The infinite alleles model revisited: a Gibbs sampling approach

Marc Manceau

August 31, 2021

Introduction

Motivation & Goal

Data:

- A huge number of SARS-CoV-2 sequences have accumulated on GISAID since January 2020.
- The mutation rate is quite low but the sampling is excellent.
- We thus observe lots of sequences that are similar.

This motivates the development of new phylodynamic methods
better tailored to analyze big genomic data characterized by a low diversity

Goals:

- design a simpler data-generating model as compared to current phylodynamic methods,
- with quantities that we are interested in: pop size, sampling intensity, mutation rate,
- and an appropriate inference method to recover these quantities from the data.

Can we recover information on the population size and sampling intensity in the past
while working under a very simple infinite alleles model ?

Introduction
Motivation & Goal

Data:

- A huge number of SARS-CoV-2 sequences have accumulated on GISAID since January 2020.
- The mutation rate is quite low but the sampling is excellent.
- We thus observe lots of sequences that are similar.

This motivates the development of new phylodynamic methods
better tailored to analyze big genomic data characterized by a low diversity

Goals:

- design a simpler data-generating model as compared to current phylodynamic methods,
- with quantities that we are interested in: pop size, sampling intensity, mutation rate,
- and an appropriate inference method to recover these quantities from the data.

Can we recover information on the population size and sampling intensity in the past
while working under a very simple infinite alleles model ?

Introduction

Motivation & Goal

Data:

- A huge number of SARS-CoV-2 sequences have accumulated on GISAID since January 2020.
- The mutation rate is quite low but the sampling is excellent.
- We thus observe lots of sequences that are similar.

This motivates the development of new phylodynamic methods
better tailored to analyze big genomic data characterized by a low diversity

Goals:

- design a simpler data-generating model as compared to current phylodynamic methods,
- with quantities that we are interested in: pop size, sampling intensity, mutation rate,
- and an appropriate inference method to recover these quantities from the data.

Can we recover information on the population size and sampling intensity in the past
while working under a very simple infinite alleles model ?

Introduction

Motivation & Goal

Data:

- A huge number of SARS-CoV-2 sequences have accumulated on GISAID since January 2020.
- The mutation rate is quite low but the sampling is excellent.
- We thus observe lots of sequences that are similar.

This motivates the development of new phylodynamic methods
better tailored to analyze big genomic data characterized by a low diversity

Goals:

- design a simpler data-generating model as compared to current phylodynamic methods,
- with quantities that we are interested in: pop size, sampling intensity, mutation rate,
- and an appropriate inference method to recover these quantities from the data.

Can we recover information on the population size and sampling intensity in the past
while working under a very simple infinite alleles model ?

Introduction
Motivation & Goal

Data:

- A huge number of SARS-CoV-2 sequences have accumulated on GISAID since January 2020.
- The mutation rate is quite low but the sampling is excellent.
- We thus observe lots of sequences that are similar.

This motivates the development of new phylodynamic methods
better tailored to analyze big genomic data characterized by a low diversity

Goals:

- design a simpler data-generating model as compared to current phylodynamic methods,
- with quantities that we are interested in: pop size, sampling intensity, mutation rate,
- and an appropriate inference method to recover these quantities from the data.

Can we recover information on the population size and sampling intensity in the past
while working under a very simple infinite alleles model ?

Introduction

Motivation & Goal

Data:

- A huge number of SARS-CoV-2 sequences have accumulated on GISAID since January 2020.
- The mutation rate is quite low but the sampling is excellent.
- We thus observe lots of sequences that are similar.

This motivates the development of new phylodynamic methods
better tailored to analyze big genomic data characterized by a low diversity

Goals:

- design a simpler data-generating model as compared to current phylodynamic methods,
- with quantities that we are interested in: pop size, sampling intensity, mutation rate,
- and an appropriate inference method to recover these quantities from the data.

Can we recover information on the population size and sampling intensity in the past
while working under a very simple infinite alleles model ?

Introduction

Motivation & Goal

Data:

- A huge number of SARS-CoV-2 sequences have accumulated on GISAID since January 2020.
- The mutation rate is quite low but the sampling is excellent.
- We thus observe lots of sequences that are similar.

This motivates the development of new phylodynamic methods
better tailored to analyze big genomic data characterized by a low diversity

Goals:

- design a simpler data-generating model as compared to current phylodynamic methods,
- with quantities that we are interested in: pop size, sampling intensity, mutation rate,
- and an appropriate inference method to recover these quantities from the data.

Can we recover information on the population size and sampling intensity in the past
while working under a very simple infinite alleles model ?

Introduction

Motivation & Goal

Data:

- A huge number of SARS-CoV-2 sequences have accumulated on GISAID since January 2020.
- The mutation rate is quite low but the sampling is excellent.
- We thus observe lots of sequences that are similar.

This motivates the development of new phylodynamic methods
better tailored to analyze big genomic data characterized by a low diversity

Goals:

- design a simpler data-generating model as compared to current phylodynamic methods,
- with quantities that we are interested in: pop size, sampling intensity, mutation rate,
- and an appropriate inference method to recover these quantities from the data.

Can we recover information on the population size and sampling intensity in the past
while working under a very simple infinite alleles model ?

Introduction

Motivation & Goal

Data:

- A huge number of SARS-CoV-2 sequences have accumulated on GISAID since January 2020.
- The mutation rate is quite low but the sampling is excellent.
- We thus observe lots of sequences that are similar.

This motivates the development of new phylodynamic methods
better tailored to analyze big genomic data characterized by a low diversity

Goals:

- design a simpler data-generating model as compared to current phylodynamic methods,
- with quantities that we are interested in: pop size, sampling intensity, mutation rate,
- and an appropriate inference method to recover these quantities from the data.

Can we recover information on the population size and sampling intensity in the past
while working under a very simple infinite alleles model ?

Model assumptions

Introduction

Model assumptions
- Parameters of the model
- Sampling and coalescent history
- Priors and summary

Inference method
- Gibbs sampling strategy
- Prior conjugacy properties for the parameters
- Data augmentation with past coalescent history
- Summary of the Gibbs sampler

Results
- Validation of the augmentation with the past coalescent history
- Validation of the MCMC by SBC
- Estimation of the running time
- Illustration on the SARS-CoV-2 dataset

Discussion
- Opportunities for future developments
- Conclusion

Model assumptions

Introduction

Model assumptions

- Parameters of the model
- Sampling and coalescent history
- Priors and summary

Inference method

- Gibbs sampling strategy
- Prior conjugacy properties for the parameters
- Data augmentation with past coalescent history
- Summary of the Gibbs sampler

Results

- Validation of the augmentation with the past coalescent history
- Validation of the MCMC by SBC
- Estimation of the running time
- Illustration on the SARS-CoV-2 dataset

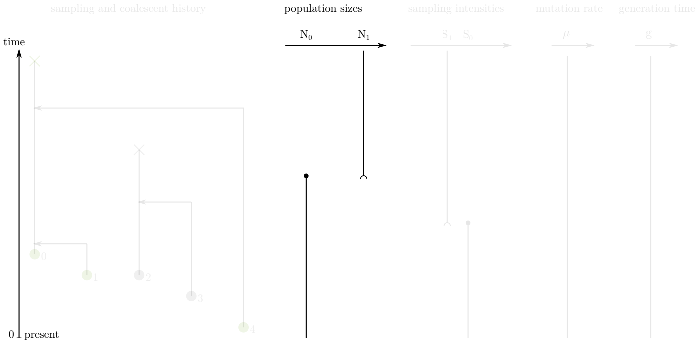
Discussion

- Opportunities for future developments
- Conclusion

Model assumptions

Parameters of the model – Past effective population sizes N

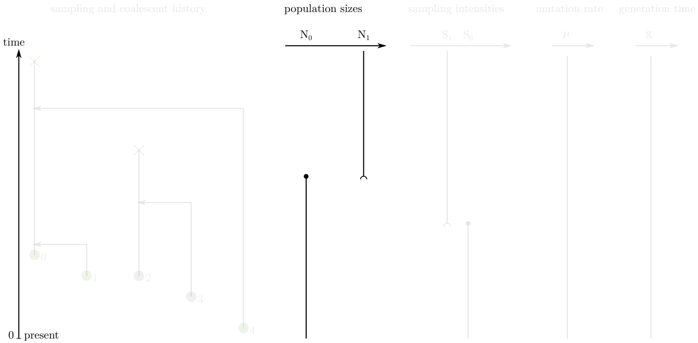
- N is piecewise-constant on a partition $(\Delta_j^{(N)})_{j=0}^{p}$ of $(0, +\infty)$.
- a priori, $N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$.
- (GIG: Generalized Inverse Gaussian distribution, and we'll see a bit later why.)



Model assumptions

Parameters of the model – Past effective population sizes N

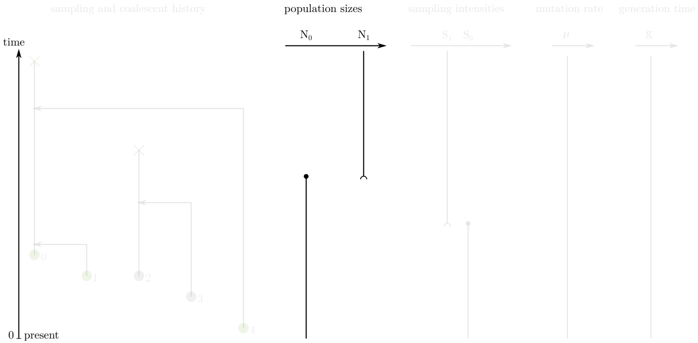
- N is piecewise-constant on a partition $(\Delta_j^{(N)})_{j=0}^{p}$ of $(0, +\infty)$.
- a priori, $N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$.
- (GIG: Generalized Inverse Gaussian distribution, and we'll see a bit later why.)



Model assumptions

Parameters of the model – Past effective population sizes N

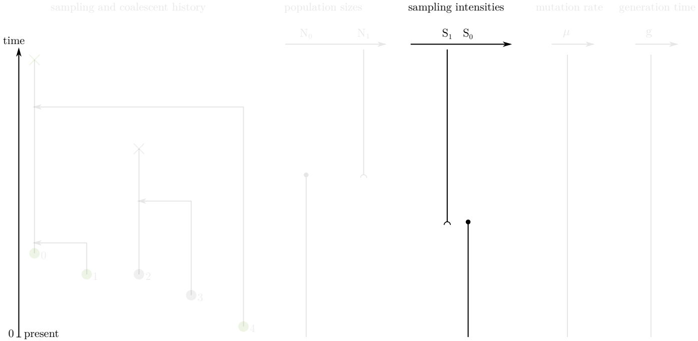
- N is piecewise-constant on a partition $(\Delta_j^{(N)})_{j=0}^{p}$ of $(0, +\infty)$.
- a priori, $N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$.
- (GIG: Generalized Inverse Gaussian distribution, and we'll see a bit later why.)



Model assumptions

Parameters of the model – Past sampling intensities S

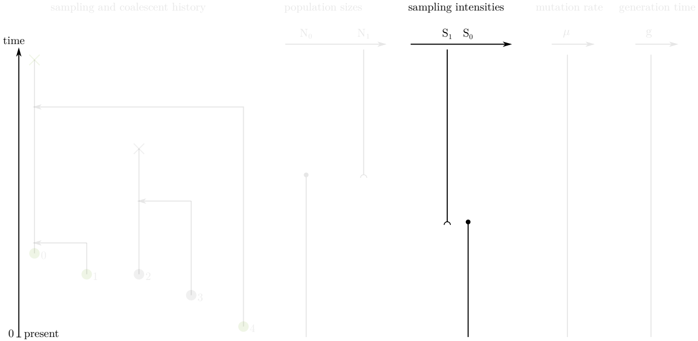
- S is piecewise-constant through time, on a partition $(\Delta_j^{(S)})_{j=0}^{p'}$ of $(0, \infty)$.
- a priori, $S_j \sim \Gamma(\alpha_S, \beta_S)$.
- (and we'll see a bit later why.)



Model assumptions

Parameters of the model – Past sampling intensities S

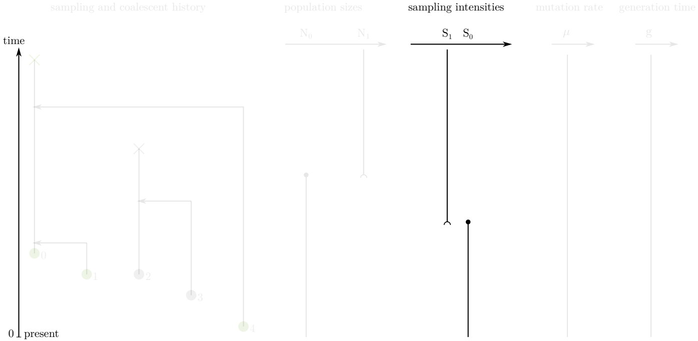
- S is piecewise-constant through time, on a partition $(\Delta_j^{(S)})_{j=0}^{p'}$ of $(0, \infty)$.
- a priori, $S_j \sim \Gamma(\alpha_S, \beta_S)$.
- (and we'll see a bit later why.)



Model assumptions

Parameters of the model – Past sampling intensities S

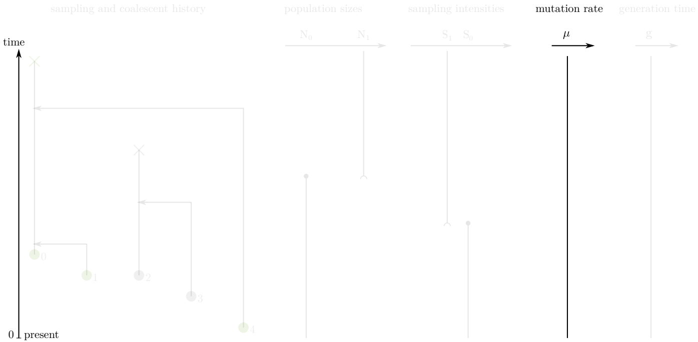
- S is piecewise-constant through time, on a partition $(\Delta_j^{(S)})_{j=0}^{p'}$ of $(0, \infty)$.
- a priori, $S_j \sim \Gamma(\alpha_S, \beta_S)$.
- (and we'll see a bit later why.)



Model assumptions

Parameters of the model – Mutation rate μ

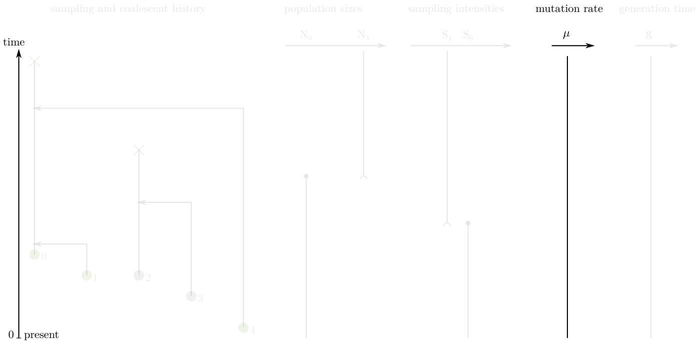
- μ is constant through time,
- a priori, $\mu \sim \Gamma(\alpha_\mu, \beta_\mu)$.
- (and we'll see a bit later why.)



Model assumptions

Parameters of the model – Mutation rate μ

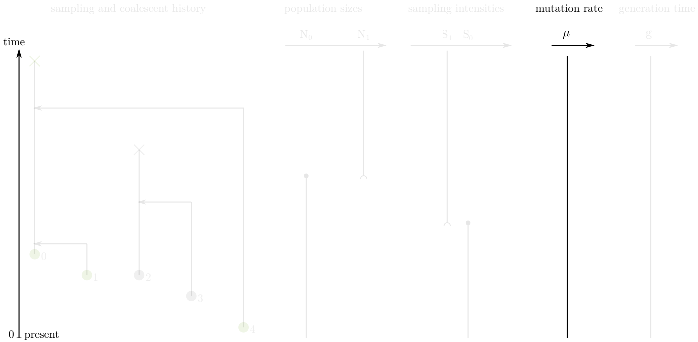
- μ is constant through time,
- a priori, $\mu \sim \Gamma(\alpha_\mu, \beta_\mu)$.
- (and we'll see a bit later why.)



Model assumptions

Parameters of the model – Mutation rate μ

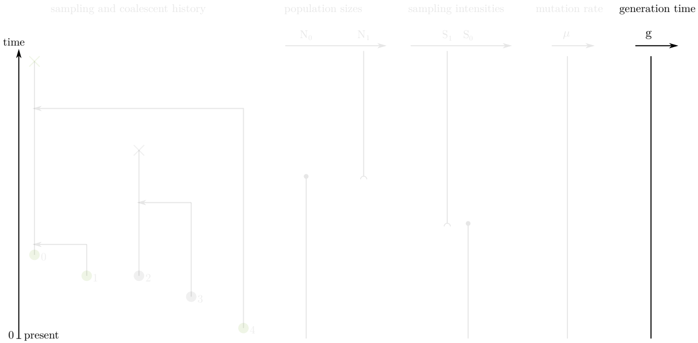
- μ is constant through time,
- a priori, $\mu \sim \Gamma(\alpha_\mu, \beta_\mu)$.
- (and we'll see a bit later why.)



Model assumptions

Parameters of the model – Generation time g

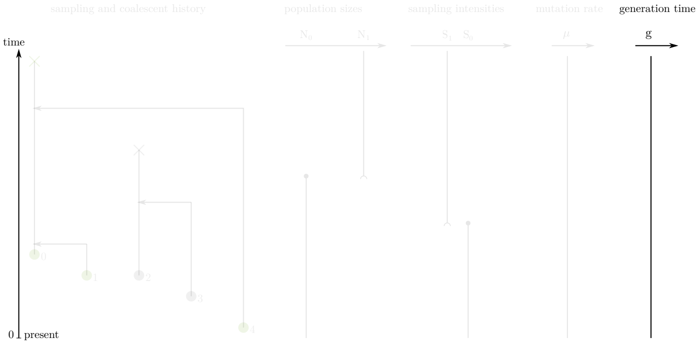
- g is constant through time,
- a priori, $g \sim \Gamma^{-1}(\alpha_g, \beta_g)$.
- (and we'll see a bit later why.)



Model assumptions

Parameters of the model – Generation time g

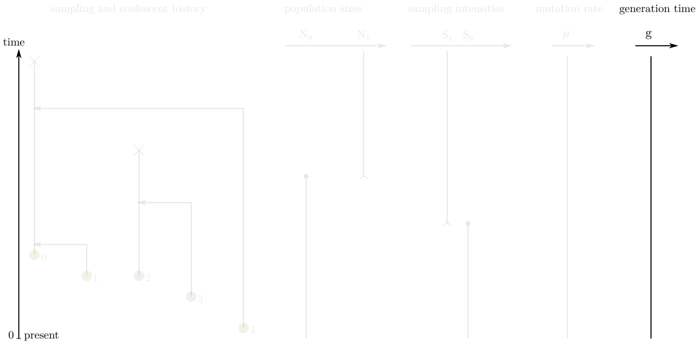
- g is constant through time,
- a priori, $g \sim \Gamma^{-1}(\alpha_g, \beta_g)$.
- (and we'll see a bit later why.)



Model assumptions

Parameters of the model – Generation time g

- g is constant through time,
- a priori, $g \sim \Gamma^{-1}(\alpha_g, \beta_g)$.
- (and we'll see a bit later why.)



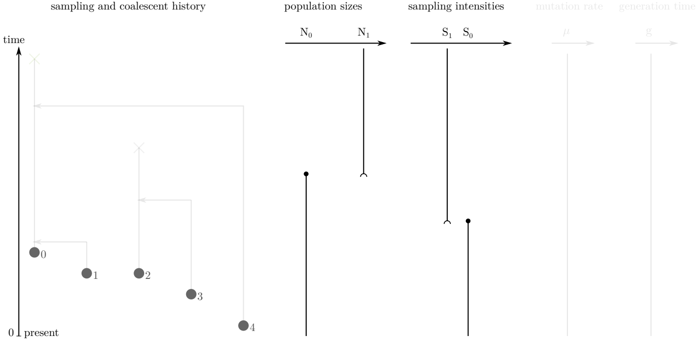
Model assumptions

Sampling and coalescent history – law of the sampling history $\mathcal{B}$

- the sampling history is given by a Poisson point process with rate

$$\lambda_t^{(b)} := S_t N_t$$

- It generates the set of ordered sampling times of our individuals $\mathcal{B} = (b_i)_{i=0}^{B-1}$.



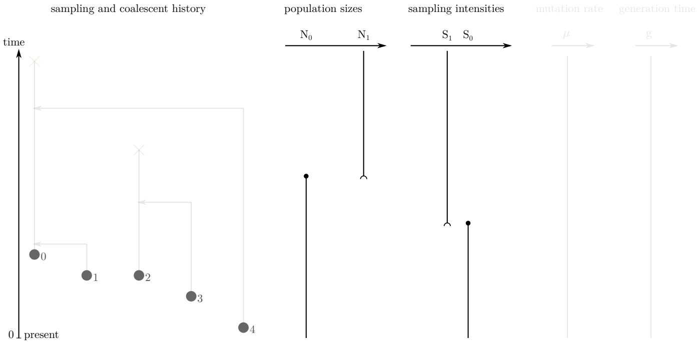
Model assumptions

Sampling and coalescent history – law of the sampling history $\mathcal{B}$

- the sampling history is given by a Poisson point process with rate

$$\lambda_t^{(b)} := S_t N_t$$

- It generates the set of ordered sampling times of our individuals $\mathcal{B} = (b_i)_{i=0}^{B-1}$.



Model assumptions

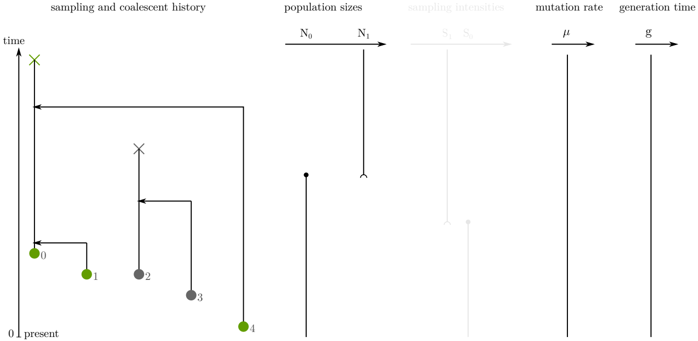
Sampling and coalescent history – law of the coalescent history $\mathcal{H}$

- While k_t lineages are alive in the process, the next coalescent/differentiation event happens with rate

$$\lambda_t^{(c)} := \binom{k_t}{2} (gN_t)^{-1}$$

$$\lambda_t^{(d)} := \mu k_t$$

- It generates a record of death events $\mathcal{H}$, and a partition of our B individuals into D alleles: $\mathcal{A}$.



Model assumptions

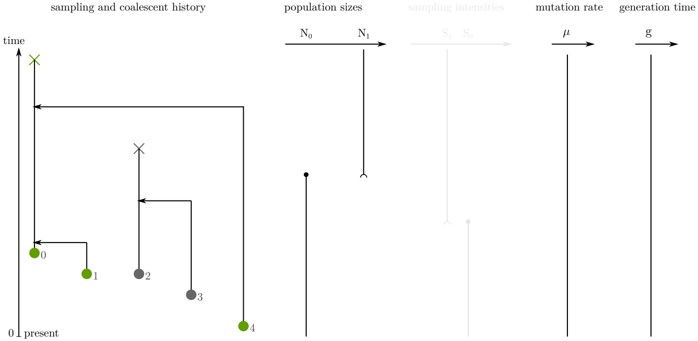
Sampling and coalescent history – law of the coalescent history $\mathcal{H}$

- While k_t lineages are alive in the process, the next coalescent/differentiation event happens with rate

$$\lambda_t^{(c)} := \binom{k_t}{2} (gN_t)^{-1}$$

$$\lambda_t^{(d)} := \mu k_t$$

- It generates a record of death events $\mathcal{H}$, and a partition of our B individuals into D alleles: $\mathcal{A}$.

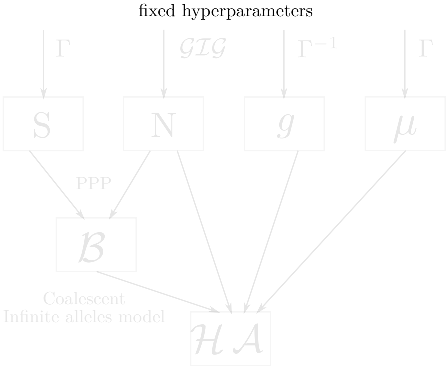


Model assumptions

Priors and summary

- The likelihood of the complete history is known and is an exponential family,

$$\mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) = \left(\prod_{i=0}^{B-1} \lambda^{(b)}_{b_i} (\lambda^{(c)}_{h_i} \mathbb{1}_{o_i \neq i} + \lambda^{(d)}_{h_i} \mathbb{1}_{o_i = i}) \right) \exp \left(- \int_0^\infty (\lambda^{(b)}_t + \lambda^{(c)}_t + \lambda^{(d)}_t) dt \right)$$

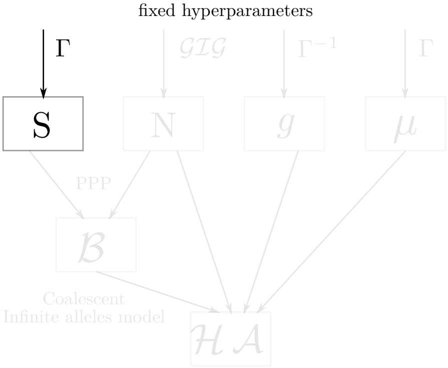


Model assumptions

Priors and summary

- The likelihood of the complete history is known and is an exponential family,

$$\mathbb{P}(\mathcal{B},\mathcal{H} \mid N,S,\mu,g) = \left(\prod_{i=0}^{B-1} \lambda_{b_i}^{(b)}(\lambda_{h_i}^{(c)}\mathbb{1}_{o_i\neq i} + \lambda_{h_i}^{(d)}\mathbb{1}_{o_i=i})\right) \exp\left(-\int_0^\infty (\lambda_t^{(b)}+\lambda_t^{(c)}+\lambda_t^{(d)})dt\right)$$

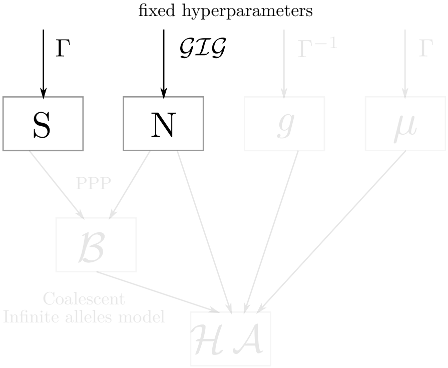


Model assumptions

Priors and summary

- The likelihood of the complete history is known and is an exponential family,

$$\mathbb{P}(\mathcal{B},\mathcal{H} \mid N,S,\mu,g) = \left(\prod_{i=0}^{B-1} \lambda_{b_i}^{(b)}(\lambda_{h_i}^{(c)}\mathbb{1}_{o_i\neq i} + \lambda_{h_i}^{(d)}\mathbb{1}_{o_i=i})\right) \exp\left(-\int_0^\infty (\lambda_t^{(b)}+\lambda_t^{(c)}+\lambda_t^{(d)})dt\right)$$

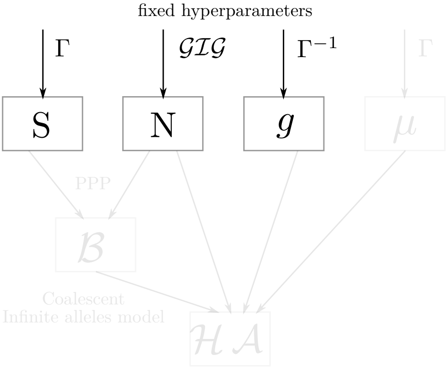


Model assumptions

Priors and summary

- The likelihood of the complete history is known and is an exponential family,

$$\mathbb{P}(\mathcal{B},\mathcal{H} \mid N,S,\mu,g) = \left(\prod_{i=0}^{B-1} \lambda_{b_i}^{(b)}(\lambda_{h_i}^{(c)}\mathbb{1}_{o_i\neq i} + \lambda_{h_i}^{(d)}\mathbb{1}_{o_i=i})\right)\exp\left(-\int_0^\infty (\lambda_t^{(b)}+\lambda_t^{(c)}+\lambda_t^{(d)})dt\right)$$

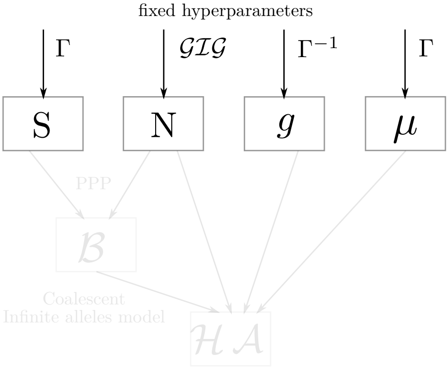


Model assumptions

Priors and summary

- The likelihood of the complete history is known and is an exponential family,

$$\mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) = \left(\prod_{i=0}^{B-1} \lambda_{b_i}^{(b)} (\lambda_{h_i}^{(c)} \mathbb{1}_{o_i \neq i} + \lambda_{h_i}^{(d)} \mathbb{1}_{o_i = i}) \right) \exp \left(- \int_0^\infty (\lambda_t^{(b)} + \lambda_t^{(c)} + \lambda_t^{(d)}) dt \right)$$

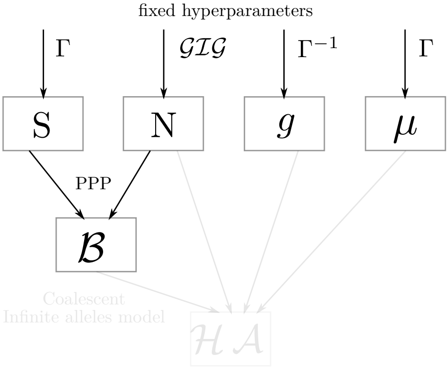


Model assumptions

Priors and summary

- The likelihood of the complete history is known and is an exponential family,

$$\mathbb{P}(\mathcal{B},\mathcal{H} \mid N,S,\mu,g) = \left(\prod_{i=0}^{B-1} \lambda_{b_i}^{(b)}(\lambda_{h_i}^{(c)}\mathbb{1}_{o_i\neq i} + \lambda_{h_i}^{(d)}\mathbb{1}_{o_i=i})\right) \exp\left(-\int_0^\infty (\lambda_t^{(b)} + \lambda_t^{(c)} + \lambda_t^{(d)})dt\right)$$

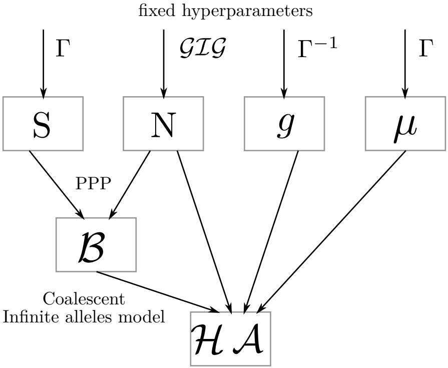


Model assumptions

Priors and summary

- The likelihood of the complete history is known and is an exponential family,

$$\mathbb{P}(\mathcal{B},\mathcal{H} \mid N,S,\mu,g) = \left(\prod_{i=0}^{B-1} \lambda_{b_i}^{(b)}(\lambda_{h_i}^{(c)}\mathbb{1}_{o_i\neq i} + \lambda_{h_i}^{(d)}\mathbb{1}_{o_i=i})\right) \exp\left(-\int_0^\infty (\lambda_t^{(b)} + \lambda_t^{(c)} + \lambda_t^{(d)})dt\right)$$



Inference method

Introduction

Model assumptions
- Parameters of the model
- Sampling and coalescent history
- Priors and summary

Inference method
- Gibbs sampling strategy
- Prior conjugacy properties for the parameters
- Data augmentation with past coalescent history
- Summary of the Gibbs sampler

Results
- Validation of the augmentation with the past coalescent history
- Validation of the MCMC by SBC
- Estimation of the running time
- Illustration on the SARS-CoV-2 dataset

Discussion
- Opportunities for future developments
- Conclusion

Inference method

Gibbs sampling strategy

Aim Infer the posterior distribution of N, S, μ, g,

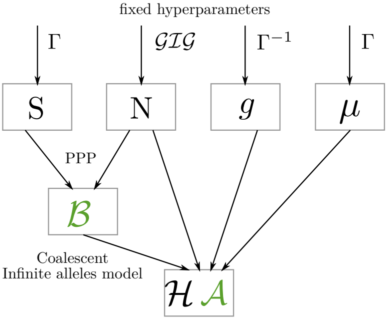
$$\mathbb{P}(N, S, \mu, g \mid \mathcal{A}, \mathcal{B}) = \int_{\mathcal{H}} \mathbb{P}(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B})$$

Strategy Design a MCMC with a Gibbs sampling approach, converging to the stationary distribution of the augmented target distribution,

$$\mathbb{P}(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B})$$

Subtargets Derive efficient ways to alternatively sample from,

$$\mathbb{P}(N_i | N_{-i}, S, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(S_i | N, S_{-i}, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mu | N, S, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(g | N, S, \mu, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mathcal{H}_i | N, S, \mu, \mathcal{A}, \mathcal{B}, \mathcal{H}_{-i})$$



Inference method

Gibbs sampling strategy

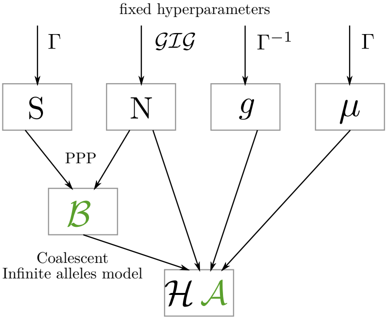
Aim Infer the posterior distribution of N, S, μ, g,

$$\mathbb{P}\left(N, S, \mu, g \mid \mathcal{A}, \mathcal{B}\right)$$

$$= \int_{\mathcal{H}} \mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$

Strategy Design a MCMC with a Gibbs sampling approach, converging to the stationary distribution of the augmented target distribution,

$$\mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$



Subtargets Derive efficient ways to alternatively sample from,

$$\mathbb{P}(N_i | N_{-i}, S, \mu, g, \mathcal{B}, \mathcal{H})$$

$$\mathbb{P}(S_i | N, S_{-i}, \mu, g, \mathcal{B}, \mathcal{H})$$

$$\mathbb{P}(\mu | N, S, g, \mathcal{B}, \mathcal{H})$$

$$\mathbb{P}(g | N, S, \mu, \mathcal{B}, \mathcal{H})$$

$$\mathbb{P}(\mathcal{H}_i | N, S, \mu, \mathcal{A}, \mathcal{B}, \mathcal{H}_{-i})$$

Inference method

Gibbs sampling strategy

Aim Infer the posterior distribution of N, S, μ, g,

$$\mathbb{P}\left(N, S, \mu, g \mid \mathcal{A}, \mathcal{B}\right)$$

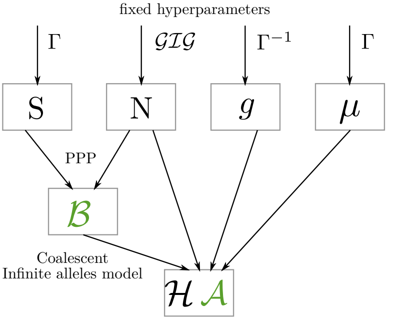
$$= \int_{\mathcal{H}} \mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$

Strategy Design a MCMC with a Gibbs sampling approach, converging to the stationary distribution of the augmented target distribution,

$$\mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$

Subtargets Derive efficient ways to alternatively sample from,

$$\mathbb{P}(N_i | N_{-i}, S, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(S_i | N, S_{-i}, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mu | N, S, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(g | N, S, \mu, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mathcal{H}_i | N, S, \mu, \mathcal{A}, \mathcal{B}, \mathcal{H}_{-i})$$



Inference method

Gibbs sampling strategy

Aim Infer the posterior distribution of N, S, μ, g,

$$\mathbb{P}(N, S, \mu, g \mid \mathcal{A}, \mathcal{B})$$

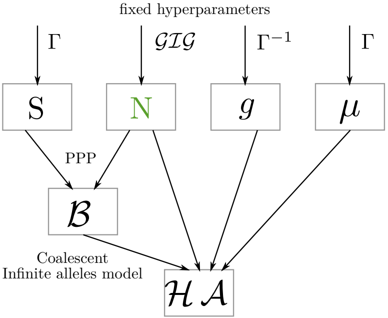
$$= \int_{\mathcal{H}} \mathbb{P}(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B})$$

Strategy Design a MCMC with a Gibbs sampling approach, converging to the stationary distribution of the augmented target distribution,

$$\mathbb{P}(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B})$$

Subtargets Derive efficient ways to alternatively sample from,

$$\mathbb{P}(N_i | N_{-i}, S, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(S_i | N, S_{-i}, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mu | N, S, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(g | N, S, \mu, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mathcal{H}_i | N, S, \mu, \mathcal{A}, \mathcal{B}, \mathcal{H}_{-i})$$



Inference method

Gibbs sampling strategy

Aim Infer the posterior distribution of N, S, μ, g,

$$\mathbb{P}\left(N, S, \mu, g \mid \mathcal{A}, \mathcal{B}\right)$$

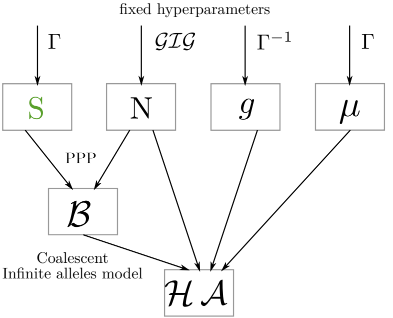
$$= \int_{\mathcal{H}} \mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$

Strategy Design a MCMC with a Gibbs sampling approach, converging to the stationary distribution of the augmented target distribution,

$$\mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$

Subtargets Derive efficient ways to alternatively sample from,

$$\mathbb{P}(N_i | N_{-i}, S, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(S_i | N, S_{-i}, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mu | N, S, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(g | N, S, \mu, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mathcal{H}_i | N, S, \mu, \mathcal{A}, \mathcal{B}, \mathcal{H}_{-i})$$



Inference method

Gibbs sampling strategy

Aim Infer the posterior distribution of N, S, μ, g,

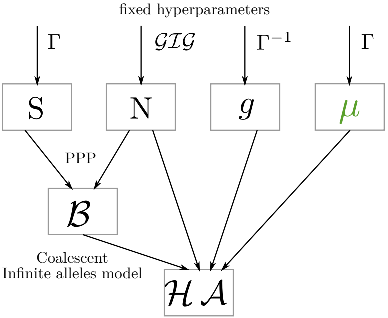
$$\mathbb{P}\left(N, S, \mu, g \mid \mathcal{A}, \mathcal{B}\right)$$
$$= \int_{\mathcal{H}} \mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$

Strategy Design a MCMC with a Gibbs sampling approach, converging to the stationary distribution of the augmented target distribution,

$$\mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$

Subtargets Derive efficient ways to alternatively sample from,

$$\mathbb{P}(N_i | N_{-i}, S, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(S_i | N, S_{-i}, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mu | N, S, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(g | N, S, \mu, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mathcal{H}_i | N, S, \mu, \mathcal{A}, \mathcal{B}, \mathcal{H}_{-i})$$



Inference method

Gibbs sampling strategy

Aim Infer the posterior distribution of N, S, μ, g,

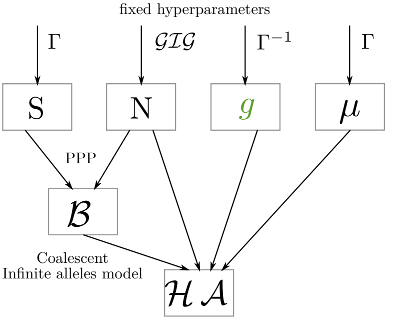
$$\mathbb{P}\left(N, S, \mu, g \mid \mathcal{A}, \mathcal{B}\right)$$
$$= \int_{\mathcal{H}} \mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$

Strategy Design a MCMC with a Gibbs sampling approach, converging to the stationary distribution of the augmented target distribution,

$$\mathbb{P}\left(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B}\right)$$

Subtargets Derive efficient ways to alternatively sample from,

$$\mathbb{P}(N_i | N_{-i}, S, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(S_i | N, S_{-i}, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mu | N, S, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(g | N, S, \mu, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mathcal{H}_i | N, S, \mu, \mathcal{A}, \mathcal{B}, \mathcal{H}_{-i})$$



Inference method

Gibbs sampling strategy

Aim Infer the posterior distribution of N, S, μ, g,

$$\mathbb{P}(N, S, \mu, g \mid \mathcal{A}, \mathcal{B})$$

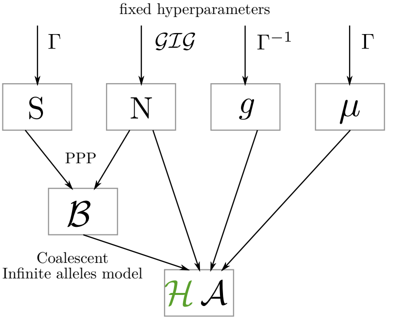
$$= \int_{\mathcal{H}} \mathbb{P}(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B})$$

Strategy Design a MCMC with a Gibbs sampling approach, converging to the stationary distribution of the augmented target distribution,

$$\mathbb{P}(N, S, \mu, g, \mathcal{H} \mid \mathcal{A}, \mathcal{B})$$

Subtargets Derive efficient ways to alternatively sample from,

$$\mathbb{P}(N_i | N_{-i}, S, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(S_i | N, S_{-i}, \mu, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mu | N, S, g, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(g | N, S, \mu, \mathcal{B}, \mathcal{H})$$
$$\mathbb{P}(\mathcal{H}_i | N, S, \mu, \mathcal{A}, \mathcal{B}, \mathcal{H}_{-i})$$



Inference method

Prior conjugacy properties for the parameters – effective population sizes N

- Assume that a priori, $N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$.
- The posterior is thus given by,

$$\begin{aligned}\mathbb{P}(N_j \mid N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(N_j)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto N_j^{\lambda-1}\ \exp\left(-\frac{1}{2}(\chi N_j^{-1} + \psi N_j)\right) \\ &\quad N_j^{B(\Delta_j^{(N)})-C(\Delta_j^{(N)})} \\ &\quad \exp\left(-N_j^{-1} g^{-1} \sum_{l=0}^{2B} \binom{k_l}{2} |\Delta_l \cap \Delta_j^{(N)}| - N_j \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)\end{aligned}$$

Conclusion : the prior and posterior of N_j are conjugate distributions, with $N_j | N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H}\ \sim$

$$\mathcal{GIG}\left(\lambda + B(\Delta_j^{(N)}) - C(\Delta_j^{(N)})\ ,\ \chi + g^{-1} \sum_{l=0}^{2B} k_l(k_l-1)|\Delta_l \cap \Delta_j^{(N)}|\ ,\ \psi + 2\sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)$$

where $C(\Delta_j^{(N)})$ and $B(\Delta_j^{(N)})$ are respectively the number of coalescent and sampling events happening over the interval $\Delta_j^{(N)}$.

Inference method

Prior conjugacy properties for the parameters – effective population sizes N

- Assume that a priori, $N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$.
- The posterior is thus given by,

$$\begin{aligned}
\mathbb{P}(N_j \mid N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(N_j)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\
&\propto N_j^{\lambda-1}\ \exp\left(-\frac{1}{2}(\chi N_j^{-1} + \psi N_j)\right) \\
&\quad N_j^{B(\Delta_j^{(N)}) - C(\Delta_j^{(N)})} \\
&\quad \exp\left(-N_j^{-1} g^{-1} \sum_{l=0}^{2B} \binom{k_l}{2} |\Delta_l \cap \Delta_j^{(N)}| - N_j \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)
\end{aligned}$$

Conclusion : the prior and posterior of N_j are conjugate distributions, with $N_j | N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H} \ \sim$

$$\mathcal{GIG}\left(\lambda + B(\Delta_j^{(N)}) - C(\Delta_j^{(N)})\ ,\ \chi + g^{-1} \sum_{l=0}^{2B} k_l(k_l - 1)|\Delta_l \cap \Delta_j^{(N)}|\ ,\ \psi + 2 \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)$$

where $C(\Delta_j^{(N)})$ and $B(\Delta_j^{(N)})$ are respectively the number of coalescent and sampling events happening over the interval $\Delta_j^{(N)}$.

Inference method

Prior conjugacy properties for the parameters – effective population sizes N

- Assume that a priori, $N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$.
- The posterior is thus given by,

$$\mathbb{P}(N_j \mid N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H}) \propto \mathbb{P}(N_j)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g)$$

$$\propto N_j^{\lambda-1}\ \exp\left(-\frac{1}{2}(\chi N_j^{-1} + \psi N_j)\right)$$

$$N_j^{B(\Delta_j^{(N)})-C(\Delta_j^{(N)})}$$

$$\exp\left(-N_j^{-1} g^{-1} \sum_{l=0}^{2B} \binom{k_l}{2} |\Delta_l \cap \Delta_j^{(N)}| - N_j \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)$$

Conclusion : the prior and posterior of N_j are conjugate distributions, with $N_j | N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H} \sim$

$$\mathcal{GIG}\left(\lambda + B(\Delta_j^{(N)}) - C(\Delta_j^{(N)})\ ,\ \chi + g^{-1} \sum_{l=0}^{2B} k_l(k_l - 1)|\Delta_l \cap \Delta_j^{(N)}|\ ,\ \psi + 2 \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)$$

where $C(\Delta_j^{(N)})$ and $B(\Delta_j^{(N)})$ are respectively the number of coalescent and sampling events happening over the interval $\Delta_j^{(N)}$.

Inference method

Prior conjugacy properties for the parameters – effective population sizes N

- Assume that a priori, $N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$.
- The posterior is thus given by,

$$\begin{aligned}\mathbb{P}(N_j \mid N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(N_j)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto N_j^{\lambda-1}\ \exp\left(-\frac{1}{2}(\chi N_j^{-1} + \psi N_j)\right) \\ &\quad N_j^{B(\Delta_j^{(N)})-C(\Delta_j^{(N)})} \\ &\quad \exp\left(-N_j^{-1} g^{-1} \sum_{l=0}^{2B} \binom{k_l}{2} |\Delta_l \cap \Delta_j^{(N)}| - N_j \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)\end{aligned}$$

Conclusion : the prior and posterior of N_j are conjugate distributions, with $N_j | N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H} \ \sim$

$$\mathcal{GIG}\left(\lambda + B(\Delta_j^{(N)}) - C(\Delta_j^{(N)})\ ,\ \chi + g^{-1} \sum_{l=0}^{2B} k_l(k_l-1)|\Delta_l \cap \Delta_j^{(N)}|\ ,\ \psi + 2\sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)$$

where $C(\Delta_j^{(N)})$ and $B(\Delta_j^{(N)})$ are respectively the number of coalescent and sampling events happening over the interval $\Delta_j^{(N)}$.

Inference method

Prior conjugacy properties for the parameters – effective population sizes N

- Assume that a priori, $N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$.
- The posterior is thus given by,

$$\begin{aligned}\mathbb{P}(N_j \mid N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(N_j)\, \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto N_j^{\lambda - 1} \, \exp\left(-\frac{1}{2}(\chi N_j^{-1} + \psi N_j)\right) \\ &\quad N_j^{B(\Delta_j^{(N)}) - C(\Delta_j^{(N)})} \\ &\quad \exp\left(-N_j^{-1} g^{-1} \sum_{l=0}^{2B} \binom{k_l}{2} |\Delta_l \cap \Delta_j^{(N)}| - N_j \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)\end{aligned}$$

Conclusion : the prior and posterior of N_j are conjugate distributions, with $N_j | N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H} \ \sim$

$$\mathcal{GIG}\left(\lambda + B(\Delta_j^{(N)}) - C(\Delta_j^{(N)}) \ , \ \chi + g^{-1} \sum_{l=0}^{2B} k_l(k_l - 1)|\Delta_l \cap \Delta_j^{(N)}| \ , \ \psi + 2 \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)$$

where $C(\Delta_j^{(N)})$ and $B(\Delta_j^{(N)})$ are respectively the number of coalescent and sampling events happening over the interval $\Delta_j^{(N)}$.

Inference method

Prior conjugacy properties for the parameters – effective population sizes N

- Assume that a priori, $N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$.
- The posterior is thus given by,

$$\begin{aligned}\mathbb{P}(N_j \mid N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(N_j)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto N_j^{\lambda-1}\ \exp\left(-\frac{1}{2}(\chi N_j^{-1} + \psi N_j)\right) \\ &\quad N_j^{B(\Delta_j^{(N)}) - C(\Delta_j^{(N)})} \\ &\quad \exp\left(-N_j^{-1} g^{-1} \sum_{l=0}^{2B} \binom{k_l}{2} |\Delta_l \cap \Delta_j^{(N)}| - N_j \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)\end{aligned}$$

Conclusion : the prior and posterior of N_j are conjugate distributions, with $N_j | N_{-j}, S, \mu, g, \mathcal{B}, \mathcal{H} \ \sim$

$$\mathcal{GIG}\left(\lambda + B(\Delta_j^{(N)}) - C(\Delta_j^{(N)})\ ,\ \chi + g^{-1} \sum_{l=0}^{2B} k_l(k_l - 1)|\Delta_l \cap \Delta_j^{(N)}|\ ,\ \psi + 2 \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}|\right)$$

where $C(\Delta_j^{(N)})$ and $B(\Delta_j^{(N)})$ are respectively the number of coalescent and sampling events happening over the interval $\Delta_j^{(N)}$.

Inference method

Prior conjugacy properties for the parameters – sampling intensities S

- Assume that a priori, $S_j \sim \Gamma(\alpha, \beta)$.
- Its posterior is thus given by,

$$\begin{aligned}\mathbb{P}(S_j \mid N, S_{-j}, \mu, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(S_j)\,\mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto S_j^{\alpha-1}\ \exp\left(-\beta S_j\right) \\ &\quad S_j^{B(\Delta_j^{(S)})}\ \exp\left(-S_j \sum_{k=0}^{p-1} N_k |\Delta_k^{(N)} \cap \Delta_j^{(S)}|\right)\end{aligned}$$

Conclusion : the prior and posterior of S_j are conjugate distributions, with,

$$S_j \mid N, S_{-j}, \mu, g, \mathcal{B}, \mathcal{H}\ \sim\ \Gamma\left(\alpha + B(\Delta_j^{(S)})\ ,\ \beta + \sum_{k=0}^{p-1} N_k |\Delta_k^{(N)} \cap \Delta_j^{(S)}|\right) \quad .$$

where $B(\Delta_j^{(S)})$ is the number of birth events happening over interval $\Delta_j^{(S)}$.

Inference method

Prior conjugacy properties for the parameters – sampling intensities S

- Assume that a priori, $S_j \sim \Gamma(\alpha, \beta)$.
- Its posterior is thus given by,

$$\begin{aligned}\mathbb{P}(S_j \mid N, S_{-j}, \mu, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(S_j)\; \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto S_j^{\alpha-1} \; \exp\left(-\beta S_j\right) \\ &\quad\; S_j^{B(\Delta_j^{(S)})} \; \exp\left(-S_j \sum_{k=0}^{p-1} N_k |\Delta_k^{(N)} \cap \Delta_j^{(S)}|\right)\end{aligned}$$

Conclusion : the prior and posterior of S_j are conjugate distributions, with,

$$S_j \mid N, S_{-j}, \mu, g, \mathcal{B}, \mathcal{H} \;\sim\; \Gamma\left(\alpha + B(\Delta_j^{(S)}) \;,\; \beta + \sum_{k=0}^{p-1} N_k |\Delta_k^{(N)} \cap \Delta_j^{(S)}|\right) \quad .$$

where $B(\Delta_j^{(S)})$ is the number of birth events happening over interval $\Delta_j^{(S)}$.

Inference method

Prior conjugacy properties for the parameters – sampling intensities S

- Assume that a priori, $S_j \sim \Gamma(\alpha, \beta)$.
- Its posterior is thus given by,

$$\begin{aligned}\mathbb{P}(S_j \mid N, S_{-j}, \mu, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(S_j)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto S_j^{\alpha-1}\ \exp\left(-\beta S_j\right) \\ &\quad S_j^{B(\Delta_j^{(S)})}\ \exp\left(-S_j \sum_{k=0}^{p-1} N_k |\Delta_k^{(N)} \cap \Delta_j^{(S)}|\right)\end{aligned}$$

Conclusion : the prior and posterior of S_j are conjugate distributions, with,

$$S_j \mid N, S_{-j}, \mu, g, \mathcal{B}, \mathcal{H} \ \sim\ \Gamma\left(\alpha + B(\Delta_j^{(S)})\ ,\ \beta + \sum_{k=0}^{p-1} N_k |\Delta_k^{(N)} \cap \Delta_j^{(S)}|\right) \quad .$$

where $B(\Delta_j^{(S)})$ is the number of birth events happening over interval $\Delta_j^{(S)}$.

Inference method

Prior conjugacy properties for the parameters – mutation rate μ

- Assume that, a priori, $\mu \sim \Gamma(\alpha, \beta)$.
- Its posterior is given by,

$$\begin{aligned}\mathbb{P}(\mu \mid N, S, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(\mu)\, \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto \mu^{\alpha - 1} \ \exp\left(-\beta\mu\right) \\ &\quad \mu^{D} \ \exp\left(-\mu \sum_{l=0}^{2B} k_l |\Delta_l|\right)\end{aligned}$$

Conclusion : the prior and posterior of μ are conjugate distributions, with,

$$\mu \mid N, S, g, \mathcal{B}, \mathcal{H} \ \sim \ \Gamma\left(\alpha + D \ , \ \beta + \sum_{l=0}^{2B} k_l |\Delta_l|\right)$$

where D is the total number of alleles.

Inference method

Prior conjugacy properties for the parameters – mutation rate μ

- Assume that, a priori, $\mu \sim \Gamma(\alpha, \beta)$.
- Its posterior is given by,

$$\begin{aligned}\mathbb{P}(\mu \mid N, S, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(\mu)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto \mu^{\alpha-1}\ \exp\left(-\beta\mu\right) \\ &\quad\ \mu^{D}\ \exp\left(-\mu \sum_{l=0}^{2B} k_l |\Delta_l|\right)\end{aligned}$$

Conclusion : the prior and posterior of μ are conjugate distributions, with,

$$\mu \mid N, S, g, \mathcal{B}, \mathcal{H}\ \sim\ \Gamma\left(\alpha + D\ ,\ \beta + \sum_{l=0}^{2B} k_l |\Delta_l|\right)$$

where D is the total number of alleles.

Inference method

Prior conjugacy properties for the parameters – mutation rate μ

- Assume that, a priori, $\mu \sim \Gamma(\alpha, \beta)$.
- Its posterior is given by,

$$\begin{aligned}\mathbb{P}(\mu \mid N, S, g, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(\mu)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto \mu^{\alpha-1}\ \exp\left(-\beta\mu\right) \\ &\quad\ \mu^{D}\ \exp\left(-\mu \sum_{l=0}^{2B} k_l |\Delta_l|\right)\end{aligned}$$

Conclusion : the prior and posterior of μ are conjugate distributions, with,

$$\mu \mid N, S, g, \mathcal{B}, \mathcal{H}\ \sim\ \Gamma\left(\alpha + D\ ,\ \beta + \sum_{l=0}^{2B} k_l |\Delta_l|\right)$$

where D is the total number of alleles.

Inference method

Prior conjugacy properties for the parameters – generation time g

- Assume that a priori, $g \sim \Gamma^{-1}(\alpha, \beta)$.
- Its posterior is given by,

$$\begin{aligned}\mathbb{P}(g \mid N, S, \mu, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(g)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g)\\ &\propto g^{-(\alpha+1)}\ \exp\left(-\beta g^{-1}\right)\\ &\quad g^{-(B-D)}\ \exp\left(-g^{-1}\sum_{l=0}^{2B}\sum_{j=0}^{p-1}\binom{k_l}{2} N_j^{-1}|\Delta_l \cap \Delta_j^{(N)}|\right)\end{aligned}$$

Conclusion : the prior and posterior of g are conjugate distributions, with,

$$g \mid N, S, \mu, \mathcal{B}, \mathcal{H} \ \sim\ \Gamma^{-1}\left(\alpha + B - D\ ,\ \beta + \sum_{l=0}^{2B}\sum_{j=0}^{p-1}\binom{k_l}{2} N_j^{-1}|\Delta_l \cap \Delta_j^{(N)}|\right)$$

Inference method

Prior conjugacy properties for the parameters – generation time g

- Assume that a priori, $g \sim \Gamma^{-1}(\alpha, \beta)$.
- Its posterior is given by,

$$\begin{aligned}\mathbb{P}(g \mid N, S, \mu, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(g)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto g^{-(\alpha+1)}\ \exp\left(-\beta g^{-1}\right) \\ &\quad g^{-(B-D)}\ \exp\left(-g^{-1} \sum_{l=0}^{2B} \sum_{j=0}^{p-1} \binom{k_l}{2} N_j^{-1} |\Delta_l \cap \Delta_j^{(N)}|\right)\end{aligned}$$

Conclusion : the prior and posterior of g are conjugate distributions, with,

$$g \mid N, S, \mu, \mathcal{B}, \mathcal{H}\ \sim\ \Gamma^{-1}\left(\alpha + B - D\ ,\ \beta + \sum_{l=0}^{2B} \sum_{j=0}^{p-1} \binom{k_l}{2} N_j^{-1} |\Delta_l \cap \Delta_j^{(N)}|\right)$$

Inference method

Prior conjugacy properties for the parameters – generation time g

- Assume that a priori, $g \sim \Gamma^{-1}(\alpha, \beta)$.
- Its posterior is given by,

$$\begin{aligned}\mathbb{P}(g \mid N, S, \mu, \mathcal{B}, \mathcal{H}) &\propto \mathbb{P}(g)\ \mathbb{P}(\mathcal{B}, \mathcal{H} \mid N, S, \mu, g) \\ &\propto g^{-(\alpha+1)}\ \exp\left(-\beta g^{-1}\right) \\ &\quad g^{-(B-D)}\ \exp\left(-g^{-1} \sum_{l=0}^{2B} \sum_{j=0}^{p-1} \binom{k_l}{2} N_j^{-1} |\Delta_l \cap \Delta_j^{(N)}|\right)\end{aligned}$$

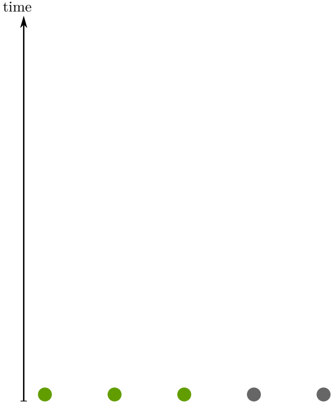
Conclusion : the prior and posterior of g are conjugate distributions, with,

$$g \mid N, S, \mu, \mathcal{B}, \mathcal{H} \ \sim\ \Gamma^{-1}\left(\alpha + B - D\ ,\ \beta + \sum_{l=0}^{2B} \sum_{j=0}^{p-1} \binom{k_l}{2} N_j^{-1} |\Delta_l \cap \Delta_j^{(N)}|\right)$$

Inference method

Data augmentation with past coalescent history when all samples are taken at present

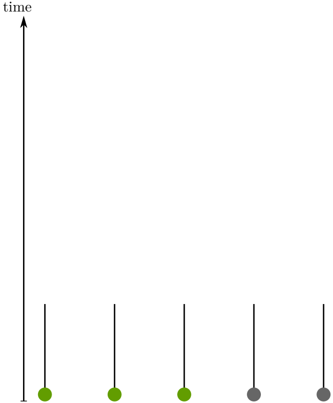
1. simulate $T_k \sim \mathcal{E}(k(\theta + k - 1)/2)$,
2. choose one of the k living lineages uniformly at random and,
   if it is a singleton in a_k, there is a mutation and this lineage is killed,
   if not, choose uniformly another lineage in the same allele and make them coalesce.



Inference method

Data augmentation with past coalescent history when all samples are taken at present

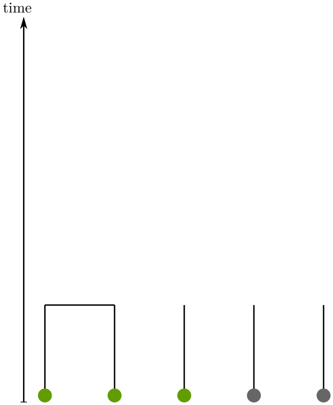
1. simulate $T_k \sim \mathcal{E}(k(\theta + k - 1)/2)$,
2. choose one of the k living lineages uniformly at random and,
   if it is a singleton in a_k, there is a mutation and this lineage is killed,
   if not, choose uniformly another lineage in the same allele and make them coalesce.



Inference method

Data augmentation with past coalescent history when all samples are taken at present

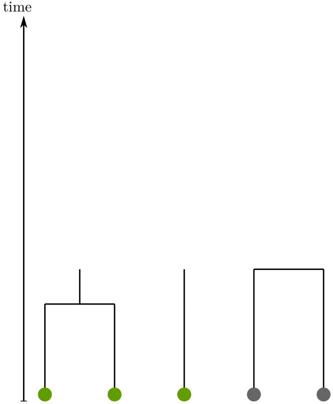
1. simulate $T_k \sim \mathcal{E}(k(\theta + k - 1)/2)$,
2. choose one of the k living lineages uniformly at random and,
   if it is a singleton in a_k, there is a mutation and this lineage is killed,
   if not, choose uniformly another lineage in the same allele and make them coalesce.



Inference method

Data augmentation with past coalescent history when all samples are taken at present

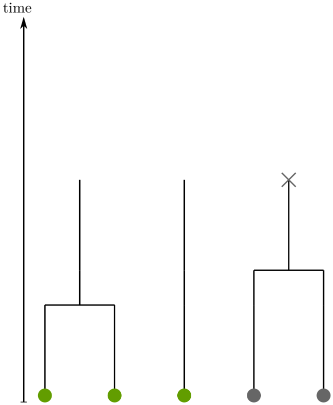
1. simulate $T_k \sim \mathcal{E}(k(\theta + k - 1)/2)$,
2. choose one of the k living lineages uniformly at random and,
   if it is a singleton in a_k, there is a mutation and this lineage is killed,
   if not, choose uniformly another lineage in the same allele and make them coalesce.



Inference method

Data augmentation with past coalescent history when all samples are taken at present

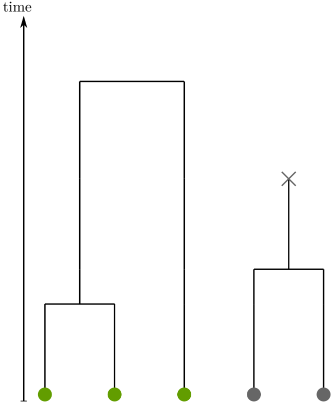
1. simulate $T_k \sim \mathcal{E}(k(\theta + k - 1)/2)$,
2. choose one of the k living lineages uniformly at random and,
   if it is a singleton in a_k, there is a mutation and this lineage is killed,
   if not, choose uniformly another lineage in the same allele and make them coalesce.



Inference method

Data augmentation with past coalescent history when all samples are taken at present

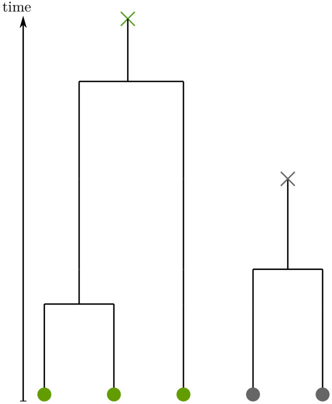
1. simulate $T_k \sim \mathcal{E}(k(\theta + k - 1)/2)$,
2. choose one of the k living lineages uniformly at random and,
   if it is a singleton in a_k, there is a mutation and this lineage is killed,
   if not, choose uniformly another lineage in the same allele and make them coalesce.



Inference method

Data augmentation with past coalescent history when all samples are taken at present

1. simulate $T_k \sim \mathcal{E}(k(\theta + k - 1)/2)$,
2. choose one of the k living lineages uniformly at random and,
   if it is a singleton in a_k, there is a mutation and this lineage is killed,
   if not, choose uniformly another lineage in the same allele and make them coalesce.



Inference method

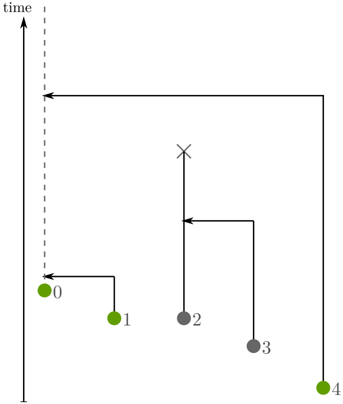
Data augmentation with past coalescent history with heterochronous sampling

- so far I didn’t succeed in finding such an elegant simulation of the past,
- but I can compute the death time of one focal individual conditioned on everything else.

Inference method

Data augmentation with past coalescent history with heterochronous sampling

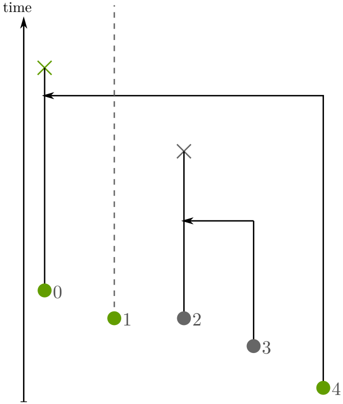
- so far I didn't succeed in finding such an elegant simulation of the past,
- but I can compute the death time of one focal individual conditioned on everything else.



Inference method

Data augmentation with past coalescent history with heterochronous sampling

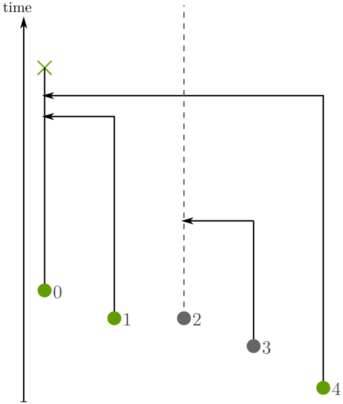
- so far I didn't succeed in finding such an elegant simulation of the past,
- but I can compute the death time of one focal individual conditioned on everything else.



Inference method

Data augmentation with past coalescent history with heterochronous sampling

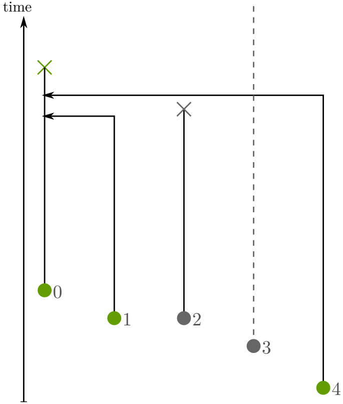
- so far I didn't succeed in finding such an elegant simulation of the past,
- but I can compute the death time of one focal individual conditioned on everything else.



Inference method

Data augmentation with past coalescent history with heterochronous sampling

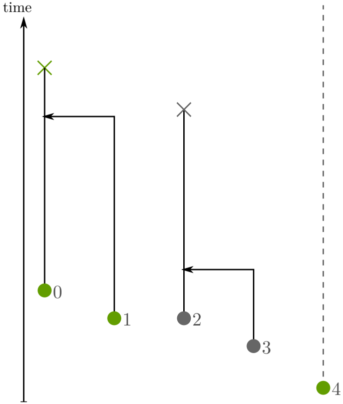
- so far I didn't succeed in finding such an elegant simulation of the past,
- but I can compute the death time of one focal individual conditioned on everything else.



Inference method

Data augmentation with past coalescent history with heterochronous sampling

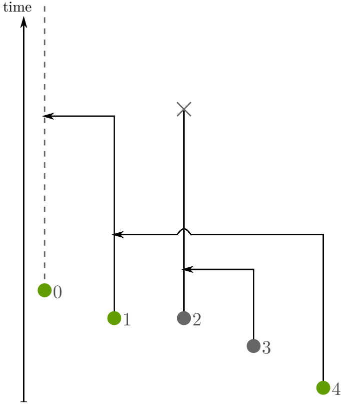
- so far I didn't succeed in finding such an elegant simulation of the past,
- but I can compute the death time of one focal individual conditioned on everything else.



Inference method

Data augmentation with past coalescent history with heterochronous sampling

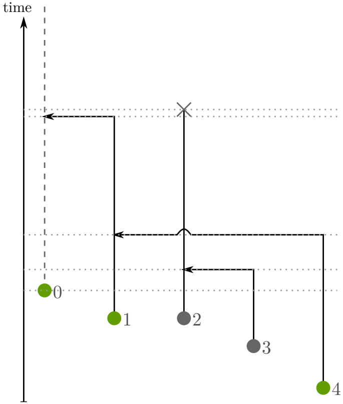
- so far I didn't succeed in finding such an elegant simulation of the past,
- but I can compute the death time of one focal individual conditioned on everything else.



Inference method

Data augmentation with past coalescent history with heterochronous sampling

- so far I didn't succeed in finding such an elegant simulation of the past,
- but I can compute the death time of one focal individual conditioned on everything else.



Inference method

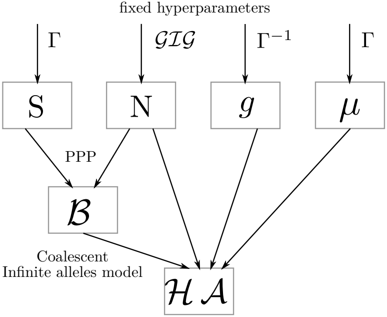
Summary of the Gibbs sampler

Initialization:

- Fix $\forall i, H_i = b_i$ and $O_i = \min\{j \in a_i\}$.
- Draw $\forall j, N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$,
- Draw $\forall j, S_j \sim \Gamma(\alpha_S, \beta_S)$,
- Draw $\mu \sim \Gamma(\alpha_\mu, \beta_\mu)$,
- Draw $g \sim \Gamma^{-1}(\alpha_g, \beta_g)$,

One step in the chain:

- Draw each H_i, O_i in turn using $H_{-i}, O_{-i}, N, S, \mu, g$,
- Draw each N_j using its $\mathcal{GIG}$ posterior,
- Draw each S_j using its Γ posterior,
- Draw μ using its Γ posterior,
- Draw g using its Γ^{-1} posterior.



Inference method

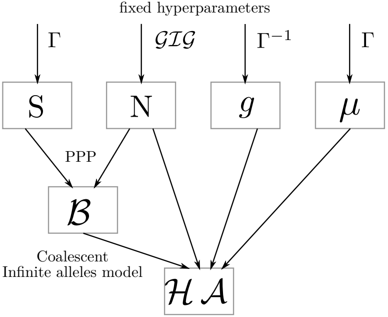
Summary of the Gibbs sampler

Initialization:

- Fix $\forall i, H_i = b_i$ and $O_i = \min\{j \in a_i\}$.
- Draw $\forall j, N_j \sim \mathcal{GIG}(\lambda, \chi, \psi)$,
- Draw $\forall j, S_j \sim \Gamma(\alpha_S, \beta_S)$,
- Draw $\mu \sim \Gamma(\alpha_\mu, \beta_\mu)$,
- Draw $g \sim \Gamma^{-1}(\alpha_g, \beta_g)$,

One step in the chain:

- Draw each H_i, O_i in turn using $H_{-i}, O_{-i}, N, S, \mu, g$,
- Draw each N_j using its $\mathcal{GIG}$ posterior,
- Draw each S_j using its Γ posterior,
- Draw μ using its Γ posterior,
- Draw g using its Γ^{-1} posterior.



Results

Introduction

Model assumptions
- Parameters of the model
- Sampling and coalescent history
- Priors and summary

Inference method
- Gibbs sampling strategy
- Prior conjugacy properties for the parameters
- Data augmentation with past coalescent history
- Summary of the Gibbs sampler

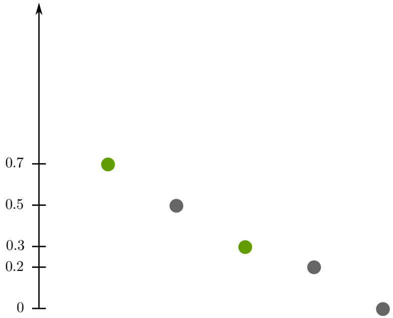
Results
- Validation of the augmentation with the past coalescent history
- Validation of the MCMC by SBC
- Estimation of the running time
- Illustration on the SARS-CoV-2 dataset

Discussion
- Opportunities for future developments
- Conclusion

Results

Validation of the augmentation with the past coalescent history

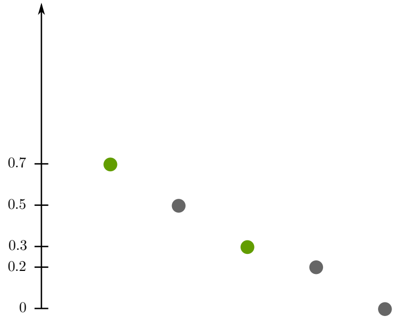
- Fix a very small dataset $\mathcal{A}, \mathcal{B}$ and all parameters.
- Wrap up the data augmentation in a minimalist Gibbs sampler without parameter updates.
- Compare $\mathcal{H}$ to what is obtained by naive rejection sampling on 10^4 samples.



Results

Validation of the augmentation with the past coalescent history

- Fix a very small dataset $\mathcal{A}, \mathcal{B}$ and all parameters.
- Wrap up the data augmentation in a minimalist Gibbs sampler without parameter updates.
- Compare $\mathcal{H}$ to what is obtained by naive rejection sampling on 10^4 samples.

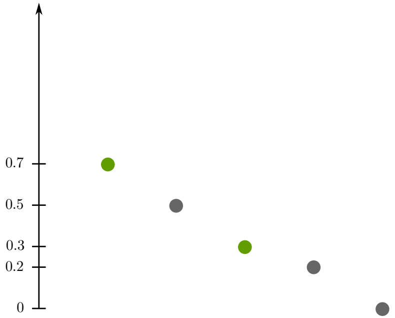


$N = 10$
$\mu = 1.5$
$g = 0.1$

Results

Validation of the augmentation with the past coalescent history

- Fix a very small dataset $\mathcal{A}, \mathcal{B}$ and all parameters.
- Wrap up the data augmentation in a minimalist Gibbs sampler without parameter updates.
- Compare $\mathcal{H}$ to what is obtained by naive rejection sampling on 10^4 samples.



$N = 10$

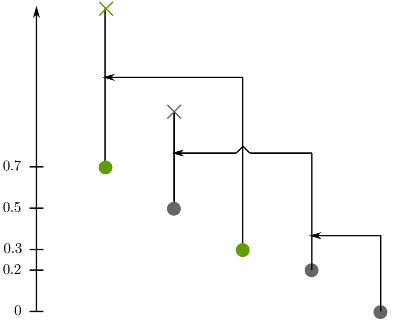
$\mu = 1.5$

$g = 0.1$

Results

Validation of the augmentation with the past coalescent history

- Fix a very small dataset $\mathcal{A}, \mathcal{B}$ and all parameters.
- Wrap up the data augmentation in a minimalist Gibbs sampler without parameter updates.
- Compare $\mathcal{H}$ to what is obtained by naive rejection sampling on 10^4 samples.

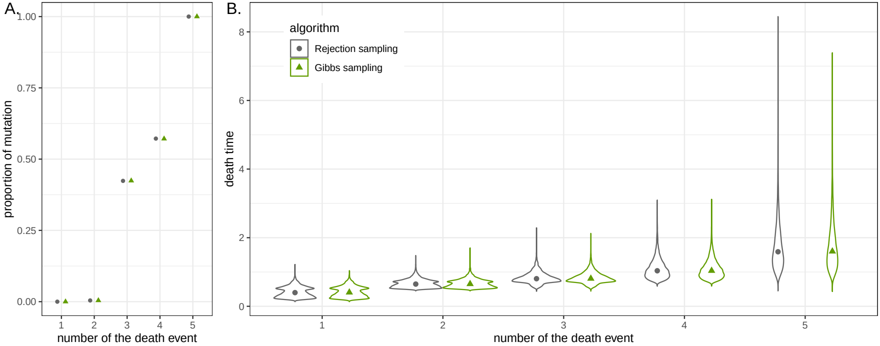


$N = 10$
$\mu = 1.5$
$g = 0.1$

Results

Validation of the augmentation with the past coalescent history

- Fix a very small dataset $\mathcal{A}, \mathcal{B}$ and all parameters.
- Wrap up the data augmentation in a minimalist Gibbs sampler without parameter updates.
- Compare $\mathcal{H}$ to what is obtained by naive rejection sampling on 10^4 samples.



Results

Validation of the MCMC by SBC – following Talts et al. 2018

- Fix a set of hyperparameters.
- Sample 10^4 complete datasets $N, S, \mu, g, \mathcal{B}, \mathcal{A}$.
- On each dataset, compute the posterior of $N, S, \mu, g \mid \mathcal{A}, \mathcal{B}$.
- Check that the quantiles of the prior and posterior are the same.
- Check that the histogram of rank statistics is uniform
  – where the rank statistic is the nb of samples from the posterior being less than the true value.

Results

Validation of the MCMC by SBC – following Talts et al. 2018

- Fix a set of hyperparameters.
- Sample 10^4 complete datasets $N, S, \mu, g, \mathcal{B}, \mathcal{A}$.
- On each dataset, compute the posterior of $N, S, \mu, g \mid \mathcal{A}, \mathcal{B}$.
- Check that the quantiles of the prior and posterior are the same.
- Check that the histogram of rank statistics is uniform
  – where the rank statistic is the nb of samples from the posterior being less than the true value.

Results

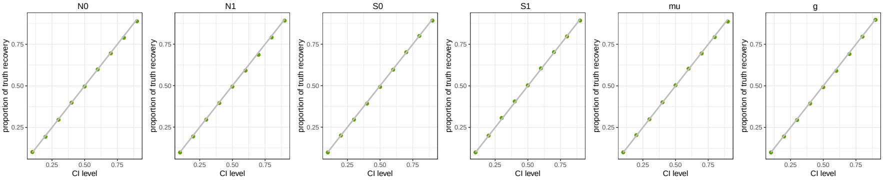
Validation of the MCMC by SBC – following Talts et al. 2018

- Fix a set of hyperparameters.
- Sample 10^4 complete datasets $N, S, \mu, g, \mathcal{B}, \mathcal{A}$.
- On each dataset, compute the posterior of $N, S, \mu, g \mid \mathcal{A}, \mathcal{B}$.
- Check that the quantiles of the prior and posterior are the same.
- Check that the histogram of rank statistics is uniform
  – where the rank statistic is the nb of samples from the posterior being less than the true value.

Results

Validation of the MCMC by SBC – following Talts et al. 2018

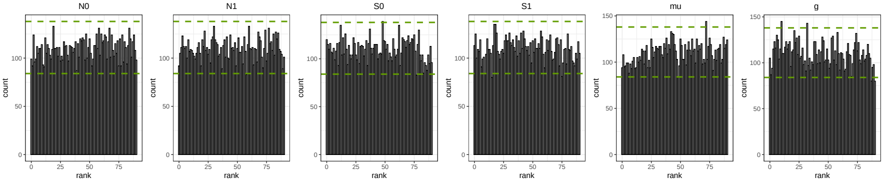
- Fix a set of hyperparameters.
- Sample 10^4 complete datasets $N, S, \mu, g, \mathcal{B}, \mathcal{A}$.
- On each dataset, compute the posterior of $N, S, \mu, g \mid \mathcal{A}, \mathcal{B}$.
- Check that the quantiles of the prior and posterior are the same.
- Check that the histogram of rank statistics is uniform
  – where the rank statistic is the nb of samples from the posterior being less than the true value.



Results

Validation of the MCMC by SBC – following Talts et al. 2018

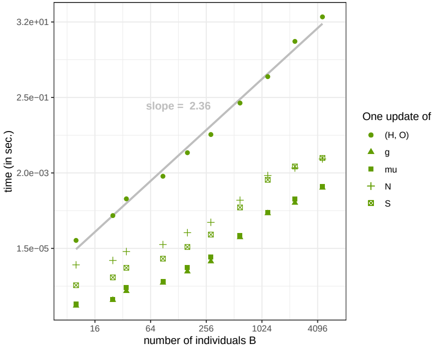
- Fix a set of hyperparameters.
- Sample 10^4 complete datasets $N, S, \mu, g, \mathcal{B}, \mathcal{A}$.
- On each dataset, compute the posterior of $N, S, \mu, g \mid \mathcal{A}, \mathcal{B}$.
- Check that the quantiles of the prior and posterior are the same.
- Check that the histogram of rank statistics is uniform
  – where the rank statistic is the nb of samples from the posterior being less than the true value.

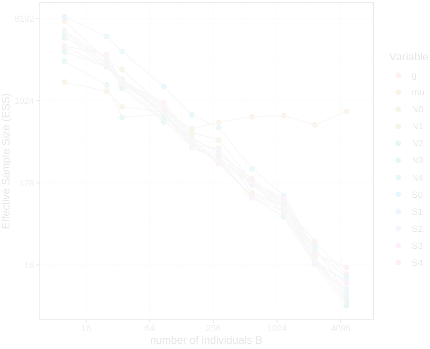


Results

Estimation of the running time

- The update of $\mathcal{H}$ runs in $O(B^2)$.
- One also needs to run the MCMC long enough to get reasonable ESS values.
- This first naive implementation seems reasonable to be used on up to $\sim 10^4$ samples.

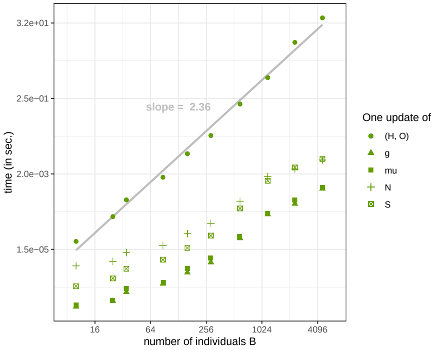


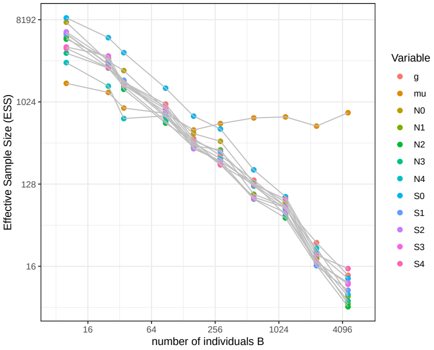


Results

Estimation of the running time

- The update of $\mathcal{H}$ runs in $O(B^2)$.
- One also needs to run the MCMC long enough to get reasonable ESS values.
- This first naive implementation seems reasonable to be used on up to $\sim 10^4$ samples.

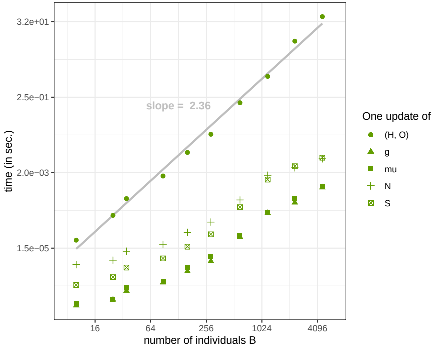


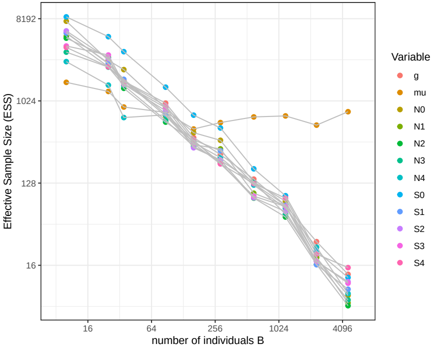


Results

Estimation of the running time

- The update of $\mathcal{H}$ runs in $O(B^2)$.
- One also needs to run the MCMC long enough to get reasonable ESS values.
- This first naive implementation seems reasonable to be used on up to $\sim 10^4$ samples.





Results

Illustration on the SARS-CoV-2 dataset

- Sequences from GISAID until 1st of June 2020.

  CH 1284 genomes in 627 alleles,
  DE 1673 genomes in 886 alleles,
  FR 1919 genomes in 1166 alleles,
  IT 1314 genomes in 750 alleles.

- Fixed $\mu = 0.065$ mutations per genome per d, $g = 5$ d and timeline with 5 periods of 4 weeks each.
- To fix hyperparameters, imagine a period with few data,
  Make a guess on the order of magnitude of $N \sim 10^4$ and $S \sim 4 \times 10^{-5}$.
  One would then observe 4 birth on 10 days, and this fixes $\lambda, \chi, \psi, \alpha_S, \beta_S$.

Results

Illustration on the SARS-CoV-2 dataset

- Sequences from GISAID until 1st of June 2020.

  CH 1284 genomes in 627 alleles,
  DE 1673 genomes in 886 alleles,
  FR 1919 genomes in 1166 alleles,
  IT 1314 genomes in 750 alleles.

- Fixed $\mu = 0.065$ mutations per genome per d, $g = 5$ d and timeline with 5 periods of 4 weeks each.

- To fix hyperparameters, imagine a period with few data,
  Make a guess on the order of magnitude of $N \sim 10^4$ and $S \sim 4 \times 10^{-5}$.
  One would then observe 4 birth on 10 days, and this fixes $\lambda, \chi, \psi, \alpha_S, \beta_S$.

Results

Illustration on the SARS-CoV-2 dataset

- Sequences from GISAID until 1st of June 2020.

  CH 1284 genomes in 627 alleles,
  DE 1673 genomes in 886 alleles,
  FR 1919 genomes in 1166 alleles,
  IT 1314 genomes in 750 alleles.

- Fixed $\mu = 0.065$ mutations per genome per d, $g = 5$ d and timeline with 5 periods of 4 weeks each.
- To fix hyperparameters, imagine a period with few data,
  Make a guess on the order of magnitude of $N \sim 10^4$ and $S \sim 4 \times 10^{-5}$.
  One would then observe 4 birth on 10 days, and this fixes $\lambda, \chi, \psi, \alpha_S, \beta_S$.

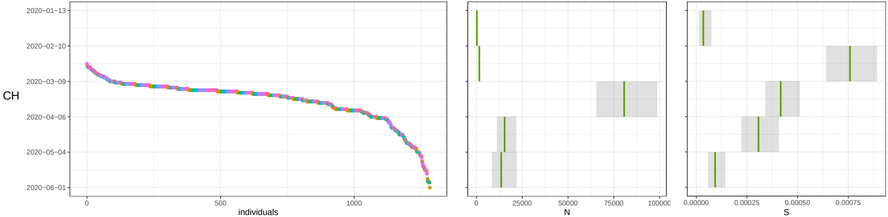
Results

Illustration on the SARS-CoV-2 dataset

- Sequences from GISAID until 1st of June 2020.

  CH 1284 genomes in 627 alleles,
  DE 1673 genomes in 886 alleles,
  FR 1919 genomes in 1166 alleles,
  IT 1314 genomes in 750 alleles.

- Fixed $\mu = 0.065$ mutations per genome per d, $g = 5$ d and timeline with 5 periods of 4 weeks each.
- To fix hyperparameters, imagine a period with few data,
  Make a guess on the order of magnitude of $N \sim 10^4$ and $S \sim 4 \times 10^{-5}$.
  One would then observe 4 birth on 10 days, and this fixes $\lambda, \chi, \psi, \alpha_S, \beta_S$.



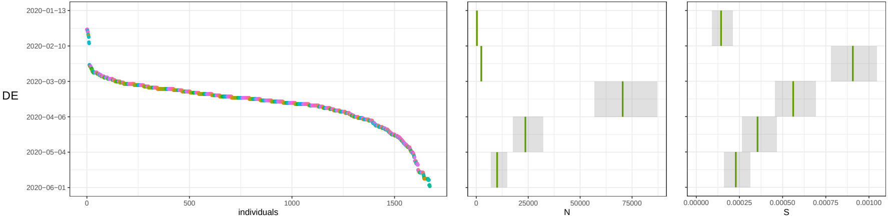
Results

Illustration on the SARS-CoV-2 dataset

- Sequences from GISAID until 1st of June 2020.

  CH 1284 genomes in 627 alleles,
  DE 1673 genomes in 886 alleles,
  FR 1919 genomes in 1166 alleles,
  IT 1314 genomes in 750 alleles.

- Fixed $\mu = 0.065$ mutations per genome per d, $g = 5$ d and timeline with 5 periods of 4 weeks each.
- To fix hyperparameters, imagine a period with few data,
  Make a guess on the order of magnitude of $N \sim 10^4$ and $S \sim 4 \times 10^{-5}$.
  One would then observe 4 birth on 10 days, and this fixes $\lambda, \chi, \psi, \alpha_S, \beta_S$.



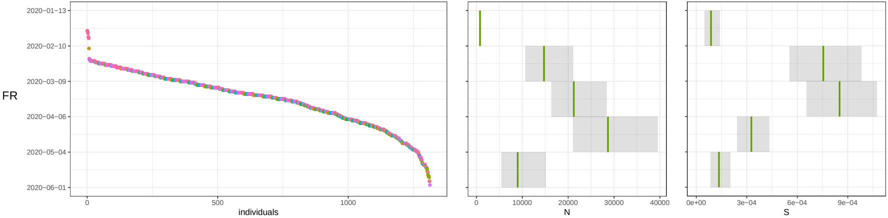
Results

Illustration on the SARS-CoV-2 dataset

- Sequences from GISAID until 1st of June 2020.

  CH 1284 genomes in 627 alleles,
  DE 1673 genomes in 886 alleles,
  FR 1919 genomes in 1166 alleles,
  IT 1314 genomes in 750 alleles.

- Fixed $\mu = 0.065$ mutations per genome per d, $g = 5$ d and timeline with 5 periods of 4 weeks each.
- To fix hyperparameters, imagine a period with few data,
  Make a guess on the order of magnitude of $N \sim 10^4$ and $S \sim 4 \times 10^{-5}$.
  One would then observe 4 birth on 10 days, and this fixes $\lambda, \chi, \psi, \alpha_S, \beta_S$.



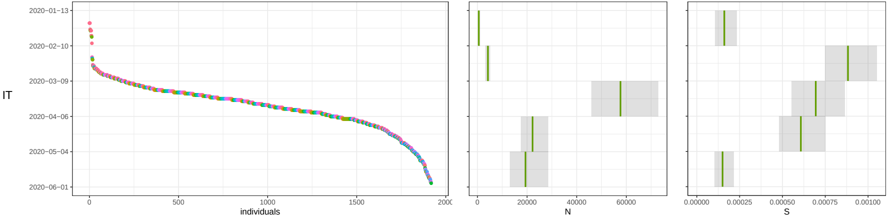
Results

Illustration on the SARS-CoV-2 dataset

- Sequences from GISAID until 1st of June 2020.

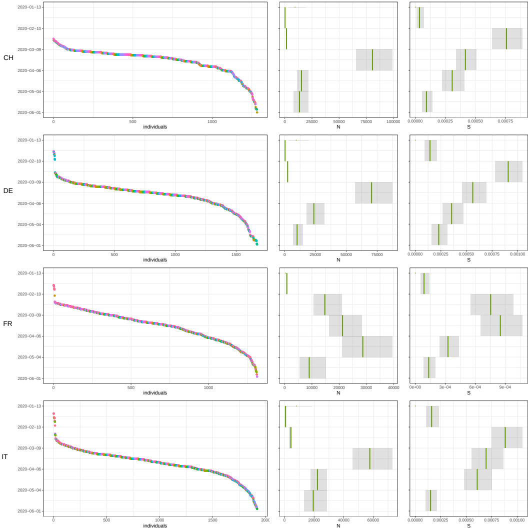
  CH 1284 genomes in 627 alleles,
  DE 1673 genomes in 886 alleles,
  FR 1919 genomes in 1166 alleles,
  IT 1314 genomes in 750 alleles.

- Fixed $\mu = 0.065$ mutations per genome per d, $g = 5$ d and timeline with 5 periods of 4 weeks each.
- To fix hyperparameters, imagine a period with few data,
  Make a guess on the order of magnitude of $N \sim 10^4$ and $S \sim 4 \times 10^{-5}$.
  One would then observe 4 birth on 10 days, and this fixes $\lambda, \chi, \psi, \alpha_S, \beta_S$.



Results

Illustration on the SARS-CoV-2 dataset



Discussion

Introduction

Model assumptions
- Parameters of the model
- Sampling and coalescent history
- Priors and summary

Inference method
- Gibbs sampling strategy
- Prior conjugacy properties for the parameters
- Data augmentation with past coalescent history
- Summary of the Gibbs sampler

Results
- Validation of the augmentation with the past coalescent history
- Validation of the MCMC by SBC
- Estimation of the running time
- Illustration on the SARS-CoV-2 dataset

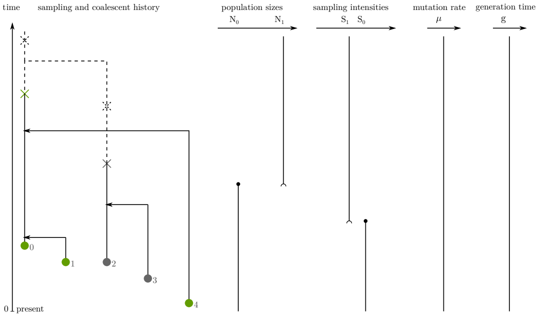
Discussion
- Opportunities for future developments
- Conclusion

Discussion

Opportunities for future developments – simulation study

How does it compare in terms of statistical power with a finite sites model ?
What signal do we loose by forgetting about the coalescent history above the first mutation ?
When does the trade-off between computation time and precision turn in favor of an infinite alleles model ?

- This could be assessed based on inferences on simulations.



- We can imagine some datasets with large allele families AND very different alleles.

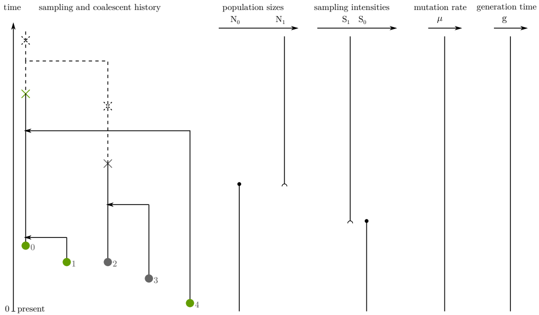
Can we combine in a clever way an infinite alleles model near the tips
and a finite sites model deeper in the tree ?

Discussion

Opportunities for future developments – simulation study

How does it compare in terms of statistical power with a finite sites model ?
What signal do we loose by forgetting about the coalescent history above the first mutation ?
When does the trade-off between computation time and precision turn in favor of an infinite alleles model ?

- This could be assessed based on inferences on simulations.



- We can imagine some datasets with large allele families AND very different alleles.

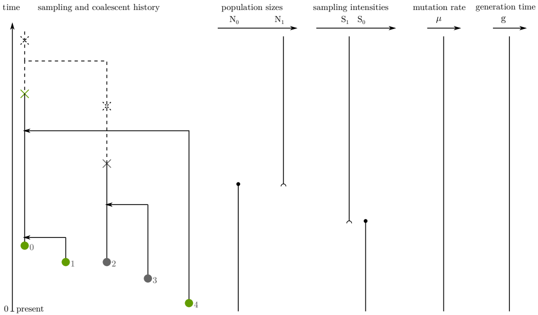
Can we combine in a clever way an infinite alleles model near the tips
and a finite sites model deeper in the tree ?

Discussion

Opportunities for future developments – simulation study

How does it compare in terms of statistical power with a finite sites model ?
What signal do we loose by forgetting about the coalescent history above the first mutation ?
When does the trade-off between computation time and precision turn in favor of an infinite alleles model ?

- This could be assessed based on inferences on simulations.



- We can imagine some datasets with large allele families AND very different alleles.

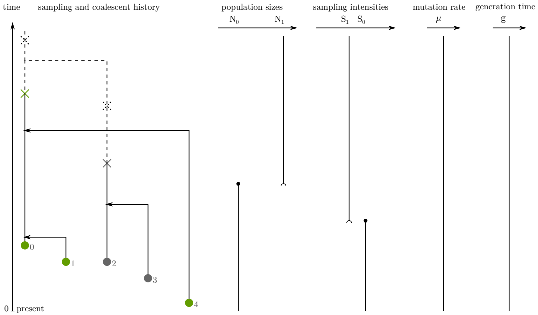
Can we combine in a clever way an infinite alleles model near the tips
and a finite sites model deeper in the tree ?

Discussion

Opportunities for future developments – simulation study

How does it compare in terms of statistical power with a finite sites model ?
What signal do we loose by forgetting about the coalescent history above the first mutation ?
When does the trade-off between computation time and precision turn in favor of an infinite alleles model ?

- This could be assessed based on inferences on simulations.



- We can imagine some datasets with large allele families AND very different alleles.

Can we combine in a clever way an infinite alleles model near the tips
and a finite sites model deeper in the tree ?

Discussion

Opportunities for future developments – Focus on running time

- It is still way too slow to apply to large (SARS-CoV-2 like) datasets.

Can we improve the running time and hope to take into account much larger datasets ?

1. Work on basic numerical optimization / parallelize chains.
2. Find a more optimized update of $\mathcal{H}$, possibly using approximations.
3. Abandon the slow Bayesian approach for a faster ML approach, possibly with EM.

Discussion

Opportunities for future developments – Focus on running time

- It is still way too slow to apply to large (SARS-CoV-2 like) datasets.

Can we improve the running time and hope to take into account much larger datasets ?

1. Work on basic numerical optimization / parallelize chains.
2. Find a more optimized update of $\mathcal{H}$, possibly using approximations.
3. Abandon the slow Bayesian approach for a faster ML approach, possibly with EM.

Discussion

Opportunities for future developments – Focus on running time

- It is still way too slow to apply to large (SARS-CoV-2 like) datasets.

Can we improve the running time and hope to take into account much larger datasets ?

1. Work on basic numerical optimization / parallelize chains.
2. Find a more optimized update of $\mathcal{H}$, possibly using approximations.
3. Abandon the slow Bayesian approach for a faster ML approach, possibly with EM.

Discussion

Opportunities for future developments – Focus on running time

- It is still way too slow to apply to large (SARS-CoV-2 like) datasets.

Can we improve the running time and hope to take into account much larger datasets ?

1. Work on basic numerical optimization / parallelize chains.
2. Find a more optimized update of $\mathcal{H}$, possibly using approximations.
3. Abandon the slow Bayesian approach for a faster ML approach, possibly with EM.

Discussion

Opportunities for future developments – Focus on running time

- It is still way too slow to apply to large (SARS-CoV-2 like) datasets.

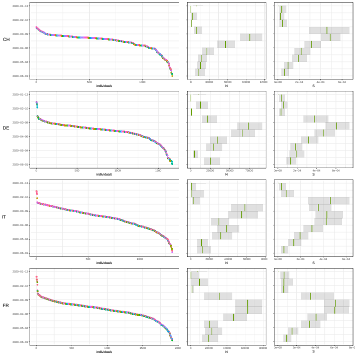
Can we improve the running time and hope to take into account much larger datasets ?

1. Work on basic numerical optimization / parallelize chains.
2. Find a more optimized update of $\mathcal{H}$, possibly using approximations.
3. Abandon the slow Bayesian approach for a faster ML approach, possibly with EM.

Discussion

Opportunities for future developments – Smoothing priors on S and N

- We don't want to believe in huge steps from a time period to another.



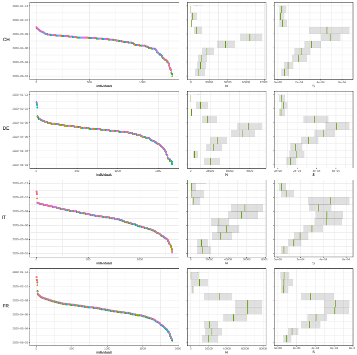
Can we incorporate smoothing priors with nice properties in this framework ?

1. They could be agnostic about the process, chosen because they satisfy nice conjugacy properties.
2. Or be rooted in epidemiology thinking, more in a Cori-Re-style for example.

Discussion

Opportunities for future developments – Smoothing priors on S and N

- We don't want to believe in huge steps from a time period to another.



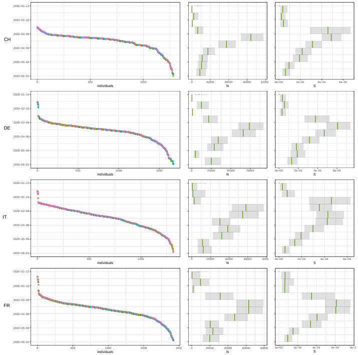
Can we incorporate smoothing priors with nice properties in this framework ?

1. They could be agnostic about the process, chosen because they satisfy nice conjugacy properties.
2. Or be rooted in epidemiology thinking, more in a Cori-Re-style for example.

Discussion

Opportunities for future developments – Smoothing priors on S and N

- We don't want to believe in huge steps from a time period to another.



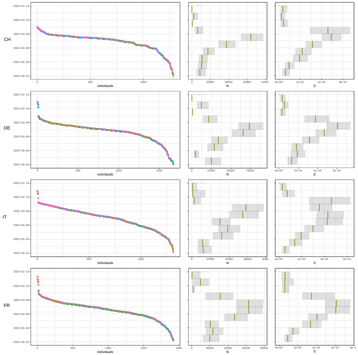
Can we incorporate smoothing priors with nice properties in this framework ?

1. They could be agnostic about the process, chosen because they satisfy nice conjugacy properties.
2. Or be rooted in epidemiology thinking, more in a Cori-Re-style for example.

Discussion

Opportunities for future developments – Smoothing priors on S and N

- We don't want to believe in huge steps from a time period to another.



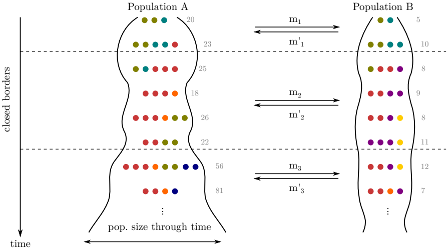
Can we incorporate smoothing priors with nice properties in this framework ?

1. They could be agnostic about the process, chosen because they satisfy nice conjugacy properties.
2. Or be rooted in epidemiology thinking, more in a Cori-Re-style for example.

Discussion

Opportunities for future developments – Demes and migrations

- The patterns of alleles across borders, with sampling through time, could inform on migration patterns.



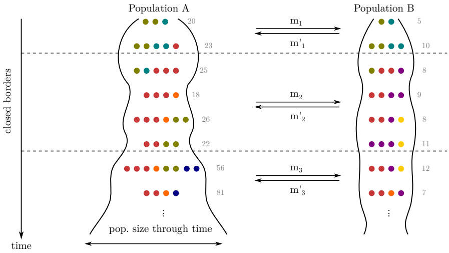
How can we extend this work to a coalescent with demes and migrations between demes ?

1. The same augmentation strategy (one genome at a time) is likely to work as well.
2. This could offer a model-based alternative to the study of "infection chains".
3. This was actually the original motivation for this project.

Discussion

Opportunities for future developments – Demes and migrations

- The patterns of alleles across borders, with sampling through time, could inform on migration patterns.



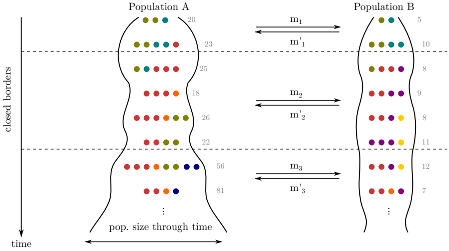
How can we extend this work to a coalescent with demes and migrations between demes ?

1. The same augmentation strategy (one genome at a time) is likely to work as well.
2. This could offer a model-based alternative to the study of "infection chains".
3. This was actually the original motivation for this project.

Discussion

Opportunities for future developments – Demes and migrations

- The patterns of alleles across borders, with sampling through time, could inform on migration patterns.



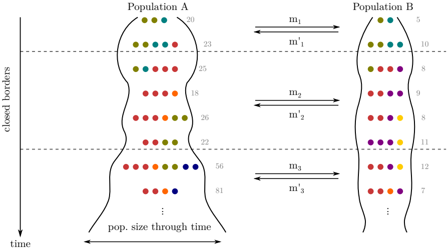
How can we extend this work to a coalescent with demes and migrations between demes ?

1. The same augmentation strategy (one genome at a time) is likely to work as well.
2. This could offer a model-based alternative to the study of "infection chains".
3. This was actually the original motivation for this project.

Discussion

Opportunities for future developments – Demes and migrations

- The patterns of alleles across borders, with sampling through time, could inform on migration patterns.



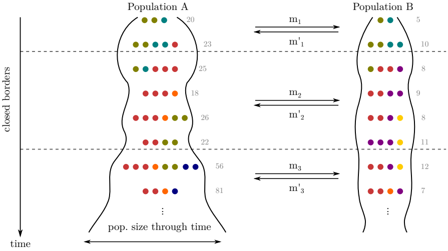
How can we extend this work to a coalescent with demes and migrations between demes ?

1. The same augmentation strategy (one genome at a time) is likely to work as well.
2. This could offer a model-based alternative to the study of "infection chains".
3. This was actually the original motivation for this project.

Discussion

Opportunities for future developments – Demes and migrations

- The patterns of alleles across borders, with sampling through time, could inform on migration patterns.



How can we extend this work to a coalescent with demes and migrations between demes ?

1. The same augmentation strategy (one genome at a time) is likely to work as well.
2. This could offer a model-based alternative to the study of "infection chains".
3. This was actually the original motivation for this project.

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention

And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics

Discussion

Take-home message on the project:

1. Input data: an allele partition of sequences sampled through time,
2. Output inference: N, S, μ, g,
3. Elegant conjugacy properties provide a good intuition on the inference process,
4. The Gibbs sampling algorithm also benefits from conjugacy properties,
5. It is illustrated on SARS-CoV-2 data from the first wave in Europe.

Opportunities for future work:

1. Simulation study to understand the benefits of using $\mathcal{B}$ vs. $\mathcal{A}, \mathcal{B}$ vs. full alignment.
2. Joint use with a classic finite sites model on different parts of the tree,
3. Developing clever approximations or turning to an EM algorithm instead of the MCMC approach,
4. Extension with smoothing priors for S and N,
5. Extension with demes and migrations,

Thank you for your attention
And please join if you are interested in any of these topics