Extracting and Exploiting Signals in Genetic Sequences

A thesis presented in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

in Mathematics

at Massey University

Walton Timothy James White

2011

Abstract

As DNA databases continue to grow at an exponential rate, the need for more efficient solutions to basic problems in computational biology grows ever more pressing. These problems range from the principal questions driving evolutionary science—How can we accurately infer the history of genes, individuals and species? How can we separate the signal from the noise in our data? How can we visualise that signal?—to the purely practical: How can we efficiently store all this data? With these goals in mind, this thesis mounts a computational combination attack on a variety of topics in bioinformatics and phylogenetics:

- A program is designed and implemented for solving the Maximum Parsimony problem—in essence, finding phylogenetic trees having the fewest mutations. This program generally outperforms existing highly optimised programs when using a single CPU, and unlike these earlier programs, offers highly efficient parallelisation across multiple CPUs for further speedup.

- A program is designed and implemented for compressing databases of DNA sequences. This program outperforms general-purpose compression by taking advantage of the special "treelike" structure of DNA databases, using a novel data structure, the "leaky move-to-front hashtable", to achieve speed gains.

- A data visualisation technique is introduced that concisely summarises the "treelikeness" of phylogenetic datasets on a ternary plot. Each dataset is represented by a single point, allowing multiple datasets, or multiple treatments of a dataset, to be displayed on a single diagram.

- We demonstrate problems with a standard phylogenetic analysis methodology in which a single tree is assumed *a priori*. We argue for a shift towards network methods that can in principle reject the hypothesis of a single tree.

- Motivated by a phylogenetic problem, a fast new algorithm is developed for finding the mode(s) of a multinomial distribution, and an exact analysis of its complexity is given.

Acknowledgements

I would like to thank my supervisors, Mike Hendy, David Penny and Barbara Holland, for their support (both financial and moral), guidance and patience during my candidature. I feel extremely lucky to have had the chance to work with such all-round high-calibre people. Several years ago, in a combination attack of their own, these three gradually persuaded me (against my own (worse) judgment) to upgrade from a Masters degree to a PhD. I'm grateful that they knew me well enough, and thought it worthwhile, to keep nudging me until I realised what was good for me!

Many people have had a positive impact on my life during the course of my PhD. A somewhat biased random sample includes:

Past and present members of the Allan Wilson Centre at Palmerston North, whom I have found to be without exception friendly, intelligent and helpful.

My flatmates at the venerable 53 Te Awe Awe St. estate—Matt, Rachel, Michael, both Karens, Richard and Scott—for providing a thoroughly enjoyable place to live, where I always felt comfortable, *even* when Richard was around. As always, it seems fitting to mention that Scott knows a guy who once slept *on a door*.

Tony Dale of the BlueFern supercomputer group, for always responding quickly and helpfully when something went wrong.

The loose association of People Who Have Dinner and Play Board Games on Saturday Nights—to a first approximation, Rogerio, Klaus, Simon, Gillian, Bennet, Barbara and Robin—for tolerating my sense of humour, and occasionally affording me the chance to repay their gastronomic generosity with a few least significant bits of my own.

My good friend Sylvia, for many things, including for teaching me that people from very different backgrounds can find the same kinds of people extremely funny-looking.

Jing, for being an ideal officemate. Trish, for being a tireless force for good. Joy, for returning my wisecracks with interest.

Everyone over the years who helped me *test the graphics cards* on the computers in the Allan Wilson Centre.

And of course my family, whose love and support has been a constant in my life.

Contents