Lecture 2: Non-phylogenetic transmission reconstruction

Epidemiological vs genomic outbreak reconstruction

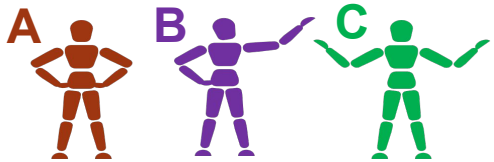
Epidemiological outbreak data alone can be used for outbreak reconstruction (i.e. contact tracing), but genomic data offer a high-resolution source of information

What can genomic data offer?

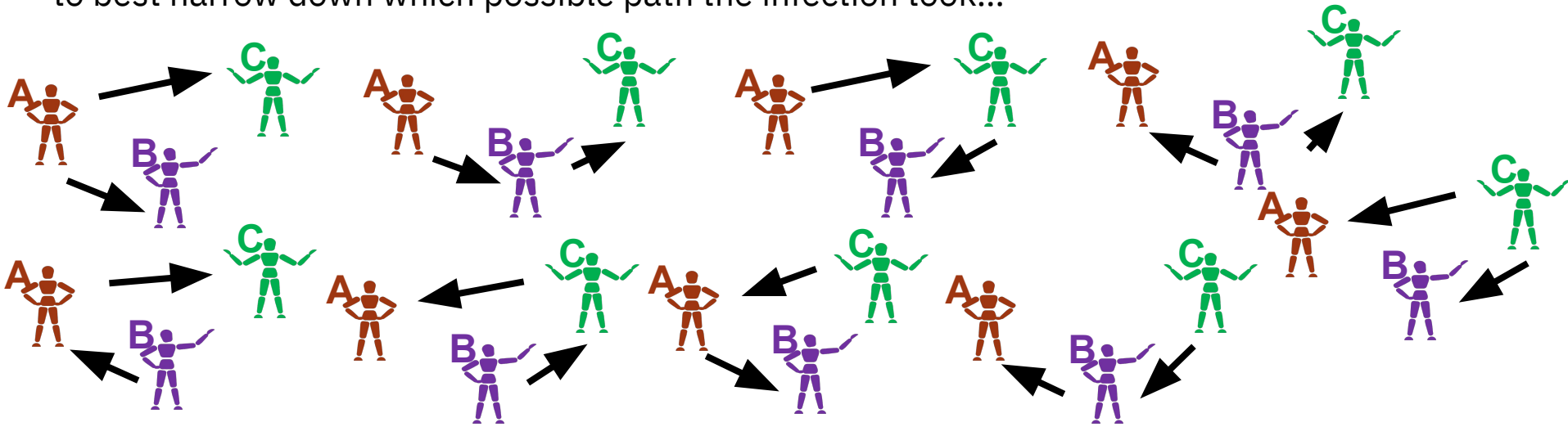
- Extra detail
- Resolve transmission where epi data are hard to get and/or have gaps
- Genomic data becoming ever easier, cheaper and faster to obtain (e.g. real-time sequencing becoming possible)

Challenge: create a single framework/likelihood incorporating genomic + epidemiological data

Imagine we have 3 people infected in an outbreak...



We want to combine our genomic information and our epidemiological information, to best narrow down which possible path the infection took...



2 of the earliest approaches

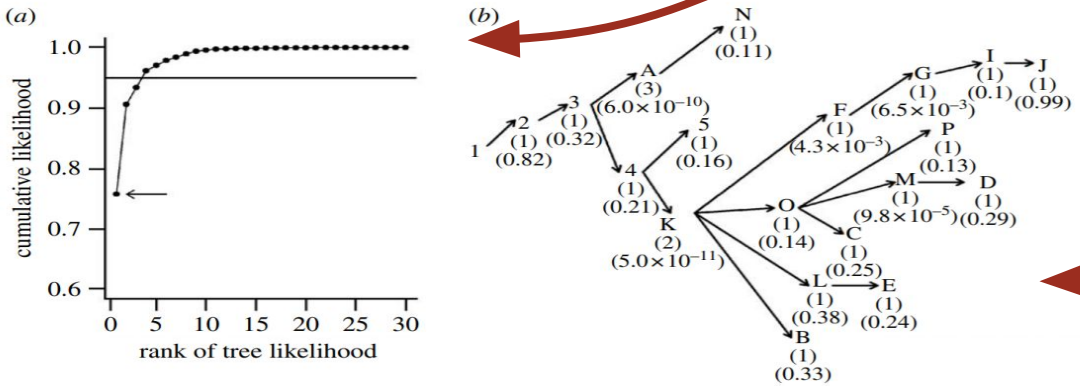
Many of the earliest methods tackled the 2 data streams separately

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2008) **275**, 887–895
doi:10.1098/rspb.2007.1442
Published online 29 January 2008

Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus

Eleanor M. Cottam[1,2], Gaël Thébaud[2,†], Jemma Wadsworth[1], John Gloster[3,‡], Leonard Mansley[4], David J. Paton[1], Donald P. King[1] and Daniel T. Haydon[2,*]



Cottam et al

- 3-step maximum likelihood approach
- Rank the likelihood of the set of plausible trees

Applied to 20 farms from 2001 UK Foot-and-mouth disease outbreak, to obtain a most likely transmission tree

Many of the earliest methods tackled the 2 data streams separately - Cottam et al process

Begin with a set of **all** possible transmission trees given the set of sampled cases

1. Select only trees that are consistent with **known infection pairs**
2. Select only remaining trees that are consistent with **the genomic data**
3. Calculate the **likelihood** of each remaining tree based on the **epi information** – describing both the chance each host (farm) was infected on a given day and able to infect others on a given day.

Many of the earliest methods tackled the 2 data streams separately - Cottam et al process

Begin with a set of **all** possible transmission trees given the set of sampled cases

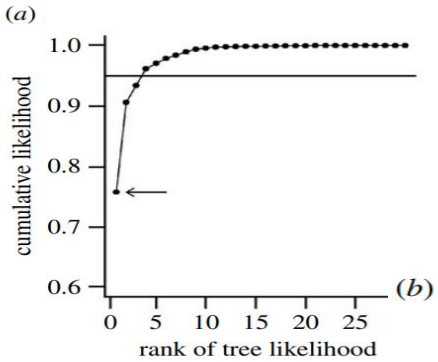
1. Select only trees that are consistent with **known infection pairs**
2. Select only remaining trees that are consistent with **the genomic data**
3. Calculate the **likelihood** of each remaining tree based on the **epi information** – describing both the chance each host (farm) was infected on a given day and able to infect others on a given day.

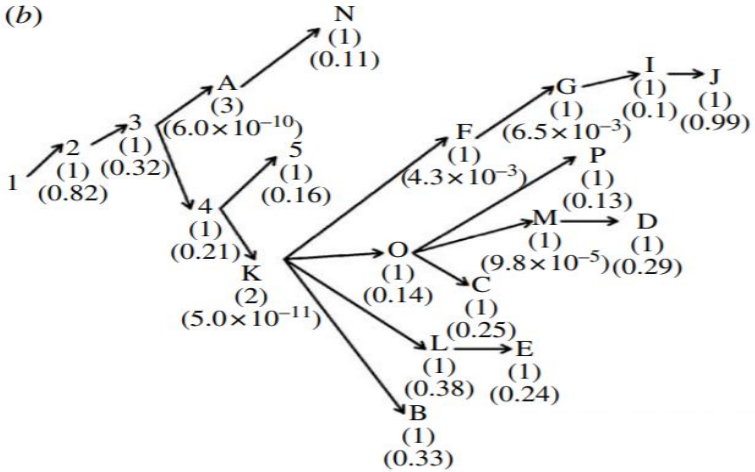
← Potentially demanding

← Also potentially demanding

Depending on the size of the data and how many trees you were able to exclude

Many of the earliest methods tackled the 2 data streams separately - Cottam et al process





Finally, either

(a) pick 1 optimal tree, or

(b) pick a set of optimal trees (and look for similarities between them)

This is done by ranking the remaining trees by their likelihood

SeqTrack - a graph based approach

This method also tackles the genomic and then the epidemiological information

Heredity (2011) 106, 383–390

www.nature.com/hdy

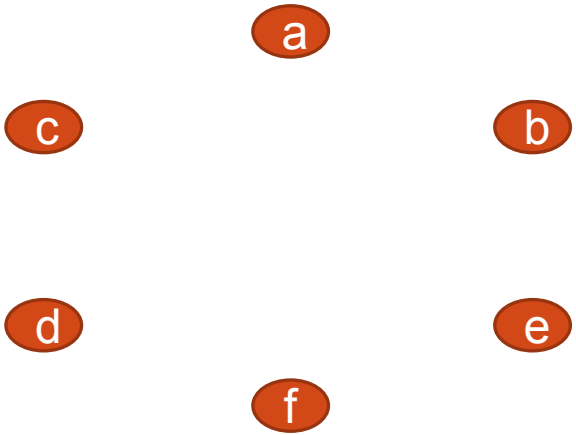
ORIGINAL ARTICLE

Reconstructing disease outbreaks from genetic data: a graph approach

T Jombart, RM Eggo, PJ Dodd and F Balloux
Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analysis and Modelling, Imperial College Faculty of Medicine, London, UK

- **Graph theory** approach to find 'genetically parsimonious' transmission trees
- Algorithm *SeqTrack* finds the optimum branching in a directed graph

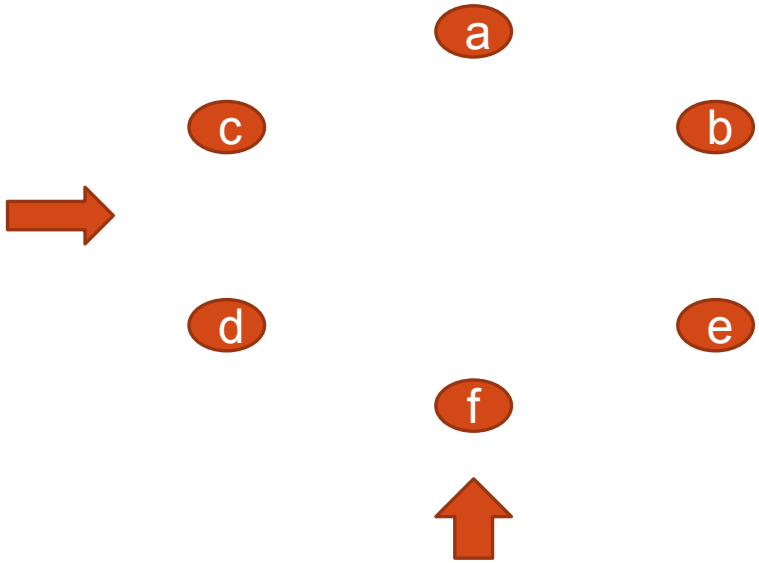
Imagine we have an outbreak with 6 cases, a:f



Imagine we have an outbreak with 6 cases, a:f

Genomic distance matrix

	a	b	c	d	e	f
a	0	1	3	2	5	9
b		0	2	4	7	5
c			0	1	4	12
d				0	1	3
e					0	8
f						0



Sample collection dates:

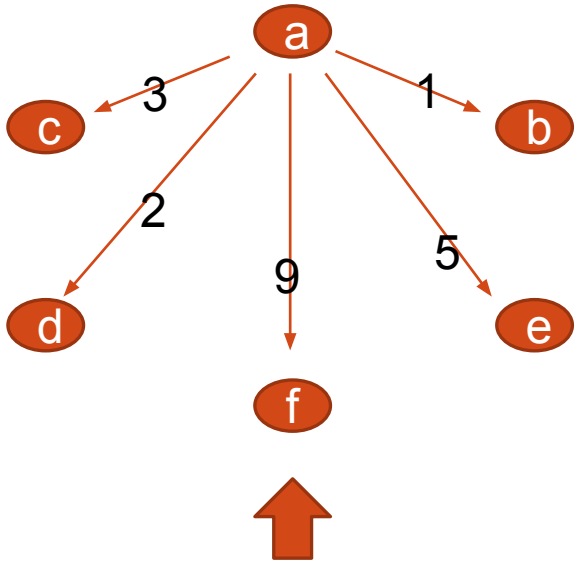
a: t=3
b: t=1
c: t=2
d: t=5
e: t=4
f: t=7

Imagine we have an outbreak with 6 cases, a:f

Genomic distance matrix

	a	b	c	d	e	f
a	0	1	3	2	5	9
b		0	2	4	7	5
c			0	1	4	12
d				0	1	3
e					0	8
f						0

(i) Create a connected, directed graph with weights w_{ij} equal to the genetic distance



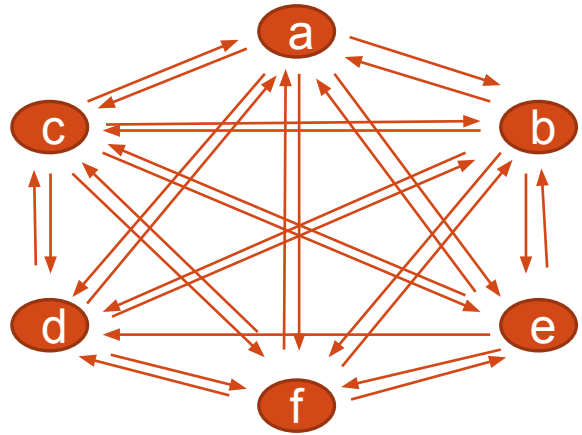
Sample collection dates:

a: t=3
b: t=1
c: t=2
d: t=5
e: t=4
f: t=7

Imagine we have an outbreak with 6 cases, a:f

Genomic distance matrix

	a	b	c	d	e	f
a	0	1	3	2	5	9
b		0	2	4	7	5
c			0	1	4	12
d				0	1	3
e					0	8
f						0



(i) Create a connected, directed graph with weights w_{ij} equal to the genetic distance

We do this every node, but lets restrict to (a) for simplicity…

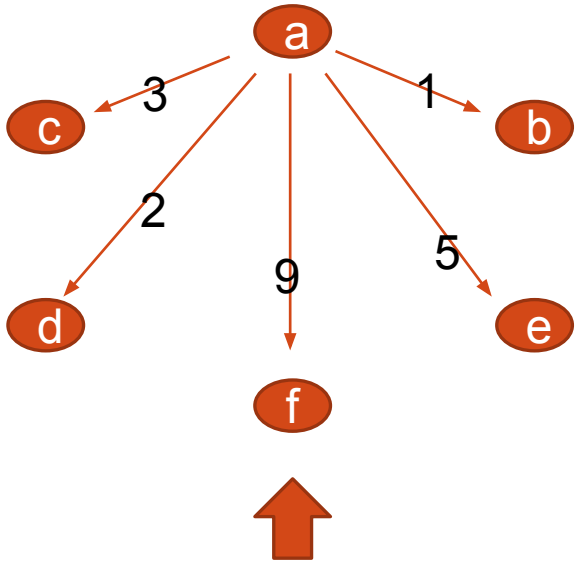
Sample collection dates:

a: t=3
b: t=1
c: t=2
d: t=5
e: t=4
f: t=7

Imagine we have an outbreak with 6 cases, a:f

Genomic distance matrix

	a	b	c	d	e	f
a	0	1	3	2	5	9
b		0	2	4	7	5
c			0	1	4	12
d				0	1	3
e					0	8
f						0



(i) Create a connected, directed graph with weights w_{ij} equal to the genetic distance
(ii) Remove edge ij if $t_j < t_i$

Sample collection dates:

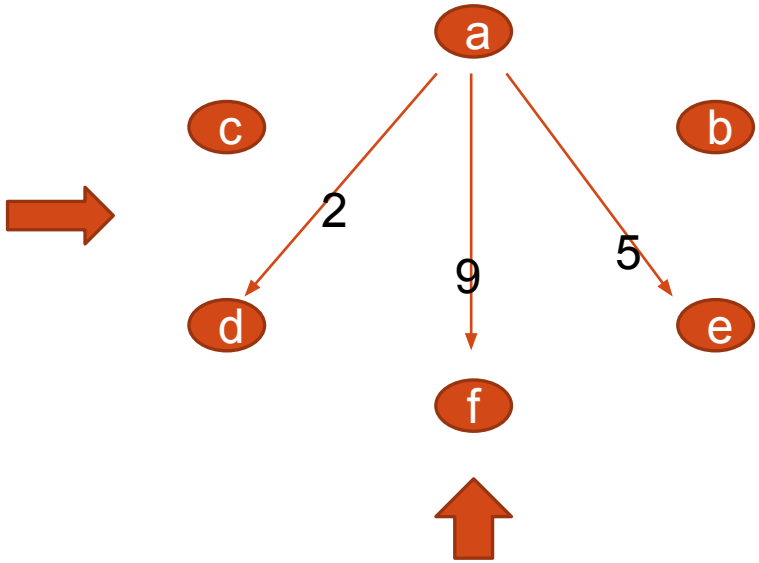
a: t=3
b: t=1
c: t=2
d: t=5
e: t=4
f: t=7

Imagine we have an outbreak with 6 cases, a:f

Genomic distance matrix

	a	b	c	d	e	f
a	0	1	3	2	5	9
b		0	2	4	7	5
c			0	1	4	12
d				0	1	3
e					0	8
f						0

(i) Create a connected, directed graph with weights w_{ij} equal to the genetic distance
(ii) Remove edge ij if $t_j < t_i$



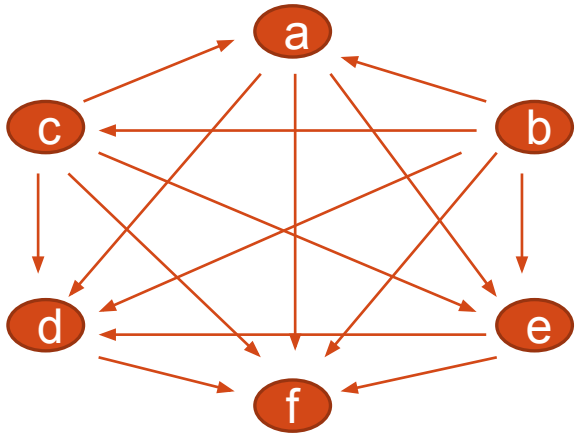
Sample collection dates:

a: t=3
b: t=1
c: t=2
d: t=5
e: t=4
f: t=7

Imagine we have an outbreak with 6 cases, a:f

Genomic distance matrix

	a	b	c	d	e	f
a	0	1	3	2	5	9
b		0	2	4	7	5
c			0	1	4	12
d				0	1	3
e					0	8
f						0



(i) Create a connected, directed graph with weights w_{ij} equal to the genetic distance
(ii) Remove edge ij if $t_j < t_i$

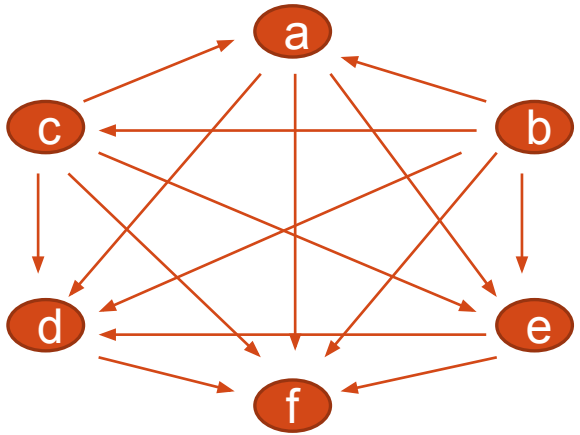
Sample collection dates:

a: t=3
b: t=1
c: t=2
d: t=5
e: t=4
f: t=7

Imagine we have an outbreak with 6 cases, a:f

Genomic distance matrix

	a	b	c	d	e	f
a	0	1	3	2	5	9
b		0	2	4	7	5
c			0	1	4	12
d				0	1	3
e					0	8
f						0



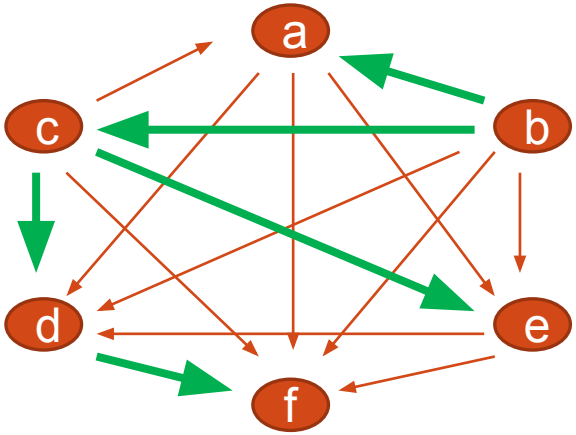
(i) Create a connected, directed graph with weights w_{ij} equal to the genetic distance
(ii) Remove edge ij if $t_j < t_i$
(iii) Find the spanning directed tree optimizing (i.e. minimizing) $\sum w_{ij}$

Sample collection dates:

a: t=3
b: t=1
c: t=2
d: t=5
e: t=4
f: t=7

'This problem has been solved by Edmonds (1967) and Chu and Liu (1965), …The algorithm proceeds by identifying optimum ancestors for each node at the exception of the root (the oldest isolate), and then recursively removes possible cycles. However, in our case, cycles are impossible as ancestries cannot go back in time, which greatly simplifies computations.'

	a	b	c	d	e	f
a	0	1	3	2	5	9
b		0	2	4	7	5
c			0	1	4	12
d				0	1	3
e					0	8
f						0



Sample collection dates:

a: t=3
b: t=1
c: t=2
d: t=5
e: t=4
f: t=7

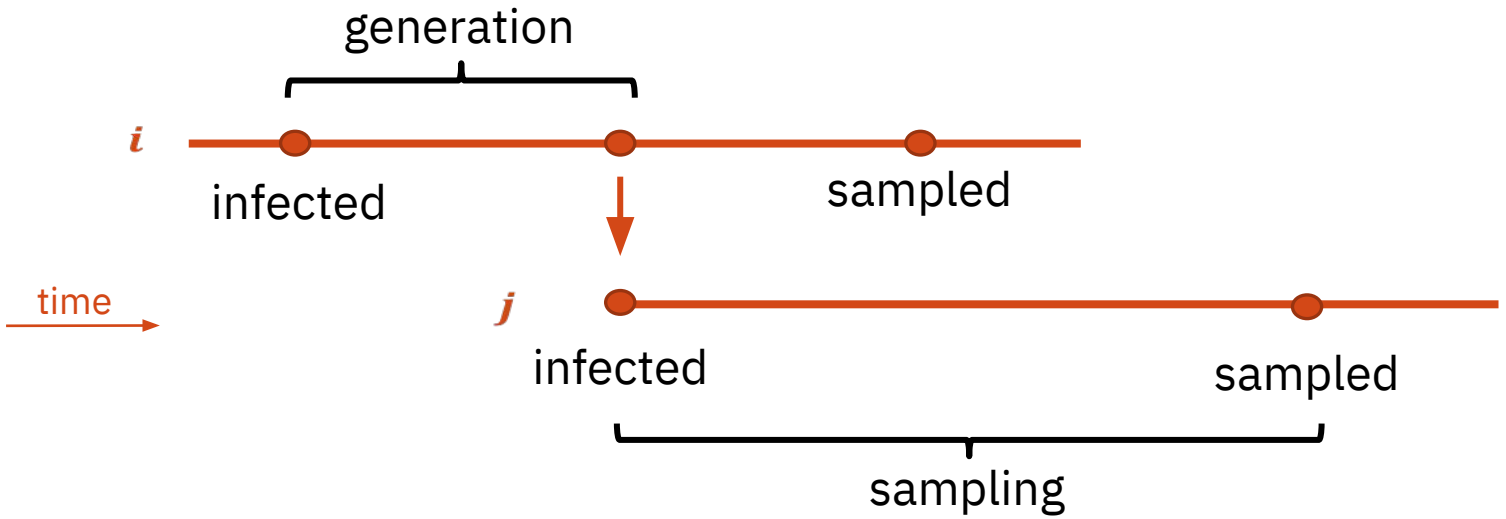
Some limitations:
- All cases come from single index case e.g. a single sampled ancestor
- All cases are known and sampled
- Sampling times not used in weighting

SeqTrack also:
- Assumes that individuals became infectious in the order they are sampled
- Has no uncertainty in the output transmission tree or probabilistic parameters

But
- Fast, simple, explore all the possibilities

2 short primers for lecture 2

A quick primer 1: generation time and sampling time



Generation time = the time interval between the infection of an individual and their seeding of new secondary cases.

Sampling time = the time interval between infection and collection of an isolate.

A quick primer 2: Markov Chain Monte Carlo (MCMC)

A popular computational method for exploring complex and/or high-dimensional spaces – e.g. transmission trees

The main idea:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

likelihood

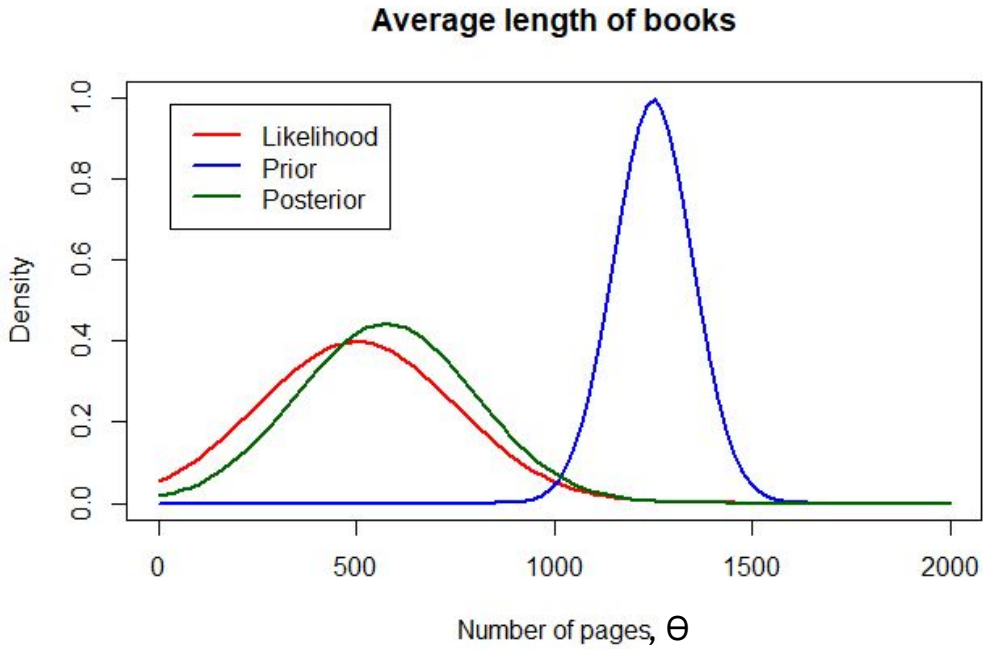
prior

Posterior distribution– the probability of our model parameters Θ given the data *y*

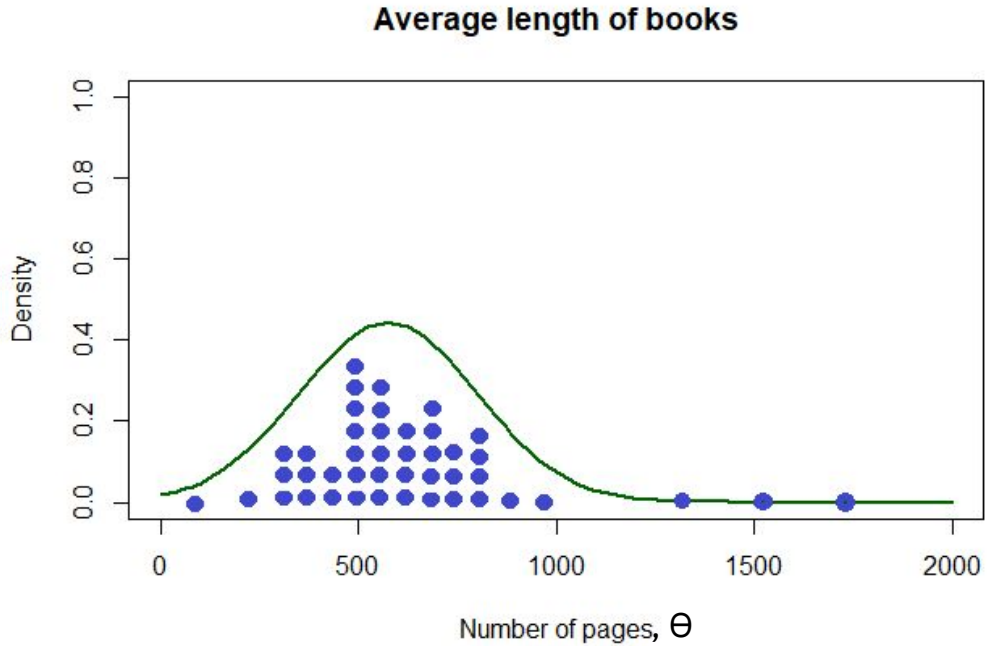
For us, we may have e.g.
Θ = transmission tree and parameters controlling it
y = sequences and epi data

When this quantity (the posterior) is hard to maximise directly, we instead form a Markov chain with equilibrium distribution equal to the posterior distribution, and take many samples from this chain.

A quick primer 2: Markov Chain Monte Carlo (MCMC)



A quick primer 2: Markov Chain Monte Carlo (MCMC)



Essentially, we approximate the posterior distribution by random sampling from a probabilistic space (of all possible books or all possible transmission trees).

A quick primer 2: Markov Chain Monte Carlo (MCMC)

Data-augmented MCMC is a method for dealing with missing data within an MCMC algorithm. As well as sampling from the parameter space at each step of the Markov chain, we also sample values for the missing data.

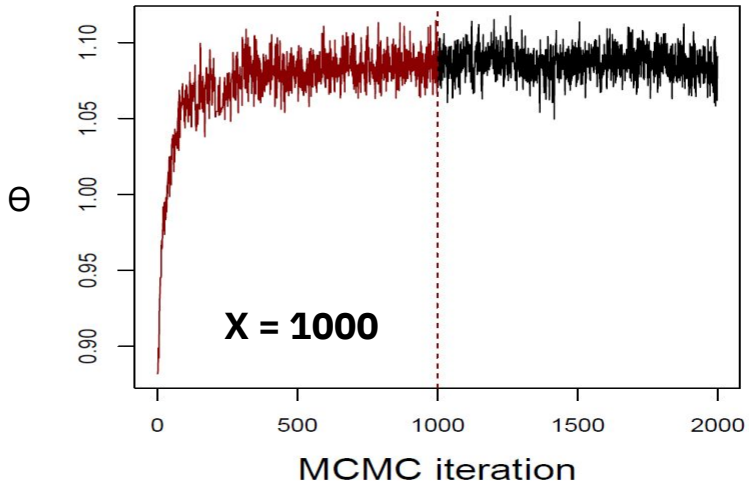
In transmission inference, missing data might be the time of infection of the cases (since typically we only know sampling times) or the number of unsampled cases, for example.

A quick primer 2: Markov Chain Monte Carlo (MCMC)

In actuality, the ‘random’ samples we collect in MCMC are not independent – they form a chain with *equilibrium* distribution equal to the target distribution.

The set of **X** samples at the start of the MCMC run are often discarded – it takes some time to reach an area of the state space with good posterior support. We call this initial set **X** the **burn-in**

Trace plot diagnostics



Transmission reconstruction with *outbreaker(2)*

outbreaker and outbreaker2

We’re going to look at these 2 methods in detail – and will be using them in the next exercise

These also create a unified likelihood for genetic + epidemiological data, but within a Bayesian framework, that allows more estimation and greater flexibility.

outbreaker vs outbreaker2

outbreaker2 is a more customisable version of outbreaker

We’re mainly going to focus on the core outbreaker model…

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data

Thibaut Jombart*, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser*, Neil Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

Data:

N sampled cases, each with genetic sequence s_i and time of sampling t_i

Quantities:

$d(s_i, s_j)$ = number of mutations (distance) between sequences i and j
$l(s_i, s_j)$ = number of nucleotide positions which can be compared i and j
w = distribution of the generation time
f = distribution of the sampling time

```
> 1:1999-08-01
GCACCCATTCCCGCCTGGAGAT
> 2:2007-11-01
GCACCCATTCCCGCCTAGAGAT
```

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data

Thibaut Jombart*, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser*, Neil Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

Data:

N sampled cases, each with genetic sequence s_i and time of sampling t_i

Quantities:

$d(s_i, s_j)$ = number of mutations (distance) between sequences i and j

$l(s_i, s_j)$ = number of nucleotide positions which can be compared i and j

} Derived

w = distribution of the generation time

f = distribution of the sampling time

} Assumed

```
> 1:1999-08-01
GCACCCATTCCCGCCTGGAGAT
> 2:2007-11-01
GCACCCATTCCCGCCTAGAGAT
```

Goal: find the most likely transmission tree

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data

Thibaut Jombart*, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser*, Neil Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

Data:

N sampled cases, each with genetic sequence s_i and time of sampling t_i

Quantities:

$\mathrm{d}(s_i, s_j)$ = number of mutations (distance) between sequences i and j
$\mathrm{l}(s_i, s_j)$ = number of nucleotide positions which can be compared i and j
w = distribution of the generation time
f = distribution of the sampling time

Augmented data:

α_i = index of the most recent sampled ancestor of i
κ_i = number of (Sampled and unsampled) generations between i and α_i
T_i^{inf} = date of infection of i

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data

Thibaut Jombart*, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser*, Neil Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

Data:

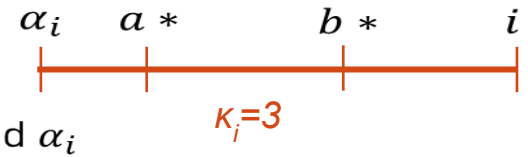
N sampled cases, each with genetic sequence s_i and time of sampling t_i

Quantities:

$\mathrm{d}(s_i, s_j)$ = number of mutations (distance) between sequences i and j
$\mathrm{l}(s_i, s_j)$ = number of nucleotide positions which can be compared i and j
w = distribution of the generation time
f = distribution of the sampling time

Augmented data:

α_i = index of the most recent sampled ancestor of i
κ_i = number of (Sampled and unsampled) generations between i and α_i
T_i^{inf} = date of infection of i



In addition to obtaining MCMC samples of the augmented data, we estimate 2 parameters

Parameters:

μ = mutation rate, per site per generation of infection
π = proportion of unsampled cases
are estimated as well as the transmission tree

In addition to obtaining MCMC samples of the augmented data, we estimate 2 parameters

Parameters:

μ = mutation rate, per site per generation of infection
π = proportion of unsampled cases
are estimated as well as the transmission tree

Posterior distribution:

$$P(A, \theta \mid D) = \frac{P(D, A \mid \theta) P(\theta)}{P(D)} \propto p\left(\left\{s_i, t_i, \alpha_i, \kappa_i, T_i^{\inf}\right\}_{i=1,\ldots,N} \,\middle|\, \mu, \pi\right) \times p(\mu, \pi).$$

In addition to obtaining MCMC samples of the augmented data, we estimate 2 parameters

Parameters:

μ = mutation rate, per site per generation of infection
π = proportion of unsampled cases
are estimated as well as the transmission tree

Posterior distribution:

$$P(A, \theta \mid D) = \frac{P(D, A \mid \theta) P(\theta)}{P(D)} \propto \underbrace{p}_{\text{likelihood}} \Big(\underbrace{\{s_i, t_i}_{\text{data}}, \underbrace{\alpha_i, \kappa_i, T_i^{\inf}\}_{i=1,\ldots,N}}_{\text{augmented data}} \Big| \underbrace{\mu, \pi}_{\text{given the param.s}} \Big) \times \underbrace{p(\mu, \pi)}_{\text{prior on param.s}}.$$

In addition to obtaining MCMC samples of the augmented data, we estimate 2 parameters

Parameters:

μ = mutation rate, per site per generation of infection
π = proportion of unsampled cases
are estimated as well as the transmission tree

Posterior distribution:

$$P(A,\theta\,|D) = \frac{P(D,A\,|\theta)P(\theta)}{P(D)} \propto p\left(\left\{s_i,t_i,\alpha_i,\kappa_i,T_i^{\inf}\right\}_{i=1,\dots,N}\,\middle|\,\mu,\pi\right)\times p(\mu,\pi).$$

All cases are assumed to be conditionally independent, given the identity of their most recent sampled ancestor, so the likelihood decomposes to:

$$p\left(\left\{s_i,t_i,\alpha_i,\kappa_i,T_i^{\inf}\right\}_{i=1,\dots,N}\,\middle|\,\mu,\pi\right) = \prod_{i=2}^{N} p\left(s_i,t_i,\alpha_i,\kappa_i,T_i^{\inf}\,|s_{\alpha_i},t_{\alpha_i},T_{\alpha_i}^{\inf},\mu,\pi\right)\times$$
$$p(t_1\,|T_1^{\inf})p(s_1)p\left(T_1^{\inf}\right)p(\alpha_1)p(\kappa_1)$$

In addition to obtaining MCMC samples of the augmented data, we estimate 2 parameters

Parameters:

μ = mutation rate, per site per generation of infection
π = proportion of unsampled cases
are estimated as well as the transmission tree

Posterior distribution:

$$P(A, \theta \mid D) = \frac{P(D, A \mid \theta) P(\theta)}{P(D)} \propto p\left(\left\{s_i, t_i, \alpha_i, \kappa_i, T_i^{\inf}\right\}_{i=1,\ldots,N} \,\middle|\, \mu, \pi\right) \times p(\mu, \pi).$$

All cases are assumed to be conditionally independent, given the identity of their most recent sampled ancestor, so the likelihood decomposes to:

$$p\left(\left\{s_i, t_i, \alpha_i, \kappa_i, T_i^{\inf}\right\}_{i=1,\ldots,N} \,\middle|\, \mu, \pi\right) = \prod_{i=2}^{N} p\left(s_i, t_i, \alpha_i, \kappa_i, T_i^{\inf} \mid s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{\inf}, \mu, \pi\right) \times$$
$$p(t_1 \mid T_1^{\inf}) p(s_1) p\left(T_1^{\inf}\right) p(\alpha_1) p(\kappa_1)$$

} These terms relate only to initial case

One point to note: since cases may share a common unsampled ancestry, this is technically a composite (approximate) likelihood

$$p\left(\left\{s_i, t_i, \alpha_i, \kappa_i, T_i^{\inf}\right\}_{i=1,\ldots,N} \,\middle|\, \mu, \pi\right) = \prod_{i=2}^{N} p\left(s_i, t_i, \alpha_i, \kappa_i, T_i^{\inf} \,|s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{\inf}, \mu, \pi\right) \times p(t_1 \,|T_1^{\inf}) p(s_1) p\left(T_1^{\inf}\right) p(\alpha_1) p(\kappa_1)$$

The pseudo-likelihood is further decomposed into genetic and epidemiological components. For each case $i = 1, \ldots, \mathrm{N}$:

$$p\left(s_i, t_i, \alpha_i, \kappa_i, T_i^{\inf} \mid s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{\inf}, \mu, \pi\right)$$
$$= p\left(s_i \mid \alpha_i, s_{\alpha_i}, \kappa_i, \mu\right) \times p\left(t_i \mid T_i^{\inf}\right) p\left(T_i^{\inf} \,\middle|\, \alpha_i, T_{\alpha_i}^{\inf}, \kappa_i\right) p(\kappa_i \mid \pi) p(\alpha_i)$$

Genetic part

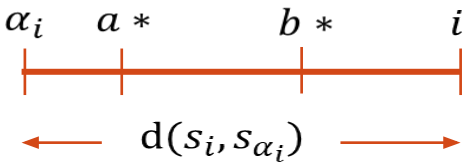
Epidemiological part

Constant

Genetic part

The outbreaker genetic model assumes no within-host genetic diversity, and so mutations are direct features of transmission events. All transmission events are assumed independent, and the genetic pseudo-likelihood is very fast to compute.

Genetic pseudo-likelihood of case i = the probability of observing genetic distance $\mathrm{d}(s_i, s_{\alpha_i})$ between sequence s_i and the ancestral sequence s_{α_i} with i and α_i separated by κ_i generations.



As a method designed for shorter timescale outbreaks, reverse mutations are considered negligible.

Genetic pseudolikelihood = $$\mu^{\mathrm{d}(s_i, s_{\alpha_i})} (1-\mu)^{\kappa_i \times \mathrm{l}(s_i, s_{\alpha_i}) - \mathrm{d}(s_i, s_{\alpha_i})}$$

Epidemiological part

Remember:
w = distribution of the generation time
f = distribution of the sampling time

Describes the probability of…

Time of sampling given time of infection

Time of infection given knowledge of infector

Number of missing cases given rate of missing cases

$$p\left(t_i \mid T_i^{\inf}\right) p\left(T_i^{\inf} \mid \alpha_i, T_{\alpha_i}^{\inf}, \kappa_i\right) p(\kappa_i \mid \pi)$$

$$=$$

$$f\left(t_i - T_i^{\inf}\right) \times w^{\kappa_i}\left(T_i^{\inf} - T_{\alpha_i}^{\inf}\right) \times \mathrm{NB}(1 \mid \kappa_i\text{-}1, \pi)$$

probability of obtaining one 'success' (sampling a case) after $\kappa_i - 1$ 'failures' (unobserved cases), with probability of success π.

$$p\left(s_i, t_i, \alpha_i, \kappa_i, T_i^{\inf} \mid s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{\inf}, \mu, \pi\right)$$
$$= \underbrace{p(s_i \mid \alpha_i, s_{\alpha_i}, \kappa_i, \mu)}_{\text{Genetic part}} \times \underbrace{p\left(t_i \mid T_i^{\inf}\right) p\left(T_i^{\inf} \,\middle|\, \alpha_i, T_{\alpha_i}^{\inf}, \kappa_i\right) p(\kappa_i \mid \pi)}_{\text{Epidemiological part}} \underbrace{p(\alpha_i)}_{\text{Constant}}$$

That forms the core of the outbreaker model.

The likelihood expressions introduced in the previous slides are combined with priors for the mutation rate μ and proportion of unsampled cases π.

In outbreaker 1:

μ is given a uniform prior on [0,1] – corresponding to an assumption of scarce prior information on this
π is given a beta distributed prior with parameters controlled by the user of outbreaker. This is a flexible prior which can reflect different levels of prior knowledge for different datasets.

The authors also introduce a method for detecting imported cases – i.e. cases that are not descended from another case in the outbreak.

In an initial step of the model, genetic outliers are detected, relative to the other samples in the dataset. A 'global influence' GI_i is calculated for each sampled case, defined as

$$GI_i = \mathbb{E}\left(\sum_{j=1, j\neq i}^{n} GPL_J\right) - \mathbb{E}\left(\sum_{i=1}^{n} GPL_i\right)$$

where GPL is the genetic pseudo-likelihood. This is calculated over the first few samples of the MCMC, say 50.

A large value of the GI_i implies unlikely numbers of mutations i.e. a 'distant' sequence. Cases with a global influence more than 5 times the average across all cases are considered outliers.

The authors also introduce a method for detecting imported cases – i.e. cases that are not descended from another case in the outbreak.

In an initial step of the model, genetic outliers are detected, relative to the other samples in the dataset. A 'global influence' GI_i is calculated for each sampled case, defined as

$$GI_i = \mathbb{E}\left(\sum_{j=1, j\neq i}^{n} GPL_J\right) - \mathbb{E}\left(\sum_{i=1}^{n} GPL_i\right)$$

Describes what proportion *i*'s GPL is of the total GPL

where GPL is the genetic pseudo-likelihood. This is calculated over the first few samples of the MCMC, say 50.

A large value of the GI_i implies unlikely numbers of mutations i.e. a 'distant' sequence. Cases with a global influence more than 5 times the average across all cases are considered outliers.

An application from

Data from 2003 Singaporean Severe Acute Respiratory Syndrome (SARS) outbreak.
13 genomes with <15 mutations between all pairs.

Generation time = Γ(mean 8.4, SD 3.8)
Same sampling time

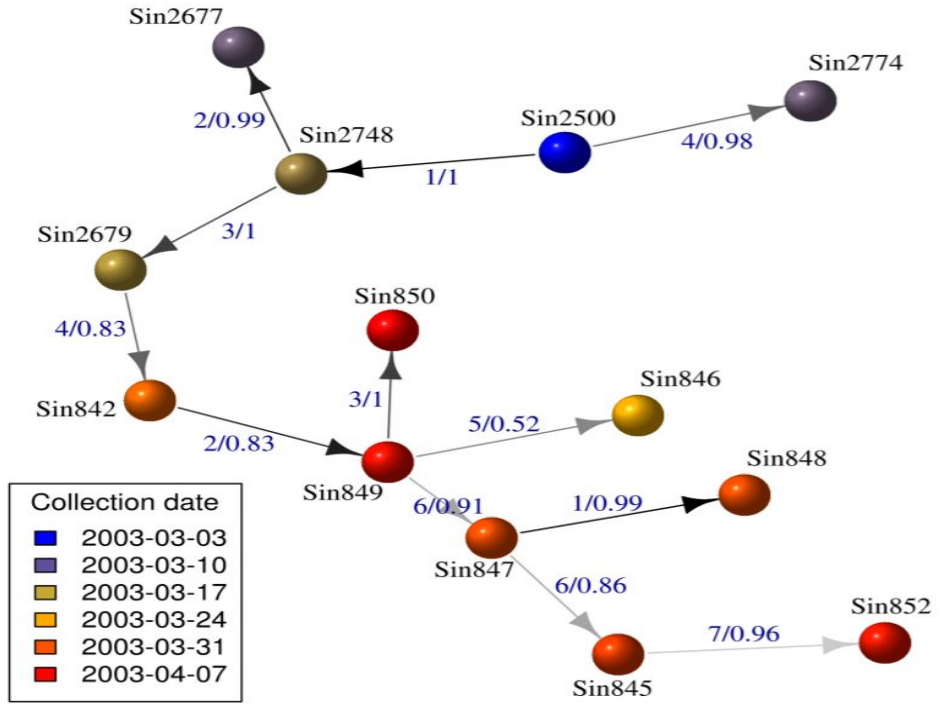
OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data

Thibaut Jombart*, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser*, Neil Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom



An application from

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data

Thibaut Jombart*, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser*, Neil Ferguson

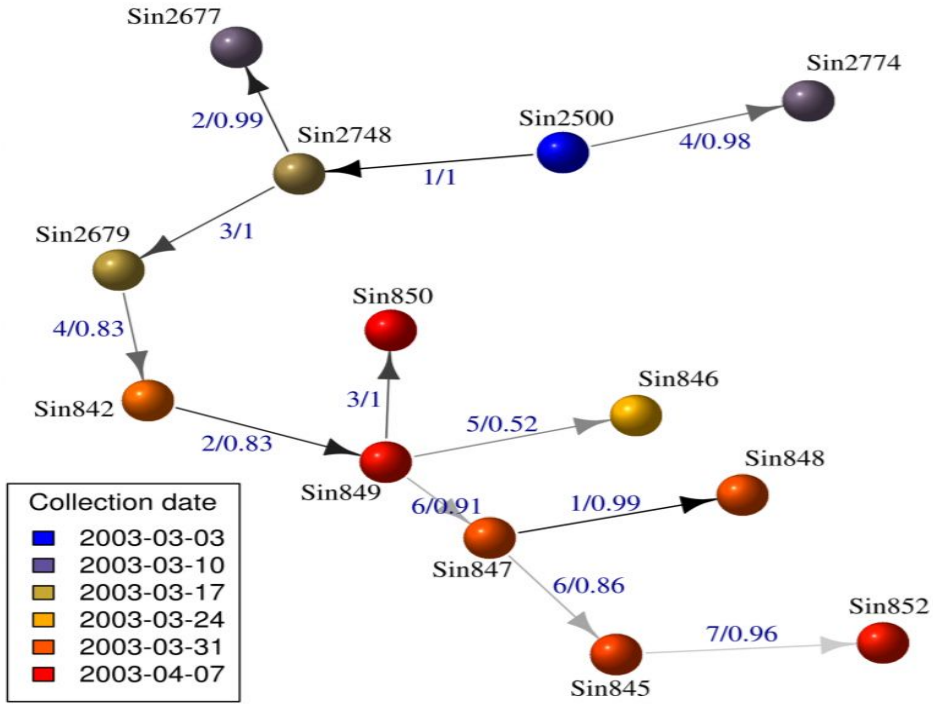
MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

Data from 2003 Singaporean Severe Acute Respiratory Syndrome (SARS) outbreak.
13 genomes with <15 mutations between all pairs.

Generation time = Γ(mean 8.4, SD 3.8)
Same sampling time

How to get here from the posterior expression?

1. Run MCMC to sample many trees (and many μ, π, … values)
2. Pick a consensus tree that best represents the set



outbreaker2: extensions

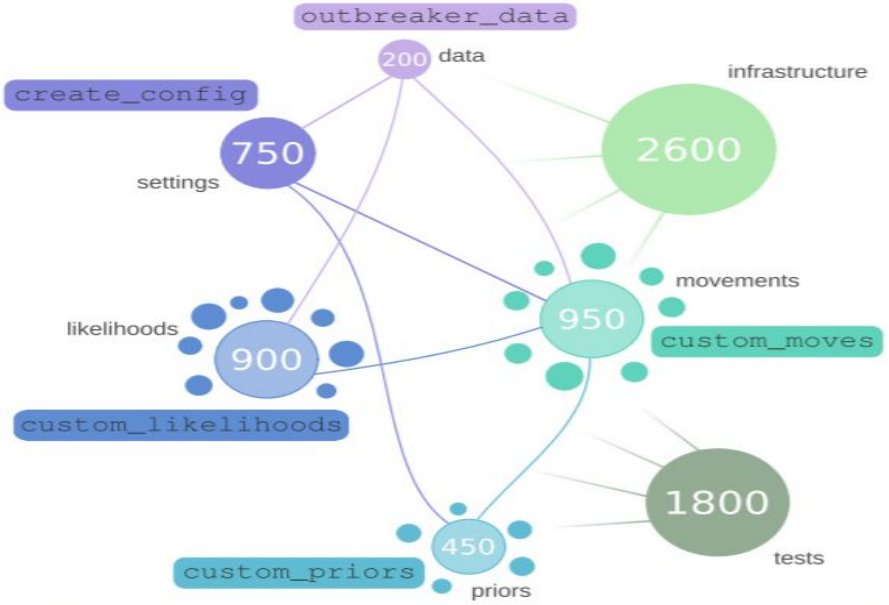
Campbell *et al. BMC Bioinformatics* 2018, **19**(Suppl 11):363
https://doi.org/10.1186/s12859-018-2330-z

BMC Bioinformatics

SOFTWARE Open Access

outbreaker2: a modular platform for outbreak reconstruction

Finlay Campbell, Xavier Didelot, Rich Fitzjohn, Neil Ferguson, Anne Cori and Thibaut Jombart*



numbers = lines of code

- Combines an R package with C++ code for efficiency, through Rcpp
- Can customise all these facets of the package
- For example, they implemented the TransPhylo methodology, which we will work with tomorrow, in outbreaker2.