Coalescents and Genealogies

Jay Taylor

School of Mathematical and Statistical Sciences
Center for Evolutionary Medicine and Informatics
Arizona State University
mtDNA genealogy of humans, Neanderthals and Denisovans

(Krause et al., 2010)
Epidemiological processes shape pathogen genealogies

Grenfell et al. (2004): *Unifying the Epidemiological and Evolutionary Dynamics of Pathogens*
Pedigrees vs. Genealogies

- Pedigrees show genetic relationships between individuals. The number of ancestors increases backwards in time (two parents, four grand parents, etc.).
- Genealogies (gene genealogies) show how individual copies of the same gene/locus are evolutionarily related. The number of ancestors decreases backwards in time.
The Wright-Fisher Model: Birth and Death in Finite Populations

Assumptions:

1. Non-overlapping generations: adults reproduce and then die.
2. Constant adult population size: exactly $N$ haploid individuals survive to adulthood.
3. Neutrality: all individuals have the same fitness.
4. No mutation (for now).
5. Each individual alive in generation $t$ gives birth to a very large number of offspring, each of which is equally likely to survive to reproductive maturity in generation $t + 1$.

Equivalently: each individual alive in generation $t + 1$ independently chooses its parent from the preceding generation uniformly at random and with replacement.
Simulation of the Wright-Fisher Model \((N = 10)\)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model \((N = 10)\)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model \((N = 10)\)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model ($N = 10$)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model (N = 10)
Simulation of the Wright-Fisher Model ($N = 10$)
Simulation of the Wright-Fisher Model (N = 10)
Pairwise Coalescent Times

Suppose that two individuals are sampled at random and let $T_{mrca}$ be the number of generations back to their most recent common ancestor. Our objective is to determine the distribution of $T_{mrca}$.

We begin by calculating the probability that $T_{mrca} = 1$, i.e., that the two individuals share the same parent. Since there are $N$ possible parents, we have:

$$\mathbb{P}(T_{mrca} = 1) = \sum_{i=1}^{N} \mathbb{P}($$ they both have individual $i$ as a parent $$)$$

$$= \sum_{i=1}^{N} \frac{1}{N^2} = \frac{1}{N}.$$
To find the probability that $T_{mrca} = 2$, notice that two conditions must be satisfied:

1. The sampled individuals must have different parents: $T_{mrca} > 1$;
2. Those two parents must have the same parent.

Consequently, by conditioning on the first event, we obtain

$$
P(T_{mrca} = 2) = P(T_{mrca} > 1) \times P(T_{mrca} = 2 | T_{mrca} > 1)
\begin{align*}
&= \left(1 - \frac{1}{N}\right) \times \frac{1}{N}.
\end{align*}
$$
By similar arguments, we can show that for any \( t \geq 1 \), we have

\[
P(T_{mrca} = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}.
\]

In other words, \( T_{mrca} \) is a geometrically-distributed random variable with parameter \( 1/N \). Furthermore, the mean and the variance of \( T_{mrca} \) are:

\[
E[T_{mrca}] = N
\]
\[
Var(T_{mrca}) = N(N - 1)
\]
Distribution of Pairwise Coalescent Times

- Density vs. $T_{\text{mrca}}$
- Graph showing data for $N = 200$, $N = 500$, and $N = 1000$
- Box plots illustrating variation in $T_{\text{mrca}}$ for different $N$ values
Now suppose that we have sampled \( k > 2 \) individuals at random from the population. In this case, our task is to describe the distribution of the random genealogy \( G \) of the sample. As in the previous section, it suffices to consider what can happen in a single generation. There are three possibilities:

1. None of the \( k \) individuals share a common ancestor in the previous generation.
2. Exactly two of the \( k \) individuals share a common ancestor in the previous generation. In this case, we say that their lineages **coalesce**.
3. More than two of the individuals share a common ancestor in the previous generation.

Throughout we will assume that the population size \( N \) is large, say \( N > 100 \), and that the sample size \( k \) is small relative to \( N \), say \( k < N/10 \).
For none of the \( k \) individuals to share a common ancestor in the previous generation, each one must ‘choose’ a different parent. Since there are \( N \) choices for the first individual, \( N - 1 \) choices for the second individual, \( N - 2 \) choices for the third and so forth, the probability of this event is:

\[
P(\text{no common ancestors}) = \frac{N}{N} \frac{N - 1}{N} \frac{N - 2}{N} \cdots \frac{N - (k - 1)}{N}
\]

\[
= 1 \left( 1 - \frac{1}{N} \right) \left( 1 - \frac{2}{N} \right) \cdots \left( 1 - \frac{k - 1}{N} \right)
\]

\[
= 1 - \frac{k(k - 1)}{2} \frac{1}{N} + O \left( \frac{1}{N^2} \right).
\]
Similarly, for exactly two of the \( k \) individuals to share a common ancestor in the previous generation, those two must choose the same parent, whilst each of the remaining \( k - 2 \) individuals must choose a different parent. Since there are \( \binom{k}{2} = k(k - 1)/2 \) distinct pairs of sampled individuals, it follows that the probability of this event is:

\[
\mathbb{P}\text{(one pair of sibs)} = \frac{k(k - 1)}{2} \cdot \frac{1}{N} \cdot \frac{N - 1}{N} \cdot \frac{N - 2}{N} \cdots \frac{N - (k - 2)}{N} = \frac{k(k - 1)}{2} \cdot \frac{1}{N} + O\left(\frac{1}{N^2}\right).
\]

Furthermore, each pair of individuals is equally likely to be the one with the common ancestor.
For more than two individuals to share a common ancestor in the previous generation, at least one of the following events must occur:

- **Multiple merger:** three or more individuals share the same common ancestor;
- **Simultaneous mergers:** there are two or more distinct pairs of sibs.

However, it can be shown that the collective probability of all of these events is of order $O(1/N^2)$. Since this is small relative to $\binom{k}{2} \frac{1}{N}$ when $N$ is much larger than $k$, it can also be shown that the sample genealogy is unlikely to contain any complex mergers, i.e., $G$ is likely to be strictly bifurcating.
Let us write $\tau_k$ for the random time until some pair of lineages ancestral to our sample first finds a common ancestor. In the jargon of the field, we say that the two lineages coalesce. If we ignore complex mergers, then there are two possibilities in each generation:

1. With probability $\binom{k}{2} \frac{1}{N}$, some pair of lineages coalesces. Furthermore, that pair is chosen uniformly at random from among the $\binom{k}{2}$ pairs.

2. Otherwise, with probability $1 - \binom{k}{2} \frac{1}{N}$, each ancestral lineage is extended one more generation into the past.

Since the parent-child relationships are chosen independently in each generation, it follows that $\tau_k$ is a geometric random variable with parameter $\binom{k}{2} \frac{1}{N}$, i.e.,

**Distribution of $\tau_n$**

$$\mathbb{P}(\tau_k = t) = \left(1 - \binom{k}{2} \frac{1}{N}\right)^{t-1} \binom{k}{2} \frac{1}{N}.$$
So far we have described the genealogy up until the first pairwise coalescent event. However, because of the inter-generational independence of the Wright-Fisher model, this actually characterizes the entire genealogy:

1. The first coalescent event occurs after time $\tau_k$, at which point the number of lineages ancestral to the sample is reduced from $k$ to $k - 1$.

2. The next coalescent event occurs after an additional time $\tau_{k-1}$, which is independent of $\tau_k$ and which has distribution

$$\mathbb{P}(\tau_{k-1} = t) \approx \left(1 - \binom{k-1}{2} \frac{1}{N}\right)^{t-1} \binom{k-1}{2} \frac{1}{N}.$$ 

3. In general, once the genealogy contains $i$ ancestral lineages, the time until the next pairwise coalescent event is $\tau_i$, with distribution

$$\mathbb{P}(\tau_i = t) \approx \left(1 - \binom{i}{2} \frac{1}{N}\right)^{t-1} \binom{i}{2} \frac{1}{N}.$$ 

Furthermore, $\tau_i$ is independent of $\tau_k, \tau_{k-1}, \ldots, \tau_{i+1}$.

4. The process terminates when only one ancestral lineage remains.
There is another approximation that can be made when $N$ is large. In this case, the discreteness of the coalescent time $\tau_k$ is not important and we can approximate its distribution by an exponential distribution:

\[
P(\tau_k > t) \approx \exp \left( - \frac{k}{2} \frac{t}{N} \right).
\]

In fact, if we let $\tau_k^{(N)} = \frac{1}{N} \tau_k$ be the coalescent time measured in units of $N$ generations, then it can be shown that the variables $\tau_k^{(N)}$ converge in distribution to an exponential variable:

\[
\lim_{N \to \infty} P(\tau_k^{(N)} > t) = \exp \left( - \frac{k}{2} t \right).
\]
These observations lead to the following approximate description of the genealogy of a random sample of $k$ individuals from a haploid population of size $N$. As shown in Kingman (1982a, b), this description is exact in the limit as $N \to \infty$.

**Kingman’s Coalescent**

The sample genealogy $G_n$ can be realized by running the following continuous-time Markov chain until only one ancestral lineage remains.

1. Simulate a random time $\tau_k$ for the first coalescent event using the exponential distribution with rate $\binom{k}{2}/N$:

$$
\mathbb{P}(\tau_k > t) = \exp \left( -\binom{k}{2} \frac{t}{N} \right)
$$

2. Choose two of the individuals in the sample uniformly at random and draw the top part of the tree so that these two share a MRCA at time $\tau_k$ in the past.

3. If $k > 1$, then replace $k$ by $k - 1$ and go back to step 1. If $k = 1$, then the MRCA has been reached and the process terminates.
Random genealogies for samples of 20 individuals
Although genealogies generated under Kingman’s coalescent are random, they tend to have a characteristic shape in which recent branches are much shorter than ancient branches. Here are two examples for a sample containing $k = 100$ individuals:
The shape of these trees is a consequence of the combinatorial properties of Kingman’s coalescent. In particular, the mean branch length is inversely proportional to the square of the number of coexisting lineages.

### Branch length statistics

\[
\mathbb{E}[\tau_k] = \frac{2N}{k(k - 1)} \\
\text{Var}(\tau_k) = \frac{4N^2}{k^2(k - 1)^2}
\]
The quadratic decrease in the branch length has a profound influence on the distribution of the time to the most recent common ancestor, which is equal to the sum of the branch lengths:

\[ T_{\text{mrca}}^{(k)} = \tau_k + \tau_{k-1} + \cdots + \tau_2 \]

In fact, it can be shown that the variables \( \tau_k \) decrease so rapidly that both the mean and the variance of \( T_{\text{mrca}}^{(k)}/N \) have finite positive limits as \( k \to \infty \).

**Mean and variance of \( T_{\text{mrca}}^{(k)} \)**

\[
\begin{align*}
\mathbb{E}[T_{\text{mrca}}^{(k)}] &= 2N \left( 1 - \frac{1}{k} \right) \quad k \to \infty \quad 2N \\
\text{Var}(T_{\text{mrca}}^{(k)}) &= 4N^2 \sum_{i=2}^{k} \frac{1}{i^2(i-1)^2} \quad k \to \infty \quad 1.16N^2.
\end{align*}
\]
The rapid coalescence of terminal branches also means that the variability of $T_{mrca}^{(n)}$ is not appreciably reduced once the sample size exceeds 20.

![Graph showing the relationship between sample size and $T_{mrca} / N$.]
Kingman’s Coalescent: Summary of Properties

1. Kingman’s coalescent is a continuous-time Markov chain which characterizes the genealogy of a random sample from a Wright-Fisher population.

2. The rate of coalescence is inversely proportional to the population size $N$.

3. Because the rate of coalescence is proportional to the square of the number of branches in the genealogy, branches near the top of the tree tend to be much shorter than branches near the base.

4. The time to the most recent common ancestor only weakly depends on the sample size.

5. Because the mean and the standard deviation of $T^{(n)}_{\text{mrca}}$ are of comparable magnitude, it is not uncommon for unlinked loci to have very different times to their most recent common ancestors.
Robustness and the Scope of Kingman’s Coalescent

Although it was convenient to use the Wright-Fisher model in the derivation, the importance of Kingman’s coalescent stems from the fact that it is relevant to a much larger class of population genetical models satisfying these conditions:

1. The population size is constant.
2. The population is panmictic.
3. Evolution is neutral.
4. The within-generation fecundity variance is not too large.
5. The sample size is small relative to the population size.

Under these conditions, it can be shown that by choosing an appropriate scaling of time the distribution of the genealogy of a random sample can be approximated by Kingman’s coalescent. What changes is the relationship between the pairwise coalescent rate and the population size.
We first consider a class of models that combine the generality of the reproductive mechanism of a Galton-Watson branching process with the population regulation step of the Wright-Fisher model.

1. Non-overlapping generations: adults reproduce and then die.
2. Constant adult population size: $N$ haploid individuals survive to adulthood.
3. The number of offspring born to the $i$’th individual alive in generation $t$ is a random variable $\nu_{i,t}$. We will assume that these variables are independent and identically distributed with distribution $\nu$.
4. $N$ of the offspring are sampled uniformly at random and without replacement to form the next generation.
Two Examples

$\nu(1) = 0.9, \nu(2) = 0.1$  
$\nu(0) = 0.75, \nu(5) = 0.25$
Two Examples

\( \nu(1) = 0.9, \nu(2) = 0.1 \)

\( \nu(0) = 0.75, \nu(5) = 0.25 \)
Two Examples

\[ \nu(1) = 0.9, \nu(2) = 0.1 \]

\[ \nu(0) = 0.75, \nu(5) = 0.25 \]
Two Examples

\( \nu(1) = 0.9, \nu(2) = 0.1 \)  \hspace{1cm}  \( \nu(0) = 0.75, \nu(5) = 0.25 \)
Two Examples

\[ \nu(1) = 0.9, \nu(2) = 0.1 \]

\[ \nu(0) = 0.75, \nu(5) = 0.25 \]
Two Examples

$\nu(1) = 0.9, \nu(2) = 0.1$

$\nu(0) = 0.75, \nu(5) = 0.25$
Two Examples

\[ \nu(1) = 0.9, \nu(2) = 0.1 \]

\[ \nu(0) = 0.75, \nu(5) = 0.25 \]
Robustness and Effective Population Size

Two Examples

\( \nu(1) = 0.9, \nu(2) = 0.1 \)

\( \nu(0) = 0.75, \nu(5) = 0.25 \)
Two Examples

\[ \nu(1) = 0.9, \nu(2) = 0.1 \]

\[ \nu(0) = 0.75, \nu(5) = 0.25 \]
Two Examples

\[ \nu(1) = 0.9, \nu(2) = 0.1 \]

\[ \nu(0) = 0.75, \nu(5) = 0.25 \]
Two Examples

\[ \nu(1) = 0.9, \nu(2) = 0.1 \]

\[ \nu(0) = 0.75, \nu(5) = 0.25 \]
Two Examples

\[ \nu(1) = 0.9, \nu(2) = 0.1 \]

\[ \nu(0) = 0.75, \nu(5) = 0.25 \]
Robustness and Effective Population Size

Two Examples

\[ \nu(1) = 0.9, \nu(2) = 0.1 \]

\[ \nu(0) = 0.75, \nu(5) = 0.25 \]
Robustness and Effective Population Size

Two Examples

\[ \nu(1) = 0.9, \nu(2) = 0.1 \]

\[ \nu(0) = 0.75, \nu(5) = 0.25 \]
Provided that the variance of $\nu$ is finite, it can be shown that for large $N$ the genealogy of a random sample can be approximated by Kingman’s coalescent. However, the distribution of the time until the next coalescent event when there are $k$ lineages is

$$\mathbb{P}(\tau_k > t) \approx \exp \left( -\binom{k}{2} \frac{t}{N e} \right)$$

where the **effective population size** $N_e$ is

**Effective population size of a regulated Galton-Watson model**

$$N_e \equiv \frac{N}{\left( 1 - \frac{1}{m} + \frac{\sigma^2}{m} \right)} \quad \text{with} \quad m = \mathbb{E}[\nu] \quad \text{and} \quad \sigma^2 = \text{Var}(\nu).$$

Thus, as the within-generation fecundity variance $\sigma^2$ increases, the time between coalescent events will tend to decrease. Note, however, that the coalescent rate is still inversely proportional to the census population size $N$. 
The Moran Model

The Moran model is a continuous-time model of birth and death in a finite population with overlapping generations. It makes the following assumptions:

1. The population contains $N$ haploid individuals.
2. Each individual reproduces independently of the others at rate 1.
3. When an individual reproduces, they give birth to a single offspring. To keep the population size constant, one of the remaining $N-1$ individuals is chosen uniformly at random and removed from the population (i.e., dies).

For this model, the genealogy of a random sample is exactly described by Kingman’s coalescent, even when $N$ is finite. However, in this case, the effective population size is $N_e = N/2$:

$$\mathbb{P}(\tau_k > t) = \exp \left( - \frac{k}{2} \frac{2t}{N} \right).$$
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model ($N = 10$)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model ($N = 10$)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
Simulation of the Moran Model (N = 10)
To understand why the effective population size of the Moran model is half the census population size, let $\nu$ be the random number of offspring produced by an individual over their lifetime and let $T$ be this lifetime. Then

1. $T$ is exponentially distributed with mean and variance equal to 1.

2. Conditional on $T$, $\nu$ is Poisson distributed with mean and variance $\mathbb{E}[\nu|T] = \text{Var}(\nu|T) = T$.

3. The average number of offspring produced by an individual over their lifetime can be calculated using the law of total expectation:

$$m = \mathbb{E}[\nu] = \mathbb{E}[\mathbb{E}[\nu|T]] = \mathbb{E}[T] = 1.$$ 

4. Similarly, the variance of $\nu$ can be calculated using the law of total variance:

$$\sigma^2 = \text{Var}(\nu) = \mathbb{E}[\text{Var}(\nu|T)] + \text{Var}(\mathbb{E}[\nu|T])$$

$$= \mathbb{E}[T] + \text{Var}(T)$$

$$= 2.$$
The formula derived for the effective population size of a regulated Galton-Watson process is also valid in this case. Since $m = 1$ and $\sigma^2 = 2$, we have:

\[
N_e = \frac{N}{\left(1 - \frac{1}{m} + \frac{\sigma^2}{m}\right)} = \frac{N}{\left(1 - 1 + 2\right)} = \frac{N}{2}.
\]

In this case, the increased fecundity variance is a consequence of the variance of the life span: some individuals leave more offspring than others because they live longer.

In general, $N_e$ depends on a multitude of life-history traits, including longevity, fecundity, age-structuring and sex ratio.
Coalescents in Large Metapopulations

Under some conditions, Kingman’s coalescent can be used to model genealogies from subdivided populations (Wakeley & Aliacar 2001). Consider the following version of Wright’s island model:

1. The population is subdivided into $D$ subpopulations, each containing $N$ haploid individuals.
2. Each individual reproduces at rate 1, independently of all others, and gives birth to a single offspring.
3. With probability $(1 - m)$, the offspring remains within the subpopulation where it was born, in which case one of the $N$ adults in that subpopulation is chosen at random and dies.
4. Otherwise, with probability $m$, the offspring migrates to another subpopulation, which is chosen uniformly at random, where it replaces a random chosen adult.

The quantity $m$ can be thought of as the migration rate.
A Metapopulation with 15 Demes
Kingman’s coalescent will provide an accurate approximation to the genealogy of a random sample from a metapopulation so long as the following conditions are satisfied:

1. Each lineage must be sampled from a separate deme;
2. The number of demes must be much greater than the (effective) population size within any single deme \((D \gg N)\).

Together, these two conditions lead to a simplification of the genealogical process for this model because they result in a separation of time scales: whilst coalescence occurs rapidly between lineages that occupy the same deme \((O(N)\) generations), the process gathering lineages into the same deme is much slower \((O(D)\) generations).
The rate of coalescence in the metapopulation is inversely proportional to:

### Effective population size in the island model

\[ N_e = D \left( N + \frac{1 - m}{m} \right) \geq DN. \]

The magnitude of \( N_e \) depends on whether \( Nm \) is less than or greater than 1.

- When \( Nm \gg 1 \), the population is nearly panmictic and \( N_e \approx DN \).
- When \( Nm \ll 1 \), the demes are isolated and \( N_e \approx D/m \).
Kingman’s coalescent can also be extended to models in which the effective population size changes over time. The main requirement is that the population size is never very small, say $N_e(t) > 20$.

- In this case, the ancestral process is a **time-inhomogeneous continuous time Markov chain**.
- The rate of coalescence of lineages at time $t$ in the past is inversely proportional to the population size $N_e(t)$ at that time. In particular,
  \[ P(t \leq \tau_k < t + \delta t) = \binom{k}{2} \frac{\delta t}{N_e(t)}. \]
- Branches tend to be shorter during periods when the population size was historically small.
Random genealogies for samples of 20 individuals when $N_e(t) = N_0 e^{-2t}$
Random genealogies for samples of 100 individuals when $N_e(t) = N_0 e^{-2t}$
The relationship between ancestral effective population size and coalescent rate can be used to infer the demographic history of a population from genetic sequence data. Both parametric and semi-parametric approaches are available.

- Parametric approaches assume a parametric model for the demography (e.g., linear, exponential or logistic population growth) and then jointly infer the genealogy $G$ and the unknown parameters $\theta_{dem}$.

- Semi-parametric approaches include the classical skyline plot (Pybus et al., 2005) and Bayesian skyline plots (Drummond et al. 2005), which allow for either a fixed or variable number of changes of population size that coincide with inferred coalescent events.

- Semi-parametric approaches are more flexible and allow for non-monotonic changes but can also be more computationally demanding.
In its original conception, the skyline plot uses a method-of-moments approach to estimate ancestral population sizes:

\[ \mathbb{E}[\tau_k] = \frac{N_k}{\binom{k}{2}} \rightarrow \hat{N}_k = \binom{k}{2} \tau_k. \]

- Population sizes between coalescent events can either be constant or linearly interpolated.
- In a classic skyline plot, the genealogy is inferred first and then used to estimate the genealogy.
- In Bayesian skyline plots, both the genealogy and the demography are inferred together.
Bayesian skyline plot for a global sample of *P. vivax* mtDNA genomes.

Taylor et al. (2013)
Genealogical inference of demographic history is subject to several important limitations:

1. Estimates derived from a single locus are usually very imprecise.
2. The relationship between the effective population size inferred using genealogical methods and the census population size is usually unknown.
3. Estimates of the effective population size are also confounded by the generation time, which might vary over time.
4. These issues are especially problematic for parasites, where the effective population size and generation time depend on the prevalence and the incidence of the infection, respectively (Frost & Volz 2010, Volz 2012, Koelle & Rasmussen 2012).
5. The relative importance of selection and genetic drift is still unresolved and probably varies greatly between species with different ecologies.
Bayesian Phylogenetics

Bayesian methods have proven to be especially useful in the analysis of genetic sequence data. In these problems, the unknown parameters can often be divided into three categories:

- the unknown tree, $T$
- the parameters of the demographic model, $\Theta_{dem}$
- the parameters of the substitution model, $\Theta_{subst}$.

Then, given a sequence alignment $D$, the analytical problem is to calculate the posterior distribution of all of the unknowns:

\[
p(T, \Theta_{dem}, \Theta_{subst} | D) = p(T, \Theta_{dem}, \Theta_{subst}) \cdot \left( \frac{p(D | T, \Theta_{dem}, \Theta_{subst})}{p(D)} \right)
\]
If the substitution process is assumed to be neutral, then the prior distribution and the likelihood functions can be simplified:

**Bayes’ formula for phylogenetic inference with neutral data**

\[
p(\mathcal{T}, \Theta_{subst}, \Theta_{dem}|D) = p(\Theta_{subst}) \cdot p(\Theta_{dem}) \cdot p(\mathcal{T}|\Theta_{dem}) \cdot \left( \frac{p(D|\mathcal{T}, \Theta_{subst})}{p(D)} \right)
\]

This is a consequence of the following assumptions:

- Under the prior distribution, the parameters of the substitution process \( \Theta_{subst} \) are independent of the tree \( \mathcal{T} \) and the demographic parameters \( \Theta_{dem} \).
- The demographic parameters \( \Theta_{dem} \) typically determine the conditional distribution of the genealogy \( \mathcal{T} \), e.g., through a coalescent model.
- Conditional on \( \mathcal{T} \) and \( \Theta_{subst} \), the sequence data \( D \) is independent of \( \Theta_{dem} \).
Example: Bayesian Inference of Effective Population Size

Suppose that our data $D$ consists of $n$ randomly sampled individuals that have been sequenced at a neutral locus and that our objective is to estimate the effective population size $N_e$. For simplicity, we will use the Jukes-Cantor model for the substitution process with a strict molecular clock and we will assume that the demography can be described by the constant population size coalescent.

To carry out a Bayesian analysis, we need to specify $p(\mu)$, $p(N_e)$ and $p(T|N_e)$.

- $p(\mu)$ should be chosen to reflect what we know about the mutation rate, e.g., we could use a lognormal distribution with mean $u$ and variance $\sigma^2$.
- Since $N_e$ is a scale parameter in the coalescent, it is common practice to use the Jeffreys prior $p(N_e) \propto 1/N_e$.
- $p(T|N_e)$ is then determined by Kingman’s coalescent.
Assuming that we are only interested in $N_e$, then $\mu$ and $\mathcal{T}$ are nuisance parameters and so we need to calculate the marginal posterior distribution of $N_e$ by integrating over $\mu$ and $\mathcal{T}$:

$$p(N_e|D) = \int \int p(\mathcal{T}, N_e, \mu|D)d\mu d\mathcal{T}$$

$$= \frac{p(N_e)}{p(D)} \int \int p(D|\mathcal{T}, \mu)p(\mu)p(\mathcal{T}|N_E)d\mu d\mathcal{T}. $$

However, unless $n$ is quite small, integration over $\mathcal{T}$ is not feasible. For example, if our sample contains 20 sequences, then there are approximately $8 \times 10^{21}$ possible trees to be considered. Even with the fastest computers available, this is an impossible calculation.
Until recently, Bayesian methods were regarded as impractical for many problems because of the computational difficulty of evaluating the posterior distribution. In particular, to use Bayes’ formula,

\[ p(\theta|D) = p(\theta) \cdot \left( \frac{p(D|\theta)}{p(D)} \right), \]

we need to evaluate the marginal probability of the data

\[ p(D) = \int p(D|\theta)p(\theta)d\theta, \]

which requires integration of the likelihood function over the parameter space. Except in special cases, this integration must be performed numerically, but sometimes even this is very difficult.
An alternative is to use Monte Carlo methods to sample from the posterior distribution. Here the idea is to generate a random sample from the distribution and then use the empirical distribution of that sample to approximate $p(\theta|D)$:

$$
p(\theta|D) \approx \frac{1}{N} \sum_{i=1}^{N} \delta_{\Theta_i} \quad \text{where} \quad \Theta_1, \ldots, \Theta_N \sim p(\theta|D)
$$

For example, the figure shows two histogram estimators for the Beta(8, 4) density generated using either 100 (left) or 1000 (right) independent samples:
In particular, the empirical distribution can be used to estimate probabilities and expectations under $p(\theta|D)$:

$$
\mathbb{P}(\theta \in A|D) \approx \frac{1}{N} \sum_{i=1}^{N} 1_{A}(\Theta_i)
$$

$$
\mathbb{E}[f(\theta)|D] \approx \frac{1}{N} \sum_{i=1}^{N} f(\Theta_i).
$$

What makes this approach difficult is the need to generate random samples from distributions that are only known up to a constant of proportionality (e.g., $p(D)$). This is where Markov chain Monte Carlo methods come in.
Markov Chains

A discrete-time Markov chain is a stochastic process $X_0, X_1, \cdots$ with the property that the future behavior of the process only depends on its current state. More precisely, this means that for every set $A$ and $t, s \geq 0$,

$$
\mathbb{P}(X_{t+s} \in A | \underbrace{X_t, X_{t-1}, \cdots, X_0}_{\text{future}}, \underbrace{X_t, X_{t-1}, \cdots}_{\text{present}}, X_0) = \mathbb{P}(X_{t+s} \in A | X_t).
$$

Markov chains have numerous applications in biology. Some familiar examples include:

- random walks
- branching processes
- the Wright-Fisher model
- chain-binomial models (Reed-Frost)
One consequence of the **Markov property** is that many Markov chains have a tendency to ‘forget’ their initial state as time progresses. More precisely, 

**Asymptotic behavior of Markov chains**

Many Markov chains have the property that there is a probability distribution $\pi$, called the **stationary distribution** of the chain, such that for large $t$ the distribution of $X_t$ approaches $\pi$, i.e., for all states $x$ and sets $A$,

$$\lim_{t \to \infty} \mathbb{P}(X_t \in A|X_0 = x) = \pi(A).$$

**Stationary behavior of the Wright-Fisher process:**

$(N = 100, \mu = 0.02)$
Some Markov chains satisfy an even stronger property, called **ergodicity**.

**Ergodicity**

A Markov chain with stationary distribution $\pi$ is said to be **ergodic** if for every initial state $X_0 = x$ and every set $A$, we have

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} 1_A(X_t) = \pi(A).$$

In other words, if we run the chain for a long time, then the proportion of time spent visiting the set $A$ is approximately $\pi(A)$.

Ergodic behavior of the Wright-Fisher process:

$(N = 100, \mu = 0.02)$
Markov Chain Monte Carlo: General Approach

The central idea in Markov Chain Monte Carlo is to use an ergodic Markov chain to generate a random sample from the target distribution $\pi$.

1. The first step is to select an ergodic Markov chain which has $\pi$ as its stationary distribution. There are different methods for doing this, including the Metropolis-Hastings algorithm and the Gibbs Sampler.

2. We then need to simulate the Markov chain until the distribution is close to the target distribution. This initial period is often called the burn-in period.

3. We continue simulating the chain, but because successive values are highly correlated, it is common practice to only collect a sample every $T$ generations, so as to reduce the correlations between the sampled states (thinning).

4. We can then use these samples to approximate the target distribution:

\[ \frac{1}{N} \sum_{n=1}^{N} \delta_{X_{B+nT}} \approx \pi. \]
The Metropolis-Hastings Algorithm

Given a probability distribution $\pi$, the Metropolis-Hastings algorithm can be used to explicitly construct a Markov chain that has $\pi$ as its stationary distribution.

Implementation requires the following elements:

- the **target distribution**, $\pi$, known up to a constant of proportionality;
- a family of **proposal distributions**, $Q(y|x)$, and a way to efficiently sample from these distributions.

The MH algorithm is based on a more general idea known as **rejection sampling**. Instead of sampling directly from $\pi$, we propose values using a distribution $Q(y|x)$ that we can easily sample but then we reject values that are unlikely under $\pi$. 
The Metropolis-Hastings algorithm consists of repeated application of the following three steps. Suppose that $X_n = x$ is the current state of the Markov chain. Then the next state is chosen as follows:

**Step 1:** We first propose a new value for the chain by sampling $y$ with probability $Q(y|x)$.

**Step 2:** We then calculate the acceptance probability of the new state:

$$\alpha(x; y) = \min \left\{ \frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)}, 1 \right\}$$

**Step 3:** With probability $\alpha(x; y)$, set $X_{n+1} = y$. Otherwise, set $X_{n+1} = x$.

**Remark**

Because $\pi$ enters into $\alpha$ as a ratio $\pi(y)/\pi(x)$, we only need to know $\pi$ up to a constant of proportionality. This is why the MH algorithm is so well suited for Bayesian analysis.
The choice of the proposal distribution $Q$ can have a profound impact on the performance of the MH algorithm. While there is no universal procedure for selecting a ‘good’ proposal distribution, the following considerations are important.

- $Q$ should be chosen so that the chain rapidly converges to its stationary distribution.
- $Q$ should also be chosen so that it is easy to sample from.
- There is usually a tradeoff between these two conditions and sometimes it is necessary to try out different proposal distributions to identify one with good properties.

Many implementations of MH (e.g., BEAST, MIGRATE) offer the user some control over the proposal distribution.
One of the most challenging issues in MCMC is knowing for how long to run the chain. There are two related considerations.

1. We need to run the chain until its distribution is sufficiently close to the target distribution (convergence).
2. We then need to collect a large enough number of samples that we can estimate any quantities of interest (e.g., the mean TMRCA) sufficiently accurately (mixing).

Unfortunately, there is no universally-valid, fool-proof way to guarantee that either one of these conditions is satisfied. However, there are a number of convergence diagnostics that can indicate when there are problems.
Convergence Diagnostics: Trace Plots

Trace plots show how the value of a parameter changes over the course of a simulation. In general, what we want to see is that the mean and the variance of the parameter are fairly constant over the duration of the trace plot, as in the two examples shown below.
Problems with convergence or mixing may be revealed by trends or sudden changes in the behavior of the trace plot.

The increasing trend indicates that the chain has not yet converged. The sudden changes in mean indicate that the chain is poorly mixing.
Trace Plots: Some Guidelines

1. You should examine the trace plot of every parameter of interest, including the likelihood and the posterior probability. **If any of the trace plots look problematic, then all of the results are suspect.**

2. The fact that a trace plot appears to have converged is not conclusive proof that it has. Especially in high-dimensional problems, a chain that appears to be stationary for the first 500 million generations may well show a sudden change in behavior in the next.

3. The program **Tracer** (http://tree.bio.ed.ac.uk/software/tracer/) can be used to display and analyze trace plots generated by BEAST, MrBayes and LAMARC.
Convergence Diagnostics: Effective Sample Size

Because successive states visited by a Markov chain are correlated, an estimate derived using $N$ values generated by such a chain will usually be less precise than an estimate derived using $N$ independent samples. This motivates the following definition.

Effective Sample Size (ESS)

The effective sample size of a sample of $N$ correlated random variables is equal to the number of independent samples that would estimate the mean with the same variance.

For a stationary Markov chain with autocorrelation coefficients $\rho_k$, the ESS of $N$ successive samples is equal to

$$ESS = \frac{N}{1 + 2 \sum_{k=1}^{\infty} \rho_k}.$$
Effective Sample Size: Guidelines

1. Each parameter has its own ESS and these can differ between parameters by more than order of magnitude.

2. Parameters with small ESS’s indicate that a chain either has not converged or is slowly mixing. As a rule of thumb, the ESS of every parameter should exceed 1000 and larger values are even better.

3. Thinning by itself will not increase the ESS. However, we can increase the ESS by simultaneously thinning and increasing the duration of the chain, e.g., collecting a 1000 samples from a chain lasting 100000 generations is better than collecting a 1000 samples from a chain lasting 10000 generations.

4. The ESS of a parameter can usually only be estimated from its trace. For this reason, large ESS’s do not guarantee that the chain has converged.


