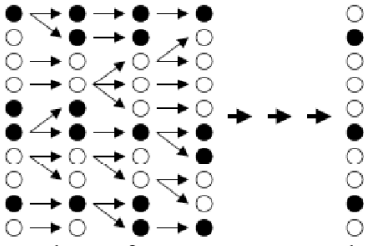
Contents

Part V
The Coalescent

1 Introduction

1.1 Moving Backwards

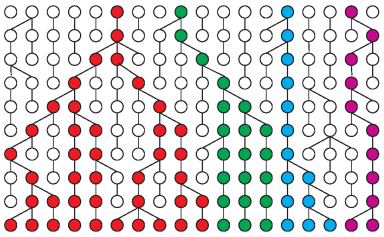
Wright-Fisher Model



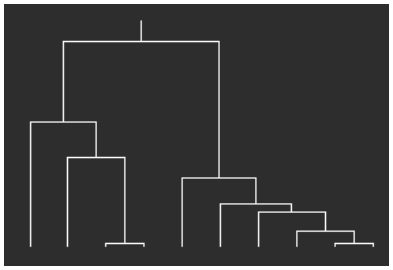
Our old friend the Wright-Fisher model envisions populations moving forward in time, each generation sampling allele counts X_t according to a Binomial distribution

$$P\left(X_{t+1} = j \mid X_t = i\right) = \frac{(2N)!}{j!(2N-j)!}\left(\frac{i}{2N}\right)^j\left(1-\frac{i}{2N}\right)^{2N-j}$$

Coalescent Model



- Two alleles are IBD with respect to the preceding generation with probability $\frac{1}{2N_e}$.
- By chance some alleles are copied multiple times, and some are not copied at all.
- Therefore, the number of alleles in generation $t-1$ that have descendents in generation t is always less than or equal to the number of alleles in generation $t-1$.
- Moving back in time, the number of alleles with surviving descendents is dwindling. Lineages, traced back in time, *coalesce* again and again until only one common ancestor (the MRCA) exists of all extant alleles.



There is a coalescent tree that gives rise to the sample of n sequences.

There are dependencies among all individuals in the sample. In other words, individuals sampled from a population are *not* independent.

Often little is gained by increasing the sample size. Depending on the strength of the dependencies, there may be little new information provided by an additional sampled individual.

Usefulness of the Coalescent Model

- Powerful simulation tool: we don't need to simulate the *whole* population, just the parts that left descendents.
- Hypothesis testing and statistical estimation of demographic parameters from molecular data: we have already inferred some connections between genetic variation and population demographics (e.g. $f = \frac{1}{1+4N_e\mu}$), but the coalescent has facilitated a veritable explosion of new theory to match the ever-growing piles of molecular data.

Failing to accommodate the statistical variation introduced by the randomness of the "experiment" can easily lead to over-interpretation of the data. In other words, if you just account for statistical sampling variation, but not genetic sampling variation, your confidence in conclusions will be inflated.

The coalescent is a stochastic model of genetic transmission in populations. It was invented in 1980, extends the classical genetic models that you have been studying, and provides a convenient and simple framework for explicitly modeling this pesky statistical variation.

While it may be premature to assess its importance yet, it has been called one of the single greatest advances in genetics and is often cited as the latest good example of how biology can profoundly benefit from mathematical and statistical techniques/approaches.

Quote from Nordborg

> I consider a basic understanding of coalescent theory to be extremely valuable – even essential – for anyone analyzing genetic polymorphism data from populations... When intuition is not enough, the coalescent provides a simple and powerful tool for exploratory data analysis through the generation of simulated data. Comparison of observed data with data simulated under various assumptions can give considerable insight.
>
> – Nordborg, 2001

Relation to Phylogenetic Trees

Phylogenetic methods use (mostly) sequence data to infer the evolutionary relationships between taxa.

Phylogenetic methods assume a model for the mutations that occur in molecular sequences, but they generally assume no model for the shape of the tree *except* that it is bifurcating and perhaps satisfies the molecular clock assumption. (Note, some Bayesian methods assume a "coalescent prior" on tree shape.)

If the Wright-Fisher model assumptions apply, then the phylogenetic relationships are those imposed by the coalescent tree, so the phylogeny inferred is an estimate of the coalescent tree.

However, note that most phylogenies are inferred for multiple species. The coalescent assumes that all sampled individuals come from the same population, where all lineages can merge with all others, but lineages

from different species cannot merge (if species are taken to be organisms that cannot mate). Modifications of the coalescent model are needed to handle species level evolution.

In short, *the coalescent is a model for the expected shape and characteristics of phylogenetic trees that connects population demographics (and related parameters, e.g. N_e) to phylogenies*.

Variations on the Coalescent

- **Mutation.** Selectively neutral mutations (i.e. those that are not selected) do not impact the reproduction process, i.e. they happen independent of the coalescent events. They only depend on the amount of time that has passed (i.e. more time, more opportunities for mutation). Thus mutations can be added very easily conditional on a coalescent tree.
- **Reproduction process.** Random variation in reproductive success, skewed sex ratios, age structure (where individuals are not all the same age and reproduce at different rates conditional on age) change the rate of coalescent. We have analyzed some of these and found they only alter the process through the *effective population size* N_e.
- **Population size changes.** Population structure, growth and decline change the shape of the coalescent tree.
- **Recombination.** Produces a random graph, rather than a random coalescent tree.
- **Selection.** The tricky one, for now lineages no longer pick their parents at random. Instead, they tend to pick those parents which are more fit.

Coalescent, Mutation and Molecular Data

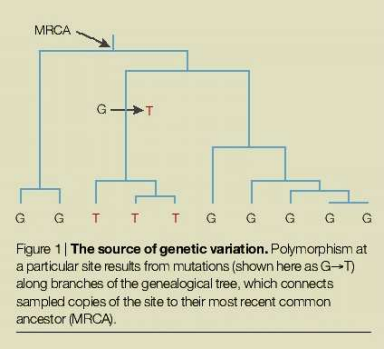


Figure 1 | **The source of genetic variation.** Polymorphism at a particular site results from mutations (shown here as G→T) along branches of the genealogical tree, which connects sampled copies of the site to their most recent common ancestor (MRCA).

Pattern of polymorphism at a single site in the genome depends on the historical pattern of *coalescence* and mutation.

No variation could mean strong *purifying selection* or a sample of highly related individuals.

Figure from Rosenberg and Nordborg (2002).

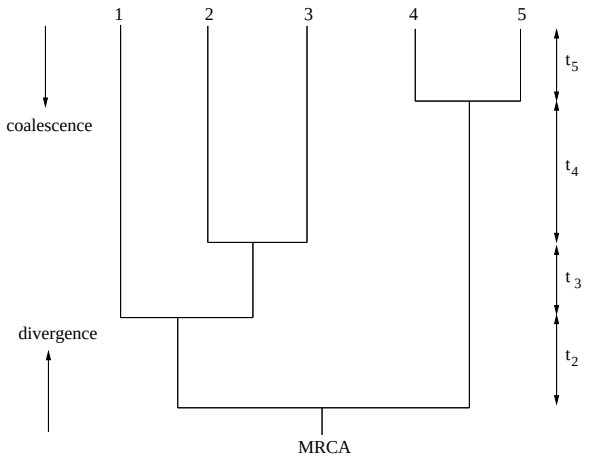
Introductory Coalescent Vocabulary

You sample a collection of *lineages*. A lineage is synonymous with an allele or a haplotype (if multiple loci have been sampled). Sampled lineages, under some circumstances, can be viewed as randomly selecting their parent lineages from the previous generation.

A *coalescent* event occurs when two sampled lineages select the same parent lineage, for that means they were on the same chromosome in the preceding generation. Eventually, all lineages will coalesce into the *most recent common ancestor* (MRCA).

The rate of coalescence (and therefore the shape of the tree) depends on the many factors. For example, if there are more lineages, there are more opportunities for two lineages to pick the same parents. Or if the population size is smaller (where the population size determines how many parents are available to choose from the preceding generation), then coalescent events will occur more frequently.

Notation and Terminology for a Coalescent Tree



The coalescent time t_i indicates the amount of time in the history of these sequences that i sample lineages persist.

1.2 Coalescent Simulation

Coalescent and Simulation

The genetics of infinite populations are well understood. The difficulty is understanding finite populations. Trying to understand finite populations led us to lots and lots of approximations (e.g. Wright diffusion model).

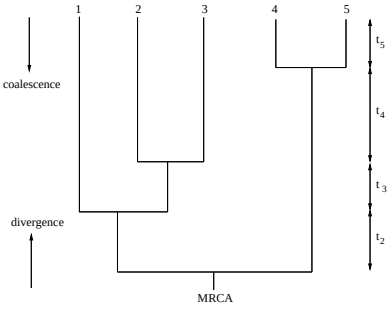
One can simulate data using classic population genetics models. We discussed such simulations when introducing finite populations and genetic drift, but we could only handle relatively small populations. In general, the coalescent is a much more powerful simulation approach because you need not simulate the *whole population*.

- **Forward simulation (classical genetics).** The classical genetics models use forward simulation. You start with the base population of usually substantial size N and follow the population forward in time.
- **Backward simulation (coalescent).** With the coalescent you start with the sample collected today (of reasonable size $n \ll N$) and trace it back until the MRCA.

Forward Simulation

- Randomly generate your base population of size N_0.
- Produce gametes, allow gametes to mutate, randomly unite gametes to produce early pre-selection adults. Often it is assume the population size throughout this stage is infinite which eases computations. You just need to compute the probability of each outcome since the infinite population will have each allele and genotype in these exact proportions. However, if any stage here is finite, you need to track each allele (gamete) or genotype (individual) and apply random events to it.
- Randomly select N_{t+1} surviving adults from the desired probability distribution to represent the next generation.
- Repeat for T total generations to the present day.
- Randomly sample n individuals from the N_T present day individuals to simulate statistical sampling.
- Much waste: the MRCA of the sample may occur exist in generation $T_m \gg 0$, making the first T_m generations wasted computations.

Backward Simulation



The simulation procedure is

- Randomly generate a sample of size n.
- Randomly generate the time of the first coalescent event t_n.
- Randomly select the lineages that will coalescent.
- Repeat until the last two lineages are chosen to coalesce, t_2 time after the penultimate coalescence.
- Construct the coalescent tree from the coalescent times $t_i, i = n, n-1, \ldots, 2$ and the coalescing lineages.

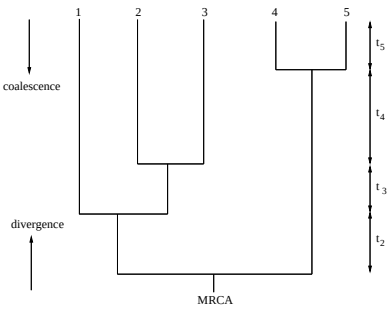
Parametric Bootstrap Using Coalescent Simulation

You collect data that you think is somehow odd (non-neutral, for example). You don't want to fall in the trap of overstating the significance of your data. How do you assess significance? Use coalescent simulation to perform *parametric bootstrapping*.

- Collect data and compute some statistic $\hat{\theta}$, e.g. the number of conserved sites among L sequenced in a sample of n individuals.
- Make assumptions about the population and history of the population from which you obtained your sample. This is your null hypothesis H_0.
- Set up a coalescent model satisfying these assumptions.
- Simulate the coalescent model many times and compute the same statistic $\hat{\theta}_i$ for each of the $i = 1, 2, \ldots, M$ parametric bootstraps.
- The p-value for the rejecting H_0 is $p = \frac{\#\{\theta_i \text{ more extreme than } \hat{\theta}\}}{M}$, where "more extreme than" can be $\leq$ or $\geq$, depending on the circumstance.

2 The Coalescent Model

Theory



The structure of the tree is completely determined by the t_i and the pairs of lineages that merge at each coalescent event. To define the coalescent process, then we need to know what the t_i are and we need to know how lineages are selected to merge.

Lineages are selected to merge randomly (consequence of the random selection of parents).

The coalescent times t_i are also random numbers (this is a stochastic process after all). What we will derive now is the distribution of t_i. Different assumptions about the population will lead to different distributions for t_i. We start with the basic Wright-Fisher model.

2.1 Coalescent Time Distributions

Wright-Fisher Coalescent Times

Let $p_k(i)$ be the probability that a random sample of k alleles in generation i come from $k-1$ alleles in generation $i+1$. In other words, there is a coalescent event in going back from generation i to generation $i+1$ or there has been a duplication in going from generation $i+1$ to i, forward in time.

We know that in a diploid population with N individuals, there are $2N$ alleles. We also know that the probability any two alleles are ibd copies from the previous generation is

$$p_2(i) = \frac{1}{2N},$$

for all i.

$1 - p_k(i)$ is the probability that none of the k lineages coalesce between generation i and $i+1$. In other words, this is the probability that no two of the k alleles in generation i pick the same parent from generation $i+1$. We know

$$1 - p_k(i) = \left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right)\cdots\left(1 - \frac{k-1}{2N}\right) = \prod_{i=1}^{k-1}\left(1 - \frac{i}{2N}\right),$$

When $k \ll 2N$, so the number of current lineages is much smaller than the population size (which is true whenever the sample is much smaller than the census (or effective) population size), then

$$\begin{aligned} 1 - p_k(i) &= 1 - \frac{1}{2N} - \frac{2}{2N} - \cdots - \frac{k-1}{2N} + o(\frac{1}{N}) \\ &\approx 1 - \frac{1}{2N}\left(1 + 2 + \cdots + k - 1\right) \\ &= 1 - \frac{1}{2N}\left[\frac{k(k-1)}{2}\right]. \end{aligned}$$

Rewriting, we have

$$p_k(i) = \frac{k(k-1)}{4N}$$

We'll use this to find the distribution of t_n, namely $P(t_n = t)$, the probability that n lineages present at generation 0 coalesce into $n-1$ lineages at precisely generation t. In other words, we need the probability of no coalescent events for $t-1$ generations, followed by a coalescent event at generation t, but this is

$$P(t_n = t) = \left[1 - p_n(1)\right]\cdots\left[1 - p_n(t-1)\right]p_n(t),$$

but it is clear that $p_n(i)$ does not depend on the generation i (it only depends on the number of lineages n), so

$$P(t_n = t) = \left[1 - p_n\right]^{t-1} p_n,$$

where $p_n(i) = p_n$ for all i.

What is this distribution?

$$P(t_n = t) = \left[1 - p_n\right]^{t-1} p_n.$$

It can be approximated by the exponential distribution

$$\begin{aligned} P(t_n = t) &\approx p_n e^{-p_n t} \\ &= \frac{n(n-1)}{4N} e^{-\frac{n(n-1)t}{4N}}, \end{aligned}$$

where we have substituted our previous result $p_n = \frac{n(n-1)}{4N}$. The approximation is good when the per-generation probability of coalescence p_n is very small, i.e. the sample size is much smaller than the census size (something we already assumed).

Interpretation of Coalescent Time t_n Distribution

Hence, the Wright-Fisher coalescent times t_i follow an approximate exponential distribution with mean $\frac{4N}{i(i-1)}$ depending on the census population size N an the current number of sampled lineages i.

- Wait times are independent, i.e. the time it takes for 3 lineages to coalesce t_3 does not depend on the time t_4 it took for 4 lineages to coalesce to 3.
- The wait times are *memoryless* (a property of the exponential). This concept can be hard to grasp, because it seems unintuitive, but it is a fact of the Wright-Fisher model. *Memoryless* means that if I have waited 5 million years for the i present lineages to coalesce, that doesn't mean that a coalescent event is imminent. On average, I will have to wait the same amount of time $\frac{4N}{i(i-1)}$ that I was expecting to wait when the 5 million years started.
- The wait time increases as the number of lineages decreases. So, as I go back in time, I wait longer and longer for the coalescent events. This is a consequence of the fact the fewer lineages have less opportunity to choose the same parents. It also means that Wright-Fisher coalescent trees have a particular shape, long trunks with lots of leaves.
- Notice that $\mathrm{E}(t_2) = \frac{4N}{2} = 2N$, so we are expected to wait $2N$ generations before any two randomly selected individuals coalesce into their MRCA. That's a pretty long time.
- **Coalescent time.** Define a new time scale $\tau = \frac{t}{2N}$. We can define the coalescent times $t_n, t_{n-1}, \ldots$ on this new time as $\tau_n, \tau_{n-1}, \ldots$. Notice that $\mathrm{E}(\tau_2) = 1$, so any two randomly sampled lineages are expected to coalescent in one unit of scaled time.

Scales of Time

- **Real time.** Measured in regular units, years, days, hours, minutes, seconds, etc.
- **Generations.** In the basic coalescent, t is measured in generations.
- **Coalescent time.** If generation time is scaled by the population size $\tau = \frac{t}{2N}$, then it is called *coalescent time*, and τ is measured in units of average coalescent time for two lineages.
- **Scaled time.** If time is scaled (e.g. to match some population process like $u = \frac{t}{\sigma^2}$ or $v = \frac{\tau}{\sigma^2}$), then we speak of scaled time (either scaled generations u or scaled coalescent time v).

2.2 Adding Mutation

Adding Mutation to Wright-Fisher Coalescent

Neutral mutation is very easy to add to the basic coalescent model because it occurs independent of the coalescent process except for a dependence on overall time $t_n + t_{n-1} + \cdots + t_2$.

Suppose neutral mutation occurs with probability μ during each replication cycle. We generally assume μ is very small, while t_i are very large. This is conveniently modeled by a Poisson distribution. Specifically,

$$P(l \text{ mutations along branch of length } t) = \frac{(\mu t)^l e^{-\mu t}}{l!},$$

where μt is the expected number of mutations and is just equal to the mutation rate times the number of generations along a branch of length t.

The probability that there is no mutation in all k lineages for the current generation is

$$(1-\mu)^k \approx 1 - k\mu.$$

The probability that there is exactly one mutation in 1 of the k lineages $t+1$ generations ago is

$$k\mu(1-k\mu)^t \approx k\mu e^{-k\mu t}.$$

In other words, the wait time for the next mutation (going backwards in time) is again an exponentially distributed random variable with mean $\frac{1}{k\mu}$.

We can think of coalescent events and mutation events as competing with each other. Their wait times are independent. Once one occurs, the wait times reset (memoryless property) and the competition starts again.

Simulating Coalescent with Mutation

- Start with your sample of n haplotypes.
- Suppose there are currently k lineages.
- Generate

$$\begin{aligned} x_{\text{coalescent}} &\sim \text{Exponential}\left(\frac{k(k-1)}{4N}\right) \\ x_{\text{mutation}} &\sim \text{Exponential}(k\mu) \end{aligned}$$

- If $x_{\text{coalescent}} < x_{\text{mutation}}$, the coalescent process won.
  - Set $t_k = x_{\text{coalescent}}$
  - Randomly choose two existing lineages to merge.
  - Decrement k by 1.
- Otherwise, if $x_{\text{mutation}} < x_{\text{coalescent}}$, the mutation process won.
  - Randomly select a lineage l to mutate from the k available.
  - Make note of the mutation and the branch where the mutation applies.
- Repeat from step 2 until you reach the MRCA.
- Randomly simulate a haplotype for the MRCA. Make two identical copies to evolve down each of the descendent lineages.
- Randomly generate the mutations assigned to each of the descent branches of the MRCA (you must have a mutation model in mind).
- The resulting sequences are the ancestor sequences at the next split.
- Repeat until you reach the present day sequences. You will have one sequence simulated for each of the n sampled sequences.

Simulating the Coalescent with Mutation (Method II)

- Select your sample size n.
- Suppose there are currently $k \leq n$ lineages.
- Generate the next coalescent time:

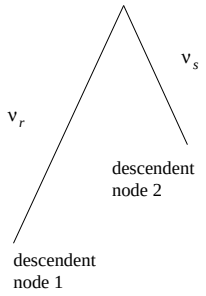
$$t_k \sim \text{Exponential}\left(\frac{k(k-1)}{4N}\right)$$

- Randomly choose two of the existing lineages to merge.
- Decrement k by 1.
- Repeat until $k = 1$. Then, $t_n + t_{n-1} + \cdots + t_2$ is the time of the MRCA (most recent common ancestor).

- For each branch j in the coalescent tree, compute it's branch length ν_j.
- Generate the number of mutations along this branch:

$$u_j = \text{Poisson}(\mu\nu_j).$$

- Generate the MRCA haplotype (e.g. $AAAGAGA\ldots$)
- For each coalescent fork with descending branches r and s, generate two copies of the current ancestral haplotype. Randomly apply u_r mutations to one and u_s mutations to the other. These are the ancestors of the next descendent nodes.
- Repeat until the terminal sample of n is reached.



2.3 Generalizing Wright-Fisher Coalescent

Effective Population Size

It turns out that various violations of Wright-Fisher model of population growth can be made into the basic coalescent process by scaling time. If N_e is the effective population size, then if the generation time t is scaled as

$$\frac{t}{2N_e},$$

we obtain a mapping from the non-Wright-Fisher population to the standard coalescent on the coalescent time scale.

Important implications are

- we can use the coalescent process even when our population does not satisfy Wright-Fisher assumptions, but
- *we cannot use polymorphism data alone to infer the biological phenomena that represent scale changes in coalescent time*. In other words, many, many processes could produce the same time scale change.

2.3.1 Variable Population Size

Variable Population Size

Suppose the population size at time t is $N(t)$ (here t measured in generations increases as we move back in real time).

Clearly, coalescent events happen more rapidly when $N(t)$ is small. If we continuously re-scale time appropriately, we can restore the standard coalescent process.

The amount of coalescent time traversed in going from generation i to $i+1$ is $\frac{1}{2N(i)}$.

And the total amount of coalescent time traversed in going from generation 1 to t is

$$g(t) = \sum_{i=1}^{t} \frac{1}{2N(i)}.$$

$g(t)$ is a strictly increasing function, so we can invert it and compute the number of generations $t = g^{-1}(\tau)$ corresponding to τ units of coalescent time.

Simulating Variable Population Size

Thus, if it is known how the population size $N(t)$ changes from generation to generation, then we can simulate a standard fixed-population size coalescent in coalescent time $\tau_k = \frac{t_k}{2N}$, where t_k are the exponentially distributed random variables we derived previously. Then map the coalescent times to generations in the variable size population via $g^{-1}(\tau)$ to draw our coalescent tree for a variable-sized population.

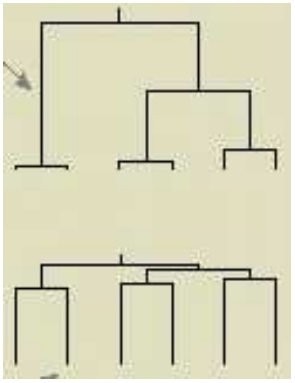
As an example, suppose $N(t) = N(0)e^{-\beta t}$, then

$$g(t) \approx \int_0^t \frac{1}{2N(s)} ds = \frac{e^{\beta t} - 1}{2\beta N(0)}.$$

The inverted function is

$$g^{-1}(\tau) \approx \frac{\log\left[1 + 2N(0)\beta\tau\right]}{\beta}.$$

Exponential Growth



Most coalescent events occur when the population is small, i.e. early in the history. Therefore, the result tree looks star-like.

2.3.2 Structured Coalescent

Introduction to Structured Coalescent

Consider a population of size N and suppose it is subdivided into m equal subpopulations of size N_i, such that $N = \sum_{i=1}^{m} N_i$. Allow migration between subpopulations (at the infinite gamete stage) with probability m_{ij} that a gamete starting in subpopulation i ends up in subpopulation j.

The backward probability b_{ij}, that an individual of subpopulation i came from subpopulation j is

$$b_{ij} = \frac{N_j m_{ji}}{\sum_{k=1}^{m} N_k m_{ki}}$$

The number of offspring in subpopulation j that descend from an parent of subpopulation i has distribution

$$\text{Bin}\left(N_j, \frac{b_{ji}}{N_i}\right)$$

Coalescent Events with Structure

Lineages are now associated with subpopulations.

A lineage in subpopulation i can coalesce with a lineage from subpopulation j ending up in subpopulation k if both lineages pick the same parent from subpopulation k, which happens with probability

$$\frac{b_{ik}b_{jk}}{N_k}$$

This model can be approximated by a continuous time coalescent process as before.

Structured Coalescent

Let

$$\begin{aligned} c_i &= \frac{N_i}{N} \\ B_{ij} &= 2Nb_{ij}, i \neq j \end{aligned}$$

Then it is possible to show that a pair of lineages in subpopulation i coalesces backwards in time with rate $\frac{1}{c_i}$ and each lineage in i migrates (backwards in time) to subpopulation j at rate $\frac{B_{ij}}{2}$. These processes are independent, so it is easy to envision how one could simulate a labeled coalescent tree, where each branch is labeled with the subpopulation in which it exists.

Simulating Structured Coalescent

Specifically, if k_i lineages currently exist in subpopulation i, then the rate at which either a coalescent or a migration happens is

$$h(k_1, \ldots, k_m) = \sum_{i=1}^{m} \left[\frac{\binom{k_i}{2}}{c_i} + \sum_{j \neq i}^{m} k_i \frac{B_{ij}}{2} \right]$$

Generate an exponential random variable with this rate to determine the time of the *next event*.

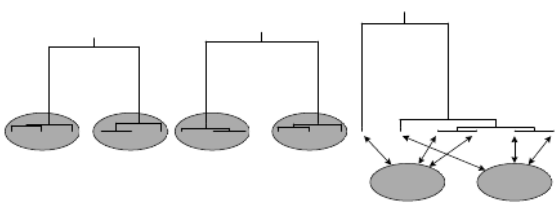
Then, the next event is a coalescence in subpopulation i with probability

$$\frac{\binom{k_i}{2}/c_i}{h(k_1, \ldots, k_m)}$$

or it is a migration from i to j with probability

$$\frac{k_i B_{ij}/2}{h(k_1, \ldots, k_m)}$$

Realization of Structured Coalescent



2.4 Coalescent Properties

Properties of Coalescent Tree

- The probability that a sample of size n contains the MRCA of the *whole population* is $\frac{n-1}{n+1} \approx 1$ for reasonable n.

Saunders, Tavaré, Watterson (1984) *Adv. Appl. Prob.* **16**:471.

- Thus, increasing sample size tends to add short "twigs" to the coalescent tree. For each additional sequence sampled,
  - relatively little evolutionary time is added to the history,
  - relatively few additional mutations are observed in the new data, and
  - relatively little is learned about ancient events occurring at the base of the tree.

Lesson: If our interest is to estimate time to a common ancestor or mutation rate, adding more sequences provides little extra information. We will formalize this argument next.

2.4.1 Expected Tree Length

Length of the Tree

The total expected sum of branch lengths of the tree

$$T_{\text{tot}}(n) = \mathrm{E}\left(\sum_{k=2}^{n} k\tau_k\right) = \sum_{k=1}^{n-1} \frac{2}{k} \approx 2\left(\gamma + \log n\right),$$

as $n \to \infty$, where $\gamma \approx 0.577216$ is Euler's constant.

Since the number of mutations depends on $T_{\text{tot}}(n)$, sampling n individuals only improves estimation of mutation rate μ, *and all other population parameters related to* μ, as sampling $\log n$ independent samples would. This quantitates the cost of the dependence among the samples.

2.4.2 Expected Age of MRCA

Age of the MRCA

Let t be the random age of the MRCA of the current sample, so

$$t = t_n + t_{n-1} + \cdots + t_2.$$

The expected age of the MRCA is

$$\begin{aligned} \mathrm{E}(t) &= \sum_{k=2}^{n} \mathrm{E}(t_k) \\ &= \sum_{k=2}^{n} \frac{4N}{k(k-1)} \\ &= \frac{4N(n-1)}{n}. \end{aligned}$$

and the variance (use independence of t_n and variance of exponential distribution) is

$$\mathrm{Var}(t) = 16N^2 \sum_{k=2}^{n} \frac{1}{k^2(k-1)^2}.$$

	$\mathbf{E}(t)$ ($\mathbf{Var}(t)$)			
n	$N = 10$	$N = 100$	$N = 1000$	$N = 10000$
2	20 (400)	200 (40000)	2000 (4×10^6)	20000 (4×10^8)
10	36 (11.6)	360 (115.8)	3600 (1158)	36000 (11581)
100	39.6 (11.6)	396 (115.9)	3960 (1159)	39600 (11595)
1000	40.0 (11.6)	400 (115.9)	3996 (1159)	39960 (11595)

Interpretation:

- Small samples will tend to have an MRCA that is younger than the MRCA of the whole population. However, the estimate rises and plateaus quickly with sample size n, reflecting the Saunders *et. al* results.
- The proportion of the underestimate is independent of the population size N.
- The uncertainty is large, regardless of sample size, because the genetic variance of that last coalescent time t_2 is large and dominating.

Application: How Old are Humans?

How long ago did the MRCA of us all exist? Substitute $n = N$ in the expectation formula:

$$\mathrm{E}(t_{\text{total}}) = \frac{4N(N-1)}{N} = 4(N-1) \approx 4N.$$

where N is about 6 billion. But, we'd be wrong.

We need to consider exponential (or some form of non-constant) population growth to get an accurate estimate. And, we'd really need to know and plug in the effective population size N_e.

3 Inference under Coalescent

3.1 Total Mutations

Expected Total Number of Mutations

Let K be the total number of mutations occurring in the history of the n sampled sequences. Conditioning on the coalescent tree, the expected number of mutations throughout the tree is

$$\mathrm{E}(K \mid t_n, t_{n-1}, \ldots, t_2) = \mu \left(2t_2 + \cdots + nt_n\right).$$

Taking expectation again, we have the total number of expected mutation is

$$\begin{aligned} \mathrm{E}(K) &= \mathrm{E}\left[\mathrm{E}(K \mid t_n, t_{n-1}, \ldots, t_2)\right] \\ &= \mu \sum_{k=2}^{n} k\mathrm{E}(t_k) \\ &= \mu \sum_{k=2}^{n} \frac{4N}{k-1} \\ &= 4N\mu a_n = a_n\theta, \end{aligned}$$

with $a_n = \sum_{k=2}^{n} \frac{1}{k-1}$ a constant depending on the sample size n and $\theta := 4N\mu$ is the "population genetics parameter".

Variance in Number of Mutations

Let K_k be the number of mutations occurring in the k lineages during time t_k. Then,

$$\begin{aligned}
\text{Var}(K_k) &= \text{E}\left(K_k^2\right) - \text{E}^2\left(K_k\right) \\
&= \text{E}\left[\text{E}\left(K_k^2 \mid t_k\right)\right] - \text{E}^2\left[\text{E}\left(K_k \mid t_k\right)\right] \\
&= \text{E}\left[\text{Var}\left(K_k \mid t_k\right) + \text{E}^2\left(K_k \mid t_k\right)\right] - \text{E}^2\left[\text{E}\left(K_k \mid t_k\right)\right] \\
&= \text{E}\left[\mu k t_k + \mu^2 k^2 t_k^2\right] - \mu^2 k^2 \text{E}^2\left(t_k\right) \\
&= \mu k \text{E}(t_k) + \mu^2 k^2 \text{E}(t_k^2) - \mu^2 k^2 \text{E}^2(t_k) \\
&= \mu k \frac{4N}{k(k-1)} + \mu^2 k^2 \left[\text{Var}(t_k) + \text{E}^2(t_k)\right] - \mu^2 k^2 \frac{16N^2}{k^2(k-1)^2} \\
&= \frac{4N\mu}{k-1} + 2\mu^2 k^2 \frac{16N^2}{k^2(k-1)^2} - \mu^2 k^2 \frac{16N^2}{k^2(k-1)^2} \\
&= \frac{4N\mu}{k-1} + \frac{16N^2\mu^2}{(k-1)^2}.
\end{aligned}$$

Of course, the total variance is

$$\text{Var}(K) = \sum_{k=2}^{n} \text{Var}(K_k).$$

Let

$$b_n = \sum_{k=2}^{n} \frac{1}{(k-1)^2},$$

then

$$\text{Var}(K) = a_n\theta + b_n\theta^2,$$

where $\theta = 4N\mu$.

Infinite Sites Model

The infinite sites model (Watterson, 1975) posits

- each locus has infinitely many sites
- when a mutation occurs, it will occur at a site that has not previously mutated

and it makes life easy because the assumptions ensure that all mutations that occur along a lineage after the MRCA of a sample are observed so that

$$K = \text{the number of segregating sites in a sample of size } n$$

It is a good approximation to loci if

- there is little divergence between individuals, and
- loci represent a quite lengthy sequence.

Method of Moments Estimator of θ

If K segregating sites are observed, then

$$\hat{\theta} = \frac{K}{a_n} = \frac{K}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

Example: ADH locus of *Drosophila*.

A sample $n = 11$ alleles are sampled from populations in Florida, Washington, Africa, Japan, and France. A total of $K = 14$ sites showed some variation in the sample. Thus,

$$\hat{\theta} = 4.78.$$

Remember this estimate is based on the assumption of neutral mutations. Thirteen of the 14 mutations were synonymous. One changed the protein sequence. Often, researchers use only the synonymous mutations to estimate θ, throwing out nonsynonymous mutations as possibly subject to selection.

Finite Sites Model

Let d_{ij} be the number of nucleotide changes between two sequences i and j, then by applying the segregating site formulas for $n = 2$,

$$\mathrm{E}(d_{ij}) = \theta \quad \text{and} \quad \mathrm{Var}(d_{ij}) = \theta + \theta^2.$$

To improve your estimate, you may wish to sample $n > 2$ sequences. How can you improve the estimator? Let π be the average pairwise distance

$$\pi = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}.$$

Then, $\mathrm{E}(\pi) = \mathrm{E}(d_{ij}) = \theta$ and (not derived)

$$\mathrm{Var}(\pi) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2.$$

The advantage of the pairwise distance approach is that CTMC models can be used to handle the reality of *finite sites models* where multiple mutations can affect the same site.

Interpretation

The more polymorphic the data, the larger the pairwise distances d_{ij}, so the larger the estimate θ.

Turn the argument around and you see that as the population size increases, $\theta = 4N\mu$ increases and you will expect greater diversity in your data. Remember that we know diversity decreases because of the loss of alleles in finite (small) populations.

In addition, and logically, the higher the mutation rate, the more diversity you expect to see in your data.

3.2 Tajima's D

Tajima's D Statistic

We have just discussed two estimators of θ:

$$\begin{aligned} \hat{\theta}_K &= \frac{K}{a_n} \\ \hat{\theta}_\pi &= \pi \end{aligned}$$

Tajima (1989) defined

$$D = \hat{\theta}_\pi - \hat{\theta}_K$$

and argued that when genetic drift and neutral mutation are at equilibrium $D = 0$. Tajima D statistics significantly different from 0 can signify many different processes:

Effect of Non-Neutral Mutation or Non-equilibrium

- **Overdominant selection.** Consider a locus with two alleles. One can show that the average pairwise distance is maximized when both alleles have frequency 0.5. Overdominant selection drives allele frequencies toward 0.5. However, the number of alleles, and thus segregating sites, remains unchanged as long as both alleles persist at some positive frequency in the population. Thus, $D > 0$.
- **Population bottleneck.** Alleles are lost, so the number of segregating sites decreases immediately, but the average pairwise distance declines more slowly. $D > 0$.
- **Purifying selection.** Purifying selection removes mutants, so their frequencies will tend to be low. Thus, there may be segregating sites, but average pairwise diversity is quite low. $D < 0$.
- **Population expansion.** Population expansion allows mutations to persist that would not normally persist (think of a branching process where every individual, including mutants are producing more offspring). However, neutral mutations take a long time to fix in the population with no selection to help them, so overall mutants will have low frequency. $D < 0$.

Var(D)

$$\text{Var}(D) = AK - BK(K-1)$$

where

$$\begin{aligned} A &= \frac{\frac{n+1}{3(n-1)} - \frac{1}{a_n}}{a_n} \\ B &= \frac{\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2}}{a_n^2 + b_n} \end{aligned}$$

3.3 Ancestral Population Size

Ancestral Population Size

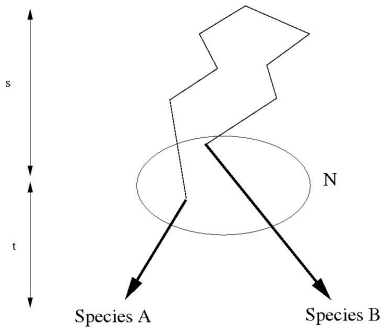


Figure. The history of two *orthologous* genes. Orthologous genes are copies of a gene created by speciation, as opposed to *paralogous* genes that are copies of genes produce by gene duplication events.

Suppose that species A and B diverged t generations ago when the population size was N (unknown). The coalescent time of these two genes is $t + s$, where s is the amount of time before the two copies of the gene in the ancestral species took to coalesce in the ancestral species.

Let K be the number of mutational differences between the two sampled genes in species A and B. Let μ be the mutation rate, assumed to be constant throughout the history of these two genes since their MRCA. Then,

$$P(K = k \mid s) = \frac{[2\mu(t+s)]^k}{k!} e^{-2\mu(t+s)},$$

from the Poisson distribution.

The coalescent time s in the ancestral species is unknown, but we know its distribution (exponential). Integrate it out.

$$\begin{aligned} P(K) &= \int_0^\infty P(K \mid s)P(s)ds \\ &= \left(\frac{1}{1+\theta}\right)\left(\frac{\theta}{1+\theta}\right)^K e^{-2\mu t}\sum_{i=1}^{K}\frac{1}{i!}\left[\frac{2\mu(1-\theta)}{\theta}\right]^i. \end{aligned}$$

The above can be viewed as the likelihood of the observed data K conditional on the model and model parameters. Maximize this likelihood over the unknown N and report $\hat{N}$ as the estimate of the ancestral population size.

An estimate of μ is required.

Takahata et al. (1995) studied 13 orthologous sequences from humans and chimpanzees, 7 orthologous sequences from humans and gorillas, and 7 orthologous sequences from gorilla and chimpanzee. They assumed a mutation rate of $\mu = 1 \times 10^{-9}$ per site per year. They assume the average generation length is 15 years, which allowed them to translate coalescent time into real time.

Species Pair	$\hat{N}$
Human/Chimpanzee	83,000
Human/Gorilla	77,000
Chimpanzee/Gorilla	42,000

3.4 Examples

3.4.1 Bottleneck

Example - Bottleneck in Malaria

Hughes and Vera (2001) *Proc. Roy. Soc. Lond., B* 268(1478):1855-1860.

Previous work had shown that *Plasmodium falciparum* (malaria) is very homogeneous, which could indicate it had recently experienced a bottleneck. A very recent bottleneck would manifest as a small effective population size N_e since there would have been lots of inbreeding in a recently very small population. Vaccine and treatment efforts could be more likely successful if there had been a bottleneck, so the authors wish to estimate N_e.

Mine the sequence database for malaria sequences and select data for loci that are polymorphic (at least two different alleles in the database for these loci) and under neutral evolution (how they determined this is beyond the scope of this lecture).

We know

$$\mathrm{E}\left(t_{\mathrm{MRCA}}\right) = \frac{4N_e(n-1)}{n},$$

where we have substituted the effective population size in for the census population size N.

We also know the MRCA of any collection of sequences is very likely to be the MRCA of the whole population.

The authors take the two most divergent sequences and estimate their coalescent time $\hat{t}_2$. Then,

$$\hat{t}_2 \approx t_{\mathrm{MRCA}} \approx \frac{4N_e}{2},$$

so they estimate N_e as

$$\hat{N}_e = \frac{\hat{t}_2}{2}.$$

Example - Malaria

They also estimated the effective population size N_e in a second way via

$$4N_e\mu = \theta.$$

To do so, they needed an estimate of θ and μ.

The average pairwise distance between sequences provides an estimate of θ

$$\hat{\theta} = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}.$$

For the mutation rate, they compared the malaria that infects humans with the malaria that infects chimpanzees. These two parasites are thought to have diverged when humans and chimpanzees diverged. There are actual time estimates of when humans and chimpanzees diverged, call this time t_{hc}. Convert this to generation scale by assuming g generations per year, so $t'_{hc} = t_{hc}g$.

Let M be the number of mutations observed between the human and chimp malaria, then

$$\hat{\mu} = \frac{M}{2t'_{hc}}.$$

Do you know why there is a 2 in this formula?

Finally, a second estimate of the effective population size is

$$\hat{N_e} = \frac{\hat{\theta}}{4\hat{\mu}}.$$

Via all methods of estimation, they found N_e to be quite large. They conclude there has been no recent bottleneck affecting the malaria parasite.

3.5 Likelihood-Based Inference

Let τ represent the topology of the coalescent tree. Let $t = (t_2, \ldots, t_n)$ represent the coalescent times of the coalescent tree. Let θ represent the parameters of the model, most simply $\theta = 4N\mu$, but more complex variations of the coalescent would include other parameters. Let D represent the data (perhaps sequence data).

Typically, we are interested in inferring the population parameters θ. The topology τ and branch lengths t are nuisance parameters.

Then, the likelihood of the data is

$$L(\theta) = P(D \mid \theta) = \int_{t_2} \cdots \int_{t_n} \sum_{\tau} P(D \mid \tau, t, \theta) P(\tau, t \mid \theta) dt_2 \cdots dt_n$$

- $P(D \mid \tau, t)$ is the likelihood of data given a phylogenetic tree (see Misha' lectures).
- $P(\tau, t \mid \theta)$ is the coalescent model.

Monte Carlo Sampling

If $X \sim f_X(x)$, then the mean of any function $g(X)$ can be approximated by simulating

$$X_1, X_2, \ldots, X_M \sim f_X(x)$$

and computing the sample mean

$$E[g(X)] = \int g(x) f_X(x) dx \approx \frac{1}{M} \sum_{i=1}^{M} g(X_i)$$

In the context of likelihood inference under the coalescent, we have

$$L(\theta) \approx \frac{1}{M} \sum_{i=1}^{M} P(D \mid \tau_i, t_i, \theta).$$

where $(\tau_i, t_i) \sim P(\tau, t \mid \theta)$ are topologies and branch lengths simulated from the coalescent model (something that you know is quite easy to do).

Further Troubles

But, of course there is a problem. Straight Monte Carlo sampling is not efficient because many of the coalescent trees (τ_i, t_i) simulated from $P(\tau_i, t_i \mid \theta)$ lead to very low likelihoods $P(D \mid \tau_i, t_i, \theta)$.

The numerical methods that have been used successfully for coalescent simulation include

- **Importance Sampling**. Sample (τ, t) from $Q_\theta(\tau, t)$ instead of $P(\tau, t \mid, \theta)$, where $Q(\cdot)$ is chosen to make "better" choices.
- **Markov Chain Monte Carlo**. See software by Kuhner, Beerli, and Felsenstein.

Estimating Tree Properties

If properties of the tree are of interest, then they can be recovered from the sample (τ_i, t_i). Weights are needed

- **Importance Sampling**. Define the weight w_i associated with tree (τ_i, t_i) be

  $$w_i = \frac{W_i}{\sum_{j=1}^{M} W_j}$$

  where

  $$W_j = P(D \mid \tau_i, t_i \theta) \frac{P(\tau_i, t_i \mid \theta)}{Q(\tau_i, t_i)}$$

- **MCMC**. $w_j = 1$

For example, to obtain an estimate of the time until the MRCA from the likelihood analysis, use

$$E(T_{\text{MRCA}}) \approx \sum_{i=1}^{M} w_i T_{\text{MRCA}}(\tau_i, t_i)$$