Consensus sequence estimation methods for DNA sequence data sets

Sarah Ellen James

Thesis submitted for the degree of

Master of Philosophy

in

Statistics

at

The University of Adelaide

(Faculty of Engineering, Computer and Mathematical Sciences)

School of Mathematical Sciences

November 28, 2017

Contents

Abstract

A combination of ancient and modern DNA sequences and geographical information is used in phylogenetics, to study the evolutionary relationships between individuals or species. A common problem with using ancient DNA sequences in phylogenetic analyses is that ancient DNA sequence data sets can often contain damaged or missing bases. This limits the accuracy of the analysis and reduces the number of statistical methods available for use.

Since DNA sequencing technologies have improved, more informative statistical techniques have been developed to estimate the DNA sequence for a single individual. In the 1990s, several DNA sequence estimation methods were developed using only the reads contained in the alignment. Current statistical methods use both the reads in the alignment and the qualities associated with each read, to estimate the consensus sequence. We then obtain the DNA sequence by removing any gaps in the consensus sequence.

A limitation of these DNA sequence estimation methods is that these methods may return estimated DNA sequences with missing bases, due to low coverage or poor quality data. DNA sequences with missing bases are problematic when used in statistical analyses, and this can limit the accuracy of the analysis. Therefore, in this thesis we focus on developing a consensus sequence estimation method that uses information from the alignment as well as outside sources of information. In particular, our consensus sequence estimation method estimates bases for the sites

in the DNA sequence that may have otherwise been assigned a missing base, due to low coverage or poor quality data.

We develop two consensus sequence estimation methods based on the EM algorithm and Gibbs sampling respectively. Our consensus sequence estimation method based on the EM algorithm uses the alignment data as well as the associated quality data to estimate the consensus sequence. Our consensus sequence estimation method based on Gibbs sampling uses the alignment data, quality data, and cytosine deamination rates to also estimate sites that were damaged by ancient DNA damage.

Since we often do not know the true DNA sequence for an individual, we use simulated DNA sequences and alignment data to assess the accuracy of our consensus sequence estimation methods. Using a DNA sequence distance metric, we compare our estimated DNA sequences to the true DNA sequence for each consensus sequence estimation method presented in this thesis. We also consider the entropy at each site along the consensus sequence to quantify the amount of uncertainty in the estimated consensus sequence. Based on these results, we make recommendations on which consensus sequence estimation methods are suitable for particular DNA sequence data sets.

An advantage of the Gibbs sampling approach over the EM algorithm approach is that we have informative prior distributions for sites with low coverage or poor quality data. Hence, the estimated bases at these sites are more informative than those estimated using the EM algorithm approach. Our consensus sequence estimation methods generate estimates for all sites, including those that may have otherwise been allocated missing bases. Therefore, our consensus sequence estimation methods reduce the problem of using DNA sequences with missing bases in phylogenetic analyses.

Signed Statement

I, Sarah James, certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

SIGNED: ..... DATE: .....

Dedication

I dedicate this thesis to my loving family. I greatly appreciate your encouragement and support while I completed this thesis.

Acknowledgements

Firstly, I would like to sincerely thank my supervisors Professor Nigel Bean and Dr Simon Tuke for their endless support and encouragement over the past few years. Your depth of knowledge and ability to explain even the most challenging of topics sets the bar and is something I hope to aspire to.

To my parents, Ralda and Steven. You are the most loving parents anyone could ever ask for. Thank you for all of the advice across the years, your constant support and unconditional love. You both made me feel like nothing is beyond my reach and challenged me to break the limits of knowledge, and doing so has no doubt helped me to complete this thesis. To my siblings, Abbie and Stephen. Thank you for your support and for always being there whenever I needed someone to chat to and have a laugh with. Both of you always encouraged me to go further and believe in the possibilities of every dream I had. You two are the best siblings anyone could ever have.

To my closest friends, Amber, Bianca, Danielle, Jacqui, Kate, Laura and Max. Thank you for all the fun and wonderful times we have had since we first met. It is an absolute privilege to know you all and I look forward to many more enjoyable and exciting times. To all of my friends and colleagues at university, thank you for making university so enjoyable over the past few years.

To my senior school mathematics teacher, Mr Constantin Naum. Thank you for showing me how enjoyable and rewarding studying mathematics can be, and encouraging me to study mathematics at university.

Finally, to everyone at the Australian Centre for Ancient DNA. Thank you for answering every question I had about ancient DNA. Your help has been invaluable to me during my postgraduate study. To everyone in the School of Mathematical Sciences, it has been an absolute pleasure getting to know you all. Thank you for all the support and friendly chats in the office and the hallways.

This work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide.