MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

Declaration Confirming Content of Digital Version of Thesis

I confirm that the content of the digital version of this thesis

Title: Lineage Specific Evolution and Phylogenetic Analysis

is the final amended version following the examination process and is identical to this hard bound paper copy.

Student's Name: Liat Shavit Grievink

Student's Signature: Liat Shavit Grievink

Date: 12/8/09

LINEAGE SPECIFIC EVOLUTION AND PHYLOGENETIC ANALYSIS

A thesis presented in partial fulfillment of the requirements for the degree

of

Doctor of Philosophy

in

Biomathematics

at Massey University, Palmerston North,

New Zealand.

Liat Shavit Grievink

2009

ABSTRACT

Phylogenetic models generally assume a homogeneous, time reversible, stationary process. These assumptions are often violated by the real, far more complex, evolutionary process. This thesis is centered on non-homogeneous, lineage-specific, properties of molecular sequences. It consist several related but independent studies. LineageSpecificSeqgen, an extension to the Seq-Gen program, which allows generation of sequences with changes in the proportion of variable sites, is introduced. This program is then used in a simulation study showing that changes in the proportion of variable sites can hinder tree estimation accuracy, and that tree reconstruction under the best-fit model chosen using a relative test can result in a wrong tree. In this case, the less commonly used absolute model-fit was a better predictor of tree estimation accuracy. This study found that increased taxon sampling of lineages that have undergone a change in the proportion of variable sites was critical for accurate tree reconstruction and that, in contrast to some earlier findings, the accuracy of maximum parsimony is adversely affected by such changes.

This thesis also addresses the well-known long-branch attraction artifact. A non-parametric bootstrap test to identify changes in the substitution process is introduced, validated, and applied to the case of Microsporidia, a highly reduced intracellular parasite. Microsporidia was first thought to be an early branching eukaryote, but is now believed to be sister to, or included within, fungi. Its apparent basal eukaryote position is considered a result of long-branch attraction due to an elevated evolutionary rate in the microsporidian lineage. This study shows that long-branch estimates and basal positioning of Microsporidia both correlate with increased proportions of radical substitutions in the microsporidian lineage. In simulated data, such increased proportions of radical substitutions leads to erroneous long-branch estimates. These results suggest that the long microsporidian branch is likely to be a result of an increased proportion of radical substitutions on that branch, rather than increased evolutionary rate *per se*.

The focus of the last study is the intriguing case of Mesostigma, a fresh water green alga for which contradicting phylogenetic relationships were inferred. While some studies placed Mesostigma within the Streptophyta lineage (which includes land plants), others placed it as the deepest green algae divergence. This basal positioning is regarded as a result of long-branch attraction due to poor taxon sampling. Reinvestigation of a 13-taxon mitochondrial amino acid dataset and a sub-dataset of 8 taxa reveals that site sampling, and in particular the treatment of missing data, is just as important a factor for accurate tree reconstruction as taxon sampling. This study identifies a difficulty in recreating the long-branch attraction observed for the 8-taxon dataset in simulated data. The cause is likely to be the smaller number of amino acid characters per site in simulated data compared to real data, highlighting the fact that there are properties of the evolutionary process that are yet to be accurately modeled.

ACKNOWLEDGMENT

First and foremost, I would like to thank my supervisors Dr. Barbra Holland, Prof. David Penny, and Prof. Mike Hendy for allowing me to join their research group, for their expertise, guidance, suggestions, and encouragement. I am grateful for the opportunities they have offered me and helped me realize. This work would not have been possible without their open-door policy and immense patience. I would also like to acknowledge the financial support from the Marsden fund given to Barbara.

I am extremely grateful to Prof. Pete Lockhart for his interest in my study, extensive discussions, many valuable ideas and suggestions, and for his inspiring enthusiasm.

My sincere thanks go to Prof. Bill Martin for his involvement in this work, for sharing his ideas with me, for his kind hospitality, and for the financial support that has enabled my two visits to Dusseldorf.

Thanks also go to Dr. Tal Dagan, who has kindly shared her office in Dusseldorf with me, and other members of Bill's lab for helping me find my way around Dusseldorf.

I am grateful to Prof. David Bryant for his contribution to this study, and his great efforts to explain things clearly and simply.

Special thanks are due to Dr. Klaus Schliep for his keen help with R and Latex, Warwick Allen for getting me started with Perl, and Tim White for programming advice. I am also thankful to my office-mates Angela, Atheer, and Bennet, and other colleagues, for their company, discussions, and coffee breaks. Thanks also go to Susan Adams, Joy Wood, and Karen Sinclair for their kind help and support.

Constructive comments and suggestions from many participants of the NZ phylogenetic meeting, in the past 3 years, were very much appreciated. In particular, I would like to thank Prof. Mike Steel for helpful discussions.

I am most grateful to my friends, in Israel, The Netherlands, and here in New Zealand, who supported me throughout this study, and who accepted my limited social (and e-mailing) time slots. Special gratitude goes to Ofir, Evelyn, Nell, Aurelie, Estelle, and the lunch-time gang for the emotional support and their very much valued friendship.

Last, but not least, I would like to thank my close and extended family. I am greatly indebted to my husband, Hilbert. He has kept me grounded and sane through this journey, helped me keep things in perspective, celebrated my small successes with me, stood by me, and provided me with love, support, and encouragement, all while undertaking his own PhD study. I thank my parents, Rachel and Meir, for their love and support throughout my life and for their understanding when I decided to study on the other side of the world. Their belief in my ability to do anything I put my mind to gave me the determination to complete this study. This thesis is humbly dedicated to them.

CONTENTS