Introduction to Coalescent Theory

Jon Wilkins
Santa Fe Institute
wilkins@santafe.edu

Beijing CSSS 2007

Overview

- Intro to Coalescent Theory (Today)
- Genomic Imprinting, Mathematical Modeling, and Notions of Optimality in Evolution (Wednesday)
- Statistical Inference in Complex Systems

Ingredients of Natural Selection

- Heritable variation
- Differential reproductive success
- Causal connection between the two

Population Genetics

- How is variation generated and maintained in a population?
- What can patterns of genetic diversity tell us about the history of a population?
  - Demography (migration, reproduction, etc.)
  - Molecular events (mutation, recombination, etc.)
  - Natural selection (directional, purifying, etc.)

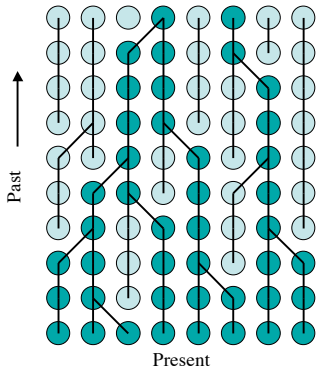
Why diversity?

- Muller - mutation drives deviations from the optimal phenotype
- Dobzhansky - heterogeneous environments / frequency dependent effects
- Lewontin-Hubby experiments (mid 1960s)
  - Too much variation for either explanation
- Kimura - neutral theory

Neutral Theory

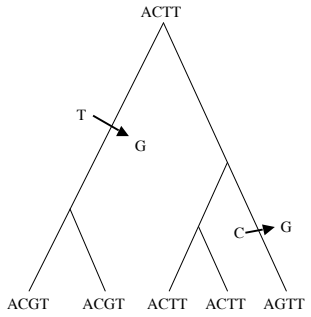
- Not just selective neutrality
- Constant population size
- Well mixed population (panmictic)
- This is always true

Sampling with Replacement



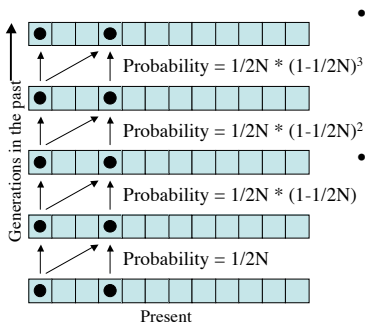
- Some alleles pass on no copies to the next generation, while some pass on more than one
- All that we care about are the ancestors of sequences present in our dataset

The Coalescent



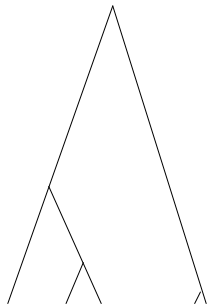
- Homologous genes share a common ancestor
- DNA sequence diversity is shaped by genealogical history
- Genealogies are shaped by chance, demography, selection

Balls in Boxes



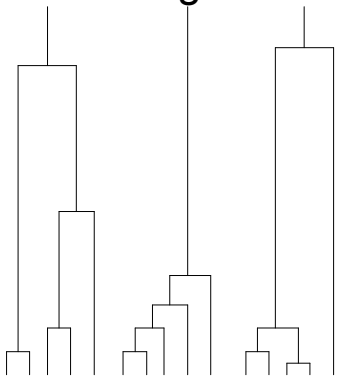
- The coalescent models genealogies backwards in time
- Follow ancestral lineages back until the most recent common ancestor (MRCA) is reached

The Shapes of Genealogies



- Time to the MRCA of a pair of sequences is exponentially distributed with mean time of 2N generations
- Time to the next coalescent event for a sample of n sequences is exponential with mean $2N/\binom{n}{2}$ generations

Genealogies are highly variable



- The variance on the length of each portion of the genealogy is large, on the order of N^2
- Variation in topology as well
- Mutations are random on top of the genealogy

The problem

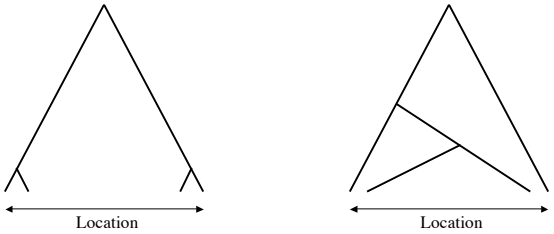
- Want to infer the underlying processes that have shaped genetic diversity, but
- The inherent stochasticity means that any given genealogy is consistent with a wide range of demographic processes
- How do we estimate parameters, and how do we know how good our estimates are?

Estimating N

- Expected pairwise distance (π)
  - 2N times 2μ (= θ)
- Expected number of polymorphisms (S)

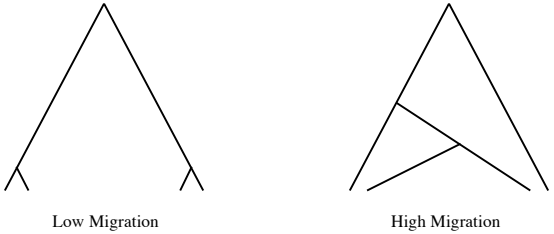
$$\Theta \sum_{i=1}^{n-1} \frac{1}{i}$$

The Structured Coalescent



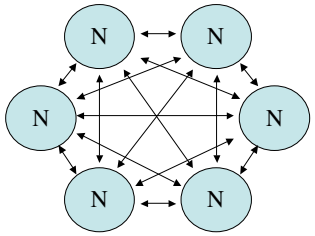
- With geographically structured populations, all topologies are not equally likely

The Structured Coalescent



- The relationship between genealogy and geography can be used to make inferences

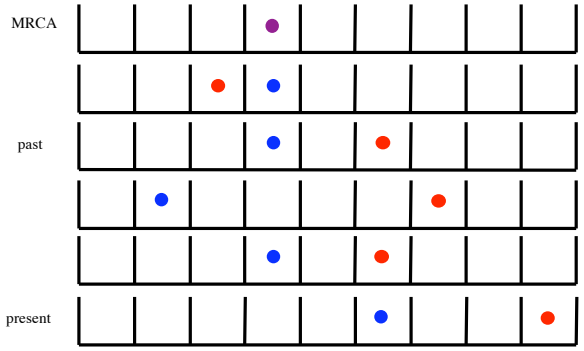
The Island Model



- Each migrant is equally likely to come from any deme
- Population structure, but no geography

$$F_{ST} = \frac{\pi - \pi_0}{\pi} = \frac{1}{1 + 4Nm}$$

A finite, linear habitat



The Solution

$$U_x(x,t) = \sum_{i=1}^{\infty} \frac{\alpha_i f_i(x_0)\cos(\alpha_i x)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)} e^{-2\sigma_m^2 \alpha_i^2 (t-\frac{1}{2})} + \sum_{j=1}^{\infty} \frac{\alpha_j^* f_j^*(x_0)\sin(\alpha_j^* x)}{\alpha_j^* - \sin(\alpha_j^*)\cos(\alpha_j^*)} e^{-2\sigma_m^2 \alpha_j^{*2} (t-\frac{1}{2})}$$

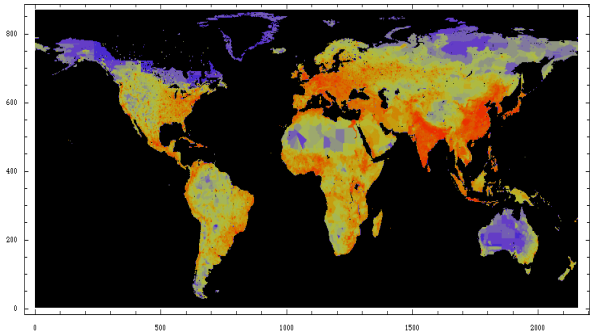
$$U_y(y,t) = \sum_{i=1}^{\infty} \frac{\alpha_i f_i(y_0)\cos(\alpha_i y)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)} e^{-2\sigma_m^2 \alpha_i^2 (t-\frac{1}{2})} + \sum_{j=1}^{\infty} \frac{\alpha_j^* f_j^*(y_0)\sin(\alpha_j^* y)}{\alpha_j^* - \sin(\alpha_j^*)\cos(\alpha_j^*)} e^{-2\sigma_m^2 \alpha_j^{*2} (t-\frac{1}{2})}$$

$$\cot(\alpha_i) = \frac{4N\sigma_m^2 \alpha_i + \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} (-1)^n (2\sigma_m \alpha_i)^{2n-1} \frac{(n-1)!}{(2n-1)!}}{1 + \sum_{n=1}^{\infty} (-\sigma_m^2 \alpha_i^2)^n \Big/ \prod_{m=1}^{n} 2m} \qquad f_i(x_0) = \frac{1}{\sqrt{4\pi\sigma_m^2}} \int_{-1}^{1} \cos(\alpha_i x)(e^{-(x-x_0)^2/4\sigma_m^2} + e^{-(2-x-x_0)^2/4\sigma_m^2}) dx$$

$$-\tan(\alpha_j^*) = \frac{4N\sigma_m^2 \alpha_j^* + \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} (-1)^n (2\sigma_m \alpha_j^*)^{2n-1} \frac{(n-1)!}{(2n-1)!}}{1 + \sum_{n=1}^{\infty} (-\sigma_m^2 \alpha_j^{*2})^n \Big/ \prod_{m=1}^{n} 2m} \qquad f_j^*(x_0) = \frac{1}{\sqrt{4\pi\sigma_m^2}} \int_{-1}^{1} \sin(\alpha_j^* x)(e^{-(x-x_0)^2/4\sigma_m^2} + e^{-(2-x-x_0)^2/4\sigma_m^2}) dx$$

Not trivial to extend to > 1 dimension
Not trivial to extend to > 2 sequences

Realistic Geography



Coalescent Simulations

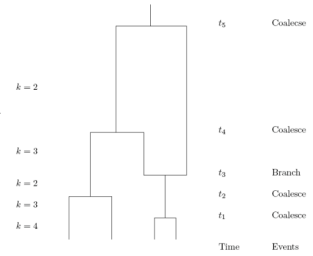
- In most systems of interest, analytic solutions are too cumbersome
- The coalescent provides an efficient framework in which to do simulations
- Must understand how to relate the forward-time system to a corresponding backward-time process

Coalescent Simulations

- In the case of selective neutrality, the genealogical and mutational processes are separable
  - we can produce the genealogy, and then simply place mutations on the tree afterwards
- But, if there is selection, then the reproductive success of an individual depends on its type

Ancestral Selection Graph

- Dual Process
- Lineages coalesce with probability $\binom{k}{2}/2N$
- A lineage splits with probability sk/2



Extracting the genealogy

- Move down the graph, allowing mutations to occur
- Choose the incoming branch with higher fitness

Take-home messages

- The coalescent provides a convenient approach to modeling evolutionary processes
  - Well suited to dealing with data
- Analytic results are accessible only for very simple models
  - In other cases, it produces efficient simulations
- Leaves the question of how to make inferences
  - Come back on Friday