Master's thesis

Designing and implementing a coalescence tool based on phase-type distributions

Tobias Bovbjerg Røikjer, 201505878 (tobiasroikjer@gmail.com)

Advisor: Kasper Munch Terkelsen

Aarhus University
Aarhus, Denmark

Master's degree in Bioinformatics

June 2020

Abstract

Large parts of the inference of population demographics and history is based on analysing the current genetic diversity sampled from individuals in the populations of interest. Underlying evolutionary models often used are based on coalescence theory and the distribution of coalescence trees. Knowing and working with the full distribution of coalescence trees gives us more insight into the cause of genetic diversity. However, the state-space of such distributions is often very large. Many approaches are therefore based on approximations and sampling from the distribution.

We consider approaches from computer science and apply them to continuous and discrete phase-type distributions modelling coalescence trees. Formalizing the computation of distributions of coalescence trees as graph problems allows us to compute both the state-space and properties of the distribution of coalescence trees faster and using less memory without approximating the statistical distributions.

We develop and implement algorithms that can construct and reward transform phase-type distributions as a graph. This generalization allows us to reduce the state-space depending on the properties of interest.

We show fast algorithms for computing the expected value and variance of acyclic continuous phase-type distributions without first reward transforming them. This allows us to find the variance of popular estimators of the population mutation rate θ and show an unbiased linear minimum variance estimator of θ. Similarly, we demonstrate an algorithm that reduces an acyclic continuous phase-type distribution to an exponential mixture distribution, also without reward transforming first, assuming unique rates.

We show how discrete phase-type distributions can be used to model mutations in a coalescence model, and how we can find the probability mass function of certain estimators of θ, e.g. Watterson's and the pairwise estimator.

We also expand the demographics that might be modelled by phase-type distributions by an isolation with migration model, with an approximation of the isolation period. This gives the marginal distribution of sites in a joint site frequency spectrum. In comparison, the expected value is very similar to results found by the program $\delta a \delta i$.

Resumé

Analyse af den nuværende genetiske diversitet samplet fra populationers individer danner ofte grundlaget for inferens af populationernes demografi og historie. Underlæggende evolutionsmodeller er ofte baseret på coalescenceteorien og fordelingen af coalescencetræer. Kendskabet og arbejdet med den fulde fordeling af coalescencetræer giver større indsigt i årsagen til genetisk diversitet. Dog er state-spacet af sådanne fordelinger ofte meget store, og mange tilgange er derfor baseret på approksimationer og sampling fra fordelingen.

Vi ser på fremgangsmåder fra datalogi og anvender dem på kontinuære og diskrete fase-type fordelinger, som modellerer coalescencetræer. Beregningsformaliseringen for fordelinger af coalescencetræer som grafproblemer giver os mulighed for at beregne både state-spacet og fordelingens egenskaber hurtigere og med mindre hukommelse. Dette uden at approksimere de statistiske fordelinger.

Vi udvikler og implementerer algoritmer, som kan konstruere og reward-transformere fase-type fordelinger som en graf. Generaliseringen giver os mulighed for at reducere state-spacet, afhængigt af de egenskaber som har interesse.

Vi viser hurtige algoritmer til beregning af den forventede værdi og varians af acykliske kontinuære fase-type fordelingen uden først at have reward-transformeret dem. Dette gør, at vi kan finde variansen af populære estimatorer af populationsmutationsraten θ, og vi viser en lineær estimator med minimumsvarians uden bias for θ. Lignende viser vi en algoritme, der reducerer en acyklisk kontinuær fase-type fordeling til en eksponentiel mixture-fordeling. Denne er også uden at reward transformere først, under antagelse af unikke rates.

Vi viser, hvordan diskrete fase-type fordelinger kan bruges til at modellere mutationer i en coalescencemodel, og hvordan vi kan finde sandsynlighedsfunktionen for visse estimatorer af θ, fx Wattersons og den parvise estimator.

Vi udvider også demografierne som fase-type fordelingerne kunne modellere med en isolation-med-migrationsmodel, som har en approksimation af isolationsperioden. Dette giver den marginale fordeling af sites i et joint site frequency spectrum. Til sammenligning er den forventede værdi meget tæt på resultaterne fundet via programmet $\delta a \delta i$.

Acknowledgments

I am thankful for the advice, time, and motivation of my supervisor Kasper Munch Terkelsen. The culture at the Bioinformatics Research Centre is amazing, and I appreciate the community formed by students and faculty alike. Daily life at BiRC has been wonderful with all the students from very different backgrounds. Thanks to my office, the student union, the Christmas committee, the board games and beers. A special thanks to my wife Lærke, my true purpose in life.

Notation

iff	if and only if, sometimes written $\Leftrightarrow$
convolution	sum of independent stochastic variables
$\boldsymbol{\alpha}$	bold lowercase Greek letters are row vectors
$\boldsymbol{s}$	bold lowercase Latin characters are column vectors
α_i, s_i	entry i of $\boldsymbol{\alpha}$ and $\boldsymbol{s}$
$\boldsymbol{e}$	a vector with all entries of 1
$\triangle(\boldsymbol{\alpha})$	a diagonal matrix with entries of the vector $\boldsymbol{\alpha}$
$\boldsymbol{S}$	bold uppercase letters are matrices
$\boldsymbol{S}^T$	matrix transpose
$e^{\boldsymbol{S}}$	matrix exponential of the matrix $\boldsymbol{S}$
$\boldsymbol{S}_{i,\cdot}$	row i of the matrix $\boldsymbol{S}$
$\boldsymbol{S}_{\cdot,j}$	column j of the matrix $\boldsymbol{S}$
$s_{i,j}$	entry (i,j) of the matrix $\boldsymbol{S}$
$\boldsymbol{I}$	the identity matrix
$\boldsymbol{0}$	zero matrix or zero vector
$\mathbb{N}$	the set $\{0,1,2,\dots\}$, the natural numbers
$\mathbb{R}$	the real numbers
$\mathbb{N}^n$, $\mathbb{R}^n$	the natural and real numbers in n dimensions
$\lvert V \rvert$	number of elements in the finite set V
$V \setminus v$	the set V where the element v is removed
$\emptyset$	the empty set
$a \leftarrow a + 1$	an algorithm increments the variable a by one
FUNC$(1, "abc")$	Invocation of the algorithm FUNC with the arguments 1 and "abc"
$(v \to u)$	an edge from the vertex v to u
$w(v \to u)$	the associated weight of the edge v to u
DAG	directed acyclic graph
Exp, Geo	exponential distribution, geometric distribution
PH, DPH	continuous phase-type distribution, discrete phase-type distribution
PDF, CDF, PMF	probability density function, cumulative distribution function, probability mass function
$f(t)$, $F(t)$, $\bar{F}(t)$	the letter is mostly used for the PDF, the CDF, and survival function (1-CDF)
$\hat{\theta}$	an estimator of the parameter θ

Contents

1 Introduction **1**
- 1.1 Background . . . 1
- 1.2 Kingman coalescence . . . 2
- 1.3 Summary statistics of genetic diversity . . . 3
- 1.4 Phase-type distributions . . . 3
- 1.5 Reward transformation . . . 5
- 1.6 Phase-type distributions in population genetics . . . 5
- 1.7 Graph-based approach . . . 6
- 1.8 Decomposition of phase-type distributions . . . 7
- 1.9 Acyclic phase-type distributions . . . 9

2 Generalization of state-space generation **13**
- 2.1 Generic state-space algorithm . . . 16
- 2.2 Reward transformations and the state-space . . . 16
- 2.3 Combining state-space, reward transformation, and properties . . . 18
- 2.4 Unlabelled lineages . . . 19
- 2.5 Implementation of the unlabelled Kingman state-space . . . 21
- 2.6 Can the state-space be reduced? . . . 22

3 Moments of acyclic multivariate phase-types **24**
- 3.1 An unbiased linear minimum variance estimator of θ . . . 24
- 3.2 Computing the moments . . . 28
- 3.3 Comparison with other algorithms . . . 29

4 The distribution function of acyclic phase-types **31**
- 4.1 Cumani's basic path decomposition . . . 31
- 4.2 Exponential mixture distribution . . . 32
  - 4.2.1 Path-based algorithm . . . 33
  - 4.2.2 Algorithm based on form-invariance under convolutions . . . 34
- 4.3 Comparison with other algorithms . . . 36
- 4.4 Implementation of weight finding algorithm . . . 36
- 4.5 Random sampling . . . 38
- 4.6 Acyclic multivariate phase-types . . . 38

5 Mutations on the coalescence tree **41**
- 5.1 Marginal distribution of mutations . . . 41
  - 5.1.1 Discrete phase-type distributions . . . 41
    - 5.1.1.1 The Hobolth-Siri-Jégousse-Bladt DPH construction . . . 42
    - 5.1.1.2 Acyclic DPH construction . . . 42
  - 5.1.2 Non-DPH approaches . . . 43
    - 5.1.2.1 Acyclic dynamic programming on geometric distributions . . . 43
    - 5.1.2.2 Acyclic graphs with same path length . . . 44
    - 5.1.2.3 Mixture of geometric distributions . . . 44
- 5.2 Joint distribution of mutations . . . 45
  - 5.2.1 The Navarro approach . . . 45

5.2.2 Acyclic dynamic programming on multivariate geometric distributions 45
5.2.3 Multivariate normal approximation 45
5.3 Weighting the sites in the SFS 46
5.3.1 Simpler approach for acyclic graphs with same path length 46
5.4 Distribution of statistics on the SFS 47

6 State-space algorithm for isolation with migration **51**
6.1 Parameterization of the state-space 51
6.2 State representation 52
6.3 State-space algorithm 55
6.4 State-space size 56
6.5 Comparison with other algorithms 56
6.6 Converting discretized time in number of coalescence events to continuous time in years 57
6.7 A bias in the model from the isolation period 57

7 Disallowing back-migrations in isolation with migration **59**
7.1 Distribution of branch lengths 59
7.2 Distribution of sites 60
7.3 Is no back-migrations a good assumption? 60

8 Marginal distribution of sites in the joint site frequency spectrum **63**
8.1 Comparing expected value to $\delta a \delta i$ 63
8.2 The distribution of sites against migration rate and isolation time 65

9 Does the isolation with migration model match reality? **66**
9.1 Discretized time 66
9.2 No recombination 69

10 Discussion **70**
10.1 Future work 71
10.2 Conclusion 72

11 Methods **77**

12 Supplementary **79**