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 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 θ_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

Suppose we sequence the genomes of a set of pathogenic bacterial samples. At

an unknown point in time a particular strain acquired a mutation which conferred resistance to a widely used antibiotic. This increases the strain's fitness and enables it to undergo a period of rapid growth leading to a clonal expansion. Assuming that this increase in fitness occurs in a short time span, the clade – a set of lineages sharing the same common ancestor, of this strain will behave differently in the phylogenetic tree. The clade corresponding to this strain will have a coalescent rate corresponding to a rapidly expanding population starting from a very small number of individuals.
The problem of identifying hidden population structure has been proposed in [1], where a testing based approach was used to identify structure in a phylogeny, as well as in [2], where a birth-death type model was used.
In our approach, we will build upon the standard coalescent model, modifying it as to allow for change points located on the branches of the phylogeny, marking the event when a particular clade starts behaving according to a different population size function than its parent clade.
In our 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 clades, each behaving under its own growth function.
Similar models have been used in epidemiology to track outbreaks [3], or transmission chains [4]. These models are often called structured coalescent process, effectively adding a colour property to the vertices of phylogenies.

1.1.1 Model

1.2 Preliminaries

A given genealogy $\mathbf{g} = (V_{\mathbf{g}}, E_{\mathbf{g}}, t_{\mathbf{g}})$ is an incomplete, empirical sample of the underlying process.
It consists of nodes $V_{\mathbf{g}}$, directed edges $E_{\mathbf{g}}$, and node labels $t_{\mathbf{g}}$ corresponding to event times.
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 $\alpha \triangleq \{\alpha_j(t)\}_{1\leq j\leq M}$. Let $Y(t)$ be a CTMC with the state space:

$$\Sigma = \{\mathbf{s} \in \mathbb{Z}_+ : |\mathbf{s}| \geq 1\} \tag{1.1}$$

and the transition rates

$$\mathbf{s} \to \mathbf{s} - \mathbf{e_j} \qquad \binom{s_j}{2}\alpha_j^{-1}(t) \qquad 1 \leq j \leq M \tag{1.2}$$

$$\mathbf{s} \to \mathbf{s} - \mathbf{e_j} + \mathbf{e_k} \qquad \delta_{1,j}\beta s_k \qquad 1 \leq j,k \leq M \tag{1.3}$$

Where β is an unknown rate.

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 of any other clade that is extant at a given time.
Under the hypothesis, along each of the edges from the parent of a clade MRCA to the MRCA lies a point in that characterises the time of divergence of the clade, after which the clade starts to undergo clonal expansion.

1.3 Full Generative Model

We now proceed to define the full model that will be used for inference and simulations. First we note that as the coalescent process is backwards in time, our model will have to be backwards in time as well.
Assume k individuals $x_1, x_2, ..., x_k$ have been sampled, with k fixed and known. The k individuals can be partitioned into $n+1$ colours corresponding to clades, with $n \sim \texttt{poi}(\theta)$. We denote the indicator of colouring of an individual with $I_{c_j}(\cdot)$ Each individual is assigned to clade $j \in 1, ..., n+1$ with probability p_j. In other words $P[I_{c_j}(x_i) = 1] = p_j$. The probability vector $p = (p_i)_{1,..,n+1}$ is drawn from a dirichlet distribution with a given concentration α, i.e. $p \sim \texttt{dirichlet}(\alpha)$. Next, expansion parameters such as carrying capacity, expansion rate, and time of divergence time of expansion are sampled. The case of background population can be viewed as an expansion that has happened sufficiently long ago that it is effectively constant. As such for background population we only sample carrying capacity. Finally, with all parameters sampled, we proceed to simulate a genealogy under model described in 1.2.1.

Chapter 2

References

Bibliography

[1] 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).

[2] Joëlle Barido-Sottani, Timothy G. Vaughan, and Tanja Stadler. "A Multitype Birth–Death Model for Bayesian Inference of Lineage-Specific Birth and Death Rates". In: *Systematic Biology* 69.5 (Sept. 1, 2020). Publisher: Oxford Academic, pp. 973–986. ISSN: 1063-5157. DOI: 10.1093/sysbio/syaa016. URL: https://academic.oup.com/sysbio/article/69/5/973/5762626 (visited on 08/28/2020).

[3] 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).

[4] Xavier Didelot et al. "Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks". In: *Molecular Biology and Evolution* 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).