The University of Melbourne

Department of Mathematics and Statistics

Population History Estimation: A Coalescent Theory Approach

Author:
Xuehua Lan

Supervisor:
Nathan Ross

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

May 2017

Dedicated to all my teachers
and
to my parents.

Acknowledgments

My foremost gratitude goes to my supervisor, Nathan Ross.

Contents

List of Figures

List of Notations

$\mathbb{R}$ real number.

$\lfloor x \rfloor$ the biggest integer less or equal to x.

$\lceil x \rceil$ the smallest integer greater or equal to x.

$\mathbf{1}_{\{\cdot\}}$ indicator function.

$\mathcal{D} := (\mathcal{D}, d)$ pseudo-metric or metric space of set $\mathcal{D}$ with metric d.

$(\Omega, \mathcal{F}, \mathbf{P})$ probability space, where $\mathcal{P}$ is a family of probability measure, and $\mathcal{F}$ is a σ-algebra.

$co(\mathcal{P})$ the convex hull of $\mathcal{P}$.

$\mathbb{E}(\cdot)$ expectation of the interest.

$\theta(P)$ parameter of interest with values in $(\mathcal{D}, d)$.

$\hat{\theta} = \hat{\theta}(X)$ an estimator of $\theta(P)$ based on an X with distribution P.

$d_{TV}(P, Q)$ the total variation distance between P and Q.

$d_H(P, Q)$ the Hellinger distance between P and Q.

$d_{KL}(P||Q)$ the Kullback Leibler distance from Q to P, also known as relative entropy.

List of Abbreviations

RVs	Random variables
SNP	Single Nucleotide Polymorphism
MRCA	Most Recent Common Ancestor
WF	Wright-Fisher, refer to a coalescent model
KL	Kullback-Leibler, refer to a statistical distance
AFS	Allele frequency spectrum, some may refer as site frequency spectrum
SFS	Sample frequency spectrum
AG	Ancient genealogy
TG	Truncated genealogy

Preface

This presented thesis is the research project of the Master of Science degree. In this thesis, the error of estimating the population history on a coalescent inference approach will be presented.

Coalescent theory is one tool in theoretical population genetics to mathematically model the process of evolution, it is a stochastic process that describes the probability distribution on ancestors at the genetic levels and formulates the genetic drift backwards in time. Along the process, one can reconstruct past genealogy from present day genetic data. Given a past population shape, the distributions of times of coalescence and the expected sample frequency spectrum (SFS) can be used to infer this history. In this thesis, we look at the statistical error made when using these two methods for inference. First we derive bounds on the error made in the setting of populations with a bottleneck, in terms of the amount of data used. We then expand our interests to population shapes that have split from one population, constructing the density of coalescent times and then the distribution of expected SFS. Finally we use these derivations to obtain bounds on the error when using these two methods for inference, again in terms of the amount of data used.

This thesis consists of 6 chapters. Firstly, we introduce the developments in coalescent theory in chapter 1. Chapter 2 is devoted to prior knowledge in metrics and functions, probability theory, information theory, minimax theory, and coalescent theory. Then in chapter 3 some useful techniques and results that can help us to derived outcomes in chapter 4 will be explored. Although this chapter focuses on inferring in a single population, there are broader insights one can gained for further inference between multiple populations. We start looking at the population history estimation in chapter 4 and focus in a single population in this chapter. Two developed methods will be explored to construct the distributions that can reveal the population size functions, based on these two methods, we derived minimax bounds and compared in terms of the amount of data. In chapter 5, we turn our interests to population shapes that have split from one population, which is also known as diverging populations. In terms of two distribution methods explored in chapter 4, we construct these two distributions with respect to a diverging population, and output a minimax lower bounds between distributions for comparison. A general conclusion will be given in chapter 6, with a discussion on strengths and limitations.

Chapter 1

Introductions

In this first chapter, we introduced the developments in coalescent processes.

We firstly introduced the Kingman coalescent, and briefly look at the drift of gene itself, to other forces varies the processes, such as mutation, recombination, and flows of population. Along with discussing the limitations of Kingman coalescent, other types of coalescent will be introduced.

One powerful strategy for extracting information from genetic data is to mathematically model the process of evolution. Theoretic population genetic was formed at the starts of the 20's, half century after Darwin found the evolution of traits. In the processes on formulating the evolution of genomes with polymorphism data, the foundations are grounded on stochastic formalization by Fisher [6], Wright [22], and Haldane [8] in 1930's. Wright-Fisher is a cornerstone model in population genetics, with binary selection scheme underlies, it assuming in each generation at most one coalescent event happen to one pair of individual, with $2N$ haploid individuals, or N diploid individuals. All individuals in each generation are considered to reproduced next generations and then die at the end of each generation.

Diffusion approach was firstly developed to capture the neutral net movements between genes in 1950. Not until 1982, Kingman [11] introduced the coalescent approach, this approach allow us to reconstruct the genealogy from present and relax preferred selection in the process.

Kingman coalescent rely on Wright-Fisher model and taking limits in continuous time, the coalescent process can be considered as a homogenous pure death processes start at n lineages (states), see figure 1.1 for example, and eventually coalescent into 1 lineages, which in coalescent theory known the most recent common ancestor (MRCA). The associated death rate in the each state is actually the frequency of the coalescent, and the statistics of the process can be formulated through the analysis of this frequency at each number of lineages (states). Under the binary coalescent setting as a pure death processes, the finite states (lineages) give advantages in inferences. Many existing results can be applied without consider conditional assumptions in the case of explosive continuous time process.

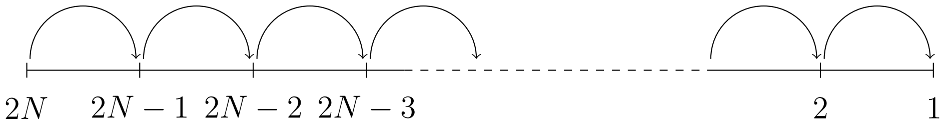

Figure 1.1: A pure death process for Kingman coalescent.

Wright-Fisher model is not the only model to approach the Kingman coalescent, it can also be approached by Moran model, introduced by Moran [14] in 1958. Both Wright-Fisher and Moran model are discrete and all the individuals waiting to coalescent are considered random selected and so are exchangeable, which introduced a generalized discrete model, the Canning model. As two special cases of the Caning model, the main differences between them is that Moran model allows overlapping in generations while Wright-Fisher model is not, and consequently the coalescent time of Moran model is $1/2N$ $(1/N)$ slower than WF model in haploid (diploid) individuals.

The models mentioned above are basic formulations of individuals' births and deaths, there are other forces along with the drift. Mutations are considered as one of the important forces of the coalescent processes, which occur occasionally in the processes by changing the gene information. To mimic the processes, infinite alleles models and infinite sites models are developed. These two models are based on different accessible genetic information, on the alleles and the segregating sites respectively. Infinite alleles models create a new allele in each mutation, so one should count number of alleles types and number of presentations. Mathematical proof shows the number of allele types converges in a reciprocal logarithm, but it is sufficient to provide all the information. Infinite sites model, on the other hand, mutation affects a site that was previously unaffected before, with only two bound nucleotides (adenine (A) paired with thymine (T), and guanine (G) paired with cytosine (C)), one can tells the mutant genes under infinite site setting.

There are other intrinsic forces such as recombination and natural selections. Not all these forces are mutually happen, in fact, in reality, allele have higher possibility to happen in neighborhood regions; and there are some loci are considered to have higher mutant rate than others, which are known as hot spot.

Extrinsic forces like population flows due to wars and diseases, industrial revolutions and economic booms, and migration are also factors that affect the drift of the coalescent processes, Kingman coalescent allows us to relax these forces in it.

Coalescent theory allow us to track back past genealogies, the applications of coalescent can be widely used as disease gene mapping, given that the distribution can be constructed on the ancestor of a group of sample genes, one can calculate the distributions of forces like mutations and recombination, to calculated the theoretical proof of the processes. It gives theoretic heterozygous of the population.

Readers may find our works mainly based on Kingman coalescent with Wright-Fisher model underlies, but should also note the limitations based on it. In addition that population N large enough, Kingman coalescent requires the number of offspring produced by each parent is small compared to the population size N in order to satisfies small fluctuation. Furthermore, Kingman coalescent limits on the binary setting, for each pair to coalescent in each generation. In this setting, violations exist in real life application, to tackle this, multiple merger of one coalescent with two or more lineages are introduced. See figure 1.2 for motivation.

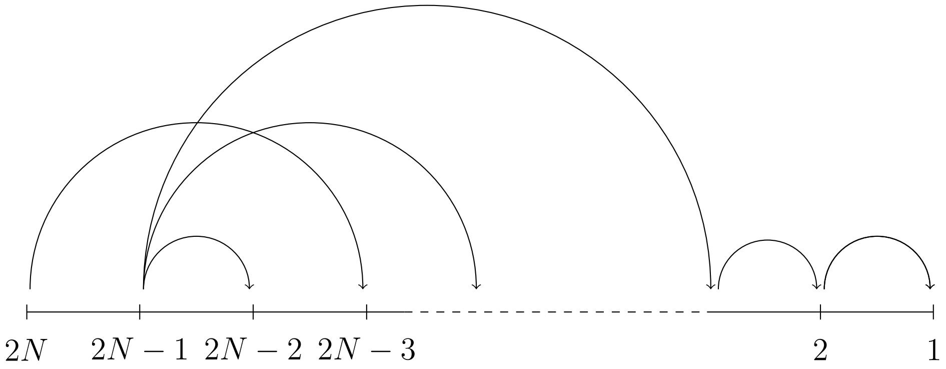

Figure 1.2: A pure death process for multiple merger coalescent.

Chapter 2

Preliminaries

This chapter is devoted to some mathematical facts, definitions and results in probability, information, and coalescence theory.

Appendix A

Supplement Notes and Proofs

A.1 Sets and Urn Models

Bibliography

[1] Blum, M, G, B., and Rosenberg, N, A. (2007). Estimating the Number of Ancestral Lineages Using a Maximum-Likelihood Method Based on Rejection Sampling. *Genetics* **176** 3.

[2] Bhaskar, A., and Song, Y. S. (2014). Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Annals of Statistics* **42** 2469-2493.

[3] Chen, H. (2012). The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theoretical Population Biology* **81** 179-195.

[4] Cover, T, M., and Thomas, J, A. (1991) *Elements of Information Theory.* first edition, Wiley New York.

[5] Durrett, R. (2008). *Probability Models for DNA Sequence Evolution.* second edition, Springer New York.

[6] Fisher, R, A. (1930). *The Genetical Theory of Natural Selection.* second edition, Clarendon Press, Oxford, UK.

[7] Griffiths, R. C., amd Tavare, S. (1998). The age of mutation in the general coalescent tree. *Stochastic Models* **14** 273-295.

[8] Haldane, J, B, S. (1932). *The causes of evolution.* Longmans, Green and Co., London.

[9] Kim, J., Mossel, E., Rácz, M. Z. and Ross, N. (2015). Can one hear the shape of a population history? *Theoretical Population Biology* **100** 26-38.

[10] Kimura, M., and Ohta, T. (1973). The age of a neutral mutant persisting in a finete population. *Genetics* **75** 199-212.

[11] Kingman, J, F, C. (1982). The coalescent. *Stochastic Processes and their Applications* **13** 235-248.

[12] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Annuals of Statistics* **1** 38-53.

[13] Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475** 493-496.

[14] Moran, P, A, P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* **1** 60-71.

[15] Myers, S., Fefferman, C., and Patterson, N. (2008) Can one learn history from the allelic spectrum? *Theoretical Population Biology* **73(3)** 342-348.

[16] Polanski, A., Bobrowski, A., and Kimmel, M. (2003). A note on distributions of times to coalescence, under time-dependent population size. *Theoretical Population Biology* **63** 33-40.

[17] Polanski, A., and Kimmel, M. (2003). New Explicit Expressions for Relative Frequencies of Single-Nucleotide Polyorphisms with application to statistical inference on population growth. *Genetics* **165** 427-436.

[18] Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal* **27** 379–423, 623–656.

[19] Sheehan, S., harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genome Research* **15** 1576-1583.

[20] Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology* **26(2)** 119-164.

[21] Terhorst, J., and Song, Y. S. (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences of the United States of America* **112**(25) 7677-7682.

[22] Wright, S. (1931). Ecolution in Mendelian populations. *Genetics* **16** 0097-0159.

[23] Yu, B. (1997). Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, & G. L. Yang (Ed.), *Festschrigt for Lucien Le Cam* (pp. 423-435). Springer New York.