

**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

Signed Statement	xxiii
Acknowledgements	xxv
Dedication	xxix
Abstract	xxxii
1 Introduction	1
2 Background	7
2.1 DNA	7
2.1.1 Sequence structure	8
2.1.2 Mutation	8
2.1.3 Substitutions and natural selection	10
2.1.4 Multiple sequence alignments	11
2.2 Inferring phylogenetic trees	12
2.2.1 Maximum parsimony	13
2.2.2 Neighbour joining	14
2.2.3 Maximum likelihood	14
2.2.4 Bayesian Markov-chain Monte-Carlo	15
2.3 Models of sequence evolution	17
2.3.1 Heterogeneous models	20

3	Heterotachy	25
3.1	Simulating heterotachously-evolved alignments	28
3.2	Maximum parsimony	30
3.2.1	Definitions	30
3.2.2	Elucidation of events F and A	32
3.2.3	Evidence for and against the correct tree	35
3.3	Neighbour joining	41
3.4	Maximum likelihood	48
3.4.1	The Expected Dataset	50
3.5	Model misspecification	52
4	Modeling heterotachous evolution	55
4.1	Inference with heterotachously-evolved data	55
4.1.1	Sequence alignments	55
4.1.2	The proposed model: JC+I+H2	56
4.1.3	Implementation in R	57
4.1.4	Performance	59
4.2	The GHOST model	64
4.3	IQ-TREE Development	65
4.4	Inferring phylogenies with IQ-TREE	66
4.4.1	Parameter optimization in IQ-TREE	66
4.4.2	Searching tree space	69
4.5	Implementation of the GHOST model in IQ-TREE	69
4.5.1	Optimizing branch lengths and substitution model parameters of the GHOST model	69
4.5.2	Optimization of weights for the GHOST model	73
5	Validation of the GHOST model in IQ-TREE	75
5.1	Tree topology recovery	75
5.1.1	Experiment 1	76

5.1.2	Experiment 2	78
5.1.3	Experiment 3	79
5.2	Parameter recovery	80
5.2.1	Specific Case	81
5.2.2	General case	87
5.3	Soft classification of sites to classes	92
6	The Convergent Evolution of Electric Fishes	95
6.1	Background	95
6.2	Data	96
6.3	Identifying the optimal GHOST model	97
6.4	Analysis of classes inferred by ML-GTR+H4 model	99
6.5	Soft classification of sites to classes	103
6.6	ML-GHOST vs comparable models and methods	105
7	Conclusion	109
	Bibliography	113

List of Tables

2.1	The Genetic Code: The 64 codons and the amino acids (AA) they encode. Some amino acids correspond to as many as six codons. Consequently not all nucleotide substitutions will result in an amino acid replacement. The process by which the codons produce the stated amino acid is more complex, involving transcription of the DNA into ribonucleic acid (RNA), followed by the translation of RNA into the amino acids. The details of these processes are beyond the scope of the thesis.	9
2.2	Common types of mutation. Substitution: the nucleotide G at the second site in the sequence has been substituted to a T. Insertion: an additional nucleotide, T, has been inserted after the second site in the sequence. Deletion: the nucleotide, G, at the second site in the sequence has been deleted.	10
2.3	An example of a multiple sequence alignment (MSA) for nucleotide data. The pattern of nucleotides at a particular site is known as a site pattern. For example, the site pattern at site four is GTTC. . . .	12
2.4	Number of unique unrooted phylogenetic trees for a given number of taxa	16

3.1	Parsimony scores of the 15 generic site patterns for the correct AB CD topology and the incorrect AD BC topology. Highlighted in blue, <i>xyyy</i> is the only site pattern for which the correct tree is more parsimonious than the incorrect tree, whereas highlighted in red, <i>xyyx</i> is the only site pattern for which the incorrect tree is more parsimonious than the correct tree. All other site patterns are uninformative.	34
3.2	Possible substitution combinations that will result in the site pattern <i>xyyy</i> , given a substitution has occurred along the internal edge, <i>k</i> . The weight, <i>W</i> , indicates the proportion of time that the described substitution combination will result in the <i>xyyy</i> site pattern.	37
3.3	Possible substitution combinations that will result in the site pattern <i>xyyx</i> , given a substitution has not occurred along the internal edge, <i>k</i> . The weight, <i>W</i> , indicates the proportion of time that the described substitution combination will result in the <i>xyyx</i> site pattern.	38
5.1	General Case simulation results - Summary statistics for the rate score (RS), frequency score (FS), branch score (BS) and weight score (WS) for comparisons of the GHOST model and the partition model to the true parameters for the 1000 simulations conducted under the General Case.	93
6.1	The 11 fish species in the dataset and the GenBank accession numbers for the <i>Na_v1.4a</i> gene of each species.	97
6.2	The relative substitution rates inferred by ML-GTR+H4 for the electric fish dataset. Rates are shown relative to the G↔T substitution rate which is fixed at 1.	101
6.3	The base frequencies inferred by ML-GTR+H4 for the electric fish dataset.	102
6.4	The relative frequency of codon position for each of the four inferred classes.	105

6.5	The ten sites in the alignment with the highest probability of belonging to the convergent class. Note the over-representation of codon position 1, suggesting these sites are likely to have non-synonymous substitutions present.	106
6.6	The sequence alignment corresponding to the ten sites identified in Table 6.5, ordered from highest probability of belonging to the convergent class to lowest. Clearly the overwhelming majority of substitutions (highlighted in magenta) occur in the electric fish.	107
6.7	The results show that when applied to the electric fish dataset the GHOST model in IQ-TREE clearly outperformed all of the Pagel & Meade models in Bayes Phylogenies. Their best fitting model, the PMRB4, was inferior to GHOST in terms of AIC by 117 units and it took approximately 140 times longer to run.	108

List of Figures

1.1	Darwin's original sketch from On the Origin of Species (Darwin, 1859), showing him formulating the concept of a phylogenetic tree.	2
2.1	Flowchart depicting the relationships of the models of sequence evolution to each other. Models on the right hand side are those that maintain equal base frequencies while those on the left allow for unequal base frequencies.	21
2.2	Schematic displaying the increasing complexity of the arrangement of substitution rate parameters for a selection of models of nucleotide sequence evolution. The JC model is the simplest, all substitutions share the same rate parameter α . The K80 introduces a second rate parameter, so that transitions occur at rate α and transversions at rate β . The K81 introduces a third parameter so that transversions now occur either at rate β or rate γ . Finally the GTR model defines a separate rate parameter for each of the six unique pairs of nucleotides. This is the most general model possible while still maintaining the desirable property of time reversibility.	22
3.1	Four taxon tree with two events, e_1 and e_2 at which a certain percentage of invariable sites are switched to become variable. Branch lengths used for the simulations were: $a = 0.4, b_1 = 0.1, b_2 = 0.3, c_1 = 0.1, c_2 = 0.3, d = 0.4$ and $k = 0.1$	29

3.2	MP results for the four taxa simulation study on heterotachously-evolved 100,000bp MSAs. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 at increments of 0.008. At each value, 100 MSAs were simulated. The y-axis reports the fraction of MSAs for which MP inferred the correct tree topology.	31
3.3	MP results for the 4-taxa simulation study on heterotachously-evolved sequence data. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 at increments of 0.008. At each value, 100 MSAs were simulated. The y-axis reports the fraction of MSAs for which MP inferred the correct tree topology. The theoretical proportion of heterotachous sites at which MP should fail to recover the correct topology is shown by the dashed red line. Clearly the calculations concur with the empirical data.	42
3.4	NJ results for the 4-taxa simulation study on heterotachously-evolved sequence data. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 at increments of 0.008. At each value 100 replicate MSAs were simulated. The y-axis reports the fraction of replicates for which NJ inferred the correct tree topology.	43
3.5	$E[Q_{ij}]$ scores, as a function of p_{het} . The x-axis displays proportion of heterotachous sites in the alignment. The y-axis displays the $E[Q_{ij}]$ scores for each topology. The minimum Q_{ij} dictates which pair of taxa will be clustered together, which fully resolves the topology in the four taxa case. It is difficult to see in detail which topology is the minimum for some values of p_{het} . To further investigate the plot is reproduced in Figure 3.6 with the trend removed.	45

3.6 Detrended $E[Q_{ij}]$ scores, as a function of p_{het} . The x-axis displays proportion of heterotachous sites in the alignment. The y-axis displays the detrended $E[Q_{ij}]$ scores, that is $E[Q_{ij}] - \frac{1}{3}(E[Q_{AB}] + E[Q_{AD}] + E[Q_{AC}])$. The minimum Q_{ij} dictates which pair of taxa will be clustered together, which fully resolves the topology in the four taxa case. We can see that, consistent with the empirical results seen in Figure 3.4, we expect to infer the AB|CD tree for low and high values of p_{het} , and the AD|BC tree in between. 46

3.7 NJ results for the four taxa simulation study on heterotachously-evolved sequence data. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value 100 MSAs were simulated. The y-axis reports the fraction of MSAs for which NJ inferred the correct tree topology. The dashed red lines indicate the theoretical transition points at which the NJ method should switch from inferring the correct tree to the incorrect tree, and then from the incorrect tree back to the correct tree. Clearly the calculated transition points concur with the empirical data. . . . 47

3.8 ML results for the four taxa simulation study on heterotachously-evolved sequence data. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value 100 MSAs were simulated. The y-axis reports the fraction of MSAs for which ML inferred the correct tree topology. 48

3.9	Comparison of conditional likelihoods of the simulated datasets. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value 100 replicate MSAs were simulated. The y-axis reports the difference in log likelihood. The solid lines show the mean difference in conditional likelihoods between the correct topology and the two incorrect topologies. The dashed lines indicate minimum and maximum difference over the 100 MSAs.	49
3.10	Comparison of conditional likelihoods of the expected datasets. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was constructed. The y-axis shows the difference in conditional likelihoods between the correct topology and the two incorrect topologies. Both curves correspond closely with the empirical evidence shown Figure 3.9.	52
4.1	The difference in conditional maximum likelihood scores between the correct topology and the two incorrect topologies. D_{ABAD} refers to the difference between the maximum likelihood conditional on the AB CD topology and the maximum likelihood conditional on the AD BC topology. D_{ABAC} refers to the difference between the maximum likelihood conditional on the AB CD topology and the maximum likelihood conditional on the AC BD topology. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was generated as described in Section 3.4.1. The fact that the differences increase as p_{het} increases suggests that as the influence of heterotachy increases the JC+I+H2 model becomes more likely to infer the correct topology.	60

4.2	Branch lengths inferred by R under the JC+I+H2 model for the variable class of the expected MSAs. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was generated as described in Section 3.4.1. All branch lengths appear to be slightly overestimated, particularly the branches leading to taxa A and D. The magnitude of the overestimation appears to increase as p_{het} increases.	61
4.3	Branch lengths inferred by R under the JC+I+H2 model for the heterotachous class of the expected MSAs. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was generated as described in Section 3.4.1. All branch lengths appear to be recovered reasonably accurately, particularly the branches leading to taxa B and C.	62
4.4	Class weights inferred by R under the JC+I+H2 model for the expected MSAs. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was generated as described in Section 3.4.1. The proportion of invariable sites is recovered with a high degree of accuracy. The proportion of variable sites appears to be slightly underestimated while the proportion of heterotachous sites appears to be slightly overestimated. The magnitude of these errors appears to increase as p_{het} increases.	63
4.5	Flow chart for IQ-TREE's core optimization algorithm, largely reproduced from Figure 3 of Nguyen <i>et al.</i> (2015).	67
4.6	Schematic of a phylogenetic tree. The circles represent two nodes on the tree, a and b , connected by a branch of length λ . The triangles represent subtrees.	68

4.7	Flow chart detailing the Candidate Tree Set Algorithm (CTSA). *If \mathbb{C} contains less than 100 unique topologies at this point in the algorithm then \mathbb{C} is populated with random unique topologies until it contains the lesser of 100 or all possible topologies.	70
4.8	Flow chart for the Hill-climbing Nearest Neighbour Interchange Algorithm (HNNIA). *A valid nearest neighbour interchange (NNI) is any NNI upon the initial iteration, or any NNI on an inner edge within 2 branches of a tagged edge upon subsequent iterations.	71
5.1	The two symmetric, 4-taxa trees of identical topology used in the simulation studies of K&T. The branch lengths were constructed such that each tree comprised of one pair of non-sister long branches and one pair of non-sister short branches.	76
5.2	Performance of ML-JC+H2, ML-JC and MP for data generated under strong heterotachy, $p=0.75$ and $q=0.05$. The length of the internal branch, r , is displayed on the x-axis and was varied between 0.01 and 0.4 with 200 replicates at each value of r . The y-axis displays the fraction of the 200 replicates that recovered the correct topology. The results for MP and ML-JC were identical to the results of K&T, neither performed adequately but MP is able to recover the correct topology for shorter r than ML-JC. However ML-JC+H2 was able to reliably recover the tree topology for this data even when the internal branch is very short.	77

5.3 Results of K&T's Experiment 2, assessing the performance of MP, ML-JC and ML-JC+H2 for different combinations of p and q . On the x-axis are three different values of p and three different values of q are displayed in the separate facets. On the y-axis is BL_{50} , defined by K&T as the minimum internal branch length required for the method to recover the correct tree topology at least 50% of the time, for a sequence length of 10,000bp. Small values of BL_{50} indicate that the model is less likely to infer the incorrect topology given the heterotachously-evolved data. The ML-JC+H2 model clearly outperforms MP and ML-JC over the range of heterotachous conditions tested by K&T. The only cases in which MP and ML perform comparably to ML-JC+H2 is when p and q are similar, that is when the data is not particularly heterotachous, (e.g. $p = 0.3$ & $q = 0.4$, or $p = 0.5$ & $q = 0.4$). 79

5.4 The mean inferred base frequency parameters for Class 1 of the Specific Case. The weight of Class 1 is shown on the x-axis, the base frequency is shown on the y-axis. The data points indicate the mean base frequency inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the base frequencies used to simulate the Class 1 component of the MSAs. The results indicate that IQ-TREE was able to accurately recover the base frequencies for Class 1 of the Specific Case simulations. 82

5.5 The mean inferred base frequency parameters for Class 2 of the Specific Case. The weight of Class 1 is shown on the x-axis, the base frequency is shown on the y-axis. The data points indicate the mean base frequency inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the base frequencies used to simulate the Class 2 component of the MSAs. The results indicate that IQ-TREE was able to accurately recover the base frequencies for Class 2 of the Specific Case simulations. 83

5.6 The mean inferred substitution rate parameters for Class 1 of the Specific Case. The weight of Class 1 is shown on the x-axis, the substitution rate is shown on the y-axis. The data points indicate the mean substitution rate inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the substitution rates used to simulate the Class 1 component of the MSAs. All rates are recovered by IQ-TREE with a reasonable level of accuracy. The error appears to be greater for substitution rates with higher true values, most notably the A↔T rate. The fact that the error decreases as w_1 increases suggests that it is primarily an artefact of stochastic variation in the simulation process, the effect is diminished as the length of the Class 1 component of the MSA increases. 85

5.7 The mean inferred substitution rate parameters for Class 2 of the Specific Case. The weight of Class 1 is shown on the x-axis, the substitution rate is shown on the y-axis. The data points indicate the mean substitution rate inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the substitution rates used to simulate the Class 2 component of the MSAs. All rates are recovered by IQ-TREE with a reasonable level of accuracy. The error appears to be greater for substitution rates with higher true values, most notably the C↔T rate. The fact that the error decreases as w_1 increases suggests that it is primarily an artefact of stochastic variation in the simulation process, the effect is diminished as the length of the Class 1 component of the MSA increases. 86

5.8 The mean inferred Branch Score (BS) for Class 1 of the Specific Case, for both the GHOST and partition models. The weight of Class 1 is shown on the x-axis, the BS is shown on the y-axis. The data points indicate the mean BS inferred by IQ-TREE using ML-GTR+H2 or ML-GTR+PART over the 20 replicate MSAs at that Class 1 weight. The difference in BS between the partition and the GHOST models is small in comparison to the magnitude of the partition model BS, suggesting that with respect to branch length recovery IQ-TREE using the GHOST model performs as well as we could expect. Furthermore this distance decreases as w_1 increases (as the sequence generated under Class 1 becomes longer), implying consistency. 88

5.9	The mean inferred Branch Score (BS) for Class 2 of the Specific Case, for both the GHOST and partition models. The weight of Class 1 is shown on the x-axis, the BS is shown on the y-axis. The data points indicate the mean BS inferred by IQ-TREE using ML-GTR+H2 or ML-GTR+PART over the 20 replicate MSAs at that Class 1 weight. The difference in BS between the partition and the GHOST models is small in comparison to the magnitude of the partition model BS, suggesting that with respect to branch length recovery IQ-TREE using the GHOST model performs as well as we could expect. Furthermore this distance increases as w_1 increases (as the sequence generated under Class 2 becomes shorter), implying consistency.	89
5.10	The mean inferred weights for Classes 1 and 2 of the Specific Case. The weight of Class 1 is shown on the x-axis, the inferred weight is shown on the y-axis. The data points indicate the mean weight inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the weights for Class 1 and Class 2. The results indicate that IQ-TREE was able to accurately recover the weights of the two classes for Specific Case simulations.	90
5.11	Soft classification of sites to classes - the probability of a site belonging to Class 1 is shown on the y-axis, the two Classes are shown on the x-axis. The boxplots clearly show that sites generated under Class 1 parameters are classified as having a higher probability of belonging to Class 1 than sites generated under Class 2.	94
6.1	The AIC scores achieved by varying the number of classes while fitting an ML-GTR+H model. The results indicate that 4 is the optimal number of classes for this dataset.	98

6.2	The AIC scores achieved by varying the number of classes while fitting an ML-GTR+H model. For each class, m , 100 ML-GTR+H m models were fitted to the data independently.	99
6.3	The four trees obtained from fitting the ML-GTR+H4 model to the electric fish data. The inferred weight of each class is indicated above each tree. Note the different scales for each tree, the dominant class (by weight) is much slower evolving than the three smaller classes. An indication of this is the total tree length (TTL) for the four classes: $TTL_1 = 0.19$, $TTL_2 = 5.12$, $TTL_3 = 3.35$ and $TTL_4 = 1.90$	100
6.4	The convergent class inferred by ML-GTR+H4. The 11 fish species comprised four South American electric fish (blue), one African electric fish (red), and six non-electric fish from various locations. The smallest class from the GHOST4 model shows that in comparison to the electric fish the non-electric species are relatively conserved.	102
6.5	Probability of sites belonging to the convergent class by codon position. The amino acid positions selected correspond with those identified by Zakon <i>et al.</i> as being critical to the inactivation of the Na^+ gene. The line at 0.1218 represents the average probability of belonging to the convergent class over all sites in the alignment. Sites at which nucleotide substitutions lead to functionally important amino acid replacements have a high probability of belonging to the convergent class. For example, at amino acid site 647 an otherwise conserved proline (codon CCN) is replaced by a valine (GTN) in the Pintailed Knifefish and a cysteine (TGY) in the Electric Eel. Substitutions at codon position 1 and 2 are necessary for both of these amino acid replacements and we find these sites have a high probability of belonging to the convergent class.	104

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.