

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.

SOME APPLICATIONS OF STATISTICAL PHYLOGENETICS

A thesis presented in partial  
fulfilment of the requirements

for the degree

of Doctor of Philosophy

in Biomathematics at  
Massey University

Klaus Peter Schliep  
2009

Copyright © 2009 by Klaus Peter Schliep



## Abstract

The increasing availability of molecular data means that phylogenetic studies nowadays often use datasets which combine a large number of loci for many different species. This leads to a trade-off. On the one hand more complex models are preferred to account for heterogeneity in evolutionary processes. On the other hand simple models that can answer biological questions of interest that are easy to interpret and can be computed in reasonable time are favoured. This thesis focuses on four cases of phylogenetic analysis which arise from this conflict.

- It is shown that edge weight estimates can be non-identifiable if the data are simulated under a mixture model. Even if the underlying process is known the estimation and interpretation may be difficult due to the high variance of the parameters of interest.
- Partition models are commonly used to account for heterogeneity in data sets. Novel methods are presented here which allow grouping of genes under similar evolutionary constraints. A data set, containing 14 genes of the chloroplast from 19 anciently diverged species is used to find groups of co-evolving genes. The prospects and limitations of such methods are discussed.
- Penalised likelihood estimation is a useful tool for improving the performance of models and allowing for variable selection. A novel approach is presented that uses pairwise dissimilarities to visualise the data as a network. It is further shown how penalised likelihood can be used to decrease the variance of parameter estimates for mixture and partition models, allowing a more reliable analysis. Estimates for the variance and the expected number of parameters of penalised likelihood estimates are derived.
- Tree shape statistics are used to describe speciation events in macroevolution. A new tree shape statistic is introduced and the biases of different cluster methods on tree shape statistics are discussed.



## Acknowledgements

I would like to thank my supervisors Michael Hendy, Barbara Holland, David Penny and Peter Waddell for their support and advice during the time of my studies. I have been blessed to have supervisors with an enormous enthusiasm for science in general and phylogenetics in particular.

I also have to thank the Marsden Fund and the Allan Wilson Centre for financial support, which made it possible for me to study in New Zealand.

I would like to acknowledge Trish McLenachan and Gillian Gibb for their heroic effort, together with my supervisors, to proof-read this thesis and fight back my German grammar and spelling.

Many people contributed with ideas, data to the different chapters of this thesis. I have to thank Peter Lockhart and Ellen Nisbet and all other biologists who came up with challenging biological problems or supplying data. I thank Elisabeth Allman and Mark Pagel for helpful discussions about multiple optima and mixture models and Berwin Turlach for some advice on the LASSO. Some of the ideas were born or enhanced during numerous discussions with Bhalchandra, Matt, Tim, Warwick, Scott and many others, involving even more coffee.

I want to thank all the assistance from AWC staff (Joy, Susan, Karen) and IMBS (Ann, Cynthia) for doing a fabulous job. Special thanks to Tim, Warwick, Jing and Nat for taking care of my computers and software.

Thanks to all the members and visitors of the Allan Wilson centre, especially the ‘boffin lounge’, for creating such a friendly, multidisciplinary working environment during all my studies.

I must thank all the people who made my stay in Palmerston North such an enjoyable time. First I want to thank all the Latin Americans by passport, spouse or soul in Palmy. First of all my flatmate Rogerio, who put up with me for such a long time. Katia, Paul, Carlos, Matt and many others for all the good times at salsa classes or parties, churrascos or just at a coffee and cheesecake. Furthermore all members of the ‘monkeys uncle’ volleyball team and everybody I have been walking across Tongariro

with (I can't mention them all here).

I want to thank to my friends in Munich who have kept in contact with me through all this time, especially those calling during night times. The main thanks goes to my family, including a new addition, who will be mostly unaware how important their role was during all the challenges in my studies.

# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Contents</b>	<b>vii</b>
<b>List of Figures</b>	<b>xii</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Abbreviations</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Structure of the thesis . . . . .	1
1.2 Background . . . . .	5
1.2.1 Graphs, trees and networks . . . . .	5
1.2.2 Data . . . . .	8
1.2.3 Methods of tree estimation . . . . .	8
1.2.4 Tree rearrangements . . . . .	9
1.2.5 Markov models of character evolution . . . . .	11
1.3 Maximum likelihood estimation in phylogenetics . . . . .	12
1.3.1 Optimising the likelihood . . . . .	13
1.3.2 Hypothesis testing . . . . .	15
1.4 Hadamard conjugation . . . . .	17

1.4.1	Maximum Likelihood Estimation using the Hadamard conjugation	18
1.4.2	Distance Hadamard . . . . .	19
1.5	Data sets . . . . .	20
<b>2</b>	<b>Mixture models</b>	<b>22</b>
2.1	Background . . . . .	23
2.1.1	General theory of mixture models . . . . .	23
2.1.2	Identifiability of mixture models . . . . .	25
2.1.3	Mixtures to model rate heterogeneity . . . . .	26
2.1.4	Mixtures of sets of edge lengths and topologies . . . . .	28
2.1.5	Detecting partitions . . . . .	28
2.2	Methods . . . . .	29
2.3	Results . . . . .	31
2.3.1	Mixture of two trees . . . . .	31
2.3.2	Model misspecification of mixture models . . . . .	39
2.4	Conclusion . . . . .	40
<b>3</b>	<b>Multiple Optima</b>	<b>41</b>
3.1	Background . . . . .	42
3.1.1	Multiple optima in general functions . . . . .	42
3.1.2	Multiple optima on four taxon trees . . . . .	44
3.1.3	Parameter correlations and multiple optima . . . . .	47
3.2	Methods . . . . .	52
3.3	Results . . . . .	53
3.3.1	Constructing counter-examples from mixture models . . . . .	53
3.3.2	Finding multiple optima with maximum likelihood and Bayesian methods . . . . .	56
3.4	Conclusions . . . . .	58
<b>4</b>	<b>Partition Models models for multi-gene datasets</b>	<b>61</b>
4.1	Methods . . . . .	62

4.1.1	Stochastic Partitioning . . . . .	62
	Choosing the number of clusters . . . . .	65
4.1.2	GO-analysis . . . . .	66
	Relationship to other methods . . . . .	67
4.1.3	Other approaches to clustering genes . . . . .	68
4.1.4	Hadamard and distance Hadamard . . . . .	68
4.2	Results . . . . .	69
4.2.1	Yeast data . . . . .	69
4.2.2	Comparison with Gene Ontology . . . . .	75
4.2.3	Exploring relationships between genes within the chloroplast . .	78
4.3	Summary . . . . .	80
<b>5</b>	<b>Penalized least-squares and phylogenetic networks</b>	<b>83</b>
5.1	Background . . . . .	84
5.1.1	Overview of distance based methods . . . . .	85
5.1.2	Ridge regression . . . . .	89
5.1.3	The LASSO . . . . .	91
5.2	Methods . . . . .	92
5.2.1	Constructing phylogenetic network using the LASSO . . . . .	92
5.2.2	Distance Hadamard . . . . .	93
5.2.3	Choosing the number of splits . . . . .	95
5.3	Results . . . . .	96
5.4	Conclusions . . . . .	102
<b>6</b>	<b>Penalized ML for phylogenetic partitions</b>	<b>103</b>
6.1	Methods . . . . .	103
6.1.1	Example . . . . .	104
6.1.2	Moments of penalised likelihood estimates . . . . .	107
6.1.3	Optimising the penalty . . . . .	108
6.2	Results . . . . .	109

6.3	Summary . . . . .	114
<b>7</b>	<b>Biases in hierarchical clustering</b>	<b>115</b>
7.1	Methods . . . . .	116
7.1.1	Random tree generation . . . . .	116
7.1.2	Clustering Methods . . . . .	118
7.1.3	Tree Comparison Measures . . . . .	121
7.2	Results . . . . .	125
7.2.1	Simulation Study . . . . .	125
7.2.2	Case Studies . . . . .	127
7.3	Discussion . . . . .	135
<b>A</b>	<b>The R-package phangorn</b>	<b>137</b>
A.1	Mixture models . . . . .	138
A.2	Multiple optima . . . . .	139
A.3	Partition models . . . . .	139
A.4	Distance methods and penalized likelihood . . . . .	140
	<b>References</b>	<b>143</b>

# List of Figures

1.1	Trees and networks . . . . .	6
1.2	Circular splits and splits graph . . . . .	7
1.3	NNI and SPR tree rearrangements . . . . .	10
1.4	The two most frequent gene tree topologies for the Yeast data set . . . . .	21
2.1	Density and distribution function of the gamma function . . . . .	27
2.2	Likelihood for different mixtures . . . . .	32
2.3	Mixtures of trees . . . . .	32
2.4	Bootstrap . . . . .	34
2.5	Correlation matrix of the edge lengths . . . . .	36
2.6	Correlation matrix of the edge lengths for Bayesian analysis . . . . .	37
2.7	Posterior probability of the mixtures . . . . .	38
2.8	Likelihood for different mixtures . . . . .	39
3.1	A function with infinite number of multiple optima . . . . .	43
3.2	Schematic of the different possibilities for maximum likelihood optima . . . . .	45
3.3	A four taxon tree on the topology $T_{12 34}$ . . . . .	49
3.4	Estimated trees for different mixtures . . . . .	51
3.5	Mixture of two trees and resulting multiple optima . . . . .	54
3.6	Splits graph for mixture data . . . . .	55
3.7	Multiple optima for simulated data . . . . .	56
3.8	A posteriori distribution of edge weight on a multiple optima tree. . . . .	57
3.9	A posteriori distribution of edge weight on a multiple optima tree. . . . .	59

4.1	Stochastic partitioning of genes . . . . .	64
4.2	Likelihood, AIC and BIC for different partitions models . . . . .	71
4.3	Principal components for edge spectra . . . . .	72
4.4	Directed acyclic graph of the gene ontology . . . . .	76
4.5	Trees of estimated classes . . . . .	79
4.6	Principal components for edge spectra . . . . .	81
5.1	Schematic representation of bias-variance trade-off . . . . .	84
5.2	Example trees with 5 taxa . . . . .	87
5.3	Plot Edge weights in dependence of the LASSO penalty . . . . .	96
5.4	Networks for different LASSO penalties . . . . .	98
5.5	The paths of the edge weights for the distance Hadamard . . . . .	99
5.6	Comparison of splits graphs . . . . .	100
5.7	Comparison of splits graphs . . . . .	101
6.1	Three 3-taxon trees as an example to set up the penalty matrix . . . . .	104
6.2	Trees for PML. . . . .	110
6.3	Dependence between degrees of freedom and the penalty term. . . . .	111
6.4	AIC, BIC and CV for partion models. . . . .	112
6.5	Penalized Likelihood . . . . .	113
7.1	Empirical cumulative distribution function for the path length . . . . .	124
7.2	Correlations of tree measures . . . . .	126
7.3	Robinson-Foulds distances . . . . .	128
7.4	Robinson-Foulds distances . . . . .	129
7.5	Parsimony score . . . . .	130
7.6	Parsimony score . . . . .	131
7.7	Differences in the number of cherries. . . . .	132
7.8	Sackin index . . . . .	133

# List of Tables

1.1	10 sites of an alignment of 8 species of yeast. . . . .	8
3.1	Site patterns and sequence spectra . . . . .	48
3.2	Site pattern and sequence spectra . . . . .	50
3.3	Correlation matrix of edge weights . . . . .	50
4.1	Summary of runs for the stochastic partitioning algorithm. . . . .	70
4.2	Shimodaira-Hasegawa test . . . . .	74
4.3	Biological function associated to the clusters . . . . .	77
4.4	Summary of 14 different amino acid sequences of the chloroplast . . . . .	78
5.1	Design matrix for an unrooted tree and network . . . . .	86
5.2	Design matrix and contrast matrix for a rooted tree . . . . .	86
5.3	Least-squares representation for different distance methods . . . . .	88
5.4	Mallows' $C_p$ for different sized network . . . . .	97
7.1	Parsimony score and numbers of cherries for the Nickrent et al. (2002) data set . . . . .	127
7.2	Parsimony score and numbers of cherries for the trees generated by the different methods for the human mitochondrial DNA data. . . . .	134



## Abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
$C_p$	Mallows $C_p$
JC	Jukes-Cantor (model of nucleotide substitution)
EM-algorithm	Estimation-maximisation algorithm
GLS	General Least-Squares
GO	Gene ontology
GTR	general time-reversible (model of nucleotide substitution)
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LS	Least-Squares
MCMC	Markov Chain Monte Carlo
MDS	Multidimensional scaling
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MP	Maximum Parsimony
NJ	neighbour joining
NNI	Nearest-Neighbor Interchange
NR	Newton-Raphson
PDA	Proportional to Distinguishable Arrangements
PML	Penalised Maximum Likelihood
PNJ	Parsimony Neighbour Joining
SPR	Subtree Pruning and Regrafting
UPGMA	Unweigthed Pair-Group Mean Average
WPGMA	Weighted Pair-Group Mean Average
WLS	Weighted Least-Squares