

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

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. Charlton*

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

<b>Acknowledgements</b>	<b>ii</b>
<b>Dedication</b>	<b>iii</b>
<b>Table of Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xii</b>
<b>List of Algorithms</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Organization of this thesis . . . . .	1
1.2 The Problem . . . . .	4
1.3 This Study . . . . .	4
1.4 Definitions . . . . .	6
1.4.1 Mathematical terminology . . . . .	6
1.4.2 Phylogenetic terminology . . . . .	10
1.5 Models of evolution . . . . .	13
<b>2 Phylogenetic Methods</b>	<b>17</b>
2.1 Introduction . . . . .	17
2.2 Desirable Characteristics of Phylogenetic Methods . . . . .	18
2.2.1 Accuracy . . . . .	18
2.2.2 Consistency . . . . .	18
2.2.3 Efficiency . . . . .	18
2.2.4 Falsifiability . . . . .	20
2.2.5 Robustness . . . . .	20



2.3	General Classes of Phylogenetic Methods . . . . .	21
2.4	Constructive methods . . . . .	23
2.4.1	Unweighted Pair-Group Method with Arithmetic Mean . . . . .	25
2.4.2	Transformed Distance method (TD) . . . . .	25
2.4.3	Neighbourliness (ST) . . . . .	26
2.4.4	Neighbour-joining (NJ) . . . . .	27
2.5	Search methods . . . . .	28
2.5.1	Closest Tree (CT) . . . . .	28
2.5.2	Compatibility method (Co) . . . . .	29
2.5.3	Maximum Parsimony (MP) . . . . .	30
2.5.4	Branch and bound implementations . . . . .	31
2.6	A note on some other methods . . . . .	39
<b>3</b>	<b>New Methods</b>	<b>43</b>
3.1	Why find more methods ? . . . . .	43
3.2	The Distance Spectrum . . . . .	44
3.3	Compatibility — again . . . . .	46
3.4	SL . . . . .	47
3.5	NJa . . . . .	48
<b>4</b>	<b>Experimental methods</b>	<b>49</b>
4.1	Models of sequence evolution used . . . . .	49
4.2	Deviations of the data from a model . . . . .	50
4.2.1	Inadequate and contradictory models . . . . .	50
4.2.2	Sampling error . . . . .	51
4.2.3	“White noise” and “Pink noise” . . . . .	51
4.3	General approach . . . . .	52
4.4	Small $n$ . . . . .	53
4.4.1	Choosing the tree topology and other parameters . . . . .	53
4.4.2	Choosing the edge lengths . . . . .	57
4.4.3	Calculation of the expected bipartition frequencies . . . . .	57
4.4.4	Sampling from the expected bipartition spectrum . . . . .	58
4.4.5	Distance Calculation . . . . .	60
4.4.6	Inferring edge lengths . . . . .	60
4.4.7	Example . . . . .	62

4.5	Large $n$ . . . . .	66
4.5.1	Choosing a random tree . . . . .	67
4.5.2	Deriving the ancestral sequence . . . . .	69
4.5.3	Growing the data . . . . .	70
4.5.4	Parameters . . . . .	71
<b>5</b>	<b>Results 1: “Small <math>n</math>”</b>	<b>75</b>
5.1	Introduction . . . . .	75
5.2	Representation of findings . . . . .	76
5.3	Agreement between methods . . . . .	76
5.4	Sampling error . . . . .	78
5.5	Tree topology . . . . .	87
5.6	Edge length distribution . . . . .	91
5.7	Number of taxa . . . . .	108
5.7.1	Required growth in sequence length with number of taxa . . . . .	108
5.7.2	Optimal number of taxa for inferring a given edge . . . . .	111
5.8	Use of the Distance Spectrum . . . . .	113
5.9	White noise . . . . .	115
5.10	Pink Noise . . . . .	126
5.11	Summary . . . . .	131
<b>6</b>	<b>Results 2: “Large <math>n</math>”</b>	<b>135</b>
6.1	Computational considerations . . . . .	136
6.2	Sampling error . . . . .	136
6.3	Number of taxa . . . . .	138
6.4	Overall evolutionary time . . . . .	141
6.5	Time to last bifurcation event . . . . .	142
6.6	Edge length distribution . . . . .	145
6.6.1	Type of distribution . . . . .	145
6.6.2	Variance of the distribution . . . . .	146
6.7	Depth of edges . . . . .	148
6.8	Summary . . . . .	150
<b>7</b>	<b>Discussion</b>	<b>153</b>
7.1	Introduction . . . . .	153
7.2	Summary of results . . . . .	154

7.2.1	Sampling error . . . . .	154
7.2.2	Tree topology . . . . .	155
7.2.3	Number of taxa . . . . .	156
7.2.4	Edge length probability distribution . . . . .	157
7.2.5	Treatment of observed data . . . . .	157
7.2.6	Relationship between phylogenetic methods . . . . .	158
7.2.7	White noise . . . . .	158
7.2.8	Pink noise . . . . .	159
7.3	Comparison with some other studies . . . . .	159
7.4	That which may be . . . . .	163
7.4.1	Sources of error in finite data sets . . . . .	163
7.4.2	Properties of selection criteria . . . . .	164
7.4.3	Search strategies . . . . .	165
<b>A</b>	<b>Tree Topology Description Notation</b>	<b>171</b>
<b>B</b>	<b>Proofs of theorems</b>	<b>181</b>
<b>C</b>	<b>Pseudocode</b>	<b>195</b>
C.1	Introduction . . . . .	195
C.2	General functions . . . . .	196
C.3	Functions used in <code>sim.c</code> . . . . .	202
C.3.1	Clustering methods . . . . .	214
C.3.2	Search methods . . . . .	227
C.4	Main program structure of <code>sim.c</code> . . . . .	242
C.5	Functions used in <code>big.c</code> . . . . .	244
C.6	Main program structure of <code>big.c</code> . . . . .	250
<b>D</b>	<b>Dangers of Computer Simulation</b>	<b>253</b>
D.1	Tied Decisions . . . . .	253
D.2	Rounding error . . . . .	255
D.3	Programming errors . . . . .	256
	<b>Bibliography</b>	<b>257</b>

# List of Figures

1.1	Example of a graph . . . . .	8
4.1	Typical generating tree used in simulations . . . . .	63
4.2	Example incorrect tree inferred by some methods . . . . .	63
4.3	Spectra of edge lengths for an example tree . . . . .	66
4.4	Choosing a rooted binary tree from a given permutation. . . . .	68
4.5	Two rooted binary trees on three pendant vertices . . . . .	69
4.6	An example of a rooted tree used in <code>big.c</code> . . . . .	72
5.1	The agreement of phylogenetic methods . . . . .	77
5.2	UPGMA vs. $c$ with $n = 10$ and all trees equally likely . . . . .	80
5.3	TD vs. $c$ with $n = 10$ and all trees equally likely . . . . .	81
5.4	NJ vs. $c$ with $n = 10$ and all trees equally likely . . . . .	82
5.5	Mean proportion of trials in which the correct tree is inferred . . . .	85
5.6	Performance of NJ with all tree topologies of the same diameter, with 10 taxa . . . . .	88
5.7	Effect of varying $r$ with maximum path length $\sigma = 0.112$ . . . . .	93
5.8	Effect of varying $r$ with maximum path length $\sigma = 0.35$ . . . . .	95
5.9	Effect of varying $r$ with maximum path length $\sigma = 1.12$ . . . . .	97
5.10	Effect of varying maximum path length $\sigma$ with $r = 0.16$ . . . . .	101
5.11	Effect of varying maximum path length $\sigma$ with $r = 0.5$ . . . . .	103
5.12	Effect of varying maximum path length $\sigma$ with $r = 1$ . . . . .	105
5.13	Minimum required growth rate in $c$ with $n$ for 85% confidence in inferred tree . . . . .	109
5.14	The mean number of edges wrongly inferred with increasing $n$ . . . .	112
5.15	Compatibility methods with sequencing error rate $e = 0.025$ . . . . .	117
5.16	Compatibility methods with sequencing error rate $e = 0.1$ . . . . .	118
5.17	Closest tree methods with sequencing error rate $e = 0.04$ . . . . .	119

5.18	Closest tree methods with sequencing error rate $e = 0.064$ . . . . .	120
5.19	Maximum parsimony methods with sequencing error rate $e = 0.1$ . .	121
5.20	Distance Hadamard methods with sequencing error rate $e = 0.064$ .	122
5.21	Sequence Hadamard methods with sequencing error rate $e = 0.025$ .	123
5.22	Sequence Hadamard methods with sequencing error rate $e = 0.1$ . .	124
5.23	Constructive methods with sequencing error rate $e = 0.064$ . . . . .	125
5.24	Effect of amalgamating data from two trees . . . . .	129
6.1	Effect of sequence length with $n = 30$ . . . . .	137
6.2	Effect of number of taxa on the accuracy of constructive methods .	139
6.3	Distance from $T_G$ of inferred trees . . . . .	140
6.4	Effect of time from the first bifurcation to the present . . . . .	141
6.5	Inferred time of divergence of two sequences . . . . .	143
6.6	Effect of divergence time factor $f$ . . . . .	144
6.7	Effect of variance of edge length probability distribution . . . . .	147
6.8	Effect of edge depth on the probability of its correct inference . . .	149
7.1	Possible visualisation of the sources of error in inference of amount of evolutionary change . . . . .	164
7.2	Possible move in the Hitch-hiking heuristic search strategy . . . . .	167
A.1	The three binary unrooted trees on 4 pendant vertices. . . . .	172
A.2	Example tree $T$ for which the TTDN is sought . . . . .	173
A.3	Skeleton of tree $T$ in the previous figure. . . . .	173
A.4	The first stage in reconstructing a tree from its TTDN. . . . .	174
A.5	The second stage in reconstructing a tree from its TTDN. . . . .	175
A.6	The third stage in reconstructing a tree from its TTDN. . . . .	175
B.1	Tree $T$ used in the proof that TD is inconsistent. . . . .	184
B.2	The three unrooted binary trees on four pendant vertices . . . . .	190
D.1	An example labelled tree . . . . .	255

# List of Tables

2.1	The proportion of trees tested using branch and bound. . . . .	40
3.1	A classification of some phylogenetic methods . . . . .	44
3.2	The bipartitions and even-ordered subsets of $\{1, \dots, 6\}$ . . . . .	45
4.1	The number of trees with each topology for $4 \leq n \leq 10$ . . . . .	55
4.2	Edge-length probability distributions for different diameters of generating tree. . . . .	57
4.3	Number of operations required to generate a bipartition frequency spectrum. . . . .	59
4.4	Typical distance matrices, true, observed and inferred. . . . .	64
4.5	Typical expected, observed and inferred edge lengths . . . . .	65
5.1	Half the mean partition distance between inferred trees of different methods . . . . .	78
5.2	The net disagreement of some phylogenetic methods . . . . .	79
5.3	Effect of tree topology on performance of phylogenetic methods . . . . .	90
5.4	Variance of edge lengths as inferred from the distance and bipartition spectra . . . . .	114
6.1	Effect of edge length probability distribution . . . . .	146
6.2	Number of edges with depth $k$ over all binary trees on 26 pendant vertices . . . . .	148
6.3	Effect of depth of edges on their correct inference . . . . .	150

# List of Algorithms

2.1	Clustering process . . . . .	24
2.2	Branch and bound for Co and CT . . . . .	35
2.3	Branch and bound for MP . . . . .	38
4.1	get_distances . . . . .	60
5.1	variance.c . . . . .	113
7.1	Example of Hitch-hiking . . . . .	168
C.1	compare_sets( $A, B, max$ ) . . . . .	196
C.2	Hadamard( $v, w$ ) . . . . .	197
C.3	HexpH(inv,outv) . . . . .	198
C.4	HlnH . . . . .	198
C.5	permute( $x, perm$ ) . . . . .	199
C.6	rough_exp( $i$ ) . . . . .	199
C.7	permutation_to_tree( $perm, output\_tree$ ) . . . . .	200
C.8	sample_uniform( $mean, range$ ) . . . . .	201
C.9	sample_normal( $mean, std\_dev$ ) . . . . .	201
C.10	sample_log_normal( $mean, std\_dev$ ) . . . . .	202
C.11	compat( $x, f, A$ ) . . . . .	202
C.12	choose_topology( $topnumber, edge\_set$ ) . . . . .	203
C.13	bipartitions_to_distances( $v$ ) . . . . .	205
C.14	choose_tree( $x, edge\_set$ ) . . . . .	206
C.15	correct_distances . . . . .	207
C.16	get_pathsets( $D$ ) . . . . .	208
C.17	number_of_bipartitions . . . . .	209
C.18	random_edge_lengths( $edge\_set, outv$ ) . . . . .	210
C.19	sample_bipartitions( $length, error\_rate, inv, outv$ ) . . . . .	211
C.20	sort_vector_descending( $inv, outv$ ) . . . . .	213
C.21	NJ( <i>averaging</i> ) . . . . .	214

C.22	ST . . . . .	217
C.23	UPGMA( <i>version</i> ) . . . . .	220
C.24	TD( <i>version</i> ) . . . . .	224
C.25	CT( <i>v, distances, use_Hadamard</i> ) . . . . .	227
C.26	Co( <i>v, distances, use_Hadamard</i> ) . . . . .	231
C.27	set_up_first_tree( <i>A</i> ) . . . . .	234
C.28	add_taxon( <i>taxon, position, A, new_node</i> ) . . . . .	234
C.29	remove_taxon( <i>taxon, position, A, node</i> ) . . . . .	234
C.30	convert_edges_to_tree( <i>edge_set, tree_array</i> ) . . . . .	235
C.31	convert_tree_to_edges( <i>tree_array, edge_set</i> ) . . . . .	236
C.32	Fitch( <i>A, nodes, v, bound</i> ) . . . . .	237
C.33	MP( <i>v, distances, use_Hadamard</i> ) . . . . .	239
C.34	sim.c . . . . .	242
C.35	get_bif_times( <i>output_bif_time</i> ) . . . . .	244
C.36	ones_count( <i>z</i> ) . . . . .	244
C.37	get_whole_tree . . . . .	245
C.38	grow_data . . . . .	247
C.39	generalJC . . . . .	249
C.40	big.c . . . . .	250
D.1	Typical loop to find $\{i, j\}$ to maximise $f(i, j)$ . . . . .	254