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. Straightforward 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 Maaike 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

List of Figures

List of Tables