Species Tree Inference

Laura Kubatko
Departments of Statistics and Evolution, Ecology, and Organismal Biology
The Ohio State University
kubatko.2@osu.edu

July 27, 2012
Relationship between population genetics and phylogenetics

- **Population genetics**: Study of genetic variation within a population

- **Phylogenetics**: Use genetic variation between taxa (species, populations) to infer evolutionary relationships

- Often assume:
  - Each taxon is represented by a single sequence – this is often called “exemplar sampling”
  - We have data for a single gene and wish to estimate the evolutionary history for that gene (the *gene tree* or *gene phylogeny*)
Given current technology, we could do much more:

- Sample many individuals within each taxon (species, population, etc.)
- Sequence many genes for all individuals

Need models at two levels:

- Model what happens within each population (standard population genetics)
- Apply within-population models to each population represented on a phylogeny (more recent work)
Wright-Fisher Model

Assumptions:

- Population of $2N$ gene copies
- Discrete, non-overlapping generations of equal size
- Parents of next generation of $2N$ genes are picked randomly with replacement from preceding generation (genetic differences have no fitness consequences)
- Probability of a specific parent for a gene in the next generation is $\frac{1}{2N}$
Wright-Fisher Model
Wright-Fisher Model
Wright-Fisher Model
Wright-Fisher Model
Wright-Fisher Model
Wright-Fisher Model
Wright-Fisher Model
Wright-Fisher Model
Wright-Fisher Model
Wright-Fisher Model
The Coalescent Model

- **Discrete Time Coalescent**
  - \( P(\text{two genes have same parent in the previous generation}) = \frac{1}{2N} \)
  - Number of generations since two genes first shared a common ancestor \( \sim \text{Geometric}\left(\frac{1}{2N}\right) \)
  - Number of generations since at least two genes in a sample of \( k \) shared a common ancestor \( \sim \text{Geometric}\left(\frac{k(k-1)}{4N}\right) \)
Coalescent Theory

- Use the coalescent model to link our population genetics model to our phylogenetic model – recall a few important ideas about Kingman’s coalescent from Peter’s lecture:

- The time to coalescence of two lineages in a sample of $n$ lineages is exponentially distributed with rate $(\frac{n}{2})$ (in coalescent units)

- The probability that $u$ lineages coalesce into $v$ lineages in time $t$ is given by (Tavare, 1984; Watterson, 1984; Takahata and Nei, 1985; Rosenberg, 2002)

$$P_{uv}(t) = \sum_{j=\nu}^{u} e^{-j(j-1)t/2} \frac{(2j-1)(-1)^{j-\nu}}{\nu!(j-\nu)!(\nu+j-1)} \prod_{y=0}^{j-1} \frac{(\nu+y)(u-y)}{u+y}$$
Coalescent Theory

- When \( u \) and \( v \) are small, these are easy to compute. For example,

\[
P_{21}(t) = \text{probability that 2 lineages coalesce to 1 lineage in time } t
\]

\[
= \text{probability of 1 coalescent event in time } t \text{ when } n=2
\]

\[
= P(T \leq t), \text{ where } T \sim \operatorname{Exp}(\mu = 1)
\]

\[
= \int_{0}^{t} e^{-x} \, dx = 1 - e^{-t}
\]

- Similarly,

\[
P_{22}(t) = \text{prob. of no coalescence in time } t \text{ for 2 lineages}
\]

\[
= P(T > t)
\]

\[
= \int_{t}^{\infty} e^{-x} \, dx = e^{-t}
\]
The Coalescent Model Along a Species Tree

- So far, we’ve considered the coalescent process within a single population.
- A phylogenetic tree consists of many populations followed throughout evolutionary time:

![Phylogenetic Tree Diagram]
The Coalescent Model Along a Species Tree

- Goal is to apply coalescent model across the phylogeny. The basic assumption is that events that occur in one population are independent of what happens in other populations within the phylogeny.

- More specifically, given the number of lineages entering and leaving a population, coalescent events within populations are independent of one another.

- It is also important to recall an assumption we “inherit” from our population genetics model: all pairs of lineages are equally likely to coalesce within a population.
The Coalescent Model Along a Species Tree

- When talking about gene tree distributions, there are two cases of interest:
  - The gene tree topology distribution
  - The joint distribution of topologies and branch lengths

- Start with the simple case of 3 species with 1 lineage sampled in each and look at the gene tree topology distribution
Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

Example of gene tree probability computation (for simplicity, let’s use coalescent units for our time scale):

(a) Prob = $1 - e^{-t}$; (b), (c), (d) Prob = $\frac{1}{3} e^{-t}$
Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

Thus, we have the following probabilities:

- Gene tree \((A,(B,C))\): \(\text{prob} = 1 - e^{-t} + \frac{1}{3} e^{-t} = 1 - \frac{2}{3} e^{-t}\)
- Gene tree \((B,(A,C))\): \(\text{prob} = \frac{1}{3} e^{-t}\)
- Gene tree \((C,(A,B))\): \(\text{prob} = \frac{1}{3} e^{-t}\)

Note: There are two ways to get the first gene tree. We call these histories.

The probability associated with a gene tree topology will be the sum over all histories that have that topology.
Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

- What are these probabilities like as a function of $t$, the length of time between speciation events?

(b) $\text{prob} = 1 - \exp(-t)$

$\text{prob} = (1/3)\exp(-t)$

$\text{prob} = (1/3)\exp(-t)$

$\text{prob} = (1/3)\exp(-t)$

(c) 

$\text{Top} \text{ology } \text{Probability}$

$t$ (Coalescent Units)
Consider 4 taxa – the human-chimp-gorilla problem
Coalescent Histories for the 4-taxon Example

- There are 5 possibilities for this example:
In the general case, we have the following:
The probability of gene tree $g$ given species tree $S$ is given by

$$P\{G = g | S\} = \sum_{\text{histories}} P\{G = g, \text{history} | S\}$$
In the general case, we have the following:

The probability of gene tree $g$ given species tree $S$ is given by

$$P\{ G = g \mid S \} = \sum_{\text{histories}} P\{ G = g, \text{history} \mid S \}$$

- Implemented in the software COAL (Degnan and Salter, Evolution, 2005)
- A more efficient method has been proposed (Wu, Evolution, 2012)

Examined 23,210 distinct alignments for 5 primate taxa: Human, Chimp, Gorilla, Orangutan, Rhesus

Looked at distribution of gene trees among these taxa - observed strongly supported incongruence only among the Human-Chimp-Gorilla clade.
Applications of the Topology Distribution - Example 1

Chimp
  Human
  Gorilla
  Orangutan

Gorilla
  Human
  Chimp
  Orangutan

Gorilla
  Chimp
  Human
  Orangutan
Applications of the Topology Distribution - Example 1

Observed proportions of each gene tree among ML phylogenies

- Chimp
  - Human
  - Gorilla
  - Orangutan
  - 76.6%

- Gorilla
  - Human
  - Chimpanzee
  - Orangutan
  - 11.4%

- Gorilla
  - Chimp
  - Human
  - Orangutan
  - 11.5%
Applications of the Topology Distribution - Example 1

- Observed proportions of each gene tree among ML phylogenies
  - Chimp
    - Human
    - Gorilla
    - Orangutan
    - 76.6%
    - 79.1%
  - Gorilla
    - Human
    - Chimp
    - Orangutan
    - 11.4%
    - 9.9%
  - Gorilla
    - Chimp
    - Human
    - Orangutan
    - 11.5%
    - 9.9%

- Predicted proportions using parameters from Rannala & Yang, 2003.
Applications of the Topology Distribution - Example 2

- In the previous example, one topology is clearly preferred
- Must the distribution always look this way?
- Examine entire distribution when the number of taxa is small
Applications of the Topology Distribution - Example 2

- Consider 4 taxa: A, B, C, and D
- Species tree:

```
A   B   C   D
\   \   \   \n  x   y   z
```

- Look at probabilities of all 15 tree topologies for values of x, y, and z
Applications of the Topology Distribution - Example 2

\[ y=1, \quad x=1 \]

\[ y=0.01, \quad x=0.01 \]

\[ y=1, \quad x=0.001 \]
Applications of the Topology Distribution - Example 2

y=1, x=1

y=0.01, x=0.01

y=1, x=0.001

A

B

C

D
Applications of the Topology Distribution - Example 2

y=1, x=1

y=0.01, x=0.01

y=1, x=0.001
The existence of anomalous gene trees has implications for the inference of species trees.

Rosenberg & Tao, *Systematic Biology*, 2008
What about mutation? How does this affect data analysis?

The coalescent gives a model for determining gene tree probabilities for each gene.

View DNA sequence data as the result of a two-stage process:

- Coalescent process generates a gene tree topology.
- Given this gene tree topology, DNA sequences evolve along the tree.
Given this model, how should inference be carried out?

**Hypothesis:** As more data (genes) are added, the process of estimating species trees from concatenated data can be **statistically inconsistent**.

- May fail to converge to any single tree topology if there are many equally likely trees.
- May converge to the wrong tree when a gene tree that is topologically incongruent with the species tree has the highest probability.
Applications of the Topology Distribution - Example 3

Species Tree

A → Generate gene trees in COAL → Generate sequence data in Seq-Gen → Estimate tree using concatenation

Repeat 100 times
Applications of the Topology Distribution - Example 3

Simulation Study 1

A  \( x = 0.01, y = 2.0 \)

B  \( x = 0.05, y = 1.0 \)

C  \( x = 0.1, y = 1.0 \)

D  \( x = 0.1568, y = 0.1568 \)

E  \( x = 0.01, y = 1.0 \)

F  \( x = 0.05, y = 0.05 \)

G  \( x = 0.1, y = 0.05 \)

H  \( x = 0.25, y = 0.01 \)
Applications of the Topology Distribution - Example 3

Simulation Study 2

![Graph showing the relative frequency of inferring the topology distribution for different branch lengths. The graph compares the relative frequency of inferring the true topology (MT) to the false topology (S1) for different number of genes (100, 50, 20, 10). The x-axis represents branch length, and the y-axis represents the relative frequency of inferring the true topology.](image-url)
Performance of the Concatenation Approach:

- Can be statistically inconsistent when branch lengths in the species phylogeny are sufficiently small
- May perform poorly even when branch lengths are only moderately short
- Bootstrap procedure can be positively misled in this situation

Question: How does the bootstrap perform in these cases?
Hypothesis: The bootstrap may provide strong support for the incorrect tree when gene trees that are incongruent with the species tree are fairly probable.

Simulation study to examine the performance of the bootstrap:

- \( n = 100 \) loci
- \( x = 0.01, y = 1.0 \)
- \( \theta = 0.001 \)
- \( B = 200 \) bootstrap samples per repetition
- Repeated 500 times
The Concatenation Approach – Performance of the Bootstrap

A

**ML Topology is MT**

B

**ML Topology is MT**

---

**ML Topology is S1**

---

**ML Topology is ST**

---

Percent Bootstrap Support for (A,B) Percent Bootstrap Support for (B,C,D)

---
The Concatenation Approach – Performance of the Bootstrap

- The bootstrap can be *positively misleading* – show strong support for an incorrect clade

- **Important note:** This is NOT a failing of the bootstrap methodology; the observed “poor” performance is due to the use of an incorrect model (concatenation)

- **Question:** Is there a better way to estimate species phylogenies?
The Concatenation Approach – Performance of the Bootstrap

The bootstrap can be *positively misleading* – show strong support for an incorrect clade

Important note: This is NOT a failing of the bootstrap methodology; the observed “poor” performance is due to the use of an incorrect model (concatenation)

Question: Is there a better way to estimate species phylogenies?

Explicitly model the coalescent process!
Coalescent-based Methods for Species Tree Inference

- **Summary statistic methods:** Start with estimated gene trees
  - Using estimated branch lengths:
    - STEM (Kubatko et al. 2009)
    - STAR, STEAC (Liu et al. 2009)
  - Using topology information only:
    - Minimize Deep Coalescences (PhyloNet; Than & Nakhleh 2009)
    - MP-EST (Liu et al. 2010)
    - ST-ABC (Fan and Kubatko 2011)
    - STELLS (Wu 2011)

- **Methods that utilize the full data:** Input is aligned sequences
  - BEST (Liu and Pearl 2007)
  - *BEAST (Heled and Drummond 2010)
  - New method based on algebraic statistics (Chifman and Kubatko 2012)
Comparison of approaches:

- **Summary statistic methods**
  - Advantage: Quick
  - Disadvantage: Ignore information in data
  - Most current implementations do not easily allow for assessment of uncertainty

- **Full data methods**
  - Advantage: Fully model-based framework
  - Disadvantage: Computationally intensive, sometimes prohibitively so
  - Both BEST and *BEAST utilize a Bayesian framework and involve MCMC
Likelihood Function

- Suppose that we have available alignments for $N$ genes, denoted by $D_1, D_2, \ldots, D_N$

- We would like to find the likelihood of the species phylogeny given these $N$ alignments, assuming that
  - individual gene trees are randomly generated according to the coalescent model
  - evolution of sequences along fixed gene trees occurs following a standard nucleotide-based Markov model
  - the data for the genes are independent given the species tree and associated parameters
Maddison (1997) provided the likelihood function in this context:

\[
L(S|D_1, D_2, ... D_N) = \prod_{i=1}^{N} P(D_i|S) = \prod_{i=1}^{N} \sum_{g_j}^{G} P(D_i|g_j)f(g_j|S)
\]

where \( S \) is the species tree (topology and branch lengths) and \( g_j \) represents a gene tree.

This likelihood is difficult to evaluate directly, because of the dimension of the inner sum (which is really an integral).

To deal with this, either assume gene trees are known (summary statistics), use Bayesian techniques, or think about small problems.
A simpler problem is to suppose that our data consist of a set of gene trees

Let $g_1, g_2, \ldots, g_N$ be a set of $N$ gene trees with branch lengths (assume known without error)

Consider a species tree, $S$ (topology and branch lengths)

The likelihood function is

$$L(S|g_1, g_2, \ldots, g_N) = \prod_{j=1}^{N} P(G = g_j | S)$$

where $P(G = g | S)$ is given by Rannala and Yang (2003).
Assumptions

- No recombination within loci
- Free recombination between loci
- No gene flow following speciation
- Only source of variability in single-gene histories is due to the coalescent process
- There is a single $\theta$ for the entire tree, for each locus
- Evolutionary rates may vary across loci
Liu et al. (2009) showed that the ML estimate of the species tree can be computed by sequentially clustering minimum observed divergence times between pairs of species across genes.

They have shown that when gene trees are known without error, the ML species tree is a consistent estimator.

A similar result was obtained by Roch & Mossel (2010) – they call their estimator the GLASS tree (an acronym for Global LAteSt Split, based on the algorithm they developed to compute it).

STEM computes the ML estimate of the species tree this way.
The results of Liu et al. (2009) can be extended to derive the ML estimates of the speciation times for an arbitrary species tree.

Thus, the likelihood of any species tree can be readily computed by using this result to obtain ML branch lengths.

This is important in that it allows us to compare alternative phylogenetic hypotheses.
A simulated annealing algorithm is used to search the space of all species trees for trees that have high likelihoods.

The $k$ best trees found during the search are saved and printed to a file ($k$ is set by the user).

Exploration of the likelihood surface is particularly important for many of these problems.

The details of the simulated annealing algorithm are similar to those given in Salter & Pearl (2001).
Features of STEM

- No limits (that I know of) on the number of taxa or the number of loci.
- Can handle intraspecific sampling.
- Allows information concerning mutation rate for each locus to be used in the analysis.
- Can handle different taxon samples across genes.
- Fairly robust to missing data.
- Bootstrapping can be used to provide measures of support on the clades in the species tree.
- Version 2.0 is written in Java (using Clojure).
STEM Data Preparation - Gene Trees

- STEM takes as its input one gene tree for each locus.

- Thus, a first step in an analysis using STEM is to estimate gene trees with branch lengths for each locus.

- Any method can be used to do this, but note a couple requirements:
  - Branch lengths are assumed to be in units of expected number of substitutions per site per unit time.
  - Branch lengths must be estimated subject to a molecular clock. This is not checked by the program.
  - Gene trees must be fully resolved; however, polytomies can be included by setting branch lengths to 0 for an arbitrary resolution of the polytomy.
A value of the parameter $\theta = 4N\mu$ must be provided. Note that this is the “per-site $\theta$”, not a “per-locus” value as used by other population genetics programs.

This will be used to convert gene tree branch lengths to coalescent units (number of $2N$ generations) by dividing all gene tree branch lengths by $\theta$.

Estimates of $\theta$ could be obtained by standard methods. Typical values of $\theta$ will be between 0.001 and 0.1.

Each locus can also be given a rate multiplier.

These can adjust for
- Variation in mutation rate across loci.
- Ploidy (e.g., haploid loci – mtDNA – should be given a rate of 0.5).
STEM: Strengths and Weaknesses

STEM makes some strong assumptions:

- Error in estimating gene trees and branch lengths is not incorporated (but the bootstrap helps with this).
- Information in the sequence data is not used directly; it is only used as summarized by estimated gene divergence times.
- There is a single value of $\theta$ for the entire tree.
STEM: Strengths and Weaknesses

STEM makes some strong assumptions:

- Error in estimating gene trees and branch lengths is not incorporated (but the bootstrap helps with this).
- Information in the sequence data is not used directly; it is only used as summarized by estimated gene divergence times.
- There is a single value of $\theta$ for the entire tree.

There are trade-offs involved, and STEM does some things well:

- It is quick (even the tree search does not take long).
- It can handle missing data easily and intuitively.
- Simulations demonstrate reasonable performance (unlikely to be misleading; may be uninformative).
STEM – Performance with Missing Data

- Look at simulation to assess performance of STEM in the presence of substantial missing data
STEM Extensions

- STEM is used for species delimitation in SpedeSTEM (Ence and Carstens, 2011)

- STEM-hy uses the STEM likelihood calculation to examine hybridization (Kubatko 2009)

- Newest version (which includes analyzes using hybridization) is available at: http://www.stat.osu.edu/~lkubatko/software/STEM-hy.

- There is an important difference in how STEM v1.1 and STEM-hy/STEM v2.0 treat observed divergence times of 0. In version 1.1, this divergence time is assigned a very small value (0.000001), while in version 2.0, it is treated as missing data.
Full Data Methods: BEST and *BEAST

- Model the entire process of data generation:
  - Species tree → gene trees [coalescent process]
  - Gene trees → sequence data [standard nucleotide substitution models]

- Goal of both methods is to estimate the posterior distribution of the species tree and associated model parameters

- BEST and *BEAST use slightly different algorithms – we will briefly discuss the main ideas of each
BEST: Overview of Method

Hierarchical model

- Species trees
  - Prob(S) Uniform Distribution
- Gene trees
  - Prob(G|S) Coalescent Theory
- Sequence Data
  - Prob(D|G) Markovian substitution model
BEST: Overview of Method

Assumptions:

- Given the species tree, the gene trees are conditionally independent.
- Given the gene tree, the DNA sequences are conditionally independent of the species tree.
- Random mating in each population.
- No gene flow after species divergence.
- No recombination within a locus.

Source: Lecture 7, Stat 882 Spring 2010, Dennis Pearl
BEST: Overview of Method

- BEST uses MCMC to sample from the joint posterior distribution of the gene trees and the species tree:

\[
f(S, G|D) = \frac{f(D|G)F(G|S)f(S)}{f(D)}
\]

Source: Lecture 7, Stat 882 Spring 2010, Dennis Pearl
BEST: Overview of Method

Implementation: MrBayes with BEST

- Step 1: Use MrBayes to propose vectors of joint gene trees (unlinked).
- Step 2: Given those gene trees: propose a compatible species tree.
- Step 3: Implement the chain fully within MrBayes using the usual properties of the MCMC as proposed by the user.

Source: Lecture 7, Stat 882 Spring 2010, Dennis Pearl
BEST: Overview of Method

- **BEST algorithm:**
  - DNA sequences
  - MrBayes (preset best=1)
    - Unrestricted gene tree moves
    - Restricted species tree moves
  - Posterior of G,S
    - Annealing step to speed convergence
    - Neighborhood of MT to increase acceptance rate

Source: Lecture 7, Stat 882 Spring 2010, Dennis Pearl
The MCMC algorithm in *BEAST also samples both gene trees and the species tree at each step of the algorithm \(\implies\) same goal as BEST (to get posterior distribution of species tree).

The method of moving in gene trees is one reason that *BEAST works more quickly than BEST.

Authors have pointed out substantial increases in speed [we'll discuss an empirical example a little later].
Both methods use MCMC, which means:

- Need to think carefully about setting prior distributions.
- Need to check carefully for convergence, in all parameters.
- Need to choose methods of summarizing the estimated posterior distribution carefully, and interpret these summaries correctly.
- etc., .......
Comparison of STEM with *BEAST (He, Lanier, and Knowles, submitted)
Comparison of Methods – Some Simulation Results

Comparison of *BEAST and STEM in the missing data study

![Comparison of Methods: BEAST and STEM](image)
Putting It All Together (Plus Some): An Empirical Example

- North American Rattlesnakes - Joint work with Dr. Lisle Gibbs (EEOB at OSU)
- Of interest evolutionarily because of the diversity of venoms present in the various species and subspecies.
- Of conservation interest because population sizes in the eastern subspecies are very small.

Pictures by Jimmy Chiucchi and Brian Fedorko
Geographic Distribution of Snake Populations

[Map showing the distribution of snake populations across the United States, labeled with different colors and regions designated as Smm, Sm, Smb, Sce, Scm, Sct, and Scc.]
Sistrurus rattlesnakes

- Data: 7 (sub)species, 26 sequences, 19 genes

<table>
<thead>
<tr>
<th>Species</th>
<th>Location</th>
<th>No. of individuals per gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. catenatus catenatus</td>
<td>Eastern U.S. and Canada</td>
<td>9</td>
</tr>
<tr>
<td>S. c. edwardsii</td>
<td>Western U.S.</td>
<td>4</td>
</tr>
<tr>
<td>S. c. tergeminus</td>
<td>Western and Central U.S.</td>
<td>5</td>
</tr>
<tr>
<td>S. miliarius miliarius</td>
<td>Southeastern U.S.</td>
<td>1</td>
</tr>
<tr>
<td>S. m. barbouri</td>
<td>Southeastern U.S.</td>
<td>3</td>
</tr>
<tr>
<td>S. m. streckeri</td>
<td>Southeastern U.S.</td>
<td>2</td>
</tr>
<tr>
<td>Agkistrodon sp. (outgroup)</td>
<td>U.S.</td>
<td>2</td>
</tr>
</tbody>
</table>
Individual Gene Tree Estimates

Some are very informative:
Individual Gene Tree Estimates

Some are a little informative:
Individual Gene Tree Estimates

And then there are others .....
Example: Sistrurus Rattlesnakes ... Species Tree Estimation

STEM, STEAC

BEAST (concatenated data), *BEAST

BEST, Parsimony & MrBayes (concatenated data)

PhyloNet, STAR
Example: Sistrurus Rattlesnakes ... Species Tree Estimation

Some observations:

- Estimate from PhyloNet places *S. c. catenatus* as sister to the entire clade – it turns out this is due to only two gene trees. If those genes are removed, the estimate agrees with STEM.

- The portion of the tree that differs between STEM, *BEAST*, and BEST is the arrangement of the *S. miliarius* subspecies – all three arrangements are observed.

- **Both** BEST and *BEAST* have trouble converging: BEST did not converge in the branch length parameters, while *BEAST* did not converge in the effective population size parameters, especially for the tip species (same problem?).

- *BEAST* was much faster than BEST (days vs. months for ~ 350 million iterations) – but with an older version of BEST.
Example: Sistrurus Rattlesnakes ... Species Tree Estimation

Evaluate alternative trees:

(a) $-\log L = 11070.48$

\[
\begin{array}{c}
\text{Smb} \\
\text{Sms} \\
\text{Smm} \\
\text{Sct} \\
\text{Sce} \\
\text{Scc} \\
\text{Ac}
\end{array}
\]

(b) $-\log L = 11071.87$

\[
\begin{array}{c}
\text{Smm} \\
\text{Sms} \\
\text{Smb} \\
\text{Sct} \\
\text{Sce} \\
\text{Scc} \\
\text{Ac}
\end{array}
\]

(c) $-\log L = 11071.87$

\[
\begin{array}{c}
\text{Smm} \\
\text{Smb} \\
\text{Sms} \\
\text{Sct} \\
\text{Sce} \\
\text{Scc} \\
\text{Ac}
\end{array}
\]

(d) $-\log L = 11082.32$

\[
\begin{array}{c}
\text{Sms} \\
\text{Smb} \\
\text{Smm} \\
\text{Sct} \\
\text{Scc} \\
\text{Sce} \\
\text{Ac}
\end{array}
\]

(e) $-\log L = 11083.71$

\[
\begin{array}{c}
\text{Smm} \\
\text{Smb} \\
\text{Sms} \\
\text{Sct} \\
\text{Sce} \\
\text{Scc} \\
\text{Ac}
\end{array}
\]

(f) $-\log L = 11083.71$

\[
\begin{array}{c}
\text{Sms} \\
\text{Smb} \\
\text{Smm} \\
\text{Sct} \\
\text{Scc} \\
\text{Sce} \\
\text{Ac}
\end{array}
\]

(g) $-\log L = 11082.32$

\[
\begin{array}{c}
\text{Smm} \\
\text{Smb} \\
\text{Sms} \\
\text{Sct} \\
\text{Sce} \\
\text{Scc} \\
\text{Ac}
\end{array}
\]

(h) $-\log L = 11083.71$

\[
\begin{array}{c}
\text{Smm} \\
\text{Smb} \\
\text{Sms} \\
\text{Sct} \\
\text{Scc} \\
\text{Sce} \\
\text{Ac}
\end{array}
\]
Example: Sistrurus Rattlesnakes ... Species Tree Estimation

Evaluate alternative trees:

- *BEAST posterior probabilities were $< 95\%$
- *BEAST estimated short branches in this area of the tree
A New Method Based on Algebraic Statistics

Data: DNA sequences for gene $i$

Example:

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Human</td>
<td>GCCGATGCGCGATGCGCGAA</td>
</tr>
<tr>
<td>(B) Chimp</td>
<td>GCCGTTGCGCGTTGCGCGTT</td>
</tr>
<tr>
<td>(C) Gorilla</td>
<td>GCGGAAGCGGAAGCGGAAGCGCGAA</td>
</tr>
</tbody>
</table>
A New Method Based on Algebraic Statistics

- Data: DNA sequences for gene $i$, $D_i$
- Example:

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Sequence</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Human</td>
<td>GCCG</td>
<td>A</td>
</tr>
<tr>
<td>(B) Chimp</td>
<td>GCCG</td>
<td>T</td>
</tr>
<tr>
<td>(C) Gorilla</td>
<td>GCGG</td>
<td>A</td>
</tr>
</tbody>
</table>

- Assume each site in the sequence evolves independently of other sites
- Data are assumed to be an iid sample of sites: $(D_i)_j = \text{data at the tips of the tree for site } j \text{ in gene } i$
- Consider site pattern probabilities – for example, $p_{ATA}$
Let $T$ be an $n$-leaf, rooted, binary tree with distribution of states $\Pi = (\pi_1, \pi_2, \cdots, \pi_k)$ at the root.

Edges $e$ of $T$ are labeled by $k \times k$ transition probability matrices $M_e$ that give the probabilities of changes in state from a node to its child.

Let $X_H$ be the state of taxon $H$.

Together $(T, \{M_e\}, \Pi)$ define the joint distribution at the leaves of tree:

$$ p_{ijk} = P(X_A = i, X_B = j, X_C = k) $$

$$ = \sum_r \sum_s \pi_r M_1(r, s) M_2(s, i) M_3(s, j) M_4(r, k) $$
A New Method Based on Algebraic Statistics

- This gives site pattern probability distribution on a gene tree.
- We want the site pattern probability distribution on a species tree under the coalescent model.
- To get this:
  - Find a good post-doc, and allow her to spend a month coding all histories [in mathematica]
  - Evaluate the species tree likelihood function using mathematica for small problems (3 and 4 taxa) and simple models (Jukes-Cantor)
  - Hope you see a pattern
**Definition:** A split of a set is a bipartition of the set of taxa into two groups. A split $A \mid B$ of the leaves of a tree $T$ is valid for $T$ if the induced tree $T\mid_A$ and $T\mid_B$ do not intersect.

- **Valid:** 12|34
- **Not valid:** 13|24
**Definition:** flattenings

\[ p_{ijkl} = P(X_1 = i, X_2 = j, X_3 = k, X_4 = l) \]

\[ \text{Flat}_{12|34}(P) = \begin{pmatrix}
p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\
p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\
p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\
p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\
p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots \\
\end{pmatrix} \]

**Theorem** (Chifman and Kubatko 2012):

- If \( A|B \) is a valid split for \( T \), the \( \text{rank}(\text{Flat}_{A|B}(P)) \leq 10 \).
- If \( C|D \) is not a valid split for \( T \), then generically \( \text{rank}(\text{Flat}_{C|D}(P)) = 16 \).
What does this mean?

- When the Flat matrix is constructed for a true tree, some columns will be “duplicates” of the others in some sense.

- When the Flat matrix is constructed for a false tree, all columns are independent of the others.

- We do not have the $p_{ijkl}$ directly (they come from the true underlying tree with its branch lengths and substitution model parameters), but we can estimate them with data.

- **Idea:** Construct an estimate Flat matrix, $\hat{Flat}(P)$, and use a measure of whether all columns are independent. We use *singular value decomposition* (SVD).
Species tree inference

- Would like to use these ideas to estimate a species tree when given multi-locus data for \( L \) genes, \( D_1, D_2, \ldots D_L \)

- Issue 1:
  
  - The model assumes each site has its own gene tree, i.e.,
    \( (D_1)_1 \) arises from gene tree \( (G_1)_1 \), \( (D_1)_2 \) arises from \( (G_1)_2 \), etc. . . .
    \( (D_2)_1 \) arises from gene tree \( (G_2)_1 \), \( (D_2)_2 \) arises from \( (G_2)_2 \), etc. . . .
    . . .
    \( (D_L)_1 \) arises from gene tree \( (G_L)_1 \), \( (D_L)_2 \) arises from \( (G_L)_2 \), etc. . . .

  - Multilocus phylogenetics generally assumes that each gene has a single underlying tree, i.e.,
    \( (D_1)_1, (D_1)_2, \ldots \) arise as iid observations from \( G_1 \)
    \( (D_2)_1, (D_2)_2, \ldots \) arise as iid observations from \( G_2 \)
    . . .
    \( (D_L)_1, (D_L)_2, \ldots \) arise as iid observations from \( G_L \)
Goal 2: Species tree inference

- Issue 2:
  - Appropriate for SNPs (=single nucleotide polymorphism)
  - May need to worry about ascertainment – SNP data commonly include only variable sites
  - For our example data:

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Human</td>
<td>GCCGATGCCGATGCCGAA</td>
</tr>
<tr>
<td>(B) Chimp</td>
<td>GCCGGTTGCCGTTGCGTT</td>
</tr>
<tr>
<td>(C) Gorilla</td>
<td>GCGGAAAGCGGAAAGCGGAA</td>
</tr>
</tbody>
</table>

this would be

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Human</td>
<td>CATCATCAA</td>
</tr>
<tr>
<td>(B) Chimp</td>
<td>CTTCTTCTT</td>
</tr>
<tr>
<td>(C) Gorilla</td>
<td>GAAGAAGAA</td>
</tr>
</tbody>
</table>
Simulation 1: Effect of substitution model

Simulate 5,000,000 base pairs from JC and GTR models
Simulation 2: Effect of sample size

Vary the number of bp simulated [5,000; 10,000; 50,000; 100,000]
Simulation 3a: Effect of ascertainment

Remove constant sites from datasets - 5 million bp

SVD Score, 5 million bp

- ((AB),(CD))
- ((AC),(BD))
- ((AD),(BC))

All Sites
Non-constant sites
Simulation 3b: Effect of ascertainment

Remove constant sites from datasets - 5,000 bp

SVD Score, 5,000 bp

((AB),(CD)) ((AC),(BD)) ((AD),(BC))
All Sites
Non-constant sites
SVD Score, 5,000 bp
0.0 0.5 1.0 1.5 2.0 2.5 3.0

((AB),(CD)) ((AC),(BD)) ((AD),(BC))

[Legend: All Sites, Non-constant sites]
Simulation 4: “SNP” vs. “Multilocus” data

“SNP” dataset contains 5,000,000 bp
“Multilocus” dataset has 10,000 genes with 500bp per gene
Simulation 5: Effect of sample size for multilocus data

Vary the number of genes [10; 20; 100; 200], each with 500bp per gene
Compare effect of sample size for “SNP” vs. “Multilocus” data
Apply Method to Empirical Example: Sistrurus Rattlesnakes

19 genes, \( \sim 8500 \) bp

\[
\begin{array}{cccc}
Scc & Sce & Sms & Smm \\
SVD \text{ score} &=& 3.84 \times 10^{-14} \\
Scc & Sms & Sce & Smm \\
SVD \text{ score} &=& 5.395 \\
Sce & Sms & Scc & Smm \\
SVD \text{ score} &=& 5.396
\end{array}
\]
The only theoretical results thus far are for 4 taxa, but the use of flattenings in the single gene case extends beyond 4 taxa.

A straightforward method is to consider all (or a sample of) quartets, and use a quartet-method to reconstruct the species tree.

We’ll examine this type of method next week.
What about methods that do not assume the coalescent process?

- Bayesian Concordance Analysis (BCA) – implemented in BUCKy (Bayesian Untangling of Concordance Knots)

  Idea: Estimate the proportion of the genome that has a certain clade

  Build a tree that consists of those clades that are inferred to be true for a high proportion of the genome
First step: Run MrBayes to get estimated posterior distribution for each gene.

Run BCA, a second MCMC step which utilizes the individual gene posterior distributions as input.

The BCA method clusters genes into a number of groups, so that genes in the same group are assumed to share the same gene tree.
The user specifies a prior distribution on the number of groups. This controls how much discordance among gene trees is expected.

- At one extreme, all gene trees are assumed to be the same, and the method mimics concatenation.
- At the other, all gene trees are assumed to be completely independent, and the method mimics a consensus method.

Goal is to estimate concordance factors for all possible clades – percent of genes for which that clade is true. These are often displayed in a primary concordance tree.
Results using the Ebersberger et al. data we looked at earlier:

![Tree Diagram]

- Numbers above nodes are genome-wide concordance factors
- Intervals below nodes are 95% credibility intervals.
BCA: Application to Rattlesnakes

The Primary Concordance Tree from BUCKy
Example: Sistrurus Rattlesnakes ... Species Tree Estimation

- Agreement across methods concerning relationships within subspecies S. catenatus
- Relationships with S. miliarius are much more poorly resolved – all methods indicate uncertainty in these relationships
Example: Sistrurus Rattlesnakes ... Species Tree Estimation

- Agreement across methods concerning relationships within subspecies *S. catenatus*
- Relationships with *S. miliarius* are much more poorly resolved – all methods indicate uncertainty in these relationships

But have we learned anything?

Gloyd, 1940
The individual gene tree estimates contain important information about the species history. How else might they be used to learn about species history?

Example question of interest: Within *S. catenatus*, is the eastern subspecies distinct from the western subspecies?

This will impact conservation status of this group.
Recall that we observed monophyly of the sequences for the eastern subspecies for at least a few genes. Such monophyly could happen just by chance, or it could reflect distinctiveness between the eastern and western subspecies within *S. catenatus*. 
Example: Sistrurus Rattlesnakes .... Beyond the Species Tree

- **Idea:** [Rosenberg (2007)] Given an observation of monophyly, what’s the chance of seeing such an event under a null model of random branching within the species?

- Consider $c$ total lineages, with $a$ lineages from group $A$ and $b$ lineages from groups other than $A$. Under the Yule model, the probability of the lineages being monophyletic is given by

  $$P_A(a, b) = \frac{2}{\binom{a+b}{a}} \frac{a + b}{a(a + 1)}$$

- The probability of groups $A$ and $B$ with $a$ and $b$ lineages being reciprocally monophyletic is

  $$P_{AB}(a, b) = \frac{2}{\binom{a+b}{a}} \frac{1}{a + b - 1}$$
Within *S. catenatus*, there are 9 individuals from the eastern subspecies (Scc), and 9 individuals from the western subspecies (Sce, Sct) – so there are 18 sequences each, for a total of 36 sequences.

Using the formulas of Rosenberg (2007), we have

- $P_A(18, 18) = 2.32 \times 10^{-11}$
- $P_{AB}(18, 18) = 6.30 \times 10^{-12}$
So it would be unusual to observe monophyly or reciprocal monophyly of the eastern and western subspecies if in fact they were not distinct.
So it would be unusual to observe monophyly or reciprocal monophyly of the eastern and western subspecies if in fact they were not distinct.

We observe monophyly of this group in 7 of the 19 genes.
So it would be unusual to observe monophyly or reciprocal monophyly of the eastern and western subspecies if in fact they were not distinct.

We observe monophyly of this group in 7 of the 19 genes.

The probability of observing monophyly of the samples from the eastern subspecies in seven or more of the 19 genes under the null model of random branching is

\[ \sum_{j=7}^{19} \binom{19}{j} (2.32 \times 10^{-11})^j (1 - 2.32 \times 10^{-11})^{19-j} = 1.82 \times 10^{-70} \approx 0 \]
The observed levels of monophyly would be extremely unlikely under a model of random branching. There is strong evidence for distinctiveness of these subspecies.
The observed levels of monophyly would be extremely unlikely under a model of random branching. There is strong evidence for distinctiveness of these subspecies.

Look at results for other subspecies:

<table>
<thead>
<tr>
<th>Subspecies</th>
<th>No. of loci supporting monophyly</th>
<th>No. of loci contradicting monophyly</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. c. catenatus</td>
<td>7</td>
<td>5</td>
<td>$1.82 \times 10^{-70}$</td>
</tr>
<tr>
<td>S. c. edwardsii</td>
<td>1</td>
<td>11</td>
<td>0.00014</td>
</tr>
<tr>
<td>S. m. streckeri</td>
<td>5</td>
<td>3</td>
<td>$9.46 \times 10^{-10}$</td>
</tr>
<tr>
<td>S. m. barbouri</td>
<td>2</td>
<td>6</td>
<td>0.02715</td>
</tr>
</tbody>
</table>
Beyond the Species Tree – the gsi Statistic

- Another idea: [Cummings et al. 2008] Genealogical sorting index
- Quantify the extent of exclusive ancestry within a group – derive a measure on a scale of 0-1, where 1 indicates complete monophyly
- Intuition of measure: minimum number of nodes on a fully resolved tree required to unite a group, divided by the actual number of nodes required to unite the group

\[ gs = \frac{n}{\sum_{u=1}^{U} (d_u - 2)} \]

where \( d_u \) is the degree of node \( u \)
and \( n \) is minimum number of nodes required to unite \( n + 1 \) taxa
Beyond the Species Tree – the gsi Statistic

Note that

\[ \max(\text{gs}) = 1 \]

and

\[ \min(\text{gs}) = \frac{n}{\sum_{i=1}^{l} (d_u - 2)} \]

where \( l \) is the total number of nodes on the tree.

Note that the minimum depends on the number of taxa, while the max does not.

Use the statistic:

\[ gsi = \frac{gs - \min(\text{gs})}{\max(\text{gs}) - \min(\text{gs})} \]
Beyond the Species Tree – the gsi Statistic

- Carry out a test of significance to determine whether the extent of genealogical exclusivity deviates significantly from what would be expected by chance. A p-value is computed using permutation.

- Implemented in R, able to combine information across loci

- Results for all subspecies across all 19 genes:
Beyond the Species Tree – the gsi Statistic

- Carry out a test of significance to determine whether the extent of genealogical exclusivity deviates significantly from what would be expected by chance. A p-value is computed using permutation.

- Implemented in R, able to combine information across loci

- Results for all subspecies across all 19 genes:

<table>
<thead>
<tr>
<th>Subspecies</th>
<th>egsi</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. c. catenatus</td>
<td>0.7357</td>
<td>&lt;0.00000001</td>
</tr>
<tr>
<td>S. c. edwardsii</td>
<td>0.4318</td>
<td>&lt;0.00000001</td>
</tr>
<tr>
<td>S. c. tergeminus</td>
<td>0.4097</td>
<td>&lt;0.00000001</td>
</tr>
<tr>
<td>S. m. miliarius</td>
<td>0.5470</td>
<td>0.0015</td>
</tr>
<tr>
<td>S. m. streckeri</td>
<td>0.5498</td>
<td>&lt;0.00000001</td>
</tr>
<tr>
<td>S. m. barbouri</td>
<td>0.6464</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
Conclusions – Rattlesnakes

- Much evidence that the eastern subspecies (S. c. catenatus) is distinct from the western subspecies. We suggest that S. c. catenatus be elevated to full species status.
Conclusions – Rattlesnakes

- Much evidence that the eastern subspecies (S. c. catenatus) is distinct from the western subspecies. We suggest that S. c. catenatus be elevated to full species status.

- Relationships within S. miliarius are much less resolved. We need to collect more data to determine whether speciation has occurred within this group.
Conclusions – Rattlesnakes

- Much evidence that the eastern subspecies (S. c. catenatus) is distinct from the western subspecies. We suggest that S. c. catenatus be elevated to full species status.

- Relationships within S. miliarius are much less resolved. We need to collect more data to determine whether speciation has occurred within this group.

- For these data, a combination of approaches was useful: both formal species tree inference, and examination of relationships within and among gene trees.
Conclusions – Understanding Species Histories

- Failure to incorporate the coalescent model in estimation of the species tree can lead to statistical inconsistency, even when a method that is statistically consistent is applied.

- Many new methods for inferring species trees are being developed – each has its advantages and disadvantages.

- In addition, we should continue to think about other ways of using multi-locus data to its full advantage.... beyond estimation of the species tree.

- Lots of areas emerging: species delimitation, incorporating horizontal events along the phylogeny, etc. – get involved and have fun!