Coalescent Modeling for Distributions of Alleles

Biostatistics 666
Last Lecture: Introduction to the Coalescent

- Coalescent approach
  - Proceed backwards through time.
  - Model the genealogy of sample of sequences.

- Infinite sites model
  - All mutations distinguishable.
  - No reverse mutation.
Some key ideas ...

- Probability of coalescence events
- Length of genealogy and its branches
- Expected number of mutations
- Parameter $\theta$ which combines population size and mutation rate
Building Blocks...

- Probability of sampling distinct ancestors for \( n \) sequences

\[
P(n) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{N}\right) \approx 1 - \frac{n \choose 2}{N}
\]

- Coalescence time \( t \) is approximately exponentially distributed
Some Key Results...

- Coalescence Time (population size units)
  
  \[ E(T_j) = \frac{1}{\binom{j}{2}} \]

- Total Tree Length (population size units)
  
  \[ E(T_{tot}) = \sum_{i=1}^{n-1} \frac{2}{i} \]
Some More Key Results ...

- **Expected Number of Polymorphisms**

  For a diploid sample
  \[
  E(S) = 4N\mu \sum_{i=1}^{n-1} 1/i = \theta \sum_{i=1}^{n-1} 1/i
  \]

  For an haploid sample
  \[
  E(S) = 2N\mu \sum_{i=1}^{n-1} 1/i = \theta \sum_{i=1}^{n-1} 1/i
  \]
Estimating $\theta$

- Number of variants $S$ can be used to estimate $\theta$
  - Expected $S$ is simply $\theta E(T_{tot})$
  - To estimate $\theta$, divide by $S$ 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
Alternative Estimator for $\theta$ ...

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

$$\tilde{\theta} = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{n} S_{ij}$$
$$\text{Var}(\hat{\theta}) \text{ as a function of } n$$

Parameters

$$N = 10,000 \text{ individuals}$$

$$\mu = 10^{-4}$$

$$\theta = 4$$
Var($\hat{\theta}$) as a function of $n$

Parameters

$N = 10,000$ individuals
$\mu = 10^{-4}$
$\theta = 4$

If larger samples don’t help, how else could we improve inferences about $\theta$?
Today ...

- More applications of the coalescent
- Predicting allele frequency distributions
  - Using simulations
- Modeling the distribution of S
  - Using analytical calculations
A Coalescent Simulation ...

- Let’s consider tracing the ancestry of 4 sequences
When $n = 4$

Probability of Coalescent Event

$P(4) \approx \binom{4}{2}/2N$

Time to Next Coalescent Event

$T(4) \approx 2N/\binom{4}{2}$

Sample time from exponential distribution

Pick two sequences at random to coalesce
Next $n = 3$ ...

Let’s assume that sequences 3 and 4 are selected ...

Then, we repeat the process for a sample of 3 sequences
Next $n = 2$ ...

Let’s assume that sequences 1 and 2 are selected to coalesce

Then, we repeat the process for a sample of 2 sequences
At this point, we could place mutations in genealogy. Most often, these would fall in longer branches.
A Coalescent Simulation ...

Mutations in these branches affect a pair of sequences
A Coalescent Simulation ...

Mutations in these branches affect a single sequence
Repeating the simulation multiple times, would give us a predicted mutation spectrum.
Frequency Spectrum (n = 10)
Frequency Spectrum

- Constant size population
- Exponentially growing population

Most variants are rare

- For $n = 100$, \(\sim 44\%\) of variants occur \(< 5/100\).
- For $n = 10$, \(\sim 35\%\) of variants observed once.
Mutation Spectrum

- Depends on genealogy
  - Population Size
  - Population Growth
  - Population Subdivision

- Does not depend on
  - Mutation rate!
Deviations from Neutral Spectrum

- When would you expect deviations from the spectra we described?

- What would you expect for …
  - A rapidly growing population?
  - A population whose size is decreasing?

- Why?
Effect of Polymorphism Type

Data from Cargill et al, 1999
Frequency Spectrum of Protein Altering Variants

Spectrum of Coding Variation (in 12,000 individuals)

- Non-Synonymous
- Splice Variants
- Stop

Exome Chip Consortium (2011)
Number of Mutations

- Can be derived from coalescent tree
  - What are the key features?

- Analytical results possible
  - Trace back in time until MRCA, tracking mutation events
Sample of Two Sequences

- Track coalescences and mutations
  - Probability of a coalescent event?
    - Depends on population size …
  - Probability of a mutation?
    - Depends on mutation rate …

- Proceed backwards until either occurs…
  - Conditional probability for each outcome?
Two Identical Sequences

\[ P_2(S \text{ is 0}) \approx \frac{P_{CA}}{P_{CA} + P_{mut}} = \frac{1/2N}{1/2N + 2\mu} = \frac{1}{1 + \theta} \]
Full distribution of $S$...

- Probability that first $j$ events are mutations...

\[
P_2(j) = \left(\frac{\theta}{1 + \theta}\right)^j \left(\frac{1}{1 + \theta}\right)
\]
Example...

- 2 sequences
- Population size $N = 25,000$
- Mutation rate $\mu = 10^{-5}$

- Probability of 0, 1, 2, 3… mutations
And for multiple sequences...

- Describe number of mutations until the next coalescence event

- Proceed back in time, until:
  - One of $n$ sequences mutates…
  - A coalescent event occurs…
  - Then track mutations in (n-1) sequences
Formulae ...

\[ Q_n(j) = \left( \frac{n\mu}{n\mu + \frac{2}{2N}} \right)^j \frac{\binom{n}{2}}{2N} = \left( \frac{\theta}{\theta + n - 1} \right)^j \frac{n - 1}{\theta + n - 1} \]

\[ P_n(j) = \sum_{i=0}^{j} P_{n-1}(j-i)Q_n(i) \]
Example...

- 3 sequences
- Population size $N = 25,000$
- Mutation rate $\mu = 10^{-5}$

- Probability of 0, 1, 2, 3… mutations
Number of Mutations

N = 10,000
n = 10
μ = 2 * 10^{-5}

Approximately 2kb of sequence, sequenced in 10 individuals
So far ...

- One homogeneous population
  - Coalescence times
  - Number of mutations
    - Expectation
    - Distribution
  - Spectrum of mutations
- Several assumptions, including ...
  - No recombination
Next: Models w/ Recombination

- No recombination
  - Single genealogy

- Free recombination
  - Two independent genealogies
  - Same population history

- Intermediate case
  - Correlated genealogies
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.