GHOST: A time-reversible mixture model for recovering phylogenetic signal from heterotachously-evolved sequence alignments.

Stephen Crotty

Thesis submitted for the degree of

Doctor of Philosophy

in

Applied Mathematics

at

The University of Adelaide

(Faculty of Mathematical and Computer Sciences)

Department of Applied Mathematics

January 13, 2017

Contents

List of Tables

List of Figures

Signed Statement

I 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 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.

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

Acknowledgements

Much of the credit for this Degree must go to my wife Jennifer, who has contributed as much indirectly as I have directly. There are several sound arguments against commencing a PhD at the same time as bringing twins into the world. It places a ceiling on earning ability at a time when expenses rise dramatically. It requires attention when free time has become a faded memory. It drains mental and emotional resources that are already in short supply. Over the last four years we have had to overcome all of these challenges, and more. At no stage was it easy, but we have survived. Thank you for your support at every step of the way. I hope that all of our efforts bear ample fruit for our children and for ourselves.

My Mum and Dad, to whom I owe everything I have. Your love, support and generosity has been a constant throughout my life, regardless of any decisions I have made. Your perpetual hard work and sacrifice is the only reason that I was in a position to consider commencing a PhD at the age of 33 with a young family on the way. You have taught me to stand on my own two feet, all the while making sure that I never had to.

To my robust supervisory panel: Professor Nigel Bean, Doctor Jonathan Tuke, Associate Professor Barbara Holland and Doctor Lars Jermiin. As in the movies, the team of superheroes all have their own particular skill set. Nigel, I could not have asked for a better principal supervisor. You were always patient and understanding of the external pressures that often took precedence over my PhD. Without your compassion in this area I doubt that I would have been able to successfully complete. Your generosity in funding my trip to Vienna last year provided the opportunity

to form a collaboration that strengthened my PhD and resulted in my ongoing employment. Your mathematical assistance and mentorship also exceeded anything I could have hoped for. Our weekly meetings were essential to me, helping me to organise my thoughts and give focus to the project. Many times I called past your office unannounced, seeking enlightenment on some trivial concept or result. I know you must often have been very busy on far more important matters, yet you never made me feel like I was interrupting. Jono, in contrast, you always made me feel like I was interrupting even when I had an appointment. When I returned to tertiary study seven years ago, entering a formula in the cell of an Excel spreadsheet was about the extent of my coding expertise. Any progress I have made in this area has been greatly facilitated by your advice, guidance and example. You often made great improvements to my code in a matter of seconds ('select all + delete' usually sufficed), you taught my son a valuable lesson about trusting Mancunians and you never let maths spoil a fun meeting. Barbara, there have been a handful of times over the last four years that my understanding of the subject matter has undertaken a quantum leap forward. These occasions coincided with the all too rare times where I was able to sit down with you at a conference and discuss the material in depth. I left every conference with renewed clarity of thought and enthusiasm for the project. There is no doubt in my mind that had I been lucky enough to have you as a local supervisor, I would have found the project easier and the end result stronger. Lars, your enthusiasm was contagious and always provided me with motivational boosts when I needed them most. Often I would enter a supervisor meeting to report what I believed to be mundane results, only to have you convince me that they were exciting, interesting and novel. This encouragement was vital, particularly in the early stages of the Degree when it was very easy to feel that the task was too great. To all of my supervisors, thank you for your efforts over the past four years. You have all made significant contributions that have shaped not only the project but myself as a researcher.

To Professor Arndt von Haeseler and Doctor Bui Quang Minh, the collaboration

that you facilitated marked a major turning point in my project. Without your help the results would have been primarily theoretical, and the thesis significantly weaker. The implementation of the GHOST model in IQ-TREE has enabled its application to a wide variety of biological problems, hopefully ensuring its relevance long after my PhD is complete. I also must thank you for the faith you have shown in me by your offer of employment when I was only half way through my PhD. As a student with a young family, I had anticipated the uncertain transition from PhD to Post Doc employment as a very stressful time. The security provided by your employment offer has minimised this stress for myself and my family.

To Ben, history will show that the first couple of phylogeneticists to come out of the School of Mathematics at the University of Adelaide were of outstanding quality, on average. It is no coincidence that over the years I have requested your assistance far more regularly than you have requested mine. The times we have shared at conferences have been a highlight, some more memorable than others. I look forward to catching up with you at these events long into the future.

Dedication

For Emily and Daniel, may you find happiness wherever life takes you.

Abstract

The accuracy and reliability of phylogenetic inference is compromised by the adoption of models of sequence evolution that don't adequately reflect the dynamic nature of evolution by natural selection. Heterotachy refers to variation in the rate of evolution of a particular site across lineages on a tree. We carry out simulations, showing that phylogenetic inference using popular methods and models is unreliable when the data evolved under the influence of heterotachy. We carry out a theoretical analysis of these methods and models, concluding that their failure was inevitable given the nature of the data.

To remedy this we introduce the General Heterogeneous evolution On a Single Topology (GHOST) model. We implement the GHOST model under a maximum-likelihood (ML) framework in the phylogenetic inference program IQ-TREE. We perform extensive simulation studies, showing that the GHOST model can successfully recover the tree topology, branch lengths and substitution model parameters from heterotachously-evolved sequences. We apply our model to a real dataset and identify a subtle phylogenetic signal linked to the convergent evolution of the electric organ in two geographically distinct lineages of electric fish. Furthermore, we use the model to successfully identify specific sites in the alignment that are pivotal to the effective function of the electric organ.

The GHOST model and its implementation in IQ-TREE provide the most flexible mixture model currently available for performing phylogenetic inference in a ML framework. This increased flexibility better equips the GHOST model to represent the process of evolution by natural selection. We show that the GHOST model is

able to highlight subtleties in evolutionary relationships that coarser models cannot. We foresee the GHOST model having potential uses in a variety of applications: helping to resolve disputed topologies; focusing the efforts of biologists by identifying alignment sites of functional importance; bringing to light evidence of convergent evolution; and investigating the coevolution that occurs between disease and immune cells, or hosts and parasites. As computing resources continue to grow and phylogenetic algorithms are revised and improved, the GHOST model will be applicable to ever larger MSAs, ultimately assisting in illuminating the history of life on earth.