Phylogenies
Phylogenies describe history
Phylogenies describe history

Haeckel. 1879.
Phylogenies describe history

Phylogenies are the result of branching processes
Timeseries and phylogeny are dual outcomes of an infectious process
Epidemic process
Epidemic process

Count

Time
Can ask for the probability of observing this timeseries given epidemiological parameters $\beta$ and $\gamma$. 
Epidemic process
Epidemic process

Sample some individuals
Epidemic branching process
Epidemic branching process
Epidemic branching process

Can ask for the probability of observing this tree given epidemiological parameters $\beta$ and $\gamma$. 
The coalescent

Assume equilibrium number of infecteds. Call this equilibrium $N$. 
The coalescent

Sample some individuals
The coalescent

Each generation, there is a small chance for coalescence for each pair

\[ Pr(\text{coal}|i = 2) = \frac{1}{N} \]
The coalescent

Probability of coalescence scales quadratically with lineage count

\[ \Pr(\text{coal}) = \binom{i}{2} \frac{1}{N} = \frac{i(i - 1)}{2N} \]
The coalescent
The coalescent
The coalescent
The coalescent

\[ T_i \sim \text{Exponential} \left( \frac{2N}{i(i-1)} \right) \]
Demo
Population size affects tree shape

The rate of coalescence decreases linearly with the population size $N$. 
rates across lineages

recent methods designed to estimate the variation in and populations, but will also help test the validity of ancient DNA will not only give us a more detailed the rate of molecular evolution have been at the level at this taxonomic level. As a result, most estimates of evolution within species have not been possible because of the difficulty in assigning fossils to specific lineages should be measurably evolving. This expectation has

Fig. I (a coalescence event), and increases when sampled individuals are follow the number of lineages in the genealogy in each generation. This value decreases when two lineages share a common ancestor with large con.

Changing population size

ARTICLE IN PRESS

The coalescent is a variable process, so it often produces estimates of the demographic history of a large population and the shared ancestry of individuals subsequently extended to measurably evolving populations by Rodrigo populations sampled at one time point, the coalescent model has been applied to estimate the demographic history of natural populations. This approach is based on a population genetic model that assumes a constant population size and operates by the following processes: (1) mutation, (2) selection, (3) drift or by the strong directional selection exerted by host immune responses, resulting in a very fast rate of substitution of the order of 10^2 substitutions site y. Although these papers consider the substitution rate will be 23. Finally, heterochronous sequences contain independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple

Extra sequence (Box 2). Figure I illustrates the shared ancestry of individuals randomly sampled from it, as represented by a genealogical tree. This tree, in turn, determines the pattern of genetic diversity seen in sampled sequences (Box 2). The coalescent and measurably evolving populations can be estimated in calendar time units (generations or years). This estimation is based on the demographic function (e.g. constant size, exponential growth or logistic growth) that describes population-size change, and the likelihood of this function can be calculated given a distribution depends on a demographic function (e.g. constant size,

http://tree.trends.com

http://tree.trends.com

http://tree.trends.com

ARTICLE IN PRESS

The coalescent is a variable process, so it often produces estimates of the demographic history of a large population and the shared ancestry of individuals subsequently extended to measurably evolving populations by Rodrigo populations sampled at one time point, the coalescent model has been applied to estimate the demographic history of natural populations. This approach is based on a population genetic model that assumes a constant population size and operates by the following processes: (1) mutation, (2) selection, (3) drift or by the strong directional selection exerted by host immune responses, resulting in a very fast rate of substitution of the order of 10^2 substitutions site y. Although these papers consider the substitution rate will be 23. Finally, heterochronous sequences contain independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple

Extra sequence (Box 2). Figure I illustrates the shared ancestry of individuals randomly sampled from it, as represented by a genealogical tree. This tree, in turn, determines the pattern of genetic diversity seen in sampled sequences (Box 2). The coalescent and measurably evolving populations can be estimated in calendar time units (generations or years). This estimation is based on the demographic function (e.g. constant size, exponential growth or logistic growth) that describes population-size change, and the likelihood of this function can be calculated given a distribution depends on a demographic function (e.g. constant size,

Extra sequence (Box 2). Figure I illustrates the shared ancestry of individuals randomly sampled from it, as represented by a genealogical tree. This tree, in turn, determines the pattern of genetic diversity seen in sampled sequences (Box 2). The coalescent and measurably evolving populations can be estimated in calendar time units (generations or years). This estimation is based on the demographic function (e.g. constant size, exponential growth or logistic growth) that describes population-size change, and the likelihood of this function can be calculated given a distribution depends on a demographic function (e.g. constant size,
and populations, but will also help test the validity of ancient DNA. Will not only give us a more detailed picture of the difficulty in assigning fossils to specific lineages. Evolution within species have not been possible because sources of bacteria successful isolation of DNA from exceptionally ancient populations sampled at one time point, the coalescent model has been generalized by Griffiths and Felsenstein [46].

The coalescent describes the relationship between the demographic history of a large population and the shared ancestry of individuals randomly sampled from it, as represented by a genealogical tree. This tree, in turn, determines the pattern of genetic diversity seen in sampled sequences (Box 2). Figure I illustrates the shared ancestry of individuals from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process). This value decreases when two lineages share a common ancestor. Changing population size.

Fig. I

- Constant size
- Growing population

The coalescent model provides a probability distribution of times between the coalescence events in the sample genealogy. This distribution is a function of the length of time that has elapsed since the most recent common ancestor of the sample sequences, as well as the size and growth rate of the population. This distribution can be used to calculate the likelihood of observing a particular pattern of coalescence, and can be used to estimate the demographic history of the population.

The coalescent is a variable process, so it often produces estimates that are not consistent with the true demographic history of the population. This is because the coalescent model assumes that the population is large and that the mutation rate is low. In reality, populations are often small and the mutation rate is high, so the coalescent model may not be able to accurately estimate the demographic history of the population.

As a group, RNA viruses encompass such well known pathogens as HIV, influenza and foot and mouth disease, and are characterized by populations that continuously generate huge numbers of mutations owing to their large numbers, very short generation times and the error-prone nature of their replication machinery. These mutations are carried to fixation by random genetic drift or by the strong directional selection exerted by host immune responses, resulting in a very fast rate of measurable evolution.
Given a phylogeny, how can we learn about the evolutionary process that underlies it?

Generally, we want to know:

\[ p(\text{model} | \text{data}) \]

Bayes rule:

\[ p(\text{model} | \text{data}) \propto p(\text{data} | \text{model}) \, p(\text{model}) \]

Often referred to as:

posterior \propto \text{likelihood} \times \text{prior}
\( \lambda \) – coalescent model  \( D \) – sequence data
\( \tau \) – phylogeny  \( \mu \) – mutation model

In this case, we have:

\[
p(\lambda|\tau) \propto p(\tau|\lambda) \ p(\lambda)
\]

However, we don’t observe the tree directly:

\[
p(\tau, \mu|D) \propto p(D|\tau, \mu) \ p(\tau) \ p(\mu)
\]

We integrate over uncertainty:

\[
p(\lambda|D) \propto \int p(D|\tau, \mu) \ p(\tau|\lambda) \ p(\lambda) \ p(\mu) \ d\tau \ d\mu
\]
BEAST: Bayesian Evolutionary Analysis by Sampling Trees
Integration through Markov chain Monte Carlo
Integration through Markov chain Monte Carlo
Metropolis-Hastings algorithm

Starting from state $\theta$ propose a new state $\theta^*$. For the following, this proposal must to symmetric, i.e. $Q(\theta \to \theta^*) = Q(\theta^* \to \theta)$

If new state is more likely, always accept. If new state is less likely, accept with probability proportional to ratio of new state to old state.

Acceptance probability: $\min \left( 1, \frac{p(\theta^*)}{p(\theta)} \right)$

Simple example:

$p(x) = 0.2$  \quad  \quad  p(y) = 0.8$

$A(x \to y) = 0.8/0.2 = 1$  \quad  A(y \to x) = 0.2/0.8 = 0.25$

Mass moving from $x$ to $y$: $p(x) A(x \to y) = 0.2 \times 1 = 0.2$

Mass moving from $y$ to $x$: $p(y) A(y \to x) = 0.8 \times 0.25 = 0.2$
BEAST will produce samples from:

\[\lambda\] – coalescent model

\[\mu\] – mutation model

\[\tau\] – phylogeny
Use a ‘skyline’ demographic model
Use a ‘skyline’ demographic model
Practical part 1
Estimating $R_0$ from timeseries data

$r(0) = \beta - \gamma$

$r = 0.20$ per day for 1918 influenza

We know the approximate recovery rate

$\gamma \approx 0.25$

We can solve for $\beta$ and hence $R_0$

$\beta = r + \gamma \approx 0.45$

$R_0 = \frac{\beta}{\gamma} \approx \frac{0.45}{0.25} \approx 1.8$
Growth rate of pandemic H1N1

\[ r = 0.11 \text{ per day} \]
\[ \beta = 0.11 + 0.33 = 0.44 \text{ per day} \]
\[ R_0 = \frac{0.44}{0.33} = 1.33 \]
Generation time $\tau$ of infection

At the beginning of the epidemic, new infections emerge at rate $\beta$.

$$\tau = \frac{1}{2\beta S(0)} = \frac{1}{2 \times 0.36} = 1.39$$

Final susceptible fraction:

$$S(\infty) = e^{-R_0(1 - S(\infty))}$$

At the end of the epidemic:

$$\tau = \frac{1}{2\beta S(\infty)} = \frac{1}{2 \times 0.36 \times 0.84} = 1.65$$
Effective population sizes of flu vs measles

$N_e = 7.2$ years
$N_e = 1050$ infections (duration of infection of 5 days)
$N = 70$ million infections (prevalence)
Off by a factor of 6,700

$N_e = 124.6$ years
$N_e = 8270$ infections (duration of infection of 11 days)
$N = 0.9$ million infections (prevalence)
Off by a factor of 110
Practical part 2
Continuous time Markov chains (CTMCs)

\[ p_{t \to \infty}(A) = \frac{\mu_{BA}}{\mu_{AB} + \mu_{BA}} \]

\[ p_{t \to \infty}(B) = \frac{\mu_{AB}}{\mu_{AB} + \mu_{BA}} \]

\[ \mu_{AB} = 3 \quad q(A) = 0.25 \]

\[ \mu_{BA} = 1 \quad q(B) = 0.75 \]
CTMCs on trees

Transition matrix with $\mu_{AB} = 3$ $\mu_{BA} = 1$ $t = 0.2$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.59</td>
<td>0.41</td>
</tr>
<tr>
<td>B</td>
<td>0.14</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Integrate over internal states

Transition matrix with \( \mu_{AB} = 3 \) \( \mu_{BA} = 1 \) \( t = 0.2 \)

<table>
<thead>
<tr>
<th></th>
<th>( A )</th>
<th>( B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>0.59</td>
<td>0.41</td>
</tr>
<tr>
<td>( B )</td>
<td>0.14</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Integrate over internal states

Transition matrix with $\mu_{AB} = 3$ $\mu_{BA} = 1$ $t = 0.2$

Pr = 0.0211

Pr = 0.0036

Pr = 0.0073

Pr = 0.0109
Integrate over internal states

\[ p(D|\tau, \mu) = 0.0211 + 0.0073 + 0.0036 + 0.0109 = 0.0429 \]
Practical part 3