Notes

David Helekal

July 31, 2020

Contents

Chapter 1

Motivation

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 [1]. 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 [1, 2, 3, 4]. This often happens when a particular strain or individual obtains a variant of a particular gene that confers evolutionary advantange, for example, antibiotic resistance [5, 6, 4].

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 [1, 2]. The problem of detecting hidden population structure corresponding to clonal expansions has become a problem of interest in epidemiology and outbreak surveillance [7].

While methods to detect inhomogeneities in the population structure and size have been of interest since the early days of genetic sequencing [1, 2], the interest in the problem increased with whole genome sequencing becoming more accessible and affordable [5, 8, 9].

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 [7], 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.

Chapter 2

Questions

.

Chapter 3

Methods

3.1 Coalescent Preliminaries

We shall begin with an overview of the standard Kingman's Coalescent process[10]. This process is often used to characterise evolutionary histories of populations [11, 12, 13]. The coalescent is a CTMC defined on the set $\{1...n\}$, parametrised via the coalescent rate, in our case $1/Neg(t)$, where g is a scale parameter and $Neg(t)$ the effective population size at time t [11, 12]. Define $\alpha = Neg$.

The transition rates of the coalescent process are given by

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

The waiting times in the homogenous case are exponentially distributed

$$P[W_j \leq s] = 1 - \exp\left(-s\frac{\binom{j}{2}}{\alpha(t)}\right)$$

Furthermore, the waiting times for individual coalescent events, conditioned on being less than the time between two consecutive sampling events Δt are distributed as follows

$$P[W_j \leq s \mid W_j \leq \Delta t] = \frac{P[W_j \leq s]}{P[W_j \leq \Delta t]} \quad \forall s \leq \Delta t \tag{3.1}$$

In the case of time-inhomogenous effective population size, 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

$$\exp\left(-\int_t^{t+s} E_j(\tau)d\tau\right) = \exp\left(-\int_0^s E_j(t+\tau)d\tau\right) \tag{3.2}$$

The waiting times are defined as

$$W_j(t) = \inf\{s : X(t+s) \neq j \mid X(t) = j\} \tag{3.3}$$

As such

$$W_j(t) > s \Rightarrow \forall \tau \in [t, t+s] \quad X(\tau) = j \tag{3.4}$$

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

$$P[W_j(t) \leq s] = 1 - \exp\left(-\int_0^s \frac{\binom{j}{2}}{\alpha(t+\tau)}d\tau\right) \tag{3.5}$$

Note, the waiting times are still memoryless:

$$P\left[W_j(t) > s+u \mid W_j(t) > s\right] = P\left[W_j(t) > s+u \mid X(s) = j\right] \tag{3.6}$$

By markov property

$$P\left[W_j(t) > s+u \mid X(s) = j\right] = P\left[W_j(t+s) > u\right] \tag{3.7}$$

3.1.1 Homogenous case

The simulation of the coalescent process conditioned on sampling times follows a modified gillespie scheme. In order to facilitate the computation of the likelihoods of the individual simulated trees, it is preferred to avoid rejection sampling. As such we require sampling the conditional likelihood 3.1. This is achieved by inverse transform sampling. Let:

$$u \sim U([0,1])$$

$$T(u) : P[T(u) \leq s] = \frac{P[T(u) \leq s]}{P[T(u) \leq \Delta t]} \quad \forall s \leq \Delta t \tag{3.8}$$

Where $T(u)$ is assumed to be monotone increasing and invertible.

$$P[T(u) \leq s] = P[u \leq T^{-1}(s)]$$

$$\Rightarrow \qquad P[u \leq T^{-1}(s)] = \frac{\int_0^s \lambda \exp(-\lambda t) dt}{\int_0^{\Delta t} \lambda \exp(-\lambda t) dt}$$

$$\Rightarrow \qquad T^{-1}(s) = \frac{1 - \exp(-\lambda s)}{1 - \exp(-\lambda \Delta t)}$$

Defining $y \triangleq T^{-1}(s)$, we obtain the transform:

$$T(y) = \frac{-1}{\lambda} \log[1 - y(1 - \exp(-\lambda \Delta t))] \qquad (3.9)$$

The corresponding pdf evaluated at u is

$$f_{\mathbf{T}(u)}(T(u)) = \lambda \left(\frac{1}{1 - \exp(\lambda \Delta t)} - u \right) \qquad (3.10)$$
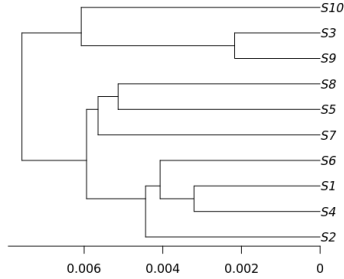
6

Figure 3.1: An example simulated coalescent tree

```
f <- (sampling_times, Ne): //Sampling times in descending order
  extant_lineages <- 1
  future_lineages <- length(sampling_times)-1
  t <- sampling_times[1]
  idx <- 1
  log_lh <- 0

  while extant_lineages > 1 or future_lineages > 0:
    if extant_lineages < 2:
      idx++
      t <- sampling_times[idx]
      extant_lineages++
      future_lineages--
    else:
      delta_t <- t-sampling_times[idx+1]
      rate <- binom(extant_lineages,2)/Ne

      p_coal <- 1-exp(-rate*delta_t)
      r_c ~ U[0,1]

      if r_c < p_coal:

        log_lh += log(p_coal)

        coalesce_lineages
        extant_lineages--

        r_w ~ U[0,1]
        w_t <- (-1/rate)*log(1-r_w*(1-exp(-rate*delta_t)))
        t <- t+w_t
```

```
cond_lh <- rate*(1/(1-exp(-rate*delta_t)) - r_w)
log_lh += log(cond_lh)

else:
    log_lh += log(1-p_coal)
    idx++
    t <- sampling_times[idx]
    extant_lineages++
    future_lineages--

return: coalescent_times, log_lh
```

3.1.2   Inhomogenous Case

In the inhomogenous case, the scheme is similar, with the key difference that the sampling times now follow a much more complex distribution. We proceed with a modified conditional sampling scheme as in 3.8. In order to obtain draws $w_j(t)$, several approaches exist. While draws from standard exponential w_j are rescaled, akin to algorithm described in [12] Pg 98., we instead proceed by deriving the appropriate distributions and simulation schemes thereof.

Waiting times distribution

Consider the time interval $[t_i, s_i]$ with $s_j = min\,\{s \in S : s > t_i\}$. Define $\Delta t_i \triangleq s_i - t_i$. The probability that no coalescent events happens within this interval is

$$P[W_j(t_i) > \Delta t_i] = \exp\left(-\int_0^{\Delta t_i} \frac{\binom{j}{2}}{\alpha(t+\tau)}d\tau\right)$$

analogously, the probability of waiting times conditioned on that the waiting time is less than Δt_i is:

$$P[W_j(t_i) < s \mid W_j(t_i) < \Delta t_i] = \frac{1 - \exp\left(-\int_0^s \frac{\binom{j}{2}}{\alpha(t+\tau)}d\tau\right)}{1 - P[W_j(t_i) > \Delta t_i]} \tag{3.11}$$

Simulating

In order to simulate $W_j(t_i)$ we proceed with an inverse transform sampling scheme, derived from the base samples W_j. First, assume W_j are distributed according to

$$P[W_j < s \mid W_j < u] = \frac{1 - \exp\left(-\frac{\binom{j}{2}}{\alpha}\right)}{1 - P[W_j > u]} \tag{3.12}$$

8

Where u is chosen such that

$$P[W_j > u] = P[W_j(t_i) > \Delta t_i] \tag{3.13}$$

Then the function

$$F(W_j; t_i): \quad P[F(W_j; t_i) < s \mid W_j < u] = P[W_j(t_i) < s \mid W_j(t_i) < \Delta t_i] \tag{3.14}$$

Is given by the inverse with respect to s of

$$G(s; t_i) = \int_0^s \frac{\alpha}{\alpha(t_i + \tau)} d\tau \tag{3.15}$$

Which exists for any biologically sensible choice of $\alpha(t)$

Likelihood

Let $\{t_i\}_{i \in S \subset \mathbb{N}}$ denote the times of events in increasing order. Using notational convention from [11], let $t_Y \triangleq \{t_i\}_{i \in Y}$ denote times of coalescent events and $t_I \triangleq \{t_i\}_{i \in I}$ denote times of sampling events, where Y, I are disjoint partitions of the index set S with the property that $S = Y \cup I$. The likelihood of a particular genealogy is then given by:

$$\mathcal{L}(g \mid \alpha) = \prod_{i \in S \backslash 1} \left(\mathbb{1}_Y(i) \frac{\binom{k_i}{2}}{\alpha(t_i)} + \mathbb{1}_I(i) \right) \exp\left(-\int_{t_{i-1}}^{t_i} \frac{\binom{k_i}{2}}{\alpha(\tau)} d\tau \right) \tag{3.16}$$

The log-likelihood is:

$$\log \mathcal{L}(g \mid \alpha) = -\sum_{i \in S \backslash 1} \int_{t_{i-1}}^{t_i} \frac{\binom{k_i}{2}}{\alpha(\tau) d\tau} + \sum_{i \in Y} \log \frac{\binom{k_i}{2}}{\alpha(t_i)} \tag{3.17}$$

***Skygrid* and other families of population functions**

Following subsection NOT CURRENTLY RELEVANT for project

Skygrid is an approach introduced in [14]. It considers a family of time-dependent effective population size functions specified as follows. Let $T \subset \mathbb{R}$ be the interval under consideration in coalescent time. Consider an arbitrary population size function F:

$$F: \quad T \subset \mathbb{R} \to \mathbb{R}^+ \tag{3.18}$$

Given a mutually disjoint family of time intervals $\{D_i\}_{i \in N} \subseteq T$, $D_j = [d_{j-1}, d_j)$, with $\bigcup_{i \in N} D_i = T$, the *skygrid* family of functions is then given by

$$F_{skygrid} \triangleq \left\{ \alpha \in F \mid \forall i \in N, \quad \forall t \in D_i, \quad \alpha(t) = c_i, \quad c_i \in \mathbb{R}^+ \right\} \tag{3.19}$$

This can be easily extended to any function G

$$G(t) \triangleq \sum_{i \in N} \mathbb{1}_{D_i}(t)g_i(t) \tag{3.20}$$

Where $g_i(t)$ are arbitrary positive integrable functions. Such formulation makes computation of likelihood straightforward.

$$\begin{aligned}
\log \mathcal{L}\left(g \mid G(t)\right) = &- \sum_{i \in S\backslash 1} \binom{k_i}{2} \sum_{j \in N} \mathbb{1}_{D_j}(t_{i-1}) \int_{t_{i-1}}^{\min\{d_j, t_i\}} g_j^{-1}(\tau)d\tau \\
&+ \sum_{i \in Y} \log \binom{k_i}{2} - \sum_{i \in Y} \log \sum_{j \in N} \mathbb{1}_{D_j}(t_i)g_j(t_i)
\end{aligned} \tag{3.21}$$

3.2 Inhomogenous Lineages and Outbreaks

Suppose we sample genomes of a set of pathogenic bacterial samples. At an unknow point in time a particular strain acquires a mutation conferring it resistance to a widely used antibiotic. This increases the particular strains fitness and enables it to undergo a period of rapid growth leading to a clonal expansion. If we assume that this increase in fitness occurs in a short time span, the lineage of this strain will behave differently in the phylogenetic tree, having a coalescent rate corresponding to a rapidly expanding population, starting from a very small number of individuals CITE ME . This problem of identifying hidden population structure has been proposed in [7], where a testing based approach was used to indentify structure in a phylogeny.

We seek a modification to the coalescent model that will allow for change points located on the branches of the phylogeny, marking the event when a particular lineage starts behaving according to a different population size function than its parent lineage.

In this model, coalescent nodes have an added colour property, and each colour coalesces according to a colour specific, time dependent case. Nodes of non-identical colour can coalesce iff at least one of them is the last remaining node of a given colour. Different colours correspond to different lineages, each behaving under its own growth function.

Similar models, refered to as Structured Coalescent, adding colouring to vertices of phylogenies have been used in epidemiology to track outbreaks [15], or transmission chains [16].

3.2.1 Model

A given genealogy $\mathbf{g} = (V_\mathbf{g}, E_\mathbf{g}, t_\mathbf{g})$ is an incomplete, empirical sample of the underlying process CITE ME .

It consists of nodes $V_{\mathbf{g}}$, directed edges $E_{\mathbf{g}}$, and node labels $t_{\mathbf{g}}$ corresponding to event times.

As in section 3.1.2 $\mathbf{g}$ shall be indexed by an index set $S = 1 \leq i \leq N \subset \mathbb{N}$, with $Y \subset S$ corresponding to coalescent events and $I \subset S$ corresponding to sampling events.

For convenience, assume that all edges are in the forwards time direction, i.e.:

$$\forall k, l \in S : (k, l) \in E_{\mathbf{g}} \Rightarrow t_k < t_l$$

Furthermore, all event times are ordered in descending (backwards) time order, i.e.:

$$\forall k, l \in S : k < l \Rightarrow t_k > t_l$$

Under the assumption that $\mathbf{g}$ is a genealogy of a given sample, with each edge in $E_{\mathbf{g}}$ there is an associated unobserved set of individuals descending from one another CITE ME . At some point along an edge a node of one lineage will undergo colour change and change its type to that of another lineage.

We seek a model which enable us to investigate this proposed behaviour.

Definition 3.2.1 (Multiple Lineage Coalescent)**.** Given M colours, M population size functions $\alpha \triangleq \{\alpha_j(t)\}_{1 \leq j \leq M}$. Let $Y(t)$ be a CTMC with the state space:

$$\Sigma = \{\mathbf{s} \in \mathbb{Z}_+ : |\mathbf{s}| \geq 1\} \tag{3.22}$$

and the transition rates

$$\mathbf{s} \to \mathbf{s} - \mathbf{e_j} \qquad \binom{s_j}{2} \alpha_j^{-1}(t) \qquad 1 \leq j \leq M \tag{3.23}$$

$$\mathbf{s} \to \mathbf{s} - \mathbf{e_j} + \mathbf{e_k} \qquad \delta_{1,j} s_k \qquad 1 \leq j, k \leq M \tag{3.24}$$

The interpretation of this model in backwards (coalescent) time is that each node corresponds to a single specific lineage (colour). Nodes of the same lineage coalesce at i.i.d rates, according to a lineage specific growth functions, until reaching the MRCA of given lineage. The MRCA then changes type (colour) to that of a different lineage.

Under the hypothesis, along each of the edges from the parent of a lineage MRCA to the MRCA lies a point in time that characterises the time of divergence of the lineage, after which the lineage starts to undergo clonal expansion.

3.2.2 Inference

Assume that we have been given a genealogy as described in previous section. The number of individua lineages M, the effective population size functions

α, the placement and divergence events in descending order $\Xi = \{\xi_i\}_{1 \leq i \leq M}$, are unknown. In effect we seek to identify terms in the factorisation of the posterior:

$$\mathcal{L}(\alpha, \Xi, M \mid \mathbf{g}) \propto \mathcal{L}(\mathbf{g} \mid \Xi, M, \alpha)\mathcal{L}(\alpha \mid \Xi, M)\mathcal{L}(\Xi, M) \qquad (3.25)$$

Definition 3.2.2 (Divergence Event). A divergence event ξ associated with a lineage is a tuple:

$$\xi = (e', \tau) \qquad (3.26)$$

where τ denotes the time of the divergence event, and $e' = (v^{pa}, v^{MRCA})$ denotes the edge along which the divergence event happens, with v^{MRCA} being the MRCA of the newly diverging lineage.

In order to identify the likelihood $\mathcal{L}(\mathbf{g} \mid \Xi, M, \alpha)$, the genealogy needs to be first partitinioned into separate subtrees, each corresponding to an individual lineage.

Definition 3.2.3 (Descendant Set).

$$\mathcal{D}(i) \triangleq \{j \in S \mid j \text{ is a descendant of } i\} \qquad (3.27)$$

Definition 3.2.4 (Lineage Subtree). A subtree $W(\xi_i)$ associated with a divergence event ξ_i consists of all vertices, edges and times associated with a given lineage.

$$W(\xi_i) \triangleq \mathcal{D}(v_i^{pa}) \setminus \bigcup_{j<i}(\mathcal{D}(v_j^{pa})) \qquad (3.28)$$

As members of a lineage can only coalesce with one another, the coalescence rate is independent of other lineages, and the time of divergence has been given, the total likelihood is equal to the product of individual lineage specific likelihoods. A final step before deriving lineage specific likelihoods is required, that is, the divergence times have to be included along with the times contained within the lineage subtree.

Denote the set of all divergence times associated with a Lineage Subtree $W(\xi_j)$ $\tau_{W(\xi_j)}$

$$\tau_{W(\xi_j)} \triangleq \tau_i \in \xi_i : v_i^{pa} \in W(\xi_j), \quad \forall 1 \leq i \leq M$$

Finally define the set of all times associated with a particular lineage $T(\xi_j) = \tau_{W(\xi_j)} \cup t_{W(\xi_j)}$, indexed by set $S_{T(\xi_j)} = \{1...|T(\xi_j)|\}$ such that

$$\forall t_k, t_l \in T(\xi_j), \quad k > l \Rightarrow t_k \geq t_l$$

Let k_{i,ξ_j} denote the number of extant individuals corresponding to subtree lineage ξ_j at during the time interval $[t_{i-1}, t_i]$. The total likelihood is equal to:

$$\mathcal{L}(\mathbf{g} \mid \Xi, M, \alpha) = \prod_{j=1}^{M} \mathcal{L}(T(\xi_j) \mid \alpha_j)$$

$$\mathcal{L}(T(\xi_j) \mid \alpha_j) = \prod_{t_i \in T(\xi_j)} \left(\mathbb{1}_{t_Y}(t_i) \frac{\binom{k_{i,\xi_j}}{2}}{\alpha(t_i)} + \mathbb{1}_{t_I \cup \tau_{W(\xi_j)}}(t_i) \right) \exp\left(-\int_{t_{i-1}}^{t_i} \frac{\binom{k_{i,j}}{2}}{\alpha(\tau)} d\tau \right)$$

(3.29)

3.2.3   Choice of Population Size Functions

SIS Model

The dynamics of an immunity non-confering pathogen with the host population constant can be described by the SIS model[17]:

$$\frac{dI}{dt'} = (N - I)\beta I/N - \gamma I \tag{3.30}$$

By defining $A = \beta - \gamma$ and $B = \beta/N$, we obtain the solution This equation admits a solution of the form:

$$I(t') = \frac{A}{B + (A/I_0 - B)\exp(-t'A)} \tag{3.31}$$

We assume $I_0 = 1$. This can be rearranged and contrasted with standard logistic growth with carrying capacity equal to A/B and growth rate equal to A:

$$I(t') = \frac{A/B}{1 + (A/B - 1)\exp(-t'A)} \tag{3.32}$$

As our hypothesis posutlates that a divergence event is a new lineage being born and undergoing rapid growth, we require $\alpha(0) = 0$. Hence we take $\alpha(t') = I(t') - 1$.

Finally in order to convert into population decline in coalescent time, perform the change of variables $-t' + T_{max} - T_{div} = \tau$, while also defining $K = A/B$:

$$\begin{aligned} \alpha(\tau) &= \frac{K}{1 + (K-1)\exp((\tau - T_{max} + T_{div})A)} - 1 \\ &= \frac{(K-1) - (K-1)\exp((\tau - T_{max} + T_{div})A)}{1 + (K-1)\exp((\tau - T_{max} + T_{div})A)} \end{aligned} \tag{3.33}$$

And

$$\alpha^{-1}(\tau) = \frac{1 + (K-1)\exp((\tau - T_{max} + T_{div})A)}{(K-1) - (K-1)\exp((\tau - T_{max} + T_{div})A)}$$

$$\Rightarrow \frac{1}{(K-1)(1 - \exp((\tau - T_{max} + T_{div})A))}$$

$$+ \frac{1}{\exp(-(\tau - T_{max} + T_{div})A) - 1} \tag{3.34}$$

As such:

$$\int_t^{t+s} \alpha^{-1}(\tau)\,d\tau = \left[\frac{-1}{(K-1)A}\log(\exp(-(\tau - T_{max} + T_{div})A) - 1)\right]_t^{t+s}$$

$$+ \left[\frac{1}{A}\log(1 - \exp((\tau - T_{max} + T_{div})A))\right]_t^{t+s}$$

$$\Rightarrow \frac{1}{A}\left[\log\left(\frac{1 - \exp((\tau - T_{max} + T_{div})A)}{(\exp(-(\tau - T_{max} + T_{div})A) - 1)^{\frac{1}{K-1}}}\right)\right]_t^{t+s}$$

$$\tag{3.35}$$

Define $Q = \frac{K-2}{(K-1)A}$, $T = T_{max} + T_{div}$, expanding the integral:

$$Q\left[\log\left(\frac{1 - \exp((t + s - T)A)}{\exp(-(t + s - T)A) - 1}\right) - \log\left(\frac{1 - \exp((t - T)A)}{\exp(-(t - T)A) - 1}\right)\right]$$

$$= Q\left[\log\left(\frac{(1 - \exp((t + s - T)A))(\exp(-(t - T)A) - 1)}{(\exp(-(t + s - T)A) - 1)(1 - \exp((t - T)A))}\right)\right]$$

$$= Q\left[\log\left(\frac{\exp(-(t - T)A) - \exp(sA) + \exp((t + s - T)A)) - 1}{-\exp(-sA) + \exp((t - T)A) + \exp(-(t + s - T)A) - 1}\right)\right]$$

$$= Q\left[\log\left(\frac{\exp(sA)(\exp((t - T)A) - 1) + \exp(-(t - T)A) - 1}{\exp(-sA)(\exp(-(t - T)A) - 1) + \exp((t - T)A) - 1}\right)\right]$$

$$= Q\left[\log\left(\frac{\exp(sA)(\exp((t - T)A) - 1) - \exp(-(t - T)A)(\exp((t - T)A) - 1)}{\exp(-sA)(\exp(-(t - T)A) - 1) - \exp((t - T)A)(\exp(-(t - T)A) - 1)}\right)\right]$$

$$= Q\left[\log\left(\frac{(\exp((t - T)A) - 1)(\exp(sA) - \exp(-(t - T)A))}{(\exp(-(t - T)A) - 1)(\exp(-sA) - \exp((t - T)A))}\right)\right]$$

$$= Q\left[\log\left(\frac{(\exp((t - T)A) - 1)}{\exp(-(t - T)A) - 1}\right) + \log\left(\frac{\exp(sA) - \exp(-(t - T)A)}{\exp(-sA) - \exp((t - T)A)}\right)\right]$$

$$\tag{3.36}$$

14

Set $\rho = \frac{(\exp((t-T)A)-1)}{\exp(-(t-T)A)-1}$, $Y = \int_t^{t+s} \alpha^{-1}(\tau)\,d\tau$, $\phi = -\exp(Y/Q - \log(\rho))$, $\psi = \exp((t-T)A)) - \exp(-(t-T)A)$

$$
\begin{aligned}
Y/Q - \log(\rho) &= \log\left(\frac{\exp(sA) - \exp(-(t-T)A)}{\exp(-sA) - \exp((t-T)A)}\right) \\
\exp(Y/Q - \log(\rho)) &= \frac{\exp(sA) - \exp(-(t-T)A)}{\exp(-sA) - \exp((t-T)A)} \\
-\phi(\exp(-sA) - \exp((t-T)A)) &= \exp(sA) - \exp(-(t-T)A) \\
\exp(sA) + \psi + \phi(\exp(-sA)) &= 0 \\
\exp(2sA) + \psi\exp(sA) + \phi &= 0
\end{aligned}
\tag{3.37}
$$

By defining $x = \exp(sA)$, this becomes a quadratic equation. yielding the solution:

$$
\begin{aligned}
x_{1,2} &= \frac{-\psi \pm \sqrt{\psi^2 - 4\phi}}{2} \\
\Rightarrow \quad s &= \frac{1}{A}\log(x)
\end{aligned}
\tag{3.38}
$$

Gompertz Curve

3.2.4 Priors

As a prior for change points $\mathcal{L}(\Xi, M)$ a Poisson Point Process is used. Denote by B the total branch length of the genealogy. Then

$$
M \sim \frac{(\lambda B)^M}{M!}\exp(-\lambda B)
\tag{3.39}
$$

MCMC

Due to the variable nature of the dimensionality of the parameter space, it will be necessary to use Reversible Jump MCMC (rjMCMC) [18, 19].

Chapter 4

Results

4.1 Simulations

4.1.1 Exponetial Growth

An example of a population under exponential growth, $\alpha(t) = N * \exp(-\lambda t)$ consisting of 100 sampling events between 0 and 10 years before present has been simulated with a rate parameter λ and final size parameter N drawn from a uniform densities on intervals $[0.1, 10]$, and $[1, 100]$ respectively.

A Metropolis-Hastings MCMC scheme was then used to infer the parameters λ and N. A zero-centred laplace distribution with rate equal to one was chosen as the prior for λ, whereas for N the exponential distribution with rate one was used.

One million iterations were used and the first One hundred thousand discarded as burn in time. To further validate the fit, both a maximum likelihood (MLE) and maximum *a-posteriori* (MAP) estimates were computed and plotted against the posterior marginals inferred by the MCMC.
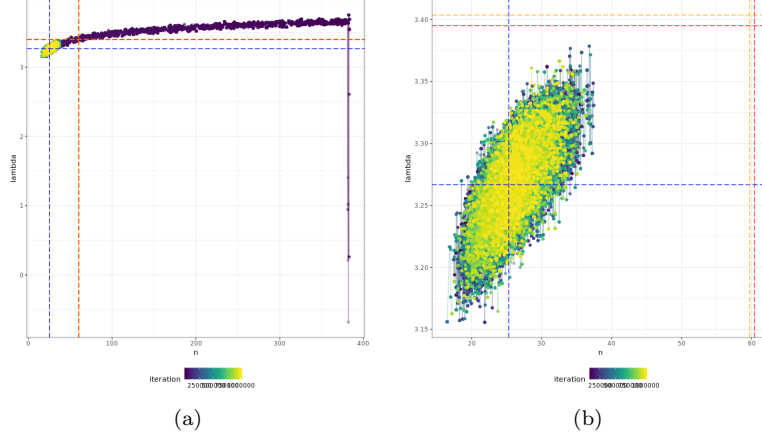
Figure 4.1: Trace plots for the markov chain. Red lines denote true parameter values. MLE marked by blue lines. MAP marked by orange lines. 4.1a Shows the entire trace of the chain. 4.1b Shows the trace with the first 100000 iterations discarded
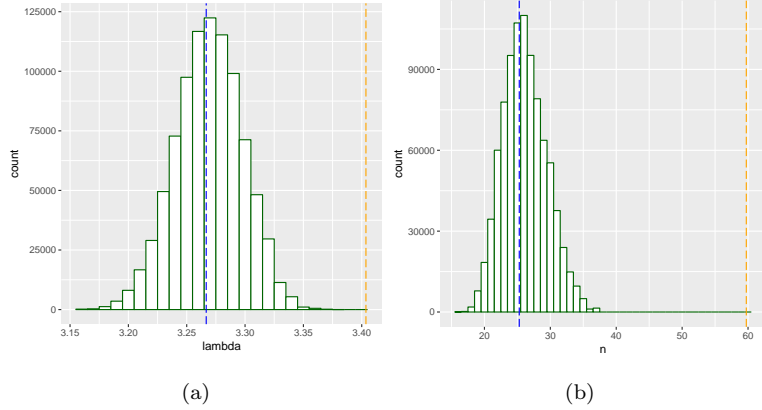

Figure 4.2: Histograms of the posterior marginals. MLE marked by blue line. MAP marked by orange line. 4.2a λ marginal 4.2b N marginal
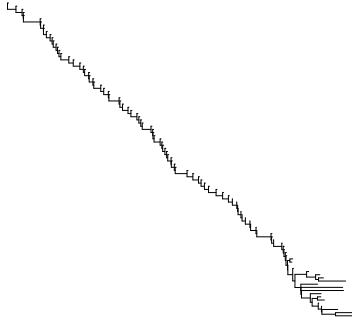
17

Figure 4.3: The simulated genealogy used for this example.

4.2 Previous Work

A framework utilising sampling intensity in order to extract more information
is proposed in [20]

4.3 Bibliography

Bibliography

[1] 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).

[2] 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).

[3] 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).

[4] 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).

[5] 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.

[6] 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).

[7] 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).

[8] 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).

[9] 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).

[10] 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).

[11] 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: IN-VESTIGATIONS, pp. 1307–1320. ISSN: 0016-6731, 1943-2631. URL: https://www.genetics.org/content/161/3/1307 (visited on 07/02/2020).

[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] Simon Y. W. Ho and Beth Shapiro. "Skyline-plot methods for estimating demographic history from nucleotide sequences". In: *Molecular Ecology Resources* 11.3 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-0998.2011.02988.x, pp. 423–434. ISSN: 1755-0998. DOI: 10.1111/j.1755-0998.2011.02988.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2011.02988.x (visited on 06/24/2020).

[14] 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).

[15] Nicola De Maio, Chieh-Hsi Wu, and Daniel J. Wilson. "SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent". In: *PLOS Computational Biology* 12.9 (Sept. 28, 2016). Publisher: Public Library of Science, e1005130. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005130. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005130 (visited on 07/22/2020).

[16] Xavier Didelot et al. "Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks". In: *Molecular Biology and Evolu-*

tion 34.4 (Apr. 1, 2017). Publisher: Oxford Academic, pp. 997–1007. ISSN: 0737-4038. DOI: 10.1093/molbev/msw275. URL: https://academic.oup.com/mbe/article/34/4/997/2919386 (visited on 07/15/2020).

[17]   Matt J. Keeling and Pejman Rohani. "Introduction to Simple Epidemic Models". In: *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008, pp. 15–53. ISBN: 978-0-691-11617-4. DOI: 10.2307/j.ctvcm4gk0.5. URL: https://www.jstor.org/stable/j.ctvcm4gk0.5 (visited on 07/22/2020).

[18]   Y. Fan and S. A. Sisson. "Reversible jump Markov chain Monte Carlo". In: *arXiv:1001.2055 [stat]* (Jan. 12, 2010). arXiv: 1001.2055. URL: http://arxiv.org/abs/1001.2055 (visited on 07/13/2020).

[19]   Peter J. Green. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4 (Dec. 1, 1995). Publisher: Oxford Academic, pp. 711–732. ISSN: 0006-3444. DOI: 10.1093/biomet/82.4.711. URL: https://academic.oup.com/biomet/article/82/4/711/252058 (visited on 07/14/2020).

[20]   Kris V. Parag, Louis du Plessis, and Oliver G. Pybus. "Jointly Inferring the Dynamics of Population Size and Sampling Intensity from Molecular Sequences". In: *Molecular Biology and Evolution* (). DOI: 10.1093/molbev/msaa016. URL: https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msaa016/5719057 (visited on 06/19/2020).