

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Statistical Methods of Phylogenetic Analysis:
Including Hadamard Conjugations, LogDet Transforms,
and Maximum Likelihood

A thesis presented in partial fulfillment
of the requirements for
the degree of Ph.D. in
Biology at Massey University

Peter J. Waddell
1995

Abstract

This thesis studies phylogenetics from a biological-statistical perspective. Chapter 1 offers an overview of the field, with particular emphasis upon the classification and interrelationships of phylogenetic methods. Separating tree selection criteria from 'corrections' for multiple hits is crucial to understanding the behaviour of different methods. Chapter 2 extends Hadamard conjugations to allow for a distribution of unequal rates at different sites in a DNA sequence. This can be done, with minimal additional computational effort, assuming a gamma, lognormal etc. distribution of site rates. The result is either 'correction' of observed sequences assuming a certain distribution of rates, or prediction of sequence probabilities given a distribution of rates and a tree. A new set of faster Hadamard conjugations for correcting four state data are presented. These conjugations also allow unequal rates across sites, transition to transversion weighting and fixing the transition to transversion ratio.

Chapter 3 considers the more general time reversible and LogDet-Paralinear distances. These are extended to accommodate unequal rates across sites. It is shown that removing a proportion of constant sites gives the LogDet a high degree of robustness to unequal rates across sites even if the true model is not invariant sites plus identical rates. Analyses of 16S-like rRNA with constant site removal (CSR) LogDet reveals surprising results, including good evidence that Microsporidia are the most distantly related (i.e. first branch) eukaryotes. Chapter 4 deals with understanding the sampling properties of transformations, especially the Hadamard conjugation. Results include forcing the Hadamard conjugation to the Kimura 2ST and Jukes Cantor models, thereby reducing sampling variance. In doing this families of tree informative linear invariants were found. It is also shown that replacing log functions with truncated power series can reduce sampling errors (RMSE) substantially.

Chapter 5 deals with tree selection criteria. Studies reveal some interesting inter-relationships between Hadamard conjugation, distance and maximum likelihood (ML) based methods. Calculation of likelihoods with unequal rates across sites (e.g. a gamma distribution) are also developed. This can be done quickly with Hadamard conjugations, and a variety of sequences and models are studied. ML solutions to inferring reticulate phylogenies are described, and in an application are used to infer the population size of our ancestors with chimps and gorillas. A wide variety of methods, including ML, are shown to be inconsistent in the Felsenstein zone when site rates are unequal (in a similar situation ML is also seen to be inconsistent under a molecular clock). Overcorrecting the data is also a potential pitfall, and the concept of the 'anti-Felsenstein zone' is introduced, illustrated, and developed. A related phenomena is that two or more optimal binary trees can predict exactly the same sequences when rates across sites are unequal, and examples are provided. Chapter 6 describes new statistical tests. These include faster model based resampling to evaluate fit of model to data and tests of whether two data sets came from the same tree. A Bayesian view of support for different trees is presented. The thesis is large, but well illustrated, and looking at the figures alone should provide a useful overview of new results.

Acknowledgments

I would firstly like to thank my supervisors Professors David Penny and Mike Hendy for inviting me to participate in the phylogenetic research at Massey University, and allowing me to pursue a wide field of study. Both have been generous with their time and patience over a long period. David's enthusiasm for biology is always encouraging, as is Mike's love of mathematics. Other members of the group, especially Peter Lockhart and Mike Steel have provided assistance and encouragement in too many ways to mention. Thanks to you all.

Many statisticians have indulged my interest in the topic, especially Greg Arnold (a co-supervisor) and Terry Moore, and at other important times Brian McArdle and Chris Triggs. Graeme Wake helped tremendously by hosting me for much of the time in the Department of Mathematics and Statistics. Thanks also to the people I meet on my sojourn to the USA, especially David Swofford, Jaxk Reeves, David Hillis, Dick Hudson, Joe Felsenstein, Terry Speed, Walter Fitch, Jeff Thorne, Nick Goldman, Mike Donoghue, Arend Sidow, Mitch Sogin, Masatoshi Nei, Ken Kidd, John Hartigan, Adrian Gibbs and Rebecca Cann,. Your hospitality and enthusiasm are well marked.

Special thanks to Trish McLenachan, without whose encouragement and proof reading this thesis might never have been finished.

My family were always supportive in the background, which unfortunately is the way it has been for a long time now; certainly one of the personal costs of the work.

To everyone else who has aided me through this long period I extend my thanks and apologize that I cannot mention you all here. I look forward to rejoining society and enjoying your company once again.

Contents

Abstract	iii
Acknowledgments	iv
Table of Contents	v
Glossary of Symbols and Abbreviations	xiii

1 Estimating evolutionary trees from DNA sequences

1.1 INTRODUCTION	1
1.2 THE RATIONALE FOR STATISTICS	2
1.2.1 Popperian ideals	2
1.2.2 The question Bayes tried to answer	2
1.3 TERMINOLOGY ASSOCIATED WITH EVOLUTIONARY TREES	4
1.3.1 A first taste of Hadamard conjugations	7
1.4 STOCHASTIC MODELS OF EVOLUTION	8
1.5 STEPS IN RECONSTRUCTING TREES FROM SEQUENCES	11
1.5.1 The logical structure of phylogenetic analysis	11
1.5.2 Statistical testing of phylogenetic hypotheses	15
1.6 DESIRABLE CHARACTERISTICS OF TREE BUILDING METHODS	18
1.7 CURRENT METHODS OF TREE ESTIMATION	21
1.7.1 Maximum likelihood on sequences	22
1.7.2 Parsimony, compatibility and closest tree	23
1.7.3 Distance based methods	25
1.7.4 Phylogenetic invariants	26
1.7.5 Invariants by invertible transformation of sequences	27
1.7.6 Other ways to classify tree estimation methods	28
1.8 MAIN RESULTS OF THIS THESIS	30
1.9 DATA SETS ANALYSED IN THIS THESIS	32
1.9.1 A subset of Lake's alignment of rRNA molecules	32
1.9.2 A long stretch of mtDNA from apes	33
1.9.3 Gouy an Li's alignment of diverse 16S-like rRNA sequences	34
1.10 OVERVIEW AND COMPUTER SOFTWARE USED	34

2 Extending Hadamard conjugations to model unequal rates across sites

2.1 INTRODUCTION	37
2.2 HADAMARD CONJUGATIONS REVIEWED	39
2.2.1 Definitions and a worked example	40
2.3 HADAMARD CONJUGATIONS WITH UNEQUAL RATES ACROSS SITES	44
2.3.1 Discrete distributions	44
2.3.2 Continuous distributions	45

2.3.3 Closed form pathset correction formulae.....	47
2.3.4 Multi-modal distributions of rates across sites	53
2.3.5 Analysing transversional changes with extended Hadamard conjugations	56
2.4 RATES ACROSS SITES 4-STATE HADAMARD CONJUGATIONS	59
2.4.1 A review of the i.r. 4-state Hadamard conjugation	59
2.4.2 Consistency of the extended 4-state Hadamard conjugation	62
2.4.3 Unequal rates across sites causing inconsistency of tree selection.....	62
2.5 SEQUENCE DATA ANALYSED WITH EXTENDED 4-STATE CONJUGATIONS	65
2.5.1 Analysis of 5kb of mtDNA relating to human origins	65
2.5.2 Analysis of anciently diverged rRNA sequences.....	68
2.6 SEPARATING SITES INTO RATE CLASSES TO AVOID INCONSISTENCY	70
2.7 DISCUSSION	71

APPENDICES TO CHAPTER 2:

A2.1 Proof of the inconsistency of Hadamard conjugations if sites change their relative rates	74
A2.2 Proof of equation 2.3.2-3: The consistency of extended Hadamard conjugations.	75
A2.3 Deriving moment generating functions while fixing the mean to one	76
A2.4 The moment generating function of translated distributions	77
A2.5 A closed form correction formula for a trimodal distribution.....	78
A2.6 Order 2^{l-1} Hadamard conjugations for 4-state data.....	79
A2.6.1 Corrected pathset lengths under different models.....	80
A2.6.2 Multiplication of corrected pathset length vectors to obtain γ vectors	82
A2.6.3 Counting changes on higher order pathsets, and proving consistency.....	83
A2.6.4 Applications to data.....	84

3 Modifying LogDet distances to cope with unequal rates across sites

3.1 INTRODUCTION	87
3.2 FUNDAMENTAL EQUATIONS OF A MARKOV PROCESS ON A TREE.....	89
3.2.1 A general distance estimate for time reversible models.....	92
3.2.2 A distribution of rates across sites with the general time reversible distance	94
3.3 DISTANCE ESTIMATION UNDER NON-STATIONARY MODELS.....	97
3.3.1 LogDet distance measures including new results on their interpretation	97
3.3.2 Approximate methods to give robustness with varying base composition	105
3.4 CONSISTENCY AND ROBUSTNESS UNDER A NON-STATIONARY MODEL.....	108
3.4.1 A model of non-stationary evolution	108
3.4.2 Inconsistency when using the Barry and Hartigan asynchronous distance.....	110
3.4.3 Oh No! Long edges can repel	112
3.4.4 A brief history of LogDet distances.....	116
3.5 MAKING LOGDET DISTANCES ROBUST TO RATES ACROSS SITES	118
3.5.1 Four ways to modify distances to be consistent under invariant sites models	118

3.5.2 A direct look at the robustness of the invariant sites-LogDet method	121
3.6. IMPORTANT PRELIMINARY STEPS IN ANALYSING SEQUENCES	125
3.6.1 Studying the base composition of rRNA	126
3.6.2 Five different types of method to infer the number of invariant sites.....	130
3.6.3 A new capture-recapture method suitable for rRNA.....	130
3.6.4 Using “observed” numbers of changes to infer p_{inv}	135
3.6.5 Inferring p_{inv} with a ML model of sequence evolution	138
3.6.6 Estimating p_{inv} by directly measuring additivity of distances on a tree	140
3.6.7 The Bealey theorem inequality.....	141
3.6.8 Summary of diagnosing this 16S-like rRNA.....	143
3.7 FIELD TRIALS OF THE INVARIANT SITES-LOGDET TRANSFORMATION	143
3.7.1 Six prespecified hypotheses about the “tree of life”	144
3.7.2 Using the bootstrap as a guide to statistical support	146
3.7.3 Support for our six hypotheses with the invariant sites-LogDet transform	147
3.7.4 The overall invariant sites LogDet “tree of life”	153
3.7.5 The relative performance of ML and parsimony methods on this data.....	158
3.7.6 Does an analysis of just transversional changes help?	160
3.7.7 The validity of grouping transversions.....	162
3.8 CHECKING THE INVARIANT SITES-LOGDET TREE RESULTS	163
3.8.1 Using just the most conserved informative sites to avoid model uncertainties ...	163
3.8.2 What level of bootstrap support is significant on our tree?	168
3.8.3 Other sequences supporting Microsporidia as earliest diverging eukaryotes	169
3.8.4 Results of the application of split decomposition to this data.....	170
3.9 ROBUSTNESS VIA CLASSIFICATION OF SITES INTO RATE CLASSES	173
3.10 DISCUSSION	175
3.10.1 Earliest eukaryotic evolution reconsidered	175
3.10.2 The need to study variances and bias	178
3.10.3 Miscellaneous discussion	179
3.10.4 Speculation on compositional bias effects in proteins and rRNA.....	182
3.10.5 Invariant sites LogDet transforms: A most useful distance estimate.....	183
APPENDICES TO CHAPTER 3:	
A3.1 Proof that all 2-state transition matrices can be considered the result of a continuous time process.....	184
A3.2 Proof of the identity of averaged “asynchronous distances”, LogDet and paralinear distances.....	184
A3.3 Proof that F is symmetric under time reversibility and the clock	185
A3.4 Proof that any two distance matrices additive on the same unweighted tree are still additive when combined.....	186

4 Sampling errors associated with transformed data

4.1 INTRODUCTION	187
4.2 CALCULATING THE VARIANCE-COVARIANCE MATRIX OF PHYLOGENETIC SPECTRA	190
4.2.1 Our illustrative model	190
4.2.2 The variance-covariance matrix $V[\hat{S}]$ of the sequence spectrum \hat{S}	191
4.2.3 The calculation of \hat{r} ($= H\hat{S}$) and its covariance matrix $V[\hat{r}]$	192
4.2.4 The covariance matrix $V[\hat{\rho}]$ of the estimated path lengths $\hat{\rho}$	193
4.2.5 The covariance and correlation matrix of $\hat{\gamma}$, the corrected data	195
4.3 THE MARGINAL DISTRIBUTIONS OF ENTRIES IN $\hat{\gamma}$	199
4.4 PROPERTIES OF DELTA METHOD COVARIANCE MATRICES	202
4.4.1 Bias in entries in $V[\hat{\gamma}]$ and $C[\hat{\gamma}]$, estimated with $s(T)$	202
4.4.2 Error and bias in $\hat{V}[\hat{\gamma}]$ estimated from random samples, \hat{S}	203
4.5 ESTIMATING $V[\hat{\gamma}]$ WHEN COMPENSATING FOR UNEQUAL RATES ACROSS SITES	204
4.5.1 First derivatives of closed form URAS correction formulae.....	205
4.5.2 How unequal rates across sites affect accurate distance estimation	206
4.5.3 Knowing the model, we can estimate even very large distances accurately	213
4.5.4 Can data editing improve consistent tree building methods?	216
4.6 NEW 4-STATE HADAMARD CONJUGATIONS TO REDUCE VARIANCE	222
4.6.1 Kimura 2ST and Jukes-Cantor 4^{t-1} Hadamard conjugations.....	222
4.6.2 Linear tree invariants in the Kimura 2ST and Jukes-Cantor model	226
4.6.3 Calculating the covariance matrix of $\hat{\gamma}_{K2}$ and $\hat{\gamma}_{IP}$	228
4.6.4 Testing difference in fit between \hat{S}_{K2} , \hat{S}_{IP} , and \hat{S}_{K3}	230
4.6.5 Reduced variance and bias by using new pathlength transformations.....	232
4.7 SAMPLING ERRORS OF $\hat{\gamma}_D$, THE DISTANCE HADAMARD.....	233
4.7.1 The covariance matrix of $\hat{\rho}_D$	234
4.7.2 A comparison of the structure in $V[\hat{\gamma}_D]$ vs $V[\hat{\gamma}]$	237
4.7.3 The statistical structure of $\hat{\gamma}_D$ vs $\hat{\gamma}$ evaluated on a six taxon tree	239
4.7.4 Variations on the distance Hadamard	245
4.8 THE MEANING OF $\hat{\gamma}$ AND ESTIMATING TREE SELECTION PROBABILITY	246
4.8.1 Are Hadamard conjugations ML estimators?	246
4.8.2 Tree selection probabilities estimated via the sampling distribution of $\hat{\gamma}$	250
4.9 CONCLUSION.....	254
APPENDICES TO CHAPTER 4:	
A4.1 The calculation of HVH.....	257
A4.2 An unbiased and reduced variance transformation of $\hat{r} \rightarrow \hat{\rho}$	257

A4.2.1	New pathlength transformations applicable when \hat{r}_i is negative.....	258
A4.2.2	The contribution of bias to stochastic error in pathlength estimators	259
A4.2.3	The reason for the often large RMSE of the rb estimator	263
A4.2.4	The region where the rb estimator has the best RMSE	264
A4.2.5	Accuracy of estimating large distances with small sequences	271
A4.2.6	Accuracy of delta method variance estimates for very short sequences ...	273
A4.2.7	Discussion.....	274

5 Properties of tree selection criteria

5.1	INTRODUCTION.....	277
5.2	TREE SELECTION OPTIMALITY CRITERIA FOR $\hat{\gamma}$	279
5.2.1	Some important properties of $\hat{\gamma}$ with respect to tree selection	279
5.2.2	Some real data to illustrate tree selection criteria.....	280
5.2.3	Ordinary (or unweighted) Least Squares (OLS).....	283
5.2.4	What is closest tree?	284
5.2.5	Weighted Least Squares (WLS)	285
5.2.6	Generalised Least Squares (GLS).....	287
5.2.7	Maximum likelihood tree selection from $\hat{\gamma}$	292
5.2.8	How many likelihood optima per tree?	295
5.2.9	Comparing GLS on sequences with GLS on distances	296
5.2.10	Statistical properties of compatibility and parsimony applied to gamma.	299
5.2.11	Non-iterated likelihood and non-iterated X^2	300
5.2.12	Statistically efficient criteria to choose amongst the best trees.....	301
5.2.13	Using these methods to select a consensus tree from bootstrap proportions	302
5.3	ML TREE SELECTION FROM THE OBSERVED SEQUENCES	302
5.3.1	Calculating likelihood via Hadamard conjugations.....	303
5.3.2	Finding the maximum likelihood point of a specific tree.....	309
5.3.3	Branch and bound of maximum likelihood	312
5.3.4	Maximum Likelihood with a distribution of rates across sites.....	319
5.3.5	ML models where sites change their rate class	326
5.3.6	Results with ML models that allow distributions of rates across sites.....	330
5.3.7	ML analysis of four ancient rRNA sequences.....	332
5.3.8	Properties of parameter estimates under URAS ML models.....	336
5.3.9	ML analysis of Hominoid mtDNA	340
5.3.9.1	Other results on this mtDNA data	349
5.3.9.2	Concluding remarks to these ML single tree analyses	352
5.3.10	Approximate likelihood via approximations to Hadamard conjugations	352
5.4	RETICULATE EVOLUTION IN PHYLOGENETICS	354
5.4.1	A likelihood model of reticulate evolution.....	354
5.4.2	ML methods to estimate degrees of ancestral polymorphism	352

5.4.2.1 Examining the human-chimp-gorilla divergence	358
5.4.2.2 Estimating ancestral diversity free of the effect of multiple hits	359
5.4.2.3 Solving for ancestral population size	363
5.4.2.4 Testing the adequacy of this model and our conclusions	365
5.4.2.5 What this population size estimate may be telling us about human evolution.....	366
5.5 ROBUSTNESS OF TREE SELECTION IN THE FELSENSTEIN ZONE.....	369
5.5.1 Robustness to URAS of parsimony and neighbor joining	370
5.5.2 Robustness of WLS methods of tree selection from $\hat{\gamma}$ and δ	374
5.5.3 Maximum likelihood is inconsistent when there are invariant sites	377
5.5.4 Inconsistency with a continuous distribution of unequal rates across sites	380
5.5.5 Different trees can give identical sequences!.....	384
5.6 ROBUSTNESS OF TREE SELECTION CRITERIA IN THE ANTI-FELSENSTEIN ZONE	385
5.6.1 The anti-Felsenstein zone.....	386
5.6.2 Long edges repel effects with simple criteria applied to γ , $\gamma(d)$ and δ	391
5.6.3 Performance of weighted least squares methods from γ and δ	393
5.6.4 "Goodness-of-fit criteria" measured on the observed sequences	394
5.6.5 Summary of tree selection in the anti-Felsenstein zone and its implications	395
5.7. INCONSISTENCY OF ML IN THE HENDY-PENNY ZONE.....	399
5.7.1 The Hendy-Penny zone	399
5.7.2 The Hendy-Penny zone with unequal rates of change across sites	401
5.7.3 Showing ML to be inconsistent in the Hendy-Penny zone with URAS	401
5.8 STATISTICAL EFFICIENCY OF TREE SELECTION ON s , $\gamma(s)$ AND $\gamma(D)$	404
5.8.1 A six taxon tree model to evaluate tree selection procedures	405
5.8.2 Features of tree selection from \hat{s}	405
5.8.2 Comparative performance of tree selection from $\hat{\gamma}(s)$ and $\hat{\gamma}(d)$	407
5.8.3 Tree selection on \hat{s} , $\hat{\gamma}(s)$, and $\hat{\gamma}(d)$ when rates at sites are unequal.....	413
5.9 OPTIMISATION AND TREE SELECTION WITH URAS	415
5.9.1 General trends in fitting a distribution of rates across sites	417
5.9.2 Optimising the shape of a distribution of rates across sites using $\hat{\gamma}$ and δ	418
5.9.3 Optimisation by fit measured at the s level.....	420
5.10 DISCUSSION	422
APPENDICES TO CHAPTER 5:	
A5.1 Two or more trees can predict identical sequence data	426
A5.1.1 Different trees can give the same sequences: A simple example with 4 taxa.....	426
A5.1.2 Different binary trees can give the same sequences!	428
A5.1.3 Simplifying 4-taxon binary trees to have fewer unique pathset lengths...	430
A5.1.4 What happens with more taxa, or using just pairwise distances?	432

A5.1.5 Where can correction curves cross	433
A5.1.6 Discussion (to appendix)	434

6 Statistical tests

6.1 INTRODUCTION.....	437
6.2 OVERALL FIT OF DATA TO THE MODEL.....	437
6.2.1 Measuring overall fit	437
6.2.2 Factors which distort the asymptotic distribution of fit statistics.....	439
6.2.3 Ways to overcome sparseness distorting asymptotic expectations	443
6.2.4 Overall goodness-of-fit statistics not necessarily reliable	445
6.2.5 Some aspects of overall fit of data to $\hat{\gamma}$	446
6.2.6 Guides to selecting a well fitting model	446
6.2.7 Modifying Monte Carlo simulations to avoid possible parameter biases	447
6.3 TESTS OF THE GENERAL SUITABILITY OF A PHYLOGENETIC METHOD	448
6.3.1 The expectation of equally well-fitting suboptimal trees.....	448
6.3.2 The general distribution of the likelihoods of trees under a reliable model.....	450
6.3.3 Extensions to split decomposition	451
6.3.4 A sign test for the fit of $\hat{\gamma}$ to model expectations	452
6.4 COMMENTS ON THE BOOTSTRAP	453
6.4.1 Approximately estimating the bias in bootstrap support for edges in a tree	453
6.4.2 The number of alternative trees and bootstrap bias.....	454
6.4.3 Subtree extraction to counter conservatism when adding extra taxa	455
6.5 TESTING FOR SPECIFIC DEPARTURES FROM THE MODEL	456
6.5.1 The fit of individual entries in $\hat{\gamma}$	456
6.5.2 Comparing actual and predicted numbers of observed changes per site.....	459
6.5.3 Testing for an excess of changes predicted on external edges in the tree	460
6.5.4 Evaluating numbers of parallel and convergent substitutions	461
6.5.5 The number of states shown at each site	461
6.5.6 Testing for evidence of trapped ancestral polymorphism	462
6.5.7 Testing the molecular clock.....	462
6.6 OBTAINING A CONFIDENCE SET OF TREES	463
6.7 TESTING WHETHER TWO DATA SETS EVOLVED BY THE SAME TREE	465
6.7.1 Did two sets of data evolve by the same processes?	466
6.7.2 Testing: "Did two data sets evolve according to the same weighted tree?"	467
6.7.3 Did two data sets evolve on a weighted tree with the same relative edge lengths?.....	468
6.7.4 "Did these data sets evolve on the same tree?"	469
6.8 CONFIDENCE LIMITS ON FEATURES OF EVOLUTIONARY MODELS	471
6.8.1 Confidence limits parameters associated with the substitution mechanism.....	471
6.8.2 Confidence limits on features of weighted trees	472
6.8.3 Confidence limits for a ratio of edge lengths	472

6.8.4 Differences in p_{inv} or shape parameters from different data sets	473
6.9 COMPREHENSIVE STANDARD ERRORS FOR DIVERGENCE TIMES.....	473
6.10 A BAYESIAN VIEW OF PHYLOGENETIC ANALYSES	476
6.10.1 The need to integrate different sources of knowledge	476
6.10.2 Setting up the prior.....	476
6.10.3 A worked example based on the archaebacteria question.....	477
6.10.4 Integrating prior and experimental results to update hypothesis support	477
6.10.5 Using resampling schemes to asses the 'likelihood' of different trees	478
6.11 DISCUSSION	480

7 Discussion and overview

7.1 INTRODUCTION	483
7.2 QUESTIONS FOR THE FUTURE.....	484

Bibliography	487
---------------------------	------------

Glossary of symbols and abbreviations

Generally, special symbols have a common meaning throughout the thesis, although some, of necessity, have multiple uses. If a generally defined variable such as c (usually being the sequence length) is used in a different context (e.g. as the shape parameter of the Weibull distribution), then this second usage will be indicated specifically in the text. A short list of symbols specific to chapter 3 is given at the end of section 3.1.

<i>Symbol</i>	<i>Definition</i>
α	The probability of a type 1 error in a statistical test (that is, rejecting the null hypothesis when it is correct).
δ	A transformed distance, or an additive distance (if in bold a distance matrix) (except appendix 3.1, where it refers to the delta method approximation)
γ	A vector description of pattern frequencies taking account of the effect of multiple hits with the Hadamard conjugation (or in the case of γ_D an approximation to this)
$\gamma(T)$	A vector description of a weighted tree
γ_D	A spectrum estimated from just pairwise distances
$\hat{\gamma}$	An estimate of γ based on a sample of sites
$\hat{\gamma}(T)$	A tree inferred from $\hat{\gamma}$
Γ	The gamma probability distribution (usually of the λ_j)
λ_j	The relative substitution rate of sites in set j
c	The sequence length (or shape parameter of the Weibull distribution, section 2.3.3)
c.v.	Coefficient of variation (standard deviation / mean)
d	The shape parameter of the inverse Gaussian distribution
d	A distance
d_{obs}	An observed distance
d.f.	Degrees of freedom in the χ^2 distribution
\mathbf{f}	A vector of observed site pattern frequencies = cs
k	The shape parameter of the Γ distribution
kb	One thousand nucleotide base pairs
G^2	The log-likelihood ratio goodness of fit statistic (the G statistic of Sokal and Rohlf, 1981).
H	A Hadamard matrix
i.r.	Identical rates of substitution across sites
M	A moment generating function, e.g. $M_\lambda(t) = E[e^{t\lambda}]$, with inverse M^{-1}
ML	Maximum likelihood

p_{inv}	A proportion of invariant sites (sites which cannot undergo substitution)
\mathbf{s}	A vector of observed site pattern probabilities (which sum to 1) = \mathbf{f} / c
$\mathbf{s}(T)$	A vector of observed site pattern probabilities generated by a particular tree evolutionary model
$\hat{\mathbf{s}}$	Observed site pattern probabilities (proportions) estimated from a sample
s.d.	Standard deviation
t	The number of taxa (or a dummy variable for a moment generating function in chapter 2, or a time scalar in chapter 3, as specifically indicated)(as a superscript to a matrix it means transpose)
tr/tv	Transition to transversion ratio
T_{12}	The tree ((1,2), 3, 4)
T_{star}	The unresolved tree (1, 2, 3, 4)
\mathbf{V}	A variance-covariance matrix
var	Variance
equipfrequency	The states are in equal proportions
OLS	Ordinary (unweighted) least squares
WLS	Weighted (usually by the inverse of the variance) least squares
GLS	Weighted least squares, taking account of correlations
SS	Sum of squares
URAS	Unequal rates across sites