



Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand). This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.
<http://researchspace.auckland.ac.nz/feedback>

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library [Thesis Consent Form](#)

Stochastic Geometry, Data Structures and
Applications of Ancestral Selection Graphs

Nicoleen Cloete

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Mathematics at the University of Auckland, 2005

To Danie, Renier and Nicola

May all your dreams come true

Contents

Notation	iii
Abstract	ix
Acknowledgment	xi

Notation

n : Size of sample taken from a population of organisms, 19

N : Size of the population of organisms, 19

λ_0 : Type 0 birth rate, 19

λ_1 : Type 1 birth rate, 19

UA : Ultimate ancestor, 19, 22

$MRC A$: Most recent common ancestor, 3, 29, 75, 77

t_{UA} : Time to ultimate ancestor, 20, 22

$Z_0(t_0)$: Number of allele type 0 at time t , 20

$Z_1(t_0)$: Number of allele type 1 at time t , 20

θ : Mutation parameter, 20

σ : Selection parameter, 20

z : Frequency of allele type 0, 20

$\eta(z)$: Density of the diffusion process, 20

K : Normalizing constant, 20

$\mathcal{G}_n(t)$: Ancestral selection graph process with n leaves, 21

\mathcal{A} : Set of labels assigned to the sample, 21

$G = (V, E)$: Directed graph with vertex labels V and edge labels E , 21

M : Total number of events in one realization of a graph, 21

t_v : Time at vertex v , 24

$\mathcal{A}(t)$: Set-valued process denoting the set of particles at time t , 21

\mathcal{R}_m : Jump process of times at which events occur, 21

$|\mathcal{A}(t)|$: Cardinality of the set \mathcal{A} at time t , 22

k : Number of edges between two events, 22, 26

B_v : Particle type at vertex $v \in V$ where $B \in \{0, 1\}$, 23

V_L : Set of leaf vertex labels, 23

V_A : Set of internal or ancestral vertex labels, 23

V_c : Set of vertex labels representing coalescent events, 24

V_b : Set of vertex labels representing branching events, 24

E_g : Set of edge labels, 24

E_c : Set of edge labels with coalescent events at the top, 24

E_b : Set of edge labels with branching events at the top, 24

N_c : Total number of coalescent events in a graph g , 24

N_b : Total number of branching events in a graph g , 24

d_{in} : In-degree of a vertex, 25

d_{out} : Out-degree of a vertex, 25

$\chi_{V,E}$: Space of admissible vertex times for a given topology V, E , 25

S_i : Gating value of edge i , 25

$\mathcal{S}_{V,E}$: Set of admissible assignments of selection gating values, 25

$\Gamma_{N_b, n}^*$: Set of admissible ancestral graph topologies V, E on N_b branching events
and n leaves, 25

(Γ, \mathcal{F}, P) : Space of unlabeled ancestral graphs with σ algebra of subsets \mathcal{F}
and probability measure P , 26

ρ_m : Total rate of coalescent and branching events in the interval t_{m-a}, t_m , 26

$v^{(i)}$: Vertex at the top of edge i , 28

$v_{(i)}$: Vertex at the bottom of edge i , 28

$\langle v_{(i)}, v^{(i)} \rangle$: Edge connecting vertex $v_{(i)}$ with vertex $v^{(i)}$, 28

B_{UA} : Ancestral type of the ultimate ancestor, 31

p_0 : Probability that ultimate ancestor is of type 0, 31

p_1 : Probability that ultimate ancestor is of type 1, 31

τ_i : Length of edge i , 31

$P_{B_{v_{(i)}}, B_{v^{(i)}}}$: 2×2 transition matrix associated with edge $\langle v_{(i)}, v^{(i)} \rangle$, 31

B_I : Set of allele types of the individuals at the leaf tips, 32

B_Y : Set of allele types at the internal vertices, 32

$f_G(g|\sigma)$: Population density on the space Γ of unlabeled graphs, 27

$P(B|\theta, g)$: Likelihood of the leaf types given θ and g , 31

$h(g, B_Y, \theta, \sigma|B_I)$: Posterior density, 33

$\Phi_n = (g, B_Y, \theta, \sigma)$: The n^{th} update of the Markov Chain, 36

p^\dagger : Probability of deleting an edge, 38

p^* : Probability of adding an edge, 38

$|E|$: Total number of edges in a graph, 41

$|E_D|$: Total number of deletable edges in a graph, 42

V_{AR} : Label of the vertex above the root, 56

$Tr_{(V,E)}$: Genealogy of an ancestral selection graph with topology (V, E) , 79

\mathcal{E} : Set of events in a graph, 87

(β_m, γ_m) : Label process describing the topology of the graph, 89

β_m : The label of the edge branching if the m^{th} event is a branching event, 89

γ_m : The labels of the two coalescent edges if the m^{th} event is a coalescent event, 89

Σ : The set of selection parameter values used for experiments, 102

$w(g, B_Y, \sigma, \theta | B_I, Tr)$, 102

$k(g, B_Y | \sigma \in \Sigma, B_I)$, 102

LI : Ladder index, 107

ζ : Statistic of interest, 126

τ_ζ : Integrated autocorrelation time, 126

$v_\zeta(s)$: Auto-covariance at lag s , 126

\hat{var}_ζ : Estimated variance of statistic ζ , 126

t_{FE} : Time to the first event, 127

D : Data at the leaf tips, 136

$Tr_{0.5}$: Genealogy when $\sigma = 0.5$, 138

Tr_{exp} : Genealogy for exponential growth, 148

$Tr_{0.5}^{(15)}$: Genealogy when true $\sigma = 0.5$ and $\sigma \sim U(1.5)$, 152

$D_{0.5}$: Data simulated when $\sigma = 0.5$, 138

$w(\sigma)$: Marginal posterior density of σ , 140

se : Standard error, 152

acct : Acceptance rate, 152

Abstract

The genealogy of a random sample of a population of organisms can be represented as a rooted binary tree. Population dynamics determine a distribution over sample genealogies. For large populations of constant size and in the absence of selection effects, the coalescent process of Kingman determines a suitable distribution. Neuhauser and Krone gave a stochastic model generalising the Kingman coalescent in a natural way to include the effects of selection.

The model of Neuhauser and Krone, determines a distribution over a class of graphs of randomly variable vertex number, known as ancestral selection graphs. Because vertices have associated scalar ages, realisations of the ancestral selection graph process have randomly variable dimensions.

A Markov chain Monte Carlo method is used to simulate the posterior distribution for population parameters of interest. The state of the Markov chain Monte Carlo is a random graph, with random dimension and equilibrium distribution equal to the posterior distribution.

The aim of the project is to determine if the data is informative of the selection parameter by fitting the model to synthetic data.

Acknowledgment

I want to acknowledge my supervisors Dr. Geoff Nicholls from the Department of Mathematics and Assoc. Prof. David Scott of the Department of Statistics for their help, suggestions and encouragement throughout the project. Our weekly discussions will be missed. Dr Nicholls helped me in formulating the project and spent hours helping me to find bugs in the code. Assoc. Prof Scott was very helpful especially at the end when Dr. Nicholls left the University of Auckland. His suggestions on the content of the thesis are greatly appreciated.

I wish to thank Prof. Allen Rodrigo of the Bio-informatics Institute of the University of Auckland for his involvement during the initial stages of this project.

As the recipient of a Bright Future Top Achiever Doctoral Scholarship, I want to thank the Foundation of Science and Industry of New Zealand for the financial support. Without the support this thesis would not have been possible.

I also want to thank all my friends - there are too many to name - but espe-

cially Prof. John Butcher for the moral support throughout my PhD.

The last but certainly not the least, thank you to the most important people in my life, my husband Danie and my two children Renier and Nicola. You guys are better than A1. Sometimes the going was tough, but still you supported me and were always there with smiles.