FACTORS AFFECTING THE PERFORMANCE OF PHYLOGENETIC METHODS

A THESIS PRESENTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF PH.D. IN
MATHEMATICS AT
MASSEY UNIVERSITY

Michael A. Charleston

1994

Massey University Library
Thesis Copyright Form

Title of thesis: A Factors Affecting The Performance of Phylogenetic Methods

(1) (a) I give permission for my thesis to be made available to readers in Massey University Library under conditions determined by the Librarian.

(b) I do not wish my thesis to be made available to readers without my written consent for .3. months.

(2) (a) I agree that my thesis, or a copy, may be sent to another institution under conditions determined by the Librarian.

(b) I do not wish my thesis, or a copy, to be sent to another institution without my written consent for .3. months.

(3) (a) I agree that my thesis may be copied for Library use.

(b) I do not wish my thesis to be copied for Library use for .3. months.

Signed M. Charleston

Date 26/1/94

The copyright of this thesis belongs to the author. Readers must sign their name in the space below to show that they recognise this. They are asked to add their permanent address.

NAME AND ADDRESS DATE

Acknowledgements

I am indebted to my supervisors, Michael Hendy and David Penny, who have with patience and wisdom guided me through the last three years' research. Without their enthusiasm, understanding and skill I am certain I would not have achieved this opus.

I am indebted also to my colleagues here in the Department of Mathematics and further afield, who have aided my study in countless ways. In particular, I would like to thank Richard Rayner for his computing support, Mark Byrne for his patience with my inane questions, Todd Cochrane for his tranquility, Peter Frizzell for his appreciation of classical music, Shane Dye for his enthusiasm, Marijcke Vlieg for her tolerance, Mark Johnston for his humility, John Giffin for his cynicism, and Mike Steel for his inspirational insight.

My friends, who have not forsaken me, I thank also. Nigel Green, Maree Sleeman, Lon Teal, Marian Trembath, Kathryn Hurr, and countless others have kept me more or less in touch with reality, and I appreciate it immensely.

I thank Julie Spicer for her love and understanding, treasured forever.

I thank my parents for their love and support throughout this endeavour, and for keeping me almost as sane as I was when I started.

To everyone else who has aided me through this period I give my thanks, and apologies that I cannot mention you all here.

Dedication

This thesis is dedicated to my parents: a small token, but heartfelt.

Abstract

This thesis comprises several computer simulation experiments in which the performance of a selection of phylogenetic methods was assessed. Data were generated according to a known model and used as input for the phylogenetic methods. Some new methods were introduced, and their performance compared with extant methods. Performance was judged by several criteria, being *accuracy*, *consistency*, *efficiency*, *falsifiability* and *robustness*.

The experiments were designed to be biologically relevant, and yet computationally tractible. Hence the models of evolution used were simple, to allow a wide range of parameters to be tested for their effects within the bounds of available computing resources.

The experiments were divided into two main types, the "small n" with up to 10 taxa, and the "large n" with from 10 to 30 taxa. Parameters which were allowed to vary in the "small n" case included number of taxa (n), sequence length, tree topology, edge length probability distribution, and purity of data. In the "large n" case, number of taxa, sequence length, and edge length probability distribution were varied.

The simulation experiments show that the accuracy of phylogenetic methods decreases with increasing n, and that the mean number of internal edges of the generating tree which are incorrectly inferred increases at least linearly with n. The rate at which the sequence length must increase with n, to retain a fixed confidence in the inferred tree, is shown to be at least linear in n.

All the methods are approximately as susceptible as each other to sampling error, which is exacerbated by the generating tree having very short or very long internal edges, and by finite sequence length. All the methods are susceptible to random error such as sequencing error, but provided such error is small, the effect is not great.

One type of method, using edge lengths inferred by the Hadamard conjugation

process, is shown to be much more robust to impure data and to sequencing error than are the other methods.

With $n \geq 10$ only the fastest methods were used. Increasing n again decreased the accuracy of the methods. Varying the "molecular-clockness" of the generating tree was shown to have a much greater effect upon those methods inconsistent with data which do not satisfy the molecular clock hypothesis.

All the methods used are described algorithmically, and their computational complexity is discussed. New proofs are provided of the consistency/inconsistency of several methods with the models of evolution used.

A notation is introduced to characterize all tree topologies, and used throughout this thesis.

Pseudocode is provided for all the major algorithms used in the simulation experiments.

Contents

List of Figures

List of Tables

List of Algorithms