

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.

EVOLUTIONARY ANALYSES OF LARGE DATA SETS: TREES AND
BEYOND

A thesis presented in partial
fulfilment of the requirements

for the degree

of Doctor of Philosophy

in Mathematics at
Massey University

Barbara Ruth Holland
2001

Abstract

The increasing amount of molecular data available for phylogenetic studies means that larger, often intra-species, data sets are being analysed. Treating such data sets with methods designed for small interspecies data may not be useful. This thesis comprises four projects within the field of phylogenetics that focus on cases where the application of current tree estimation methods is not sufficient to answer the biological questions of interest.

- A simulation study contrasts the accuracy of several tree estimation methods for a particular class of five-taxon, equal-rate, trees. This study highlights several difficulties with tree estimation, including the fact that some tree topologies produce “misleading” patterns that are incorrectly interpreted; that correction for multiple changes does not always increase accuracy, because of increased variance; and the difficulty of correctly placing outgroup taxa.
- A mitochondrial DNA data set, containing over 400 modern and ancient Adélie penguin samples, is used to estimate the rate of evolution. Straight-forward tree-estimation is unhelpful because the amount of homoplasy in the data makes the construction of a single reliable tree impossible. Instead the data is represented by a network.
- A method, that extends statistical geometry, assesses whether or not a data set can be well-represented by a tree. The “tree-likeness” of each quartet in the data is evaluated and displayed visually, either for the entire data set or by taxon. This aids in identifying reticulate (or simply noisy) data sets, and also particular taxa that confound tree-like signal.
- Novel methods are developed that use pairwise dissimilarities between isolates in intra-species microbial data sets, to identify strains that are good representatives of their species or subspecies.

Acknowledgements

First and foremost my thanks go to Michael Hendy and David Penny for sharing with me their wealth of ideas and passion for the subject of phylogenetics. If this thesis is at all readable, it is due to their tireless proof-reading efforts and insistence that I put forward my ideas in a semi-intelligible form.

I'd like to acknowledge the support of the many people who made this venture financially viable. Mike and David for the Marsden funded scholarship. The DAAD for making my first trip to Europe possible. Andreas Dress for supporting my stay in Bielefeld, Germany. Vincent Moulton and the STINT grant that made possible two trips to Sundsvall, Sweden. Allen Rodrigo for supporting a visit to Auckland.

I was most fortunate to have the opportunity to meet and work with a wide range of people over the last three years. Thanks to all my collaborators, you provided me with inspiration, interesting problems to work on and the benefit of your wide knowledge. In order of latitude, I express my gratitude to the Sundsvall crowd: Vincent Moulton and Katharina Huber for their fantastic hospitality, and Sverker Edvardsson for lending me his Athlon. In Greifswald I'd like to thank Dietmar Cieslik and Professor Kugelmann. In Bielefeld thanks to Andreas and Heidi Dress for their kind hospitality. Thanks also to Jan Weyer-Menkhoff for his heroic attempts to improve my Deutsch. Thanks to Jack Koolen, and in Duesseldorf, Bill Martin. Moving now to the southern hemisphere, in Auckland thanks go to Allen Rodrigo and Alexei Drummond for great advice on the "penguin chapter". Here at Massey it has been my great pleasure to work with David Lambert, Peter Ritchie and Jan Schmid, whose enthusiasm for penguins, penguins and microbes respectively, was highly infectious.

Thanks to two people who have been a great help on numerous matters, Peter Lockhart and Abby Harrison. Thanks also to the rest of the Thursday lunchtime gang, for sharing your work and for providing helpful comments on mine. I mustn't forget the maths grads, thanks for creating such a friendly atmosphere in which to

work.

A big thank you goes out to all of my friends, those who prevented me from going insane, *and* those that prevented me from going sane. To pick on a few by name, thanks to Agnieszka Szremska for persuading me that Maths and Biology made an interesting combination, and also for answering my dumb questions about genetics. Thanks to Maaïke Bendall for always being a sympathetic ear and for your statistical know-how. Thanks to Paul Gardner for always being willing to bounce ideas, your handy computer hints and all those coffee breaks.

Behind everything I do there is always my family. Thanks Mum for your belief; Dad for the walks and talks at the beach; and my extended family for not being afraid to ask “so what is this phylogenetics stuff anyway?”, and for providing food and shelter. Lastly thanks to my sister Miranda for proof reading, not complaining too vociferously when the dishes remained undone, and judicious application of hugs.

Contents

Abstract	iii
Acknowledgements	v
1 Introduction	1
1.1 Overview	1
1.2 Linking Themes	4
1.2.1 Trees and Beyond	4
1.2.2 Large data sets	5
1.2.3 Using simulation as a tool	6
1.3 Basic Concepts	7
1.3.1 Trees and networks	7
1.3.2 Tree estimation methods	10
Input Data	12
1.3.3 Models of character substitution	15
2 Tree estimation with equal rates	16
2.1 Introduction	16
2.2 Background	18
2.3 Models and Methods	20
2.3.1 Models of sequence evolution	21
2.3.2 Tree estimation methods	22
2.3.3 The tree and generated sequence alignments	24

2.3.4	From sequences to distances	25
2.3.5	The simulation process	27
2.4	Results	27
2.4.1	Accuracy of the Methods	27
	General effects	27
	Correcting for multiple changes	32
	Differences in two-state and four-state results	33
	Split decomposition	34
	Asymmetry of internal edges	37
	Summary	37
2.4.2	Methods can be Consistent but Misleading	37
2.4.3	Classes of Error in Placing the Root	47
2.5	Discussion	49
2.6	Appendix	51
	Derivation of equation 2.2	51
	NJ is unaffected by the length of the outgroup edge	52
3	Detecting evolution in Adélie Penguins	53
3.1	Introduction	53
3.2	Background	54
3.3	Simulations	58
3.4	Haplotype sampling	63
3.5	Phylogenetic Analysis of the Adélie Data	67
3.6	Median Networks of the Adélie Data	72
	3.6.1 Geographic analysis of subgroups	75
3.7	Calculating the rate	78
	3.7.1 Is there a measurable difference in diversity?	78
	3.7.2 Using the median network to estimate a rate.	79
	Rate estimation method	80

3.8	Conclusions	85
4	δ-plots: A tool for visualising tree-likeness	87
4.1	Introduction	87
4.2	Background	88
4.3	δ -plots	92
4.4	Simulations	94
	A sample input file for Treevolve	96
4.5	Identifying “troublesome” taxa	99
4.5.1	Removing “troublesome” taxa	100
4.5.2	Dependence of δ on topology	103
4.5.3	Identifying recombinant taxa	104
4.6	Case Study: <i>Candida albicans</i>	110
4.6.1	$\bar{\delta}_x$ for <i>C. Albicans</i> data	114
4.7	Discussion	116
	Directions for future research	117
5	Selecting Good Model Strains	118
5.1	Introduction	118
5.2	Motivation	119
5.3	Methods	122
5.3.1	Dissimilarity Based Methods	123
5.3.2	Quartet Based Method	126
5.3.3	Graph theoretic approach	128
5.3.4	Greedy algorithms	131
5.4	Analysis of example data sets	133
5.4.1	<i>Pseudomonas aeruginosa</i>	134
5.4.2	<i>Helicobacter pylori</i>	135
5.4.3	<i>Candida albicans</i>	142
5.5	Discussion	143

List of Figures

1.1	Basic concepts with graphs and trees	8
2.1	Common errors in tree construction	17
2.2	Including an outgroup can cause errors	20
2.3	Generating tree for five-taxon simulations	25
2.4	Flowchart of the simulation process	28
2.5	Example plot showing accuracy of NJ with $c = 100$	29
2.6	Accuracy with two-state data	30
2.7	Accuracy with four-state data	31
2.8	Accuracy of split decomposition with four-state data	35
2.9	The misleading zone for MP	40
2.10	Expected frequencies of the non-trivial splits	44
2.11	Frequencies of the different types of error	48
3.1	Ancestor-descendent pairs	57
3.2	Discovery curve for the Adélie penguin samples	64
3.3	The best fit theoretical discovery curve for the Adélie penguins	66
3.4	Non-consensus plot for the Adélie penguin sequence alignment	68
3.5	Majority-rule consensus tree for the Adélie penguin data	70
3.6	A common pattern within HVRI	71
3.7	Overview diagram for median network subgroups	75
3.8	Median networks for each subgroup	76
3.9	Resolving ambiguity in the rate estimation method	82

4.1	The four point condition	90
4.2	A metric on four taxa	91
4.3	Example δ -plots for random and sequence data	93
4.4	The δ -plots for a mammal, viral, and a yeast data set	95
4.5	$\bar{\delta}$ versus n	97
4.6	$\bar{\delta}$ versus sequence length for five different levels of recombination per nucleotide	98
4.7	$\bar{\delta}$ versus sequence length for three different levels of recombination per sequence	98
4.8	$\bar{\delta}_x$ for the mammal, and virus data sets	100
4.9	The effect of random versus $\bar{\delta}_x$ -directed taxon removal orders on four measures of tree-likeness	102
4.10	The caterpillar and balanced tree topologies	105
4.11	δ -plots for the caterpillar and balanced trees	105
4.12	$\bar{\delta}_x$ for the caterpillar and balanced trees	106
4.13	Trees used to generate recombinant sequences	107
4.14	$\bar{\delta}_x$ for six types of recombinant alignment	108
4.15	$\bar{\delta}_x$ for six different combinations of sequence length and proportions of contribution from recombinant parents	109
4.16	δ -plots for <i>C. albicans</i> AFLP data	112
4.17	δ -plots for <i>C. albicans</i> RFLP data	112
4.18	The p -value distribution for the linkage analysis of <i>C. albicans</i>	115
4.19	$\bar{\delta}_x$ for the <i>C. albicans</i> AFLP data	115
5.1	Example for dissimilarity based criteria	125
5.2	Example for quartet based criterion	128
5.3	Example for dominating set based criterion	130
5.4	Neighbor-joining tree for <i>P. aeruginosa</i>	136
5.5	Neighbor-joining tree for <i>H. pylori</i>	140

5.6	Neighbor-joining tree for <i>C. albicans</i>	144
-----	--	-----

List of Tables

1.1	Example sequence alignment	13
1.2	Example distance matrix	14
2.1	Notation for tree estimation methods	22
2.2	Generating tree for five-taxon simulations	24
2.3	Summary of the accuracy of the methods with two-state and four-state data.	34
2.4	Accuracy of split decomposition with four-state data	36
2.5	Accuracy of methods just outside the misleading zone	41
2.6	Accuracy of MP at the boundary of consistency	43
2.7	Star tree simulation	46
3.1	Birth and Death probabilities used in the simulation of Adélie penguin populations.	61
3.2	Results of Adélie population simulations	62
3.3	Location of the modern penguin samples by subgroup	77
3.4	P-values for the test of independence between subgroup and location.	78
3.5	Haplotype diversity of the ancient samples compared to the modern samples	79
3.6	Results of the rate estimation method by subgroup	83
3.7	Results of the randomisation test by subgroup	85
4.1	Parameters for the removal order simulation	101

4.2	Summary of the linkage analysis for <i>C. albicans</i>	114
4.3	$\bar{\delta}$ for five categories of quartets	116
5.1	Exact and Greedy choices of model strains for <i>P. aeruginosa</i> using criteria DC1, DC2, and DC3	137
5.2	r^* for different number of model strains k , and threshold values T , for <i>P. aeruginosa</i>	138
5.3	Best model strain for <i>P. aeruginosa</i> , with $k = 1$, using DSC	138
5.4	Exact and Greedy choices of model strains for <i>H. pylori</i> using criteria DC1, DC2, and DC3.	141
5.5	r^* for different number of model strains k , and threshold values T , for <i>H. pylori</i>	142
5.6	Exact and Greedy choices of model strains for <i>C. albicans</i> using criteria DC1, DC2, and DC3	145
5.7	r^* for different number of model strains k , and threshold values T , for <i>C. albicans</i>	146