Statistics 877
Statistical Methods for Molecular Biology
Phylogenetics

Cecile Ane

Departments of Botany and of Statistics
University of Wisconsin - Madison

Spring 2014 - Feb. 11 & 18
Outline

1. Trait evolution on a phylogeny
2. Molecular evolution on a tree: Maximum Likelihood Estimation
3. The coalescent model for gene trees
Outline

1. Trait evolution on a phylogeny
   - Examples
   - Non-independence problem
   - Brownian motion model
Trait evolution on a phylogeny

Sometimes we don’t care about the tree, but we need it.

- In antelopes, are brain size and group size correlated?
- In plants, did the capacity to shed leaves evolve before or after the move to freezing winters?
- In flowers, is the color red correlated with pollination by hummingbirds?
Is red correlated with hummingbirds?

Pollinators in *Iochroma*: hummingbirds, ants, butterflies, flies.

Pollinator importance = visitation rate \times \# ovules potentially pollinated per visit

Predictors: flower color, reward, corolla length, etc.

Smith, Ané & Baum (2008) Evolution
Problem: units are not independent

A sample of species do not form a “random” sample because of a lack of independence.

\[ Y = X\beta + \epsilon \]

but \( \epsilon \)'s are not iid: identically distributed (id) perhaps, but not independent (i).
Using the tree to parametrize the non-independence for quantitative traits:

- Brownian motion model: neutral evolution
- Ornstein-Uhlenbeck model: natural selection towards an selection optimum $\mu$
Brownian motion model for the residuals

\[ B_t - B_s \sim \mathcal{N}(0, \sigma^2(t - s)) \]

\[ \epsilon \sim \mathcal{N}(0, V_\theta) \]

\( V_{\theta, ij} = \sigma^2 \times \text{time shared between tips } i \text{ and } j \)

\[
V_\theta = \sigma^2 \begin{pmatrix}
1 & .8 & 0 & 0 \\
.8 & 1 & 0 & 0 \\
0 & 0 & 1 & .4 \\
0 & 0 & .4 & 1
\end{pmatrix}
\]

\[
V_\theta = \sigma^2 \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]
Likelihood for Brownian motion model

\[ Y = X\beta + \epsilon, \epsilon \sim \mathcal{N}(0, V_\theta) \]

Likelihood:

\[ -2 \log f(Y|\theta) = n \log(2\pi) + \log |V_\theta| + (Y - X\beta)' V_\theta^{-1} (Y - X\beta) \]

Looks simple, but tricky for a tree with 1000’s species. What is the size of \( V_\theta \)?
Areas for current research

- Linear time algorithm: for BM model, but also phylogenetic logistic or Poisson regression (Ho & Ané, 2014, Syst. Biol)
- Model selection: need new tools
Molecular evolution on a tree: Maximum Likelihood Estimation
- Maximum Likelihood Estimation for 2 species
- Likelihood Calculations on Trees
- Search for Maximum Likelihood
Distance Between Pairs of Taxa

- In a two-taxon tree, the distance between two taxa can be estimated under any model by maximum likelihood.
- If the distance is \( t \) and at site \( i \) one species has base A and the other has base C, the contribution to the likelihood at this site \( j \) is
  \[
  L_j(t) = \pi_A P_{AC}(t) = \pi_C P_{CA}(t)
  \]
  for a time-reversible model.
- The overall likelihood is
  \[
  L(t) = \prod_j L_j(t)
  \]
- and the log-likelihood is
  \[
  \ell(t) = \sum_j \log L_j(t) = \sum_j \left( \log \pi_{x[j]} + \log P_{x[j]y[j]}(t) \right)
  \]
Distance Between Pairs of Taxa

- Models with free $\pi$: it is common to estimate $\pi$ with observed base frequencies.
- Other parameters usually estimated by maximum likelihood.
- The simplest models have closed form solutions, others require numerical optimization.
Alignment of \( m \) taxa and \( n \) sites: \( mn \) nucleotide bases.

- Observed base for the \( i \)th taxon and the \( j \)th site: \( x_{ij} \).
Notation for the Tree

- With a time-reversible model, the location of a root (where the CTMC begins at stationarity) does not affect the likelihood calculation.
- We can only infer an unrooted tree. For calculation purposes, we denote node $\rho$ (which could be any node) as the root.
- An unrooted tree with $m$ taxa has $m - 2$ internal nodes. Number the nodes $i = 1, \ldots, 2m - 2$ with the first $m$ for leaf nodes and the last $m - 2$ for internal nodes.
- There are $2m - 3$ edges in the tree, numbered $e = 1, \ldots, 2m - 3$.
- Relative to root node $\rho$, edge $e$ connects parent node $p(e)$ and child node $c(e)$ where $p(e)$ is closer to $\rho$ than $c(e)$.
- Edge $e$ has length $t_e$. 
Notation for Unobserved Data

The likelihood for a tree is computed by summing over all possible bases at the internal nodes for each of the \( n \) sites.

For each site, there are \( 4^{m-2} \) possible allocations of bases at internal nodes we will index by \( k \).

Internal node \( i \) is set to nucleotide \( b_{ik} \) at the \( k \)th allocation, \( i = m + 1, \ldots, 2m - 2 \).

Let \( z(i, j, k) \) be the nucleotide at node \( i \), site \( j \), and allocation \( k \).

\[
z(i, j, k) = \begin{cases} 
x_{ij} & \text{if } i \leq m \ (i \text{ is a leaf node}) \\
b_{ik} & \text{if } i > m \ (i \text{ is an internal node}) 
\end{cases}
\]
Likelihood of a Tree

- Let $P(t)$ be the $4 \times 4$ probability transition matrix over an edge of length $t$.
- The likelihood of the tree is

$$\prod_j \sum_k \left( \prod_z \pi_{z(p, j, k)} \prod_e P_z(p(e), j, k) z(c(e), j, k)(t_e) \right)$$

- Notice that the sum is over the $4^{m-2}$ possible allocations.
- A naive calculation would not be tractible for large trees.
Felsenstein’s Pruning Algorithm

- Felsenstein’s pruning algorithm is an example of dynamic programming.
- By saving partial calculations, the time complexity of the likelihood evaluation grows linearly with the number of sites, not exponentially.
- For each site and node, the algorithm depends on calculating the probability in the subtree rooted at that node for each possible base.
- The algorithm begins at the leaves of the tree and recurses to the root.
- The likelihood of the site is a weighted average of the conditional subtree probabilities at the root weighted by the stationary distribution.
For a single tree topology, the ML estimation requires optimization of branch lengths and of any parameters in the substitution model.

Numerical optimization methods are required even for simple models and small trees.
The search for the maximum likelihood tree conceptually requires obtaining the maximum likelihood for each possible tree topology and then picking the best of these. For more than a dozen or so taxa, exhaustive search is non-feasible. Heuristic search algorithms typically define a neighborhood structure for possible topologies. The search goes through neighbors and jumps to the first neighbor with a higher likelihood. When all neighbors are inferior to the current tree, the search stops. Much improvement has been made in recent years (RAxML and GARLI are two modern ML programs).
The coalescent model for gene trees
- The coalescent model in one population
- The coalescent model along a species tree
- A Dirichlet model for gene tree discordance
The coalescent model

- Biologists used to build species trees from 1 or 2 genes. Now: all genomes are becoming available. It brings fundamental challenges.
- Genes do not all have the same ancestry! Recombination is sexual organisms, and through various processes in other organisms.
- 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 . . .
- This phenomenon can be explained really simply. No need to invoke hybridization or any kind of 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

- Take a sample of \( n = 2 \) individuals, then go backward in time. When do they first share a common ancestor—or coalesce?

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

Geometric distribution. On average, it takes \( g = N \) generations.

- If \( N \) is very large, rescale \( g \) into a continuous time scale:

\[
t = \# \text{ generations} / N
\]

then:

\[
P\{a \text{ and } b \text{ coalesce at time } t\} \approx e^{-t} dt
\]

the coalescent time has an exponential distribution with rate 1.
The coalescent probabilities between $n$ individuals

Take a random sample of $n$ individuals.

- Each pair of individuals has an equal probability to share a common ancestor at the previous generation: $1/N$.
- With 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:

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

\(T\): 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 $g/N$.

$$\mathbb{P}\{\text{gene } T|\tau\} = \prod_i \mathbb{P}\{\text{coalescent events in population } i\}$$
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 $g/N$.

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

$$\mathbb{P}\{\text{gene } T|\tau\} = \prod_i \mathbb{P}\{\text{coalescent events in population } i\}$$
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 $g/N$.

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

\[ \mathbb{P}\{\text{gene } T|\tau\} = \prod_i \mathbb{P}\{\text{coalescent events in population } i\} \]
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 \( g/N \).

\[
e^{-t_1} \ast 3e^{-3t_2} \frac{1}{3} e^{-(t_3-t_2)} \ast e^{-(s_2-s_1)} \ast 1 \\
\ast 3e^{-3(t_4-s_2)} \frac{1}{3} e^{-(t_5-t_4)}
\]

\[
\mathbb{P}\{\text{gene } T|\tau}\} = \prod_i \mathbb{P}\{\text{coalescent events in population } i\}
\]
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:
\[ P\{ac|b\} = e^{-t_1}/3, \ P\{bc|a\} = e^{-t_1}/3 \]
and \[ P\{T = \tau = ab|c\} = 1 - 2/3 * e^{-t_1} \]
Gene tree topology probabilities

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,

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

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

Regardless of where coalescences occur,

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

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

Regardless of where coalescences occur,

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

\[ p_{22}(t_1) \cdot \frac{1}{3} p_{31}(t_2) \cdot 1 + p_{22}(t_1) \cdot \frac{1}{3} p_{32}(t_2) \cdot \frac{1}{3} + p_{22}(t_1) \cdot p_{33}(t_2) \cdot \frac{1}{6} \cdot \frac{1}{3} \]
Using the coalescent model to estimate species trees

Active research area

- **Maximum likelihood**: species tree from \( n \) gene trees
  - Assumes known ‘perfect’ gene trees.
  - No general ML method for species tree given sequences.
- **Bayesian**: posterior distribution of the species tree given sequences. MCMC approach.
  - Very heavy computational burden. Convergence issues.
- **Dubious assumptions**:
  - Branch lengths: in substitutions/site in gene trees, coalescent units in species tree \( (g/N) \). Constant substitution rate assumed for each gene, need ad-hoc way to scale branch lengths.
  - Placement of the root in gene trees assumed perfectly known.
A Dirichlet model for gene tree discordance

Alternative for gene tree / species tree reconstruction.

- Bayesian approach
- does not use the coalescent model: no hypothesis about source of discordance
Prior on gene trees: Dirichlet Process

Cluster = group of genes with the same tree topology

\[
\mathbb{P}\{T_1, \ldots, T_k | \alpha\} = f(\# \text{ clusters } c, \text{ cluster sizes } k_1, \ldots, k_c, \alpha)
= \frac{1}{A_k(\alpha)} \prod_{\text{cluster } j=1}^{c} A_{k_j}(\alpha / T)

T = \text{total # trees, } \\
A_k(x) = x(x+1) \cdots (x+k-1).
\]

\(\alpha = 0\): 1 cluster.
\(\alpha = \infty\): independent trees.

posterior distribution of gene trees:

\[
\mathbb{P}\{T_1, \ldots, T_k | D_1, \ldots, D_k, \alpha\} = \frac{\mathbb{P}\{T_1, \ldots, T_k | \alpha\} \prod_{i=1}^{k} \mathbb{P}\{D_i | T_i\}}{\mathbb{P}\{D_1, \ldots, D_k\}}
\]

implemented in BUCKy.
The concordance approach

Example: 30,040 genes on 5 species: human, chimp, gorilla, orangutan, rhesus macaque.

- No restriction on gene tree branch lengths, runs fast.
- can be used to test the coalescent model: here, not rejected.
Current active research areas

- Improvements to BUCKy: improve prior distribution and accuracy of the computational approach.
- More general tests of the coalescent model
- Estimate a species network instead of species tree: to include hybridization or ‘gene flow’ events