is no quadruple a, b, c, d of elements of X with $ab|_\Sigma cd$ and $ac|_\Sigma bd$, and (ii) for every distinct triple $a, b, c \in X$, $aa|_\Sigma bc$. This is the unrooted analogue of Lemma 2.3.

It is also possible to characterise when an arbitrary quaternary relation on X (i.e., a subset ℓ of X^4 , where we write $ab\ell cd$ when $(a, b, c, d) \in \ell$) can be written as $\ell = |_T$ for some (unique) phylogenetic X -tree. As might be expected (by analogy with the ternary case), such a characterization involves conditions with up to five elements of X . This was first described in a mathematical psychology paper from 1981 (by H. Colonius and H. H. Schultze); for references and a more modern treatment, see [20].

Another way to encode a phylogenetic X -tree $T \in P(X)$ is via the graphical distances between each pair of leaves. For $x, y \in X$, let $d_T(x, y)$ be the number of edges on the unique shortest path in T connecting x and y . The fact that d_T determines any tree $T \in P(X)$ up to equivalence is an old result (~1965). Moreover, for each binary tree $T \in B(n)$, just $(2n - 3)$ carefully chosen d_T values suffice to determine T , a result from 1983 found in [82]. We will explore this encoding in greater generality (allowing edges to take weights other than 1) in Chapter 6.

Circular split systems and circular orderings of phylogenies. There is yet another way to encode phylogenies based on circular orderings, which will also be relevant in Chapter 6. Let $\pi = (x_1, x_2, \dots, x_n)$ be a cyclic permutation of X . A set Σ of X -splits is said to be a *circular split system* on X for π if, for each split $A|\bar{A} \in \Sigma$, A is a consecutive subsequence of x_i values in the cyclic permutation π (note that x_{n-1}, x_n, x_1 would be such a consecutive subsequence).

A circular split system has size at most $\binom{n}{2}$. Moreover, this size is realized by the *full circular split system* obtained by placing the element of X at the corners of a regular n -gon according in the circular order specified by π , and considering all splits that arise when two edges of the regular n -gon are deleted (this disconnects the n -gon into two connected components, and so the elements of X are partitioned accordingly into two

⁶For example, the Buneman graph is a *median graph*, in other words, for every three vertices, v_1, v_2 , and v_3 there is a unique vertex v that lies on the shortest paths joining v_1 and v_2 , v_1 and v_3 , and v_2 and v_3 .

parts). Visually, we can think of the splits as the result of slicing through any two edges of the n -gon with a straight line in all possible ways.

For a phylogeny $T \in P(X)$, we say that π is a *circular ordering* of T if $\Sigma(T)$ is a circular split system for π . Informally, this means one can draw the tree in the plane in such a way that the order in which the leaves are encountered as one traces around the tree in a clockwise fashion is x_1, x_2, \dots, x_n . An example is shown in Fig. 2.2. The number of circular orderings for any phylogeny depends only on the degrees of the interior vertices according to the following result.

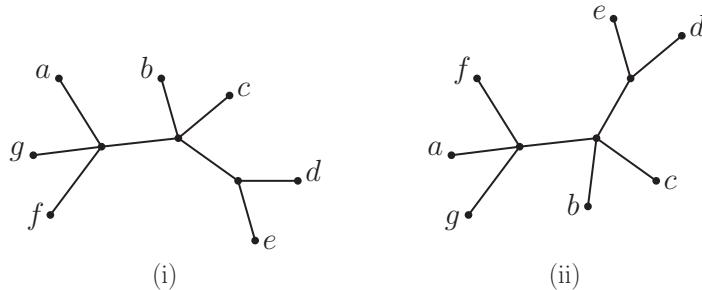


Figure 2.2. (i) A phylogeny for which the cyclic permutation (a, b, c, d, e, f, g) is a circular ordering; (ii) the same phylogeny with a different circular ordering, namely (f, e, d, c, b, g, a) .

Lemma 2.6. For $T \in P(X)$, where $n = |X| \geq 3$, the number of distinct circular orderings of T is

$$\prod_{v \in \overset{\circ}{V}(T)} (d(v)-1)!$$

Proof. We use induction on the number $v_T = |\overset{\circ}{V}(T)|$ of interior vertices of T . Since $|X| \geq 3$, we have $v_T \geq 1$; moreover, $v_T = 1$ precisely if T is a star tree. Notice that each of the $(n-1)!$ possible cyclic permutations of X provides a circular ordering for T , so the lemma holds in the base case where $v_T = 1$. Now suppose that the lemma holds when $v_T = k-1 \geq 1$ and that $T \in P(X)$ has k interior vertices. In this case, T has a vertex u that is adjacent to $l \geq 2$ leaves—say, x_1, x_2, \dots, x_l —and also to one other interior vertex. Notice that $d(u) = l+1$. Let T' be a phylogenetic tree obtained from T by deleting the l leaves of T that are adjacent to u and labeling the newly created leaf u . By induction, this tree has $\prod_{v \in \overset{\circ}{V}(T')} (d(v)-1)!$ circular orderings. Moreover, for each circular ordering of T' and each of the $l!$ permutations σ of the leaf set x_1, x_2, \dots, x_l , we obtain a circular ordering π of T by replacing the occurrence of u in the circular ordering of T' by the permutation σ of x_1, x_2, \dots, x_l ; in addition, this association is bijective. In this way, the number of circular orderings of T is $\prod_{v \in \overset{\circ}{V}(T')} (d(v)-1)! \cdot l! = \prod_{v \in \overset{\circ}{V}(T)} (d(v)-1)!$, thereby establishing the induction step. ■

Corollary 2.7. For any cyclic permutation π of $[n]$, the number of phylogenies T in $B(n)$ for which π is a circular ordering is the (Catalan) number:

$$\frac{1}{n-1} \binom{2n-4}{n-2}.$$

Proof. Let S denote the set of pairs (T, π) , where $T \in B(X)$ and π is a circular ordering for T . We count S in two ways. The number of choices of T is $b(n) = (2n-4)!/[(n-2)!2^{n-2}]$, and since each such T has $n-2$ interior vertices, each of which has degree 3, the number of choices for π is 2^{n-2} , by Lemma 2.6. Let us now count S by selecting π first: there are exactly $(n-1)!$ cyclic permutations on X , and the number of trees $T \in B(X)$ for which $(T, \pi) \in S$ is the number we want. Equating these two counts of S gives the expression in part (ii). ■

If we let $o(T)$ denote the set of circular orderings of T , then this set determines T . This follows from a stronger result, namely that one phylogeny on X refines another precisely if every circular ordering of the latter is a circular ordering of the former. A proof of this result can be found in [316].

Proposition 2.8. *For any two phylogenies T and T' in $P(X)$, $T \preceq T'$ holds if and only if $o(T') \subseteq o(T)$. Thus $T \cong T'$ if and only if T and T' possess exactly the same set of circular orderings.*

2.5 • Tree rearrangement metrics

2.5.1 • Surgery operations on trees (NNI, SPR, TBR)

From any binary phylogeny T , it is possible to generate a set of neighboring trees by simple “cut-and-paste” type operations. These play an important role in phylogenetics; for example, in optimization algorithms that attempt to find a “best tree” for data under some scoring function, and in the statement of some propositions in coming chapters. These tree rearrangement operations also provide a way of comparing trees, in which the distance between two binary phylogenies is the minimum number of “rearrangement” events required to transform one phylogeny into the other. The resulting metrics are quite different from the RF metric restricted to binary phylogenies.

The simplest rearrangement event on a tree $T \in B(X)$ is *nearest neighbor interchange* (NNI). Such an operation on T selects an interior edge of T and swaps two of the four rooted trees incident with the opposite ends of that edge to produce a tree $T' \in B(X)$. Equivalently, two trees T and T' in $B(X)$ are related by a single NNI operation if they share all but one nontrivial split. Two such trees are said to be NNI neighbors. The NNI distance between two trees T and T' , denoted $d_{\text{NNI}}(T, T')$, is defined as the smallest number of such moves required to convert T to T' (i.e., the minimal k for which there is a sequence $T = T_0, T_1, \dots, T_k = T'$, where T_i and T_{i+1} are NNI neighbors). It is easily checked that d_{NNI} is a metric on $B(n)$ (in particular, any tree in $B(n)$ can be converted into any other by a finite number of NNI operations).

Exercise: Let $T_0 \in B(n)$ be a caterpillar tree. Show that for any tree T in $B(n)$ there is a sequence of NNI operations that converts T to T_0 . Deduce that for any two trees $T, T' \in B(n)$ there is a sequence of NNI operations that converts T to T' .

Notice that the NNI metric and RF metric are related by the inequality

$$d_{\text{RF}}(T, T') \leq 2d_{\text{NNI}}(T, T'),$$

since T and T' are NNI neighbors precisely if $d_{\text{RF}}(T, T') = 2$, coupled with the triangle inequality for d_{RF} .

For each tree $T \in B(n)$, there are exactly $2(n - 3)$ other trees that can be obtained from T by a single NNI operation, since each tree $T \in B(n)$ has $n - 3$ interior edges and there are two rearrangements about each interior edge that lead to different trees; moreover, rearranging about a tree around one interior edge cannot produce the same tree as rearranging the tree about another interior edge.

Two further tree rearrangement operations are *subtree prune and regraft* (SPR) and *tree bisection and reconnection* (TBR). An SPR operation on $T \in B(n)$ is specified by an ordered pair of edges (e, e') . The edge e is cut and e' is subdivided to create a new vertex v of degree 2 (see Fig. 2.3). The maximal rooted binary subtree T_e of $T - e$ that is disconnected from e' is reattached by its root to v by a new edge to create a tree $T' \in B(n)$. Remarkably, each tree $T \in B(n)$ has exactly the same number of SPR neighbors, namely $2(n - 3)(2n - 7)$ [6].

TBR is the most general tree rearrangement operation, which is defined in exactly the same way as SPR, except that in attaching T_e to v , we have the freedom to either use the root of T_e (as in SPR), or to suppress this root vertex and subdivide any edge within T_e and attach T_e to v by a new edge that goes from this subdivision vertex to v (see Fig. 2.3).

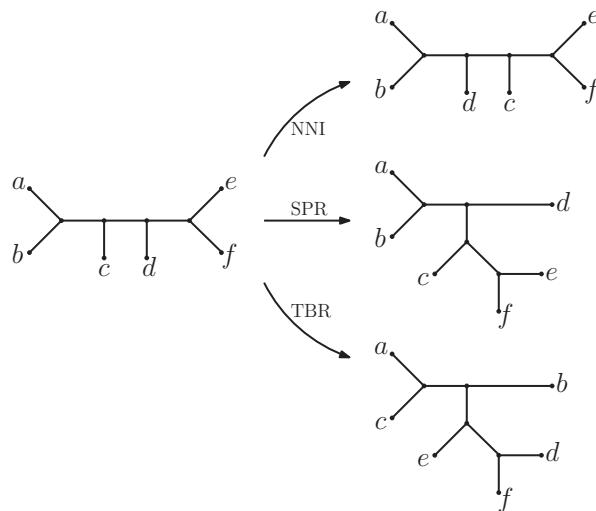


Figure 2.3. Illustration of the three types of tree rearrangement operations. The NNI operation involves the central interior edge. The SPR operation cuts the rightmost interior edge and attaches it to the edge leading to c . The TBR operation cuts the central interior edge and attaches it by a different rooting to the edge leading to b .

Notice that all three operations (NNI, SPR, and TBR) are reversible: if a phylogeny T' can be obtained by a single operation (NNI, SPR, or TBR) from T , then T can also be obtained by a single operation of the same type from T' . It is easy to see that every TBR operation on T results in a tree that can be obtained by making, at most, two SPR operations starting from T . On the other hand, a single SPR operation on T may require a large number (linear in n) of NNI operations to produce the same tree. Despite this similarity of TBR to SPR, there is one important difference. Unlike the neighborhood sizes for NNI and SPR, the number of TBR neighbors of a tree depends on the shape of T . Fortunately, there is still an explicit description of the size of the TBR neighborhood

of a tree, given in [201]. Let

$$\Gamma(T) = \sum_{A|B \in \dot{\Sigma}(T)} |A| \cdot |B|,$$

where the sum is taken over all nontrivial splits $A|B$ of T . The number of trees that can be obtained by one TBR operation on a given tree T in $B(n)$ is then given by

$$4\Gamma(T) - (4n - 2)(n - 3).$$

Moreover, the authors in [201] show that the trees with the largest neighborhood size are precisely the caterpillar trees, for which there are $\frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2$ neighboring trees. This is established by a connection between $\Gamma(T)$ and a more classical index (studied in chemical graph theory) associated with a tree called its *Wiener index* $W(T)$, which is the sum of the distances between all pairs of vertices. For any tree T in $B(n)$, $\Gamma(T)$ is proportional to $W(T)$ plus an additive term that is dependent only on n , so the binary trees that have the largest number of TBR neighbors are precisely the binary trees with the largest Wiener indices, which are known to be caterpillar trees. It is also possible to characterize the trees that have the minimum number of TBR neighbors; this class includes the perfect binary trees with $n = 2^b$ leaves, for which the neighborhood size is

$$4n^2 \lfloor \log_2(n) \rfloor + O(n^2).$$

In summary, the TBR neighborhood of a binary tree with n leaves has a size that lies between $\Theta(n^2 \log(n))$ and $\Theta(n^3)$, depending on the shape of the tree.

The distance between two trees under SPR or TBR is defined in the same way as for NNI: it is the minimum number of operations of the given type (SPR or TBR) required to convert one tree into the other. For $\theta \in \{NNI, SPR, TBR\}$, the associated distance function d_θ is a metric on the set $B(n)$. In contrast to the RF metric, computing these metrics is NP-hard; however, this has not stopped mathematicians from studying them. One of the most illuminating findings has been an equivalent description of TBR in terms of “maximum agreement forests.”

Agreement forests. An *agreement forest* for $T_1, T_2 \in B(X)$ (where $|X| \geq 4$) is a binary phylogenetic forest $\mathcal{F} = \{t_1, \dots, t_k\}$ (in other words, a collection of binary phylogenetic trees), in which the leaf sets partition X , so that if X_j is the leaf set of t_j , then

- (i) $T_1|X_j = T_2|X_j = t_j$ for all $j \in \{1, \dots, k\}$; and
- (ii) the subtrees $T_i|X_j$, where $j \in \{1, \dots, k\}$, are vertex disjoint subtrees of T_i for each $i \in \{1, 2\}$.

An example is shown in Fig. 2.4.

It is clear that any two trees in $B(n)$ have an agreement forest; for example we can obtain one simply by taking $k = |X|$ and letting the sets X_i be singletons (i.e., $X_i = \{i\}$ for each $i \in X$). Moreover, $T = T'$ precisely when T and T' have an agreement forest with $k = 1$. The smallest value of k for which two trees have such an agreement forest \mathcal{F} is denoted $m(T_1, T_2)$, and this \mathcal{F} is referred to as a *maximum agreement forest* (MAF). Informally, $m(T_1, T_2) - 1$ is the smallest number of edges that need to be deleted from each tree (T_1 and T_2) so that the resulting phylogenetic forests agree (once any unlabeled vertices of degree less than 3 are removed). The relevance of MAFs is given by the following result from [6].

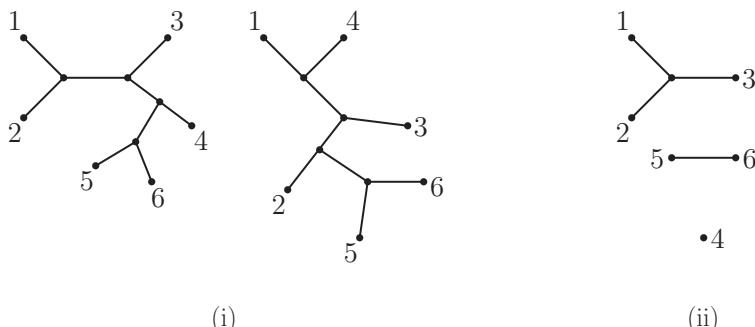


Figure 2.4. For the two trees from $B(6)$ in (i), an agreement forest is shown in (ii). This is also a maximum agreement forest, and so either tree in (i) can be obtained from the other by two TBR operations, but not fewer, by Proposition 2.9.

Proposition 2.9. *For any two trees $T, T' \in B(X)$,*

$$d_{\text{TRB}}(T, T') \equiv m(T, T') - 1.$$

This equivalence has proved useful in obtaining a better understanding of the computational and mathematical properties of d_{TRB} .

It is also possible to define NNI and SPR operations on rooted binary phylogenetic trees. The definition of the rooted NNI operation on $RB(X)$ is straightforward (T and T' are one move apart if they are equivalent upon contracting an edge in each tree).

For rooted subtree and prune (rSPR) operation on $T \in RB(n)$ an edge leading to a pendant subtree T' is cut. This subtree T' is then reattached to either a subdividing vertex of some edge that is present in the other part of T , or T' is attached to a new degree-2 root vertex (with the other outgoing edge from this new vertex attaching to the original root of T). Computing the rSPR distance is also NP-hard; however, on the positive side, the problem of computing d_{rSPR} for two trees is fixed parameter tractable in the rSPR distance between the trees [49]. Rooted SPR distance has a characterization in terms of maximum agreement forests of rooted trees (analogous to Proposition 2.9 for TBR); for details, see [49]. Fixed parameter algorithms for finding such forests have recently led to improved algorithms for calculating rSPR distances [371].

By contrast, ordinary SPR on unrooted binary trees appears to stubbornly resist having a close connection with agreement forests. Another way in which rSPR is not closely related to SPR is that it is possible for two rooted trees, T_1 and T_2 in $RB(n)$, to have rSPR distance $\Omega(n)$, and yet for the associated unrooted trees $T_1^{-\rho}$ and $T_2^{-\rho}$ to be identical (i.e., have SPR distance of 0) [17].

2.5.2 • Properties of (discrete) tree space

For $\theta \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$, θ -tree space is the graph $G_\theta(n)$ with vertex set $B(n)$ and with an edge between two trees precisely if they are neighbors under a single θ -operation. Thus the distance d_θ between two trees in $B(n)$ is simply the length of the shortest path in $G_\theta(n)$ between the two trees. We will also let $G_{\text{rSPR}}(n)$ denote the (connected) graph with vertex set $RB(n)$ and an edge between two trees if they are rSPR neighbors. Each of these four graphs is a connected graph. In this section, we will explore some basic properties of θ -tree space.

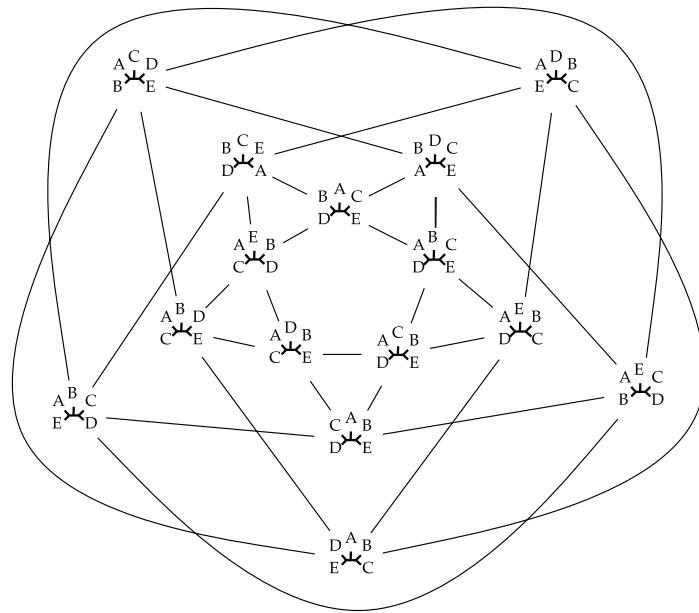


Figure 2.5. NNI-tree space for $B(5)$. This graph is isomorphic to the line graph of the Petersen graph. Reprinted with kind permission of Joseph Felsenstein [136].

Starting with $G_{\text{NNI}}(n)$, this graph is regular (each vertex has the same degree, namely $2(n - 3)$). When $n = 4$ the graph is simply a 3-cycle, while for $n = 5$ it has a more interesting structure as shown in Fig. 2.5. This graph is closely connected with a famous graph called the “Petersen graph” (this has 10 vertices, each of degree 3, and 15 edges). The Petersen graph can be viewed as the graph whose vertex sets consist of the 10 nontrivial splits of $\{1, 2, 3, 4, 5\}$, and with two vertices being adjacent if the two associated splits are compatible.⁷ Thus the 15 edges of the Petersen graph correspond to the 15 trees in $B(5)$ since each edge (tree) is specified by its two end-vertices (two nontrivial splits); moreover, two edges of this representation of the Petersen graph are incident if their associated trees are NNI neighbors. Stated more formally, $G_{\text{NNI}}(5)$ is isomorphic to the “line graph” of the Petersen graph.⁸

For all $n \geq 3$, $G_{\text{NNI}}(n)$ has the remarkable property of possessing a Hamiltonian path. In other words, it is possible to start at one tree and, by making single NNI moves, visit each tree in $B(n)$ exactly once. A constructive proof of this was recently described in [157].

$G_{\text{NNI}}(n)$ harbors another surprise. If two trees T and T' from $B(n)$ share a nontrivial split $A|B$, then it might be supposed that there would be a shortest path in $G_{\text{NNI}}(n)$ for which all trees in that path also have the split $A|B$ (i.e., an optimal path should exist that just carries out rearrangements on either side of the shared split). Despite its apparent plausibility—enhanced somewhat by being stated as a “theorem” in a paper from the 1970s—it turns out not to be true; the existence of a counterexample to this claim, for sufficiently large values of n , was established in [228].

⁷This follows from viewing the Petersen graph as the Kneser graph $KG_{5,2}$.

⁸The line graph of a graph $G = (V, E)$ has vertex set E , with $\{e, e'\}$ forming an edge if e and e' are distinct elements of E that are incident with the same vertex of G .

$G_{\text{SPR}}(n)$ is also a regular graph, and with a remarkable property: Given any two trees $T, T' \in B(n)$ there is a path from T to T' in $G_{\text{SPR}}(n)$ for which the RF distance d_{RF} between the k th tree in the path and T' is strictly decreasing with k (see [47], Theorem 3.1). Although $G_{\text{NNI}}(n)$ and $G_{\text{SPR}}(n)$ are regular graphs, they are not vertex-transitive, except for small values of n .⁹ For example, $G_{\text{NNI}}(n)$ is vertex-transitive for $n \leq 5$ but not for $n \geq 6$, since the number of trees at an NNI distance of 2 from $T \in B(n)$ is exactly $2n^2 - 10n + 4c(T)$, where $c(T)$ is the number of cherries of T [104]. This dependence on the shape of the trees precludes vertex transitivity. Similarly, the size of the second neighborhood of T under SPR is influenced by the shape of T [104]. On the other hand, if T and T' have the same shape, then there is an automorphism of $G_{\theta}(n)$ that maps T to T' : namely applying the same permutation of the leaf set required to convert T to T' to all the trees in $B(n)$.

Diameters of θ -tree spaces. Two basic questions about $G_{\theta}(n)$ follow:

- (i) What is the diameter of these graphs (i.e., how far apart can two trees be under d_{θ})?
- (ii) How quickly does a random walk on this graph approach its stationary distribution?

Regarding the diameter $\Delta[G_{\theta}(n)]$ of θ -tree space, first observe that

$$\Delta[G_{\text{TBR}}(n)] \leq \Delta[G_{\text{SPR}}(n)] \leq 2\Delta[G_{\text{TBR}}(n)], \quad (2.11)$$

since each SPR is a TBR, and each TBR can be realized by, at most, two SPRs.

A crude counting argument gives us a first-order determination of the diameter of these two graphs. First, it is straightforward to see how to transform any tree T in $B(n)$ into a given caterpillar tree $T_0 \in B(n)$ using just $N = O(n)$ SPR operations. It follows that for any two trees T and T' in $B(n)$, there is a chain of $2N = O(n)$ SPR operations that converts T to T' . Thus $G_{\text{SPR}}(n)$ has diameter $O(n)$, and so $G_{\text{TBR}}(n)$ does also by inequality (2.11). Next, we establish a linear lower bound: $\Delta[G_{\text{TBR}}(n)] = \Omega(n)$. We saw above that the number of TBR neighbors of any tree T in $B(n)$ is, at most, cn^3 for some constant c , so the number $N(k)$ of trees reachable from $T \in B(n)$ by, at most, k TBR operations is bounded above by

$$\sum_{i=0}^k (cn^3)^i \leq (k+1)C^k n^{3k}, \quad (2.12)$$

where $C = \max\{c, 1\}$. Now, if $d = \Delta[G_{\text{TBR}}(n)]$, then $N(d) \geq b(n)$, and so with $k = d$ in eqn. (2.12) we obtain

$$(d+1)C^d n^{3d} \geq b(n) \sim 2^{n \log_2 n + O(n)}. \quad (2.13)$$

Taking the logarithm of both sides gives $d = \Omega(n)$, which establishes the claimed linear lower bound, so $G_{\text{TBR}}(n)$ has diameter $\Theta(n)$ (and $G_{\text{SPR}}(n)$ does too, by inequality (2.11)). However, using the agreement forest connection of Proposition 2.9, it is possible to tease apart this simple first-order behavior to get a much more precise asymptotic, as stated in the following result from [17] and [111].

Theorem 2.10. *Each of $G_{\text{SPR}}(n)$, $G_{\text{TBR}}(n)$, and $G_{\text{rSPR}}(n)$ has diameter $n - \Theta(\sqrt{n})$.*

⁹A graph G is vertex-transitive if for any two vertices u and v of G , there is an automorphism of G that maps u to v ; roughly speaking, this means the graph looks the same when viewed from any vertex.

The same $n - \Theta(\sqrt{n})$ relationship holds for all three choices of rearrangement operation θ if we replace the diameter of $G_\theta(n)$ by the potentially smaller value

$$\min_{T \in B(n)} \max_{T' \in B(n)} \{d_\theta(T, T')\}.$$

Moreover, if two random trees \mathcal{T} and \mathcal{T}' are selected independently and uniformly at random from $B(n)$, then the expected distance between them (under any of the three operations) is described as follows:

$$\mathbb{E}[d_\theta(\mathcal{T}, \mathcal{T}')] = n - \Theta(n^{2/3}).$$

For further details, see [17].

What about the diameter of NNI? If we try the same lower bound trick with NNI (using the linear upper bound on the number of NNI neighbors of $T \in B(n)$ rather than the cubic upper bound), we still get that $\Delta[G_{\text{NNI}}(n)] = \Omega(n)$. Unlike SPR and TBR, however, this lower bound asymptotic can be improved. It turns out that $\Delta[G_{\text{NNI}}(n)] = \Theta(n \log n)$. This result, due to [228], follows directly from an unusual upper bound (also presented in that paper) for the number of trees at an NNI distance of, at most, k from any given T , namely $3^{n-1}4^{2k}$ (this upper bound was established by exploiting an elegant connection between NNI operations on binary phylogenies and flip operations on planar triangulations from [319]). Moreover, [228] also showed that $\Delta[G_{\text{NNI}}(n)] \leq n \log n + O(n)$ and so

$$\Delta[G_{\text{NNI}}(n)] = \Theta(n \log n).$$

Random walks in tree space. Tree rearrangement operations are used to “explore” the space of trees and form the basis of various search algorithms to locate an optimal tree for the given data under certain optimization criteria (discussed later). A natural question at this stage is what happens if we perform a simple random walk in $G_\theta(n)$ starting from some given tree T_0 . In other words, at each step, if we are at tree T , then we select one of the θ -neighbors of T uniformly at random and move to this tree.

Let $T_\theta(k)$ be the current tree after k steps of the walk. Thus $T_\theta(k)$ forms a discrete Markov chain with state space $B(n)$ (for $\theta \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$) and $RB(n)$ (for $\theta = \text{rSPR}$). Since this process is a simple random walk on a graph, it is a reversible Markov chain under the stationary distribution which assigns to each tree T a probability that is equal to the number of θ -neighbors it has, divided by twice the number of edges in the graph $G_\theta(n)$. In addition, $G_\theta(n)$ is a connected graph (for each θ), and for $n \geq 4$ there are walks of lengths 2 and 3 from any tree back to itself. Consequently, the Markov chain $T_\theta(k), k \geq 0$, is both irreducible and aperiodic, and so $T_\theta(k)$ converges to the stationary distribution described (as $k \rightarrow \infty$), regardless of the starting tree T_0 .

In the cases where $\theta = \text{NNI}$ or SPR , the graph $G_\theta(n)$ is regular, and so the stationary distribution corresponds precisely to the uniform distribution on $B(n)$. In other words, if one starts at any tree $T_0 \in B(n)$ and performs a large enough sequence of random NNI operations (or random SPR operations), then the probability of being at any given tree at the end of this sequence of operations is close to the uniform distribution that assigns each $T \in B(n)$ the probability $1/b(n)$. An obvious question is, how long does it take to be “close” to uniform; in particular, how does this time scale with n ?

This question was studied in [310] for a more restricted form of SPR in which the subtree being transferred was just a single leaf (this again led to a regular graph, again with a uniform limiting distribution). The “relaxation time” (roughly speaking, the number

of operations until one is ϵ -close to a uniform distribution on $B(n)$ for any given ϵ) grows with n at rate not faster than $O(n^2)$, confirming an earlier conjecture of David Aldous, who had established an $O(n^3)$ bound. A more recent paper [321] considered a modification of SPR and NNI on $RB(n)$ that also have uniform equilibrium distributions (the SPR analogue is slightly different from the usual rSPR one) and have established relaxation times of order $O(n^{5/2})$ for SPR and $O(n^4)$ for NNI. Further properties of $G_{rSPR}(n)$, and the behavior of random walks in this graph, have recently been investigated in [372], based on the concept of “Ricci–Ollivier curvature.”

2.6 • Consensus functions

Earlier in this chapter, we saw that the majority rule tree provides a way to summarize a set of different trees into a single tree. There are many other such “consensus” methods, and in this final section we explore this topic in more detail.

The mathematical study of consensus methods in biomathematics and group choice was pioneered by the mathematician Fred R. McMorris and colleagues from the 1970s onwards ([99] provides an nice overview). This led to numerous results showing how various consensus methods can be characterised by certain axiomatic properties, as well as “impossibility” theorems akin to Kenneth Arrow’s famous result in voting theory. Here we just give a brief taste of some aspects of consensus theory, including two recently studied consensus methods.

A *consensus function* for rooted phylogenetic X -trees is a function

$$\varphi : \bigcup_{k \geq 1} RP(X)^k \rightarrow RP(X).$$

Thus φ takes a k -tuple of rooted phylogenetic X -trees $\mathcal{P} = (T_1, \dots, T_k)$ and returns a single rooted phylogeny on the same leaf set. The sequence \mathcal{P} is often referred to as a *profile* of trees. Consensus functions on unrooted phylogenetic X -trees are defined similarly (i.e., as a function $\varphi : \bigcup_{k \geq 1} P(X)^k \rightarrow P(X)$).

Because of the equivalence between hierarchies on X and rooted phylogenies on X , we can also view a consensus function as one that takes a profile of rooted phylogenetic X -trees and returns a hierarchy on X . In the first part of this section, it is more convenient to adopt this more general viewpoint.

Given a profile $\mathcal{P} = (T_1, \dots, T_k) \in RP(X)^k$, let

- $\bigcup \mathcal{P} = \mathcal{H}_1 \cup \dots \cup \mathcal{H}_k$, where \mathcal{H}_i is the hierarchy associated with T_i ;
- for any set $A \in \bigcup \mathcal{P}$, $cf(A)$ be the proportion of i for which A is a cluster of T_i (cf refers to “concordance factor”).

The strict consensus method, mentioned earlier, is defined by the equation $\varphi_{SC}(\mathcal{P}) = \{A \in \bigcup \mathcal{P} : cf(A) = 1\}$ (i.e., the set of clusters present in every tree in \mathcal{P}). Notice that here (and in what follows) we are regarding the output hierarchy as equivalent to a rooted phylogenetic X -tree. The majority rule consensus method, also described earlier, is defined by $\varphi_{MR}(\mathcal{P}) = \{A \in \bigcup \mathcal{P} : cf(A) > 1/2\}$.

To define two further consensus methods (studied in [114, 205], and the references therein), we need some further notation: For two subsets A and B of X , let us write $A \perp B$ if $A \cap B \notin \{A, B, \emptyset\}$ (i.e., A and B are *incompatible* in the sense that there is no tree that contains both sets as clusters). In addition, for a tree $T \in RP(X)$, we write $A \perp T$ if $A \perp B$ for some cluster B of T .

- (1) The *majority (+) consensus* φ_{M+} : Given a profile $\mathcal{P} \in RP(X)^k$, let

$$\varphi_{M+}(\mathcal{P}) = \left\{ A \in \bigcup \mathcal{P} : cf(A) > \frac{\#j \in [k] : A \perp T_j}{k} \right\}.$$

In other words, this is the collection of clusters A that are present in more trees in the profile than there are trees that contain a cluster that is incompatible with A . This set system forms a hierarchy since if $A \perp B$ and A is present in m_1 trees, whereas B is present in m_2 trees (these sets of trees must be disjoint since $A \perp B$), then if $A \in \varphi_{M+}(\mathcal{P})$, then $m_1 > m_2$, whereas if $B \in \varphi_{M+}(\mathcal{P})$, then $m_2 > m_1$, hence both of these cannot hold.

- (2) The *frequency-difference consensus* φ_{FD} : Given a profile $\mathcal{P} \in RP(X)^k$, let

$$\varphi_{FD}(\mathcal{P}) = \left\{ A \in \bigcup \mathcal{P} : cf(A) > cf(B) \text{ for all } B \in \bigcup \mathcal{P} \text{ for which } B \perp A \right\}.$$

In other words, this is the set of clusters A that are present in more trees in that profile than is the case for any cluster that is incompatible with A . Once again this set system forms a hierarchy; otherwise, if $A \perp B$, and both A and B were in $\varphi_{FD}(\mathcal{P})$ then both $cf(A) > cf(B)$ and $cf(B) > cf(A)$ would have to hold, which is impossible.

These four methods are illustrated in Fig. 2.6.

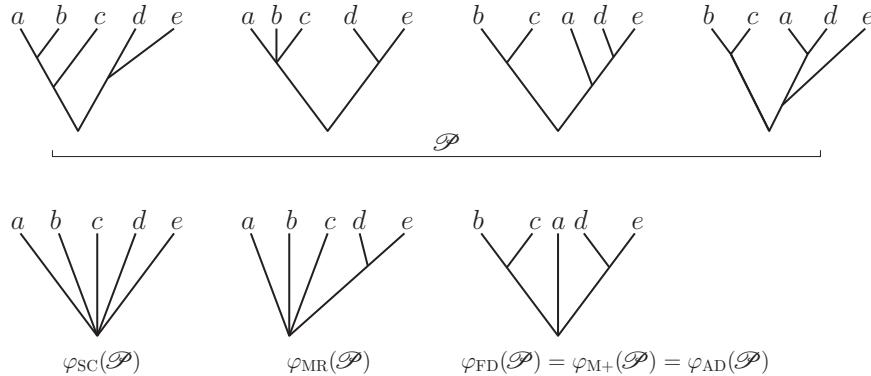


Figure 2.6. Top: A profile of four trees. Bottom: the associated consensus trees.

What is the relationship between these different consensus approaches? The following proposition shows that the hierarchies returned by these methods form a nested sequence, meaning that the associated consensus trees have the property that each refines the earlier consensus tree in the sequence.

Proposition 2.11. *For any profile \mathcal{P} , we have*

$$\varphi_{SC}(\mathcal{P}) \subseteq \varphi_{MR}(\mathcal{P}) \subseteq \varphi_{M+}(\mathcal{P}) \subseteq \varphi_{FD}(\mathcal{P}).$$

Proof. The first containment is clear: A set that is in all the hierarchies is, by default, in a majority of them. For the second containment, suppose that A is a cluster of the majority rule tree for \mathcal{P} (i.e., $A \in \varphi_{MR}(\mathcal{P})$). Then since no hierarchy in the profile can contain a

pair of incompatible sets from $\bigcup \mathcal{P}$, and since A lies in more than half the hierarchies, it follows that $cf(A) > \#\{j \in [k] : A \perp T_j\}$, so A is a cluster of the majority (+) consensus tree (i.e., $A \in \varphi_{M+}(\mathcal{P})$).

Finally, suppose that A is a cluster of the majority (+) consensus tree for \mathcal{P} (i.e., $A \in \varphi_{M+}(\mathcal{P})$). Suppose that $A, B \in \bigcup \mathcal{P}$ are incompatible (i.e., $B \perp A$). By definition, $cf(B) \leq \#\{j \in [k] : A \perp T_j\}$. Since A is a cluster of the majority (+) consensus tree, we have $cf(B) < \#\{j \in [k] : A \perp T_j\}$. Combining these two inequalities gives $cf(A) > cf(B)$. Since this holds for all choices of B from $\bigcup \mathcal{P}$ that are incompatible with A , it follows that A is a cluster in the frequency-difference consensus tree (i.e., $A \in \varphi_{FD}(\mathcal{P})$). ■

Another consensus method, *loose consensus*, is defined by $\varphi_{LC}(\mathcal{P}) = \mathcal{H}[\bigcup \mathcal{P}]$ (i.e., the clusters in at least one tree that are compatible with all other clusters; cf. eqn. (2.8)). This does not fit so cleanly into the total ordering of consensus methods in Proposition 2.11, since although it is clearly more refined than strict consensus, the loose consensus tree can contain clusters that are not in the majority rule tree. Nevertheless, $\varphi_{LC}(\mathcal{P}) \subseteq \varphi_{M+}(\mathcal{P})$.

Adams consensus. All of the consensus functions described so far have natural analogues for profiles of unrooted trees (replacing hierarchies with pairwise compatible split systems). However, there is a further consensus approach—the “Adams tree”—which relies heavily on the trees being rooted. Its description relies on the following notion: Given a sequence $(\Pi_1, \Pi_2, \dots, \Pi_k)$ of partitions of X , the associated *product partition*, denoted $\otimes_{i=1}^k \Pi_i$, is the partition of X in which the blocks are the sets that comprise the nonempty intersections of representative blocks of each partition. In other words, $\otimes_{i=1}^k \Pi_i$ is the set of all nonempty sets of the form $B_1 \cap B_2 \cap \dots \cap B_k$, where $B_i \in \Pi_i$ for $i = 1, \dots, k$.

Given a hierarchy \mathcal{H} on X , the maximal proper subsets of \mathcal{H} form a partition of X , which we can write as $\Pi(\mathcal{H})$. Thus, given a sequence $(\mathcal{H}_1, \dots, \mathcal{H}_k)$ of hierarchies on X , consider the product partition $\otimes_{i=1}^k \Pi(\mathcal{H}_i)$.

The *Adams consensus* of \mathcal{P} , $\varphi_{AD}(\mathcal{P})$, corresponds to the hierarchy on X that we describe recursively using a sequential partitioning process (starting with X and ending at singleton leaves). First, for the base case, $|X| = 1$, all the trees in \mathcal{P} consist of just one leaf x , in which case the Adams consensus returns the trivial hierarchy $\{\{x\}\}$. So now suppose that $|X| \geq 2$. Then X is the maximal set in $\varphi_{AD}(\mathcal{P})$, and the maximal proper subsets are the blocks B_1, \dots, B_m of $\otimes_{i=1}^k \Pi(\mathcal{H}_i)$. The remaining sets in $\varphi_{AD}(\mathcal{P})$ are obtained by continuing this approach on each block B_i by replacing $(\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_k)$ by

$$(\mathcal{H}_1|B_i, \mathcal{H}_2|B_i, \dots, \mathcal{H}_k|B_i),$$

where $\mathcal{H}_j|B_i = \{A \cap B_i : A \in H_j, A \cap B_i \neq \emptyset\}$, and continuing in this way until we arrive at the base case. For example, the Adams consensus of the four trees of \mathcal{P} in Fig. 2.6 is shown on the bottom right.

Adams consensus has a remarkable property. To explain this, recall the definition of least common ancestor (LCA) from Section 1.2.3. Given two nonempty subsets A and B of X , A is said to *nest in* B if the LCA of A in T is a strict descendant of the LCA of B in T . Then Adams consensus satisfies the following two nesting conditions:

- (A1) If A nests in B for each tree in the profile, then A nests in B in the Adams tree.
- (A2) If A and B are clusters of the Adams tree, and if A nests in B , then A nests in B in every tree in the profile.

For example, for the profile \mathcal{P} of Fig. 2.6, $\{d, e\}$ nests in $\{b, d\}$ for each tree in \mathcal{P} , and so, by (A1), this holds also for $\varphi_{AD}(\mathcal{P})$.

Now suppose that $ab|_{T_i} b$ holds for each tree T_i in the profile \mathcal{P} . Then since $\{a, b\}$ nests in $\{a, c\}$ for all the trees in the profile, property (A1) ensures that the Adams tree T' for the profile must also satisfy $ab|_{T'} b$. Let us summarize this as follows.

Proposition 2.12. *For a profile $\mathcal{P} = (T_1, \dots, T_k) \in RP(X)^k$, let $T = \varphi_{AD}(\mathcal{P})$. If $ab|_{T_i} c$ holds for all i , then $ab|_T c$ also holds.*

A partial converse also holds [61]: If $ab|_T c$ holds, then $ab|_{T_i} c$ holds for at least one i .

Proposition 2.12 has the following corollary: The strict consensus tree for any profile \mathcal{P} is equal to or strictly refined by the Adams tree for \mathcal{P} (in other words, $\varphi_{SC}(\mathcal{P})$ is a subset of $\varphi_{AD}(\mathcal{P})$). This is because if A is a proper cluster of each tree T_i in \mathcal{P} , then $aa'|_{T_i} b$ holds for all $a, a' \in A, b \in X - A$ and all i , and so (by Proposition 2.12) $aa'|_T b$ holds for $T = \varphi_{AD}(\mathcal{P})$, for all $a, a' \in A, b \in X - A$. Thus A is a proper cluster of T .

Exercise⁺: Consider the following property for a consensus method φ : For every profile \mathcal{P} of rooted phylogenies on X ,

$$\varphi(\mathcal{P}) \in RB(X) \Leftrightarrow \mathcal{P} = (T, T, \dots, T) \text{ for some } T \in RB(X).$$

Does strict consensus satisfy this property? What about majority rule consensus and Adams consensus?

Axioms and impossibilities. Proposition 2.12 is not satisfied by the other consensus methods described earlier. The condition in Proposition 2.12 that all input trees resolve a, b, c in the same way is rather restrictive. It is tempting to ask whether or not there is a consensus function φ that satisfies the following more general property:

- (*) For $\mathcal{P} = (T_1, \dots, T_k) \in RP(X)^k$ and $a, b, c \in X$, if (i) $ab|_{T_i} c$ for at least one i , and
(ii) there is no j with $ac|_{T_j} b$ or $bc|_{T_j} a$, then $ab|_T c$ where $T = \varphi(\mathcal{P})$.

In other words, if at least one tree supports $ab|c$ and none of the other trees support an alternative resolution, then the consensus tree should also support $ab|c$. However, there is no consensus function that satisfies property (*) in general. To see why, consider the four trees in $RP(5)$, each of which has one of the following corresponding nontrivial clusters: $\{1, 2\}$, $\{2, 3\}$, $\{3, 4\}$, and $\{4, 5\}$. Then property (*) would require $12|_T 5$, $23|_T 5$, $34|_T 1$, and $45|_T 1$, and there is no tree $T \in RP(5)$ that satisfies all four of these ternary relations (this can be verified by a case analysis, though it also follows easily from the BUILD algorithm of Chapter 4).

Each one of the consensus functions φ that we have so far considered satisfies the following two axiomatic properties.

[**Anonymity**] φ is invariant to any permutation of the trees in \mathcal{P} .

[**Neutrality**] φ is equivariant to any permutation of X .

Anonymity is the condition that for any profile $\mathcal{P} = (T_1, T_2, \dots, T_k)$ and any permutation ρ on $\{1, \dots, k\}$, we have $\varphi(\mathcal{P}) = \varphi(\rho(\mathcal{P}))$, where $\rho(\mathcal{P})$ is the profile $(T_{\rho(1)}, \dots, T_{\rho(k)})$. In other words, the order of the input trees does not matter.

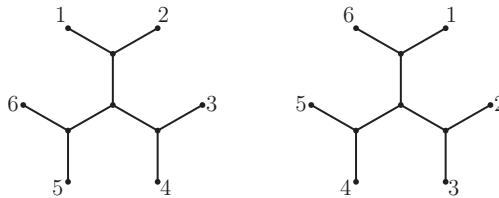


Figure 2.7. An example of a profile \mathcal{P} of two trees for which no consensus function simultaneously satisfies anonymity and neutrality and preserves the quaternary relationship. Notice that the tree on the right is obtained by rotating the leaves in the tree on the left by one-sixth of a full turn.

Neutrality is the condition that for any profile $\mathcal{P} = (T_1, T_2, \dots, T_k)$ and any permutation σ of X we have $\varphi(\mathcal{P}^\sigma) = \varphi(\mathcal{P})^\sigma$, where \mathcal{P}^σ is the profile $(T_1^\sigma, T_2^\sigma, \dots, T_k^\sigma)$, and where $T^\sigma \in RP(X)$ refers to the tree obtained from T by permuting the leaves according to σ . In other words, the names given to the objects labeling the leaves of the tree should not play any special role in the consensus function. For example, if “dog” and “cat” were interchanged in each input tree, then the output tree would just be the same consensus tree but with “dog” and “cat” interchanged.

As well as the two properties above, Adams consensus also preserves the ternary relation (Proposition 2.12) on rooted phylogenies. What about unrooted phylogenies? In this case the analogue of the ternary relation is the quaternary relation (described in Section 2.4.2). A consensus function φ on unrooted phylogenies preserves the quaternary relation if for every profile $\mathcal{P} = (T_1, \dots, T_k) \in P(X)^k$, the following condition holds:

(***) If $ab|_{T_i} cd$ holds for all i , then $ab|_T cd$ also holds, where $T = \varphi(\mathcal{P})$.

A natural question is whether there is some analogue of the Adams consensus that applies in the unrooted setting, and which satisfies anonymity, neutrality, and preserves the quaternary relation (**). It turns out that no such consensus method exists, as the following “impossibility” result.

Proposition 2.13. For $|X| \geq 6$ there is no consensus function which satisfies anonymity and neutrality and preserves the quaternary relationship (**), even when restricted to pairs of trees.

Proof. Suppose there were such a consensus function φ . Let $\mathcal{P} = (T_1, T_2)$ be the profile consisting of the two trees from $B(6)$ shown in Fig. 2.7. In this case, $12|_{T'} 45, 34|_{T'} 16$ and $56|_{T'} 23$ hold for both $T' = T_1$ and $T' = T_2$, so for the tree $T = \varphi(\mathcal{P})$, we must also have $12|_T 45, 34|_T 16$ and $56|_T 23$. However, it turns out that only two trees in $B(6)$ satisfy this property, namely the two trees in \mathcal{P} . Thus $\varphi(\mathcal{P})$ must be one of the two trees in \mathcal{P} . Suppose it is T_i . Notice that the cyclic permutation $\sigma = (123456)$ gives a permutation ρ that interchanges T_1 and T_2 so, by the anonymity condition, $\varphi(\mathcal{P}^\sigma) = \varphi(\rho(\mathcal{P})) = \varphi(\mathcal{P}) = T_i$. However, by the neutrality condition, we have $\varphi(\mathcal{P}^\sigma) = T_i^\sigma \neq T_i$, which provides the required contradiction. It is straightforward to extend this example to cases where $|X| > 6$. ■

A different and much earlier Arrow-type theorem for consensus on unrooted tree was discovered by F.R. McMorris [249]. A very recent axiomatic treatment of consensus (and “supertree”) functions appears in [247].

Chapter 3

Tree shape and random discrete phylogenies

3.1 • Tree shapes

If we ignore the labeling of the leaves of a rooted or unrooted phylogenetic tree, we obtain a “tree shape.” For example, when $n = 4$, there are two rooted binary tree shapes: the fork tree shape and the caterpillar tree shape, shown in Figs. 3.1 (a) and (b). Biologists are interested in the shapes of trees, since they shed light on the process of speciation and extinction in evolution (a topic we will explore further in Chapter 9).

Elementary group theory provides a nice trick to count the number of phylogenetic X -trees of a given shape using the “orbit-stabilizer theorem.” Given a finite group G , which acts on a set S , and an element $s \in S$, let

$$O(s) = \{g \cdot s : g \in G\} \subseteq S$$

denote the orbit of s under the action of G , and let

$$\text{Stab}(s) = \{g \in G : g \cdot s = s\} \subseteq G$$

be the stabilizer subgroup of G . The orbit-stabilizer theorem then provides a bijection between the orbit of s and the cosets of $\text{Stab}(s)$ in G , so, in particular,

$$|O(s)| = \frac{|G|}{|\text{Stab}(s)|}. \quad (3.1)$$

There is a natural action of the symmetric group Σ_n of permutations on $[n]$ on the set $RP(n)$: given $\sigma \in \Sigma_n$, simply permute the leaves of each tree T by replacing leaf x by leaf $\sigma(x)$. This action restricts to an action on the set $RB(n)$ of rooted binary trees, and so, by (3.1), the number of trees in $RB(n)$ that have the same shape as some tree T is $n! / |\text{Stab}(T)|$. Now $\text{Stab}(T)$ is just the symmetry group of T (i.e., $S(T)$ from Section 1.2.4). In particular, for rooted binary trees, this is a group of order $2^{s(T)}$, where $s(T)$ is the number of *symmetry vertices* of T —these are interior vertices for which the two subtrees of T that the vertex separates from the root have the same shape. For example, for a phylogenetic tree with the fork tree shape in Fig. 3.1(a), T has three symmetry vertices, and so $\text{Stab}(T)$ is a group of order 8. This group is isomorphic to the dihedral group of rotational and reflectional symmetries of a square, as illustrated in Figs. 3.1 (c) and (d). In particular, for any set X of size 4, there are precisely $4! / 2^3 = 3$ rooted binary X -trees that have the shape of the fork tree; by contrast, the caterpillar tree (Fig. 3.1(b)) has only one symmetry vertex, and so there are 12 rooted binary phylogenetic X -trees of this shape. For unrooted

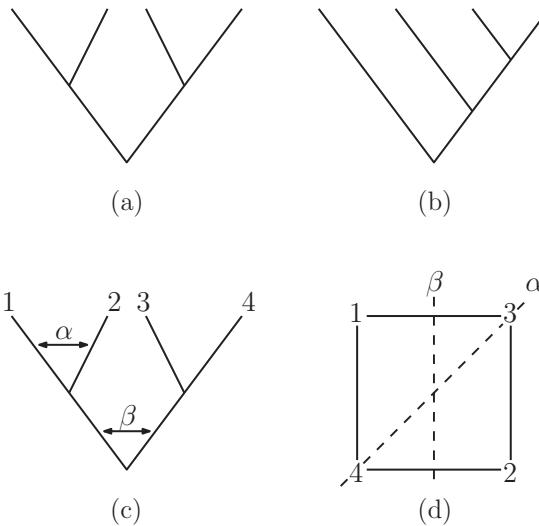


Figure 3.1. Top: The two tree shapes for rooted binary trees on four leaves: (a) the fork and (b) the caterpillar tree shape. The stabilizer subgroup of a phylogenetic tree having the fork shape is isomorphic to the dihedral group of symmetries of a square. Bottom: The two symmetries shown (α and β in (c)) correspond to reflections in (d) that generate this group.

binary trees and nonbinary trees, similar formulae apply, though more complex symmetries arise, as described in Section 1.2.4. For example, an unrooted binary tree can have a twofold symmetry about a central edge and a symmetry of order 3! about a central vertex. Only the size of the symmetry group is important for us here, rather than a detailed description of its algebraic structure (which relies on the notion of wreath product).

Notice that the number of symmetry vertices in a rooted binary tree with n leaves is, at most, the number of powers of 2 in $n!$. Moreover, at least one rooted binary tree achieves this bound, by a simple parity argument: the number of binary phylogenetic trees of shape T is $n!/2^{s(T)}$, and the sum of the numbers of equivalent-shaped trees in $RB(n)$ is equal to the size of $RB(n)$ which is $(2n-3)!!$, which, in turn, is an odd number, so there must be an odd number (and hence at least one) phylogeny T for which $n!/2^{s(T)}$ is odd.

We end this first section by considering the number \tilde{r}_n of rooted binary tree shapes on n leaves. We saw in Chapter 2 that if the leaves of these shapes are labeled (to give the trees in $RB(n)$), there is a simple and explicit formula for the size of this class, namely $r b(n) = (2n-3)!!$. However, for \tilde{r}_n , one must be content with recursions and asymptotic results, which can be derived from a functional equation. To describe this, let

$$\tilde{r}(x) = \sum_{n \geq 1} \tilde{r}_n x^n = x + x^2 + x^3 + 2x^4 + 3x^5 + 6x^6 + 11x^7 + \dots,$$

which is the (ordinary) generating function for the numbers \tilde{r}_n (called the Wedderburn–Etherington numbers). Then $\tilde{r}(x)$ is fully determined by the following equation:

$$\tilde{r}(x) = x + \frac{1}{2}(\tilde{r}^2(x) + \tilde{r}(x^2)). \quad (3.2)$$

To see why (3.2) holds, consider the three terms on the right-hand side. The term x accounts for the value $\tilde{r}_1 = 1$. For $n \geq 2$, a rooted tree shape has two subtree shapes adjacent to the root, and the order of the subtrees is irrelevant. Thus, $\tilde{r}^2(x)$ counts all those tree shapes in which the two shapes are not isomorphic twice and those in which the subtrees are isomorphic once, whereas $\tilde{r}(x^2)$ counts those tree shapes in which the two subtrees are isomorphic.

An extension of eqn. (3.2) to count (rooted and unrooted) “multilabeled trees” (a class that includes both binary tree shapes and binary phylogenetic trees as extreme cases) was recently presented in [93].

Another related problem is to count the number of “tanglegrams” up to equivalence [30]. Roughly speaking, a tanglegram can be viewed as an ordered pair of trees T and S in which the leaves are matched up (these are used in comparing trees in the study of “co-phylogeny”). More precisely, a *tanglegram* is a triple (T, σ, S) , where $T, S \in RB(n)$ and σ is a permutation of $[n]$, and where an edge is placed between leaf i in T and leaf $\sigma(i)$ in S for all $i \in [n]$. To count the “shapes” of tanglegrams we regard (T, σ, S) and (T', σ', S') as having equivalent shapes if T and T' have the same shape, S and S' have the same shape, and for some rooted graph isomorphism $\pi : T \rightarrow T'$ and $\pi' : S' \rightarrow S$ we have $\sigma(x) = \pi'(\sigma'(\pi(x)))$ for all $x \in X$. If t_n denotes the number of resulting equivalence classes, then $t_3 = 2$ (the two possibilities are shown in Fig. 3.2), while $t_4 = 13$. An elegant exact expression for t_n is derived in [30], along with the asymptotic equivalence

$$\frac{t_n}{n!} \sim \frac{e^{\frac{1}{8}} 4^{n-1}}{\pi n^3}.$$

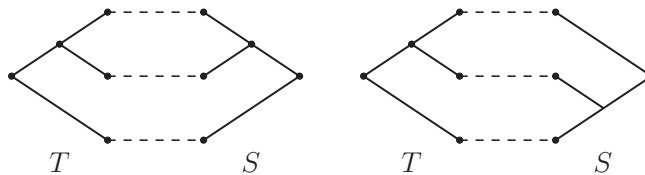


Figure 3.2. The two tanglegram shapes for $n = 3$.

3.2 • The shape of evolving trees

As well as speciation, extinction has also played a major role in the history of life; after all, most species are extinct.¹⁰ Suppose we sample some subset X of species that are present today (species a – e in Fig. 3.3(i)) and then consider the minimal tree linking these species. This results in the so-called “reconstructed tree” illustrated in Fig. 3.3(ii). Let us view this as a rooted phylogenetic X -tree (ignoring the length of the edges). It turns out that, under very general assumptions concerning the speciation-extinction process, many models predict an identical and simple discrete probability distribution on $RB(X)$ [225] (this process will play an important role in Chapter 9). Moreover this discrete probability distribution can be easily described and is called the *Yule-Harding (YH) model* (or distribution).

¹⁰It is estimated that current plant and animal diversity preserves at most 1–2% of the species that have existed over the past 600 million years [128].

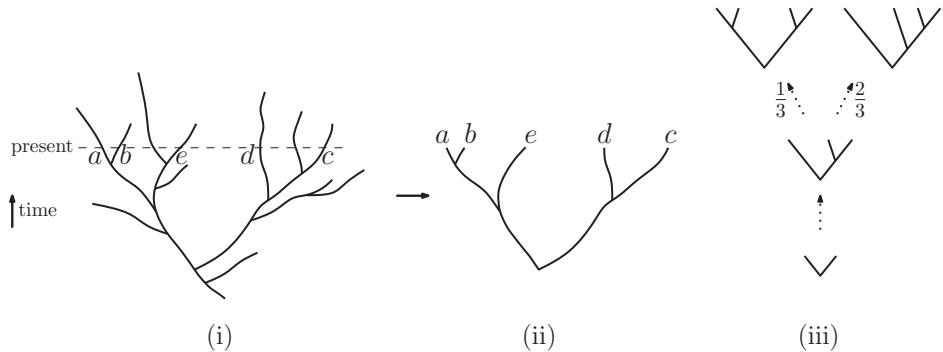


Figure 3.3. (i) A birth-death tree showing speciation and extinction. (ii) The associated discrete “reconstructed tree.” (iii) Growing a tree by the YH process.

To obtain a binary tree shape under the YH model, we start with a tree shape on two leaves and sequentially attach leaves, attaching a new leaf at each step to one of the leaf edges chosen uniformly at random from the tree constructed so far. For example, the probabilities of generating the fork and caterpillar tree shapes are $\frac{1}{3}$ and $\frac{2}{3}$, respectively, since from the (unique) tree shape on three leaves, we can attach a new leaf to exactly one of the three leaf edges to obtain a fork tree shape or to any two of these leaf edges to obtain a caterpillar tree shape (see Fig. 3.3(iii)).

Once we have built up a tree with n leaves in this way, we obtain a random tree shape on n leaves, and we can now label the leaves of this tree shape according to a permutation on $\{1, 2, \dots, n\}$ chosen uniformly at random. This is the YH probability distribution on $RB(n)$.

We now explain how to compute the probability of a YH tree shape and that of any rooted phylogenetic tree with this shape. First, let us grow a tree under the YH process until it has n leaves and then randomly select one of the two subtrees incident with the root (say, the “left-hand one,” since the orientation in the plane plays no role) and let Z_n denote the number of leaves in this tree. Remarkably, Z_n has a completely flat distribution.

Lemma 3.1. Z_n has a uniform distribution between 1 and $n - 1$, so

$$\mathbb{P}(Z_n = i) = \frac{1}{n-1}, \text{ for } i = 1, \dots, n-1.$$

Proof: The random process Z_1, Z_2, \dots can be exactly described as a special case of a classical process in probability called *Pólya’s urn*. This consists of an urn that initially has a blue balls and b red balls. At each step, a ball is sampled uniformly at random and returned to the urn along with another ball of the same color. In our setting, $a = b = 1$, and “blue” corresponds to the left-hand subtree and “red” to the right-hand subtree in the YH tree. At each step, the uniform process of leaf attachment ensures that Z_n has exactly the same probability distribution as the number of blue balls in the urn after $n - 2$ steps. It is well known, and easily shown by induction, that in Pólya’s urn with $a = b = 1$, the proportion of blue balls has a uniform distribution. ■

Lemma 3.1 provides the key to computing the YH probability of a tree.

Proposition 3.2. *For any particular tree $T \in RB(n)$, the probability $\mathbb{P}_{\text{YH}}(T)$ of generating T under the YH model is given by*

$$\mathbb{P}_{\text{YH}}(T) = \frac{2^{n-1}}{n! \prod_{v \in \overset{\circ}{V}(T)} \lambda_v},$$

where $\overset{\circ}{V}(T)$ is the set of interior vertices of T and where λ_v is number of leaves of T that are descendants of v , minus 1.

Proof: Suppose that the two maximal subtrees T_1 and T_2 of T are of size k and $n-k$, where we may assume that $2k \leq n$. By Lemma 3.1, the probability of such a size distribution is $2/(n-1)$ if $2k < n$ and $1/(n-1)$ if $2k = n$. Conditional on this division, the number of ways to select leaf sets for T_1 and T_2 that partition $[n]$ is $\binom{n}{k}$ when $2k < n$, and $\frac{1}{2} \binom{n}{k}$ when $2k = n$ (the factor of $\frac{1}{2}$ recognizes that the order of T_1 and T_2 is interchangeable in T when they have the same number of leaves). By the Markovian nature of the YH process, each of these two subtrees also follows the YH distribution. This leads to the recursion

$$\mathbb{P}_{\text{YH}}(T) = \frac{2}{n-1} \binom{n}{k}^{-1} \mathbb{P}_{\text{YH}}(T_1) \mathbb{P}_{\text{YH}}(T_2),$$

from which Proposition 3.2 now follows by induction. ■

To illustrate Proposition 3.2, consider the tree in Fig. 1.4(a). Then we have $\mathbb{P}_{\text{YH}}(T) = \frac{2^4}{5! \times 4 \times 3 \times 1^2} = \frac{1}{90}$, while the tree in Fig. 3.3(ii) gives $\mathbb{P}_{\text{YH}}(T) = \frac{1}{60}$.

Exercise: Find a general formula for the probability that a random tree \mathcal{T} in $RB(n)$ generated by the YH model has the shape of a rooted caterpillar tree. What is the probability that $\mathcal{T} = T$ for a particular caterpillar tree $T \in RB(n)$?

Curiously, a quite different process that arises in population genetics, and which proceeds backward in time (rather than forward, as in Fig. 3.3(iii)), also leads to the YH distribution when we ignore the length of the edges and the associated ranking of interior vertices. This is the celebrated *coalescent* process most usually associated with Sir John Kingman and which was developed in the early 1980s. We will describe this process in continuous time briefly in Chapter 9, but as a discrete process, the coalescent starts with the set X and selects uniformly at random a pair of elements to join (these form the “cherry” of the tree that is closest to the leaves). These two leaves are then regarded as a single element in a set of size $|X| - 1$, and the process is repeated. This discrete coalescent process generates a ranked binary phylogeny, which is often referred to as a *labeled history*. This consists of a pair (T, r) , where $T \in RB(X)$ and r is a *ranking* of the interior vertices of T , that is, a bijective function $r : \overset{\circ}{V}(T) \rightarrow \{-1, -2, \dots, -(n-1)\}$ with the property that if u is a descendant of v , then $r(u) > r(v)$ (thus the root is the vertex assigned an r value of $-(n-1)$). The function r describes the order in which the coalescent events occur, so the first cherry to form in the process has rank -1 , for instance.

By symmetry, the discrete coalescent process generates each pair (T, r) with equal probability, so the probability of generating any particular pair is simply 1 divided by the number of such pairs. Counting such pairs is easier than it might first seem, since we have $\binom{n}{2}$ choices for the first coalescence, then $\binom{n-1}{2}$ for the second, and so forth. Consequently,

the total number of such pairs (T, r) is just

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

Suppose we select a ranked phylogeny (T', r) uniformly at random and then ignore r , to thereby focus just on T' . What is the probability that we produce a particular tree $T \in RB(X)$? This is simply the number of rankings of T divided by the number of pairs (T', r) . Fortunately, there is an exact formula for the number of rankings of any rooted phylogenetic X -tree T , even if it is nonbinary. There are precisely

$$\frac{|\overset{\circ}{V}(T)|!}{\prod_{v \in \overset{\circ}{V}(T)} \lambda_v}$$

such rankings, where λ_v is the number of interior vertices v' of T that are descended from v (including v itself). This expression follows from classical results in enumerative combinatorics for counting the number of linear extensions of particular posets (for details, see [315] and the references therein). For rooted binary trees, the formula simplifies to $\frac{(n-1)!}{\prod_{v \in V(T)} \lambda_v}$, and in this case λ_v is also just the number of leaves descended from v minus 1, as in Proposition 3.2.

Consequently, the probability that the coalescent process produces a given T in $RB(X)$ is

$$\frac{\# \text{rankings } r \text{ for } T}{\#(T', r)} = \frac{(n-1)!}{\prod_{v \in \overset{\circ}{V}(T)} \lambda_v} \Bigg/ \frac{n!(n-1)!}{2^{n-1}}.$$

This simplifies to the expression $\frac{2^{n-1}}{n! \prod_{v \in \overset{\circ}{V}(T)} \lambda_v}$ for $\mathbb{P}_{YH}(T)$ in Proposition 3.2. In summary, we see formally that another distribution (selecting a labeled history and ignore the ranking) leads to the YH distribution.

3.2.1 ■ The big picture

We have seen how the YH distribution can be realized either by a uniform distribution on ranked phylogenies (forgetting the ranking) or by a random process of speciation. These links are shown by the arrows labeled E and F in Fig. 3.4. However, there are further links to additional classes of trees (the boxes on the right in Fig. 3.4), made explicit by Amaury Lambert and Tanja Stadler [225]. This requires introducing two new notions: “ranked oriented trees,” and “unlabeled ranked trees.”

A *ranked oriented tree* is a rooted ranked binary tree shape in which, for each interior vertex v , the two edges directed away from v receive an orientation (left or right). Thus the leaves of such a tree can be viewed as being labeled $0, 1, \dots, n-1$, ordered from left to right.

Remarkably, the set of ranked oriented trees on n leaves is in bijective correspondence with the set of permutations on the numbers $1, \dots, n-1$. To see how, place the numbers $0, 1, \dots, n-1$ in ascending order along a horizontal axis, draw vertical lines l_1, \dots, l_{n-1} of different lengths below $1, \dots, n-1$, and place a vertical line of infinite length below 0. Next, for each line l_i , $i \geq 1$, draw a horizontal line to the left of l_i until it intersects the first vertical line it meets (i.e., the largest $j < i$ for which $l_j > l_i$). In this way, we

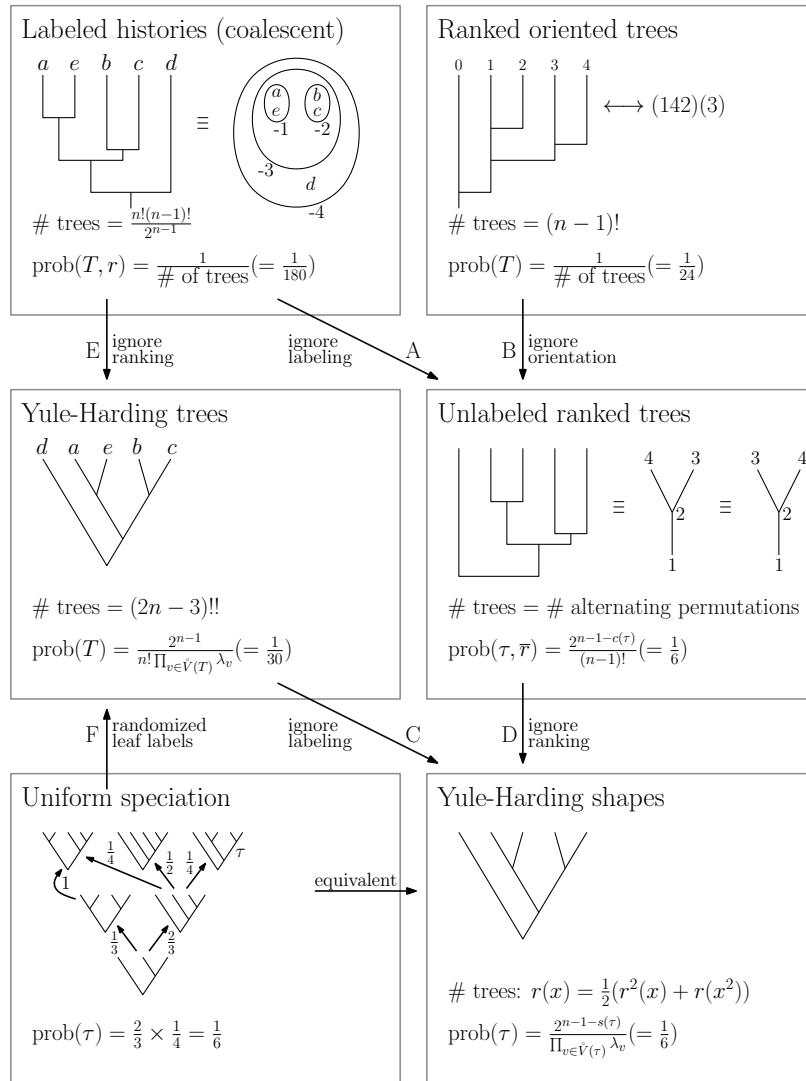


Figure 3.4. Relationships between different classes of trees: The left-hand column shows how the YH tree can be realized either by a coalescent process (backwards in time) or a uniform speciation process (forward in time). The processes in the right column provide linkages to further processes (see text for details). The probability of each tree shown is given in parentheses.

obtain a ranked oriented tree on leaf set $\{0, 1, \dots, n-1\}$; moreover, the set of such trees is purely determined by the relative sizes of the lines l_1, \dots, l_{n-1} , so we obtain the promised bijection from the set of ranked oriented trees to the set of permutations on $\{1, \dots, n-1\}$. For example, in Fig. 3.4 the tree in the top right panel has the ordering $l_2 < l_4 < l_3 < l_1$ and so corresponds to the permutation $2 \mapsto 1, 4 \mapsto 2, 3 \mapsto 3, 1 \mapsto 4$ (i.e., $(142)(3)$ in disjoint cycle notation).

Thus there are $(n-1)!$ ranked oriented trees with n leaves, and the uniform distribution on this class selects any given representative with probability $1/(n-1)!$. This

construction of ranked oriented trees (where it arises via the so-called “coalescent point process”) will be explored further in Chapter 9.

We now describe the second class, “unlabeled ranked trees,” and we see that we get the same distribution on this class if we take either the uniform distribution on ranked oriented trees and ignore the orientation, or the uniform distribution on labeled histories and ignore the leaf labeling (cf. Fig. 3.4). To describe this class and its associated probability distribution, given a binary tree shape τ , consider the set of rankings r of the interior vertices of τ (where, as before, $r(v) < r(v')$ whenever v' is a proper descendant of v in τ). Let us regard two rankings r and r' as equivalent if there is a symmetry of τ that maps r to r' . An *unlabeled ranked binary tree* is then a pair (τ, \bar{r}) consisting of a binary tree shape τ together with an equivalence class \bar{r} of a ranking.

Let \mathcal{R}_n denote the set of unlabeled ranked binary trees on n leaves and let $R_n = |\mathcal{R}_n|$; for example, $R_4 = 2$ (the rooted caterpillar has a unique ranking, and for the fork tree the two possible rankings are equivalent since the leaves are unlabeled). The sequence R_n for $n \geq 2$ is 1, 1, 2, 5, 16, 61, 272, This sequence of numbers is described by a delightful formula involving trigonometric functions. To see this, set $E_n = R_{n+1}$ for all $n \geq 1$ and set $E_0 = 1$. Thus the n in E_n counts the number of interior vertices of the tree rather than the number of leaves. Now, if we remove the root of a binary tree, we get two rooted binary trees (one of which might be a leaf with no interior vertex), and by counting how we can distribute the n interior vertex rankings between the two trees (noting that the root is always ranked first), we get the following recurrence:

$$E_{n+1} = \frac{1}{2} \sum_{k=0}^n \binom{n}{k} E_k E_{n-k}. \quad (3.3)$$

Here the $\frac{1}{2}$ factor is because the trees are unlabeled, so we have a twofold symmetry about the root, and $k = 0, k = n$ covers the case where a leaf is adjacent to the root; also, the ranking on one subtree incident with the root does not at all restrict the ranking on the other subtree.

Consider the (exponential) generating function

$$E(x) = \sum_{n \geq 0} E_n \frac{x^n}{n!}.$$

The recursion (3.3) can be restated as

$$\frac{dE(x)}{dx} = \frac{1}{2}(E(x)^2 + 1),$$

which is a separable differential equation that has the solution

$$E(x) = \tan\left(\frac{x}{2} + C\right)$$

for a constant C . The initial condition $E(0) = E_0 = 1$ gives $C = \pi/4$, and a standard trigonometric identity leads to the final expression

$$E(x) = \tan(x) + \sec(x).$$

Notice that the odd powers of x are contributed by \tan and the even terms by \sec . From this equation, asymptotic results can be derived for E_n and hence R_n , as well as expressions for R_n in terms of Euler numbers and Bernoulli numbers.¹¹

¹¹The coefficients in $E(x)$ also appear in a classic problem in combinatorics of counting “alternating permutations,” and a specific bijection between these permutations and a class of rooted trees is described in [113].

However, we are interested not in the uniform distribution on \mathcal{R}_n but in the distribution on this set induced by selecting either a ranked oriented tree uniformly at random and ignoring the ranking, or a labeled history uniformly at random and ignoring the leaf labeling (arrows labeled A and B in Fig. 3.4). These two distributions on \mathcal{R}_n coincide [225], as we now state more formally.

Proposition 3.3. *Let \mathcal{T} and \mathcal{T}' be random unlabeled ranked trees obtained as follows: For \mathcal{T} select a labeled history on n leaves uniformly at random and ignore the leaf labeling. For \mathcal{T}' select a ranked oriented tree from \mathcal{R}_n uniformly at random and ignore the leaf labeling. Then \mathcal{T} and \mathcal{T}' have the same probability distribution, with*

$$\mathbb{P}(\mathcal{T} = (\tau, \bar{r})) = \mathbb{P}(\mathcal{T}' = (\tau, \bar{r})) = \frac{2^{n-1-c(\tau)}}{(n-1)!},$$

where $c(\tau)$ is the number of cherries in τ .

Proof: First, suppose that a ranked oriented tree is selected uniformly at random. For each interior vertex v , there are then two choices as to the orientation of the two sister lineages, and these choices lead to different trees, except where v is adjacent to the leaves of a cherry (note that the ranking of the interior vertices of the tree destroys other symmetries that would be normally counted). Now if a shape τ of the ranked oriented tree chosen has $c(\tau)$ cherries, then it has $n-1-c(\tau)$ interior vertices that are not incident with cherries, and so the probability of obtaining a given unlabeled ranked tree $(\tau, \bar{r}) \in \mathcal{R}_n$ by this process is $2^{n-1-c(\tau)} \cdot \frac{1}{(n-1)!}$. The alternative route (starting from a uniformly sampled labeled history and ignoring the leaf labeling) gives the same probability for (τ, \bar{r}) , since each labeled history has probability $\frac{2^{n-1}}{n!(n-1)!}$, and the number of these that produce the same unlabeled ranked tree (τ, \bar{r}) is $\frac{n!}{2^{c(\tau)}}$ (by the orbit-stabilizer theorem); the product of these two expressions again gives $\frac{2^{n-1-c(\tau)}}{(n-1)!}$. ■

Arrows C and D in Fig. 3.4 also give rise to the same probability distribution on Yule-Harding shapes; this is formalized as follows.

Proposition 3.4. *Let \mathcal{T} and \mathcal{T}' be random rooted binary tree shapes obtained as follows: For \mathcal{T} select a Yule-Harding tree on n leaves and ignore the leaf labeling. Let \mathcal{T}' select an unlabeled ranked tree according to the distribution described in Proposition 3.3, and ignore the ranking. Then \mathcal{T} and \mathcal{T}' have the same probability distribution, with*

$$\mathbb{P}(\mathcal{T} = \tau) = \mathbb{P}(\mathcal{T}' = \tau) = \frac{2^{n-1-s(\tau)}}{\prod_{v \in \overset{\circ}{V}(\tau)} \lambda_v}.$$

Proof: $\mathbb{P}(\mathcal{T} = \tau)$ is simply the sum of $\mathbb{P}_{\text{YH}}(\mathcal{T} = T)$ over all $T \in \text{RB}(n)$ that have shape τ , and so, by Proposition 3.2, $\mathbb{P}_{\text{YH}}(\mathcal{T} = T) = \frac{2^{n-1}}{n! \prod_{v \in \overset{\circ}{V}(T)} \lambda_v}$. Notice that this depends only on the shape τ , and the number of trees T of this shape is $n!/2^{s(\tau)}$. Multiplying these two quantities together gives the expression in Proposition 3.4. To see that \mathcal{T}' has the same distribution, notice that \mathcal{T}' is the result of selecting a labeled history uniformly at random and ignoring the labeling (arrow A in Fig. 3.4) and then ignoring the ranking (arrow D). But this process is identical to selecting a labeled history uniformly at random and ignoring the ranking (arrow E) and then ignoring the labeling (arrow C), which is the distribution of \mathcal{T} . ■

3.2.2 • Properties of the YH and uniform models

Notice that the YH process leads to a different probability distribution on $RB(n)$ from that obtained by simply selecting a tree uniformly at random from $RB(n)$, which would assign each tree $T \in RB(n)$ probability $1/(2n-3)!!$. We call this later distribution the *uniform model*, though biologists sometimes refer to it as the “proportional-to-distinguishable arrangements” model.

To see that this is different from the YH distribution, observe that the probability of obtaining a tree with the fork shape in Fig. 3.1(a) is equal to $\frac{1}{5}$ under a uniform distribution on $RB(4)$ (since only 3 of the 15 trees in $RB(4)$ have that shape) and $\frac{1}{3}$ under YH. There are several further differences between the uniform and YH distributions. For instance, in YH trees on n leaves, the expected number of edges between the root and a randomly selected leaf grows at the rate $\log(n)$, while for uniform binary trees, it grows at the rate \sqrt{n} .

YH trees also tend to be more “balanced” than uniform trees, where balance refers to the average difference between the sizes of the two daughter subtrees in the tree, as one ranges over the interior vertices of the tree. For example, Proposition 3.2 shows that the probability that a tree with n leaves generated under the YH distribution has a single leaf adjacent to the root is $\frac{2}{n-1}$; for the uniform distribution, the corresponding probability is $\binom{n}{1} \frac{rb(n-1)}{rb(n)} = \frac{n}{2n-3}$, which converges to $\frac{1}{2}$ as n grows. It turns out that many “real” phylogenetic trees tend to have a degree of balance somewhere between that predicted by the YH and uniform distributions. Several interesting new mathematical and statistical insights have helped explain this phenomenon [3, 33, 173, 225].

For a phylogenetic model θ (e.g., $\theta = \text{YH}$ or $\theta = \text{U}$ (= uniform)), a central question is the following: What is the probability $p_\theta(A)$ that a given subset A of X is a cluster of a randomly generated tree $\mathcal{T} \in RB(X)$? For the YH or the uniform model, this probability depends only on $n = |X|$ and the size k of A . For the uniform model, the probability is

$$p_U(A) = \frac{rb(k) \cdot rb(n-k+1)}{rb(n)} = \frac{\binom{n-1}{k-1}}{\binom{2n-2}{2k-2}}. \quad (3.4)$$

To see why the first equality in eqn. (3.4) holds, we need to count the subset S_A of trees in $RB(X)$ that have A as a cluster. Let x_A be a new label that is not in X . One then has a bijection from $RB(A) \times RB((X-A) \cup \{x_A\})$ to S_A , in which (T, T') is mapped to the tree obtained by attaching T' to T by the process of identifying the root of T' with the vertex x_A . The second equality in eqn. (3.4) is straightforward algebra using eqn. (2.2).

For the YH model, the corresponding probability $p_{\text{YH}}(A)$ is even simpler:

$$p_{\text{YH}}(A) = \frac{2n}{k(k+1)} \binom{n}{k}^{-1} \quad (3.5)$$

for each $k = 1, \dots, n-1$. To justify this identity, let $Y_n(k)$ be the number of proper subsets of X of size k that are clusters of a YH tree on n leaves. An inductive argument (on n) using Lemma 3.1 shows that

$$\mathbb{E}[Y_n(k)] = \frac{2n}{k(k+1)}, \text{ for } 1 \leq k \leq n-1. \quad (3.6)$$

Now, for $1 \leq k \leq n - 1$, each subset A of X of size k has the same probability of being a cluster under the YH model, so

$$p_{\text{YH}}(A) = \sum_{m \geq 0} \mathbb{P}(\mathcal{T} \text{ has } m \text{ clusters of size } k) \cdot \frac{m}{\binom{n}{k}} = \mathbb{E}[Y_n(k)] \binom{n}{k}^{-1},$$

from which eqn. (3.5) follows from eqn. (3.6).

The probability that A is a cluster in a YH tree is identical to the probability that A is a cluster in a tree produced by Kingman's coalescent process from population genetics (this follows by the linkages described earlier and illustrated in Fig. 3.4). In this way, eqn. (3.5) and its extensions to pairs of clusters form a null model in which to test hypotheses of the monophyly of species in statistical phylogenetics [303].

Notice that if we generate a YH tree \mathcal{T} with n leaves and select one of the $n - 1$ interior vertices of \mathcal{T} uniformly at random, then the probability that the cluster associated with this vertex is a particular set $A \in \binom{X}{k}$ is $\frac{1}{n-1} p_{\text{YH}}(A)$. In particular, the probability that a non-singleton cluster chosen uniformly from a YH tree has size k is precisely $\frac{1}{n-1} \binom{n}{k} p_{\text{YH}}(A) = \frac{2^n}{(n-1)k(k+1)}$.

Notice also that although the size of a (randomly chosen) subtree of a YH tree \mathcal{T} incident with the root has a uniform distribution, if we select a leaf x from X and then generate a YH tree \mathcal{T} and ask for the size \tilde{Z}_n of the subtree incident with the root that contains x , then this distribution favors larger subtrees. This is just a simple consequence of Bayes' rule, since $\mathbb{P}(\tilde{Z}_n = i)$ is the conditional probability $\mathbb{P}(Z_n = i | x \in Z_n)$ from which it can readily be shown that

$$\mathbb{P}(\tilde{Z}_n = i) = \frac{2^i}{n(n-1)}. \quad (3.7)$$

Exercise: Prove eqn. (3.7), using Lemma 3.1.

Various other quantities of interest can be readily derived for the YH process and we mention just two here. First, suppose we select an element x from X and then generate a YH tree $\mathcal{T} \in RB(X)$. Let M_n be the size of the minimal cluster of \mathcal{T}_n that strictly contains x (thus M_n is the number of leaves descended from the parent vertex of x). The distribution of M_n under the YH model has been investigated in [31] (in the context of the equivalent coalescent model), obtaining the following inverse cubic power law for $k = 2, \dots, n - 1$:

$$\mathbb{P}_{\text{YH}}(M_n = k) = \frac{4}{k(k^2 - 1)}, \quad (3.8)$$

which turns into an inverse square at $k = n$, where $\mathbb{P}_{\text{YH}}(M_n = n) = \frac{2}{n(n-1)}$. This last equality is just eqn. (3.7) for the value $i = 1$, and we can derive the less obvious eqn. (3.8) by again using eqn. (3.5). Select a leaf x . Now, $\mathbb{P}_{\text{YH}}(M_n = k)$ is the probability that \mathcal{T} has a cluster A of size k that contains x , and with x adjacent to the interior vertex v of T associated with A (i.e., $c_T(v) = A$). This probability is simply

$$\sum_{A \subseteq \binom{X}{k}: x \in A} \frac{2}{k(k-1)} \cdot p_{\text{YH}}(A),$$

and eqn. (3.8) follows from the expression for $p_{\text{YH}}(A)$ in eqn. (3.5).

A second example is motivated by the setting in which a biologist wishes to estimate some early ancestral state in a tree that is at (or near) the root, which raises the question of how many species should be sampled under a null phylogenetic model such as YH. Suppose we were to sample a set \mathcal{Y} of $k \geq 2$ leaves uniformly at random from a YH tree \mathcal{T} having n leaves and ask how close the least common ancestor (LCA) of these leaves is to the root (i.e., the number of edges between the root and $\text{lca}_{\mathcal{T}}(\mathcal{Y})$). Of course, it is possible that this LCA vertex is the root itself and it can be shown that this probability is precisely

$$1 - \frac{2(n-k)}{(k+1)(n-1)},$$

which is asymptotic (as $n \rightarrow \infty$) to $1 - \frac{2}{k+1}$. Thus for a large YH tree, around $\frac{1}{3}$ of all pairs of leaves are connected by a path that passes through the root. Moreover, as n grows, the distance $D_n(k)$ from the root to the LCA of k randomly chosen leaves converges to a geometric distribution with parameter $2/(k+1)$ [337]. In other words, for $k \geq 2$ and $r \geq 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_n(k) \geq r) = \left(\frac{2}{k+1} \right)^r.$$

3.2.3 • Exchangeability and sampling consistency

Although the YH and uniform models lead to quite different predictions regarding tree shape, they also share some basic properties. First, for either model, if two trees from $RB(n)$ have the same (unlabeled) shape, then they have the same probability (i.e., the models are invariant to any permutation of the leaf labels). More generally, for any stochastic model θ on rooted binary phylogenies we say the model satisfies the *exchangeability property* (EP) if, for every rooted phylogeny T and any permutation σ of the leaves of T , $p_\theta(T) = p_\theta(T^\sigma)$, where $p_\theta(T) := \mathbb{P}_\theta(\mathcal{T} = T)$.

As well as the exchangeability property, the YH and uniform models also share the following property. Let $\mathcal{T} \in RB(X)$ be the random tree generated by $\theta = \text{YH}$ or uniform, and let \mathcal{T}_1 and \mathcal{T}_2 be two trees obtained by deleting the root of T and its incident edges, and let $\mathcal{L}_1, \mathcal{L}_2$ be their respective leaf sets. Conditional on \mathcal{L}_i , the tree \mathcal{T}_i is also distributed according to the same θ model on $RB(\mathcal{L}_i)$; moreover, \mathcal{T}_1 and \mathcal{T}_2 are conditionally independent, given (X_1, X_2) . In other words, for any partition $\{X_1, X_2\}$ of X and any two trees $T_1 \in RB(X_1)$ and $T_2 \in RB(X_2)$,

$$\mathbb{P}_\theta(\mathcal{T}_1 = T_1 \wedge \mathcal{T}_2 = T_2 | \mathcal{L}_1 = X_1 \wedge \mathcal{L}_2 = X_2) = p_\theta(T_1) \cdot p_\theta(T_2). \quad (3.9)$$

We will refer to this as the *Markovian property*.

A further *sampling consistency* property of either model is that if we select a subset S of X and consider the tree obtained by restricting \mathcal{T} to just the taxa in S (this idea is formalized in the next chapter), then this random tree in $RB(S)$ inherits the same distribution (YH or uniform) as the parent tree.

There are many other models beyond the YH and uniform that satisfy the properties described so far in this section (we will describe another one in Section 3.3.2), but the last property on our list seem to be very strong and is called *group elimination*. For an exchangeable model, this states that conditional on some subset Y of X being a cluster of \mathcal{T} , the rooted binary phylogeny obtained by deleting this cluster from \mathcal{T} has the same probability distribution as the model prescribes for trees on $RB(X - Y)$.

Group elimination implies sampling consistency, and it is satisfied by the YH and the uniform model, and it also holds for one other (nonbiological) model, the “comb”

model, which assigns strictly positive probability only to trees that have the shape of a rooted caterpillar, with each such tree assigned the same probability on any given leaf set. No other exchangeable model satisfying group elimination has been found; the reason may simply be that there are none, as David Aldous conjectured in 1995.

Conjecture [2]. The YH, uniform, and comb models are the only exchangeable models that satisfy group elimination.

3.3 ■ Measuring and modeling tree shape

3.3.1 ■ Balance indices (Colless and Sackin)

Various measures have been proposed to capture the extent to which a tree $T \in RB(X)$ is “balanced” or not. In this section we describe two of them. Informally, balance refers to how evenly the leaves descended from each vertex are split as one moves from the root to the tips. For example, a perfect rooted binary tree is highly balanced, while a rooted caterpillar tree is quite unbalanced. Phylogenetic trees in biology exhibit varying degrees of balance between these two extremes. In this section, we describe two indices for tree balance; these are specific to rooted phylogenies, and moving the root to a different location in the tree will change their values.

Perhaps the most intuitive measure of tree balance is the *Colless index*. This assigns each rooted binary phylogeny T a score C_T that is equal to the sum over the interior vertices of T of the difference in the number of leaves between the two subtrees descended from that vertex. In other words,

$$C_T = \sum_{v \in V(T)} |L_v - R_v|,$$

where L_v (respectively, R_v) is the number of leaves in the left-hand (respectively right-hand) subtree of T that are descended from v (note that the left-right orientation here is arbitrary, and interchanging them leaves C_T invariant). For example, tree $T \in RB(n)$ has $C_T = 0$ if and only if T is a perfect rooted binary tree (on $n = 2^b$ leaves, for some $b \geq 0$). At the other extreme, for a caterpillar tree $T \in RB(n)$, we have $C_T = (n-1)(n-2)/2$.

A second measure of tree balance is the *Sackin index*. This assigns each rooted binary phylogeny T a score S_T that is equal to the sum of the lengths of the paths from the leaves to the root of T . More formally, this is expressed as follows:

$$S_T = \sum_{x \in X} \ell(x),$$

where $\ell(x)$ is the number of edges in the path from the root ρ of T to leaf x . In certain literature, the “Sackin index” can also refer to the average path length S_T/n .

There is an equivalent way to write S_T as a summation, as noted by various authors (here, we follow [355]) as follows:

$$S_T = \sum_{e \in E} |c_T(e)| = \sum_{v \in V - \{\rho\}} |c_T(v)|. \quad (3.10)$$

Recall that $c_T(v)$ is the set of leaves that are descendants of v , and for $e = (u, v) \in E$, let $c_T(e) = c_T(v)$ denote the set of leaves that are descendants of the endpoint vertex (v) of e . To establish eqn. (3.10), first notice that we can write $\ell(x)$ as $\sum_{e \in E} \mathbb{I}_{\{x \in c_T(e)\}}$, where \mathbb{I}_A

is the variable that takes the value 1 if event A holds, and the value 0 otherwise. The first equality in eqn. (3.10) is found by reversing the order of summation:

$$\begin{aligned} S_T &= \sum_{x \in X} \ell(x) = \sum_{x \in X} \sum_{e \in E} \mathbb{I}_{\{x \in c_T(e)\}} \\ &= \sum_{e \in E} \sum_{x \in X} \mathbb{I}_{\{x \in c_T(e)\}} = \sum_{e \in E} |c_T(e)|. \end{aligned}$$

The second equality in eqn. (3.10) holds simply because we have $c_T(e) = c_T(v)$ for each $e = (u, v)$; however, the root vertex ρ (for which $|c_T(\rho)| = n$) is counted when we sum over all $v \in V$, but ρ is not the endpoint of any edge of T .

Equation (3.10) provides the key to deriving an explicit formula for the expected value of S_T when T is sampled according to *any* distribution on $RB(X)$ that satisfies the exchangeability property (EP) described earlier in this chapter. This formula involves just the probability that a given subset A of size i ($i \in \{1, \dots, n-1\}$) is a cluster of T according to the exchangeable distribution. The following result is from [355] (Lemma 6).

Proposition 3.5. *For any probability model θ on $RB(X)$ that satisfies the exchangeability property (EP), let S_n denote the Sackin index of the sampled tree. We then have*

$$\mathbb{E}_\theta[S_n] = \sum_{i=1}^{n-1} \binom{n}{i} i p_n(i),$$

where $n = |X|$ and $p_n(i)$ is the probability under θ that a given set A of size i is a cluster of the sampled tree.

Proof:

$$\mathbb{E}_\theta[S_n] = \sum_{T \in RB(X)} S_T p_\theta(T) = \sum_{T \in RB(X)} p_\theta(T) \sum_{v \in V(T) - \{\rho\}} |c_T(v)|,$$

where the second equality is from eqn. (3.10). Notice that we can write the last double sum as

$$\sum_{T \in RB(X)} p_\theta(T) \sum_{A \subset X, A \neq \emptyset} |A| \cdot \mathbb{I}_{A \in \mathcal{C}(T)},$$

where $\mathcal{C}(T)$ is the set of clusters of T . By stratifying the sets A by their possible sizes, and interchanging the order of summation, this last expression can be written as

$$\sum_{i=1}^{n-1} \sum_{A \in \binom{X}{i}} \sum_{T \in RB(X)} i \cdot p_\theta(T) \cdot \mathbb{I}_{A \in \mathcal{C}(T)}.$$

Finally, by exchangeability, $\sum_{T \in RB(X)} p_\theta(T) \cdot \mathbb{I}_{A \in \mathcal{C}(T)} = p_n(i)$, which leads directly to $\mathbb{E}_\theta[S_n] = \sum_{i=1}^{n-1} i p_n(i) \times |\{A \in \binom{X}{i}\}|$, and thereby the claimed identity. ■

The YH model provides a tailor-made application of Proposition 3.5, since we saw in eqn. (3.5) that $p_n(i) = \frac{2n}{i(i+1)} \binom{n}{i}^{-1}$ for all i in $\{1, \dots, n-1\}$. If we substitute this into the equation in Proposition 3.5, then all the troublesome terms conveniently cancel to yield a classical identity derived from the early 1990s [181, 218]:

$$\mathbb{E}_{YH}[S_n] = 2n \sum_{j=2}^n \frac{1}{j}. \quad (3.11)$$

It follows that $\mathbb{E}_{\text{YH}}[S_n] = 2n \ln n + (2\gamma - 2)n + o(n)$, where γ is Euler's constant; moreover, $\text{Var}_{\text{YH}}[S_n] \sim (7 - 2\frac{\pi^2}{3})n^2$ as n grows [32, 34].

Let us continue to concentrate on the YH model and consider the values S_n of $S_{\mathcal{T}}$, and C_n of $C_{\mathcal{T}}$ for a random tree $\mathcal{T} \in RB(n)$ selected according to this model. The analysis of the random variable S_n benefits from the fact that, by Lemma 3.1 and eqn. (3.9), S_n satisfies the following stochastic recurrence equation:

$$S_n = S_J + S_{n-J} + n, \quad (3.12)$$

where J is a uniform random variable over the subset $\{1, \dots, n-1\}$, and where S_J and S_{n-J} are conditionally independent given J . Equation (3.12) also allows for the computation of the mean and variance of S_n , while other more sophisticated techniques show that the normalized Sackin index $(S_n - \mathbb{E}[S_n])/n$ converges in distribution as n become large [32]. A similar analysis applies for C_n , where the analogue of eqn. (3.12) is

$$C_n = C_J + C_{n-J} + |n - 2J|.$$

Asymptotic results for the mean and variance of C_n (from [32, 34]) are

$$\mathbb{E}_{\text{YH}}[C_n] = n \log n + (\gamma - 1 - \log 2)n + o(n),$$

$$\text{Var}_{\text{YH}}[C_n] \sim \left(3 - \frac{\pi^2}{6} - \log 2\right)n^2,$$

while the correlation coefficient of the two variables S_n and C_n in the YH model (i.e., $\text{Cor}_{\text{YH}}[S_n, C_n]$) is asymptotic to an explicit constant that is close to 1 (≈ 0.98) [34].

For the uniform model (U), corresponding results for the Colless and Sackin indices are

$$\mathbb{E}_U[C_n] \text{ and } \mathbb{E}_U[S_n] \text{ are both asymptotic to } \sqrt{\pi}n^{3/2},$$

$$\text{Var}_U[C_n] \text{ and } \text{Var}_U[S_n] \text{ are both asymptotic to } \frac{10 - 3\pi}{3}n^3, \text{ and}$$

$$\text{Cor}_U[S_n, C_n] \sim 1,$$

as $n \rightarrow \infty$ [34].

Notice that the expected Colless and Sackin indices of a tree generated by the uniform model both grow at a faster rate ($n^{3/2}$) than those for the YH model ($n \log n$), reflecting the imbalance in tree shapes predicted by these two distributions. It is curious that S_n and C_n have asymptotically equivalent means (and variances) for the uniform model, even though they measure quite different quantities. For the YH model, these means (and variances) grow at the same rate, though they are asymptotically different, with the expected value of S_n converging to twice that of C_n for the YH model.

We will return to the Sackin index in Chapter 9, where it reappears in formulae for measuring the degree of reconciliation between a gene tree and a species tree.

3.3.2 • The Aldous β -splitting model

In a pioneering paper, David Aldous [2] introduced a one-parameter distribution for generating rooted binary phylogenies. The parameter (β) allows control over the degree of balance of the generated trees, and includes the uniform, YH, and comb models as special cases. It produces a distribution on rooted phylogenies that satisfies the exchangeability property (EP) as well as the sampling consistency property described earlier.

To motivate this model, imagine that we select n points independently and uniformly at random on the interval $(0, 1)$. Suppose that f is some continuous density on $(0, 1)$ which is symmetric (i.e., $f(x) = f(1-x)$), and let us cut $(0, 1)$ at a point Y sampled from the density f . A random number J of the n uniformly selected points will then lie to the left of Y . Conditional on $Y = y$, the distribution of J is binomial with parameters n and y (i.e., $\mathbb{P}(J = j|Y = y) = \binom{n}{j} y^j (1-y)^{n-j}$) so the unconditional probability that $J = j$ is given by

$$\mathbb{P}(J = j) = \mathbb{E}_Y[\mathbb{P}(J = j|Y)] = \binom{n}{j} \int_0^1 x^j (1-x)^{n-j} f(x) dx.$$

Now, we want to only consider cuts where at least one of the n points is on each side of the cut, so for $j \in \{1, \dots, n-1\}$, let $q_n(j)$ denote the conditional probability $\mathbb{P}(J = j|J \neq 0, n)$. Clearly,

$$q_n(j) = a_n^{-1} \mathbb{P}(J = j), \text{ where } a_n = \mathbb{P}(J \neq 0, n) = 1 - 2 \int_0^1 x^n f(x) dx.$$

With this in hand, we can now construct a tree by splitting the n points placed randomly on the interval into two subsets of sizes $J \in \{1, \dots, n\}$ and $n-J$ by this process of cutting the interval according to the (conditional) distribution $q_n(j)$. We then start to build a tree with J leaves on one side of the root and $n-J$ on the other. This splitting process is then repeated independently on each of these subsets (where n is now replaced by J and $n-J$) and the process is continued until a tree with n leaves results; these leaves are then labeled by a random permutation from $[n]$ to a random tree in $RB(n)$ having a distribution that satisfies the exchangeability property (EP). This is illustrated in Fig. 3.5.

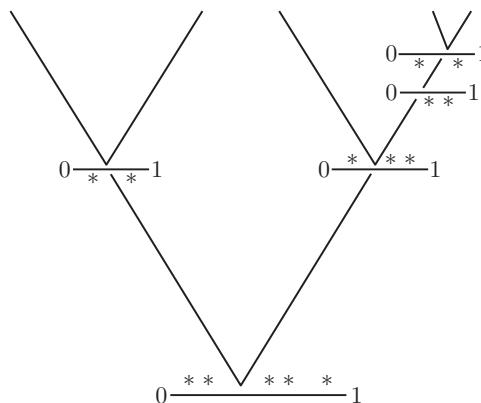


Figure 3.5. A schematic illustration of the generation of a discrete rooted binary phylogeny under the β -splitting model. Starting with n points at the root, these points are distributed uniformly on the unit interval and the interval cut according to a continuous density f . The two sets of points are then passed up to the next level where the process is repeated (if all points lie on one side of the cut, the process is repeated until a genuine division occurs, as illustrated here also).

In the case where $f(x)$ is the uniform distribution, this leads precisely to the YH distribution. More generally, if f has a beta distribution (i.e., $f(x) = c \cdot x^\beta (1-x)^\beta$, where $\beta \in (-1, \infty)$), then

$$q_n(j) = \frac{1}{a_n(\beta)} \frac{\Gamma(\beta + j + 1)\Gamma(\beta + n - j + 1)}{\Gamma(j + 1)\Gamma(n - j + 1)}, \quad 1 \leq j \leq n - 1, \quad (3.13)$$

where $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ is the “gamma function” and $a_n(\beta)$ is the normalizing constant (to ensure the $q_n(j)$ values sum to 1). Equation (3.13) is well defined in the range $-2 \leq \beta \leq \infty$ (this is slightly larger than the range that the beta distribution is a valid probability distribution on). Over this extended range, the *β -splitting model* is the distribution on rooted phylogenies in the tree construction process described using eqn. (3.13) to determine the splitting process.

Notice that the YH model corresponds to the case $\beta = 0$ since $f(x) = c \cdot x^\beta (1-x)^\beta$ is the uniform continuous distribution $f(x) = 1$ on $(0, 1)$, and we saw in Lemma 3.1 that the YH model involves uniform splitting. As for the uniform model (U), this corresponds to $\beta = -3/2$, while the comb model corresponds to the lower limiting value $\beta = -2$. Thus as β decreases from $\beta = 0$ (YH model), the trees become increasingly unbalanced. Several empirical studies have supported an estimate of $\beta \approx -1$ for phylogenies reconstructed from biological data [3, 33]. The value of β for one or more phylogenies is typically done by maximum likelihood estimation using numerical analysis techniques (available in the R package “apTreeshape”). The choice of basing q_n on a beta distribution may at first seem somewhat arbitrary; however, the resulting model has some desirable properties—including the Markovian property and sampling consistency—and the model turns out to be the unique one that satisfies these two properties and a third one.¹²

3.3.3 • Models on unrooted phylogenies

So far, we have considered distributions only on rooted binary trees. However, if we generate a rooted tree $T \in RB(n)$ according to some distribution and suppress the location of the root, then we obtain an unrooted tree $T^{-\rho} \in B(n)$. In the case of the uniform and Yule distribution on $RB(n)$ there is a simple way to describe the induced probability distribution on $B(n)$. For the uniform distribution on $RB(n)$, the associated distribution on $B(n)$ is again uniform; the probability of any tree $T \in B(n)$ is simply $1/b(n)$. This distribution can also be described constructively in a way similar to that for the YH model on rooted trees, as follows: We start with the tree $T \in B(2)$ and add the leaves $3, 4, \dots, n$ sequentially (or in random order) according to the following rule: Given the tree so far constructed, select an edge uniformly at random, subdivide this edge, and attach the next available leaf in the sequence to the tree by placing an edge between the leaf and the newly created subdivision vertex.

For the YH model on $RB(n)$, we will call the induced model on $B(n)$ obtained by suppressing the root vertex the *unrooted YH model*. It can be described in a more direct way (as a process on unrooted trees only) as follows: First, select a uniform random permutation x_1, x_2, \dots, x_n of $[n]$ and start with the tree $T \in B(\{x_1, x_2\})$. Add the leaves x_3, x_4, \dots, x_n sequentially according to the following rule: From the present tree, select a *pendant* edge uniformly at random, subdivide this edge, and attach the next available leaf in the sequence to the tree by placing an edge between the leaf and the newly created subdivision vertex.

An interesting question now arises. Suppose that a rooted tree $T \in RB(X)$ is generated under some exchangeable model (e.g., uniform or YH) but we only see the unrooted version of this tree $T^{-\rho} \in B(X)$. Can we estimate which edge of this tree originally contained the root of T (i.e. which edge of $T^{-\rho}$ corresponds to identifying the two edges of T that were incident with the root, when this root vertex was suppressed)? Let us call this unknown edge of $T^{-\rho}$ the *root edge*.

¹²Namely, that $q_n(i)$ can be written as proportional to $w(i)w(n-i)$ for some function w (see [244], Theorem 2).

In the case of the uniform model on $RB(n)$, each edge of $T^{-\rho}$ has equal probability of being the root edge, so estimation can be no more accurate than picking an edge at random. In particular, as n grows, the probability of identifying root edge T goes to zero. However, for the YH process, the situation is quite different: in this case, a maximum likelihood estimate (MLE) of the root edge is the edge for which the resulting rooting confers the largest number of rankings on the resulting rooted binary tree. Such an MLE edge can therefore be readily identified: when the centroid of the tree consists of two vertices, this is precisely the MLE edge, while if the centroid is a single vertex v , then the MLE edge(s) join v to the adjacent subtree(s) with the largest number of leaves. Moreover, for a YH tree \mathcal{T} with n leaves, the probability that the unknown root edge of $\mathcal{T}^{-\rho}$ is within a few edges distant from the MLE edge remains high (> 0.5) even as $n \rightarrow \infty$ (for details, see [337]).

3.4 • Cherries and extended Pólya urn models

The number $c(T)$ of cherries in a tree provides a further measure of tree balance. In this last section, we begin by investigating the distribution of $c(T)$ under the YH and uniform models, a topic which connects with some classical theory in discrete probability. While $c(T)$ is not a particularly discriminating measure of tree shape, nevertheless it has one property in its favor: it is robust to the location of the root of a phylogeny $T \in RP(X)$, since it is also defined on unrooted trees and $c(T^{-\rho})$ is either equal to $c(T)$ or $c(T) + 1$. By contrast, the Colless and Sackin indices are defined only on rooted trees, and they can be quite variable as to where the tree is rooted (we will see in Chapter 7 that determining the location of the root of a tree from data is not straightforward).

An extension of the simple Pólya urn model described earlier in this chapter provides a convenient way to calculate the distribution of cherries in random trees under both the YH model (on $RB(n)$) and the uniform model on $B(n)$.

We start first with the YH model on $RB(n)$. Think of the pendant edges as two types: *red* and *blue*: a pendant edge is red if it is part of a cherry, and blue otherwise. In the YH model a pendant edge is selected uniformly at random from the tree constructed so far, and this edge is replaced by a cherry. Notice that if a red edge $e = (u, x)$ is selected, the number of pendant edges in a cherry remains the same and the number of blue edges increases by one, because, while e is replaced by two red edges, the other red edge $e' = (u, x')$ (for which $\{x, x'\}$ formed a cherry) becomes a blue edge; alternatively, if a blue edge e is selected, then the number of blue edges decreases by one but the number of red edges increases by two. We can model the number of red and blue edges in a YH tree by an *extended Pólya urn model* in which we start with an urn containing two red balls (corresponding to two red pendant edges in the tree $T \in RB(2)$) and repeatedly apply the following rule:

Draw a ball uniformly at random from the urn: if it is red, return a red and a blue ball to the urn; if it is blue, return two red balls to the urn.

Notice that after $n - 2$ draws, the urn will contain n balls and the number of red balls will be an even number that is distributed identically to twice the number of cherries in a random tree produced by the YH process. The process is illustrated in Fig. 3.6.

This connection makes it easy to find the mean and variance for the number of cherries χ_n for a YH tree with n leaves, giving

$$\mathbb{E}[\chi_n] = \frac{n}{3} (n \geq 3) \text{ and } \text{Var}[\chi_n] = \frac{2n}{45} (n \geq 5).$$

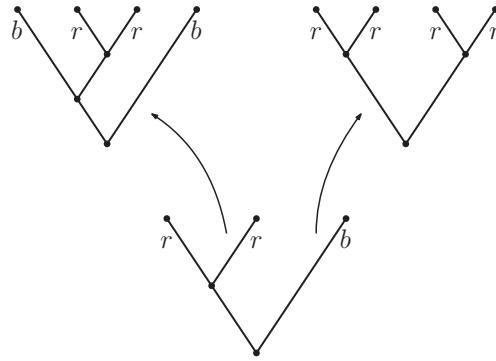


Figure 3.6. The effect of selecting a red (r) cherry edge, or a blue (b) noncherry pendant edge in the extended Pólya urn scheme for the YH model on rooted binary trees.

Moreover, there is a general theory (see, e.g., [237]) which applies for certain extended Pólya urn models, including the type we have described here, and which shows that the number of balls of given colors has a limiting multinormal distribution. In particular, the following convergence in distribution applies:

$$\frac{\chi_n - n/3}{\sqrt{2n/45}} \xrightarrow{D} \mathcal{N}(0, 1). \quad (3.14)$$

Exercise: Show that for $n \geq 2$, $\chi_{n+1} = \chi_n + W$, where W is a Bernoulli random variable that takes the value 1 with probability $(n-2\chi_n)/n$ and is zero otherwise. Deduce that $\mu_n := \mathbb{E}[\chi_n]$ satisfies the recurrence $\mu_{n+1} = 1 + \mu_n(1 - \frac{2}{n})$ for all $n \geq 2$. Deduce that $\mu_n = n/3$ for $n \geq 3$.

A similar analysis based on an extended Pólya urn model applies for the number of cherries χ'_n for the uniform model on $B(n)$. However, we need to make one important adjustment: Instead of just two colors for the balls in the urn, we require a third color—say, green—to mark the interior edges. This is because, as we saw in Section 3.3.3, uniformly distributed trees in $B(n)$ are generated by a similar random process in which we start with $T \in B(4)$ and repeatedly apply the following process: From the tree generated so far, select one of the edges uniformly at random, subdivide the edge, and attach a new leaf to be adjacent to this subdivision vertex.

In this way, we can model the distribution of the edge types (and thereby χ'_n) for a uniform tree on $B(n)$ by starting with an urn containing four red balls and one green ball (i.e., the four pendant edges in a cherry and the single interior edge in a tree in $B(4)$) and repeatedly applying the following rule:

Draw a ball uniformly at random from the urn: if it is red, then replace it by one blue, one red, and one green ball; if it is blue, then replace it by two red balls and one green ball; if it is green, then replace it by two green balls and one blue ball.

Notice that in each draw, the number of balls increases by 2 (corresponding to the increase in the number of edges of the tree); Fig. 3.7 illustrates the process.

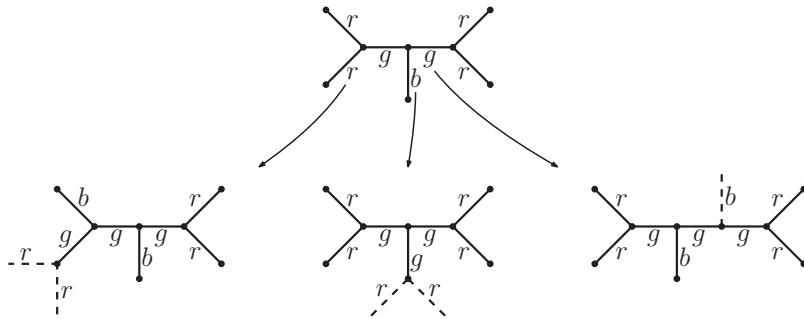


Figure 3.7. The effect of selecting a red (r) cherry edge, or a blue (b) noncherry pendant edge, or a green (g) interior edge in the extended Pólya urn scheme for the uniform model on unrooted binary trees.

Again, by using the extended Pólya urn theory, it can be shown that

$$\mathbb{E}[\chi'_n] \sim \frac{n}{4}, \quad \text{Var}[\chi'_n] \sim \frac{n}{16} \quad \text{and} \quad \frac{\chi'_n - n/4}{\sqrt{n/16}} \xrightarrow{D} \mathcal{N}(0, 1). \quad (3.15)$$

Thus we have an asymptotic result for YH on rooted trees and for the uniform model on unrooted trees. However, for each model, exactly the same asymptotic results apply for rooted or unrooted trees, since the number of cherries in $T \in R\bar{B}(n)$ and in the tree $T' \in B(n)$ obtained by suppressing the root vertex differ by, at most, 1. For more details on the results described here, see [246]. The distribution of pendant subtrees of size larger than 2 has been investigated in [302] and [32].

For the uniform distribution on $B(n)$, we gave asymptotic formulae for the mean and variance of the random variable χ'_n ; however, exact formulae exist. Moreover, the exact distribution of χ'_n is known and it can be derived by direct combinatorial arguments, in the following result from [104].

Proposition 3.6. For $n \geq 4$,

$$\mathbb{P}(\chi'_n = k) = \frac{\binom{n-2}{n-2k} \binom{2k-2}{k-2}}{\binom{2n-4}{n-4}} \cdot 2^{n-2k}$$

for $2 \leq k \leq n/2$, with $\mathbb{P}(\chi'_n = k) = 0$ otherwise.

Proof: Let $B(n, k)$ denote the set of binary trees T on leaf set $[n]$ that have exactly c cherries. We will show that for $n \geq 4$,

$$|B(n, k)| = \frac{n!(n-4)!}{k!(k-2)!(n-2k)!2^{2k-2}} \quad (3.16)$$

for $2 \leq k \leq n/2$, and with $|B(n, k)| = 0$ otherwise. Equation (3.16) was first stated in [185] using an inductive argument. Here, we provide a constructive proof from [104]. The number of ways to choose the $2k$ leaves from $[n]$ to form the k cherries of T is $\binom{n}{2k}$, and the number of ways to pair up these $2k$ leaves into a set of k (unordered) pairs is $\frac{(2k)!}{k!2^k}$. Thus there are

$$M = \binom{n}{2k} \frac{(2k)!}{k!2^k} = \frac{n!}{k!(n-2k)!2^k}$$

ways to select the k cherries from $[n]$. Let Y be any one choice of such k pairs. There are $b(k)$ trees in $B(Y)$, and any tree in $B(n, k)$ is obtained by (i) selecting the set of pairs Y (in M ways), (ii) selecting a tree T_Y in $B(Y)$ (in $b(k)$ ways), and (iii) attaching each of the remaining $n - 2k$ leaves of $[n]$ to one of the $2k - 3$ edges of $T \in B(Y)$ (so as not to create any further cherries). To count cases for step (iii), recall that $\binom{a+b-1}{b}$ is the number of ways to place b unlabeled objects into a labeled bins, and $b!$ is the number of ways to assign b distinct labels bijectively to b unlabeled objects. Applying this to the $a = 2c - 3$ edges of T_Y and to the $b = n - 2k$ leaves to be attached to the edges of this tree, the number of ways to perform step (iii) is $(n - 2k)! \binom{(n-2k)+(2k-3)-1}{n-2k} = \frac{(n-4)!}{(2k-4)!}$. Notice that different choices of Y (or a different choice of T_Y) lead to a different set of trees, so the total number of trees in $B(n, k)$ is the product over the three counts for steps (i), (ii), and (iii) above, namely

$$M \cdot b(k) \cdot \frac{(n-4)!}{(2k-4)!},$$

which simplifies to the expression in eqn. (3.16). Finally, $\mathbb{P}(\chi'_n = k) = |B(n, k)|/b(n)$, and some tedious but straightforward algebra leads to the expression in Proposition 3.6. ■

3.4.1 • The Robinson–Foulds distance to a random tree

Cherries also play a central role in describing a distribution that arises from the Robinson–Foulds (RF) metric d_{RF} from the previous chapter. Given an unrooted binary phylogenetic X -tree, suppose that a second such tree \mathcal{T} is drawn uniformly at random from $B(X)$. Then $d_{\text{RF}}(\mathcal{T}, T) = 2(n - 3 - s(\mathcal{T}, T))$, where $s(\mathcal{T}, T)$ is the number of *nontrivial* splits shared by \mathcal{T} and T , and $n = |X|$. It can be shown that as n grows, $s(\mathcal{T}, T)$ approaches a Poisson distribution with mean $\lambda_T = c(T)/2n$, in other words, one quarter of the proportion of leaves of T that lie in cherries. A more precise version of this last sentence is that

$$\sum_k \left| \mathbb{P}(s(\mathcal{T}, T) = k) - e^{-\lambda_T} \frac{\lambda_T^k}{k!} \right| \leq \frac{A}{n}$$

for an absolute constant A , and so the left-hand side of the inequality converges (uniformly) to 0 as $n \rightarrow \infty$ [70]. The reason for this Poisson limit is that the only nontrivial splits that a random binary tree will (asymptotically) share with T are of the form $\{x, x'\}|(X - \{x, x'\})$ (these “cherry splits” correspond to a common cherry in both trees) and all other shared nontrivial splits combined have vanishing probability in the limit of large n . Thus one can focus asymptotically on just the shared cherries. As n grows, the probability a given cherry split is shared by T , and \mathcal{T} vanishes, but the number of such cherries grows, provided that λ_T is bounded away from zero. Although the shared cherry splits are not exactly independent for any given n , nevertheless the number of shared cherry splits is still asymptotically Poisson distributed.

In addition, we can allow the second tree T to be a random tree \mathcal{T}' also chosen uniformly at random from $B(n)$ and independently of \mathcal{T} . From eqns. (3.14) and (3.15) for a YH distribution, $\lambda_{\mathcal{T}'}$ converges in probability to $\frac{1}{6}$, and for the uniform distribution, $\lambda_{\mathcal{T}'}$ converges in probability to $\frac{1}{8}$. Thus the probability that two trees from $B(n)$ chosen independently and uniformly at random share exactly k nontrivial splits is $e^{-1/8}/(k!8^k)$.

A nonasymptotic bound, which does not even mention n or the number of cherries in T , is the following (from [104]): For any unrooted phylogenetic X -tree T , and a tree \mathcal{T} selected uniformly at random from $B(X)$, the probability that these two trees share at

least k nontrivial splits is at most $2^{-k}/k!$. Thus, the probability that a uniformly selected binary phylogenetic X -tree shares four or more splits with any given phylogenetic X -tree is less than 0.003.

Exercise⁺: Suppose \mathcal{T} and \mathcal{T}' are sampled independently from $B(n)$ according to the uniform model. Show that the event that \mathcal{T} and \mathcal{T}' share one or more splits $A|B$, where A and B are both of size at least 3, has a probability that converges to zero as n grows.

Chapter 4

Pulling trees apart and putting trees together

In many areas of mathematics, structures are studied in terms of their substructures. For example, in algebra, groups are analyzed and described in terms of subgroups; fields in terms of subfields. In a similar spirit, a phylogeny induces a subphylogeny by restricting attention to any subset of its leaves. These subphylogenies encode the original tree, even for small subsets (of sizes 3 and 4 for rooted and unrooted phylogenies, respectively). This provides yet another way to view a phylogeny (i.e., beyond being either a graph, a hierarchy, or a set of splits). The exploration of the links between these three equivalent ways of encoding phylogenetic trees, along with a fourth that we will learn about in Chapter 6, forms the basis of an emerging area, namely “phylogenetic combinatorics” (see [117] for more details).

For biologists, the primary interest in the topic of this chapter has been the inverse process: rather than encoding a tree into subtrees, how can one combine phylogenies that classify overlapping sets of species into a “supertree” where every species appears as a leaf in a single tree? This raises many interesting combinatorial questions, such as when phylogenies will “fit together” into a tree, and if they do, when is there just one possible supertree? These questions have led to some of the deeper mathematical results in phylogenetics.

4.1 • Restriction and display

Suppose T is a (rooted or unrooted) phylogenetic X -tree and Y is a subset of X . We will assume throughout this chapter that Y (and so also X) have at least two elements. The *restriction* of T to Y , denoted $T|Y$ and referred to as a *restricted* subtree of T , is a phylogenetic Y -tree that is of the same type (rooted or unrooted). This notion can be defined either graphically or in terms of set systems (the former is more intuitive; the latter is easier to state precisely). Graphically, for $T \in P(X)$, $T|Y$ is obtained from T by taking the minimal subtree of T that connects the leaves in Y and suppressing any resulting degree-2 vertices. When $T \in RP(X)$, the same applies, except that it is the resulting vertices of in-degree 1 and out-degree 1 that are suppressed; the root of $T|Y$ is the unique vertex in the resulting tree that has in-degree 0. Alternatively, $T|Y$ can also be described more precisely in terms of splits (for unrooted phylogenies) or hierarchies (for rooted phylogenies) as follows. For $T \in P(X)$, $T|Y$ is the unrooted phylogeny on Y associated with the split system

$$\{A \cap Y | B \cap Y : A | B \in \Sigma(T), A \cap Y, B \cap Y \neq \emptyset\}.$$

For a rooted phylogeny corresponding to some hierarchy \mathcal{H} on X , $T|Y$ is the rooted phylogeny on Y associated with the hierarchy

$$\{A \cap Y : A \in \mathcal{H}, A \cap Y \neq \emptyset\}.$$

Roughly speaking, $T|Y$ is the tree one obtains from T by ignoring all the species that are not in Y and their connecting branches, in other words, by tracing out the subtree of T just for the particular subset Y of the species of interest.

Notice that each vertex v of $T|Y$ corresponds to a unique vertex v' of T . This association can be described by the well-defined one-to-one mapping from $V(T|Y)$ to $V(T)$: $v = \text{med}_{T|Y}(a, b, c) \mapsto v' = \text{med}_T(a, b, c)$ for $a, b, c \in Y$. Notice also that for any $T \in P(X)$, there is a subset Y of X with $T|Y \in B(Y)$; moreover, a largest set Y with the property that $T|Y$ is binary can be found by a simple polynomial-time algorithm (we will leave the details as an exercise for the reader).

A tree T' in $P(X)$ is said to *display* a tree T in $P(Y)$ if $T'|Y$ is equivalent to T or strictly refines it. Thus if T is binary, then T' displays T if and only if $T'|Y \cong T$. The reader may wonder why the definition of displays allows T' to refine T rather than impose the stronger requirement that $T'|Y \cong T$. This is primarily to reflect the convention in evolutionary biology of interpreting vertices of high degree as uncertainty as to how lineages may have split at that vertex (a “soft polytomy”), rather than certainty that there was a simultaneous separation into multiple lineages (a “hard polytomy”). This choice also has desirable mathematical consequences (for instance, it allows a unified treatment of related notions, such as perfect phylogeny, discussed in the next chapter). Exactly analogous definitions of “display” apply in the rooted case; this concept is illustrated in Fig. 4.1.

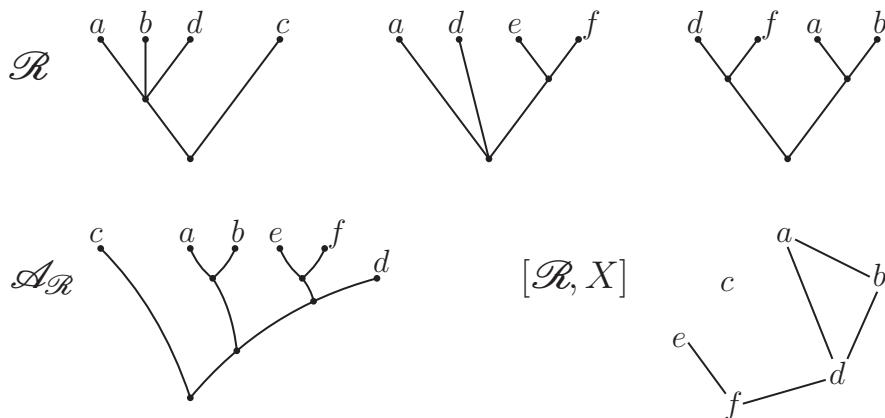


Figure 4.1. Each of the three rooted trees in \mathcal{R} is displayed by the tree $\mathcal{A}_{\mathcal{R}}$. The graph $[\mathcal{R}, X]$ (from Section 4.2.1) is also shown.

How many binary phylogenies on $n \geq 2$ leaves display a given binary phylogenetic tree T on some subset Y of $[n]$ of size k ? The answer is surprisingly simple. For any values $n \geq k \geq 2$, and any $T \in B(Y)$, with $Y \in \binom{[n]}{k}$,

$$\#\{T' \in B(n) : T'|Y = T\} = \frac{b(n)}{b(k)}. \quad (4.1)$$

The proof is a straightforward extension of the argument from Chapter 2 used to count $B(n)$. Starting from T , this tree has $(2k - 3)$ edges to which a first leaf of $[n] - Y$ can be

attached; from this new tree, there are then $(2k - 1)$ edges for attaching the second leaf. Continuing in this way gives a total of $(2k - 3)(2k - 1) \cdots (2n - 5) = \frac{b(n)}{b(k)}$ ways to construct $T' \in B(n)$ from T .

Similarly, the number of trees T' in $RB(n)$ that display a given tree T' in $RB(Y)$ for $Y \subseteq \binom{[n]}{k}$ is $r b(n)/rb(k)$.

Suppose that we now have two binary phylogenies, T_1 and T_2 , which have disjoint leaf sets Y_1 and Y_2 with union $[n]$. Let $N(T_1, T_2)$ be the number of phylogenies $T \in B(n)$ with $T|Y_1 = T_1$ and $T|Y_2 = T_2$. If $N(T_1, T_2)$ depended just on n , and $k_i = |Y_i|$ ($i = 1, 2$), then this would require the equality $N(T_1, T_2) = \frac{b(n)}{b(k_1)b(k_2)}$. This identity holds whenever $\min\{k_1, k_2\} \leq 5$; however, in general, $N(T_1, T_2)$ depends on the “shapes” of the two trees, as these influence how many ways the two trees “fit together” into a parent tree. For example, when $n = 12$ and $k_1 = k_2 = 6$, there is no pair of trees $T_1 \in B(Y_1)$ and $T_2 \in B(Y_2)$ with $N(T_1, T_2) = \frac{b(12)}{b(6)^2}$ for the very simple reason that this rational number is not an integer. Nevertheless, a simple argument reveals that for any tree $T_1 \in B(Y_1)$ and any random tree T_2 selected uniformly at random from $B(Y_2)$ we have

$$\mathbb{E}[N(T_1, T_2)] = \frac{b(n)}{b(k_1)b(k_2)}, \quad (4.2)$$

where $k_1 = |Y_1|$ and $k_2 = |Y_2|$, and $Y_1 \cap Y_2 = \emptyset$, as before.

Exercise: Establish eqn. (4.2).

4.1.1 • The span of a set of trees, and compatibility

The following definitions all apply equally for rooted and unrooted phylogenetic trees. Given a collection $\mathcal{P} = \{T_1, T_2, \dots, T_k\}$ of phylogenies we say that a phylogenetic tree T displays \mathcal{P} if T displays each tree $T_i \in \mathcal{P}$. Moreover, if we let $\mathcal{L}(\mathcal{P}) = \bigcup_{i=1}^k X_i$, where X_i is the leaf set of T_i , the set of all phylogenetic trees on $\mathcal{L}(\mathcal{P})$ that display \mathcal{P} is called the (compatible) *span* of \mathcal{P} and is denoted $\langle \mathcal{P} \rangle$ (an alternative notation that has been used elsewhere is $\text{co}(\mathcal{P})$). Similarly, the set of all binary phylogenetic X -trees that display \mathcal{P} is called the *binary span* of \mathcal{P} and is denoted $\langle \mathcal{P} \rangle_B$. Notice that $\langle \mathcal{P} \rangle_B \subseteq \langle \mathcal{P} \rangle$ and that if either set is empty, then so is the other. Notice also that $\langle \mathcal{P} \rangle$ is closed under refinement (i.e., if $T \in \langle \mathcal{P} \rangle$ and T' refines T , then $T' \in \langle \mathcal{P} \rangle$). A collection \mathcal{P} of trees is said to be *compatible* if there is a phylogeny that displays each tree (i.e., if $\langle \mathcal{P} \rangle \neq \emptyset$).

4.1.2 • Quartet trees and rooted triples

Recall that a *quartet tree* is a binary phylogeny with four leaves, while a *rooted triple* is a rooted binary phylogeny with three leaves. Let $\mathcal{Q}(T)$ be the set of all quartet trees that are displayed by T . That is,

$$\mathcal{Q}(T) = \left\{ T|Y : Y \in \binom{X}{4}, T|Y \in B(Y) \right\}.$$

Notice that for any $T \in P(X)$, the set $\mathcal{L}(\mathcal{Q}(T))$ (the union of the leaf sets of the quartet trees in \mathcal{Q}) equals X except in the special case where T is a star tree, in which case $\mathcal{Q}(T)$

is the empty set. Notice also that

$$\mathcal{Q}(T) = \{aa'|bb' : \exists A|B \in \overset{\circ}{\Sigma}(T) : a, a' \in A, b, b' \in B; a \neq a', b \neq b'\},$$

and that if $Y \subseteq X$, then $\mathcal{Q}(T|Y) = \{ab|cd \in \mathcal{Q}(T) : \{a, b, c, d\} \in \binom{Y}{4}\}$. In terms of the quaternary relation $|_T$ introduced in Chapter 2, $ab|_T cd \Leftrightarrow ab|cd \in \mathcal{Q}(T)$ for every four distinct elements a, b, c, d of X . $\mathcal{Q}(T)$ provides yet another way to encode phylogenies, as the following result shows.

Lemma 4.1. *Let T and T' be two phylogenetic X -trees. Then T' refines T if and only if $\mathcal{Q}(T) \subseteq \mathcal{Q}(T')$, and $T \cong T'$ if and only if $\mathcal{Q}(T) = \mathcal{Q}(T')$.*

Proof: For any unrooted phylogenetic X -tree T , $\overset{\circ}{\Sigma}(T)$ is precisely the set of X -splits $A|B$ for which $aa'|bb' \in \mathcal{Q}(T)$ for all $\{a, a'\} \in \binom{A}{2}$ and $\{b, b'\} \in \binom{B}{2}$. So if $\mathcal{Q}(T) \subseteq \mathcal{Q}(T')$, then $\overset{\circ}{\Sigma}(T) \subseteq \overset{\circ}{\Sigma}(T')$, and since T and T' have the same set of trivial splits, $\Sigma(T) \subseteq \Sigma(T')$, which means that T' refines T . Conversely, if T' does not refine T , then either T strictly refines T' or T has a split that is incompatible with some split of T' . In either case this leads to the existence of a quartet $\{a, b, c, d\}$ from X , with $T|\{a, b, c, d\} = ab|cd$ and $T'|\{a, b, c, d\} \neq ab|cd$, so $\mathcal{Q}(T)$ is not a subset of $\mathcal{Q}(T')$. The last claim in the statement of the lemma follows from the fact that phylogenies have the same set of splits precisely if they are equivalent. ■

Exercise: Show that for every set \mathcal{P} consisting of two phylogenetic X -trees T and T' , \mathcal{P} is compatible if and only if no four elements a, b, c, d in X exist with $ab|cd \in \mathcal{Q}(T)$ and $ac|bd \in \mathcal{Q}(T')$.

For any set \mathcal{Q} of quartet trees (compatible or not) with $\mathcal{L}(\mathcal{Q}) = X$ consider the set of X -splits $A|B$ for which $aa'|bb' \in \mathcal{Q}$ for all $a, a' \in A, b, b' \in B$ (with $a \neq a', b \neq b'$). It is an easy exercise to show that this collection of X -splits is either empty or is pairwise compatible. In either case if we add to this set all the trivial X -splits, we obtain the splits of a unique X -tree, which we will denote by $T_{\mathcal{Q}}$ (this is sometimes called the \mathcal{Q}^* tree). For example, for any $T \in P(X)$ we have $T_{\mathcal{Q}(T)} \cong T$.

Let us now consider a collection $\mathcal{P} = \{T_1, T_2, \dots, T_k\}$ of unrooted phylogenies. Let

$$q(\mathcal{P}) := \bigcup_{i=1}^k \mathcal{Q}(T_i),$$

which is the set of quartet trees that are displayed by at least one tree in \mathcal{P} . Notice that $\mathcal{L}(q(\mathcal{P})) \subseteq \mathcal{L}(\mathcal{P})$, with equality if and only if each leaf in $\mathcal{L}(\mathcal{P})$ is present in at least one nonstar tree, in which case we say that \mathcal{P} is *nondegenerate*. By Lemma 4.1, a phylogenetic tree T on $\mathcal{L}(\mathcal{P})$ displays each tree in \mathcal{P} if and only if T displays each quartet tree that is displayed by at least one tree in \mathcal{P} . We can express this more succinctly in terms of the notion of span introduced above. First, in the case where \mathcal{P} is nondegenerate, we have

$$\langle \mathcal{P} \rangle = \langle q(\mathcal{P}) \rangle. \quad (4.3)$$

For a general (possibly degenerate) collection \mathcal{P} of phylogenies with $X = \mathcal{L}(\mathcal{P})$ and $Y = \mathcal{L}(q(\mathcal{P}))$, the span $\langle \mathcal{P} \rangle$ equals the set $\{T \in P(X) : T|Y \in \langle q(\mathcal{P}) \rangle\}$. In particular, $\langle \mathcal{P} \rangle$ is nonempty (i.e., \mathcal{P} is compatible) if and only if $\langle q(\mathcal{P}) \rangle$ is nonempty (i.e., $q(\mathcal{P})$ is

compatible). In that case, if s is the number of leaves in \mathcal{P} that occur only in star trees, then, from eqn. (4.1),

$$\#\langle \mathcal{P} \rangle_B = \#\langle q(\mathcal{P}) \rangle_B \cdot \frac{b(n)}{b(n-s)},$$

where $n = \#\mathcal{L}(\mathcal{P})$.

The span of a set of quartet trees can be large and consist of only binary phylogenies. Indeed, a construction is described in [36] for all $k \geq 1$ of a set \mathcal{Q}_k of $4k$ quartet trees with $\#\mathcal{L}(\mathcal{Q}_k) = 4k + 3$ and for which $\langle \mathcal{Q}_k \rangle$ consists of 2^k phylogenies, which are all binary.

Rooted triples. Once again, analogous results apply in the rooted setting. A rooted triple $ab|c \in RB(\{a, b, c\})$ is displayed by T if $T|Y = ab|c$, which, in terms of the ternary relation $|_T$ introduced in Chapter 2 is equivalent to the condition $ab|_T c$. For T in $RP(X)$, we let

$$\mathcal{R}(T) = \left\{ T|Y : Y \in \binom{X}{3}, T|Y \in RB(Y) \right\}$$

be the set of all the rooted triples displayed by T . For a set \mathcal{R} of rooted phylogenies, let $r(\mathcal{R})$ be the union of $\mathcal{R}(T)$ over all T in \mathcal{R} . For a nondegenerate collection \mathcal{R} of rooted phylogenies and the associated set $r(\mathcal{R})$ of rooted triples, we then have

$$\langle \mathcal{R} \rangle = \langle r(\mathcal{R}) \rangle.$$

There is a way to build a rooted phylogeny from an arbitrary set \mathcal{R} of rooted triples, analogous to $T_{\mathcal{R}}$. However, in Section 4.2.1 we will concentrate on an alternative approach which always constructs a canonical tree in $\langle \mathcal{R} \rangle$ when this set is nonempty.

4.2 ■ When is a collection of trees compatible?

Given a collection \mathcal{P} of phylogenies, the question of whether or not there is a tree that displays \mathcal{P} is a basic one that also links to other questions in phylogenetics, as we will see in the subsequent chapters.

Let us start with the easiest case where \mathcal{P} consists of just two phylogenies T_1 and T_2 (either both rooted or both unrooted) on overlapping leaf sets. Let Y denote the set of leaves common to both trees. Then $\{T_1, T_2\}$ is compatible if and only if $|Y| \leq 1$ or the set $\{T_1|Y, T_2|Y\}$ is a pair of compatible phylogenies on Y . Since the latter set involves trees on the same leaf set (Y), the results from Section 2.3 apply (two rooted phylogenetic trees on the same leaf set are compatible if and only if the union of their clusters is a hierarchy). When \mathcal{P} has more than two trees, the compatibility question becomes more interesting, as we now explain.

4.2.1 ■ The $\mathcal{A}_{\mathcal{R}}$ tree for rooted phylogenies

So far, rooted and unrooted phylogenies have seemed to be rather interchangeable, and the results for one setting seem to carry over naturally to the other. However, this smooth interplay can fail, and we have seen an example of this already, in the study of consensus functions (Proposition 2.13). With tree compatibility, we meet another abrupt disconnect between rooted and unrooted settings. While the compatibility question is NP-hard to decide for collections of unrooted phylogenies on differing leaf sets, it turns out to be easy to decide for rooted phylogenies.¹³

¹³We will shortly meet another example where rooted and unrooted results do not exactly match up in Proposition 4.12.

The simple algorithm for determining the compatibility of any collection \mathcal{R} of rooted phylogenies directly constructs a tree that displays \mathcal{R} when one exists [1]. It relies on the following key concept. Let \mathcal{R} be a collection of rooted phylogenies, and let $\mathcal{L}(\mathcal{R})$ be the union of the leaf sets of the trees. For a subset S of $\mathcal{L}(\mathcal{R})$, let $[\mathcal{R}, S]$ be the graph with vertex set S and with an edge between any two elements $x, x' \in S$ for which there exists an element $y \in S$ with $T|\{x, x', y\} = xx'|y$ for some phylogeny T in \mathcal{R} . For example, for the set \mathcal{R} shown in Fig. 4.1, a and d are adjacent in $[\mathcal{R}, X]$ but not in $[\mathcal{R}, \{a, b, d, e, f\}]$.

With this in hand, we can state a precise characterization for when a set of rooted phylogenies is compatible. Moreover, the proof reveals an algorithm for actually constructing a tree to display the set when one exists.

Proposition 4.2. *A collection \mathcal{R} of rooted phylogenies is compatible if and only if $[\mathcal{R}, S]$ is disconnected for every subset S of $\mathcal{L}(\mathcal{R})$ of size at least 2.*

Proof. First, suppose that \mathcal{R} is compatible, in which case there is a rooted phylogeny T on $X = \mathcal{L}(\mathcal{R})$ that displays \mathcal{R} . For any subset S of X of size at least 2, let $T' = T|S$. Since $\#S \geq 2$, T' has at least two maximal subtrees adjacent to its root, and so we can select leaves x and y from different maximal proper subtrees. Notice that if two elements—say x' and x'' —from X are adjacent in $[\mathcal{R}, S]$, then x' and x'' must lie in the same maximal proper subtree of T' . It follows by transitivity that any two elements of S that are connected by a path in $[\mathcal{R}, S]$ must also lie in the same maximal proper subtree of T' . Since the leaves x and y are from different maximal proper subtrees of T' , $[\mathcal{R}, S]$ must be disconnected.

For the converse result we apply induction on $N = |X|$, where $X = \mathcal{L}(\mathcal{R})$. For $N \leq 3$, the result clearly holds, so we now suppose that it holds for all $N < n$ ($n \geq 4$), and that $|X| = n$. By the assumed condition, $[\mathcal{R}, X]$ is disconnected. Furthermore, if C_1, \dots, C_k , $k \geq 2$, are the components of $[\mathcal{R}, X]$, let $\mathcal{R}_i = \{T|C_i : T \in \mathcal{R}\}$ for $i = 1, \dots, k$. Notice that \mathcal{R}_i is a set of rooted phylogenies on C_i and that for any subset S of C_i ($= \mathcal{L}(\mathcal{R}_i)$), we have $[\mathcal{R}_i, S] = [\mathcal{R}, S]$. By the induction hypothesis, this is disconnected whenever $\#S \geq 2$, so there is a tree $T_i \in RP(C_i)$ that displays \mathcal{R}_i . If we now join the roots of T_1, \dots, T_k to a new root vertex, we obtain a tree $T \in RP(X)$ that displays \mathcal{R} . ■

The algorithm that is implicit in the proof of the “if” part of Proposition 4.2 is called the BUILD algorithm (from [1]) and has a similar flavor to the Adams consensus function of being a sequential partitioning process (starting with X and ending at singleton leaves). One starts by computing the connected components of $[\mathcal{R}, X]$ which form the maximal proper clusters of a hierarchy, and then recursively repeats this procedure on each of these clusters C by replacing $[\mathcal{R}, X]$ with $[\mathcal{R}, C]$ and so on, until either all the clusters obtained are singletons (in which case the tree associated with this hierarchy displays \mathcal{R}) or a cluster C' of size 2 or more occurs, for which $[\mathcal{R}, C']$ is connected (in which case \mathcal{R} is not compatible). The resulting phylogeny is denoted $\mathcal{A}_{\mathcal{R}}$. Figure 4.1 (top) illustrates $\mathcal{A}_{\mathcal{R}}$ for the three input phylogenies, along with the graph $[\mathcal{R}, X]$.

Exercise⁺: Using Proposition 4.2, show that if each of $\mathcal{R} \cup \{ab|c\}$ and $\mathcal{R} \cup \{ac|b\}$ are compatible, then so too is $\mathcal{R} \cup \{bc|a\}$ (the analogous result for quartet trees is not true [164]).

The following result of David Bryant [61] provides an elegant way to view $\mathcal{A}_{\mathcal{R}}$ in terms of the Adams consensus function from Chapter 2. We state it without proof.

Proposition 4.3. *For any compatible set \mathcal{R} of rooted triples, $\mathcal{A}_{\mathcal{R}}$ is the Adams consensus tree of $\langle \mathcal{R} \rangle$.*

It can be shown that the tree $\mathcal{A}_{\mathcal{R}}$ is minimal under refinement [312]. In particular, $\mathcal{A}_{\mathcal{R}}$ is a binary tree if and only if $\#(\mathcal{R}) = 1$. However, in general, $\mathcal{A}_{\mathcal{R}}$ may not be the only tree that displays \mathcal{R} . Charles Semple [312] described a simple construction that allows for exponentially many such minimal trees. Let $\mathcal{R}_1 = \{ab|c, ac|d\}$ and $\mathcal{R}_2 = \{ab|x_1, ab|x_2, \dots, ab|x_k\}$, where the leaf labels $a, b, c, d, x_1, \dots, x_k$ are all distinct, and let $\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2$. Notice that $\langle \mathcal{R}_1 \rangle = \{((a, b), c), d\}$ and that there are 2^k minimal trees that display \mathcal{R} since each leaf x_i can be attached in a minimal tree to either the root or the interior vertex descended from it, independently of the placement of the other x_j leaves.

A further intriguing property of $\mathcal{A}_{\mathcal{R}}$, which is connected to the NNI (nearest neighbor interchange) operation of Section 2.5, was established in Magnus Bordewich's PhD thesis [46].

Theorem 4.4. *For any two rooted binary phylogenies $T, T' \in \langle \mathcal{R} \rangle_B$, there is a sequence of (rooted) NNI operations that transforms T into T' and for which each intermediate tree also lies in $\langle \mathcal{R} \rangle_B$.*

The analogue of this theorem for unrooted trees fails to hold; the two trees in Fig. 2.7 and the set $\mathcal{Q} = \{12|45, 34|16, 56|23\}$ of quartet trees provide a counterexample.

A faster algorithm than BUILD for determining the compatibility of any set \mathcal{R} of rooted phylogenies has recently been described in [105], which requires $O(M \log^2 M)$ steps, where M is the total number of vertices and edges across all the trees in \mathcal{R} .

What if the trees in \mathcal{R} are not compatible? In practice, phylogenies inferred from data on different sets of taxa may not be compatible, and so the BUILD algorithm will not return a tree. Nevertheless, biologists would still like to infer a tree for the entire set of species under study. This has led to the development of various “supertree” approaches, some of which relax the BUILD algorithm, or which otherwise seek to find a tree that comes close to displaying the input trees. For example, a graph-theoretic approach developed in [27] for combining (possibly incompatible) trees on different taxonomic levels was recently applied in the reconstruction of a large “tree of life” in [189]. In the next chapter, we will describe a popular supertree method matrix representation with parsimony (MRP), which is based on coding up the input trees by sequences.

4.2.2 • Compatibility of unrooted phylogenies

Although it is easy to determine whether or not a set of two unrooted phylogenies is compatible, the compatibility question for sets of phylogenies is NP-hard in general even for sets of quartet trees. However, there are special cases where compatibility can be determined in polynomial time. For example, if there is a leaf present in all of the trees in \mathcal{P} , then the BUILD algorithm from Section 4.2.1 can be applied to the rooted trees obtained by deleting leaf x from each tree in \mathcal{P} . A slight relaxation of this would require that for two leaves $x, y \in \mathcal{L}(\mathcal{P})$, each tree in \mathcal{P} must contain either x or y (or both). Even in this relaxed setting, the question of whether or not \mathcal{P} is compatible becomes NP-complete, including for quartet trees (for a proof, see [315], Theorem 6.5.1). Notice that \mathcal{P} is compatible if and only if the associated set $q(\mathcal{P})$ of quartet trees is compatible, and so the compatibility question for unrooted trees is often studied for sets of quartet trees.

Given a sequence $\mathcal{Q} = (q_1, \dots, q_k)$ of quartet trees, if $q_i = ab|cd$ let $S_i = \{\{a, b\}, \{c, d\}\}$, and let $G[\mathcal{Q}]$ be the graph with vertex set $\{(i, A) : A \in S_i, i = 1, \dots, k\}$ and with an edge joining (i, A) and (j, B) whenever $A \cap B \neq \emptyset$. An example of this (stratified) intersection graph is shown in Fig. 4.2 for $\mathcal{Q} = (12|34, 13|45, 24|56)$. It can be shown that \mathcal{Q} is compatible if and only if $G[\mathcal{Q}]$ is chordal or can be made chordal by adding edges of the form (q, A) and (q', B) for $q \neq q'$ [327]. For the quartet trees $\mathcal{Q} = (12|34, 13|45, 24|56)$, the graph $G[\mathcal{Q}]$ is shown in Fig. 4.2. This graph is not chordal, but by adding the allowed edge 13–24 the resulting graph is chordal, which ensures that \mathcal{Q} is compatible.

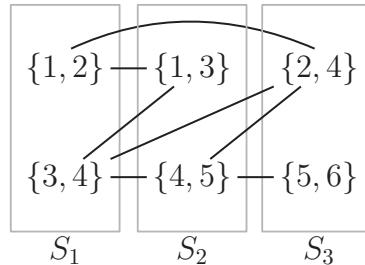


Figure 4.2. The graph $G[\mathcal{Q}]$ for $\mathcal{Q} = (12|34, 13|45, 24|56)$. This graph is not chordal, since the cycle 12–13–34–24–12 has no chord; however, if we add the edge 13–24, then the resulting graph is chordal.

More recently, a different combinatorial characterization has been established for when an arbitrary set \mathcal{Q} of quartet trees is compatible in terms of the existence of a certain “unification operation” on a graph associated with \mathcal{Q} [162]. A further combinatorial result was discovered by Stefan Grünewald [161], who found an elegant sufficient condition for the compatibility of a set of unrooted binary phylogenies which involves just the pattern of intersection of the leaf label sets, not the actual identity of the trees themselves. Informally, the condition requires any subset of trees from the set to contain sufficiently many leaves between them. More formally, given a set \mathcal{P} of binary phylogenies, define the *excess* of \mathcal{P} , denoted $\text{exc}(\mathcal{P})$, by the equation

$$\text{exc}(\mathcal{P}) := |\mathcal{L}(\mathcal{P})| - 3 - \sum_{T \in \mathcal{P}} |\mathring{E}(T)|, \quad (4.4)$$

where $\mathring{E}(T)$ is the set of interior edges of T . One way to interpret this is that $\text{exc}(\mathcal{P})$ is the difference between the number $(|\mathcal{L}(\mathcal{P})| - 3)$ of interior edges in any binary tree that displays the trees in \mathcal{P} , and the total number of interior edges present in the trees in \mathcal{P} . In particular, $\text{exc}(\{T\}) = 0$ for any binary phylogeny T .

The excess index allows us to define a new notion: We say that a collection \mathcal{P} of phylogenies is *slim* if $\text{exc}(\mathcal{P}') \geq 0$ for every nonempty subset \mathcal{P}' of \mathcal{P} . For example, the set of quartet trees $\mathcal{Q} = \{ab|cd, ac|de, bc|ef\}$ is slim, but the set $\mathcal{Q}' = \{ab|cd, ac|de, ae|bc\}$ is not. The following charming result is from [161] (see also [117]), the proof of which is quite intricate.

Theorem 4.5. Every slim set of binary phylogenies is compatible.

As a simple application of Theorem 4.5, suppose that \mathcal{Q} is a set of quartet trees that can be ordered so that each tree introduces at least one leaf not present in the earlier trees. Then \mathcal{Q} is slim (and thus is compatible), since if we take a subset \mathcal{Q}' of \mathcal{Q} of size r , then \mathcal{Q}' must contain at least $r + 3$ leaves, so $\text{exc}(\mathcal{Q}') \geq (r + 3) - 3 - r = 0$. Of course, in this

special case it is fairly easy to see that \mathcal{Q} will be compatible, since we can construct a tree by adding taxa sequentially according to the given order of the quartet trees specified.

Theorem 4.5 provides the key to establishing another nontrivial theorem that we describe in Section 4.3.2.

4.2.3 • The display graph for a set of trees

Let $\mathcal{P} = (T_1, \dots, T_k)$ be a sequence of (unrooted) phylogenies on arbitrary leaf sets (which may or may not overlap), and let X be the union of these leaf sets. Consider the forest consisting of the k disjoint trees T_1, \dots, T_k . Now, for an element $x \in X$ that is a leaf in two or more trees from \mathcal{P} , identify all the leaves labeled x . Repeat this identification for all elements of X present in two or more trees from \mathcal{P} . The resulting graph is called the *display graph for \mathcal{P}* , denoted $G(\mathcal{P})$. An example, taken from [362], is shown in Fig. 4.3. The display graph provides a further way to study compatibility.

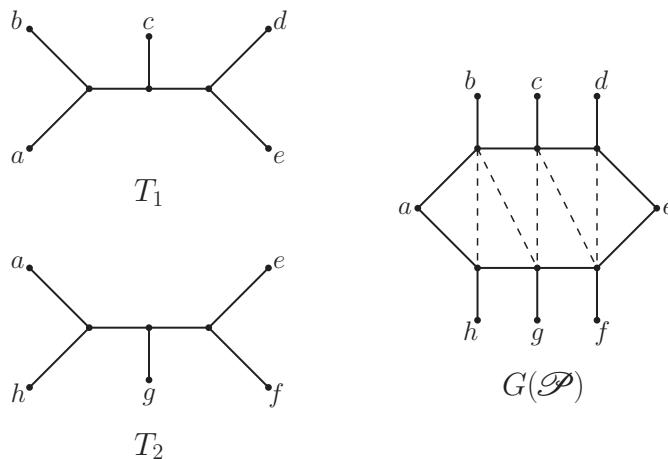


Figure 4.3. Two trees on overlapping leaf sets and their associated display graph $G(\mathcal{P})$ (solid edges). A legal triangulation adds the dashed edges.

This graph was described by [67] and, for quartet trees, by [162] (and further studied by [362] and [213]). One important property of the display graph for \mathcal{P} , from [67], bounds the treewidth of the display graph of any compatible sequence of trees by the number of trees in the sequence.

Proposition 4.6. If $\mathcal{P} = (T_1, T_2, \dots, T_k)$ is compatible, then $\text{tw}(G(\mathcal{P})) \leq k$.

Using this result, determining whether or not \mathcal{P} is compatible is fixed-parameter tractable (FPT) in k (the number of trees in \mathcal{P}) [67]. Here, FPT means that the compatibility question for k phylogenies on a total of n leaves can be determined by an algorithm with a running time of order $f(k)n^s$ for some function f and number s .¹⁴ More recently, [159] established the following restricted converse to Proposition 4.6.

¹⁴This is due to a celebrated result in algorithmic graph theory (Courcelle's Theorem), which states that problems that can be expressed as properties of graphs in a certain way (within monadic second-order logic) can be solved by algorithms with a running time that is polynomial (indeed, linear) in n (but usually exponential in k) for graphs of treewidth at most k .

Proposition 4.7. If $\mathcal{P} = (T_1, T_2, \dots, T_k)$ satisfies $\text{tw}(G(\mathcal{P})) \leq 2$, then \mathcal{P} is compatible.

The same authors of this last result showed that it does not extend further by exhibiting an incompatible set \mathcal{P} of trees for which $\text{tw}(G(\mathcal{P})) = 3$. Nevertheless, in the special case when each tree in $\mathcal{P} = (T_1, T_2, \dots, T_k)$ has the same leaf set, \mathcal{P} is compatible if and only if $\text{tw}(G(\mathcal{P})) \leq k$ (see [364], Theorem 4.1).

A further bound on the treewidth applies when two trees are related by a single tree bisection and reconnection (TBR) operation (cf. Section 2.5). The following result from [213] also has algorithmic implications.

Proposition 4.8. If $\mathcal{P} = \{T_1, T_2\}$ consists of two trees from $B(X)$, then

$$\text{tw}(G(\mathcal{P})) \leq d_{\text{TBR}}(T_1, T_2) + 2.$$

The display graph also provides a way to characterize the compatibility of a set of trees, as shown in [362]. Recall from Chapter 1 that a chordal graph is a graph in which every cycle of length 4 or more has a chord. Clearly, every graph can be made chordal (or “triangulated”) by adding in enough edges, since a complete graph is chordal. However, the condition for the compatibility of trees to hold requires the triangulation to be restricted in a certain way. To describe this restriction in the case of the display graph $G(\mathcal{P})$, we first state two further definitions.

Let us say that an edge e of $G(\mathcal{P})$ is an *internal edge* of $G(\mathcal{P})$ if, in the input tree where it originated, both endpoints of e were interior vertices. A vertex of $G(\mathcal{P})$ is an *internal vertex* of $G(\mathcal{P})$ if it does not correspond to one of the leaf vertices of the trees in \mathcal{P} . A *legal triangulation* of $G(\mathcal{P})$ is the addition of edges to $G(\mathcal{P})$ to produce a chordal graph G' in which the following two restrictions hold:

- (i) Edges of G' that are not in $G(\mathcal{P})$ can only have internal vertices as their endpoints.
- (ii) A clique in G' that contains an internal edge of $G(\mathcal{P})$ can contain no other edge from $G(\mathcal{P})$.

An example is shown in Fig. 4.3. We can now state the main result of [362].

Theorem 4.9. A profile \mathcal{P} of unrooted trees is compatible if and only if the display graph $G(\mathcal{P})$ of \mathcal{P} has a legal triangulation.

Two further characterizations of the compatibility of \mathcal{P} have also recently been described, one in terms of the existence of certain types of cuts in $G(\mathcal{P})$ [363], and the other in terms of restricted triangulations of an intersection graph associated with \mathcal{P} [170].

4.3 • Sets of trees that “define” and “identify” a phylogeny

4.3.1 • Defining a tree

Given a set \mathcal{P} of phylogenetic trees on subsets of X , we say that \mathcal{P} *defines* a phylogenetic X -tree T if T is the unique (up to equivalence) phylogeny on $\mathcal{L}(\mathcal{P})$ that displays each tree in \mathcal{P} ; in other words, $\langle \mathcal{P} \rangle = \{T\}$. A necessary condition for T to be defined by \mathcal{P} is that T must be binary; otherwise there would be a phylogeny T' that strictly refines T , so T' would accompany T in $\langle \mathcal{P} \rangle$, which is impossible if this set has size 1. A set \mathcal{P} of trees

from $P(X)$ with $\#\mathcal{L}(\mathcal{P}) > 3$ defines a phylogeny T if and only if \mathcal{P} is nondegenerate and $q(\mathcal{P})$ defines T .

It is instructive to start by considering the case where \mathcal{P} consists of just two phylogenies, $T_1 \in P(X_1)$ and $T_2 \in P(X_2)$. Let $Y = X_1 \cap X_2$ denote the set of leaves common to both trees. In the simplest setting where $X_1 = X_2 (= Y)$, the phylogenies \mathcal{P} define a tree if and only if T_1 and T_2 are compatible and the union of the splits of these two phylogenies correspond to a binary phylogeny $T' \in B(Y)$. In general, the condition that $T_1|Y$ and $T_2|Y$ must define a phylogeny is necessary but not sufficient for T_1 and T_2 to define a tree. For a characterization we require two further conditions, and the following result is part of the “tree-joining theorem” from the PhD thesis of Sebastian Böcker [35]. Note that in part (iii), we write $\sigma' \leq \sigma$ for σ' a split of Y , and $\sigma = A|B$ is a split of X provided that $A \cap Y|B \cap Y = \sigma'$.

Proposition 4.10. *A set $\mathcal{P} = \{T_1, T_2\}$ of two phylogenies defines a tree if and only if the following three conditions hold:*

- (i) $T_1|Y$ and $T_2|Y$ define a phylogeny T' on Y ;
- (ii) $|\Sigma(T_i)| - |\Sigma(T_i|Y)| = 2|X_i| - 2|Y|$ for $i = 1, 2$; and
- (iii) for every $\sigma' \in \Sigma(T')$, we have

$$\#\{\sigma \in \Sigma(T_1) : \sigma' \leq \sigma\} \leq 1 \text{ or } \#\{\sigma \in \Sigma(T_2) : \sigma' \leq \sigma\} \leq 1.$$

Informally, condition (ii) states that, except for the restricted trees $T_1|Y$ and $T_2|Y$, the remainder of T_1 and T_2 must be binary, while condition (iii) says that a pendant subtree of T_1 and a pendant subtree of T_2 do not both attach to the same edge of T' . This is illustrated in Fig. 4.4. Proposition 4.10 provides a polynomial-time algorithm to determine whether or not an arbitrary pair of phylogenies defines a phylogeny.

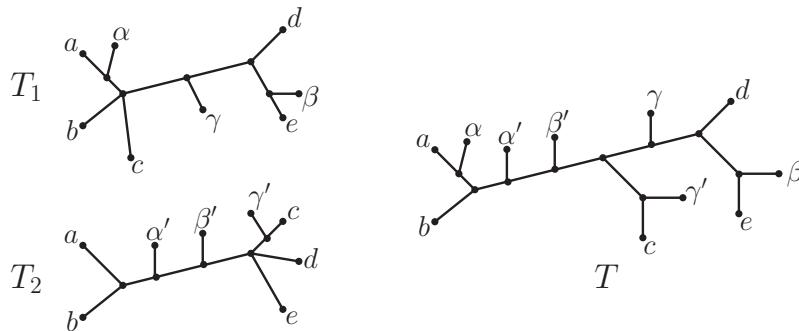


Figure 4.4. The two trees T_1 and T_2 define the tree T . The set of leaves shared by T_1 and T_2 is $Y = \{a, b, c, d, e\}$. However, if T_1 had an additional leaf (say δ) attached to the pendant edge leading to c , then this tree and T_2 would no longer define any tree since δ and γ' could attach in various ways to the edge of T leading to c .

Notice that when T_1 and T_2 are both binary trees, Proposition 4.10 can be simplified as follows.

Corollary 4.11. *A set $\mathcal{P} = \{T_1, T_2\}$ of two binary phylogenies defines a tree if and only if $T_1|Y = T_2|Y$ and condition (iii) of Proposition 4.10 holds.*

Corollary 4.11 provides the basis of a simple polynomial-time solution for the following *maximal defining label set* problem in the special case $k = 2$: Given a set $\mathcal{P} = \{T_1, T_2, \dots, T_k\}$ of binary phylogenies, find a largest subset W of leaves for which $\mathcal{P}|W$ defines a binary phylogeny [307]. The complexity of this problem for values of k greater than 2 is currently open.

Notice that if a quartet tree $q = ab|cd$ is displayed by a phylogeny T , then there is at least one interior edge of T whose deletion results in one component containing a and b , and another component containing c and d . Provided that there is only one such interior edge e , we say that e is *distinguished* by the quartet tree $ab|cd$. Equivalently, this means that the path in T connecting a to b intersects e at one endpoint of this edge, and the path connecting c to d intersects e at its other endpoint.

Proposition 4.12.

- (i) If a set \mathcal{Q} of quartet trees defines a phylogeny $T \in B(X)$, then \mathcal{Q} distinguishes each interior edge of T .
- (ii) A set \mathcal{R} of rooted triples defines a phylogeny $T \in RB(X)$ if and only if $T \in \langle \mathcal{R} \rangle$ and each interior edge of T is distinguished by \mathcal{R} .
- (iii) If \mathcal{Q} is a set of quartet trees, each of which contains a fixed leaf x , then \mathcal{Q} defines a phylogeny $T \in B(X)$ if and only if $T \in \langle \mathcal{Q} \rangle$ and each interior edge of T is distinguished by \mathcal{Q} .

The proof of part (i) of Proposition 4.12 follows from the observation that if e is not distinguished, then \mathcal{Q} cannot define T , since the tree obtained by collapsing e would also display \mathcal{Q} . The same argument applies for the “only if” direction of part (ii), while the “if” direction can be established by induction on the height of T [327]. Part (iii) follows from part (ii) via the bijection o_x described in Section 2.1.

A consequence of part (i) of Proposition 4.12 is that if \mathcal{Q} defines $T \in B(n)$, then $|\mathcal{Q}| \geq n - 3$ since a binary tree with n leaves has $n - 3$ interior edges, and each quartet can distinguish, at most, one interior edge. Moreover, part (iii) shows that sets of this minimal size exist and are easy to construct.

Exercise⁺: For any binary phylogeny T on $[n]$, construct a sequence of displayed quartet trees q_1, q_2, \dots, q_{n-3} that collectively define T , such that for all values of i with $2 \leq i \leq n - 3$, q_i has precisely one leaf that is not present in the earlier quartet trees [*Hint:* consider a cherry of T .]

Although $n - 3$ well-chosen quartet trees are sufficient to define a tree T in $B(n)$, the choice of these depends on knowing T . If one is required to ask a (truthful) oracle a series of queries of the form “what is the quartet tree $T|Y_i$?” for a sequence of subsets Y_i of subsets of X of size 4 (where the choice of Y_i can depend on the answers given to the questions posed earlier), then a phylogeny $T \in P(n)$ can be reconstructed in $O(n^2)$ queries. In addition, if T is binary, then just $O(n \log n)$ queries suffice [288]. By contrast, the smallest subset S of $\binom{X}{4}$ for which the quartet trees $\{T|Y : Y \in S\}$ define T for every tree $T \in B(X)$ is exactly $\binom{n}{4} - (n - 4)$ [260].

For any rooted phylogeny $T \in RB(n)$, any minimal set of rooted triples that defines T has exactly the same size, namely $n - 2$, by Proposition 4.12(ii). However,

when we move to unrooted binary trees, the minimal sets of quartet trees that define $T \in B(n)$ can be larger than the minimum possible size of $n - 3$. For example, $\mathcal{Q} = \{12|35, 24|57, 13|47, 34|56, 15|67\}$ defines the caterpillar tree $12|345|67$, but no proper subset of it does. In this example, $|\mathcal{Q}| = n - 2$ (i.e., it has one quartet tree more than is strictly required to define the caterpillar tree).

The intriguing question of how large a minimal defining set of quartet trees can be has been studied in [110], where the following result was established: Any largest set \mathcal{Q} of quartet trees on $[n]$ that (i) defines a phylogeny on $[n]$ and (ii) is minimal in the sense that no proper subset of the set defines that tree, has a size that lies between order n^2 and n^3 . More precisely, there is such a set \mathcal{Q} satisfying properties (i) and (ii), and with $|\mathcal{Q}| = \sum_{i=1}^{n-3} \lceil i/2 \rceil = \frac{1}{4}n^2 + O(n)$, and any set satisfying properties (i) and (ii) must also satisfy the bound $|\mathcal{Q}| \leq \sum_{i=1}^{n-3} i(i+1) = O(n^3)$.

We have seen that just because a set of restricted quartet trees distinguishes every edge of an unrooted binary phylogenetic tree, \mathcal{Q} need not define that tree. Proposition 4.12(iii) provided one case where this holds (when each quartet tree contains a fixed leaf). However, if we strengthen the condition on \mathcal{Q} that it distinguishes not only the interior edges but also all interior paths, then we obtain a set that defines a tree. More formally, given a tree $T \in B(n)$, a collection $\mathcal{Q} \subseteq \mathcal{Q}(T)$ of the displayed quartet trees of T is said to be a *generous cover* for T if for every two interior vertices u, v , of T , there is a quartet tree $q = ij|kl \in \mathcal{Q}$ for which $u = \text{med}_T(i, j, v)$ and $v = \text{med}_T(k, l, u)$. In other words, the interior vertex of q that is adjacent to i and j corresponds to u , while the interior vertex of q that is adjacent to k and l corresponds to v . This stronger condition turns out to be enough to ensure that \mathcal{Q} defines T . More precisely, for a set \mathcal{Q} of quartet trees on X , let $\text{cl}_1(\mathcal{Q})$ be the minimal set \mathcal{Q}' of quartet trees containing \mathcal{Q} that satisfies the following dyadic closure rule:

If $ab|cd$ and $ab|ce$ are both present in \mathcal{Q}' , then $ab|de$ is also present in \mathcal{Q}' .

In practice, $\text{cl}_1(\mathcal{Q})$ is obtained by starting with \mathcal{Q} and just adding quartet trees according to this closure rule until no further quartet trees can be added. The following result is from [109] and will play an important role at one point in the proof of Theorem 8.4.

Proposition 4.13. *If T is a binary phylogeny and $\mathcal{Q} \subseteq \mathcal{Q}(T)$ is a generous cover for T , then $\text{cl}_1(\mathcal{Q}) = \mathcal{Q}(T)$. In particular, \mathcal{Q} defines T .*

This last result also shows that T can be recovered from \mathcal{Q} by a polynomial-time algorithm.

4.3.2 ■ The Böcker–Dress–Grünewald theorem

Suppose that a set \mathcal{Q} of quartet trees on X defines a binary tree $T \in B(X)$. By Proposition 4.12(i), we must have $|\mathcal{Q}| \geq |X| - 3$, an inequality that can be restated in terms of the excess function defined above (eqn. (4.4)) as

$$\text{exc}(\mathcal{Q}) \leq 0.$$

A set \mathcal{Q} for which $\text{exc}(\mathcal{Q}) = 0$ is said to be *excess-free*. In particular, if a set \mathcal{Q} is excess-free and defines a tree, then each interior edge of T is distinguished by precisely one quartet tree in \mathcal{Q} . Moreover, for any subset \mathcal{Q}_0 of $\mathcal{Q}(T)$ that distinguishes every interior edge of T and is excess-free, the collection χ of excess-free subsets of \mathcal{Q}_0 has the following combinatorial property:

$$\mathcal{Q}, \mathcal{Q}' \in \chi \text{ and } \mathcal{Q} \cap \mathcal{Q}' \neq \emptyset \implies \mathcal{Q} \cap \mathcal{Q}' \in \chi \text{ and } \mathcal{Q} \cup \mathcal{Q}' \in \chi.$$

This *patchwork* property is one of several ingredients in the following theorem, which occupied much of the PhD thesis of Sebastian Böcker [35] in joint work with his supervisor, Andreas Dress. A more simplified (though still quite nontrivial) proof was recently found by Stefan Grünewald, in which his result above (Theorem 4.5) played a key role. For details of this proof, the reader is referred to either [161] or [117].

Theorem 4.14. *Suppose that a set \mathcal{Q} of two or more quartet trees is excess-free and defines a binary tree. Then \mathcal{Q} contains a maximal hierarchy of excess-free subsets, each of which defines a tree.*

One way to restate this theorem is to call an excess-free set of quartet trees that defines a binary tree a *tight set*; with this terminology, Theorem 4.14 is equivalent to the statement that any tight set of size 2 or more is the disjoint union of two tight sets. This means that one can readily construct from \mathcal{Q} the tree T that \mathcal{Q} defines. To do so, one begins by selecting a tight set of size 2 and replacing it by the binary tree on five leaves that this set defines. Continuing this process, at each step, one identifies two of the binary trees present that define a further binary tree (using Corollary 4.11) and replaces these two trees by the tree they define. Figure 4.5 illustrates a simple example of this.

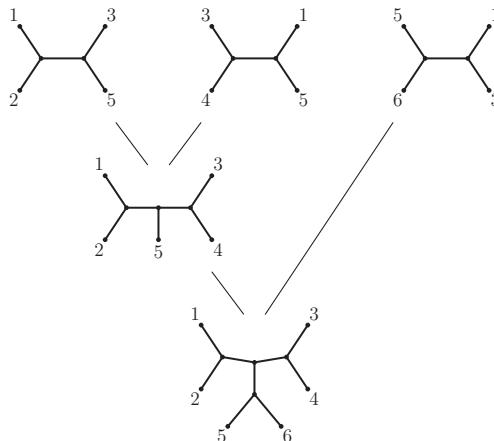


Figure 4.5. The set \mathcal{Q} , consisting of the three quartet trees at the top, forms an excess-free set that defines a tree. Consequently, the tree it defines can be constructed by sequentially combining pairs of trees (see text for details).

Theorem 4.14 is easy to prove in two special cases:

- (i) If there is a leaf x common to all the quartet trees in \mathcal{Q} (in which case the trees can be regarded as rooted, and so defining a tree is equivalent to distinguishing every interior edge), and
- (ii) \mathcal{Q} can be ordered so that each quartet tree in the series after the first introduces exactly one new leaf.

Theorem 4.14 has some strong algorithmic consequences. First, suppose that \mathcal{Q} is a set of quartet trees. Then if \mathcal{Q} is excess-free, it is possible to determine in polynomial-time whether or not \mathcal{Q} defines a tree and, if so, to find that tree. This is pertinent since the computational problem of deciding whether or not an arbitrary set of compatible

quartet trees defines a tree has been shown to be hard (NP-hard [172] and, independently, co-NP-complete [41]).

What if the set \mathcal{Q} set of quartet trees is not excess-free? In that case, it is still possible to determine in polynomial-time whether or not \mathcal{Q} contains an excess-free subset that defines some (unknown) phylogenetic X -tree. This relies on the concept of a second dyadic closure rule. For a set \mathcal{Q} of quartet trees on X , let $\text{cl}_2(\mathcal{Q})$ be the minimal set \mathcal{Q}' of quartet trees containing \mathcal{Q} that satisfies the following dyadic closure rule:

If $ab|cd$ and $ac|de$ are both present in \mathcal{Q}' , then $ab|ce$, $ab|de$, and $bc|ed$ are also present in \mathcal{Q}' .

In practice, $\text{cl}_2(\mathcal{Q})$ is obtained by starting with \mathcal{Q} and just adding quartet trees according to this closure rule until no further quartet trees can be added. If \mathcal{Q} is a compatible set of quartet trees on X that is excess-free and defines some tree $T \in B(X)$, then Theorem 4.14 tells us that $\text{cl}_2(\mathcal{Q}) = \mathcal{Q}(T)$. Let us now suppose that \mathcal{Q}' is some arbitrary collection of quartet trees on X . If $\text{cl}_2(\mathcal{Q}')$ contains a contradictory pair of quartet trees (i.e., $ab|cd$ and $ac|bd$ for some $a, b, c, d \in X$), then \mathcal{Q}' is incompatible. On the other hand, if $\text{cl}_2(\mathcal{Q}') = \mathcal{Q}(T)$, then \mathcal{Q}' is compatible and defines T . Theorem 4.14 ensures that one of these two outcomes must occur whenever \mathcal{Q}' contains an (unknown) excess-free subset that defines some phylogenetic X -tree T . Of course, it is also possible that $\text{cl}_2(\mathcal{Q}')$ does not contain a contradictory pair of quartet trees and yet also has size less than $\binom{n}{4}$ for $n = |X|$, in which case, no conclusion can be drawn.

Leaf addition sequences and “short quartets.” A special case of an excess-free set of quartet trees that define a binary tree $T \in B(n)$ is any set \mathcal{Q} that (i) can be ordered so that each quartet tree after the first introduces one new leaf (not present in the earlier quartets), and (ii) distinguishes every interior edge of T . Note that property (i) alone suffices for \mathcal{Q} to be compatible.

The set of quartet trees in Fig. 4.5 is of this form, but not all excess-free sets of quartet trees that define a phylogeny can be ordered so as to satisfy property (i). One case where they do is the following. Let $\sigma(T)$ consists of those subsets of leaves that appear as one of the two blocks of a nontrivial split of T (i.e., $\sigma(T) = \{A \subset X : A|\overline{A} \in \dot{\Sigma}(T)\}$). If we delete from T an interior edge e along with the four edges that are incident with e , then the resulting four connected components of T partition X into the four sets A_e, A'_e, B_e, B'_e , each lying in $\sigma(T)$, and with e corresponding to the split $A_e \cup A'_e | B_e \cup B'_e$. The property that \mathcal{Q} distinguish each interior edge e of T is equivalent to the existence of a function $\psi : \sigma(T) \rightarrow X$ that satisfies $\psi(A) \in A$ for all $A \in \sigma(T)$, and for which the quartet tree $\psi(A_e)\psi(A'_e)|\psi(B_e)\psi(B'_e)$ lies in \mathcal{Q} for each interior edge e . Strengthening these conditions on ψ by appending a “stability” condition gives a defining set of quartet trees for T [38].

Proposition 4.15. *For any unrooted binary phylogenetic X -tree T , consider any function $\psi : \sigma(T) \rightarrow X$ that satisfies the following two properties:*

- $\psi(A) \in A$ for all $A \in \sigma(T)$; and
- if $A, B \in \sigma(T)$, with $A \subset B$ and $\psi(B) \in A$, then $\psi(A) = \psi(B)$.

If a subset \mathcal{Q} of $\mathcal{Q}(T)$ contains the quartet tree $\psi(A_e)\psi(A'_e)|\psi(B_e)\psi(B'_e)$ for each interior edge e of T , then \mathcal{Q} defines T .

An example of a function ψ that satisfies these properties is the following. First, impose an arbitrary total order on X . Next, for a set $A \in \sigma(T)$, where $A|\overline{A}$ corresponds to

interior edge e , let $\psi(A)$ be the minimal element of X (under the imposed total order) of those leaves of A that are closest to edge e . For example, this is how the three quartet trees in Fig. 4.5 were generated from the tree T they define (using the natural ordering of [6]). Proposition 4.15 assures us that this set of “short quartets” defines T . This result has proved useful for establishing upper bounds on the number of characters required to accurately reconstruct a phylogeny under Markovian models of character evolution (a topic addressed further in Chapter 8).

4.3.3 • Identifying a tree

There is also a more general notion of “define” that applies to all phylogenies (including nonbinary ones). Let \mathcal{P} be a set of unrooted phylogenies, and let X be the union of their leaf sets. We say that \mathcal{P} identifies a tree $T \in P(X)$ if \mathcal{P} is compatible and T is the unique minimally refined phylogenetic X -tree that displays \mathcal{P} . In other words, $\langle \mathcal{P} \rangle$ coincides with the set of phylogenetic X -trees that refine T . If T is binary, then the two notions—define and identify—are equivalent. There are analogous notions of “define” and “identify” for rooted phylogenetic trees also.

A nonempty subset \mathcal{Q} of $\mathcal{Q}(T)$ identifies T if and only if

$$\mathcal{Q}(T) = \bigcap_{T' \in \langle \mathcal{Q} \rangle} \mathcal{Q}(T'). \quad (4.5)$$

This result (from [164]) follows by observing that since $T \in \langle \mathcal{Q} \rangle$, eqn. (4.5) holds if and only if $\mathcal{Q}(T) \subseteq \mathcal{Q}(T')$ for all $T' \in \langle \mathcal{Q} \rangle$, which holds if and only if each tree T' that displays \mathcal{Q} refines T (by Lemma 4.1). The analogue of eqn. (4.5) holds for rooted phylogenies and rooted triples.

Notice that if a set \mathcal{R} of rooted triples identifies a rooted phylogeny T , then T must be the tree $\mathcal{A}_{\mathcal{R}}$ constructed by the BUILD algorithm (since $\mathcal{A}_{\mathcal{R}}$ is a minimal tree that displays \mathcal{R}). The question of whether or not \mathcal{R} identifies a rooted phylogeny can be solved in polynomial-time (using BUILD) as the following exercise shows.

Exercise⁺: Show that a compatible set \mathcal{R} of rooted phylogenies identifies a tree in $R(X)$ if and only if, for every rooted triple $xy|z$ that is displayed by $\mathcal{A}_{\mathcal{R}}$, both $\mathcal{R} \cup \{xz|y\}$ and $\mathcal{R} \cup \{yz|x\}$ are incompatible.

When $T \in P(X)$ has at least one interior edge (i.e., is not a star tree), then $\mathcal{Q}(T)$ identifies T , since $\mathcal{L}(\mathcal{Q}(T)) = X$ and any phylogeny T' on X that displays $\mathcal{Q}(T)$ necessarily satisfies $\mathcal{Q}(T) \subseteq \mathcal{Q}(T')$, which implies that T' refines T by Lemma 4.1. Similarly, for any tree $T \in RP(X)$ that has at least one interior edge, the set $\mathcal{R}(T)$ of rooted triples displayed by T identifies T . But, in general, we do not require all the quartet trees (or rooted triples) to identify a tree (just as we saw for the corresponding notion of define). In the rooted case, the following theorem is from [164], where d^+ refers to the out-degree (of a vertex).

Theorem 4.16. *For any rooted phylogenetic X -tree T with at least one interior edge, any smallest set of rooted triples that identifies T has size*

$$\sum_{(u,v) \in \overset{\circ}{E}(T)} (d^+(u)-1)(d^+(v)-1).$$

Notice that for a rooted binary tree, where $d^+(u) = 2$ for each interior vertex u , this gives $|\mathring{E}(T)| = n - 2$, in agreement with Proposition 4.12(ii). There is a directly analogous notion of identifying an unrooted phylogenetic tree for subsets of restricted quartet trees, and the following unrooted analogue of Theorem 4.16 is from [162].

Theorem 4.17. *For any (unrooted) phylogenetic X -tree T with at least one interior edge, any smallest set of quartet trees that identifies T has size*

$$\tilde{q}(T) := \sum_{\{u,v\} \in \mathring{E}(T)} q(d(u)-1, d(v)-1),$$

where $q(r,s) := \lceil \frac{r(s-1)}{2} \rceil$ for all $r,s \geq 2$.

An interesting consequence follows: $\tilde{q}(T)$ has a minimum value of $n - 3$ and a maximum value of $\lfloor \lceil \frac{n}{2} - 1 \rceil^2 \rfloor$ over all trees in $P(n)$. Moreover, the trees that realize these minimum and maximum values have been precisely characterized [162].

4.4 ▪ Agreement subtrees

Given two phylogenies on the same leaf set (both rooted or both unrooted) we have described various ways to measure how similar they are. For example, we can ask how many clusters (or splits for unrooted trees) they share or how many subtree transfers or NNI moves are needed to convert one tree into another. However, there is another very natural way to measure similarity; namely how many leaves need to be removed in order for the two trees to agree? More formally, if $T_1, T_2 \in P(X)$, or $T_1, T_2 \in RP(X)$, then a *maximal agreement set* (MAS) for this pair of trees is a largest subset Y of X for which $T_1|Y = T_2|Y$, in which case such a tree is called a *maximal agreement subtree* (MAST) for T_1 and T_2 .

In general, there can be (exponentially) many MASTs; however, an MAST for any two trees both in $P(n)$ or both in $RP(n)$ can be computed in polynomial-time by a dynamic programming algorithm. Moreover, the notion of an MAST extends beyond a pair of phylogenies on $[n]$ to a subset \mathcal{P} of k phylogenies on $[n]$ in the obvious way. When $k = 3$, finding an MAST is already NP-hard. However, provided that at least one phylogeny in \mathcal{P} has maximal vertex degree d , there is an algorithm for computing an MAST for this subset that is polynomial in n and k (for fixed values of d) [134].

MASTs also give rise to interesting mathematical questions.

Q1 What is the smallest MAST for two trees in $B(n)$?

Q2 If two trees in $B(n)$ are chosen independently and randomly according to some probability distribution (e.g., uniform or YH), what is the expected MAST size?

Notice that Q1 has a trivial answer if we consider rooted rather than unrooted binary trees. In this case, it is easy to construct pairs of rooted caterpillar trees for each $n = |X|$ that agree only on subsets of X of size 2. However, for unrooted trees, the question is more interesting. For example, any pair of caterpillar trees on a common set of n leaves has a MAST of order \sqrt{n} by the classic combinatorial result that any permutation of the numbers from 1 to $n^2 + 1$ must have a monotone (increasing or decreasing) subsequence of length $n + 1$.¹⁵ However, it is also clear that the MAST for trees in $B(n)$ can be smaller

¹⁵This is a special case of the celebrated Erdős-Szekeres theorem, which states that any permutation of the numbers $1, \dots, (r-1)(s-1) + 1$ has a monotone increasing subsequence of length r or a monotone decreasing subsequence of length s .

than \sqrt{n} . For example, for any caterpillar tree and any perfect binary tree (where each leaf is the same number of edges from some vertex) the MAST has size $O(\log n)$. This has led to the following more precise formulation of Q1:

Q1' Is the smallest MAST for two distinct trees in $B(n)$ of size $\Omega(\log n)$?

In other words, is there a constant $c > 0$ so that every two distinct phylogenies on X have an MAST of size at least $c \cdot \log |X|$? At present, the best known result is that every pair of different trees in $B(n)$ has a MAST of size $\Omega(\sqrt{\log n})$, a result from [239], which established a fundamental Ramsey-type result that every sufficiently large binary tree contains either a caterpillar of given size or a perfect rooted subtree of given size. This paper also established that the answer to Q1' is “yes,” provided one of the two trees is a caterpillar tree or both of the trees are perfect rooted phylogenies (in the latter case an improved bound applies).

For Q2, the expected MAST size for both the uniform and YH distributions for two trees in $RB(n)$ is known to grow no faster than $O(\sqrt{n})$ and at least as fast as n^α for values of α in $(0, 0.5)$ [26].

There is also a weaker notion than MAST which arises in the *maximum compatible tree* (MCT) problem. Given a set \mathcal{P} of k phylogenies on $[n]$ this problem asks for a largest subset Y of $[n]$ for which $\{T|Y : T \in \mathcal{P}\}$ is compatible. If the trees in \mathcal{P} are all binary, this problem is clearly equivalent to MAST. However, in contrast to MAST, the MCT problem turns out to be NP-hard even when $\#\mathcal{P} = 2$. But once again, a polynomial-time algorithm can be devised for MCT if one bounds both $\#\mathcal{P}$ and the maximal vertex degree of the trees in \mathcal{P} (for further details, see [166] and the references therein).

4.4.1 ■ The quartet metric

Induced quartet trees provide yet another metric to compare two unrooted phylogenies. The *quartet metric* on $P(n)$, $d_{\mathcal{Q}}$, is the number of restricted quartet trees that lie in exactly one of the two trees (i.e., the symmetric difference of the set of restricted quartet trees), so $d_{\mathcal{Q}}(T, T') = |\mathcal{Q}(T) \Delta \mathcal{Q}(T')|$. The quartet metric is clearly computable in polynomial-time and has the advantage over the Robinson-Foulds (RF) metric d_{RF} of being less sensitive to perturbations to a tree such as moving a leaf across the tree (such a move can drastically alter d_{RF} but has a relatively minor impact on $d_{\mathcal{Q}}$). An interesting property of the quartet metric is that for any two trees $T, T' \in B(n)$ there is a sequence of subtree prune and regraft (SPR) moves that converts T into T' in such a way that quartet distance $d_{\mathcal{Q}}$ between T' and the k th tree in the sequence is strictly decreasing with k (see [47], Theorem 4.1).

The expected distance between any tree $T \in B(n)$ and a random tree \mathcal{T} selected from $B(n)$ under any exchangeable model (cf. Section 3.2.3) is also easily seen to be $\frac{2}{3} \binom{n}{4}$ since we can write $d_{\mathcal{Q}}(T, \mathcal{T})$ as $\sum_{Y \in \binom{[n]}{4}} \mathbb{I}_{T|Y \neq \mathcal{T}|Y}$, and so

$$\mathbb{E}[d_{\mathcal{Q}}(T, \mathcal{T})] = \sum_{Y \in \binom{[n]}{4}} \mathbb{E}[\mathbb{I}_{T|Y \neq \mathcal{T}|Y}] = \sum_{Y \in \binom{[n]}{4}} \mathbb{P}(T|Y \neq \mathcal{T}|Y) = \frac{2}{3} \binom{n}{4},$$

where the last equality follows from the exchangeability assumption (the same result holds if T is also chosen randomly and independently of \mathcal{T} under such a model).

An interesting and nontrivial question in extremal combinatorics concerns the diameter of $d_{\mathcal{Q}}$ on $B(n)$ and, in particular, its limiting value as $n \rightarrow \infty$. Let

$$d(n) = \max_{T, T' \in B(n)} \{d_{\mathcal{Q}}(T, T')\} / \binom{n}{4}.$$

Clearly, $d(n) \leq 1$; moreover, $d(6) < 1$ (i.e., any two binary trees on six leaves must share at least one restricted quartet tree), and $d(n)$ is monotone decreasing with n [20]. This monotonic decline is equivalent to the monotone increase with n in the smallest proportion ν_n of quartet trees shared by two trees from $B(n)$. To establish this monotonicity, suppose that T and T' are two trees in $B(n)$ that minimize ν_n . Let us count the number of pairs (S, Y) , where (i) S is a subset of $[n]$ of size k , (ii) Y is a subset of $[n]$ with $T|Y = T'|Y$, and (iii) $Y \subseteq S$. Since there are $\nu_n \binom{n}{4}$ such choices of Y satisfying (ii) and $\binom{n-4}{k-4}$ choices of S for any given $Y \in \binom{[n]}{4}$ satisfying (i) and (iii), we obtain

$$\#(S, Y) = \nu_n \binom{n}{4} \binom{n-4}{k-4}. \quad (4.6)$$

On the other hand, if we first fix S , then the number of four-element subsets Y of S for which $T|Y = T'|Y$ is the same as the number of sets $Y \in \binom{S}{4}$ for which $(T|S)|Y = (T'|S)|Y$, and since $T|S$ and $T'|S$ are binary trees on the same set of k leaves, the number of choices for Y is at least $\nu_k \binom{k}{4}$. Thus since there are $\binom{n}{k}$ choices for S , we have

$$\#(S, Y) \geq \nu_k \binom{k}{4} \binom{n}{k}. \quad (4.7)$$

Combining eqns. (4.6) and (4.7) gives

$$\nu_n \geq \nu_k \cdot \frac{\binom{k}{4} \binom{n}{k}}{\binom{n}{4} \binom{n-4}{k-4}} = \nu_k,$$

where equality arises because the fraction term on the right is 1, since both the numerator and denominator count the number of pairs (S, Y) , where S is a subset of $[n]$ of size k and Y is a subset of S of size 4.

Bandelt and Dress conjectured in 1986 that $\lim_{n \rightarrow \infty} d(n) = \frac{2}{3}$; in other words, the proportion of quartet trees on which two binary trees with n leaves differ is asymptotically the same as for two random trees [20]. Resolving this conjecture has not been easy; however, [13] provided some evidence in its favor by establishing that the maximum diameter is at most $(0.69 + o(1)) \binom{n}{4}$, and for pairs of caterpillar trees it is $(\frac{2}{3} + o(1)) \binom{n}{4}$, as predicted by the conjecture.

A related question applies when \mathcal{Q} is an arbitrary collection of quartet trees on $[n]$, with one quartet tree for each four-element subset of $[n]$. In this setting we can ask, what is the largest subset of \mathcal{Q} that is compatible? In general, this is a hard problem, but there is always a phylogeny that displays at least $\frac{1}{3}$ of the quartet trees in \mathcal{Q} . To see why, select a tree $\mathcal{T} \in B(X)$ uniformly at random, and let $Y_{\mathcal{Q}}$ denote the number of quartet trees from \mathcal{Q} that are displayed by \mathcal{T} . The expected value of the random variable $Y_{\mathcal{Q}}$ can be readily computed as follows:

$$\mathbb{E}[Y_{\mathcal{Q}}] = \sum_{ab|cd \in \mathcal{Q}} \mathbb{P}(\mathcal{T}| \{a, b, c, d\} = ab|cd) = \frac{1}{3} |\mathcal{Q}|.$$

Since $\mathbb{E}[Y_{\mathcal{Q}}] = \frac{1}{3} |\mathcal{Q}|$, at least one tree T in $B(X)$ must display at least $\frac{1}{3}$ of the quartet trees in \mathcal{Q} . A recent paper [14] explored several related questions, and we present just one result here.

Theorem 4.18. A set \mathcal{Q} of $\Theta(n \log n)$ quartet trees exists such that every subset $\mathcal{Q}' \subseteq \mathcal{Q}$ of size $\Theta\left(\frac{n}{\log n}\right)$ is compatible, yet no subset of \mathcal{Q} containing more than $\frac{1}{3} + \epsilon$ proportion of the quartet trees in \mathcal{Q} is compatible.

4.5 • Phylogenetic decisiveness and terraces

Suppose that χ is a collection of subsets X_1, \dots, X_k of X . For any tree T in $B(X)$, let $T|\chi = \{T|X_1, \dots, T|X_k\}$. The collection χ is said to be *decisive* for a tree $T \in B(X)$ if $T|\chi$ defines T . There is an analogous notion for rooted phylogenies. We say that a collection $\chi = \{X_1, \dots, X_k\}$ of subsets of X is *decisive for unrooted phylogenies* (respectively, for rooted phylogenies) if χ is decisive for every $T \in B(X)$ (respectively, if χ is decisive for every $T \in RB(X)$). In this section, we study the conditions on χ for it to be decisive in this strong sense.

The biological context for this question is that the distribution of genes across species is often patchy, but if one could determine the (unknown) phylogeny for each (restricted) subtree, then when might these fit together uniquely into a “supertree”? To make progress, we need one more definition: let us say that χ *covers 3-sets* if for every subset Y of X of size 3, there is a set $X_i \in \chi$ with $Y \subseteq X_i$.

Proposition 4.19. Let χ be a collection of k subsets of X .

- (i) If χ is decisive for (rooted or unrooted) phylogenies, then χ covers 3-sets.
- (ii) χ is decisive for rooted phylogenies if and only if χ covers 3-sets.
- (iii) If the intersection of the sets in χ is nonempty, then χ is decisive for unrooted phylogenies if and only if χ covers 3-sets.
- (iv) If χ is decisive for (rooted or unrooted) phylogenies and if each set contains, at most, the proportion f of all the species present in X , then $k \geq \frac{1}{f^3}$.

Proof: To see why part (i) holds, suppose $Y = \{a, b, c\} \subseteq X$ is not contained within any set X_i from χ . Let T be any binary phylogeny that has a, b as a cherry, and for which the vertex v adjacent to a and to b is also adjacent to the vertex v' to which c is adjacent. Let T' be the phylogeny obtained by interchanging leaves b and c . Then $T|\chi = T'|\chi$, so χ cannot be decisive. Part (ii) is due to Mareike Fischer; it follows from part (i) and (for the converse) from Proposition 4.12(iii). Part (iii) follows from part (ii), and part (iv) is left as an exercise. ■

Exercise: Use part (i) of Proposition 4.19 to prove part (iv).

Notice the asymmetry between decisiveness for rooted and unrooted phylogenies in Proposition 4.19. An example, due to Peter Humphries [197], shows that part (ii) does not apply to unrooted phylogenies. Take $X = \{1, 2, \dots, 8\}$, and let χ_8 be the set of subsets of X of size 4 which can be formed by taking any one of the sets $\{1, 2\}, \{3, 4\}, \{5, 6\}$, and $\{7, 8\}$ and adding any two additional elements (i.e., $4 \times \binom{6}{2} = 60$ sets in total). Then χ_8 covers all 3-sets but it is not decisive for unrooted phylogenies, since there are three perfect binary trees in $B(8)$ with cherries $\{1, 2\}, \{3, 4\}, \{5, 6\}$, and $\{7, 8\}$, and each of them gives rise to the same restricted subtrees on the subsets in χ_8 .

Nevertheless, it is still possible to provide a different type of mathematical characterization of decisiveness, as shown in the following result from [340].

Theorem 4.20. *A collection χ of subsets of X is decisive for unrooted phylogenies if and only if for every partition Π of X into four blocks, there exist elements x_1, x_2, x_3, x_4 , one from each of the four blocks of Π , for which $\{x_1, x_2, x_3, x_4\} \subseteq Y$ for some $Y \in \chi$.*

This theorem shows immediately that the collection χ_8 above is not decisive, as it violates the condition stated in Theorem 4.20 for the partition $\Pi = \{\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\}\}$. The complexity of determining whether a collection of subsets of X is decisive for unrooted phylogenies is currently unresolved, and it may well be NP-complete.

Phylogenetic terraces. Patchy taxon coverage has a direct combinatorial consequence for the tree reconstruction methods that seek to optimize (e.g., minimize) some objective function based on how well the data “fit” each tree. This can lead to large collections of equally optimal trees (i.e., a flat landscape of trees), referred to as a (phylogenetic) *terrace* [307]. To describe this connection with terraces, we first describe some properties of commonly used scoring functions.

Let X be a set of species. For a phylogeny $T \in B(X)$ and a gene G that is present in some or all of the species in X , consider a scoring function s that assigns a positive real value to the pair (G, T) . If X_G denotes the subset of X consisting of the species that possess gene G , then most scoring functions s in computational biology satisfy the following equation:

$$s(G, T) = s(G, T|X_G). \quad (4.8)$$

This condition essentially says that the presence of species for which the gene is not present should not affect how well the data for the gene “fit” the tree under consideration.

Now suppose the data consists of a sequence of genes $\mathcal{G} = (G_1, G_2, \dots, G_k)$. Given the score $s(G_i, T)$ for each i , how might we combine them to obtain a score $s(\mathcal{G}, T)$ for how well this collection of genes “fits” T ? A natural option is simply to form a linear sum and let

$$s(\mathcal{G}, T) = \sum_{i=1}^k s(G_i, T).$$

We say that any such scoring scheme is *linear*. Given the data \mathcal{G} , we seek to find a tree that minimizes $s(\mathcal{G}, T)$. While linearity may seem a strong condition to impose, it turns out that certain standard phylogenetic methods select a tree that minimizes a linear scoring scheme. These include a method we describe in the next chapter (maximum parsimony) and a method from Chapter 8 called maximum likelihood (provided that parameters such as the branch lengths of a tree are optimized independently from gene to gene) for which $s(G_i, T)$ is equal to minus the log-likelihood of T for G_i ; for further details see [307]. Both of these methods also satisfy eqn. (4.8), as do several others that fail linearity (e.g., maximum likelihood with branch length parameters linked across genes).

Now suppose that T is a binary tree that has some particular score (e.g., the optimal score) for \mathcal{G} under some scoring function s . Let $X_i = X_{G_i}$ for $i = 1, \dots, k$, let χ be the sequence (X_1, \dots, X_k) , and let $T|\chi = \{T|X_1, \dots, T|X_k\}$.

Proposition 4.21. *If s is a linear scoring function that satisfies eqn. (4.8), then each phylogeny in $\langle T|\chi \rangle$ has exactly the same score.*

Proof. If $T' \in \langle T | \chi \rangle$, then $T'|X_i = T|X_i$ for all $i = 1, \dots, k$ (since T is binary), and so $s(\mathcal{G}, T')$ can be written as

$$\sum_{i=1}^k s(G_i, T') = \sum_{i=1}^k s(G_i, T'|X_i) = \sum_{i=1}^k s(G_i, T|X_i) = \sum_{i=1}^k s(G_i, T) = s(\mathcal{G}, T). \quad \blacksquare$$

Thus, in general, there will be a set of equally optimal trees (the “terrace” containing an optimal tree) or, indeed, a set of trees having any given fixed (suboptimal) score. In real applications, this number can be very large, for example, 61 million equally optimal (maximum likelihood) trees for a data set consisting of 298 species of grasses on three genes [307]. The existence of large, flat landscapes of trees can make the search for optimal trees by hill-climbing approaches more problematic.

4.5.1 ■ Decisiveness for random taxon coverage

Suppose a sequence $\chi_{n,k,p} = (\mathcal{X}_1, \dots, \mathcal{X}_k)$ of subsets of X is generated by the following stochastic process. The elements of \mathcal{X}_i are generated independently by placing each element x of X in \mathcal{X}_i independently with probability p . In other words, the events $\{\mathbb{I}_{x \in \mathcal{X}_i} : x \in X, i \in [k]\}$ are stochastically independent, and $\mathbb{P}(\mathbb{I}_{x \in \mathcal{X}_i}) = p$. This can be viewed as the simplest “null model” for partial taxon coverage, and it allows us to answer some basic questions under such a model, such as which values of p and k is it likely that $\chi_{n,k,p}$ will be phylogenetically decisive for rooted (or unrooted) phylogenies?

For rooted phylogenies, the following results show that the required value of k grows at the rate $\Theta(\log n)$, and as p approaches zero it also grows at the rate $1/p^3$ (cf. Proposition 4.19(iv)).

Proposition 4.22. *For all $p \in (0, 1)$, $\epsilon \in (0, 1]$, and $r > 0$, the probability that $\chi_{n,k,p}$ is phylogenetically decisive for rooted phylogenies is*

- (i) *at least $1 - \epsilon$ if k is at least $\frac{\ln(\binom{n}{3}/\epsilon)}{-\ln(1-p^3)}$,*
- (ii) *at most e^{-r} if k is less or equal to $\frac{\ln(\lfloor n/3 \rfloor)}{-\ln(1-p^3)}$.*

Proof. For each $i \in [k]$, any given subset Y of X of size 3 is contained in \mathcal{X}_i with probability p^3 , so the probability that none of the k sets in $\chi_{n,k,p}$ contains Y is $(1 - p^3)^k$. Let \mathcal{E} be the event that at least one such subset $Y \in \binom{X}{3}$ has the property that it fails to be a subset of any of the sets in $\chi_{n,k,p}$. By Boole’s inequality,

$$\mathbb{P}(\mathcal{E}) \leq \binom{n}{3} (1 - p^3)^k.$$

Now $\chi_{n,k,p}$ is decisive for rooted phylogenies under the complementary event $\overline{\mathcal{E}}$ (by Proposition 4.19(ii)) and $\binom{n}{3} (1 - p^3)^k \leq \epsilon$ for the values of k in part (i) of Proposition 4.22.

For part (ii), let $m = \lfloor n/3 \rfloor$, and let Y_1, \dots, Y_m be m disjoint subsets of X of size 3. The probability that Y_i is a subset of at least one of the k sets from $\chi_{n,k,p}$ is $1 - (1 - p^3)^k$, and since the m sets (Y_i) are disjoint, the probability of the event \mathcal{E}' that all of the Y_i sets are subsets of sets in $\chi_{n,k,p}$ is $(1 - (1 - p^3)^k)^m$. Thus

$$\mathbb{P}(\mathcal{E}') \leq e^{-r} \Leftrightarrow 1 - (1 - p^3)^k \leq e^{-r/m} \Leftrightarrow k \leq \frac{-\ln(1 - e^{-r/m})}{-\ln(1 - p^3)}.$$

By the inequality $-\ln(1-e^{-x}) \geq -\ln(x)$ for $x = r/m > 0$, we now get $\mathbb{P}(\mathcal{E}') \leq e^{-r}$ when k is less than or equal to the expression in the second part of Proposition 4.22. Part (ii) now follows by noting that \mathcal{E}' is a necessary condition for $\chi_{n,k,p}$ to be decisive, again by Proposition 4.22(ii). ■

What if we just require a sequence of subsets of X that is decisive for a given binary phylogeny T , rather than being decisive for all trees in $RB(X)$? Presumably this will require fewer sets, but how many fewer? The results for this question (the analogues of Proposition 4.22) were derived in [306], based on the characterization for when a set of rooted triples defines a rooted phylogeny (Proposition 4.12(ii)). The probability that $\chi_{n,k,p}$ is decisive for T is at least $1 - \epsilon$ if k is at least

$$\frac{\ln((n-2)/\epsilon)}{-\ln(1-p^3)}.$$

Notice that this is indeed smaller than the expression in part (i) of Proposition 4.22), but not by much. For example, if $n = 50$ and $p = 0.5$, then $k = 64$ sets suffice to define a given tree with a probability of at least 0.99, and just $k = 109$ suffice to define all trees (i.e., to be decisive for rooted phylogenies) with a probability of at least 0.99.

It might be expected that the lower bound on k in part (ii) of Proposition 4.22 might also need to be decreased if we wish to define only a given tree. However, exactly the same lower bound applies. Regarding unrooted phylogenies, the results in Proposition 4.22 are the same, apart from a small modification to part (i), where p^3 is replaced by p^4 (the proof uses Proposition 4.19(ii)).

A different but related question is to fix two subsets X_1 and X_2 of X , select T from $B(X)$ uniformly at random, and ask for the probability that $\mathcal{P} = \{T|X_1, T|X_2\}$ defines T . This probability, $p(X_1, X_2)$, is close to 1 precisely when the size of the symmetric difference $|X_1 \Delta X_2|$ is sufficiently small in comparison to the size of $|X_1 \cap X_2|$ [340].

4.5.2 • Disentangling trees

We have seen how each tree $T \in B(n)$ is defined by the set of all its $\binom{n}{4}$ restricted quartet trees (or from just $\Theta(n)$ -sized subsets of this set). In this section, we extend the quartet determination result from single trees to subsets of trees from $B(n)$. The motivation for this question comes from considering so-called “phylogenetic mixture models,” which we will consider in Chapter 7. As a starting case, suppose we have an unknown subset $\mathcal{P} = \{T_1, T_2\}$ of $B(n)$ consisting of two trees, and that $\mathcal{P}|Y := \{T_1|Y, T_2|Y\}$ is given for all subsets Y of X of size k where $k \geq 4$. Can we then determine the set \mathcal{P} ?

Since single binary phylogenies are determined by their restricted quartet trees, we might suspect that $k = 4$ would work. However, it doesn’t, and neither does $k = 5$, as the example shown in Fig. 4.6 reveals. Nevertheless, when we get to $k = 6$, the collections $\{T_1|Y, T_2|Y\}$ for every subset Y of $[n]$ of size 6 uniquely determine the trees in \mathcal{P} (indeed, \mathcal{P} can be reconstructed by an algorithm that is polynomial in n). The important point here is that this holds for all values of $n \geq 6$, and so this result (from [242]) is sometimes referred to as the “six-to-infinity theorem.” Moreover, there is a generalization of this result to more than two trees (from [197] and [346]) which we now describe briefly.

Suppose $\mathcal{P} = \{T_1, \dots, T_k\}$ is a collection of k trees from $B(n)$ (i.e., $\mathcal{P} \in \binom{B(n)}{k}$), where $n \geq k$. For a subset Y of X , consider the set of restricted trees $\mathcal{P}|Y = \{T_1|Y, \dots, T_k|Y\}$, which may be a set of size less than k . For $n \geq k$, we say that $\binom{[n]}{m}$ disentangles $\binom{B(n)}{k}$ if for any two sets $\mathcal{P}, \mathcal{P}' \in \binom{B(n)}{k}$, $\mathcal{P} \neq \mathcal{P}'$ implies that $\mathcal{P}|Y \neq \mathcal{P}'|Y$ for some subset Y .

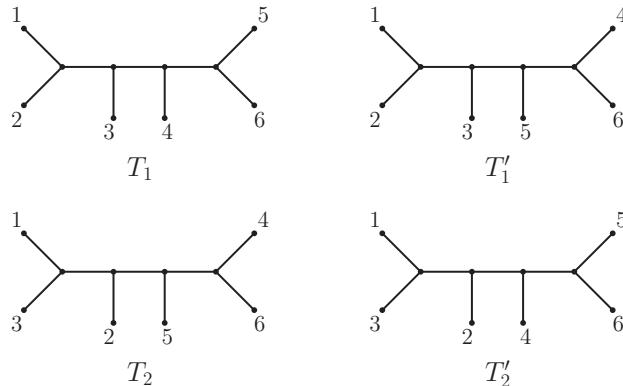


Figure 4.6. For $\mathcal{P} = \{T_1, T_2\}$ and $\mathcal{P}' = \{T'_1, T'_2\}$, we have $\mathcal{P}|Y = \mathcal{P}'|Y$ for all subsets Y of $\{1, 2, 3, 4, 5, 6\}$ of size 5. Notice also that $\Sigma(T_1) \cup \Sigma(T_2) = \Sigma(T'_1) \cup \Sigma(T'_2)$.

of $[n]$ of size m . In other words, the function

$$\binom{B(n)}{k} \times \binom{[n]}{m} \rightarrow 2^{B(m)},$$

$$(\mathcal{P}, Y) \mapsto \mathcal{P}|Y$$

is one-to-one and, in particular, \mathcal{P} is uniquely determined by the sets $\mathcal{P}|Y$ for all subsets Y of $[n]$ of size m .

For $k \geq 1$, let $D(k)$ be the smallest value of m for which all $\binom{[n]}{m}$ disentangles $\binom{B(n)}{k}$. Thus, $D(1) = 4$ and the six-to-infinity theorem (and Fig. 4.6) shows that $D(2) = 6$. However, for $k > 2$ it is not immediately obvious that $D(k)$ should even be finite. Nevertheless, [197] showed that this is indeed the case, with $D(k) \leq 3k$ and also that $D(k)$ increases monotonically with k , and $D(k) = \Omega(\log k)$. In [346], a much tighter (logarithmic) upper bound was described for a value $\tilde{D}(k)$ that is defined in a similar way to $D(k)$ but where the multiplicities (i.e., how many trees in \mathcal{P} restrict to each given tree in $\mathcal{P}|Y$) are also available. Using a novel argument which applies a result from polytope theory at a key step, [346] shows that $\tilde{D}(k) \leq 3(\lfloor \log_2 k \rfloor + 1) + 1$; a more direct argument for this inequality appears in [236].

Exercise: If a phylogeny $T \in B(9)$ is defined by a pair of trees in $B(k)$, explain why k must be at least 6. Can every tree $T \in B(9)$ be defined by a pair of trees from $B(6)$? [Hint: Consider a tree in $B(9)$ that has a central symmetry vertex (as in Fig. 1.3(c)).]

Chapter 5

Phylogenies based on discrete characters

5.1 • Characters, homoplasy, and perfect phylogeny

A function f from the set of species X into some set S of states is referred to by biologists as a *character* on X .¹⁶ In the case when f takes at most r distinct values on X , we say that f is an *r-state character* (this holds if $|S| = r$ but it may also hold when $|S| > r$). For example, $f(x)$ might be a morphological character that describes the number of legs that species x has, or a genetic character that describes the nucleotide at a particular position in a genetic sequence for species x . That is, $f(x)$ describes some “characteristic” of x that we compare across other species in X . When $r = 2$, we refer to f as a *binary character*. If a phylogenetic X -tree describes the evolution of a set of species, a character tells us the states of the species at the leaves, but not of the hypothetical ancestral species that correspond to the interior vertices of the tree.

There are typically a myriad of ways to explain how the character could have evolved in the tree from some ancestral state at the root. It is possible that in a path from the root to a leaf, a *reversal* occurs, where a state s_1 changes to state s_2 and later back to s_1 ; for example, in birds, wings first evolved, but then in some species (e.g., kiwi) they subsequently disappeared. It is also possible for “convergent evolution” to occur, where state s_1 at some vertex changes to s_2 down two edge-disjoint paths that start from that vertex. Again, wings provide an example: from an ancestor of birds and mammals that lacked wings (state s_1), convergent evolution in birds and in mammalian bat lineages led to wings (state s_2). A character whose evolution on a given tree can be explained without postulating any reversal or convergent events is said to be “homoplasy-free.”

The notion of being homoplasy-free can be defined more easily if we suppress the rooting of the tree and consider unrooted trees. Formally, we will say that a character f on X is *homoplasy-free* on an unrooted phylogenetic X -tree T if $f : X \rightarrow S$ has an extension $F : V \rightarrow S$ to the set V of all vertices of T so that F is constant on the path between any two vertices with the same F value (equivalently, for each $\alpha \in f(X)$, the subgraph of T induced by the set of vertices v with $F(v) = \alpha$ is connected). Any such F is said to be a *valid extension*. Notice that the homoplasy-free condition is considerably weaker than requiring that the actual evolution of the character on some rooting of the tree involve no reversals or convergent evolution; it merely requires that the character *could* have evolved in this way.

¹⁶A more general concept of a character on X is a function $\chi : X' \rightarrow S$ [315], where $X' \subseteq X$; however, we do not require this extra generality.

Here are two other ways to characterize when a character $f : X \rightarrow S$ is homoplasy-free on a phylogenetic X -tree T :

- f has an extension F to all the vertices of T for which F assigns different states to the endpoints of at most (and therefore exactly) $|f(X)| - 1$ edges.
- For any four leaves x, y, w, z with $f(x) = f(y) \neq f(w) = f(z)$ the two paths $P(T; x, y)$ and $P(T; w, z)$ have no vertex in common.

For a third characterization, we need to extend the notion of the splits of a tree discussed in Chapter 2. Given a partition Π of X into two or more blocks, and a block $B \in \Pi$, let $T[B]$ denote the minimal subtree of T that connects the leaves in Π . We say that Π is a *convex* on T if the collection of subtrees $\{T[B] : B \in \Pi\}$ is vertex-disjoint.¹⁷ Given a character f let $\Pi(f)$ be the partition of X induced by the equivalence relation $x \approx x'$ if and only if $f(x) = f(x')$. Then a character $f : X \rightarrow S$ is homoplasy-free on a phylogenetic X -tree T if and only if

- the partition $\Pi(f)$ is convex on T .

These concepts are illustrated in Fig. 5.1.

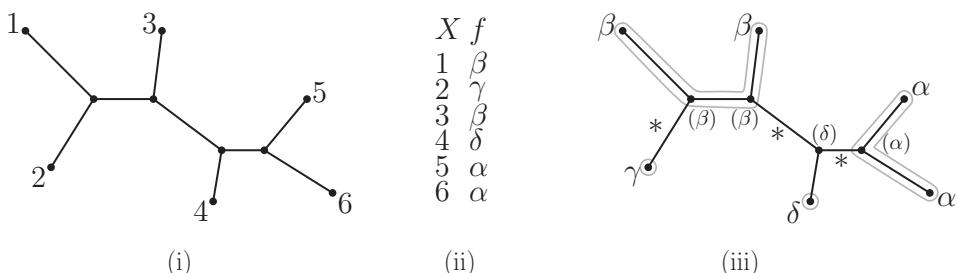


Figure 5.1. (i) A binary phylogenetic tree $T \in B(6)$, and (ii) a four-state character f that is homoplasy-free on T . In (iii), an extension F of f on T is shown by the values in parentheses for the interior vertices, and for this extension the three edges that have different states assigned to their endpoints is indicated by *. This extension F is not unique, since the interior vertex assigned state δ could instead have been assigned state α or β . The induced partition $\Pi(f)$ is convex on T since the four circled subtrees $T[B], B \in \Pi(f)$ are disjoint. However if leaf 2 had been assigned state α instead of state γ , then the resulting character would no longer be homoplasy-free on T .

A sequence $\mathcal{C} = (f_1, f_2, \dots, f_k)$ of characters on X is said to have a *perfect phylogeny* if and only if there is a phylogenetic X -tree on which each character is homoplasy-free (the tree is said to be a perfect phylogeny for those characters). An alternative phrase, popular in biology, and in earlier literature, is to say that the characters are “compatible.”

A perfect phylogeny represents an ideal scenario in which the data (sequence of characters) have evolved without reversals or convergent evolution. This type of homoplasy-free evolution might be expected to hold (i) when the state changes occur very rarely; and/or (ii) when the number of potential states is very large, so each change is likely to be to a new state (for example, the order of genes on a chromosome under random rearrangement operations); or (iii) for certain genomic insertion data such as “short interspersed elements” (SINEs) in mammalian genetics (which helped establish that the hippopotamus is the closest living species to the whale [278]). These and other types of (largely) homoplasy-free

¹⁷An equivalent condition for this is that for some subset E' of edges from T , Π is the set of equivalence classes of X under the equivalence relation $x \sim y$ precisely if x and y are connected in the graph $T - E'$.

“retroposon” data have been used more recently to help reconstruct phylogenies for other Eukaryotic taxa (e.g., birds, fish, etc). Perfect phylogenies have also found application in other areas of classification, including language evolution, and the study of the mutational history of tumor cells [175]. For other types of data (e.g., aligned DNA sequence data) homoplasy-free evolution is unlikely to hold, and Chapters 7 and 8 describe stochastic techniques for inferring a tree from such data.

For any sequence $\mathcal{C} = (f_1, f_2, \dots, f_k)$ of **binary** characters,

- (i) \mathcal{C} has a perfect phylogeny if and only if each pair of characters does;
- (ii) when \mathcal{C} has a perfect phylogeny, there is a unique minimal one (under the refinement partial order), and the interior edges of this phylogeny correspond to the non-trivial splits in the collection: $\bigcup_{i=1}^k \{B|\bar{B} : B \in \Pi(f_i)\}$.

Neither (i) nor (ii) applies in the general setting of r -state characters, except in certain special cases. One notable exception is when each pair of characters f_i and f_j from \mathcal{C} is *strongly compatible*, which means that either $\Pi(f_i) = \Pi(f_j)$ or there is a block of $\Pi(f_i)$ and a block of $\Pi(f_j)$ whose union is X . When \mathcal{C} is a sequence of pairwise strongly compatible characters, then both the above properties (i) and (ii) hold. Notice that when each character in \mathcal{C} takes exactly two states, pairwise strong compatibility is equivalent to requiring that the set of bipartitions of X in $\{\Pi(f_i), i = 1, \dots, k\}$ comprise a pairwise compatible set of X -splits; in particular, the set $\{\Pi(f_i), i = 1, \dots, k\}$ has size at most $2n - 3$, where $n = |X| \geq 2$. By comparison, if each character in \mathcal{C} takes at least two states then if \mathcal{C} is pairwise strongly compatible, the set $\{\Pi(f_i), i = 1, \dots, k\}$ has size at most $3n - 5$ for $n \geq 2$ (Proposition 3.2 of [187]).

Exercise⁺: Given a sequence $\mathcal{C} = (f_1, f_2, \dots, f_k)$ of binary characters on X , describe a polynomial-time algorithm to determine whether there exist two trees $T, T' \in P(X)$ with the property that for each f_i character in \mathcal{C} , f_i is homoplasy-free on T or T' (or both) [234].

A character f on X is *trivial* if all of the blocks of $\Pi(f)$, except perhaps one, are singleton sets. Every trivial character f on X is homoplasy-free on any tree $T \in P(X)$, and so given a sequence of characters $\mathcal{C} = (f_1, \dots, f_k)$, \mathcal{C} has a perfect phylogeny if and only if the pruned sequence obtained by deleting every trivial character from \mathcal{C} has one. Thus in most of the statements that follow in this chapter we can assume, without loss of generality, that \mathcal{C} consists of only nontrivial characters.

Exercise: Show that a character f on X is trivial if and only if (f, g) has a perfect phylogeny for every character g on X .

We next describe three quite different mathematical characterizations for a sequence of characters to have a perfect phylogeny. The first provides a direct link to the previous chapter, in particular the span $\langle \mathcal{Q} \rangle$ for a set of quartet trees \mathcal{Q} . For any sequence \mathcal{C} of characters on X , let $q(\mathcal{C})$ be the set of all induced quartet trees $aa'|bb'$, where a, a' lies in one block of a character from \mathcal{C} and b, b' lies in another block of the same character (note that a, a', b, b' are all distinct).

Proposition 5.1. *Let \mathcal{C} be a sequence of characters on X . Then $T \in P(X)$ is a perfect phylogeny for \mathcal{C} if and only if T displays every quartet tree in $q(\mathcal{C})$. In particular, \mathcal{C} has a perfect phylogeny if and only if either $q(\mathcal{C}) = \emptyset$ (i.e., all characters are trivial) or $\langle q(\mathcal{C}) \rangle \neq \emptyset$.*

The proof amounts to showing that T displays each quartet tree induced by a character f_i from \mathcal{C} if and only if the partition $\Pi(f_i)$ is convex on T .

Proposition 5.1 shows that determining whether an arbitrary set of characters has a perfect phylogeny is NP-hard, since (i) we saw in the last chapter that the problem of determining whether or not $\langle \mathcal{Q} \rangle \neq \emptyset$ is NP-hard, and (ii) for any set \mathcal{Q} of quartet trees we can replace each quartet tree $ab|cd \in \mathcal{Q}$ by a character f that has $\{a, b\}, \{c, d\}$ as the only two nontrivial blocks in $\Pi(f)$, resulting in a sequence \mathcal{C} of characters with $\mathcal{Q} = q(\mathcal{C})$.

The existence of a perfect phylogeny for a sequence $\mathcal{C} = (f_1, \dots, f_k)$ of characters on X also has an attractive graph-theoretic characterization involving the *partition intersection graph* of \mathcal{C} , denoted $\text{int}(\mathcal{C})$. This is the graph that has the vertex set $V_{\mathcal{C}} = \bigcup_{i=1}^k \{(f_i, B) : B \in \Pi(f_i)\}$, with an edge between (f_i, A) and (f_j, B) precisely if $A \cap B \neq \emptyset$. Notice that the existence of such an edge requires $i \neq j$, so $\text{int}(\mathcal{C})$ is a k -partite graph. An example of $\text{int}(\mathcal{C})$ for a set \mathcal{C} of three characters is shown in Fig. 5.2(i).

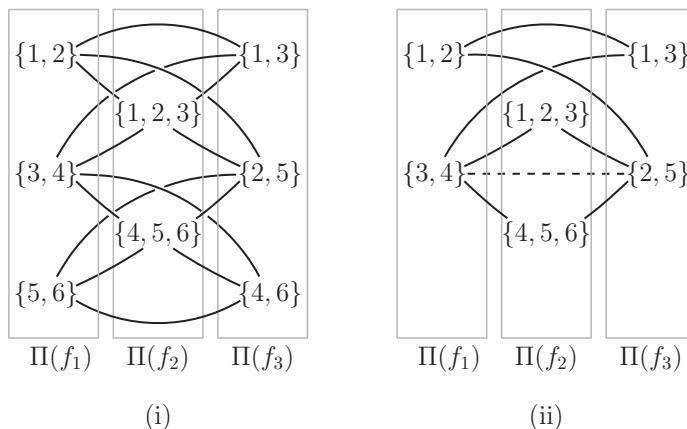


Figure 5.2. (i) The partition intersection graph $\text{int}(\mathcal{C})$ for a set \mathcal{C} consisting of two 3-state characters f_1, f_3 and one binary character f_2 , where $\Pi(f_1) = \{\{1,2\}, \{3,4\}, \{5,6\}\}$ (the three vertices at left), $\Pi(f_2) = \{\{1,2,3\}, \{4,5,6\}\}$ (middle two vertices), and $\Pi(f_3) = \{\{1,3\}, \{2,5\}, \{4,6\}\}$ (the three vertices at right); (ii) $\text{int}(\mathcal{C})$ has the 4-cycle $\{3,4\} - \{1,2,3\} - \{2,5\} - \{4,5,6\} - \{3,4\}$, and since an additional chord between $\{1,2,3\}$ and $\{4,5,6\}$ cannot be added, we are forced to introduce the chord $\{3,4\} - \{2,5\}$ (dashed line). This, in turn, creates a further 4-cycle: $\{1,2\} - \{1,3\} - \{3,4\} - \{2,5\} - \{1,2\}$, and neither of the two possible chords ($\{1,2\} - \{3,4\}$ and $\{1,3\} - \{2,5\}$) can be added. This shows that \mathcal{C} has no perfect phylogeny.

For binary characters, it is not hard to show that \mathcal{C} has a perfect phylogeny if and only if $\text{int}(\mathcal{C})$ is chordal.

Exercise⁺: Show that a sequence \mathcal{C} of binary characters has a perfect phylogeny if and only if $\text{int}(\mathcal{C})$ is a chordal graph.

In general (i.e., for binary or nonbinary characters) if $\text{int}(\mathcal{C})$ is chordal, then \mathcal{C} has a perfect phylogeny. The converse is not true; if \mathcal{C} has a perfect phylogeny, then $\text{int}(\mathcal{C})$ need not be chordal. However, in that case $\text{int}(\mathcal{C})$ can be made chordal by adding edges, subject to the constraint that no edge is added between two blocks of the same partition.

To see why, suppose that T is a perfect phylogeny for \mathcal{C} . Let $\text{int}_T(\mathcal{C})$ be the graph obtained from $\text{int}(\mathcal{C})$ adding an edge between (f_i, A) and (f_j, B) whenever $A \cap B = \emptyset$ and $T[A]$ and $T[B]$ have at least one vertex in common (thus $i \neq j$ since T is a perfect

phylogeny for \mathcal{C}). Then the vertices of $\text{int}_T(\mathcal{C})$ can be regarded as a collection of subtrees of T (by the association $(j, B) \mapsto T[B]$), and there is an edge between any two vertices of $\text{int}_T(\mathcal{C})$ precisely if those subtrees have a vertex in common. From Section 1.2.2, the intersection graph of a collection of subtrees of any tree forms a chordal graph, and so $\text{int}_T(\mathcal{C})$ is chordal.

Figure 5.2(ii) illustrates how this necessary condition can be used to show that a particular set \mathcal{C} has no perfect phylogeny. With further work one can obtain the following precise characterization (for details, see [315]).

Theorem 5.2. *\mathcal{C} has a perfect phylogeny if and only if $\text{int}(\mathcal{C})$ is chordal or can be made chordal by adding edges, subject to the constraint that no edge is added between two blocks of the same partition.*

This second characterization provides a nice proof of a classic result found by Fred (Buck) McMorris and George Estabrook in the mid-1970s: $\mathcal{C} = (f_1, f_2)$ has a perfect phylogeny if and only if the intersection graph of $\Pi(f_1) \cup \Pi(f_2)$ is acyclic. To see why this follows, simply notice that (i) if $\Pi(f_1) \cup \Pi(f_2)$ is acyclic, it is already chordal; and (ii) if $\Pi(f_1) \cup \Pi(f_2)$ has a cycle and if we introduce additional edges to make this graph chordal, we create a 3-cycle, which is not possible in a bipartite graph without destroying the bipartite property.

Another direct consequence of this second characterization concerns the treewidth of $\text{int}(\mathcal{C})$: If a sequence $\mathcal{C} = (f_1, \dots, f_k)$ of characters on X has a perfect phylogeny, then

$$\text{tw}(\text{int}(\mathcal{C})) \leq k. \quad (5.1)$$

To see why this inequality must hold, recall that if \mathcal{C} has a perfect phylogeny T , then the graph $\text{int}_T(\mathcal{C})$ (defined above) is a chordal graph, and this graph is just $\text{int}(\mathcal{C})$ with zero or more additional edges added. By construction, $\text{int}_T(\mathcal{C})$ has no clique larger than k , since it retains the property of $\text{int}(\mathcal{C})$ of being k -partite. The inequality now follows from the definition of treewidth in Chapter 1 that involves chordal completion.

Exercise: Suppose that $\mathcal{C} = (f_1, f_2, \dots, f_k)$ is a sequence of characters on X , and that each character assigns a particular state, say 0, to strictly more than half the elements in X . Suppose T is a perfect phylogeny for \mathcal{C} and that F_i is any valid extension of f_i for each i . Show that T has vertex v for which $F_i(v) = 0$ for all $i = 1, \dots, k$. [Hint: The Helly property for a collection of subtrees of a tree may be helpful.]

Our third characterization of the existence of a perfect phylogeny for \mathcal{C} relies on the slight generalization of phylogenetic trees to the notion of X -trees, described near the end of Section 1.3.

Proposition 5.3. *A sequence of characters $\mathcal{C} = (f_1, \dots, f_k)$ on X has a perfect phylogeny if and only if there is an X -tree $T_i = (V, E, \phi)$ with $\{\phi^{-1}(v) : v \in V\} - \{\emptyset\} = \Pi(f_i)$ for each $i \in \{1, \dots, k\}$, and $\bigcup_{i=1}^k \Sigma(T_i)$ constitutes a (pairwise) compatible set of X -splits.¹⁸*

We will describe an application of this result shortly.

¹⁸The set $\Sigma(T)$ of splits of an X -tree T is defined in the same way as for a phylogenetic tree, namely, as the bipartitions of X that arise by deleting an edge of T .

Algorithms and special cases. The computational problem of determining whether or not a collection of characters has a perfect phylogeny is NP-complete in general, but polynomial-time algorithms exist when a bound is placed on either the number of characters [251] or the number of states per character (r) [210]. In the special cases where $r = 2$ and $r = 3$, a collection of r -state characters has a perfect phylogeny if and only if every subset of size r of the characters has one. However, the “if” direction fails for larger values of r , as there is a set of $\lfloor \frac{r}{2} \rfloor \cdot \lceil \frac{r}{2} \rceil + 1$ characters on $r \geq 4$ states that do not have a perfect phylogeny even though every proper subset does [318].

We now consider the special case where $\mathcal{C} = (f_1, f_2, \dots, f_k)$ is a sequence of characters each of which takes either two or three states. In this case, the existence of a perfect phylogeny has a simple characterization by a connection with a classical problem in logic called 2-SAT, which can be solved in linear time. An instance of 2-SAT is any conjunction of clauses, each involving at most two Boolean variables (possibly negated), and it is said to have a satisfying assignment if there are truth values for each Boolean variable so that every clause in the conjunction is true. For example, suppose that, in a court case, witnesses have stated the following three opinions as to who may or may not have been involved in a crime: “Peter or Susan,” “John or not Peter,” and “not John or not Susan.” Then this has a satisfying assignment (i.e., all three statements can be correct) if John and Peter were both involved and Susan was not.

The connection between the character compatibility for at most three states and 2-SAT was made explicit in [169]; the idea is to associate a Boolean variable x_{is} with each character f_i and each state s that f_i takes, and code up the compatibility question as an instance of 2-SAT. First, if character f_i takes exactly two states (s and s' , say) form the binary clause $(x_{is} \wedge x_{is'})$. If f_i takes three states (e.g., s , s' , and s''), form the following conjunction of binary clauses $(x_{is} \vee x_{is'}) \wedge (x_{is} \vee x_{is''}) \wedge (x_{is'} \vee x_{is''})$. Notice that the satisfiability of these three clauses under some truth assignment (“true” or “false”) to each Boolean variable is equivalent to the condition that at least two of the variables x_{is} , $x_{is'}$, and $x_{is''}$ are true. Finally, for each pair of characters f_i and f_j , if the X -splits $f_i^{-1}(s)|(X - f_i^{-1}(s))$ and $f_j^{-1}(s')|(X - f_j^{-1}(s'))$ are incompatible, then add the clauses $(\neg x_{is} \vee \neg x_{js'})$. Here, a satisfying assignment ensures that x_{is} and $x_{js'}$ are not both true. It then follows from Proposition 5.3 that the characters $\mathcal{C} = (f_1, f_2, \dots, f_k)$ have a perfect phylogeny if and only if the conjunction of all of these clauses has a satisfying assignment.

5.1.1 • Capturing a perfect phylogeny

When a sequence of characters has a perfect phylogeny T , we can also ask when it is unique, up to equivalence, in which case we say that the sequence of characters *captures* T (in some literature the word “defines” is alternatively used). A necessary condition for this is that T is binary; otherwise, we could arbitrarily resolve any vertex of T of degree greater than 3, and obtain a different tree on which all the characters were homoplasy-free.

An interesting question now arises: What is the smallest number $h(n)$ so that for each tree T in $B(n)$, there is a sequence of $h(n)$ characters on $[n]$ that captures T ? If we restrict ourselves to binary characters, then $h(n) = n - 3$, since for any tree T in $B(n)$, the nontrivial characters that are homoplasy-free on T correspond (via $f \mapsto \Pi(f)$) precisely to the nontrivial splits of T . Moreover, T is the unique perfect phylogeny for a sequence of such characters provided that all $n - 3$ nontrivial splits of T are represented (if one was missing, we could contract the corresponding edge and still obtain a tree on which the characters were homoplasy-free).

But what if we do not insist on restricting ourselves to 2-state characters? For r -state characters for any fixed r , the required number of characters to capture T needs to grow linearly with n . More precisely, suppose that a sequence \mathcal{C} of k characters on X , each of which takes at most r states, captures a phylogenetic X -tree with n leaves; then

$$k \geq \left\lceil \frac{n-3}{r-1} \right\rceil. \quad (5.2)$$

This result (Proposition 4.2 of [314]) can be established by observing that each interior edge of the captured tree needs to be distinguished by at least one quartet tree $ab|cd$ for which $f_i(a) = f_i(b) \neq f_i(c) = f_i(d)$ for some character f_i from \mathcal{C} . Remarkably, the lower bound on k in (5.2) was recently shown in [53] to be sharp for every value of $r \geq 2$, provided that $n \geq n_r$, where n_r is some (increasing) function of r .

Proposition 5.4. *For each positive integer $r > 1$, there is a positive integer n_r such that for all binary phylogenetic X -trees T with $n = |X| \geq n_r$, there is a collection \mathcal{C} of r -state characters of size*

$$|\mathcal{C}| = \left\lceil \frac{n-3}{r-1} \right\rceil$$

that captures T .

For binary characters, $n_2 = 3$ (i.e., each binary tree can be captured by $n-3$ binary characters, as already noted). For 4-state characters, $n_4 = 13$ is the smallest value for which Proposition 5.4 holds.

The question now arises, if we do not fix r , is it possible that $h(n)$ might grow more slowly than linearly with n ? Perhaps \sqrt{n} or even $\log(n)$ characters might suffice? Surprisingly, it turns out that $h(n)$ is never more than 4.

Theorem 5.5 (four characters suffice). *For any binary phylogenetic X -tree T , there is a set \mathcal{C}_T of four characters at most for which T is the unique perfect phylogeny for \mathcal{C}_T .*

An example of the set \mathcal{C}_T from Theorem 5.5 is provided by the four hypothetical characters across a set of eight well-known species shown in the inset table of Fig. 5.3. It is easily seen that the tree T shown in Fig. 5.3 is a perfect phylogeny for this data set (this unrooted tree, incidentally, is the one biologists generally accept), and T is the only such perfect phylogeny for these four characters. Moreover, the states at the interior vertices (shown in brackets) are uniquely determined by the homoplasy-free condition (i.e., each character f_i has a unique valid extension F_i).

A recipe to generate the set \mathcal{C}_T for this example is indicated by the letters l, r, l', r' on the edges of the tree. These correspond to alternating “left” (l, l') and “right” (r, r') orientations as one moves up the tree under an arbitrary planar embedding. Now, suppose that any edge on which l is placed causes a state change for the first character, any edge on which r is placed causes a state change for the second character, and, similarly, any edge on which l' (respectively, r') is placed causes a state change for the third (respectively, fourth) character. State changes are always to a new state for that character (to ensure the homoplasy-free condition; a state present in one character is free to reappear in a different one). For example, the bottom-most l causes TRUE to change to CRUE.

More generally, for any binary phylogeny T , this procedure produces a sequence \mathcal{C}_T of (at most) four characters that are clearly homoplasy-free on T , since $\Pi(f)$ is convex on T for each character f in \mathcal{C}_T . The nontrivial step is to show that no other tree has this property (i.e., that \mathcal{C}_T captures T).

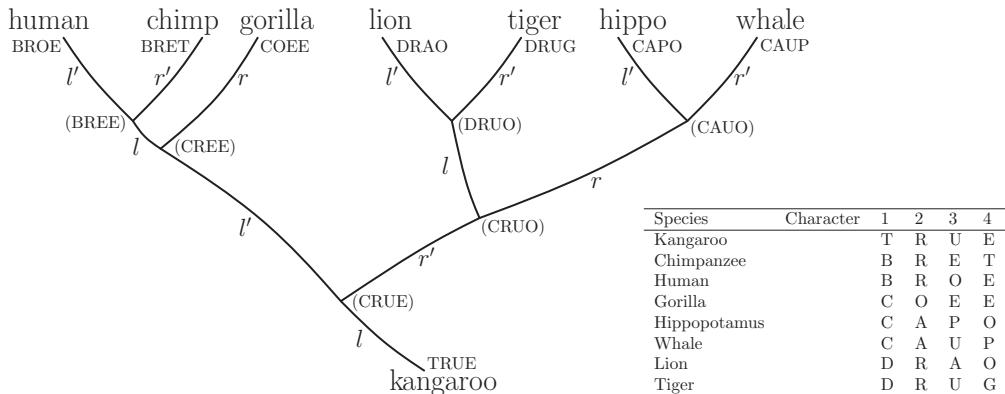


Figure 5.3. The unique perfect phylogeny T for the four characters described in the table shown. The assignment of the ancestral states (in brackets) is also determined by the unique valid extension of each character.

One way to establish this is via the quartet closure rule cl_2 that we described in Chapter 4. Recall that for a set \mathcal{Q} of quartet trees, $\text{cl}_2(\mathcal{Q})$ is the minimal set \mathcal{Q}' of quartet trees containing \mathcal{Q} that satisfies the following closure rule:

$$(\text{cl}_2) \text{ If } ab|cd, ac|de \in \mathcal{Q}', \text{ then } ab|de, bc|de, ab|ec \in \mathcal{Q}'.$$

In practice, $\text{cl}_2(\mathcal{Q})$ is obtained by starting with \mathcal{Q} and adding quartet trees according to this closure rule until no further quartet trees can be added. The point about the closure rule cl_2 is that any tree that displays the two quartet trees that satisfy the requirements of cl_2 must necessarily also display the three quartet trees that this rule implies. Consequently,

$$\langle \mathcal{Q} \rangle = \langle \text{cl}_2(\mathcal{Q}) \rangle. \quad (5.3)$$

It can be shown that when $\mathcal{C} = \mathcal{C}_T$, this closure of $q(\mathcal{C}_T)$ generates *all* the restricted quartet trees of T (i.e., $\text{cl}_2(q(\mathcal{C}_T)) = \mathcal{Q}(T)$). Since $\mathcal{Q}(T)$ defines T , eqn. (5.3) (with $\mathcal{Q} = q(\mathcal{C}_T)$) gives

$$\langle q(\mathcal{C}_T) \rangle = \langle \text{cl}_2(q(\mathcal{C}_T)) \rangle = \langle \mathcal{Q}(T) \rangle = \{T\},$$

and so \mathcal{C}_T captures T by Proposition 5.1. For further details, see [193].

The four-character result (Theorem 5.5) is part of a more general result that asks how many characters are required to “identify” a tree. We met this notion in the previous chapter, albeit in the context of restricted subtrees rather than characters. However, a similar idea applies here. Given a sequence \mathcal{C} of characters on X and a phylogenetic X -tree T , \mathcal{C} (*phylogenetically*) identifies T precisely if the set of phylogenetic X -trees on which the characters are all homoplasy-free coincides with the set of trees that can be obtained from T under refinement. Thus if T is binary, then this is equivalent to the characters capturing T . The following result is a consequence of a more general result (phrased in terms of X -trees) from [57].

Theorem 5.6. For any phylogenetic X -tree T , let d be the maximum degree of any vertex in T . If $k = 4\lceil \log_2(d-2) \rceil + 4$, there is a collection of k characters that identifies T .

Notice that this theorem generalizes Theorem 5.5 since if T is a binary phylogenetic tree, then $d = 3$, and so Theorem 5.6 assures us that $k = 4$ characters identify T , which implies that these characters also capture it, since T is binary.

5.1.2 ▪ Two enumeration questions

How many binary trees make a given character f homoplasy-free? This question has a simple explicit solution thanks to a result from Chapter 2.

Proposition 5.7. *Let f be a character on $[n]$. The number of trees in $B(n)$ for which f is homoplasy-free is*

$$\frac{b(n)}{b(n-k+2)} \prod_{i=1}^k r b(n_i),$$

where n_1, n_2, \dots, n_k are the sizes of the blocks of $\Pi(f)$.

Proof: We first establish this result in the special case where $n_i > 1$ for all i . There are $\prod_{i=1}^k b(n_i)$ choices of unrooted binary phylogenetic forests that can be constructed from the leaf sets that form the blocks of $\Pi(f)$. The trees $T \in B(n)$ for which f is homoplasy-free are precisely the trees that can be constructed from any one of these unrooted binary phylogenetic forests by sequentially adding edges (this follows from the convexity characterization of homoplasy-free). Equation (2.7) from Chapter 2 tells us that for each such forest, there are $\frac{b(n)}{b(n-k+2)} \prod_{i=1}^k (2n_i - 3)$ ways to construct a tree $T \in B(n)$ by joining the trees in these forests together with additional edges. Since different forests on the leaf partition $\Pi(f)$ give rise to distinct trees in $B(n)$ by any such edge addition process and since $(2n_i - 3)b(n_i) = r b(n_i)$, there are thus $\frac{b(n)}{b(n-k+2)} \prod_{i=1}^k r b(n_i)$ trees in $B(n)$ on which f is homoplasy-free. This establishes the result in the case where $n_i > 1$ for all i . Now suppose that $n_i > 1$ for $s \geq 0$ values of i . If $s = 0$, then $k = n$, and so

$$\frac{b(n)}{b(n-k+2)} \prod_{i=1}^k r b(n_i) = \frac{b(n)}{b(2)} \prod_{i=1}^k r b(1) = b(n),$$

so the result holds. On the other hand, if $s > 0$, then we may suppose, without loss of generality, that $n_1, \dots, n_s > 1$, and that if Y is the union of these blocks of $\Pi(f)$ and $n' = |Y|$, then $n' = n - (k - s)$. From the special case we established in the first half of the proof there are $\frac{b(n')}{b(n'-s+2)} \prod_{i=1}^s r b(n_i)$ trees on leaf set Y on which T is homoplasy-free. The trees in $B(n)$ on which f is homoplasy-free are precisely the trees $T \in B(n)$ for which f is homoplasy free on $T|Y$ and, by eqn. (4.1), there are $b(n)/b(n')$ such choices of T for each $T|Y$. Since $n' - s = n - k$ and $r b(1) = 1$, there is therefore a total of

$$\frac{b(n')}{b(n'-s+2)} \prod_{i=1}^s r b(n_i) \times \frac{b(n)}{b(n')} = \frac{b(n)}{b(n-k+2)} \prod_{i=1}^k r b(n_i)$$

trees from $B(n)$ on which f is homoplasy-free. ■

Notice that a 1-state character is homoplasy-free on all trees; so, too, is a character in which each element of X takes a different state. Thus the “most informative” characters (i.e., the ones that minimize the number of the trees on which the character is homoplasy-free) will lie between these two extremes. By using Proposition 5.7, it is fairly easy to identify such characters. For example, for $X = \{1, 2, \dots, 120\}$, a character that is homoplasy-free on the *fewest* trees in $B(120)$ is a 24-state character in which each state is shared by five elements from X . For any such character, the proportion of binary phylogenies on which it is homoplasy-free is less than 10^{-100} .¹⁹

¹⁹This is, incidentally, an argument for the merits of mathematics over simulations, since one would never find these trees by sampling $B(120)$ uniformly at random.

Counting characters for a given tree. A second enumeration question can be viewed as a dual version of the first: For a given binary tree, T in $B(X)$ with $|X| = n$, and set S , how many characters $f : X \rightarrow S$ that take r distinct states are homoplasy-free on T ?

The main step is to count the number of partitions of $[n]$ into r blocks that are convex on T . Again, there is a concise answer that depends only on n and r and not the shape of T . There are precisely

$$N(T, r) = \binom{2n - r - 1}{r - 1} \quad (5.4)$$

such partitions [327]. To see this, select a pair of leaves (say, x and x') that form a cherry of T . Notice that x and x' can only belong to different blocks of a partition that is convex on T if one or both of them forms singleton sets in the partition. We can therefore divide the set of partitions that are convex on T into three disjoint classes:

- C_1 : Partitions for which x and x' appear in the same block.
- C_2 : Partitions for which $\{x\}$ and $\{x'\}$ are both (singleton) blocks.
- C_3 : Partitions for which exactly one of $\{x\}$ and $\{x'\}$ is a (singleton) block.

Let $X' = X - \{x'\}$ and $X'' = X - \{x, x'\}$, and let $T' = T|X'$ and $T'' = T|X''$. Thus $|C_1| = N(T', r)$. Similarly, C_2 is in bijective correspondence with the partitions of X'' into $r - 2$ blocks that are convex on T'' . Thus $|C_2| = N(T'', r - 2)$. Finally, $|C_3| = 2(N(T', r - 1) - N(T'', r - 2))$ and thus

$$N(T, r) = N(T', r) + 2N(T', r - 1) - N(T'', r - 2). \quad (5.5)$$

Using this last equation, one can then verify two facts simultaneously by induction on n : first, that $N(T, r)$ depends solely on n and r (and not the shape of T), and also that $N(T, r)$ satisfies eqn. (5.4). The latter claim comes down to verifying the following binomial coefficient identity: $\binom{a}{b+1} + 2\binom{a+1}{b} - \binom{a}{b+1} = \binom{a+2}{b+1}$, which follows from two applications of the more familiar equality $\binom{c}{d} + \binom{c}{d-1} = \binom{c+1}{d}$.

By eqn. (5.4), the number of choices possible for a character $f : X \rightarrow S$ (where S is a fixed set of size s) to take r states and be homoplasy-free on T is given by the expression $\frac{s!}{(s-r)!} N(T, r) = \frac{s!(2n-r-1)!}{(s-r)!(r-1)!(2n-2r)!}$.

Equation (5.4) also leads to a nice connection with the Fibonacci sequence. Let F_k be the k th Fibonacci number (starting with $F_1 = F_2 = 1$). Then eqn. (5.4) implies that there are precisely

$$\sum_{r=1}^n \binom{2n - r - 1}{r - 1} = F_{2n-1} \quad (5.6)$$

partitions of $[n]$ that are convex on $T \in B(n)$. Again, this is independent of the shape of T .

This connection with Fibonacci numbers was made explicit by [212] where further interesting related results were discovered and applied. We describe one of these here. For $i \geq 1$, let $g_i(T, r)$ be the number of partitions of $[n]$ into r blocks, where each block is of size at least i , and where the partitions are convex on T , and let $g_i(T) = \sum_{r=1}^n g_i(T, r)$, which is the number of partitions of $[n]$ in which each block is of size at least i and which are convex on T . Thus $g_1(T, r)$ and $g_1(T)$ are given by eqns. (5.4) and (5.6), respectively. The following result is from [212].

Proposition 5.8. For any tree $T \in B(n)$, with $n \geq 2, r \geq 1$, $g_2(T, r) = \binom{n-r-1}{r-1}$ and so $g_2(T) = F_{n-1}$. In particular, for even values of n , $g_2(T)$ equals the g_1 value of any binary tree with half as many leaves.

Proposition 5.8, from [212], also showed that the curious invariance of $g_i(T)$ and $g_i(T, r)$ to the shape of T does not extend beyond $i = 2$; for $i = 3$, examples of two different trees from $B(9)$ with identical $g_3(T)$ values exist.

Exercise: How many partitions of $[n]$ are convex on the star tree $T \in P(n)$?

5.1.3 • Random binary characters

Suppose we generate a binary character \tilde{f} on X purely at random; that is, each element x in X is assigned the state 0 or 1 with equal probability $(\frac{1}{2})$ independently across the difference choices of x from X . We can then ask the following:

- Given a fixed character f , what is the probability that f and \tilde{f} have a perfect phylogeny?
- What is the probability that two independently generated random characters on X are compatible?
- Given a fixed phylogeny T on X , what is the probability that \tilde{f} is homoplasy-free on T ?

It turns out that the first two questions have a precise answer, and the third has a simple upper bound [311]. As might be expected, the first two probabilities converge to zero exponentially fast as $n = |X|$ grows.

- (a) If f is a binary character on $[n]$ for which $\Pi(f)$ has blocks of size r and $n - r$, and \tilde{f} is a random binary character on $[n]$, then the probability that $\{f, \tilde{f}\}$ have a perfect phylogeny is

$$\left(\frac{1}{2}\right)^{r-1} + \left(\frac{1}{2}\right)^{n-r-1} - \left(\frac{1}{2}\right)^{n-2},$$

whenever $1 \leq r \leq n - 1$ (the probability equals 1 if $r \in \{0, 1, n - 1, n\}$).

- (b) The probability p_n that two independent random binary characters on $[n]$ have a perfect phylogeny is given by

$$p_n = 4\left(\frac{3}{4}\right)^n - \left(\frac{1}{2}\right)^{n-1} \left(3 - \left(\frac{1}{2}\right)^{n-1}\right).$$

Thus p_n is bounded above by $4\left(\frac{3}{4}\right)^n$ and is asymptotically equivalent to it as n grows.

- (c) For any tree T in $P(n)$ with k interior edges, the probability that a random binary character on $[n]$ is homoplasy-free on T is bounded above by $k\left(\frac{1}{2}\right)^{n-3} + \left(\frac{1}{2}\right)^k$.

When n is sufficiently large, these results allow for the rapid and accurate identification of a tree T from $P(n)$ that has a moderate number of interior edges from a sequence \mathcal{C} of binary characters on $[n]$, each of which is either homoplasy-free on T or is random [311].

5.1.4 • Extensions of the binary perfect phylogeny problem

There are numerous variations on the perfect phylogeny problem. We outline four of these now.

PPP Let $\mathcal{C} = (f_1, \dots, f_k)$ be a sequence of binary characters on X , each with state space $\{0, 1\}$. A *persistent perfect phylogeny (PPP)* for \mathcal{C} is a phylogeny $T \in P(X \cup \rho)$ (here, ρ is an external “root” leaf and all edges of T are directed away from ρ) and, for each i , there is an extension F_i of f_i to all the vertices of T so that

- (i) $F_i(\rho) = 0$, and
- (ii) there is at most one edge $e = (u, v)$ for which $F_i(u) = 0$ and $F_i(v) = 1$ and at most one edge $e' = (u', v')$ for which $F_i(u') = 1$ and $F_i(v') = 0$.

Thus a persistent phylogeny for \mathcal{C} is a phylogeny that allows, for each character, the ancestral state 0 to evolve to the derived state 1 just once in the tree and the derived state 1 to revert to the ancestral state 0 at most once.

Notice that if both edges e and e' referred to in (ii) are present in T , then e must lie on the path between the root and e' . Of course, \mathcal{C} may not have a PPP, raising the questions of how to characterize this and whether or not there is a polynomial-time algorithm to determine the existence of a PPP (and find such a tree when it exists). The existence of a PPP of \mathcal{C} can be recast in terms of whether an associated graph has a certain property, and algorithms for PPP (along with some variations of this problem) have been pioneered by Paola Bonizzoni and colleagues [43, 44, 45]. At present, the computational complexity of determining whether a sequence of binary characters \mathcal{C} has a persistent perfect phylogeny is still unresolved.

There is a close connection between the PPP problem and chordal graphs. Recall that the intersection graph of a sequence of (ordinary) binary characters is chordal if and only if the characters form a perfect phylogeny. For PPP, there is a corresponding related result. If a PPP exists for $\mathcal{C} = (f_1, \dots, f_k)$, then the intersection graph of the sets $\{f_i^{-1}(s) : i = 1, \dots, k\}$ is chordal for $s = 0$ and for $s = 1$ [290]. PPP is also closely related to another phylogeny problem studied in [156]. Given a sequence \mathcal{C} of binary characters on X and integers $l_1, l_2 \geq 1$, \mathcal{C} has an (l_1, l_2) -phylogeny if there is an (unrooted) phylogeny $T \in P(X)$ and, for each i , there is an extension F_i of f_i to all the vertices of T so that the subgraph of T induced by $F_i^{-1}(1)$ has at most l_1 components and the subtree T induced by $F_i^{-1}(0)$ has at most l_2 components. Let $\mathcal{C}^{+\rho}$ be the sequence of characters on $X \cup \{\rho\}$ obtained by extending each character f_i in \mathcal{C} to ρ so that the extension maps ρ to 0. Then \mathcal{C} has a PPP precisely if $\mathcal{C}^{+\rho}$ has a $(1, 2)$ -phylogeny.

Partial binary characters

A *partial binary character* on X is a function $f : X \rightarrow \{0, 1, ?\}$, where “?” refers to a state denoting ambiguity or missing data. For example, if f refers to some binary feature of a gene that is present across certain species, then $f(x) = ?$ if the state at the leaf species x is uncertain or yet to be determined, or if species x does not have this gene at all.

The notions of convexity, homoplasy-free, and perfect phylogeny carry over directly to partial binary characters. For example a partial character f on X is convex on a phylogeny $T \in P(X)$ if the two minimal subtrees of T connecting the leaves in $f^{-1}(0)$ and the leaves in $f^{-1}(1)$ are vertex-disjoint. A perfect phylogeny for a sequence of partial binary characters is a phylogeny for which all the characters are convex in this sense.

Moving from ordinary binary characters to partial ones does have a major impact on some results, however. For example, although it is easy to determine whether a set of binary characters has a perfect phylogeny, this problem becomes NP-complete for partial binary characters.

To see why, we first define some notation. Given a split $A|B$ of a subset X' of X let $f_{(A|B)}$ be the character on X with state space $\{0, 1, ?\}$ defined by

$$f_{(A|B)}(x) = \begin{cases} 0 & \text{if } x \in A, \\ 1 & \text{if } x \in B, \\ ? & \text{if } x \in X - (A \cup B). \end{cases}$$

Now, recall from Section 4.2.2 that the problem of determining whether or not a set of quartet trees is compatible (i.e., whether or not $\langle \mathcal{Q} \rangle \neq \emptyset$) is NP-complete. Therefore, given any instance of \mathcal{Q} for this question, consider the associated set $\mathcal{C}_{\mathcal{Q}} = \{f_{(\{a, a'\}|\{b, b'\})} : aa'|bb' \in \mathcal{Q}\}$ of partial binary characters. It is an easy exercise now to show that \mathcal{Q} is compatible if and only if $\mathcal{C}_{\mathcal{Q}}$ has a perfect phylogeny.

Exercise: Show that a sequence \mathcal{C} of characters on X has a perfect phylogeny if and only if the associated collection of partial binary characters $\{f_{(A|B)} : A, B \in \Pi(f), f \in \mathcal{C}\}$ has a perfect phylogeny.

The concept of a partial binary character is central to the next two variations on perfect phylogeny.

PPH Suppose that $\mathcal{C} = (f_1, \dots, f_k)$ is a sequence of k partial binary characters on X . The *perfect phylogeny haplotyping (PPH) problem* asks where there is an associated pair of binary sequences $f'_i, f''_i : X \rightarrow \{0, 1\}$ so that $f'_i(x) = f''_i(x)$ for all $x \in X$ for which $f_i(x) \in \{0, 1\}$ and $f'_i(x) \neq f''_i(x)$ for all $x \in X$ for which $f_i(x) = ?$, and so that the resulting sequence of $2k$ characters $f'_i, f''_i : i = 1, \dots, k$ has a perfect phylogeny [19]. This problem, relevant in computational genomics, was formalized and studied by Dan Gusfield and colleagues. Over several years, increasingly more efficient algorithms for its solution, including linear-time ones, have been developed [112, 42].

IDPP Let $\mathcal{C} = (f_1, \dots, f_k)$ be a sequence of k partial binary characters on X . The *incomplete directed perfect phylogeny (IDPP) problem* asks whether there is a rooted phylogeny T on X and, for each $i \in \{1, \dots, k\}$, a function F_i from the vertices of T to $\{0, 1\}$ so that (i) if $f_i(x) \in \{0, 1\}$ for $x \in X$, then $F_i(x) = f_i(x)$; (ii) at most one edge (u, v) of T has $F_i(u) = 0$ and $F_i(v) = 1$ and (iii) no edge (u, v) of T has $F_i(u) = 1$ and $F_i(v) = 0$. In other words, the model requires that the known states (0 and 1) at the leaves can be described by evolution in which state 1 evolves at most once in the tree and is never lost (though the state at some descendant leaf may be unknown). An efficient solution for this problem was discovered by [289]. It can also be analyzed and solved by using the techniques from Section 4.2.1. Let $\mathcal{R}_{\mathcal{C}}$ be the set of rooted triples $ab|c$ for which there is a character f_i with $f_i(a) = f_i(b) = 1, f_i(c) = 0$. Then an instance \mathcal{C} of IDPP has an incomplete directed phylogeny and is supported by T

if and only if T displays $\mathcal{R}_{\mathcal{C}}$. In particular, a solution exists if and only if $\langle \mathcal{R}_{\mathcal{C}} \rangle \neq \emptyset$, which can be solved by applying the BUILD algorithm.

5.2 • Minimal evolution (maximum parsimony (MP))

The homoplasy-free condition is strong and will often be violated for certain types of real data. In that case, one approach is to simply seek a largest subset of characters that form a perfect phylogeny. This *maximum compatibility* approach, while conceptually simple and sometimes useful for certain types of data, often has limited power in extracting phylogenetic signal (e.g., under the sorts of processes we will consider in Chapter 7). It is also computationally difficult (NP-hard), though certain restricted versions of it have polynomial-time solutions. For example, suppose we have a sequence \mathcal{C} of binary characters on X and we wish to determine (i) whether there exists a *binary* phylogeny $T \in B(X)$ for which each split is present in $\Pi(f)$ for at least one character f in \mathcal{C} , and (ii) among the set of such trees finding one that maximizes the subset of characters on which the tree is homoplasy-free. Then this problem can be solved in polynomial-time by a dynamic programming approach [60].

An alternative approach to the “all-or-nothing” treatment of characters by maximum compatibility (in dealing with homoplasy) is to score each character by how well it fails to “fit” a given tree, with homoplasy-free characters having the smallest penalty. Given a character f and a phylogenetic tree T , the simplest way to score f is by the smallest possible number of edges of T which need to have differently assigned states at their endpoints in order to extend f to all the vertices of T . This score is called the *parsimony score* of the character on T , denoted, $\text{ps}(f, T)$. An extension F of f to all the vertices of T which minimizes the number of “ F -change edges” (i.e., edges $e = \{u, v\}$ for which $F(u) \neq F(v)$) is called a *minimal extension* of f on T . Thus, $\text{ps}(f, T) \geq |f(X)| - 1$, with equality if and only if f is homoplasy-free.

Since the number of extensions F of f to T grows exponentially with the number of interior vertices of T , it might be suspected that computing $\text{ps}(f, T)$ is hard. However, in 1971, the evolutionary biochemist Walter Fitch proposed a fast ($O(|X| \times |S|)$) algorithm, which was formally verified by mathematician John Hartigan in 1973. This *Fitch-Hartigan algorithm* has a “forward pass” step, which proceeds from the leaves to the root (this computes $\text{ps}(f, T)$) and a “backward pass” step that provides an explicit minimal extension function F .

When T is a rooted binary tree, then $\text{ps}(f, T)$ has a particularly simple description. Let $*$ denote the commutative and nonassociative binary *parsimony operation* on $2^S - \emptyset$ by

$$A * B = \begin{cases} A \cap B & \text{if } A \cap B \neq \emptyset, \\ A \cup B & \text{if } A \cap B = \emptyset. \end{cases}$$

Starting with the singleton set $\{f(x)\}$ at each leaf $x \in X$, and for each vertex v of T for which sets (say, A and B) have been assigned to the two immediate descendants of v , one assigns v the set $A * B$. For any binary tree, $\text{ps}(f, T)$ equals the number of vertices of T for which the empty intersection case arises when applying $*$ to assign a set of states at the vertex; moreover, the set assigned to the root of T is precisely the set of states which are possible at the root for at least one minimal extension F of f . This is illustrated in Fig. 5.4.

We will use this property of $\text{ps}(f, T)$ in Section 5.2.3 and Chapter 8.

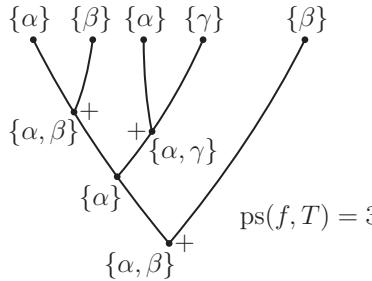


Figure 5.4. An application of the parsimony operation applied to a 3-state character f . Starting with the states at the leaves, represented as singleton sets, there are three vertices (marked by +) where the intersection of the sets of states at the two immediate descendants of the vertex is empty, and thus $\text{ps}(f, T) = 3$. There is at least one minimal extension of f on T for which the root is in state α and at least one for which the root is in state β but none with the root in state γ .

Before considering the complications of inferring a “maximum parsimony (MP) tree” from a sequence of characters, we first delve further into the combinatorial properties of this scoring procedure.

5.2.1 ■ The combinatorics of parsimony

For any character f on X , $\text{ps}(f, T)$ is the same as the minimum number of edges that need to be deleted from T in order to separate any two leaves that are in different states. Thus, for 2-state characters, it follows from Menger’s min-max theorem in graph theory that $\text{ps}(f, T)$ equals the maximum number of edge-disjoint paths that can be placed in T so that each path connects a pair of leaves of T that are assigned different states by f . This min-max connection turns out to be particularly useful in certain arguments. Here, we describe a simple application (another is provided in Chapter 7).

Proposition 5.9. If T is an unrooted binary phylogenetic X -tree with $n = 2m$ leaves, then the number of binary characters f on X with $\text{ps}(f, T) = m$ is 2^m .

Proof. For each binary tree T on $2m$ leaves, there is a unique set of m edge-disjoint paths that connects the $2m$ leaves (this can be proved by induction on m , by considering a cherry x, y of T and the associated restricted tree $T|(X - \{x, y\})$). By the min-max connection with Menger’s theorem, the characters of parsimony score m are precisely the characters which assign different states to the leaves at the ends of each of these m paths. Since there are two choices for each path we obtain 2^m choices altogether. ■

Proposition 5.9 is a special case of the more general result: for any tree T in $B(n)$ the number of characters $f : [n] \rightarrow \{0, 1\}$ for which $\text{ps}(f, T) = k$ is

$$\frac{2n-3k}{k} \binom{n-k-1}{k-1} 2^k \quad (5.7)$$

for all k between 1 and $n/2$ (and is zero otherwise) [328]. A remarkable feature of this expression is that it does not depend on the shape of T ; for nonbinary characters, this invariance to tree shape no longer applies.

A related question is the following. Suppose we assign either state 0 or 1 independently to each leaf of a binary phylogenetic tree T with n leaves, according to the toss

of a fair coin. Under this scenario, what is the expected parsimony score μ_n of this random binary character on T ? Various (correct) recursive formulae for this quantity have been published, one of which runs over several lines. However, it turns out that μ_n has a remarkably simple and explicit formula, namely

$$\mu_n = \frac{3n - 2 - (-\frac{1}{2})^{n-1}}{9} \sim \frac{n}{3}. \quad (5.8)$$

A short and direct proof of this expression follows. Select any cherry x, y of T . Either (i) x and y receive the same state under f or (ii) x and y receive different states. Notice that each of these two outcomes occurs with the same probability, namely $\frac{1}{2}$. In case (i), the parsimony score $s(T)$ of T equals the parsimony score $s(T')$ of the binary phylogeny $T' = T|(X - \{x\})$ with the inherited leaf states. In case (ii), $s(T) = s(T'') + 1$, where $T'' = T|(X - \{x, y\})$, with the inherited leaf states. In this way, we can write

$$s(T) = s(T') \cdot \mathbb{I}_{f(x)=f(y)} + (s(T'') + 1) \cdot \mathbb{I}_{f(x) \neq f(y)}. \quad (5.9)$$

Taking the expectation of both sides (and noting that $s(T')$, as well as $s(T'')$, is independent of $\mathbb{I}_{f(x)=f(y)}$ and $\mathbb{I}_{f(x) \neq f(y)}$), we see by induction that μ_n depends only on n and not the shape of T , and that for $n \geq 3$,

$$\mu_n = \frac{1}{2}\mu_{n-1} + \frac{1}{2}(\mu_{n-2} + 1).$$

Solving this simple linear recursion subject to the initial conditions $\mu_1 = 0, \mu_2 = \frac{1}{2}$ gives eqn. (5.8). With further work, eqn. (5.9) can be used to derive eqn. (5.7).

The parsimony score of a character on a tree also has an alternative “subdivision” interpretation. Consider the partition $\Pi(f)$ of X that f induces. Then f is homoplasy-free on T if and only if $\Pi(f)$ is convex on T . Moreover, for any character f , there is clearly some refinement Π of $\Pi(f)$ for which Π is convex on T (e.g., we could take Π to be simply the trivial partition in which each block is a singleton). The following result, from [59], shows that the parsimony score can be viewed as the number of blocks in a minimal refinement of $\Pi(f)$ that is convex on T , minus 1. That is,

$$\text{ps}(f, T) = \min\{\#\Pi : \Pi \text{ is convex on } T, \text{ and } \Pi \text{ refines } \Pi(f)\} - 1.$$

Let us now define $h(f, T) = \text{ps}(f, T) - (|f(X)| - 1)$, the *homoplasy score* of f on T . To the biologist, $h(f, T)$ has a natural interpretation: regardless of where we may wish to root the tree T , $h(f, T)$ is the minimal number of events of reverse evolution and/or convergent evolution required in total to explain the evolution of f on the rooted version of T (for details, see [315]). Some elegant connections exist between this homoplasy score and the tree rearrangement metrics described in Section 2.5, most of which were pioneered by David Bryant. First observe that if T is a binary phylogeny and T' is obtained from T by a single rearrangement operation θ (e.g., $\theta = \text{TBR}, \text{SPR}$ or NNI), then

$$|h(f, T) - h(f, T')| \leq 1. \quad (5.10)$$

Exercise: Prove inequality (5.10).

From inequality (5.10), the triangle inequality shows that if f is homoplasy-free on some tree T' (i.e., $b(f, T') = 0$), then $b(f, T) \leq d_\theta(T, T')$. Since this holds for all such trees T' ,

$$b(f, T) \leq \min\{d_\theta(T, T') : b(f, T') = 0\}.$$

Remarkably, this last inequality is an equality for SPR and TBR (but not for NNI). That is, we have the following result from [59].

Theorem 5.10. *For any character $f : X \rightarrow S$ and any tree $T \in B(X)$, we have*

$$b(f, T) = \min\{d_\theta(T, T') : b(f, T') = 0\},$$

when $\theta = \text{SPR}$ or $\theta = \text{TBR}$.

We turn now to a second connection between the homoplasy score and tree rearrangement operations, from [62]. For $\theta = \text{TBR}$, SPR , or NNI , an unrooted phylogenetic X -tree, and an integer $r \geq 0$, let $\Sigma_\theta(T, r)$ be the set of X -splits that are present in at least one phylogeny that is r or fewer θ -rearrangement operations from T . Formally,

$$\Sigma_\theta(T, r) := \bigcup_{T' \in B(X) : d_\theta(T, T') \leq r} \Sigma(T').$$

It is clear that $\Sigma_{\text{SPR}}(T, r) \subseteq \Sigma_{\text{TBR}}(T, r)$ since every SPR operation is also a TBR operation. The following result shows that, somewhat surprisingly, these two sets are equivalent; moreover the splits in this shared system are precisely those that correspond to the binary characters that have a homoplasy score of at most r . More precisely, we have the following theorem (from [62]).

Theorem 5.11. *The following are equivalent:*

- (i) $A|B \in \Sigma_{\text{SPR}}(T, r)$;
- (ii) $A|B \in \Sigma_{\text{TBR}}(T, r)$;
- (iii) $b(f_{A|B}, T) \leq r$,

where $f_{A|B}$ is any binary character with $\Pi(f_{A|B}) = \{A, B\}$.

By invoking eqn. (5.7), these equivalences allow us to count the number of X -splits in the θ -neighborhood of T :

$$|\Sigma_{\text{SPR}}(T, r)| = |\Sigma_{\text{TBR}}(T, r)| = \sum_{k=1}^{r+1} \frac{(2n-3k)}{k} \binom{n-k-1}{k-1} 2^k.$$

A characterization of $\Sigma_{\text{NNI}}(T, r)$ was also provided in [62].

The parsimony score is also the basis for a recently proposed and studied metric d_{MP} on phylogenies defined by setting

$$d_{MP}(T, T') = \max_f |\text{ps}(f, T) - \text{ps}(f, T')| = \max_f |b(f, T) - b(f, T')|$$

for all $T, T' \in P(n)$, with f ranging over all characters on X . d_{MP} is a metric on $P(X)$ and has a number of attractive properties; for example, the maximization need only range over the characters f for which either $b(f, T) = 0$ or $b(f, T') = 0$. However, the number of states for f cannot be bounded in the maximization search (for all n), and the computation of d_{MP} has been shown to be NP-hard in general. For further details, see [140].

5.2.2 • Ancestral state reconstruction

Given a character $f : X \rightarrow S$ and a rooted phylogenetic tree T in $RP(X)$, biologists are often interested in estimating an ancestral state at one or more interior vertices of T , such as the root vertex, or some particular interior vertex of interest, or even for all interior vertices simultaneously. We will first consider the estimation of the state at the root.

The simplest method is to ignore the tree T completely and just estimate the root state as being the state in S that occurs most frequently (with ties broken arbitrarily), a method known as *majority rule*. However, it is usual to try to use the tree structure in the estimation. A simple approach is to select a state for the root that occurs in a minimal extension of f to the vertices of T in the MP approach. In other words, the root is assigned a state that allows for the evolution of the character f on T with as little homoplasy as possible.²⁰

The MP approach can lead to ancestral state estimates that are quite different from those that majority rule predicts. This is easy to see if T is a rooted caterpillar tree. In this case, if the two leaves closest to the root are in the same state, then the root will be assigned that state under MP, regardless of the states at all the other leaves of T . Of course, in this case, there is a great degree of asymmetry in the tree—some leaves are very close to the root and others are far away—so it is instructive to consider a perfect rooted phylogeny T of height b (i.e., $T_b \in RB(2^b)$) for which each leaf is the same number of edges distant from the root. Even when there are just two states (say, $S = \{0, 1\}$), it turns out that the parsimony approach can estimate state 0 for the root even when the proportion of leaves in state 0 is arbitrarily small (for large enough values of b).

To see this, let v_b denote the minimal number of leaves that can be in state 0 for which the unique most parsimonious state for the root is 0. It can then be shown that

$$v_b = v_{b-1} + v_{b-2},$$

with $v_1 = 2, v_2 = 3$, which implies that v_b is (a translation of) the Fibonacci sequence. Thus v_b grows at the rate of the “golden ratio” $((1 + \sqrt{5})/2 \approx 1.618\dots)$ raised to the power of b , so we have $v_b/2^b \rightarrow 0$ as $b \rightarrow \infty$, as claimed.

MP can also be used to estimate the states at all interior vertices of the tree. One issue is that the character f on X may have several minimal extensions; indeed, the number of these can grow exponentially with the number of leaves of T , even for binary characters. For example, consider a caterpillar tree $T \in B(2n)$ with the leaves ordered sequentially $1, 2, 3, \dots, 2n$ in a planar drawing (with 1, 2 and $2n-1, 2n$ forming cherries). For the binary character $f_{2n} : [2n] \rightarrow \{\alpha, \beta\}$ be given by $f(i) = \alpha$ if $i = 0, 1 \pmod{4}$ and $f(i) = \beta$ if $i = 2, 3 \pmod{4}$. Once again, the Fibonacci sequence appears, since the number of minimal extensions of f_{2n} can be shown to be precisely the Fibonacci number F_{n+1} (with $F_1 = 1, F_2 = 1$).

Well-spaced state changes can be reconstructed. There is one special condition where ancestral reconstruction using MP is unique and, in a certain sense, provably accurate. Consider a character f that has evolved on a tree $T \in RP(X)$, and let G denote the (unknown) function that describes the states at every vertex of the tree (i.e., an extension of f to $V(T)$). Let us suppose that any two edges on which state changes occur under G (i.e., edges whose endpoints receive different states under G) are separated by at least three edges of T . For example, G might describe the evolution in the tree of some characteristic

²⁰More sophisticated approaches for ancestral state reconstruction are based on stochastic models of character evolution, described in Chapter 8, and the use of maximum likelihood or Bayesian estimation methods.

that changes only very rarely and in distant parts of the tree. Although G is not known, the function $f = G|X$ is, since it describes the resulting character that we observe at the leaves of T . It can be shown that there is a unique minimal extension F of f on $T^{-\rho}$; moreover $F(v) = G(v)$ for every vertex v of $T^{-\rho}$ [339]. In other words, if we know that a character has evolved in such a way on a tree (with “well-separated” changes of state) then we can exactly reconstruct the states at all the interior vertices of this tree, except perhaps for the state at the root if this has degree 2.

5.2.3 • Counting minimal evolution trees

Let $n_k(a, b)$ denote the number of trees T in $B(n)$ with $\text{ps}(f_{a,b}, T) = k$, where $f_{a,b}$ is a binary character on $[n]$ that assigns state 0 to a leaves, and state 1 to b leaves. For example, $n_1(2, 2) = 1$ and $n_2(2, 2) = 2$. Remarkably, there is an exact expression for $n_k(a, b)$. To describe it, we first need to count forests of rooted binary trees.

Let $N(m, k)$ denote the number of forests consisting of k rooted binary phylogenetic trees on leaf sets that partition $[n]$. For example, $N(m, 1)$ is just $rb(m)$, the number of rooted binary phylogenetic trees on m leaves, while $N(m, 2)$ also equals $rb(m)$ when $m > 1$, since the deletion of the root of a tree in $RB(m)$ induces a bijection from $RB(m)$ to the set of forests of size 2 on $[m]$. At the other end of the spectrum, $N(m, m) = 1$. There is a concise formula for $N(m, k)$, namely

$$N(m, k) = \frac{(2m - k - 1)!}{(m - k)!(k - 1)!2^{m-k}}, \quad (5.11)$$

for $k = 1, \dots, m$ ($N(m, k) = 0$ otherwise). The proof consists of a simple application of the Lagrange inversion formula to find the coefficient of x^m in $(1 - \sqrt{1 - 2x})^k$ (the k th power of the exponential generating function of $rb(n)$), and noting that this coefficient is simply $k!N(m, k)/m!$ (for details, see [315]).

Returning to $n_k(a, b)$, notice that $\sum_k n_k(a, b) = b(n)$, and that $n_k(a, b) = 0$ if k exceeds $\min\{a, b\}$. Also, if either a or b equals zero, then $f_k(a, b)$ is zero when $k > 0$ and is equal to $b(n)$ when $k = 0$. Thus we may assume that $a, b, k \geq 1$, in which case the formula for $n_k(a, b)$ is given as follows.

Theorem 5.12. *For $n = a + b$, where $a, b, k \geq 1$, we have*

$$n_k(a, b) = \frac{(2n - 3k)}{k} \times \frac{b(n)}{b(n - k + 2)} \times k!N(a, k)N(b, k).$$

In view of eqn. (5.11) and the expression for $b(m)$, it follows that $n_k(a, b)$ can be written as a ratio of terms, each of which is either a factorial or a power of 2, along with the term $(2n - 3k)$ in the numerator.²¹

Notice that $k!N(a, k)N(b, k)$ is the number of forests of k unrooted trees that can be formed by matching up the trees in a forest of k rooted binary trees on the a leaves assigned one state by f with the trees in a forest of b leaves being assigned the other state of f . Moreover, if we join these k unrooted trees to form a tree $T \in B(n)$ by adding additional edges, then the parsimony score of f of this tree must be at least k by Menger’s theorem. By refining these ideas further (and using eqn. (2.7) from Chapter 2), it is possible to stitch together a constructive proof of Theorem 5.12 (for details, see [315]). The original proof

²¹Legend has it that this remarkable closed form expression for $n_k(a, b)$ was conjectured through trial and error by the biologist David Penny during some of the less riveting televised segments of the 1984 Olympic Games.

of Theorem 5.12 from [75] was very different, however, and relied totally on generating function techniques and the judicious application of a piece of machinery from algebraic combinatorics. We outline a simplified version of this argument now.

Although the quantity $n_k(a, b)$ counts unrooted phylogenies, we need to work with rooted binary trees for the recursions to work. In that way we can relate a rooted tree (and its parsimony score) to the two maximal subtrees incident with the root (and their parsimony scores), and so the generating functions that follow can be viewed as refinements of the equation $\varphi(x) = \frac{1}{2}\varphi(x)^2 + x$ (eqn. (2.4)) for counting $RB(n)$.

We also need to take into account the set of states that the root of a tree can take across all minimal extensions of a character, which was described by the “parsimony operation” * discussed in Section 5.2. This operation also determines the parsimony score of T for a character: it is either the sum of the parsimony scores of the two subtrees incident with the root, or this sum plus 1 in the case where $A \cap B = \emptyset$.

We will define a quadratic system of three functions ψ_0, ψ_1, ψ_2 , each of which is a formal power series in three variables x_0, x_1, x_2 . Here x_0 counts the number of leaves in state 0, x_1 the number of leaves in state 1, and x_2 the parsimony score of the character on the tree. For $i \in \{0, 1\}$ let $\tilde{n}_k(a, b, i)$ be the number of rooted trees $T \in RB(n)$ for which the parsimony score of $f_{a,b}$ is k and where i is the only state that the root of T can take across all minimal extensions of $f_{a,b}$ to T . For $i = 2$ let $\tilde{n}_k(a, b, i)$ be defined in the same way, except the count is for the setting where both of the two states for root (0 or 1) occur in at least one minimal extension of $f_{a,b}$. Let

$$\psi_i := \sum_{a,b \geq 1, k \geq 0} \tilde{n}_k(a, b, i) \frac{x_0^a x_1^b}{a! b!} x_2^k.$$

Lemma 5.13. *The generating functions ψ_0, ψ_1 , and ψ_2 satisfy the simultaneous quadratic system*

$$\psi_0 = \frac{1}{2}\psi_0^2 + \psi_0\psi_2 + x_0, \quad \psi_1 = \frac{1}{2}\psi_1^2 + \psi_1\psi_2 + x_1, \quad \text{and} \quad \psi_2 = \frac{1}{2}\psi_2^2 + x_2\psi_0\psi_1.$$

Proof: The equation for ψ_0 holds since a rooted binary tree for which $i = 0$ is either (i) a single leaf that is assigned state 0 (the term x_0); or (ii) a tree that has two subtrees incident with the root, each with $i = 0$ (the term $\frac{1}{2}\psi_0^2$); or (iii) a tree for which one subtree incident with the root has $i = 0$ and the other has $i = 2$ (the term $\psi_0\psi_2$); moreover, the sum of the parsimony scores of these two maximal subtrees equals the parsimony score for the tree. A similar argument holds for ψ_1 . For ψ_2 the appearance of x_2 in this last equation is because x_2 counts the parsimony score of the character on the tree and this increases by 1 above the sum of the scores of the two subtrees of T that are incident with the root precisely when one subtree has $i = 0$ and the other has $i = 1$. ■

The next step turns out to be particularly fortuitous: for each choice of i from $\{0, 1, 2\}$ Lemma 5.13 reveals that we can write

$$\psi_i = x_i G_i(\psi_0, \psi_1, \psi_2), \text{ where} \tag{5.12}$$

$$G_0 = \left(1 - \frac{1}{2}\psi_0 - \psi_2\right)^{-1}, \quad G_1 = \left(1 - \frac{1}{2}\psi_1 - \psi_2\right)^{-1},$$

$$\text{and } G_2 = \psi_0\psi_1 \left(1 - \frac{1}{2}\psi_2\right)^{-1}.$$

Notice also that $\tilde{n}_k(a, b, i) = a!b![x_0^a x_1^b x_2^k] \psi_i(x_0, x_1, x_2)$, where $[*]\psi$ asks for the coefficient of the monomial $*$ in ψ . Thus, since a tree $T \in B(n)$ has $(2n-3)$ edges on which a root can be placed, we have

$$(2n-3)n_k(a, b) = a!b![x_0^a x_1^b x_2^k](\psi_0 + \psi_1 + \psi_2). \quad (5.13)$$

Consequently, if we can extract the coefficient of $x_0^a x_1^b x_2^k$ in each of the ψ_u functions, then we can readily obtain $n_k(a, b)$. Fortunately, there is an explicit way to find such coefficients when the functions satisfy identities of the type shown in (5.12) by applying the multivariate Lagrange inversion formula for monomials (Corollary 1.2.13 of [158]). Although the calculations are slightly tedious, as they involve computing various 3×3 determinants, they can still be done by hand (the original proof used a more general form of the multivariate Lagrange formula and required computer algebra). In this way, eqn. (5.13) leads to the expression in Theorem 5.12.

By using Theorem 5.12, it is possible to derive the asymptotic distribution of binary phylogenies according to their parsimony score. If a tree is selected according to the uniform model on $B(n)$, and Z is the parsimony score of a character, in which a_n leaves are in state 0 and b_n leaves in state 1, then if $a_n/n \rightarrow \alpha$ and $b_n/n \rightarrow \beta$, then Z is asymptotically normally distributed with mean μn and standard deviation $s\sqrt{n}$, where

$$\mu = \frac{2}{3} \left(1 - \sqrt{1 - 3\alpha\beta} \right) \text{ and } s = \frac{\mu\sqrt{1-\mu}}{2-3\mu}.$$

For a more precise statement and a proof of these results, see [262]. A variation of this result is where each leaf $x \in X$ is assigned state 0 with a fixed probability p and state 1 with probability $1-p$ (leading to a normal distribution with the same asymptotic mean but a variance that is asymptotically larger when $p \neq \frac{1}{2}$) [254]. A further variant arises for the case where T_n is any given tree in $B(n)$ and each leaf x of T_n is independently assigned a state $\alpha \in S$ with probability $p_x(\alpha) \geq \epsilon > 0$; again, a central limit theorem applies as n grows [334].

5.3 ■ Minimal evolution trees for a sequence of characters

Given a sequence of characters on X , a *maximum parsimony tree* for these data is a phylogenetic tree T that minimizes the sum of the parsimony scores of the characters. As a technique for reconstructing phylogenies from character data, MP was once popular, though for genetic and genomic data it has largely been eclipsed by techniques we will describe in the next two chapters. MP is still used, however, for analyzing particular types of character data, such as morphological and fossil data.²²

Finding a maximum parsimony tree (MP tree) is, in general, an NP-hard problem. Most algorithms are therefore heuristic, and typically search the space of trees by various “hill-climbing” techniques using the rearrangement operations (NNI, SPR, TBR) that we discussed in Chapter 2.

In certain cases, it turns out to be easy to find an MP tree on certain inputs. For example, when $\mathcal{C} = (f_1, \dots, f_k)$ is a sequence of binary characters that correspond to the splits of a phylogenetic X -tree T (i.e., $\{\Pi(f_i) : i \in [k]\} = \Sigma(T)$), then T is the unique minimally resolved MP tree for \mathcal{C} . It is therefore of interest to consider a relaxation of this where each character corresponds to a split in T or a split in a different tree T' . In

²²A small but vocal group of biologists, associated with the Willi Hennig society, advocates maximum parsimony as the phylogenetic method of choice, though on philosophical rather than scientific grounds.

general, finding an MP tree for such data is NP-hard [163]; however, there is a special case where it is easy: if T and T' are binary phylogenies on X , and T' is related to T by a single TBR operation. In that case, an MP tree can be computed in polynomial time. This is relevant to settings in biology where there has been a hybridization event, whereby the evolution of each character may well be described by one of two trees, which are related to each other by a single rearrangement operation.

Another special case that is easily solvable is when there are just two characters. For $\mathcal{C} = (f_1, f_2)$, let $b(\mathcal{C}) = \min_{T \in P(X)} \{h(f_1, T) + h(f_2, T)\}$, which is the sum of the homoplasy scores of these characters on an MP tree. The following result from [84] (and, independently, [59]) says that the homoplasy score of the characters is equal to the cyclomatic number of the partition intersection graph of $\{f_1, f_2\}$.

Theorem 5.14. *When \mathcal{C} consists of just two characters, $b(\mathcal{C}) = \text{cy}(\text{int}(\mathcal{C}))$.*

A further result from [59] provides a fast and simple way to construct a maximum parsimony tree for any pair f_1, f_2 of characters on X . Let G be the complete graph on vertex set X with the edge between x and x' being weighted by 0, 1, or 2 depending on the number of i values for which $f_i(x) \neq f_i(x')$. Then any minimal length spanning tree²³ of G corresponds to an MP tree for $\{f_1, f_2\}$.

It is sometimes also easy to identify a split that must be in at least one (or all) maximum parsimony trees, based on the following result of Bryant [61].

Proposition 5.15. *Suppose \mathcal{C} is a sequence of binary characters and that $A|B$ is an X -split that is compatible with every X -split induced by the characters in \mathcal{C} . Then an MP tree for \mathcal{C} exists which contains the split $A|B$. If, in addition, $A|B$ is the partition induced by at least one character in \mathcal{C} , then every MP tree for \mathcal{C} contains the split $A|B$.*

There is a direct constructive way to prove this last result; however, it is also a consequence of a deeper theorem which states that MP trees are embedded in the Buneman graph that we met in Chapter 2.

An interesting stochastic question is, what is the distribution of MP trees for a sequence of random characters? For example, suppose that \mathcal{C} consists of a sequence of k independent random binary characters, with each element x from X being assigned state 0 or 1 with equal probability (as considered in Section 5.1.3). Then the expected parsimony score of \mathcal{C} on any binary tree X -tree T is just $k\mu_n$, where μ_n is from eqn. (5.8). Moreover, the probability distribution on this parsimony score of \mathcal{C} is the same for all trees $T \in B(X)$ (this follows from the dependence of eqn. (5.7) on n and k only). It might be suspected, therefore, that each tree $T \in B(n)$ has equal probability of being an MP tree for such a \mathcal{C} . This is almost but not quite correct, even for $k = 2$ and $n = 6$, where a caterpillar tree has a slightly higher probability of being an MP tree for \mathcal{C} than is the case for a perfect tree. However, for any two trees T and T' in $B(n)$, each tree has, in the limit as k grows, the same probability ($\frac{1}{2}$) of being more parsimonious than the other [139].

5.3.1 ■ Short encodings and supertrees

We saw in Section 5.1.1 that no sequence of fewer than $n - 3$ binary characters can have a unique perfect phylogeny. But can we get away with fewer characters if we just want a unique maximum parsimony tree? That is, for each tree $T \in B(n)$, is there a sequence

²³Every connected graph $G = (V, E)$ has a *spanning tree*, which is a subgraph $G' = (V, E')$ where $E' \subseteq E$, for which G' is a tree; its “length” is the sum of assigned weights of the edges in E' .

$\eta(T)$ of 2-state characters of length $k = k(n)$ that grows sublinearly with n and for which T is the unique most parsimonious tree? A simple counting argument sets an absolute lower bound on k . Let $S(n, k)$ be the set of sequences of 2-state characters on $[n]$ of length k . Then k must be at least large enough for the function $T \mapsto \eta(T)$ from $B(n)$ to $S(n, k)$ to be one-to-one. Since $S(n, k) = 2^{nk}$, this requires that $|B(n)| \leq 2^{nk}$, which can be rewritten as $k \geq \frac{1}{n} \log_2 |B(n)|$. If we now invoke eqn. (2.2) and Stirling's approximation for $n!$ to calculate $|B(n)|$, we see that k must grow at least at the rate $\log(n)$. Remarkably, it was recently shown [80] that this primitive logarithmic growth rate can be achieved—and by a function η that can be constructively implemented. Moreover, the homoplasy score per character of the resulting sequences $\eta(T)$ on T necessarily tends to infinity as n grows, so this encoding is very “far” from supporting a perfect phylogeny. For example, for a perfect tree T on $2^{15} = 32768$ leaves, $\eta(T)$ consists of just 57 characters which have an average homoplasy score on the unique maximum parsimony tree of more than 1000 per character.

As well as phylogeny reconstruction from characters, MP has another use: it is the basis of a popular “supertree method” (a method for reconstructing a phylogeny on X from phylogenies on subsets of X). This method is called *matrix representation with parsimony* (MRP) and it takes any sequence $\mathcal{P} = (T_1, \dots, T_k)$ of phylogenies, possibly on different leaf sets X_1, \dots, X_k , and returns a phylogeny on $X = \bigcup_{i=1}^k X_i$. We first describe the simplest case where $X_i = X$ for all i . For each $j \in [k]$ and each $\sigma \in \Sigma(T_j)$, let $f_{(\sigma, j)}$ be a binary character on X with $\Pi(f_{(\sigma, j)}) = \sigma$. Then MRP first constructs the set of MP trees for the sequence of characters $\mathcal{C}_{\mathcal{P}} = (f_{(\sigma, j)} : \sigma \in \Sigma(T_j), j \in [k])$. If we now take the strict consensus of this set of MP trees, we obtain the *strict consensus MRP supertree* for \mathcal{P} . In this special setting where the sets X_i are all equal, this provides yet another consensus function on phylogenies, as studied in the last section of Chapter 2. Moreover, Proposition 5.15 provides a link with another consensus function, strict consensus: any split in the strict consensus tree is necessarily in the strict consensus MRP supertree.

The extension to the supertree setting where the sets X_i are not equal is achieved by moving from binary characters to partial binary characters. In that case, given a sequence of phylogenies on overlapping leaf sets, each tree is replaced by the sequence of partial characters $f_{(A|B)}$ corresponding to each split $A|B$ of that tree. These sequences of partial binary characters are then concatenated to give a data set for which a maximum parsimony tree is sought (the parsimony score of a partial binary character f on T is the parsimony score of the binary character $f|X_i$ on $T|X_i$, though it can also be computed directly on T quite easily).

Chapter 6

Continuous phylogenies and distance-based tree reconstruction

6.1 • Metrics from trees with edge lengths

So far, we have regarded the edges of phylogenies as having no particular length; however, it is useful—both in biology and mathematics—to assign a positive length to each edge (often called a “branch length” in biology). In biology, the length of an edge could correspond to evolutionary time, or to some measure of the amount of genetic change along that edge. Assigning lengths to edges introduces further mathematical structure to help study and reconstruct trees. This is pivotal to approaches for inferring trees from data that try to estimate an “evolutionary distance” between pairs of species, as well as for the statistical methods that we will discuss in Chapters 7 and 8.

Suppose that we have a phylogenetic X -tree T (rooted or unrooted) and some function l that assigns a real-valued length to each edge of the tree. Define $d = d_{(T,l)} : X \times X \rightarrow \mathbb{R}$ by letting $d(x,y)$ be the sum of the lengths of the edges on the path in T connecting x and y for each pair x,y in X . Notice that $d(x,x) = 0$ for all $x \in X$ and d is symmetric (i.e., $d(x,y) = d(y,x)$ for all x,y). Unless stated otherwise, we will assume that $l(e) > 0$ for each edge e (we will denote this by writing $l > 0$), in which case $d_{(T,l)}$ is a metric on X ; in particular, d satisfies the triangle inequality and the property $d(x,y) = 0 \Leftrightarrow x = y$. When d can be represented by a tree in this way (and with $l > 0$) we say that it has a *tree representation* (on T) or, more briefly, that it is a *tree metric*.²⁴ This leads to two natural questions:

- Does every metric on X have a tree representation?
- Is the choice of T and l in a tree representation unique?

The answers to these questions are “no” and (for unrooted trees) “yes,” respectively.

Let us consider the first question. When $|X| = 3$, it is an easy exercise to show that every metric d on X can be represented as a tree metric. However, this result is particular to $|X| = 3$ and already runs into problems when $|X| = 4$. It is instructive to see why. Consider the three pairwise sums

$$d(x,y) + d(w,z), \quad d(x,z) + d(y,w), \quad \text{and} \quad d(x,w) + d(y,z).$$

²⁴In certain literature, the phrase “is additive (on T)” is also used.

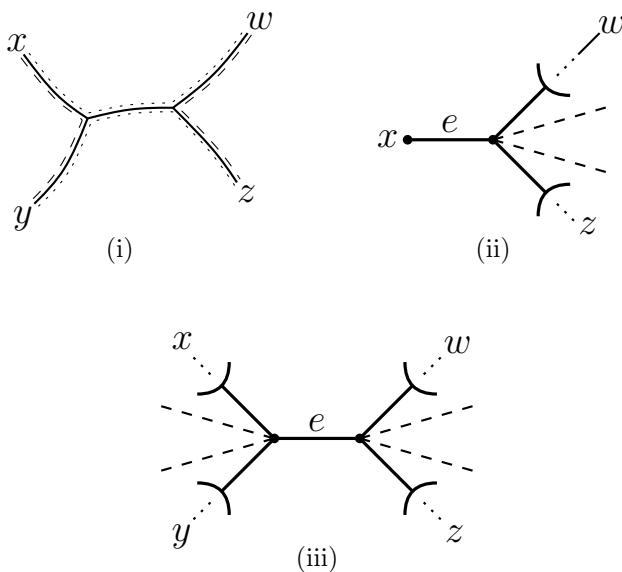


Figure 6.1. (i) The sum $d(x, y) + d(w, z)$ (dashed paths) is smaller than $d(x, w) + d(y, z)$ (dotted paths), which, in turn, equals $d(x, z) + d(y, w)$. (ii) In a general phylogeny, the edge length $l(e)$ can be easily written as a linear combination of certain d values: If e is pendant (top), $l(e) = \frac{1}{2}(d(x, w) + d(x, z) - d(w, z))$; if e is interior (bottom), $l(e) = \frac{1}{2}(d(x, w) + d(y, z) - d(x, y) - d(w, z))$. Dashed lines in (ii) and (iii) indicate the possibility of other edges joining the ends of e .

If d has a tree representation (i.e., $d = d_{(T,l)}$), then two of these pairwise sums must be equal, and at least as large as the third, regardless of the choice of T . This is illustrated in Fig. 6.1(i). This condition is not usually satisfied by an arbitrary metric d on a set of size 4, but when it is, it turns out that d can be represented on a tree. What is much more remarkable is that, for a set X of any size, the four-point condition holds for all subsets of X of size 4 if *and only if* d has a tree representation. This result, in various forms, dates back to the mid-1960s and early 1970s [72, 383]. More formally, d satisfies the *four-point condition* if, for any four elements x, y, z , and w from X (not necessarily distinct), we have $d(x,y) + d(w,z) \leq \max\{d(x,z) + d(y,w), d(x,w) + d(y,z)\}$.

Theorem 6.1. *A metric d on X has a tree representation if and only if d satisfies the four-point condition.*

In fact, Theorem 6.1 holds if we weaken the four-point condition by restricting it to any fixed value of x , and for all values y, z, w that are different from x and from each other.

Consider now the second question: the uniqueness of a tree representation. As before, this question was resolved many decades ago, and the uniqueness of both the unrooted tree and the strictly positive edge lengths holds; in other words, for any two trees T and T' in $P(X)$, with strictly positive edge lengths l and l' , respectively, we have

$$d_{(T,l)} = d_{(T',l')} \implies T \cong T' \text{ and } l = l'. \quad (6.1)$$

The conclusion that $T \cong T'$ follows from the next exercise, and the further result that $l = l'$ can be seen in Fig. 6.1 (ii) and (iii).

Exercise⁺: Suppose that $d = d_{(T,l)}$, where $l > 0$. Show that the set of X -splits $A|B$ for which $d(a,a') + d(b,b') < d(a,b) + d(a',b')$ for all $a,a' \in A$ and $b,b' \in B$ is precisely $\Sigma(T)$. Deduce that d determines T up to equivalence.

To reconstruct a phylogeny $T \in P(n)$ and its edge lengths, we do not usually need all the $\binom{n}{2}$ possible d values. For example, for a binary tree $T \in B(n)$, a subset of $2n - 3$ carefully chosen pairs of elements from $[n]$ suffice to uniquely determine both T and l from the value of $d_{(T,l)}$ for those pairs (we explore this issue further in Section 6.2.3).

Notice that any positive linear combination of tree metrics represented on the same tree T is also a tree metric on T (since $\sum_i c_i d_{(T,l_i)} = d_{(T,l)}$ where $l = \sum_i c_i l_i$). More generally, $\mathcal{P} = (T_1, T_2, \dots, T_k)$ is a sequence of pairwise compatible trees from $P(X)$, and if d_i is a metric that has a tree representation on T_i for each i , then $\sum_{i=1}^k d_i$ is also a tree metric. A metric d' has a tree representation on the star tree T' if and only if $d'(x,y) = c(x) + c(y)$ for some function $c : X \rightarrow \mathbb{R}^{>0}$, in which case, for any $T \in P(X)$, $d_{(T,l)} + d'$ also has a tree representation on T .

Before proceeding, we state a key definition. A *distance function* δ on X is any function $\delta : X \times X \rightarrow \mathbb{R}^{>0}$ which is symmetric (i.e., $\delta(x,y) = \delta(y,x)$ for all $x,y \in X$) and vanishes only on the diagonal (i.e., $\delta(x,y) = 0 \Leftrightarrow x = y$). Thus a distance function is a metric if and only if it satisfies the triangle inequality. In particular, Theorem 6.1 holds if we require only that d is a distance function (since the four-point condition also implies the triangle inequality if we take two of the points as being equal).

In this chapter, will use d for a metric (e.g., a tree metric) and δ for a distance function. We can view the set of all distance functions on X as a full-dimensional closed polyhedral cone in $\mathbb{R}^{\binom{n}{2}}$, where $n = |X|$, within which the set of all metrics lies, and also as a full-dimensional closed polyhedral cone.

Exercise: Show that $d = d_{(T,l)}$ can be represented as $d = \sum_{A|B \in \Sigma(T)} l_{A|B} \delta_{A|B}$, where $l_{A|B} = l(e)$ for the edge e of T that corresponds to split $A|B$, and where $\delta_{A|B}$ is the distance function on X defined by $\delta_{A|B}(x,y) = 1$ if x and y lie in different blocks of $\{A,B\}$, with $\delta_{A|B}(x,y) = 0$ otherwise.

6.1.1 • Ultrametrics and the Gromov–Farris transform

An *ultrametric* is any metric d on X which satisfies the additional property that for any three pairwise distinct elements $x,y,z \in X$, two of the values $d(x,y), d(x,z)$, and $d(y,z)$ are equal and at least as large as the third (equivalently, $d(x,y) \leq \max\{d(x,z), d(y,z)\}$). This property can be viewed as a strengthening of the triangle inequality.

Ultrametrics arise in many mathematical settings, including graph theory, optimization, and p -adic analysis. In phylogenetics, ultrametrics relate to rooted phylogenies in the same way that tree metrics do to unrooted trees. There is a simple way to navigate between these two representations, as we explain below. Notice that every ultrametric satisfies the four-point condition (and thus is a tree metric by Theorem 6.1). Ultrametrics also satisfy some properties that elude tree metrics in general; for example, if d is an ultrametric on X and $f : \mathbb{R} \rightarrow \mathbb{R}$ is any monotonically increasing function, then the composition $f \circ d$ is also an ultrametric on X .

We now describe how an ultrametric d on X can be represented by a rooted phylogenetic X -tree. For any tree T in $RP(X)$, assign each edge e of T a length $l(e) > 0$ so that the sum of the lengths from the root ρ of T to each leaf is the same. Such an assignment of edge lengths is said to produce *ultrametric* edge lengths since it is readily verified that $d = d_{(T,l)}$ is an ultrametric.

Notice that for any ultrametric edge length assignment l on T , we can associate with it a function $f : V(T) \rightarrow \mathbb{R}^{\geq 0}$ by setting $f(v) = 0$ for each leaf v , and, for each interior vertex v , letting $f(v) = 2 \sum_{e \in P(T;v,x)} l(e)$, where $x \in c_T(v)$ is any leaf that is a descendant of v . We need to worry about whether f is well defined (in this second case), given that we have a choice in selecting x ; it is here that the ultrametric constraint imposed on l comes to our rescue and ensures that f is indeed well defined. With f in hand, it now follows that

- (i) $d(x,y) = f(\text{lca}_T(x,y))$ for all x,y ; and
- (ii) for any (directed) edge (u,v) of T , $f(u) > f(v)$.

This alternative way of representing d , using a function f satisfying properties (i) and (ii), is called a *vertex representation* of d , and there is a direct equivalence between vertex representations and tree representations based on ultrametric edge lengths.

We now describe how to associate a rooted phylogeny on X to any distance function $\delta : X \times X \rightarrow \mathbb{R}$. Let $\mathcal{A}[\delta]$ be the set X together with the collection of subsets A of X for which we have

$$\delta(a,a') < \delta(a,x) \text{ for all } a,a' \in A \text{ and } x \in X - A. \quad (6.2)$$

$\mathcal{A}[\delta]$ is the collection of *Apresjan clusters* of δ . This set can be constructed in $O(n^2)$ ($n = |X|$) time from δ [65]; moreover, it leads directly to a rooted phylogeny as follows.

Lemma 6.2. *For any distance function $\delta : X \times X \rightarrow \mathbb{R}$, $\mathcal{A}[\delta]$ is a hierarchy on X .*

Proof: Suppose that a pair $A,B \in \mathcal{A}[\delta]$ contravenes the nesting property (H1) (cf. Section 2.2.1). We will show this leads to a contradiction. By assumption, there are elements $a \in A - B$, $b \in B - A$ and $c \in A \cap B$. Since $a,c \in A \in \mathcal{A}[\delta]$ and $b \in X - A$, we have $\delta(c,a) < \delta(c,b)$, but since $b,c \in B \in \mathcal{A}[\delta]$ and $a \in X - B$, we have $\delta(c,b) < \delta(c,a)$. Combining these two inequalities gives $\delta(c,a) < \delta(c,b) < \delta(c,a) \Rightarrow \delta(c,a) < \delta(c,a)$, a contradiction. Thus (H1) must hold. Regarding the second hierarchy condition (H2), observe that $\{x\} \in \mathcal{A}[\delta]$ for all $x \in X$, since $0 = \delta(x,x) < \delta(x,y)$ for all $y \neq x$. Since X is also included in $\mathcal{A}[\delta]$ by definition, $\mathcal{A}[\delta]$ contains all the trivial clusters of X . ■

The first part of the following fundamental result can be proved by induction. Instead, we give a direct constructive argument. As we will see at the end of this section, it also allows for an easy proof of Theorem 6.1.

Proposition 6.3. *Any ultrametric d on X has a vertex representation on a rooted phylogenetic X -tree. Moreover, the only tree (up to equivalence) that provides such a representation is the one that corresponds to the hierarchy $\mathcal{A}[d]$.*

Proof: For any ultrametric d on X , let $T \in RP(X)$ be the rooted phylogeny corresponding to the hierarchy $\mathcal{A}[d]$ (cf. Lemma 6.2). For a vertex v of T , set $f(v)$ equal to the maximal value of $d(x,y)$ over all leaves x,y that are descendants of v (i.e., elements of the cluster $c_T(v)$).

Notice that if v is a proper descendant of u (i.e., $c_T(v)$ is a strict subset of $c_T(u)$), then $f(v) < f(u)$ since if z is a leaf in the cluster $c_T(u)$ that is not in $c_T(v)$, then since $c_T(v) \in \mathcal{A}[d]$, we must have

$$f(v) = \max\{d(x, y) : x, y \in c_T(v)\} < d(x, z) \leq f(u).$$

This implies that f satisfies condition (ii) for a vertex representation; also, if $x, y \in c_T(v)$ and $d(x, y) = f(v)$, then $v = \text{lca}_T(x, y)$. In particular, without loss of generality, we have $f(v) = d(x_1, x_2)$ for some leaf $x_1 \in C_1$ and some leaf $x_2 \in C_2$, where $\mathcal{C} = \{C_1, \dots, C_k\}$ is the set of maximal proper clusters of T contained in $c_T(v)$. To establish condition (i) for a vertex representation, it suffices to show that

$$v = \text{lca}_T(x, y) \Rightarrow d(x, y) = f(v). \quad (6.3)$$

To this end, first suppose that $x \in C_i$ and $y, y' \in C_j$ for some $i, j \in \{1, \dots, k\}$, and let $t = d(x, y)$. Since d is an ultrametric, and $d(y, y') < t$ (since $\text{lca}_T(y, y') = v'$ for a proper descendant v' of v), it follows that $d(x, y') = t$. Applying the same argument for $x' \in C_i$ it follows that $d(x', y') = t$ for all $x' \in C_i$ and $y' \in C_j$. We claim that $d(x, y) = f(v)$ for all x and y chosen from different clusters from \mathcal{C} ; otherwise, consider the smallest value $t' < f(v)$ for which $d(x, y) = t'$ for some x and y lying in different clusters from \mathcal{C} , and let \mathcal{C}' be the strict subset of \mathcal{C} for which $d(x, y) = t'$ for all leaves x, y chosen from different clusters in \mathcal{C}' . It follows that $d(x, y) \leq t'$ for all $x, y \in \cup \mathcal{C}'$, and $d(x, w) > t'$ for any $x \in \mathcal{C}'$ and $w \in X - \cup \mathcal{C}'$; however, this would imply that $\cup \mathcal{C}' \in \mathcal{A}[d]$, in violation of the assumption that \mathcal{C} comprises the set of maximal proper clusters of T in $c_T(v)$. This justifies (6.3), and thereby condition (i) for a vertex representation.

It remains to show that no other tree $T \in RP(X)$ furnishes a vertex representation for d . However, for any such tree T , any nontrivial cluster C of T must satisfy condition (6.2) and so belong to $\mathcal{A}[d]$. Moreover, property (ii) of a vertex representation ensures that no proper subset of $\mathcal{A}[d]$ can be the hierarchy associated with a tree that provides a vertex representation for d . Thus T has precisely $\mathcal{A}[d]$ as its set of clusters and therefore is unique up to equivalence. ■

Remarks. Proposition 6.3 shows that ultrametrics are in bijective correspondence with the pairs (\mathcal{H}, f) , where \mathcal{H} is a hierarchy on X and f is a nonnegative real-valued function on \mathcal{H} which is zero on singleton sets and which satisfies the additional property $A \subset B \Rightarrow f(A) \leq f(B)$.

The proof of Proposition 6.3 suggests a natural way to associate any distance function $\delta : X \times X \rightarrow \mathbb{R}^{\geq 0}$ with an ultrametric δ_U . Simply construct the rooted phylogeny T in $RP(X)$ corresponding to the hierarchy $\mathcal{A}[\delta]$ and then set

$$\delta_U(x, y) := \min\{\delta(x', y') : \text{lca}_T(x', y') = \text{lca}_T(x, y)\}$$

for all $x, y \in X$. Thus δ_U is an ultrametric on X and satisfies the additional property that $\delta_U(x, y) \leq \delta(x, y)$ for all $x, y \in X$. However, there is a different canonical ultrametric that provides a better approximation to δ , as we now explain.

The subdominant ultrametric. For a distance function $\delta : X \times X \rightarrow \mathbb{R}^{\geq 0}$, and $t \in \Delta := \{\delta(x, y) : x, y \in X\}$, consider the graph G_t having vertex set X and edge set $E_t = \{(x, y) : \delta(x, y) \leq t\}$. For $x, y \in X$, let

$$\delta_{SD}(x, y) = \min_{t \in \Delta} \{t : x \text{ and } y \text{ are in the same connected component of } G_t\}.$$

Notice that the collection of connected components of G_t , as t ranges over Δ , forms a hierarchy on X , and that δ_{SD} is an ultrametric that has a representation on the tree T_{SD} that corresponds to this hierarchy. Moreover, $\delta_{SD}(x,y) \leq \delta(x,y)$ for all $x,y \in X$ and δ_{SD} . It is clear that both δ_{SD} and δ_U can be computed efficiently (in polynomial time in $|X|$) from δ .

Exercise: Show that if δ is an ultrametric, then $\delta = \delta_{SD} = \delta_U$.

What is remarkable about δ_{SD} is that it is the (unique) largest ultrametric that lies below δ . More formally, if δ' is an ultrametric on X that satisfies $\delta'(x,y) \leq \delta(x,y)$ for all $x,y \in X$, then $\delta'(x,y) \leq \delta_{SD}(x,y)$ for all $x,y \in X$. For this reason, δ_{SD} is called the *subdominant ultrametric* associated with δ . In particular, $\delta_U(x,y) \leq \delta_{SD}(x,y)$ for all $x,y \in X$, and in general this inequality can be strict, as the following example shows.

Example: Consider the distance function on $X = \{x,y,w,z\}$ defined by $\delta(x,y) = \delta(y,w) = 1$ and $\delta(u,v) = 2$ for all other choices of $u,v \in X$, $u \neq v$. Then δ_{SD} agrees with δ except on the pair $\{x,w\}$ for which $\delta_{SD}(x,w) = 1 \neq 2 = \delta(x,w)$ and so is represented by the tree that has the nontrivial cluster $\{x,y,w\}$. By contrast, there are no nontrivial Apresjan clusters of δ , and so $\delta_U(u,v) = 1$ for all $u,v \in X, u \neq v$.

We can say a little more about the relationship between δ_U and δ_{SD} . Recall that the tree representing δ_U corresponds to the hierarchy of Apresjan clusters for δ . In fact, this tree is refined by the tree T_{SD} that provides a representation of δ_{SD} , as follows.

Proposition 6.4. *For any distance function δ on X , every Apresjan cluster of δ is a cluster of the tree T_{SD} that provides a representation for δ_{SD} .*

Proof: Suppose that C is an Apresjan cluster for δ , and let $t = t_C \in \Delta$ be the smallest value of t for which C is connected in G_t (the graph defined at the start of this section on subdominant ultrametrics). We will show that C is a connected component of G_t , and so C is a cluster of the tree T_{SD} . Since C is connected in G_t if C was not a connected component, then there would be some $x \in X - C$ and some element $a_x \in C$ with $d(a_x, x) \leq t$. But since C is an Apresjan cluster for δ , $d(a, a') < d(a_x, x)$ for all $a, a' \in C$. Thus there exists a smaller value $t' < t$ for which C is connected in $G_{t'}$, which contradicts the assumption that $t = t_C$ was chosen minimal with this property. ■

A further property of δ_{SD} is that, following a translation, it provides the closest ultrametric to δ under the l_∞ metric.²⁵ Let $\ell = \|\delta - \delta_{SD}\|_\infty$, where $\|*\|_\infty$ refers to the l_∞ metric (i.e., the largest difference between the two distance functions over all pairs from X), and let $\delta_{SD}^{+\frac{1}{2}\ell}$ be the ultrametric on X obtained by adding $\frac{1}{2}\ell$ to all values of $\delta(x,y)$ for which $x \neq y$ (for $x = y$, $\delta_{SD}^{+\frac{1}{2}\ell}(x,y) = 0$). The following result is from [86].

Proposition 6.5. *Suppose that δ is a distance function on X . Then $\delta_{SD}^{+\frac{1}{2}\ell}$ is the ultrametric on X that is closest to δ under the l_∞ metric.*

²⁵The problem of finding the closest tree metric to a distance function is NP-hard in general.

The Gromov–Farris transform. Notice that if $d : X \times X \rightarrow \mathbb{R}$ is any symmetric function (i.e., $d(x, y) = d(y, x)$ for all $x, y \in X$) and we select any element $r \in X$, then we can define the function $\tilde{d}_r : X \times X \rightarrow \mathbb{R}$ by setting

$$\tilde{d}_r(x, y) = \frac{1}{2}(d(x, y) - d(x, r) - d(y, r)).$$

When d is a tree metric $-\tilde{d}_r(x, y)$ has a simple interpretation: it is the length of the path from r to $\text{lca}_r(r, x, y)$, as indicated in Fig. 6.2. Consequently, \tilde{d}_r satisfies the ultrametric property $\tilde{d}_r(x, y) \leq \max\{\tilde{d}_r(x, z), \tilde{d}_r(y, z)\}$ for all distinct elements x, y, z in X .

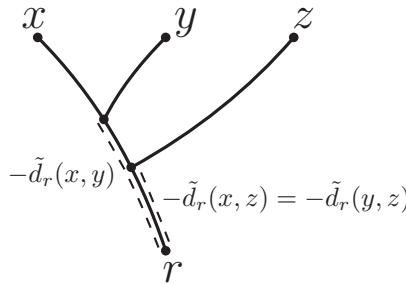


Figure 6.2. The quantities $-\tilde{d}_r(i, j)$ for $i, j = x, y, z$ correspond to the lengths of the dashed paths shown. Other edges and leaves that may be present in this tree are not shown.

The transformation $d \mapsto \tilde{d}_r$ is referred to as the *Gromov product*. Notice that even when d is a metric, $\tilde{d}_r(x, y) < 0$ for all $x, y \in X - \{r\}$ (including $x = y$), so \tilde{d}_r is not itself a distance function. Fortunately, this can be easily remedied, as follows. Let c be any constant that is greater than or equal to $\max\{d(x, r) : x \in X\}$. For $x, y \in X - \{r\}$, the function

$$d_r(x, y) = \begin{cases} c + \tilde{d}_r(x, y) & \text{if } x \neq y, \\ 0 & \text{if } x = y \end{cases}$$

is then a distance function on X . The transformation $d \mapsto d_r$ is referred to as the “Farris transform” in biology [118]. It is easily verified that if d is a metric, then so too is d_r , and that if d satisfies the four-point condition, then d_r is an ultrametric on $X - \{r\}$ for every choice $r \in X$.

By Proposition 6.3, d_r has a vertex representation on a tree $T \in RP(X - \{r\})$, and from this it is straightforward to construct a representation of d on the tree $T' \in P(X)$ obtained by the operation o_{+r} of adding r as an outgroup to the root. This not only establishes the nontrivial part of Theorem 6.1 but also shows that the four-point condition need only hold for all four points that contain a given leaf r .

Exercise: Prove that if d satisfies the four-point condition, then d_r is an ultrametric on $X - \{r\}$ for every choice $r \in X$.

6.1.2 • Symbolic ultrametrics

The theory of symbolic ultrametrics, developed by Sebastian Böcker and Andreas Dress [37], provides an elegant and incisive technique for establishing certain existence and

uniqueness theorems in phylogenetics. It has also led to the development of some novel approaches in molecular systematics. For example, [183] applied symbolic ultrametric theory (along with a connection to cographs) to propose a new way of inferring phylogenetic trees from certain types of genomic data.

The idea of symbolic ultrametrics is to abstract away the notion of distances taking real values, to consider more generally a (formal) “distance” function taking values in *any* set M , and then to find minimal condition(s) under which such a “distance” function can be represented (uniquely) on a rooted tree, with elements of M assigned to the vertices. Here we give a brief summary of the key definitions and main result.

A function $\partial : X \times X \rightarrow M$ is said to be a *symbolic ultrametric* on X if the following three conditions hold:

- (U1) ∂ is symmetric (i.e., $\partial(x,y) = \partial(y,x)$ for all $x,y \in X$);
- (U2) ∂ has the property that, for all pairwise distinct elements $x,y,z \in X$, at least two of the values $\partial(x,y), \partial(x,z), \partial(y,z)$ are equal; and
- (U3) ∂ satisfies the condition that there are no pairwise distinct elements x,y,w,z for which

$$\partial(x,y) = \partial(y,w) = \partial(w,z) \neq \partial(y,z) = \partial(z,x) = \partial(x,w). \quad (6.4)$$

In the special case where $M = \mathbb{R}^{\geq 0}$, any ultrametric satisfies these three properties. However, the converse is not true. For example, consider the function $\partial_s : X \times X \rightarrow \mathbb{R}^{\geq 0}$, where $X = \{a,b,c,d\}$, and with $\partial_s(x,y) = 2$ if $\{x,y\} = \{a,c\}$ or $\{b,c\}$, $\partial_s(x,x) = 0$, for all $x \in X$ and $\partial_s(x,y) = 1$ otherwise. Then ∂_s is a symbolic ultrametric but it fails to be an ultrametric.

We say that a rooted phylogenetic X -tree $T = (V,A)$ with a map $t : V \rightarrow M$ provides a *discriminating symbolic representation* of a function $\partial : X \times X \rightarrow M$ if the following two conditions hold:

- (i) $\partial(x,y) = t(\text{lca}_T(x,y))$ for all $x,y \in X$,
- (ii) $t(u) \neq t(v)$ for every interior edge (u,v) of T .

An example of a discriminating symbolic representation of the symbolic ultrametric ∂_s (mentioned above) is shown in Fig. 6.3.

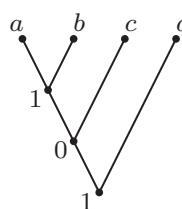


Figure 6.3. The t values shown on the interior vertices (together with $t(x) = 0$ for each leaf x) provides a discriminating symbolic representation of the symbolic ultrametric ∂_s .

It can be checked that any function $\partial : X \times X \rightarrow M$ that has a discriminating symbolic representation is also a symbolic ultrametric (i.e., it satisfies the properties (U1)–(U3) above). There is a strong converse to this, which is summarized in the following result from [37].

Theorem 6.6. $\partial : X \times X \rightarrow M$ is a symbolic ultrametric if and only if ∂ has a discriminating symbolic representation; moreover, in this case ∂ uniquely determines both the associated rooted phylogenetic tree T (up to equivalence) and map t .

A rooted phylogenetic tree on X that provides a representation for ∂ is simply the BUILD tree $\mathcal{A}_{\mathcal{R}}$ (from Chapter 4) for the set \mathcal{R} of rooted triples $xy|z$ ($x, y, z \in X$) for which

$$\partial(x, y) \neq \partial(x, z) = \partial(y, z).$$

In this way, it is possible to construct a discriminating symbolic representation of any symbolic ultrametric in polynomial time (for details, see [315], Section 7.6).

6.1.3 ■ Distances versus characters

In Chapter 5, we studied the encoding of trees via characters. A natural question is to ask how this notion relates to encoding trees by distances. For binary characters there is a one-way direct link. Let $\mathcal{C} = (f_1, \dots, f_k)$ be a sequence of binary characters, and let $\delta_{\mathcal{C}} : X \times X \rightarrow \mathbb{R}^{\geq 0}$ denote the (normalized) *Hamming distance* on X induced by \mathcal{C} . That is, $\delta_{\mathcal{C}}$ is the proportion of characters in \mathcal{C} for which x and y take different states. Formally,

$$\delta_{\mathcal{C}}(x, y) = \frac{1}{k} \times \#\{j \in \{1, \dots, k\} : f_j(x) \neq f_j(y)\}.$$

It is easily shown that if a sequence $\mathcal{C} = (f_1, \dots, f_k)$ of binary characters has a perfect phylogeny, then $\delta_{\mathcal{C}} = d_{(T, l)}$, where $T \in P(X)$ is the unique minimal (under refinement) perfect phylogeny for \mathcal{C} and where l is strictly positive on the interior edges of T and nonnegative on the pendant edges (notice that this is a slightly more general notion of tree representation). This connection does not go the other way, however: it is entirely possible for a sequence \mathcal{C} of binary characters that do not have a perfect phylogeny to possess a tree representation for $\delta_{\mathcal{C}}$.

Moreover, when we move away from binary characters, even the one-way connection we have just described evaporates and is replaced by a rather startling contrast: there are sequences of characters that have a unique perfect phylogeny, for which the induced (normalized) Hamming distances also have a representation on a unique tree, although these two trees can be arbitrarily different. In addition, the representation on the second tree can be chosen to be an ultrametric one. This is summarized in the following result from [21].

Proposition 6.7. For any $n \geq 4$ and any two distinct trees $T, T' \in B(n)$, there is a sequence \mathcal{C} of 3-state characters for which

- T is a unique perfect phylogeny for \mathcal{C} ;
- $\delta_{\mathcal{C}}$ has a tree representation on T' and is an ultrametric (and so its representation on T' uses ultrametric edge lengths).

A simple example of this result in the case where $n = 4$ is illustrated in Fig. 6.4. Notice that even if we transform $\delta_{\mathcal{C}}$ by any monotone increasing transformation to try to “correct” the distance values, then the resulting distances will still have a representation on T' (only) since $\delta_{\mathcal{C}}$ is an ultrametric.

Despite Proposition 6.7, we will see in the next chapter that the expected Hamming distance between species, derived from a sequence \mathcal{C} of characters that have evolved on

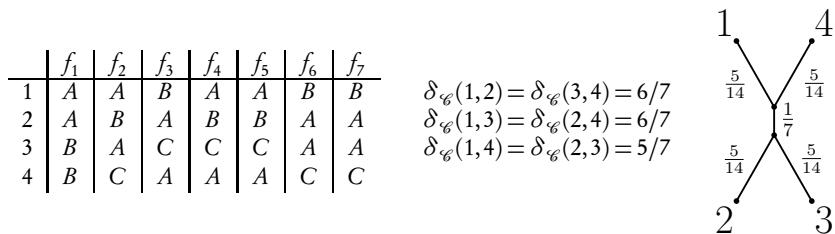


Figure 6.4. The characters f_1, \dots, f_7 capture the quartet tree $12|34$ as a unique perfect phylogeny; however, the normalized Hamming distance they induce has a tree representation on a different quartet tree, namely $14|23$, and with ultrametric edge lengths once it is rooted at the midpoint of its central edge.

a phylogenetic tree T under a stochastic model, can be “transformed” into distances that have a tree representation on T .

Exercise: Suppose that \mathcal{C} is a sequence of binary characters on a set X of size 4. Show that $ab|cd$ is an MP (maximum parsimony) tree for \mathcal{C} if and only if $\delta_{\mathcal{C}}(a,b) + \delta_{\mathcal{C}}(c,d) \leq \min\{\delta_{\mathcal{C}}(a,c) + \delta_{\mathcal{C}}(b,d), \delta_{\mathcal{C}}(a,d) + \delta_{\mathcal{C}}(b,c)\}$.

6.1.4 • Distances from genomic data

In this section, we briefly describe some of the ways that biologists use genetic and genomic data to estimate an “evolutionary distance” between pairs of species.

The most widespread approach is to compute distances from aligned DNA or amino acid sequences. Here, “aligned” means that the sites in the sequences have been “matched up” across the species (perhaps by introducing gaps in places) so that each site can be viewed as a homologous discrete character on the set X of species. A measure of dissimilarity (such as the normalized Hamming distance, described in the previous section) can then be computed from these discrete characters.

Notice that if we knew the average (across the characters) of the number of state changes that occurred along the paths linking pairs of species, then since these “evolutionary distances” have, by definition, a tree representation on the underlying tree, we would be able to infer the correct tree. The problem is that these evolutionary distances cannot be directly observed. Instead, the normalized Hamming distances provide merely an underestimate of them, due to the fact that certain changes are “hidden” (for example, a change from state A to B to C along a path is seen only as a single change at the leaves (from A to C) while a change from A to B to A appears at the leaves to be no change at all. Thus the normalized Hamming distances need to be transformed or “corrected” to give an estimate of an evolutionary distance (based on the models similar to those we will discuss in Chapter 7), and from these distances a phylogeny can then be estimated by techniques that we will consider in the next section.

There are other ways to estimate evolutionary distance, and we mention some topical ones here. The first is a class of methods that estimate evolutionary distances based on raw (unaligned) DNA sequence data. One approach measures the distance between two species by how much they differ in the distribution of counts of (consecutive) substrings of length k (called k -mers). For example, for $k = 6$, the sequence AACATG will appear a certain number of times in the DNA of two different species, as will the other $4^6 - 1$

possible 6-mers. Two species that have only recently diverged are likely to have very similar counts for most of the 6-mers. However, for species that are well separated in the tree, events such as site substitutions (e.g., an A changing to a G at a position), insertions (of new nucleotides, or blocks of nucleotides), and deletions will cause these counts to diverge over time. Numerous ways have been proposed for converting these counts into estimates of evolutionary distance, though most lack an explicit model or compelling justification. Alternative approaches to unaligned sequence data include various heuristics based on information theory or data compression concepts. A recent survey is provided in [178].

A quite different approach for estimating evolutionary distance is to compare the order of conserved regions (we will loosely call these “genes” here) on chromosomes between the different species. Chromosomes can be viewed as a sequence of genes arranged in a linear (or sometimes circular) ordering; moreover, each gene has a particular orientation. This arrangement of genes can be disrupted by genome rearrangement events. These events may alter the order and orientation of a sequence of genes (inversions), or move blocks of sequences (translocations), but they can also result in chromosomes being fused together or cut (fusion and fission events). The use of gene order and arrangement to estimate evolutionary distance, pioneered by David Sankoff and Pavel Pevzner and colleagues, initially aimed to compute the smallest number of rearrangement events of a given type to convert a single chromosome into another, or to consider some more coarse (but easily computed) statistic, such as the number of “breakpoints” between the two sequences.

While these approaches led to some impressive algorithmic results (notably the Hannenhalli–Pevzner theory from the mid-1990s), a discovery in the mid-2000s [378] changed the focus to a distance measure that was remarkably easy to compute, yet also allowed for several of the rearrangement events mentioned above to be handled at once. This is the *double-cut-and-join* (DCJ) operation. The algorithmic and mathematical properties of the corresponding distance $d_{\text{DCJ}}(G_1, G_2)$ (the minimal number of DCJ operations required to convert one multichromosome genome G_1 into another G_2) were developed further by Anne Bergeron, Jens Stoye, and colleagues (see, for example, [25]). An algebraic description of DCJ was provided by [253] and developed further by Andrew Francis and colleagues. In this way, a multichromosome genome consisting of n oriented genes in linear and/or circular chromosomes can be viewed as a permutation of $\{1, 2, \dots, 2n\}$, and DCJ can be viewed as a group action on the genome; this leads to an elegant group-theoretic expression for $d_{\text{DCJ}}(G_1, G_2)$ (for details, see [28], and for further applications of group theory to gene order, see [144]).

It is worth noting that these various ways to define distances between genomes, based on minimizing rearrangement operations, are not based on any particular stochastic model of genome evolution, a topic that has so far received less attention.

6.2 ■ Distance-based tree reconstruction methods

A variety of fast (polynomial time) methods have been devised for building a phylogenetic X -tree from an arbitrary distance function d on X . The most popular, by far, is *neighbor joining* (NJ) (the paper [305] that described this heuristic algorithm has now been cited more than 40000 times). A desirable property of such methods is that when a distance function has a tree representation, the method will return the underlying tree and its associated edge lengths. This consistency property can be formalized as follows.

Consider any tree reconstruction method that takes any distance function δ on X and returns a tree $T_\delta \in P(X)$ along with associated edge lengths l_δ . The method is said to be *consistent* if it satisfies the following condition: Whenever δ has a tree representation on some phylogeny $T \in P(X)$ (i.e., $\delta = d_{(T,l)}$, with $l > 0$), then $T_\delta = T$ and $l_\delta = l$.

From what we have learned already in this chapter, it's easy to concoct a consistent tree reconstruction method (the challenge is to find one that performs well when δ is only “close” to a tree metric). For example, we could select some element $r \in X$, apply the Gromov–Farris transformation, and construct $\mathcal{A}[\delta_r]$ (which is a hierarchy by Lemma 6.2), add back in r as a leaf, and then estimate the $l(e)$ values using the formulae in the caption to Fig. 6.1.

A more systematic approach is the following method. Given δ this method searches for a pair (T, l) to minimize the total length of the tree $\sum_{e \in E(T)} l(e)$, subject to the constraints that $l(e) \geq 0$ for all edges e of T and $d_{(T,l)}(x, y) \geq \delta(x, y)$ for all $x, y \in T$.²⁶ Note that if $l(e) = 0$, then e is collapsed in the reconstructed tree to ensure that all edges in that tree have strictly positive length. For a fixed phylogeny T , the optimization of l is an easily solved linear programming problem. However, finding an optimal tree T is hard. Although it is not immediately obvious, it can be shown (by various arguments) that this tree reconstruction method is consistent in returning (T, l) when applied to $d_{(T,l)}$.

A further approach is to search for a pair (T, l_δ) where l_δ is an assignment of edge lengths l to T that is chosen so that $d_{(T,l)}$ is as “close” to δ as possible. T is then chosen according to some further optimization criterion. For example, one may ask for the value l_δ that minimizes a sum of squares criterion or, more generally, a weighted sum of squares criterion. In other words, given δ and a phylogeny T , $l = l_\delta$ is selected to minimize

$$\Delta_{(T,l)}(\delta) := \sum_{ij} w_{ij} (\delta(i,j) - d_{(T,l)}(i,j))^2.$$

One could then search for the pair (T, l_δ) that minimizes $\Delta_{(T,l)}(\delta)$ or that minimizes $\sum_e \tilde{w}(e)$, where $\tilde{w}(e)$ could be taken throughout to be either $l(e)$ or $|l(e)|$ or $\max\{l(e), 0\}$. The mathematical analysis of these approaches, particularly the estimation of l_δ used to minimize $\Delta_{(T,l)}(\delta)$ for a given T , and the conditions under which the resulting tree reconstruction is consistent, has lead to a rich combinatorial story, culminating in two high-profile papers [257] and [283]. We refer readers to these papers for further details and references, and move on to consider two alternative approaches that are widely used in biology.

6.2.1 • Neighbor joining (NJ)

As an introduction to NJ, observe that if d is a tree metric on X , then for $r \in X$, any pair x, y with $x \neq y$ that minimizes the Gromov product $\tilde{d}_r(x, y)$ (or $d_r(x, y)$) is necessarily adjacent to a common vertex u (i.e., the pair x, y forms a cherry). Moreover, the distance from u to x, y and from u to every other element in $X - \{r\}$ can be readily calculated from d .

To build a tree from an arbitrary distance function δ on X , NJ starts with a star tree on X and proceeds iteratively, by sequential identification of pairs of leaves that are to be new cherries in the tree constructed thus far, along with the re-estimation of distances and edge

²⁶This last condition suggests a similarity to maximum parsimony in the character setting since, in both cases, the methods seek to minimize the amount of “change” in the tree, subject to the constraint that a certain amount of change is required to explain the data.

lengths. However, rather than selecting pairs that minimize δ_r , to select a cherry based on an arbitrary choice of r , it is desirable, for statistical purposes, to apply an averaging approach over the possible choices of r . Thus the pair x, y is selected to minimize

$$Q(x, y) := \delta(x, y) - \frac{1}{n-2} \sum_{r \in X} [\delta(x, r) + \delta(y, r)]. \quad (6.5)$$

We could consider other functions \hat{Q} of the δ values to minimize. Nevertheless, $\hat{Q} = Q$ satisfies the following three desirable properties (properties (2) and (3) are clear, but (1) requires a short proof [63]).

- (1) If $\delta = d_{(T, l)}$ and x, y minimizes \hat{Q} , then x, y is a cherry of T .
- (2) \hat{Q} is a linear function of δ .
- (3) If x, y minimizes \hat{Q} for δ , then for any permutation σ of X , $(\sigma(x), \sigma(y))$ minimizes \hat{Q} for δ^σ .

In part (3), $\delta^\sigma(x, y) = \delta(\sigma(x), \sigma(y))$ for all $x, y \in X$, so condition (3) is analogous to the neutrality condition for consensus functions in Chapter 2; both conditions require that the names of the species should not play any special role in the operation of the algorithm. The main result from [63] is that *any* function \hat{Q} that satisfies properties (1)–(3) will make the same choices as Q . Moreover, for such a function \hat{Q} , one necessarily has $\hat{Q} = aQ + b$, where a and b depend only on $n = |X|$, and $a > 0$.

Starting with a star tree with unspecified edge lengths, the NJ algorithm selects a pair x, y to minimize $Q(x, y)$ as given by eqn. (6.5).²⁷ The branches leading to x and y are replaced with a new branch leading to a new vertex u , and with new edges from u to x and u to y . Lengths are now specified for the latter two new edges as follows:

$$l(\{u, x\}) = \frac{1}{2} \delta(x, y) + \frac{1}{2(n-2)} \left[\sum_{r \in X} \delta(x, r) - \sum_{r \in X} \delta(y, r) \right],$$

and $l(\{u, y\})$ is obtained by symmetry. NJ then replaces x, y by u in X to obtain a set X' , which is one element smaller, and δ is replaced by the distance function δ' on X' that is the restriction of δ for pairs that do not contain u . For $x' \in X - \{x, y\}$,

$$\delta(u, x') = \frac{1}{2} [\delta(x, x') + \delta(y, x') - \delta(x, u) - \delta(y, u)].$$

We now have a tree in which the numbers of interior edges and leaves have both reduced by one. The process described is now repeated on this tree (with n reduced to $n-1$ in the above formulae) and continued until just three taxa are present. The edge lengths of this three-leaf tree are determined by the final three δ values. By tracing back through the process, the edge lengths of all edges in the phylogeny on X have also been determined (any that have a length that is less than or equal to zero are collapsed). This gives the final output tree $T \in P(X)$, with its associated edge lengths, in a total running time of $O(n^3)$ for $n = |X|$.

²⁷The formula for Q , and the further formulae for $l(\{u, x\})$, are the ones from [344] which modified the original expressions from [305]; however, they are essentially equivalent and reconstruct the same tree and the same edge weights, as shown in [148].

ℓ_∞ safety radius and discontinuity. While consistency is a desirable property for any distance-based tree reconstruction method, it only applies when the distance function δ is exactly equal to a tree metric. A more robust notion is that a distance tree reconstruction method should return T when δ is “sufficiently close” to $d_{(T,l)}$. For this notion to make sense, we need to restrict T to be binary, since if T has a vertex of degree greater than 3, then for any associated edge lengths $l > 0$, any tree T' that refines T has edge lengths l' for which $d_{(T',l')}$ comes arbitrarily close to $d_{(T,l)}$. Moreover, if l_{\min} is the smallest interior edge length of T , then what is required for “sufficiently close” is dependent on this value. Accordingly, a distance-based tree reconstruction method is said to have *safety radius* r if the method is guaranteed to return each binary tree T whenever δ differs from $d = d_{(T,l)}$ by less than $r \cdot l_{\min}$ on each pair of leaves.

Notice that we are here using the ℓ_∞ metric in comparing δ with $d_{(T,l)}$. This notion was introduced by Kevin Atteson [18], who showed that the safety radius of NJ is exactly $r = \frac{1}{2}$. It is not hard show that $\frac{1}{2}$ is the largest possible safety radius for any distance-based tree reconstruction method (see the following exercise); some other tree reconstruction methods also achieve this optimal value, but not all; indeed, for some methods the safety radius converges to zero as n grows.

Exercise⁺: Show that for any $T \in P(X)$ with edge lengths $l > 0$, there is a different phylogeny $T' \in P(X)$ with edge lengths $l' > 0$, with $\epsilon := l_{\min} = l'_{\min}$ and $|d_{(T,l)}(x,y) - d_{(T',l')}(x,y)| \leq \epsilon$ for all $x,y \in X$.

The safety radius requirement that $|d(x,y) - d_{(T,l)}(x,y)| < r \cdot l_{\min}$ holds for *all* pairs $x,y \in X$ is very restrictive, particularly as n grows, since it fails if just one “outlying” pair x,y violates the condition. Thus a more robust notion of a safety radius may be one based on the l_2 (or l_1) metric, a topic initiated in [125]. An alternative concept of a “stochastic safety radius” in which the $\delta(x,y)$ values are regarded as independent random Gaussian variables centered on $d_{(T,l)}(x,y)$ was recently introduced and studied in [152].

Curiously, although distance data (generated from simulated data derived from some underlying tree T with edge lengths) often violates the safety radius condition required for NJ, this method still succeeds in correctly returning T . In a ground-breaking paper [256], more relaxed underlying combinatorial conditions were shown to suffice for NJ to successfully return a given tree T , along with the proof of a general result that entails Atteson’s original safety radius result for NJ as a corollary.

The authors of [256] also resolved a conjecture of Atteson that had been open for 20 years concerning a variant of safety radius, which had been formulated to address another limitation of the original safety radius definition that is also exacerbated as n grows. Namely, l_{\min} is likely to become small because of the increasing possibility of there being at least one very short edge in a tree with many edges. This motivates the following notion: A distance-based tree reconstruction method has *edge safety radius* r' if whenever $|\delta(x,y) - d_{(T,l)}(x,y)| < r' \cdot l(e)$ for all x,y , and some interior edge e of T , the method applied to δ will return a tree that has a split that corresponds to edge e in T . In other words, edges that are not too small will be recovered by the method (more precisely, the splits they induce will be recovered). Atteson conjectured that NJ has edge safety radius $\frac{1}{4}$, and this was established in [256].

Even when a distance-based tree reconstruction method has a positive safety radius, it is still possible for an arbitrarily small slight perturbation of δ to cause the method to “jump” to a quite different tree, when δ is not close to a tree metric. Indeed, NJ

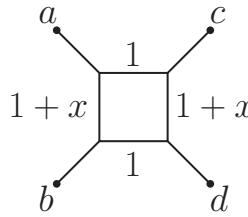


Figure 6.5. A distance function $\delta^{(x)}$ on which NJ exhibits a discontinuity around $x = 0$.

exhibits such discontinuities. For any $\epsilon \in (0, 1)$ and $x \in (-\epsilon, \epsilon) - \{0\}$, consider the distance function $\delta^{(x)}$ on $\{a, b, c, d\}$ that corresponds to the shortest path distance in the graph shown in Fig. 6.5. A positive safety radius for a method (such as NJ) ensures that such large “jumps” only occur when applied to a distance function that is far from a tree metric.

As $x \rightarrow 0$, either from above or from below, NJ applied to $\delta^{(x)}$ returns a quartet tree with an interior edge length that converges to $\frac{1}{2}$. However, the particular quartet tree returned from $\delta^{(x)}$ is $ab|cd$ as x increases to 0 from below, although it is $ac|bd$ as x declines to zero 0 from above. In other words, the reconstructed phylogeny “flips” from one quartet tree to the other without the interior edge shrinking to zero in the process.

6.2.2 • Balanced minimum evolution (BME)

Given a phylogenetic X -tree T with an edge length assignment l , consider the sum $L = \sum_{e \in E(T)} l(e)$ of all the edge lengths across the tree. L can be expressed as a positive linear combination of $d(i, j)$ values in various ways. In particular, suppose that a cyclic permutation (x_1, x_2, \dots, x_n) of X is a circular ordering for T (cf. Section 2.4.2). In this case

$$L = \frac{1}{2} [d(x_1, x_2) + d(x_2, x_3) + \dots + d(x_{n-1}, x_n) + d(x_n, x_1)], \quad (6.6)$$

since for each edge e of T its length $l(e)$ appears in exactly two of the d values on the right-hand side.²⁸ For example, for the tree in Fig. 6.6, we can write

$$L = \frac{1}{2} [d(a, b) + d(b, c) + d(c, d) + d(d, e) + d(e, f) + d(f, g) + d(g, a)], \quad (6.7)$$

since the cyclic permutation (a, b, c, d, e, f, g) traverses the tree in a clockwise direction and so covers every edge exactly twice. However, as we saw in Section 2.4.2, a phylogeny will generally have several circular orderings. For example, Lemma 2.6(i) tells us that the tree in Fig. 6.6 has $3! \times 3! \times 2! = 72$ circular orderings, and that these alternatives

²⁸In addition, if the cyclic permutation (x_1, x_2, \dots, x_n) of X is not a circular ordering for T , then L is strictly less than the expression on the right of eqn. (6.6).

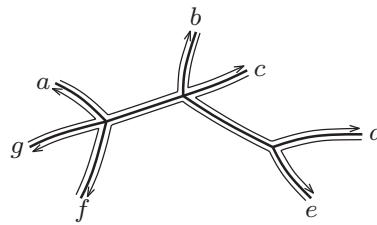


Figure 6.6. The circular ordering (a, b, c, d, e, f, g) covers every edge of the tree exactly twice.

(e.g., (a, d, e, c, b, g, f)) can lead to different, but equally valid, expressions for the same quantity L than that given in eqn. (6.7).

It is mathematically natural to ask for an expression for L that is not dependent on the arbitrary choice of a particular circular ordering. This is also helpful for applications, where we are provided with estimates of the d values rather than exact values, since different expressions of L will lead to different estimates of this quantity.

To obtain such a canonical expression for L is easy: we can simply average all of the different expressions for L (one for each circular ordering of T) given by eqn. (6.6) over all possible circular orderings of T . On the left, we still have L , and the equation we obtain can then clearly be written as

$$L = \sum_{\{x,y\}} \lambda_T(x,y) d(x,y), \quad (6.8)$$

for some nonnegative coefficients $\lambda_T(x,y)$. What are these coefficients? They turn out to have a nice description in terms of the number and degrees of the vertices in T on the path between x and y . If T is a binary tree, then $\lambda_T(x,y) = (\frac{1}{2})^{|I(x,y)|}$, where $I(x,y)$ is the set of interior vertices in the path between x and y in T . For instance, in the case of the quartet tree $xy|wz$ (Fig. 6.1(i)), this gives

$$L = \frac{1}{2}d(x,y) + \frac{1}{2}d(w,z) + \frac{1}{4}(d(x,w) + d(x,z) + d(y,w) + d(y,z)).$$

More generally, for any phylogenetic tree T , it can be shown [316] that

$$\lambda_T(x,y) = \prod_{v \in I(x,y)} (d(v)-1)^{-1}. \quad (6.9)$$

The identity in eqn. (6.8) suggests a new way to build phylogenetic trees from distances, which is called *balanced minimum evolution* (BME), proposed by the biologist Yves Pauplin [287] (based on the coefficients $(\frac{1}{2})^{|I(x,y)|}$ for binary trees). Given an arbitrary distance function (not necessarily a tree metric) δ on X , BME scores each phylogenetic X -tree T by the value

$$L(T, \delta) = \sum_{\{x,y\}} \lambda_T(x,y) \delta(x,y),$$

and searches for a tree T that has the smallest $L(T, \delta)$ score. If δ has a tree representation on some tree T , then this tree has the smallest $L(T, \delta)$ score. BME is a consistent tree reconstruction method, and, like NJ, BME has the largest possible safety radius of $\frac{1}{2}$. BME can be viewed algebraically as a type of “weighted least squares” method [106] and is also closely connected with the NJ method, both algorithmically (NJ can be viewed as a

locally greedy algorithm for constructing a BME tree [149]) and via polyhedra geometry [179].

Searching for a BME tree is typically done using tree rearrangement operations, such as NNI and SPR (from Chapter 2). At each stage of this search a neighboring tree that has smallest $L(T, \delta)$ score is selected as the next tree in the sequence. One reason for the popularity of BME is that there are fast algorithms to recompute the BME score of trees under tree rearrangements (see [107]). This is particularly pertinent because of a nontrivial result from [47], which shows that when δ is sufficiently close to $d_{(T,l)}$ for some binary tree T with $l > 0$, an SPR strategy based on locally optimizing the BME score and applied to an arbitrary starting tree T' is guaranteed to converge on T .

6.2.3 • Tree reconstruction from partial distances

In Chapter 4, we saw how a tree can be reconstructed from just some of its “building blocks” (quartet trees, or rooted triplets). Likewise, we can ask here if we really need distances for all pairs of elements from X in order to recover a tree and/or its edge lengths. Apart from its intrinsic interest, this question is motivated in part by the fact that accurate empirical estimates of distances between species may only be known for certain pairs of species. This may hold, for example, if the distances are estimated from various genes, each of which may only be present (or available) for some subset of species.

Let \mathcal{L} denote a subset of $\binom{X}{2}$ and, for a metric d on X , let $d|_{\mathcal{L}}$ denote the restriction of d to \mathcal{L} (i.e., $d(x,y)$ is specified for just those pairs x,y for which $\{x,y\} \in \mathcal{L}$). Given a phylogeny T in $P(X)$, we say that \mathcal{L} determines the edge lengths of T if, for all $l > 0$ and $l' > 0$, we have

$$d_{(T,l)}|_{\mathcal{L}} = d_{(T,l')}|_{\mathcal{L}} \implies l = l'.$$

In addition, we say that \mathcal{L} determines the topology of T if, for all $l > 0$, and $T' \in P(X)$ and $l' > 0$, we have

$$d_{(T,l)}|_{\mathcal{L}} = d_{(T',l')}|_{\mathcal{L}} \implies T = T'.$$

We say that \mathcal{L} determines both T and its edge lengths if \mathcal{L} determines both the topology of T and the edge lengths of T .²⁹

Some of the basic results concerning these notions (mostly from [119]) are the following.

- For every binary tree $T \in B(n)$, there is a set \mathcal{L}_T of size $2n - 3$ that determines both the topology of T and its edge lengths, and a set \mathcal{L}'_T of size $2n - 4$ that determines the topology of T . Moreover, these are the smallest possible sizes of such sets. If x_1, \dots, x_n is any circular ordering of T (cf. Section 2.4.2), then we can take

$$\mathcal{L}_T = \{\{x_1, x_i\}, i = 2, \dots, n\} \cup \{\{x_{i-1}, x_i\}, i = 3, \dots, n\}.$$

- At the other extreme, the only subset \mathcal{L} of $\binom{X}{2}$ that determines the topology of the star tree on X is the full set $\binom{X}{2}$.
- It is possible for a set \mathcal{L} to determine the edge lengths of a binary phylogeny T but not determine its topology; an example is shown in Fig. 6.7.

²⁹The word “lasso” has been used in earlier work; however, given the alternative uses of that word in other settings, we use the more neutral word “determine” here.

- For $T \in P(X)$ let $\mathcal{M}(T)$ denote the collection of all minimal subsets \mathcal{L} of $\binom{X}{2}$ that determine the edge lengths of T . Then each such \mathcal{L} has size exactly equal to the number of edges of T (by elementary linear algebra). Moreover, the map $T \mapsto \mathcal{M}(T)$ from $P(X)$ to $2^{\binom{X}{2}}$ is one-to-one (i.e., it is possible to reconstruct T from $\mathcal{M}(T)$).

Some further insight into the question of whether or not \mathcal{L} determines the topology or the edge lengths of some tree is provided by considering the graph (X, \mathcal{L}) with vertex set X , and edge set \mathcal{L} , as the following result shows.

Proposition 6.8.

- If $|X| \geq 4$, and \mathcal{L} determines the topology of T , then (X, \mathcal{L}) is connected.
- If \mathcal{L} determines the edge lengths of T , then (X, \mathcal{L}) is strongly nonbipartite (i.e., every connected component contains a cycle of odd length).
- In particular, (X, \mathcal{L}) must be connected and strongly nonbipartite if \mathcal{L} determines both the topology and the edge lengths of some tree.

Figure 6.7 illustrates that the necessary (connectivity) condition for \mathcal{L} to determine the topology of T is not sufficient.

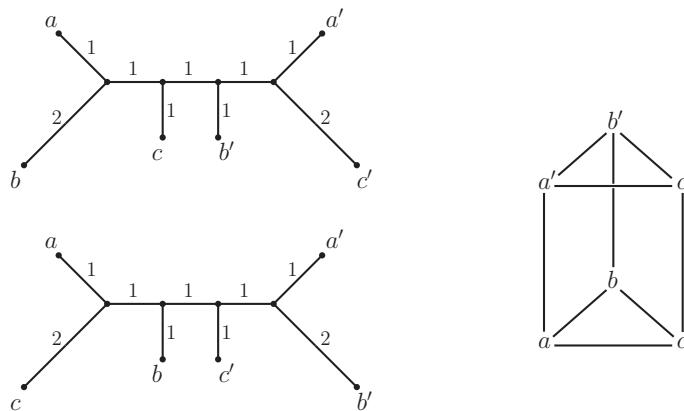


Figure 6.7. Left: The set $\mathcal{L} = \{ab, ac, bc, a'b', a'c', b'b', cc'\}$ determines the edge lengths of the top tree, but it does not determine its topology, since the lower tree has edge lengths for which the induced distances agree for the pairs in \mathcal{L} . Right: The associated graph (X, \mathcal{L}) .

For binary phylogenies, a particular class of sets \mathcal{L} has been considered in detail, the so-called “triplet cover” sets. A subset \mathcal{L} of $\binom{X}{2}$ is said to be a *triplet cover* for $T \in B(X)$ if, for every interior vertex v of T , three leaves $x, y, z \in X$ exist for which $v = \text{med}_T(x, y, z)$ and $\binom{\{x, y, z\}}{2} \subseteq \mathcal{L}$. In other words, a triplet cover for T is obtained by selecting, for each interior vertex v of T , a leaf of T from each of the three components of $T - v$, and forming all three pairs from those three leaves.

It is easily shown that any triplet cover of a binary tree T determines its edge lengths. However, the question of whether a triplet cover always determines the topology of T is much less straightforward. It is relatively easy to show that if \mathcal{L} is a triplet cover for two binary trees T and T' , then T and T' must be equivalent, but this is weaker than showing

that a triplet cover \mathcal{L} of T determines the topology of T . Nevertheless the stronger result has been established in various special cases.

Exercise⁺: Suppose that we are given \mathcal{L} , and we know that \mathcal{L} is a triplet cover for an unknown binary phylogeny T . Can one always determine whether or not $\{x, y\}$ is a cherry of T , given $d|\mathcal{L}$?

The determination story becomes much clearer if the edge lengths are constrained to be ultrametric (for both T and the tree T' in the earlier definitions), in which case the induced distances will be ultrametrics, a setting studied by Katharina Huber and colleagues. The key object is a graph $G(\mathcal{L}, v)$ defined for each interior vertex v of T and with a vertex set comprising the set E_v of directed edges of T that are outgoing edges of v , with $e, e' \in E_v$ forming an edge of $G(\mathcal{L}, v)$ precisely if there is a pair $\{a, b\} \in \mathcal{L}$ for which the (undirected) path in T connecting these two leaves uses both e and e' . The following results are from [194].

Theorem 6.9.

- (i) A nonempty subset \mathcal{L} of $\binom{X}{2}$ determines the edge lengths on a rooted tree T (assuming that these have to be ultrametric) if and only if for each interior vertex v of T , the graph $G(\mathcal{L}, v)$ contains at least one edge.
- (ii) A nonempty subset \mathcal{L} of $\binom{X}{2}$ determines the (rooted) topology of T (assuming that the edge lengths of any tree are required to be ultrametric) if and only if, for each interior vertex v of T , the graph $G(\mathcal{L}, v)$ is a clique.

It follows immediately that if \mathcal{L} determines the topology of T , then it also determines the edge lengths of T (this fails without the ultrametric edge length assumption, as noted earlier). Moreover, if T is a rooted binary X -tree, then \mathcal{L} determines the edge lengths of T if and only if \mathcal{L} determines the topology of T . For further details and extensions of the results concerning the ultrametric edge length setting, the reader is referred to [194, 191, 216].

6.3 ■ Generalizations and geometry

6.3.1 ■ Indexed pyramids and Kalmanson metrics

We have seen how the Gromov–Farris transform provides an almost seamless translation between ultrametric and tree metrics (and thereby between rooted phylogenies with ultrametric edge lengths and unrooted phylogenies with unconstrained edge lengths). This connection between the “affine” and “projective” settings extends to more general classes of metrics. Our presentation here follows a small part of an elegant story (which involves further connections involving PQ -trees and PC -trees) from Kleinman et al. [219].

A distance function δ on X is said to satisfy the *Kalmanson* property if there is a cyclic permutation (x_1, x_2, \dots, x_n) so that for all $1 \leq i < j < k < l \leq n$, the following condition holds:

$$\max\{d(x_i, x_j) + d(x_k, x_l), d(x_l, x_i) + d(x_j, x_k)\} \leq d(x_i, x_k) + d(x_j, x_l). \quad (6.10)$$

Notice that this condition is a relaxation of the four-point condition, so it holds for any tree metric. Moreover, something stronger holds in this case, as the next exercise makes clear.

Exercise: If d is a tree metric (say $d = d_{(T,l)}$), and we take any circular ordering (x_1, \dots, x_n) for T , show that condition (6.10) holds by virtue of being an equality.

Metrics that satisfy the Kalmanson property are therefore a more general class than tree metrics, and are important in the study of the famous “traveling salesman problem.” Kalmanson metrics also have a simple characterization in terms of circular split systems (cf. Section 2.4.2), as the following shows. For any X -split $A|B$, let $\delta_{A|B}$ be the distance function on X that takes the value 1 for the pair $i, j \in X$ precisely if i is in one block of the split and j is in the other, and takes the value zero otherwise.

Proposition 6.10. *A metric d on X satisfies the Kalmanson condition if and only if there is a circular split system Σ on X and a weight function $w : \Sigma \rightarrow \mathbb{R}^{>0}$ such that*

$$d = \sum_{A|B \in \Sigma} w(A|B) \delta_{A|B}.$$

Moreover, the decomposition (choice of Σ and w) is unique.

Proposition 6.10 is due to [85] (see also [219]); it provides an analogue of the classic earlier results that a metric satisfies the four-point condition if and only if the metric has a tree representation, and this representation is unique. There is now a rooted analogue of Proposition 6.10 (based again on the Gromov–Farris transform) in which weighted splits are replaced by (indexed) subsets of X that generalize hierarchies. The appropriate set system is called a *pyramid* on X . This is a collection \mathcal{P} of nonempty subsets of X containing X and $\{x\}$ for all $x \in X$ (i.e., condition **H1**) and for which

- (i) $A, B \in \mathcal{P}, A \cap B \neq \emptyset \Rightarrow A \cap B \in \mathcal{P}$;
- (ii) there is a linear ordering on X for which $A \in \mathcal{P}$ is an interval with respect to that ordering.

Notice that any hierarchy on X is also a pyramid (by ordering the leaves of the associated rooted tree from left to right in any planar drawing). A pair (\mathcal{P}, f) where \mathcal{P} is a pyramid on X and f is a function $f : \mathcal{P} \rightarrow \mathbb{R}$ that satisfies the property $A \subset B \Rightarrow f(A) < f(B)$ is called an *indexed pyramid*. The Gromov–Farris transform then provides a link between the class of weighted circular split systems on X and the set of indexed pyramids on X . The corresponding link at the metric level (i.e., the analogue of the tree metric to ultrametric link) is from Kalmanson metrics to a further class of metrics called “Robinsonian,” for further details, the reader is referred to [219].

6.3.2 ■ The geometry of tree space

For each $T \in P(X)$, let $\Lambda(T) = (0, \infty)^{|E(T)|}$, which is the set of all possible assignments of strictly positive edge lengths for T . We can think of $\Lambda(T)$ as nestled within the (tree-independent) space $(0, \infty)^{\Sigma(X)}$, where $\Sigma(X)$ is the set of X -splits. Here, the natural embedding $\iota : \Lambda(T) \rightarrow (0, \infty)^{\Sigma(X)}$ assigns $l \in \Lambda(T)$ to $\iota(l) = l_T : \Sigma(X) \rightarrow (0, \infty)$, where

$$l_T(\sigma) = \begin{cases} l(e) & \text{if } \sigma \text{ is a split of } T \text{ that corresponds to } e, \\ 0 & \text{if } \sigma \text{ is not a split of } T. \end{cases}$$

For two different trees T and $T' \in P(X)$, the image sets $\iota(\Lambda(T))$ and $\iota(\Lambda(T'))$ are disjoint, but their closures have nonempty intersection.³⁰ Now consider the (disjoint) union \mathbb{T}_X of $\iota(\Lambda(T))$ over all phylogenetic X -trees. Thus,

$$\mathbb{T}_X := \bigcup_{T \in P(X)} \iota(\Lambda(T)),$$

which can be thought of as the space of all trees with edge lengths.

Various metrics on this space are possible, such as l_1 , l_2 , and l_∞ . For example, if we take the l_∞ metric, then this distance would generalize the RF distance between T and T' (when $l(e) = 1$ and $l'(e') = 1$ for all edges e of T and e' of T'). However, it is desirable that any metric have the property that the shortest path (geodesic) linking points in \mathbb{T}_X lies entirely within \mathbb{T}_X (i.e., all points on the path correspond to any tree with edge lengths). Thus another metric has been proposed, in an important paper from 2001 [29]. The distance between two points $l_T \in \Lambda(T)$ and $l_{T'} \in \Lambda(T')$ in \mathbb{T}_X is taken to be the shortest length (under the l_2 metric) of any path that lies entirely within \mathbb{T}_X from one point to the other. Remarkably, this geometry leads to a so-called CAT(0) space (a space that has nonpositive curvature everywhere), which implies that between any two points, there is a unique shortest path (a geodesic).

This tree space geometry has subsequently been named the *BHV space* (after the authors of the original paper, Billera, Holmes, and Vogtmann [29]). Two mathematical results from that paper state that if a nontrivial split is present in trees T and T' , then all points on the geodesic linking l_T and $l_{T'}$ must also involve a tree with this split. On the other hand, if T and T' do not share any nontrivial split, then the geodesic between them passes through a point that corresponds to the star tree. However, in general, the shortest path joining $l_T \in \Lambda(T)$ and $l_{T'} \in \Lambda(T')$ may also pass through points that correspond to trees that have a split present in neither T nor T' . More recently, polynomial-time algorithms have been developed to compute shortest paths in BHV space between two trees [280], and some related centroid measures.

The topology of \mathbb{T}_X and some of its subspaces are also of interest, and they connect with some classic results involving the “shellability” of simplicial complexes. Notice that \mathbb{T}_X is contractable, since we can continuously shrink all the interior edges of each tree to zero to arrive at the star tree, and then massage all the pendant edge lengths to a fixed length 1, and so arrive at a fixed point in \mathbb{T}_X . However, if we let \mathbb{T}_X^* be the image of ι on the edge lengths of trees from $P(X)$ that have at least one interior edge, then the space is no longer contractable. Accordingly, let $\tilde{\mathbb{T}}_X$ be the subspace of \mathbb{T}_X^* in which, for every tree,

- (i) each pendant edge has length 1,
- (ii) each interior edge e has a length $l(e)$ satisfying $0 < l(e) \leq 1$, and
- (iii) the sum of the interior edge lengths equals 1.

An alternative choice³¹ to (iii) that leads to the same topology for $\tilde{\mathbb{T}}_X$ is

- (iii)' $l(e) = 1$ for at least one interior edge.

³⁰If T and T' have identical matching pendant edge lengths, then as we shrink to zero the lengths of the interior edges that give a split present in just one tree, then we approach the same point in $(0, \infty)^{\Sigma(X)}$.

³¹The space $\tilde{\mathbb{T}}_X$ described by (iii) is essentially the “link of the origin” in [29], while that described by (iii)' has been called the space of “fully grown trees.”

The point about condition (i) is that the pendant edge lengths play a trivial role in the geometry of BHV tree space, which can be viewed as a product of a space in which (i) holds, with the space $(0, \infty)^{|X|}$ (i.e., there is one free variable for every pendant edge). For $|X| = 4$, $\tilde{\mathbb{T}}_X$ consists of three isolated points; however, for $n = |X| \geq 5$, conditions (ii) and (iii) ensure that $\tilde{\mathbb{T}}_X$ is a compact connected simplicial complex of dimension $n - 4$. This complex turns out to have the property in combinatorial topology of being “shellable.” The reason for having dimension $n - 4$ is simply that the binary phylogenies have $n - 3$ interior edges (other phylogenies have fewer); however, condition (iii) (alternatively, (iii)') reduces the number of free edge parameters by one. For example, for $|X| = 5$, $\tilde{\mathbb{T}}_X$ is a connected one-dimensional simplicial complex, so it corresponds to a graph. This graph turns out to be the Petersen graph mentioned in Section 2.5.2, the line graph of which is isomorphic to $G_{\text{NNI}}(5)$.

The compact space $\tilde{\mathbb{T}}_X$ and the unbounded space \mathbb{T}_X^* that it lies in are quite different (even topologically), but they are the same through the eyes of homotopy (much as a sphere of dimension $n - 1$ has the same homotopy type as n -dimensional Euclidean space minus the origin). To see this, given the edge lengths l for a phylogeny T , with $l(e) > 0$ on all interior edges, let \tilde{l} be the modified edge lengths that (a) assign the length 1 to all pendant edges and (b) for each interior edge e , divides $l(e)$ by $\sum_{e \in \hat{E}(T)} l(e)$ in case (iii) or $\max_{e \in \hat{E}(T)} \{l(e)\}$ in case (iii)' (note that these quantities are both nonzero by assumption). With this definition, $\tilde{\mathbb{T}}_X$ is now a strong deformation retract³² of \mathbb{T}_X^* , by virtue of the following homotopy:

$$F(l_T, s) = s\tilde{l}_T + (1-s)l_T, \quad 0 \leq s \leq 1.$$

It follows that \mathbb{T}_X^* and $\tilde{\mathbb{T}}_X$ have the same homotopy type. The structure of the latter space has been investigated mathematically in various papers (e.g., [295, 359]) and has the homotopy type of a wedge of $(n - 2)!$ spheres, each of dimension $n - 4$. For example, consider the special case where $X = \{1, 2, 3, 4, 5\}$, where $\tilde{\mathbb{T}}_X$ corresponds to the Petersen graph, viewed as a one-dimensional simplicial complex (note that $n - 4 = 1$ in this case); the reason for this can be traced back to our discussion of the NNI graph on $B(5)$ (Fig. 2.5) in Chapter 2. Any finite connected graph viewed in this way is homotopic to a wedge of 1-spheres (circles), and the number of these is the cyclomatic number of the graph, which, in the case of the Petersen graph, with 15 edges and 10 vertices, is 6 ($= 15 - 10 + 1$). This is illustrated in Fig. 6.8.

Notice that the association $(T, l) \mapsto d_{(T, l)}$ provides a natural (topological) embedding of \mathbb{T}_X into the subspace $\mathcal{D}(X)$ of all distance functions on X (a closed polyhedral cone in $\mathbb{R}^{\binom{n}{2}}$ for $n = |X|$). Therefore, if we let \mathbb{D}_X be the image of \mathbb{T}_X under this map, then any consistent distance-based tree reconstruction method can be viewed as a function from $\mathcal{D}(X)$ to \mathbb{D}_X which fixes \mathbb{D}_X . We saw at the end of Section 6.2.1 that if we view NJ in this way, it is a discontinuous function. This raises the question of whether there is a continuous map from $\mathcal{D}(X)$ onto \mathbb{D}_X that fixes \mathbb{D}_X . In other words, is \mathbb{D}_X a topological retraction of $\mathcal{D}(X)$? It turns out that such retractive maps do indeed exist [271]. A further result, due to Mikhail Gromov, quantifies how close to \mathbb{D}_X an arbitrary metric $d \in \mathcal{D}(X)$ must be, based on the *hyperbolicity* of δ : if d is “close” to satisfying the four-point condition,

³²For readers less familiar with topology, saying that a subspace A is a strong deformation retract of a space B amounts to saying that we can continuously “squash up” B while keeping A fixed until all we are left with is A . For example, if B is a disc with a small hole in the center and A is the circular boundary of B , then A is a strong deformation retract of B by the process of gradually enlarging the hole out to the boundary.

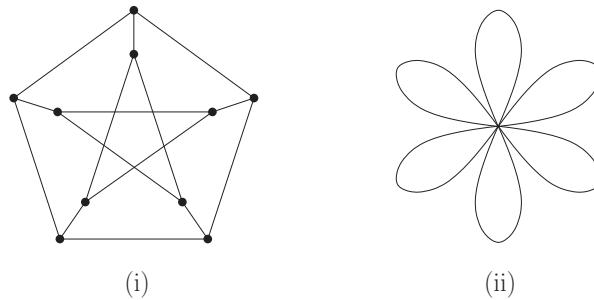


Figure 6.8. The Peterson graph (i) is homotopic to a wedge of six 1-spheres (ii).

in the sense that

$$d(x, y) + d(w, z) \leq \max\{d(x, w) + d(y, z), d(x, z) + d(y, w)\} + \epsilon,$$

for all $x, y, w, z \in X$, then $\|d - d'\|_\infty \leq \epsilon \cdot (1 + \log_2 n)$ for some tree metric $d' \in \mathbb{D}_X$. Further details and references for these last results are found in [271].

The geometry of ultrametric tree spaces (related to BHV space) has also recently become a hot topic [153, 229], where connections to tropical geometry can be usefully exploited (for details, see [229]).

In Chapter 8 (Section 8.2.2), we will consider an alternative topology to the BHV space in which (T, l) and (T', l') are close if they lead to similar probability distributions on characters under Markovian evolution.

6.4 ■ Phylogenetic diversity

As well as considering the total sum $L = L(T, l)$ of the edge lengths of a tree, we can also consider how much of this total is spanned by different subsets of leaves. This measure is called *phylogenetic diversity* (PD), a notion that is particularly relevant to biodiversity conservation [291], though it has also been applied to other areas such as genomics [284] and genotype date [386]. In biodiversity conservation, PD is used to quantify how much of the “tree of life” is spanned by a particular group of species (either in terms of evolutionary time or genetic diversity). This, in turn, can help quantify how much of this diversity might be lost if some or all of these species were to become extinct in the near future, because of the current (human-influenced) mass extinction event.

PD is also applied to provide rankings of species or sets of species, and to help formulate strategies to try to maximize future biodiversity [370]. PD can be defined for both rooted and unrooted phylogenies. We will consider the latter first, since it is more general, and includes the former as a special case.

PD for unrooted phylogenies. Formally, given an unrooted phylogenetic X -tree T and a nonnegative³³ assignment l of edge lengths, we can associate to each nonempty subset Y of X a value, denoted $PD_{(T,l)}(Y)$ —or, more briefly, just $PD(Y)$ —which is equal to the sum of the lengths of the edges of the minimal subtree of T that connect the leaves in Y . Thus $PD_{(T,l)}$ is a function from 2^X to $\mathbb{R}^{\geq 0}$.

Notice that each edge e of $T|Y$ corresponds to a set $E_Y(e)$ of edges of T that form a path in T , namely the set of edges f of T for which the associated X -split $A_f|B_f$ has the

³³For PD, we can allow edges to have zero length.

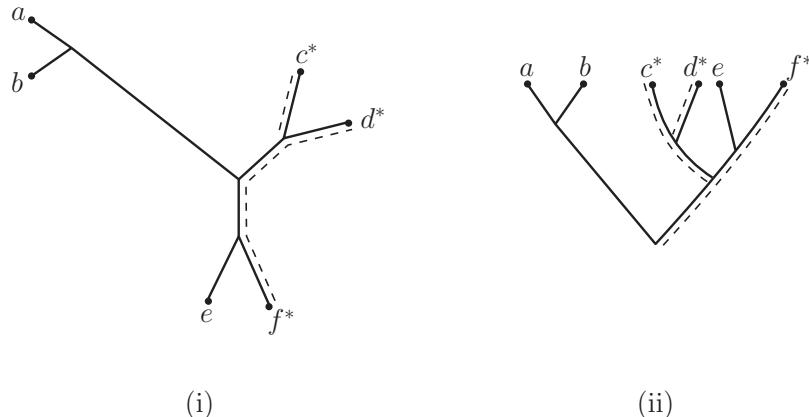


Figure 6.9. PD of $\{c, d, f\}$ as the sum of the length of the edges with dashed lines, for an unrooted phylogeny (i) and for a rooted phylogeny (ii). In this example, the extinction (i.e., removal) of either one of the species a or b (in either tree) would lead to less PD loss than the extinction of any other single species; however, the extinction of both a and b would lead to a greater loss of PD than the extinction of any other pair of species. In this example, the edge lengths are ultrametric and the trees are binary, but, in general, neither of these conditions is assumed.

property that $A_f \cap Y \neq \emptyset$ and $B_f \cap Y \neq \emptyset$. Thus we can define an edge length function l_Y for $T|Y$ by setting $l_Y(e) = \sum_{f \in E_Y(e)} l(f)$. In this way,

$$PD_{(T,l)}(Y) = \sum_{e \in E(T|Y)} l_Y(e).$$

When $\#Y \leq 1$, we have $PD(Y) = 0$; at the other end of the scale (i.e., $Y = X$), we have $PD(Y) = L$. Notice that when $\#Y = 2$ (say, $Y = \{x, y\}$) then $PD(Y) = d_{(T,l)}(x, y)$. Just as the PD scores of subsets of size $k = 2$ (i.e., tree metrics) can be used to reconstruct a tree, so can the PD scores of subsets of size k for any k up to (but not exceeding) $\lceil n/2 \rceil$ [281]. An example illustrating PD on unrooted phylogenies is shown in Fig. 6.9(i).

The function PD is monotone in the sense that $PD(Y') \leq PD(Y)$ for any nonempty subset Y' of Y . Moreover, PD enjoys the following “strong exchange” property, from [330].

Proposition 6.11. *For any $T \in P(X)$, any subset W' of X , with $|W'| \geq 2$, and any subset W of X that is larger in size than W' , there is always an element $x \in W - W'$ for which*

$$PD_{(T,l)}(W' \cup \{x\}) + PD_{(T,l)}(W - \{x\}) \geq PD_{(T,l)}(W) + PD_{(T,l)}(W'), \quad (6.11)$$

for all edge lengths $l \geq 0$.

Proof: We will write $PD(*) = PD_{(T,l)}(*)$ (noting that our choice of x in the argument that follows does not depend on l). Consider the restricted phylogeny $T|W$ in $P(W)$, which has at least three leaves (since $|W| > |W'| \geq 2$). We can regard T as being obtained from $T|W$ by attaching a set \mathcal{P} of subtrees of T to certain vertices and (subdivisions of) edges of $T|W$. Because $|W'| < |W|$ and $T|W$ has at least three exterior edges, there is at least one exterior edge e of $T|W$ with the property that e and any subtree in \mathcal{P} that attaches to e do not contain any leaf in W' . Let $x \in W - W'$ be the leaf of $T|W$ incident with e .

We then have

$$PD(W) = PD(W - \{x\}) + l_W(e), \quad (6.12)$$

where $l_W(e)$ is the sum of the lengths of the set of edges ($E_W(e)$) of T that correspond to e in $T|W$. Because e was chosen in $T|W$ so that any subtree in \mathcal{P} that contains a leaf in W' must attach either to some edge of $T|W$ that is different from e or to a vertex of $T|W$, we have

$$PD(W' \cup \{x\}) \geq PD(W') + l_W(e). \quad (6.13)$$

Combining eqns. (6.12) and (6.13) gives inequality (6.11). ■

Proposition 6.11 justifies a simple and fast strategy for finding a subset Y of X of any given size k that is guaranteed to have maximal PD for sets of that size. The strategy is simply the greedy one: first select any two leaves x, y that are furthest apart in the tree (i.e., that maximize $d_{(T,l)}(x, y)$) and then sequentially add a leaf to the tree constructed thus far that increases the PD score by the maximum amount, until k leaves are present [284]. Formally, the collections of subsets of X that have a maximal PD score for their cardinality form a set system that is referred to in combinatorics as a “greedoid.”

To see that the greedy algorithm is justified by (6.11), let PD_m , $m \geq 1$, denote the collection of subsets Y of X of size m that have maximal PD. Consider any sets $W \in PD_k$ and $W' \in PD_{k-1}$ for $k \geq 3$. For any leaf $x \in W - W'$, the definitions of PD_k and PD_{k-1} ensure that $PD(W - \{x\}) \leq PD(W')$ with equality if and only if $W - \{x\} \in PD_{k-1}$, and that $PD(W' \cup \{x\}) \leq PD(W)$ with equality if and only if $W \cup \{x\} \in PD_k$. Combining these two statements, we see that

$$PD(W' \cup \{x\}) + PD(W - \{x\}) \leq PD(W) + PD(W'), \quad (6.14)$$

with equality if and only if

$$W' \cup \{x\} \in PD_k \text{ and } W - \{x\} \in PD_{k-1}. \quad (6.15)$$

Suppose now that $x \in W - W'$ is chosen so as to satisfy inequality (6.14). Inequality (6.14) must then be an equality, so eqn. (6.15) must hold. Consequently, by the first statement in (6.15), any optimal set $W' \in PD_k$ can be extended to an optimal set $W \in PD_k$ and, by the second statement, every optimal set W can be obtained in that way. For further details, see [284] or [330].

Variations on PD have also been proposed by some biologists. One example is $\delta PD(Y) := PD(X) - PD(X - Y)$, which is how much PD is lost if the species in Y were to disappear. Finding a subset of a given size to maximize δPD is no longer guaranteed by the greedy algorithm.

Another diversity measure applies for any metric on species (i.e., not necessarily corresponding to a tree with edge lengths). For any metric d on X and for subsets Y of X of size at least 2, let

$$MD(Y) := \min\{d(x, x') : x, x' \in Y, x \neq x'\}.$$

Given positive integer k , the *maximum minimum distance* (MMD) problem is to find a subset Y of X that maximizes $MD(Y)$ among all subsets of X of size k . This problem, studied in [48], is NP-hard in general (it is closely related to a problem in classical coding theory). However, when d is a tree metric (i.e., $d = d_{(T,l)}$), the MMD problem has a polynomial-time solution, though it is more complex than the greedy algorithm. Moreover, the sets Y of size k that maximize $MD(Y)$ can be different from the sets that maximize $PD(Y)$ (on the same pair (T, l)). An interesting example to demonstrate this nonequivalence when $k = 3$ was described in [48] for an expanded version of Carl Woese's

classic 1987 small-subunit ribosomal RNA “tree of life.” In general (i.e., for any metric d), the greedy algorithm produces a solution to the MMD problem that is, at worst, at least half of the global optimal score [48].

It is also possible to extend PD to more than one tree (e.g., to the set of trees, with edge lengths, that arise in the posterior estimate of a phylogenetic tree under Bayesian tree reconstruction methods which we will discuss in Chapter 8). The following “weighted average PD on t trees” ($WAPD_t$) problem was studied in [56]. In this problem, there is a collection (T_i, l_i) , $i = 1, \dots, t$, of phylogenetic X -trees with associated edge lengths, along with an associated collection λ_i , $i = 1, \dots, t$ of nonnegative real-valued tree lengths, and a positive integer k . The $WAPD_t$ problem is to find a subset Y of X that maximizes $\sum_{i=1}^t \lambda_i PD_i(Y)$, where $PD_i(Y)$ is the phylogenetic diversity of Y on (T_i, l_i) . We may assume that $\lambda_i = 1$ for all i by simply replacing the edge lengths described by l_i for T_i by $\lambda_i \cdot l_i$. For $t = 1$, this problem is equivalent to the classic PD optimization problem described above, so it has a linear time in ($n = |X|$) solution via the greedy algorithm. For $t > 2$, the problem has been shown to be NP-hard. This leaves the case $t = 2$, which turns out to have an elegant polynomial-time ($O(n^3 \log^2 n)$) algorithm based on the theory of network flows in an edge-weighted graph derived from the two trees. For details, see [56].

Further extensions of PD for unrooted phylogenies are also possible. A particularly general setting considers an (arbitrary) collection Σ of X -splits and a weighting function $w : \Sigma \rightarrow \mathbb{R}^{>0}$. For any subset Y of X , the *PD of Y on Σ* can be defined as

$$PD_{(\Sigma, w)}(Y) = \sum_{\substack{A|B \in \Sigma: \\ A \cap Y, B \cap Y \neq \emptyset}} w(A|B).$$

When Σ is the set of splits of a phylogenetic X -tree and $w(A|B)$ is the length of the edge that corresponds to the split $A|B$, this is just the familiar notion of PD in phylogenetics (more generally, if Σ is the union of the splits of two or more trees, then $PD_{(\Sigma, w)}$ describes the setting in the previous paragraph). In general, finding a subset Y of X of size k to maximize $PD_{(\Sigma, w)}$ is NP-hard. However, a polynomial-time algorithm exists for finding a set of size k that has a PD score that is at least $(1 - 1/e) \approx 0.632$ times the globally optimal value, and this approximation is the best possible unless $P = NP$ [52].

Exercise: Given (T, l) suppose we wish to select a subset Y of X of size k containing a given subset Y_0 of X and so that Y maximizes $PD_{(T, l)}(Y) + \sum_{y \in Y} w(y)$ for some nonnegative weight function w on X . Show how this problem can also be solved via a greedy algorithm.

6.4.1 • PD optimization and diversity indices for rooted trees

So far, we have considered PD on unrooted phylogenies. Moving to rooted trees, we face a choice. We could define $PD(Y)$ to be the sum of the edge lengths of the minimal tree connecting Y , or we could insist that this minimal tree also connect the root. We will adopt the latter notion for several reasons: (i) it is the one usually adopted by biologists, (ii) it leads to simpler and nicer mathematical results, and (iii) there is a natural link to the unrooted definition. Regarding this last point, notice that if we add another “outgroup” leaf η adjacent to the root of a tree, and put zero length on this new edge, then the PD score of Y on the original rooted phylogeny T equals the PD score of $Y \cup \{\eta\}$ on the associated unrooted phylogeny $T^{+\eta}$.

Accordingly, for a rooted tree $T \in RP(X)$ with the edge lengths l and any subset Y of X , we let

$$PD_{(T,l)}(Y) = \sum_{e \in E(T): c_T(e) \cap Y \neq \emptyset} l(e), \quad (6.16)$$

where $c_T(e)$ is the set of leaves of T descended from e . An example illustrating the notion of PD for rooted phylogenies is shown in Fig. 6.9(ii).

Notice that PD on rooted trees also satisfies the strong exchange property (eqn. (6.11)), since, as noted already, $PD(Y)$ on a rooted tree T equals $PD(Y)$ on the associated unrooted tree $T^{+\eta}$, and any element $x \in W - W'$ (in eqn. (6.11)) cannot be the leaf η of $T^{+\eta}$. It follows that the greedy algorithm for finding an optimal PD set among all choices of k leaves still applies. Indeed, the beginning step is easier: rather than having to find two leaves at a maximal distance, we start by simply finding a leaf that is at a maximal distance from the root. In this rooted setting, the MMD and PD maximization problems, which seemed somewhat unconnected in the previous setting, turn out to coincide for a rooted tree in which the edge lengths are ultrametric (i.e., the distance from the root to each leaf is the same), as the following result from [48] reveals.

Proposition 6.12. *Suppose that $T \in RP(X)$ has ultrametric edge lengths, and $d = d_{(T,l)}$ is the associated ultrametric. Then the greedy algorithm for MMD finds an optimal solution Y of size k . This solution also maximises PD over all subsets Y of X of size k .*

PD on rooted trees enjoys a further combinatorial property that is absent in the unrooted setting. Recall that a function $f : 2^X \rightarrow \mathbb{R}^{\geq 0}$ is said to be *submodular* if it satisfies the property that $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$, for all subsets A and B of X .

Proposition 6.13. *PD on rooted trees is a submodular function on 2^X .*

The proof uses an elementary result, which will also be useful later.

Lemma 6.14. *Consider any linear function $y(\mathbf{x}) = \sum_{i=1}^m c_i x_i$, where $c_i \in \mathbb{R}$ and $\mathbf{x} = (x_1, \dots, x_m)$ is a sequence of variables. Then $y(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in (\mathbb{R}^{\geq 0})^m$ if and only if $y(\mathbf{x}) \geq 0$ for all the m choices of \mathbf{x} in which x_j is zero for all but one value of j (say, i) and $x_i = 1$. The same applies if $y \geq 0$ is replaced by $y = 0$ throughout.*

The “only if” direction of this lemma is trivial; the “if” direction follows by observing that $y = \sum_{i=1}^m y_i x_i$, where y_i is the value of y when $x_i = 1$ and $x_j = 0$ for all $j \neq i$.

Proof of Proposition 6.13: Notice that

$$y_{A,B} := PD(A) + PD(B) - PD(A \cup B) - PD(A \cap B)$$

is a linear function of l on a fixed phylogeny T . By Lemma 6.14, submodularity holds if and only if, for each edge e of T , $y_{A,B} \geq 0$ when $l(e) = 1$ and $l(e') = 0$ for all $e' \neq e$. In that case $PD(Y)$ is either 1 or 0 depending on whether or not $c_T(e)$ contains a leaf of Y , so $y_{A,B} = 0$ except when $c_T(e)$ contains a leaf of both A and B and does not contain a leaf of $A \cap B$, in which case, $y_{A,B} = 1$. ■

Submodularity can fail for PD on unrooted trees, for example, if A and B are disjoint singleton sets, or if the subtree connecting A is separated from the subtree connecting B by a sufficiently long path.

We have seen that finding a subset of the leaf set of a given size to maximize PD has a simple solution via the greedy algorithm. However, conservation managers may instead focus on conserving certain “regions,” each of which will contain a subset of species. This raises the question of which regions to conserve to maximize the PD of the species that lie in at least one conserved region. Here, “conserve” means ensuring that none of the species in the region becomes extinct (e.g., by converting the region into a nature reserve). Of course, if cost were no issue, it would always be optimal to conserve all the regions; but when a choice has to be made, the problem becomes less trivial mathematically (not to mention politically).

More formally, suppose we have the following:

- a rooted phylogenetic X -tree T with nonnegative edge lengths l ;
- a collection \mathcal{A} of subsets of X (each set in \mathcal{A} corresponds to the subset of species present within some region);
- a nonnegative integer cost function c on the sets in \mathcal{A} (the cost of conserving the associated region);
- a total budget B .

The *budgeted nature reserve problem* (BNRP) seeks a subset \mathcal{A}' of \mathcal{A} that maximizes the PD score of the set $\bigcup_{A \in \mathcal{A}'} A$ on T so that $\sum_{A \in \mathcal{A}'} c(A) \leq B$. In other words, the problem aims to select regions to conserve, within the allowed budget, so that the PD score of the set of all species found in one or more selected regions is maximal. This problem was formalized and studied in [51]. Although the problem is NP-hard, it turns out that the submodular property of (rooted) PD described above allows for a greedy $(1 - 1/e) \approx 0.632$ approximation algorithm for BNRP. Moreover, this can be extended to several rooted trees with associated edge lengths or, more generally, to weighted split systems [52]. A helpful technique for such problems is to invoke a general result in combinatorial optimization theory [351]. Roughly speaking, this result states that given (i) a function $f : 2^X \rightarrow \mathbb{R}^{>0}$ that is nondecreasing (i.e., $f(A) \leq f(A')$ whenever $A \subseteq A'$) and submodular, (ii) a cost function $c : X \rightarrow \mathbb{Z}^{>0}$, and (iii) a fixed positive value B , it is possible to find a subset A of X for which $f(A)$ is at least a $(1 - 1/e) \approx 0.632$ fraction of the maximal possible value of f over all subsets of X that satisfy $\sum_{a \in A} c(a) \leq B$.

Diversity indices. We now consider ways in which the total length of a tree can be apportioned additively to the species at the leaves of the tree. One motivation for this is that conservation managers typically wish to rank species by taking into account how much they “contribute” to total (phylogenetic) diversity. To make this more precise, we are interested here in associating to each tree $T \in R(X)$ a function $\psi_T : X \rightarrow \mathbb{R}$ for which

$$\sum_{x \in X} \psi_T(x) = PD_{(T,l)}(X). \quad (6.17)$$

We will refer to any such function ψ_T as a *PD index*. Notice that the right-hand side of (6.17) is a linear function of l , so it seems reasonable to restrict our attention to functions ψ_T that are also linear functions of l , and thus can be written

$$\psi_T(x) = \sum_{e \in E(T)} c_{T,e}(x)l(e), \quad (6.18)$$

for coefficients $c_{T,e}(x)$ that do not depend on l . In this case, we say that ψ_T is a *linear* PD index.

What choices are possible for these coefficients ($c_{T,e}(x)$) in order to obtain a valid PD index (i.e., satisfying (6.17) for all l)? This has an easy answer. If we substitute eqn. (6.18) into eqn. (6.17) and interchange the order of summation, it is clear that the coefficients satisfy these two equations for all l if and only if

$$\sum_{x \in X} c_{T,e}(x) = 1 \text{ for each edge } e \text{ of } T. \quad (6.19)$$

Exercise: Prove that eqn. (6.19) is both necessary and sufficient for eqns. (6.17) and (6.18) to hold for all $l \geq 0$.

Notice also that $c_{T,e}(x)$ is the value of $\psi_T(x)$ for the tree on which e has length 1, while all other edges of T have length 0. This will be useful shortly for showing that two apparently different PD indices are equivalent. We now introduce two simple diversity indices.

Fair proportion: For each leaf $x \in X$, let

$$FP_T(x) = \sum_{e \in P(T; \rho, x)} \frac{1}{|c_T(e)|} l(e)$$

(recall that $c_T(e)$ is the set of leaves of T descended from e). The function FP_T is referred to as the *fair proportion* (FP) index or sometimes also as *evolutionary distinctiveness* (ED). Essentially, the length of each edge is distributed equally among the leaves that are descendants of that edge. An example of its application to a real phylogeny is shown in Fig. 6.10. In the context of eqn. (6.18), we have

$$c_{T,e}(x) = \begin{cases} \frac{1}{|c_T(e)|} & \text{if } e \in P(T; \rho, x), \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that eqn. (6.19) is satisfied, so FP_T is a valid PD index. Figure 6.11(i) shows how it is possible for a leaf with a maximal FP score to lie on a shortest pendant edge, even when the tree is binary and the edge lengths are ultrametric.

Equal splits: For each leaf $x \in X$, let

$$ES_T(x) = \sum_{e \in P(T; \rho, x)} \frac{1}{\Pi(e, x)} l(e),$$

where $\Pi(e, x) = 1$ if e is a pendant edge; otherwise, if $e = (u, v)$, then $\Pi(e, x)$ is the product of the outdegrees of the interior vertices of T on the directed path from v to leaf x . For example, if T is binary and there are k interior edges separating e and x , then $\Pi(e, x) = 2^k$. The function ES_T is referred to as the *equal splits* (ES) index. Essentially, the length of each edge is evenly distributed at each branching point (regardless of how many leaves are in each subtree). It is different from the FP index, and an example illustrating this shown in Fig. 6.11(i). Equal splits is a valid PD index (i.e., $\sum_x c_{T,e}(x) = 1$ for every edge e of T) by virtue of the identity

$$\sum_{x \in c_T(e)} \frac{1}{\Pi(e, x)} = 1.$$

This identity can be proved directly by induction (it also follows from the generalized Pauplin formula (eqns. (6.8) and 6.9)).

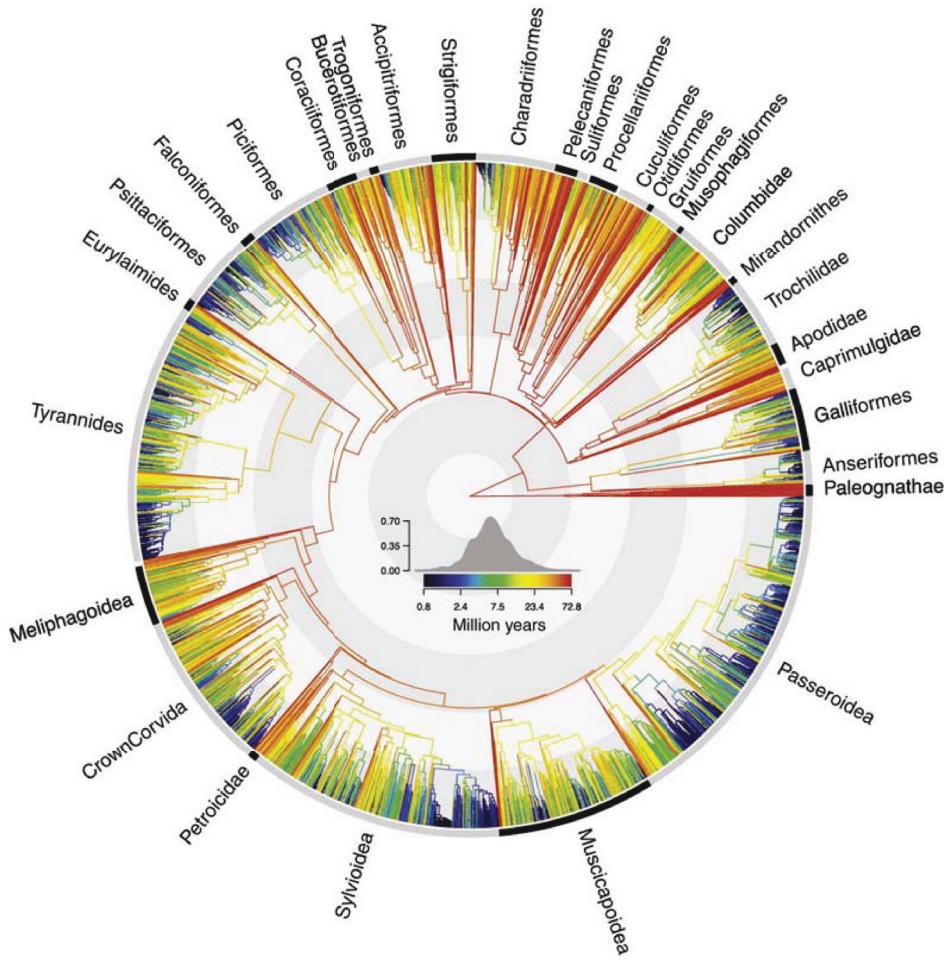


Figure 6.10. A phylogenetic tree of nearly all the (approximately) 10000 species of birds, in which the edge lengths correspond to time. The FP index of species is indicated via colors. The loss of 575 species identified as “imperiled” would constitute a total loss of approximately 2.7 billion years of evolution as measured by PD. Reprinted with permission from Elsevier [207].

Notice that the index $\psi = \text{FP}$ satisfies the following continuity condition: If e is an interior edge of a phylogenetic tree and T/e is the tree obtained from T by collapsing edge e , then

$$\lim_{l(e) \rightarrow 0} \psi_T(a) = \psi_{T/e}(a).$$

However, taking $\psi = \text{ES}$, this continuity condition can fail, even for binary trees with ultrametric edge lengths. A counterexample is provided in Fig. 6.11(iii), where $\lim_{\epsilon \rightarrow 0} E\text{S}_T(c) = 1 + \frac{1}{2} \neq 1 + \frac{1}{3} = E\text{S}_{T/e}(c)$.

The Shapley value. A quite different type of index was introduced by [171], which applied an earlier concept from cooperative game theory to phylogenetics. This paper discussed PD in the context of unrooted trees (where, as we have seen, there is a slight difference in how the PD of subsets is defined). Following [146], we define it first for rooted trees, as this allows us to state the main theorem of this section.

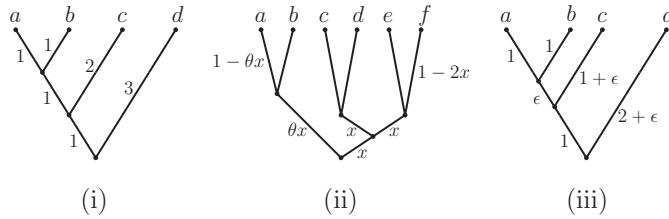


Figure 6.11. (i) For this tree T , with all interior edge lengths equal to 1, leaf a has $FP_T(a) = 1 + \frac{1}{2} + \frac{1}{3}$ and $ES_T(a) = 1 + \frac{1}{2} + \frac{1}{4}$. (ii) For the tree shown, leaves a and b tie for the largest FP (or ES) score when $\theta \in (2, 2 + \frac{1}{2})$, even though both a and b lie on the shortest pendant edges. Notice that the edge lengths in these trees are ultrametric. (iii) A tree in which equal splits exhibits a discontinuity as $\epsilon \rightarrow 0$.

Given a subset S of X and an element $x \in S$, let

$$\Delta_T(S, x) := PD(S) - PD(S - \{x\}).$$

In other words, $\Delta_T(S, x)$ describes how much PD is lost if leaf x is removed from set S (i.e., if the species x becomes extinct but the remaining species in S do not).

Now suppose we select “at random” a subset S of X containing leaf x . We need to be clear what “at random” means here: it does not mean uniformly at random (i.e., by taking each possible subset with equal probability). Instead we first select a size K for the set S by choosing a number between 1 and n uniformly at random. We then select a subset S of X that contains x of size K uniformly at random from the collection of all such sets. The *Shapley value* of x , denoted $SV_T(x)$, is the expected value of $\Delta_T(S, x)$. By the law of total expectation, we have

$$SV_T(x) = \mathbb{E}[\Delta_T(S, x)] = \mathbb{E}[\mathbb{E}[\Delta_T(S, x)|K]] = \frac{1}{n} \sum_{k=1}^n \mathbb{E}[\Delta_T(S, x)|K=k],$$

where

$$\mathbb{E}[\Delta_T(S, x)|K=k] = \frac{1}{\binom{n-1}{k-1}} \sum_{S \subseteq X : x \in S, |S|=k} \Delta_T(S, x).$$

The term in front of the summation arises because there are $\binom{n-1}{k-1}$ ways to select S so that $x \in S$ and $|S| = k$. Combining the last two equations leads to the following expression for the Shapley index [171]:

$$SV_T(x) = \frac{1}{n!} \sum_{S, x \in S} (|S|-1)!(n-|S|)! \Delta_T(S, x).$$

There is an interesting and not immediately obvious connection between the Shapley value (for rooted trees) and FP: they are exactly identical. This remarkable result was only proven recently by [146]. We establish the formal equivalence with a slightly more direct argument than that from [146], starting with the following lemma.

Lemma 6.15. $\frac{1}{n!} \sum_{j=0}^{n-r} \binom{n-r}{j} j!(n-1-j)! = \frac{1}{r}$, for $1 \leq r \leq n$.

Proof: We can rewrite the expression on the left-hand side as

$$\frac{(n-r)!(r-1)!}{n!} \times \sum_{j=0}^{n-r} \binom{n-1-j}{r-1} = \frac{(n-r)!(r-1)!}{n!} \times \binom{n}{r} = \frac{1}{r},$$

where the first equality is from the standard identity $\sum_{j=0}^{a-b} \binom{a-j}{b} = \binom{a+1}{b+1}$, applied to $a = n-1, b = r-1$. ■

Theorem 6.16. *For any rooted phylogenetic X -tree T , $FP_T(x) = SV_T(x)$ for all $x \in X$.*

Proof. Since $FP_T(x)$ and $SV_T(x)$ are both linear functions of l , so too is $FP_T(x) - SV_T(x)$. Therefore, by Lemma 6.14, it suffices to show that $FP_T(x) - SV_T(x) = 0$ when any given edge (say, e) has length 1 and all other edges have zero length. Clearly, such a tree

$$FP_T(x) = \begin{cases} \frac{1}{|c_T(e)|} & \text{if } e \in P(T; \rho, x), \\ 0 & \text{otherwise.} \end{cases} \quad (6.20)$$

Now consider $SV_T(x)$, again with $l(e) = 1, l(f) = 0$ for all $f \neq e$. Observe that the term $\Delta_T(S, x)$ is always zero if e is not in the path from ρ to x and that therefore $SV_T(x) = 0$ in that case also. On the other hand, if $e \in P(T; \rho, x)$, then $\Delta_T(S, x)$ is 1 if (i) $x \in S$ and (ii) S contains no other leaf that is a descendant of e ; the difference is 0 otherwise. Thus when $e \in P(T; \rho, x)$, the quantity $SV_T(x)$ can be written as

$$\frac{1}{n!} \sum_{\substack{S: x \in S \\ S - \{x\} \subseteq X - c_T(e)}} (|S| - 1)! (n - |S|)! \cdot 1 = \frac{1}{n!} \sum_{k=1}^{n - |c_T(e)| + 1} \binom{n - |c_T(e)|}{k-1} (k-1)! (n-k)!.$$

If we now apply Lemma 6.15 (with $j = k-1, r = |c_T(e)|$) the last expression equals $\frac{1}{|c_T(e)|}$. Comparing these two cases for $SV_T(x)$ with (6.20), we see that $SV_T(x) = FP_T(x)$ for the tree that has $l(e) = 1$ and $l(f) = 0$ for all $f \neq e$. Since this holds for all choices of e , $SV_T(x) - FP_T(x) = 0$ by Lemma 6.14. ■

One can also define a Shapley value on the unrooted tree obtained from the rooted tree T by suppressing the root vertex. However, the PD score of a set Y changes when we unroot the tree (since the PD of a subset may no longer include contributions from edges that led to the root), so the resulting Shapley value is not expected to agree with the version on the original tree. In [176], Klaas Hartmann showed that this Shapley value on unrooted trees, while different from FP on the original rooted tree, is nevertheless closely correlated with it. The associations between FP and various versions of the Shapley value are explored further in [373], while the computation and application of these and other associated biodiversity indices based on a phylogeny are described further in [241].

6.4.2 • Biodiversity conservation (“Noah’s Ark”)

Given a rooted phylogenetic X -tree T with edge lengths, suppose that a random set of species $\mathcal{X} \subseteq X$ is selected according to some stochastic process that models extinction in the near future. For example, if X is a set of species extant at present, then \mathcal{X} might be the subset of species that are still extant, say, 100 years from now. In that case, by eqn. (6.16), the PD score of \mathcal{X} on T is a random variable that can be written in the following form:

$$PD(\mathcal{X}) = \sum_{e \in E(T)} l(e) \cdot \mathbb{I}_{\mathcal{X} \cap c_T(e) \neq \emptyset}, \text{ and so}$$

$$\mathbb{E}[PD(\mathcal{X})] = \sum_e l(e) \cdot \mathbb{P}(\mathcal{X} \cap c_T(e) \neq \emptyset).$$

In the (generalized) *field of bullets* model, \mathcal{X} is obtained by independently including each element $x \in X$ in \mathcal{X} with some survival probability s_x .³⁴ By the (strong) independence assumption, $\mathbb{P}(\mathcal{X} \cap c_T(e) \neq \emptyset) = 1 - \prod_{x \in c_T(e)} (1 - s_x)$, and therefore

$$\mathbb{E}[PD(\mathcal{X})] = \sum_e l(e) \left(1 - \prod_{x \in c_T(e)} (1 - s_x) \right). \quad (6.21)$$

For the *simple field of bullets model*, which corresponds to the special case where $s_x = s$ for all $x \in X$, let $\psi(s) = \mathbb{E}[PD(\mathcal{X})]$. It then follows from eqn. (6.21) that $\frac{d\psi}{ds} \geq 0$ and $\frac{d^2\psi}{ds^2} \leq 0$ for all $s \in [0, 1]$. Thus ψ is an increasing concave function of s .

Exercise: For the simple field of bullets model, show that (i) the slope of the curve $\psi(s)$ as s approaches 1 is equal to the sum of the lengths of the pendant edges of T , and (ii) $\frac{d^2\psi}{ds^2} = 0$ for some $s \in [0, 1]$ if and only if there is no interior edge of T with a strictly positive length (in which case, $\psi(s)$ is linear in s).

For the simple field of bullets model, an alternative way to write eqn. (6.21), from [279], is

$$\mathbb{E}[PD(\mathcal{X})] = \sum_{k \geq 1} S(k)(1 - (1 - s)^k), \quad (6.22)$$

where $S(k)$ is the sum of lengths of those edges that have exactly k descendant leaves.

Suppose now that in the simple field of bullets model the survival probability is a function of time, so $s = s(t)$ for some (necessarily nonincreasing) function of t . For example, if extinction is modeled by a constant-rate pure-death process, then $s = e^{-rt}$ for some $r > 0$. In this case, elementary calculus shows that $\psi(s)$ is a decreasing and convex function of time. If we now move to the more realistic model of the generalized field of bullets model, where s_x is allowed to be variable (for example, let us suppose that $s_x = e^{-r_x t}$ for $r_x > 0$), then $\mathbb{E}[PD(\mathcal{X})]$ is still a decreasing function of t . However, $\mathbb{E}[PD(\mathcal{X})]$ can exhibit more complex behavior for the second derivative; for example, it can transition from concave to convex with increasing t .

Expected PD can also be used to assign indices of conservation value to species based on proposed extinction models. For example, for a species $x \in X$, the quantity

$$\mathbb{E}[PD(\mathcal{X} \cup \{x\}) - PD(\mathcal{X} - \{x\})]$$

describes how much extra PD is expected to be present if species x survives some future extinction event than if it disappears. Such indices are easy to calculate under the field of bullet models [241].

Regarding the actual distribution of the random variable $PD(\mathcal{X})$, it is possible to compute its variance and, indeed, its full distribution under the (generalized) field of bullets model. Also, provided that the survival probabilities of the species are all bounded away from 0 and 1, and under assumptions that exclude extreme distributions of edge lengths in the underlying tree, it can also be shown that $PD(\mathcal{X})$ has a limiting Gaussian distribution as $n = |X|$ grows (for details, see [132]).

The field of bullets models is, of course, very simple in that it assumes that extinction events are independent between species. For more complex models, in which survival

³⁴Survival probabilities in conservation biology have sometimes been estimated by using the International Union for Conservation of Nature (IUCN) classification scheme for endangered species.

probabilities depend on traits (characters) that evolve in the tree by models like those discussed in Chapter 7, the expected loss of PD is always greater than or equal to the corresponding value under a generalized field of bullets model with the same marginal probability for survival of each species [133]. Models that incorporate ecological dependencies (e.g., food web structures) have also been studied in modeling PD loss.

The Noah’s Ark problem. The field of bullets model also leads to various optimization questions, in particular the *Noah’s Ark problem* (NAP). Introduced by Martin Weitzman in 1998, the problem is, broadly speaking, how to allocate resources so as to maximize future expected PD subject to limited resources. More precisely, for a (rooted) X -phylogenetic T with edge lengths, suppose that each species $x \in X$ has an estimated probability a_x of being extant at a certain time in the future, and that this probability can be increased to b_x given a certain intervention cost c_x to conserve this species.

Suppose the total budget B is fixed. The problem is to determine which set $S \subseteq X$ of species should be conserved within the allowed budget B (i.e., so that $\sum_{x \in S} c_x \leq B$) so as to maximize the expected PD score of the subset \mathcal{X} of species from X that survive the extinction event. From eqn. (6.21), we have

$$\mathbb{E}[PD(\mathcal{X})|S] = \sum_e l(e) \left(1 - \prod_{x \in c_T(e) \cap S} (1 - b_x) \prod_{y \in c_T(e) - S} (1 - a_y) \right).$$

This optimization problem is referred to as the $a_x \xrightarrow{c_x} b_x$ NAP. Efficient solutions have been obtained for special cases (i.e., placing restrictions on either the parameters a_x, b_x, c_x , or the tree, or its edge lengths). Notice that the simplest version, where $a_x = 0, b_x = c_x = 1$ for all x (i.e., the $0 \xrightarrow{1} 1$ NAP) is just the original simple PD optimization problem we considered at the start of this chapter, which is solved by the greedy algorithm.

Although the general problem $a_x \xrightarrow{c_x} b_x$ NAP is computationally difficult, there is nevertheless a polynomial-time approximation scheme for solving it (i.e., for each $\epsilon > 0$, there is a polynomial-time algorithm that returns a solution which scores at least $1 - \epsilon$ times the score of the maximal solution; for details, see [188]).

6.4.3 ■ Extensions of PD

We end this chapter by describing two further ways in which the PD concept has been extended and abstracted to areas both within and beyond phylogenetics.

PD over Abelian groups. Distances and diversities also have a clear mathematical meaning if the edge lengths of a tree take nonzero real values (possibly negative) or, more generally, values in an arbitrary Abelian group \mathcal{G} different from the identity element $1_{\mathcal{G}}$. In these cases, $PD(Y)$ is simply the sum (in the group) of the \mathcal{G} -lengths of the edges that lie in the minimal subtree of T that connect the leaves in Y .³⁵

Provided that the Abelian group \mathcal{G} has no elements of order 2, the classic uniqueness and existence results from earlier in this chapter concerning tree-based metrics still hold in almost identical form. More precisely,

- (i) a tree with edge length values in \mathcal{G} (not equal to $1_{\mathcal{G}}$) are uniquely determined by the induced function $d : X \times X \rightarrow \mathcal{G}, (x, y) \mapsto PD(\{x, y\})$, and

³⁵The notion can even be extended further to non-Abelian groups, but the lack of commutativity requires us to distinguish $d(x, y)$ from $d(y, x)$ (for details, see [315]).

- (ii) an arbitrary symmetric function $d : X \times X \rightarrow \mathcal{G}$ can be represented on a phylogeny with the edges taking length values in \mathcal{G} if and only if d satisfies the following natural analogue of the four-point condition: For every $x, y, w, z \in X$ (not all distinct), two of the three partial sums

$$d(x, y) + d(w, z), \quad d(x, z) + d(y, w), \quad \text{and} \quad d(x, w) + d(y, z)$$

are equal, and this common sum is equal to the third plus $2g$ for some element $g \in \mathcal{G}$ (dependent on x, y, w, z).

However, there is a “fly in the ointment” for both statements (i) and (ii) when \mathcal{G} has elements of order 2. First, the uniqueness of the tree representation in statement (i) fails. To see why, consider the 15 perfect binary trees in $B(6)$ with each edge assigned the non-identity element 1 of $\mathcal{G} = (\{0, 1\}, +)$. In this case, $d(x, y) = 0$ for all $x, y \in \{1, 2, 3, 4, 5, 6\}$, so none of these 15 trees can be distinguished from each other. Nevertheless, uniqueness can be restored by moving from distances to diversities, where not just pairs but also triples of leaves are considered. Similarly, for statement (ii), the existence of a tree representation of an arbitrary $d : X \times X \rightarrow \mathcal{G}$, when \mathcal{G} has elements of order 2, can be characterized by an extension of the four-point condition that involves diversities on pairs and triplets, along with a five-point condition (for further details, see [121]).

Abstract diversity theory. Some of the features of PD have been abstracted well beyond phylogenetics to areas of pure mathematics (such as metric geometry and T-theory) by David Bryant and Paul Tupper (see, e.g., [71]). These authors refer to a *diversity* as a pair (Y, δ) where Y is a (possibly infinite) set and δ is a function from the finite subsets of Y to \mathbb{R} satisfying just two axioms:

- $\delta(A) \geq 0$, and $\delta(A) = 0$ if and only if $|A| \leq 1$;
- if $B \neq \emptyset$, then $\delta(A \cup C) \leq \delta(A \cup B) + \delta(B \cup C)$,

for all finite $A, B, C \subseteq Y$. From these two axioms it can be shown that δ is monotonic (i.e., $A \subseteq B$ then $\delta(A) \leq \delta(B)$) and that the induced function $\delta' : Y \times Y \rightarrow \mathbb{R}$ defined by $\delta'(y, y') = \delta(\{y, y'\})$ is a metric on Y . A variety of further results can be derived and a structured theory built just from the two axioms stated.

It is easily seen that PD can be regarded as a diversity in this sense if we take $Y = X$ and, for a tree $T \in P(X)$ with weighted edges, we let $\delta(A) = PD(A)$. Abstract diversities that can be realized in this way can be characterized (for details, see [71]). A second example of diversities that are relevant to phylogenetics is the following. Let (X, d) be a metric space. For each finite subset A of X , let $\delta(A)$ denote the minimum length of a Steiner tree within X connecting elements in A . In this case, (X, δ) is also a diversity. Diversities also appear in several other settings. For example, given any metric space (X, d) and any finite subset A of X consider the “diameter” function $\delta(A) = \max\{d(a, a') : a, a' \in A\}$. Once again, (X, δ) is a diversity.

In [71], the authors extend some classical theory (that describes how a metric space embeds naturally into a minimal “hyperconvex” space) from metric spaces to the more general notion of diversities, and thereby establish a number of new and far-reaching results.

Chapter 7

Evolution on a tree: Part one

A major advance in phylogenetics has been the development of stochastic models to describe the evolution of discrete characters as they evolve along the branches of an evolutionary tree from some unknown ancestral state. For genetic sequences, stochastic models typically describe point substitutions that occur at sites in the DNA sequence that codes for some particular gene. Models for protein sequences and structure, and gene order on chromosomes, have also been developed.

Using statistically based techniques, biologists convert the genetic sequences we observe today at the leaves of the tree into an estimate of the phylogenetic tree that describes the evolution of the gene (and perhaps the tree's edge lengths, or ancestral states within the tree). By combining these "gene trees," one can, in turn, estimate the "species tree" (a topic we will discuss in Chapter 9). Stochastic models also allow biologists to study the reliability of tree reconstruction methods—both those described in earlier chapters, and more stochastically based methods described in this chapter and the next.

The rise of "statistical phylogenetics" was pioneered in the 1960s and 70s by Anthony Edwards, Joseph Felsenstein, Jerzy Neyman, and others (including David Sankoff, with a visionary early paper [308]). Stochastic models of character evolution typically assume that characters evolve independently on a tree and that the evolution of each character is described by some Markovian process, or by a mixture of such processes (e.g., to allow some characters to evolve more rapidly than others). Given a sequence of characters on X as data, *maximum likelihood* (ML) estimation asks for the phylogenetic X -tree T that maximizes the probability of this data, assuming that each character has been generated independently according to a fixed stochastic model on T (usually, this involves also maximizing other continuous parameters associated with the model).

One of the catalysts that ushered in this stochastic approach was a landmark 1978 paper by Joseph Felsenstein [135]. He showed that if characters evolve independently under a simple stochastic process, then primitive character-based methods like maximum parsimony (MP, discussed in Chapter 5) or distance-based methods using (uncorrected) Hamming distances (discussed in Chapter 6) can be seriously misled. Therefore, as the number of characters increases, it becomes increasingly certain that MP tree will differ from the "true" tree (i.e., the one on which the characters evolved). By contrast, other methods (like ML) are, under certain conditions, provably statistically consistent; in other words, they converge on the true tree as the number of characters grows.

Today, most phylogenetic tree reconstruction methods rely on some underlying stochastic process to model the evolution of characters. Using such models, simple empirical

distances (such as the normalized Hamming distance mentioned in Section 6.1.3) can be transformed so that methods like NJ applied to these transformed distances are (statistically) consistent estimators of phylogenies and branch lengths. Although such distance-based methods are fast, more sophisticated approaches based on maximum likelihood and Bayesian estimation are now widely used, with programs such as RAxML and PhyML (for maximum likelihood), and MrBayes and BEAST (for Bayesian methods) being currently popular. Model selection criteria (such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)) are also increasingly being used, and programs such as ModelTest are widely used for selecting an appropriate process of character evolution based on such criteria.

7.1 ■ Nonhomogeneous Markov chains

Markov processes on trees are an extension of a class of stochastic process that are widely used throughout science: the Markov chain. Normally, these are introduced in the setting of homogeneous processes (i.e., with a fixed transition matrix). Here, we begin by providing a tailored crash course in nonhomogeneous Markov process theory in the setting of a discrete state space. We will emphasize aspects of relevance to phylogenetics. Indeed, many results here apply directly to Markov processes on trees, since a path in a tree can be viewed as a Markov chain consisting of the states at the vertices along that path.

Let Y_0, Y_1, \dots, Y_m be a sequence of random variables, each taking values in a set S of discrete states. We say that $(Y_i; i = 0, \dots, m)$ is a (nonhomogeneous) *Markov chain* if for each $i > 0$, Y_{i+1} is conditionally independent of Y_0, \dots, Y_{i-1} given Y_i . In other words, the present “screens off” the future from the past.

In this chapter, we will assume that the state space S is finite and will denote $\#S$ by r . Also, by way of notation, we will henceforth write $[x_\alpha]$ to denote a row vector of length r indexed by the elements α of S , and $[x_{\alpha\beta}]$ to denote an $r \times r$ matrix with the rows indexed by $\alpha \in S$ and the columns indexed by $\beta \in S$.

Let $\pi = [\pi_\alpha]$ describe the probability distribution of states of Y_0 (i.e., $\pi_\alpha = \mathbb{P}(Y_0 = \alpha)$ for each state $\alpha \in S$). We will let $P^{(i)} = [P_{\alpha\beta}^{(i)}]$ denote the $r \times r$ *transition matrix* of conditional probabilities:

$$P_{\alpha\beta}^{(i)} := \mathbb{P}(Y_{i+1} = \beta | Y_i = \alpha).$$

Notice that each row of $P^{(i)}$ sums to 1.

We will assume henceforth that for each i , the matrix $P = P^{(i)}$ is strictly positive (i.e., $P_{\alpha\beta} > 0$ for all $\alpha, \beta \in S$) and that π is also strictly positive (i.e., $\pi_\alpha > 0$ for all $\alpha \in S$, which we write more compactly as $\pi > 0$). These assumptions imply that the “reversed” stochastic process $Y_m, Y_{m-1}, \dots, Y_1, Y_0$ is also a Markov chain.

Connection with linear algebra. Since each row of $P^{(i)}$ sums to 1, we have $P\mathbf{1} = \mathbf{1}$ for the column vector $\mathbf{1}$ consisting entirely of 1s, and so $P^{(i)}$ has an eigenvalue of 1. Any transition matrix P has $-1 \leq \det(P) \leq 1$, and the transition matrices that realize the extreme values -1 and +1 are precisely the “permutation matrices,” which are obtained from the identity matrix by permuting its rows. Thus, since we are assuming that the entries of $P^{(i)}$ are strictly positive, the determinant of P , $\det(P)$, lies strictly between -1 and +1. In addition, Perron–Frobenius theory tells us that all the eigenvalues of P apart from 1 have a modulus less than 1 and that there is a unique distribution $\tilde{\pi}$ for

which $\tilde{\pi}P = \tilde{\pi}$, which is called the *stationary distribution* for P . It is also clear from the assumption on P that $\tilde{\pi}$ is strictly positive (i.e., $\tilde{\pi}_\alpha > 0$ for all states α). Notice that if all matrices in the chain have the same stationary distribution and this equals the distribution π of Y_0 , then π is the marginal distribution of every random variable Y_i in the chain.

The joint distribution of any initial sequence of the Markov chain Y_0, Y_1, \dots, Y_m is entirely determined by the initial state π together with the matrices $P^{(0)}, \dots, P^{(m-1)}$. The formal justification of this claim is via the product rule in elementary probability theory: For any sequence of events A_0, A_1, \dots, A_m , the probability that all these events occur is

$$\mathbb{P}(A_0)\mathbb{P}(A_1|A_0)\mathbb{P}(A_2|A_0 \wedge A_1) \cdots \mathbb{P}(A_m | \wedge_{i=0}^{m-1} A_i). \quad (7.1)$$

Therefore, if we let A_i be the event that $Y_i = y_i$, then the conditional probability of A_i , given the earlier events, simplifies to $\mathbb{P}(A_i | A_{i-1}) = P_{y_{i-1}, y_i}^{(i-1)}$. As a corollary, the marginal distribution of Y_m is also determined just by π and the transition matrices. By the law of total probability, the probability that $Y_m = \beta$ is given by

$$\mathbb{P}(Y_m = \beta) = \sum_{\substack{F: \{0, \dots, m\} \rightarrow S \\ F(m) = \beta}} \pi_{F(0)} \prod_{i=0}^{m-1} P_{F(i)F(i+1)}^{(i)}. \quad (7.2)$$

This formidable-looking expression becomes much tamer when viewed through the lens of elementary linear algebra: the row vector $[\mathbb{P}(Y_m = \beta)]$ is obtained by multiplying row vector π by the product of the transition matrices in the given ordering. In other words, the marginal probability distribution of Y_m is given by the row vector $\pi P^{(0)} P^{(1)} \cdots P^{(m-1)}$. Similarly, the m -step transition matrix $P^{(0,m)} = [\mathbb{P}(Y_m = \beta | Y_0 = \alpha)]$ is given by the matrix product $P^{(0)} P^{(1)} \cdots P^{(m-1)}$.

Example: Consider the simple model involving just two states and symmetric transition probabilities between them. In this case,

$$P^{(i)} = \begin{bmatrix} 1-p_i & p_i \\ p_i & 1-p_i \end{bmatrix}.$$

This process can be viewed as one that flips between two states with a probability p_i that may vary between time steps. This process has the uniform stationary distribution $\pi = [0.5, 0.5]$. The m -step transition matrix $P^{(0,m)}$ is also symmetric (being a product of symmetric 2×2 matrices), and its off-diagonal entry $p^{(m)}$ is given by

$$p^{(m)} = \frac{1}{2} \left(1 - \prod_{i=0}^{m-1} (1 - 2p_i) \right). \quad (7.3)$$

Exercise: Establish eqn. (7.3), either (i) by induction, or (ii) by noticing that the matrices $P^{(i)}$ are diagonalized by the same matrix, or (iii) by computing the matrix determinant of both sides of the equation $P^{(0,m)} = P^{(0)} P^{(1)} \cdots P^{(m-1)}$.

For the symmetric 2-state model, eqn. (7.3) reveals that if p_i is bounded away from 0 and 1, then $p^{(m)}$ converges to $\frac{1}{2}$ as m grows. Additionally, the random number N of state changes over the first m steps of the chain has expected value $\mathbb{E}[N] = \sum_{i=0}^{m-1} p_i$, and $Y_m \neq Y_0$ if and only if N is odd. It might therefore be suspected that if m is chosen so that $\mathbb{E}[N]$ is equal (or close) to an odd number, then $\mathbb{P}(Y_m = Y_0) > \frac{1}{2}$. However, this need not be the case; notice that when $p_i < \frac{1}{2}$ for all i , eqn. (7.3) shows that $p^{(m)} < \frac{1}{2}$ for all m . For example, suppose we flip a light switch according to the roll of a fair die, with a flip occurring every time the number 6 comes up. If we perform six rolls of the die, the expected number of flips is $6 \times \frac{1}{6} = 1$, but the light is still more likely than not to be in the same state as it was before the experiment.

Markov processes that can be described in continuous time. A transition matrix P is frequently represented as the net effect of a continuous-time Markovian process operating for a fixed duration $t > 0$. In this setting, a *rate matrix* is an $r \times r$ matrix Q with nonnegative real off-diagonal entries and with each row summing to zero. Under a continuous-time Markov process, the rate matrix Q operates for duration $t \geq 0$. The resulting transition matrix $P(t)$ is the solution to the linear system $dP(t)/dt = P(t)Q$, subject to the initial condition $P(0) = I$. Consequently,

$$P(t) = \exp(Qt), \quad (7.4)$$

where \exp denotes the exponential of a matrix (i.e., $\exp(Qt) = I + \sum_{i \geq 1} Q^i t^i / i!$). When Q can be diagonalized, then $P(t)$ can be computed by replacing the eigenvalues in the diagonal matrix by their exponentiated values.

A transition matrix P in a Markov chain has a *continuous realization* if $P = \exp(Qt)$ for some rate matrix Q and value $t > 0$. In this case, notice that $\det P > 0$, by Jacobi's identity, which states that for any matrix A ,

$$\det \exp(A) = e^{\text{tr}(A)}, \quad (7.5)$$

where “tr” refers to the trace of the matrix. Consequently, not every transition matrix has a continuous realization, for example, if $\det P \leq 0$.

Equation (7.4) has some redundancy in that any rate matrix Q can be multiplied by a scaling factor f to get a rate matrix $Q' = fQ$ for which $Q'(t/f) = Qt$. So if eqn.(7.4) holds for some rate matrix, we can assume it holds either for $t = 1$ or (which is more usual) for a rate matrix that has been scaled so that the expected rate of substitutions per unit time is 1.

Notice that if $\pi Q = 0$, then $\pi P = \pi$ (since $\pi \exp(Qt) = \pi(I + \sum_{i \geq 1} Q^i t^i / i!) = \pi I + 0 = \pi$), so π is a stationary distribution for P . In this case, if a state is selected according to this distribution π and the rate matrix Q acts for duration t , then the expected number of substitutions μ (often called the *evolutionary distance* in molecular phylogenetics) is proportional to time but is scaled by the trace of a matrix determined by Q as follows:

$$\mu = \sum_{\alpha \in S} \pi_\alpha \sum_{\beta \neq \alpha} q_{\alpha\beta} t = -\text{tr}(\text{diag}(\pi)Q)t, \quad (7.6)$$

where $\text{diag}(\pi)$ is the diagonal matrix with the entries of π on the diagonal.

Example: 2-state model (continued). Continuing our 2-state symmetric example, each transition matrix $P^{(i)}$ has a continuous realization for the fixed matrix

$$Q = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Writing $P^{(i)} = \exp(Q t_i)$ gives $p_i = \frac{1}{2}(1 - \exp(-2t_i))$, and so $t_i = -\frac{1}{2}\ln(1 - 2p_i)$. By eqn. (7.6), the expected number of substitutions when Q acts for time t_i is t_i (i.e., “duration” corresponds to the expected number of state changes). In addition, $p_i \sim t_i$ as $t_i \rightarrow 0$. However, since p_i is a concave function of t_i with an asymptote at $\frac{1}{2}$ as $t_i \rightarrow \infty$, p_i behaves very differently from t_i as t_i becomes large.

The function $t \mapsto P(t) = \exp(Qt)$ is one-to-one. This follows from eqn. (7.5) which gives the identity: $\det P(t) = \det \exp(Qt) = \exp(\text{tr}(Q)t)$, and taking the natural logarithm of both sides gives

$$t = \frac{-\ln \det(P(t))}{-\text{tr}(Q)}. \quad (7.7)$$

Thus for a fixed Q , the function $-\ln \det(P)$ provides a natural time scale for a Markov process realized by Q . More generally, for any Markov chain, the log-determinant function $\varphi(P) = -\log |\det(P)|$ provides a nonnegative natural time scale in the sense that

$$\varphi(P^{(0,m)}) = \varphi(P^{(0)}) + \cdots + \varphi(P^{(m-1)}),$$

provided that $\det P^{(i)} \neq 0$ for every i . See [160] for further discussion and extensions of this idea.

The requirement that P is strictly positive can still allow for Q to have zero entries, as follows. Given a rate matrix Q , consider the directed graph G_Q which has the set S of states as its vertex set and a directed edge (α, β) provided that $Q_{\alpha\beta} > 0$. The Markov process described by Q is said to be *irreducible* if the resulting digraph is strongly connected (i.e., from any one state, it is possible to reach any other state by following some directed path in G_Q). Clearly, this holds if $Q_{\alpha\beta} > 0$ for all distinct states $\alpha, \beta \in S$ but it can also hold when many entries in Q are zero. When Q is irreducible, all entries in $P(t) = \exp(Qt)$ are strictly positive, for all $t > 0$.

An example of an irreducible process in which Q has some zero entries is a simple formulation of the *covariation model* for DNA (or protein sequence) evolution, which involves a Markov process on an extended state space $\tilde{S} = S \times \{0, 1\}$. Here, transitions on $S \times \{1\}$ proceed according to some continuous-time irreducible Markov process on S ; transitions from $(s, 0) \rightarrow (s, 1)$ and $(s, 1) \rightarrow (s, 0)$ proceed according to an independent fixed 2-state continuous-time Markov process; and all other transitions have zero rate. This models a situation where a site that is “on” (i.e., $(s, 1)$) at some point is free to change to a different state according to Q or to turn “off” (i.e., change to $(s, 0)$); sites that are “off” (i.e., $(s, 0)$) can only stay in this state or turn “on” (i.e., change to $(s, 1)$). For this model, the associated $2|S| \times 2|S|$ rate matrix Q has zero entries, but it is irreducible, and so $P(t) = \exp(Qt)$, for $t > 0$, has all its entries strictly positive.

Stationary and time-reversible processes. A Markov chain Y_i in which each transition matrix has a continuous realization based on a fixed rate matrix Q and which has a stationary distribution π (i.e., $\pi Q = 0$) is said to be *stationary*. If, in addition, the

condition

$$\pi_\alpha q_{\alpha\beta} = \pi_\beta q_{\beta\alpha} \quad (7.8)$$

holds for all distinct states α and β in S , then the chain is (*time*) *reversible*. This condition can also be stated more algebraically as the condition that $\text{diag}(\pi)Q$ is a symmetric matrix. The reversibility condition implies that

$$\pi_\alpha P_{\alpha\beta}^{(i)} = \pi_\beta P_{\beta\alpha}^{(i)}$$

from the series expansion of $\exp(Qt)$ (this equality also allows for a definition of reversibility for a discrete, rather than continuous-time, process). Reversibility simply says that if we run the chain in the reverse direction, then we can use the same transition matrix P to compute this joint distribution. The reversibility of Q also has another well-known algebraic characterization due to Kolmogorov: Q gives rise to a reversible process if and only if the product of transition rates around any cycle of states (not necessarily distinct) of length 3 or more is the same, regardless of whether the cycle is traversed in a clockwise or counterclockwise direction. It follows that all 2-state stationary Markov processes are reversible.

Reversible Markov processes have several attractive mathematical properties, stemming from spectral decomposition theory. For example, in a time-reversible process, all the eigenvalues of Q are real, so the transition probabilities of $P_{\alpha\beta}(t)$ of the transition matrix $P = \exp(Qt)$ can be written as

$$P_{\alpha\beta}(t) = \pi_\beta + \sum_{i=1}^{k-1} c_{\alpha\beta}^{(i)} e^{-\lambda_i t}, \quad (7.9)$$

where $-\lambda_i < 0$ are the nonzero eigenvalues of Q , and $c_{\alpha\beta}^{(i)}$ are constants determined by the eigenvectors of Q . Moreover, when $\alpha = \beta$, these c constants are all nonnegative, so $P_{\alpha\alpha}(t)$ is a monotone decreasing function of t for any reversible process and any state α .³⁶

It is clear from eqn. (7.8) that the sum, average, or, more generally, convex linear combination of any collection of time-reversible rate matrices, each of which has the same fixed stationary distribution π , is also a time-reversible rate matrix with π as its stationary distribution.

The concept of time-reversibility also applies to Markov chains generally (regardless of whether or not the transition matrices have a continuous realization). If π is a stationary distribution for each transition matrix $P^{(i)}$ and $\text{diag}(\pi)P^{(i)} = P^{(i)}\text{diag}(\pi)$, then the chain is said to be time-reversible. This means that the reversed stochastic process Y_m, Y_{m-1}, \dots, Y_0 —which is also a Markov chain—has a transition matrix for the step from Y_{i+1} to Y_i that is the same matrix as that for the forward step from Y_i to Y_{i+1} .

Exercise: If $(Y_i; i = 0, \dots, m)$ is a stationary Markov chain, use Bayes' formula to write the conditional probability $\mathbb{P}(Y_j = \beta | Y_{j+1} = \alpha)$ of the reversed process $(Y_{m-i}; i = 0, \dots, m)$ in terms of the entries of $P^{(j)}$ and π .

³⁶However, for certain reversible Markov processes it is also possible for $P_{\alpha\alpha}(t) < P_{\alpha\beta}(t)$ for certain values of t and states $\alpha \neq \beta$.

7.2 ■ From Markov chains to processes on trees

It is relatively straightforward to extend the notion of a Markov chain to a Markov process on a rooted tree T . Again, we will assume throughout that $\pi > 0$, and that each edge $e = (u, v)$ of T has an associated transition matrix $P^{(e)} = [P_{\alpha\beta}^{(e)}] = [\mathbb{P}(Y_v = \beta | Y_u = \alpha)]$ with all its entries strictly positive.

Suppose that we place a total order (\leq) on the vertices of T that is compatible with the descendancy ordering (i.e., if $u \preceq_T v$, then $u \leq v$). For example, we could view \leq as arising from some ordering on a time scale, though this is not necessary. Then the collection of random variables $(Y_v; v \in V(T))$ forms a *Markov process* on T if, for each edge (v, w) of T , Y_v renders Y_w conditionally independent of the (“earlier”) random variables $(Y_u; u < v)$. The comparison with Markov chains is illustrated in Fig. 7.1.

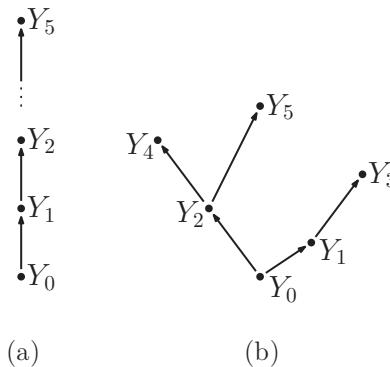


Figure 7.1. (a) A Markov chain and (b) a Markov process on a tree. In both cases if we condition on the state Y_i ($i > 0$) at a vertex, then the state at a child of that vertex becomes conditionally independent of the “earlier” states Y_0, \dots, Y_{i-1} .

This definition might seem to depend on the choice of the total order (\leq). However, we will see shortly that if $(Y_v; v \in V(T))$ is a Markov process on T for some total ordering compatible with the descendancy ordering \leq_T , then the same is true for any other such total ordering. To see why, apply eqn. (7.1) to the setting where the events (A_i) refer to the states of vertices ordered according to the total order \leq . Then we can express the joint probability distribution of $(Y_v; v \in V(T))$ purely in terms of the transition matrices associated with the edges of T and the distribution of states π at the root vertex. More explicitly, if $F : V(T) \rightarrow S$ is an assignment of states to the vertices of T , then applying the product rule (eqn. (7.1)), we obtain

$$\mathbb{P}(\bigwedge_{v \in V(T)} \{Y_v = F(v)\}) = \pi_{F(v_0)} \prod_{e=(u,v) \in E(T)} P_{F(u)F(v)}^{(e)}. \quad (7.10)$$

It follows that for any character $f : X \rightarrow S$, if T has leaf set X , the probability $p(f)$ that the leaves in X are in the states specified by f is given by

$$p(f) = \sum_{\substack{F: V(T) \rightarrow S, \\ F|X=f}} \pi_{F(v_0)} \prod_{e=(u,v)} P_{F(u)F(v)}^{(e)}. \quad (7.11)$$

Notice that this reduces to eqn. (7.2) in the special case where T is a linear tree on the vertices $0, \dots, m$.

Equation (7.11) does not look particularly friendly for calculations, since it involves a summation over the set of all functions from X to S , and this set grows exponentially with $n = |X|$ for $|S| > 1$. The same situation occurred for a Markov chain, where matrix multiplication provided us with an easy way to calculate these probabilities in polynomial time. For a Markov process on a tree, a similar trick is possible, based on dynamical programming that starts at the leaves and works progressively towards the root. The key is to keep a running tally of the probabilities $p(f; v, \alpha)$ of generating the character that f specifies at the leaves descended from v given that $Y_v = \alpha$.

Exercise: If vertex v has outgoing edges, $e_1 = (v, w_1), \dots, e_m = (v, w_m)$ in T , write an equation for $p(f; v, \alpha)$ in terms of the $m \times |S|$ values $p(f; w_i, \alpha_i)$ ($\alpha_i \in S$) and the entries of the m transition matrices $P^{(e_i)}$ for $i = 1, \dots, m$.

We saw that a Markov chain $(Y_i; i = 1, \dots, m)$ remains a Markov chain if we consider the reversed sequence $(Y_{m-i}; i = 1, \dots, m)$. A similar situation applies for a Markov process on a tree. We can reroot the tree at any vertex and redirect the edges away from this vertex, and the original process will still be a Markov process on T with this new orientation (the transition matrix associated with each edge may change unless the process is stationary and time-reversible).

Thus, although we have defined a Markov process on a tree using a rooted tree, it can also be viewed as an unrooted process. The following lemma provides a case in point, where the rooting of the tree and orientation of edges is irrelevant: If we specify the state at any interior vertex of a tree T , then the states within the subtrees that attach to any interior vertex v become conditionally independent between the subtrees (cf. Fig. 7.2(i)).³⁷

Lemma 7.1. Suppose that $(Y_v; v \in V(T))$ forms a Markov process on a rooted tree T . Let v be an interior vertex of T and consider the connected components of $T - v$. If V_1, V_2, \dots, V_k are the vertex sets of these components, and $W_i = (Y_v; v \in V_i)$ is the associated collection of random variables, then W_1, W_2, \dots, W_k are conditionally independent, given Y_v .

Proof. First reroot the tree on v (which may change the transition matrices) and then apply eqn. (7.10). ■

When no further assumptions are placed on a Markov process on a tree, we refer to this as the *general Markov model* (GMM). If \mathcal{M} is the GMM or any submodel of it, we let $p_{(T, \theta)}$ denote the probability distribution on characters $f : X \rightarrow S$ evolving on T with the associated parameters θ . For GMM, θ corresponds to the entries in the transition matrices on the edges of T along with the distribution π at some particular vertex. However, in submodels, fewer parameters are required, typically just one for each edge length of the tree, together with perhaps some additional parameters that specify further aspects of the model.

We will denote the probability of any particular character $f : X \rightarrow S$ by $p_{(T, \theta)}(f)$ or $p(f|T, \theta)$, or just by $p(f)$ when this is clear. Before we consider particular models, we will describe a generic property of the GMM arising from linear algebra.

³⁷Another way to define a Markov process on an unrooted tree T is as a special case of a “Markov random field,” which requires the state at each vertex v , conditional on the states at the immediate neighbors of v , to be independent of the states at the other vertices of T .

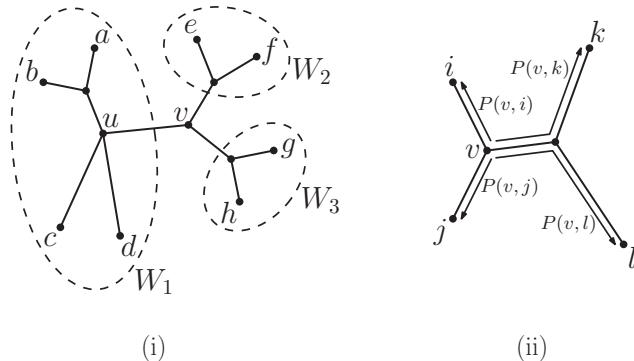


Figure 7.2. (i) By specifying the state at the vertex v , the random variables W_1 , W_2 , and W_3 (the states at the vertices of the three subtrees comprising $T - v$) become conditionally independent; (ii) the transition probabilities in the derivation of eqn. (7.13).

7.2.1 • Identifiability of the phylogeny

A Markov process on a phylogenetic tree T determines a probability distribution p on characters. For tree reconstruction, we need to ask the inverse question. Namely, does p determine the phylogeny? Surprisingly, it does, up to the placement of the root, under the mild assumption that none of the transition matrices are singular (recall that we are also assuming throughout that π and the transition matrices are all strictly positive).

To see why, consider the joint probability distribution for each pair of leaves. For leaves i and j , let J^{ij} be the $r \times r$ matrix, with rows and columns indexed by the states of S and with $J^{ij} = [J_{\alpha\beta}^{ij}]$, where $J_{\alpha\beta}^{ij} = \mathbb{P}(Y_i = \alpha, Y_j = \beta)$. Notice that if v is any interior vertex on the path between i and j in T , then we can write

$$J^{ij} = P(v, i)^T D_v P(v, j), \quad (7.12)$$

where $P(v, w)$ (for $w = i$ and $w = j$) is the matrix with the $\alpha\beta$ entry $\mathbb{P}(Y_w = \beta | Y_v = \alpha)$, and where D_v is the diagonal matrix with the α entry $\mathbb{P}(Y_v = \alpha)$. To see why eqn. (7.12) holds, simply observe that Y_v renders Y_i and Y_j conditionally independent, and so

$$\mathbb{P}(Y_i = \alpha \wedge Y_j = \beta) = \sum_{\gamma \in S} \mathbb{P}(Y_i = \alpha | Y_v = \gamma) \mathbb{P}(Y_j = \beta | Y_v = \gamma) \mathbb{P}(Y_v = \gamma).$$

Repackaged in the language of matrix algebra, this corresponds to eqn. (7.12). Consequently, if for leaves i, j, k , and l of T , the restricted tree $T|\{i, j, k, l\}$ is either the quartet tree $ij|kl$ or the star tree, and we select v to be (say) the median vertex $\text{med}_T(i, j, k)$ (see Fig. 7.2(ii)), then $J^{ik} J^{jl}$ and $J^{il} J^{jk}$ involve the product of exactly the same four matrices $P(v, w)$ ($w \in \{i, j, k, l\}$) and the matrix D_V , just in a different order (cf. Fig. 7.2(ii)). Consequently, provided the transition matrices have nonzero determinants, then we can let $\delta(x, y) = -\ln |\det J^{xy}|$ (i.e., this quantity is finite) and obtain

$$\delta(i, k) + \delta(j, l) = \delta(i, l) + \delta(j, k). \quad (7.13)$$

Moreover, it can also be shown that when $T|\{i, j, k, l\} = ij|kl$, the third pairwise sum ($\delta(i, j) + \delta(k, l)$) is strictly less than the other two [329]. In other words, if we extend δ to a distance function on $[n]$ by setting $\delta(i, i) = 0$ for each leaf $i \in [n]$, then δ satisfies the

four-point condition for T and thus has a tree representation on T with strictly positive edge lengths. Consequently, from Chapter 6 (eqn. (6.1)) the pairwise joint distributions J^{ij} suffice to determine the phylogeny T uniquely up to the placement of the root (i.e., $T^{-\rho}$ can be recovered from δ). In practice, other methods (like ML) are used for analyzing genetic sequence data; however, establishing that the tree is uniquely determined by the probability distribution on characters ensures that methods like ML will be statistically consistent and converge on the correct tree as the number of characters grows (we study this further in Chapter 8).

If we impose slightly stronger constraints on the transition matrices for T , then these matrices can also be recovered from p [83]. Curiously though, it is no longer always enough to consider the joint pairwise distributions (i.e., the J^{ij} matrices); instead, the joint distribution at **triples** of leaves is generally required.

7.2.2 • Models in molecular phylogenetics

In molecular phylogenetics, Markovian models are most widely applied to aligned DNA sequence data, where $S = \{A, C, G, T\}$, which is the set of nucleotide bases. The models are also applied to larger state spaces also, such as protein sequences where S consists of the 20 amino acids, or to model “codon” evolution where S is the set of $4^3 = 64$ triples of nucleotide bases.

In all these settings, most models start by assuming a stationary and time-reversible continuous-time Markov process on a tree. Without any further restriction, this model is often referred to as the *general time-reversible (GTR) model*. For DNA sequence data, where $|S| = 4$, the rate matrix Q for this model has nine parameters (or six if π is fixed); however, if we scale Q so that the expected substitution rate ($-\text{tr}(\Pi Q)$) equals 1, then this reduces the number of parameters by 1. Often, this model is constrained further to reflect underlying symmetries in the substitution pattern of DNA. The simplest model is the one-parameter *Jukes–Cantor model* (JC69) in which all the substitution rates are assumed to be equal. This is the 4-state analogue of the 2-state symmetric model that we considered above. Although this model leads to some tractable and interesting mathematical results, it is usually too constrained for most biological data.

Exercise⁺: For the GTR model on r states, show that the $r \times r$ matrix J^{ij} (the joint probability of states at i and j) is a symmetric matrix for each pair of leaves i, j .

A slightly more general process is Kimura’s two-substitution type model (from 1980) which is referred to here as the *K2ST model*. This allows a common rate for “transitions” ($A \rightarrow G$ and $G \rightarrow A$ (purine to purine); $C \rightarrow T$ and $T \rightarrow C$ (pyrimidine to pyrimidine)) and a possibly different but common rate for all the other eight state changes (“transversions”). Kimura’s three-substitution type model (*K3ST model*) relaxes this further and allows separate rates for two types of transversions.

Each rate matrix of these classes (JC69, K2ST, and K3ST) has the uniform distribution $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ as its stationary distribution. This may be appropriate for modeling certain sequence data, but for others the frequency of the bases deviates significantly from a uniform distribution (e.g., in 18S rRNA, bird and mammal sequences are rich in G and C). In general, models that do not describe the distribution of nucleotide bases for a set of species tend to perform poorly when used in methods to infer a phylogenetic tree for those species.

A plethora of other models are used in molecular phylogenetics, and we will not try to describe all of them here (but we will mention some additional ones shortly that have desirable mathematical properties). In all these models, transitions between states are typically referred to as “substitution” events. Such an event is a mutation at a DNA site that spreads through the population by genetic drift or selection and becomes “fixed” (though it may subsequently be followed by a similar event).

The edge lengths in the tree are also usually allowed to vary freely; however, sometimes they are constrained to be ultrametric. This applies, for example, if the length of an edge corresponds to time, and there is assumed to be a constant rate of substitution across the tree in a stationary model (this constant-rate assumption is often called the *molecular clock* hypothesis). Without a molecular clock constraint on the edge lengths, different placements of the root in the tree can lead to indistinguishable probability distributions on characters under time-reversible Markov processes. In that case, other methods are required to locate the root in phylogeny reconstructed from character data using a model, such as the inclusion of an “outgroup taxon” (i.e., a species s so distant from those under study that the root is likely to lie on the pendant edge incident with s).

7.3 • Classes and properties of models

Let \mathcal{M} be a class of transition matrices. If the condition that $P, P' \in \mathcal{M} \implies PP' \in \mathcal{M}$, holds, then \mathcal{M} is said to be *multiplicatively closed*.³⁸ This condition is relevant in molecular phylogenetics when the substitution process of evolution may change through time or when additional species are included or excluded from study. Of course, the class of all transition matrices is multiplicatively closed, as are the simple subclasses obtained by appending one or both of the following conditions: $\det(P) \neq 0$ or $\det(P) > 0$. A particularly relevant multiplicatively closed class arises from any fixed rate matrix Q by letting $\mathcal{M}_Q = \{\exp(Qt) : t \geq 0\}$. Thus \mathcal{M}_Q is the set of transition matrices that have a continuous realization based on Q . It is easy to see that \mathcal{M}_Q is closed, since $\exp(Qt)\exp(Qt') = \exp(Q(t+t'))$.

More generally, suppose that \mathcal{Q} is a set of commuting rate matrices. In other words, for all $Q, Q' \in \mathcal{Q}$, we have $QQ' = Q'Q$ or, equivalently,

$$[Q, Q'] := QQ' - Q'Q = 0.$$

When this holds, the *Baker–Campbell–Hausdorff (BCH) formula*³⁹ (or a direct linear algebra argument) shows that

$$\exp(Q_1 t_1) \exp(Q_2 t_2) = \exp(Q_1 t_1 + Q_2 t_2).$$

The matrix on the right of this last equation has a continuous realization, by taking $Q = \frac{t_1}{t_1+t_2} Q_1 + \frac{t_2}{t_1+t_2} Q_2$ and $t = t_1 + t_2$. Thus if \mathcal{Q} is a collection of rate matrices that is closed under addition, and multiplication by nonnegative reals, and which satisfies $[Q, Q'] = 0$ for all $Q, Q' \in \mathcal{Q}$, then the associated set $\mathcal{M}_{\mathcal{Q}} = \{\exp(Qt) : Q \in \mathcal{Q}, t > 0\}$ of transition matrices is multiplicatively closed. Several models in molecular phylogenetics satisfy the commutativity property $[Q, Q'] = 0$, such as JC69, K2ST, and K3ST (even when the transition and transversion rates vary between the matrices).

³⁸In other words, \mathcal{M} and the identity matrix form a semigroup.

³⁹This is an identity of the form $\exp(A)\exp(B) = \exp(C)$ for $C = A + B + \frac{1}{2}[A, B] + \frac{1}{12}([A, [A, B]] + [B, [B, A]] - \frac{1}{24}[B, [A, [A, B]]]) + \dots$.

Commutativity is a strong condition to require. One can ask if a weaker condition suffices for closure. First, the set of all transition matrices that have a continuous realization is **not** multiplicatively closed. That is, the product $\exp(Q_1 t_1) \exp(Q_2 t_2)$ cannot in general be written as $\exp(\overline{Q}(t_1 + t_2))$ for a valid rate matrix \overline{Q} , even when $t_1 = t_2 = 1$. More precisely, though it may be possible to write $\exp(Q_1) \exp(Q_2) = \exp(\overline{Q})$ for some matrix \overline{Q} (using the BCH formula) this matrix may contain some off-diagonal entries that are negative, and thus it is not a valid rate matrix.

Thus continuous realization is too general a class for closure, but what if we just consider the subclass of rate matrices that are time-reversible? This has particular relevance for molecular phylogenetics because of increasing interest in models that allow the substitution process (i.e., the matrix Q) to vary across the tree. It turns out that for the class \mathcal{Q}_{rev} of time-reversible rate matrices, $\mathcal{M}_{\mathcal{Q}_{\text{rev}}}$ fails to be closed; moreover, the same applies if \mathcal{Q}_{rev} is replaced by the class $\mathcal{Q}_{\text{rev}}[\pi]$ of time-reversible rate matrices that have a fixed stationary distribution π [347]. This nonclosure can have important implications for phylogenetic estimation of edge lengths and substitution rates, particularly the sensitivity of estimates to the inclusion or exclusion of unrelated species (for details, see [348]).

The search and classification of more general classes of rate matrices that lead to closed models has therefore concentrated on classifying the linear spaces \mathcal{Q} generated by sets of rate matrices that satisfy the property of a *Lie algebra*:

$$Q, Q' \in \mathcal{Q} \Rightarrow [Q, Q'] \in \mathcal{Q}.$$

This has led to a concise and elegant classification of allowable models in the four-state case of most interest in molecular systematics (i.e., $S = \{A, C, G, T\}$); for details, see [138, 347, 377]. The simplest noncommuting model that forms a Lie algebra is a model that we will consider in the next section (the “equal input model”), but extended slightly to allow π to vary between transition matrices.

7.3.1 • The equal input model

We now consider a class of models that generalizes a number of simpler models, but also has particularly tractable mathematical properties. Suppose we have a Markov process on a rooted tree with state space S , and with a distribution π of states at a root vertex v_0 . The process is said to be an *equal input model* if each transition matrix $P^{(e)}$ satisfies

$$P_{\alpha\beta}^{(e)} = \pi_\beta \cdot \theta_e, \quad (7.14)$$

for some $\theta_e \in [0, 1]$, and for all states $\alpha, \beta \in S$ with $\alpha \neq \beta$.

The defining property of the model is that the probability of a transition from α to β (two distinct states) is the same, regardless of the initial state α ($\neq \beta$). In other words, all off-diagonal entries in each column of $P^{(e)}$ are equal. In the case $r = 4$ with S being the four nucleotide bases, the model is known as the “Felsenstein 1981 model” (F81), which includes JC69 as a special case when π is the uniform distribution. More generally, for a state space S of arbitrary size r , if π is uniform, the model is called the *r-state fully symmetric model*⁴⁰—here all of the off-diagonal entries in the transition matrix (or rate matrix) are equal—and we will study this in the special case of $r = 2$ in the next section.

⁴⁰This is also sometimes referred to in other fields as the Potts model, and was studied by Jerzy Neyman in molecular evolution in 1971. For $r = 2$ we will just call it the 2-state symmetric model.

Lemma 7.2. *The following properties hold for the equal input model:*

- (i) $P_{\alpha\alpha}^{(e)} = 1 - \theta_e + \pi_\alpha \theta_e$.
- (ii) π is a stationary distribution for each vertex v of the T .
- (iii) The process is time-reversible (i.e., for each edge e , $\pi_\alpha P_{\alpha\beta}^{(e)} = \pi_\beta P_{\beta\alpha}^{(e)}$).
- (iv) The process is multiplicatively closed. Specifically, for $\alpha \neq \beta$, $(P^{(e)} P^{(e')})_{\alpha\beta} = \pi_\beta \theta$, for $\theta = 1 - (1 - \theta_e)(1 - \theta_{e'})$.

Proof: For (i), $P_{\alpha\alpha}^{(e)} = 1 - \sum_{\beta \neq \alpha} P_{\alpha\beta}^{(e)} = 1 - \theta_e \sum_{\beta \neq \alpha} \pi_\beta = 1 - \theta_e (1 - \pi_\alpha)$. For (ii), it suffices to show that if (u, v) is a directed edge and u has stationary distribution π , then v does too. We have

$$\mathbb{P}(Y(v) = \beta) = \sum_{\gamma} \pi_\gamma P_{\gamma\beta}^{(e)} = \pi_\beta P_{\beta\beta}^{(e)} + \sum_{\gamma \neq \beta} \pi_\gamma P_{\gamma\beta}^{(e)} = \pi_\beta,$$

where the last equality uses part (i). For (iii) the result clearly holds if $\alpha = \beta$, so suppose $\alpha \neq \beta$. Then

$$\pi_\alpha P_{\alpha\beta}^{(e)} = \pi_\alpha (\pi_\beta \theta_e) = \pi_\beta (\pi_\alpha \theta_e) = \pi_\beta P_{\beta\alpha}^{(e)}.$$

Part (iv) is left as an exercise. ■

Exercise: Establish part (iv) of Lemma 7.2.

For any equal input model on r states, the transition matrix $P^{(e)}$ has the eigenvalue 1 with multiplicity 1 and has the eigenvalue $1 - \theta_e$ with multiplicity $r - 1$. Also, for a fixed π (but variable θ_e) the matrices $P^{(e)}$ commute, as they can be simultaneously diagonalized by a fixed matrix (which depends on π).

For two vertices u, v of T , let $p(u, v)$ be the probability that the u and v receive different states under the equal input model:

$$p(u, v) = \gamma \left(1 - \prod_{e \in P(T; u, v)} (1 - \theta_e) \right), \quad (7.15)$$

where $\gamma(e) = (1 - \sum_\alpha \pi_\alpha^2)$. To see this, first suppose that $e = (u, v)$. Then

$$p(u, v) = \sum_{\alpha} \pi_\alpha \sum_{\beta \neq \alpha} P_{\alpha\beta}^{(e)} = \sum_{\alpha} \pi_\alpha \sum_{\beta \neq \alpha} \pi_\beta \theta_e,$$

which simplifies to the expression in (7.15). The general case where u and v are the endpoints of a path now follows from part (iv) of Lemma 7.2.

The equal input model also has a continuous realization with rate matrix Q defined by its off-diagonal entries as follows:

$$Q_{\alpha\beta} = \pi_\beta \text{ for all } \alpha, \beta \in S, \quad \alpha \neq \beta,$$

where the diagonal entries are determined by the requirement that each row of Q sums to 0. In this case, $P^{(e)} = \exp(Q t_e)$ for $t_e = -\ln(1 - \theta_e)$, so $\theta_e = 1 - \exp(-t_e)$. Thus,

by eqn. (7.15), the probability that the leaves x and y are in different states is given by $p(x, y) = \gamma(1 - \exp(-\sum_{e \in P(T; u, v)} t_e))$. If $\mu(x, y)$ is the evolutionary distance between leaves x and y (i.e., the expected number of changes along the path in T connecting these two leaves) then from eqn. (7.6) we have the invertible pair of relationships:

$$p(x, y) = \gamma \left(1 - \exp \left(-\frac{\mu(x, y)}{\gamma} \right) \right) \text{ and } \mu(x, y) = -\gamma \ln \left(1 - \frac{p(x, y)}{\gamma} \right). \quad (7.16)$$

In the case where π is uniform, the equal input model reduces to the fully symmetric model in which all substitution events have equal probability. For example, for the JC69 model, eqn. (7.16) gives the well-known JC69 distance correction formula

$$\mu(x, y) = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p(x, y) \right). \quad (7.17)$$

One feature of the equal input model that fails for most other Markov processes on trees is the following. Let Π be any partition of the state space S and, for a state $s \in S$, let $[s]$ denote the corresponding block of Π containing s . For an equal input process Y on a phylogenetic tree T , let \tilde{Y} be the induced stochastic process on T , defined by $\tilde{Y}_v = [Y_v]$ for all vertices v of T .

Proposition 7.3. *For any equal input model with parameters π and $\{\theta_e\}$, and any partition σ of S , \tilde{Y} is also an equal input Markov process on T , with parameters $\tilde{\pi}$ and $\{\theta_e\}$, where, for each block B of σ , $\tilde{\pi}_B := \sum_{\beta \in B} \pi_\beta$.*

Proof. For $(\tilde{Y}_v, v \in V(T))$ to be a Markov process on T it suffices to check the “lumpability” criterion (cf. Theorem 6.3.2 of [214]) that for any two choices $\alpha, \alpha' \in A \in \Pi$ and block $B \in \Pi$, and edge (u, v) of T , we have

$$\mathbb{P}(Y_v \in B | Y_u = \alpha) = \mathbb{P}(Y_v \in B | Y_u = \alpha').$$

For each $B \neq A$, this last equality is clear from eqn. (7.14). Since $\mathbb{P}(Y_v \in A | Y_u = \alpha) = 1 - \sum_{B \in \Pi, B \neq A} \mathbb{P}(Y_v \in B | Y_u = \alpha)$, the criterion also holds for the case where $B = A$. Finally, for $B \neq A$, $\mathbb{P}(\tilde{Y}_v = B | \tilde{Y}_u = A) = \sum_{\beta \in B} (\pi_\beta \theta_e) = \tilde{\pi}_B \theta_e$. ■

As a simple application of Proposition 7.3 for the JC69 model (the symmetric model on $S = \{A, C, G, T\}$), any partition of S into two blocks of size 2 leads to a symmetric 2-state model, while partitioning S into blocks of sizes 3 and 1 gives an equal input model on two states with $\pi = (3/4, 1/4)$.

The equal input model as an instance of the random cluster model. A further description of the equal input model is as an instance of the (finite) “random cluster model” (defined shortly) on trees.⁴¹

First, we state a simple lemma that helps us to establish the link between these models, and which will be used later in this chapter. This lemma can be phrased and proved purely in the language of probability theory (cf. [293], Theorem 12), but we describe it algebraically here.

Lemma 7.4. *For variables x_1, x_2, \dots, x_n , let $h(\mathbf{x}) \in \mathbb{R}[x_1, \dots, x_n]$ be a polynomial of the form*

$$h(\mathbf{x}) = \sum_{A \subseteq [n]} c_A \prod_{j \in A} x_j,$$

⁴¹Random cluster models are also used to study processes on graphs, such as the “Ising model” in physics.

where $c_A \in \mathbb{R}$. Then $c_A = 0$ for all $A \subseteq [n]$ (i.e., $b \equiv 0$) if and only if for any $t \neq 0$, $h(\mathbf{x}) = 0$ for all $\mathbf{x} \in \{0, t\}^n$.

Proof: The “only if” part holds automatically; for the “if” direction, given any subset B of $[n]$, let $h(B) = h(\mathbf{x}^B)$, where $x_i^B = t$ if $i \in B$ and $x_i^B = 0$ otherwise. Then $h(B) = 0$ by hypothesis, and $h(B) = \sum_{A \subseteq B} c_A t^{|A|}$ by definition. Applying the (generalized) principle of inclusion and exclusion it follows that, for each $A \subseteq [n]$, $c_A t^{|A|} = \sum_{B \subseteq A} (-1)^{|A-B|} h(B) = 0$, so $c_A = 0$. ■

In the *random cluster model* on an unrooted phylogeny $T \in P(X)$, each edge e of T is deleted independently with probability p_e . The leaves in each connected component of the resulting disconnected graph are then all assigned the same state α with probability π_α , independently of the assignments to the other components (see Fig. 7.3).

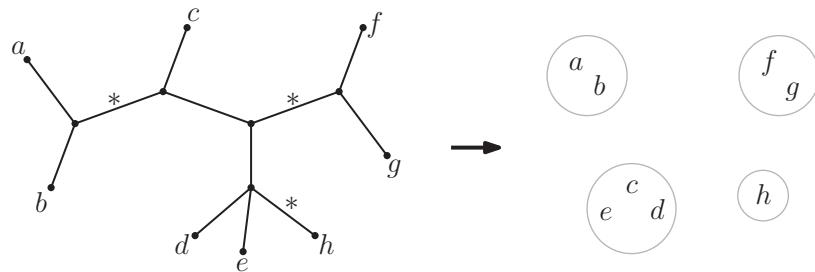


Figure 7.3. Cutting the three edges marked * in the tree on the left leads to the partition of X shown at right. Under the random cluster model, these four blocks are independently assigned states from the distribution π .

The following result shows that the equal input model (a Markov process on a rooted tree) is essentially equivalent to the random cluster model (a stochastic process described on an unrooted tree).

Proposition 7.5. *For any phylogenetic tree T , the equal input model with parameters π and $\{\theta_e\}$ produces an identical probability distribution on characters as the random cluster model with parameters π and $p_e = \theta_e$ for each edge e of T .*

Proof: We will use Lemma 7.4, taking $t = 1$, the x variables to be the θ_e parameters, and $h(\mathbf{x})$ to be the difference in probability of an arbitrary character f between the two models. We can apply this lemma since in both the equal input model and the random cluster model, the probability of a character is a first-order function of each θ_e value. By the lemma, it suffices to show that the two models produce identical probability distributions on characters when $\theta_e \in \{0, 1\}$ for each edge e . In the equal input model, if $e = (u, v)$ and $\theta_e = 0$, then u and v are in the same state with probability 1. If $\theta_e = 1$, then under the equal input model v is assigned state β with probability π_β independently of the state of vertex u . In summary, the leaves in each component of the tree obtained by deleting the edges with $\theta_e = 1$ are all the same state, and that state is assigned independently between the components according to π . This is precisely what the random cluster model does for the same set of θ_e values in the set $\{0, 1\}$. The general case follows by Lemma 7.4. ■

As mentioned previously, when π is uniform, the equal input model reduces to the fully symmetric model. The latter model allows an interesting link to the MP score that

we discussed in Chapter 6. Let $p(f|T, \theta)$ denote the probability of generating character f under the equal input model with edge parameters $\theta = (\theta_e : e \in E(T))$ and π uniform.

Proposition 7.6. *For the fully symmetric model on a set S of r states, and any character $f : X \rightarrow S$,*

$$\sup_{\theta} p(f|T, \theta) = \left(\frac{1}{r}\right)^{\text{ps}(f, T)+1}.$$

Proposition 7.6 has an easy proof when $r = 2$ by a result (related to Menger's theorem) from Section 5.2.1 as follows. Recall that $\text{ps}(f, T)$ equals the maximum number of edge-disjoint paths that can be placed in T so that each path connects a pair of leaves of T that are assigned different states by f . The probability that the two endpoints of any path in T are in different states is at most $\frac{1}{2}$. Since these events (involving edge-disjoint paths) are independent, it follows that, for any leaf x ,

$$p(f|T, \theta) \leq \pi_{f(x)} \cdot \left(\frac{1}{2}\right)^{\text{ps}(f, T)} = \left(\frac{1}{2}\right)^{\text{ps}(f, T)+1}. \quad (7.18)$$

Conversely, consider any minimal extension F of f . For each edge $e = \{u, v\}$ for which $F(u) \neq F(v)$, put $\theta_e^{(\epsilon)} = 1 - \epsilon$, and for all other edges e , set $\theta_e^{(\epsilon)} = \epsilon$. In this case,

$$\lim_{\epsilon \rightarrow 0} p(f|T, \theta^{(\epsilon)}) = \left(\frac{1}{2}\right)^{\text{ps}(f, T)+1}. \quad (7.19)$$

Combining eqns. (7.18) and (7.19) establishes Proposition 7.6 in the special case of $r = 2$. For the general case of $r \geq 2$, a more delicate argument is required (for further details, applications, and extensions, see [361] and [141]).

7.3.2 ■ Symmetries within models

The existence of symmetries in a model can be particularly useful for deriving exact analytical results that do not hold in more general settings. In this section, we describe two different notions of how a group \mathcal{G} can express the symmetries in a Markov process on a tree.

\mathcal{G} -equivariant models. This is a relatively recent notion, investigated in [115], and developed further by Marta Casanellas and colleagues (see, e.g., [77]). Models with this property have a tractable algebraic structure (a topic we will explore further in the next chapter).

Let \mathcal{G} be a group of permutations on the set S of states of a Markov model \mathcal{M} . Then \mathcal{M} is said to be \mathcal{G} -equivariant if it satisfies the property that for all $g \in \mathcal{G}$, and each transition matrix P in \mathcal{M} and distribution of states π allowed by \mathcal{M} , the following conditions hold for all states $\alpha, \beta \in S$:

$$P_{\alpha\beta} = P_{g(\alpha)g(\beta)} \text{ and } \pi_{g(\alpha)} = \pi_\alpha. \quad (7.20)$$

In other words, each permutation of the states in \mathcal{G} corresponds to a symmetry in the underlying Markov process (both in the transition matrices and the stationary distribution). An equivalent way to state the condition in eqn. (7.20) is that $K_g P K_g^{-1} = P$, and $\pi K_g = \pi$.

for all $g \in \mathcal{G}$, where K_g is the permutation matrix with $(K_g)_{\alpha\beta} = 1$ if $g(\beta) = \alpha$ and 0 otherwise. The requirement that π is fixed by each K_g does not necessarily uniquely determine π , and so the process need not be stationary in the tree. An example where this occurs is in the *strand-symmetric model* (SSM), which has the nucleotide state set $S = \{A, C, G, T\}$ and where \mathcal{G} is the cyclic group of order 2 generated by the element (AT)(CG) of order 2. In this case, the \mathcal{G} -invariance of π gives two equations (namely $\pi_A = \pi_T$ and $\pi_C = \pi_G$) which, together with $\pi_A + \pi_C + \pi_G + \pi_T = 1$, leaves one free parameter for specifying π . Consequently, π can vary across the tree, in which case the model would not be stationary.

The set $\mathcal{M}_{\mathcal{G}}$ of \mathcal{G} -equivariant transition matrices is also multiplicatively closed. To see why, suppose that $P, P' \in \mathcal{M}_{\mathcal{G}}$. For any $g \in \mathcal{G}$ (and permutation matrix K_g as defined above),

$$K_g(PP')K_g^{-1} = (K_gPK_g^{-1})(K_gP'K_g^{-1}) = PP'.$$

Examples of \mathcal{G} -equivariant models for $S = \{A, C, G, T\}$ and the numbers (k_1, k_2) of free parameters that specify π and P can be described using group notation⁴² as follows:

- JC69, corresponding to the full symmetric group $\mathcal{G} = \Sigma(S)$ of order 24; and a uniform root distribution; with (0, 1) parameters.
- K2ST, corresponding to the dihedral group $\mathcal{G} = \langle (ACGT), (AG) \rangle$ of order 8; and a uniform root distribution; with (0, 2) parameters.
- K3ST, corresponding to the Klein group $\mathcal{G} = \langle (AT)(CG), (AG)(CT) \rangle$ of order 4; and a uniform root distribution; with (0, 3) parameters.
- SSM, corresponding to the group $\mathcal{G} = \langle (AT)(CG) \rangle$ of order 2; and a root distribution with $\pi_A = \pi_T, \pi_C = \pi_G$; with (1, 6) parameters.
- GMM, corresponding to the trivial group $\mathcal{G} = \langle e \rangle$, and an arbitrary root distribution (3 parameters); with (3, 12) parameters.

There is also a weaker notion than \mathcal{G} -equivariance, which requires only that the model \mathcal{M} , viewed as a collection of transition matrices and vectors π , is invariant as a set under the action of \mathcal{G} , as discussed in [138].

Group-based models. Let $\mathcal{G} = (S, +)$ be an Abelian group. In other words, there is some commutative operation (+) on the state space S that satisfies the usual group axioms. A set \mathcal{M} of transition matrices is said to form a *group-based model* (based on \mathcal{G}) if for each $P \in \mathcal{M}$, we have (in additive group notation)

$$P_{\alpha\beta} = f(\alpha - \beta),$$

for all $\alpha, \beta \in S$ where the function $f = f_P : \mathcal{G} \rightarrow [0, 1]$ may depend on P . Notice that f must necessarily satisfy the condition $\sum_{\gamma \in S} f(\gamma) = 1$.

Notable examples of group-based models include the 2-state symmetric model (for $\mathcal{G} = \mathbb{Z}_2$), and K3ST (for $\mathcal{G} = \mathbb{Z}_2 \times \mathbb{Z}_2$), and its submodels (JC69 and K2ST). JC69 can also be viewed as a group-based model for the cyclic group $\mathcal{G} = \mathbb{Z}_4$. More generally, a fully symmetric model on r states can be viewed as a group-based model on \mathbb{Z}_r , by placing the additional restriction that $f(\gamma)$ takes the same value for all nonidentity elements γ of \mathcal{G} .

⁴²So $\langle * \rangle$ refers to the finite group generated by the elements $*$, (ACGT) is the 4-cycle $A \rightarrow C \rightarrow G \rightarrow T \rightarrow A$, while the remaining elements (e.g., (AG)(CT)) are order-2 elements that interchange states (e.g., $A \leftrightarrow G$ and $C \leftrightarrow T$).

Some of the key properties of group-based models are the following:

- (i) Any transition matrix in a group-based model is doubly stochastic (i.e., not only does each row sum to 1, but each column does also). Consequently, each group-based model has the uniform distribution as a stationary distribution, so a group-based model \mathcal{M} is time-reversible precisely if P is symmetric. This holds whenever \mathcal{G} is an elementary Abelian group (i.e., each element of \mathcal{G} has order 2), which applies, for example, for the K3ST model where $\mathcal{G} = \mathbb{Z}_2 \times \mathbb{Z}_2$, or more generally for any group \mathcal{G} provided that $f(\gamma) = f(-\gamma)$ for all $\gamma \in \mathcal{G}$.
- (ii) If \mathcal{M} is based on \mathcal{G} , and has a uniform distribution for π , then \mathcal{M} is \mathcal{G} -equivariant under the natural action of \mathcal{G} on S defined by $\gamma(\gamma') = \gamma + \gamma'$, since

$$P_{\gamma(\alpha)\gamma(\beta)} = f((\gamma + \alpha) - (\gamma + \beta)) = P_{\alpha\beta}.$$

However, there are models that are \mathcal{G} -equivariant but not group-based, such as the SSM and the general Markov model.

- (iii) The set of transition matrices based on \mathcal{G} is closed. This follows from part (ii).

Exercise: Prove that for any transition matrix P in a group-based model, each column of P sums to 1.

Notice that if a rate matrix Q satisfies $Q_{\alpha\beta} = h(\alpha - \beta)$ for all α, β , where h is a nonnegative function on the nonidentity elements of \mathcal{G} and $\sum_{\gamma} h(\gamma) = 0$, then Q is the rate matrix for a model that is also based on \mathcal{G} .

Group-based models have the limitation that their stationary distribution is uniform. However, they have a number of attractive properties and we mention one now (we will discuss another shortly in connection with discrete Fourier analysis). Call two characters $f, f' : X \rightarrow \mathcal{G}$ equivalent if $f'(x) = f(x) + \gamma$ holds for all $x \in X$ and some fixed $\gamma \in \mathcal{G}$. We will then refer to equivalence classes as *patterns*. When \mathcal{G} is a group-based model with uniform π , characters in the same equivalence class have equal probability. Moreover, when \mathcal{G} is an elementary Abelian group and π is not required to be uniform, then the probability distribution on patterns is independent of the distribution $\hat{\pi}$ of states at any (root) vertex of T , as well as the location of that root.

The relationship between the different classes of models we have discussed so far is represented in Fig. 7.4. In molecular phylogenetics, the particular choice of model depends on the type and properties of the data under study. Statistically based model selection criteria are typically employed for making this choice.

7.4 • The Hadamard story

A special case of the equal input model is the 2-state symmetric model. This simple process is familiar in coding theory as the “binary symmetric channel.” In phylogenetics, each edge e of the tree has a certain probability p_e of a change of state between its endpoints; these change events are treated independently across the edges. We will refer to p_e as the *substitution probability* for edge e . As in coding theory, it is usually assumed that p_e lies strictly between 0 and 0.5. The model usually assumes that the (marginal) state at any given leaf is uniform (i.e., no state is “preferred”). In biology, the model has been variously referred to as the Cavendar–Farris–Neyman (CFN) model or the Neyman 2-state model.

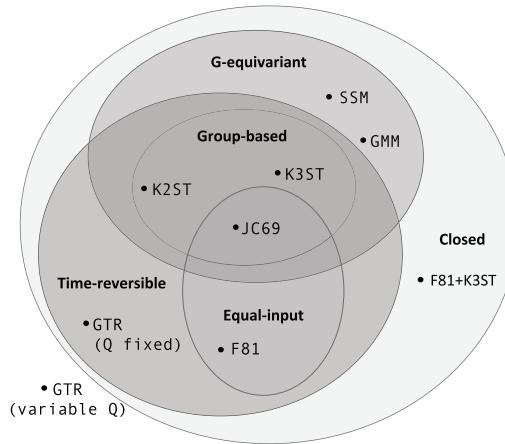


Figure 7.4. A schematic diagram illustrating the relationship between the different classes of models described, and the location of some particular models (SSM = strand symmetric model; K2ST, K3ST = Kimura 2- and 3-parameter models; JC69 = Jukes–Cantor model; F81 = Felsenstein’s 1981 model; GMM = general Markov model; GTR = general time-reversible model; F81 + K3ST = a Lie algebra model described in [348, 347]).

There are 2^n different ways to assign the two states (say, α and β) to the set $[n]$ but if we regard complementary assignments (obtained by interchanging α and β) as being equivalent, we get 2^{n-1} distinct equivalence classes. Each such equivalence class can be encoded as the subset A of $[n-1]$ consisting of those elements of $[n]$ having a state different from that assigned to n . We will refer to a subset of $[n-1]$ as a *pattern* on $[n]$.

For the 2-state symmetric model on T and a subset A of $[n-1]$, we let p_A be the probability of generating pattern A at the leaves of the tree. For example, p_\emptyset is the probability that all leaves are in the same state (i.e., all α or all β). For the tree in Fig. 7.5, a check of the four possible pairs of states at the two interior vertices gives

$$\begin{aligned} p_\emptyset &= (1-p_1)(1-p_2)(1-p_3)(1-p_4)(1-p_5) + p_1 p_2 p_3 p_4 (1-p_5) \\ &\quad + p_1 p_2 p_5 (1-p_3)(1-p_4) + p_3 p_4 p_5 (1-p_1)(1-p_2). \end{aligned}$$

In fact, for any tree $T \in P(n)$ and any subset A of $[n-1]$, p_A is independent of the probability distribution π chosen for any root vertex. In particular, π need **not** be uniform; however, if π is uniform, then each of the two characters on $[n]$ that induce pattern A has equal probability.

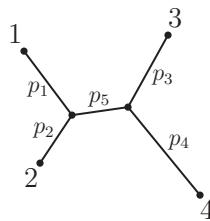


Figure 7.5. Assignment of substitution probabilities p_i to the edges of T . For the 2-state model p_i is related to the edge length (expected number of substitutions under a continuous-time realization) by $p_i = \frac{1}{2}(1 - \exp(-2t_i))$.

There are various ways to compute the p_A values, but one particularly elegant way holds for any phylogenetic tree with n leaves. To describe it, we need to make a short combinatorial detour.

Combinatorics of path systems in a tree. Let $\mathcal{E}(n)$ denote the 2^{n-1} subsets B of $[n]$ of even cardinality. For any phylogenetic T in $P(n)$ and any set B in $\mathcal{E}(n)$, let $P(T, B)$ denote those edges e of T for which B contains an odd number of leaves in each of the two components of $T - e$.

There is an equivalent way to view $P(T, B)$. Let β_1, \dots, β_k be an arbitrary partition of B into $k = \frac{1}{2}|B|$ disjoint pairs. Then $\Delta_{i=1}^k P(T, \beta_i)$ (the symmetric difference of the paths $P(T, \beta_i)$) is the set of edges e of T that lie in an odd number of these paths, and so $\Delta_{i=1}^k P(T, \beta_i)$ is equal to $P(T, B)$. This equivalence shows that this second path-description of $P(T, B)$ is well defined (i.e., it is independent of the choice of β_1, \dots, β_k). Clearly β_i can be chosen so that the corresponding paths in T are edge-disjoint (and also vertex-disjoint if the tree is binary).⁴³

For a subset G of the edges of T , let $\alpha_{T,G} \subseteq [n-1]$ be the pattern generated when a state change occurs on each edge in G and not on any other edges of T . The following result is the key combinatorial ingredient in a result that will follow.

Lemma 7.7. *For any $T \in P(n), B \in \mathcal{E}(n)$, and $G \subseteq E(T)$,*

$$|P(T, B) \cap G| = |\alpha_{T,G} \cap B| \pmod{2}.$$

Proof: For a pair β of leaves, the set $P(T, \beta) \cap G$ has odd cardinality if and only if $\alpha_{T,G}$ contains exactly one of the two leaves in β . Write $P(T, B) = \Delta_{i=1}^k P(T, \beta_i)$ where the paths $P(T, \beta_i)$ are edge-disjoint. Thus, using the generic equalities $|\Delta_i Y_i| = \sum_i |Y_i| \pmod{2}$ and $(\Delta_i Y_i) \cap W = \Delta_i(Y_i \cap W)$ we have the following equivalences (mod 2):

$$\begin{aligned} |P(T, B) \cap G| &= |(\Delta_{i=1}^k P(T, \beta_i)) \cap G| = \sum_{i=1}^k |P(T, \beta_i) \cap G| \\ &= \sum_{i=1}^k |\alpha_{T,G} \cap \beta_i| = |\alpha_{T,G} \cap B| \pmod{2}, \end{aligned}$$

which establishes Lemma 7.7. ■

We now present the alternative description of p_A , discovered by Mike Hendy [184], and which at first sight may seem slightly mysterious.

Theorem 7.8.

$$p_A = \frac{1}{2^{n-1}} \sum_{B \in \mathcal{E}(n)} (-1)^{|A \cap B|} \prod_{e \in P(T, B)} (1 - 2p_e).$$

For the tree in Fig. 7.5, if we let $\omega_i = 1 - 2p_i$ and take $A = \emptyset$ (so that $(-1)^{|A \cap B|} = 1$ for all B in Theorem 7.8), then we obtain the following expression for p_\emptyset :

$$\frac{1}{8}(1 + \omega_1\omega_2 + \omega_3\omega_4 + \omega_1\omega_3\omega_5 + \omega_2\omega_3\omega_5 + \omega_1\omega_4\omega_5 + \omega_2\omega_3\omega_5 + \omega_1\omega_2\omega_3\omega_4). \quad (7.21)$$

⁴³As a combinatorial aside, the problem, “Given a phylogeny with edge lengths, (T, l) , find an even cardinality subset B of leaves to maximize $\sum_{e \in P(T, B)} l(e)$,” turns out to have a polynomial-time solution [16].

All other p_A values are obtained from the right-hand side of eqn. (7.21) by replacing + with $-$ for exactly half the terms: If pattern A arises from state changes on (just) a certain subset E' of edges, then p_A is obtained by the transformations $\omega_e \mapsto -\omega_e$ for each $e \in E'$ (the resulting expression is independent of the particular choice of E' for A).

Theorem 7.8 has been generalized to the K3ST model using discrete Fourier analysis by Evans and Speed [130]. This was further extended to all group-based models (on any state space) by Székely et al. [352]. When the group \mathcal{G} is an elementary Abelian group, this representation is always based on Hadamard matrices, but for groups that have elements of order greater than 2, the corresponding matrices are unitary matrices with complex number entries that correspond to the character table of \mathcal{G} (of course the final probabilities are always real). An alternative and novel way of viewing the Hadamard representation is described in [64].

To prove Theorem 7.8 we require a second lemma. Recall that a square $k \times k$ matrix H is a *Hadamard matrix* if its entries consist of ± 1 and $HH^T = kI$ (i.e., the rows of H are orthogonal).

Lemma 7.9. *For subsets A and A' of $[n-1]$, with $A \neq A'$,*

$$\sum_{B \in \mathcal{E}(n)} (-1)^{|A \cap B| + |A' \cap B|} = 0. \quad (7.22)$$

In other words, $H = [(-1)^{|A \cap B|}]$ is a Hadamard matrix.

Proof: For each $A \in 2^{[n-1]}$, the function $\varphi_A : \mathcal{E}(n) \rightarrow \{\pm 1\}$ defined by $\varphi_A(B) = (-1)^{|A \cap B|}$ is a group homomorphism from $(\mathcal{E}(n), \Delta)$ to $\mathbb{Z}_2 = (\{\pm 1\}, \cdot)$. Moreover, if $A' \neq A$, then the homomorphism

$$\begin{aligned} \varphi_A \varphi_{A'} : (\mathcal{E}(n), \Delta) &\rightarrow \mathbb{Z}_2, \\ B &\mapsto (-1)^{|A \cap B| + |A' \cap B|} \end{aligned}$$

maps onto $\{\pm 1\}$ and if $A \neq A'$, then there is an element x in one set (say A) that is not in the other set, in which case $\varphi_A(\{x, n\}) = -1$ and $\varphi_{A'}(\{x, n\}) = 1$ and so $\varphi_A \varphi_{A'}(\{x, n\}) = -1$. Thus, by Lagrange's theorem, exactly half of the elements of $\mathcal{E}(n)$ map to -1 under $\varphi_A \varphi_{A'}$, so

$$\sum_{B \in \mathcal{E}(n)} (-1)^{|A \cap B| + |A' \cap B|} = 0;$$

in other words, the rows of H are orthogonal. ■

Proof of Theorem 7.8. From eqn. (7.11), the direct Markov process description, p_A is

$$p_A = \sum_{\alpha \in S} \pi_\alpha \sum_F \prod_{e \in \text{ch}(F)} p_e \prod_{e \in E(T) - \text{ch}(F)} (1 - p_e), \quad (7.23)$$

where F ranges over all functions from the set of vertices of T into the state space S with $F(v_0) = \alpha$ and $A = \{x \in [n] : F(x) \neq F(n)\}$; additionally, $\text{ch}(F)$ is the set of edges $\{u, v\}$ of T for which $F(u) \neq F(v)$ (i.e., the edges on which a state change occurs).

To show that this expression is equivalent to the Hadamard description, we will again use Lemma 7.4. For this lemma we take $t = 1$, the x variables to be the p_e parameters, and b to be the difference in probability of an arbitrary pattern under the 2-state symmetric model (eqn. 7.23) and the Hadamard description.⁴⁴ Let G be any subset of edges of T ,

⁴⁴In this proof we therefore relax the restriction that $p_e < \frac{1}{2}$.

and suppose that $p_e = 1$ for all $e \in G$ and $p_e = 0$ otherwise. Then for every subset A of $[n-1]$, eqn. (7.23) gives

$$p_A = \begin{cases} 1 & \text{if } \alpha_{T,G} = A, \\ 0 & \text{otherwise.} \end{cases} \quad (7.24)$$

With this same set G of edges notice that $(1 - 2p_e) = -1$ for all $e \in G$ while $(1 - 2p_e) = 1$ otherwise. Thus the Hadamard expression for p_A becomes

$$\frac{1}{2^{n-1}} \sum_{B \in \mathcal{E}(n)} (-1)^{|A \cap B|} (-1)^{|P(T,B) \cap G|} = \frac{1}{2^{n-1}} \sum_{B \in \mathcal{E}(n)} (-1)^{|A \cap B| + |\alpha_{T,G} \cap B|},$$

where the second equality is by Lemma 7.7. By Lemma 7.9 this last expression is 1 when $\alpha_{T,G} = A$ and is zero otherwise, exactly as for eqn. (7.24). This shows that the two expressions for p_A agree when $p_e \in \{0, 1\}$ for all e , and since both expressions are first order in each variable p_e the two expressions are equal in general by Lemma 7.4. This complete the proof of Theorem 7.8. ■

Notice that the Hadamard representation (Theorem 7.8) makes it clear that (7.23) does not depend on π (even though F in the second sum refers to the state α at the root vertex v_0). Notice also that Theorem 7.8 can be written

$$\mathbf{p} = \mathbf{p}(\omega) = \frac{1}{2^{n-1}} H \omega, \quad (7.25)$$

where H is the Hadamard matrix in Lemma 7.9, \mathbf{p} is the column vector indexed by $2^{[n-1]}$ with A entry p_A , and ω is the column vector indexed by $\mathcal{E}(n)$ with B entry $\omega_B = \prod_{e \in P(T,B)} (1 - 2p_e)$. Since H^T is the inverse of $\frac{1}{2^{n-1}} H$, eqn. (7.25) gives

$$\omega = H^T \mathbf{p}. \quad (7.26)$$

One immediate consequence of this identity is to provide two algebraically independent quadratic polynomials for the tree T in Fig. 7.5. For this tree, we have

$$\omega_{\{1,3\}} \omega_{\{2,4\}} - \omega_{\{1,4\}} \omega_{\{2,3\}} = 0 \text{ and } \omega_{\{1,2,3,4\}} - \omega_{\{1,2\}} \omega_{\{3,4\}} = 0. \quad (7.27)$$

By eqn. (7.26), these two equalities translate into quadratic polynomial expressions in the p_A values that vanish for all parameters for T (i.e., they are called “phylogenetic invariants” for T , a topic we explore further in the next chapter).

Another interesting consequence of eqn. (7.25) is the inner product identity:

$$\langle \mathbf{p}(\omega), \mathbf{p}(\omega') \rangle = p_\emptyset(\omega \omega'), \quad (7.28)$$

where $(\omega \omega')_e := \omega_e \omega'_e$ for each $e \in E(T)$. To see why, just notice that

$$\langle \mathbf{p}(\omega), \mathbf{p}(\omega') \rangle = \mathbf{p}(\omega)^T \mathbf{p}(\omega') = \omega^T \frac{1}{2^{n-2}} H^T H \omega' = \frac{1}{2^{n-1}} \omega^T \omega' = p_\emptyset(\omega \omega').$$

If we take $\omega = \omega'$ in eqn. (7.28), the left-hand side is simply the sum of the squares of the p_A values, in other words, the probability that two characters generated independently on T are identical. Equation (7.28) reveals that this probability coincides with the probability of generating the pattern \emptyset (i.e., all leaves having the same state) on T when the edge lengths ($q_e = -\frac{1}{2} \ln(1 - 2p_e)$) are doubled.

Exercise⁺: Under the 2-state symmetric model on a tree $T \in P(n)$, suppose that $\sum_A c_A p_A = 0$ holds for all choices of $(p_e : e \in E(T))$. Show that $c_A = 0$ for all $A \subseteq [n-1]$.

There is a further twist to the Hadamard story. It is possible to re-express eqn. (7.25) in the form

$$\mathbf{p} = \tilde{H}^{-1} \exp(\tilde{H}\gamma),$$

where $\tilde{H} = [(-1)^{|A \cap B|}]$ is a symmetric Hadamard matrix of rank 2^{n-1} in which the rows and columns are both indexed by the subsets of $[n-1]$, and where \exp is applied componentwise. The vector γ is indexed by the subsets of $[n-1]$ with $\gamma_A = 0$ unless either (i) $A|([n]-A)$ is a split of T , in which case, $\gamma_A = -\frac{1}{2} \ln(1-2p_e)$ for the edge e that corresponds to the split $A|([n]-A)$; or (ii) $A = \emptyset$, in which case $\gamma_\emptyset = -\sum_{A \neq \emptyset} \gamma_A$. This “conjugation” representation of \mathbf{p} has two remarkable properties: The tree only enters into the equation via the vector γ (the Hadamard matrix is independent of the tree), and the representation allows for an immediate (tree-independent) inversion formula, namely,

$$\gamma = \tilde{H}^{-1} \ln(\tilde{H}\mathbf{p}).$$

The Hadamard representation for the 2-state symmetric model, and its extension to group-based models, has been useful in providing streamlined proofs of a number of phylogenetic results. One recent example was to show that the pattern probabilities p_A for just the five splits $A|([n]-A)$ of a four-leaf binary tree suffice to determine the five edge parameters p_e [89].⁴⁵

Another example is the extension of the Hadamard representation of the 2-state symmetric model to provide a link between phylogenetic diversity (Chapter 6) and the expected proportion of segregating (nonconstant) characters by [66].

Here, we give a more standard application, by showing how the Hadamard representation can be used to determine conditions under which the types of methods discussed in earlier chapters (MP, maximum compatibility, and uncorrected distance methods) can converge on an incorrect tree as the number of characters increases.

7.4.1 • Application: The Felsenstein zone

Consider the tree shown in Fig. 7.6(b). We can imagine it as a tree in which there has been an accelerated rate of evolution (resulting in higher probabilities of change) in two nonadjacent lineages. It can also be realized on a rooted tree as in Fig. 7.6(a), with a single rate increase in one short branch (the branch leading to 1), and a distant outgroup species (4). Denote the probabilities of change on the edges of the tree in Fig. 7.6(b) by the values p_1, \dots, p_5 , as shown.

Theorem 7.10. *For a character generated on tree T in Fig. 7.6(b) under the 2-state symmetric model with $p_1 = p_4 = p$ and $p_2 = p_3 = p_5 = p'$, the expected parsimony score for T is larger than for the tree T' in Fig. 7.6(c) precisely when $p^2 > p'(1-p')$.*

Proof: The only 2-state characters that have different parsimony scores on T and T' are those that correspond to patterns which we will denote by f_{12} and f_{23} , and for which

⁴⁵Using the inverse function theorem, a “local” version of this result holds (i.e., for p_e sufficiently small) for trees with any number of leaves; whether the result holds globally is still open, even for $n=5$.

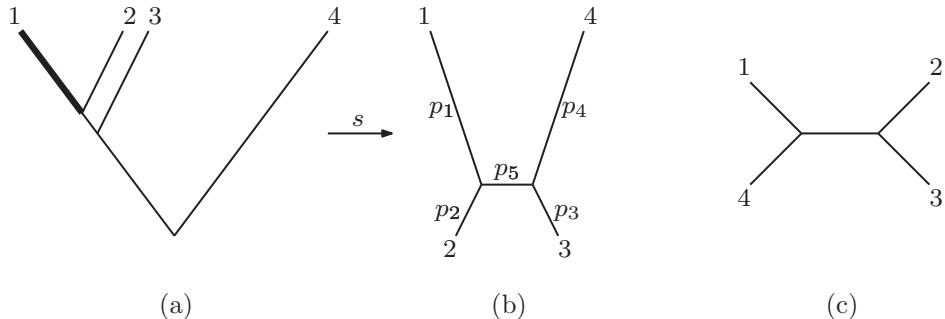


Figure 7.6. (a) A high rate of evolution on the lineage leading to species 1 and a distant outgroup species (4) can be modeled by a Markov process on the associated unrooted tree (obtained by suppressing the root) in (b); for this tree T , if p_1 and p_4 are large enough relative to the other p_i values, the MP tree for a large number of characters generated on T is likely to be the tree T' shown in (c).

$f_{12}(1) = f_{12}(2) \neq f_{12}(3) = f_{12}(4)$ and $f_{23}(2) = f_{23}(3) \neq f_{23}(1) = f_{23}(4)$. Notice that a character of type f_{12} has a parsimony score of 1 on T and 2 on T' , while a character of type f_{23} has a parsimony score of 1 on T' and 2 on T . Moreover, under the 2-state symmetric model on T , the probabilities of generating the patterns f_{12} and f_{23} are $p_{\{1,2\}}$ and $p_{\{2,3\}}$, respectively. For a character generated by this model on T , consider the random variable Δ that is the parsimony score of that character on T minus the parsimony score of that character on T' . The expected value of Δ , denoted $\mathbb{E}[\Delta]$, satisfies

$$\mathbb{E}[\Delta] = p_{\{2,3\}} - p_{\{1,2\}}. \quad (7.29)$$

If we now apply Theorem 7.8 for $n = 4$, with $\omega_i = (1 - 2p_i)$, and with $A = \{1, 2\}$ and $\{2, 3\}$, then

$$p_{\{1,2\}} = \frac{1}{8}(1 + \omega_1\omega_2 + \omega_3\omega_4 - \omega_1\omega_3\omega_5 - \omega_2\omega_3\omega_5 - \omega_1\omega_4\omega_5 - \omega_2\omega_4\omega_5 + \omega_1\omega_2\omega_3\omega_4),$$

$$p_{\{2,3\}} = \frac{1}{8}(1 - \omega_1\omega_2 - \omega_3\omega_4 - \omega_1\omega_3\omega_5 + \omega_2\omega_3\omega_5 + \omega_1\omega_4\omega_5 - \omega_2\omega_4\omega_5 + \omega_1\omega_2\omega_3\omega_4).$$

Substituting these identities into eqn. (7.29) gives $\mathbb{E}[\Delta] = \frac{1}{4}(-\omega_1\omega_2 - \omega_3\omega_4 + \omega_2\omega_3\omega_5 + \omega_1\omega_4\omega_5)$. If we now set $\omega_1 = \omega_4 = u = (1 - 2p)$ and $\omega_2 = \omega_3 = \omega_5 = v = (1 - 2p')$, we obtain

$$\mathbb{E}[\Delta] = \frac{v}{4}[u^2 + v^2 - 2u] = v(p^2 - p'(1 - p')),$$

and so $\mathbb{E}[\Delta] > 0$ precisely if $p^2 > p'(1 - p')$. This completes the proof. ■

Theorem 7.10, together with the law of large numbers (or the central limit theorem), ensures that for k characters generated by the T (with these p_i values), a different tree, namely T' , will have a lower parsimony score than T , with probability converging to 1 as k grows. Intuitively, parallel changes on the two long branches of T become more probable than a single change on the short edges. In other words, through the eyes of parsimony, it is more optimal to join these two edges together in the reconstruction. This phenomenon of “long branch attraction” has been observed in biological data [196]. Similar zones of inconsistency for MP exist under other models, such as GMM; for example, when MP is applied to 18S rRNA sequence data it tends to group birds with mammals

rather than with crocodilians, because the first two groups have independently become rich in G and C bases. This can be viewed as a Markov process on the tree in Fig. 7.6(ii), where the pendant edges leading to leaves 1 and 4 have a different type of transition matrix from the other three edges.

Note that for any sequence \mathcal{C} of binary characters on four leaves, the MP tree(s) coincide with the phylogenies that maximize the number of homoplasy-free characters. Moreover, these are the same trees that would be produced by applying most distance-based tree reconstruction methods (such as neighbor joining (NJ)) to the Hamming distance $\delta_{\mathcal{C}}(x, y)$.

Notice also that Theorem 7.10 allows for statistical inconsistency of MP even when all the substitution probabilities on the edges are small; in particular, p can be arbitrarily small, provided that p' is (quadratically) smaller. Alternatively, the probabilities can all be large also: p can be arbitrarily close to $\frac{1}{2}$ provided that p' is correspondingly sufficiently close to $\frac{1}{2}$.

Exercise⁺: Suppose that $p_1 = p_2 = p$, and that $p_3 = p_4 = p'$ in the tree T in Fig. 7.5(b). Can MP converge to a tree that is different from T as the number of characters grows, for some selection of values $p, p', p_5 \in (0, 0.5)$?

Although parsimony can fail to recover the true tree, there are methods for inferring it that are statistically consistent as the number k of characters grows. A particularly simple one for the 2-state symmetric model relies on the following distance function on $[n]$. For x, y in $[n]$, let

$$\hat{\mu}(x, y) = -\frac{1}{2} \ln(1 - 2\hat{p}(x, y)),$$

where $\hat{p}(x, y)$ is the proportion of characters that assign different states to x and y . Provided we apply a distance-based tree reconstruction method with a positive safety radius (cf. Section 6.1), we are guaranteed to recover the underlying (unrooted) tree T from k independently evolved characters as k grows. The reason is that, as $k \rightarrow \infty$, the law of large numbers ensures that $\hat{p}(x, y)$ converges to the probability $p(x, y)$ that leaves x and y are in different states, so $\hat{\mu}(x, y)$ converges to $\mu(x, y) = -\frac{1}{2} \ln(1 - 2p(x, y))$. It is then an easy exercise to show that μ has a tree representation on the true tree T with the edge weighting $w(e) = -\frac{1}{2} \ln(1 - 2p_e)$. In other words, $\mu = d_{(T, w)}$, for $w > 0$. The implication in (6.1) then ensures the reconstruction of both the unrooted tree and the edge weights from μ (and thereby from $\hat{\mu}$ for k sufficiently large). We will explore this topic further in the next chapter.

7.5 ■ Phylogenetic mixture models

The assumption that all characters evolve according to a fixed Markov process on a tree is a very strong one. In reality, some characters may evolve at a higher rate than others, or even according to a different process. In a *phylogenetic mixture model* on a fixed tree T , each character f is generated independently by a Markov process under a model \mathcal{M} with parameters (T, θ) , but where $\theta = \theta(\lambda)$ depends on a variable λ that is selected independently for each character from some fixed distribution \mathcal{D} .

The distribution \mathcal{D} can be either continuous or discrete. In the latter case, we will assume that \mathcal{D} has finite support. In other words, λ takes only a finite set of possible values $(\lambda_1, \lambda_2, \dots, \lambda_m)$ for some $m \geq 1$. In this case, the probability distribution p on

characters is given by

$$p = \sum_{i=1}^m a_i p_{(T, \theta_i)}, \quad (7.30)$$

where $a_i > 0$ is the probability of selecting λ_i , and so satisfies $\sum_i a_i = 1$, and $\theta_i = \theta(\lambda_i)$. In this case, we say that p is a *phylogenetic mixture* of m classes on T under model \mathcal{M} .

In molecular phylogenetics, mixtures may simply scale the rate matrix Q (i.e., site-to-site rate variation models, discussed shortly), while in other studies phylogenetic mixtures may also allow different characters to have different suites of edge lengths (since the transition rates for some characters may change in different parts of the tree, a phenomenon referred to as “heterotachy”).

Example: Consider a continuous-time 2-state symmetric model on a phylogeny $T \in B(n)$, $n \geq 4$, for which half the characters evolve at rate $\lambda_1 = 1$ and half evolve at rate $\lambda_2 = 5$. In terms of eqn. (7.30),

$$p = 0.5 p_{(T, \omega)} + 0.5 p_{(T, \omega')},$$

where (in the Hadamard representation of Section 7.4) we have $\omega_e = (1 - 2p_e)$ and $\omega'_e = \omega_e^5$. This phylogenetic mixture is no longer described by a Markov process on T under the 2-state symmetric model.

The simple mixture example we have highlighted (with $\lambda_1 = 1$ and $\lambda_2 = 5$) suffices to demonstrate two subtle but important points:

- (i) a mixture of Markov processes on T will not, in general, be a Markov process on T ;
- (ii) a phylogenetic mixture on T involving a given model may give a probability distribution on characters than cannot be realized by that model (in the unmixed setting) on that tree.

To see why, we first state a general inequality from elementary probability theory.⁴⁶

Lemma 7.11. Suppose that A_1 , A_2 , and B are events in a probability space, and

- (a) A_1 and A_2 are conditionally independent, when conditioned on either B or \overline{B} ;

⁴⁶This result formalizes the well-known warning that correlations between two events (e.g., ice cream sales and drownings) may not be due to direct causality but may arise via a third factor (e.g., weather). The proof of the lemma uses condition (a) to verify that $\mathbb{P}(A_1 \wedge A_2) - \mathbb{P}(A_1)\mathbb{P}(A_2)$ can be identified with the expansion of

$$\frac{1}{2} \sum_{(x,y):x,y \in \{B,\overline{B}\}} (\mathbb{P}(A_1|x) - \mathbb{P}(A_1|y))(\mathbb{P}(A_2|x) - \mathbb{P}(A_2|y))\mathbb{P}(x)\mathbb{P}(y).$$

Conditions (b) and (c) then ensure that this sum is strictly positive.

- (b) $\mathbb{P}(A_i|B) > \mathbb{P}(A_i|\overline{B})$ for $i = 1, 2$; and
- (c) $0 < \mathbb{P}(B) < 1$.

Then $\mathbb{P}(A_1 \wedge A_2) > \mathbb{P}(A_1)\mathbb{P}(A_2)$.

Returning to point (i), consider a cherry x_1, x_2 in T , select one of the two states for the model (say α), let A_i be the event that leaf x_i is in state α , and let B be the event that the character evolves on T under the slower rate $\lambda = 1$. Then if the probabilities in Lemma 7.11 are all calculated conditional on the event that the vertex v adjacent to x_1, x_2 is in state α , then conditions (a)–(c) of the lemma apply. It follows that the states at x_1 and x_2 are not conditionally independent given the state at v , and this violates a property of a Markov process on a tree (cf. Lemma 7.1).

To establish point (ii), suppose that T has at least four leaves, and so T has two edge-disjoint paths, each of which connects a pair of leaves. Let A_1 be the event that the first pair of leaves are in the same state, and let A_2 be the event that the second pair of leaves are in the same state. Under the 2-state symmetric model, A_1 and A_2 are independent. However, under the mixture process these events are strictly positively correlated. To see why, let B be the event that the character is generated by the model with the slower rate $\lambda_1 = 1$. In this case, conditions (a)–(c) in Lemma 7.11 hold, and so A_1 and A_2 are positively correlated, as claimed.

Exercise⁺: Consider a phylogenetic mixture of m equal input models, each on T , and with the same π vector and the same θ_e values, except for one edge e_0 . Show that the resulting phylogenetic mixture is described exactly by a single equal input model.

Identifiability of the discrete tree parameter. For mixtures of Markov processes, the identifiability of a binary tree can easily be lost, even for mixtures of two classes. A striking example of this loss was provided for the 2-state symmetric model [243]: if 50% of DNA sites evolve on a quartet tree T with one carefully chosen set of edge lengths, and 50% evolve on the same tree under a different chosen collection of edge lengths, then the probability distribution on characters is exactly identical to that in which all sites evolve on a different quartet tree with appropriately chosen edge lengths. Notice that in this case, we have more parameters in the generating model ($5 + 5 = 10$ edge lengths) than the dimension ($= 7$) of the space of probabilities on binary patterns.

However, we will see from algebraic considerations in the next chapter (Section 8.3.2) that for certain models (including JC69 and K2ST) the tree is still an identifiable parameter, even under arbitrarily complex phylogenetic mixtures. Moreover, for other models (including GMM on $r > 2$ states) the tree is “generically” identifiable for mixture where the number of processes (on the same tree) is not too large.

Site-to-site rate variation models. The simplest and most widely applied mixture models in phylogenetics are the ones that simply scale the substitution rate uniformly across the tree. In other words, if we have a stationary time-reversible rate matrix Q , then the transition matrix associated with each edge e can be written

$$P^{(e)} = \exp(Q\lambda l(e)), \quad (7.31)$$

where $l(e)$ is the length of edge e and λ is drawn independently from some discrete or continuous distribution \mathcal{D} . In the simple example earlier in this section, $\lambda = 1$ and $\lambda = 5$ with equal probability.

For such a model, if \mathcal{D} is known, then both T and l are identifiable parameters from the probability distribution on characters. If \mathcal{D} is unknown (e.g., containing parameters to be estimated), then the situation is much less clear. One general case where T is an identifiable parameter is when the edge lengths are ultrametric. To see why, observe that, in this case, the expected (normalized) Hamming distance between two leaves x and y is a monotone increasing function of the path length between x and y (by eqn. (7.9) and the comment that follows it), so this expected distance is an ultrametric having a representation on the underlying tree. Thus, for this very special case, the identifiability of T is robust to both the nature of \mathcal{D} and whether this is known or not.

There is another special case where \mathcal{D} does not complicate the identifiability question for T , which is for models that have “linear topology invariants” (e.g., JC69, K2ST), but we will defer discussion of this to the next chapter.

Apart from these two special cases, what can one infer about the binary tree T from a mixture distribution p (given by eqn. (7.30)) when \mathcal{D} is unknown? In general, it may be impossible to determine anything about T , as the following early result from [341] shows.

Proposition 7.12. *For the 2-state symmetric model, each binary phylogeny T in $B(n)$ ($n \geq 4$) has an associated set of edge lengths $\theta = \theta^T$, and a finite-support rate distribution $\mathcal{D} = \mathcal{D}^T$ for λ for which the mixture distribution (given by eqns. (7.30) and (7.31)) is identical for every tree T in $B(n)$.*

In other words, in such a setting it is impossible to tell which tree generated the data, even from infinitely long sequences (we should remember, however, that much “fine tuning” of parameters is needed for this to occur). A complementary positive result, established by [265], states that, roughly speaking, for a fixed (but unknown) site-specific rate distribution, \mathcal{D} , the phylogenetic tree parameter can typically be distinguished from p , provided that n (the number of leaves of T) is sufficiently large.

In molecular phylogenetics, the distribution \mathcal{D} is often taken to be a gamma distribution (either exactly or, more commonly, a discrete approximation of it). This leads to the widely used *general time-reversible with gamma-distributed rates model* (GTR + Γ). An extension of this model allows $\lambda = 0$ (i.e., an invariable character) with a fixed positive probability i_0 and, with probability $1 - i_0$, λ is chosen from a gamma distribution (the *general time-reversible model with gamma-distributed rates and invariable sites model* (GTR + $\Gamma + I$). Since the gamma distribution can be taken to have mean 1, only one (“shape”) parameter is required to specify it, and a second parameter is required to specify the proportion of invariable sites.

When these two parameters are known, the tree and edge lengths can be easily recovered from the mixture distribution p . However, when one or both of these parameters for \mathcal{D} is unknown, the situation is more delicate. This led to various partial results, culminating in the result from [81] that the tree T , its edge lengths l , the shape parameter for the (continuous) gamma, and the proportion i_0 of invariable sites are generically identifiable parameters under GTR + $\Gamma + I$. Indeed, these parameters can be identified provided that either (i) the interspecies distances (i.e., $d_{(T,l)}$) take at least three distinct values or (ii) if the tree has only two distinct interspecies distances and Q has three distinct eigenvalues.

It is also possible to consider a phylogenetic mixture where T is not fixed. In other words, we can consider mixtures of $p_{(T_i, \theta_{ij})}$ involving two or more binary phylogenetic X -tree T_1, \dots, T_s and a set of continuous parameters θ_{ij} for each tree T_i . Determining the conditions under which such models support the identifiability of the set of trees T_i that

appear in such a mixture is considerably more complex (some results in the case $s = 2$ are presented in [242], based on the notion of “disentangling trees” from Chapter 4).

We end this section with a brief technical comment. Notice that the set of all phylogenetic mixtures on $T \in P(n)$ with finite support for some model \mathcal{M} on state space S is the convex hull⁴⁷ of the vectors $p_{(T,\theta)} \in \mathbb{R}^N$, where θ ranges over all parameter values permitted under \mathcal{M} , and $N = |S|^n$ is the number of characters $f : [n] \rightarrow S$. Similarly, the set of all phylogenetic mixtures across all trees in $P(n)$ is a convex hull lying within the same vector space. It follows from *Carathéodory’s theorem*⁴⁸ that mixtures with finite support (on a fixed tree, or across a set of trees) need never be based on more than $N + 1$ classes.

⁴⁷The *convex hull* of a subset B of a vector space over \mathbb{R} is the set of all finite linear combinations $\sum_{i=1}^k a_i \cdot b_i$, where $a_i \in \mathbb{R}^{\geq 0}$, $b_i \in B$, $\sum_{i=1}^k a_i = 1$, and $k \geq 1$.

⁴⁸Carathéodory’s theorem states that if B is a subset of a d -dimensional vector space over \mathbb{R} , then every point in the convex hull of B can be expressed as a convex combination of not more than $d + 1$ points of B .

Chapter 8

Evolution on a tree: Part two

8.1 • Preliminaries

8.1.1 • Probability metrics and information

There are various ways to quantify the similarity of two probability distributions on the same set. These measures are particularly relevant for estimating how much data one needs to reconstruct a parameter of interest (such as a phylogeny, or its edge lengths, or a state at an ancestral vertex) from a sequence of independent samples. Essentially, when two probability distributions are “close” to each other, more data are required to distinguish which distribution generated the given data. In this section, we describe three related measures that compare two probability distributions, and describe, in a general setting, their relevance for inference methods such as maximum likelihood (ML).

Consider the space \mathcal{P} of all probability distributions on a finite set U ; in other words, $\mathcal{P} = \{p : U \rightarrow [0, 1] : \sum_u p(u) = 1\}$. For $p, q \in \mathcal{P}$, let

$$d_1(p, q) = \sum_{u \in U} |p(u) - q(u)|$$

and let

$$d_H(p, q) = \sqrt{\sum_{u \in U} (\sqrt{p(u)} - \sqrt{q(u)})^2} = \sqrt{2 - 2 \sum_{u \in U} \sqrt{p(u)q(u)}}.$$

Here, d_1 is the l_1 distance between p and q , and it has a very natural interpretation: $\frac{1}{2}d_1(p, q)$ —the total variation distance between p and q —can easily be shown to equal the largest possible difference between the probability of any event A under the two probability distributions p and q (i.e., $\frac{1}{2}d_1(p, q) = \max_{A \subset U} |p(A) - q(A)|$, where $p(A) = \sum_{u \in A} p(u)$ and $q(A) = \sum_{u \in A} q(u)$).

The metric d_H is called the *Hellinger distance*. Its definition looks rather odd, but it has one remarkable property. Let $p^{(k)}$ be the probability distribution on U^k corresponding to k independent samples of U drawn according to distribution p . Define $q^{(k)}$ similarly. It is easily shown that

$$1 - \frac{1}{2}d_H^2(p^{(k)}, q^{(k)}) = \left(1 - \frac{1}{2}d_H^2(p, q)\right)^k. \quad (8.1)$$

In other words, there is an exact relationship between the Hellinger distance involving k -fold independent samples and the Hellinger distance involving a single sample.⁴⁹ The metrics d_1 and d_H satisfy the inequalities $d_H^2(p, q) \leq d_1(p, q) \leq 2d_H(p, q)$, so if either d value is small, then the other is as well.

A basic result is that the number of samples k required to determine, with a given accuracy, whether all the samples were drawn according to the probability distribution p or q is at least a constant times $\frac{1}{d_H^2(p, q)}$, and this holds for any estimation procedure (see, e.g., [342]). The constant referred to depends on the degree of accuracy required.

For example, suppose that coin A has a slight bias for heads (H) (say, $\mathbb{P}(H) = 0.5 + \delta$) and coin B has a bias for tails (say, $\mathbb{P}(H) = 0.5 - \delta$), and let p_A and p_B be the resulting probability distributions on $\{H, T\}$. Suppose that one of these two coins is selected uniformly at random and tossed repeatedly. If we wish to accurately determine which coin (A or B) was chosen, then since $d_H^2(p_A, p_B) = \Theta(\delta)$, k has to grow at the rate $\frac{1}{\delta^2}$ as $\delta \rightarrow 0$. Of course, the standard statistical test (based on the central limit theorem) assures us that this rate is also sufficient. The same is true if, for the second coin, we replace $0.5 + \delta$ by 0.5. However, if one coin has $\mathbb{P}(H) = 0$ and the other has $\mathbb{P}(H) = \delta$, then for the resulting probability distributions q_A and q_B on $\{H, T\}$, $d_H^2(q_A, q_B)$ now grows only at the rate $\frac{1}{\delta}$, and a test based on the Poisson distribution ensures that this lower rate is now sufficient for accurately inferring the generating coin. Notice that $d_1(p_A, p_B)$ and $d_1(q_A, q_B)$ are both of order δ , so this metric fails to detect this difference in data requirements between these scenarios.

Exercise: Prove eqn. (8.1).

A third distance measure is not a metric, since it fails to be symmetric. However, it is closely connected to ML estimation and to information theory. This is the *Kullback–Leibler* divergence $d_{KL}(p||q)$, defined by $d_{KL}(p||q) = \sum_{u \in U} p(u) \ln\left(\frac{p(u)}{q(u)}\right)$, provided that $q(u) \neq 0$ for all $u \in U$ with $p(u) > 0$ (if $q(u) = p(u) = 0$, we may interpret $0 \ln\left(\frac{0}{0}\right) = 0$). *Pinsker’s inequality* states that $d_{KL}(p||q) \geq \frac{1}{2}d_1(p, q)^2$. This shows that $d_{KL}(p||q) \geq 0$, with equality if and only if $p = q$; the inequality also reveals that if d_{KL} is small, then so are d_1 and d_H . However, it is possible for the latter two metrics to be close to zero and d_{KL} arbitrarily large.

The *mutual information* of two random variables X and Y , denoted $I(X, Y)$, is $d_{KL}(p_{XY}||p_X p_Y)$, where p_{XY} is the joint probability distribution of (X, Y) and $p_X p_Y$ is the product of the marginal distributions. Thus,

$$I(X, Y) = \sum_{x,y} \mathbb{P}(X = x, Y = y) \ln\left(\frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)\mathbb{P}(Y = y)}\right),$$

from which it is clear that $I(X, Y) = I(Y, X)$.

Pinsker’s inequality shows that $I(X, Y) = 0$ if and only if X and Y are independent, and $I(X, Y)$ can be interpreted, roughly speaking, as a measure of the extent to which learning the state of one variable increases the ability to estimate the state of the other. *Fano’s lemma* provides a precise formulation of the claim that if $I(X, Y)$ is small, then knowing the state of X provides only a small increase in the accuracy of estimating Y .

⁴⁹Equation (8.1) also extends directly to allow the distributions to vary between samples.

8.1.2 ■ Maximum likelihood (ML) and variants

Suppose that we have a discrete parameter taking values in a finite set \mathcal{A} of discrete parameters (e.g., \mathcal{A} equal to the set of phylogenetic X -trees) and each element $a \in \mathcal{A}$ generates an associated probability distribution p_a on a finite set U (e.g., characters on X). A *reconstruction method* m is a (deterministic or randomized) function from U^k to \mathcal{A} , where $m(\mathbf{u}^{(k)})$ is the estimate of the element $a \in \mathcal{A}$ that generated the observed data $\mathbf{u}^{(k)} = (u_1, u_2, \dots, u_k)$. In the case where the reconstruction method m is based on an optimization process (e.g., MP or ML in phylogenetics), we will break ties uniformly.

Suppose a random element $\tilde{a} \in \mathcal{A}$ is selected according to some prior probability distribution μ on \mathcal{A} (i.e., $\mu(a) = \mathbb{P}(\tilde{a} = a)$) and that a sequence of k samples $\mathbf{u}^{(k)} = (u_1, u_2, \dots, u_k)$ is independently generated according to the probability distribution $p_{\tilde{a}}$. The *accuracy* of m on k -tuple samples is the expected probability that m will make a correct selection; in other words, the value

$$\sum_{a \in \mathcal{A}} \mathbb{P}(m(\mathbf{u}^{(k)}) = a | \tilde{a} = a) \mu(a).$$

Given $\mathbf{u}^{(k)}$, ML estimation selects from an element $a \in \mathcal{A}$ that maximizes $\prod_{i=1}^k p_a(u_i)$ (i.e., the probability of generating \mathbf{u} from the independent samples under p_a). ML has a number of nice properties, and we highlight a few here. For $\mathbf{u}^{(k)} = (u_1, \dots, u_k) \in U^k$, let \hat{p} be the empirical distribution on U defined by $\hat{p}_u = \frac{1}{k} |\{j : u_j = u\}|$.

Proposition 8.1.

- (i) *ML selects the parameter a to minimize $d_{KL}(\hat{p} || p_a)$.*
- (ii) *Provided that $d_1(p_a, p_b) > 0$ for all distinct $a, b \in \mathcal{A}$, the accuracy of ML on k -tuples tends to 1 as $k \rightarrow \infty$.*
- (iii) *If μ is uniform, then ML has the largest possible accuracy among all reconstruction methods.*
- (iv) *If μ is not uniform, then the reconstruction method that selects the element a with the largest posterior probability, $\mu(a) \cdot \mathbb{P}(\mathbf{u}^{(k)} | \tilde{a} = a)$, achieves the maximum accuracy.*

Part (i) is an easy exercise, noting that $d_{KL}(\hat{p} || p_a) = C - \frac{1}{k} \ln \left(\prod_{i=1}^k p_a(u_i) \right)$ where C does not involve p_a . Part (ii) asserts that ML is statistically consistent in this discrete setting and is easily proven [342]. Part (iii) is a special case of (iv), which is a standard result in probability theory (see Theorem 17.2 of [165], or see [342]).

The distribution on U is sometimes determined not only by the discrete parameter $a \in \mathcal{A}$ but also by associated “nuisance” or hidden parameters, which are generally continuous (e.g., edge lengths in phylogenetics). In this case, Proposition 8.1 requires some refinement. Thus, suppose that each element a of A has an associated set $\Theta(a)$ of additional continuous parameters, and let $p_{(a, \theta)}$, $\theta \in \Theta(a)$, denote the associated probability distribution on U associated with the pair (a, θ) .

For a sequence of observations $\mathbf{u} \in U^k$, ML selects a pair (a, θ) (with $a \in \mathcal{A}, \theta \in \Theta(a)$) to maximize $\prod_{i=1}^k p_{(a, \theta)}(u_i)$. Suppose first that we are just interested in estimating a (i.e., θ is a “nuisance parameter”), in which case we first find the optimal pair (a, θ) and then ignore θ . Provided that the following *identifiability condition* holds for all $a, a' \in \mathcal{A}$ and $\theta \in \Theta(a)$:

$$\inf_{\theta' \in \Theta(a')} \{d_1(p_{(a, \theta)}, p_{(a', \theta')})\} = 0 \implies a = a', \quad (8.2)$$

(where “inf” refers to infimum), the ML procedure for estimating the discrete parameter a has an accuracy that converges to 1 as k grows.

Example. In the phylogenetic setting, condition (8.2) applies when $\mathcal{A} = B(n)$ and when $p_{(T,\theta)}$ is the probability distribution on 2-state characters on $[n]$ under the 2-state symmetric model on T with $\theta = (p_e : e \in E(T)) \in (0, 0.5)^{E(T)}$. However, condition (8.2) fails if we include nonbinary trees by taking $\mathcal{A} = P(n)$, since the probability distribution induced by a nonbinary tree T can be approximated arbitrarily closely by a binary tree T' that refines T with p_e values being sufficiently close to zero on each edge e of T' that corresponds to a split that is absent from T . Similarly, condition (8.2) fails if we take $\mathcal{A} = RB(n)$, and, in this case, for the stronger reason that one can have $d_1(p_{(a,\theta)}, p_{(a',\theta')}) = 0$ for $a \neq a'$ (i.e., two different placements of the root on the same tree can lead to identical probability distributions on patterns).

Despite this last example involving $A = P(n)$, notice that, in that case, we still satisfy a slightly different “identifiability” condition: For all $a, a' \in \mathcal{A}, \theta \in \Theta(a)$ and $\theta' \in \Theta(a')$,

$$d_1(p_{(a,\theta)}, p_{(a',\theta')}) = 0 \implies a = a' \text{ and } \theta = \theta'. \quad (8.3)$$

This is weaker than (8.2) if we remove the second condition ($\theta = \theta'$) but, in general, is neither weaker nor stronger than (8.2).

Sometimes θ is not really a “nuisance” parameter, and we wish to estimate the pair (a, θ) . We can no longer hope to do this exactly, since θ is typically continuous. However, given a metric d on $\Omega = \{(a, \theta) : a \in \mathcal{A}, \theta \in \Theta(a)\}$ and provided that $(a, \theta) \mapsto p_{(a,\theta)}$ is continuous, the second identifiability condition (8.3) implies that the distance under the metric d between the ML estimate (a', θ') calculated from $\mathbf{u}^{(k)}$ and the actual pair (a, θ) that generated $\mathbf{u}^{(k)}$ converges in probability to zero as k grows [83, 360].

8.2 ■ Phylogeny reconstruction methods and properties

In Section 7.2.1, we saw that it is possible to recover the underlying tree under the general Markov model (GMM). For instance, one can apply the LogDet transformation from sufficiently long sequences. While this is nice in theory, one practical drawback is that the logarithmic transformation can exhibit high variance unless the number of characters is very large. Therefore, more direct and standard statistical techniques, such as ML and Bayesian methods, are generally preferred.

The use of ML to reconstruct trees from discrete characters (e.g., DNA) was pioneered by Joseph Felsenstein, with a highly cited paper in 1981. Given a model \mathcal{M} of character evolution, the standard implementation of ML aims to find a pair (T, θ) that maximizes the probability of generating the observed k characters independently under the model \mathcal{M} with parameters (T, θ) . In other words, given a sequence $\mathcal{C} = (f_1, f_2, \dots, f_k)$ of characters on X , the pair (T, θ) is selected with the aim of maximizing the *likelihood score* $\prod_{i=1}^k p_{(T,\theta)}(f_i)$. The phylogenetic tree T in an optimal pair (T, θ) is referred to as a *maximum likelihood tree* (ML tree) for \mathcal{C} (in practice, the trees returned by ML software packages may not have a global maximal likelihood score, and so may differ from the true ML tree).

Section 8.1.2 provides sufficient conditions for the *statistical consistency* of ML in phylogenetics. First, if $\mathcal{A} = B(n)$ and if a model \mathcal{M} satisfies the identifiability condition

of (8.2), then the probability that the ML phylogeny matches the generating binary phylogeny converges to 1 as k grows.⁵⁰ Second, if \mathcal{A} is the larger set $P(n)$, and \mathcal{M} satisfies the second identifiability condition of (8.3), then the distance (under some metric d) between the generating tree with its associated continuous parameters, and the reconstructed ML tree with parameters, converges to zero as k grows. Thus the statistical consistency question hinges primarily on checking the relevant identifiability condition. For unrooted phylogenies, GMM satisfies condition (8.2) and so do all submodels of GMM. However, more general “mixture” models can fail to satisfy condition (8.2), as Proposition 7.12 showed.

Computing the likelihood score of the given data on a given tree (once the θ -parameters have been specified) is straightforward and can be carried out in polynomial time in the number of characters and number of leaves (cf. Section 7.2). However, finding the ML tree for the given data turns out to be NP-hard, even for the 2-state symmetric model [297].⁵¹

There are two variations on the “standard implementation” of ML. In the first, for each character $f_i : X \rightarrow S$, we could consider an extension F_i of f_i to all the vertices of T , and then select the pair (T, θ) that maximizes the joint probability of generating the extensions F_1, F_2, \dots, F_k independently on T . This method is called “ancestral ML” or “most-parsimonious likelihood.” It is rarely used, and has recently been shown to be statistically inconsistent [266]. By contrast, the standard implementation of ML effectively averages over all possible extensions (weighting each by its probability).

A second nonstandard approach is to allow θ to vary freely from character to character in the optimization step. This “no common mechanism” approach is also statistically inconsistent, and it turns out to be exactly equivalent to MP (on any sequence of r -state characters) when the model in this (modified) likelihood calculation is taken to be the fully symmetric r -state model. This follows directly from Proposition 7.6; for further details and related results, see [361] and [141].

Performance of reconstruction methods on different tree types. We saw at the end of the last chapter that MP is a statistically inconsistent estimator of phylogeny for the “Felsenstein (zone) tree” in Fig. 8.1(i) when the edge length l is sufficiently small in relation to L . Our proof was in the special case of the 2-state symmetric model; however, the same result holds for most other models.

By contrast, the statistical consistency of ML ensures that for a sequence $\mathcal{C}^{(k)}$ of k characters generated on the tree in Fig. 8.1(i), we have

$$\lim_{L \rightarrow \infty} \lim_{k \rightarrow \infty} \mathbb{P}(ML(\mathcal{C}^{(k)}) = ab|cd) = 1. \quad (8.4)$$

However, at finite sequence lengths, ML can also suffer from a form of “long branch attraction.” More precisely, for any given k and l , if L is made large enough, then when ML is applied to characters generated on the tree shown in Fig. 8.1(i), the method is more likely to return the tree $ac|bd$ rather than the correct phylogeny $ab|cd$. This was investigated recently in [286]. Indeed, a simple symmetry argument reveals the following variation of eqn. (8.4), where the order of the limits is reversed:

$$\lim_{k \rightarrow \infty} \lim_{L \rightarrow \infty} \mathbb{P}(ML(\mathcal{C}^{(k)}) = ab|cd) \leq \frac{1}{2}. \quad (8.5)$$

⁵⁰There is also a “strong law” version of this statement: with probability 1 there exists a number K so that for every $k \geq K$, ML will return the generating tree from $\mathbf{u}^{(k)}$.

⁵¹The complexity of finding optimal ML θ parameters for a fixed tree T and given data is currently unclear, even for the 2-state model.

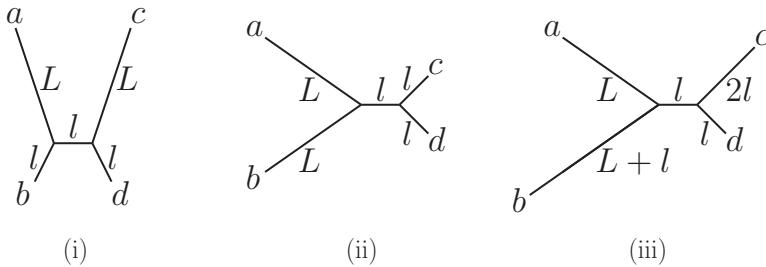


Figure 8.1. (i) The “Felsenstein zone” tree, (ii) the “Farris zone” tree, and (iii) the “twisted Farris zone” tree.

In fact, the simulations in [286] suggest that the value on the left is quite a bit smaller than $\frac{1}{2}$, though this seems less straightforward to establish analytically. Of course, if ML is replaced by MP in (8.5), then the corresponding limit is 0.

There are other ways to arrange “long and short” branches in a tree, and one is shown in Fig. 8.1(ii). Suppose that l is very small and L is large. In this case, MP will return the correct phylogeny $ab|cd$, and will do so from a small number of characters. For example, under the 2-state symmetric model, it is an easy exercise to show that

$$\lim_{l \rightarrow 0, L \rightarrow \infty} \mathbb{P}(MP(\mathcal{C}^{(k)}) = ab|cd) \geq 1 - \left(\frac{3}{4}\right)^k. \quad (8.6)$$

By contrast, ML will require many more characters than MP to recover the tree in Fig. 8.1(ii) correctly when L is large and l is small. In other words, while ML is statistically consistent and MP is not, this does not mean that ML has uniformly higher accuracy for inferring the phylogeny on every combination of phylogenies and edge lengths. Indeed, for the tree in Fig. 8.1(ii), simulations showed that the accuracy of ML initially declines with increasing k , leading to an incorrect published claim that ML was statistically inconsistent. The mathematical proof of the consistency of ML tells us that the accuracy must eventually climb towards 1 for large enough values of k , which was indeed confirmed in a subsequent simulation study.

We need to be careful with how we interpret eqn. (8.6). Although this might be seen as a desirable property of MP, it really reveals an inherent bias of this method. In the limit as $L \rightarrow \infty$, the character states at leaves a and b are completely random, so there is no reason to strongly support any particular phylogeny.

Exercise⁺: Prove inequalities (8.5) and (8.6) for the 2-state symmetric model.

The statistical consistency of tree reconstruction using ML or corrected distances assumes that the model that generated the data is identical to the model that is used in the tree reconstruction (by ML or corrected distances). But what if the two models differ? The consequences of “model misspecification” have been widely studied in molecular phylogenetics. We mention just one result here. Suppose that normalized Hamming distances for a sequence of characters evolved under a simple model (e.g., the equal input model) on a quartet tree are corrected under a model that is different from the model that generated the data. Then for either tree (i) or (iii) in Fig. 8.1, there exist values L and l so that neighbor joining (NJ) will return an incorrect tree in the limit as the number k of characters tends to infinity (for a more precise statement, and details, see [252]).

Bayesian methods and properties. Bayesian estimation techniques have become widespread in many areas of science, with software based on Markov chain Monte Carlo (MCMC) sampling, and other heuristics (e.g., approximate Bayesian computation (ABC)). Phylogenetics is no exception. Bayesian phylogenetic methods usually involve a prior distribution μ on phylogenetic X -trees and, for each tree T , a prior density function v_T on the continuous parameters θ associated with T under some model \mathcal{M} . Typically, T would be a sampled according to the Yule–Harding (YH) model studied in Chapter 3, and the edge lengths of T are sampled independently from an exponential distribution. The posterior probability of a phylogeny T given a sequence of characters $\mathcal{C} = (f_1, \dots, f_k)$ is then

$$\mu(T) \cdot \mathbb{E}_{v_T} \left[\prod_{i=1}^k p_{(T,\theta)}(f_i) \right] = \mu(T) \cdot \int \left(\prod_{i=1}^k p_{(T,\theta)}(f_i) \right) v_T(\theta) d\theta.$$

One advantage of Bayesian methods is that they can be used to estimate parameters without having to decide on a particular phylogenetic tree. For example, a biologist may be interested in estimating how many millions of years ago the lineages leading to birds and mammals separated, or whether *Amborella* is the sister group to all flowering plants. Questions like these can be studied by sampling a very large number of phylogenies (weighted by their posterior probability) and considering what each tree says about the question under consideration. In this sense, the tree is itself treated almost as a “nuisance parameter” to be averaged over.

Some interesting mathematical properties of Bayesian phylogenetics have been established. One is an analytical result from a paper in *Science* [270] showing that the mixing times of MCMC methods can grow exponentially with the number of characters under a simple phylogenetic mixture model. For this model, each character is generated independently according to a simple process (e.g., Jukes–Cantor (JC69)) on one of two phylogenies on the same set of $n \geq 5$ leaves, selected with equal probability, and with edge lengths for each tree chosen carefully.

A second area where mathematics has helped to clarify arguments that arose from inconclusive simulation studies concerns the so-called *Bayesian star “paradox.”* Suppose that we evolve k characters under a simple model such as the 2-state symmetric model on a star tree with a small number of leaves (e.g., three leaves if we assume a molecular clock; four leaves otherwise). Given this (random) data, let us now consider the posterior probabilities for each phylogeny under a prior μ that gives uniform probability to the three binary trees (and zero probability to the star tree), and with a prior v on edge lengths consisting of independent samples from an exponential distribution with a fixed mean. We might expect that as k becomes large, the posterior probability of each of the three trees would tend to $\frac{1}{3}$ with a probability that converges to 1 as k grows. However, it can be formally proven that this is not the case. Indeed, a stronger result holds: for any $\epsilon > 0$, the posterior probability of a particular binary tree is at least $1 - \epsilon$ with a probability⁵² that does not converge to 0 as $k \rightarrow \infty$. While this is not really a true paradox, and not particular to phylogenetics, the phenomenon is real, despite a published paper claiming otherwise. For further details, see [336, 349, 379].

For models that satisfy the usual identifiability condition (8.2), the maximum posterior probability phylogeny is a statistically consistent estimator of the binary tree parameter [331]. Further statistical properties of Bayesian phylogenetic methods, including the long branch attraction phenomenon that we mentioned for ML, have been explored in [350].

⁵²Recall that each data set of length k is generated on the star tree with a corresponding probability.

8.2.1 ■ Information-theoretic bounds

Each character that has evolved on some (perhaps unknown) phylogeny carries some information or “signal” about the identity of this tree, and about the possible ancestral state at some interior vertex (e.g., the root). It is possible to quantify this information mathematically in various ways (see e.g., [150, 268, 358]). In this section, we describe some results, which help address basic questions such as,

- How many randomly evolved characters are needed to accurately reconstruct a (binary) phylogenetic tree?
- When and how can we accurately estimate an ancestral state from the evolved states at the leaves?

These two questions are quite different—one concerns a sequence of characters; the other a single character—but we will see that there is a close link between them.

Trees can avoid the curse of information loss. We start with a general remark that is relevant to both questions. Under fairly general conditions, Markovian evolution can be regarded as an “information-destroying” process. More precisely, suppose that $X(t)$ is any (possibly nonhomogeneous) Markov process on a finite set of states S , and that for some fixed $l > 0$ and $\delta > 0$,

$$\mathbb{P}(X(t+l)=j|X(t)=i)\geq\delta$$

for all ordered pairs (i, j) of states from S , and all $t > 0$. In this case, the mutual information between the process at two different times, $I(X(t), X(t'))$, converges to zero at an exponential rate as $|t' - t|$ increases [320]. This condition applies for all of the models we have considered in the previous chapter. Thus a single leaf or even a fixed number of leaves in a tree will eventually be unable to estimate an ancestral state at a vertex that is sufficiently far in the past. However, a Markov process on a tree T , based on state space S , has one advantage over a Markov chain: the state space for the joint distribution of states at the leaf set X of a tree is S^X , so it can grow in size (exponentially) with the number of leaves of the tree. Thus, while each leaf provides vanishing information about the state at the root (or a vertex near it) as the depth of the tree grows, the collection of states at the leaves may still retain this signal even as the depth of the tree tends to infinity. These considerations may be particularly relevant to resolving divergences in the distant past (e.g., the origin of Metazoa (animals) ~ 550 – 600 million years ago, the origin of photosynthesis at least 2 billion years ago, or the earliest divergences within the “tree of life” ~ 3.5 billion years ago).

Let us return to the first question above (tree reconstruction from characters) for which several factors come into play. First there is the accuracy we require; we will assume throughout that a method is required to return the correct tree with probability at least $1 - \epsilon$ for some fixed (small) values of ϵ . Next, an interior edge e with a very short length $l(e) = f$ in a binary tree T will require a large number of characters, since the distance between the probability distribution on characters induced by T and by a tree T' obtained by an NNI rearrangement about e in T will converge to zero as $f \rightarrow 0$. Using Hellinger distance (cf. Section 8.1.1), it can be shown that sequence length (i.e., the number of characters) k required to reconstruct the tree (with all other edge lengths fixed) grows at the rate

$$k = \Theta\left(\frac{1}{f^2}\right) \text{ as } f \rightarrow 0. \quad (8.7)$$

Similarly, if T has a pendant edge e for which the length $l(e) = l$ is very long, then as l becomes large, the state at the leaf incident with e is effectively independent of the states at the remaining leaves of T . Consequently, the Hellinger distance between the induced probability distribution on the characters and that of any tree T' obtained from T by pruning and regrafting leaf x to some other location in T (using a new edge, also of length l) will converge to 0 as $l \rightarrow \infty$. It can thus be shown that the sequence length k required to reconstruct the tree (with all the other edge lengths fixed) grows at the rate

$$k = \Theta(e^{cl}) \text{ as } l \rightarrow \infty, \quad (8.8)$$

for some constant $c > 0$, dependent on the model.

A natural question is how the effects of short and long edges described by eqns. (8.7) and (8.8) interact. It turns out that the effect is multiplicative; in other words, if there is an interior edge of length f beside an exterior edge of length l then the required sequence length grows at the rate $k = \Theta\left(\frac{e^{cl}}{f^2}\right)$.

Exercise⁺: Consider the 2-state symmetric model on the quartet trees in $B(4)$, with $p_e = f$ for each edge e . Show that the number k of characters required to reconstruct each quartet tree accurately grows at the rate $k = \Theta\left(\frac{1}{f}\right)$ as $f \rightarrow 0$.

Dependence of k on the number n of species. The question of how the number n of species (leaves of T) affects the required number of evolved characters for accurate tree reconstruction is particularly interesting. In Chapter 5, we saw that for r -state characters (for a fixed r), the minimal number of characters needed to capture a binary tree $T \in B(n)$ (as the unique perfect phylogeny) grows linearly with n . But what if the characters have instead evolved on a tree under a stochastic process? In that case, we can ask how many characters k are required in order for each tree to be returned correctly with at least some given high probability value.

A very simple lower bound on k is as follows. Suppose that for each tree, we generate r -state characters according to some stochastic model—this model need not necessarily involve any Markovian or independence assumptions—all we require is that the number of characters (k) should be sufficiently large that each tree is returned by some method with a probability that is strictly greater than 0.5. It is easily shown that this requires the number of data sets (consisting of k r -state characters) to be at least as large as $b(n)$ (i.e., there needs to be at least as many data sets as there are binary trees).

Since there are r^n characters $f : [n] \rightarrow S$ where $|S| = r$ and since we can write $b(n) = 2^{n \log_2 n + O(n)}$ (cf. Chapter 2), we are led to the inequality

$$r^{nk} \geq B(n) = 2^{n \log_2 n + O(n)}.$$

Taking logarithms, this gives $k = \Omega(\log(n))$. In other words, under the most minimal assumptions, tree reconstruction would require at least a logarithmic growth in the number of characters with n in order to achieve better than 50% accuracy.

One of the remarkable discoveries of theoretical phylogenetics is that, at least for simple Markovian models (and independently generated characters), this trivial logarithmic lower bound growth in k is not only necessary but also sufficient for accurate tree reconstruction.⁵³

⁵³So, the number of “randomly evolved” r -state characters needed to infer a tree accurately can grow much more slowly than the minimal number of r -state characters required to capture the tree as defined in Chapter 5.

To make this more precise, consider the 2-state symmetric model, and suppose that each edge e of T has a substitution probability p_e that lies in the range $0 < f \leq p_e \leq g < \frac{1}{2}$. An early result from the late 1990s showed that for nearly all trees $T \in B(n)$, accurate tree reconstruction was possible for sequences of evolved binary characters of length $\varphi(n)/f^2$ where $\varphi(n)$ was proportional to $\log(n)$ for certain trees (e.g., caterpillars) and to a power of $\log(n)$ for nearly all trees in $B(n)$. However, for “worst” case trees (including, for example, perfect trees), $\varphi(n)$ was required to grow polynomially with n . These results relied on a combinatorial tool from Chapter 4, concerning “short quartets” (cf. Proposition 4.15 and the subsequent discussion).

The clue for how this might be improved came from an important paper [131] that studied the question of estimating the root state of a perfect rooted binary tree from the states at the $n = 2^h$ leaves under the symmetric 2-state model. Suppose that each edge has the same substitution probability p . As the height of the tree increases, each leaf provides progressively less information about the root state; however, the number of leaves is also increasing exponentially with h . Which of these two competing factors wins out depends on whether p is greater or less than a critical value:

$$p_+ = \frac{1}{2} \left(1 - \frac{1}{\sqrt{2}} \right) \approx 0.146.$$

When $p > p_+$, the mutual information between the root state and the collection of leaf states converges to 0 with increasing tree height (we explain why shortly). This implies, by Fano’s lemma, that the ancestral state at the root cannot be recovered with an accuracy that is better than the toss of a fair coin in the limit as h tends to infinity. However, for $p < p_+$, information is retained.⁵⁴

This led to the following conjecture. Suppose that k characters are generated under the 2-state model on a binary tree for which the substitution probabilities lie in the range $0 < f \leq p_e \leq g$. Provided that $g < p_+$ the value of k that is both necessary and sufficient for reconstructing each tree in $B(n)$ correctly with probability at least $1 - \epsilon$ (where $\epsilon > 0$ is any fixed value) is

$$k = \Theta\left(\frac{\log n}{f^2}\right). \quad (8.9)$$

Here, the constant implicit in Θ depends on g and ϵ .

This conjecture turned out to be particularly difficult to resolve but was eventually proven in [97], based on a novel tree reconstruction algorithm. By contrast, the popular NJ method described in Chapter 6 would require k to grow exponentially with n to achieve comparable accuracy [223]. However, a more subtle distance-based tree reconstruction method has recently been shown to be accurate when k is given by eqn. (8.9) [298]. When $p > \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)$, the logarithmic dependence of k on n in eqn. (8.9) changes to polynomial dependence. ML is also accurate when k is given by eqn. (8.9), but this result is far from trivial and was only recently established in [299].

The reader may be curious where the mysterious critical value $p_+ = \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)$ comes from. Again, information theory is helpful. Consider the 2-state symmetric model on a binary tree $T \in R(n)$ and the mutual information $I(Y_\rho, (Y_j : j \in [n]))$ between the root state Y_ρ (determined by the toss of a fair coin) and the collective assignment of the states $(Y_j : j \in [n])$ at the leaves of T . In [131], it was shown that if we make all the paths from

⁵⁴Using maximum parsimony to estimate ancestral states on a perfect binary trees under the symmetric 2-state model has a smaller transition value at $p = \frac{1}{8}$.

the root to the leaves disjoint (i.e., convert T to a star tree and keep the expected number of substitutions down each path to a leaf the same) then the resulting mutual information between the root state and the collective leaf states does not decrease.⁵⁵ From this, it follows (by a standard inequality in information theory) that

$$I(Y_\rho, (Y_j : j \in [n])) \leq \sum_{j=1}^n I(Y_j, Y_\rho). \quad (8.10)$$

For $j \in [n]$, let $\omega_j = \prod_{e \in P(T; \rho, j)} (1 - 2p_e)$. As we saw in the examples at the start of Chapter 7, the probability that the root ρ is in the same (respectively, different) state to leaf j is given by $\frac{1}{2}(1 + \omega_j)$ (respectively, $\frac{1}{2}(1 - \omega_j)$). It now follows from the inequality $\frac{1}{2}[(1+x)\ln(1+x) + (1-x)\ln(1-x)] \leq x^2$ applied to $x = \omega_j$ that

$$I(Y_\rho, Y_j) \leq \omega_j^2. \quad (8.11)$$

Thus, if T is a perfect binary tree of height h , for which each edge has $p_e = p$, then since T has $n = 2^h$ leaves, eqns. (8.10) and (8.11) combine to give

$$I(Y_\rho, (Y_j : j \in [n])) \leq n[(1-2p)^h]^2 = [2(1-2p)^2]^h.$$

Notice that the term on the right of this last inequality converges to zero precisely when $2(1-2p)^2 < 1$; in other words, at the critical value p_+ . Therefore, when $p > p_+$, information concerning the root state is asymptotically lost as $h \rightarrow \infty$.

Estimating ancestral states. While we have focussed on tree reconstruction, there are many other inference questions in phylogenetics. For example, given some proposed phylogenetic tree T for a set of species, we may wish to estimate the state at some ancestral vertex of T by using the value of a character f on the leaf set X of T . If this character has evolved according to some stochastic process on T and we know the continuous parameters of that model (e.g., the edge lengths of T), and all ancestral states have equal prior probability, then Proposition 8.1(iii) (with $k = 1$) assures us that an ML-based approach has optimal reconstruction accuracy.

What if we just know the tree T and little about the model? To simplify the following discussion let us suppose the character evolved under a fully symmetric model. Then Proposition 7.6 shows that the ML estimate of an ancestral state (in which the edge lengths are now treated as nuisance parameters in the ML optimization) coincides with a MP estimate.⁵⁶ It is instructive to compare the performance of MP with a much simpler method, *majority rule* (MR), which estimates the ancestral state at the vertex by the most frequently occurring state among the descendant leaves.

For example, consider a rooted caterpillar tree T_n with n leaves, and ultrametric edge lengths, with h being the height of the tree (the length from the root to each leaf) and l being the sum of the interior edge lengths, as in Fig. 8.2(i). For MP, the estimation of the root state is dominated by the pendant edges close to the root (e.g., if the two pendant edges nearest to the root of T_n lead to leaves that are in a different state from the root state, then MP will estimate the root state incorrectly). Assuming the character evolves under a fully symmetric model on r states, the accuracy of MP as $n \rightarrow \infty$ and with h and l fixed, can be shown to be $\frac{1}{r} + O(e^{-ch})$ for a constant $c > 0$. However, the simpler

⁵⁵This does not extend to the fully symmetric model on an arbitrary number of states.

⁵⁶This form of ML is no longer one that has highest reconstruction accuracy on all trees.

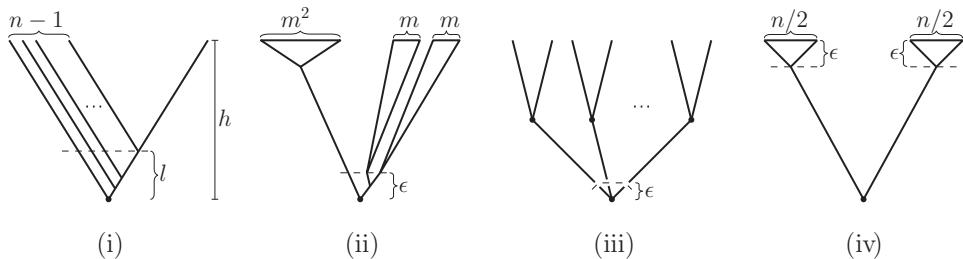


Figure 8.2. Under the symmetric model on r -states (i) shows a tree for which majority rule (MR) can have high accuracy for estimating the root state but MP does not while, conversely, (ii) shows tree where MP can have high accuracy for estimating the root state but MR does not (see text for details). (iii) When all of the edges are long and n is sufficient large, then none of the other interior vertices of this tree can be accurately estimated (by any method) while the root state can be. (iv) Conversely, when the two edges incident with the root are sufficiently long, and the other edges are sufficiently short, then the root state cannot be accurately estimated (by any method), but all of the other interior vertices can be.

method MR has an accuracy that for a fixed h converges to 1 as $n \rightarrow \infty$ and $l \rightarrow 0$ (by the central limit theorem). In other words, by selecting n and h to be sufficiently large and $l > 0$ to be sufficiently small, there is a tree for which MP has an accuracy that is close to $\frac{1}{r}$, while MR has an accuracy close to 1.

It is a little surprising that a method that ignores the tree structure entirely (MR) can have higher accuracy in estimating an ancestral state than MP, which explicitly takes the tree structure into account. Moreover, we will see in Chapter 9 that this can occur for “typical” trees rather than just for the contrived setup in our last example. However, MR can also perform more poorly than MP trees. An example is shown in Fig. 8.2(ii). Here, the trees with m leaves are perfect trees and m is large enough relative to the height h of the tree that each edge in these two trees has a sufficiently small substitution probability, and h is large enough that the expected number of substitutions from the root to the subtree with m^2 leaves is large [151].

Estimating the state at an ancestral vertex v accurately from the states at the leaves involves a subtle trade-off between the number of leaves n , the height of the tree h , and the distribution of edge lengths. It is possible to define a notion of statistical consistency (where it is now the number of leaves n that is tending to infinity, with a fixed value of h , and some assumption is placed on the distribution of edge lengths). This was studied in [150] where the estimation technique depends on the type of model considered.

Finally, it might be suspected that ancestral state estimation will be least accurate for the deepest vertices, such as the root. However, Fig. 8.2(iii) shows that this need not be the case; though, of course, it sometime is (as shown in part (iv) of the figure).

8.2.2 • The space of “phylogenetic oranges”

If one views phylogenies with edge lengths from the perspective of the probabilities they confer on characters under Markov-type models, then the resulting “tree space” looks quite different from the BHV tree space discussed in Chapter 6. It is still true that when certain interior edge lengths collapse to zero, trees that differ on the corresponding splits can look identical; however, at the other extreme—when various (interior and/or pendant) edges become very long—different phylogenies can also look very similar (rather than looking more and more different as in the case for the BHV space). This concept was discussed in a prescient paper by the biologist Junhyong Kim in 2000, who drew the

analogy with an orange: the “segments” (i.e., different phylogenies) all meet at a point (where all edges are of zero length) and also at an antipodal point (where all edges have infinite length) [217]. A formal definition and analysis of the topological structure of this tree space was explored further in [272] and [155]; here we give a tailored account here.

For convenience, we will consider a continuous-time Markov process, that is stationary and time-reversible, with an otherwise arbitrary but fixed rate matrix Q , and stationary distribution π . For definiteness, we could consider the r -state symmetric model; however, the topology of the space we describe is the same for any other model of the type described operating on a tree. Thus for each tree T with edge length assignment l , we have a corresponding probability distribution $p_{(T,l)}$ on characters. In the “phylogenetic orange” space, we view (T, l) and (T', l') as being “close” if $p_{(T,l)}$ is close to $p_{(T',l')}$ under some natural metric on probability distributions, such as d_1 or d_H .

It helps at this point to let $\omega_e = e^{-l(e)}$ for each edge e of T so that we can reparametrize this distribution by writing $p_{(T,\omega)}$. Notice that $l(e) \rightarrow \infty$, we have $\omega_e \rightarrow 0$, and as $l(e) \rightarrow 0$, then $\omega_e \rightarrow 1$. To investigate what happens at these limits we allow these limiting values ($\omega_e = 0$ and 1).

For a fixed tree T the map $\omega \mapsto p_{T,\omega}$ is continuous and one-to-one on $(0, 1]^{E(T)}$, but is no longer one-to-one if we extend the domain of ω to $[0, 1]^{E(T)}$ (i.e., allow $\omega_e = 0$, which corresponds to the limiting value $l(e) \rightarrow \infty$). Also, although the images of the maps $\omega \mapsto p_{(T,\omega)}$ are disjoint across trees when we restrict ω to be in $(0, 1)^{E(T)}$, these images can overlap (for different trees) if we allow $\omega_e = 0$ or $\omega_e = 1$.

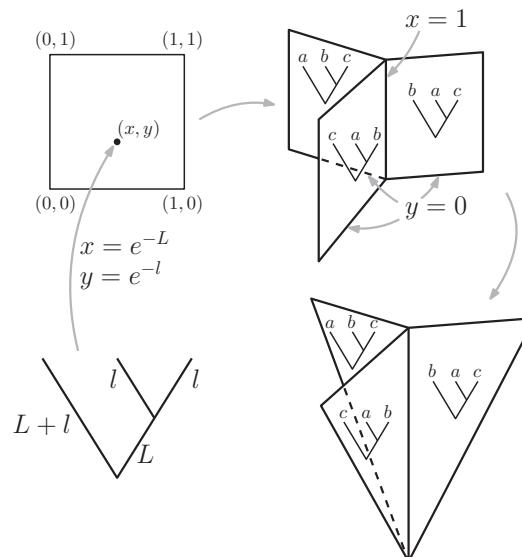


Figure 8.3. When edge lengths are constrained to be ultrametric, the associated “phylogenetic orange” space on three species corresponds to a “paper dart” cell complex.

It is easiest to visualize this in the ultrametric setting (which will assume here only for this visualization example, before lifting this constraint). Consider the three rooted triples $ab|c$, $ac|b$ and $bc|a$, with ultrametric edge lengths, with length L for the interior edge and length l for a pendant edge in a cherry (see Fig. 8.3). For each tree T , the composition function $(L, l) \mapsto (x = e^{-L}, y = e^{-l}) \mapsto p_{(T,(x,y))}$ maps the open square $(0, \infty)^2$ homeomorphically, and the three images (open squares) are disjoint across the three trees.

But when $L = 0$ the three images are “glued together” in Fig. 8.3 along a shared boundary (corresponding to $x = 1$). Moreover, as $l \rightarrow \infty$, the three trees produce probability distributions that converge on the same distribution (regardless of L), namely the distribution in which all three leaves are independently assigned states from the stationary distribution π . This results in the “pinching” to a point of the three lines that correspond to $y = 0$, thereby producing the final “paper dart” space in Fig. 8.3.

Let us now remove the restriction that the edge lengths are ultrametric. In that case, the probability distribution on characters is based on the single unrooted phylogeny on three leaves and is now three-dimensional. Although there is only one tree, there is still some identification (“pinching”) that occurs when one or more of the ω_e values in the three-dimensional cube converge to zero.

More generally, for any phylogeny $T \in P(n)$ let $N = |S|^n$, and let

$$o_T : [0, 1]^{E(T)} \rightarrow \mathbb{R}^N,$$

$$(T, \omega) \mapsto p_{(T, \omega)}$$

assign (T, ω) to the resulting probability distribution on characters. Now let \mathbb{O}_T be the image of o_T ; in other words, the set $\{p_{(T, \omega)} : \omega \in [0, 1]^{E(T)}\}$ of all probability distributions arising from the Markov process on T over all values for the ω_e parameters, including their limit values $\omega_e = 0$ and $\omega_e = 1$. This is the *phylogenetic orange space* for T . It is clear that \mathbb{O}_T is compact (being the continuous image of a compact space). It is also contractable, since the map

$$F : \mathbb{O}_T \times [0, 1] \rightarrow \mathbb{O}_T,$$

$$F(p_{(T, \omega)}, s) = p_{(T, (1-s)\omega)}$$

is a homotopy that fixes each point of \mathbb{O}_T at $s = 0$ and collapses \mathbb{O}_T to a single point at $s = 1$, namely $p_{(T, \omega^*)}$, where $\omega^*(e) = 0$ for all edges e of T (this corresponds to all edges having infinite length, and so each leaf is independently assigned a random state from the stationary distribution π). We can also consider the union of \mathbb{O}_T over all trees in $P(n)$, which is the *space of phylogenetic oranges*, denoted \mathbb{O}_n . The space \mathbb{O}_n is also identical to the union of \mathbb{O}_T over just the trees in $B(n)$, since setting $\omega_e = 1$ is equivalent to collapsing edge e . Notice that (T, ω) and (T', ω') are mapped to the same point $p \in \mathbb{O}_n$ if and only if $p_{(T, \omega)} = p_{(T', \omega')}$. It is not hard to show that this holds precisely if and only if

$$\prod_{e \in P(T; x, y)} \omega_e = \prod_{e \in P(T'; x, y)} \omega'_e,$$

for all $x, y \in [n]$. This gives an alternative topological description of \mathbb{O}_n , called the “edge-product space,” which allows for a more tractable mathematical analysis. The topology of these spaces is particularly interesting. Here is a sample of three properties:

- (i) \mathbb{O}_T is homeomorphic to a closed ball.
- (ii) \mathbb{O}_n is a regular CW complex and is homeomorphic to the geometric realization of a poset of X -forests (the *Tuffley poset*).
- (iii) For any point in $p \in \mathbb{O}_T$, the subset $o_T^{-1}(p)$ of $[0, 1]^{E(T)}$ is a contractible regular CW complex.⁵⁷

⁵⁷When $p = p_{(T, \omega^*)}$ this space has dimension equal to the number of interior vertices of T .

Part (ii) provides a precise combinatorial description of the topology of \mathbb{O}_n once we define the Tuffley poset.⁵⁸ Recall from Section 1.3 the definition of an X -tree, which can also be viewed as a collection of pairwise compatible X -splits. An X -forest \mathbb{F} is simply a collection of X_i -trees, for $i = 1, \dots, k$, where $k \geq 1$ and where the sets X_1, \dots, X_k partition X . Given two X -forests \mathbb{F} and \mathbb{F}' , we write $\mathbb{F} \prec \mathbb{F}'$ if \mathbb{F} can be obtained from \mathbb{F}' by a sequence of operations, each of which either collapses an edge (with the label sets of the identified vertices combined) or deletes an edge (and suppresses any resulting unlabelled vertices of degree 2). For example, the cover digraph of the Tuffley poset for $X = \{1, 2, 3\}$ is shown in Fig. 8.4.

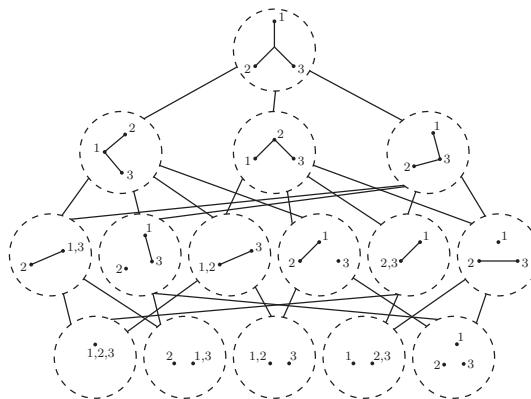


Figure 8.4. The Tuffley poset of X -forests for $X = \{1, 2, 3\}$.

In general, the Tuffley poset has a number of desirable properties: It is pure (i.e., all maximal chains are of the same length), it is graded (by the number $\rho(\mathbb{F})$ of edges in the forest \mathbb{F}), the maximal elements correspond to the trees in $B(X)$, it is “thin” (i.e., all intervals of length two contain exactly four elements), it has a “recursive coatom ordering” (the combinatorial analogue of shellability), and its Möbius function is described by $\mu(\mathbb{F}, \mathbb{F}') = (-1)^{\rho(\mathbb{F}) - \rho(\mathbb{F}')}$. For more details, see [126, 155, 272, 387].

The geometry of \mathbb{O}_T and \mathbb{O}_n is also of interest, but this has been less explored than the topology. The geometry will be influenced, at least in part, by the type of Markov process considered and the metric used on the probability distribution on characters. Some clues as to the l_2 (Euclidean) geometry of \mathbb{O}_T in the case of the symmetric 2-state model is provided by eqn. (7.28).

8.3 • Algebraic analysis of Markov models

One of the most interesting mathematical developments in the study of Markov processes on trees has been the application of techniques from commutative algebra and algebraic geometry to study and classify “phylogenetic invariants.” This is part of a broader emerging area called “algebraic statistics” [122].

For any Markov processes on a tree, T , the resulting probability distribution p on characters satisfies various polynomial equations that are independent of the underlying continuous parameters of the model (e.g., transition matrices and π). These invariants, are not only of theoretical interest; they can also be used to establish further identifiability results that ensure the consistency of methods like ML [7, 10]. Moreover, invari-

⁵⁸Named after Christopher Tuffley, who defined this poset and investigated its properties in his MSc thesis.

ants have recently also shown considerable promise as an alternative way to reconstruct a phylogenetic tree which (unlike ML) does not attempt to estimate the continuous model parameters (e.g., those related to the transition matrices) [137]. Here, we provide only a brief introduction to this active field of research, beginning with a short summary of some key definitions and classical results from commutative algebra.

Ideals and varieties. Given a field k (e.g., $k = \mathbb{R}$ or \mathbb{C}), consider the ring $k[x_1, \dots, x_n]$ of polynomials in the variables x_1, \dots, x_n with coefficients in k . A subset I of this ring is said to form an *ideal* if it satisfies the following two properties:

- If $g, h \in I$, then $cg + c'h \in I$ for any $c, c' \in k$.
- If $g \in I$ and $h \in k[x_1, \dots, x_n]$, then $gh \in I$.

One important example of an ideal is the set of polynomials in $k[x_1, \dots, x_n]$ that vanish on every point in some subset W of k^n . We will let $I(W)$ denote this ideal. Thus

$$I(W) = \{g \in k[x_1, x_2, \dots, x_n] : g(\mathbf{x}) = 0 \text{ for all } \mathbf{x} \in W\}. \quad (8.12)$$

Two classical results concerning the ideals of polynomial rings over fields are the following. *Hilbert's basis theorem* for polynomial rings states that any ideal I of $k[x_1, \dots, x_n]$ has a finite set of generators. That is, there exists a finite subset $\{g_1, g_2, \dots, g_N\}$ of $k[x_1, \dots, x_n]$ for which every element g of I can be written in the form $g = \sum_{i=1}^N h_i g_i$ for $h_i \in k[x_1, \dots, x_n]$. Gröbner basis techniques provide a way to calculate a finite set of generators for a given ideal.

For the second result, we will suppose that k is algebraically closed (e.g., $k = \mathbb{C}$). Given a subset P of polynomials from $k[x_1, \dots, x_n]$, the set

$$V(P) = \{\mathbf{x} \in k^n : g(\mathbf{x}) = 0 \text{ for all } g \in P\}$$

is an *algebraic variety* (or “algebraic set”). Notice that $V(P)$ is equal to $V(I_P)$ where I_P is the ideal of $k[x_1, \dots, x_n]$ generated by P .

For an ideal J of $k[x_1, \dots, x_n]$, consider the associated algebraic variety $V(J)$ and the resulting ideal $I(V(J))$ of polynomials $g \in k[x_1, \dots, x_n]$ that vanish on every point in $V(J)$. A first guess might be that these two ideals of $k[x_1, \dots, x_n]$ (J and the resulting $I(V(J))$) are identical but this turns out to be not quite right. It is true that $J \subseteq I(V(J))$; however, the containment can be strict. For example, if $n = 1$ and $J = \langle x_1^2 \rangle$ (in which case $J = \{x_1^2 b : b \in k[x_1]\}$), then $V(J) = \{0\}$ and so $I(V(J)) = \langle x_1 \rangle$. *Hilbert's Nullstellensatz* states that $I(V(J))$ consists precisely of those $g \in k[x_1, \dots, x_n]$ for which there exists some positive integer p with $g^p \in J$.

8.3.1 • Phylogenetic invariants and inequalities

Let $k[\mathbf{x}]$ be the set of polynomials with coefficients in a field k ($= \mathbb{R}$ or \mathbb{C}) and variables x_f indexed over all characters $f : X \rightarrow S$. Let \mathcal{M} be any class of Markov models on trees (or mixtures of Markov models) and, for each character $f : X \rightarrow S$, let $p(f|T, \theta)$ be the probability of generating character f on T under model \mathcal{M} with the associated parameter θ (the transition matrix entries, and any other continuous parameters needed to specify \mathcal{M}). The polynomial $g(\mathbf{x}) \in k[\mathbf{x}]$ is said to be a *phylogenetic invariant* for T under model \mathcal{M} if, for all continuous parameters θ , if we set $x_f = p(f|T, \theta)$ for all $f : X \rightarrow S$, then $g(\mathbf{x}) = 0$. In other words, g is a polynomial that vanishes whenever we substitute in the probability distribution on characters arising from T for any choice of the continuous parameters for the model.

Example: For the 2-state symmetric model on the quartet tree $T = 12|34$, eqn. (7.27) provides two quadratic identities

$$\omega_{\{1,3\}}\omega_{\{2,4\}} - \omega_{\{1,4\}}\omega_{\{2,3\}} = 0 \text{ and } \omega_{\{1,2,3,4\}} - \omega_{\{1,2\}}\omega_{\{3,4\}} = 0,$$

where each ω value is a linear combination $p(f|T, \theta)$ values. These two equations provide a pair of quadratic phylogenetic invariants for T under this model.

Notice that the set of phylogenetic invariants for T under a given model \mathcal{M} forms an ideal—the *phylogenetic ideal* for T , denoted I_T —and we can view this ideal in the context of eqn. (8.12) as the set $I_T = I(\text{Im}\phi_T)$ where $\text{Im}\phi_T$ is the image of the function ϕ_T that maps θ to $p_{(T,\theta)}$. The associated *phylogenetic variety* is the set $V_T = V(I_T)$ of points in \mathbb{C}^N that are zero on all the phylogenetic invariants for T . V_T contains $\text{Im}\phi_T$ as a strict subset.

There has been considerable interest in constructing and classifying phylogenetic invariants, and in exploring the properties of phylogenetic ideals and varieties, using computational algebra techniques (e.g., Gröbner basis algorithms). In general, the phylogenetic ideal for a given model and tree can grow very quickly with n [78]; for example, for the simple K3ST model on a quartet tree, I_T has a minimal generating set of 8002 polynomials! However, for an arbitrary tree T and model \mathcal{M} , one can sometimes find a much smaller set S_T of polynomials that define the algebraic variety V_T set-theoretically (i.e., $V(S_T) = V_T$) and for applications, this may be sufficient [76] (indeed, for the K3ST model on a quartet tree that we just discussed, only 48 invariants suffice to define V_T set-theoretically).

Model invariants and topology invariants. If g is a phylogenetic invariant for every phylogenetic tree on X , then g is said to be a *model invariant*. We will call a phylogenetic invariant of T that is not a model invariant a *topology invariant* (for T) since it can potentially be used to exclude certain tree topologies (i.e., phylogenetic trees $T' \neq T$, ignoring transition matrix entries) as possible candidates for generating the observed data. The two quadratic phylogenetic invariants described above for the 2-state symmetric model on $T = 12|34$ are both examples of topology invariants.

Exercise: Suppose that g is a phylogenetic invariant for $T \in P(X)$ under model \mathcal{M} , and that $T' \in P(X)$ is refined by T . Show that g is also a phylogenetic invariant for T' . Deduce that every phylogenetic invariant is valid also for the star tree.

The LogDet identity from Section 7.2.1 provides a phylogenetic invariant for the most general Markov process on a tree, the GMM. It follows from the four-point condition that if T , when restricted to the set of leaves $\{i, j, k, l\}$, gives rise to either the quartet tree $ij|kl$ or a star tree, then

$$\det J^{ik}J^{jl} - \det J^{il}J^{jk} = 0.$$

This is a phylogenetic invariant for the GMM on T , since the entries of the J matrices are linear functions of p and the matrix determinant is a polynomial function of the entries in the matrix. This invariant is a multilinear function of the x_f values (i.e., is degree 1 in each variable). Moreover, it is a topology invariant, and has degree $2r$ where $r = |S|$. We next consider low-degree phylogenetic invariants.

Linear phylogenetic invariants. We will say that a phylogenetic invariant $g(\mathbf{x})$ in $\mathbb{R}[\mathbf{x}]$ for T under some model is *linear* if each term involves just one variable, and it has degree one (i.e., $g(\mathbf{x}) = \sum_i a_i x_i$). Linear invariants for a model \mathcal{M} are of particular interest because they also vanish if the process from \mathcal{M} varies between characters (i.e., phylogenetic mixtures, described in Chapter 7, and which we will discuss further shortly).

The *trivial invariant* $g(\mathbf{x}) = x_1 + x_2 + \cdots + x_N - 1$ is a phylogenetic invariant for all trees and all models, but, technically, it is not a linear one as defined here, because it has a term (-1) of order zero. But this is essentially the only exception, since $\sum_i a_i x_i + c$ is a phylogenetic invariant (for any tree and model) if and only if $\sum_i (a_i + c)x_i$ is, and so all phylogenetic invariants of the first type can be described by homogeneous ones of the second type (i.e., linear invariants under the definition here).

The set \mathcal{L}_T of linear invariants g of T under model \mathcal{M} now forms a vector space (i.e., if g and g' are in \mathcal{L}_T , then so is $c_1 g + c_2 g'$ for all constants c_1, c_2). An obvious question asks what the dimension of this space is for different models. For the JC69 model, for instance, if $T \in B(n)$, then

$$\dim[\mathcal{L}_T] = 4^n - F_{2n-1},$$

where F_k is the k th Fibonacci number (with $F_1 = F_2 = 1$) [333]. For the subspace of model invariants, $\mathcal{L}_n := \bigcap_{T \in P(n)} \mathcal{L}_T$, the dimension is given by

$$\dim[\mathcal{L}_n] = \frac{1}{6}(23 \cdot 4^{n-1} - 3 \cdot 2^{n-1} - 2).$$

For the fully symmetric model, there are numerous linear relationships among the probabilities of different characters. If we write $p_{\alpha_1 \alpha_2 \dots \alpha_n}$ for the probability that the leaves $1, 2, \dots, n$ are in the states $\alpha_1, \alpha_2, \dots, \alpha_n$, then it is clear that $p_{\alpha\alpha\beta} - p_{\beta\beta\alpha} = 0$, for example. These linear phylogenetic invariants hold regardless of the underlying tree T , so they are model invariants. In other words, they reflect symmetries in the underlying model without providing any information about the underlying tree T . Thus it is of interest to ask whether there are topology invariants for the fully symmetric model or more general models. When $|S| = 2$, there are none; however, for larger state spaces, they exist. An important example is the linear invariants discovered by James Lake in 1987 [224] for $|S| = 4$. Suppose that S contains (at least) four distinct states $\alpha, \beta, \gamma, \delta$ and consider the linear function

$$L(x) = x_{\alpha\beta\alpha\beta} + x_{\alpha\beta\delta\gamma} - x_{\alpha\beta\delta\beta} - x_{\alpha\beta\gamma\alpha}. \quad (8.13)$$

If T is the quartet tree $12|34$ (or the star tree) on $\{1, 2, 3, 4\}$, then $L(p) = 0$ whenever p is the probability distribution arising on T for any selection of edge parameters under the r -state symmetric model for $r \geq 4$ (e.g. the JC69 model); the same holds for more general models with certain symmetries, as we will see shortly.

By contrast, if T is one of the other two quartet trees on $\{1, 2, 3, 4\}$, then $L(p) \neq 0$ in general, so L is a topology invariant for $12|34$ (and for the star tree). There is a second linear topology invariant for $12|34$ or the star tree, namely

$$L'(x) = x_{\alpha\beta\beta\alpha} + x_{\alpha\beta\gamma\delta} - x_{\alpha\beta\beta\delta} - x_{\alpha\beta\gamma\alpha}.$$

To better understand eqn. (8.13), notice that all four terms in L are of the form $x_{\alpha\beta**}$, and that in T ($12|34$ or the star tree on $\{1, 2, 3, 4\}$) the states α and β label a cherry. Also, if v is the interior vertex of T that is adjacent to leaves 3 and 4, then the Markov assumption implies that the joint states at leaves 1 and 2, the state at leaf 3, and the state at leaf 4 become

conditionally independent once we specify the state at v . Thus we can write $p(f)$ as

$$\sum_{s \in S} \left[\mathbb{P}(Y_1 = f(1) \wedge Y_2 = f(2) | Y_v = s) \prod_{i=3}^4 \mathbb{P}(Y_i = f(i) | Y_v = s) \right] \mathbb{P}(Y_v = s).$$

Consequently, we can write

$$L(p) = \sum_{s \in S} \Delta_s \cdot \mathbb{P}(Y_1 = \alpha \wedge Y_2 = \beta | Y_v = s) \cdot \mathbb{P}(Y_v = s), \quad (8.14)$$

where Δ_s equals

$$\begin{aligned} & \mathbb{P}(Y_3 = \alpha | Y_v = s) \mathbb{P}(Y_4 = \beta | Y_v = s) + \mathbb{P}(Y_3 = \delta | Y_v = s) \mathbb{P}(Y_4 = \gamma | Y_v = s) \\ & - \mathbb{P}(Y_3 = \delta | Y_v = s) \mathbb{P}(Y_4 = \beta | Y_v = s) - \mathbb{P}(Y_3 = \alpha | Y_v = s) \mathbb{P}(Y_4 = \gamma | Y_v = s). \end{aligned}$$

A case analysis now shows that for each $s \in S$, $\Delta_s = 0$ holds under an equal input model on $r \geq 4$ states when $\pi_\alpha = \pi_\delta$ and $\pi_\beta = \pi_\gamma$ (e.g., the fully symmetric model). It can also be shown that $\Delta_s = 0$ for the Kimura 2ST model (for certain choices, such as $\alpha = A$, $\beta = C$, $\gamma = G$, $\delta = T$), so L vanishes on p by eqn. (8.14). For details, and extensions, see [79].

Flattening. We now describe an idea that has provided an important source of phylogenetic invariants. Elizabeth Allman and John Rhodes developed this idea (following an early suggestion by Bernd Sturmfels), and used it efficiently to solve a variety of otherwise formidable identifiability questions.

Suppose that $e = \{v, v'\}$ is an edge of a phylogeny on X and let $A|B$ be the split of T that corresponds to e . For a Markov process on T , let p be the probability distribution on characters, and let $\text{flat}_{A|B}(p)$ be the matrix with rows indexed by characters $f : A \rightarrow S$ and columns indexed by characters $g : B \rightarrow S$, and with the (f, g) entry being the probability of generating character f on A and g on B according to the probability distribution p .

For example, for the quartet tree $T = 12|34$ and the GMM model on two states (say 0 and 1), if we take the split $A|B = \{1, 2\}|\{3, 4\}$ then $\text{flat}_{A|B}(p)$ is a 4×4 matrix:

$$\text{flat}_{\{1,2\}|\{3,4\}}(p) = \begin{bmatrix} p_{0000} & p_{0010} & p_{0001} & p_{0011} \\ p_{1000} & p_{1010} & p_{1001} & p_{1011} \\ p_{0100} & p_{0110} & p_{0101} & p_{0111} \\ p_{1100} & p_{1110} & p_{1101} & p_{1111} \end{bmatrix}.$$

In general, if $|S| = r$, then $\text{flat}_{A|B}(p)$ is a $r^{|A|} \times r^{|B|}$ matrix. Such a matrix would typically have rank⁵⁹ equal to $\min\{r^{|A|}, r^{|B|}\}$. Crucially, however, $\text{flat}_{A|B}(p)$ has small rank, as follows.

Proposition 8.2. *Consider the GMM with state space S on a phylogeny T . If $A|B$ is a split of T then $\text{flat}_{A|B}(p)$ has rank at most $r = |S|$.*

Proof: By Lemma 7.1, we can write the (f, g) entry of $\text{flat}_{A|B}(p)$ as follows:

$$\sum_{\alpha \in S} \mathbb{P}(Y_v = \alpha) \cdot \mathbb{P}(\wedge_{x \in A} \{Y_x = f(x)\} | Y_v = \alpha) \cdot \mathbb{P}(\wedge_{x \in B} \{Y_x = g(x)\} | Y_v = \alpha). \quad (8.15)$$

⁵⁹The *rank* of a matrix is size of the largest subset of linearly independent rows (or, equivalently, columns).

Let P_A be the matrix with rows indexed by characters $f : A \rightarrow S$ and columns indexed by states α of S , and with $P_A = [\mathbb{P}(\wedge_{x \in A}\{Y_x = f(x)\}|Y_v = \alpha)]$. Define a matrix P_B with rows indexed by characters $g : B \rightarrow S$ similarly (i.e., $P_B = [\mathbb{P}(\wedge_{x \in B}\{Y_x = g(x)\}|Y_v = \alpha)]$). Equation (8.15) can then be written more simply as a matrix factorization of this simpler “flattened” stochastic process, as follows:

$$\text{flat}_{A|B}(p) = P_A^T D_v P_B, \quad (8.16)$$

where T denotes matrix transpose, and D_v is the diagonal $r \times r$ matrix with rows and columns indexed over S and with an α entry equal to $\mathbb{P}(Y_v = \alpha)$. Thus $\text{flat}_{A|B}(p)$ is a product of three matrices, so, by elementary linear algebra, its rank is, at most, the largest rank of any one of these matrices. Now the middle matrix D has just r rows, so its rank is at most r also. ■

How do phylogenetic invariants arise from Proposition 8.2? The link is that if a matrix M has a rank of at most r , then for any $(r+1) \times (r+1)$ matrix M' obtained from M by selecting $r+1$ rows and $r+1$ columns, then we must have $\det M' = 0$. This leads immediately to a large class of phylogenetic invariants for the GMM: all the $(r+1) \times (r+1)$ minors of $\text{flat}_{A|B}(p)$ equal zero. These “edge invariants” are topology invariants for T . Indeed, in the 2-state GMM setting and any binary phylogeny T , the set of these 3×3 minors (across all the edges of T), together with the trivial invariant, generates the algebraic ideal of all the phylogenetic invariants for T (Theorem 4 of [7]). In our example above, these invariants correspond to the 16 determinants of the 3×3 submatrices of the 4×4 matrix $\text{flat}_{\{1,2\}|\{3,4\}}(p)$.

In summary, the edge invariants for a binary phylogenetic X -tree T all vanish on $p = p_{(T,\theta)}$, while the phylogenetic invariants for X -splits that are not splits of T are (generically) nonzero. This means that, generically (i.e., apart from a set of θ parameters of measure zero), p determines the splits of T , and so (from Chapter 2) T is also determined, up to equivalence, from p .

A further property of the 2-state GMM model applies under the (mild) restriction that all transition matrices have a positive determinant. Suppose we select any state $\alpha \in S$ and we let $p(A)$ be the probability that A is the set of leaves in state α . For any phylogenetic X -tree, the set function

$$\varphi : 2^X \mapsto \mathbb{R}, A \mapsto -\log p(A)$$

is submodular, so $\varphi(A) + \varphi(B) \geq \varphi(A \cup B) + \varphi(A \cap B)$. This translates into the quadratic inequality $p(A \cup B)p(A \cap B) - p(A)p(B) \geq 0$ [332, 387].

8.3.2 • Invariants for mixture models

If a model \mathcal{M} has linear topology invariants (e.g., JC69, K2ST), then it is clear that these invariants are also valid for any mixture process on T under \mathcal{M} . In this way, models such as JC69 and K2ST are immune to the type of nonidentifiability result described in the final section of Chapter 7 for models such as the symmetric 2-state model. An important dichotomy result (Theorem 4 of [326]) says, roughly speaking, that under a mild assumption concerning the model \mathcal{M} , either linear topology invariants exist (and these suffice to determine the phylogeny) or there are finite-support mixture distributions on two trees under \mathcal{M} that are identical (so the phylogeny cannot always be uniquely determined).

Classifying the space of linear invariants for such mixture models has been elegantly figured out in [77] for the group-equivariant models on $S = \{A, C, G, T\}$ that were discussed in Chapter 7.

For models that lack linear topology invariants, such as K3ST and GMM, the algebraic techniques from the Section (8.3.1) provide an elegant way to study the identifiability question. It turns out that the discrete binary phylogeny is an identifiable parameter, at least “generically” (except on a set of measure zero), for phylogenetic mixtures under GMM (or submodels) when the number of classes m is not too large. A central tool in these types of arguments is the following.

Proposition 8.3. *If p is a mixture of m classes on T under the GMM with a state space of size r , then for any split $A|B$ of T , $\text{flat}_{A|B}(p)$ has rank at most $m \cdot r$. In particular, all the $(mr + 1) \times (mr + 1)$ minors of this matrix vanish.*

Proof: If $p = \sum_{i=1}^m a_i \cdot p_{(T, \theta_i)}$, then $\text{flat}_{A|B}(p) = \sum_{i=1}^m a_i \cdot \text{flat}_{A|B} p_{(T, \theta_i)}$. Matrix rank is a subadditive function; in other words, the rank of a sum of matrices is never larger than the sum of the ranks. Thus, the result now follows since the rank of $\text{flat}_{A|B} p_{(T, \theta_i)}$ is at most r by Proposition 8.2. ■

As an example to illustrate Proposition 8.3, consider a phylogenetic mixture of three GMM processes, involving four states and a quartet tree T , and let $A|B$ to correspond to the nontrivial split of T . In that case, the matrix $\text{flat}_{A|B}(p)$ is then a 16×16 matrix. However, by Proposition 8.3, the rank of this matrix is $3r = 12$ at most, and the phylogenetic invariants corresponding its 13×13 minors are topology invariants. Notice, however, that for the GMM on two states, we do not obtain a phylogenetic invariant, even for a mixture of just two classes. Further algebraic techniques for the study of identifiability questions appear in [12, 294]. The application of phylogenetic invariant techniques for accurately inferring phylogenies under mixture models has recently been demonstrated in [137].

Using algebraic methods (such as phylogenetic invariants, and ratios of determinants) it is also possible to show that for the $GM + I$ model, the tree, continuous parameters for the model (up to symmetries), and the proportion of invariable sites are also identifiable parameters [8]. Again using algebraic techniques, the generic identifiability of the tree and associated continuous parameters under the covarion model (mentioned in Section 7.1) has been established [9].

8.4 • The infinite-state random cluster model

In the random cluster model description of the equal input model (Proposition 7.5), each edge e of T is cut with some probability θ_e to obtain a resulting partition Π of the leaf set X . Each block is then assigned a state independently according to the distribution π . However, we can omit the second step and just consider the partition Π itself as the output of this process (rather than assigning states, which has the effect of combining some blocks together when they receive the same state). We call this the *infinite-state random cluster model*, denoted RC_∞ .

This model has a natural interpretation as the limiting distribution on partitions induced by the equal input model as the number of states k in S tends to infinity and the states have at least roughly similar probabilities. More precisely, under the random cluster description of the equal input model, the probability that there are two (or more) blocks of Π assigned a shared state is at most $\binom{n}{2} \sum_{\alpha \in S} \pi_\alpha^2$, by Boole’s inequality (note that there are at most n blocks in Π). Suppose that $\pi_\alpha \in [a/k, b/k]$ for some fixed a, b ; then as $k = |S| \rightarrow \infty$, all blocks of Π receive distinct states with probability converging to 1 (and so the corresponding characters are homoplasy-free on the generating tree T). The RC_∞

model is sometimes referred to as the “Kimura infinite alleles” model in phylogenetics and it was studied mathematically in [267].

Let $p_{(T,\theta)}(\Pi)$ denote the probability of generating the partition Π under the RC_∞ model on T with cut probabilities $\theta = (\theta_e)$. A first observation is that $p_{(T,\theta)}(\Pi) = 0$ unless Π is convex on T (recall from Section 5.1 that a partition Π of $[n]$ is convex on $T \in P(X)$ if the collection of induced subtrees $\{T[B] : B \in \Pi\}$ are vertex-disjoint). In general, $p_{(T,\theta)}(\Pi)$ can be computed efficiently (in polynomial time); however, there is also an elegant explicit description of it due to Steve Evans, which we now describe.

Consider the poset of partitions of X under the partial order of refinement, and write $\Pi \leq \Pi'$ if Π refines Π' (i.e., the blocks of Π' are unions of the blocks in Π). Let $Q(\Pi) = \prod_{B \in \Pi} \prod_{e \in T[B]} (1 - \theta_e)$. In words, $Q(\Pi)$ is the probability that none of the edges in any of the minimal subtrees of T that connect each block of Π is cut. It follows that

$$\sum_{\Pi' : \Pi \leq \Pi'} p_{(T,\theta)}(\Pi') = Q(\Pi).$$

We can now use (the dual form of) Möbius inversion to write

$$p_{(T,\theta)}(\Pi) = \sum_{\Pi' : \Pi \leq \Pi'} \mu(\Pi, \Pi') Q(\Pi').$$

Here, the Möbius function $\mu(\Pi, \Pi')$ for $\Pi \leq \Pi'$ is calculated as follows. If Π' has m blocks and the i th block of Π' is equal to the union of n_i blocks of Π then $\mu(\Pi, \Pi') = \prod_{i=1}^m (-1)^{n_i-1} (n_i - 1)!$.

As an example, consider the quartet tree $12|34$, where the pendant edge incident with leaf i has cut probability θ_i and the central edge has cut probability θ_5 . For $\Pi = \{\{1, 2\}, \{3, 4\}\}$, there are two partitions Π' with $\Pi \leq \Pi'$, namely $\Pi' = \Pi$ and $\Pi' = \{\{1, 2, 3, 4\}\}$, with associated $\mu(\Pi, \Pi')$ values $+1$ and -1 respectively; so, writing $q_i = 1 - p_i$, we have

$$p_{(T,\theta)}(\Pi) = q_1 q_2 q_3 q_4 - q_5 q_1 q_2 q_3 q_4 = p_5 q_1 q_2 q_3 q_4.$$

Earlier, we mentioned a deep and difficult result that $k = \Theta(\log(n))$ characters suffice for tree reconstruction under simple finite-state models, provided the substitution probabilities on the edges are not too large, and they are bounded away from 0. The corresponding result for RC_∞ also holds—we state this shortly—but it is much easier to prove. An interesting feature of the result for RC_∞ is that the dependence of k on the smallest edge probability f grows at the rate $\frac{1}{f}$ as $f \rightarrow 0$, in contrast to the finite state setting, where the growth rate is $\frac{1}{f^2}$ from eqn. (8.7). The following result is from [267].

Theorem 8.4. *Suppose that k partitions are generated under the RC_∞ model on any tree $T \in B(X)$ with associated edge parameters θ_e that satisfy $0 < f \leq \theta_e \leq g$, where $g < \frac{1}{2}$ is any fixed value. Then*

$$k = \left\lceil \frac{2(1-g)^4}{f(1-2g)^4} \ln \left(\frac{n}{\sqrt{\epsilon}} \right) \right\rceil$$

partitions suffice for reconstructing T correctly with probability $1 - \epsilon$.

Proof. The proof exploits a result from Chapter 4. Proposition 4.13 stated that a generous cover \mathcal{Q} for T satisfies $\text{cl}_1(\mathcal{Q}) = \mathcal{Q}(T)$. In particular, \mathcal{Q} defines T . To apply this result we first require a lemma.

Lemma 8.5. Consider the RC_∞ model on $T' \in RB(X)$, with $\theta_e \leq g < \frac{1}{2}$. The probability P_g that there is at least one leaf x of T' for which none of the edges from the root of T' to x are cut is at least $\frac{(1-2g)}{(1-g)^2}$.

Proof: Lemma 8.5 can be proved directly, but readers familiar with the classic *Galton–Watson branching process* will perhaps prefer the following argument. Let us first convert T' into a fully balanced binary tree of height h that is as large as (or even larger than) the longest path to any leaf. Assign a cut probability g to all the edges. The probability P that there is an uncut leaf path in this modified tree is now a lower bound on P_g . Consider a Galton–Watson branching process in which each individual has 0, 1, or 2 children with probabilities g^2 , $2g(1-g)$, and $(1-g)^2$, respectively. The connection to P is to regard each vertex of T' as giving “birth” to 0, 1, or 2 vertices that are joined by “uncut” edges (i.e., a cut extinguishes the remainder of the lineage). Since the cuts are independent, P is simply the probability that at least one individual in the branching process is still alive in generation h of the branching process (see Fig. 8.5(a)). For $g < \frac{1}{2}$, the expected number of children of each vertex is greater than 1, so the probability of the eventual extinction of the process q is strictly positive and is given by $q = \frac{g^2}{(1-g)^2}$. Thus $P_g \geq P \geq 1 - q = \frac{1-2g}{(1-g)^2}$, establishing the lemma. ■

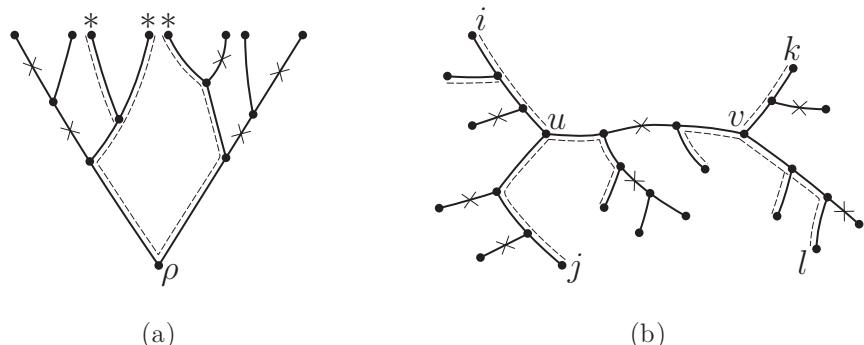


Figure 8.5. (a) Under the RC_∞ model on a perfect binary tree, in which each edge has a constant cut probability, the number of leaves that trace back to the root without crossing a cut edge can be modeled exactly by a Galton–Watson branching process. In this instance, the root vertex ρ has two “live” children, whereas each of the children of ρ just has one “live” child, resulting in three “live” leaves (indicated by *). (b) For the pattern of cuts shown for RC_∞ , two leaves (i and j) lie in one block of the resulting partition (and in different subtrees incident with u), and two leaves (k and l) also lie in one block (and in different subtrees incident with v), and these blocks are different, due to the cut on the path between u and v .

Returning to the proof of Theorem 8.4, given two interior vertices u and v of T , consider the event \mathcal{E}_{uv} illustrated in Fig. 8.5(b) where

- (i) there exists $i, j \in [n]$, where i and j are in different subtrees incident with u (neither containing v) and with no cut edge on the paths from u to i and to j ;
- (ii) there exists $k, l \in [n]$, where k and l are in different subtrees incident with v (neither contains u) and with no cut edge on the paths to v to k and to l ; and
- (iii) there is at least one cut on the path from u to v .

Lemma 8.5 ensures that for any two distinct interior vertices u, v of T ,

$$\mathbb{P}(\mathcal{E}_{uv}) \geq f \cdot [(1-g)P_g]^4. \quad (8.17)$$

Now consider a sequence \mathcal{C}_k of k partitions generated by the model. Let $q(\mathcal{C}_k)$ be the set of quartet trees $xy|wz$ for which there is at least one partition in $q(\mathcal{C}_k)$ with x, y in one block and w, z in a different block. The probability P_k that $q(\mathcal{C}_k)$ is a generous cover for T is 1 minus the probability that \mathcal{E}_{uv} fails for one of the $\binom{n-2}{2}$ pairs of interior vertices u, v of T . Thus

$$P_k \geq 1 - \binom{n-2}{2}(1-f(1-g)^4 P_g^4)^k.$$

Routine algebra, using the inequalities $\binom{n-2}{2} \leq n^2$ and $-\ln(1-x) \geq x$ for $x \in (0, 1)$, now shows that $P_k \geq 1 - \epsilon$ by taking k to be as large as described.

Since all the partitions in \mathcal{C}_k are convex on T (with probability 1), Proposition 4.13 ensures that $\text{cl}_1(q(\mathcal{C}_k)) = \mathcal{Q}(T)$ and so, with a probability of at least $1 - \epsilon$, T is the only tree on which each partition in \mathcal{C}_k is convex. This completes the proof. ■

Notice the reconstruction of T from \mathcal{C}_k in the proof of Theorem 8.4 is possible in polynomial time in n . When the θ_e values are larger than $1/2$, the number of partitions required grows polynomially rather than logarithmically with n [267].

Phylogenetic invariants for the RC_∞ model. The linear invariants for RC_∞ are particularly easy to describe. When T is a binary tree, precisely F_{2n-1} partitions of $[n]$ are convex on T , where F_k is the k th Fibonacci number, starting with $F_1 = F_2 = 1$ (this follows from eqn. (5.6)). By contrast, for a star tree on $[n]$, there are precisely $2^n - n$ partitions of $[n]$ that are convex on this tree.

Let $\text{co}(T)$ and $\text{Inc}(T)$ be the sets of partitions of $[n]$ that are convex on T , and not convex on T (i.e., they are “incompatible” with T), respectively. Under the RC_∞ model it can be shown that:

$$\Pi \in \text{Inc}(T) \implies p_{(T, \theta)}(\Pi) = 0 \text{ for all } \theta \quad (8.18)$$

and

$$\Pi \in \text{co}(T) \implies p_{(T, \theta)}(\Pi) > 0 \text{ whenever } \theta_e > 0 \text{ for all } e. \quad (8.19)$$

Exercise: Prove eqns. (8.18) and (8.19).

It follows from eqn. (8.18) that the vector space \mathcal{L}_T of homogeneous linear phylogenetic invariants of T are generated by the system of equations $\{x_\Pi = 0 : \Pi \in \text{Inc}(T)\}$, which has dimension

$$|\text{Inc}(T)| = B_n - |\text{co}(T)|,$$

where the “Bell number” B_n is the total number of partitions of the set $[n]$. The dimension of the space of all phylogenetic mixtures on T has dimension $|\text{co}(T)| - 1$ and the space of all phylogenetic mixtures on all trees under the RC_∞ model has dimension $B_n - 1$.

The construction of (quadratic) phylogenetic invariants for RC_∞ is also quite easy, and we leave the following result as an exercise to the reader.

Exercise: Under the RC_∞ model on T , let $p(x,y)$ denote the probability that x and y are in the same block of the random partition produced. Show that if $T|\{x,y,w,z\}$ is either $xy|wz$ or the star tree, then

$$p(x,w)p(y,z) - p(x,z)p(y,w) = 0.$$

It is also possible to define a “phylogenetic orange” space for RC_∞ ; topologically, these are identical to the spaces \mathbb{O}_T and \mathbb{O}_n described earlier.

Does “testing a tree” require less data than finding it? In Section 8.2.1, we presented a simple and general argument to show that the number of r -state characters k required for accurate tree reconstruction needs to grow at a minimal rate of $\log(n)$ in the number n of leaves of the tree. Moreover, we also described some deeper results that showed how this low growth rate is also achievable. Now suppose that we are given a candidate binary tree $T \in B(n)$ along with the data (sequence of characters), and are merely asked to decide whether or not this is the tree from $B(n)$ that generated the given data. We might call this “testing a tree,” rather than “reconstructing a tree.” Accuracy now means that whatever answer we give (“yes” or “no”) is correct with high probability.

This raises an interesting question: Does the number of characters required to test a tree still need to grow at a minimal rate of $\log(n)$ or can we get away with a lesser rate (note that the original argument for the $\log(n)$ lower bound no longer applies)? Using information-theoretic arguments, it can be shown that for finite-state Markov processes on trees, the $\log(n)$ lower bound still holds for testing a tree [343]. However, for the infinite-state random cluster model the situation is quite different. For any given accuracy, the number of characters required to “test” $T \in B(n)$ is bounded above by a constant C , independent of n and dependent only on the accuracy required, and the bounds f and g on the edge parameters in Theorem 8.4.

The test is very simple. Given a sequence \mathcal{C} of k partitions and a candidate tree T , reject T if any of the partitions is not convex on T ; otherwise, accept T . It is clear that if T is rejected, then it cannot have been the tree that produced the partitions under RC_∞ (e.g., by eqn. (8.18)). The nontrivial direction is that if T is not the tree that generated \mathcal{C} , then it will be rejected with a probability of at least $1 - \epsilon$ for $k = O(1)$ partitions (where the constant in O depends just on ϵ and not n). The proof requires a similar argument to that used in Theorem 8.4 (for details, see [343]).

8.4.1 • An application using the probabilistic method

We end our study of random processes on trees by showing how they relate to an earlier topic from Chapter 5. Proposition 5.4 (from [53]) stated that for any fixed number of states r , the minimum number k of r -state characters required to capture a binary phylogeny $T \in B(n)$ is $\lceil \frac{n-3}{r-1} \rceil$ provided that $n \geq n_r$ for a certain increasing sequence n_r . Let us consider how small k can be when r is not fixed but is allowed to depend on n . From Theorem 5.5, we know that there is a set \mathcal{C} of $k = 4$ characters for which the associated number n_r of states satisfies $n/r_n = O(1)$. Thus we focus on the setting where both r_n and n/r_n grow with increasing n .

More precisely, suppose that we want a set \mathcal{C}_n of $k_n = \lfloor n^\alpha \rfloor$ characters on $[n]$, each taking $r_n = \lfloor n^\beta \rfloor$ states at most, to capture some phylogenetic X -tree, where $\alpha, \beta > 0$. Notice that the inequality $k \geq \lceil (n-3)/(r-1) \rceil$ implies that k_n must exceed $n^{1-\beta}$ for n sufficiently large, thus $\alpha + \beta > 1$. We show here that any value of $\alpha, \beta > 0$ with

$\alpha + \beta > 1$ allows for such a set \mathcal{C}_n for any binary tree T . This result is independent of Proposition 5.4, in the sense that neither result directly implies the other. The proof involves a simple application of the probabilistic method, using Theorem 8.4.

Theorem 8.6. *For any two values $\alpha, \beta \in (0, 1)$ for which $\alpha + \beta > 1$, there is a value N so that for all $n \geq N$, the following holds: For any $T \in B(n)$, there is a set \mathcal{C}_n of $k_n = \lfloor n^\alpha \rfloor$ characters on $[n]$ that capture T , where each character takes (at most) $r_n = \lfloor n^\beta \rfloor$ distinct states.*

Proof: Consider the RC_∞ model on T , with each edge having cut probability $\theta_n = r_n/4n$. We will associate each such randomly generated partition Π with a character f_Π that induces the same partition of $[n]$ as Π . Let Y denote the random number of edges of T that are cut. Y has a binomial distribution $Y \sim \text{Bin}(2n-3, \theta_n)$, which has mean $\mu_n = (2n-3)\theta_n = (\frac{1}{2} - o(1))n^\beta$. By a multiplicative form of the Chernoff bound in probability theory (cf. [174], eqn. (6) with $f = 1$), we have $\mathbb{P}(Y \geq 2\mu_n) \leq \exp(-\mu_n/3)$. Since $r_n > 2\mu_n$, we obtain

$$\mathbb{P}(Y \geq r_n) \leq \exp(-\mu_n/3). \quad (8.20)$$

The number of blocks of the partition of $[n]$ under RC_∞ is at most $Y + 1$. Thus the probability that a character, generated by the model described, takes strictly more than r_n distinct states is at most $\mathbb{P}(Y + 1 > r_n) = \mathbb{P}(Y \geq r_n) \leq \exp(-\mu_n/3)$, by (8.20).

Let us generate a set \mathcal{C}_n of k_n such characters independently by the process described (i.e., constructing partitions of $[n]$ and giving an associated character for each partition). The probability that at least one of these characters has more than r_n distinct states is, by Boole's inequality, at most

$$n^\alpha \exp(-\mu_n/3) = n^\alpha \exp\left(-\frac{1}{3}\left(\frac{1}{2} - o(1)\right)n^\beta\right) \rightarrow 0$$

as $n \rightarrow \infty$ (recall $\beta > 0$). Thus, there exists some value N_1 for which, for any $n \geq N_1$, at least one character in \mathcal{C}_n takes more than r_n distinct states with probability at most $1/3$.

What is the probability that \mathcal{C}_n captures T ? By Theorem 8.4, \mathcal{C}_n captures T with probability at least $1 - \epsilon$ provided that $k = \lceil \gamma_n \ln(n^2/\epsilon) \rceil$, where $\gamma_n = \frac{1}{\theta_n} \left(\frac{1-\theta_n}{1-2\theta_n}\right)^4$. Now, as $n \rightarrow \infty$,

$$\gamma_n \sim \frac{1}{\theta_n} = \frac{4n}{r_n} \sim \frac{4n}{n^\beta} = 4n^{1-\beta}.$$

Since $\alpha + \beta > 1$, it follows that for any $\epsilon > 0$

$$\gamma_n \ln(n^2/\epsilon)/n^\alpha \sim 4n^{1-\beta} \ln(n^2/\epsilon)/n^\alpha \rightarrow 0 \text{ as } n \rightarrow \infty.$$

So, if we take $\epsilon = 1/3$, there is a value N_2 for which, for any $n \geq N_2$, $\lceil \gamma_n \ln(n^2/\epsilon) \rceil \leq \lfloor n^\alpha \rfloor$ holds for all $n \geq N_2$. Thus, with $k_n = \lfloor n^\alpha \rfloor$ where $n \geq N_2$, \mathcal{C}_n fails to capture T with a probability of at most $1/3$.

Combining these two observations, if we set $N = \max\{N_1, N_2\}$ then for all $n \geq N$, the probability that a set \mathcal{C}_n of $\lfloor n^\alpha \rfloor$ randomly generated characters satisfies at least one of the following properties:

- (i) \mathcal{C}_n contains a character that takes more than r_n distinct states, or
- (ii) \mathcal{C}_n fails to capture T ,

is at most $\frac{1}{3} + \frac{1}{3} = \frac{2}{3}$, by Boole's inequality. Thus there is a strictly positive probability that \mathcal{C}_n satisfies neither of conditions (i) and (ii), so a set of $\lfloor n^\alpha \rfloor$ characters which captures T must exist, where each character takes at most $\lfloor n^\beta \rfloor$ states, which captures T . This completes the proof. ■

8.5 • Additional topics

In our two chapters on Markov processes on trees, many topics have been omitted, including resampling methods (e.g., bootstrap) and model selection (AIC, BIC), the detailed theory underlying the MCMC machinery in Bayesian phylogenetics, and the evolution of continuous characters based on stochastic models such as Brownian motion and the Ornstein–Uhlenbeck process. Fortunately, most of these topics have been discussed in other books or survey papers (e.g., for continuous characters, an overview is provided in [136]; for more recent results, see [190]).

We have also regarded molecular data as consisting of discrete characters that correspond to the aligned sites in genetic (or protein) sequences. As such we have largely ignored the preprocessing alignment step which is typically achieved by introducing “gaps” at various places in certain sequences. An alternative approach is to directly use a stochastic model that allows for substitutions at sites as well as insertion and deletion events. These “statistical alignment” models were introduced in [357], and the analysis of these types of models was further pioneered by Jotun Hein, István Miklós, and colleagues to a high degree of sophistication. These models provide a way to carry out sequence alignment and phylogeny reconstruction simultaneously, and doing so directly from sequence data within a likelihood or Bayesian setting [345]. As such, these methods are more directly model-based than the distance-based approaches to raw sequence data (based on k -mers and related notions) described in Section 6.1.4. Two recent papers on this topic are [58] and [186].

Rather than dealing further with these topics, we turn next to the evolution of phylogenetic trees.

Chapter 9

Evolution of trees

Until now, we have mostly treated a phylogenetic tree T as a fixed entity to be analyzed, or as a parameter to be reconstructed from data. However, the tree itself can be viewed as a random variable in two ways. First, at the species level, it is the product of a process of speciation and extinction over macroevolutionary timescales. Second, at the level of individual genes, the ancestry of a gene sampled from the tips of a given (fixed) species tree is again a random variable due to processes such as lineage sorting and lateral gene transfer (discussed below).

We will start with the species-level phylogeny, where Chapter 3 provided some insights into the discrete shape of the resulting tree. We remarked there that many processes lead to the YH model for the “reconstructed tree.” We will now study speciation and extinction processes in more detail (as “birth-death” models), and consider not just the shape of the resulting tree, but also the distribution on its edge lengths. This provides a second (continuous) dimension to tree shape and the study of what these two aspects of tree shape say about the underlying evolutionary processes is sometimes referred to as *phylogenetics*.

As well as the tree-balance statistics considered in Chapter 3, biologists are interested in how the number of lineages have changed through time (often represented by a “lineage-through-time” plot, usually on a log-log scale), or in the distribution of edge lengths (e.g., whether edges near the root or near the tips tend to be longer or shorter than other edges on average). In particular, phylogenetic trees reconstructed from genetic data (by techniques described in earlier chapters) can be used to study the processes of speciation and extinction that led to a given group of present-day species [177, 263, 322, 323].

The study of continuous-time processes for modeling phylogeny traces back to a pioneering paper by George Udny Yule in 1925 [382]. Yule was asked by a colleague if he could explain the distribution of species across genera in various data sets (e.g., many genera of snakes had just one species, fewer had two, and so on, leading to a somewhat “long-tailed” distribution). Yule considered a simple model—now called the “Yule pure-birth” process—in which each species can give rise to a new species according to a Poisson process at rate λ . This was coupled with a second pure-birth process in which each species can also give rise to a new genus at rate g . In this way, Yule was able to describe the distribution of species across genera in a way that depends, asymptotically, just on the ratio of g to λ .

9.1 ■ Yule pure-birth trees: The simplest model

The simplest model for generating a rooted phylogeny in continuous time is the Yule pure-birth model with a constant speciation rate λ . In this model, we start at time $t = 0$ either with a single lineage of length zero, or a single bifurcation (two lineages, both of length zero) and the process proceeds forward in time. These two starting scenarios—the *single-start* and the *double-start*—are closely related, though sometimes one is slightly easier to analyze than the other; we will generally assume the former unless stated otherwise. These two scenarios are illustrated in Fig. 9.1.

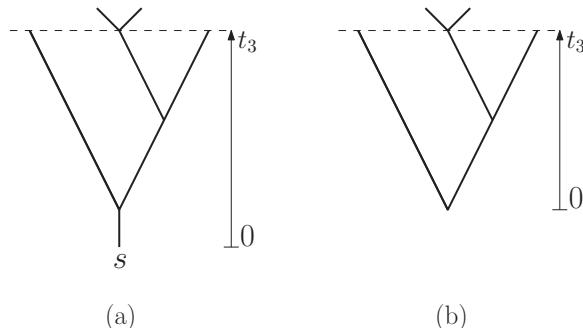


Figure 9.1. A Yule pure-birth tree (single-start (a) and double-start (b)) grown until the time t_3 when it last had three leaves.

At each time $t > 0$, the tip of each of the existing lineages can bifurcate (we will say “speciate”) according to an exponential distribution with a constant rate λ , independently of other lineages. Thus, the probability that a lineage bifurcates into two lineages between time t and $t + \delta$ is $\lambda\delta + O(\delta^2)$, where the $O(\delta^2)$ term accounts for the possibility of two or more bifurcations occurring in the time period δ . Let $N(t)$ be the (random variable) number of lineages at time t . Since $\mathbb{E}[N(t + \delta)|N(t)] = (1 + \lambda\delta)N(t) + O(\delta^2)$, if we let $v(t) = \mathbb{E}[N(t)]$, then the law of total expectation gives

$$v(t + \delta) = \mathbb{E}[\mathbb{E}[N(t + \delta)|N(t)]] = (1 + \lambda\delta)v(t) + O(\delta^2).$$

Consequently,

$$\frac{dv(t)}{dt} = \lambda v(t),$$

which, when coupled with the initial condition $v(0) = 1$ (single-start scenario), gives $v(t) = e^{\lambda t}$. What can we say about the actual distribution of $N(t)$? We will see in Section 9.2.3 that for the single-start scenario $N(t)$ has a geometric distribution with parameter $p = e^{-\lambda t}$; in other words,

$$\mathbb{P}(N(t) = n) = (1 - e^{-\lambda t})^{n-1} e^{-\lambda t}.$$

This distribution has mean $1/p = e^{\lambda t}$, in agreement with previous calculation, and $\mathbb{P}(N(t) = 1) = e^{-\lambda t}$. Two standard features of this geometric distribution are worth noting. First, $N(t)$ remains well dispersed about its mean, in the sense that the ratio $N_t/\mathbb{E}[N_t]$ does not converge in probability to 1 as t grows. Second, the maximum likelihood estimate of t for the event $N(t) = n$ is $\frac{\ln(n)}{\lambda}$, as can be seen by solving the equation $\frac{d}{dt}\mathbb{P}(N(t) = n) = 0$ for t . If t is known but λ is not, then the same argument shows

that the maximum likelihood estimate of the speciation rate λ is $\frac{\ln(n)}{t}$. In the double-start scenario, n is replaced by $n/2$.⁶⁰

A Yule pure-birth tree can be generated in several ways. The most obvious is to grow the tree for some fixed period of time t . In this case, the number of leaves, $N(t)$, is a random variable. A second option is to grow a Yule pure-birth tree for time t and also conditional on the event that $N(t) = n$. A third option is to sample the tree when it has n leaves. This last case is slightly ambiguous; it could mean “when it first has n leaves” (i.e., the minimal value of t for which $N(t) = n$) or the last moment when $N(t) = n$, or at some random time in between (we will make this clear later). To summarize, one can sample according to these three scenarios:

- conditional just on t ;
- conditional just on n ;
- conditional on n and t .

Let us consider the distribution of the lengths of the edges of a Yule pure-birth tree, starting with the simplest case of conditioning just on t . First, let $L(t)$ denote the sum of all the edge lengths of the resulting tree \mathcal{T}_t .

Between t and $t + \delta$, the sum of the edge lengths grows by δ for each of the $N(t)$ extant edges. In addition, the probability of any speciation event in this interval is of order δ and any such event contributes order δ additional length. This allows us to write $L(t + \delta) = L(t) + \delta N(t) + O(\delta^2)$. Therefore, if we let $\Lambda(t) = \mathbb{E}[L(t)]$, then the usual infinitesimal argument gives

$$\frac{d\Lambda(t)}{dt} = \mathbb{E}[N(t)] = e^{\lambda t}.$$

With the initial condition $\Lambda(0) = 0$, integration gives

$$\Lambda(t) = \frac{1}{\lambda}(e^{\lambda t} - 1). \quad (9.1)$$

Let us now consider separately the sum $P(t)$ of the expected lengths of the pendant edges that are incident with the leaves of the tree at time slice t , and $I(t)$ the sum of the expected lengths of the remaining interior edges of the tree (note that the initial lineage will count as a P -edge up until the first speciation event, after which it counts as an I -edge). Thus if the tree has $N(t) = n$, then the number of P -type and I -type edges is n and $n - 1$, respectively. The connection between $I(t)$ and $P(t)$ is particularly succinct:

$$\frac{dI(t)}{dt} = \lambda P(t). \quad (9.2)$$

Since $P(t) = \Lambda(t) - I(t)$, eqn. (9.1) furnishes the following first-order linear differential equation for $I(t)$:

$$\frac{dI(t)}{dt} + \lambda I(t) = e^{\lambda t} - 1.$$

⁶⁰If $\tilde{N}(t)$ is the number of lineages in the double-start scenario at time t , then $\tilde{N}(t)$ is a sum of two independent copies of $N(t)$ and so has distribution $\mathbb{P}(\tilde{N}(t) = n) = (n-1)(1-e^{-\lambda t})^{(n-2)}e^{-2\lambda t}$. Solving $\frac{d}{d\lambda}\mathbb{P}(\tilde{N}(t) = n) = 0$ for λ gives $\lambda = \frac{\ln(n/2)}{t}$.

Its solution, with $I(0) = 0$, is $I(t) = \frac{1}{2\lambda}(e^{\lambda t} + e^{-\lambda t} - 2)$ and so $P(t) = \frac{1}{2\lambda}(e^{\lambda t} - e^{-\lambda t})$. Notice that if we divide $I(t)$ and $P(t)$ by the expected number of interior and pendant edges of \mathcal{T}_t respectively, then these ratios both converge to $\frac{1}{2\lambda}$ at an exponential rate as $t \rightarrow \infty$.

Exercise⁺: Formally establish eqn. (9.2).

9.1.1 • Conditioning on n : A curiously exact result

In this section, we calculate the expected length of randomly selected edge in a Yule pure-birth tree in which the speciation rate is λ and the tree is sampled when it has n leaves. It might be supposed that the average length of the pendant edges would be a bit shorter than the interior ones, since they have been sampled at some “time-slice” and would have continued to grow if we had not sampled them. For the interior edges, we might expect the average length to be $\frac{1}{\lambda}$, since an exponential distribution with rate λ has mean $\frac{1}{\lambda}$. We will see that both these initial guesses are wrong. Of course, the exponential distribution leads to other surprises, such as the well-known “waiting paradox.”⁶¹

Let $\mathcal{T}_{(n)}$ be the tree obtained by growing a (double-start) Yule pure-birth tree \mathcal{T}_t until the smallest value of $t = t_n$ for which $N(t) = n + 1$. Since, at this value of t , the two new pendant edges both have length 0, we can regard this tree as really having just n leaves, and with the edge incident with this leaf being a pendant edge. In other words, $\mathcal{T}_{(n)}$ is the tree \mathcal{T}_t at the last moment that it still has n leaves with positive edge lengths. This is illustrated in Fig. 9.1(b).

Let L_n be the sum of the lengths of the edges of $\mathcal{T}_{(n)}$. Notice that we can write this as

$$L_n = 2X_2 + 3X_3 + \cdots + nX_n,$$

where X_i is the length of time in T_t for which there are i lineages. Since X_i is the minimum of i independent exponential random variables (each having mean $\frac{1}{\lambda}$) it also has an exponential distribution, and with mean $\frac{1}{i\lambda}$. Consequently, $\mathbb{E}[iX_i] = \frac{1}{i\lambda}$ for all i , and so

$$\mathbb{E}[L_n] = \frac{n-1}{\lambda}. \quad (9.3)$$

Since $2X_2, 3X_3, \dots, nX_n$ are independent random variables having the same exponential distribution, L_n has a gamma distribution (with mean $\frac{n-1}{\lambda}$ and variance $\frac{n-1}{\lambda^2}$). In particular, L_n is asymptotically normally distributed as n grows, by the central limit theorem.

Let us now select one of the $2n - 2$ edges of T uniformly at random, and let ℓ_n be its length, and let \bar{l}_n be the average length of all the edges. Thus, $\bar{l}_n = \mathbb{E}[\ell_n | T]$ where expectation is taken with respect to a uniform distribution on the edges, and so $\mathbb{E}[\ell_n] = \mathbb{E}[E[\ell_n | T]] = \mathbb{E}[\bar{l}_n]$. Equation (9.3) gives

$$\mathbb{E}[\bar{l}_n] = \frac{1}{2\lambda}. \quad (9.4)$$

This expression is interesting for two reasons. First, the right-hand side is completely independent of n . Second, it might have been suspected that if the expected time for

⁶¹If buses arrive at a bus stop regularly every 20 minutes and we go to the bus stop at a random time, our expected waiting time is 10 minutes. But if buses arrive randomly according to an exponential distribution with a mean time of 20 minutes between arrivals, and if we again go to the stop at a random time, then our expected waiting time is now 20 minutes.

a lineage to speciate is $1/\lambda$, then the expected length of an edge selected uniformly at random in a Yule pure-birth tree conditional only on n would also be $1/\lambda$. However, eqn. (9.4) shows that it is exactly half this value. An explanation is that the $1/\lambda$ expectation holds when speciation events are occurring along a line, but here, they are occurring on a bifurcating tree. For the single-start scenario, we obtain a similar result to eqn. (9.4), with the minor adjustment that n now plays a (asymptotically negligible) role.

This prompts a second question. Does a pendant edge picked uniformly at random, have the same expected length as an interior edge picked uniformly at random? Again, the double-start scenario leads to a pleasing result: the two values are exactly identical [338].

Proposition 9.1. *Let ι_n and p_n denote the lengths of a randomly selected interior and pendant edge in $\mathcal{T}_{(n)}$, respectively. For all $n \geq 2$,*

$$\mathbb{E}[\iota_n] = \mathbb{E}[p_n] = \mathbb{E}[\ell_n] = \frac{1}{2\lambda}.$$

Proof. Let I_n (respectively, P_n) be the sum of the lengths of the interior edges (respectively, pendant edges) of $\mathcal{T}_{(n)}$. Then

$$\mathbb{E}[I_n] = \mathbb{E}[I_{n-1}] + \frac{1}{n-1} \mathbb{E}[P_{n-1}],$$

since at the instant when the number of species increases from $n-1$ to n , one of the $n-1$ pendant edges (selected uniformly at random) is converted to an interior edge, and so contributes to the existing sum of the $n-1$ interior edge lengths. Since $\mathbb{E}[I_n] + \mathbb{E}[P_n] = \mathbb{E}[I_n + P_n] = \mathbb{E}[L_n] = (n-1)/\lambda$ (eqn. (9.4)), we obtain the identity

$$\mathbb{E}[P_n] = \mathbb{E}[P_{n-1}] \left(1 - \frac{1}{n-1}\right) + \frac{1}{\lambda},$$

from which $\mathbb{E}[P_2] = \frac{1}{\lambda}$ gives $\mathbb{E}[P_n] = \frac{n}{2\lambda}$. Since there are n pendant edges in the T_n , a uniformly chosen pendant edge has expected length $\frac{1}{2\lambda}$. This is the same as for a uniformly chosen edge (by eqn. (9.4)), and so the interior edges also have exactly the same expected length on average, namely $\frac{1}{2\lambda}$. ■

Not only do ι_n and p_n have the same mean under the double-start scenario, but they also have the same exponential distribution ([324], Corollary 3.2 and Theorem 3.3). Notice, however, that the lengths of edges in $\mathcal{T}_{(n)}$ are not independent random variables, since they are ultrametric, so the distance from the root to each leaf is the same. Nor are the lengths of edges in $\mathcal{T}_{(n)}$ identically distributed; for example, if we select uniformly at random one of the two edges incident with the root of a Yule tree under the double-start scenario, then the expected length of this edge, conditional on the tree having n leaves, is

$$\frac{1}{\lambda} \left(1 - \frac{1}{n}\right).$$

A similar $\frac{1}{\lambda}$ asymptotic applies if we condition on t rather than n . In other words, for large values of n (or t), the average length of the two edges incident with the root has about twice the expected value of a randomly selected (pendant or interior) edge. For details, see [324].

9.1.2 • Ancestral state reconstruction

Suppose that we grow a (single-start) Yule pure-birth tree for time t and then evolve a character on this tree from the root to the leaves according to a Markov process on a tree, as in Chapter 7. For convenience, consider the 2-state symmetric model on the state set $S = \{\alpha, \beta\}$, with a constant substitution rate ν across the tree. The expected number of substitutions on an edge e having length ℓ_e is $\nu\ell_e$. Thus the probability of a net substitution between the endpoints of e is $\frac{1}{2}(1 - \exp(-2\nu\ell_e))$.

Suppose that the initial species s at time 0 has either state α or β with equal probability, and that this state evolves on the tree towards the leaves by the model described. In Chapter 8, we studied the relative performance of different ancestral state estimation methods on fixed trees; in particular, the methods MR (majority rule), MP (maximum parsimony), and ML (maximum likelihood). Here, we consider what happens for these methods when the tree is a random tree generated by a Yule pure-birth process.

Notice that there are two random processes at play here: first, the model that generates the tree (Yule pure-birth); second, the substitutional process that occurs along the edges of this tree (the 2-state symmetric model). Our first result shows that the ratio of speciation to substitution plays a key role in determining whether ancestral state reconstruction is any better than guessing in the limit of large t .

Let \mathcal{T}_t be a (single-start) Yule pure-birth model with speciation rate λ grown for time t from a founding species s . For the two-state symmetric model operating on \mathcal{T}_t at a constant rate ν , and any method ϕ that estimates the state at s from the character f of at the leaves of \mathcal{T}_t (and perhaps also \mathcal{T}_t), let $P_t(\phi)$ be the probability that ϕ infers the correct state that was assigned to s .

Proposition 9.2. *Let $R = \lambda/\nu$ be the ratio of speciation to substitution.*

- (i) *If $R < 4$, then $\lim_{t \rightarrow \infty} P_t(\phi) = \frac{1}{2}$ for any ancestral state estimation method ϕ .*
- (ii) *For $\phi = \text{MR}$, if $R > 4$, then for all $t > 0$, $P_t(\text{MR}) > 1 - \frac{2}{R}$. In particular, $\lim_{t \rightarrow \infty} P_t(\text{MR}) \geq 1 - \frac{2}{R} > \frac{1}{2}$.*

In other words, if the ratio of speciation to substitution is less than 4, then no ancestral state reconstruction method (including ML, with edge lengths given) will have an asymptotic accuracy better than guessing. However, for a ratio that is greater than 4, the simple MR method, which does not even use the tree or its edge lengths, has an asymptotic accuracy bounded above that of a random guess. Part (i) of Theorem 9.2 is from [150] and follows from information-theoretic arguments (similar to those employed in Chapter 8). Part (ii), from [269], relies on a second moment calculation, coupling, and a novel application of a reflection principle.

It is interesting to compare the accuracy of MR with MP, and, in particular, the speciation-to-substitution rate R at which MP becomes asymptotically no better than guessing. We will see shortly that MP requires a higher value of R to be asymptotically accurate, and simulations confirm its inferior performance to MR on Yule pure-birth trees [151]. This is perhaps surprising, given that MP (but not MR) uses the tree as well as the leaf states to estimate the ancestral state.

The analysis of $P_t(\phi)$ for $\phi = \text{MP}$ under the 2-state symmetric model turns out to be slightly easier than that for MR. Let S_t (respectively D_t) denote the probability that there is a unique MP state for the ancestor and that this state coincides with (respectively, is different from) the ancestral state, and let E_t denote the probability that both states

are equally parsimonious estimates of the ancestral state. When both states are equally parsimonious, we select one of them with uniform probability, so

$$P_t(\text{MP}) = S_t + \frac{1}{2}E_t. \quad (9.5)$$

If we now let $D_t = 1 - S_t - E_t$ then the three quantities S_t, E_t and D_t satisfy the following system of simultaneous quadratic first-order differential equations:

$$\begin{aligned} \frac{dS_t}{dt} &= -(\lambda + \nu)S_t + \nu D_t + \lambda(S_t^2 + 2S_t E_t); \\ \frac{dD_t}{dt} &= -(\lambda + \nu)D_t + \nu S_t + \lambda(D_t^2 + 2D_t E_t); \\ \frac{dE_t}{dt} &= -\lambda E_t + \lambda(E_t^2 + 2S_t D_t). \end{aligned}$$

The justification of this system replies on the parsimony operation description of $\text{ps}(f, T)$ from Section 5.2, alongside an infinitesimal argument. The system can be readily transformed from a three- to two-dimensional one, for which standard techniques from dynamical systems theory can then be applied. Together with eqn. (9.5), this leads to the following result from [150].

Proposition 9.3. *In the setting of Proposition 9.2, take $\phi = \text{MP}$ and again let $R = \lambda/\nu$ be the ratio of speciation to substitution.*

- (i) *If $R \leq 6$, then $\lim_{t \rightarrow \infty} P_t(\phi) = \frac{1}{2}$.*
- (ii) *If $R > 6$, then $P_t(\phi)$ is at least equal to the value*

$$\lim_{t \rightarrow \infty} P_t(\phi) = \frac{1}{2} \left(1 + \sqrt{\left(1 - \frac{6}{R}\right) \left(1 - \frac{2}{R}\right)} \right) > 1 - \frac{3}{R},$$

which is strictly greater than $\frac{1}{2}$.

9.2 ■ Birth-death models

We now leave the simple world of pure-birth speciation and consider the impact of also allowing extinction. Thus we have a speciation rate $\lambda > 0$ and an extinction rate μ . It is often assumed that $\lambda > \mu$; otherwise eventual extinction is certain, though one can also consider models with $\lambda \leq \mu$ by conditioning on the reconstructed tree having at least one species at time τ .

Starting from a single lineage at time $t = 0$, let $N(t)$ be the number of species at time t for a birth-death process with the parameters (λ, μ) . Let $p_i(t)$ be the probability that $N(t) = i$. This probability distribution was worked out many decades ago [215] and can be described as follows. Let $F(s, t)$ be the probability generating function

$$F(s, t) = \sum_{i \geq 0} p_i(t) s^i. \quad (9.6)$$

By considering the leading terms in $N(t + \delta) - N(t)$ as $\delta \rightarrow 0$, it can be shown that $F(s, t)$ obeys the following Riccati differential equation:

$$\frac{dF(s, t)}{dt} = \mu - (\mu + \lambda)F(s, t) + \lambda F^2(s, t), \quad (9.7)$$

which satisfies the initial condition $F(s, 0) = s$.

If we now let $M(t) = \mathbb{E}[N(t)]$, then $M(t) = \sum_{i \geq 0} i p_i(t) = \frac{\partial}{\partial s} F(s, t)|_{s=1}$. Therefore, if we differentiate eqn. (9.7) with respect to s and set $s = 1$ (noting that $F(1, t) = 1$), we get

$$\frac{dM(t)}{dt} = (\lambda - \mu)M(t).$$

Thus,

$$\mathbb{E}[N(t)] = e^{(\lambda - \mu)t} = e^{rt}, \quad (9.8)$$

where $r := \lambda - \mu$ is the net *diversification rate* (the speciation rate minus the extinction rate).

The actual distribution $p_i(t) = \mathbb{P}(N(t) = i)$ can also be determined. The case $i = 0$ for $p_i(t)$ is the solution $F(0, t)$ to the differential equation in (9.7) when $s = 0$. Once this is solved, the functions $p_i(t)$, for $i > 0$ form a system of linear differential equations, for which the stated solutions can be verified by induction on $i \geq 1$. This leads to the following expressions:

$$p_i(t) = \begin{cases} \frac{\mu}{\lambda} q(t), & i = 0; \\ e^{-rt} (1 - p_0(t))^2 q(t)^{i-1}, & i > 0, \end{cases} \quad (9.9)$$

where

$$q(t) = \begin{cases} \lambda(1 - e^{-rt}) / (\lambda - \mu e^{-rt}), & \text{if } r \neq 0; \\ \lambda t / (1 + \lambda t), & \text{if } r = 0. \end{cases} \quad (9.10)$$

Equation (9.9) reveals that, conditional on the event $N(t) > 0$, $N(t)$ has a geometric distribution with parameter $(1 - q(t))$, so $\mathbb{E}[N(t)|N(t) > 0] = \frac{1}{1 - q(t)}$. In particular, for a Yule pure-birth model, the unconditional distribution of $N(t)$ is also geometric, since $N(t) > 0$ with probability 1. Notice also that $\lim_{t \rightarrow \infty} p_0(t) = \mu/\lambda$, which is the probability that the tree eventually stops growing (i.e., all lineages die off).

For the critical case ($\lambda = \mu$),

$$p_i(t) = \begin{cases} \lambda t / (1 + \lambda t), & i = 0, \\ \left(\frac{\lambda t}{1 + \lambda t}\right)^{i-1} / (1 + \lambda t)^2, & i > 0. \end{cases} \quad (9.11)$$

In this case, the tree is certain to eventually stop growing; however, the expected time for this to occur is infinite.⁶² For all birth-death processes, it can be shown that $N(t)$ converges almost surely to 0 or to infinity (i.e., the process either dies out, or, if not, the number of leaves tends to infinity); this “extinction or explosion” dichotomy holds even for more general processes (not necessarily Markovian) by a general theorem of Peter Jagers (Theorem 2 of [204]).

⁶²If $D = \inf\{t \geq 0 : N(t) = 0\}$ then $\mathbb{E}[D] = \int_0^\infty \mathbb{P}(D > t) dt = \int_0^\infty (1 - p_0(t)) dt = \int_0^\infty \frac{1}{1 + \lambda t} dt = \infty$.

Exercise⁺: For any constant-rate birth-death process and integer $n > 0$, consider the set $\{t : N(t) = n\}$, which is either empty or comprises a collection of half-open intervals. Show that the number and total length of these intervals is finite with probability 1.

9.2.1 ■ The complete tree and reconstructed tree

Suppose that a phylogenetic tree evolves under some model, starting from a single species s . Consider the resulting random tree \mathcal{T}_t where t varies continuously from 0 to some chosen (or random) time τ when the tree is sampled (e.g., the present). Some of the lineages may have become extinct before τ , but if we include these lineages (and the initial lineage from s to the first speciation event), we refer to this tree, with its edge lengths, as the *complete tree*. In the species-level setup, we will assume that edge lengths correspond to time and thus are ultrametric.

If we now delete from \mathcal{T}_t all lineages that do not have a descendant leaf extant at the sampling time τ , and from this resulting tree we further delete the edge from s to the first remaining speciation event, we obtain the *reconstructed tree*, denoted $\tilde{\mathcal{T}}_t$ (note that this tree depends not just on t but on τ). The distinction between the complete tree and reconstructed tree is illustrated in Fig. 9.2. In the particular case where there is just one species extant at τ then one can take the reconstructed tree to mean just this single species (with no edge). Notice that species that have become extinct, and the initial species s do not appear in this tree, and the reconstructed tree inherits its edge lengths from \mathcal{T}_t in usual way (i.e., the length of an edge e in the reconstructed tree is the sum of the lengths of the edges in the path of the complete tree that corresponds to e). Note also that the number of leaves of $\tilde{\mathcal{T}}_t$ is less or equal to that of \mathcal{T}_t and the two numbers agree at $t = \tau$. When $t = \tau$ we will just write \mathcal{T}_t and $\tilde{\mathcal{T}}_t$ as \mathcal{T} and $\tilde{\mathcal{T}}$, respectively.

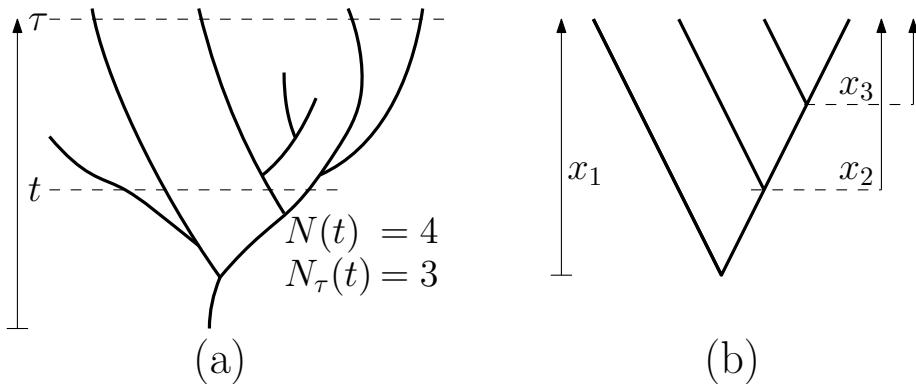


Figure 9.2. (a) the complete tree \mathcal{T} , and (b) the reconstructed tree $\tilde{\mathcal{T}}$. In (a) there are $N(t) = 4$ lineages present at the time slice shown. However, only three of these lineages survive to the sampling time τ and so $N_\tau(t) = 3$. In (b), x_i refers to the time from the i th speciation event and the sampling time τ .

As in the pure-birth setting, we may condition on the number of leaves of the reconstructed tree, its height x_1 (the time from the first speciation event to the present) or both. If we condition just on n , then we face a similar (but more subtle) issue concerning how

exactly the tree should be sampled. When both speciation and extinction can occur, it is possible for $N(t)$ to equal n for several (disconnected) intervals for t , though the number and total length of these intervals is finite (with probability 1).

One way around this problem is to place a prior on the height t of the tree that is uniform on $[0, \tau]$ and then sample trees with height t and $N_t = n$ (where t is selected uniformly on $[0, \tau]$). This still involves a dependence on τ ; however, one can take the limit as $\tau \rightarrow \infty$ (an “improper prior”) to obtain a distribution on the reconstructed tree conditional on n leaves that depends just on n , λ and μ . It turns out that for the Yule pure-birth process, this is precisely the tree $\mathcal{T}_{(n)}$ defined in Section 9.1.1. There, we saw that the expected length of a randomly selected pendant edge in the reconstructed tree is exactly $\frac{1}{2\lambda}$. When $\mu > 0$, the pendant edges are longer on average than for pure-birth; more precisely, if $\mathbb{E}[p_n]$ is the expected length of a randomly selected pendant edge in a reconstructed birth-death tree on n extant species, then for $0 < \mu < \lambda$,

$$\mathbb{E}[p_n] = \frac{\mu + (\lambda - \mu) \ln(1 - \mu/\lambda)}{\mu^2},$$

which increases towards the critical case $\mu = \lambda$, for which

$$\mathbb{E}[p_n] = \frac{1}{\lambda}. \quad (9.12)$$

The expected number of lineages with surviving descendants. Let $N_\tau(t)$ denote the number of species that are present in the complete tree at time t for $0 \leq t \leq \tau$ that have at least one surviving species at time τ (note that when $N_\tau(t) > 1$ this is the same as the number of species in the reconstructed tree \mathcal{T}_τ at time t). The distinction between $N_\tau(t)$ and $N(t)$ is illustrated in Fig. 9.2(a). Notice that $N_\tau(t)$ can be written as follows:

$$N_\tau(t) = Y_1(t) + \cdots + Y_{N(t)},$$

where Y_i is the binary (0,1) random variable that takes the value 1 if a given species present at time t has at least one descendant at time τ , which is an event that has probability $1 - p_0(\tau - t)$. Thus,

$$\mathbb{E}[N_\tau(t)|N(t)] = N(t) \cdot (1 - p_0(\tau - t)),$$

and so, by the law of total expectation,

$$\mathbb{E}[N_\tau(t)] = \mathbb{E}[\mathbb{E}[N_\tau(t)|N(t)]] = \mathbb{E}[N(t)] \cdot (1 - p_0(\tau - t)).$$

Since $\mathbb{E}[N(t)] = e^{rt}$ from eqn. (9.8), this gives

$$\mathbb{E}[N_\tau(t)] = e^{rt} \cdot (1 - p_0(\tau - t)). \quad (9.13)$$

This will be useful in the next section.

Two helpful properties of birth-death models. For the general birth-death setting, a fundamental property of \mathcal{T} is that, conditional on the number of leaves, and the height x_1 , the random times $x_2 > \cdots > x_{n-1}$ of the speciation events (cf. Fig. 9.2(b)) are the order statistics of $n-2$ independent and identically distributed variables. This makes it possible to calculate the probability density function of \mathcal{T} noting that (i) each labeled history is

equally probable for these models (cf. Section 3.2.1), and (ii) given a labeled history along with x_1 and the speciation times x_2, \dots, x_{n-1} , these values fully determine the edge lengths. The reason for this connection to order statistics will become apparent in Section 9.2.3.

Another interesting property of birth-death model is that if we prune leaves at random from the reconstructed tree, then the resulting reconstructed tree can often also be described by the same birth-death model, but with adjusted speciation and extinction rates. More precisely, suppose that $\tilde{\mathcal{T}}$ is the reconstructed tree in a birth-death process with speciation and extinction rates λ and μ , and that each leaf of $\tilde{\mathcal{T}}$ is sampled independently with probability f . If \mathcal{Y} is the set of leaves sampled, the restricted tree $\tilde{\mathcal{T}}|\mathcal{Y}$, with its induced edge lengths, has exactly the same distribution as a reconstructed tree generated under a birth-death model with the modified parameters $(\hat{\lambda}, \hat{\mu})$ provided that $0 < f \leq 1$ and $1 \geq \frac{\mu}{\lambda} \geq 1 - f$. The modified parameters are defined by

$$\hat{\lambda} = f\lambda, \text{ and } \hat{\mu} = \mu - \lambda(1-f).$$

Notice that the diversification rates of the two processes (i.e., $\hat{\lambda} - \hat{\mu}$ and $\lambda - \mu$) are equal, and the restriction $\frac{\mu}{\lambda} \geq 1 - f$ is required in order that $\hat{\mu}$ is nonnegative. Moreover, this equivalence holds under all three sampling scenarios (conditioning on the number of extant leaves or on the time from the first speciation event to the present, or both); for details, see [324].

9.2.2 • The “pull of the present” and “push of the past”

The unusual title of this section refers to two predictions of birth-death models in phylogenetics. One of these predictions (the “pull of the present”) asserts, roughly speaking, that the rate of diversification (speciation minus extinction) inferred from the reconstructed tree increases as we approach the present. A similar burst (concave rather than convex) near the root of the complete tree is also predicted (the “push of the past”). These phenomena and their possible applications were explored by Sean Nee and colleagues in the mid-1990s [275]. The relevance for biologists is in providing a possible way to decouple speciation and extinction rates (λ and μ) so that these can be estimated separately from empirical phylogenies (rather than via the average slope of (logarithmic) lineage-through-time plots which reveal only the difference $r = \lambda - \mu$).

In this section, we condition on the event \mathcal{E}_τ that $N(\tau) > 0$ (i.e., at least one species is extant at the sampling time τ). Notice that $\mathbb{P}(\mathcal{E}_\tau) = 1 - p_0(\tau) = r / (\lambda - \mu \exp(-r\tau))$. From eqn. (9.8), $\log \mathbb{E}[N(t)] = rt$ grows linearly with t . How does conditioning on \mathcal{E}_τ affect this growth rate for $N(t)$ and $N_\tau(t)$? Let us consider the latter first. By the law of total expectation,

$$\mathbb{E}[N_\tau(t)] = \mathbb{E}[N_\tau(t)|\mathcal{E}_\tau]\mathbb{P}(\mathcal{E}_\tau) + \mathbb{E}[N_\tau(t)|\overline{\mathcal{E}}_\tau]\mathbb{P}(\overline{\mathcal{E}}_\tau).$$

Notice that when the complementary event $\overline{\mathcal{E}}_\tau$ occurs (i.e., all species have become extinct by time τ), we have $N_\tau(t) = 0$ with probability 1 for all $0 \leq t \leq \tau$. Thus, since $\mathbb{P}(\mathcal{E}_\tau) = 1 - p_0(\tau)$, we have $\mathbb{E}[N_\tau(t)|\mathcal{E}_\tau] = \frac{\mathbb{E}[N_\tau(t)]}{(1 - p_0(\tau))}$. If we now combine this equation with eqn. (9.13), we get

$$\mathbb{E}[N_\tau(t)|\mathcal{E}_\tau] = e^{rt} \cdot \frac{1 - p_0(\tau - t)}{1 - p_0(\tau)}, \quad (9.14)$$

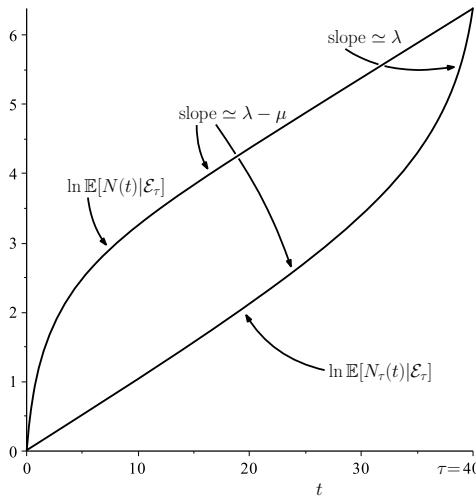


Figure 9.3. Plot of the natural log of the expected number of lineages of the complete tree (top curve) and of the lineages that have at least one extant descendant at time $\tau = 40$ (bottom curve), both conditioned on survival to the present time $\tau = 40$ under a birth-death model with $\lambda = 1.1$ and $\mu = 1$.

which is eqn. (2) in [177]. In the critical case where $\lambda = \mu$, eqn. (9.11) gives a result from [277]:

$$\mathbb{E}[N_\tau(t)|\mathcal{E}_\tau] = \frac{1 + \lambda\tau}{1 + \lambda(\tau - t)}. \quad (9.15)$$

In this case, $\mathbb{E}[N_\tau(\tau)] = 1$ for all τ ; however, the conditional expectation $\mathbb{E}[N_\tau(\tau)|\mathcal{E}_\tau]$ grows linearly with τ .

A similar treatment applies for the complete tree and the expected number of species extant at time t , given that there is at least one species surviving at time τ . In this case, one can perform a similar (though slightly more complex) calculation for $\mathbb{E}[N(t)|\mathcal{E}_\tau]$. From [177], we have

$$\mathbb{E}[N(t)|\mathcal{E}_\tau] = \frac{e^{rt}}{1 - p_0(\tau)} - \frac{p_0(\tau) - p_0(t)}{(1 - p_0(t))(1 - p_0(\tau - t))}. \quad (9.16)$$

Consider the logarithm of the expected lineage-through-time plots of the complete tree and the lineages in the complete tree that have at least one descendant species extant at time τ , in both cases conditioned on nonextinction at the present (i.e., \mathcal{E}_τ):

$$\ln \mathbb{E}[N_\tau(t)|\mathcal{E}_\tau] \text{ and } \ln \mathbb{E}[N(t)|\mathcal{E}_\tau],$$

In contrast to $\ln[\mathbb{E}[N(t)]]$, these last two functions are no longer linear in t . This nonlinearity becomes significant for $\ln \mathbb{E}[N_\tau(t)|\mathcal{E}_\tau]$ near $t = \tau$, where the slope of the expected lineage-through-time plot increases from $r = \lambda - \mu$ to λ (the “pull of the present”), and for the complete tree near $t = 0$, where the slope increases from $r = \lambda - \mu$ to a value greater than λ as we approach $t = 0$ (the “push of the past”). This is illustrated in Fig. 9.3.

Using the equations above, it can be shown that for a birth-death model with parameters (λ, μ) , with $\lambda > \mu$, starting with a single species at time 0 and conditioned on survival at some time τ ,

$$\lim_{t \rightarrow \tau^-} \frac{d}{dt} \ln \mathbb{E}[N_\tau(t)|\mathcal{E}_\tau] = \lambda. \quad (9.17)$$

An informal explanation of the increase in the slope $\ln \mathbb{E}[N_\tau(t)|\mathcal{E}_\tau]$ as we approach the sampling time τ , is that new species have formed but have not yet become extinct. The detection of this “pull of the present” prediction in real phylogenies seems to be compromised by other factors, including protracted speciation effects [129].

Exercise: Verify eqn. (9.17) for the critical birth-death model using eqn. (9.15).

9.2.3 • Coalescent point process models, and unlabeled ranked trees

So far, we have described birth-death trees as proceeding forward in time, then being sampled according to various scenarios, with the focus on the resulting reconstructed tree. However, there is an alternative and more direct way to describe this random reconstructed tree. This description is both elegant and perfectly tailored for certain calculations (such as the estimation of parameters or predicting the properties of different models) as well as for simulations. This alternative representation was described for certain particular processes (e.g., the conditioned critical branch process [4]) and was developed in greater generality more recently by Amaury Lambert and Tanja Stadler [225].

The idea is to view the reconstructed tree as being described by a *coalescent point process* (CPP) (not to be confused with the well-known Kingman’s coalescent process, discussed below), in which we have a continuous random variable H described by some density function f on $(0, \infty)$, which is dependent on the model of tree evolution and its parameters. Let us sample H independently several times to get a sequence H_1, H_2, \dots , of independent and identically distributed (i.i.d.) random variables. For a range of models that we will describe shortly (including the birth-death models discussed earlier), and for two of the scenarios for the reconstructed tree (conditioning on t or on n and t) the probability distribution on the reconstructed tree can be described by the following generic tree-building procedure (with minor modifications between the two scenarios).

Given H_1, H_2, \dots, H_m , we construct a tree with m leaves by a similar procedure to that described in Section 3.2.1. Place the numbers $0, 1, \dots, m - 1$ in this order along a horizontal axis, draw vertical lines l_1, \dots, l_m of length H_1, \dots, H_m below $1, \dots, m$ and place a vertical line of infinite length below 0. For each line l_i , $i \geq 1$, draw a horizontal line to the left of l_i until it intersects the first vertical line it meets (i.e., the largest $j < i$ for which $H_j > H_i$); this is illustrated in Fig. 9.4. In this way, we obtain a ranked oriented tree on the leaf set $\{0, 1, \dots, m\}$.

We now explain how birth-death trees with conditioning on t , or on n and t are obtained. For conditioning the reconstructed tree to have height t we take m (in H_1, \dots, H_m) to be the last value j for which $H_j < t$ (see Fig. 9.4). For conditioning on both n and t , from the sequence H_1, H_2, \dots , we ignore any outcomes H_i that are greater than t and take the first n values that are less than t (equivalently, we use the modified distribution of H_i that conditions on the event $H_i \leq t$).

Notice how conditioning on t immediately shows that the number $N(t)$ of leaves in the reconstructed tree at time t has a geometric distribution (since $\mathbb{P}(N(t) = n)$ is the probability that the first $n - 1$ i.i.d. samples of H are less than t , but the n th is greater).

Which models have the property that the reconstructed tree can be exactly modeled stochastically by a CPP for an appropriate choice of f ? The class includes the birth-death processes with the fixed parameters (λ, μ) we considered earlier, including the pure-birth model with $\mu = 0$ and the critical model with $\mu = \lambda$. In the case of the pure-birth model, with birth rate λ , the density for H is the exponential distribution with mean $\frac{1}{\lambda}$.

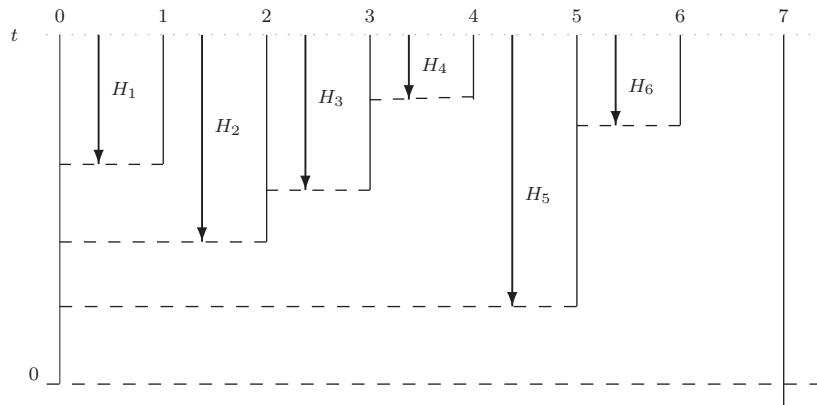


Figure 9.4. Illustration of a CPP showing the node depths H_1, \dots, H_6 for each of the six consecutive pairs of tips. The node depth H_7 is the first one which is larger than t .

The class of models describable by a CPP also includes more general birth-death processes that allow the speciation and extinction rates to depend on time. Allowing time-variable rates recognizes that evolution has not always progressed at a uniform rate. For example, there have been several short bursts of mass extinction (e.g., at the KT boundary around 65 million years ago) and other bursts of speciation (e.g., the Cambrian explosion more than 500 million years ago). One can also sample each leaf with a given probability f and so form the “pruned” reconstructed tree on this subset. Again, this will be described by a CPP but with an adjusted density f .

As well as time dependence in the speciation and extinction rates, the CPP model also allows the extinction to depend on the *age* of the species. This quantity needs some explanation. Assuming that the initial species s_0 has age 0, as time increases each species in the tree increases its age linearly with time, until it becomes extinct (at which point its age remains fixed) or it reaches a speciation event. In the latter case, the age of one of the child lineages is reset to zero (it is the “new” species that branches off the existing lineage) while the other lineage continues to increase its age linearly with time. In summary, the CPP allows for the following:

- The speciation rate of the tips of the complete tree at each time point t is allowed to vary with t . We denote this rate as $\lambda(t)$.
- The extinction rate μ of any tip of the complete tree at each time point t can vary both with time t , and the length a of the pendant edge incident with that tip (a is the (asymmetric) “age” of that tip described above). We denote this rate as $\mu(t, a)$.
- Each leaf at the present can be sampled independently with probability f and the reconstructed tree can be restricted to this subset to give a pruned reconstructed tree.

Notice how much simpler simulations are under CPP. For example, if we want to sample a reconstructed birth-death tree with n leaves and height t , then the “pedestrian” approach would require one to simulate the evolution of a complete tree for time t , form the reconstructed tree, and then throw away all of the samples that did not have exactly n leaves. By contrast, the CPP just requires one to draw $n - 1$ samples from the same distribution and construct the tree by the simple procedure described above. The CPP

representation is also ideal for performing likelihood calculations (and estimating parameters such as μ and λ), and for establishing further analytic properties of models that can be described in this way [225].

Notice also that a CPP leads to the YH distribution on binary trees shapes when we ignore the edge lengths of the tree. To see why, it helps to consult Fig. 3.4 from Chapter 3. The tree shape generated by the CPP is obtained by first sampling a ranked oriented tree under the uniform distribution and ignoring the orientation to give a distribution on unlabeled ranked trees. We will call the distribution the unlabeled ranked tree (URT) distribution. If we now also ignore the ranking we obtain a tree shape that is identical to that of YH shapes (cf. arrows B, C, and D in Fig. 3.4). The URT distribution is the same as the distribution obtained by taking the uniform distribution on labeled histories and ignoring the labeling (cf. arrow A in Fig. 3.4), and so this distribution has also been called the “uniform distribution on ranked tree shapes.”

Thus for reconstructed trees generated by any CPP model or a tree generated by the Kingman coalescent process (discussed in more detail below), the discrete ranked tree (i.e., ignoring edge lengths) follows the URT distribution. In [225], the authors show that the class of speciation and extinction models that lead to the URT distribution for the reconstructed tree is quite a bit larger than CPP. Recall that for CPP, it was sufficient that $\lambda = \lambda(t)$, and $\mu = \mu(t, a)$; however, if we also want one or both of λ and μ to also depend on N_t , the number of species present at time t in the complete tree, then we obtain the URT distribution on the discrete reconstructed tree. This extra generality of “density-dependent” speciation and extinction rates can lead to tree distributions that are no longer described by the CPP.

The ubiquitous nature of the URT distribution under a range of speciation and extinction scenarios was noted in an early paper by Aldous in 1995 [2], and discussed further in [323]. In this later paper, a simple argument is presented to show that the URT distribution arises for the reconstructed tree (and for any tree obtained from it by uniform random sampling of the species at the present) under a general assumption that includes all of the models discussed above. This is the assumption of *species-speciation-exchangeability* which requires (i) at any point in time given a speciation event, each species is equally likely to be the one that speciates, and (ii) the discrete speciation-extinction pattern descending from a speciation event is independent of which species was the one speciating.

However, the link to URT disappears if we allow the speciation rate λ to depend on the age of the species (a). The link also disappears even when the speciation rate is constant and the age is defined symmetrically for each child of a vertex (an example from Section 2.2.1 of [225] shows this). Further results concerning the statistical properties of time-dependent birth-death trees in phylogenetics have been derived recently in [282].

Kingman coalescent trees. In population genetics, the famous *Kingman coalescent process* is a stochastic model that describes how the ancestry of a gene from a sample of individuals traces back in time through generations of a population to a common ancestor. By default, the model also produces a rooted binary tree with edge lengths, so an obvious question is how these relate to the trees described by the processes we have described so far.

The motivation and detailed analysis of the Kingman coalescent model can be found in any book on population genetics (e.g., [182]), so we give only the briefest summary here. Starting with a set X of size n , the process produces a sequence of sets $X = X_n, X_{n-1}, \dots, X_1$ of reducing size, where $X_{i-1} = X_i - \{x, x'\} \cup \{\{x, x'\}\}$ for some $x, x' \in X_i$. Thus at each step, two elements are combined (they “coalesce”) to form the next set in the sequence, reducing the set size by 1, until just one element remains in X_1 . It is assumed that each

pair of elements $x, x' \in X_i$ coalesces independently at a constant rate 1, so the time T_i until the coalescent event that converts X_i to X_{i-1} is the minimum of $\binom{i}{2}$ independent exponential random variables with a mean of 1. The variables T_n, T_{n-1}, \dots, T_2 are assumed to be independent. This model gives rise to a rooted binary tree with (ultrametric) edge lengths which is uniform on labeled histories (see Fig. 3.4).

Kingman coalescent trees share one characteristic with CPP models; if one ignores the edge lengths, then the Kingman coalescent produces the URT distribution and hence gives rise to the YH distribution on rooted binary tree shapes. However, when one considers edge lengths, Kingman coalescent trees tend to look very different from (say) birth-death trees as n grows, except in certain extreme cases (e.g., when $\mu = \lambda$ and the tree is conditioned on having n leaves). Although Yule pure-birth trees tend to have edges of approximately similar lengths, a Kingman coalescent tree with large n has many very short edges near the leaves, and two long edges incident with the root.

More precisely, if $\mathcal{H}_n^{\text{KC}}$ is the total height of the Kingman coalescent tree, and $\mathcal{L}_n^{\text{KC}}$ is the total sum of edge lengths, then we can write $\mathcal{H}_n^{\text{KC}}$ and $\mathcal{L}_n^{\text{KC}}$ as linear functions of the coalescent times T_i (recall that these are independent and that T_i has an exponential distribution with mean $\binom{i}{2}$). Specifically,

$$\mathcal{H}_n^{\text{KC}} = \sum_{i=2}^n T_i \text{ and } \mathcal{L}_n^{\text{KC}} = \sum_{i=2}^n i T_i.$$

Consequently,

$$\mathbb{E}[\mathcal{H}_n^{\text{KC}}] = 2 \sum_{i=2}^n \frac{1}{i(i-1)} = 2 \left(1 - \frac{1}{n}\right)$$

and

$$\mathbb{E}[\mathcal{L}_n^{\text{KC}}] = 2 \sum_{i=2}^n \frac{1}{i-1} \sim 2 \ln n.$$

Thus, the expected length of a randomly selected edge in a Kingman coalescent tree for large n is (asymptotically) $\frac{\ln n}{n} \rightarrow 0$, whereas the shorter of the two edges incident with the root of the tree has an expected length of 1.

Although the Kingman coalescent tree is different from the CPP models, there is a remarkable link to the constant-rate birth-death model for which $\mu = \lambda$. Suppose that we condition on a critical birth-death tree having n leaves at the present (and take a uniform improper prior on the height of the tree as noted earlier). This *conditioned critical branching process* (cCBP) was developed and studied by David Aldous and Lea Popovic, culminating in an enlightening discussion of its relevance to questions in systematic biology in [5]. The cCBP model leads to a different distribution on the reconstructed tree than the Kingman coalescent model. However, since both produce a uniform distribution on ranked trees, we can easily compare their edge lengths by considering the times at which the number of lineages increases by 1. It turns out that if we set $\mu = \lambda = n/2$ in the cCBP, then the expected edge lengths in the two models agree [154].⁶³ Equation (9.12) then shows that the expected length of a randomly selected pendant edge in a Kingman coalescent tree is of order n^{-1} .

Gamma statistic. Given any rooted binary phylogenetic tree with temporal edge lengths, the *gamma statistic* γ , introduced to phylogenetics by Pybus and Harvey [292], provides a measure of the extent to which the longer edges are distributed near the root

⁶³The joint distribution does not, however.

($\gamma > 0$) or near the tips ($\gamma < 0$) of the tree. It is sometimes said to measure the “steminess” of the tree, in contrast to the balance measures we described in Chapter 3, which depended only on the tree shape, not the edge lengths.

For $i = 2, \dots, n$ let g_i be the length of time for which there are i lineages extant. The gamma statistic is defined by

$$\gamma = \frac{\frac{1}{n-2} \sum_{i=2}^{n-1} \sum_{k=2}^i k g_k - \frac{S_n}{2}}{Y_n}$$

where $S_n = \sum_{j=2}^n j g_j$ and $Y_n = S_n / \sqrt{12(n-2)}$.

An equivalent (and slightly more concise) expression for γ is

$$\gamma = \frac{1}{(n-2)Y_n} \sum_{i=2}^n \left(\frac{n}{2} - i + 1 \right) i g_i.$$

For a Yule pure-birth model (with a constant birth rate) $\mathbb{E}[\gamma] = 0$. Moreover, γ is (asymptotically with n) distributed according to a standard normal distribution (cf. Section 3.3 of [92] and eqn. (11) therein). By contrast, for trees generated under Kingman’s coalescent process, the expected value of γ is positive and grows with n at the rate \sqrt{n} [261].

9.2.4 • Predicting future PD loss

In Section 6.4 of Chapter 6, we considered the expected loss of phylogenetic diversity (PD) under a “field of bullets” model of extinction acting at the present. This analysis assumed that we had a fixed phylogeny and fixed edge lengths. Here we ask how much PD we expect to lose under a simple field of bullets model for a “typical tree” (i.e., one generated by a random process of speciation and extinction, of the type considered in the previous chapter). We begin with the simplest model, the Yule pure-birth process.

Suppose we grow a Yule pure-birth tree \mathcal{T} for time t . Let \mathcal{L}_t be the total length of the tree; in other words, the sum of the lengths of all the edges of \mathcal{T} . Let us now model an “extinction at the present” event by a simple field of bullets model, in which each species survives independently with probability s . Let $\mathcal{L}_t(s)$ be the resulting total length of the subtree of \mathcal{T} induced by just the leaves that survive. In other words, $\mathcal{L}_t(s) = PD_{\mathcal{T}}(\mathcal{Y})$, where \mathcal{Y} is the subset of leaves that survive the extinction event at the present. Notice that $\mathcal{L}_t(s)$ is a random variable that depends on two compound random processes: first, the random process that generates the tree; second, the random pruning of leaves at the present.

Notice also that $\mathcal{L}_t = \mathcal{L}_t(1)$. Now consider $\mu_t(s) := \mathcal{L}_t(s)/\mathcal{L}_t(1)$, which is the proportion of the total length of the tree that is still present after the extinction event at the present. Clearly, $\mu_t(s)$ takes values in $[0, 1]$ with $\mu_t(0) = 0$ and $\mu_t(1) = 1$. It can be shown that as $t \rightarrow \infty$, (i) $\mu_t(s)$ converges in probability to a constant (dependent only on s), which we will call $\mu(s)$, and (ii) $\mu(s)$ is the same as the limiting ratio of the expected value of $\mathcal{L}_t(s)$ to that of $\mathcal{L}_t(1)$. In other words, if we let $\pi_t(s) = \mathbb{E}[\mathcal{L}_t(s)]$, then $\mu(s) = \lim_{t \rightarrow \infty} \pi_t(s)/\pi_t(1)$.

There is a concise exact expression for $\pi_t(s)/\pi_t(1)$, and it is identical for the single-start and double-start scenarios. We will assume the single-start scenario (the calculations are slightly simpler). To calculate $\pi_t(s)$, an infinitesimal argument involving t (see [261]) for $\pi_t(s)$ leads to the following first-order linear differential equation:

$$\frac{d\pi_t(s)}{dt} - \lambda \pi_t(s) = \frac{s}{s + (1-s)e^{-\lambda t}}. \quad (9.18)$$

This equation has the following solution for $s \neq 1$:

$$\pi_t(s) = \frac{s}{(1-s)\lambda} \cdot [-\ln(s + (1-s)e^{-\lambda t})]. \quad (9.19)$$

For $s = 1$, the solution to (9.18) is $\pi_t(1) = \frac{1}{\lambda}(e^{\lambda t} - 1)$, in agreement with eqn. (9.1).⁶⁴ By taking the limit of the ratio $\pi_t(s)/\pi_t(1)$ as $t \rightarrow \infty$, we arrive at the following concise formula from [261], which is independent of λ . For all $s \in (0, 1)$,

$$\mu(s) = \frac{-s \ln s}{1-s}, \quad (9.20)$$

and $\mu(0) = 0$, $\mu(1) = 1$. This describes the limiting (as $t \rightarrow \infty$) proportion of the PD of the surviving leaves (each survives independently with probability s) to the total PD of the tree [226].

Notice that we can expand $\mu(s)$ from eqn. (9.20) as a power series in $q = 1 - s$ as follows:⁶⁵

$$\mu(s) = 1 - \sum_{k \geq 1} \frac{1}{k(k+1)} q^k = \sum_{k \geq 1} \frac{1}{k(k+1)} (1 - q^k). \quad (9.21)$$

It is instructive here to recall eqn. (6.22) from Chapter 6. This showed that for a given tree with branch lengths, and the random subset \mathcal{Y} of leaves that survive a simple extinction event at the present (with $q = 1 - s$ being the extinction probability per species) we have

$$\mathbb{E}[PD(\mathcal{Y})] = \sum_{k \geq 1} S(k)(1 - q^k), \quad (9.22)$$

where $S(k)$ is the sum of lengths of those edges that have exactly k descendant leaves. This equation provides another way to interpret $\mu(s)$ as given by (9.21), since, if we ignore edge lengths for a moment, Yule pure-birth trees have the discrete YH distribution, and eqn. (3.6) from Chapter 3 showed that for such (discrete) trees the expected number of edges with k descendant leaves is proportional to $\frac{1}{k(k+1)}$. Moreover, as the edges have approximately equal lengths (from Section 9.1) and so $S(k)$ in eqn. (9.22) should be proportional to $\frac{1}{k(k+1)}$ as $t \rightarrow \infty$. Of course, this interpretive argument does not constitute a formal proof of eqn. (9.20).

Extension to more general birth-death models. The analysis for PD loss on a Yule pure-birth tree is based on a very simple model. Thus it is instructive to ask how the results for PD loss extend for more complex models that allow constant or rate-dependent speciation and extinction events in the generation of the tree, followed by an extinction process at the present in the reconstructed tree.

For the class of CPPs discussed in Section 9.2.3, two general results have been derived, which we describe briefly here (for full details, see [226]). Under a CPP, consider the (random) reconstructed tree $\tilde{\mathcal{T}}$ at time t (i.e., the induced subtree based on the species that are present at time t) and let $\mathcal{L}_t(s)$ be the PD of the (random) set \mathcal{Y} of leaves of this tree that survive under a simple field of bullets model with the survival probability s . Consider a supercritical process (i.e., the expected number of species is growing with time), in which the speciation rate is independent of time (but the extinction rate may

⁶⁴This also agrees with the limit of the expression in (9.19) in the limit as $s \rightarrow 1-$.

⁶⁵The second equality is justified by the identity $\sum_{k \geq 1} \frac{1}{k(k+1)} = 1$.

depend on time as well as on the age of the species) and let us condition on $\tilde{\mathcal{T}}$ having at least one species present. In this setting, it can be shown that the ratio $\mathcal{L}_t(s)/\mathcal{L}_t(1)$ converges in probability (as $t \rightarrow \infty$) to $v(s)$, which is also (as above) the (deterministic) limit of the ratio $\mathbb{E}[L_t(s)]/\mathbb{E}[L_t(1)]$ as $t \rightarrow \infty$. For a pure-birth Yule process, we know what $v(s)$ is: it equals $\mu(s)$, as given by eqn. (9.20). For a general CPP, $v(s)$ is given explicitly as the ratio

$$v(s) = s \int_0^\infty \frac{1-F(t)}{1-(1-s)F(t)} dt \Bigg/ \int_0^\infty (1-F(t))dt.$$

Here, $F(s) = \mathbb{P}(H \leq t)$ is the probability distribution function of the node depth variable H in the CPP model (cf. Section 9.2.3). In the case of a constant-rate birth-death model, with speciation and extinction rates b and d , respectively, with $b > d$, $v(s)$ is described by the equation

$$v(s) = \frac{sx}{s+x-1} \cdot \frac{\ln(s/(1-x))}{\ln(1/(1-x))}, \quad (9.23)$$

where $x = d/b \neq 0, 1-s$. The cases where $x \in \{0, 1-s\}$ are obtained by taking the limit as x tends to those values in eqn. (9.23). Taking the limit as $x \rightarrow 0$ agrees with eqn. (9.20) for the special case where $d = 0$. An interesting feature of the curve $s \mapsto v(s)$ for this birth-death model is that it lies relatively close to the curve for $d = 0$ until d gets quite close to b ; however, in the limit as $d \rightarrow b$ (i.e., as $x \rightarrow 1$), $v(s)$ converges to the (very different) unit step function $U(s)$, where $U(0) = 0$ and $U(s) = 1$ for $s \in (0, 1]$.⁶⁶ This is illustrated in Fig. 9.5. Notice how different models can lead to quite different predictions concerning PD loss. For example, in one well-cited paper, the authors stated that “80% of the underlying tree of life can survive even when approximately 95% of species are lost” [276]. However, the model used in that study was a Kingman coalescent-type process, which had an expected edge length distribution equivalent to the limiting case as $b-d \rightarrow 0+$. By contrast, for a Yule pure-birth tree, if we lose 95% of species (randomly),

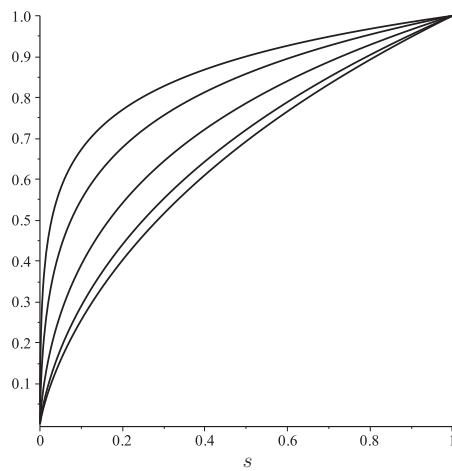


Figure 9.5. The slow progression of the curve $s \mapsto v(s)$ towards the unit step function $U(s)$ for a constant-rate birth-death process for $d/b = 0$ (the lowest curve) and the progressive curves above it where $d/b = 0.5, 0.9, 0.99$, and 0.999 .

⁶⁶This can be verified by taking the limit as $x \rightarrow 1$ in eqn. (9.23).

then we would expect to *lose* more than 84% of the PD [261]. Equation (9.23) (or Fig. 9.5) shows that even if the extinction rate were 90% of the speciation rate, the predicted PD loss would still be around 73% of PD.

A second result fixes $t = t_0$ and conditions on the number N_t of leaves of T_t (in this case, the speciation rate can depend on time). In this case, the ratio of the surviving PD to the original PD in the random tree converges almost surely to a constant $\mu'(s)$ as $N_t \rightarrow \infty$. The expression for $\mu'(s)$ is similar to the earlier expression for $v(s)$ in terms of $F(t)$: one simply replaces $F(t)$ by $F(t)/F(t_0)$ and the upper integral limits of ∞ by t_0 . For further details and proofs, see [226].

9.3 • Gene trees and species trees

When biologists construct a phylogenetic tree, it is often a tree that describes the phylogeny of a particular gene that is shared across all (or some of) the species under study. This gene will have been sampled and sequenced from an individual organism of each species. If we trace the ancestry of this gene back in time it will describe some phylogenetic tree. This *gene tree* lies within the a larger *species tree* which describes the phylogenetic relationships between the species under study.

However, the phylogeny of a gene may differ from that the species tree it lies within. For sexually reproducing diploid species, a population genetic process called *incomplete lineage sorting* (ILS) causes genes to coalesce in their ancestry in a way that is partly random, and partly constrained by the species tree (we will explain this shortly). Essentially, ILS arises because each individual receives their copy of the gene randomly from one of their two parents. For prokaryotic species (e.g., bacteria), a different process, called *lateral gene transfer* (LGT), causes gene trees to disagree with a species tree. LGT is a process whereby genes can be transferred or exchanged between lineages of the species tree. This process is often also referred to as *horizontal gene transfer* (HGT). There are other mechanisms as well, such as gene duplication and loss, which can confound the simple “gene tree as proxy for the species tree” idea.

Fortunately, new sequencing technologies have allowed biologists to base their phylogenetic estimates of species relationships not just on one gene or a few genes, but on hundreds (or even thousands) of genes. This leads to a host of interesting mathematical questions as to how best to reconcile these conflicting phylogenies to obtain a consistent estimate of the species tree (see, e.g., [10, 220, 264]). This area of “gene tree/species tree” analysis is currently one of the most active areas of research in phylogenetics (Lacey Knowles and Laura Kubatko provide a detailed overview in their 2010 book [220]).

In the previous sections of this chapter, we have treated the species tree as a random variable resulting from a stochastic process of speciation and extinction events over large timescales. Even if we condition on this “random species tree,” there is now a further random process producing the gene tree. Moreover, if the gene trees are inferred from molecular sequence data, then the models we considered in Chapter 7 and 8 mean that these sequences should be viewed as a third random variable (conditioned on the other two); in other words, at the level of DNA sequences, the process can be viewed as a compound random variable of three iterated random processes.

Notation: In this chapter, a gene tree will be denoted T_g and a species tree will be denoted T_s . These both refer to a rooted binary phylogenetic tree with the leaf set $[n]$. When the gene tree is randomly generated by a stochastic model for ILS (and later for LGT), we will denote this random tree as \mathcal{T}_g .

9.3.1 • The multispecies coalescent and ILS

The key model for studying ILS is the *multispecies coalescent* model. This is an extension of the Kingman coalescent on a single population to a process on a tree of populations that split at each speciation event. Given a phylogenetic species tree T_s , by sampling a gene g from an individual in each species, the corresponding gene tree T_g is obtained by tracing back the genealogy of this gene in time until all copies find a common ancestor. Thus, the coalescent process is constrained by the tree T_s , since two copies of g initially present in species x and y cannot coalesce until we reach the edges of T that lie on the root side of $\text{lca}_{T_s}(x, y)$.

Within each edge e of the species tree T_s , there will be a certain number n' of genes that are nearest to the present. As we travel back in time, these will be subject to the Kingman coalescent process so that at the other end of the edge e the number will be less or equal to n' . For example, in Fig. 9.6 there are $n' = 4$ gene lineages at the top end of the vertical edge shown; then (tracing back in time) two coalescence events ($n' = 2$). At the lower end of the vertical edge this increases to $n' = 3$ when the sole gene lineage from the other edge joins in.

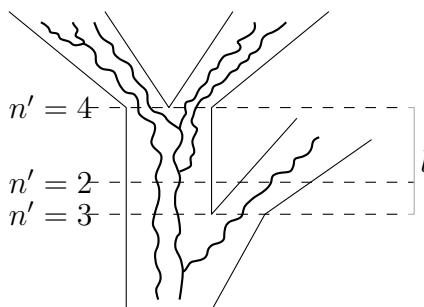


Figure 9.6. The operation of the Kingman coalescent process within the edges of the species tree. For the vertical edge, there are $n' = 4$ lineages entering the edge (closest to the present) which reduce to $n' = 2$ as we trace back in time. The number increases to $n' = 3$ when a further (single) gene lineage joins in.

The duration of the edge l is proportional to the temporal length of edge e divided by the effective population size of individuals comprising the species at any time in this edge, and is sometimes referred to as *coalescent time units*. Thus, genetic lineages are more likely to coalesce if the edge has a large temporal length or a small effective population size of that species (at any time slice across the edge) or both of these (these two effects reinforce each other); this is illustrated in Fig. 9.7.

Let $\mathbb{P}(\mathcal{T}_g = T_g | T_s, l)$ be the probability of generating the gene tree T_g on the species tree T_s with edge lengths l . This quantity can be calculated by summation over all the possible (but finite) number of “coalescent histories” (defined formally in [304]). A key quantity required is the probability that $p_{ij}(l)$ i lineages become j (where $1 \leq j \leq i$) after duration l under the Kingman coalescent process. For example, $p_{21}(l) = 1 - e^{-l}$, and $p_{22}(l) = e^{-l}$. Similarly, $p_{nn}(l) = \exp(-\binom{n}{2})$, for all $n \geq 2$. However, other expressions for $p_{ij}(l)$ are less obvious. For example, $p_{52}(l) = 2e^{-l} + \frac{30}{7}e^{-3l} + 3e^{-6l} - \frac{5}{7}e^{-10l}$. In general,

$$p_{ij}(l) = \sum_{k=i}^j c_{ij}^{(k)} \exp\left(-\binom{k}{2}l\right),$$

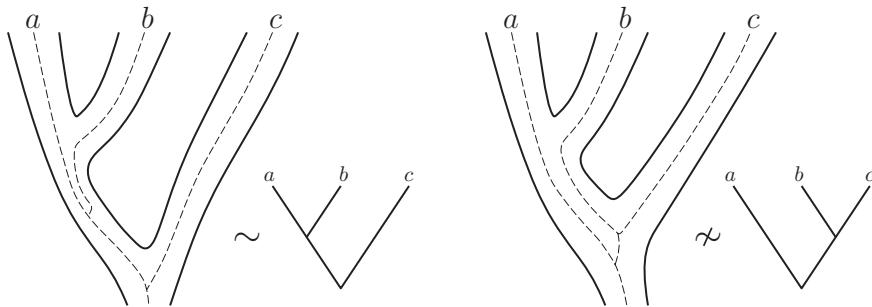


Figure 9.7. In the tree on the left, the gene tree $T_g = ab|c$ matches the species tree $T_s = ab|c$; on the right, the trees are different. A long interior edge (due to a small population size and/or a long period of time) makes a match more likely (in the left-hand tree) than what would apply for the shorter interior edge (i.e., large population size and/or shorter period of time, as shown on the right). In all cases, a match between the species tree and gene tree is more likely than the gene tree being any one particular alternative tree (however, a match may still have probability less than $\frac{1}{2}$ for a sufficiently short interior edge).

for rational coefficients $c_{ij}^{(k)}$ that have an exact expression, derived by Simon Tavaré in 1984 [353].

To apply the multispecies coalescent to ILS, let us first consider the case of three species, with $T_s = ab|c$. If l is the length of the interior edge of T_s (in coalescent time units), then

$$\mathbb{P}(\mathcal{T}_g = T_s | T_s, l) = \frac{1}{3}e^{-l} + (1 - e^{-l}) = 1 - \frac{2}{3}e^{-l}. \quad (9.24)$$

For either of the two trees $T \neq T_s$,

$$\mathbb{P}(\mathcal{T}_g = T_s | T_s, l) = \frac{1}{3}e^{-l}. \quad (9.25)$$

Equations (9.24) and (9.25) show that this model predicts that the species tree is always the most probable gene phylogeny. This means that for a large collection of rooted gene trees on the same three species, we can infer the correct species tree by simply taking a vote: estimate the species tree to be the tree that is supported by the largest number of gene trees. Assuming that each gene tree is independently generated by the multispecies coalescent process, the probability that the tree with the largest number of votes coincides with the species tree converges to 1 as the number of genes grows.

Equations (9.24) and (9.25) also reveal that (i) the two alternative trees have equal probability and (ii) the difference Δ between the most-probable tree (i.e., T_s) and any alternative gives an estimate of l , namely

$$l = -\ln(1 - \Delta).$$

Given knowledge about one of the two factors that determine l (population size and time-scale) the other can be inferred. One example concerns a large-scale phylogenetic study of primate species (human, chimpanzee, gorilla, orangutan, rhesus), in which human (H), chimpanzee (C) and gorilla (G) form a cluster. In [123], the authors investigated the relative support for the three rooted trees on these three species. It was found that around 77% of genes supported HC|G and 11–12% supported each of CG|H and HG|C. From these

proportions and the estimated temporal length between the gorilla and human-chimp divergence, an estimate of the ancestral population size can be made.

Unfortunately, when we move beyond three taxa, the species tree no longer need coincide with the most probable gene tree. For example, for a caterpillar species tree on four leaves, if we make all the interior edges of T really short, one can force the coalescences to occur earlier than the root. In that case, we are in the setting of the standard Kingman coalescent, which as we observed in Chapter 3 leads to the same distribution on trees as the YH distribution. In particular, each fork-shaped tree on four leaves has probability $\frac{1}{9}$ whereas any caterpillar-shaped tree (including T) has probability $\frac{1}{18}$. Thus, the most probable gene tree will be different from the species tree.

In a landmark paper [102], James Degnan and Noah Rosenberg established a more remarkable result: for *any* species tree with $n \geq 5$, it is possible to assign edge lengths so that a most probable gene tree will be different from the species tree. This leads to the following notion.

Definition. A gene tree T_g is said to be an *anomalous gene tree* (AGT) for a species tree T_s (where $T_g \neq T_s$) if, for some choice of edge lengths l for T_s , the following inequality holds: $\mathbb{P}(\mathcal{T}_g = T_g | T_s, l) > \mathbb{P}(\mathcal{T}_g = T_s | T_s, l)$.

The existence of AGTs means that a simple majority rule voting strategy applied to a sequence $T_{g(1)}, \dots, T_{g(m)}$ of m gene trees, independently generated under the multispecies coalescent process, no longer suffices to identify the species tree in a statistically consistent way. However, statistically consistent methods do exist. Perhaps the simplest (but maybe not the most effective) is simply to estimate $T_s|Y$ for each three-element subset Y of $[n]$ by using the voting strategy for the gene trees $T_{g(i)}|Y$. We have just seen that this approach is statistically consistent for triples. Since every rooted binary tree is determined by its rooted triples (cf. Chapter 4), we can reconstruct T_s uniquely from the rooted triple trees $T_s|Y : Y \in \binom{[n]}{3}$. This voting approach has been discussed by various authors (e.g., [100]).

AGTs tend to be more balanced than the species tree. A longstanding conjecture was that the most unbalanced phylogeny, namely a caterpillar tree, cannot be an AGT. This was established only recently using a novel approach in [101]. Various sophisticated approaches have been devised over the last decade for reconstructing species trees from gene trees (see [96] for further references).

Wicked forests. Degnan and Rosenberg carried the concept of AGTs one step further in their 2006 paper by introducing the notion of a *wicked forest*. This is a set W of two or more distinct trees from $RB(n)$ with the property that for all ordered pairs $T, T' \in W$, with $T \neq T'$, the tree T , regarded as a gene tree, is an AGT for T' , regarded as a species tree for suitably chosen edge lengths l . Wicked forests exist for $n \geq 5$, and an example of a wicked forest with three trees on eight leaves is given in [101]. Since a caterpillar tree cannot be an AGT the authors of [101] were able to justify the delightful title of their paper: *There are no caterpillars in a wicked forest*. The maximum number of trees that can form a wicked forest for $n > 5$ leaves is not currently known.

In Chapter 3 we saw that a rooted binary tree can have many possible rankings (of its interior vertices), with more balanced trees having a larger number of rankings than unbalanced trees. For example, a caterpillar has only one possible ranking, while any other binary tree has more. Indeed it was precisely for this reason that an AGT can exist for a caterpillar species tree on $n = 4$ leaves. Thus, it might be suspected that the phenomenon of AGTs might be due to simply to the fact that some gene trees have more rankings than

others. To check this, we can consider the probability of generating a particular ranked gene tree under the multispecies coalescent process on a particular species tree. When $n = 4$ the most probable ranked gene tree topology always matches the species tree topology. However, when $n = 5$ the most probable ranked gene tree can still differ from the species tree [103]. Since the “anomaly” carries over to the ranked setting as well, the existence of AGTs discussed earlier is not solely due to the differences in a gene tree having more rankings than the species tree.

We have assumed so far that it is possible to determine the gene tree up to the location of the root. However, as was noted in previous chapters, in the absence of additional assumptions (e.g., a molecular clock or the presence of an “outgroup” species), rooting a tree is problematic. A remarkable result from [10] shows that when $n \geq 5$ the probability distribution on unrooted gene trees suffices to uniquely recover the unrooted species tree, as well as the location of the root, and the lengths of the interior edges (when $n = 4$, the unrooted species tree can be uniquely recovered, but not the root location).⁶⁷

We can also ask whether the species tree is determined by the probability distribution that the clusters of the gene trees confer on subsets of $[n]$. More precisely, for a subset A of $[n]$, let $\mathbb{P}(A|T_s, l)$ denote the probability that A is a cluster of the tree \mathcal{T}_g generated under the multispecies coalescent on the species tree T_s with edge lengths l . In other words,

$$\mathbb{P}(A|T_s, l) = \sum_{T_g: A \in \mathcal{C}(T_g)} \mathbb{P}(\mathcal{T}_g = T_g | T_s, l),$$

where $\mathcal{C}(T_g)$ is the set of clusters of T_g . It is an easy exercise, using eqn. (9.24), to show that if $\mathbb{P}(A|T_s, l) > \frac{1}{3}$, then A must be a cluster of T_s [11]. However, T_s will generally contain other clusters. Using techniques based on linear invariants of clade probabilities, it can be shown that species trees with generic edge lengths can be identified just from the probability distribution on clusters induced by the gene trees [11].

Tree reconstruction when ILS conspires with the stochasticity of sequence evolution. Suppose that the (random) phylogeny for m genes is described by the multispecies coalescent process on a species tree T_s and the aligned DNA sequence sites for each gene g evolve according to a Markov process on T_g , as described in Chapters 7 and 8. A natural question is how we can reconstruct the species tree T_s from the sequence data accurately as m becomes large.

It turns out that the probability distribution on characters determines T_s up to the location of the root, at least generically, for models used in molecular systematics (e.g., the GTR model, allowing some rate variation across sites). A recent paper [87] established a number of general results in this direction, by applying algebraic techniques (such as the edge invariants arising from flattenings) discussed in Chapter 8. Although the reconstruction techniques described in [87] are of an algebraic nature, biologists have also tried more simple reconstruction techniques, such as concatenating all the sequences and treating them as one “supergene,” then applying a statistically consistent sequence-based tree reconstruction method like ML described in Chapter 8, to this long sequence of characters.

However, there is no reason to believe that such a concatenation approach should be statistically consistent (i.e., converge on the true species tree as the number of genes becomes large), since we are ignoring the fact that the different genes are described by (possibly) different gene trees under ILS. In other words, applying ML under an assumption

⁶⁷The biological relevance of this result is that the probability of each gene tree can be estimated from its empirical frequency in a large sample of gene trees.

that all sites are described by a Markov process on the same tree is tantamount to model misspecification. This, in itself, does not mean that ML will converge on an incorrect tree but simulations have suggested that the inconsistency of ML on concatenated sequences was possible [222]. This inconsistency has been subsequently established mathematically, and holds even when the Markov process on the tree is the fully symmetric model (e.g., JC69) and all genes are of the same fixed length [301].⁶⁸

Although ML on concatenated sequences is statistically inconsistent, other methods are not. Indeed, a surprisingly simple way to consistently estimate a species tree from concatenated sequences was found in [96]. For convenience, let us assume that all the m genes have the same length k and that the substitution model is JC69. Let \hat{p} be the normalized Hamming distance on the $[n]$ across the mk characters ($f_{ij} : i \in [m], j \in [k]$) in the concatenated data set. In other words, for each $x, y \in [n]$, let

$$\hat{p}(x, y) = \frac{\#\{(i, j) \in [m] \times [k] : f_{ij}(x) \neq f_{ij}(y)\}}{mk},$$

and let $\hat{\gamma}(x, y) = -\frac{3}{4} \ln(1 - \frac{4}{3} \hat{p}(x, y))$. Notice that if the gene trees are generated i.i.d. under the multispecies coalescent model, then $\hat{\gamma}(x, y)$ converges in probability as $m \rightarrow \infty$ to

$$\gamma(x, y) = -\frac{3}{4} \ln\left(1 - \frac{4}{3} \mathbb{E}[\hat{p}(x, y)]\right). \quad (9.26)$$

Equation (9.26) looks suspiciously like the distance correction formula (eqn. (7.17) for the JC69 model from Chapter 7), which described the evolutionary distance under JC69, as $\mu(x, y) = -\frac{3}{4} \ln(1 - \frac{4}{3} p(x, y))$. In that case, $p(x, y)$ is also the expected value of $\hat{p}(x, y)$ but on the JC69 model directly, without the complication of ILS. The remarkable result from [96] is that the distance function γ defined by eqn. (9.26) also satisfies the four-point condition and thus has a tree representation, with the tree that provides this representation being precisely the species tree T_s .

What is surprising here is that the correction seems to ignore the fact that the gene trees are random samples (under the multispecies coalescent) and instead acts as though the characters are sampled from a single gene tree, estimated from the concatenation of gene trees, under the JC69 model. However, in contrast to what happens with ML, this apparent model misspecification still leads to a consistent estimator of the species tree T_s , up to the location of the root (using NJ, for instance, on $\hat{\gamma}$).

Using this approach, it is even possible to set bounds on the number m of genes needed to infer the (unrooted) species tree $T_s \in RB(n)$ correctly with a probability of at least $1 - \epsilon$. If f is the smallest edge length in T_s and ϵ is fixed then

$$m = \Theta\left(\frac{\log n}{f^2}\right)$$

genes (of any length $k \geq 1$) are both necessary and sufficient, where the constant in the Θ term depends on the other parameters associated with the ILS model and the edge lengths of T_s [96]. One notable feature of this result is that the dependence of m on f as $f \rightarrow 0$ is the same inverse quadratic form that we saw with sequence length k when we were just dealing with the Markov processes on trees in the last chapter (eqn. (8.7)), without the further compounding random process of ILS. Other methods have also been developed

⁶⁸The proof exploits a link between ML and MP, and also involves an application of the Ewens sampling formula.

for inferring a species tree from sequences, such as the ASTRAL algorithm [258], which is based on solving an optimization problem involving quartet trees and is thus related to concepts from Chapter 4.

Finally, although we have treated the setting where a gene g is sampled from exactly one individual in each species, in practice, g may be sampled from several individuals in one or more species. In this case, the theory described above extends, and this enlarged “gene tree” can provide more information about the species tree, the divergence times, and the ancestral population sizes.

9.3.2 ■ ILS and deep coalescence cost

The species tree provides one firm combinatorial constraint on the way a gene tree can embed in the species tree. Namely, two gene lineages from different species cannot coalesce in the gene tree before the species do in the species tree. Subject to this combinatorial constraint, “deep coalescence cost” provides a discrete measure of how well a gene tree can fit within a proposed species tree. Roughly speaking, it counts the extent to which gene lineages fail to coalesce on edges they share within the species tree. For example, we could embed any gene tree in any species tree by forcing all the coalescence events to occur earlier than the root, but this would incur a high penalty.

We now describe these notions more formally, mostly following [354]. To keep life simple, and we will write the binary phylogenetic X -trees T_g as T and T_s as S in this section. Given $T = (V_T, E_T)$ and $S = (V_S, E_S)$, consider the function $\lambda = \lambda_{(T,S)}$ defined as follows:

$$\lambda : V_T \rightarrow V_S,$$

$$v \mapsto \text{lca}_S(c_T(v)).$$

This function λ can be referred to as the *LCA mapping* or “most recent common ancestor” (MRCA) mapping from the vertices of T to the vertices of S . In words, λ maps each vertex v of the gene tree T to the vertex of the species tree S that is the least common ancestor in T of the set of species that are descendants of v in T . Informally, λ provides a way to “reconcile” the evolution of T within S under the assumption that (tracing backwards in time) lineages in T coalesce at least as far into the past as the corresponding lineages in S joined, and λ makes these lineages coalesce as close to the species’ separations as possible.

Notice that $\lambda(x) = x$ for all $x \in X$ (i.e., it sends each leaf of T to the corresponding leaf of S) and λ also maps the root of T to the root of S .

To define the deep coalescence cost, we first need to describe a positive integer quantity for each edge e of S , which also depends on T . This quantity is called the number of *extra lineages* in e , and is denoted $\text{xl}(T, e)$.

There are several different, but equivalent, ways to define $\text{xl}(T, e)$. Perhaps the most natural is the following: $\text{xl}(T, e) = k$ if for precisely $k + 1$ of the edges (r, s) of T , the edge e of S lies on the path in S from $\lambda(r)$ to $\lambda(s)$.

For example, in Fig. 9.8, let e_1 denotes the upper interior edge of S and e_2 the lower one. The $\text{xl}(T, e_1) = 1$ since there are two edges of T for which e_1 lies on the path from the λ images of their endpoints (these two edges are the pendant edges of T incident with a and c). Similarly, $\text{xl}(T, e_2) = 1$ since e_2 lies on the λ image of the endpoints of two edges of T (the right-most interior edge of T and of the pendant edge of T incident with a).

A second (equivalent) definition for $e = (u, v)$ is

$$\text{xl}(T, e) = |c_S(e)| - I(e) - 1, \quad (9.27)$$

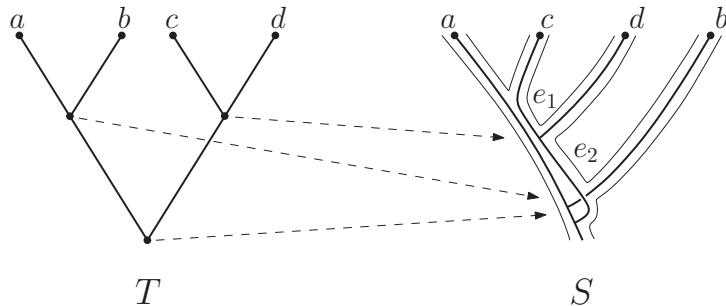


Figure 9.8. The LCA mapping from gene tree \$T\$ to species tree \$S\$. The two interior edges of \$S\$ have one extra gene lineage, and so \$\text{dc}(T, S) = 1\$ for this example.

where \$c_S(e)\$ is the set of leaves of \$S\$ descended from edge \$e\$ (equal to the cluster \$c_S(v)\$ for \$e = (u, v)\$), and where \$I(e) = |\{w \in \overset{\circ}{V}(T) : \lambda(w) \in c_S(e)\}|\$ is the number of interior vertices of \$T\$ which get mapped by \$\lambda\$ to a vertex in the subtree descended from \$e\$.

To illustrate this again for Fig. 9.8, with \$e_1, e_2\$ as before, we have \$c_S(e_1) = \{a, c\}, c_S(e_2) = \{a, c, d\}\$, and \$I(e_1) = 0, I(e_2) = 1\$, so \$\text{xl}(e_1) = 2 - 0 - 1 = 1\$ and \$\text{xl}(e_2) = 3 - 1 - 1 = 1\$.⁶⁹

The *deep coalescence cost* for reconciling tree \$T\$ within tree \$S\$, denoted \$\text{dc}(T, S)\$, is defined by

$$\text{dc}(T, S) = \sum_{e \in E(S)} \text{xl}(T, e). \quad (9.28)$$

Informally, \$\text{dc}(T, S)\$ measures the extent to which lineages need to share edges of \$S\$ in order to “fit” \$T\$ inside \$S\$. It is easily checked that \$\text{dc}(T, S) = 0\$ if and only if \$T = S\$. But as the next example shows, \$\text{dc}\$ is not symmetric.

Exercise: For \$T = (((a, b), c), d)\$ and \$S = (((a, c), d), b)\$, show that \$\text{dc}(T, S) = 3\$ and that \$\text{dc}(T, S) = 2\$.

Despite this asymmetry, if \$\text{dc}(T, S) = 1\$ then \$\text{dc}(S, T) = 1\$. Moreover, if \$T\$ and \$S\$ are two trees from \$RB(X)\$, then \$\text{dc}(T, S) = 1\$ if and only if the trees are a single rooted NNI operation apart [354].

At the other extreme, how large can \$\text{dc}(T, S)\$ be? Equations (9.27) and (9.28) show that

$$\text{dc}(T, S) \leq -(2n - 2) + \sum_{e \in E(S)} |c_S(e)|. \quad (9.29)$$

It can be shown that the summation term in (9.29) is maximized for the rooted caterpillar tree for which \$\sum_{e \in E(S)} |c_S(e)| = \frac{n(n+1)}{2} - 1\$. This leads to the inequality \$\text{dc}(T, S) \leq \binom{n}{2}\$ from [354] (Theorem 6).

Suppose that we are now given a profile \$\mathcal{P}\$ of rooted binary \$X\$-trees so that \$\mathcal{P} = (T_1, T_2, \dots, T_k)\$, where \$T_i \in RB(X)\$ for all \$i\$. A natural optimization problem asks to find a species tree \$S \in RB(X)\$ that minimizes the total deep coalescence cost (across all of the trees in the profile). In other words, we would like a species tree \$S\$ that minimizes \$\sum_{i=1}^k \text{dc}(T_i, S)\$. This turns out to be an NP-hard problem [384]; however, its solution enjoys one elegant property established by [230]. Namely, it is “Pareto on clusters,” which means that if \$A\$ is a cluster of each of the trees in the profile \$\mathcal{P}\$ then any tree \$S \in RB(X)\$ that

⁶⁹There is a third way to compute \$\text{xl}(T, e)\$ that does not explicitly use the LCA mapping; for details, see [354].

minimizes the total deep coalescence cost must also contain A as a cluster (in particular, any such tree must equal to or be a refinement of the strict consensus tree of \mathcal{P} , from Chapter 2).

There is also an interesting algebraic connection linking $\text{dc}(T, S)$ with the number of gene duplications and the number of gene losses required to explain the gene tree T by evolution on the species tree S under these two processes; for details, see [384].

We end this section by considering what happens when either the gene tree T or the species tree S is a fixed tree from $RB(X)$, and the other tree is randomly selected from $RB(X)$ by some model θ (such as the YH or uniform distribution from Chapter 3). In this case, the distribution of the random variable $\text{dc}(T, S)$, or at least its expected value, is of interest. In a recent paper [355], a number of exact formulae for $\mathbb{E}_\theta[\text{dc}(T, S)]$ are described. A general feature of these formulae is that for either a fixed choice of the gene tree or of the species tree (T or S), if the other tree is YH or uniform, the expected deep coalescence score and its maximum tend to increase with the imbalance of the fixed tree, and are maximized by rooted caterpillars. However, the variation in this expected value tends to be greater across fixed species trees than across fixed gene trees. For further details, see [355].

9.3.3 ■ Gene trees generated by LGT and related processes

There are other mechanisms by which gene trees can differ from species trees. One is *hybridization* where genes are transferred from one species into another during the formation of a new “hybrid” species. For this type of hybrid evolution, the “species tree” description is often replaced by a phylogenetic network (a topic we will explore further in the next chapter). Stochastic aspects of hybridization in concert with incomplete lineage sorting have been explored by a number of authors (see, for example, [255, 381]). Rather than describing these here, we turn to a second and somewhat different way that genes can be transferred between species, which is more relevant for haploid taxa such as prokaryotes (bacteria and archaea).⁷⁰ This process is called *lateral gene transfer (LGT)* (also called “horizontal gene transfer”). In LGT, a copy of a gene g from one prokaryotic species x is directly inserted into a different prokaryotic species y . If this copy of g replaces the version of gene g originally carried by y , then the topology of the resulting phylogenetic tree based on gene tree g may differ from a species tree for the species. This is illustrated in Fig. 9.9.

Evidence for extensive levels of LGT in bacteria has been provided by several studies. One novel approach is simply to consider the distribution of genes across species [94]. These authors considered a simple model in which each gene only arises once on a tree, but can be lost multiple times in its evolution. They argued that without significant LGTs between the edges of the tree, the observed present-day distribution of ancestral genomes would require ancestral genomes near the root of the tree to be unreasonably larger than those observed today.

To state this a little more formally, given any phylogenetic X -tree T , if gene g is present in (at least) two elements $x, x' \in X$, then the simple model described implies that gene g must also be present at every vertex in the (undirected) path in T that joins x and x' . Consequently, if \mathcal{G} is a collection of genes and $S : X \rightarrow 2^{\mathcal{G}}$ is the function that describes which subset of genes from \mathcal{G} is present in each taxon in X , then we can extend S to a function \tilde{S} from the entire vertex set V of T to $2^{\mathcal{G}}$ as follows. Let $\tilde{S}(v)$ be those genes $g \in \mathcal{G}$ that are present in $S(x)$ and $S(x')$ for at least one pair $x, x' \in X$ and for which v lies

⁷⁰The concept of how to define a species tree for prokaryotes is somewhat controversial, as discussed further in Chapter 10.

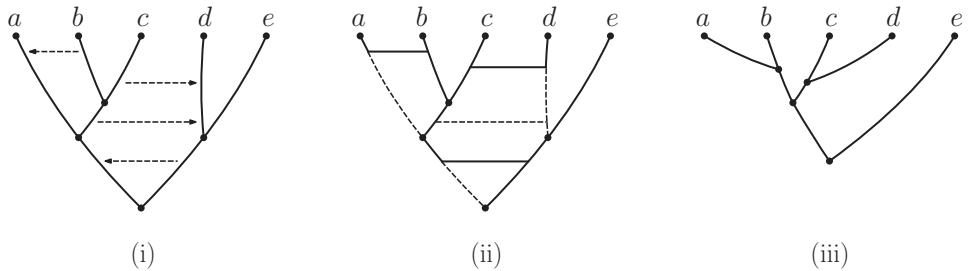


Figure 9.9. (i) Four LGT events on the tree T_s . When each event replaces the gene in a species with the transferred copy, then the resulting evolutionary history of the gene, shown in (ii), leads to the gene phylogeny T_g shown in (iii). The transition from (i) to (ii) can most easily be seen by tracing back the history of the leaves towards the root following against the direction of the dashed horizontal arrows.

on the undirected path in T connecting x and x' . Under the simple model described above (and without LGT), the subset of genes present at the ancestral taxon corresponding to vertex v in T would necessarily include $\tilde{S}(v)$. In particular, the genome size (number of genes) at vertex v would therefore need to be at least $|\tilde{S}(v)|$. A small example to illustrate this is shown in Fig. 9.10.

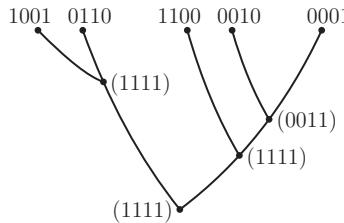


Figure 9.10. For the sequences of four genes, each of which is present (1) or absent (0) at the leaves of the tree shown, each leaf only has two genes. However, three of the ancestral vertices are forced to have four genes under a model in which a gene can only arise once but can be lost multiple times.

Notice that the function \tilde{S} can be computed readily (in polynomial time in $|X|, |\mathcal{G}|$) for each vertex v . A natural computational question asks how small the largest ancestral genome can be across all possible rooted phylogenies given just the distribution of genes across taxa in X . More precisely,

Given $S : X \rightarrow 2^{\mathcal{G}}$, determine the smallest integer k_S for which there is a rooted phylogeny $T \in RP(X)$ with $|\tilde{S}(v)| \leq k_S$ for all $v \in V(T)$.

This question turns out to be at least as hard as computing the treewidth of a graph (an NP-hard problem) even in the special case where each species in X has just two genes (as in Fig. 9.10). To see this, suppose that $G = (V, E)$ is any graph. Take $\mathcal{G} = V$ and $X = E$, and take $S : X \rightarrow 2^{\mathcal{G}}$ to simply be the identity function (i.e., the edge $\{u, v\}$ maps to the set $\{u, v\}$). It can then be shown that

$$\text{tw}(G) = k_S - 1.$$

For further details, see [369].

Modeling random LGT. To help predict the impact of LGT on gene trees, and to develop a method for inferring species trees from gene trees that are subject to LGT, it helps to model LGT by a stochastic process (in the same spirit as the multispecies coalescent was used to model ILS).

To describe this, it helps to think of a species tree $T \in RB(n)$ with ultrametric edge lengths l as a tree in which each edge e is an interval of length $l(e)$, and the total height of the tree is τ . Let us place a time scale on the tree so that each “point” p (vertex of T or point in an edge interval) has an associated “time stamp” $t(p) \in \mathbb{R}^{\geq 0}$ that corresponds to its path distance from the root ρ . Thus, $t(\rho) = 0$ and $t(x) = \tau$ for each leaf x of T . For each $t' \geq 0$, the number of points p in T with $t(p) = t'$ is finite, and this count is monotone increasing from 1 at $t = 0$ to n at τ . The simplest model of LGT assumes that transfers occur independently and that at each point p in the tree, there is a constant rate λ of a transfer event from p ; when this occurs, one of the other points p' that is contemporaneous with p (i.e., with $t(p') = t(p)$) is selected uniformly at random and the gene at p replaces the gene at p' . Thus if there are k lineages extant at time t , the rate of transfer at time t is $k\lambda$. The total number of transfers (for a given gene) in the tree is then a sum of independent Poisson random variables and so has a Poisson distribution with mean λL , where L is the sum of the edge lengths of the tree. We call this model that *uniform Poisson model for LGT*. More elaborate models may allow a transfer event to move several genes at once or to allow higher rates of transfer between points of the tree that are close together, but such models are a little harder to analyze.

Under such a model, some fundamental questions include the following. How much transfer is possible while still allowing for the correct reconstruction of the species tree topology from the induced gene tree topologies? How many gene trees are required for this?

As with ILS, it helps to start with a rooted phylogeny with just three leaves. Suppose that the species tree T_s is the rooted triple $ab|c$. Again, we will let \mathcal{T}_g and T_g denote a random gene and fixed gene tree, but this time, within the LGT model described.

Proposition 9.4. *Under the uniform Poisson LGT model on a rooted triple $ab|c$ of height τ , let $m = 3\lambda(\tau - t(\text{lca}_T(a, b)))$. Then*

$$\mathbb{P}(\mathcal{T}_g = T_s) = e^{-m} + \frac{1}{3}(1 - e^{-m}).$$

If T' is one of the other two phylogenies ($ac|b$ or $bc|a$), then

$$\mathbb{P}(\mathcal{T}_g = T') = \frac{1}{3}(1 - e^{-m}).$$

Proof: Let N denote the number of LGT events between $t(\text{lca}_T(a, b))$, and $\tau = t(a) = t(b) = t(c)$ (i.e., during the period when there are three pendant edges). N has a Poisson distribution, with expected value equal to $m = 3\lambda(\tau - t(\text{lca}_T(a, b)))$. Conditional on $N = 0$, the gene tree matches the species tree with probability 1. On the other hand, if $N \geq 1$, then for rooted triple phylogenies, the discrete gene tree is fully determined by the transfer event that occurs closest to the present; in other words, the transfer for which the source p has the largest t value. Moreover, the source of this p is equally probable to lie on any one of the three pendant edges of T , and whichever edge is chosen, the target edge is equally likely to be one of the other two pendant edges.

Therefore, $\mathbb{P}(\mathcal{T}_g = T_s | N > 0) = \frac{1}{3}$. By the law of total probability, we obtain the expression in $\mathbb{P}(\mathcal{T}_g = T_s) = e^{-m} + \frac{1}{3}(1 - e^{-m})$. The calculation of the probability for T' is similar. ■

Exercise: Suppose that we modify the uniform Poisson model for LGT so that the rate of transfer from p to p' with $t(p) = t(p')$ is not constant but increases as the path distance in T between p and p' decreases. For a rooted tree on three species, does the most probable gene tree still coincide with the species tree?

So, when $n = 3$, the most probable gene tree is the one that matches the species tree, with the other two rival gene trees having equal but lower probabilities. This is just like what happens with ILS. For $n > 3$, the similarity ends. For ILS, when $n > 3$, it is still the case that for any subset Y of three species in T_s , the most probable tree for $T_g|Y$ is still $T_s|Y$. However, this property can fail for LGT under the uniform Poisson model. Indeed, for a fixed set $Y \subset [n]$ of size 3, and a fork-shaped species tree T_s on four leaves (with ultrametric edge lengths chosen appropriately), $T_s|Y$ can be the *least* probable phylogeny for $T_g|Y$ [335]. Basically, the reason why this can occur is that additional lineage(s) can play a role because the gene can be transferred to and from the additional edge(s).

Nevertheless, the discrete species tree T_s is still an identifiable parameter from sufficiently many gene trees generated i.i.d. under the uniform Poisson model for LGT, at least provided that the rate of LGT is not too high. Indeed, a species tree generated by a Yule pure-birth model can be reconstructed from $N = \Theta(\log n)$ gene tree topologies on n species, if the expected number of transfers per gene (across the entire tree) is $O\left(\frac{n}{\log n}\right)$ [300]. The proof involves analyzing reconstruction accuracy of quartet trees for the genes. The authors of [300] also use a coupling argument to show that if the expected number of transfers per genes goes from being slightly sublinear to slightly superlinear (specifically, $\Theta(n \log \log n)$), then the species tree topology can no longer be reconstructed from $N = \Theta(\log n)$ gene tree topologies.⁷¹

Although the uniform Poisson model is overly simple, it is perhaps surprising how high the rates of transfer can be under this model and still allow for a consistent estimate of the species tree. For example, for a typical Yule pure-birth tree on $n = 200$ leaves in which the expected number of transfers per gene can be as high as 10, the species tree can be consistently inferred from sufficiently many gene trees [335]. Of course, these analyses assume that the gene tree has been correctly inferred, so (as has already been done with ILS) it would be of interest to consider the impact of the randomness in the gene tree that comes from their reconstruction from sequence data.

⁷¹This gap has been reduced further in [98] by showing that the species tree can be recovered even when the expected number of transfers grows linearly with n .

Chapter 10

Introduction to phylogenetic networks

So far, we have considered evolution to be a process that occurs on a tree (Chapters 5, 7, and 8) or that generates a tree (Chapters 3 and 9). In this final chapter, we describe a framework that allows for more complex representations of evolutionary relationships. Phylogenetic networks are graphs that display phylogenetic signals that may not fit on a single tree. In particular, directed networks can provide an explicit representation of evolution where there has been reticulation (such as the formation of hybrid species, or lateral gene transfer etc). The mathematical and computational study of networks is perhaps the most active area of modern phylogenetics. A detailed coverage of this topic can be found in two books [203, 168]; here, we describe a selection of the main concepts and results, as well as highlighting a number of more recent developments.

10.1 ▪ To tree or not to tree: Why networks?

Phylogenetic networks, as explicit representations of reticulate evolution—in particular, the formation of hybrid species—have an important role in biology. These networks can often be viewed as a tree where one (or more) branches come together to form a new lineage; in other words, as a slightly modified tree. However, in more recent times, some prominent biologists (e.g., Ford Doolittle) have argued that for prokaryotes (which include bacteria), the notion of a “species tree” is meaningless in the light of processes such as LGT (lateral gene transfer), gene gain and loss, and endosymbiosis events (where one cell becomes incorporated as a part of a new cell). However, other experts (e.g., Eugene Koonin) have argued that despite LGT, there is still a well-defined notion of a (statistical) central tendency tree for prokaryotes [221].

Part of the problem hinges on the thorny notion of what constitutes a “species,” particularly in the microbial world. Comparing the genes shared by all bacteria is difficult, since there are so few—any tree produced from these data is sometimes dismissively referred to as a “tree of one percent.” However, as we saw in Section 9.3.3, high rates of LGT still allow a tree signal to be recovered. Moreover, while the widespread disagreement of phylogenies inferred from different genes might be explained by LGT, some of these disagreements may simply be due to systematic errors in inferring the gene trees (e.g., by using an inappropriate model of character evolution) [273].

Although the extent and phylogenetic significance of LGT in prokaryotic evolution is still debated, it has clearly played a significant role, and phylogenetic networks provide a way of representing this. Moreover, in the domain of life that most nonbiologists are

more familiar with—the eukaryotes—hybridization (where two ancestral species give rise to a new lineage) has also played a significant role in the evolution of many plant and some animal species.

The last decade has seen a great deal of progress by computer scientists and mathematicians in studying the properties, construction and analysis of phylogenetic networks. In this chapter, we will begin with some definitions and then describe two types of phylogenetic networks. The first is the class of *implicit networks*, which are typically unrooted, and provide a representation of how “tree-like” (or not) the data are, and of the degree of support for conflicting splits. The second and main part of this chapter concentrates on *explicit networks*, which are rooted networks that aim to directly display reticulate evolution.

10.1.1 • Preliminaries

Definitions. Broadly speaking, a phylogenetic network N on a set X is a connected graph (either directed or undirected) for which X is a distinguished set of vertices. To make this more precise, we need to consider the undirected and directed cases separately.

- A (*undirected*) *phylogenetic network* on X is a connected graph $N = (V, E)$ for which X is the set of leaves (vertices of degree 1) and where there are no vertices of degree 2.
- A (*directed*) *phylogenetic network* on X is a connected acyclic digraph (directed graph) $N = (V, A)$, where A is the set of *arcs* (directed edges) for which (i) N has exactly one vertex of in-degree 0, the *root vertex* ρ , (ii) X is the set of vertices of out-degree 0 (called *leaves*), and (iii) there are no vertices of in-degree and out-degree both equal to 1.

Notice that, by definition, a directed phylogenetic network has no parallel arcs (since A is a set rather than a multiset). Notice also that the class of undirected (respectively, directed) phylogenetic networks on X includes the set $P(X)$ (respectively, $RP(X)$). Moreover, a phylogenetic network N is a phylogenetic tree if and only if N is a tree (in the rooted case, the root vertex is required to have out-degree at least 2).

In this chapter, we will adopt the convention of orienting directed phylogenetic networks downwards, so that the leaves X appear at the bottom of the network.

Two phylogenetic networks N and N' (both directed or both undirected) are considered to be *equivalent* if there is a (di)graph isomorphism from N to N' that maps each element of X in N to the same element in N' . We denote this by writing $N \cong N'$. This agrees with and extends the corresponding notion of equivalence when we restrict ourselves to (rooted and unrooted) phylogenetic X -trees.

10.2 • Implicit (unrooted) networks

10.2.1 • Binary unicyclic networks and undirected binary level-1 networks.

We begin by introducing the simplest class of unrooted phylogenetic networks which are binary and have just a single cycle. Formally, a *unicyclic network* on X is a graph G that has exactly one cycle (of length at least three), where the set of degree-1 vertices is X and all other vertices have degree 3. An example is illustrated in Fig. 10.1(i).

Notice that if we delete any single edge on the cycle in G and suppress the two resulting degree-2 vertices, then we obtain a binary phylogenetic X -tree. We say that G *displays* a

tree $T \in B(X)$ if T can be obtained in this way, and that G displays a profile \mathcal{P} of trees from $B(X)$ if G displays each tree in the profile.

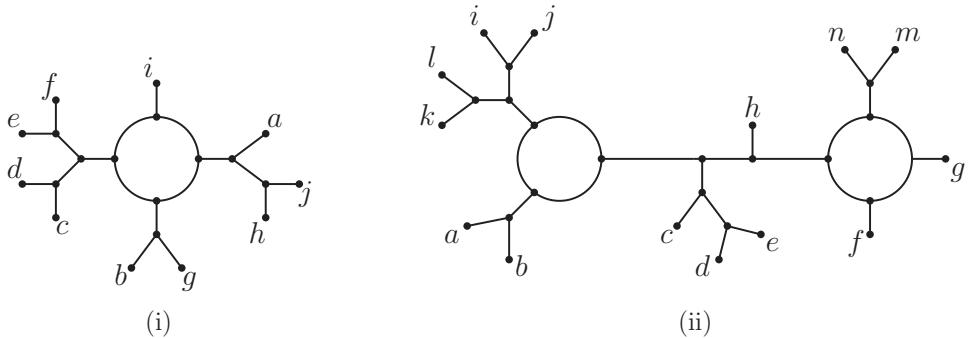


Figure 10.1. (i) A unicyclic network; (ii) an undirected binary level-1 network with two cycles.

This class of networks was studied in [317], and we summarize two results concerning this display property.

- For any two distinct trees T and $T' \in B(X)$, there is a unicyclic network that displays these trees if and only if $d_{TBR}(T, T') = 1$; moreover, in this case, $\Sigma(T) \cup \Sigma(T')$ is a circular split system.
- For any set \mathcal{P} of three or more trees from $B(X)$, there is a unicyclic network that displays each tree in \mathcal{P} if and only if, for every subset \mathcal{P}' of \mathcal{P} of size 3, there is a unicyclic network that displays each tree in \mathcal{P}' . In this case, there is a unique network on X that displays \mathcal{P} .

Let us now count the number $c(n, k)$ of unicyclic networks on $[n]$ whose unique cycle is of length k . This number is closely related to a quantity we considered in Section 5.2.3, namely the number $N(n, k)$ of rooted forests consisting of k rooted binary phylogenetic trees (allowing a singleton leaf to be a tree) on leaf sets that partition $[n]$. An exact expression for $N(n, k)$ was described in eqn. (5.11). Any unicyclic network for which the cycle has length k induces a forest in $N(n, k)$ by deleting all the edges in the cycle and their incident edges. Moreover, for any such forest F consisting of k trees, there are exactly $\frac{1}{2}(k-1)!$ unicyclic networks that induce F , since there are $(k-1)!$ cyclic permutations and each such permutation specifies an order to attach the trees around a cycle; the factor $\frac{1}{2}$ arises because reversing any permutation leads to the same network G . Consequently, $c(n, k) = \frac{1}{2}(k-1)!N(n, k)$ and so, by eqn. (5.11), we obtain for all $n, k \geq 3$

$$c(n, k) = \frac{(2n-k-1)!}{(n-k)!2^{n-k+1}}.$$

There is also an exact expression for the total number $c(n)$ of unicyclic networks on a set X of size n [317], as follows.

Proposition 10.1. For $n \geq 3$, $c(n) = (n-1)2^{n-2} - \frac{(2n-2)!}{(n-1)!2^{n-1}}$.

Proof. Let $\varphi(x) = \sum_{n \geq 1} r b(n) \frac{x^n}{n!}$, the exponential generating function for $r b(n)$. We saw in Section 2.1 that

$$\varphi(x) = 1 - \sqrt{1-2x}. \quad (10.1)$$

Let $C_k(x) = \sum_{n \geq 3} c(n, k) \frac{x^n}{n!}$, and $C(x) = \sum_{n \geq 3} \frac{x^n}{n!}$. By considering how many ways there are to form a forest of k rooted binary trees on the leaf set $[n]$ and to arrange these trees around a cycle of length k , we obtain

$$2C_k(x) = \frac{1}{k} \varphi(x)^k.$$

Consequently, $C(x) = \sum_{k \geq 3} C_k(x) = \frac{1}{3}\varphi(x)^3 + \frac{1}{4}\varphi^4 + \dots$. We now use the identity

$$-\ln(1-t) = t + \frac{1}{2}t^2 + \frac{1}{3}t^3 + \dots$$

to write

$$2C(x) = -\ln(1-\varphi(x)) - \left[\varphi(x) + \frac{1}{2}\varphi(x)^2 \right].$$

This equation seems daunting but eqn. (10.1) provides a reprieve, because $-\ln(1-\varphi(x)) = \frac{1}{2}\ln(1-2x)$. Proposition 10.1 now follows readily, since, for $n \geq 3$, $c(n)$ is just $n!$ times the coefficient of x^n in $-\varphi(x) - \frac{1}{4}\ln(1-2x)$. ■

Going beyond a single cycle, a *undirected binary level-1 network* on X is a graph for which each vertex lies in at most one cycle, every nonleaf vertex has degree 3 and the set of degree-1 vertices is X .⁷² An example is shown in Fig. 10.1(ii). If we let $g(n, k, m)$ be the number of undirected binary level-1 networks on $[n]$ containing k cycles and having a total of m edges across all cycles, then for $n, m, k \geq 0$ the following result from [317] holds:

$$g(n+2, k, m) = \frac{(2n-m+3k)!(m-2k-1)!2^{m-n-3k}}{(n-m+2k)!(m-3k)!(k-1)!k!},$$

if $3 \leq 3k \leq m \leq n+2k$ or $k=m=0$, and $g(n+2, k, m)=0$ otherwise.

Exercise: If N is an undirected binary level-1 network, show that its treewidth satisfies $\text{tw}(N) \leq 2$.

10.2.2 • Split networks and circular split systems

Chapter 2 described how a phylogenetic tree can be viewed equivalently as a set of pairwise compatible splits. When a set of splits is not pairwise compatible, we can instead represent it by a phylogenetic network. Following [203], we first need the following concept.

A *split graph* is a connected bipartite graph $G = (V, E)$ together with a surjective map $s : E \rightarrow K$ (where K is any finite set) that is *isometric* (i.e., for any two vertices (u and v , say), and for each shortest path p between u and v , s maps the edges on p one-to-one to a subset $S(p)$ of K , and with $S(p)$ constant for all such p). A fundamental property of any split graph is the following.

Proposition 10.2. *Let (G, s) be a split graph, where $G = (V, E)$ and $s : E \rightarrow K$. For any $k \in K$, let E_k be the set of edges e in G with $s(e) = k$. Then $(V, E - E_k)$ consists of precisely two connected components for each $k \in K$.*

The proof of this result exploits the properties that G is connected and bipartite, and that s is isometric (see [203] for details).

⁷²Such graphs are examples of “cactus graphs.”

With this notion in hand, we can now define a “split network.” Let Σ be a set of X -splits. Consider a split graph (G, s) where $G = (V, E)$ and $s : E \rightarrow \Sigma$. Suppose also we have a map $f : X \rightarrow V$ with the property that, for each split $\sigma = A|B \in \Sigma$, A and B are the two subsets of elements of X that get mapped by f to each of the two connected components of $(V, E - E_\sigma)$. In that case, we refer to (G, s, f) as a *split network* on X that *represents* Σ . The edges that are mapped to the same split σ by s are often called *parallel edges*, since they are typically drawn parallel. Informally, a split network that represents Σ is a graph that produces each split in Σ by cutting all the (parallel) edges of the graph that correspond to that split.

Examples:

(i) Any phylogenetic tree $T \in P(X)$ is a split network (T, s, f) where s assigns each edge of T to the corresponding X -split and f is simply the labeling of the leaves of T by elements of X . A set of splits Σ is pairwise compatible if and only if there is split network (G, s, f) that represents Σ and for which G is a tree.

(ii) We saw in Section 2.4.1 how every collection Σ of X -splits gives rise to a graph—the Buneman graph, $B(\Sigma)$ —which is connected and bipartite, and for which there is a labeling map $f : X \rightarrow V$. Moreover, each edge $e = \{v, v'\}$ of $B(\Sigma)$ corresponds to a split in Σ (namely the split on which the two vertices differ). It can then be checked that (B, s, f) is a split network that represents Σ . Fig. 10.2 shows an example of the Buneman graph $B(\Sigma_3)$ for a set of three splits, along with a simpler split network that also represents Σ_3 .

Although the Buneman graph is a mathematically natural way to represent any set of splits, it has some disadvantages. First, the graph can be very large (the number of vertices and edges can grow exponentially with $n = |X|$). Second, the graph can be particularly difficult to represent in the plane in a visually meaningful way. Both these problems can be circumvented by considering split graphs that are restricted to a special class of splits (circular split systems).

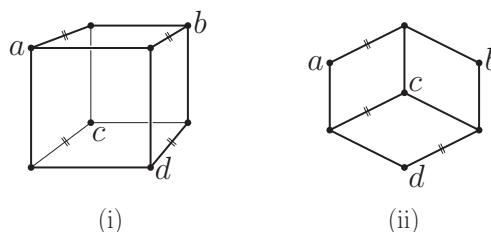


Figure 10.2. For the set of splits $\Sigma_3 = \{\{a, b\}|\{c, d\}, \{a, c\}|\{b, d\}, \{a, d\}|\{b, c\}\}$, the Buneman graph $B(\Sigma_3)$ is shown in (i). A simpler split network that also represents Σ_3 is shown in (ii). The marked parallel lines for each graph show how the split $ad|bc$ is displayed by cutting these edges.

Circular split systems. Recall from Section 2.4.2 that a circular split system on X is any set of splits that can be obtained from placing the elements of X in any order at the vertices of a regular n -gon and by considering the splits of X obtained by deleting a pair of edges of the n -gon. Note that a circular system of splits is never larger than $\binom{n}{2}$ where $n = |X|$.

A split network N on X is said to be *outer-labeled planar* if N can be drawn in the plane so that no two edges cross, edges corresponding to the same split are parallel, and the elements of X all lie on the outside of the graph. In that case, the set of splits that N represents is clearly circular. However, the converse can also be established using a result from geometry (see [203] Theorem 5.7.5, and [120]).

Proposition 10.3. *A set Σ of X -splits is circular if and only if it can be represented by a split network N that is outer-labeled planar. Moreover, N can be chosen to have $O(n^4)$ vertices and edges.*

As a simple example, consider the split system described in the caption of Fig. 10.2. Although the split network in part (ii) of this figure is planar, the leaf c is within the figure; an outer-labeled planar network is not possible in this case, since Σ_3 is not a circular split system.

It is interesting to contrast the polynomial bound in Proposition 10.3 with size of the Buneman graph for the full circular split system Σ_n on $[n]$ of size $\binom{n}{2}$. Recall from Section 2.4.1 that the number of vertices in the Buneman graph for some set Σ of splits is 1 plus the number of cliques in the graph that has vertex set Σ and an edge between each pair of splits that are incompatible. For $m = 1$, each of the $\binom{n}{2}$ elements of Σ_n constitutes a singleton clique in $B(\Sigma_n)$, while for $m \geq 2$, it is not hard to show that the subsets of Σ_n of size m that are pairwise incompatible are in one-to-one correspondence with the subsets of $2m$ edges of a regular n -gon. Since there are $\binom{n}{2m}$ ways to select $2m$ edges from the n edges of the regular n -gon, the following exact expression (from [88]) results:

$$|B(\Sigma_n)| = 1 + \sum_{m=1}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{2m} = 2^{n-1}.$$

Exercise: Suppose that T and T' are unrooted binary phylogenies on $[n]$. Is $\Sigma(T) \cup \Sigma(T')$ always a circular split system? Consider the case $n = 5$.

Neighbor-Net. Given distance data, one of the most widely used methods for constructing a phylogenetic network was developed by mathematicians David Bryant and Vincent Moulton in 2004 [68]. The *Neighbor-Net* algorithm converts a distance function δ on X into a circular split system, with associated weights, and so returns a split network (the parallel edges are assigned a length equal to the weight of the corresponding split (see Fig. 10.3)). By Proposition 10.3, the resulting network is outer-labeled planar. As well as being widely used in biology, Neighbor-Net also finds application in many other areas of classification, including language evolution, stemmatology, linguistics, and psychology.⁷³

The algorithm for constructing the cyclic split system from a distance function has a similar flavor to the NJ algorithm (Section 6.2.1), coupled with an estimation of “optimal” split weights; for the full details of the algorithm, we refer the reader to [68] or [203]. Here, we mention one important consistency property of the Neighbor-Net algorithm: if δ comes from a weighting of a circular split system, then Neighbor-Net will return the split system and the weights correctly. To make this more precise, given a set Σ of

⁷³Some novel uses of Neighbor-Net have included an application to the stock market, as well as a 2015 study of different versions of the “Little Red Riding Hood” story.

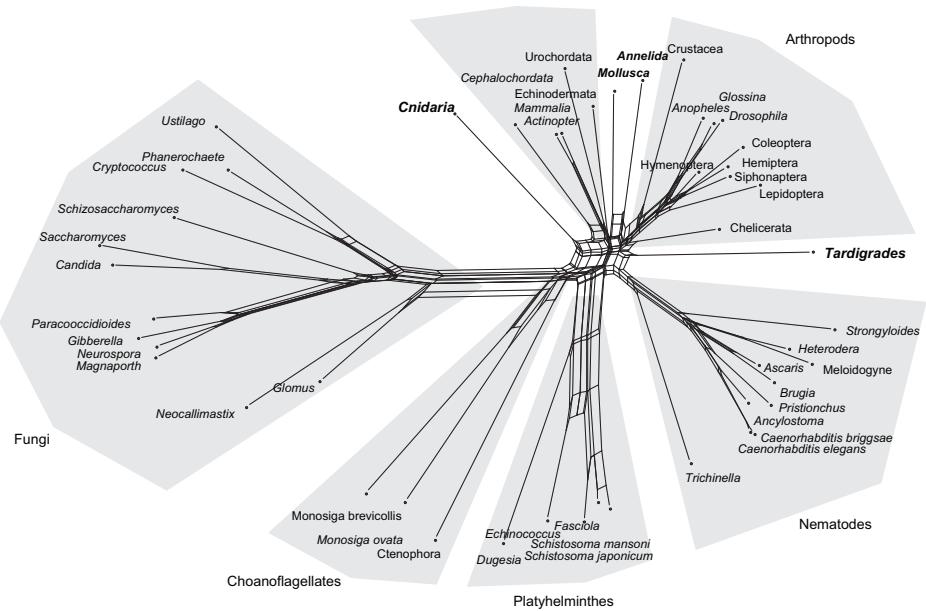


Figure 10.3. A split network (Neighbor-Net) constructed from a concatenation of 71 genes from 49 animals, fungi, and choanoflagellates. The major groupings of species are also indicated. Reprinted with permission from Oxford University Press [202].

X -splits, and a weight function $w : \Sigma \rightarrow \mathbb{R}^{>0}$, let $d = d_{(\Sigma, w)}$ be the distance function on X defined by

$$d(x, y) = \sum_{A|B \in \Sigma} w(A|B) \delta_{A|B}(x, y), \quad (10.2)$$

where $\delta_{A|B}(x, y) = 1$ precisely if x and y lie in different blocks of the split $A|B$; otherwise, $\delta_{A|B}(x, y) = 0$. When Σ is a circular split system (i.e., when d satisfies the Kalmanson condition (cf. Proposition 6.10)) we say that d is a *circular distance function* with support Σ . The following result was announced in [68], with a subsequent proof in [69] (see also [227], Theorem 29).

Theorem 10.4. If δ is a circular distance function with support Σ , then Neighbor-Net, applied to δ , returns the same circular split system Σ , with the same weights that applied in the representation of δ by eqn. (10.2).

Further mathematical, computational, and statistical aspects of the Neighbor-Net algorithm have been developed in [227]. Implementations of Neighbor-Net and many other split network techniques are available in the SplitsTree software package [202].

Circular split systems are part of a more general class called *weakly compatible split systems*. These can be viewed as the unrooted analogue of the weak hierarchies that we met in Chapter 2 (just as pairwise compatible split systems are the unrooted analogue of hierarchies). For a distance function δ induced by a positive weighting of a set Σ of weakly compatible splits, the theory of *split decomposition*, developed by Hans-Jürgen Bandelt and Andreas Dress in the early 1990s, provides a technique to recover Σ and the associated

weights uniquely from δ . In this way, δ can be represented by a split network, with lengths assigned to the parallel edges, so that $\delta(x, y)$ is the shortest path in the network connecting x and y . In general, these networks will not be outer-labeled planar, and when $n = |X|$ is large the networks tend to be poorly resolved. As a result, the use of split decomposition networks have been largely eclipsed by Neighbor-Net graphs. For further details on split decomposition, see [203] or [117].

Median (and quasi-median) networks. There is another way to describe the Buneman graph, which relates more directly to sequences of binary characters; it also leads to an unexpected connection with the maximum parsimony (MP) trees of Chapter 5. Suppose that $\mathcal{C} = (f_1, f_2, \dots, f_k)$ is a sequence of binary characters on X that induce distinct X -splits. Associate each element $x \in X$ with a sequence $s_x = (f_1(x), f_2(x), \dots, f_k(x))$ in $\tilde{S} = \prod_{i=1}^k S_i$, where S_i is the state space of character f_i .

For any three sequences $s, s', s'' \in \tilde{S}$, consider the resulting *median sequence* $m(s, s', s'')$ in \tilde{S} defined by letting $m(s, s', s'')_i$ be the majority state of s_i, s'_i and s''_i (this is well defined because the state space is binary, so one of the two states must occur as a strict majority over the other state). For example, if $s = (\text{AAC}), s' = (\text{ACT}),$ and $s'' = (\text{TAT})$, then $m(s, s', s'') = (\text{AAT})$. The *median hull* of a subset W of \tilde{S} is the minimal subset of \tilde{S} containing W that is closed under the median operation. This can be constructed from W by repeatedly applying the median sequence operation to the current set until no further elements can be added.

Let $\mathcal{M}_{\mathcal{C}}$ denote the median hull of the sequences $(s_x : x \in X)$ derived from \mathcal{C} , and let $G_{\mathcal{C}}$ be the graph with vertex set $\mathcal{M}_{\mathcal{C}}$ and with two vertices adjacent if their Hamming distance is 1 (i.e., if the two sequences differ at one position). This graph has a natural labeling map $f : X \rightarrow \mathcal{M}_{\mathcal{C}}$, defined by $x \mapsto s_x$.

In this way, the Buneman graph $B(\Sigma)$ of the splits induced by \mathcal{C} can be identified with the graph $G_{\mathcal{C}}$ (for a proof, see [315]). Suppose next that T is a maximum parsimony phylogeny for \mathcal{C} . Let $S_{\mathcal{C}}$ denote the collection of sequences at the interior vertices of T in any most parsimonious reconstruction of \mathcal{C} on T (i.e., for each character f_i , take a minimal extension F_i and consider the sequence $(F_1(v), \dots, F_k(v))$ for each vertex v of T). A remarkable result, due to Hans-Jürgen Bandelt, is that the median hull $\mathcal{M}_{\mathcal{C}}$ contains $S_{\mathcal{C}}$. In other words, the MP trees for \mathcal{C} (along with their most parsimonious reconstructions) can be regarded as sitting within the Buneman graph for the induced splits (further details, proofs and references can be found in [315]).

An extension of this result to nonbinary characters was given more recently in [22]. Again, we use a triple-wise median operation to form a median hull; however, in this case, a problem arises: there is no longer a well-defined “majority” element if the number of states is greater than 2. Fortunately, there is a simple remedy: Order the sequences arbitrarily and, for any three sequences s, s' and s'' , let $m(s, s', s'')_i$ be the majority state at position i if this exists, otherwise, in the case of a tie, take it to be the state of the minimal sequence (under the arbitrary ordering) at position i .

Given any subset W of \tilde{S} , the *quasi-median hull* of W is the minimal subset of \tilde{S} containing W that is closed with respect to this triple-wise operation. The resulting “quasi-median” graph (which generalizes the Buneman graph) for a sequence \mathcal{C} of characters again harbors a MP tree for \mathcal{C} , along with the most parsimonious reconstructions of the characters on that tree (for precise details and a proof, see [22]). There is also an elegant mathematical description of the blocks (maximal biconnected components) of this quasi-median graph, leading to a polynomial-time algorithm for computing its block decomposition [187]. This, in turn, can be potentially helpful in finding MP trees.

10.3 ▀ Explicit (directed) networks

The networks described in the previous section are typically used to show how “tree-like” (or not) particular data are. These networks can also reveal where reticulation is likely to take place, and even identify possible reticulation events, such as the formation of hybrid species. However, undirected networks are essentially implicit representations of evolution. For the rest of this chapter, we will mostly be concerned with networks that attempt to explicitly show how a group of species evolved from a common ancestor via processes that combine tree-like evolution with reticulation events. Accordingly, such networks need to be rooted with the arcs directed away from the root.

Recall that in the directed setting, a phylogenetic network on X is a connected acyclic digraph $N = (V, A)$, with a root vertex ρ of in-degree 0, a set X of vertices of out-degree 0 (leaves, which may have in-degree greater than 1), with no parallel arcs, and with no vertices having in-degree and out-degree both equal to 1.

For directed phylogenetic networks, other restrictions are also sometimes imposed; we will mention these as required.

We will let $\mathcal{N}(X)$ denote the set of all directed phylogenetic networks on X , identifying two networks if they are equivalent (as defined at the start of this chapter). In contrast to the phylogenetic tree setting, where $RP(X)$ is finite, $\mathcal{N}(X)$ is infinite for any given set X . Unless otherwise stated, “network” will henceforth mean a directed phylogenetic network.

Exercise: For any phylogenetic network $N \in \mathcal{N}(X)$ and any vertex v in N , show that there is a path from the root vertex ρ to v .

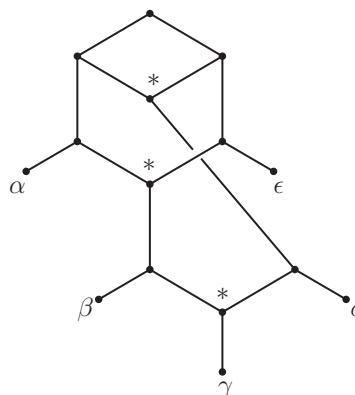


Figure 10.4. A directed phylogenetic network from a recent study into the complex hybrid evolution of a variety of species of bread wheat (modified from [238] (Fig. 3)). The three reticulation vertices indicated by * correspond to three hypothesized hybridization events. The leaf labeling corresponds to $\alpha = Triticum uartu$, $\beta = Triticum turgidum$, $\gamma = Triticum aestivum$, $\delta = Aegilops tauschii$, and $\epsilon = Aegilops speltoides$.

Definitions. For any phylogenetic network $N = (V, A)$ on X , and two vertices u and v in V we say that u is the *parent* of v , and v is the *child* of u , if $(u, v) \in A$. In addition,

- a vertex with in-degree at most one is called a *tree vertex*;
- a vertex of in-degree greater than one is called a *reticulation vertex*, or is said to be *reticulate*;

- if $e = (u, v)$ and v is reticulate, we call e a *reticulate arc*;
- if $e = (u, v)$ and v is tree vertex, we say that e is a *tree arc*.

Thus, every vertex in a phylogenetic network is either a tree vertex (i.e., a vertex of in-degree exactly 1, or the root vertex) or a reticulation vertex. We will use v , t , and r to denote the number of vertices, tree vertices, and reticulate vertices in N , respectively.

For vertices u, v of $N \in \mathcal{N}(X)$, we will write $u \preceq_N v$ precisely if $u = v$ or if there is a directed path in N from u to v . This is a partial order on the set of vertices of N with ρ being the unique minimal vertex. The direction of the arcs in the figures that follow is always downwards. An example of a (directed) phylogenetic network is shown in Fig. 10.4

Phylogenetic networks in general can be arbitrarily complex, so it helps to consider subclasses of $\mathcal{N}(X)$ that have specific properties.

10.3.1 ■ Binary phylogenetic networks

A *binary phylogenetic network* on X is a network $N \in \mathcal{N}(X)$ with the additional properties that

- the root has out-degree 2;
- the vertices in X (of out-degree 0) have in-degree 1;
- all other vertices either have in-degree 1 and out-degree 2, or in-degree 2 and out-degree 1.

We will let $\mathcal{N}_B(X)$ denote the set of binary phylogenetic networks on X . Again, this is an infinite set, even after equivalent networks are considered to be identical. Notice that for any binary phylogenetic network, the number r of reticulation vertices is also the cyclomatic number of N (regarded as a graph by ignoring the directions of the edges).

One of the reasons binary phylogenetic networks are attractive to study is that the number of vertices and arcs is fully determined by the number of leaves and reticulations. The following result is adapted from [245].

Lemma 10.5. *Let $N \in \mathcal{N}_B(X)$ have n leaves and r reticulations. Then N has $v = 2n + 2r - 1$ vertices and $3r + 2n - 2$ arcs.*

Proof. First note that N has $v = t + r$, where t is the number of tree vertices. Let b be the number of vertices that have in-degree 1 and out-degree 2. Then $t = b + n + 1$. Now, the sum of the out-degrees of the vertices of N equals the number a of arcs of N , which, in turn, equals the sum of the in-degrees of the vertices of N . This gives the equation

$$r + 2b + 2 = a = 2r + b + n,$$

from which $b = r + n - 2$ follows. The remaining equations in the lemma now follow directly. ■

Exercise: Establish the following equalities for every binary phylogenetic network N : $n + r = b + 2 = \frac{1}{2}(v + 1)$ where b is the number of vertices of in-degree 1 and out-degree 2. Deduce that any pair of the four variables n, r, b , and v (other than $\{b, v\}$) determine the other parameters, and that as v tends to infinity, both $n + r$ and b are asymptotically equivalent to $v/2$ [245].

The level of a binary phylogenetic network. In a phylogenetic network, reticulation events can be widely separated, so that the network can effectively be viewed as a tree with local reticulation events. Alternatively, a phylogenetic network could be a highly “tangled” web of reticulations. A natural way to quantify this is by an integer-valued index called the “level” of a network [90]. To formalize this notion we first need a couple of definitions.

A *biconnected component* or *block* of a rooted phylogenetic network N is a subnetwork N' of N for which (i) by removing any vertex (and its incident arcs) from N' the resulting graph is still connected (i.e., there is an undirected path connecting any two vertices in the resulting graph) and (ii) N' is maximal with respect to property (i). In this way N can be represented as a tree with vertex set being the set of blocks of N together with the set of cut vertices of N (any vertex which disconnects N when it is deleted), and the arcs assigned according to containment of cut vertices in blocks. In a rooted phylogenetic tree with two or more leaves, the biconnected components are exactly the edges of the tree, and the cut vertices are the interior vertices.

A binary phylogenetic network N is said to be a *level- k network* if every biconnected component of N has at most k reticulation vertices.⁷⁴ For example, a binary phylogenetic network N has level 0 if and only if it is a tree. A level-1 binary network is often referred to as a (rooted) *galled tree*.

Once a bound is imposed on the level of a binary phylogenetic network, the size of the network becomes bounded, as the following result from [366] (Lemma 2) shows.

Proposition 10.6. *For any fixed k , any level- k binary phylogenetic network $N = (V, A)$ on n leaves contains $O(n)$ vertices and $O(n)$ arcs. Specifically,*

$$|V| \leq 2n - 1 + k(n - 1) \quad \text{and} \quad |A| \leq 2n - 2 + \frac{3}{2}k(n - 1).$$

10.3.2 ■ Tree-child, tree-sibling, and reticulation-visible networks

Bounding the level of a network is one way to establish some control over the class of phylogenetic networks. Another way is to impose a structural condition on the network, for example, by considering how the vertices connect via paths to the leaves. This is the motivation behind three related classes of networks that are particularly amenable to analysis. Networks in these three classes need not necessarily be binary.

We start with a class that was defined fairly recently (2009), by [74], and which has turned out to be one of the most natural and important classes of networks.

A *tree-child network* is a phylogenetic network $N = (V, A)$ for which every nonleaf vertex v has a child that is a tree vertex (i.e., each nonleaf vertex has an out-going tree arc). Equivalently, a tree-child network is a network with the following property: From every vertex v of N , there is path from v to a leaf that consists only of tree vertices (except perhaps v itself).

Notice that a phylogenetic network is tree-child if and only if there is no vertex that has only reticulate children. The two ways that this can occur in the binary setting are illustrated in Fig. 10.5(iii).⁷⁵

⁷⁴The notion of level k extends to nonbinary networks by requiring that the network can be converted into a tree by deleting at most k arcs from each biconnected component.

⁷⁵This local characterization of tree-child networks was made explicit in [313]; we present another characterization of tree-child networks from the same paper, in Proposition 10.15.

For any tree-child network with n leaves, the number r of reticulation vertices satisfies the following inequality from [74]:

$$r \leq n - 1. \quad (10.3)$$

A short argument for this is to observe that if v is one of the r reticulation vertices or the root vertex, then there is a directed path P_v from v to a leaf in X that does not contain a reticulation vertex. Since these $r + 1$ paths must be vertex disjoint (since they only consist of tree vertices and so cannot merge) we must have $r + 1 \leq n$.

The bound in inequality (10.3) can be realized for every n by a tree-child network [74] (the example with $n = 2$ is shown in Fig. 10.5(i)). One can say a little more for binary tree-child networks, as the following result from [245] shows.

Proposition 10.7. *If a binary tree-child network has v vertices, n leaves, and r reticulation vertices, then*

$$r < \frac{v}{4} < n. \quad (10.4)$$

To see this, Lemma 10.5 gives $v = 2n + 2r - 1$, which, by eqn. (10.3) is both strictly more than $4r$ and strictly less than $4n$. This bound is illustrated in Fig. 10.5(i) where $r = 1$, $n = 2$, and $v = 5$.

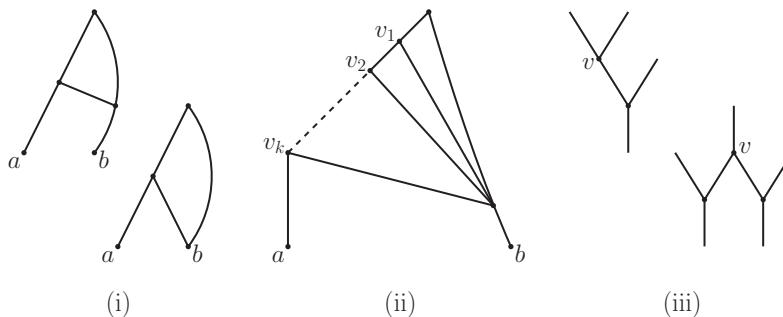


Figure 10.5. (i) Top: a binary tree-child network to illustrate inequality (10.4). The first half of this inequality does not apply to the (nonbinary) tree-child network below it. (ii) A nonbinary tree-child network with $n = 2$ leaves, which can have an arbitrarily large number $(k + 4)$ of vertices for all $k \geq 1$. (iii) Two local obstructions that prevent a binary network from being tree-child (top: the unique child of v is reticulate; bottom: the children of the tree vertex v are reticulate).

Note that the number of vertices v in a (nonbinary) tree-child network cannot be bounded just by $n = |X|$, as Fig. 10.5(ii) shows. Nevertheless, if m is the maximal indegree of any reticulation vertex, then the following bound applies (from [74], Proposition 1(c)):

$$v \leq (m + 2)(n - 1) + 1.$$

The class of tree-child networks includes all level-1 networks [74] but not all level-2 networks.

Tree-sibling networks. Two vertices v, v' in a phylogenetic network $N = (V, A)$ are said to be *siblings* if they share a parent (i.e., there is some $u \in V$ with $(u, v), (u, v') \in A$). A phylogenetic network $N = (V, A)$ is a *tree-sibling network* if every reticulation vertex v has a sibling that is a tree vertex.

Lemma 10.8. *Every tree-child network N is also a tree-sibling network.*

Proof: Suppose that N is a tree-child network and v is a reticulation vertex. Let w be a parent of v . Then w has a child v' that is a tree vertex and v' is a sibling of v . Since this holds for all such v , N is a tree-sibling network. ■

In contrast to tree-child networks, the number of reticulations in a binary tree-sibling network cannot be bounded by *any* function of n (the number of leaves), as Fig. 10.6(ii) shows.

Reticulation-visible networks. Suppose that N is a phylogenetic network on X . A vertex v is said to be *visible* if there is a leaf $l_v \in X$ so that every directed path from the root to l_v passes through v (in case v is already a leaf we can take $l_v = v$). Notice that if network N is tree-child then every vertex is visible. To see this, observe that if N is tree-child then for every nonleaf vertex $v \in V$, there is a path from v to a leaf l_v that passes through only tree vertices, so all paths from the root to l_v must pass through v . Conversely, if every vertex of a network N is visible, then it can easily be shown that N is tree-child.

This motivates the definition of the following class of networks that includes tree-child ones, where visibility is only required to hold for the reticulations vertices. A phylogenetic network $N = (V, A)$ is *reticulation-visible* if each reticulation vertex is visible.

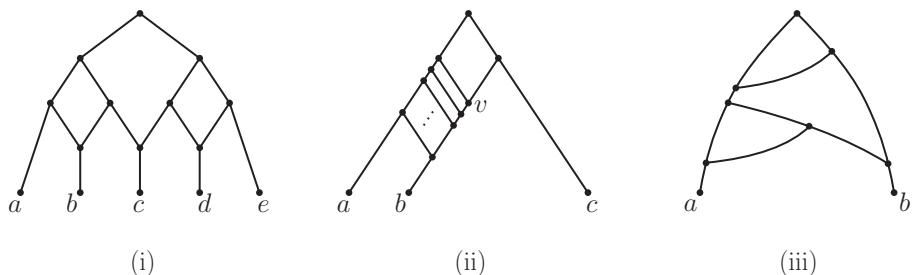


Figure 10.6. *Three binary phylogenetic networks. (i) This network is reticulation-visible (for each reticulation vertex v , take l_v to be the unique leaf below v), but is not a tree-sibling network (the parent of leaf c has only reticulate siblings). (ii) This network is tree-sibling but is not reticulation-visible (since b is the only leaf below vertex v , and there is at least one other path to b that avoids v). This network can have an arbitrary number of reticulation vertices. (iii) A network which has the maximum possible number of vertices and reticulation vertices for binary reticulation-visible networks when $n = 2$.*

Lemma 10.9. *If N is a binary reticulation-visible network, and v is a reticulation vertex of N , then each parent of v is a tree vertex.*

Proof: Suppose, to the contrary, that v has a reticulate parent w . Then any leaf l_w that is reachable from w is also reachable from v . Since v is reticulate, there is another path leading to v and from there to l_w that avoids w . Thus the vertex w violates the reticulation-visible condition for N . ■

For binary reticulation-visible networks, the number of reticulations r satisfies a linear bound with the number n of leaves that is similar to (but weaker than) the bound for tree-child networks. The following tight result is from the recent paper [54].

Theorem 10.10. Let N be a binary reticulation-visible network on X , with $|X| = n$. Then N has at most $3(n-1)$ reticulation vertices and at most $8n-7$ vertices in total. These bounds are sharp for all integers $n \geq 1$.

An example that shows these bounds are sharp for $n = 2$ is shown in Fig. 10.6(iii); this example can be used as a template for constructing extremal examples for larger values of n [54].

Exercise: Show that if N is a phylogenetic network then every nonleaf vertex v of N is visible if and only if N is a tree-child network.

10.3.3 ▪ Temporal networks

A phylogenetic network $N = (V, A)$ is said to be a *temporal network* if there is a function $t : V \rightarrow \mathbb{R}^{\geq 0}$ such that, for each arc (u, v) of N , the following two properties hold:

T1: $t(u) = t(v)$ if (u, v) is a reticulation arc.

T2: $t(u) < t(v)$ if (u, v) is a tree arc.

The map t is called a *temporal labeling* of N .

The motivation behind this concept is that if a reticulation event involves two species in the past forming a hybrid species, then those two species must have been extant at the same time (along with the hybrid they formed). In other words, on a vertical time axis, the evolution is essentially “horizontal” and the tree arcs are regarded as corresponding to “vertical evolution” from an ancestral species to a descendant species at a later time.

The notion of temporal labeling of a network seems so reasonable that it might be asked why it was not imposed from the outset on phylogenetic networks. The reason is subtle and was noted long ago by the biologist Wayne Maddison: a “horizontal” evolutionary event might have involved a lineage which later became extinct (or was not sampled at the present) but not before being involved in a further “horizontal” evolutionary event with lineages that survive to the present. To allow for such hitch-hiking on these unsampled lineages, condition T1 may need to be weakened to the requirement that $t(u) \leq t(v)$ if (u, v) is a reticulation arc; however, in that case, every phylogenetic network would have such a labeling, so the requirement for such a labeling would be vacuous. In summary, requiring temporal labeling is a reasonable condition in certain settings but not a completely compelling one in general.

Not all networks have a temporal labeling. Figure 10.7(i) provides a simple example. To see why no temporal labeling can exist for this network, suppose to the contrary that there were such a labeling. Then condition T1 implies that $t(u) = t(v) = t(w) = t_1$ (say) and that $t(u') = t(v') = t(w') = t_2$ (say). If we now apply condition T2 for the arc (u', u) , we get $t_2 < t_1$, whereas for the arc (w, w') , we obtain $t_1 < t_2$, a contradiction.

There is a simple polynomial-time algorithm to determine whether an arbitrary network N has a temporal labeling, as follows. Given $N = (V, A)$, consider the associated digraph (V', A') where V' is the set of equivalence classes of V under the equivalence relation $u \sim v$ if either $u = v$ or there is an (undirected) path in N between u and v consisting of reticulation arcs only (regarded as edges). The arc set A' is now defined by $([u], [v]) \in A'$ if there exists some $u_1 \in [u]$ and $v_1 \in [v]$ for which (u_1, v_1) is a tree arc. Then N has a temporal labeling if and only if (V', A') is acyclic [24]. This is illustrated in Fig. 10.7, where the associated digraph for the network in (i) is shown in (ii).

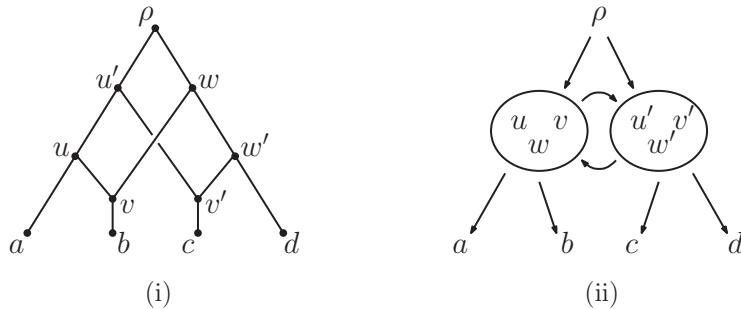


Figure 10.7. (i) A binary phylogenetic network that has no temporal labeling; (ii) the associated digraph (V', A') containing a directed cycle.

10.3.4 • Networks without redundant arcs

An arc (u, v) in a network N is said to be *redundant* if there is another directed path from u to v . Restricting networks to those without redundant arcs provides another way to impose some mathematical structure into phylogenetic networks, as we will see in this and later sections.

We first need to describe a simple “compression” operation on networks. Given a network $N \in \mathcal{N}(X)$, we will let \tilde{N} be the network in $\mathcal{N}(X)$ obtained from N by collapsing any arc $(u, v) \in A$ for which u is a vertex of out-degree 1 and v is a vertex of in-degree 1. Notice that the set of arcs of N that are collapsed in this way comprise a set of pairwise nonadjacent arcs (i.e., no two arcs share a vertex). We refer to \tilde{N} as the *compression* of N and we say that N is *compressed* if $\tilde{N} = N$. A simple example is shown in Fig. 10.5(i), where the lower network is the compression of the higher one.

Normal networks. A rooted phylogenetic network is a *normal network* if it is tree-child and has no redundant arcs. Normal networks have a number of attractive mathematical properties, which have been explored by the mathematician Stephen J. Willson. In some definitions of normal networks, it is assumed that the network is compressed; however, we will not require this here.

To delve further into the properties of normal networks, we need to discuss the notion of the least common ancestor (LCA) of two leaves in a network. Unlike the tree setting, this may not exist; but when it does, it is unique. For $N = (V, A) \in \mathcal{N}(X)$ and any two elements $x, x' \in X$, v is a *least common ancestor* of $\{x, x'\}$ if (i) $v \preceq_N x$ and $v \preceq_N x'$, and (ii) v is maximal with respect to this property (i.e., if $w \preceq_N x$ and $w \preceq_N x'$, then $w \preceq_N v$). There is at most one such vertex v (since \preceq_N is a partial order) and when it exists, we write $v = \text{lca}_N(x, x')$.

Notice that if N is normal, then its compression \tilde{N} is also a normal network on X . A fundamental property of normal networks is the following result from [374] (Theorem 3.9).

Proposition 10.11. *Suppose that N is a normal network on X . Then*

- (i) *each vertex $v \in \tilde{N}$ can be written as $v = \text{lca}_{\tilde{N}}(x, x')$ for some $x, x' \in X$ (with $x = x'$ precisely if $v \in X$);*
- (ii) *\tilde{N} has at most $\binom{n}{2} + n$ vertices, and so N has $O(n^2)$ vertices.*

An interesting aspect of part (ii) of this proposition is that, as we saw in Fig. 10.5(ii), for tree-child networks there is no upper bound on the number v of vertices in terms of just n ; however, when we couple the tree-child condition with the absence of redundant arcs, a bound emerges, which is quadratic in n .

An outline of the proof of Proposition 10.11(i) is as follows. The result clearly holds if $v \in X$. For a vertex $v \in \tilde{N}$ that is not in X , v has at least two distinct children a and b in \tilde{N} with a a tree vertex. Since N is tree-child, there is a path from a to a leaf x —and a path from b to a leaf x' —with each path passing through tree vertices only. Clearly, $v \preceq_{\tilde{N}} x$ and $v \preceq_{\tilde{N}} x'$. Moreover, because N has no redundant arcs, and the paths from v to x via a and from b to x' pass through only tree vertices it can be shown that $v = \text{lca}_{\tilde{N}}(x, x')$.

The first half of part (ii) follows immediately from part (i). For the second claim, notice that in converting $N = (V, A)$ to its compression $\tilde{N} = (\tilde{V}, \tilde{A})$, the corresponding function $f : V \rightarrow \tilde{V}$ has $\#f^{-1}(\tilde{v}) \in \{1, 2\}$ for all $\tilde{v} \in \tilde{V}$, and so N has no more than twice the number of vertices of \tilde{N} .

Exercise: Show that if N is a normal network on $[n]$ then the number r of reticulations satisfies $r \leq n - 2$, which is slightly stronger than the bound for tree-child networks given by (10.3). [Hint: First show that for a normal network, the children of the root vertex must be tree vertices.]

Any binary tree-child network that has a temporal labeling—referred to as a *hybridization network* in [235]—provides an example of a normal network. Figure 10.4 is an example of such a hybridization network.

Proposition 10.12. *Any temporal tree-child binary network is a normal network.*

Proof. Since N is tree-child, we just need to check that N has no redundant arcs. Suppose, on the contrary, that an arc (u, v) of N is redundant. In that case there is another path P from u to v , and v is reticulate; let (w, v) be the last arc on path P ($w \neq u$, since N has no parallel arcs). Since N is a temporal network, and v is reticulate, $t(u) = t(v) = t(w)$. On the other hand, path P must pass through at least one tree vertex including w , since N is tree-child and binary. This would then imply that $t(u) < t(w)$, a contradiction. ■

Regular networks. Given a phylogenetic network $N = (V, A)$ on X , the *cluster* associated with a vertex $v \in V$ is the set

$$c_N(v) = \{x \in X : v \preceq_N x\}.$$

A phylogenetic network $N = (V, A)$ is said to be *regular* if it satisfies the following three properties:

- (R1) If $u, v \in V$ with $u \neq v$, then $c_N(u) \neq c_N(v)$;
- (R2) $u \preceq_N v$ if and only if $c_N(v) \subseteq c_N(u)$;
- (R3) N has no redundant arcs.

This definition is equivalent to the condition that the map $v \mapsto c_N(v)$ is a digraph isomorphism from N to the cover digraph of the set of clusters of N .⁷⁶ In other words, N is regular precisely if it is naturally equivalent to the cover digraph of its clusters.

⁷⁶The cover digraph of a collection \mathcal{C} of subsets has vertex set \mathcal{C} and an arc (S, S') when $S' \subset S$ and there is no set $S'' \in \mathcal{C}$ with $S' \subset S'' \subset S$.

Notice that any rooted phylogenetic X -tree is a regular network. Furthermore, if N is regular, then N has no arc (u, v) where u has out-degree 1 and v has in-degree 1 (otherwise (R1) is violated); in other words, N is a compressed network (i.e., $N = \tilde{N}$).

It can be shown that every compressed normal network is a regular network; in other words, if N is normal (i.e., tree-child, and satisfying (R3)), then \tilde{N} also satisfies (R1) and (R2) [374]. However, regular networks can be much larger than normal networks on the same X of size n . We saw that any normal network has $O(n^2)$ vertices in total but the regular network that corresponds to the cover digraph of all nonempty subsets of $[n]$ has $2^n - 1$ vertices.

Cluster networks. Given a collection \mathcal{C} of subsets of X containing X , we can represent \mathcal{C} by a regular network $N_{\mathcal{C}}$ on X , namely the cover digraph of \mathcal{C} . If \mathcal{C} is a hierarchy, then we obtain a rooted phylogenetic X -tree. More generally, $N_{\mathcal{C}}$ may have reticulate vertices which have in-degree at least 2 and out-degree at least 2. For practical applications, it is usually more visually meaningful to replace $N_{\mathcal{C}}$ by its *expansion*, which is exactly the inverse of the compression operation described earlier. First if w has in-degree > 1 and out-degree > 1 , then replace w by vertices u, v and an arc (u, v) in which the arcs that led into w now lead into u and the arcs that led out of w now lead out of v . Second, if $x \in X$ is reticulate, then replace x by the arc (u, x) in which the arcs that lead into x now lead into u . The resulting network is called a *cluster network*.

Given a collection \mathcal{C} of subsets of X , one can think of a cluster network for \mathcal{C} as having each set in $\mathcal{C}' = \mathcal{C} - \{X\}$ assigned to an arc of the network that is incoming to some tree vertex, so that $A \in \mathcal{C}'$ is the subset of elements of X that are reachable by a directed path from that arc.

Counting networks. Despite the tractable formulae presented for counting undirected binary level-1 trees from Section 10.2.1, when we move to larger classes of directed networks, counting becomes much harder, and until recently few results were known, even asymptotically. This changed in 2015, with asymptotic results on the number of rooted binary phylogenetic networks of different types as a function of both the number of leaves and the total number of vertices [245]. Here, we give a sample of some of these results. Recall from Chapter 2 that we can write $|RB(n)| = 2^{n \log_2 n + O(n)}$. Let TC_n and NL_n denote the set of tree-child and normal networks with leaf set $[n]$. The authors of [245] showed that $|TC_n|$ and $|NL_n|$ can both be written in the form

$$2^{2n \log_2 n + O(n)}.$$

Thus the sizes of these two classes of networks grow at the rate $|RB(n)|^{2+o(1)}$. Moreover, as $n \rightarrow \infty$,

$$\frac{|NL_n|}{|TC_n|} \rightarrow 0.$$

In other words, nearly all tree-child networks on $[n]$ fail to be normal for large values of n . Moreover, almost all networks in TC_n and almost all networks in NL_n have $(1 + o(1))n$ reticulation vertices and $(4 + o(1))n$ vertices in total. The arguments in [245] involve techniques and classical results from random graph theory.

10.4 • Trees displayed by networks

Suppose N is a phylogenetic network on X . For each reticulation vertex, let us delete all but one of the two incoming arcs. This produces a rooted tree T' , though it may have

leaves that are not in X and it may have vertices of in-degree and out-degree both equal to 1. If we take the minimal subtree of T' that contains X and suppress any vertices of in-degree and out-degree both equal to 1, then we obtain a rooted phylogenetic X -tree T . We say that N displays T . This is illustrated in Fig. 10.8. Notice that N displays $bc|a$ in two ways, and in one case, the tree T' has a leaf (adjacent to the root) that is not in X (a so-called “dummy leaf”).

An alternative definition of display is that N contains a subgraph T' that is a subdivision of T . Notice that when N and T are rooted phylogenetic X -trees, then N displays T if and only if N and T are equivalent (thus the notion of display in this special case is stronger than corresponding notion of display in Chapter 4).

We will let $\mathcal{T}(N)$ denote the set of rooted phylogenetic X -trees displayed by N . Given a sequence \mathcal{P} of rooted phylogenetic X -trees, we say that N displays \mathcal{P} if N displays each tree in \mathcal{P} .

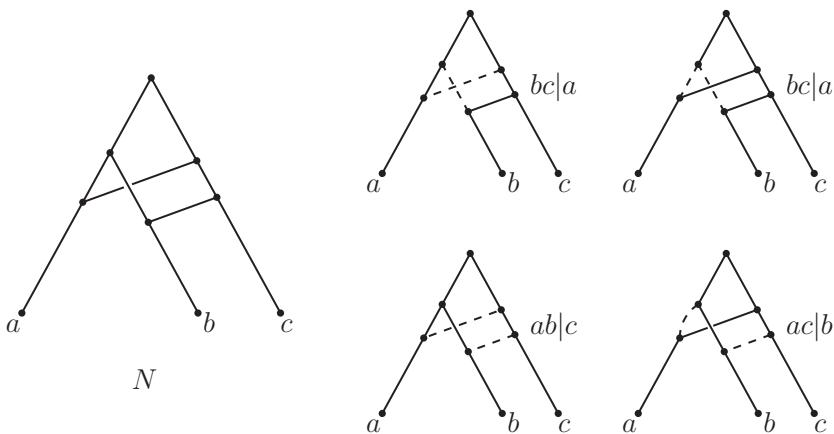


Figure 10.8. The network N displays the tree $bc|a$ in two different ways (top right), as well as each of the trees $ab|c$ and $ac|b$ (bottom right).

Exercise: Show that for any network $N \in N(X)$, $\mathcal{T}(N) = \mathcal{T}(\tilde{N})$, where \tilde{N} is the compression of N , defined at the start of Section 10.3.4.

Notice that a binary phylogenetic network N with r reticulations can display at most 2^r distinct trees, and can display less, even when N is tree-child. But if we also rule out redundant arcs, then the 2^r bound becomes exact, as the following result from [376] (Corollary 3.4; see also [91]) shows.

Proposition 10.13. *A binary normal network N with r reticulations displays exactly 2^r trees.*

For the tree-sibling network shown in Fig. 10.9 with $t - 2$ reticulations, the number of trees displayed is a Fibonacci number [233], so the proportion of displayed phylogenetic trees relative to the upper bound 2^r can be made arbitrarily small for t large enough.

When the number of phylogenetic X -trees displayed by a binary network on X is less than 2^r , this means that at least one phylogenetic X -tree must be displayed twice. An elegant characterization of a subclass of binary tree-sibling networks that displays a tree

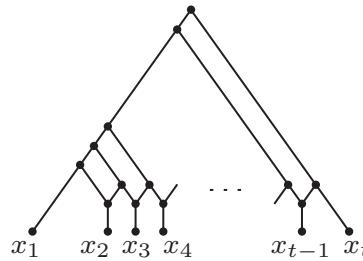


Figure 10.9. A binary phylogenetic network with t leaves, $t-2$ reticulations, and for which the number of phylogenetic trees that it displays is the Fibonacci number F_t (with $F_1 = F_2 = 1$).

twice was described in [91], along with an $O(n^2)$ algorithm ($n = |X|$) to test whether or not a given network in this subclass has this property or not.

Given any set of rooted phylogenetic X -trees \mathcal{P} , it can be shown that the regular network (or its associated cluster network) corresponding to the set $\mathcal{C} = \bigcup_{T \in \mathcal{P}} \mathcal{C}(T)$ (i.e., all clusters that are present in at least one tree from \mathcal{P}) gives a network that displays each tree in \mathcal{P} . In particular, if we take $\mathcal{P} = RP(n)$ there is a network that displays all phylogenetic X -trees. While this network is easy to describe, its number of vertices grows exponentially with n . With more care, it is not hard to construct a binary phylogenetic network N_X on X that displays all trees in $RB(X)$ using only a polynomial (in $|X|$) number of vertices and arcs.

Exercise⁺: Show that if $N_X \in \mathcal{N}_B(X)$ displays every tree $T \in RB(X)$, then the number of reticulation vertices in N_X must be at least $\Omega(n \log n)$ for $n = |X|$. Deduce that N_X cannot be chosen to be reticulation-visible for sufficiently large values of n .

10.4.1 • The tree containment problem

A basic computational question is how to decide whether a given phylogeny is displayed by a given network. More precisely, the *tree containment problem* (TCP) is the following: Given a phylogenetic network $N = (V, A)$ on X and a rooted phylogeny $T \in RB(X)$, does N display T ?

The TCP has a polynomial-time algorithm for binary tree-child networks; however, TCP is NP-hard for tree-sibling networks even when they are also further constrained to have a temporal labeling and to be regular [368]. For the class of binary level- k networks (for any fixed k), TCP also has a polynomial-time solution [368].

The question of whether or not the TCP has a polynomial-time solution for reticulation-visible binary networks remained open for five years, but was eventually solved by two groups (independently and within weeks of each other) in 2015 [54, 167]. In both cases, a cubic-time algorithm was presented; one approach developed a recursive procedure based on eliminating cherries from the tree [54], while [167] identified and exploited a structural property of reticulation-visible networks.

Next, suppose that we are given a collection \mathcal{P} of rooted binary phylogenetic X -trees. It is easy to construct a binary network that displays every tree in \mathcal{P} . But when is there a **temporal** network that displays each tree in \mathcal{P} ? This question was explored in [200], where a mathematical characterization was derived. To state this requires us to

define a particular type of ordering of X that selects the leaves according to how they appear in the trees in \mathcal{P} . A *cherry picking sequence* for \mathcal{P} is a total ordering of X (say, x_1, x_2, \dots, x_n) with the following property: for $i = 1, \dots, n-1$, x_i is a leaf of a cherry in the tree $T| \{x_i, x_{i+1}, \dots, x_n\}$ for each $T \in \mathcal{P}$. The following result is from [200] (Theorem 1).

Theorem 10.14. *Let \mathcal{P} be a set of rooted binary phylogenetic X -trees. There exists a temporal network that displays \mathcal{P} if and only if there exists a cherry-picking sequence for \mathcal{P} .*

An interesting question, apparently unresolved, is whether there is a polynomial-time algorithm for deciding whether or not a set \mathcal{P} of rooted binary phylogenetic X -trees is displayed by some temporal network, even for the special case $|\mathcal{P}| = 2$. We will return briefly to cherry picking sequences in Section 10.5.2.

The cluster containment problem. For any network $N = (V, A) \in \mathcal{N}(X)$, a subset C of X is said to be a

- a *hardwired cluster* of N if $C = c_N(v)$ for some $v \in V$ (i.e., C is the set of leaves descended from some vertex of N); or
- a *softwired cluster* if $C = c_T(v)$ for some $v \in V$ and $T \in \mathcal{T}(N)$ (i.e., C is a cluster of some tree T that is displayed by N).

Each hardwired cluster of a network is also a softwired cluster. Moreover, the set of clusters (of either type) of any level-1 binary network form a weak hierarchy (cf. Section 2.3.1), as shown in [147] (Proposition 1).

Clearly, it is easy to determine whether C is a hardwired cluster of N . A more interesting question is the *cluster containment problem*: Given $N \in \mathcal{N}(X)$ and a subset C of X , is C a softwired cluster of N ? Since a binary phylogenetic network with r reticulation vertices can display 2^r binary trees, an exhaustive algorithm will not be polynomial-time in general. In fact, the cluster containment problem is NP-complete for general binary phylogenetic networks [209], and also for tree-sibling networks (even for ones that are regular and have a temporal labeling). However, the cluster containment problem has a polynomial-time solution for reticulation-visible networks ([203], Lemma 6.12.5), as well as for binary level- k networks for fixed k [368].⁷⁷

10.4.2 • Tree-based networks

Informally, a *tree-based network* on X is any binary phylogenetic network on X that can be obtained from some rooted phylogenetic X -tree T by sequentially adding more (“linking”) arcs between the arcs of the tree. It might be suspected that *any* binary network N on X could be constructed in this way. However, this turns out not to be the case (an example was presented in 2013 [365]).

The question then arises as to which networks are tree-based and whether there is some way to quickly recognize whether a network is based on some tree (and, if so, to find a tree on which it can be based). These questions are relevant to the debate alluded to earlier about whether evolution is so complex that it cannot be represented by a “central tendency” tree with some arcs between the edges of the tree to show transfer or hybridization events. First, we formalize the concept of being tree-based more precisely, following [145].

⁷⁷A linear-time algorithm for the cluster containment problem for binary reticulation-visible networks was described recently in [167].

Let $N = (V, A)$ be a binary phylogenetic network on X with root ρ . We say that N is a *tree-based network* if there is a subset A' of arcs for which $T' = (V, A')$ is a rooted tree with the same leaf set as N (note that T' is allowed to have an out-degree 1 root). We will call T' a *subdivision tree* for N , the arcs in $A - A'$ *linking arcs* and the vertices in T' of in-degree and out-degree both equal to 1 the *subdivision points* of T . If we suppress all the subdivision points in T' , we obtain a *base tree* T for N , and we say that N is *based on* T and that the subdivision tree provides an *embedding* of T in N .

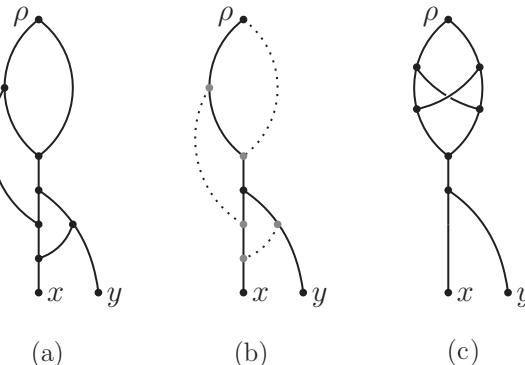


Figure 10.10. (a) A binary phylogenetic network N on two leaves; (b) the unique representation of N as a tree-based network, with the linking arcs dotted; (c) a network that is not tree-based.

Notice that all the vertices of a tree-based network N lie in any subdivision tree T' . In particular, a linking arc is never incident with another linking arc or with a leaf vertex. Figure 10.10 provides a simple example of a tree-based network (a) as well as a network that is not tree-based (c).

Since N is a binary network, if N is based on T then each vertex of T that is not in X has out-degree 2, except perhaps the root, which may have out-degree 1 (this is the case in Fig. 10.10 (b)).

Notice also that if N is based on some tree $T \in RB(X)$, then, clearly N displays T . However, there are tree-based networks N that display a phylogenetic X -tree T but cannot be based on T . For example, the network N in Fig. 10.8(i) is tree-based, with $bc|a$ as a base tree for example (the other two trees are also possible base trees); however, only one of the two possible embeddings of $bc|a$ shown on the top right of the figure provides a valid subdivision tree (the far right one does not). A recent result from [313] characterizes the class of binary networks for which every embedding of a tree provides a base tree for the network as follows.

Proposition 10.15. *If N is a binary phylogenetic network on X , then N is a tree-child network if and only if every embedded phylogenetic X -tree in N is a base tree for N .*

An alternative characterization of being tree-based is the following.

Proposition 10.16. *For $N = (V, A) \in \mathcal{N}_B(X)$, if $T' = (V, A')$ is a rooted tree, then T' is a subdivision tree for N if and only if $A - A'$ is set of pairwise nonadjacent arcs (i.e., no two arcs share a vertex). In particular, N is tree-based if and only if there is a set I of pairwise nonadjacent arcs for which $(V, A - I)$ is a rooted tree.*

Proof. If N is tree-based with a subdivision tree $T' = (V, A')$, then the set $I = A - A'$ of linking arcs is, by definition, pairwise nonadjacent. Conversely, suppose that $T' = (V, A')$ is a rooted tree, with $I = A - A'$ being pairwise nonadjacent. The subset of vertices of T' having out-degree 0 contains X , since $X \subseteq V$. Suppose there is a vertex v of T' of out-degree 0 that is not in X . Then v has in-degree 1, since T' is a rooted tree. Thus v has total degree 1 in T' and total degree 3 in N (since it is not in X and not the root). Hence two of the arcs incident with v in N are in I , which contradicts the assumption that I is pairwise nonadjacent. ■

The class of tree-based networks is rather large; for example, we will see (Corollary 10.18) that it includes all binary networks that are tree-sibling or reticulation-visible. It follows that all binary networks that are tree-child (including the binary level-1 networks) are tree-based. Moreover, any binary network with at most two reticulations is tree-based [206] and so are all binary level-2 networks. However, there are binary level-3 networks that are not tree-based (for example, Fig. 10.10(c)).

Next, we present a simple necessary condition for a network to be tree-based. An *antichain* in a digraph is a set S of vertices for which there is no directed path from any vertex in S to any other vertex in S . A network N is said to satisfy the *antichain-to-leaf property* if for any antichain \mathcal{A} of vertices N , there exist at least $|\mathcal{A}|$ arc-disjoint directed paths from \mathcal{A} to the leaf set of N . The antichain-to-leaf property is a necessary but not sufficient condition for N to be tree-based.

Notice that the necessity of the antichain-to-leaf property for tree-based networks implies that N cannot be tree-based if either (i) N has an antichain that is larger than the number of leaves, or (ii) N has a reticulate vertex with two reticulate parents (e.g., Fig. 10.10(c)). On the other hand, if every reticulation vertex has two tree vertices as parents, then N is tree-based, as we describe shortly in Corollary 10.18.

Exercise⁺: Show that if $N \in \mathcal{N}_B(X)$ is tree-based, then it satisfies the antichain-to-leaf property.

Algorithms. There are simple and fast algorithms to determine whether or not a network $N \in \mathcal{N}_B(X)$ is tree-based and, if so, to find a base tree for N . The first algorithm translates the condition of being tree-based into an instance of 2-SAT (we met 2-SAT earlier, in Section 5.1). For each arc e of N , associate a Boolean variable x_e and then, for certain arcs, assign a clause that consists of a disjunction of at most two Boolean variables or their negation. The recipe for constructing these clauses is

- (i) if $e = (u, v)$ where v has in-degree 1, then x_e is included as a (singleton) clause;
- (ii) if $e = (u, v)$ and $e' = (u', v)$ (i.e., e and e' have the same head vertex), then include the clauses $x_e \vee x_{e'}$ and $\neg x_{e'} \vee \neg x_e$;
- (iii) if $e = (u, v)$ and $e' = (u, v')$ (i.e., e and e' have the same tail vertex), then include the clause $x_e \vee x_{e'}$.

If N is tree-based, then any valid subdivision tree must contain: each arc of type (i), exactly one of each pair of arcs of type (ii), and at least one of each pair of arcs of type (iii). These requirements are encoded by the clauses described. It can be shown that the conjunction of all these clauses has a satisfying assignment if and only if N is tree-based; moreover, in that case, the Boolean variables that are assigned the value “true” in any satisfying assignment for the conjunction of these clauses are the arcs of a subdivision tree for N (which

thereby determines a base tree) [145]. Since 2-SAT has a simple linear-time solution, this provides a fast algorithm.

Characterizing tree-based networks. We first describe a famous result from combinatorics due to Philip Hall (sometimes called the “marriage theorem”). Let $G = (V \cup W, E)$ denote any bipartite graph in which V and W are disjoint sets of vertices, and $E \subseteq \{\{v, w\} : v \in V, w \in W\}$. A *matching* of $G = (V \cup W, E)$ is a subset E' of E consisting of vertex-disjoint edges. A matching E' is said to *cover* V if, for each $v \in V$, there is an edge $e \in E'$ that is incident with v . Hall’s marriage theorem states that $G = (V \cup W, E)$ has a matching that covers V if and only if, for every subset V' of V , the number of elements of W that are adjacent to a vertex in V' is at least $|V'|$.

Given a tree-based network N , let TR_N denote the set of tree vertices that have at least one reticulate child and let R_N denote the set of reticulate vertices of N . Notice that TR_N and R_N are disjoint, so we can form the bipartite graph $G[N] := (R_N \cup TR_N, E)$, where $\{u, v\} \in E$ precisely if $u \in TR_N$, $v \in R_N$ and (u, v) is an arc of N . An example is shown in Figs. 10.11 (b) and (c).

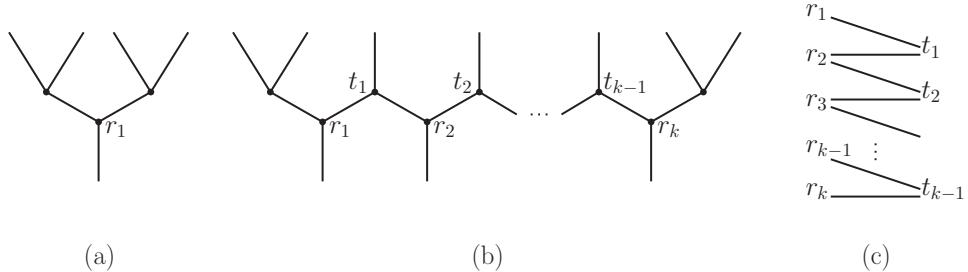


Figure 10.11. (a) A reticulation vertex r_1 with two reticulate parents cannot be present in a tree-based network. (b) Any arrangement N' of the type shown cannot be present in a tree-based network N . If it did, then $G[N]$ would contain the subgraph $G[N'] = (R_{N'} \cup TR_{N'}, E)$ (shown in (c)) as a component, and this has no matching that covers the $R_{N'}$. Here, $R_{N'} = \{r_1, r_2, \dots, r_k\}$, $TR_{N'} = \{t_1, \dots, t_{k-1}\}$, and $k \geq 2$. Note that in (b), there may also be a directed path from the child of r_i to r_j (for $j \neq i$), and the four vertices that comprise the parents of the reticulate parent of r_1 and r_k need not be distinct (also, the reticulate parent of r_k may be in $\{r_1, \dots, r_{k-1}\}$).

The following result is from [385]; related results were obtained by [206] independently.

Theorem 10.17. *Let N be a binary phylogenetic network on X . The following are equivalent.*

- (i) N is tree-based;
- (ii) $G[N]$ has a matching that covers R_N ;
- (iii) $G[N]$ contains no maximal path that starts and ends in R_N ;
- (iv) N contains neither of the configurations shown in Figs. 10.11 (a) or (b).

Proof: (i) \Leftrightarrow (ii): If N is tree-based, then each vertex in R_N has exactly one incoming linking arc and no vertex in TR_N has two outgoing linking arcs, so the set of linking arcs

provide a matching for $G[N]$ that covers R_N . Conversely, if $G[N]$ has a matching E' that covers R_N , then by deleting these arcs from N , we obtain a valid subdivision tree for N , so the arcs of N corresponding to E' can be regarded as linking arcs.

(i) \Rightarrow (iv): If N contains the configuration in Fig. 10.11(a), then N cannot be tree-based, since if it were, then exactly one of the incoming arcs into r_1 (say, (v, r_1)), would need to be a linking arc, in which case v would have out-degree 0 in the subdivision tree, but not be an element of X . Similarly, if N contains the “zig-zag” configuration in part (b) of Fig. 10.11, then if N was tree-based, the reticulate parent of r_1 forces the arcs (t_i, r_i) to be linking arcs for $i = 1, \dots, k - 1$, and which, in turn, forces r_k to have two incoming arcs from the subdivision tree, which is not possible.

(iv) \Rightarrow (iii): Suppose that $G[N]$ has a maximal path of length ℓ that starts and ends in R_N . If $\ell = 0$, then N contains the configuration in Fig. 10.11(a), while if $\ell \geq 2$, N contains the configuration shown in Fig. 10.11(b).

(iii) \Rightarrow (ii): Suppose that $G[N]$ does not have a matching that covers R_N . By Hall’s theorem, there exists a subset R' of R_N with $|R'|$ strictly greater than the number of vertices in TR_N that are adjacent to the vertices in R' . Since the bipartite graph $G[N]$ has maximum vertex degree 2, and so consists of paths and cycles, it follows that there is at least one maximal path in R_N of length $\ell \geq 0$ that starts and ends in R_N . ■

Corollary 10.18. *Let $N \in \mathcal{N}_B(X)$. Then N is tree-based if any one of the following holds:*

- (i) *N is a tree-sibling network.*
- (ii) *For each reticulation vertex of N , both parents are tree vertices.*
- (iii) *N is reticulation-visible.*

Proof. For parts (i) and (ii), observe that the arrangements shown in Figs. 10.11 (a) and (b) cannot arise for such a network N , so N is tree-based by the implication (iv) \Rightarrow (i) in Theorem 10.17. Part (iii) then follows from part (i) by Lemma 10.9. ■

Part (ii) of Corollary 10.18 can be strengthened a little, as shown in [206]. $N \in \mathcal{N}_B(X)$ is tree-based if for every reticulation vertex v of N either (i) both parents of v are tree vertices or (ii) one parent of v is a tree vertex and the sibling of v is a tree vertex. This may also be deduced from the implication (iv) \Rightarrow (i) in Theorem 10.17.

Theorem 10.17 provides further polynomial-time algorithms for determining whether or not a network $N \in \mathcal{N}_B(X)$ is tree-based. Moreover, the characterization that involves matchings in Part (ii) provides a way to identify the linking arcs and so find a base tree. An interesting question is the following: given a network N in $\mathcal{N}_B(X)$ and a tree $T \in RB(X)$ is N based on T ? A recent paper [15] has presented a proof that this question is NP-hard.

For tree-based networks, there is an alternative notion of a temporal labeling that is more appropriate when the linking arcs correspond to transfer events at some fixed point in time (e.g., for the type of lateral gene transfer (LGT) network that we considered in Section 9.3.3 (Fig. 9.9)). If a tree-based network $N = (V, A)$ has a subdivision tree $T' = (V, A')$, then this alternative notion of temporal labeling replaces conditions T1 and T2 for a temporal labeling by

T1': $t(u) = t(v)$ if (u, v) is a linking arc (i.e., $(u, v) \in A - A'$);

T2': $t(u) < t(v)$ if (u, v) is an arc of T' .

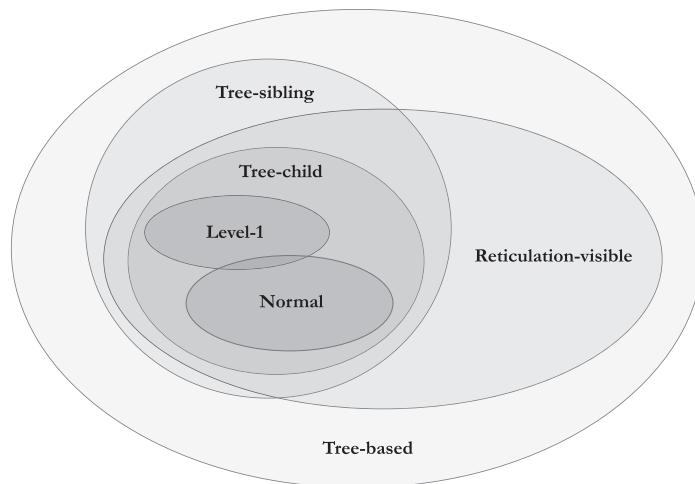


Figure 10.12. Relationships among some network properties within the class of binary phylogenetic networks.

Not every tree-based network has a subdivision tree and an associated labeling map t satisfying T1' and T2' (an example is provided by Fig. 10.7(i)).

In discussing tree-based networks, we have so far assumed that N is binary. For nonbinary networks there are various possible ways to extend the definition of tree-based, and two natural ways to do so were recently described and studied in [206]. These authors show that, under either definition, there is a polynomial-time algorithm for determining whether or not a nonbinary network is tree-based (based on matchings, though in a different bipartite graph to the characterization described above). For details, see [206].

The containment relationships among the main classes of binary phylogenetic networks that we have discussed so far is summarized in Fig. 10.12.

Universal tree-based networks. We have seen that requiring a binary phylogenetic X -tree T to be a base tree for a network is a stronger property than requiring the network to merely display T . It is not hard to construct a binary phylogenetic network N_X that not only displays all binary phylogenetic X -trees, but is also tree-based and has just a cubic number of arcs [145]. A more detailed construction leads to a binary phylogenetic network U_X that not only displays every phylogenetic X -tree, but also has every phylogenetic X -tree as a base tree (this “universal” network is thus also tree-based). For details on two independent constructions for U_X , see [180] or [385].

10.5 • Reconstructing networks

In earlier chapters, we saw how phylogenetic trees are uniquely determined by their basic substructures (e.g., rooted triples or quartets) and other information on small numbers of leaves (e.g., pairwise distances when the arcs have lengths). For phylogenetic networks, analogous results, when they exist, usually require quite stringent restrictions on the class of networks.

10.5.1 • Encoding networks by substructures

In Chapter 2, we saw that a rooted phylogeny is determined by the set of rooted triples that it displays. Thus a natural question is whether or not some analogous result holds for networks. There are several ways to make this question more precise. Earlier, we defined what it means for a network N on X to display a phylogenetic X -tree. Recalling that $\mathcal{T}(N)$ is the set of phylogenetic X -trees that are displayed by N , a natural question is the following: If $\mathcal{T}(N) = \mathcal{T}(N')$, does this imply that N and N' are equivalent?

It is clear that the answer is no, since $\mathcal{T}(N) = \mathcal{T}(\tilde{N})$, so if N is not a compressed network, its compression gives a different network that displays the same set of trees. However, $\mathcal{T}(N)$ can equal $\mathcal{T}(N')$ in other ways, as Fig. 10.13 shows.

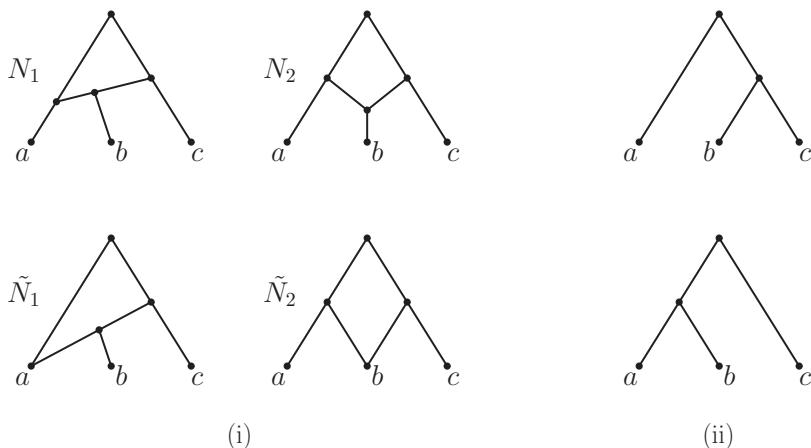


Figure 10.13. (i) Four different level-1 phylogenetic networks on $X = \{a, b, c\}$ that all display the same set of two phylogenetic X -trees shown in (ii). The networks at the bottom in (i) are compressions of the corresponding networks at the top. Networks N_1 and \tilde{N}_1 are normal, but N_2 and \tilde{N}_2 are not, due to the presence of a redundant arc. Notice that N_1 and N_2 are binary level-1 networks that have a block consisting of four vertices (there is a further such network, not shown, that also displays the same two trees). The only regular network in this figure is \tilde{N}_2 .

Figure 10.13 suggests that for regular networks, there might be some hope that N can be determined by the trees it displays. This turns out to be the case, by the following result from [375].

Theorem 10.19. *If N and N' are regular networks and $\mathcal{T}(N) = \mathcal{T}(N')$, then $N \cong N'$. Moreover, there is a polynomial-time algorithm for reconstructing N from $\mathcal{T}(N)$.*

This theorem also applies to compressed normal networks, since such networks are regular, as noted earlier.

Another class \mathcal{N} of networks, for which $N, N' \in \mathcal{N}$ and $\mathcal{T}(N) = \mathcal{T}(N') \Rightarrow N \cong N'$ is the class of binary level-1 networks that do not contain a block consisting of three or four vertices ([147], Theorem 1). Note that for a binary level-1 network a block of size 3 or 4 corresponds, as an unrooted subgraph, to a 3-cycle or a 4-cycle (respectively). The requirement that blocks of size 4 be excluded is clear from the fact that N_1 and N_2 in Fig. 10.13 displays the same set of trees, and similar but simpler examples (involving just two leaves) exclude networks with blocks of size 3.

Given a set \mathcal{P} of rooted binary phylogenetic X -trees, as we have seen, there is always a network $N \in \mathcal{N}_B(X)$ that displays each tree in \mathcal{P} . However, we can ask whether there is a network of a certain type that displays \mathcal{P} , and if so, whether we can find one that minimizes the number r of reticulation vertices. Let us consider the simplest case where \mathcal{P} consists just of rooted triples. The set \mathcal{P} is said to be *dense* if, for each three elements $x, y, z \in X$, at least one of the rooted triples $xy|z, xz|y, yz|x$ is in \mathcal{P} . The following result is from [366].

Theorem 10.20. *For $k = 1$ and $k = 2$, there is a polynomial-time algorithm to determine whether, for a dense set \mathcal{P} of rooted triples, there is a level- k binary network N that displays each tree in \mathcal{P} , and, if so, constructs such a network N with the smallest number of reticulation vertices.*

If the set \mathcal{P} of rooted triples on X is not dense, then for each $k \geq 1$, it is NP-hard to decide whether there exists a level- k binary network N on X that displays each tree in \mathcal{P} (for details, see [366] and the references therein).

In the unrooted setting, given a dense set of quartet trees on X , one can also ask if there is an unrooted level-1 network that displays each quartet tree. Using a novel approach based on techniques from linear algebra, a polynomial-time algorithm for this problem was recently described in [211].

Since phylogenetic networks cannot, in general, be uniquely reconstructed from the rooted trees they display, it is natural to ask whether the induced subnetworks of a given size suffice. First, we need to explain how a network $N = (V, A)$ on X induces a phylogenetic network on a nonempty subset Y of X . This notion for a rooted phylogenetic tree T was defined in Chapter 4 and denoted $T|Y$, so we will use the notation $N|Y$ here. The *lowest stable ancestor* of Y in N , denoted $\text{lsa}_N(Y)$, is the vertex $w \in V - X$ that (i) lies on all directed paths from the root of N to the elements of Y , and (ii) is maximal under \preceq_N with this property (the fact that this vertex is well defined is left as an exercise). The phylogenetic network $N|Y$ on Y is obtained from N by (i) deleting all vertices of N (and their incident arcs) that are not on a directed path from $\text{lsa}_N(Y)$ to some element of Y , then (ii) suppressing all vertices of in-degree and out-degree both equal to 1, and deleting any parallel arc(s). By repeating step (ii) as necessary, we eventually obtain a phylogenetic network $N|Y$ on Y .

Since rooted phylogenetic trees are determined by the subtrees induced by subsets of three leaves (i.e., rooted triples), it seems reasonable to consider subnetworks induced by subsets of three leaves. Given a network $N \in \mathcal{N}(X)$ and a subset Y of X of size 3, a *trinet* of N is the rooted phylogenetic network $N|Y$ on Y . It turns out that certain classes of networks are uniquely encoded by their induced trinets (level-1, level-2, and tree-child [367]); however, for general binary networks, this no longer holds.

This suggests that subnetworks induced by more than three leaves might need to be considered, even if one wishes to encode just the class of binary phylogenetic networks. A remarkable and somewhat surprising recent result from [195] (Theorem 2), demonstrates that such attempts are hopeless in general.

Theorem 10.21. *For all $n \geq 3$, there is a pair of nonequivalent binary phylogenetic networks N_1, N_2 on $[n]$, for which $N_1|Y \cong N_2|Y$ for all proper nonempty subsets Y of $[n]$.*

A quite different approach to the identifiability question for phylogenetic networks was developed in a recent paper by Fabio Pardi and Celine Scornavacca [285]. They considered networks in which each arc has a positive length, so any tree T displayed by N

comes equipped with a length for each arc e of the tree (corresponding to the sum of the lengths of the arcs of the network on the path that corresponds to e). It is still the case that different networks (with appropriate arc length assignments) can lead to equivalent displayed trees that have identical arc lengths.

Indeed, the equivalence $\mathcal{T}(N) = \mathcal{T}(\tilde{N})$ extends to an equivalence of induced trees with branch lengths if, for each collapsed arc, the length of the arc is added to each of its parent arcs. In fact, as shown in [285], there are further arc contraction and deletion operations that can transform any phylogenetic network $N \in \mathcal{N}(X)$ into a “canonical form” N' that has the properties that (i) N' displays the same set of trees (and with the same arc lengths) as N does, and (ii) N' has no vertex of in-degree greater than 0 and out-degree 1. Note that N' may contain parallel arcs (with different lengths).

Suppose we now impose the following constraint on the lengths of the arcs, called the *no equally long paths* (NELP) property: For any two distinct paths with the same endpoints, the sum of the lengths of the arcs along the two paths is different. This condition may seem slightly odd, since, in the language of Chapter 6, it appears to be a sort of anti-ultrametric property. However, if we think of arc lengths as estimating evolutionary distance (rather than time), the assumption is less objectionable and might be expected to hold generically. This leads to the following remarkable result from [285]: If N has the NELP property, then there is a **unique** canonical form for N among all networks satisfying this property. This, in turn, implies that two networks satisfying the NELP property have the same unique canonical form (up to equivalence) if and only if the networks display the same set of trees (with their induced arc lengths). For further details, see [285].

Paths and distances. Another approach to encoding networks is via the distribution of path lengths in the network. Suppose that $N = (V, A)$ is a phylogenetic network on $[n]$. For each interior vertex v of N let $p_i(v)$ be the number of paths from v to leaf i and let $p(v)$ be the n -tuple $(p_1(v), \dots, p_n(v))$.

For tree-child networks (not necessarily binary), [74] (Theorem 1) showed that the multiset $\{p(v) : v \in V\}$ uniquely determines N up to equivalence (within the class of tree-child networks) and that N can be reconstructed from this multiset in polynomial time.

A recent related result from [55] considers the conditions under which a network is determined by path lengths between pairs of elements from X . Specifically, for $x, y \in X$ with $x \neq y$, an *up-down path* from x to y in N is an (undirected) path from x to y that heads up (i.e., towards the root) from x to some vertex, then back down to y .⁷⁸ For each pair $x, y \in X$, let S_{xy} be the set of lengths of all the up-down paths from x to y in N , and let $S(N) := (S_{xy} : \{x, y\} \in \binom{X}{2})$ be the collection of these sets for all pairs of leaves of X .

What does this tell us about N ? Clearly, if N is a phylogenetic X -tree, then $S(N)$ determines N up to equivalence (in that case, S_{xy} consists of a single element, the length of the unique path between x and y in the tree). However, it is easy to construct a pair of inequivalent networks N and N' for which $S(N) = S(N')$. One result [55] shows that for a certain class of networks, $S(N)$ suffices to determine N .

⁷⁸More precisely, an up-down path from x to y is a sequence of distinct vertices $x = v_0, v_1, \dots, v_i, v_{i+1}, \dots, v_k = y$ ($i \leq k - 1$) where (v_j, v_{j-1}) are arcs of N for $j = 1, \dots, i$ and (v_j, v_{j+1}) are arcs of N for $j = i, \dots, k - 1$. The length of this path is k .

Proposition 10.22. *For any temporal binary tree-child network N on X , N is the unique (up to equivalence) binary phylogenetic network on X that realizes $S(N)$; moreover, N can be reconstructed from $S(N)$ in polynomial time in $|X|$.*

Reconstructing networks from characters. In Chapter 5, we saw how to assign a parsimony score $\text{ps}(f, T)$ to any character f on X and any phylogenetic tree $T \in P(X)$. This, in turn, can be used to reconstruct an optimal tree from a sequence of characters (a “most parsimonious tree”). How should we define the parsimony score of a character on a network N ? One simple option is to consider an extension of f to V that minimizes the number of arcs with different states at the endpoints. For a 2-state character, the computation of this so-called “hardwired” parsimony score can be done in polynomial time (based on network flow techniques); however, in general, it is hard for r -state characters for $r > 2$ [142]. Nevertheless, since characters are generally assumed to have evolved on a tree that is embedded within the network, a more biologically reasonable definition of the parsimony score of character f on network N is the “softwired” definition:

$$\text{ps}(f, N) = \min\{\text{ps}(f, T) : T \in \mathcal{T}(N)\}.$$

The parsimony score for a network given a sequence $\mathcal{C} = (f_1, \dots, f_k)$ of characters is then given by $\sum_{i=1}^k \text{ps}(f_i, N)$. Computing $\text{ps}(f, N)$ is NP-hard, even when f is binary and N is a binary tree-child network with a temporal labeling. On the positive side, the computation of $\text{ps}(f, N)$ for a r -state character when N is a level- k network is fixed parameter tractable in k (for details, references and further results on approximating the hardwired and softwired parsimony score, see [142]).

For the stochastic models considered in Chapters 7 and 8, the probability of generating character f on a network N with arc lengths can be computed as follows. First, sample a tree $T \in \mathcal{T}(N)$ according to some probability distribution $\mathbb{P}(T|N)$. For example, we might assume that for each reticulation vertex, just one of the incoming arcs is chosen with some probability and these events are treated independently across the reticulation vertices of N . For the resulting randomly sampled phylogenetic tree T with its induced edge lengths, l_T , we described in Chapter 7 how to compute the probability $\mathbb{P}(f|T, l_T)$ of generating f on the pair (T, l_T) under some Markov process (or mixture). Assuming that the sequence $\mathcal{C} = (f_1, \dots, f_k)$ is generated independently in this fashion, the probability of \mathcal{C} from N is given by

$$\mathbb{P}(\mathcal{C}|N) = \prod_{i=1}^k \left(\sum_{T \in \mathcal{T}(N)} \mathbb{P}(f_i|T, l_T) \mathbb{P}(T|N) \right).$$

In this way, maximum likelihood approaches can be developed for phylogenetic networks (for details, see [274, 380]).

10.5.2 ■ Minimizing reticulation

Given a phylogenetic network N on X , its *reticulation number*, denoted $r(N)$, is defined by

$$r(N) = \sum_{v \neq \rho} (d^-(v) - 1),$$

where $d^-(v)$ is the in-degree of v . Thus if all reticulation vertices have in-degree 2 (e.g., for binary phylogenetic networks), then $r(N)$ simply counts the number of reticulation

vertices. Notice also that if \overline{N} is the graph obtained from N by ignoring the direction of the arcs, then $r(N) = \text{cy}(\overline{N})$, the cyclomatic number of \overline{N} .

For a set \mathcal{P} of rooted binary phylogenetic X -trees, let $r(\mathcal{P})$, denote the minimum reticulation number of any binary network N that displays all the trees in \mathcal{P} . In other words,

$$r(\mathcal{P}) = \min\{r(N) : N \in \mathcal{N}_B(X) \text{ and } N \text{ displays } \mathcal{P}\}.$$

When \mathcal{P} consists of just two trees (T and T' , say) we will denote $r(\mathcal{P})$ by $r(T, T')$.

Computing $r(T, T')$ for an arbitrary pair of trees $T, T' \in RB(X)$ turns out to be an APX-hard problem [50]. This means that the problem has no polynomial-time approximation scheme (PTAS) unless P = NP (in other words, there is some constant c , usually very small, such that the problem cannot be approximated better than a factor c).

In the special case where $d_{rSPR}(T, T') = 1$ (i.e., T' is obtained from T by a single rooted SPR operation), we have $r(T, T') = 1$; moreover, the converse is true: $d_{rSPR}(T, T') = 1 \iff r(T, T') = 1$. In general, the following inequality (from [23]) holds:

$$d_{rSPR}(T, T') \leq r(T, T'). \quad (10.5)$$

Exercise: Suppose T and T' are rooted binary phylogenetic X -trees, and that f is a character that is homoplasy-free on T . Show that $r(T, T')$ is at least as large as the homoplasy score of f on T' . [Hint: Combine eqn. (10.5) with a (modification of) an inequality from Chapter 5.]

Inequality (10.5) follows because, as mentioned in Section 2.5, $d_{rSPR}(T, T')$ can be expressed in terms of the size of a maximal agreement forest of rooted trees [49]; $r(T, T')$ has a similar expression but now an agreement forest is required to satisfy a further (“acyclic”) condition [23, 50]. Using this connection, a simple upper bound on $r(T, T')$ across pairs of trees is $r(T, T') \leq n - 2$ and this bound is achieved for certain pairs of trees [23].

How much larger than $d_{rSPR}(T, T')$ can $r(T, T')$ be? Quite a bit: For each even integer $n \geq 4$, it is possible to find a pair of caterpillar trees $T, T' \in B(n)$ for which $d_{rSPR}(T, T') = 2$ and yet for which $r(T, T') = n/2$. The following remarkably precise result is from [198] and again exploits the connection between the rSPR metric and agreement forests of rooted trees.

Proposition 10.23. *Let $T, T' \in RB(n)$, where $n \geq 4$. Then*

- (i) $\frac{r(T, T')}{d_{rSPR}(T, T')} \leq \frac{1}{2} \lfloor \frac{n}{2} \rfloor$;
- (ii) $r(T, T') - d_{rSPR}(T, T') \leq n - \lceil 2\sqrt{n} \rceil$.

Moreover, the inequalities in (i) and (ii) are sharp for all values of $n \geq 4$.

We can also restrict the class of networks that display a set of trees. One restriction that is both biologically reasonable and mathematically convenient is to consider just temporal phylogenetic networks (cf. Section 10.3.3). Given a set \mathcal{P} of rooted binary phylogenetic X -trees, let

$$r_t(\mathcal{P}) = \min\{r(N) : N \text{ is a temporal network on } X \text{ that displays } \mathcal{P}\}.$$

In [200] and [199], the authors studied the following optimization problem: Given a set \mathcal{P} of rooted binary phylogenetic X -trees and a positive integer k , does there exist a

temporal network on X that displays \mathcal{P} and has $r(N) \leq k$? This problem turns out to be NP-hard (indeed, APX-hard) for two trees but, again, there is a connection with agreement forests, namely that $r_t(T, T') = |\mathcal{F}| - 1$ where \mathcal{F} is a so-called “temporal agreement forest” for T and T' (for details, see [199]). Moreover, the notion of a cherry-picking sequence (which played a key role in Theorem 10.14) provides a way to characterize $r_t(\mathcal{P})$, and that applies when \mathcal{P} has two or more trees. Using a simple way to score any cherry-picking sequence on \mathcal{P} , the second main result from [201] is that when \mathcal{P} is displayed by some temporal network, then a cherry-picking sequence with the minimal score gives a temporal network that has optimal reticulation score $r_t(\mathcal{P})$ (for details, see [201]). This result is of particular importance since there is currently no nice connection known between the minimal reticulation problem for \mathcal{P} and temporal agreement forests when $|\mathcal{P}| > 2$.

Exercise⁺: Suppose $T \in RB(n)$ and that \mathcal{T} is a random tree drawn from $RB(n)$ according to the uniform distribution. Show that for any constant $c \in (0, \frac{1}{2})$, the probability that $r(T, \mathcal{T}) > cn$ converges to 1 as $n \rightarrow \infty$. [Hint: Combine eqn. (10.5) with a simple upper bound estimate on the number of trees that are within rSPR distance cn from T .]

10.6 • Additional topics

We end this chapter with a brief discussion of some other aspects and varieties of phylogenetic networks. The study of phylogenetic networks is a very broad and fast-developing field, and there are many topics that are beyond the scope of this chapter. These include co-phylogeny networks, haplotype networks and ancestral recombination graphs (for the last topic, see [182] and [168]).

Comparing phylogenetic networks. Formulating a well-defined metric on phylogenetic networks that can be computed in polynomial time is much less straightforward than it is for phylogenetic trees. In general, the class of networks being considered needs to be restricted. For regular networks on X there is a direct generalization of the RF-metric: $d(N_1, N_2) = |\mathcal{C}(N_1) \Delta \mathcal{C}(N_2)|$ where $\mathcal{C}(N)$ is the set of (hardwired) clusters of N .

However, even for the seemingly well-behaved class of tree-sibling networks that have a temporal labeling (a class sometimes referred to as *tree-sibling time-consistent* (TSTC)), deciding whether two such networks are equivalent is (formally) as hard as the graph isomorphism problem [73]. Since this latter problem is believed not to have a polynomial-time solution, defining a metric on this class seems a hopeless task. If one restricts the class TSTC slightly by requiring the in-degree of each reticulate vertex to be 2,⁷⁹ then there are two well-defined metrics on this class of networks, and these can be readily computed in polynomial time; for details, see [73] and the references therein.

In Chapter 2, as well as metrics on phylogenies we also considered “spaces of trees” based on tree rearrangement operations (NNI, SPR, TBR). For (undirected) phylogenetic networks, recent work [192] has provided a similar way to define and study a “space of phylogenetic networks” based on local operations modeled on NNI. An analogue of the continuous BHV space of trees (described in Chapter 6) for circular split networks was also defined and studied recently in [108].

⁷⁹Networks with this in-degree property are sometimes called *semibinary networks*.

Closure and extension properties of networks. Let us say that a class \mathcal{N} of phylogenetic networks is *closed under sampling*, provided that whenever N is a phylogenetic network on X with $N \in \mathcal{N}$, and Y is a subset of X , then $N|Y \in \mathcal{N}$. The motivation for this concept is that we may not have considered all of the species involved in the reticulate evolution of the group of species under study (e.g., the species may have become extinct or not yet been sampled). Accordingly, we would like to know if a particular property of a network is robust to incomplete species coverage.

A natural question is which of the classes of networks that we have considered are closed under sampling. It is clear that tree-child networks and reticulation-visible networks are not. Neither is the class of tree-based networks (for example, consider $N|\{\alpha, \gamma, \epsilon\}$ for the network N in Fig. 10.4) nor the class of temporal networks. However, the class of level- k networks is closed under sampling for every k .

A dual notion to closure under sampling is the following extendability property for a class \mathcal{N} of networks. For any phylogenetic network N on X , either $N \in \mathcal{N}$ or there exists a set X' containing X and a network N' on X' with $N' \in \mathcal{N}$ so that $N'|X = N$. It is easily seen, for example, that the class of tree-child networks and the class of tree-based networks both satisfy this property. In other words, a tree-based network and a tree-child network can be generated from any network by attaching additional leaves in appropriate places. Thus, if a phylogenetic network that describes the evolution of some subgroup of species fails to be tree-based (or tree-child) it is still possible that the larger phylogenetic network for the entire group of species is tree-based (or tree-child).

Pedigrees. A pedigree represents the history of a sexually reproducing population. Formally, a *pedigree* P on X is an acyclic digraph having a vertex set that consists of two disjoint subsets M and F (male, female), with X being the set of vertices of out-degree 0, and every vertex of P either having zero in-degree (a “founder”) or having two incoming arcs, one from a vertex in M and one from a vertex in F . Technically, a pedigree is not a phylogenetic network, since there is no requirement to have a single root vertex; however, several concepts concerning phylogenetic networks can be applied to study pedigrees.

When M and F are not specified, it is possible to characterize and easily check whether a directed acyclic graph $G = (V, A)$ is a pedigree. It can be shown that G is a pedigree if and only if G has no vertices of in-degree 1 and the associated “parent” graph $M(G) = (V, E)$ is bipartite, where E is the set of pairs $\{u, v\}$ in V for which some vertex $w \in V$ exists with $(u, w), (v, w)$ both in A .

A remarkable result of a similar flavor to (but earlier and different from) Theorem 10.21 was derived in [356]: for all $n \geq 4$, there exist two finite nonisomorphic pedigrees on $[n]$ that induce isomorphic subpedigrees on all proper subsets of $[n]$ (for a precise statement and details, see [356]).

Exercise: If P is a finite pedigree, let \overline{P} be the graph obtained from P by ignoring the direction of the arcs. Show that

- \overline{P} is 3-partite.
- Any graph G obtained from \overline{P} by adding edges between M and F vertices is 4-partite. [Hint: What can one say about the subgraph of \overline{P} induced by the M (or F) vertices?]

Bibliography

- [1] A. V. AHO, Y. SAGIV, T. G. SZYMANSKI, AND J. D. ULLMAN, *Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions*, SIAM. J. Comput., 10 (1981), pp. 405–421. (Cited on p. 68)
- [2] D. ALDOUS, *Probability distributions on cladograms*, in Random Discrete Structures, The IMA Volumes in Mathematics and its Applications, Vol. 76, D. Aldous and R. Pemantle, eds., Springer, New York, 1996, pp. 1–18. (Cited on pp. 53, 55, 219)
- [3] D. J. ALDOUS, *Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today*, Stat. Sci., 16 (2001), pp. 23–34. (Cited on pp. 50, 57)
- [4] D. ALDOUS AND L. POPOVIC, *A critical branching process model for biodiversity*, Adv. Appl. Probab., 37 (2005), pp. 1094–1115. (Cited on p. 217)
- [5] D. J. ALDOUS, M. A. KRIKUN, AND L. POPOVIC, *Five statistical questions about the tree of life*, Syst. Biol., 60 (2011), pp. 318–328. (Cited on p. 220)
- [6] B. L. ALLEN AND M. STEEL, *Subtree transfer operations and their induced metrics on evolutionary trees*, Ann. Combin., 5 (2001), pp. 1–15. (Cited on pp. 30, 31)
- [7] E. S. ALLMAN AND J. A. RHODES, *Phylogenetic ideals and varieties for the general Markov model*, Adv. Appl. Math., 40 (2008), pp. 127–148. (Cited on pp. 191, 196)
- [8] E. S. ALLMAN AND J. A. RHODES, *Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites*, Math. Biosci., 211 (2008), pp. 18–33. (Cited on p. 197)
- [9] E. S. ALLMAN AND J. A. RHODES, *The identifiability of covarion models in phylogenetics*, IEEE/ACM Trans. Comput. Biol. Bioinf., 6 (2009), pp. 76–88. (Cited on p. 197)
- [10] E. S. ALLMAN, J. H. DEGNAN, AND J. A. RHODES, *Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent*, J. Math. Biol., 62 (2011), pp. 833–862. (Cited on pp. 191, 224, 228)
- [11] E. S. ALLMAN, J. H. DEGNAN, AND J. A. RHODES, *Determining species tree topologies from clade probabilities under the coalescent*, J. Theor. Biol., 289 (2011), pp. 96–106. (Cited on p. 228)
- [12] E. S. ALLMAN, J. A. RHODES, AND S. SULLIVANT, *When do phylogenetic mixture models mimic other phylogenetic models?*, Syst. Biol., 61 (2012), pp. 1049–1059. (Cited on p. 197)
- [13] N. ALON, H. NAVES, AND B. SUDAKOV, *On the maximum quartet distance between phylogenetic trees*, SIAM J. Discrete Math., 30 (2016), pp. 718–735. (Cited on p. 81)
- [14] N. ALON, S. SNIR, AND R. YUSTER, *On the compatibility of quartet trees*, SIAM J. Discrete Math., 28 (2014), pp. 1493–1507. (Cited on p. 81)

- [15] M. ANAYA *et al.*, *On determining if tree-based networks contain fixed trees*, Bull. Math. Biol., 78 (2016), pp. 961–969. (Cited on p. 260)
- [16] C. ARNOLD AND P. F. STADLER, *Polynomial algorithms for the maximal pairing problem: Efficient phylogenetic targeting on arbitrary trees*, Alg. Mol. Biol. 5 (2010), pp. 25. (Cited on p. 166)
- [17] R. ATKINS AND C. McDIARMID, *Extremal distances for subtree transfer operations in binary trees*, arXiv:1509.00669v1 [math.CO], (2015). (Cited on pp. 32, 34, 35)
- [18] K. ATTESON, *The performance of neighbor-joining methods of phylogenetic reconstruction*, Algorithmica, 25 (1999), pp. 251–278. (Cited on p. 124)
- [19] V. BAFNA, D. GUSFIELD, G. LANCIA, AND S. YOOSEPH, *Haplotyping as perfect phylogeny: A direct approach*, J. Comput. Biol., 10 (2003), pp. 323–340. (Cited on p. 99)
- [20] H.-J. BANDELT AND A. DRESS, *Reconstructing the shape of a tree from observed dissimilarity data*, Adv. Appl. Math., 7 (1986), pp. 309–343. (Cited on pp. 27, 81)
- [21] H.-J. BANDELT AND M. FISCHER, *Perfectly misleading distances from ternary characters*, Syst. Biol., 57 (2008), pp. 540–543. (Cited on p. 119)
- [22] H.-J. BANDELT AND A. RÖHL, *Quasi-median bulls in Hamming space are Steiner bulls*, Discrete Appl. Math., 157 (2009), pp. 227–233. (Cited on p. 244)
- [23] M. BARONI, S. GRÜNEWALD, V. MOULTON, AND C. SEMPLE, *Bounding the number of hybridisation events for a consistent evolutionary history*, J. Math. Biol., 51 (2005), pp. 171–182. (Cited on p. 266)
- [24] M. BARONI, C. SEMPLE, AND M. STEEL, *Hybrids in real time*, Syst. Biol., 55 (2006), pp. 46–56. (Cited on p. 250)
- [25] A. BERGERON, J. MIXTACKI, AND J. STOYE, *A unifying view of genome rearrangements*, in Algorithms in Bioinformatics (Proceedings of WABI 2006), Lecture Notes in Computer Science, Vol. 4175, P. Bücher and B. M. E. Moret, eds., Springer, Berlin/Heidelberg, 2006, pp. 163–173. (Cited on p. 121)
- [26] D. I. BERNSTEIN, L. S. T. HO, C. LONG, M. STEEL, K. ST. JOHN, AND S. SULLIVANT, *Bounds on the expected size of the maximum agreement subtree*, SIAM J. Discrete Math., 29 (2015), pp. 2065–2074. (Cited on p. 80)
- [27] V. BERRY, O. R. P. BININDA-EMONDS, AND C. SEMPLE, *Amalgamating source trees with different taxonomic levels*, Syst. Biol., 62 (2013), pp. 231–249. (Cited on p. 69)
- [28] S. BHATIA, A. EGRI-NAGY, AND A. R. FRANCIS, *Algebraic double cut and join: A group-theoretic approach to the operator on multichromosomal genomes*, J. Math. Biol., 71 (2015), pp. 1149–1178. (Cited on p. 121)
- [29] L. J. BILLERA, S. P. HOLMES, AND K. VOGTMANN, *Geometry of the space of phylogenetic trees*, Adv. Appl. Math., 27 (2001), pp. 733–767. (Cited on p. 131)
- [30] S. C. BILLEY, M. KONVALINKA, AND F. A. MATSEN IV, *On the enumeration of tanglegrams and tangled chains*, arXiv:1507.04976v1 [math.CO]. (Cited on p. 43)
- [31] M. G. B. BLUM AND O. FRANÇOIS, *Minimal clade size and external branch length under the neutral coalescent*, Adv. Appl. Probab., 37 (2005), pp. 647–662. (Cited on p. 51)
- [32] M. G. B. BLUM AND O. FRANÇOIS, *On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited*, Math. Biosci., 195 (2005), pp. 141–153. (Cited on pp. 55, 60)

- [33] M. G. B. BLUM AND O. FRANÇOIS, *Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance*, Syst. Biol., 55 (2006), pp. 685–691. (Cited on pp. 50, 57)
- [34] M. G. B. BLUM, O. FRANÇOIS, AND S. JANSON, *The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance*, Ann. Appl. Probab., 16 (2006) pp. 2195–2214. (Cited on p. 55)
- [35] S. BÖCKER, *From subtrees to supertrees*, PhD thesis, Fakultät für Mathematik, Universität Bielefeld, Germany, 1999. (Cited on pp. 73, 76)
- [36] S. BÖCKER, *Exponentially many supertrees*, Appl. Math. Lett. 15 (2002), 861–865. (Cited on p. 67)
- [37] S. BÖCKER AND A. W. M. DRESS, *Recovering symbolically dated, rooted trees from symbolic ultrametrics*, Adv. Math., 138 (1998), pp. 105–125. (Cited on pp. 22, 117, 118)
- [38] S. BÖCKER, A. W. M. DRESS, AND M. A. STEEL, *Patching up X-trees*, Ann. Combin., 3 (1999), pp. 1–12. (Cited on p. 77)
- [39] D. BOGDANOWICZ AND K. GIARO, *Matching split distance for unrooted binary phylogenetic trees*, IEEE/ACM Trans. Comput. Biol. Bioinf., 9 (2012), pp. 150–160. (Cited on p. 26)
- [40] J. A. BONDY AND U. S. R. MURTY, *Graph Theory*, Graduate Texts in Mathematics, Vol. 244, Springer, London, 2008. (Cited on pp. 7, 8)
- [41] M. L. BONET, S. LINZ, AND K. ST. JOHN, *Complexity of finding multiple solutions to betweenness and quartet compatibility*, IEEE/ACM Trans. Comput. Biol. Bioinf., 9 (2012), pp. 273–285. (Cited on p. 77)
- [42] P. BONIZZONI, *A linear-time algorithm for the perfect phylogeny haplotype problem*, Algorithmica 48 (2007), pp. 267–285. (Cited on p. 99)
- [43] P. BONIZZONI, C. BRAGHIN, R. DONDI, AND G. TRUCCO, *The binary perfect phylogeny with persistent characters*, Theor. Comput. Sci., 454 (2012), pp. 51–63. (Cited on p. 98)
- [44] P. BONIZZONI, A. P. CARRIERI, G. DELLA VEDOVA, R. DONDI, AND T. M. PRZYTYCKA, *When and how the perfect phylogeny model explains evolution*, in Discrete and Topological Models in Molecular Biology, Natural Computing Series, N. Jonoska and M. Saito, eds., Springer, Berlin/Heidelberg, 2014, pp. 67–83. (Cited on p. 98)
- [45] P. BONIZZONI, A. P. CARRIERI, G. DELLA VEDOVA, AND G. TRUCCO, *Explaining evolution via constrained persistent perfect phylogeny*, BMC Genomics, 15 (Suppl 6), (2014), S10. (Cited on p. 98)
- [46] M. BORDEWICH, *The complexity of counting and randomised approximation*, Ph.D thesis, University of Oxford, UK, 2003. (Cited on p. 69)
- [47] M. BORDEWICH, O. GASCUEL, K. T. HUBER, AND V. MOULTON, *Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic Inference*, IEEE/ACM Trans. Comput. Biol. Bioinf., 6 (2009), pp. 110–117. (Cited on pp. 34, 80, 127)
- [48] M. BORDEWICH, A. G. RODRIGO, AND C. SEMPLE, *Selecting taxa to save or sequence: Desirable criteria and a greedy solution*, Syst. Biol., 57 (2008), pp. 825–834. (Cited on pp. 135, 136, 137)
- [49] M. BORDEWICH AND C. SEMPLE, *On the computational complexity of the rooted subtree prune and regraft distance*, Ann. Combin., 8 (2004), pp. 409–423. (Cited on pp. 32, 266)

- [50] M. BORDEWICH AND C. SEMPLE, *Computing the minimum number of hybridization events for a consistent evolutionary history*, Discrete Appl. Math., 155 (2007), pp. 914–928. (Cited on p. 266)
- [51] M. BORDEWICH AND C. SEMPLE, *Nature reserve selection problem: A tight approximation algorithm*, IEEE/ACM Trans. Comput. Biol. Bioinf., 5 (2008), pp. 275–280. (Cited on p. 138)
- [52] M. BORDEWICH AND C. SEMPLE, *Budgeted Nature Reserve Selection with diversity feature loss and arbitrary split systems*, J. Math. Biol., 64 (2012), pp. 69–85. (Cited on pp. 136, 138)
- [53] M. BORDEWICH AND C. SEMPLE, *Defining a phylogenetic tree with the minimum number of r -state characters*, SIAM J. Discrete Math., 29 (2015), pp. 835–853. (Cited on pp. 93, 201)
- [54] M. BORDEWICH AND C. SEMPLE, *Reticulation-visible networks*, arXiv:1508.05424v1 [math.CO] (2015). (Cited on pp. 249, 250, 255)
- [55] M. BORDEWICH AND C. SEMPLE, *Determining phylogenetic networks from inter-taxa distances*, J. Math. Biol., 73 (2016), pp. 283–303. (Cited on p. 264)
- [56] M. BORDEWICH, C. SEMPLE, AND A. SPILLNER, *Optimizing phylogenetic diversity across two trees*, Appl. Math. Lett., 22 (2009), pp. 638–641. (Cited on p. 136)
- [57] M. BORDEWICH, C. SEMPLE, AND M. STEEL, *Identifying X -trees with few characters*, Electron. J. Combin., 13 (2006), #R83. (Cited on p. 94)
- [58] A. BOUCHARD-CÔTÉ AND M. I. JORDAN, *Evolutionary inference via the Poisson Indel process*, Proc. Natl. Acad. Sci. USA, 110 (2013), pp. 1160–1166. (Cited on p. 203)
- [59] T. C. BRUEN AND D. BRYANT, *A subdivision approach to maximum parsimony*, Ann. Combin., 12 (2008), pp. 45–51. (Cited on pp. 102, 103, 108)
- [60] D. BRYANT, *Hunting for trees in binary character sets: Efficient algorithms for extraction, enumeration, and optimization*, J. Comput. Biol. 3 (1996), 275–288. (Cited on p. 100)
- [61] D. BRYANT, *A classification of consensus methods for phylogenetics*, in BioConsensus, DIMACS Working Group Meetings on BioConsensus, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 61, M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds., American Mathematical Society, Providence, 2003, pp. 163–184. (Cited on pp. 39, 68, 108)
- [62] D. BRYANT, *The splits in the neighbourhood of a tree*, Ann. Combin., 8 (2004), pp. 1–11. (Cited on p. 103)
- [63] D. BRYANT, *On the uniqueness of the selection criterion in neighbor-joining*, J. Classif., 22 (2005), pp. 3–15. (Cited on p. 123)
- [64] D. BRYANT, *Hadamard phylogenetic methods and the n -taxon process*, Bull. Math. Biol., 71 (2009), pp. 339–351. (Cited on p. 167)
- [65] D. BRYANT AND V. BERRY, *A structured family of clustering and tree construction methods*, Adv. Appl. Math. 27 (2001), pp. 705–732. (Cited on p. 114)
- [66] D. BRYANT AND S. KLAERE, *The link between segregation and phylogenetic diversity*, J. Math. Biol., 64 (2012), pp. 149–162. (Cited on p. 169)
- [67] D. BRYANT AND J. LAGERGREN, *Compatibility of unrooted phylogenetic trees is FPT*, Theor. Comput. Sci., 351 (2006), pp. 296–302. (Cited on p. 71)
- [68] D. BRYANT AND V. MOULTON, *Neighbor-Net: An agglomerative method for the construction of phylogenetic networks*, Mol. Biol. Evol., 21 (2004), pp. 255–265. (Cited on pp. 242, 243)

- [69] D. BRYANT, V. MOULTON, AND A. SPILLNER, *Consistency of the Neighbor-Net algorithm*, *Alg. Mol. Biol.*, 2 (2007), 8. (Cited on p. 243)
- [70] D. BRYANT AND M. STEEL, *Computing the distribution of a tree metric*, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 6 (2009), pp. 420–426. (Cited on p. 61)
- [71] D. BRYANT AND P. F. TUPPER, *Hyperconvexity and tight-span theory for diversities*, *Adv. Math.*, 231 (2012), pp. 3172–3198. (Cited on p. 145)
- [72] P. BUNEMAN, *The recovery of trees from measures of dissimilarity*, in *Mathematics in the Archaeological and Historical Sciences*, F. R. Hodson, D. G., Kendall, and P. Tautu, eds., Edinburgh University Press, Edinburgh, 1971, pp. 387–395. (Cited on p. 112)
- [73] G. CARDONA, M. LLABRÉS, F. ROSELLÓ AND G. VALIENTE, *The comparison of tree-sibling time consistent phylogenetic networks is graph isomorphism-complete*, *Sci. World. J.* (2014), article ID 254279. (Cited on p. 267)
- [74] G. CARDONA, F. ROSELLÓ, AND G. VALIENTE, *Comparison of tree-child phylogenetic networks*, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 6 (2009), pp. 552–569 (Cited on pp. 247, 248, 264)
- [75] M. CARTER, M. HENDY, D. PENNY, L. A. SZÉKELY, AND N. C. WORMALD, *On the distribution of lengths of evolutionary trees*, *SIAM J. Discrete Math.*, 3 (1990), pp. 38–47. (Cited on p. 106)
- [76] M. CASANELLAS AND J. FERNÁNDEZ-SÁNCHEZ, *Relevant phylogenetic invariants of evolutionary models*, *J. Math. Pures Appl.*, 96 (2011), pp. 207–229. (Cited on p. 193)
- [77] M. CASANELLAS, J. FERNÁNDEZ-SÁNCHEZ, AND A. M. KEDZIERSKA, *The space of phylogenetic mixtures for equivariant models*, *Alg. Mol. Biol.*, 7 (2012), p. 1. (Cited on pp. 162, 196)
- [78] M. CASANELLAS, L. GARCIA, AND S. SULLIVANT, *Catalog of small trees* in Algebraic statistics for computational biology, L. Pachter and B. Sturmfels, eds., Cambridge University Press, Cambridge, 2005, pp. 291–304. (Cited on p. 193)
- [79] M. CASANELLAS AND M. STEEL, *Phylogenetic mixtures and linear invariants for equal input models*, arXiv:1602.04671. (Cited on p. 195)
- [80] J. CHAI AND E. A. HOUSWORTH, *On the number of binary characters needed to recover a phylogeny using maximum parsimony*, *Bull. Math. Biol.*, 73 (2011), pp. 1398–1411. (Cited on p. 109)
- [81] J. CHAI AND E. A. HOUSWORTH, *On Rogers' proof of identifiability for the GTR + Γ + I model*, *Syst. Biol.*, 60 (2011), pp. 713–718. (Cited on p. 174)
- [82] S. CHAIKEN, A. K. DEWDNEY, AND P. J. SLATER, *An optimal diagonal tree code*, *SIAM J. Alg. Discrete Methods*, 4 (1983), pp. 42–49. (Cited on p. 27)
- [83] J. T. CHANG, *Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency*, *Math. Biosci.*, 137 (1996), pp. 51–73. (Cited on pp. 156, 180)
- [84] D. CHEN, G. J. BURLEIGH, AND D. FERNÁNDEZ-BACA, *Spectral partitioning of phylogenetic data sets based on compatibility*, *Syst. Biol.*, 56 (2007), pp. 623–632. (Cited on p. 108)
- [85] V. CHEPOI AND B. FICHET, *A note on circular decomposable metrics*, *Geom. Dedicata*, 69 (1998), pp. 237–240. (Cited on p. 130)
- [86] V. CHEPOI AND B. FICHET, *l_∞ -approximation via subdominants*, *J. Math. Psych.*, 44 (2000), pp. 600–616. (Cited on p. 116)

- [87] J. CHIFMAN AND L. KUBATKO, *Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites*, J. Theor. Biol., 374 (2015), pp. 35–47. (Cited on p. 228)
- [88] Y. B. CHOUE, K. T. HUBER, J. H. KOOLEN, Y. S. KWON, AND V. MOULTON, *Counting vertices and cubes in median graphs of circular split systems*, Eur. J. Combin., 29 (2008), pp. 443–456. (Cited on p. 242)
- [89] B. CHOR AND M. STEEL, *Do tree split probabilities determine the branch lengths?*, J. Theor. Biol., 374 (2015), pp. 54–59. (Cited on p. 169)
- [90] C. CHOY, J. JANSSON, K. SADAKANE, AND W.-K. SUNG, *Computing the maximum agreement of phylogenetic networks*, Electr. Notes Theor. Comput. Sci. 91 (2004), pp. 134–147. (Cited on p. 247)
- [91] P. CORDUE, S. LINZ, AND C. SEMPLE, *Phylogenetic networks that display a tree twice*, Bull. Math. Biol., 76 (2014), pp. 2664–2679. (Cited on pp. 254, 255)
- [92] D. R. COX AND P. A. W. LEWIS, *The statistical analysis of series of events*, John Wiley and Sons, London, 1966. (Cited on p. 221)
- [93] É CZABARKA, P. L. ERDŐS, V. JOHNSON, AND V. MOULTON, *Generating functions for multi-labeled trees*, Discrete Appl. Math., 161 (2013), pp. 107–117. (Cited on p. 43)
- [94] T. DAGAN AND W. MARTIN, *Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution*, Proc. Natl. Acad. Sci. USA, 104 (2007), pp. 870–875. (Cited on p. 232)
- [95] F. DARWIN, *The Life and Letters of Charles Darwin, Including an Autobiographical Chapter*, Vol. 1, F. Darwin, ed., John Murray, London, 1887, pp. 46; also available online from <http://darwin-online.org.uk/content/frameset?pageseq=64&itemID=F1452.1&viewtype=side> (Cited on p. 1)
- [96] G. DASARATHY, R. NOWAK, AND S. ROCH, *Data requirement for phylogenetic inference from multiple loci: A new distance method*, IEEE/ACM Trans. Comput. Biol. Bioinf., 12 (2015), pp. 422–432. (Cited on pp. 227, 229)
- [97] C. DASKALAKIS, E. MOSSEL, AND S. ROCH, *Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel’s conjecture*, Prob. Theor. Relat. Fields, 149 (2011), pp. 149–189. (Cited on p. 186)
- [98] C. DASKALAKIS AND S. ROCH, *Species trees from gene trees despite a high rate of lateral genetic transfer: A tight bound*, in Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA ’16), SIAM, Philadelphia, 2016, pp. 1621–1630. (Cited on p. 235)
- [99] W. H. E. DAY AND F. R. McMORRIS, *Axiomatic Consensus Theory in Group Choice and Biomathematics*, Frontiers in Applied Mathematics 29, SIAM, Philadelphia, 2003. (Cited on pp. 21, 36)
- [100] J. H. DEGNAN, M. DEGIORGIO, D BRYANT, AND N. A. ROSENBERG, *Properties of consensus methods for inferring species trees from gene trees*, Syst. Biol., 58 (2009), pp. 35–54. (Cited on p. 227)
- [101] J. H. DEGNAN AND J. A. RHODES, *There are no caterpillars in a wicked forest*, Theor. Pop. Biol., 105 (2015), 17–23. (Cited on p. 227)
- [102] J. H. DEGNAN AND N. A. ROSENBERG, *Discordance of species trees with their most likely gene trees*, PLoS Genetics, 2 (2006) pp. 762–768. (Cited on p. 227)

- [103] J. H. DEGNAN, N. A. ROSENBERG, AND T. STADLER, *The probability distribution of ranked gene trees on a species tree*, Math. Biosci., 235 (2012), pp. 45–55. (Cited on p. 228)
- [104] J. V. DE JONG, J. C. MCLEOD, AND M. STEEL, *Neighbourhoods of phylogenetic trees: Exact and asymptotic counts*, arXiv:1508.03774 [q-bio.PE] (2015). (Cited on pp. 25, 34, 60, 61)
- [105] Y. DENG AND D. FERNÁNDEZ-BACA, *Fast compatibility testing for rooted phylogenetic trees*, arXiv:1510.07758v1 [cs.DS] (Cited on p. 69)
- [106] R. DESPER AND O. GASCUEL, *Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting*, Mol. Biol. Evol., 21 (2004), pp. 587–598. (Cited on p. 126)
- [107] R. DESPER AND O. GASCUEL, *The minimum evolution distance-based approach to phylogenetic inference*, in Mathematics of evolution and phylogeny, O. Gascuel, ed., Oxford University Press, Oxford, 2005, pp.1–32. (Cited on p. 127)
- [108] S. L. DEVADOSS AND S. PETTI, *A space of phylogenetic networks*, arXiv:1607.06978v1 [math.CO]. (Cited on p. 267)
- [109] T. DEZULIAN AND M. STEEL, *Phylogenetic closure operations and homoplasy-free evolution*, in Classification, Clustering, and Data Mining Applications, Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul, eds., Springer, Berlin/Heidelberg, 2004, pp. 395–416. (Cited on p. 75)
- [110] M. DIETRICH, C. MCCARTIN, AND C. SEMPLE, *Bounding the maximum size of a minimal definitive set of quartets*, Inf. Process. Lett., 112 (2012), pp. 651–655. (Cited on p. 75)
- [111] Y. DING, S. GRÜENWALD, AND P. J. HUMPHRIES, *On agreement forests*, J. Combin. Theor. A, 118 (2011), pp. 2059–2065. (Cited on p. 34)
- [112] Z. DING, V. FILKOV, AND D. GUSFIELD, *A linear-time algorithm for the perfect phylogeny haplotyping (PPH) problem*, J. Comput. Biol., 13 (2006), pp. 522–553. (Cited on p. 99)
- [113] R. DONAGHEY, *Alternating permutations and binary increasing trees*, J. Combin. Theor. A, 18 (1975), pp. 141–148. (Cited on p. 48)
- [114] J. DONG, D. FERNÁNDEZ-BACA, F. R. McMORRIS, AND R. C. POWERS, *An axiomatic study of majority-rule (+) and associated consensus functions on hierarchies*, Discrete Appl. Math. 159 (2011), pp. 2038–2044. (Cited on p. 36)
- [115] J. DRAISMA AND J. KUTTLER, *On the ideals of equivariant tree models*, Math. Ann., 344 (2009), pp. 619–644. (Cited on p. 162)
- [116] A. DRESS, K. T. HUBER, J. KOOLEN, V. MOULTON, AND A. SPILLNER, *An algorithm for computing cutpoints in finite metric spaces*, J. Classif., 27 (2010), pp. 158–172. (Not cited)
- [117] A. DRESS, K. T. HUBER, J. KOOLEN, V. MOULTON, AND A. SPILLNER, *Basic Phylogenetic Combinatorics*, Cambridge University Press, Cambridge, 2012. (Cited on pp. 63, 70, 76, 244)
- [118] A. DRESS, K. T. HUBER AND V. MOULTON, *Some uses of the Farris transform in mathematics and phylogenetics – A review*, Ann. Combin. 11 (2007), pp. 1–37. (Cited on p. 117)
- [119] A. W. M. DRESS, K. T. HUBER, AND M. STEEL, “*Lassoing* a phylogenetic tree I: Basic properties, shellings, and covers”, J. Math. Biol., 65 (2012), pp. 77–105. (Cited on p. 127)
- [120] A. W. M. DRESS AND D. H. HUSON, *Constructing splits graphs*, IEEE/ACM Trans. Comput. Biol. Bioinf., 1 (2004), pp. 109–115. (Cited on p. 242)

- [121] A. DRESS AND M. STEEL, *Phylogenetic diversity over an abelian group*, Ann. Combin., 11 (2007), pp. 143–160. (Cited on p. 145)
- [122] M. DRTON, B. STURMFELS, AND S. SULLIVANT, *Lectures on Algebraic Statistics*, Birkhäuser Basel, 2009. (Cited on p. 191)
- [123] I. EBERSBERGER, P. GALGOCZY, S. TAUDIEN, S. TAENZER, M. PLATZER, AND A. VON HAESELER, *Mapping human genetic ancestry*, Mol. Biol. Evol., 24 (2007), pp. 2266–2276. (Cited on p. 226)
- [124] J. EDMONDS AND R. GILES, *A min-max relation for submodular functions on graphs*, in Studies in Integer Programming (Proceedings of the Workshop on Programming, Bonn, 1975), Annals of Discrete Mathematics, Vol. 1, P. L. Hammer, E. L. Johnson, B. H. Korte, and G. L. Nemhauser, eds., North-Holland, Amsterdam, 1977, pp. 185–204. (Cited on p. 20)
- [125] K. EICKMEYER, P. HUGGINS, L. PACTER, AND R. YOSHIDA, *On the optimality of the neighbor-joining algorithm*, Alg. Mol. Biol., 3 (2008), 5. (Cited on p. 124)
- [126] A. ENGSTRÖM, P. HERSH, AND B. STURMFELS, *Toric cubes*, Rend. Circ. Mat. Palermo, 62 (2013), pp. 67–78. (Cited on p. 191)
- [127] P. L. ERDŐS AND L. A. SZÉKELY, *Applications of antilexicographic order. I. An enumerative theory of trees*, Adv. Appl. Math., 10 (1989), pp. 488–496. (Cited on p. 16)
- [128] D. H. ERWIN, *Extinction as the loss of evolutionary history*, Proc. Natl. Acad. Sci. USA, 105 (2008), pp. 11520–11527. (Cited on p. 43)
- [129] R. ETIENNE AND J. ROSINDELL, *Prolonging the past counteracts the pull of the present: Protracted speciation can explain observed slowdowns in diversification*, Syst. Biol., 61 (2012), pp. 204–213. (Cited on p. 217)
- [130] S. N. EVANS AND T. P. SPEED, *Invariants of some probability models used in phylogenetic inference*, Ann. Stat., 21 (1993), pp. 355–377. (Cited on p. 167)
- [131] W. EVANS, C. KENYON, Y. PERES, AND L. J. SCHULMAN, *Broadcasting on trees and the Ising model*, Ann. Appl. Probab., 10 (2000), pp. 410–433. (Cited on p. 186)
- [132] B. FALLER, F. PARDI, AND M. STEEL, *Distribution of phylogenetic diversity under random extinction*, J. Theor. Biol., 251 (2008), pp. 286–296. (Cited on p. 143)
- [133] B. FALLER AND M. STEEL, *Trait-dependent extinction leads to greater expected biodiversity loss*, SIAM J. Discrete Math., 26 (2012), pp. 472–481. (Cited on p. 144)
- [134] M. FARACH, T. M. PRZTYCKA, AND M. THORUP, *On the agreement of many trees*, Inf. Process. Lett., 55 (1995), pp. 297–301. (Cited on p. 79)
- [135] J. FELSENSTEIN, *Cases in which parsimony or compatibility methods will be positively misleading*, Syst. Zool., 27 (1978), pp. 401–410. (Cited on p. 147)
- [136] J. FELSENSTEIN, *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, 2004. (Cited on pp. 33, 203)
- [137] J. FERNÁNDEZ-SÁNCHEZ AND M. CASANELLAS, *Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages*, Syst. Biol., 65 (2016), pp. 280–291. (Cited on pp. 192, 197)
- [138] J. FERNÁNDEZ-SÁNCHEZ, J. G. SUMNER, P. D. JARVIS, AND M. D. WOODHAMS, *Lie Markov models with purine/pyrimidine symmetry*, J. Math. Biol., 70 (2015), 855–891. (Cited on pp. 158, 163)

- [139] M. FISCHER, M. GALLA, L. HERBST, AND M. STEEL, *The most parsimonious tree for random data*, Mol. Phyl. Evol., 80 (2014), pp. 165–168. (Cited on p. 108)
- [140] M. FISCHER AND S. KELK, *On the maximum parsimony distance between phylogenetic trees*, Ann. Combin., 20 (2016), pp. 87–113. (Cited on p. 103)
- [141] M. FISCHER AND B. THATTE, *Revisiting an equivalence between maximum parsimony and maximum likelihood methods in phylogenetics*, Bull. Math. Biol., 72 (2010), pp. 208–220. (Cited on pp. 162, 181)
- [142] M. FISCHER, L. VAN IERSEL, S. KELK, AND C. SCORNAVACCA, *On computing the maximum parsimony score of a phylogenetic network*, SIAM J. Discrete Math., 29 (2015), pp. 559–585. (Cited on p. 265)
- [143] L. R. FOULDS AND R. W. ROBINSON, *Enumerating phylogenetic trees with multiple labels*, Discrete Math., 72 (1988), 129–139. (Cited on p. 18)
- [144] A. R. FRANCIS, *An algebraic view of bacterial genome evolution*, J. Math. Biol., 69 (2014), pp. 1693–1718. (Cited on p. 121)
- [145] A. R. FRANCIS AND M. STEEL, *Which phylogenetic networks are merely trees with additional arcs?*, Syst. Biol., 64 (2015), pp. 768–777. (Cited on pp. 256, 259, 261)
- [146] M. FUCHS AND E. Y. JIN, *Equality of Shapley value and fair proportion index in phylogenetic trees*, J. Math. Biol., 71 (2015), pp. 1133–1147. (Cited on pp. 140, 141)
- [147] P. GAMBETTE AND K. T. HUBER, *On encodings of phylogenetic networks of bounded level*, J. Math. Biol., 65 (2012), pp. 157–180. (Cited on pp. 256, 262)
- [148] O. GASCUEL, *A note on Sattath and Tversky’s, Saitou and Nei’s and Studier and Keppler’s algorithms for inferring phylogenies from evolutionary distances*, Mol. Biol. Evol., 11 (1994), pp. 961–963. (Cited on p. 123)
- [149] O. GASCUEL AND M. STEEL, *Neighbor-joining revealed*, Mol. Biol. Evol., 23 (2006), pp. 1997–2000. (Cited on p. 127)
- [150] O. GASCUEL AND M. STEEL, *Inferring ancestral sequences in taxon-rich phylogenies*, Math. Biosci. 227 (2010), pp. 125–135. (Cited on pp. 184, 188, 210, 211)
- [151] O. GASCUEL AND M. STEEL, *Predicting the ancestral character state changes in a tree is typically easier than predicting the root state*, Syst. Biol., 63 (2014), pp. 421–435. (Cited on pp. 188, 210)
- [152] O. GASCUEL AND M. STEEL, *A “stochastic safety radius” for distance-based tree reconstruction*, Algorithmica, 74 (2016), pp. 1386–1403. (Cited on p. 124)
- [153] A. GAVRYUSHKIN AND A. DRUMMOND, *The space of ultrametric phylogenetic trees*, J. Theor. Biol., 403 (2016), pp. 197–208. (Cited on p. 133)
- [154] T. GERNHARD, *New analytic results for speciation times in neutral models*, Bull. Math. Biol., 70 (2008), pp. 1082–1097. (Cited on p. 220)
- [155] J. GILL, S. LINUSSON, V. MOULTON, AND M. STEEL, *A regular decomposition of the edge-product space of phylogenetic trees*, Adv. Appl. Math., 41 (2008), pp. 158–176. (Cited on pp. 189, 191)
- [156] L. A. GOLDBERG, P. W. GOLDBERG, C. A. PHILLIPS, E. SWEEDYK, AND T. WARNOW, *Minimizing phylogenetic number to find good evolutionary trees*, Discrete Appl. Math., 71 (1996), pp. 111–136. (Cited on p. 98)

- [157] K. GORDON, E. FORD, AND K. ST. JOHN, *Hamiltonian walks of phylogenetic treespaces*, IEEE/ACM Trans. Comput. Biol. Bioinf., 10 (2013), pp. 1076–1079. (Cited on p. 33)
- [158] I. P. GOULDEN AND D. M. JACKSON, *Combinatorial Enumeration*, John Wiley & Sons, New York, 1983. (Reprint: Dover Publications, Mineola, NY, 2004.) (Cited on p. 107)
- [159] A. GRIGORIEV, S. KELK, AND N. LEKIĆ, *On low treewidth graphs and supertrees*, J. Graph. Alg. Appl., 19 (2015), pp. 325–343. (Cited on p. 71)
- [160] I. GRONAU, S. MORAN, AND I. YAVNEH, *Towards optimal distance functions for stochastic substitution models*, J. Theor. Biol., 260 (2009), pp. 294–307. (Cited on p. 151)
- [161] S. GRÜNEWALD, *Slim sets of binary trees*, J. Combin. Theor. A, 119 (2012), pp. 323–330. (Cited on pp. 70, 76)
- [162] S. GRÜNEWALD, P. J. HUMPHRIES, AND C. SEMPLE, *Quartet compatibility and the quartet graph*, Electr. J. Combin., 15 (2008), R103. (Cited on pp. 70, 71, 79)
- [163] S. GRÜNEWALD AND V. MOULTON, *Maximum parsimony for tree mixtures*, IEEE/ACM Trans. Comput. Biol. Bioinform., 6 (2009), pp. 97–102. (Cited on p. 108)
- [164] S. GRÜNEWALD, M. STEEL, AND M. SHEL SWENSON, *Closure operations in phylogenetics*, Math. Biosci., 208 (2007), pp. 521–537. (Cited on pp. 68, 78)
- [165] S. GUIASU, *Information Theory with Applications*, McGraw-Hill, New York, 1977. (Cited on p. 179)
- [166] S. GUILLEMOT AND F. NICOLAS, *Solving the maximum agreement subtree and the maximum compatible tree problems on many bounded degree trees*, in Combinatorial Pattern Matching (Proceedings of CPM 2006), Lecture Notes in Computer Science, Vol. 4009, M. Lewenstein and G. Valiente, eds., Springer, Berlin/Heidelberg, 2006, pp. 165–176. (Cited on p. 80)
- [167] A. D. M. GUNAWAN, B. DASGUPTA, AND L. ZHANG, *Locating a tree in a reticulation-visible network in cubic time*, arXiv:1507.02119v2 [q-BIO.PE] (2015). (Cited on pp. 255, 256)
- [168] D. GUSFIELD, *ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*, MIT Press, Cambridge, 2014. (Cited on pp. 237, 267)
- [169] D. GUSFIELD AND Y. WU, *The three-state perfect phylogeny problem reduces to 2-SAT*, Commun. Inf. Syst., 9 (2009), pp. 295–302. (Cited on p. 92)
- [170] R. GYSEL, K. STEVENS, AND D. GUSFIELD, *Reducing problems in unrooted tree compatibility to restricted triangulations of intersection graphs*, in Algorithms in Bioinformatics (Proceedings of WABI 2012), Lecture Notes in Computer Science, Vol. 7534, B. Raphael and J. Tang, eds., Springer, Berlin/Heidelberg, 2012, pp. 93–105. (Cited on p. 72)
- [171] C.-J. HAAKE, A. KASHIWADA, AND F. E. SU, *The Shapley value of phylogenetic trees*, J. Math. Biol., 56 (2008), pp. 479–497. (Cited on pp. 140, 141)
- [172] M. HABIB AND J. STACHO, *Unique perfect phylogeny is NP-hard*, in Combinatorial Pattern Matching (Proceedings of CPM 2011), Lecture Notes in Computer Science, Vol. 6661, R. Giancarlo and G. Manzini, eds., Springer, Berlin/Heidelberg, 2011, pp. 132–146. (Cited on p. 77)
- [173] O. HAGEN, K. HARTMANN, M. STEEL, AND T. STADLER, *Age-dependent speciation can explain the shape of empirical phylogenies*, Syst. Biol., 64 (2015), pp. 432–440. (Cited on p. 50)
- [174] T. HAGERUP AND C. RÜB, *A guided tour of Chernoff bounds*, Inf. Process. Lett., 33 (1990), pp. 305–308. (Cited on p. 202)

- [175] I. HAJIRASOULIHA AND B. J. RAPHAEL, *Reconstructing mutational history in multiply sampled tumors using perfect phylogeny mixtures*, Algorithms in Bioinformatics (Proceedings of WABI 2014), Lecture Notes in Computer Science, Vol. 8701, D. Brown and B. Morgenstern, eds., Springer, Berlin/Heidelberg, 2014, pp. 354–367. (Cited on p. 89)
- [176] K. HARTMANN, *The equivalence of two phylogenetic biodiversity measures: The Shapley value and fair proportion index*, J. Math. Biol., 67 (2013), pp. 1163–1170. (Cited on p. 142)
- [177] P. H. HARVEY, R. M. MAY, AND S. NEE, *Phylogenies without fossils*, Evolution, 48 (1994), 523–529. (Cited on pp. 205, 216)
- [178] B. HAUBOLD, *Alignment-free phylogenetics and population genetics*, Brief. Bioinf., 15 (2014), pp. 407–418. (Cited on p. 121)
- [179] D. C. HAWS, T. L. HODGE, AND R. YOSHIDA, *Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope*, Bull. Math. Biol., 73 (2011), pp. 2627–2648. (Cited on p. 127)
- [180] M. HAYAMIZU, *On the existence of infinitely many universal tree-based networks*, J. Theor. Biol., 396 (2016), pp. 204–206. (Cited on p. 261)
- [181] S. B. HEARD, *Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees*, Evolution, 46 (1992), pp. 1818–1826. (Cited on p. 54)
- [182] J. HEIN, M. H. SCHIERUP, AND C. WIUF, *Gene genealogies, variation and evolution: A primer in coalescent theory*, Oxford University Press, Oxford, 2005. (Cited on pp. 219, 267)
- [183] M. HELLMUTH, M. HERNANDEZ-ROSALES, K. T. HUBER, V. MOULTON, P. F. STADLER, AND N. WIESKE, *Orthology relations, symbolic ultrametrics, and cographs*, J. Math. Biol., 66 (2013), pp. 399–420. (Cited on p. 118)
- [184] M. D. HENDY, *The relationship between simple evolutionary tree models and observable sequence data*, Syst. Biol., 38 (1989), pp. 310–321. (Cited on p. 166)
- [185] M. D. HENDY AND D. PENNY, *Branch and bound algorithms to determine minimal evolutionary trees*, Math. Biosci., 59 (1982), pp. 277–290. (Cited on p. 60)
- [186] J. L. HERMAN, C. J. CHALLIS, A. NOVÁK, J. HEIN, AND S. C. SCHMIDLER, *Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure*, Mol. Biol. Evol. 31 (2014), pp. 2251–2266. (Cited on p. 203)
- [187] S. HERRMANN AND V. MOULTON, *Computing the blocks of a quasi-median graph*, Discrete Appl. Math., 179 (2014), pp. 129–138. (Cited on pp. 89, 244)
- [188] G. HICKEY, P. CARMI, A. MAHESHWARI, AND N. ZEH, *NAPX: A polynomial time approximation scheme for the Noah’s Ark problem*, in Algorithms in Bioinformatics (Proceedings of WABI 2008), Lecture Notes in Computer Science, Vol. 5251, K. A. Crandall and J. Lagergren, eds., Springer, Berlin/Heidelberg, 2008, pp. 76–86. (Cited on p. 144)
- [189] C. E. HINCHLIFF ET AL., *Synthesis of phylogeny and taxonomy into a comprehensive tree of life*, Proc. Natl. Acad. Sci. USA, 112 (2015), pp. 12764–12769. (Cited on p. 69)
- [190] L. S. T. HO AND C. ANÉ, *A linear-time algorithm for Gaussian and non-Gaussian trait evolution models*, Syst. Biol., 63 (2014), pp. 397–408. (Cited on p. 203)
- [191] K.T. HUBER AND G. KETTLEBOROUGH, *Distinguished minimal topological lassos*, SIAM J. Discrete Math., 29 (2015), pp. 940–961. (Cited on p. 129)

- [192] K. T. HUBER, S. LINZ, V. MOULTON, AND T. WU, *Spaces of phylogenetic networks from generalized nearest-neighbor interchange operations*, J. Math. Biol., 72 (2016), pp. 669–725. (Cited on p. 267)
- [193] K. T. HUBER, V. MOULTON, AND M. STEEL, *Four characters suffice to convexly define a phylogenetic tree*, SIAM J. Discrete Math., 18 (2005), pp. 835–843. (Cited on p. 94)
- [194] K. T. HUBER AND A-A. POPESCU, *Lassoing and corralling rooted phylogenetic trees*, Bull. Math. Biol., 75 (2013), pp. 444–465. (Cited on p. 129)
- [195] K. T. HUBER, L. VAN IERSEL, V. MOULTON, AND T. WU, *How much information is needed to infer reticulate evolutionary histories?*, Syst. Biol., 64 (2015), pp. 102–111. (Cited on p. 263)
- [196] J. P. HUELSENBECK, *Is the Felsenstein zone a fly trap?*, Syst. Biol., 46 (1997), pp. 69–74. (Cited on p. 170)
- [197] P. J. HUMPHRIES, *Combinatorial aspects of leaf-labelled trees*, Ph.D. thesis, University of Canterbury, New Zealand, 2008. (Cited on pp. 82, 85, 86)
- [198] P. J. HUMPHRIES AND C. SEMPLE, *Note on the hybridization number and subtree distance in phylogenetics*, Appl. Math. Lett., 22 (2009), pp. 611–615. (Cited on p. 266)
- [199] P. J. HUMPHRIES, S. LINZ, AND C. SEMPLE, *On the complexity of computing the temporal hybridization number for two phylogenies*, Discrete Appl. Math., 161 (2013), pp. 871–880. (Cited on pp. 266, 267)
- [200] P. J. HUMPHRIES, S. LINZ, AND C. SEMPLE, *Cherry picking: A characterization of the temporal hybridization number for a set of phylogenies*, Bull. Math. Biol., 75 (2013), pp. 1879–1890. (Cited on pp. 255, 256, 266)
- [201] P. J. HUMPHRIES AND T. WU, *On the neighbourhoods of trees*, IEEE/ACM Trans. Comput. Biol. Bioinf., 10 (2013), pp. 721–728. (Cited on pp. 31, 267)
- [202] D. H. HUSON AND D. BRYANT, *Application of phylogenetic networks in evolutionary studies*, Mol. Biol. Evol., 23 (2006), pp. 254–267. (Cited on p. 243)
- [203] D. H. HUSON, R. RUPP, AND C. SCORNAVACCA, *Phylogenetic Networks: Concepts, Algorithms and Applications*, Cambridge University Press, 2010. (Cited on pp. ix, 237, 240, 242, 244, 256)
- [204] P. JAGERS, *Stabilities and instabilities in population dynamics*, J. Appl. Probab., 29 (1992), pp. 770–780. (Cited on p. 212)
- [205] J. JANSSON, C. SHEN, AND W.-K. SUNG, *Algorithms for the majority rule (+) consensus tree and the frequency difference consensus tree*, in Algorithms in Bioinformatics (Proceedings of WABI 2013), Lecture Notes in Computer Science, Vol. 8126, A. Darling and J. Stoye, eds., Springer, Berlin/Heidelberg, 2013, pp. 141–155. (Cited on p. 36)
- [206] L. JETTEN AND L. VAN IERSEL, *Nonbinary tree-based phylogenetic networks*, arXiv:1601.04977v1 [q-bio.PE] (2016). (Cited on pp. 258, 259, 260, 261)
- [207] W. JETZ, G. H. THOMAS, J. B. JOY, D. W. REDDING, K. HARTMANN, AND A. O. MOORES, *Global distribution and conservation of evolutionary distinctness in birds*, Current Biol., 24 (2014) pp. 919–930. (Cited on p. 140)
- [208] A. N. C. KANG AND D. A. AULT, *Some properties of a centroid of a free tree*, Inf. Process. Lett., 4 (1975), pp. 18–20. (Cited on p. 8)
- [209] I. A. KANJ, L. NAKHLEH, C. THAN, AND G. XIA, *Seeing the trees and their branches in the network is hard*, Theor. Comput. Sci., 401 (2008), pp. 153–164. (Cited on p. 256)

- [210] S. KANNAN AND T. WARNOW, *A fast algorithm for the computation and enumeration of perfect phylogenies*, 26 (1997) pp. 1749–1763. (Cited on p. 92)
- [211] J. C. M. KEIJSPER AND R. A. PENDAVINGH, *Reconstructing a phylogenetic level-1 network from quartets*, Bull. Math. Biol., 76 (2014), pp. 2517–2541. (Cited on p. 263)
- [212] S. KELK AND G. STAMOULIS, *A note on convex characters, Fibonacci numbers and exponential-time algorithms*, arXiv:1508.02598v3 [q-bio.PE] (2016). (Cited on pp. 96, 97)
- [213] S. KELK, L. VAN IERSEL, AND C. SCORNAVACCA, *Phylogenetic incongruence through the lens of monadic second order logic*, J. Graph Alg. Appl., 20 (2016), pp. 189–215. (Cited on pp. 71, 72)
- [214] J. G. KEMENY AND J. L. SNELL, *Finite Markov Chains*, Springer, New York, 1976. (Cited on p. 160)
- [215] D. G. KENDALL, *On the generalized “birth-and-death” process*, Ann. Math. Stat., 19 (1948), pp. 1–15. (Cited on p. 211)
- [216] G. KETTLEBOROUGH, J. DICKS, I. N. ROBERTS, AND K. T. HUBER, *Reconstructing (super)trees from data sets with missing distances: Not all is lost*, Mol. Biol. Evol., 32 (2015), pp. 1628–1642. (Cited on p. 129)
- [217] J. KIM, *Slicing hyperdimensional oranges: The geometry of phylogenetic estimation*, Mol. Phyl. Evol., 17 (2000), pp. 58–75. (Cited on p. 189)
- [218] M. KIRKPATRICK AND M. SLATKIN, *Searching for evolutionary patterns in the shape of a phylogenetic tree*, Evolution, 47 (1993), pp. 1171–1181. (Cited on p. 54)
- [219] A. KLEINMAN, M. HAREL, AND L. PACHTER, *Affine and projective tree metric theorems*, Ann. Combin., 17 (2013), pp. 205–228. (Cited on pp. 129, 130)
- [220] L. L. KNOWLES AND L. S. KUBATKO, *Estimating Species Trees: Practical and Theoretical Aspects*, John Wiley and Sons, Hoboken, 2010. (Cited on p. 224)
- [221] E. V. KOONIN, *The turbulent network dynamics of microbial evolution and the statistical tree of life*, J. Mol. Evol., 80 (2015), pp. 244–250. (Cited on p. 237)
- [222] L. S. KUBATKO AND J. H. DEGNAN, *Inconsistency of phylogenetic estimates from concatenated data under coalescence*, Syst. Biol., 56 (2007), pp. 17–24. (Cited on p. 229)
- [223] M. R. LACEY AND J. T. CHANG, *A signal-to-noise analysis of phylogeny estimation by neighbor joining: Insufficiency of polynomial length sequences*, Math. Biosci., 199 (2006), pp. 188–215. (Cited on p. 186)
- [224] J. A. LAKE, *A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony*, Mol. Biol. Evol., 4 (1987), pp. 167–191. (Cited on p. 194)
- [225] A. LAMBERT AND T. STADLER, *Birth-death models and coalescent point processes: The shape and probability of reconstructed phylogenies*, Theor. Pop. Biol., 90 (2013), pp. 113–128. (Cited on pp. 43, 46, 49, 50, 217, 219)
- [226] A. LAMBERT AND M. STEEL, *Predicting the loss of phylogenetic diversity under non-stationary diversification models*, J. Theor. Biol., 337 (2013), pp. 111–124. (Cited on pp. 222, 224)
- [227] D. LEVY AND L. PACHTER, *The neighbor-net algorithm*, Adv. Appl. Math., 47 (2011), pp. 240–258. (Cited on p. 243)
- [228] M. LI, J. TROMP AND L. ZHANG, *On the nearest neighbour interchange distance between evolutionary trees*, J. Theor. Biol., 182 (1996), pp. 463–467. (Cited on pp. 33, 35)

- [229] B. LIN, B STURMFELS, X. TANG, AND R. YOSHIDA, *Convexity in tree spaces*, arXiv:1510.08797v1 (2015). (Cited on p. 133)
- [230] H. T. LIN, J. G. BURLEIGH, AND O. EULENSTEIN, *Consensus properties for the deep coalescent problem and their application for scalable tree search*, BMC Bioinf., 13 (Suppl. 10):S12 (2012), doi:10.1186/1471-2105-13-S10-S12. (Cited on p. 231)
- [231] Y. LIN, V. RAJAN, AND B. M. E. MORET, *A metric for phylogenetic trees based on matching*, IEEE/ACM Trans. Comput. Biol. Bioinf., 9 (2012), pp. 1014–1022. (Cited on p. 26)
- [232] C. LINNAEUS, *Systema Naturae*, 1st ed., 1735. (Cited on p. 18)
- [233] S. LINZ, K. ST. JOHN, AND C. SEMPLE, *Counting trees in a phylogenetic network is #P-complete*, SIAM J. Comput., 42 (2013), pp. 1768–1776. (Cited on p. 254)
- [234] S. LINZ, K. ST. JOHN, AND C. SEMPLE, *Optimizing tree and character compatibility across several phylogenetic trees*, Theor. Comput. Sci., 513 (2013), pp. 129–136. (Cited on p. 89)
- [235] S. LINZ, C. SEMPLE, AND T. STADLER, *Analyzing and reconstructing reticulation networks under timing constraints*, J. Math. Biol., 61 (2010) pp. 715–737. (Cited on p. 252)
- [236] C. LONG AND S. SULLIVANT, *Identifiability of 3-class Jukes-Cantor mixtures*, Adv. Appl. Math., 64 (2015), pp. 89–110. (Cited on p. 86)
- [237] H. M. MAHMOUD, *Pólya Urn Models*, Chapman and Hall/CRC Texts in Statistical Science, CRC Press, Taylor & Francis Group, LLC, Boca Raton, 2008. (Cited on p. 59)
- [238] T. MARCUSSEN, S. R. SANDVE, L. HEIER, M. SPANNAGL, M. PFEIFER, THE INTERNATIONAL WHEAT GENOME SEQUENCING CONSORTIUM, K. S. JAKOBSEN, B. B. H. WULFF, B. STEUERNAGEL, K. F. X. MAYER, AND O.-A. OLSEN, *Ancient hybridizations among the ancestral genomes of bread wheat*, Science 345 (2014), 1250092. (Cited on p. 245)
- [239] D. M. MARTIN AND B. D. THATTE, *The maximum agreement subtree problem*, Discrete Appl. Math., 161 (2013), pp. 1805–1817. (Cited on p. 80)
- [240] W. MARTIN, *Mosaic bacterial chromosomes: A challenge en route to a tree of genomes*, Bioessays, 21 (1999), pp. 99–104. (Cited on p. 1)
- [241] I. MARTYN, T. S. KUHN, A. O. MOOERS, V. MOULTON, AND A. SPILLNER, *Computing evolutionary distinctiveness indices in large scale analysis*, Alg. Mol. Biol., 7 (2012), 6. (Cited on pp. 142, 143)
- [242] F. A. MATSEN, E. MOSSEL, AND M. STEEL, *Mixed-up trees: The structure of phylogenetic mixtures*, Bull. Math. Biol., 70 (2008), pp. 1115–1139. (Cited on pp. 85, 175)
- [243] F. A. MATSEN AND M. STEEL, *Phylogenetic mixtures on a single tree can mimic a tree of another topology*, Syst. Biol., 56 (2007), pp. 767–775. (Cited on p. 173)
- [244] P. McCULLAGH, J. PITMAN, AND M. WINKEL, *Gibbs fragmentation trees*, Bernoulli, 14 (2008), pp. 988–1002. (Cited on p. 57)
- [245] C. McDIARMID, C. SEMPLE, AND D. WELSH, *Counting phylogenetic networks*, Ann. Combin., 19 (2015), pp. 205–224. (Cited on pp. 246, 248, 253)
- [246] A. MCKENZIE AND M. STEEL, *Distributions of cherries for two models of trees*, Math. Biosci., 164 (2000), pp. 81–92. (Cited on p. 60)
- [247] F. R. McMORRIS AND R. C. POWERS, *Some axiomatic limitations for consensus and supertree functions on hierarchies*, J. Theor. Biol., 404 (2016), pp. 342–347. (Cited on p. 40)

- [248] F. R. McMORRIS, *On the compatibility of binary qualitative taxonomic characters*, Bull. Math. Biol., 39 (1977), pp. 133–138. (Cited on p. 24)
- [249] F. R. McMORRIS, *Axioms for consensus functions on undirected phylogenetic trees*, Math. Biosci., 74 (1985), pp. 17–21 (Cited on p. 40)
- [250] F. R. McMORRIS AND M. A. STEEL, *The complexity of the median procedure for binary trees*, in New Approaches in Classification and Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, eds., Springer, Berlin/Heidelberg, 1994, pp. 136–140. (Cited on p. 21)
- [251] F. R. McMORRIS, T. J. WARNOW, AND T. WIMER, *Triangulating vertex-colored graphs*, SIAM J. Discrete Math., 7 (1994), pp. 296–306. (Cited on p. 92)
- [252] E. J. MCTAVISH, M. STEEL, AND M. T. HOLDER, *Twisted trees and inconsistency of tree estimation when gaps are treated as missing data – The impact of model mis-specification in distance corrections*, Mol. Phyl. Evol., 93 (2015), pp. 289–295. (Cited on p. 182)
- [253] J. MEIDANIS AND Z. DIAS, *An alternative algebraic formalism for genome rearrangements*, Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families, Computational Biology, Vol. 1, D. Sankoff and J. H. Nadeau, eds., Springer, Dordrecht, 2000, pp. 213–223. (Cited on p. 121)
- [254] A. MEIR, J. MOON, AND M. STEEL, *A limiting theorem on 2-coloured trivalent trees*, Cong. Numer., 150 (2001), pp. 43–63. (Cited on p. 107)
- [255] C. MENG AND L. S. KUBATKO, *Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model*, Theor. Pop. Biol. 75 (2009), pp. 35–45. (Cited on p. 232)
- [256] R. MIHAESCU, D. LEVY, AND L. PACHTER, *Why neighborjoining works*, Algorithmica, 54 (2009), pp. 1–24. (Cited on p. 124)
- [257] R. MIHAESCU AND L. PACHTER, *Combinatorics of least-squares trees*, Proc. Natl. Acad. Sci. USA, 105 (2008), pp. 13206–13211. (Cited on p. 122)
- [258] S. MIRARAB, R. REAZ, MD. S. BAYZID, T. ZIMMERMANN, M. S. SWENSON, AND T. WARNOW, *ASTRAL: Genome-scale coalescent-based species tree estimation*, Bioinformatics, 30 (2014), pp. i541–i548. (Cited on p. 230)
- [259] S. L. MITCHELL, *Another characterization of the centroid of a tree*, Discrete Math., 24 (1978), pp. 277–280. (Cited on p. 8)
- [260] E. MOAN AND J. RUSINKO, *Combinatorics of linked systems of quartet trees*, Involve, 9 (2016), pp. 171–180. (Cited on p. 74)
- [261] A. MOOERS, O. GASCUEL, T. STADLER, H. LI, AND M. STEEL *Branch lengths on birth-death trees and the expected loss of phylogenetic diversity*, Syst. Biol., 61 (2012), pp. 195–203. (Cited on pp. 221, 222, 224)
- [262] J. W. MOON AND M. A. STEEL, *A limiting theorem for parsimoniously bicoloured trees*, Appl. Math. Lett., 6 (1993), pp. 5–8. (Cited on p. 107)
- [263] H. MORLON, T. L. PARSONS, AND J. B. PLOTKIN, *Reconciling molecular phylogenies with the fossil record*, Proc. Natl. Acad. Sci. USA, 108 (2011), pp. 16327–16332. (Cited on p. 205)
- [264] E. MOSSEL AND S. ROCH, *Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci*, IEEE/ACM Trans. Comput. Biol. Bioinf., 7 (2010), pp. 166–171. (Cited on p. 224)

- [265] E. MOSSEL AND S. ROCH, *Identifiability and inference of non-parametric rates-across-sites models on large-scale phylogenies*, J. Math. Biol., 67 (2013), pp. 767–797. (Cited on p. 174)
- [266] E. MOSSEL, S. ROCH, AND M. STEEL, *Shrinkage effect in ancestral maximum likelihood*, IEEE/ACM Trans. Comput. Biol. Bioinf., 6 (2009), pp. 126–133. (Cited on p. 181)
- [267] E. MOSSEL AND M. STEEL, *A phase transition for a random cluster model on phylogenetic trees*, Math. Biosci., 187 (2004), pp. 189–203. (Cited on pp. 198, 200)
- [268] E. MOSSEL AND M. STEEL, *How much can evolved characters tell us about the tree that generated them?*, in Mathematics of Evolution and Phylogeny, O. Gascuel, ed., Oxford University Press, Oxford, 2005, pp. 384–412. (Cited on p. 184)
- [269] E. MOSSEL AND M. STEEL, *Majority rule has transition ratio 4 on Yule trees under a 2-state symmetric model*, J. Theor. Biol., 360 (2014), pp. 315–318. (Cited on p. 210)
- [270] E. MOSSEL AND E. VIGODA, *Phylogenetic MCMC algorithms are misleading on mixtures of trees*, Science, 309 (2005), pp. 2207–2209. (Cited on p. 183)
- [271] V. MOULTON AND M. A. STEEL, *Retractions of finite distance functions onto tree metrics*, Discrete Appl. Math. 91 (1999), pp. 215–233. (Cited on pp. 132, 133)
- [272] V. MOULTON AND M. STEEL, *Peeling phylogenetic “oranges,”* Adv. Appl. Math., 33 (2004), pp. 710–727. (Cited on pp. 189, 191)
- [273] G. G. R. MURRAY, L. A. WINERT, E. L. RHULE, AND J. J. WELCH, *The phylogeny of *Rickettsia* using different evolutionary signatures: How tree-like is bacterial evolution?*, Syst. Biol., 65 (2016), pp. 265–279. (Cited on p. 237)
- [274] L. NAKHLEH, *Evolutionary phylogenetic networks: models and issues*, in Problem Solving Handbook in computational biology and bioinformatics, L. Heath and N. Ramakrishnan, eds., Springer, New York, 2010, pp. 125–158. (Cited on p. 265)
- [275] S. NEE, E. C. HOLMES, R. M. MAY, AND P. H. HARVEY, *Extinction rates can be estimated from molecular phylogenies*, Philos. Trans. R. Soc. London B, 344 (1994), 77–82. (Cited on p. 215)
- [276] S. NEE AND R. M. MAY, *Extinction and the loss of evolutionary history*, Science, 278 (1997), pp. 692–694. (Cited on p. 223)
- [277] S. NEE, R. M. MAY, AND P. H. HARVEY, *The reconstructed evolutionary process*, Philos. Trans. R. Soc. London B, 344 (1994), 305–311. (Cited on p. 216)
- [278] M. NIKAIDO, A. P. ROONEY, AND N. OKADA, *Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 10261–10266. (Cited on p. 88)
- [279] J. P. O'Dwyer, S. W. KEMBEL, AND T. J. SHARPTON, *Backbones of evolutionary history test biodiversity theory for microbes*, Proc. Natl. Acad. Sci. USA, 112 (2015), pp. 8356–8361. (Cited on p. 143)
- [280] M. OWEN, *Computing geodesic distances in tree space*, SIAM J. Discrete Math. 25 (2011), pp. 1506–1529. (Cited on p. 131)
- [281] L. PACHTER AND D. SPEYER, *Reconstructing trees from subtree weights*, Appl. Math. Lett., 17 (2004), pp. 615–621. (Cited on p. 134)
- [282] E. PARADIS, *The distribution of branch lengths in phylogenetic trees*, Mol. Phyl. Evol., 94 (2016), pp. 136–145. (Cited on p. 219)

- [283] F. PARDI AND O. GASCUEL *Combinatorics of distance-based tree inference*, Proc. Natl. Acad. Sci. USA, 109 (2012), pp. 16443–16448. (Cited on p. 122)
- [284] F. PARDI AND N. GOLDMAN, *Species choice for comparative genomics: Being greedy works*, PLoS Genetics, 1 (2005), e71. (Cited on pp. 133, 135)
- [285] F. PARDI AND C. SCORNAVACCA, *Reconstructible phylogenetic networks: Do not distinguish the indistinguishable*, PLoS Comput. Biol., 11 (2015), e1004135. (Cited on pp. 263, 264)
- [286] S. L. PARKS AND N. GOLDMAN, *Maximum likelihood inference of small trees in the presence of long branches*, Syst. Biol., 63 (2014), pp. 798–811. (Cited on pp. 181, 182)
- [287] Y. PAUPLIN, *Direct calculation of a tree length using a distance matrix*, J. Mol. Evol., 51 (2000), pp. 41–47. (Cited on p. 126)
- [288] J. PEARL AND M. TARSI, *Structuring causal trees*, J. Complexity, 2 (1986), pp. 60–77. (Cited on p. 74)
- [289] I. PE’ER, T. PUPKO, R. SHAMIR, AND R. SHARAN, *Incomplete directed perfect phylogeny*, SIAM J. Comput., 33 (2004), pp. 590–607. (Cited on p. 99)
- [290] T. M. PRZTYCKA, *An important connection between network motifs and parsimony models*, in Research in Computational Molecular Biology (Proceedings of RECOMB 2006), Lecture Notes in Computer Science, Vol. 3909, A. Apostolico, C. Guerra, S. Istrail, P. A. Pevzner, and M. Waterman, eds., Springer, Berlin/Heidelberg, 2006, pp 321–335. (Cited on p. 98)
- [291] A. PURVIS, P.-M. AGAPOW, J. L. GITTLEMAN, AND G. M. MACE, *Nonrandom extinction and the loss of evolutionary history*, Science, 288 (2000), pp. 328–330. (Cited on p. 133)
- [292] O. G. PYBUS AND P. H. HARVEY, *Testing macro-evolutionary models using incomplete molecular phylogenies*, Proc. R. Soc. London B, 267 (2000), pp. 2267–2272. (Cited on p. 220)
- [293] A. RÉNYI, *Probability theory*, Dover Publications, Mineola, NY, 2007. (Originally published jointly by North-Holland Publishing Company, Amsterdam and Akadémiai Kiadó, Budapest, 1970). (Cited on p. 160)
- [294] J. A. RHODES AND S. SULLIVANT, *Identifiability of large phylogenetic mixture models*, Bull. Math. Biol., 74 (2012), pp. 212–231. (Cited on p. 197)
- [295] A. ROBINSON AND S. WHITEHOUSE, *The tree representation of Σ_{n+1}* , J. Pure Appl. Algebra, 111 (1996), pp. 245–253. (Cited on p. 132)
- [296] D. F. ROBINSON AND L. R. FOULDS, *Comparison of phylogenetic trees*, Math. Biosci., 53 (1981), pp. 131–147. (Cited on p. 25)
- [297] S. ROCH, *A short proof that phylogenetic tree reconstruction by maximum likelihood is hard*, IEEE/ACM Trans. Comput. Biol. Bioinf., 3 (2006) pp. 92–94. (Cited on p. 181)
- [298] S. ROCH, *Towards extracting all phylogenetic information from matrices of evolutionary distances*, Science, 327 (2010), pp. 1376–1379. (Cited on p. 186)
- [299] S. ROCH AND A. SLY, *Phase transition in the sample complexity of likelihood-based phylogeny inference*, arXiv:1508.01964 [math.PR], (2015). (Cited on p. 186)
- [300] S. ROCH AND S. SNIR, *Recovering the treelike trend of evolution despite extensive lateral genetic transfer: A probabilistic analysis*, J. Comput. Biol., 20 (2013), pp. 93–112. (Cited on p. 235)

- [301] S. ROCH AND M. STEEL, *Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent*, Theor. Pop. Biol., 100 (2015), pp. 56–62. (Cited on p. 229)
- [302] N. A. ROSENBERG, *The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees*, Ann. Combin., 10 (2006), pp. 129–146. (Cited on p. 60)
- [303] N. A. ROSENBERG, *Statistical tests for taxonomic distinctiveness from observations of monophyly*, Evolution, 61 (2007), pp. 317–323. (Cited on p. 51)
- [304] N. A. ROSENBERG, *Counting coalescent histories*, J. Comput. Biol., 14 (2007), pp. 360–377. (Cited on p. 225)
- [305] N. SAITOU AND M. NEI, *The neighbor-joining method: a new method for reconstructing phylogenetic trees*, Mol. Biol. Evol., 4 (1987), pp. 406–425. (Cited on pp. 121, 123)
- [306] M. J. SANDERSON, M. M. MCMAHON, AND M. STEEL, *Phylogenomics with incomplete taxon coverage: the limits to inference*, BMC Evol. Biol., 10 (2010), 155. (Cited on p. 85)
- [307] M. J. SANDERSON, M. M. MCMAHON, AND M. STEEL, *Terraces in phylogenetic tree space*, Science, 333 (2011), pp. 448–450. (Cited on pp. 74, 83, 84)
- [308] D. SANKOFF, *Reconstructing the history and geography of an evolutionary tree*, Am. Math. Mon., 79 (1972), pp. 596–603. (Cited on p. 147)
- [309] E. SCHRÖDER, *Vier kombinatorische probleme*, Z. Math. Phys., 15 (1870), pp. 361–376. (Cited on p. 15)
- [310] J. SCHWEINSBERG, *An $O(n^2)$ bound for the relaxation time of a Markov chain on cladograms*, Rand. Struct. Alg., 20 (2002), pp. 59–70. (Cited on p. 35)
- [311] R. W. SCOTLAND AND M. STEEL, *Circumstances in which parsimony but not compatibility will be provably misleading*, Syst. Biol., 64 (2015), pp. 492–504. (Cited on p. 97)
- [312] C. SEMPLE, *Reconstructing minimal rooted trees*, Discrete Appl. Math., 127 (2003), pp. 489–503. (Cited on p. 69)
- [313] C. SEMPLE, *Phylogenetic networks with every embedded phylogenetic tree a base tree*, Bull. Math. Biol., 78 (2016), pp. 132–137. (Cited on pp. 247, 257)
- [314] C. SEMPLE, AND M. STEEL, *Tree reconstruction from multi-state characters*, Adv. Appl. Math., 28 (2002), pp. 169–184. (Cited on p. 93)
- [315] C. SEMPLE AND M. STEEL, *Phylogenetics*, Oxford University Press, Oxford, 2003. (Cited on pp. 12, 18, 46, 69, 87, 91, 102, 105, 119, 144, 244)
- [316] C. SEMPLE AND M. STEEL, *Cyclic permutations and evolutionary trees*, Adv. Appl. Math., 32 (2004), pp. 669–680. (Cited on pp. 29, 126)
- [317] C. SEMPLE, AND M. STEEL, *Unicyclic networks: compatibility and enumeration*, IEEE/ACM Trans. Comp. Biol. Bioinf., 3 (2006), pp. 84–91. (Cited on pp. 239, 240)
- [318] B. SHUTTERS, S. VAKATI, AND D. FERNÁNDEZ-BACA, *Incompatible quartets, triplets, and characters*, Alg. Mol. Biol., 8 (2013), 11. (Cited on p. 92)
- [319] D. D. SLEATOR, R. E. TARJAN, AND W. P. THURSTON, *Short encodings of evolving structures*, SIAM J. Discrete Math., 5 (1992), pp. 428–450. (Cited on p. 35)

- [320] E. SOBER AND M. STEEL, *Time and knowability in evolutionary processes*, Philos. Sci., 81 (2014), pp. 558–579. (Cited on p. 184)
- [321] D. A. SPADE, R. HERBEI, AND L. S. KUBATKO, *A note on the relaxation time of two Markov chains on rooted phylogenetic tree spaces*, Stat. Prob. Lett., 84 (2014), pp. 247–252. (Cited on p. 36)
- [322] T. STADLER, *Inferring speciation and extinction processes from extant species data*, Proc. Natl. Acad. Sci. USA, 108 (2011), pp. 16145–16146. (Cited on p. 205)
- [323] T. STADLER, *Recovering speciation and extinction dynamics based on phylogenies*, J. Evol. Biol., 26 (2013), pp. 1203–1219. (Cited on pp. 205, 219)
- [324] T. STADLER AND M. STEEL, *Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models*, J. Theor. Biol., 297 (2012), pp. 33–40. (Cited on pp. 209, 215)
- [325] R. P. STANLEY, *Enumerative combinatorics, Vol. 2*, Cambridge Studies in Advanced Mathematics, Vol. 62, Cambridge University Press, Cambridge, 1999. (Cited on p. 17)
- [326] D. ŠTEFAKOVÍČ AND E. VIGODA, *Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions*, J. Comput. Biol., 14 (2007), pp. 156–189. (Cited on p. 196)
- [327] M. STEEL, *The complexity of reconstructing trees from qualitative characters and subtrees*, J. Classif., 9 (1992), pp. 91–116. (Cited on pp. 70, 74, 96)
- [328] M. A. STEEL, *Distributions on bicoloured binary trees arising from the principle of parsimony*, Discrete Appl. Math., 41 (1993), pp. 245–261. (Cited on p. 101)
- [329] M. STEEL, *Recovering a tree from the leaf colourations it generates under a Markov model*, Appl. Math. Lett., 7 (1994), pp. 19–23. (Cited on p. 155)
- [330] M. STEEL, *Phylogenetic diversity and the greedy algorithm*, Syst. Biol., 54 (2005), pp. 527–529. (Cited on pp. 134, 135)
- [331] M. STEEL, *Consistency of Bayesian inference of resolved phylogenetic trees*, J. Theor. Biol., 336 (2013), pp. 246–249. (Cited on p. 183)
- [332] M. STEEL AND B. FALLER, *Markovian log-supermodularity, and its applications in phylogenetics*, Appl. Math. Lett., 22 (2009), pp. 1141–1144. (Cited on p. 196)
- [333] M. A. Steel and Y. X. Fu, *Classifying and counting linear phylogenetic invariants for the Jukes–Cantor model*, J. Comput. Biol., 2 (1995), pp. 39–47. (Cited on p. 194)
- [334] M. STEEL, L. GOLDSTEIN, AND M. S. WATERMAN, *A central limit theorem for the parsimony length of trees*, Adv. Appl. Prob., 28 (1996), pp. 1051–1071. (Cited on p. 107)
- [335] M. STEEL, S. LINZ, D. H. HUSON, AND M. J. SANDERSON, *Identifying a species tree subject to random lateral gene transfer*, J. Theor. Biol., 322 (2013), pp. 81–93. (Cited on p. 235)
- [336] M. STEEL AND E. MATSEN, *The Bayesian “star paradox” persists for long finite sequences*, Mol. Biol. Evol., 24 (2007), pp. 1075–1079. (Cited on p. 183)
- [337] M. STEEL AND A. MCKENZIE, *Properties of phylogenetic trees generated by Yule-type speciation models*, Math. Biosci., 170 (2001), pp. 91–112. (Cited on pp. 52, 58)
- [338] M. STEEL AND A. MOOERS, *The expected length of pendant and interior edges of Yule tree*, Appl. Math. Lett., 23 (2010), pp. 1315–1319. (Cited on p. 209)

- [339] M. STEEL AND D. PENNY, *Maximum parsimony and the phylogenetic information in multi-state characters*, in Parsimony, phylogeny and genomics, V. A. Albert, ed., Oxford University Press, Oxford, 2005, pp. 163–178. (Cited on p. 105)
- [340] M. STEEL AND M. J. SANDERSON, *Characterizing phylogenetically decisive taxon coverage*, Appl. Math. Lett., 23 (2010), pp. 82–86. (Cited on pp. 83, 85)
- [341] M. A. STEEL, L. A. SZÉKELY, AND M. D. HENDY, *Reconstructing trees when sequence sites evolve at variable rates*, J. Comput. Biol., 1 (1994), pp. 153–163. (Cited on p. 174)
- [342] M. A. STEEL AND L. A. SZÉKELY, *Inverting random functions II: Explicit bounds for discrete maximum likelihood estimation, with applications*, SIAM J. Discrete Math., 15 (2002), pp. 562–575. (Cited on pp. 178, 179)
- [343] M. STEEL, L. SZÉKELY, AND E. MOSEL, *Phylogenetic information complexity: Is testing a tree easier than finding it?* J. Theor. Biol., 258 (2009), pp. 95–102. (Cited on p. 201)
- [344] J. A. STUDIER AND K. J. KEPPLER, *A note on the neighbor-joining algorithm of Saitou and Nei*, Mol. Biol. Evol., 5 (1988), pp. 729–731. (Cited on p. 123)
- [345] M. A. SUCHARD AND B. D. REDELINGS, *BALi-Phy: Simultaneous Bayesian inference of alignment and phylogeny*, Bioinformatics, 22 (2006), pp. 2047–2048. (Cited on p. 203)
- [346] S. SULLIVANT, *The disentangling number for phylogenetic mixtures*, SIAM. J. Discrete Math., 26 (2012), pp. 856–859. (Cited on pp. 85, 86)
- [347] J. G. SUMNER, J. FERNÁNDEZ-SÁNCHEZ, AND P. D. JARVIS, *Lie Markov models*, J. Theor. Biol., 298 (2012), pp. 16–31. (Cited on pp. 158, 165)
- [348] J. G. SUMNER, P. D. JARVIS, J. FERNÁNDEZ-SÁNCHEZ, B. T. KAIN, M. D. WOODHAMS, AND B. R. HOLLAND, *Is the general time-reversible model bad for molecular phylogenetics*, Syst. Biol., 61 (2012), pp. 1069–1074. (Cited on pp. 158, 165)
- [349] E. SUSKO, *On the distributions of bootstrap support and posterior distributions for a star tree*, Syst. Biol., 57(2008), pp. 602–612. (Cited on p. 183)
- [350] E. SUSKO, *Bayesian long branch attraction bias and corrections*, Syst. Biol., 64 (2015), pp. 243–255. (Cited on p. 183)
- [351] M. SVIRIDENKO, *A note on maximizing a submodular set function subject to a knapsack constraint*, Oper. Res. Lett., 32 (2004), pp. 41–43. (Cited on p. 138)
- [352] L. A. SZÉKELY, M. A. STEEL, AND P. L. ERDŐS, *Fourier calculus on evolutionary trees*, Adv. Appl. Math., 14 (1993), pp. 200–216. (Cited on p. 167)
- [353] S. TAVARÉ, *Line-of-descent and genealogical processes, and their applications in population genetics models*, Theor. Pop. Biol., 26 (1984), pp. 119–164. (Cited on p. 226)
- [354] C. V. THAN AND N. A. ROSENBERG, *Mathematical properties of the deep coalescence cost*, IEEE/ACM Trans. Comput. Biol. Bioinf., 10 (2013), pp. 61–72. (Cited on pp. 230, 231)
- [355] C. V. THAN AND N. A. ROSENBERG, *Mean deep coalescence cost under exchangeable probability distributions*, Discrete Appl. Math., 174 (2014), pp. 11–26. (Cited on pp. 53, 54, 232)
- [356] B. D. THATTE, *Combinatorics of pedigrees I: Counterexamples to a reconstruction question*, SIAM J. Discrete Math., 22 (2008), pp. 961–970. (Cited on p. 268)
- [357] J. L. THORNE, H. KISHINO, AND J. FELSENSTEIN, *An evolutionary model for maximum likelihood alignment of DNA sequences*, J. Mol. Evol., 33 (1991), pp. 114–124. (Cited on p. 203)

- [358] J. P. TOWNSEND, Z. SU, AND Y. I. TEKLE, *Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny*, Syst. Biol., 61 (2012), pp. 835–849. (Cited on p. 184)
- [359] H. TRAPPMANN AND G. M. ZIEGLER, *Shellability of complexes of trees*, J. Combin. Theor. A, 82 (1998), pp. 168–178. (Cited on p. 132)
- [360] J. TRUSzkowski AND N. GOLDMAN, *Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps*, Syst. Biol., 65 (2016), pp. 328–333. (Cited on p. 180)
- [361] C. TUFFLEY AND M. STEEL, *Links between maximum likelihood and maximum parsimony under a simple model of site substitution*, Bull. Math. Biol., 59 (1997), pp. 581–607. (Cited on pp. 162, 181)
- [362] S. VAKATI AND D. FERNÁNDEZ-BACA, *Graph triangulations and the compatibility of unrooted phylogenetic trees*, Appl. Math. Lett., 24 (2011), pp. 719–723. (Cited on pp. 71, 72)
- [363] S. VAKATI AND D. FERNÁNDEZ-BACA, *Characterizing compatibility and agreement of unrooted trees via cuts in graphs*, Alg. Mol. Biol., 9 (2014), 13. (Cited on p. 72)
- [364] S. VAKATI AND D. FERNÁNDEZ-BACA, *Compatibility, incompatibility, tree-width, and forbidden phylogenetic minors*, Electr. Notes Discrete Math., 50 (2015), pp. 337–342. (Cited on p. 72)
- [365] L. VAN IERSEL, *Different topological restrictions of rooted phylogenetic networks. Which make biological sense?* <http://phylonetworks.blogspot.nl/2013/03/different-topological-restrictions-of.html> (Cited on p. 256)
- [366] L. VAN IERSEL AND S. KELK, *Constructing the simplest possible phylogenetic network from triplets*, Algorithmica, 60 (2011), pp. 207–235. (Cited on pp. 247, 263)
- [367] L. VAN IERSEL AND V. MOULTON, *Trinets encode tree-child and level-2 phylogenetic networks*, J. Math. Biol., 68 (2014), pp. 1707–1729. (Cited on p. 263)
- [368] L. VAN IERSEL, C. SEMPLE, AND M. STEEL, *Locating a tree in a phylogenetic network*, Inf. Process. Lett., 110 (2010), pp. 1037–1043. (Cited on pp. 255, 256)
- [369] L. VAN IERSEL, C. SEMPLE, AND M. STEEL, *Quantifying the extent of lateral gene transfer required to avert a “Genome of Eden”*, Bull. Math. Biol., 72 (2010), pp. 1783–1798. (Cited on p. 233)
- [370] S. VERON, T. J. DAVIES, M. W. CADOTTE, P. CLERGEAU, AND S. PAVOINE, *Predicting loss of evolutionary history: Where are we?*, Biol. Rev., to appear, (2015), DOI 10.1111/brv.12228. (Cited on p. 133)
- [371] C. WHIDDEN, R. G. BEIKO, AND N. ZEH, *Fixed-parameter algorithms for maximum agreement forests*, SIAM J. Comput., 42 (2013), pp. 1431–1466. (Cited on p. 32)
- [372] C. WHIDDEN AND F. A. MATSEN IV, *Ricci–Ollivier curvature of the rooted phylogenetic subtree-prune-regraft graph*, in 2016 Proceedings of the Thirteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO), J. A. Fill and M. D. Ward, eds., SIAM, Philadelphia, 2016, pp. 106–120. (Cited on p. 36)
- [373] K. WICKE AND M. FISCHER, *Comparing the rankings obtained from two biodiversity indices: The Fair Proportion Index and the Shapley Value*, arXiv:1507.08620v1 [q-bio.PE], (2015). (Cited on p. 142)
- [374] S. J. WILLSON, *Properties of normal phylogenetic networks*, Bull. Math. Biol., 72 (2010), pp. 340–358. (Cited on pp. 251, 253)

- [375] S. J. WILLSON, *Regular networks can be uniquely constructed from their trees*, IEEE/ACM Trans. Comput. Biol. Bioinf., 8 (2011), pp. 785–796. (Cited on p. 262)
- [376] S. J. WILLSON, *Tree-average distances on certain phylogenetic networks have their weights uniquely determined*, Alg. Mol. Biol., 7 (2012), 13. (Cited on p. 254)
- [377] M. D. WOODHAMS, J. FERNÁNDEZ-SÁNCHEZ, AND J. G. SUMNER, *A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates*, Syst. Biol., 64 (2015), pp. 638–650. (Cited on p. 158)
- [378] S. YANCOPOULOS, O. ATTIE, AND R. FRIEDBERG, *Efficient sorting of genomic permutations by translocation, inversion and block interchange*, Bioinformatics, 21 (2005), pp. 3340–3346. (Cited on p. 121)
- [379] Z. YANG, *Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics*, Mol. Biol. Evol., 24 (2007), pp. 1639–1655. (Cited on p. 183)
- [380] Y. YU, J. DONG, K. LIU, AND L. NAKHLEH, *Maximum likelihood inference of reticulate evolutionary histories*, Proc. Natl. Acad. Sci. USA, 111 (2014), pp. 16448–16453. (Cited on p. 265)
- [381] Y. YU, C. THAN, J. DEGNAN, AND L. NAKHLEH, *Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting*, Syst. Biol. 60 (2011), pp. 138–149. (Cited on p. 232)
- [382] G. U. YULE, *A mathematical theory of evolution: Based on the conclusions of Dr. J. C. Willis, F.R.S.* Philos. Trans. R. Soc. London B, 213 (1925), pp. 21–87. (Cited on p. 205)
- [383] K. A. ZARETSKII, *Constructing a tree from the set of distances between the pendant vertices*, Uspekhi Mat. Nauk, 20 (1965), pp. 90–92. (Cited on p. 112)
- [384] L. ZHANG, *From gene trees to species trees II: Species tree inference by minimizing deep coalescence events*, IEEE/ACM Trans. Comput. Biol. Bioinf., 8 (2011), pp. 1685–1691. (Cited on pp. 231, 232)
- [385] L. ZHANG, *On tree based phylogenetic networks*, J. Comput. Biol., 23 (2016), pp. 553–565. (Cited on pp. 259, 261)
- [386] P. ZHANG, X. ZHAN, N. A. ROSENBERG, AND S. ZÖLLNER, *Genotype imputation reference panel selection using maximal phylogenetic diversity*, Genetics, 195 (2013), pp. 319–330. (Cited on p. 133)
- [387] P. ZWIERNIK, *Semialgebraic statistics and latent tree models*, Monographs on Statistics and Applied Probability 146, CRC Press, Taylor & Francis Group, Boca Raton, FL, 2016. (Cited on pp. 191, 196)

Index

- age (of a species), 218
- agreement forest, 31
- Aldous β -splitting model, 55
- algebraic set/variety, 192
- antichain, 258
- Apresjan cluster, 114
- automorphism, 7
- Baker–Campbell–Hausdorff (BCH) formula, 157
- balanced minimum evolution (BME), 126
- Bayesian star paradox, 183
- biconnected component, 247
- Billera–Holmes–Vogtmann (BHV) space, 131
- block
 - of a network, 247
 - of a partition, 3
- budgeted nature reserve problem (BNRP), 138
- BUILD algorithm, 68
- Buneman graph, 26
- capture, 92
- Catalan number, 28
- caterpillar, 11
- Cayley’s formula, 15
- center, 8
- central symmetry edge, 8
- central symmetry vertex, 8
- centroid, 7
- character, 87
 - binary, 87
 - partial binary, 98
 - strongly compatible, 89
 - trivial, 89
- cherry, 4
- circular distance function, 243
- circular ordering, 28
- circular split system, 27
- clique, 3
- cluster, 18
- cluster network, 253
- coalescent point process (CPP), 217
- coalescent time units, 225
- Colless index, 53
- compatible set of trees, 21
- compressed network, 251
- conditioned critical branching process, 220
- consensus
 - Adams, 38
 - frequency-difference, 37
 - loose, 38
 - majority (+), 37
 - strict, 21
- consensus function
 - anonymity of a, 39
 - neutrality of a, 39
- consistency (of distance based tree reconstruction), 121
- continuous realization, 150
- convex partition, 88
- covariant model, 151
- cover digraph, 19, 252
- cyclic permutation, 27
- cyclomatic number, 5
- decisive sets, 82
- deep coalescent cost, 231
- define (a tree), 73
- dense set of rooted triples, 263
- determining set \mathcal{L} , 127
- digraph, 6
- disentangling trees, 85
- display (a tree), 64
 - a set of trees, 65
 - by a directed network, 254
 - by an undirected network, 239
- display graph, 71
- distance function, 113
- diversification rate r , 212
- diversity function, 145
- diversity indices, 138
- double-cut-and-join (DCJ) operation, 121
- double-start tree, 206
- edge
 - interior, 4
 - pendant, 4
 - subdividing, 4
- equal input model, 158
- equal splits (ES) index, 139
- equivalent phylogenetic networks, 238
- equivalent phylogenetic trees, 9
- evolutionary distance, 150
- evolutionary distinctiveness (ED), 139
- excess of set of trees, 70
- excess-free set of trees, 75
- exchangeability property (EP), 52
- expansion (of a network), 253
- fair proportion (FP) index, 139
- Fano’s lemma, 178
- Fibonacci sequence, 96, 104, 194
- field of bullets model, 143
 - simple, 143
- Fitch–Hartigan algorithm, 100
- forest, 4
- four-point condition, 112
- Galton–Watson branching process, 199
- gamma statistic, 220
- gene tree, 224
 - anomalous, 227
- general Markov model, 154

- general time-reversible (GTR) model, 156
- graph, 3
- bipartite, 3
 - chordal, 5
 - diameter of, 3
 - directed, 6
 - directed acyclic (DAG), 6
 - k -partite, 3
- Gromov–Farris transform, 117
- group elimination, 52
- Hadamard matrix, 167
- Hall’s theorem, 259
- Hamming distance, 119
- hardwired cluster, 256
- Hellinger distance, 177
- Helly property, 4
- hierarchy, 18
- Hilbert’s basis theorem, 192
- Hilbert’s Nullstellensatz, 192
- homoplasy score, 102
- homoplasy-free, 87
- homotopy, 132, 190
- hybridization, 232
- hybridization network, 252
- hyperbolicity, 132
- identifiability condition, 179
- identify (a phylogeny), 94
- incompatible clusters, 36
- incomplete directed perfect phylogeny haplotyping (IDPP), 99
- incomplete lineage sorting (ILS), 224
- irreducible Markov process, 151
- isometric function, 240
- Jacobi’s identity, 150
- Jukes–Cantor (JC69) model, 156
- Kalmanson metric, 129
- Kimura’s three-substitution (K3ST) model, 156
- Kimura’s two-substitution (K2ST) model, 156
- labeled history, 45
- lateral gene transfer (LGT), 224, 232
- network, 260
- leaf of a network, 238
- leaf-centroid, 8
- least common ancestor (LCA), 6
- least common ancestor (LCA)
- mapping, 230
- Linnean classification, 18
- lowest stable ancestor, 263
- majority rule, 187
- ancestral estimate, 104
 - consensus tree, 20
- Markovian property, 52
- matching, 259
- matrix representation with parsimony (MRP), 109
- maximal agreement set (MAS), 79
- maximal agreements subtree (MAST), 79
- maximum agreement forest (MAF), 31
- maximum compatibility, 100
- maximum compatible tree, 80
- maximum likelihood (ML)
- estimation, 147
 - tree, 180
- maximum minimum distance (MMD), 135
- maximum parsimony (MP) tree, 107
- median hull, 244
- median vertex, 5
- Menger’s theorem, 101, 105
- minimal extension, 100
- minimum evolution
- character-based, 100
- model invariant, 193
- molecular clock, 157
- multispecies coalescent model, 225
- multivariate Lagrange inversion, 107
- mutual information, 178
- Möbius function, 191, 198
- Möbius inversion, 198
- nearest neighbor interchange (NNI), 29
- neighbor joining (NJ), 121
- Neighbor-Net, 242
- Noah’s Ark problem (NAP), 144
- nondegenerate (set of trees), 66
- orbit-stabilizer theorem, 41
- Pólya’s urn, 44, 58
- pairwise compatibility, 23
- parallel edges, 241
- parsimony operation, 100
- parsimony score, 100
- partition intersection graph, 90
- partition of sets, 3
- patchwork, 76
- path, 3
- pattern, 165
- perfect phylogeny, 88
- perfect phylogeny haplotyping (PPH), 99
- persistent perfect phylogeny (PPP), 98
- Petersen graph, 33, 132
- phylogenetic
- ideal, 193
 - invariant, 192
 - mixture on T , 171
 - network, 238
 - terrace, 83
 - variety, 193
- phylogenetic X -tree
- binary, 10
 - restricted, 63
 - rooted, 9
 - unrooted, 10
- phylogenetic diversity (PD), 133
- index, 138
 - over Abelian groups, 144
 - unrooted, 133
- phylogenetic forest
- unrooted binary, 18
- phylogenetics, 1
- phylogeny, 12
- pigeonhole principle, 20
- product partition, 38
- profile of phylogenies, 36
- pyramid, 130
- indexed, 130
- quartet metric, 80
- quartet tree, 12, 65
- quaternary relation, 27
- random cluster model, 161
- ranked tree
- oriented, 46
 - unlabeled binary, 48
- ranking function, 45
- redundant arc, 251
- refinement
- of a partition, 3
 - of a tree, 21
- restriction (of a tree), 63
- reticulate arc, 246
- reticulation vertex, 245

- reticulation-visible network, 249
Riccati equation, 212
Robinson–Foulds (RF) distance, 20, 25
rooted subtree and prune (rSPR), 32
rooted triple, 65
Sackin index, 53
safety radius
 edge, 124
 l_∞, l_2 , 124
 stochastic, 124
sampling consistency, 52
Sauer–Shelah lemma, 23
Shapely value index, 140
Shapley value, 141
shellability, 131, 191
single-start tree, 206
site-to-site rate variation, 173
slim set of trees, 70
softwired cluster, 256
span, 65
species tree, 224
split
 trivial, 23
 X -, 23
split graph, 240
split network, 241
star tree, 4
stationary
 distribution, 149
 model, 151
Stirling’s approximation, 16
strict consensus MRP supertree, 109
subdominant ultrametric, 115
submodular function, 137
substitution probability, 164
subtree prune and regraft (SPR), 30
suppressing vertices of degree 2, 4
symbolic ultrametric, 117
symmetry group, 7
tanglegram, 43
temporal
 labeling, 250
 network, 250
ternary relation, 22
topology invariant, 193
trace of a matrix, 150
transition matrix, 148
tree, 4
 binary, 6
 perfect, 11
 refinement of a, 21, 24
 rooted, 6
 star, 4
tree arc, 246
tree bisection and reconnection (TBR), 30
tree containment problem (TCP), 255
tree metric, 111
tree representation, 111
tree vertex, 245
tree-child network, 247
tree-sibling network, 248
treewidth, 6
trinet, 263
triplet cover, 128
trivial clusters, 19
trivial invariant, 194
Tuffley poset, 190
ultrametric, 113
ultrametric edge lengths, 114
uniform model, 50
uniform Poisson model for LGT, 234
unlabeled ranked tree (URT)
 distribution, 219
valid extension (F), 87
vertex representation of an ultrametric, 114
weak hierarchy, 23
Wedderburn–Etherington number, 42
Wiener index, 31
 X -forest, 191
 X -split, 23
 X -tree, 12
Yule–Harding (YH) distribution
 model, 43
unrooted, 57

Phylogenetics is a topical and growing area of research. Phylogenies (phylogenetic trees and networks) allow biologists to study and graph evolutionary relationships between different species. These are also used to investigate other evolutionary processes—for example, how languages developed or how different strains of a virus (such as HIV or influenza) are related to each other.

This self-contained book addresses the underlying mathematical theory behind the reconstruction and analysis of phylogenies. The theory is grounded in classical concepts from discrete mathematics and probability theory as well as techniques from other branches of mathematics (algebra, topology, differential equations). The biological relevance of the results is highlighted throughout.

In *Phylogeny: Discrete and Random Processes in Evolution*, the author

- supplies proofs of key classical theorems and includes results not covered in existing books,
- emphasizes relevant mathematical results derived over the past 20 years, and
- provides numerous exercises, examples, and figures.

This book is intended for applied mathematicians, biomathematicians, discrete mathematicians, systematic biologists, computer scientists specializing in algorithms and bioinformatics, statisticians specializing in stochastic processes, researchers working in probability theory, and scholars studying the philosophy of biology.

Mike Steel is a Professor in the School of Mathematics and Statistics at the University of Canterbury and Director of its Biomathematics Research Centre. He is an elected fellow of the Royal Society of New Zealand and a recipient of the NZ Mathematical Society's annual Research Award. His research interests include combinatorics and random processes and their applications to questions in evolutionary biology and related areas of sciences, which in biology have mainly concerned phylogenetic theory and methods. Additional research interests include autocatalytic networks in origin of life, inverting random functions in mathematical statistics, and questions in the philosophy of science concerning causality and information loss. He has published approximately 240 academic papers, co-authored two books on phylogenetics, and served as associate editor of various journals, including *Bulletin of Mathematical Biology* and *Systematic Biology*.



For more information about SIAM books, journals,
conferences, memberships, or activities, contact:



Society for Industrial and Applied Mathematics
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA
+1-215-382-9800 • Fax: +1-215-386-7999
siam@siam.org • www.siam.org



Phylogeny
Discrete and Random Processes in Evolution

MIKE STEEL

Phylogeny Discrete and Random Processes in Evolution

MIKE STEEL

University of Canterbury
Christchurch, New Zealand

CBMS 89

SIAM

CBMS-NSF
REGIONAL CONFERENCE SERIES
IN APPLIED MATHEMATICS

SPONSORED BY
CONFERENCE BOARD OF
THE MATHEMATICAL SCIENCES

SUPPORTED BY
NATIONAL SCIENCE
FOUNDATION

CB89