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

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 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

List of Figures

List of Tables

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