Data Driven Model Selection and Parameter Estimation Using Semi-Automatic Approximate Bayesian Computation to Reconstruct Population Dynamics From Ancient DNA

Adam Benjamin Rohrlach

Thesis submitted for the degree of

Master of Philosophy

in

Applied Mathematics

at

The University of Adelaide

(Faculty of Engineering, Computer and Mathematical Sciences)

School of Mathematical Sciences

May 8, 2014

Contents

Abstract

Population genetics is a discipline within the biological sciences that is concerned with the change in frequency of types of individuals in a population due to natural selection, mutation, genetic drift and gene flow. Genetic drift is the part of this process explained by random sampling. Important to the process of genetic drift is population structure and so we focus on the recovery of population sizes over time, given a set of DNA sequences.

With recent advances in computational power and a growth in the amount of data available, increasingly powerful techniques are being developed for the study of sequence data. Key advances in the early 1980's centred around 'the coalescent', a continuous time approximation to the Wright-Fisher model of reproduction, and these advances resulted in Skyline Plot methods for recovering population size estimates over time. Skyline Plots suffer from large variances for the 'coalescent' event times, and sources of error common to DNA sequence sampling schemes.

Approximate Bayesian Computation (ABC) is a class of likelihood-free methods for statistical inference. ABC techniques can trace their genesis back to the biological sciences due to the complexity of the models for reproduction (and hence the intractability of likelihood calculations). Unfortunately, like Skyline Plots, ABC also suffers from many sources of error, not least of which occurs when we can not use sufficient summary statistics.

To considerably reduce the effect of the error related with the use of insufficient summary statistics, we explore a process of semi-automatic summary statistic cal-

culation through the use of 'training data' (simulated under the coalescent model). We obtain a training set of data, and fit a linear model (under a Box-Cox transformation) for each parameter of interest, using common summary statistics for DNA sequences as predictor variables. We call these linear combinations of (insufficient) summary statistics the semi-automatic summary statistics, and using a new set of simulations, we perform ABC where a simulation is retained if the predicted parameter values are 'close enough' to the predicted parameters for the observed data. We analyse three sets of coalescent simulated data from three population models; the Constant, Exponential and Migration Models, and compare our findings with the corresponding Skyline Plot analyses performed in BEAST.

When we simulate data for training our linear model, we must specify a model of population size dynamics, and we explore methods to select a population model, given our data. A common means of model comparison used with ABC analyses is called Bayes Factors. We show that Bayes Factors perform poorly for our data, and highlight a fundamental bias inherent in any model comparison where the probability of a model, given an observed summary statistic, is employed. As an alternative to Bayes Factors, we apply multiple logistic regression (MLR) to classify our observed data into one of a candidate set of possible models. In conjunction with the MLR analysis, we use principal component analysis for visualisation, and introduce a method for attempting to identify when the correct model is not in the candidate model set, or when a classification seems reasonable. We show that this method of classification performs well for the three observed data sets using sensitivity analysis.

Due to the early stage of development of our work, we can not use real world data, and so we use a different type of simulation since our method uses coalescent simulations to train the model. We obtain sequence data simulated under a 'forward simulation' framework, a type of sequence simulation that looks forward in time. We define a two-step process for analysis that begins with MLR classification, and

then, under a model chosen by the MLR classification, uses semi-automatic summary statistic calculation for parameter estimation via ABC. We correctly identify this model of population dynamics, and perform parameter estimation on the data, comparing our results with the corresponding BEAST Skyline Plot analysis.

Signed Statement

I, Adam Rohrlach, certify that this work contains no material which has been accepted for the award of any other degree or diploma 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 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 consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

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 catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

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

Dedication

I would like to dedicate my thesis to Allan and Mary Isobel Troughear.

You were the most wonderful and loving grandparents I could have hoped for. I hope I have made you as proud as you always seemed to be.

Acknowledgements

First, to my supervisors, Professor Nigel Bean and Dr Simon Tuke, thank you for seeing more in me from an early stage, than I saw in myself. It is because of your tireless encouragement, and small research projects during my undergraduate degree, that I decided to pursue post-graduate research. You are not only gifted lecturers, but two of the most patient and straight-faced mentors I could have hoped for. Our meetings have been so enjoyable, that I regularly forget that this is a job, and have begun to believe we might just take this show on the road.

To my Father, thank you for all of the advice these long years, the weekly lunches that allowed me to clear my mind, and the port. It seems you have always been waiting for me to get things right. Thank you for treating me no differently while I did not. To my Mother, thank you for the unconditional love, support, and encouragement. You have always made me feel as though nothing is beyond my reach. To my sisters, Prue and Paige, I thank you for the family meals where I was able to forget my work, and just enjoy being 'at home' again. In fact, to my entire family, I can not thank you enough for continuing to believe in me, even when all the available data suggested you should not.

To my two closest friends, Dan and Brett, thank you for coming on this five year long ride with me. It seems if I were ever in any danger of taking myself too seriously, you knew precisely when to step in. I have known you both for most of my life, and while this reflects poorly on the quality of company you keep, thank you for sticking with me during every phase of it.

To Professor Alan Cooper, and everyone at the Australian Centre for Ancient DNA, this project would not have been the same without you. Thank you for allowing me to sit in on countless meetings and answering every one of my questions. Specifically, to Julien and Oliver, thank you for taking the time to walk me through BEAST, and answering every email I sent you. I could not have done a single comparison of results without your help.

Finally, thank you to everyone in the School of Mathematical Sciences at the University of Adelaide. It has been an honour to work with all of you, and you have made the last two years a genuine pleasure. There are so many people I owe my thanks to, but I would like to thank a couple of you specifically. To David A, Heath, Jess, Kate, Nick, Nic and Stephen, thank you for all of the fun we have had, the help you have given me, and the countless coffees we have drunk over the last few years. I would be remiss not to mention three exceptional people, without whom I could not have completed this thesis with such fond memories.

To David, you have been a pleasure to work with, and I have learned so much about teaching since we 'teamed up'. Thank you for reading every word I have written, and never once berating me for them. To Mingmei, thank you for not only the last two years, but all of the time we have spent learning together. I doubt I could have succeeded without your influence, perspective, and strength. You are an amazing individual, and you pushed me harder than I would have pushed myself. To Vincent, I thank you for setting the bar so high. It is rare to meet someone so gifted, but still so easy to spend so much time with. Thank you for filling exam study periods with laughter, tea and talking british cakes.