

Phylogeny: Discrete and random processes in evolution

Mike Steel

March 21, 2016

Contents

Preface	v
1 Phylogeny	1
1.1 What is phylogenetics?	1
1.2 Preliminaries	3
1.2.1 Generic notation	3
1.2.2 Graphs and trees	3
1.2.3 Directed graphs and rooted trees	7
1.2.4 Symmetries and ‘centers’ of trees	7
1.3 Phylogenetic trees	10
1.3.1 Key phylogenetic notation	14
2 Basic combinatorics of discrete phylogenies	15
2.1 Counting trees	15
2.2 Rooted trees as nested sets of clusters	18
2.2.1 Hierarchies	18
2.2.2 First applications	20
2.3 Refinement, compatibility and encoding	22
2.3.1 Encoding phylogenies by ternary relations	22
2.4 Unrooted trees as systems of splits	24
2.4.1 Unrooted analogs of hierarchy results	26
2.4.2 Three more ways to encode an unrooted phylogeny	29
2.5 Tree rearrangement metrics	31
2.5.1 Surgery operations on trees (NNI, SPR, TBR)	31
2.5.2 Properties of (discrete) tree space	35
2.6 Consensus functions	38
3 Tree shape and random discrete phylogenies	45
3.1 Tree shapes	45
3.2 The shape of evolving trees	48
3.2.1 The big picture	51
3.2.2 Properties of the YH and uniform models	54
3.2.3 Exchangeability and sampling consistency	57
3.3 Measuring and modelling tree shape	58
3.3.1 Balance indices (Colless and Sackin)	58
3.3.2 The Aldous β -splitting model	61
3.3.3 Models on unrooted phylogenies	63
3.4 Cherries and extended Pólya urn models	64

	3.4.1	The Robinson-Foulds distance to a random tree	67
4		Pulling trees apart, and putting trees together	69
	4.1	Restriction and display	69
	4.1.1	The span of a set of trees, and compatibility	71
	4.1.2	Quartet trees and rooted triples	72
	4.2	When is a collection of trees compatible?	74
	4.2.1	The $\mathcal{A}_{\mathcal{R}}$ tree for rooted phylogenies	74
	4.2.2	Compatibility of unrooted phylogenies	76
	4.2.3	The display graph for a set of trees	78
	4.3	Sets of trees that ‘define’ and ‘identify’ a phylogeny	80
	4.3.1	Defining a tree	80
	4.3.2	The Böcker–Dress–Grünwald theorem	83
	4.3.3	Identifying a tree	86
	4.4	Agreement subtrees	87
	4.4.1	The quartet metric	88
	4.5	Phylogenetic decisiveness and terraces	90
	4.5.1	Decisiveness for random taxon coverage	93
	4.5.2	Disentangling trees	94
5		Phylogenies based on discrete characters	97
	5.1	Characters, homoplasy and perfect phylogeny	97
	5.1.1	Capturing a perfect phylogeny	103
	5.1.2	Two enumeration questions	106
	5.1.3	Random binary characters	108
	5.1.4	Extensions of the binary perfect phylogeny problem	109
	5.2	Minimal evolution (maximum parsimony (MP))	112
	5.2.1	The combinatorics of parsimony	113
	5.2.2	Ancestral state reconstruction	116
	5.2.3	Counting minimal evolution trees	117
	5.3	Minimal evolution trees for a sequence of characters	120
	5.3.1	Short encodings and super-trees	122
6		Continuous phylogenies and distance-based tree reconstruction	123
	6.1	Metrics from trees with edge lengths	123
	6.1.1	Ultrametrics and the Gromov–Farris transform	126
	6.1.2	Symbolic ultrametrics	130
	6.1.3	Distances versus characters	132
	6.1.4	Distances from genomic data	133
	6.2	Distance-based tree reconstruction methods	135
	6.2.1	Neighbor Joining (NJ)	136
	6.2.2	Balanced minimum evolution (BME)	139
	6.2.3	Tree reconstruction from partial distances	141
	6.3	Generalizations and geometry	143
	6.3.1	Indexed pyramids and Kalmanson metrics	143
	6.3.2	The geometry of tree space	145
	6.4	Phylogenetic diversity	148
	6.4.1	PD optimization and diversity indices for rooted trees	151
	6.4.2	Biodiversity conservation (‘Noah’s ark’)	158
	6.4.3	Extensions of PD	160

7	Evolution on a tree: Part one	163
7.1	Nonhomogeneous Markov chains	164
7.2	From Markov chains to processes on trees	169
7.2.1	Identifiability of the phylogeny	172
7.2.2	Models in molecular phylogenetics	173
7.3	Classes and properties of models	174
7.3.1	The equal input model	176
7.3.2	Symmetries within models	180
7.4	The Hadamard story	182
7.4.1	Application: The Felsenstein zone	188
7.5	Phylogenetic mixture models.	190
8	Evolution on a tree: Part two	195
8.1	Preliminaries	195
8.1.1	Probability metrics and information	195
8.1.2	Maximum likelihood (ML) and variants	197
8.2	Phylogeny reconstruction methods and properties	199
8.2.1	Information-theoretic bounds	203
8.2.2	The space of ‘phylogenetic oranges’	208
8.3	Algebraic analysis of Markov models	211
8.3.1	Phylogenetic invariants and inequalities.	213
8.3.2	Invariants for mixture models	217
8.4	The ‘infinite state’ random cluster model	218
8.4.1	An application using the probabilistic method	223
8.5	Additional topics	224
9	Evolution of trees	227
9.1	Yule pure-birth trees: the simplest model	228
9.1.1	Conditioning on n : A curiously exact result	230
9.1.2	Ancestral state reconstruction	232
9.2	Birth–death models	234
9.2.1	The complete tree and reconstructed tree	235
9.2.2	The ‘pull of the present’ and ‘push of the past’	238
9.2.3	Coalescent point process models, and unlabeled ranked trees	240
9.2.4	Predicting future PD loss	245
9.3	Gene trees and species trees	248
9.3.1	The multispecies coalescent and ILS	249
9.3.2	ILS and deep coalescence cost	255
9.3.3	Gene trees generated by LGT and related processes	257
10	Introduction to phylogenetic networks	263
10.1	To tree or not to tree: why networks?	263
10.1.1	Preliminaries	264
10.2	Implicit (unrooted) networks	265
10.2.1	Binary unicyclic networks and undirected binary level-1 networks.	265
10.2.2	Split networks and circular split systems	267
10.3	Explicit (directed) networks	272
10.3.1	Binary phylogenetic networks	273

10.3.2	Tree-child, tree-sibling and reticulation-visible networks	275
10.3.3	Temporal networks	278
10.3.4	Networks without redundant arcs	279
10.4	Trees displayed by networks	282
10.4.1	The tree containment problem	284
10.4.2	Tree-based networks	285
10.5	Reconstructing networks	291
10.5.1	Encoding networks by sub-structures	291
10.5.2	Minimizing reticulation	295
10.6	Additional topics	297
	Bibliography	303
	Appendix A	325
	Index	329