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

4 Sampling errors associated with transformed data

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.

Symbol	*Definition*
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
f	A vector of observed site pattern frequencies = *c***s**
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. $M_\lambda(t) = E[e^{\lambda t}]$, 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

equifrequency	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