Introduction to Coalescent Models

Biostatistics 666
Previously ...

- Allele frequencies
- Hardy Weinberg Equilibrium
- Linkage Equilibrium
  - Expected state for distant markers
- Linkage Disequilibrium
  - Association between neighboring alleles
  - Expected to decrease with distance
- Measures of linkage disequilibrium
  - $D$, $D'$ and $\Delta^2$ or $r^2$
Making predictions...

• What allele frequencies do we expect?

• How much variation in a gene?

• How are neighboring variants related?

• Are these predictions “universal”?
  • Do they depend on natural selection or the history of a population?

• How can we use genetic variation to build models of the past?
1000 Genomes Data: Variants per Genome

<table>
<thead>
<tr>
<th>Type</th>
<th>Variant sites / genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs</td>
<td>~3,800,000</td>
</tr>
<tr>
<td>Indels</td>
<td>~570,000</td>
</tr>
<tr>
<td>Mobile Element Insertions</td>
<td>~1000</td>
</tr>
<tr>
<td>Large Deletions</td>
<td>~1000</td>
</tr>
<tr>
<td>CNVs</td>
<td>~150</td>
</tr>
<tr>
<td>Inversions</td>
<td>~11</td>
</tr>
</tbody>
</table>
1000 Genomes Data: Demographic Models
Simple Approach: Simulation

1. N starting sequences
2. Sample N offspring sequences
   - Apply mutations according to $\mu$
3. Increment time
4. If enough time has passed...
   - Generate final sample
   - Stop.
5. Otherwise, return to step 1.
Simulating a Population ...
Today

• Introduce coalescent approach
  • Framework for studying genetic variation
  • Provides intuition on patterns of variation
  • Provides analytical solutions
Aim ...

• Gene genealogies:
  • Descriptions of relatedness between sequences
  • Analogous to phylogenetic trees for species

• The shape of the genealogy depends on population history, selection, etc.

• Together with mutation rate, genealogy predicts DNA variation
Genealogy

- History of a particular set of sequences
  - Describes their relatedness
  - Specifies divergence times

- Includes only a subset of the population

- Most Recent Common Ancestor (MRCA)
Coalescent approach

• Generate genealogy for a sample of sequences.
  • Introduces computational and analytical convenience.

• Instead of proceeding forward through time, go backwards!
History of the Population
Genealogy of Final Population
Levels of Complexity

• History of the population
  • Includes sequences that are “extinct”

• History of all modern sequences
  • Includes sequences that we haven’t sampled

• History of a subset of modern sequences
  • Minimalist approach!
Examples of Typical Coalescent Trees

A

B

C
Parameters we will focus on...

• Mutation rate ($\mu$)
• Population Size
  • Haploid population (N chromosomes)
  • Diploid population (2N chromosomes)
• Time (t)
• Sample size (n)
• Recombination rate ($r$)
Other Parameters

• Selection
  • For gene of interest
  • For neighboring gene

• Demographic parameters
  • Migration
  • Population Structure
  • Population Growth
Mutation Model

- The mutation process is complex
  - Rate depends on surrounding sequence
  - Reverse mutations are possible

- Two simple models are popular
  - Infinite alleles
    - Every mutation generates a different allele
  - Infinite sites
    - Every mutation occurs at a different site
Mutation Model

• Focus on infinite sites model
  • Mutation rate in genomic DNA is $\sim 10^{-8} / \text{bp}$
  • Recurrent mutations should be very rare

• Scaled mutation rate parameter, e.g.:
  • 1000 bp sequence
  • $10^{-8}$ mutations per base pair per generation
  • $\mu = 10^{-5}$ per sequence per generation
Neutral Variants

• Variants that do not affect fitness

• Accumulate inexorably through time
  • Lost through genetic drift

• Do not affect genealogy
Example:
Modeling Accumulation of Mutations

• Population of identical sequences

• Sample one descendant after \( t \) generations

• How many mutations have accumulated?
  • Hint: depends on mutation rate \( \mu \) and time \( t \)

• Tougher questions
  • How many mutations have been fixed?
  • How much variation in the total population?
So far ...

• Divergence of a single sequence
  • Accumulation of mutations
  • Depends on time $t$
  • Depends on mutation rate $\mu$
  • Does not depend on population size $N$
  • Does not depend on population growth

• Next: A pair of sequences!
A tougher example ...

- Sample of two sequences
  - 100 bp each...

- How many differences are expected?
  - Population of size, $N = 1000$
  - Mutation rate
    - $\mu = 10^{-8} / \text{bp / generation}$
    - $\mu \approx 10^{-6} / 100 \text{ bp / generation}$
Genealogy of two sequences

Mutations between MRCA and Sequence 1?
Genealogy of two sequences

MRCA

Sequence 1    Sequence 2

Time $T(2)$

Total mutations in genealogy?
Number of mutations $S$

- Distributed as Poisson, conditional on total tree length

- $E(S) = \mu E(T_{\text{tot}})$
- $\text{Var}(S) = \mu E(T_{\text{tot}}) + \mu^2 \text{Var}(T_{\text{tot}})$

- $T_{\text{tot}}$ is the total length of all branches
Estimating Coalescence Time...

- Probability that two sequences have distinct ancestors in previous generation

\[ P(2) = \frac{N - 1}{N} = 1 - \frac{1}{N} \]

- Probability of distinct ancestors for \( t \) generations is \( P(2)^t \)
Probability of MRCA at time $t+1$

$$P(2)^t (1 - P(2)) = \frac{1}{N} \left( \frac{N - 1}{N} \right)^t$$

$$= \frac{1}{N} \left( 1 - \frac{1}{N} \right)^t$$

$$\approx \frac{1}{N} e^{-\frac{1}{N^t}}$$
For $n > 2$

• Coalescence when two sequences have common ancestor
  • For simplicity, consider the possibility of multiple simultaneous coalescent events to be negligible

• Requirements for no coalescence:
  • Pick one ancestor for sequence 1
  • Pick distinct ancestor for sequence 2
  • Pick yet another ancestor for sequence 3
  • ...
Estimating $P(n)$

- Probability that $n$ sequences have $n$ distinct ancestors in previous generation

$$P(n) = \prod_{i=1}^{n-1} \frac{N-i}{N}$$

- Assume:
  - $N$ is large
  - $n$ is small
  - Terms of order $N^{-2}$ can be ignored
Probability of Coalescence at Time $t+1$

\[ P(n)^t (1 - P(n)) \approx \left( \frac{n}{2} \right)^t \left( \frac{n}{2} \right) \left( 1 - \frac{n}{N} \right)^{\frac{n}{2}} \left( \frac{n}{2} \right)^t \approx \left( \frac{n}{2} \right)^t \frac{n}{N} e^{-\frac{n}{N} t} \]
Time to next coalescent event

• Use an exponential distribution to approximate time to next coalescent event...

\[
\text{Decay Rate } \lambda = \binom{n}{2} \frac{2}{N} \\
\text{Mean } \frac{1}{\lambda} = \frac{N}{\binom{n}{2}}
\]
T(j)

• For convenience, measure time to next coalescent event in units:
  • N generations for haploids
  • 2N generations for diploids

\[ E(T_j) = \frac{1}{\binom{j}{2}} \]

• How would you calculate time to MRCA of \( n \) sequences?
Total “Time in Tree”

• Sum of all the branch lengths
• Total evolutionary time available
  • e.g. for mutations to occur

\[
E(T_{\text{tot}}) = \sum_{i=2}^{n} iT(i) = \sum_{i=2}^{n} \frac{2i}{i(i-1)}
\]

\[
= \sum_{i=2}^{n} \frac{2}{i-1} = \sum_{i=1}^{n-1} \frac{2}{i}
\]
$T_{\text{MRCA}}$ vs. $T_{\text{TOT}}$
Number of Segregating Sites

• Commonly named $S$

• Total number of mutations in genealogy
  • Assuming no recurrent mutation

• A function of the total length of the genealogy
  • $T_{\text{tot}}$
Expected number of mutations

• Factor N for haploids, 2N for diploids

\[ E(S) = 2N\mu \sum_{i=2}^{n} i E(T(i)) \]
\[ = 4N\mu \sum_{i=1}^{n-1} 1/i \]
\[ = \theta \sum_{i=1}^{n-1} 1/i \]

• Population geneticists define \( \theta = 4N\mu \) (for diploids)
  • For gene mappers, \( \theta \) is usually the recombination rate
  • For population geneticists, \( r \) is the recombination rate
Expected number of mutations

• Factor $N$ for haploids, $2N$ for diploids

$$E(S) = 2N\mu \sum_{i=2}^{n} iE(T(i))$$

$$= 4N\mu \sum_{i=1}^{n-1} 1/i$$

$$= \theta \sum_{i=1}^{n-1} 1/i$$

• Population geneticists define $\theta = 4N\mu$ (for diploids)
  • For gene mappers, $\theta$ is usually the recombination rate
  • For population geneticists, $r$ is the recombination rate
E(S) as a function of $n$

Parameters

- $N = 10,000$ individuals
- $\mu = 10^{-4}$
- $\theta = 4$
More about $S$...

• Very large variance

$$Var(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$$

• Most of the variance contributed by early coalescent events (i.e. with small $n$)
Var(S) as a function of $n$

Parameters

$N = 10,000$ individuals

$\mu = 10^{-4}$

$\theta = 4$
Inferences about $\theta$

- Could be estimated from $S$
  - Divide by expected length of genealogy

$$\hat{\theta} = \frac{S}{\sum_{i=1}^{n-1} 1/i}$$

- Could then be used to:
  - Estimate $N$, if mutation rate $\mu$ is known
  - Estimate $\mu$, if population size $N$ is known
Var(\(\hat{\theta}\)) as a function of \(n\)

Parameters

\(N = 10,000\) individuals
\(\mu = 10^{-4}\)
\(\theta = 4\)
Alternative Estimator for $\theta$ ...

- Count pairwise differences between sequences
- Compute average number of differences

$$\tilde{\theta} = \left( \frac{n}{2} \right)^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{n} S_{ij}$$
Today...

• Probability of coalescence events

• Length of genealogy and its branches

• Expected number of mutations

• Simple estimates of $\theta$
Recommended Reading

Richard R. Hudson (1990)

*Gene genealogies and the coalescent process*

Oxford Surveys in Evolutionary Biology, Vol. 7.
D. Futuyma and J. Antonovics (Eds).
Oxford University Press, New York.