Exploring Biological Sequence Space

Selected problems in sequence analysis and phylogenetics

A thesis presented in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

in Computational Biology

at Massey University

Bennet James McComish

2012

Abstract

As the volume and complexity of available sequence data continues to grow at an exponential rate, the need for new sequence analysis techniques becomes more urgent, as does the need to test and to extend the existing techniques. These include, among others, techniques for assembling raw sequence data into usable genomic sequences; for using these sequences to investigate the evolutionary history of genes and species; and for examining the mechanisms by which sequences change over evolutionary time scales. This thesis comprises three projects within the field of sequence analysis.

- It is shown that organelle genome DNA sequences can be assembled *de novo* using short Illumina reads from a mixture of samples, and deconvoluted bioinformatically, without the added cost of indexing the individual samples. In the course of this work, a novel sequence element is described, that probably could not have been detected with traditional sequencing techniques.

- The problem of multiple optima of likelihood on phylogenetic trees is examined using biological data. While the prevalence of multiple optima varies widely with real data, trees with multiple optima occur less often among the best trees. Overall, the results provide reassurance that the value of maximum likelihood as a tree selection criterion is not often compromised by the presence of multiple local optima on a single tree.

- Fundamental mechanisms of mutation are investigated by estimating nucleotide substitution rate matrices for edges of phylogenetic trees. Several large alignments are examined, and the results suggest that the situation may be more complex than we had anticipated. It is likely that genome scale alignments will have to be used to further elucidate this question.

Acknowledgements

First and foremost I would like to thank my supervisor, David Penny, for encouraging me to do a PhD after many years away from science. Along with my co-supervisor, Mike Hendy, and despite the best efforts of the University, David has spent many years building a world-class research group conveniently located in my home town. My other co-supervisor, Lesley Collins, has helped knock this thesis into shape, and fought valiantly against my inclination toward understatement.

Thank you to all those who helped with code, data and other technical stuff: Patrick Biggs, Klaus Schliep, Trish McLenachan, Robin Atherton, Simon Hills, Judith Robins, Abby Harrison, Bojian Zhong, Pete Lockhart, Eric Bapteste and others. And to Tim White for the LaTeXtemplate I used to typeset this thesis. Also to the many colleagues who asked questions and made useful suggestions at talks and conferences.

I must thank the Allan Wilson Centre and the Institute of Molecular BioSciences for financial and logistical support, and especially Katrina Ross for having, as David puts it, "the terrible failing of always being on the side of the students."

Thanks to all my friends, old and new, for making the PhD experience a social one as well as a scientific one. In particular, those who do crosswords at lunchtime (Gillian, Simon, Nick, Barbara) and those who play board games on Saturday evenings (some of the above plus Rogerio, Paul, Tim, Sylvia, Robin, Klaus, Atheer, Rick, Sam, Elsa and more). Another thank you goes to Klaus for putting me up and showing me the sights on my visits to Paris. Also to Paul and Carolyn, David and Helen, the Tappers and the Slees, for their generous help and hospitality during my visits to Melbourne.

Finally, I would like to thank my family: both my parents for always encouraging my scientific tendencies; my stepmother Heather for convincing me that doing a PhD would be a good idea; my Belgian extended family, especially Charles and Nanou, for their hospitality during my visits to Europe; and my daughter Petra for providing a good excuse to take a break every few months and for distracting me from my work. And lastly Barbara, for your love and support.

Contents