Statistics 877
Statistical Methods for Molecular Biology
Phylogenetics

Cécile Ané

Departments of Botany and of Statistics
University of Wisconsin - Madison

Spring 2017 - Feb. 16
The coalescent model for gene trees
- in one population
- along a species tree
- along a species network
The coalescent model

- Biologists used to build species trees from 1 or 2 genes. Now: whole genomes, new challenges!
- Genes do not all have the same ancestry: recombination is sexual organisms, various processes in bacteria, etc.
- Take a gene at random, found in a human, a chimp and gorilla. 76% chance that the human and the chimp genes are most closely related to each other than to the gorilla. But: 12% and 12% chance that . . .
- There is a simple explanation, no need to invoke hybridization or gene flow between human/chimp/gorilla species after they diverged.
- Coalescent: as a ‘null’ distribution of gene trees.
The coalescent model in one population

Wright-Fisher model:
- constant population size of $N$ individuals
- At generation $g + 1$, form each new individual by sampling from the $N$ individuals present at generation $g$, with replacement.
- No selection: no one individual is sampled with a higher probability than another.
- No population structure

For simplicity: think of each gene copy as 1 individuals, so each host 2 of these individuals. The total size of the gene population $= 2N$ if there are $N$ diploid individuals.
The coalescent probabilities between two individuals

sample $n = 2$ individuals, go backward in time. When do they first share a common ancestor – or coalesce?

$$\Pr\{a \text{ and } b \text{ coalesce at generation } g\} = \left(1 - \frac{1}{N}\right)^{g-1} \frac{1}{N}$$

Geometric distribution, average $g = N$ generations.

for $N$ very large, rescale $g$ into a continuous time scale:

$$t = \# \text{ generations}/N$$

$$\Pr\{a \text{ and } b \text{ coalesce at time } t\} \approx e^{-t} \, dt$$

coalescent time $\sim$ exponential distribution with rate 1
The coalescent probabilities between \( n \) individuals

random sample of \( n \) individuals

- each pair of individuals: equal probability to share a common ancestor at the previous generation: \( 1/N \).
- large \( N \): the probability that 3 individuals share the same common ancestor at the previous generation is negligible: \( n^3 \times 1/N^2 \).
- the minimum of \( n(n - 1)/2 \) exponential waiting times with rate 1 is an exponential waiting time with rate \( n(n - 1)/2 \).

One description of the coalescent process, backwards in time:

1. Pick the time \( T_1 \) to the next coalescent event from an exponential distribution with rate \( n(n - 1)/2 \).
2. At time \( T_1 \), pick a pair of individuals at random to coalesce. Now there are \( n - 1 \) individuals left.
3. Repeat with independent times between coalescent events, until all individuals have coalesced.
The coalescent probabilities in one population

Strong assumptions:
- no population structure: everyone can mate with everyone
- constant population size
- no selection on the gene we are looking at (or on a gene linked to the locus we are looking at)

**Population genetics:** this ancestral tree is a nuisance parameter. The coalescent process is typically use to integrated this tree out.

Now used more and more to compare species: what happened in past large populations left a signature on gene trees.
The coalescent model along a species tree

- Was first mostly used with only one individual per species
- now used more and more with several individuals per species across several species.
- there can be uncertainty in assignment of individuals to ‘species’.

**Task:** sequence a sample of genes, reconstruct each gene tree and the species tree.

**Tool:** see how gene trees can differ from the species tree. We will use gene tree probabilities:

\[ \mathbb{P}\{\text{gene tree } T \mid \text{species tree } \tau\} \] and \[ \mathbb{P}\{\text{sequences } D \mid \text{gene } T\} \]

T: topology with branch lengths, or topology only.
What is the probability of this gene tree, with its branch lengths?

Assume constant population size, and times $t_i$ and $s_j$ scaled as $g/N$. 

$$\mathbb{P}\{\text{gene } T|\tau\} = \prod_i \mathbb{P}\{\text{coalescent events in population } i\}$$
Example: 3-taxon species tree

What is the probability of this gene tree, with its branch lengths?

Assume constant population size, and times $t_i$ and $s_j$ scaled as $g/N$.

\[ e^{-t_1} \]

\[ \mathbb{P}\{\text{gene } T|\tau\} = \prod_i \mathbb{P}\{\text{coalescent events in population } i\} \]
What is the probability of this gene tree, with its branch lengths?

Assume constant population size, and times $t_i$ and $s_j$ scaled as $g/N$.

$$e^{-t_1} \ast 3e^{-3t_2} \frac{1}{3} e^{-(t_3-t_2)}$$

$$\mathbb{P}\{\text{gene } T|\tau\} = \prod_i \mathbb{P}\{\text{coalescent events in population } i\}$$
Example: 3-taxon species tree

What is the probability of this gene tree, with its branch lengths?

Assume constant population size, and times $t_i$ and $s_j$ scaled as $g/N$.

$$e^{-t_1} \times 3e^{-3t_2} \frac{1}{3} e^{-(t_3-t_2)} \times e^{-(s_2-s_1)} \times 1$$

$$\mathbb{P}\{\text{gene } T|\tau\} = \prod_i \mathbb{P}\{\text{coalescent events in population } i\}$$
What is the probability of this gene tree, with its branch lengths?

Assume constant population size, and times $t_i$ and $s_j$ scaled as $g/N$.

$$e^{-t_1} \times 3e^{-3t_2} \frac{1}{3} e^{-(t_3-t_2)} \times e^{-(s_2-s_1)} \times 1$$

$$\times 3e^{-3(t_4-s_2)} \frac{1}{3} e^{-(t_5-t_4)}$$

$$\mathbb{P}\{\text{gene } T|\tau\} = \prod_i \mathbb{P}\{\text{coalescent events in population } i\}$$
**Problem:** We can estimate a gene tree topology from sequences, but sequences do not provide branch lengths in number of generations. We estimate branch lengths in average number of substitutions per site.

If the substitution rate is constant, then we may include it as a parameter.

But most of the time, gene trees are not even ‘ultrametric’.

One way around: consider the likelihood of the gene tree topology, without branch lengths. But this is hard!
Example: 3-taxon species tree

1 individual per species.
What is the probability that the rooted gene $T = \tau$?

Simple on 3 species:
$\mathbb{P}\{ac|b\} = e^{-t_1}/3$, $\mathbb{P}\{bc|a\} = e^{-t_1}/3$
and $\mathbb{P}\{T = \tau = ab|c\} = 1 - 2/3 \times e^{-t_1}$
In more general cases, we need to know the probability that \( j \) individuals present at the beginning of a lineage coalesce down to \( i \) individuals at the end, after \( t \) coalescent time units:

\[
p_{ij}(t) = \sum_{k=j}^{i} e^{k(k-1)t/2} \frac{(2k - 1)(-1)^{k-j}}{j!(k-j)!(j+k-1)} \prod_{y=0}^{k-1} \frac{(j+y)(i-y)}{i+y}
\]

\[
p_{11}(t) = 1, \quad p_{22}(t) = e^{-t}, \quad p_{33}(t) = e^{-3t}
\]
\[
p_{21}(t) = 1 - e^{-t}, \quad p_{32}(t) = \frac{3}{2} e^{-t} - \frac{3}{2} e^{-3t}
\]
\[
p_{31}(t) = 1 - \frac{3}{2} e^{-t} + \frac{1}{2} e^{-3t}
\]
Example: 4-taxon species tree

Regardless of where coalescences occur,

\[ \mathbb{P}\{ T = ((a, (b, c)), d) \} = ? \]

\[ p_{22}(t_1) \frac{1}{3} p_{31}(t_2) \]
Example: 4-taxon species tree

Regardless of where coalescences occur,

\[ \Pr \{ T = ((a, (b, c)), d) \} =? \]

\[ p_{22}(t_1) \cdot \frac{1}{3} p_{31}(t_2) \cdot 1 + p_{22}(t_1) \cdot \frac{1}{3} p_{32}(t_2) \cdot \frac{1}{3} \]
Example: 4-taxon species tree

Regardless of where coalescences occur,

\[ \mathbb{P}\{ T = ((a, (b, c)), d) \} = ? \]

\[
p_{22}(t_1) \cdot \frac{1}{3} p_{31}(t_2) \cdot 1 + p_{22}(t_1) \cdot \frac{1}{3} p_{32}(t_2) \cdot \frac{1}{3} + p_{22}(t_1) \cdot p_{33}(t_2) \cdot \frac{1}{6} \cdot \frac{1}{3}
\]
Species tree estimation

Many methods based on the coalescent

- input = $n$ alignments: Bayesian, MCMC approaches. very heavy computational burden, convergence issues.

- input = $n$ gene trees: ML methods. fast, but dubious assumptions
  as is gene trees have no error
  usually rooted, as if root position has no error
  with or without branch lengths

- input = 1 alignment of $n$ SNPs: maximum pseudo-likelihood
  "SNP trees" integrated out

Alternative: Dirichlet model (Larget et al. 2010, Ané et al. 2007)
Hybridization networks
Hybridization networks
Challenge: expensive computation of likelihood

Gene tree: \(((a, c), (b_1, b_2))\)

figure: from Yu, Dong, Liu, Nakhleh (2014)
Hybridization networks

Multiple gene alignments
→ Multiple gene trees
→ Species network

Markov process
optimize: tree topology, edge lengths

Coalescent & gene transfer processes
optimize: network topology, edge lengths, inheritance probs
PhyloNetworks package

- to speed up calculations and scale to larger networks: maximum pseudolikelihood: 4 species at a time, i.e. quartets

\[
\text{pseudoL}(\text{net}, t, \gamma) = \prod_{q \in Q(\text{net})} L(q, t, \gamma)
\]

- implemented in Julia
  fast and interactive, JIT compiled
  uses NLopt package from JuliaOpt
Evolutionary network for *Xiphophorus* tropical fishes

1183 genes, 24 swordtails and platyfish, 5 hybridizations

- Loss of sword
- Loss of ability to produce sword

Solís-Lemus & Ané (2016), data from Cui et al. (2013)