The Coalescent Model

Florian Weber

23. 7. 2016
The Coalescent Model

coalessent $= „$zusammenwachsend“$
Outline

- Population Genetics and the Wright-Fisher-model
- The Coalescent
- Non-constant population-sizes
- Further extensions
- Summary
Population Genetics

(Shamelessly stealing Alexis’ slides)

- Study of polymorphisms in a population
  - What are the processes that introduce polymorphisms in the population?
  - If a polymorphism exists in a population, will it be there forever?
  - Is there some process that removes polymorphisms from the population?
  - Do the polymorphisms exhibit patterns?
  - ...
Motivation

- The coalescent is basically the Wright-Fisher-model with a lot of analysis.
Motivation

- The coalescent is basically the Wright-Fisher-model with a lot of analysis.
- It can easily do calculations about the past
Motivation

- The coalescent is basically the Wright-Fisher-model with a lot of analysis.
- It can easily do calculations about the past
- It is very fast to compute
Motivation

- The coalescent is basically the Wright-Fisher-model with a lot of analysis.
- It can easily do calculations about the past.
- It is very fast to compute.
- It can easily be extended to represent a more complex reality.
Hardy-Weinberg

- Assuming an **infinite population size**, random mating, diploid population, no selection... the allele-frequencies are constant.
Assuming an **infinite population size**, random mating, diploid population, no selection...<br>the allele-frequencies are constant

Infinity is weird... $0.3 \times \infty = \infty$

...and unrealistic
Assuming a **finite but constant population size**, random mating, non-overlapping generations, no selection... all alleles except for one will disappear over time.
Assuming a **finite but constant population size**, random mating, non-overlapping generations, no selection... all alleles except for one will disappear over time.

The likelihood for an allele to prevail is equal to its initial frequency.
Figure 1: A simulation of three alleles under the model
Wright-Fisher (Individuals)

Figure 2: An evolutionary history in the model
Figure 3: Extinct alleles removed
Wright-Fisher (Individuals)

Figure 4: Surviving Tree
Figure 5: Most Recent Common Ancestor marked
Wright-Fisher (Coalescence-Events)

Figure 6: Coalescence-Events of the green individuals
Figure 7: Two individuals and their parents
The Coalescent Model

- Likelihood for two nodes to coalesce in the previous generation:
  \[ p(P(A) = P(B)) = \frac{1}{N} \]
The Coalescent Model

- Likelihood for two nodes to coalesce in the previous generation:
  \[ p(P(A) = P(B)) = \frac{1}{N} \]

- In the previous two generations:
  \[ 1 - \left( \frac{N-1}{N} \cdot \frac{N-1}{N} \right) = 1 - \left( \frac{N-1}{N} \right)^2 \]
The Coalescent Model

- Likelihood for two nodes to coalesce in the previous generation:
  \[ p(P(A) = P(B)) = \frac{1}{N} \]

- In the previous two generations:
  \[ 1 - \left( \frac{N-1}{N} \cdot \frac{N-1}{N} \right) = 1 - \left( \frac{N-1}{N} \right)^2 \]

- In the previous three generations:
  \[ 1 - \left( \left( \frac{N-1}{N} \right)^2 \cdot \frac{N-1}{N} \right) = 1 - \left( \frac{N-1}{N} \right)^3 \]
The Coalescent Model

- Likelihood for two nodes to coalesce in the previous generation:
  \[ p(P(A) = P(B)) = \frac{1}{N} \]

- In the previous two generations:
  \[ 1 - \left( \frac{N-1}{N} \cdot \frac{N-1}{N} \right) = 1 - \left( \frac{N-1}{N} \right)^2 \]

- In the previous three generations:
  \[ 1 - \left( \left( \frac{N-1}{N} \right)^2 \cdot \frac{N-1}{N} \right) = 1 - \left( \frac{N-1}{N} \right)^3 \]

- In the previous \( t \) generations:
  \[ 1 - \left( \frac{N-1}{N} \right)^t \]
The Coalescent Model

* without loss of generality

Figure 8: Likelihood of coalescence
The Coalescent Model

- Likelihood of coalescence in the previous $t$ generations:

$$1 - \left( \frac{N - 1}{N} \right)^t$$
The Coalescent Model

- Likelihood of coalescence in the previous $t$ generations:
  
  $$1 - \left(\frac{N - 1}{N}\right)^t$$

- Likelihood for lineages to remain distinct for $t$ generations:
  
  $$\left(\frac{N - 1}{N}\right)^t$$
The Coalescent Model

- Likelihood of coalescence in the previous $t$ generations:
  
  \[ 1 - \left( \frac{N - 1}{N} \right)^t \]

- Likelihood for lineages to remain distinct for $t$ generations:
  
  \[ \left( \frac{N - 1}{N} \right)^t \]

- Expected time for coalescence: $E(t) = N$
The Coalescent Model

- Likelihood of coalescence in the previous $t$ generations:
  
  \[ 1 - \left( \frac{N - 1}{N} \right)^t \]

- Likelihood for lineages to remain distinct for $t$ generations:
  
  \( \left( \frac{N - 1}{N} \right)^t \)

- Expected time for coalescence: $E(t) = N$

- Rescale: $\tau = \frac{t}{N}$:
  
  \( \left( \frac{N - 1}{N} \right)^{\lceil N\tau \rceil} \xrightarrow[N \to \infty]{} e^{-\tau} \)
The Coalescent Model

⇒ The likelihood for two lineages to stay distinct over time is exponentially small!

![Diagram](image-url)
Moar Lineages!!

Figure 9: http://what-if.xkcd.com/13/
More Lineages

- Likelihood of no coalescence in one generation and three lineages:

\[
\frac{N - 1}{N} \times \frac{N - 2}{N}
\]
More Lineages

- Likelihood of no coalescence in one generation and three lineages:
  \[
  \frac{N - 1}{N} \times \frac{N - 2}{N}
  \]

- One generation, \(k\) lineages:
  \[
  \frac{N - 1}{N} \times \frac{N - 2}{N} \times \cdots \times \frac{N - k + 1}{N} = \prod_{i=1}^{k-1} \frac{N - i}{N}
  \]
More Lineages

- For some reason this is equal to:

\[
\prod_{i=1}^{k-1} \frac{N - i}{N} = 1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right) \approx 1 - \frac{\binom{k}{2}}{N}
\]

- \(\binom{k}{2}\) is the binomial coefficient and equates to \(\frac{k \cdot (k-1)}{2}\)
More Lineages

- For some reason this is equal to:

\[
\prod_{i=1}^{k-1} \frac{N-i}{N} = 1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right) \approx 1 - \frac{\binom{k}{2}}{N}
\]

- \(\binom{k}{2}\) is the binomial coefficient and equates to \(\frac{k \cdot (k-1)}{2}\)

- There are \(\binom{k}{2}\) ways to pick two lineages from a set of \(k\) lineages.
More Lineages

- For some reason this is equal to:

\[
\prod_{i=1}^{k-1} \frac{N - i}{N} = 1 - \frac{k}{N} + O\left(\frac{1}{N^2}\right) \approx 1 - \frac{k}{N}
\]

- \(\binom{k}{2}\) is the binomial coefficient and equates to \(\frac{k \cdot (k-1)}{2}\)

- There are \(\binom{k}{2}\) ways to pick two lineages from a set of \(k\) lineages.

- Therefore a coalescence-event is \(\binom{k}{2}\)-times as likely with \(k\) lineages than with 2.
More Lineages

- For some reason this is equal to:

\[
\prod_{i=1}^{k-1} \frac{N-i}{N} = 1 - \binom{k}{2}\frac{1}{N} + O\left(\frac{1}{N^2}\right) \approx 1 - \frac{\binom{k}{2}}{N}
\]

- \(\binom{k}{2}\) is the binomial coefficient and equates to \(\frac{k \cdot (k-1)}{2}\)

- There are \(\binom{k}{2}\) ways to pick two lineages from a set of \(k\) lineages.

- Therefore a coalescence-event is \(\binom{k}{2}\)-times as likely with \(k\) lineages than with 2

- The number of coalescence-events grows quadratically with the number of lineages!
More Lineages

Events getting exponentially rare

Figure 10: More lineages = faster coalescence
Properties

- Few deep furcations
Properties

- Few deep furcations
- Likelihood: Everything is possible but maybe unlikely
Properties

- Few deep furcations
- Likelihood: Everything is possible but maybe unlikely
- Calculation is backward in times (Wright-Fisher: forward)
Properties

- Few deep furcations
- Likelihood: Everything is possible but maybe unlikely
- Calculation is backward in times (Wright-Fisher: forward)
- Efficient: no calculation per individual or for extinct lineages
Non-constant population-sizes

Figure 11: World population - not very constant [Wikimedia]
Non-constant population-sizes

- Non-constant, but known population-size
- Coalescence is more likely in small populations
- $\Rightarrow$ Coalescence-rate changes over time
Non-constant population-sizes

- Non-constant, but known population-size
- Coalescence is more likely in small populations
- \( \Rightarrow \) Coalescence-rate changes over time
- Simply rescale time.
Rescaling Time

- Before: \( t \) Generations corresponded to \( t/N \) units of coalescence-time
- Now: \( t \) Generations correspond to

\[
\sum_{i=1}^{t} \frac{1}{N_i}
\]

units of coalescence-time
- Note: for a constant population both formulas are equal
Rescaling Time - Example

- 5 Generations, with on average 5 individuals:

\[
\tau = \frac{t}{N} = \frac{5}{5} = 1 \text{ unit of coalescence time}
\]

For non-constant \{4, 4, 5, 6, 6\} individuals:

\[
\tau = \frac{t}{\sum_{i=1}^{5} \frac{1}{N_i}} = \frac{1}{\frac{1}{4} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{6}} = \frac{31}{30}
\]

Note the lesser influence of the larger generations.

A generation with twice the size will get half the coalescence-time.
Rescaling Time - Example

- 5 Generations, with on average 5 individuals:
  
- For constant 5 individuals: \[ \tau = \frac{t}{N} = \frac{5}{5} = 1 \] unit of coalescence time

  Note the lesser influence of the larger generations

- A generation with twice the size, will get half the coalescence time
Rescaling Time - Example

- 5 Generations, with on average 5 individuals:

- For constant 5 individuals: $\tau = \frac{t}{N} = \frac{5}{5} = 1$ unit of coalescence time

- For non-constant $\{4, 4, 5, 6, 6\}$ individuals:

$$\tau = \sum_{i=1}^{t} \frac{1}{N_i} = \frac{1}{4} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{6} = \frac{31}{30}$$

note the lesser influence of the larger generations
Rescaling Time - Example

- 5 Generations, with on average 5 individuals:

- For constant 5 individuals: \( \tau = \frac{t}{N} = \frac{5}{5} = 1 \) unit of coalescence time

- For non-constant \( \{4, 4, 5, 6, 6\} \) individuals:

  \[
  \tau = \sum_{i=1}^{t} \frac{1}{N_i} = \frac{1}{4} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{6} = \frac{31}{30}
  \]

  note the lesser influence of the larger generations

- A generation with twice the size, will get halve the coalescence-time
Figure 12: Exponentially growing population versus coalescence-time
Figure 13: Exponentially growing and constant populations. Note the reverse time-scale! [Nordborg]
Rescaling Time - Applicability

- Approximation converges against theory for growing $N$
- Close enough for most purposes
Further Extensions

- Separated Populations
- Diploid Populations
- Males and Females
- Selection
- Multiple Species
- ...
Further Extensions

- Separated Populations
- Diploid Populations
- Males and Females
- Selection
- Multiple Species
- ...

Wright-Fisher:

*Assuming a finite but constant population size, random mating, non-overlapping generations, no selection...*
Further Extensions

- Separated Populations
- Diploid Populations
- Males and Females
- Selection
- Multiple Species
- ...

Wright-Fisher:

Assuming a finite but constant population size, random mating, non-overlapping generations, no selection...

Coalescent:

Assuming non-overlapping generations...
An actual example

Figure 14: Coalescent vs. Anthropological Estimates [Atkinson et al.]
Software

Software that uses the coalescent model\(^1\):
BEAST, COAL, CoaSim, DIYABC, DendroPy, GeneRecon, genetree, GENOME, IBDSim, IMa, Lamarc, Migraine, Migrate, MaCS, ms & msHOT, msms, Recodon and NetRecodon, SARG, simcoal2, TreesimJ

\(^1\)Source: https://en.wikipedia.org/wiki/Coalescent_theory
Summary

- The coalescent is the Wright-Fisher-model plus math
- Coalescent-events are, with exponential likelihood, relatively recent
- The more lineages there are, the more coalescence-events occur
- Non-Constant populations can be simulated by rescaling time
- The simulated time for a generation is anti-proportional to it’s size
References

Content

➤ Magnus Nordborg, “Coalescent Theory”, March 2000

Software-list

➤ en.wikipedia.org/wiki/Coalescent_theory

Images

➤ Fig. 09: Randal Munroe: what-if.xkcd.com/13/
➤ Fig. 11: El T: commons.wikimedia.org/wiki/File:Population_curve.svg
➤ Fig. 13: Magnus Nordborg: “Coalescent Theory”, 2000
➤ Fig. 14: Atkinson et al.: “mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory”, 2008