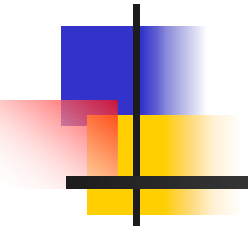
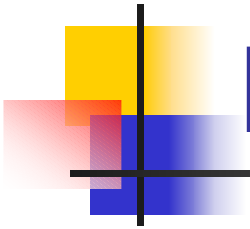


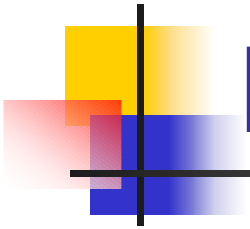
Introduction to Coalescent Models

Biostatistics 666



Last Lecture

- Haplotype Frequencies
- Linkage Equilibrium
- Linkage Disequilibrium
  - Association between neighboring alleles
  - Expected to decrease with distance
- Measures of linkage disequilbrium
  - D, D’ and Δ^2



Modeling populations

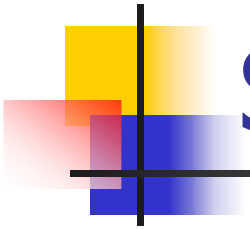
- Important Parameters
  - Mutation rate (μ)
  - Population Size
    - Haploid population (N chromosomes)
    - Diploid population (2N chromosomes)
  - Time (t)
  - Final sample size (n)
  - Also recombination rate, selection, migration

Mutation Model

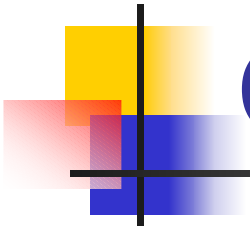
- The mutation process is complex
  - Rate depends on surrounding sequence
  - Reverse mutations are possible

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

Simple Approach: Simulation

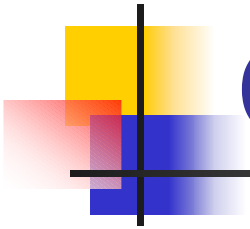


1. N starting sequences
2. Sample N offspring sequences
    1. Apply mutations according to μ
3. Increment time
4. If enough time has passed…
    1. Generate final sample
    2. Stop.
5. Return to step 1.



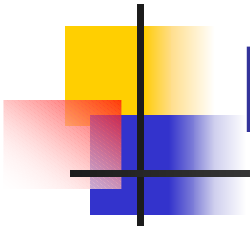
Genealogy

- History of a particular set of sequences
  - Describes their relatedness
  - Specifies divergence times
- Includes only a subset of all sequences
- Most Recent Common Ancestor (MRCA)



Coalescent approach

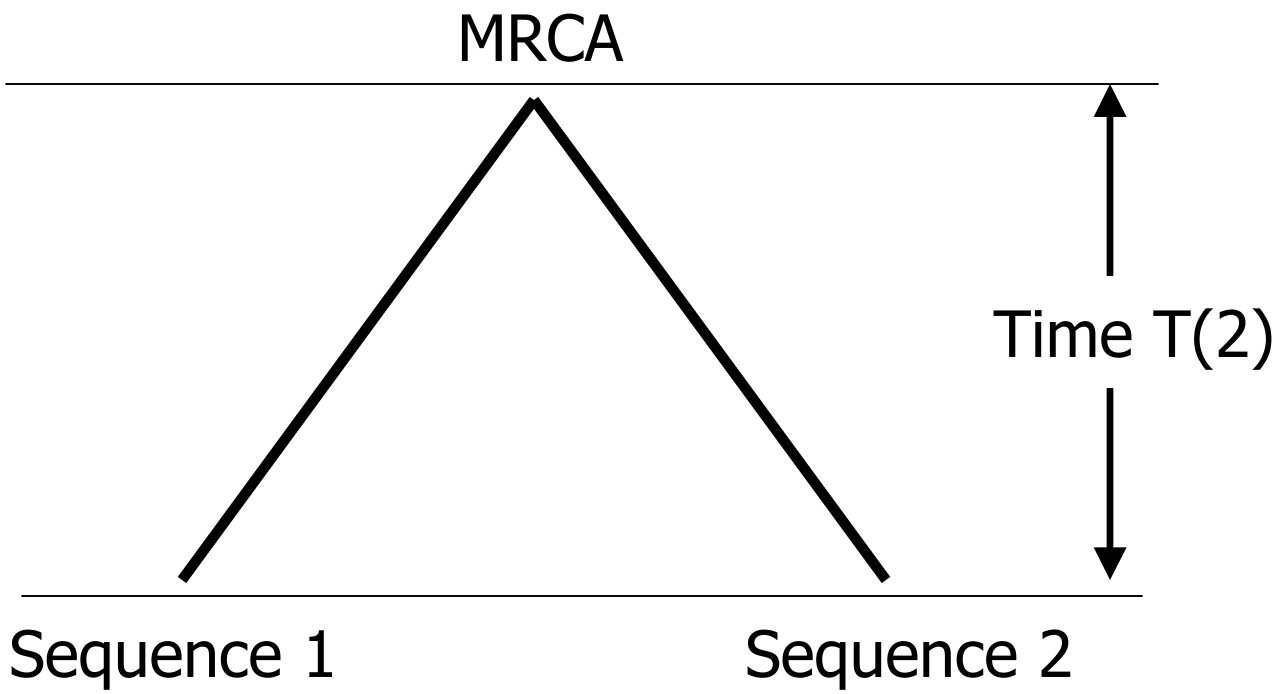
- Generate genealogy for a sample of sequences.
  - Introduces computational and analytical convenience.
- Instead of proceeding forward through time, go backwards!



Example

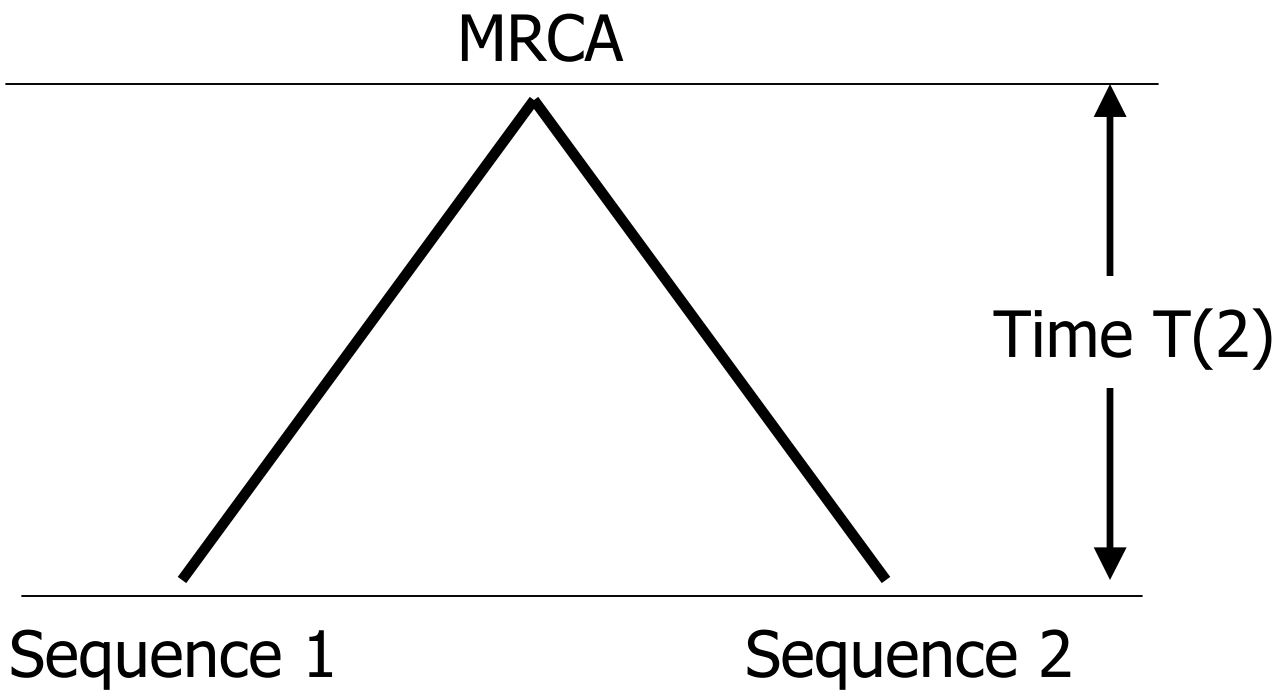
- Sample of two sequences
  - 100 bp each…
- How many differences are expected?
  - Population of size, N = 1000
  - Mutation rate
    - $\mu = 10^{-8}$ / bp / generation
    - $\mu \approx 10^{-6}$ / 100 bp / generation

Genealogy of two sequences

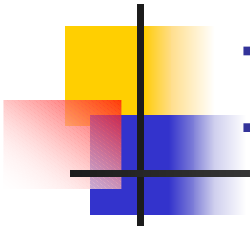


Mutations between MRCA and Sequence 1?

Genealogy of two sequences



Total mutations in genealogy?



In general…

- Number of mutations S
  - Distributed as Poisson
  - $E(S) = \mu E(T_{tot})$
  - $Var(S) = E[Var(S|T)] + Var(E(S|T)$
    $= \mu E(T_{tot}) + \mu^2 Var(T_{tot})$
- T_{tot} is the total length of all branches

Estimating T(2)

- 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

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

$$= 1/N\,(1-1/N)^t$$

$$\approx 1/N\, e^{-(1/N)t}$$

Estimating T(*n*)

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

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

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

$$\approx 1 - \frac{\binom{n}{2}}{N}$$

- Assume:
  - N is large
  - n is small
  - Terms of order N^{-2} can be ignored

Probability of Coalescence at Time t

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

$$\approx \frac{\binom{n}{2}}{N} e^{-\frac{\binom{n}{2}}{N}t}$$

Time to next coalescent event

- Use an exponential distribution to approximate time to next coalescent event…
  - Decay rate $\lambda = n(n-1)/2N$
  - Mean $1/\lambda = 2N/[n(n-1)]$

Rescaled time function T(j)

- For convenience, measure T(j) in units of N generations

$$E(T_j) = 1 / \binom{j}{2}$$

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

$$= \sum_{i=1}^{n-1} \frac{2}{i}$$

Expected number of mutations

- Factor N for diploids, 2N for haploids

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

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

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

- Population geneticists, use θ=4Nμ and *r* for recombination rate

More about S…

- Very large variance

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

- Useful to estimate θ
- Useful to estimate population size N
  - If mutation rate μ is known

Estimating θ …

- Using average number of differences between pairs of sequences

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

- Using total number of variants in sample

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