Species Tree Inference using SVDquartets

Laura Kubatko and Dave Swofford

Joint work with
Julia Chifman, American University
Colby Long, MBI at OSU

January 31, 2017

Goal of this work:

Develop a full data approach that is computationally feasible for large-scale data

How?

- Summarize data differently, so that model requires less computation
- Develop theory to infer relationships among quartets of taxa very accurately
- Use a quartet assembly method to build a large tree

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

Taxon	Sequence		
(A) Human	GCCG	A	TGCCGATGCCGAA
(B) Chimp	GCCG	T	TGCCGTTGCCGTT
(C) Gorilla	GCGG	A	AGCGGAAGCGGAA

- Assume each site in the sequence evolves independently of other sites
- Data are assumed to be an iid sample of sites:
  $(D_i)_j$ = data at the tips of the tree for site j in gene i
- Consider site pattern probabilities – for example, $\tilde{p}_{ATA}$

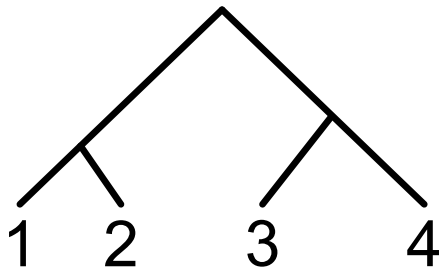
Model: Species tree $\rightarrow$ gene trees $\rightarrow$ data

- species tree $\rightarrow$ gene tree ::: coalescent process
- gene tree $\rightarrow$ data ::: nucleotide substitution models: GTR+I+Γ and submodels

Idea: compute site pattern probabilities under this model for 4 taxa by enumerating all possibilities for simple models

- Tedious, but not difficult
- Look for algebraic structure in the site pattern probabilities

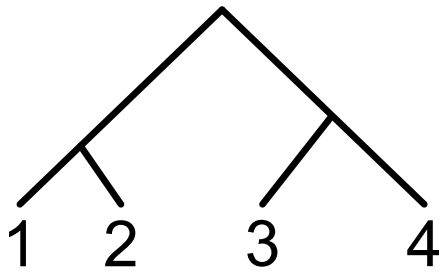
Methods – data representation



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

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

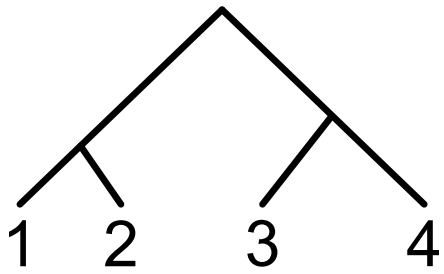
Methods – data representation



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots \\ [AA] & \mathbf{5} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

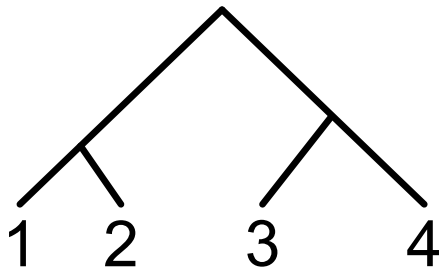
Methods – data representation



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

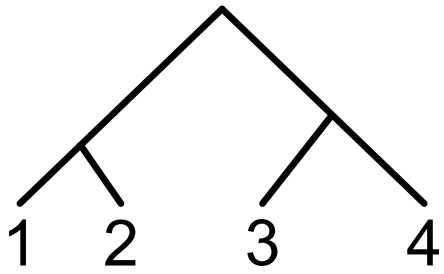
$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots \\ [AA] & \mathbf{5} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & \mathbf{2} & p_{CACA} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

Methods – data representation



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [\mathbf{AC}] & [AG] & [AT] & [\mathbf{CA}] & \cdots \\ [AA] & \mathbf{5} & p_{\mathbf{AAAC}} & p_{AAAG} & p_{AAAT} & p_{\mathbf{AACA}} & \cdots \\ [AC] & p_{ACAA} & p_{\mathbf{ACAC}} & p_{ACAG} & p_{ACAT} & p_{\mathbf{ACCA}} & \cdots \\ [AG] & p_{AGAA} & p_{\mathbf{AGAC}} & p_{AGAG} & p_{AGAT} & p_{\mathbf{AGCA}} & \cdots \\ [AT] & p_{ATAA} & p_{\mathbf{ATAC}} & p_{ATAG} & p_{ATAT} & p_{\mathbf{ATCA}} & \cdots \\ [CA] & p_{CAAA} & p_{\mathbf{CAAC}} & p_{CAAG} & \mathbf{2} & p_{\mathbf{CACA}} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [\mathbf{AC}] & [AG] & [AT] & [\mathbf{CA}] & \cdots \\ [AA] & \mathbf{5} & p_{\mathbf{AAAC}} & p_{AAAG} & p_{AAAT} & p_{\mathbf{AACA}} & \cdots \\ [AC] & p_{ACAA} & p_{\mathbf{ACAC}} & p_{ACAG} & p_{ACAT} & p_{\mathbf{ACCA}} & \cdots \\ [AG] & p_{AGAA} & p_{\mathbf{AGAC}} & p_{AGAG} & p_{AGAT} & p_{\mathbf{AGCA}} & \cdots \\ [AT] & p_{ATAA} & p_{\mathbf{ATAC}} & p_{ATAG} & p_{ATAT} & p_{\mathbf{ATCA}} & \cdots \\ [CA] & p_{CAAA} & p_{\mathbf{CAAC}} & p_{CAAG} & \mathbf{2} & p_{\mathbf{CACA}} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

These two columns are identical – matrix rank is reduced by one

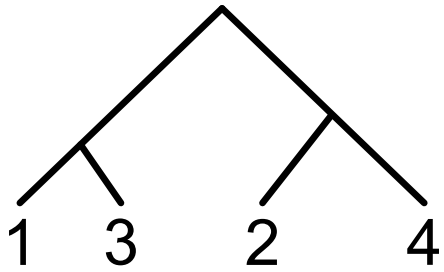
Main Result (Chifman and Kubatko, 2015):

- Species tree inference: For a flattening matrix constructed on the true four-taxon tree, **the matrix rank is 10** under the following model
  - species tree $\rightarrow$ gene tree ::: coalescent process
  - gene tree $\rightarrow$ data ::: nucleotide substitution models: GTR+I+Γ and submodels

New Result (Long and Kubatko, 2017):

- This result holds even in the absence of a molecular clock or when population sizes change along the tree

What about the incorrect tree?



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

$$\mathbf{Flat}_{13|24}(\mathbf{P}) = \begin{pmatrix} & [AA] & [\mathbf{AC}] & [AG] & [AT] & [\mathbf{CA}] & \cdots \\ [AA] & \mathbf{5} & p_{\mathbf{AAAC}} & p_{AAAG} & p_{AAAT} & p_{\mathbf{AACA}} & \cdots \\ [AC] & p_{ACAA} & p_{\mathbf{ACAC}} & p_{ACAG} & p_{ACAT} & p_{\mathbf{ACCA}} & \cdots \\ [AG] & p_{AGAA} & p_{\mathbf{AGAC}} & p_{AGAG} & p_{AGAT} & p_{\mathbf{AGCA}} & \cdots \\ [AT] & p_{ATAA} & p_{\mathbf{ATAC}} & p_{ATAG} & p_{ATAT} & p_{\mathbf{ATCA}} & \cdots \\ [CA] & p_{CAAA} & p_{\mathbf{CAAC}} & p_{CAAG} & \mathbf{2} & p_{\mathbf{CACA}} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

These two columns are no longer identical – full rank matrix in both cases (rank = 16)

Extensions of the Main Result

- Arbitrary number of states, κ, under the coalescent model:
  - If $A|B$ is a valid split for a tree T, then $rank(Flat_{A|B}(P)) \leq \binom{\kappa+1}{2}$.
  - If $C|D$ is not a valid split for a tree T, then $rank(Flat_{C|D}(P)) > \binom{\kappa+1}{2}$.
  - The species tree is completely determined by knowledge of valid splits on all quartets.

- Single underlying gene tree (no coalescent assumption):
  - If $A|B$ is a valid split for a tree T, then $rank(Flat_{A|B}(P)) \leq 4$.
  - If $C|D$ is not a valid split for a tree T, then $rank(Flat_{C|D}(P)) = 16$.
  - The species tree is completely determined by knowledge of valid splits on all quartets.

- **Basic idea:**
  - Data: aligned DNA sequences for multiple loci or for a collection of SNPs
  - Construct the flattening matrix
  - Compute some measure of how close the observed flattening matrix is to a matrix with rank 10

    We use singular value decomposition (SVD) of the flattening matrix – define the SVD score for a split $A|B$ to be

$$SVDscore(Flat_{A|B}(\hat{P})) = \sqrt{\sum_{i=11}^{16} \sigma_i^2}$$

    where σ_i^2 is the i^{th} singular value of the matrix $Flat_{A|B}(\hat{P})$.
  - Pick tree relationships that give the best value of the measure in the previous step

Multi-locus vs. SNP data

The theory is developed for the SNP setting – why do we think this might be ok for multilocus data?

Consider the case of three possible gene trees with the probabilities below under the coalescent model:

- Gene tree 1 — $p_1 = 0.4$
- Gene tree 2 — $p_2 = 0.3$
- Gene tree 3 — $p_3 = 0.3$

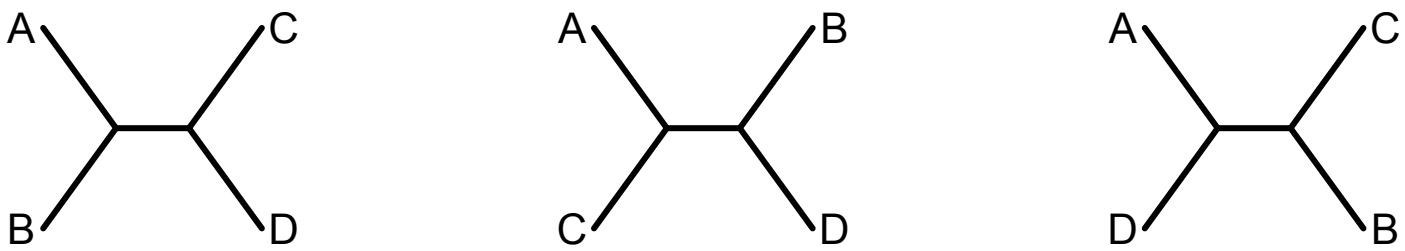
Now suppose we observe multilocus data for 1,000 genes as follows:

- Gene tree1 — 380 genes
- Gene tree 2 — 300 genes
- Gene tree 3 — 320 genes

Then, if the genes are equal in length, the proportion of sites coming from each tree is approximately what is predicted under the SNP model.

Application: Species tree estimation under the coalescent

Main idea: use the observed site pattern distribution to provide information about which of the three possible splits for a set of four taxa is the true split.

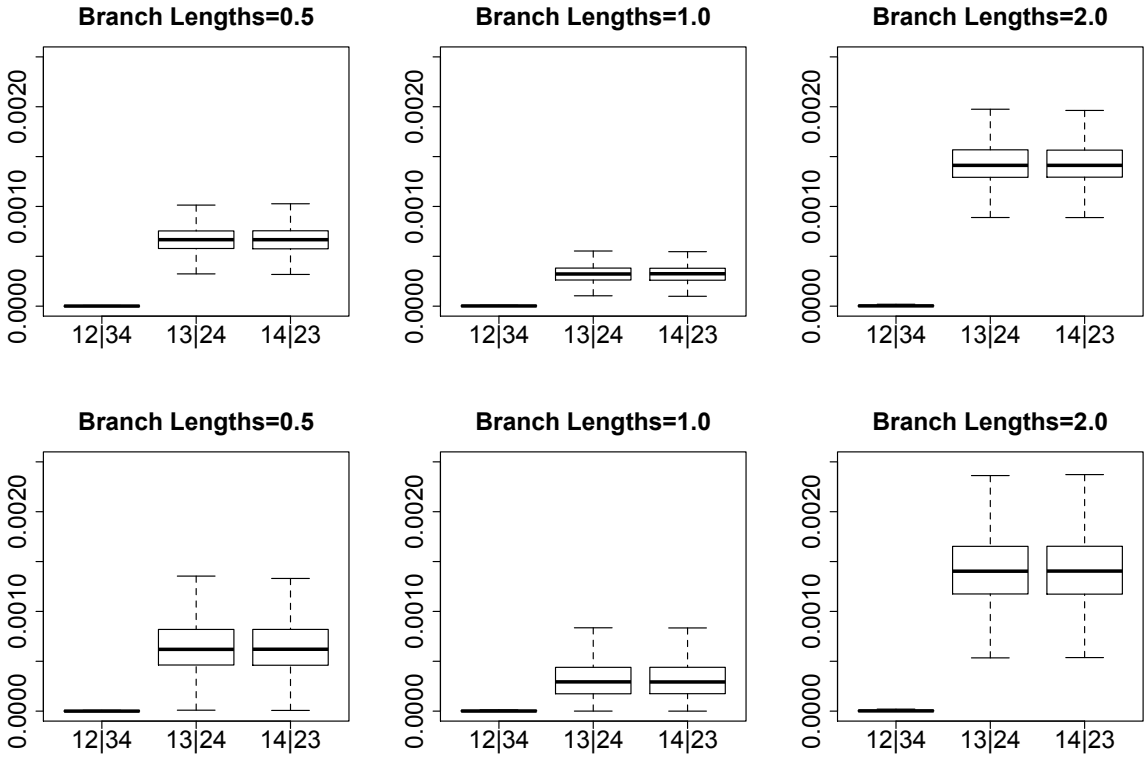


Compute a score for each split in a given quartet of taxa and choose the split with the best (lowest) score.

Simulation study 1 – can we detect the correct split?

Simulate data from the Jukes-Cantor model for a 4-taxon tree and examine split scores

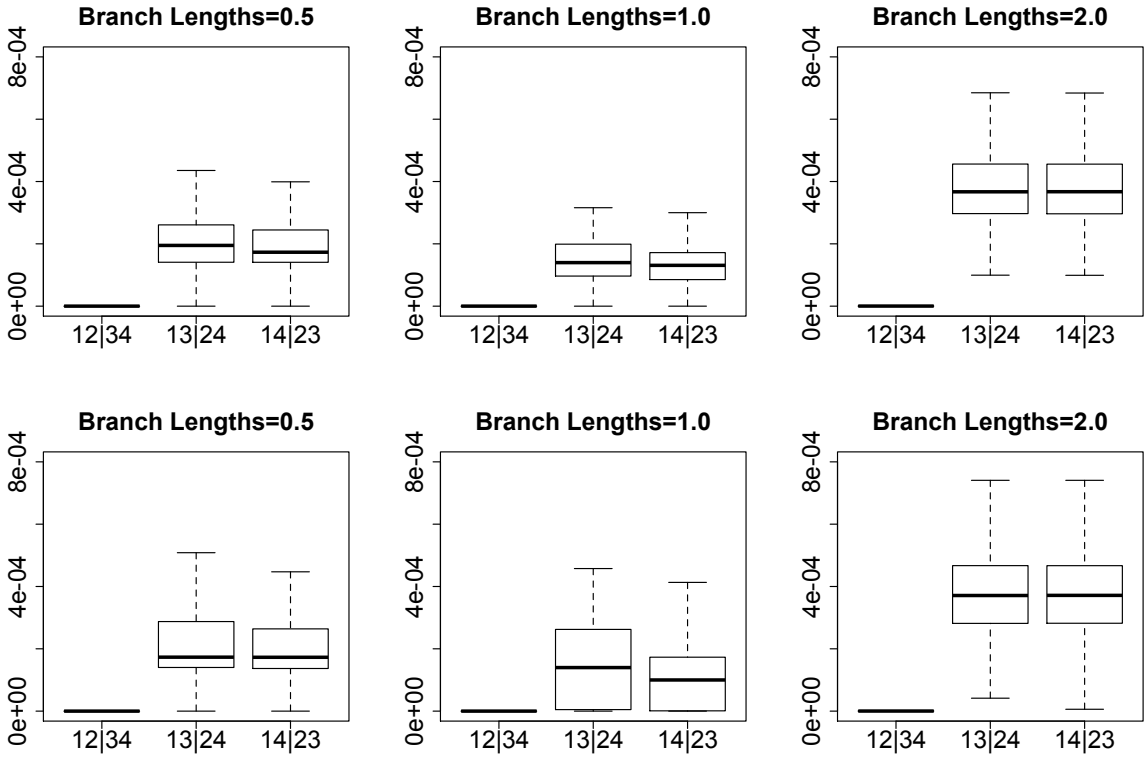
First row: 5,000 SNP sites; Second row: 10 genes of 500bp



Simulation study 1 – can we detect the correct split?

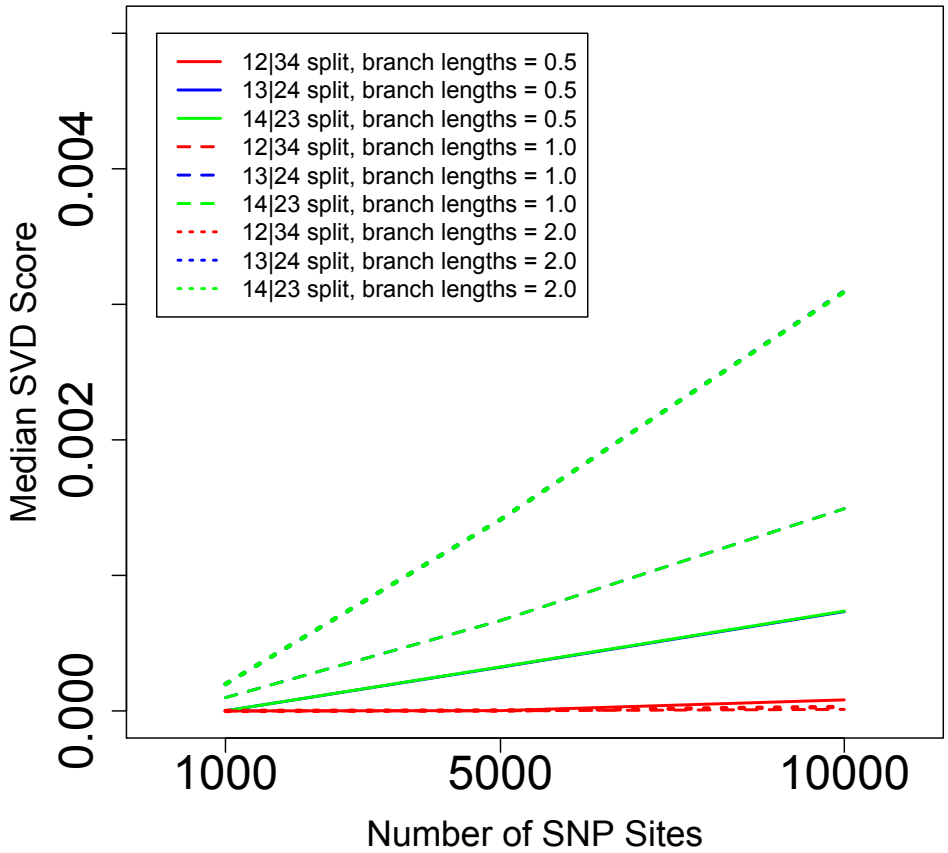
Simulate data from the GTR+I+Γ model for a 4-taxon tree and examine split scores

First row: 5,000 SNP sites; Second row: 10 genes of 500bp



Simulation study 1 – can we detect the correct split?

Change in scores as amount of data increases

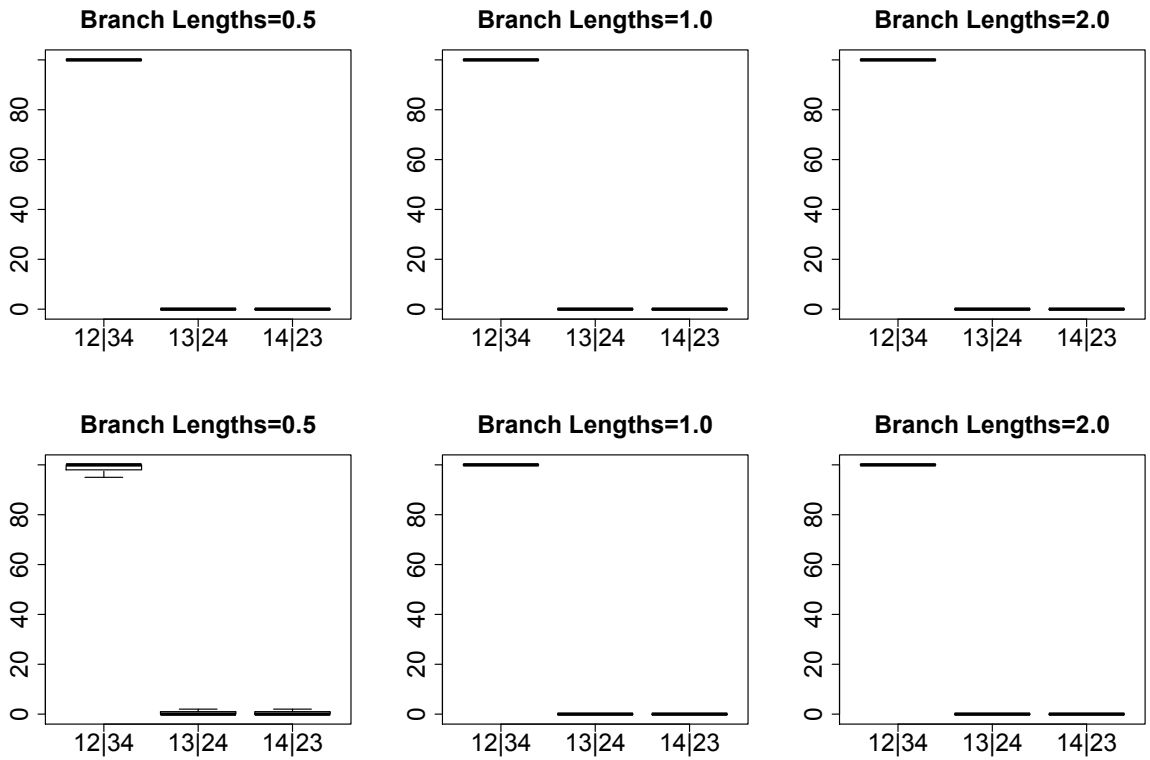


How do we assess variability?

- How can we measure confidence in the inferred split?
- Use a nonparametric bootstrap procedure
  - Generate bootstrap data sets from the original data matrix
  - Compute split scores on all three splits for each bootstrap data matrix
  - Record the number of bootstrap data sets for which each split is inferred, and use the proportion of these as a bootstrap support measure
- Evaluate performance of the bootstrap procedure using the same simulated data

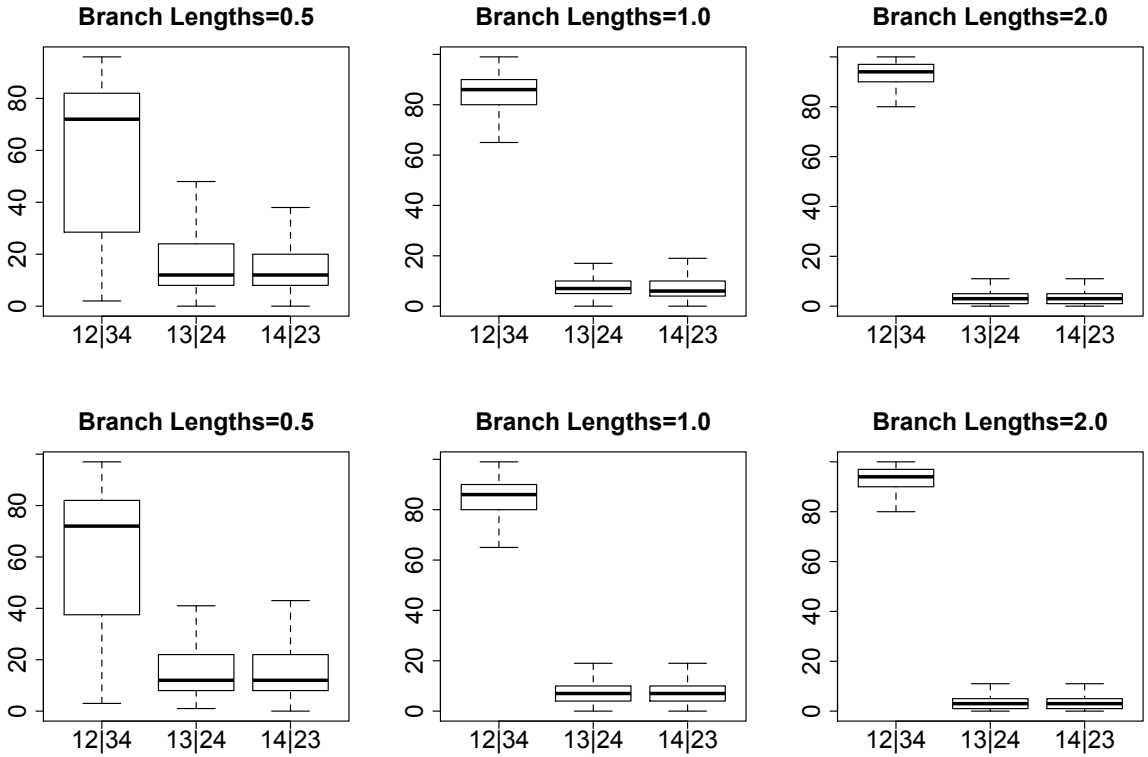
Assessing support using the bootstrap

Simulate data from the Jukes-Cantor model for a 4-taxon tree and examine bootstrap support scores



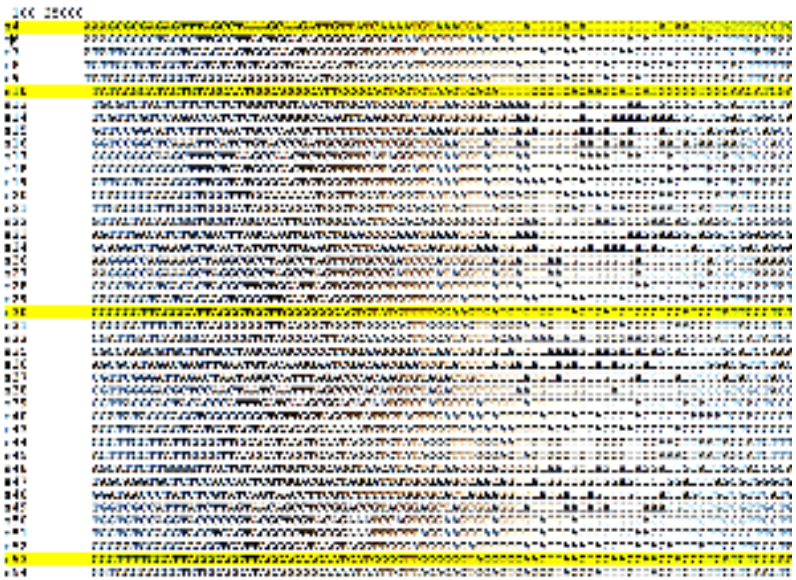
Assessing support using the bootstrap

Simulate data from the GTR+I+Γ model for a 4-taxon tree and examine bootstrap support scores



Algorithm

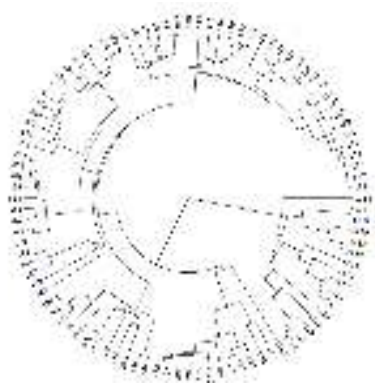
1. Generate all quartets (small problems) or sample quartets (large problems)
2. Estimate the correct quartet relationship for each sampled quartet
3. Use a quartet assembly method to build the tree - PAUP* uses the method of Reaz-Bayzid-Rahman (2014), called QFM, to build the tree.



→

1 2 | 3 4
3 5 | 2 17
19 6 | 16 1
5 22 | 3 7
. . . .

→



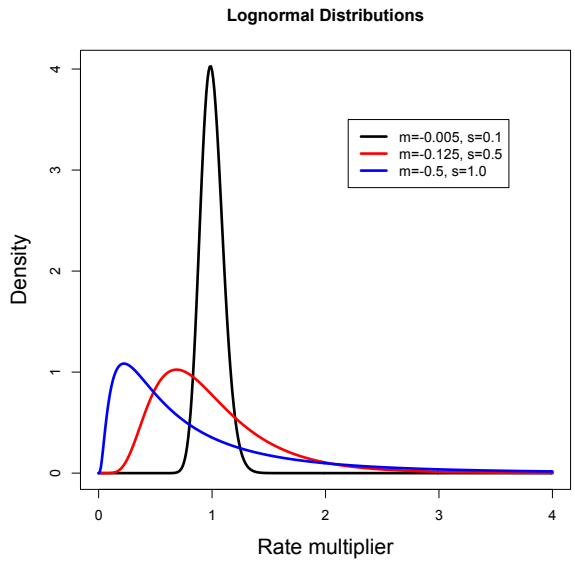
Extension to larger trees

- Multiple lineages are handled as follows:
  1. Sample four species
  2. Select one lineage at random from each species
  3. Estimate the quartet relationships among the four sampled lineages
  4. Restore the species labels (but lineage quartets are saved, too)

Simulation under more realistic scenarios

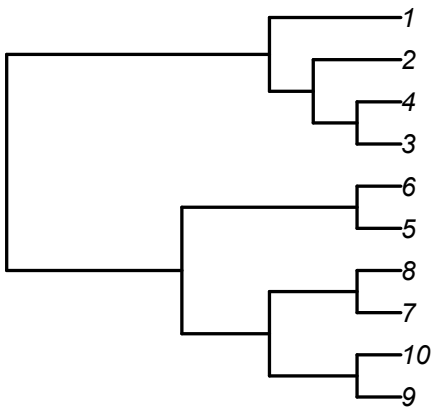
Consider the effects of:

- Larger trees: 10 species
- Multi-locus data: 10-100 genes
- Varying levels of ILS: speciation intervals of 0.5, 1.0, and 2.0
- Lineage-specific rate variation: modeled by the lognormal distribution



Simulation study 2 – larger trees with lineage-specific rate variation

Average (over 100 reps) scaled RF distance (range 0 - 1)



black = lognormal($m = -0.005, s = 0.1$)
red = lognormal($m = -0.125, s = 0.5$)
blue = lognormal($m = -0.5, s = 1.0$)

500bp per gene

	10 genes	20 genes	50 genes	100 genes
Short (0.5)	0.246 0.290 0.290	0.169 0.161 0.160	0.039 0.043 0.050	0.001 0.004 0.004
Medium (1.0)	0.117 0.107 0.099	0.024 0.027 0.001	0.001 0 0.001	0 0 0
Long (2.0)	0.016 0.017 0.011	0.001 0 0.001	0 0 0	0 0 0

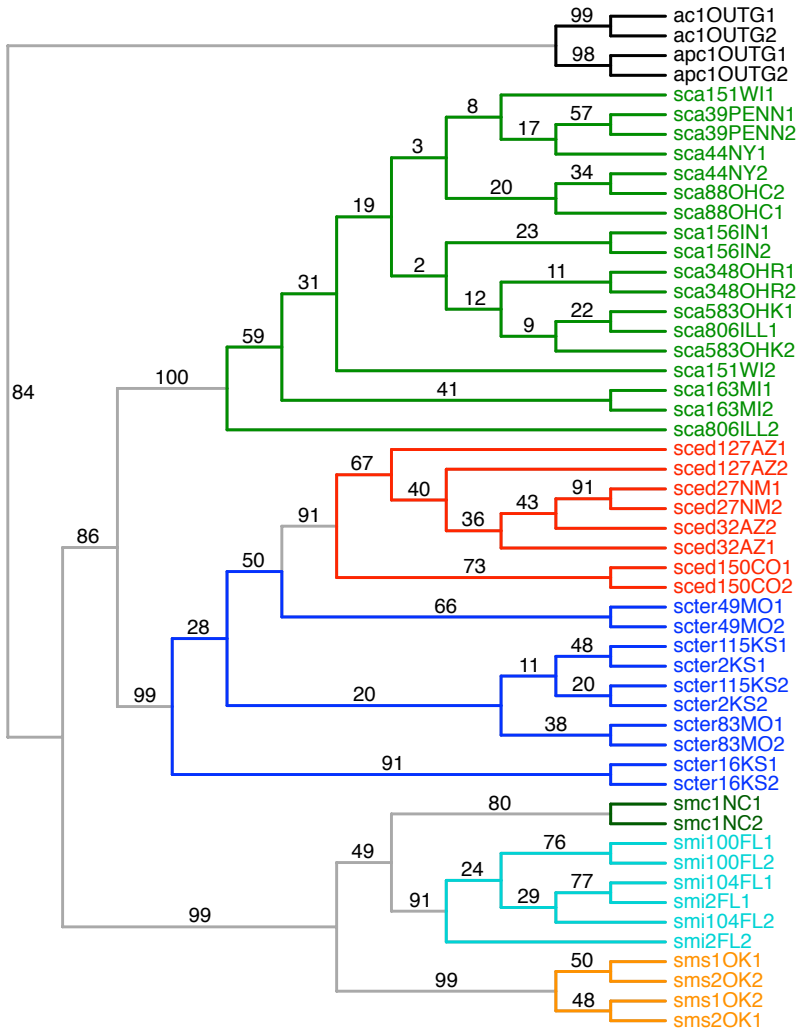
- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

Species	Location	No. of individuals per gene
S. catenatus catenatus	Eastern U.S. and Canada	9
S. c. edwardsii	Western U.S.	4
S. c. tergeminus	Western and Central U.S.	5
S. miliarius miliarius	Southeastern U.S.	1
S. m. barbouri	Southeastern U.S.	3
S. m. streckerii	Southeastern U.S.	2
Agkistrodon sp. (outgroup)	U.S.	2

Empirical example: Sistrurus rattlesnakes

All quartets and 100 bootstrap replicates

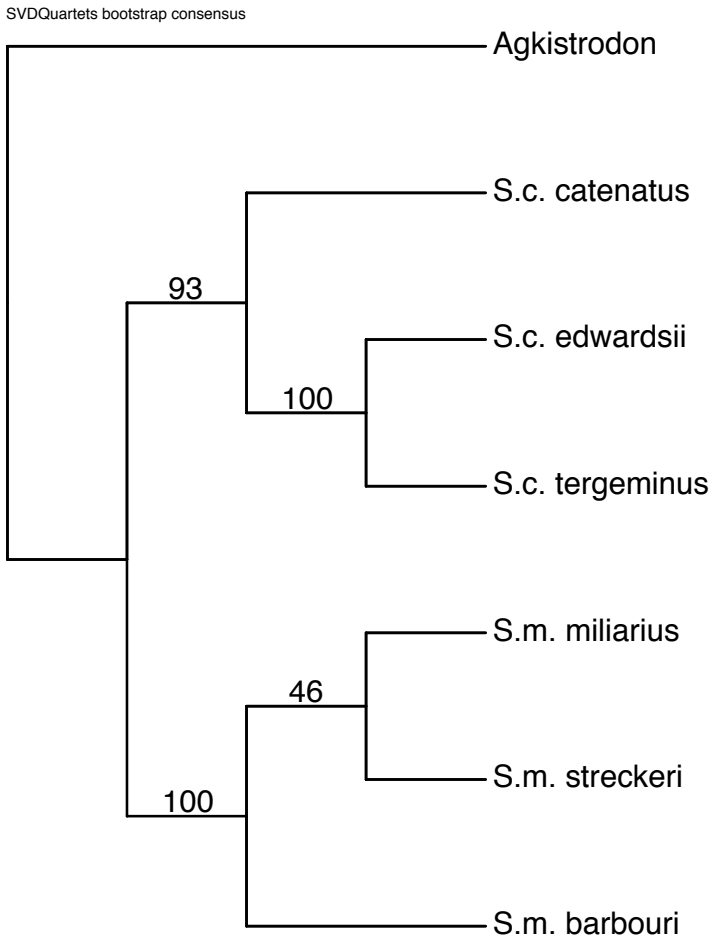
$\sim$ 11 minutes



Empirical example: Sistrurus rattlesnakes

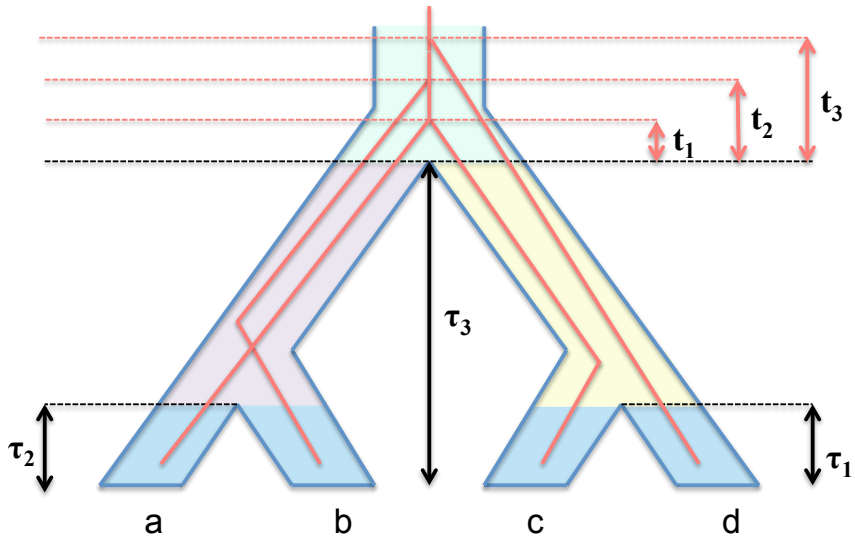
All quartets and 100 bootstrap replicates

$\sim$ 11 minutes



New features: branch length estimation

- Consider the JC69 model for the symmetric species tree with 4 taxa
- In this case, there are 9 distinct site pattern probabilities
- Chifman and Kubatko (2015) showed that these site pattern probabilities under the coalescent model could be expressed as



$$p_{i_a i_b i_c i_d|(S,\tau)} = c_0 + c_1 x_1^{2\mu} + c_2 x_2^{2\mu} + c_3 x_1^{2\mu} x_2^{2\mu} + c_4 x_3^{2\mu} + c_5 x_1^{\mu} x_3^{2\mu} + c_6 x_2^{\mu} x_3^{2\mu}$$
$$+ c_7 x_1^{\mu} x_2^{\mu} x_3^{2\mu} + c_8 x_1^{-\frac{2}{\theta}} x_2^{-\frac{2}{\theta}} x_3^{4\left(\mu+\frac{1}{\theta}\right)}$$

where $x_j = e^{-\tau_j}$ for $j = 1, 2, 3$ and the coefficients are functions of the mutation rate μ and effective population size θ.

New features: branch length estimation

- Let $\mathbf{C}_{9\times 9}$ be the matrix of coefficients. Then the above expressions for the site pattern probabilities can be written as

$$\mathbf{C}\beta = p$$

where

$$\beta' = \left(x_1^{2\mu}, x_2^{2\mu}, x_1^{2\mu}x_2^{2\mu}, x_3^{2\mu}, x_1^{\mu}x_3^{2\mu}, x_2^{\mu}x_3^{2\mu}, x_1^{\mu}x_2^{\mu}x_3^{2\mu}, (x_1x_2)^{-2/\theta}x_3^{4(\mu+1/\theta)}\right)$$

and

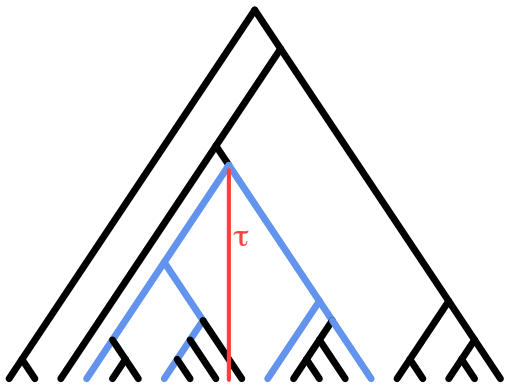
$$p' = (p_{xxxx}, p_{xxxy}, p_{xyxx}, p_{xyxy}, p_{xxyy}, p_{xxyz}, p_{yzxx}, p_{xyzx}, p_{xyzw})$$

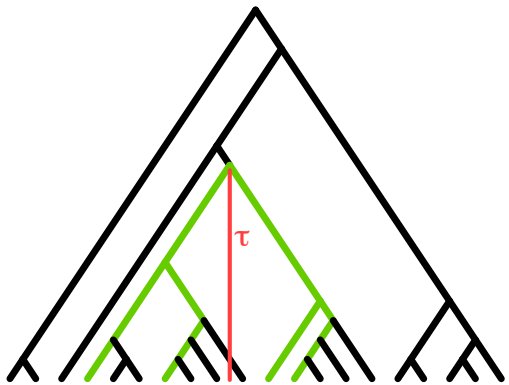
- Use this to write the likeilhood for the four-taxon case, and find maximum likelihood estimates numerically

- Asymptotic variances can be found using standard statistical theory (Fisher information matrix, etc.)

New features: branch length estimation

- For 4 taxa:
  - Test robustness and possibly use models more general than JC69
- For larger trees:
  - Combine estimates for 4 taxa





SVDquartets Summary

- Advantages:
  - Quick! And scales well to large taxon sets and next-gen sequencing data
  - Easily parallelized
  - Intuitive method for handling missing data
  - Potential for application to other data types (codons, amino acids, etc.)
- Disadvantages:
  - Estimating a matrix with 256 entries .... so may not work well with limited data

Now on to the tutorial!