Notes

David Helekal

September 4, 2020

Contents

Chapter 1

Introduction

1.1 Motivation

In epidemiology, it is often desired to be able to reconstruct the history of a pathogen population and it's structure. The problem of reconstructing the hisotry of a pathogen population can be tackles using phylodynamics. Phylodynamics utilises genomic data to assemble phylogenies, which are then used to infer the population size history. This is possible by viewing a phylogeny as a realisation of a coalescent process, wit appropriately rescaled time. This claim can be justified by viewing the coalescent as a Moran model, run backwards in time with the time rate equal to the population size [1].
Within this report we will first introduce the coalescent process for phylodynamic inference, review it's inhomogenous generalisation, and finally introduce the main result of this work, a new model capable of doing local phylodynamic inference, i.e. on a subset of the whole population capable of detecting and modelling clonal expansions.
Clonal expansions are a process in which a particular subsest of a given bacterial strain undergoes explosive population growth that can be traced back to a particular individual [2]. The presence of clonal expansions in bacterial populations have been of long-standing interest and is implicated in epidemic processes, were an outbreak can be traced to a single ancestor [2, 3, 4, 5]. This often happens when a particular strain or individual obtains a variant of a particular gene that confers evolutionary advantange, for example, antibiotic resistance [6, 7, 5].
The presence of clonal expansions leaves an imprint in the overall population structure of a given bacterial strain, the particular topology associated with this often being referred to as star-like [2, 3]. The problem of detecting hidden population structure corresponding to clonal expansions has become a problem of interest in epidemiology and outbreak surveillance [8].
While methods to detect inhomogeneities in the population structure and size have been of interest since the early days of genetic sequencing [2, 3], the in-

terest in the problem increased with whole genome sequencing becoming more accessible and affordable [6, 9, 10].
Despite the problems of inferring population size from a genealogy and detecting heterogeneities in the population size of the entire population being intrinsically tied, all but one method [8], to our knowledge, rely either on manual detection or indirect detection. We aim to propose a simulation for the formation of clonal expansions in genealogy using the structured coalescent process, and devise a fully bayesian method for joint estimation and detection of relative population size and clonal expansions.

1.2 Existing Work

1.2.1 Phylodynamic Methods

Kingman's Coalescent is a continuous time markov chain stochastic process, defined on the statespace $1, 2, 3, ..., K$ which can be interpreted as a set of a particles, where each pair of particles independently coalesces at a constant rate. This gives transition rates:

$$g(j, j-1) = \binom{j}{2}\lambda \quad \lambda \in \mathbb{R}^+ \tag{1.1}$$

[11]
By taking a backwards in time approximation of the Wright-Fisher model, the coalescent process can be modified to model evolution of genealogies, i.e. how do the ancestors of a set of individuals relate to each other backwards in time. Denote the relative population size at time t by $\alpha(t)$. Under such modification the transition rates become:

$$g(j, j-1) = \binom{j}{2} \cdot \frac{1}{\alpha(t)} \tag{1.2}$$

[1]
One way to interpret this is as a rescaling of time inversely proportional to the population size under Wright-Fisher model [12].
As the transition rates depend on the relative population size, it is possible to utilise this model for the inverse problem of determining the history of the size of a population, based on genealogies reconstructed from genomic samples. Such methods are often referred to as SkyGrid methods, have been first introduced in [13] and [14].
The framework has then been extended to allow for piecewise continuous population size functions, referred to as SkyGrid [15] and to include covariates [16].

1.2.2 Local Phylodynamics

Chapter 2

Methods

2.1 Coalescent Preliminaries

This process is a time-inhomogenous pure-death markov process, as such we conduct and re-derive the following propoerties.
The waiting times can be derived as follows. For an inhomogenous CTMC, let $E_j(t)$ be the total exit rate from state j at time t. By the markov property individual exit events from a given state only depend on the state and given time, i.e. they form a time-inhomogenous poisson process. As such the probability of no events in an interval $[t, t+s] \quad s \in \mathbb{R}^+$ is

$$\exp\left(-\int_t^{t+s} E_j(\tau)d\tau\right) = \exp\left(-\int_0^s E_j(t+\tau)d\tau\right) \tag{2.1}$$

The waiting times are defined as

$$W_j(t) = \inf\{s : X(t+s) \neq j \mid X(t) = j\} \tag{2.2}$$

As such

$$W_j(t) > s \Rightarrow \forall \tau \in [t, t+s] \quad X(\tau) = j \tag{2.3}$$

Furthermore the above relation holds iff no exit event have occured in the time interval $[t, t+s]$. As such:

$$P[W_j(t) > s] = P[\text{no exit events in } [t, t+s]] = \exp\left(-\int_0^s E_j(t+\tau)d\tau\right)$$

$$P[W_j(t) < s] = 1 - \exp\left(-\int_0^s E_j(t+\tau)d\tau\right)$$

In the case of phylodynamic coalescent this becomes

$$P[W_j(t) \leq s] = 1 - \exp\left(-\int_0^s \frac{\binom{j}{2}}{\alpha(t+\tau)}d\tau\right) \tag{2.4}$$

2.2 Inhomogenous Coalescent

2.2.1 Exponential Growth

2.3 Coalescent with Local Population Structure

Chapter 3

Results

3.1 Implementation Notes

3.2 Exponential Growth

3.2.1 Phylogeny Simulation

3.2.2 MCMC inference

3.3 Coalescent with Local Population Structure

3.3.1 Phylogeny Simulation

3.3.2 MCMC inference

Chapter 4

Discussion

Chapter 5

Bibliography

Bibliography

[1] R. C. Griffiths et al. "Sampling theory for neutral alleles in a varying environment". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 344.1310 (June 29, 1994). Publisher: Royal Society, pp. 403–410. DOI: 10.1098/rstb.1994.0079. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.1994.0079 (visited on 08/28/2020).

[2] J. M. Smith et al. "How clonal are bacteria?" In: *Proceedings of the National Academy of Sciences* 90.10 (May 15, 1993), pp. 4384–4388. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.90.10.4384. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.90.10.4384 (visited on 07/29/2020).

[3] Brian G. Spratt et al. "Displaying the relatedness among isolates of bacterial species – the eBURST approach". In: *FEMS Microbiology Letters* 241.2 (2004). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.femsle.2004.11.015, pp. 129–134. ISSN: 1574-6968. DOI: 10.1016/j.femsle.2004.11.015. URL: https://onlinelibrary.wiley.com/doi/abs/10.1016/j.femsle.2004.11.015 (visited on 07/29/2020).

[4] Christophe Fraser, William P. Hanage, and Brian G. Spratt. "Neutral microepidemic evolution of bacterial pathogens". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.6 (Feb. 8, 2005), pp. 1968–1973. ISSN: 0027-8424. DOI: 10.1073/pnas.0406993102. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC548543/ (visited on 07/29/2020).

[5] Alice Ledda et al. "Re-emergence of methicillin susceptibility in a resistant lineage of Staphylococcus aureus". In: *Journal of Antimicrobial Chemotherapy* 72.5 (May 1, 2017). Publisher: Oxford Academic, pp. 1285–1288. ISSN: 0305-7453. DOI: 10.1093/jac/dkw570. URL: https://academic.oup.com/jac/article/72/5/1285/2930201 (visited on 07/29/2020).

[6] Matthew T. G. Holden et al. "A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic". In: *Genome Research* 23.4 (Apr. 2013), pp. 653–664. ISSN: 1549-5469. DOI: 10.1101/gr.147710.112.

[7] Li-Yang Hsu et al. "Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system". In: *Genome Biology* 16.1

(Apr. 23, 2015), p. 81. ISSN: 1465-6906. DOI: 10.1186/s13059-015-0643-z. URL: https://doi.org/10.1186/s13059-015-0643-z (visited on 07/29/2020).

[8] Erik M. Volz et al. "Identification of Hidden Population Structure in Time-Scaled Phylogenies". In: *Systematic Biology* (). DOI: 10.1093/sysbio/syaa009. URL: https://academic.oup.com/sysbio/advance-article/doi/10.1093/sysbio/syaa009/5734655 (visited on 07/01/2020).

[9] Bethany L. Dearlove and Simon D. W. Frost. "Measuring Asymmetry in Time-Stamped Phylogenies". In: *PLOS Computational Biology* 11.7 (July 6, 2015). Publisher: Public Library of Science, e1004312. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004312. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004312 (visited on 07/29/2020).

[10] Vegard Eldholm et al. "Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain". In: *Nature Communications* 6.1 (May 11, 2015). Number: 1 Publisher: Nature Publishing Group, p. 7119. ISSN: 2041-1723. DOI: 10.1038/ncomms8119. URL: https://www.nature.com/articles/ncomms8119 (visited on 07/29/2020).

[11] J. F. C. Kingman. "The coalescent". In: *Stochastic Processes and their Applications* 13.3 (Sept. 1, 1982), pp. 235–248. ISSN: 0304-4149. DOI: 10.1016/0304-4149(82)90011-4. URL: http://www.sciencedirect.com/science/article/pii/0304414982900114 (visited on 07/30/2020).

[12] Jotun. Hein, Mikkel H. Schierup, and Carsten. Wiuf. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory:* Oup Oxford, Dec. 9, 2004. ISBN: 978-0-19-154615-0. URL: https://www.dawsonera.com:443/abstract/9780191546150.

[13] O G Pybus, A Rambaut, and P H Harvey. "An integrated framework for the inference of viral population history from reconstructed genealogies." In: *Genetics* 155.3 (July 2000), pp. 1429–1437. ISSN: 0016-6731. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461136/ (visited on 08/28/2020).

[14] Alexei J. Drummond et al. "Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data". In: *Genetics* 161.3 (July 1, 2002). Publisher: Genetics Section: INVESTIGATIONS, pp. 1307–1320. ISSN: 0016-6731, 1943-2631. URL: https://www.genetics.org/content/161/3/1307 (visited on 07/02/2020).

[15] Mandev S. Gill et al. "Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci". In: *Molecular Biology and Evolution* 30.3 (Mar. 1, 2013). Publisher: Oxford Academic, pp. 713–724. ISSN: 0737-4038. DOI: 10.1093/molbev/mss265. URL: https://academic.oup.com/mbe/article/30/3/713/1041171 (visited on 06/30/2020).

[16] Mandev S. Gill et al. "Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates". In: *Systematic Biology* 65.6 (Nov. 1, 2016). Publisher: Oxford Academic, pp. 1041–1056. ISSN: 1063-5157. DOI: 10.1093/sysbio/syw050. URL: https://academic.oup.com/sysbio/article/65/6/1041/2281638 (visited on 06/30/2020).