

II. Population Genetics

Allele Frequency Distributions

Lecture 13

Instructor: Su-In Lee

Last Lecture: Introduction to the Coalescent

- Coalescent approach
 - Proceed backwards through time.
 - Genealogy of a 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 θ which combines population size and mutation rate

Building Blocks ...

- Probability of sampling distinct ancestors for n sequences

$$P(n) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{2N}\right) \approx 1 - \frac{\binom{n}{2}}{2N}$$

- Coalescence time t is approximately exponentially distributed

$$P(n)^t (1 - P(n)) \approx \frac{\binom{n}{2}}{2N} \left(1 - \frac{\binom{n}{2}}{2N}\right)^t \approx \frac{\binom{n}{2}}{2N} e^{-\frac{\binom{n}{2}}{2N} t}$$

Some Key Results ...

- Coalescence Time

$$E(T_n) = \frac{2N}{\binom{n}{2}}$$

- Total Length

$$E(T_{tot}) = E\left(\sum_{i=2}^n iT_i\right) = \sum_{i=2}^n iE(T_i) = 4N \sum_{i=1}^{n-1} \frac{1}{i}$$

Some More Key Results ...

- Expected Number of Polymorphisms (Mutations)

$$E(S) = E(\mu T_{tot}) = 4N\mu \sum_{i=1}^{n-1} \frac{1}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

Inferences about Θ

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

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

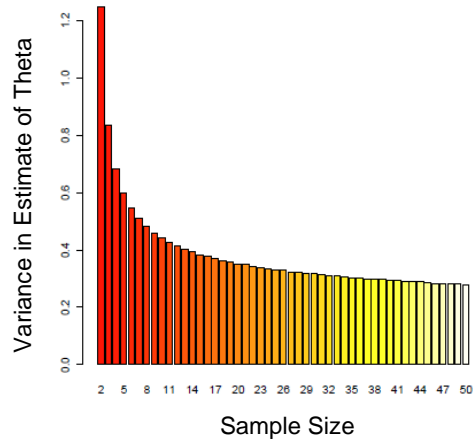
- Could be then be used to:
 - Estimate N, if mutation rate μ is known
 - Estimate μ , if population size N is known

Alternative Estimator for Θ

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

$$\tilde{\theta} = \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j=i+1}^n S_{ij}$$

Var($\hat{\Theta}$) as a function of N



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

Today's Lecture

- Applications of the coalescent
- Predicting allele frequency distributions
 - Using simulations
- The full 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

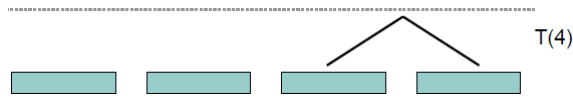
$$T(4) \approx 2N / \binom{4}{2}$$

- Sample time from exponential distribution
- Pick two sequences at random to coalesce



Next $n = 3 \dots$

- Let's assume that sequences 3 and 4 are selected ...
- Then, we repeat the process for a sample of 3 sequences

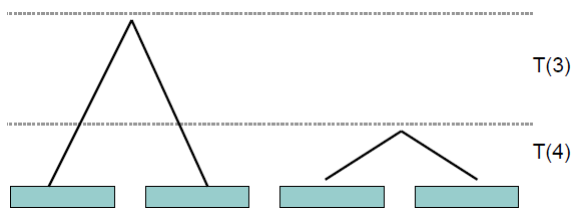


Computational Molecular Biology and Genomics, Spring 2009

13

Next $n = 2 \dots$

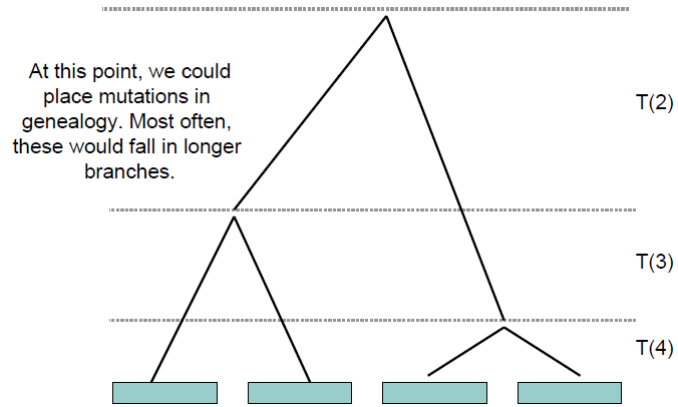
- Let's assume that sequences 1 and 2 are selected to coalesce
- Then, we repeat the process for a sample of 2 sequences



Computational Molecular Biology and Genomics, Spring 2009

14

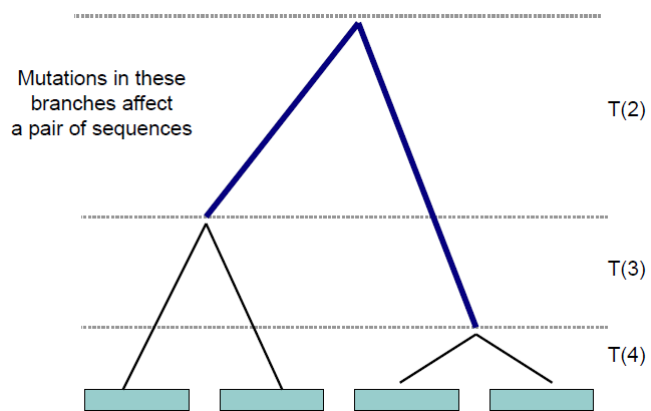
The Simulated Coalescent



Computational Molecular Biology and Genomics, Spring 2009

15

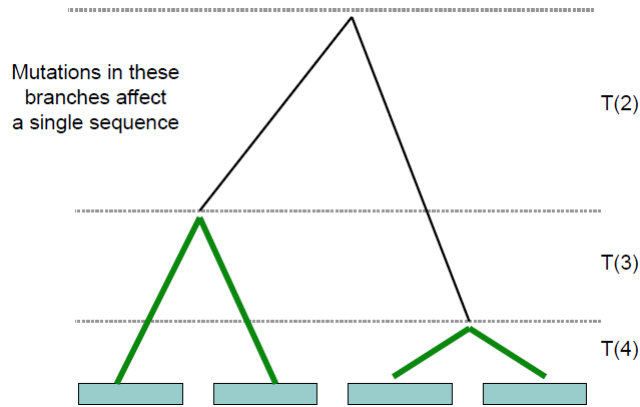
A Coalescent Simulation ...



Computational Molecular Biology and Genomics, Spring 2009

16

A Coalescent Simulation ...

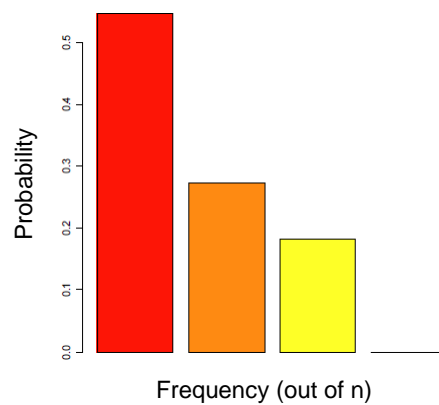


Computational Molecular Biology and Genomics, Spring 2009

17

Frequency Spectrum

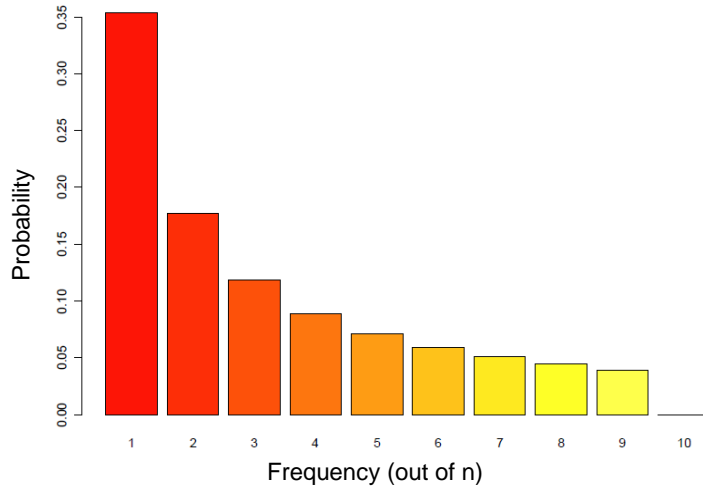
- Repeating the simulation multiple times would give us a predicted mutation spectrum.



Computational Molecular Biology and Genomics, Spring 2009

18

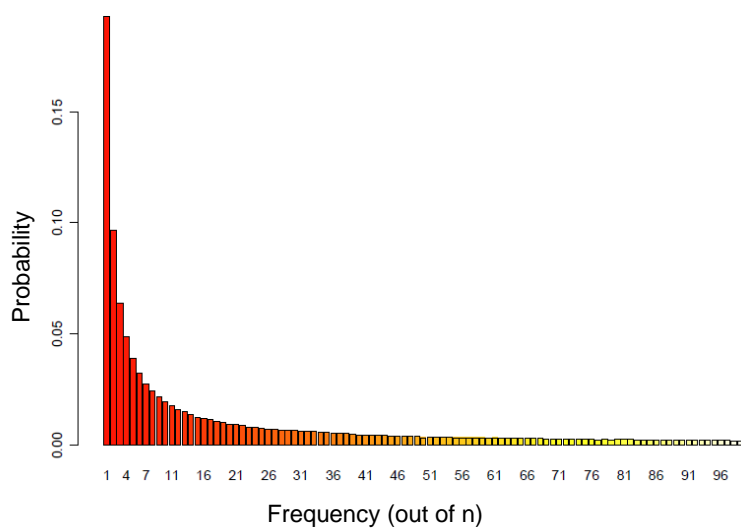
Frequency Spectrum (n = 10)



Computational Molecular Biology and Genomics, Spring 2009

19

Frequency Spectrum (n = 100)



Computational Molecular Biology and Genomics, Spring 2009

20

Frequency Spectrum

- Most variants are rare
 - For $n = 100$, ~44% of variants occur $< 5/100$.
 - For $n = 10$, ~35% of variants observed once.

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

$$\begin{aligned} P_2(S \text{ is } 0) &\approx \frac{P_{CA}}{P_{CA} + P_{mut}} \\ &= \frac{1/2N}{1/2N + 2\mu} = \frac{1}{1 + \theta} \end{aligned}$$

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 ...

Probability that mutations occur in the time in which there are n ancestral lineages

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

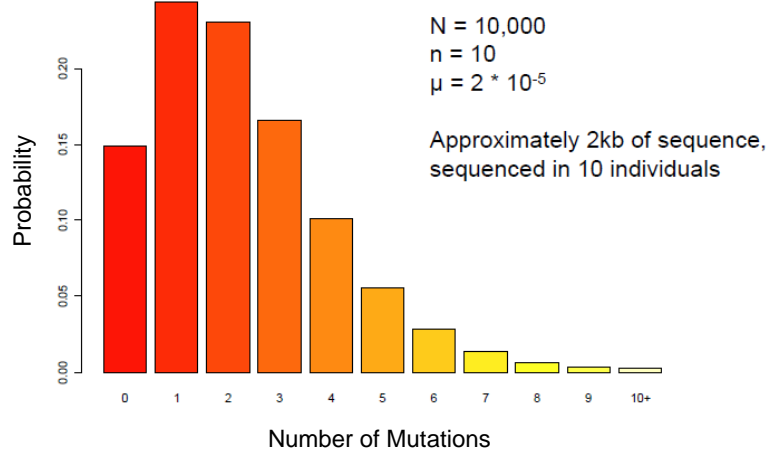
Probability of j segregating sites in a sample of size n

$$P_n(j) = \sum_{i=1}^j P_{n-1}(j-i) Q_n(i)$$

Example ...

- 3 sequences
- Population size $N = 25,000$
- Mutation rate $\mu = 10^{-5}$
- Probability of 0, 1, 2, 3... mutations

Number of Mutations



So far ...

- One homogeneous population
 - Coalescence times
 - Number of mutations
 - Expectation
 - Distribution
 - Spectrum of mutations
- Several assumptions, including ...
 - No recombination

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.

Acknowledgement

- This lecture note is based on material by Profs Goncalo Abecasis (Univ of Michigan) and Itsik Pe'er (Columbia Univ).
- Chapter 3 of *Principles of Population Genetics* by Daniel L. Hartl & Andrew Clark