DISTRIBUTIONS ON BICOLOURED EVOLUTIONARY TREES

A thesis presented in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy
in Mathematics at
Massey University

Michael Anthony Steel
February, 1989

Massey University Library

Thesis Copyright Form

Title of thesis: DISTRIBUTIONS ON BICOLOURED EVOLUTIONARY TREES

(1) (a) I give permission for my thesis to be made available to readers in the 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 _____ 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 _____ 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 _____ months.~~

Signed M A Steel

Date 1/3/89

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

M. A. Steel
c/- P&L Read No 1 Lic RD5
Palmerston North NZ

DATE

1/3/89

ABSTRACT

A central and challenging problem in contemporary biology is how to accurately reconstruct evolutionary trees from DNA sequence data. This thesis addresses three themes from this endeavour -- comparison, consistency and confidence intervals -- by analysing distributions arising from phylogenetic trees.

Toward the first theme, the distribution of the symmetric difference metric on pairs of binary and phylogenetic trees is studied, and a number of new results obtained. These theorems, as well as a result on another tree metric answer previous conjectures in this area. Also under the theme of comparison, we analyse distributions on bicoloured trees arising from the principle of parsimony. A streamlined proof is given of an elegant theorem which allows an efficient comparison of how much better a maximum parsimony tree fits given data than a randomly-chosen tree. A dual distribution, where the tree is fixed and the data varies is also analysed, answering a recent unsolved problem.

We then consider the theoretical accuracy of tree-building methods, concentrating on the statistical property of consistency. Under a simple stochastic model on bicoloured trees, conditions for the consistency of frequently-used methods based on parsimony and compatibility are examined. It is shown that even in "best possible" conditions both methods can be inconsistent, though a strong sufficient condition for compatibility is given. The analysis is extended for a molecular clock.

Finally, procedures are described for placing confidence intervals around phylogenies, and limitations on the sort of confidence intervals possible are given. Ways to efficiently implement these procedures are then considered -- in particular, approximate methods, applications to sets of taxa of size four, and simplifications under a molecular clock.

The rate that sequence data must grow as a function of the number of taxa for confidence intervals to converge to a single tree is also considered.

The arguments in this thesis are primarily combinatorial and stochastic. In the hope that their implications will also interest biologists, some space has been given to motivating and explaining the biological relevance of the results presented.

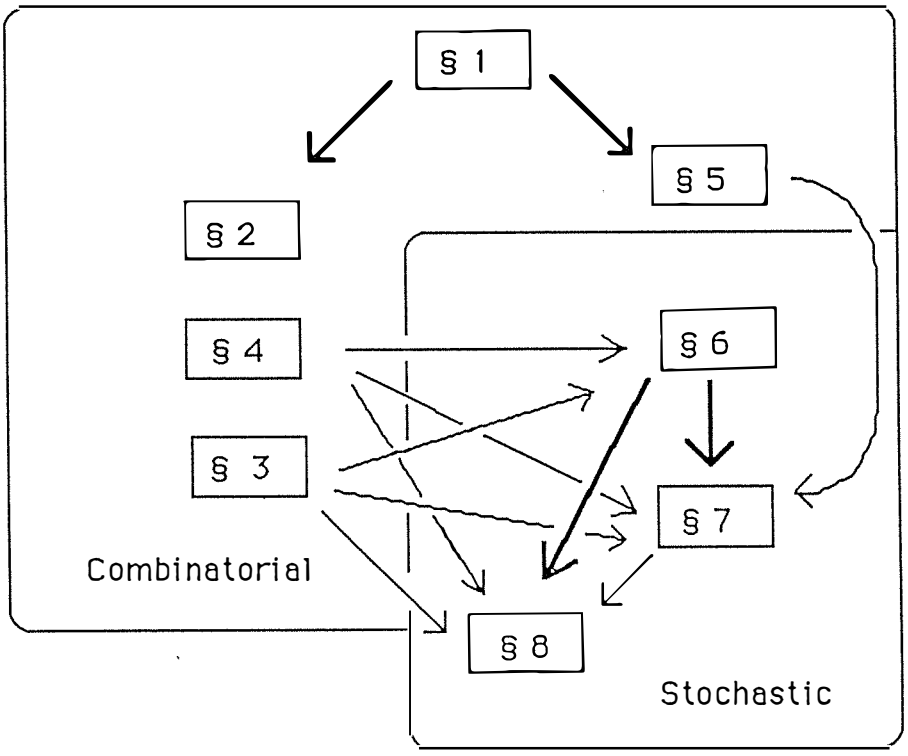

Chart illustrating the flow of results from one section to another

Contents: Page:

Introduction 1

Notation/Table of Symbols 3

Section One: Counting Trees

--trees 5
--resolved subtrees 11
--forests 15
--bicoloured trees 18

Section Two:
Distribution of the Symmetric Difference Metric

--distribution generating functions 29
--distribution on pairs of trees 34
--absolute (non-asymptotic) inequalities 38
--asymptotic range of the distribution 43
--monotonicity 46
--description of the metric from below 48
--distribution on PT(n) 51
--comparison with other metrics 54

Section Three: Subtree constraints

--induced subtrees and minimal similarity 56
--spanning sets 62
--consensus trees 66
--efficiency 68

Section Four: Sequence and Dissimilarity Data

--constraints with two colours 73
--lower bounds on information loss 76

Section Five: Combinatorics of Parsimony

--vector spaces of edge sets 82
--(weakly-) connecting trees and forests 84
--distributions arising from parsimony 90
--path/edge duality 100

Section Six: Analysis of Cavender's model

--stochastic preliminaries 108
--the model 109
--central observations 111
--invariants 113
--partition probabilities 118

Section Seven: Consistency

--selection procedures 127
--convergence and consistency 127
--consistent recovery of trees from dissimilarities 132
--consistency of parsimony and compatibility 135
--sufficient conditions 147
--consistency under a molecular clock 160

Section Eight: Confidence Intervals

--confidence intervals 168
--approximate methods 172
--exact solutions 176
--efficiency (I) 178
--efficiency (II) 180
--a χ^2 test (molecular clock) 186

Appendix 193

References 196