# Chapter 1

# Model

## 1.1 Coalescent with Local Population Structure

We now present novel local phylodynamic model. In order to help illustrate the concepts behind this model, consider the following scenario: Suppose there is a parent population of bacteria of interest evolving through time. At any point in time, a particular individual within this population gains a significant evolutionary advantage. This advantage enables the individual and its progeny to undergo a rapid clonal expansion as it gains the ability to colonise a particular niche, eventually reaching a constant population size equilibrium. We will associate the clade corresponding to the clonal expansion with a particular colouring. The clonal expansion process may repeat throughout time, giving rise to several clades. The clonal expansions may for example correspond to process such a bacterial strain gaining antibiotic resistance, or being newly introduced into a hospital.

In general, we will assume that the clonal expansion process happens relatively rarely. We will also assume the parent population is at eqillibrium. As under this scenario populations reach a constant population size in the limit this assumption can be interpreted that the clonal expansion that gave rise to the parent population has happened a very long time ago.

At present a researcher collects a random set of $N$ samples from the population of interest, containing at least some of the clades produced by the clonal expansion process described above. Each sample belongs to one of the $k$ clonal expansion clades with multinomial probability $\theta_i$, $\boldsymbol{\theta} = [\theta_1, ..., \theta_i, ..., \theta_{k-1}, \theta_k]$.

The researcher then sequences the genomes of the samples and infers the corresponding phylogenetic using any of the popular available methods. The resulting phylogeny can be viewed as a realisation of the following backwards time process.

Select $K$ clade colours, and an associated clade sampling membership probability vector $\boldsymbol{\theta}$. For each of the $N$ samples $s_i$ forming the leaves of the phylogeny, assign a colouring $c_i$. The samples of identical colour then coalesce with each other with rate governed by a colour specific population size function $\alpha_i(t)$. As it is assumed each colour is formed by a clonal expansion at time $t_i^{Div}$, $\alpha_i(t)$ vanishes to zero at the time of the expansion, and as such all coalescence within clade of colour $c_i$ happens almost surely before $t_i^{Div}$. Upon reaching the time of clonal expansion the MRCA then chnages colour to that of another clade extant at the time. The clade of the colour the MRCA changes to will be referred to as the parent clade. The MRCA is then added to the parent clade as a leaf. The process continues until all only one individual remains.

The coalescent process given leaf colouring and growth function parameters described above is characterised mathematically in the following section.

### 1.1.1 Model

## 1.2 Preliminaries

A given genealogy $\mathbf{g} = (V_\mathbf{g}, E_\mathbf{g}, t_\mathbf{g}, c_\mathbf{g})$ is an incomplete, empirical sample of the underlying process.
It consists of nodes $V_\mathbf{g}$, directed edges $E_\mathbf{g}$, node labels $t_\mathbf{g}$ corresponding to event times, and node labels $c_\mathbf{g}$ corresponding to node colouring.
The genealogy $\mathbf{g}$ 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, with the first event corresponding the the most recent sample

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

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. At some point along an edge from one lineage to another, the lineage can undergo a colour change, and become the most recent ancestor of a diverging clade. This event corresponds to this lineage somehow gaining advantage over other lineages, be it a bacterium gaining resistance against a drug, or a strain of a virus invading a completely susceptible population.

**Definition 1.2.1** (Multiple Lineage Coalescent)**.** Given $M$ colours, $M$ population size functions $\alpha \triangleq \{\alpha_j(t)\}_{1 \leq j \leq M}$, the set of $M-1$ divergence times $T_{div} = \{t_{div_j}\}_{1 \leq j < M}$, the set of $M-1$ divergence events $D = \{d_j\}_{1 \leq j < M}$, the set of tips $I \subset \mathbb{N}$, $I = \{1, 2, ..., N\}$, and the associated sampling times

$T_{sam} \triangleq \{t_i^s\}_{i \in I}$. Assume that the indexing is such that the divergence times are in ascending order.
Let $\mathbf{\Pi}(t)$ be a continuous time markov chain defined on the state space of the partitions of $I$ denoted by $\Pi$, each associated with a colouring via the function $H : \Pi \times \mathbb{R}^+ \mapsto \{1, ...M\}$. The initial state is given by $\mathbf{\Pi}(0) = \{1\}, \{2\}, ..., \{N-1\}, \{N\}$.

Denote the number of extant partitions of given colour $j$ at time $t$ by $|C_j(t)| \triangleq |\{\pi \in \Pi(t) : H(\pi, t) = j\}|$

And transition rates

$$q_{(\pi_i, \pi_j), (\pi_i \cup \pi_j)} = \lim_{h \downarrow 0} \frac{1}{h} \mathbf{P} \left[ \pi_i \cup \pi_j \in \mathbf{\Pi}(t+h) \mid \pi_i, \pi_j \in \mathbf{\Pi}(t), \pi_i \neq \pi_j \right] \tag{1.1}$$

given by

$$q_{(\pi_i, \pi_j), (\pi_i \cup \pi_j)} = \sum_{k=1}^{M} \delta_k(H(\pi_i, t)) \delta_k(H(\pi_j, t)) \frac{\binom{C_k(t)}{2}}{\alpha_k(t)} \tag{1.2}$$

Finally, $H$ is defined recursively with $\forall \pi_i \in \mathbf{\Pi}(0), H(\pi_i, 0) = c_{i,0}$ with $c_{i,0}$ given. For fixed $t$ and and $\forall \pi \in \mathbf{\Pi}(t)$, The colouring function satisfies $H(\pi, t) = H(\pi_0, t) \quad \forall \pi_0 \in \mathbf{\Pi}(0) : \pi_0 \subset \pi$. In time, $H$ is defined so that $\forall \pi \in \mathbf{\Pi}(0) : H(\pi, 0) = j, \quad \forall t > 0 \quad H(\pi, t) = \begin{cases} j & t < T_{div_j} \\ d_j & t \geq T_{div_j} \end{cases}$

While the model is defined in terms of partitions, when working with genealogies we typically work with trees. The tree corresponding to a particular realisation of the partition based model is obtained by identifying all $\pi \in \mathbf{\Pi}(t) \forall t$ with $V_{\mathbf{g}}$. The leaves $V_I$ consist of all partitions $\pi \in \boldsymbol{\pi}(0)$, while internal nodes $V_Y$ consist of all subsequent partitions $\forall t, \forall \pi \in \mathbf{\Pi}(t) : \pi \notin \mathbf{\Pi}(0)$. The edges $V_{\mathbf{g}}$ and times $t_{\mathbf{g}}$ are given by the jump chain. The colouring $c_{\mathbf{g}}$ is obtained through the colouring function $H$.

The interpretation of this model in backwards (coalescent) time is that each node corresponds to a single specific clade (colour). Nodes of the same clade coalesce at i.i.d rates, according to a clade specific growth functions, until reaching the most recent common ancestor (MRCA) of given clade. The MRCA then changes type (colour) to that given by the corresponding divergence event at corresponding divergence time time.

Within-colour transitions and their rates are independent of different colours. As such the transitions involving partitions of one colour are independent of transitions involving partitions of other colours given divergence events and times. This allows us to split the likelihood computation into a product of colour-specific likelihoods, that can be computed under coalescent with variable population size.

## 1.3 Full Model

Given a population sample of $N$ individuals indexed by $F$, $X_F = \{x_i\}_{i \in F}$, first determine the number of recent clonal expansions k:

$$k \sim \texttt{poi}(\phi) \tag{1.3}$$

Then simulate the $k+1$ clonal expansion probability vector $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \sim \texttt{dirichlet}(\psi), \quad \psi \in \mathbb{R}_+^{k+1} \tag{1.4}$$

Given the number of expansions $k$ and expansion probabilities $\boldsymbol{\theta}$, partition $F$ into $k+1$ mutually disjoint subsets $\mathbf{f} = \{f_1, ...f_k, f_{k+1}\}$ with $\bigcup_{i=1}^{k+1} f_i = F$ with the probability $P[j \in f_i] = \theta_i \quad \forall i, \forall j \in F$. Each partition corresponds to an individual colour characterising a clonal expansion. For convenience, we will set $f_{k+1}$ to be the index set of the tips corresponding to the ancestral subpopulation.

The sample along with the initial colour assignment then follows the coalescent process described above in 1.2.1, governed by population size functions $\alpha_j^{-1}(t)$. To each $\alpha_j^{-1}$ corresponds a set of parameters:

$$\begin{aligned} t_{div_j} &\in \mathbb{R}_+ \quad \text{The time of divergence} \\ r_j &\in \mathbb{R}_+ \quad \text{The growth rate of the subpopulation} \\ N_j &\in \mathbb{R}_+ \quad \text{The carrying capacity} \end{aligned} \tag{1.5}$$

The divergence events $d_i$ are simulated so that an expansion with divergence time $t_{div_j}$ merges with and thus changes colour to that of an expansion chosen uniformly at random from all expansions with time of divergence after $t_{div_j}$, i.e. an expansion that is extant at time $t_{div_j}$.

In the case of the ancestral subpopulation $f_{k+1}$, we assume that the divergence event happened a very long time ago. As by definition $\alpha(t) \to N$ as $T_{div} - t \to -\infty$, we approximate $\alpha_{k+1}(t) \approx N_{k+1}$ and as such $\alpha_{k+1}^{-1}(t) \approx 1/N_{k+1}$. The parameters $T_{div_j}, r_j, N_j$ are simulated from appropriate (prior) distributions. Finally, without loss of generality, assume that divergence events are indexed in ascending order by the time of divergence in coalescent time.

$$i > j \Leftrightarrow T_{div_i} > T_{div_j} \tag{1.6}$$

Sampling times $\mathbf{t} = \{t_i\}_{i \in F}$ are simulated from the same distribution for all $f$. It is required that for all expansions $j$, $t_i < T_{div_j}$, for all $i \in f_j$. In other words, all sampling for a given clade must happen after the clade diverges.

#### Choice of effective population size functions

While the parent clade is assumed to be at equilibrium, the population size functions of the diverging clades have to satisfy several properties.

First we introduce the variable $\tau = -t + T_{max} - T_{div}$ which denotes time relative to a divergence event of a given clade, where $T_{div}$ denotes the divergence time and $T_{max}$ denotes the time of the most recent sample. In our model we assume that the diverging subpopulation only appears after the divergence event, and as such it is required that at time $\tau = 0$ the population vanishes $\alpha(\tau) = 0$. Furthermore, we are looking for a monotone decreasing function in $\tau$, that exhibits saturating behaviour as $\tau$ grows large. Initially, functions exhibiting a period of exponential growth were investigated, however these were ill-posed numerically. Hence we arrived at the following function

$$\alpha(\tau) = K\frac{r\tau^2}{1 + r\tau^2} \tag{1.7}$$

Where $K$ is the carrying capacity and $r$ is the growth rate. This function exhibits saturating behaviour, symmetry around zero, and numerically stable behaviour due to both not relying on time-offsetting exponentials, and separately also having a more gradual decay around zero.

Figure 1.1: $\alpha(\tau)$ under different parameters, with $T_{div} = 20$

The integral of the reciprocal $\alpha^{-1}(\tau)$ under this formulation is then given by

$$\begin{aligned} &\int_t^{t+s} \alpha^{-1}(\tau)\, d\tau \\ &= \frac{1}{K}\left[-\frac{1}{r(\tau - T_{max} + T_{div})} + \tau - T_{max} + T_{div}\right]_t^{t+s} \end{aligned} \tag{1.8}$$

As $t + s$ approaches the divergence time relative to the most recent sample $T_{div} - T_{max}$, the rate integral 1.8 approaches infinity, and as such all coalescence within a clade happens before time of divergence with probability one.

**Summary**

### 1.3.1 Posterior

The posterior for the full process is:

$$\begin{aligned} P(k, \mathbf{f}, \boldsymbol{\theta}, \boldsymbol{T_{div}}, \mathbf{r}, \mathbf{N} \mid \mathbf{g}) \propto\ & \mathcal{L}(\mathbf{g} \mid \mathbf{f}, \boldsymbol{\rho}, \boldsymbol{T_{div}}, \mathbf{r}, \mathbf{N}) \\ & \times \mathcal{L}(\mathbf{f} \mid \boldsymbol{\theta}, k) \\ & \times \pi(\boldsymbol{\rho} \mid \mathbf{T_{div}}) \\ & \times \pi(\mathbf{T_{div}}, \mathbf{r}, \mathbf{N} \mid k) \\ & \times \pi(\boldsymbol{\theta} \mid k)\pi(k) \end{aligned} \tag{1.9}$$

The first likelihood term

$$\mathcal{L}(\mathbf{g} \mid \mathbf{f}, \boldsymbol{\rho}, \boldsymbol{T_{div}}, \mathbf{r}, \mathbf{N}) \tag{1.10}$$

corresponds to the likelihood of the coalescent with local population strucuture described in 1.2.1. It can be expressed as a product of likelihoods of clonal expansion subtrees with diverging clonal expansions added as leaves under varying population size coalescent process given the corresponding population size function $\alpha_i$.

$$\mathcal{L}(\mathbf{g} \mid \mathbf{f}, \boldsymbol{\rho}, \boldsymbol{T_{div}}, \mathbf{r}, \mathbf{N}) = \prod_{i=1}^{k+1} \mathcal{L}(\mathbf{g}_i \mid T_{div_i}, r_i, N_i) \tag{1.11}$$

Where $\mathbf{g_i}$ is the clonal expansion subtree corresponding to expansion $i$, contanining all the leaves in $f_i$, the entire clade specified by theses, as well as any divergence events $j$ added as leaves if $\rho_j = i$. The second term in the likelihood

$$\mathcal{L}(\mathbf{f} \mid \boldsymbol{\theta}, k) \tag{1.12}$$

Corresponds to the likelihood of samples corresponding to given clonal expansion clades, given the clonal expansion clade sampling probabilities. It is simply a multinomial probability

$$\mathcal{L}(\mathbf{f} \mid \boldsymbol{\theta}, k) = \prod_{i=1}^{k} \theta_i^{|f_i|} \tag{1.13}$$

As the model dimensionality varies with $k$, the posterior has to be integrated via RJ-MCMC. Furthermore as $\boldsymbol{\theta}$ has to be a multinomial probability vector, it must be rescaled with every transdimensional move so that it sums to one, leading to non-unitary determinants. For inference we use the following priors: The prior on the parental clades of individual clonal expansions

$$\pi(\boldsymbol{\rho} \mid \mathbf{T_{div}}) \tag{1.14}$$

is assumed to be uniform probability that the clonal expansion diverged from any other subpopulation extant at the time, i.e.:

$$\pi(\boldsymbol{\rho} \mid \mathbf{T_{div}}) \sim \prod_{i=1}^{k} \mathbb{1}_{\{x:x>i\}}(\rho_i) \frac{1}{k+1-i} \tag{1.15}$$

The prior on functions of effective population size can be expressed as

$$\pi(\mathbf{T_{div}}, \mathbf{r}, \mathbf{N} \mid k) \sim \pi(N_{k+1}) \prod_{i=1}^{k} \pi(N_i)\pi(T_{div_i})\pi(r_i) \tag{1.16}$$

For $\pi(N_i) \quad \forall i \in \{0, ..., k-1, k\}$, log-normal distributions parametrised by mean $\mu_N$ and standard deviation $\sigma_N$ provided by the decision maker are used.
$\pi(r_i) \quad \forall i \in \{1, ..., k-1, k\}$, log-normal distributions parametrised by mean $\mu_r$ and standard deviation $\sigma_r$ provided by the decision maker are used.
$\pi(T_{div_i})$ are assumed to be gamma distributed, parametrised by mean and variance supplied by the decision maker.
For prior on membership probabilities

$$\pi(\boldsymbol{\theta} \mid k) \sim \texttt{dirichlet}(\psi) \tag{1.17}$$

$k$-dimensional dirichlet distribution is used, with $\psi$ supplied by the decision maker.

For prior on the number of clonal expansions

$$\pi(k) \sim \texttt{poi}(1) \tag{1.18}$$

we use poisson distribution with rate 1

# Chapter 2

# References