Coalescent Genealogy Samplers

Mary K. Kuhner
Genome Sciences
University of Washington
Seattle WA
Outline

1. Introduction to coalescent theory
2. Practical example
3. Genealogy samplers
4. Evolutionary forces
Coalescent-based studies

- How many gray whales were there prior to whaling?
- When was the common ancestor of HIV lines in a Libyan hospital?
- Is the highland/lowland distinction in Andean ducks recent or ancient?
- Did humans wipe out the Beringian bison population?
- What proportion of HIV virions in a patient actually contribute to the breeding pool?
- What is the direction of gene flow between European rabbit populations?
All individuals release many gametes and new individuals for the next generation are formed randomly from these.
The big trick

- We have a model for the progress of a population forward in time
- What we observe is the end product: genetic data today
- We want to reverse this model so that it tells us about the past of our sequences
Sewall Wright showed that the probability that 2 gene copies come from the same gene copy in the preceding generation is

\[ \text{Prob (two genes share a parent)} = \frac{1}{2N} \]
In every generation, there is a chance of $1/2N$ to coalesce. Following the sampled lineages through generations backwards in time we realize that it follows a geometric distribution with

$$
\mathbb{E}(u) = 2N \quad \text{[the expectation of the time of coalescence $u$ of two tips is $2N$]}
$$
JFC Kingman generalized this for \( k \) gene copies.

\[
\text{Prob (} k \text{ copies are reduced to } k - 1 \text{ copies)} = \frac{k(k - 1)}{4N}
\]
Kingman’s $n$-coalescent
Kingman’s $n$-coalescent

The expectation for the time interval $u_k$ is

$$
\mathbb{E}(u_k) = \frac{4N}{k(k-1)}
$$

$$
p(G|N) = \prod_i \exp(-u_i \frac{k(k-1)}{4N}) \frac{1}{2N}
$$
The $\Theta$ parameter

- The n-coalescent is defined in terms of $N_e$ and time.
- We cannot measure time just by looking at genes, though we can measure divergence.
- We rescale the equations in terms of $N_e$, time, and the mutation rate $\mu$.
- We can no longer estimate $N_e$ but only the composite parameter $\Theta$.
- $\Theta = 4N_e\mu$ in diploids.
- Multiple time point data can separate $N_e$ and $\mu$. 
What is this coalescent thing good for?
Utopian population size estimator

1. We get the correct genealogy from an infallible oracle

2. We know that we can calculate $p(\text{Genealogy}|N)$
Utopian population size estimator

1. We get the correct genealogy from an infallible oracle

2. We remember the probability calculation

\[ p(G|N) = p(u_1|N, k) \frac{1}{2N} \times p(u_2|N, k - 1) \frac{1}{2N} \times \ldots \]
1. We get the correct genealogy from an infallible oracle

2. We remember the probability calculation

\[ p(\text{Genealogy} | N) = \prod_{j} e^{-u_{j} \frac{k_{j}(k_{j}-1)}{4N}} \frac{1}{2N} \]
Utopian population size estimator

Prob( G | N) [$\cdot 10^{-43}$]

Population size N

[Graph showing the probability function for different population sizes.]
Utopian population size estimator

Prob( G | N) [•10^{-43}]

Population size N

1000 10000 20000 30000 40000
Utopian population size estimator

\[ N = 2270 \]

\[ N = 12286 \]
Lack of infallible oracles

- We assume we know the true genealogy including branch lengths
- We don’t really know that
- We probably can’t even infer it:
  - Tree inference is hard in general
  - Population data usually don’t have enough information for good tree inference
Genealogy samplers

- Acknowledge that there is an underlying genealogy—
  - but we don’t know it
  - we can’t infer it with high certainty
  - we can’t sum over all possibilities

- A directed sample of plausible genealogies—
  - can capture much of the information in the unknown true genealogy
  - takes a long time but not forever

These are genealogy sampler methods
Outline

1. Introduction to coalescent theory
2. Practical example: red drum
3. Genealogy samplers
4. Evolutionary forces
What is the effective population size of red drum?

Red drum, *Sciaenops ocellatus*, are large fish found in the Gulf of Mexico.

Turner, Wares, and Gold
Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish
What is the effective population size of red drum?

- Census population size: 3,400,000

- Effective population size: ?

- Data set:
  - 8 microsatellite loci
  - 7 populations
  - 20 individuals per population
What is the effective population size of red drum?

Three approaches:

1. Allele frequency fluctuation from year to year
   - Measures current population size
   - May be sensitive to short-term fluctuations

2. Coalescent estimate from *Migrate*
   - Measures long-term harmonic mean of population size
   - May reflect past bottlenecks or other long-term effects

3. Demographic models
   - Attempt to infer genetic size from census size
   - Vulnerable to errors in demographic model
   - Not well established for long-lived species with high reproductive variability
Population model used for Migrate

- Multiple populations along Gulf coast
- Migration allowed only between adjacent populations
- Allowing for population structure should improve estimates of population size
What is the effective population size of red drum?

Estimates:

- Census size ($N$): 3,400,000
- Allele frequency method ($N_e$): 3,516 (1,785-18,148)
- Coalescent method ($N_e$): 1,853 (317-7,226)

The demographic model can be made consistent with these only by assuming enormous variance in reproductive success among individuals.
What is the effective population size of red drum?

- Allele frequency estimators measure current size
- Coalescent estimators measure long-term size
- Conclusion: population size and structure have been stable
What is the effective population size of red drum?

- Effective population size at least 1000 times smaller than census
- This result was highly surprising
- Red drum has the genetic liabilities of a rare species
- Turner et al. hypothesize an “estuary lottery”
- Unless the eggs are in exactly the right place, they all die
Outline

1. Introduction to coalescent theory
2. Practical example
3. Genealogy samplers
4. Evolutionary forces
Coalescent estimation of population parameters

- Mutation model: Steal a likelihood model from phylogeny inference
- Population genetics model: the Coalescent
Coalescent estimation of population parameters

$L(\Theta) = P(Data|\Theta)$
Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta) = \sum_{G} P(Data|G)P(G|\Theta)$$
Coalescent estimation of population parameters

\[
L(\Theta) = P(Data|\Theta) = \sum_G P(Data|G)P(G|\Theta)
\]

\(P(Data|G)\) comes from a mutational model
Coalescent estimation of population parameters

\[ L(\Theta) = P(Data|\Theta) = \sum_{G} P(Data|G)P(G|\Theta) \]

\[ P(G|\Theta) \text{ comes from the coalescent} \]
Coalescent estimation of population parameters

\[ L(\Theta) = P(Data|\Theta) = \sum_{G} P(Data|G)P(G|\Theta) \]

\[ \sum_{G} \text{ is a problem} \]
Can we calculate this sum over all genealogies?

<table>
<thead>
<tr>
<th>Tips</th>
<th>Topologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>180</td>
</tr>
<tr>
<td>6</td>
<td>2700</td>
</tr>
<tr>
<td>7</td>
<td>56700</td>
</tr>
<tr>
<td>8</td>
<td>1587600</td>
</tr>
<tr>
<td>9</td>
<td>57153600</td>
</tr>
<tr>
<td>10</td>
<td>2571912000</td>
</tr>
<tr>
<td>15</td>
<td>6958057668962400000</td>
</tr>
<tr>
<td>20</td>
<td>5644809895887305913369600000</td>
</tr>
<tr>
<td>30</td>
<td>4368466613103069512464680198620763891440640000000000000</td>
</tr>
<tr>
<td>40</td>
<td>3027333829948007356546303364551457200042939432053862501707888721920000000000000000</td>
</tr>
<tr>
<td>50</td>
<td>$3.28632 \times 10^{112}$</td>
</tr>
<tr>
<td>100</td>
<td>$1.37416 \times 10^{284}$</td>
</tr>
</tbody>
</table>
A solution: Markov chain Monte Carlo

- If we can’t sample all genealogies, could we try a random sample?
  - Not really.

- How about a sample which focuses on good ones?
  - What is a good genealogy?
  - How can we find them in such a big search space?
A solution: Markov chain Monte Carlo

Metropolis recipe

0. first state

1. perturb old state and calculate probability of new state

2. test if new state is better than old state: accept if ratio of new and old is larger than a random number between 0 and 1.

3. move to new state if accepted otherwise stay at old state

4. go to 1
How do we change a genealogy?
MCMC walk result

![Histogram of Tree Space Probabilities]

- Tree space
- Probability
- Probability values: 0.002, 0.005, 0.01
MCMC walk result—with problems
Outline

1. Introduction to coalescent theory
2. Genealogy samplers
3. Practical example
4. Likelihood sampler details
5. Evolutionary forces
Likelihood and Bayesian approaches

- All genealogy samplers search among genealogies
- All of them require some type of guide value (“driving value”) to determine which genealogies will be proposed
- Two major approaches: Likelihood-based and Bayesian
- Major ideological difference, relatively small practical one
Likelihood samplers

- Use arbitrary values of the parameters to guide the search
- Sample genealogies throughout the search
- At the end of the search, evaluate $P(G|\Theta)$ for sampled genealogies
- Correct for the influence of the driving values
- Iterate to improve driving values
Likelihood analysis

We will approximate:

\[ L(\Theta) = \sum_G P(Data|G)P(G|\Theta) \]
Likelihood analysis

We will approximate:

\[ L(\Theta) = \sum_{G} P(\text{Data}|G)P(G|\Theta) \]

by sampling \( n \) genealogies from \( P(\text{Data}|G)P(G|\Theta_0) \):

\[ L(\Theta) = \frac{1}{n} \sum_{G^*} \frac{P(\text{Data}|G)P(G|\Theta)}{P(\text{Data}|G)P(G|\Theta_0)/L(\Theta_0)} \]

Here the \( G^* \) are no longer random genealogies; they are sampled from a distribution that depends on the driving value \( \Theta_0 \).
Likelihood analysis

\[ L(\Theta) = \frac{1}{n} \sum_G \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)/L(\Theta_0)} \]

Isn’t this circular? We have a solution for the unknown \( L(\Theta) \) in terms of the unknown \( L(\Theta_0) \).
Likelihood analysis

\[ L(\Theta) = \frac{1}{n} \sum_{G} \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)/L(\Theta_0)} \]

Isn’t this circular? We have a solution for the unknown \( L(\Theta) \) in terms of the unknown \( L(\Theta_0) \).

\[ \frac{L(\Theta)}{L(\Theta_0)} = \frac{1}{n} \sum_{G} \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)} \]

This doesn’t give us the actual value of \( L(\Theta) \) but it does allow us to compare various values of \( \Theta \) and choose the best.
Likelihood analysis

- This approach is only asymptotically correct

- For finite sample sizes, it has a bias toward its driving value

- We can greatly reduce this:
  - Start with an arbitrary $\Theta_0$
  - Run the sampler a while and estimate the best $\Theta$
  - It will be biased toward $\Theta_0$, but...
  - Use it as the new $\Theta_0$ and start over
Likelihood versus Bayesian approaches

- Kuhner 2006: Bayes and likelihood almost identical
- Beerli 2006: Bayes has edge with sparse data

My recommendations:
- Use Bayes if you think a parameter is very close to zero
- Otherwise, with rich data either method is good
- With poor data, do you really want to be doing this analysis at all?
- When using Bayes, be careful of your priors!

If the genealogy search is inadequate, both methods will fail (and fail in similar ways)
Outline

1. Evolutionary forces
   - Genetic drift ($\Theta$)
   - Population growth/shrinkage
   - Migration
   - Recombination
   - Population divergence
   - Use of multiple time points
Genetic drift ($\Theta$)

- With one time point, we estimate $\Theta = 4N_e \mu$ in diploids
- The number estimated is $2N_e \mu$ in haploids or $N_e \mu$ in mtDNA
- Two ways to separate $N_e$ and $\mu$:
  - Dated historical data (ancient DNA, etc.)
  - External estimate of mutation rate
- For most organisms, $N_e$ is less than $N$
- Demographic models can help resolve this
Variable population size

- In a small population lineages coalesce quickly
- In a large population lineages coalesce slowly

This leaves a signature in the data. We can exploit this and estimate the population growth rate $g$ jointly with the current population size $\Theta$. 
Exponential population size expansion or shrinkage
## Grow a frog

### Mutation Rate

<table>
<thead>
<tr>
<th>Mutation Rate</th>
<th>-10000 generations</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-8}$</td>
<td>8,300,000</td>
<td>8,360,000</td>
</tr>
<tr>
<td>$10^{-7}$</td>
<td>780,000</td>
<td>836,000</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>40,500</td>
<td>83,600</td>
</tr>
</tbody>
</table>
Bayesian skyline plots
Gene flow

\[ p(G|\Theta, M) = \prod_{uj} \left( \prod_{i} g(\Theta_i, M_{i}) \right) \left\{ \frac{2}{\Theta} M_{ji} \right\} \text{ if event is a coalescence,} \]

\[ \text{if event is a migration from } j \text{ to } i. \]
Gene flow: What researchers used (and still use)
What researchers used (and still use)

Sewall Wright showed that

$$F_{ST} = \frac{1}{1 + 4Nm}$$

and that it assumes

- migration into all subpopulation is the same
- population size of each island is the same
Simulated data and Wright’s formula

\[ 4N_e m_{21} \]

\[ 4N_e m_{12} \]

**True values**

<table>
<thead>
<tr>
<th></th>
<th>Estimated values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1.14±0.77</td>
</tr>
<tr>
<td>0.01</td>
<td>7.80±22.20</td>
</tr>
<tr>
<td>0.05</td>
<td>11.46±18.54</td>
</tr>
</tbody>
</table>

**Estimated values**

<table>
<thead>
<tr>
<th></th>
<th>Estimated values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1.14±0.77</td>
</tr>
<tr>
<td>0.01</td>
<td>7.80±22.20</td>
</tr>
<tr>
<td>0.05</td>
<td>11.46±18.54</td>
</tr>
</tbody>
</table>
Maximum Likelihood method to estimate gene flow parameters

(Beerli and Felsenstein 1999)

100 two-locus datasets with 25 sampled individuals for each of 2 populations and 500 base pairs (bp) per locus.

<table>
<thead>
<tr>
<th></th>
<th>Population 1</th>
<th></th>
<th>Population 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Θ</td>
<td>0.0500</td>
<td>Θ</td>
<td>0.0050</td>
</tr>
<tr>
<td>$4N_e^{(1)}m_1$</td>
<td>10.00</td>
<td>$4N_e^{(2)}m_2$</td>
<td>1.00</td>
</tr>
<tr>
<td>Truth</td>
<td></td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.0052</td>
<td>0.0048</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td>1.09</td>
<td>1.21</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Estimation of divergence time

Wakeley and Nielsen (2001)
Estimation of divergence time

Wakeley and Nielsen (2001) Figure 7. The joint integrated likelihood surface for T and M estimated from the data by Orti et al. (1994). Darker values indicate higher likelihood.
Multiple time points

- Ancient DNA or historical samples of fast-evolving organisms

- Points must be:
  - Dated
  - Far enough apart for measurable evolution

- Advantages:
  - Separation of $\Theta$ into $N_e$ and $\mu$
  - Much better resolution of growth rates
Selection coefficient estimation

Krone and Neuhauser (1999), Felsenstein (unpubl)
Review paper

Outline

- Introduction to coalescent theory
- Genealogy samplers
- Evolutionary forces
- Practical considerations
Information content of the coalescent

What can best give us more information?

- More individuals?
- More base pairs?
- More loci?
Variability of the coalescent

10 coalescent trees generated with the same population size, $N = 10,000$
Variability of mutations
Does adding more individuals help?
The bottom line

- The information content of a single locus is limited
- Additional sequence length or individuals are only mildly helpful
- Multiple loci allow the best estimates
- If recombination is present, long sequences can partially substitute for multiple loci
- Multiple time points can also help, if significant evolution happens between them
Two publications supporting this conclusion

- Felsenstein, J (2005) Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? MBE 23: 691-700.