Simulation of Markov genealogy processes

Aaron A. King

February 28, 2024

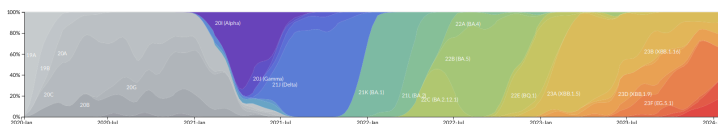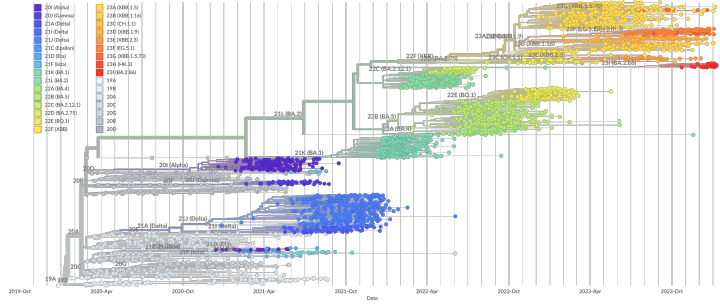## Contents

Objectives

1 Context

1.1 Example: emerging variants

Example: surveillance for emerging SARS-CoV-2 variants

$$\lambda_1 = \beta_1 \frac{I_1}{N} \qquad \lambda_2 = \beta_2 \frac{I_2}{N}$$

COVID-19 sequenced cases by variant
US + Canada, biweekly
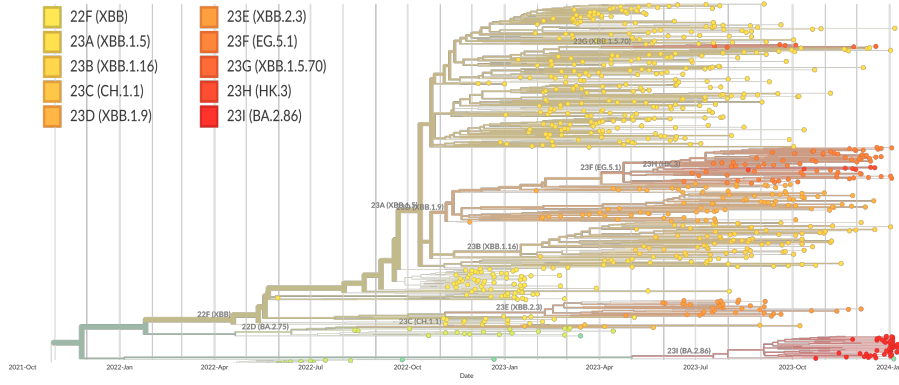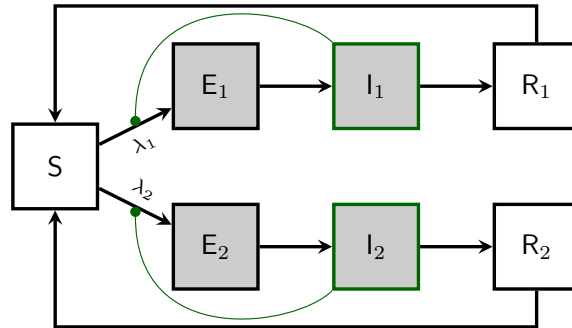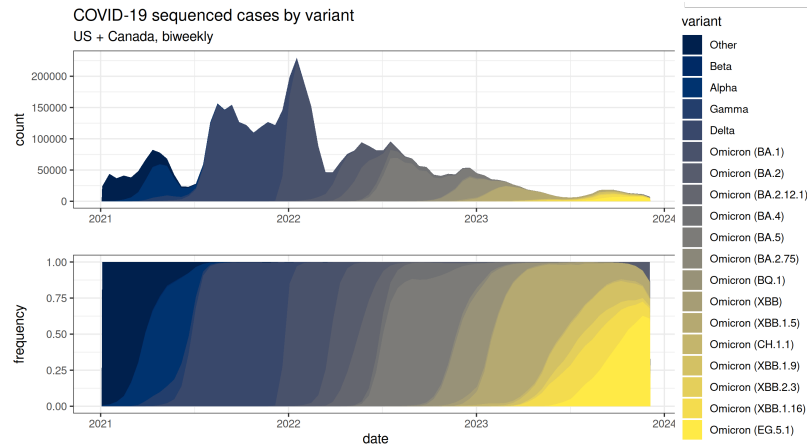
(Mathieu *et al.*, 2020)

1.2 Phylodynamics

What is phylodynamics?

Broadly: Phylodynamics is the project of inferring *determinants of epidemic spread* using *genomic data from pathogen samples.*

In this talk: Phylodynamics means using *genomic data* to infer *stochastic dynamic transmission models.*

1.3 Problems of phylodynamics

Core problems of phylodynamics

Pathogen genealogy

ACCGTATCGCCA

ACCG**GC**TCG**G**CA

Transmission tree

t

genome sequences

prediction

genealogy

stochastic transmission model
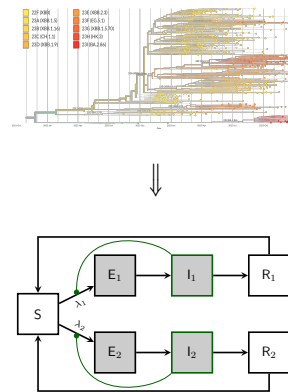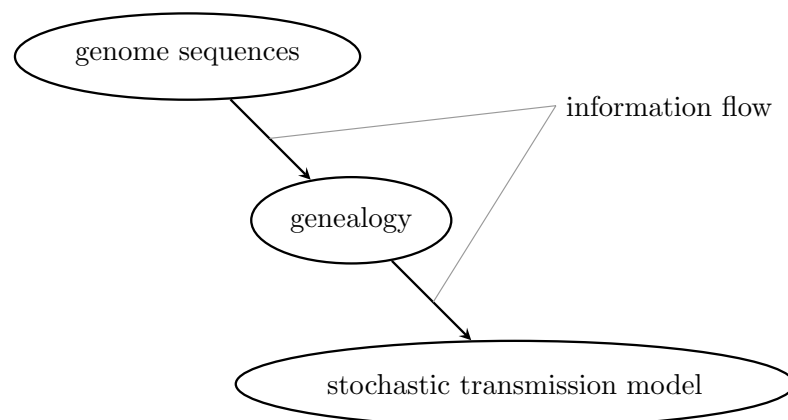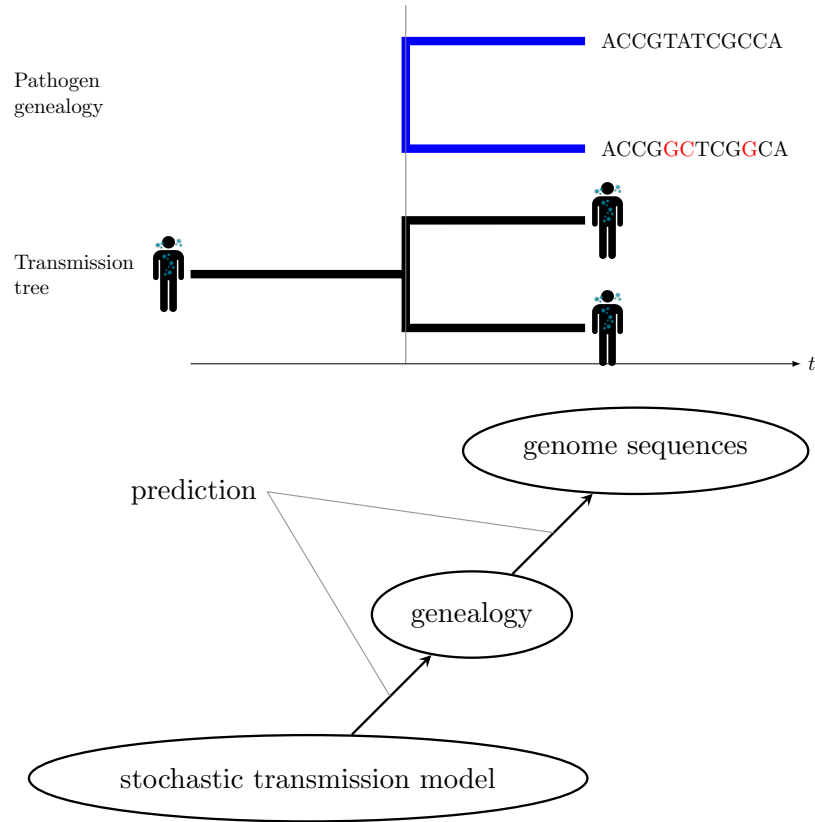
Overview

- We show how a given population process induces a unique genealogy process.

- *Pruning* and *obscuration* project a genealogy onto observable data.

- We derive the exact likelihood as the solution to a non-linear filtering problem

- This equation can be solved by standard Monte Carlo methods.

2 Population process

2.1 Examples

Population process

A

$\mathbb{D} = \{E, I\}$

B

$\mathbb{D} = \{E_1, I_1, E_2, I_2\}$

C

$\mathbb{D} = \{E, I_A, I_S\}$

D

$\mathbb{D} = \{E, I_L, I_H\}$

2.2 Formalization

Population process

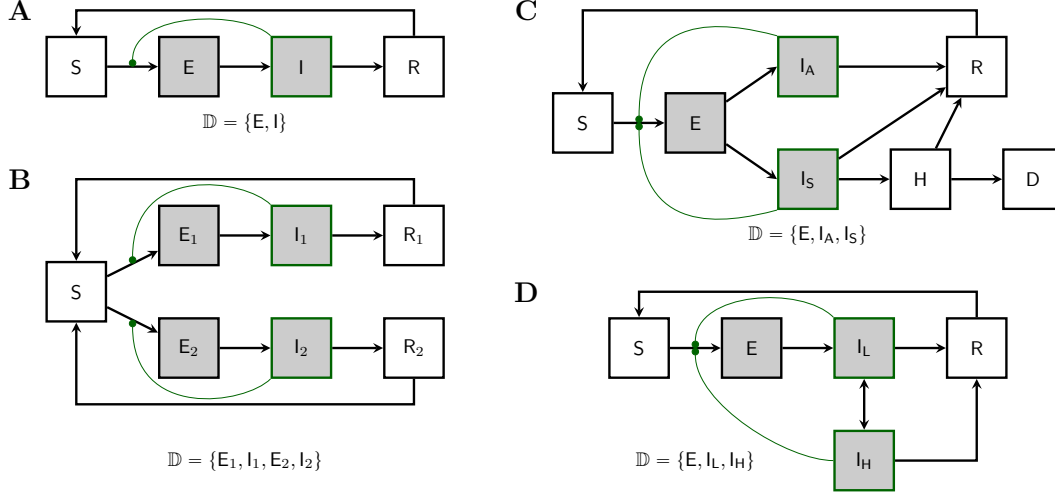- Non-explosive Markov jump process, $\mathbf{X}_t \in \mathbb{X}$, $t \in \mathbb{R}_+$: the *population process*.

- Initial-state distribution, p_0:

$$\mathsf{Prob}\left[\mathbf{X}_0 \in \mathcal{E}\right] = \int_{\mathcal{E}} p_0(x)\,\mathrm{d}x$$

- Jump rates: $\alpha(t, x, x') = $ rate of jump $x \to x'$

$$\alpha(t, x, x') \geqslant 0, \qquad \int_{\mathbb{X}} \alpha(t, x, x')\,\mathrm{d}x' < \infty$$

- Multiple events at each jump are allowed.

Kolmogorov forward equation (KFE): If

$$\frac{\partial w}{\partial t}(t, x) = \int w(t, x')\,\alpha(t, x', x)\,\mathrm{d}x' - \int w(t, x)\,\alpha(t, x, x')\,\mathrm{d}x'$$

and

$$w(0, x) = p_0(x)$$

then

$$\int_{\mathcal{E}} w(t, x)\,\mathrm{d}x = \mathsf{Prob}\left[\mathbf{X}_t \in \mathcal{E}\right].$$

KFE is sometimes called the *master equation* for $\mathbf{X}_t$.

$$\frac{\partial w}{\partial t}(t, x) = \int w(t, x')\,\alpha(t, x', x)\,\mathrm{d}x' - \int w(t, x)\,\alpha(t, x, x')\,\mathrm{d}x'$$
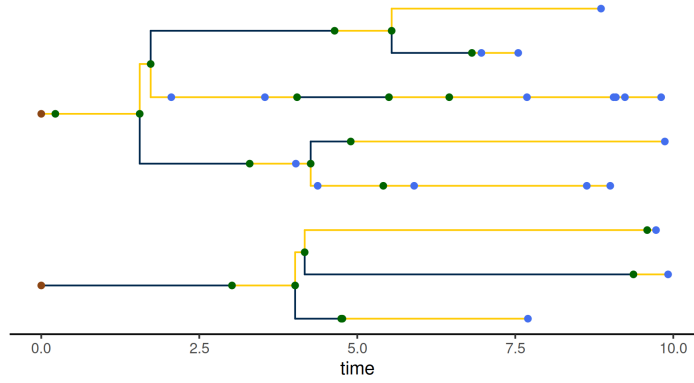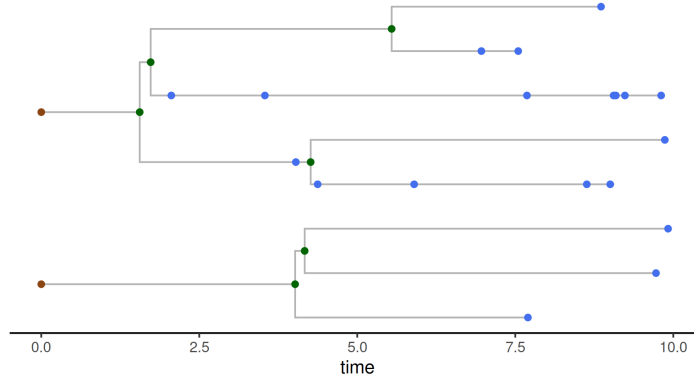
5

$$\mathbb{U} = \{\text{Trans}, \text{Prog}, \text{Recov}, \text{Wane}, \text{Sample}\}$$
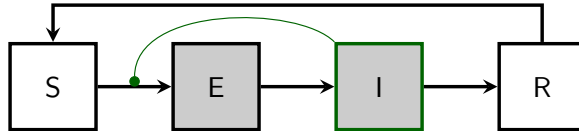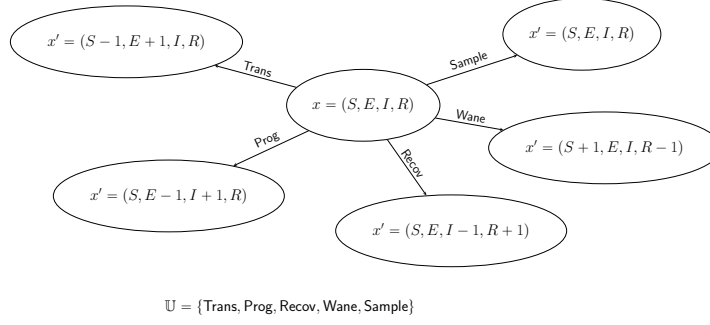
$$\frac{\partial w}{\partial t}(t, x) = \sum_{u \in \mathbb{U}} \left\{ \int w(t, x') \, \alpha_u(t, x', x) \, \mathrm{d}x' - \int w(t, x) \, \alpha_u(t, x, x') \, \mathrm{d}x' \right\}$$

$$\frac{\partial w}{\partial t}(t, S, E, I, R) = \frac{\beta(t) \, (S+1) \, I}{N} \, w(t, S+1, E-1, I, R) - \frac{\beta(t) \, S \, I}{N} \, w(t, S, E, I, R) + \sigma \, (E+1) \, w(t, S, E+1, I-1, R) - \sigma \, E \, w(t, S, E, I, R)$$
$$+ \gamma \, (I+1) \, w(t, S, E, I+1, R-1) - \gamma \, I \, w(t, S, E, I, R) + \omega \, (R+1) \, w(t, S-1, E, I, R+1) - \omega \, R \, w(t, S, E, I, R)$$

3 Genealogy process

3.1 Genealogies

What is a genealogy?
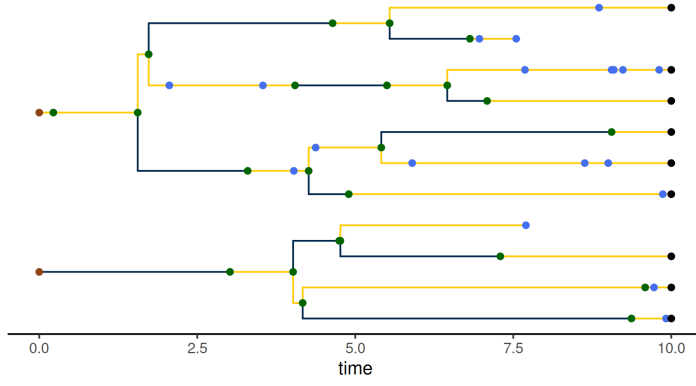
3.2 Induced genealogy process

Event types

$$\mathbb{D} = \{\mathsf{E}, \mathsf{I}\}$$

$$\mathbb{U} = \{\mathsf{Trans}, \mathsf{Prog}, \mathsf{Recov}, \mathsf{Wane}, \mathsf{Sample}\}$$
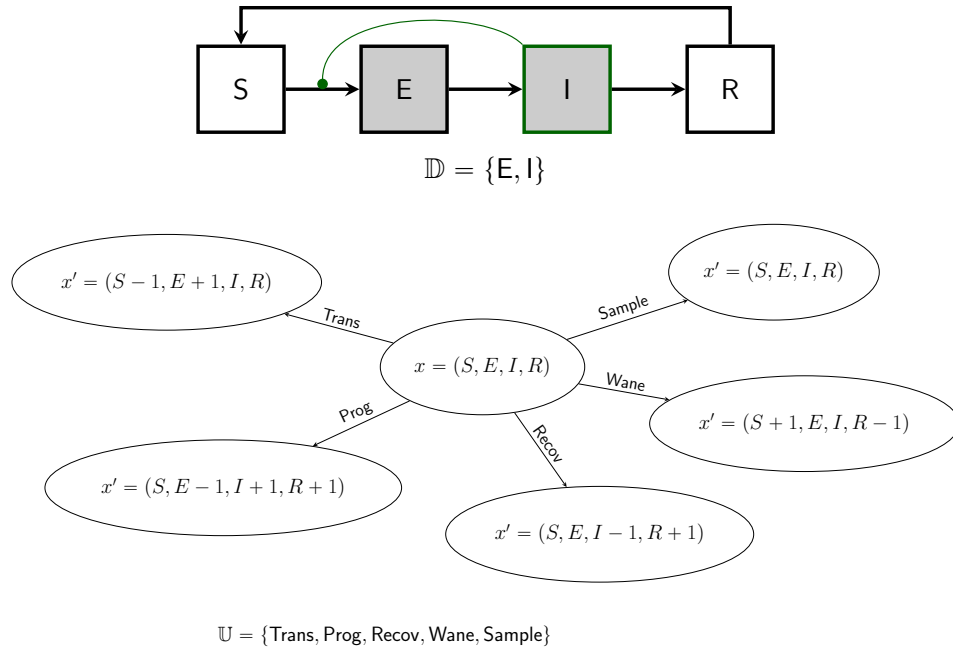
If we write

$$\alpha(t, x, x') = \sum_{u \in \mathbb{U}} \alpha_u(t, x, x'),$$

the KFE becomes

$$\frac{\partial w}{\partial t}(t, x) = \sum_u \int w(t, x')\, \alpha_u(t, x', x)\, \mathrm{d}x' - \sum_u \int w(t, x)\, \alpha_u(t, x, x')\, \mathrm{d}x'$$
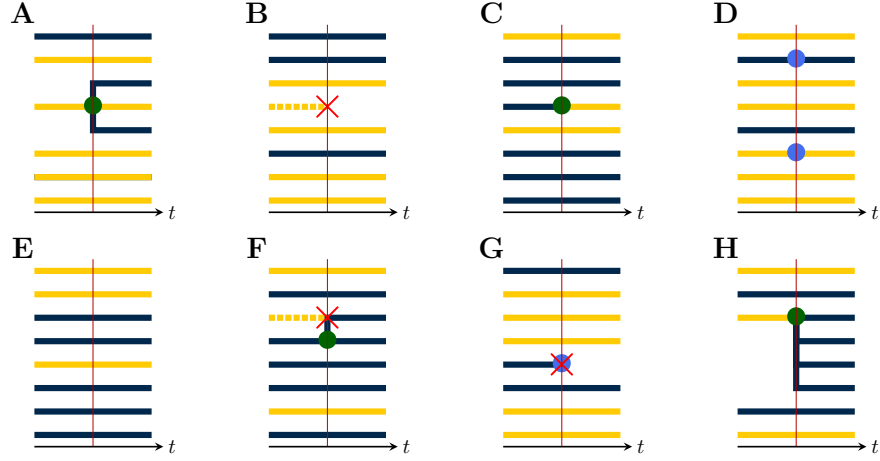
A population process induces a genealogy process

- $\mathbf{G}_t$ is a stochastic process on the space of genealogies.

- The map $\mathbf{X} \mapsto \mathbf{G}$ is random.

- **Key assumption:** Lineages within a deme are *exchangeable*. There is no more structure than is implied by the population process.

- Simulation code on github.com/kingaa/phylopomp

- Animations at https://kingaa.github.io/manuals/phylopomp/vignettes/

4 Pruned and obscured genealogies

Full genealogy
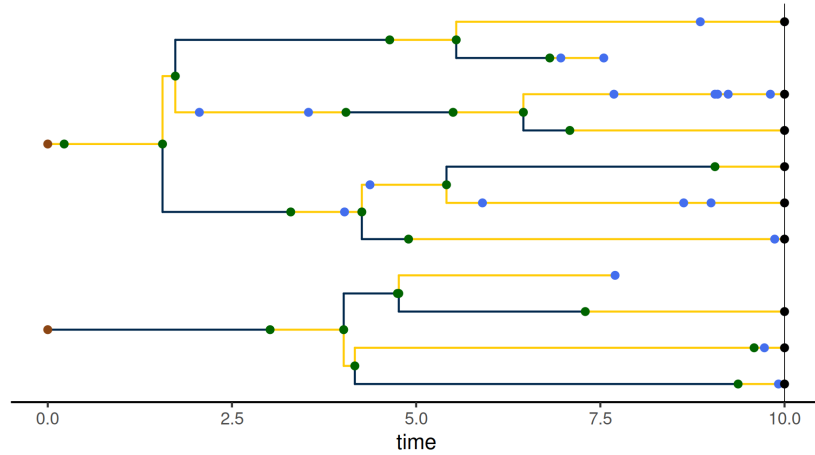
Pruned genealogy
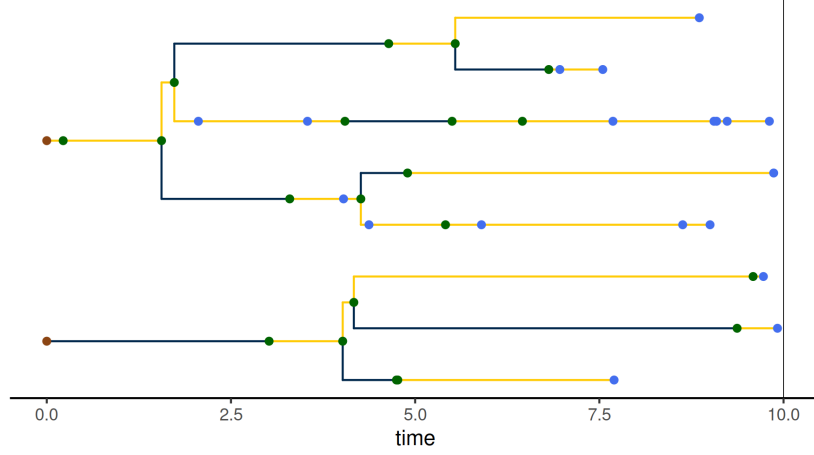
Local structure of a pruned genealogy

Top row shows the *unpruned genealogy* in neighborhood of an event. Bottom row shows the corresponding *pruned genealogy*.
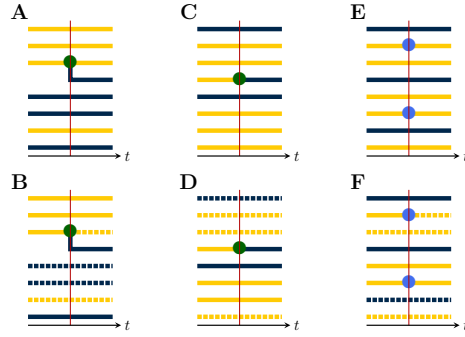
For $x \in \mathbb{X}$, $i \in \mathbb{D}$, $n_i(x)$ is the *occupancy* of deme i when the system is in state x. In panel A $n = (n_{\mathsf{blue}}, n_{\mathsf{yellow}}) = (4, 4)$; in panel C $n = (3, 5)$;
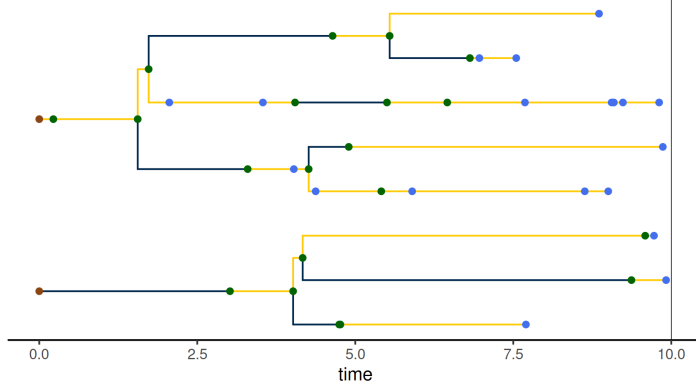
For $u \in \mathbb{U}$, $i \in \mathbb{D}$, r_i^u is the *production* of event u in deme i. In panel A, $r = (r_{\mathsf{blue}}, r_{\mathsf{yellow}}) = (1, 1)$; in panel E, $r = (0, 2)$.

The *lineage count*, $\ell_i(t)$, is the number of unpruned lineages in deme i at time t. In this case, for all panels, $\ell = (2, 2)$.

The *saturation*, s_i, is the number of unpruned lineages in deme i *descending* from the event. In panels B and D, $s = (1, 0)$; in panel F, $s = (0, 1)$.
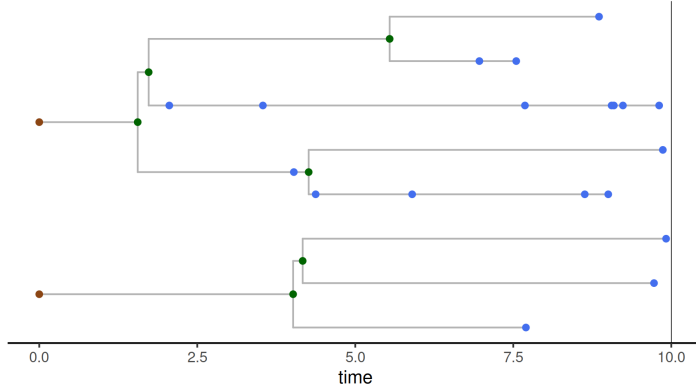
Obviously, $s_i \leqslant r_i \leqslant n_i$ and $s_i \leqslant \ell_i \leqslant n_i$.

Pruned genealogy

A pruned genealogy is specified by two functions of time, (Y, Z): Z_t gives the local topological structure; Y_t gives the local coloring.

Obscured genealogy

An obscured genealogy is specified by (T, Z).

Binomial ratio
For $n, r, \ell, s \in \mathbb{Z}_+^{\mathbb{D}}$, define the *binomial ratio*

$$
\begin{pmatrix} n & \ell \\ r & s \end{pmatrix} := \begin{cases} \prod_{i \in \mathbb{D}} \dfrac{\binom{n_i - \ell_i}{r_i - s_i}}{\binom{n_i}{r_i}}, & \text{if } \forall i \ n_i \geqslant \{\ell_i, r_i\} \geqslant s_i \geqslant 0, \\ 0, & \text{otherwise.} \end{cases}
$$

Observe that $\begin{pmatrix} n & \ell \\ r & s \end{pmatrix} \in [0, 1]$. Moreover,

$$
\sum_{s \in \mathbb{Z}_+^{\mathbb{D}}} \begin{pmatrix} n & \ell \\ r & s \end{pmatrix} \begin{pmatrix} \ell \\ s \end{pmatrix} = 1.
$$

5 Theorems

5.1 Pruned genealogies

Theorem: likelihood of a pruned genealogy
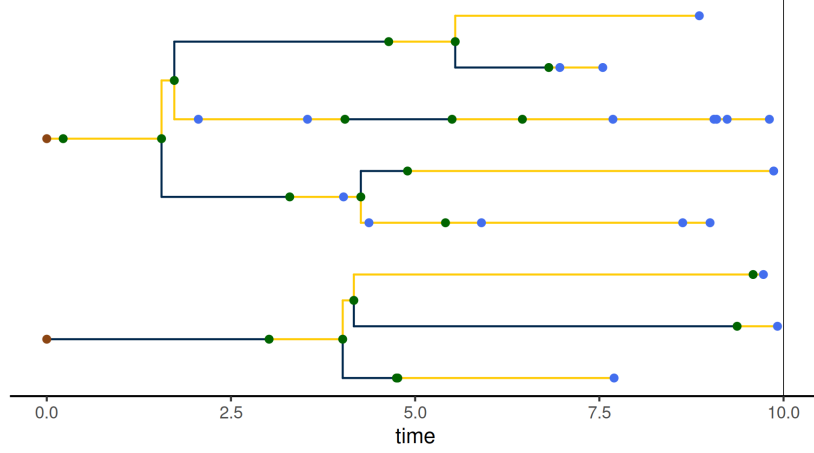
Suppose that $P = (Y, Z)$ is a given pruned genealogy with depth T.
Define

$$\phi_u(x, y, y') := \begin{pmatrix} n(x) & \ell(y') \\ r^u & s(y, y') \end{pmatrix} Q_u(y, y').$$

Here, $Q = 1$ if the local structure of P is compatible with an event of type u at that time; $Q = 0$ otherwise.
If $w = w(t, x)$ satisfies the initial condition $w(0, x) = p_0(x)$ and the filter equation
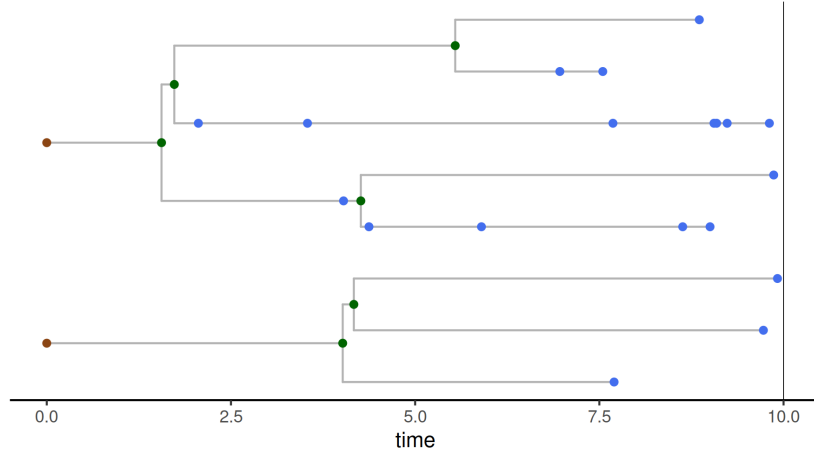
$$\frac{\partial w}{\partial t}(t, x) = \sum_u \int w(t, x') \, \alpha_u(t, x', x) \, \phi_u(x, \widetilde{Y}_t, Y_t) \, dx' - \sum_u \int w(t, x) \, \alpha_u(t, x, x') \, dx'$$

$$+ \sum_{e \in \text{ev}(P)} \delta(t, e) \left\{ \sum_u \int w(t, x') \, \alpha_u(t, x', x) \, \phi_u(x, \widetilde{Y}_t, Y_t) \, dx' - w(t, x) \right\},$$

then the likelihood of P is

$$\mathcal{L}(P) = \int w(T, x) \, dx.$$

5.2 Obscured genealogies

Theorem: likelihood of an obscured genealogy

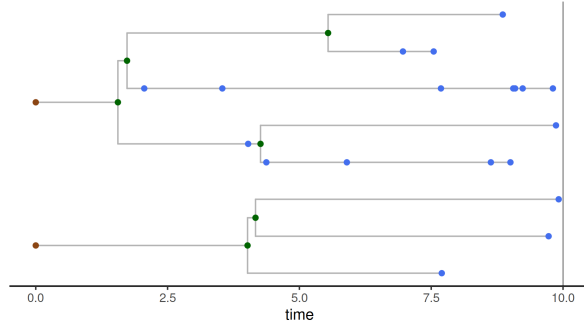Let (T, Z), be a given obscured genealogy. Then there are probability kernels π and q such that if

$$\beta_u(t, x, x', y, y') = \alpha_u(t, x, x')\, \pi_u(t, x, x', y, y'), \qquad \psi_u(t, x, x', y, y') = \frac{\phi_u(x', y, y')}{\pi_u(t, x, x', y, y')},$$

and if $w = w(t, x, y)$ satisfies the initial condition $w(0, x, y) = p_0(x)\, q(x, y)$ and the filter equation

$$\frac{\partial w}{\partial t} = \sum_{uy'} \int w(t, x', y')\, \beta_u(t, x', x, y', y)\, \psi_u(t, x', x, y', y)\, \mathrm{d}x' - \sum_{uy'} \int w(t, x, y)\, \beta_u(t, x, x', y, y')\, \mathrm{d}x'$$

$$+ \sum_{e \in \mathsf{ev}(Z)} \delta(t, e) \left\{ \sum_{uy'} \int w(t, x', y')\, \beta_u(t, x', x, y, y')\, \psi_u(t, x', x, y', y)\, \mathrm{d}x' - w(t, x, y) \right\},$$
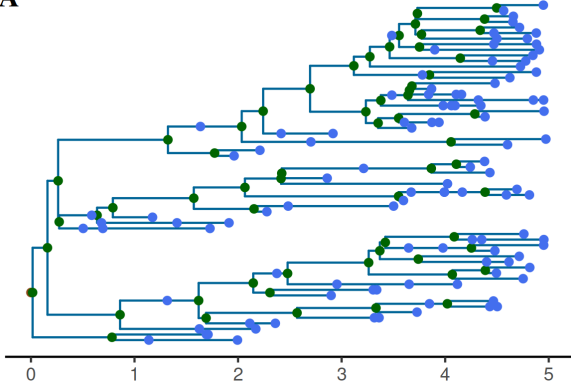
then the likelihood of Z is

$$\mathcal{L}(Z) = \sum_y \int w(T, x, y)\, \mathrm{d}x.$$
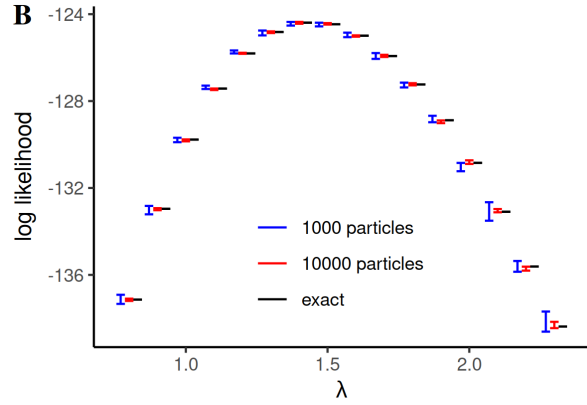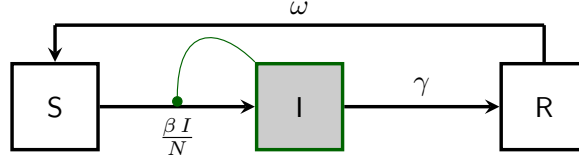

6 Examples

Linear birth-death model

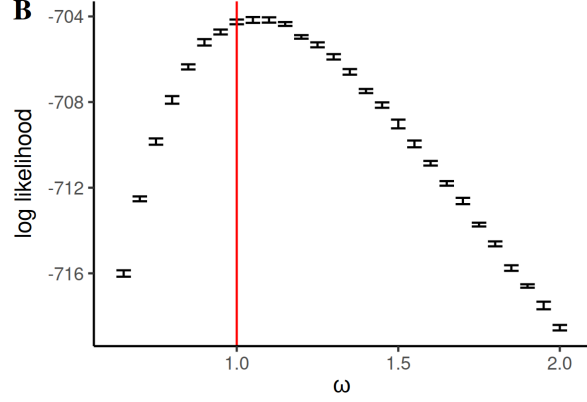Uniform sampling. Exact likelihoood is available in closed form.

SIRS model

12

Between genealogical events:

$$\frac{\partial w}{\partial t} = \frac{\beta\,(S+1)\,(I-1)}{N}\,\left(1 - \frac{\binom{\ell}{2}}{\binom{I}{2}}\right)\,w(t, S+1, I-1, R) + \gamma\,(I+1)\,w(t, S, I+1, R-1)$$

$$+\,\omega\,(R+1)\,w(t, S-1, I, R+1) - \left(\frac{\beta\,S\,I}{N} + \gamma\,I + \omega\,R + \psi\,I\right)\,w(t, S, I, R).$$
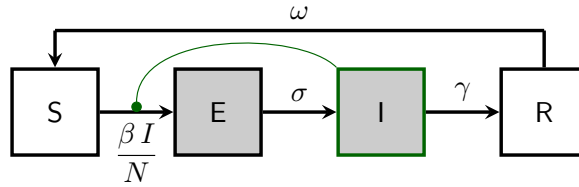
At genealogical events:

$$w(t, S, I, R) = \begin{cases} \frac{2\,\beta\,(S+1)}{I\,N}\,\widetilde{w}(t, S+1, I-1, R), & \text{branch point at } t, \\[2ex] \psi\,\widetilde{w}(t, S, I, R), & \text{internal sample at } t, \\[2ex] \psi\,(I-\ell)\,\widetilde{w}(t, S, I, R), & \text{terminal sample at } t. \end{cases}$$

A

B
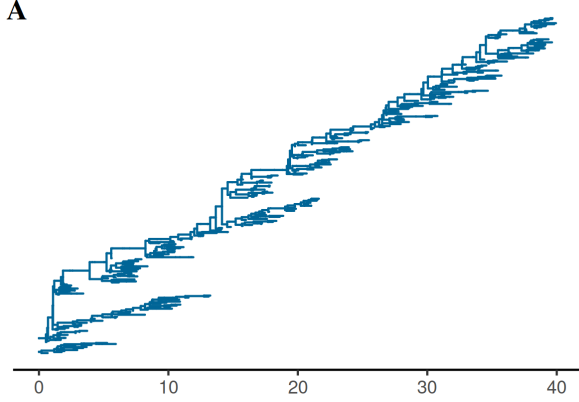
Uniform sampling. One deme only.

SEIRS model
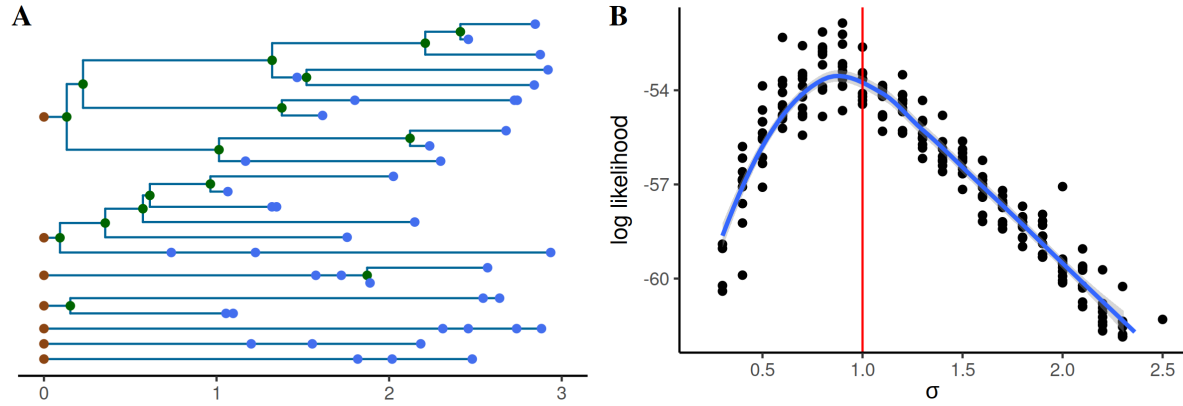
Summary and outstanding challenges

Concluding remarks

- The theory *corrects* and *strictly extends* all existing likelihood-based phylodynamic methods (e.g., Volz *et al.*, 2009; Rasmussen *et al.*, 2011; Stadler, 2010; Volz, 2012; Volz and Frost, 2014; Rasmussen *et al.*, 2014; Vaughan *et al.*, 2019).

- It eliminates the need for large population-size and small sample-fraction assumptions, as well as any dependence on linearization.

- All computations can be carried out forward in time.

- This greatly expands the class of models that can be entertained.

- Great flexibility in sampling model

- Applications in disease ecology and beyond

- The unstructured case can be found in King *et al.* (2022).

Outstanding challenges

- Choice of importance-sampling kernel

- Borrowing information from future is allowed.

- Phylogenetic uncertainty

- Efficient algorithms

- Reassortment and recombination

Summary

- A discretely structured Markov population process uniquely induces a genealogy-valued Markov process.

- The likelihood of an observed genealogy satisfies a nonlinear filtering equation, which can be efficiently computed via Feynman-Kaç (sequential Monte Carlo) algorithms.

- In principle, these results liberate us to entertain models that more closely match our biological questions, without less hindrance from inference methodology.

References

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA (2018). "Nextstrain: real-time tracking of pathogen evolution." *Bioinformatics*, **34**(23), 4121–4123. doi: 10.1093/bioinformatics/bty407. 2

King AA, Lin Q, Ionides EL (2022). "Markov genealogy processes." *Theor Popul Biol*, **143**, 77–91. doi: 10.1016/j.tpb.2021.11.003. 22

King AA, Nguyen D, Ionides EL (2016). "Statistical inference for partially observed Markov processes via the R package pomp." *J Stat Softw*, **69**(12), 1–43. doi: 10.18637/jss.v069.i12.

Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Giattino C, Hasell J, Macdonald B, Dattani S, Beltekian D, Ortiz-Ospina E, Roser M (2020). "Coronavirus pandemic (COVID-19)." *Our World in Data [Online resource]*. URL https://ourworldindata.org/coronavirus. 2

Rasmussen DA, Ratmann O, Koelle K (2011). "Inference for nonlinear epidemiological models using genealogies and time series." *PLoS Comput. Biol.*, **7**(8), e1002136. doi: 10.1371/journal.pcbi.1002136. 22

Rasmussen DA, Volz EM, Koelle K (2014). "Phylodynamic inference for structured epidemiological models." *PLoS Comput. Biol.*, **10**(4), e1003570. doi: 10.1371/journal.pcbi.1003570. 22

Stadler T (2010). "Sampling-through-time in birth-death trees." *J. Theor. Biol.*, **267**(3), 396–404. doi: 10.1016/j.jtbi.2010.09.010. 22

Vaughan TG, Leventhal GE, Rasmussen DA, Drummond AJ, Welch D, Stadler T (2019). "Estimating epidemic incidence and prevalence from genomic data." *Mol. Biol. Evol.*, **36**, 1804–1816. doi: 10.1093/molbev/msz106. 22

Volz EM (2012). "Complex population dynamics and the coalescent under neutrality." *Genetics*, **190**(1), 187–201. doi: 10.1534/genetics.111.134627. 22

Volz EM, Frost SDW (2014). "Sampling through time and phylodynamic inference with coalescent and birth-death models." *J. R. Soc. Interface*, **11**(101), 20140945. doi: 10.1098/rsif.2014.0945. 22

Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW (2009). "Phylodynamics of infectious disease epidemics." *Genetics*, **183**(4), 1421–1430. doi: 10.1534/genetics.109.106021. 22

License, acknowledgments, and links

- The materials build on previous versions of this course and related courses.

- Licensed under the Creative Commons Attribution-NonCommercial license. Please share and remix non-commercially, mentioning its origin.

- Produced with R version 4.3.2 and **pomp** version 5.6.

- Compiled on February 28, 2024.

Back to Lesson
R codes for this lesson