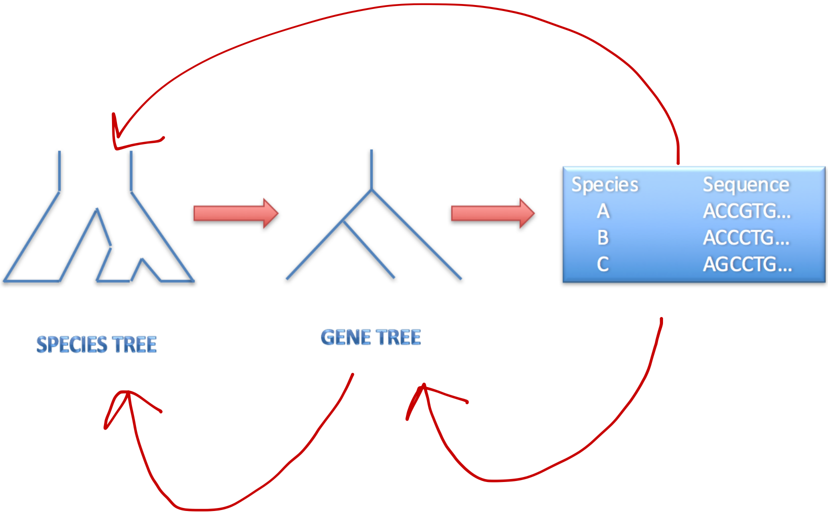
Species Tree Estimation Lab: SVDQuartets and ASTRAL

Outline

1. Species Tree Inference
2. ASTRAL Method
3. SVDQuartets Method
4. Tutorial
5. References

Species tree inference

full data methods

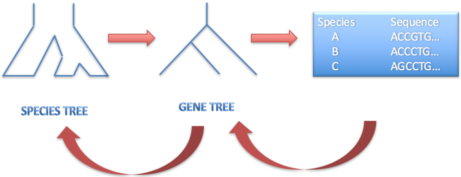


Species	Sequence
A	ACCGTG...
B	ACCCTG...
C	AGCCTG...

Summary statistic methods

ASTRAL background

- Recall our ideas about inference under the phylogenetic coalescent model

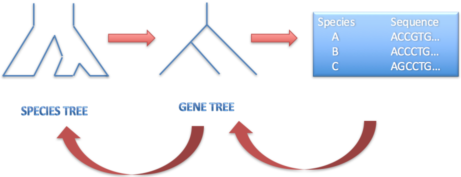


Species	Sequence
A	ACCGTG...
B	ACCCTG...
C	AGCCTG...

- **ASTRAL** is a summary statistic method for species tree estimation:
  - **Step 1.** Estimate gene trees for each locus
  - **Step 2.** Extract all quartet relationships from the estimated gene trees
  - **Step 3.** Find the species tree that “agrees” with as many quartets as possible

ASTRAL background

- Recall our ideas about inference under the phylogenetic coalescent model

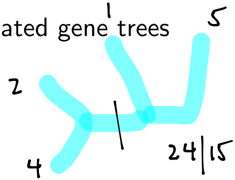


Species	Sequence
A	ACCGTG...
B	ACCCTG...
C	AGCCTG...

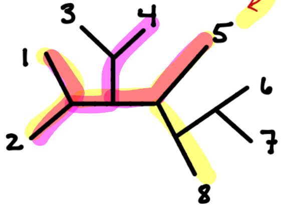
- **ASTRAL** is a summary statistic method for species tree estimation:
  - **Step 1.** Estimate gene trees for each locus ✓
  - **Step 2.** Extract all quartet relationships from the estimated gene trees
  - **Step 3.** Find the species tree that “agrees” with as many quartets as possible

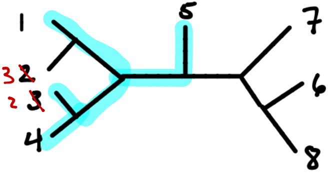
ASTRAL

- **Step 2.** Extract all quartet relationships from the estimated gene trees



Example: Input : {gene tree 1 , gene tree 2} = τ



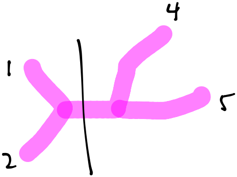


$q_{11} = 12|34$

$q_{12} = 12|58$

$q_{13} = 13|58$

⋮



$q_{21} = 13|24$

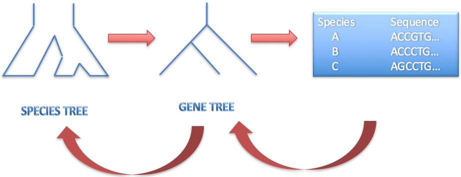
$q_{22} = 12|58$

$q_{23} = 13|58$

⋮

ASTRAL background

- Recall our ideas about inference under the phylogenetic coalescent model



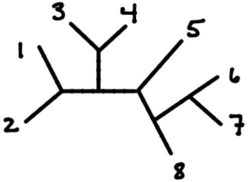
- **ASTRAL** is a summary statistic method for species tree estimation:
  - **Step 1.** Estimate gene trees for each locus ✓
  - **Step 2.** Extract all quartet relationships from the estimated gene trees ✓
  - **Step 3.** Find the species tree that "agrees" with as many quartets as possible

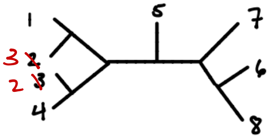
ASTRAL

- **Step 3.** Find the species tree that “agrees” with as many quartets as possible
  - This is a non-trivial problem .... recall that we expect substantial incongruence among trees
  - However, *unrooted* gene trees cannot be anomalous for four taxa in the absence of gene flow, so *if the gene trees are correct*, then this is easy
  - ASTRAL uses the Weighted Quartet Score of a candidate species tree – defined to be the number of quartets from the set of input gene trees that agree with the candidate species tree
  - Optimization problem – need to search for the species tree that maximizes the Weighted Quartet Score

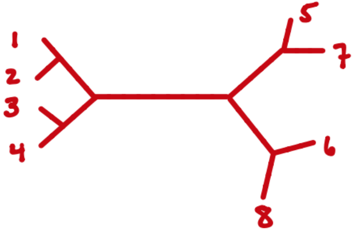
ASTRAL

- Example:





Consider the species tree T_1 =



$$\text{Score}(T_1) = \sum_{\substack{\text{quartets in} \\ \text{tree } T_1,\, q}} w(q, \tau)$$

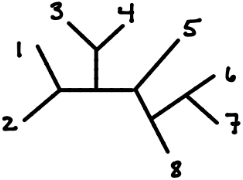
$q_1 = 12|34 \longrightarrow w(q_1, \tau) = 1$ (appears in gene tree 1)

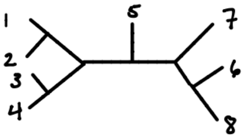
$q_2 = 12|58 \longrightarrow w(q_2, \tau) = 2$ (appears in both input gene trees)

+ all quartets in T_1

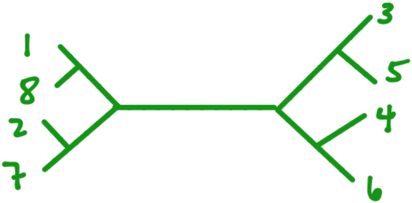
ASTRAL

- Example:





Consider the species tree T_2 =



$$\text{Score}(T_2) = \sum_{\substack{\text{quartets in} \\ \text{tree } T_2,\, q}} w(q, T)$$

$q_1 = 18|27 \longrightarrow w(q_1, T) = 0$ (doesn't appear in either input gene tree)

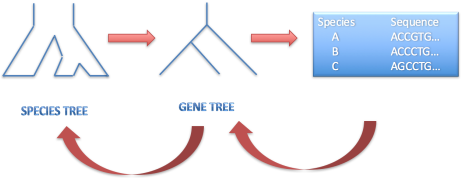
$q_2 = 12|34 \longrightarrow w(q_2, T) = 1$ (appears in gene tree 1)

$q_3 = 18|23 \longrightarrow w(q_3, T) = 0$ (doesn't appear in either input gene tree)

+ all quartets in T_2

ASTRAL background

- Recall our ideas about inference under the phylogenetic coalescent model



Species	Sequence
A	ACCGTG...
B	ACCCTG...
C	AGCCTG...

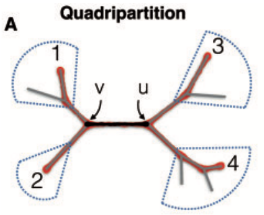
SPECIES TREE

GENE TREE

- **ASTRAL** is a summary statistic method for species tree estimation:
  - **Step 1.** Estimate gene trees for each locus ✓
  - **Step 2.** Extract all quartet relationships from the estimated gene trees ✓
  - **Step 3.** Find the species tree that “agrees” with as many quartets as possible ✓

Additional features of ASTRAL

- ASTRAL can also estimate branch lengths (in coalescent units)
- ASTRAL also provides a measure of uncertainty: *local posterior probability*



Sayyari and Mirarab, 2016

- Assume that the “clusters” on each edge of the branch under consideration are correct
- Use the gene trees to obtain quartet frequencies for the three possible arrangements of clusters
- Assume a prior distribution on the quartet trees (Yule prior with parameter λ)
- Compute the posterior probability that this branch appears in the true species tree, given the observed quartet frequencies

ASTRAL performance

- ASTRAL is statistically consistent *when the gene trees are known without error*
- ASTRAL will perform well when the gene trees can be estimated well
- Computational efficiency: the estimation of gene trees is the time-consuming step, but can be parallelized
- Crucial assumption: true unrooted quartets have higher probability than other quartet relationships
- Assessment of uncertainty: use the local posterior probability (now recommended over the bootstrap)

SVDQuartets (or just SVDQ)

Goal of this work:

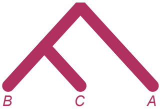
Develop a full data approach that is computationally feasible for large-scale data

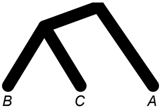
How?

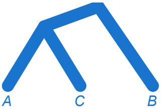
- Summarize data differently, so that model requires less computation
- Develop theory to infer relationships among quartets of taxa very accurately
- Use a quartet assembly method to build a large tree

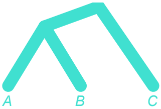
Recall the phylogenetic coalescent model with mutation

Example: Want to compute the probability that taxon A has nucleotide T, taxon B has nucleotide G and taxon C has nucleotide T – call this p_{TGT}









$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
$p^{1a}_{TGT} = 0.05$	$p^{1b}_{TGT} = 0.025$	$p^{2}_{TGT} = 0.2$	$p^{3}_{TGT} = 0.025$

$$p_{TGT} = 0.63 \times 0.05 + 0.12 \times 0.025 + 0.12 \times 0.2 + 0.12 \times 0.025 = 0.0615$$

$\Uparrow$ *For intuition only, not completely correct ...*

But .... there are a lot of histories!

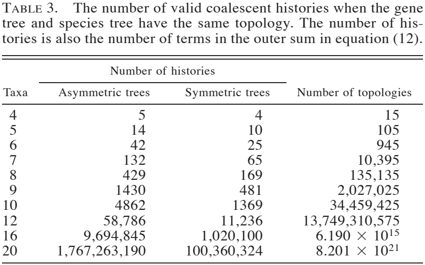
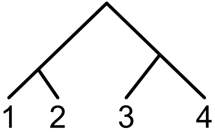


TABLE 3. The number of valid coalescent histories when the gene tree and species tree have the same topology. The number of histories is also the number of terms in the outer sum in equation (12).

	Number of histories		
Taxa	Asymmetric trees	Symmetric trees	Number of topologies
4	5	4	15
5	14	10	105
6	42	25	945
7	132	65	10,395
8	429	169	135,135
9	1430	481	2,027,025
10	4862	1369	34,459,425
12	58,786	11,236	13,749,310,575
16	9,694,845	1,020,100	6.190×10^{15}
20	1,767,263,190	100,360,324	8.201×10^{21}

- This means that calculating the likelihood – and thus using likelihood-based methods for inference – will be difficult, especially for large-scale data
- Alternative approach: compute explicitly (i.e., write formulas for) the site pattern probabilities for 4-taxon trees, and look for “structure”

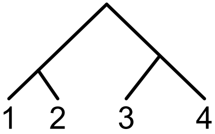
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

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

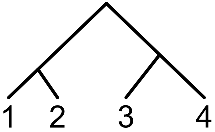
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

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

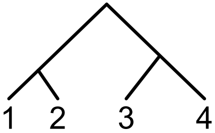
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

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

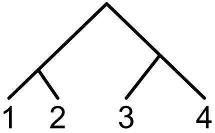
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

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

Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

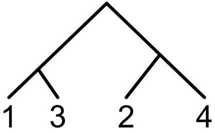
$$Flat_{12|34}(P) = \left(\begin{array}{ccccccc} & [AA] & \mathbf{[AC]} & [AG] & [AT] & \mathbf{[CA]} & \cdots \\ [AA] & \mathbf{5} & \mathbf{p_{AAAC}} & p_{AAAG} & p_{AAAT} & \mathbf{p_{AACA}} & \cdots \\ [AC] & p_{ACAA} & \mathbf{p_{ACAC}} & p_{ACAG} & p_{ACAT} & \mathbf{p_{ACCA}} & \cdots \\ [AG]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

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

Main Result:

- 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 → gene tree ::: coalescent process
    - gene tree → data ::: nucleotide substitution models: GTR+I+Γ and submodels
- **This result still holds** when the species tree violates the molecular clock and/or when there is variation in effective population size across the branches and/or when there is gene flow between sister taxa

What about the incorrect tree?



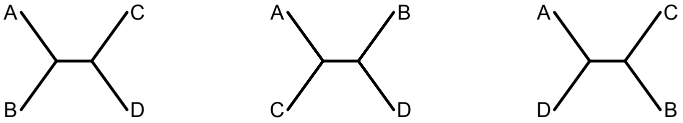
Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

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

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

How can we use these facts to estimate the species tree?

- **Basic idea:**
  - Data: aligned DNA sequences for multiple loci or for a collection of SNPs
  - Estimate the flattening matrix for each of the following trees:

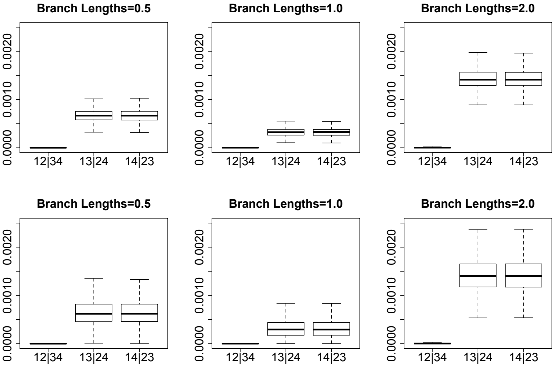


  - Compute a measure of how close each of the three observed flattening matrices is to a matrix with rank 10 – we use the SVDScore
  - Pick the tree relationship that gives the smallest SVDScore

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

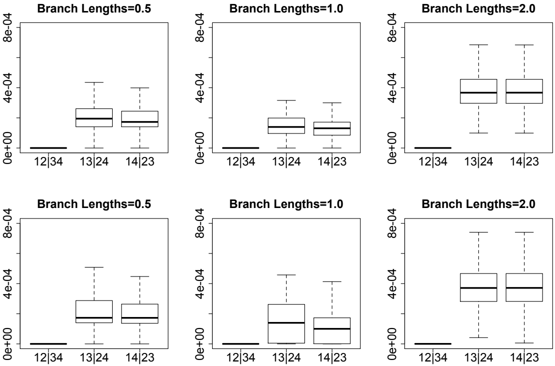
First row: 5,000 SNP sites; Second row: 10 genes of 500bp



Simulation study 1 – can we detect the correct split?

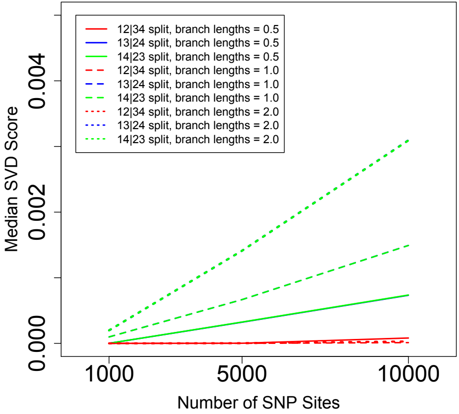
Simulate data from the GTR+I+Γ model for a 4-taxon tree and examine split scores

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?

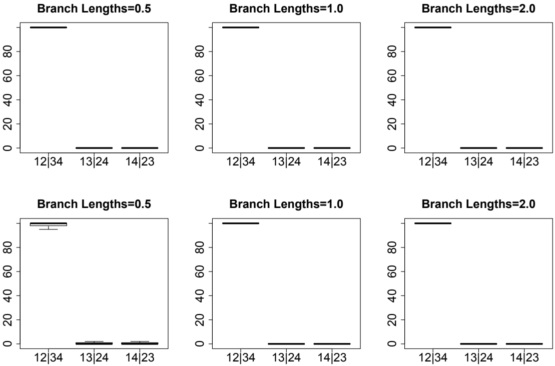
- How can we measure confidence in the inferred split?

- 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

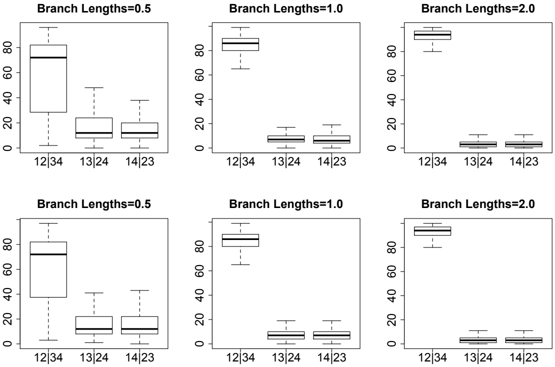
Assessing support using the bootstrap

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



Assessing support using the bootstrap

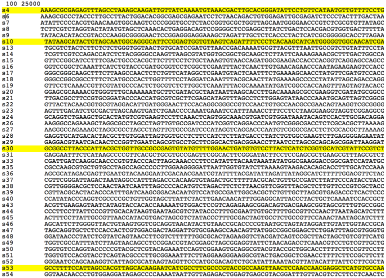
Simulate data from the GTR+I+Γ model for a 4-taxon tree and examine bootstrap support scores



Extension to larger trees

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.



```
 100  25000
s4         AAAGCGCGAGAGTTTAGCCTAAAGCAAGATTGTTATCAAAATGTAAACGACTTGCACGGGATATTCCTGTTCATAATGTTGTTTTCCTG
s6         AAAGCGCCCTACCCTTGCCTTACTGGACACGGCGAGCGAGAATCCTCTAACAGACTGTGGAGATTGCGAGATCTCCCTACTTTGACTAA
s7         ATATTCCCCACGTGAACAAGTGGCAAGTCCCCGGGTGCCTCTACCGGTGCGCTGGTTAGCAATGGGGAACCCGTCTCGCGTGTTATAGC
s8         TCTTCGTCTAGGTACTATATGTAGCTCAAACACTGAGGACAGTCCGGGGCTCCGAGGCTTATCTGACTCCACGACGTGGTGACTTTTTA
s9         TATACACATACCGTACCCAAGGCGGGGAACTCCGGAGGGAGGAAAGCTAGATTTGCTCTCACCCTACTCATCGCGGGGGCACCTTAGAA
s10         TATAAGCATACTGTAGCAATGGCAGGGCATTCCGGAGTGGTGTAAGTCAGATTTTCCCTCACAAGCATCATGGGGGTGGCAACATCGA
s13         TGCGTCTACTCTTCTCTCTGGGTGGTAACTGTGCATCCGATGTTAACCGACAAAATGCGTTACGTTTAATTGCCGTCTGTATTTCCTA
s14         TCGTTCGTCCAGACCATCTCTACGGGGCCAAGTTAAGCGTATGGTGCGGCTGCAAGCTCTTATTCAAAAGAAACGCTTTGACTGGCCA
s15         GTTCCGGCATCCTTTCGACTGCCCACGCTTCTTTCCTCTGCTAAAGTGTAACCAGATGGCGAAGACCACCCACGGTCAGGAGCCAGCC
s16         GGTCCGGCTCAAATTCATCTGTACCCGATGCGGTCCTCGGCTCGAATTTACTCAACTTGTGAACAGAACCCAATAGAAGAAGTCCACC
s17         GGGCGGGCAAGCTTTCTTAAGCCAACTCTCTGTTCATAGGCTCGAATGTACTCAATTTCTGAAAAGCAACCTATGTTCGATGTCCACC
s18         GGGCCGGCGGCCTTTCTCATGCCCACTTGGTTTTCCTAGGATCAAATTTACGCAACTTATGAAAAGCCCCCTATATTAGCAGACCAGG
s19         GTTCCTCAAACCGTTCGTATGCCCACTCTGCGTCTCTTGGCTCAAATTTACGCAAAATATGATCGGCCACCAAGTTTAGTAGTACACC
s20         GGAGCGCAAAACGTGGGTTTGCAAAAAATGCTTCGATACGCTAATAATTAAGCACCTTGACAAAAGCGCCGAAGGTCGATATGCGGCC
s21         TTGACGCCTTCCGTGCGGATACCCACCATGCGTCGCTTCGAGCATAGTTTAGCCTCTTGGAGTATAACCCAAATGCAGTTAGTCGGCG
s22         GTTACTACAACGGTGCGTAGGACATTGATGGGAACTTCCACAGGCGGGCCGTCCAACTGTGCCAACGCCGAACAGTAAGGTCGCGGGG
s23         AGTTTGACATCTGCGACTTAGCAAGTGATCTGAACCCCAAATGAAATCCATTTAAACCTTTCCTCCTAAGGAAGGTAGGTCGGAGGCG
s24         GCAGGTCTGAAGCTGCACTATGTCGTGAAGTCCTTCAAACTCAGTGGCAAACGTGATCGTGGATAAATACAGTCCACTTGCGACAGGA
s26         AAGGGCCAGAAAGCTAGGCGCCTAGCCTAGTTGTGCCGATCCCATCGGTCAGGACCCAATGTCGGGCGACCTCTCGCACGCTTAAAAG
s27         AAGGGCCAGAAAGCTAGGCGCCTAGCCTAGTTGTGCCGATCCCATAGGTCAGGACCCAATGTCGGGCGACCTCTCGCACGCTTAAAAG
s28         GAGGTGCATGACACTACGCTTGTGGATTAGTGGTGCCTCTTCCATCTCTTATGACCCTATCTGTGGCGAAGTCTTGAGGGGAGAATAT
s29         GAGGACGTAATCACAACTCCGGTTGAATCAGCGTGCCTGCTTCGGCACTGACGGAACCTTAATCCTAGCGCAGCAGCGGATTAAGGAT
s30         GCGGCCTTACCCATTACGCTGGTTGCCGCCGAGTGTATGTTTTGGAACTGATGTGTCCTTACTCATCTCGGTGCATCGTATTCCGTCT
s31         GAGGAATTTCTCATAAGCCCCGTTCACGCTGCGTGCCGAGTTCGGCACTTCGGGATACTTCCTCCGAGCGCTGAAGCGTTTAGCGTCT
s33         CGATTGATCAAGGCAACCCGTGTACCCACTTAGCAAACCCTTCCATATTTACAATAAATATATGGCGAAGGACCGGCGATCCATATGC
s35         CGCCAAGCGTGCTGTGCCTAGCCAGCCCCCTTGGAAGGGGATTAATTGGTGTATCTTAAAAGTAAGACACAGCACCTTTGCATCCAAA
s36         AGCGCATAGACGAGTTGAATGTACAAGGAATCGACAACGAATCGTATTTTACGATATTAGATTATAGGCATCTTTTTGGACGTTCTAC
s37         CGTTCGGGATTAGACTAATAGGCCATTTAGACCCACAAGTGTAGACTGCTTTTTACGCACTGTTGCCGATCTTATTCCCATACCTACC
s38         CGTTGGGGACGCTCCAACTAATCAATTTAGCCCCACATGTCTAGACTTCTCTTAACGTTCGGTTTACGCTCTTGTTCCCGTCGCTACC
s39         CGTTACGCACTACACCCATTTGATCAAGCGGACACAATGTCCATGCCGATTGGCACGCTCTGTTGCTTAGCGTGAGACCCTCACTCCC
s40         CCATATACCCAGGTCGCCCCGCTGTTGGTAGCTTATTCTACTTGAACAACATTTGGAGGCATTACCCTACTGCTCAAAATCCACCGCT
s43         ACGTTGAAGAGTAATCATAGTACCACACCAAAAATGTGGACTCCGTCAGAGAAGCGGACAGTGACGAAGCGGTAGCGTTTGTGCCGGC
s44         CCCTTCCTTATTCCCGTTGCACATAAGTGAATAGCGTCTATACCCTTTGCGAATTGTGCTTTTGCCGTTCCAACTTAATAGCATCTTT
s45         ACTTTCCTTATTGCCGTTAACCATACGTGACTAGCGTCTATACCCTTTGCGACTAGTGCCTCTGCCGTTCCAAACTAATAGCATCTTT
s46         AGCATTCTTGGGGTTACTGTAAGTGGTCGCAGGTAACATACTGTAACCGCACCCCGTAATGCTAGTATAGGATTGTATTGGTACCGCA
s47         TAGCAGGTGCTCTCCACCACTCGTGGACGACTAGGATTGTGCGAAGCCAACAGTTATGGCCGGCGGAGCTCTGGATTTAGCGTGGGTG
s48         GGATAACCCTATCTCGTATCAGTAACCTTCGTCTTTTTTGGCGCACTAGAAGGTCTACGATCAGTCCGCTTACTAGCTGTCGTTCATG
s49         TGGTCGCCATTTATCTTAGTAACAGCCTCCGGACCGTACGGAGCTAGCGGAGGTATTCTAACAGACCTCAAACGTCCTGTCGTTGCTG
s50         TGGTGTCCAGGTACCCCCGTACGCTCGTACAGAAATCCAGTCGGAAGTGCAGGCGTACGTACCGCTGGCACCTTTTCTTCCGTTCCTT
s51         TGGTGTCCACGTACCTCAGTACGCCCTTGCGGAAATTTCTTAGGAAGGGAAGGCGTACTGACGGCTCGAACCTTTTCTTCCGCTTCTT
s52         GGGAATCCAGCAAAGGTCATTAGCGCATAAGTAGGTTCCCATGCGACAGTCTGCATATTAAATACGTTATACATTTATTCCTTCGCTA
s53         GCCTTTTCCATTAGCCACGTTAGCACAAGAATCATCGCCTTCGCCCGTACCGCCAAGTTAACTCCAACCAACGAGGCTCATGTGCCAA
s54         GGTAACAACCCTGTGGAGGATAGAGCCCCAAAATAATTGTTAGAGACTGGAGTGAGCGTACGGATTTGTTACGTCTCTTCCTCAACTC
```

→

1 2 | 3 4

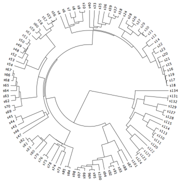
3 5 | 2 17

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)
- Quantify uncertainty using the bootstrap

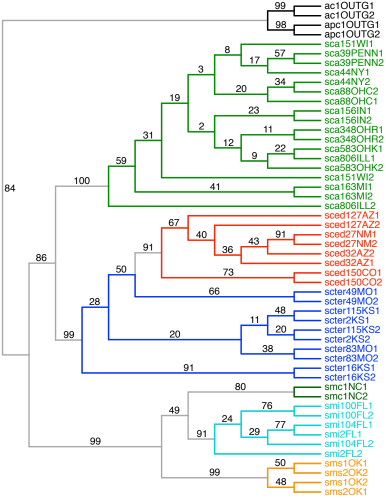
- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

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

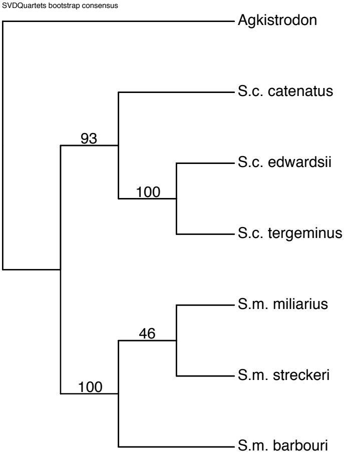
$\sim$ 11 minutes



Empirical example: Sistrurus rattlesnakes

All quartets and 100 bootstrap replicates

$\sim$ 11 minutes



Comparison between methods

- SVDQuartets
  - Statistically consistent for estimating the quartet trees
  - Will perform well when there are a lot of data (multilocus or SNP) available
  - More complex model $\Longrightarrow$ more data needed
  - Valid when the molecular clock is violated
  - Valid when there is gene flow between sister taxa
  - Computationally efficient, including bootstrapping
  - Will soon include estimates of branch lengths
- ASTRAL
  - Statistically consistent when gene trees are known without error
  - Will perform well when gene trees can be estimated well
  - Gene flow can cause the method to fail (because then quartets can be anomalous)
  - Computationally efficient after individual gene trees have been estimated
  - Can provide estimates of branch lengths
  - Local posterior probabilities used to quantify uncertainty in the data

Comparison between methods

- How do these compare to Bayesian methods, such as STARBEAST2 and BPP?
  - STARBEAST2 and BPP carry out estimation under the model, including all model components
  - Estimation of the posterior distribution provides a natural way to quantify uncertainty
  - ASTRAL and SVDQ use features of the model to assess fit of the data to the model
    - ASTRAL: gene trees
    - SVDQ: site pattern probabilities
  - Trade-offs involved in choosing among methods: computational efficiency, robustness to the model, etc.

Tutorial

Now on to the tutorial!

The tutorial is available at `http://phylosolutions.com/tutorials/wh2019-svdq-astral/species-trees-tutorial.html`

References

- ASTRAL

  - Zhang, Chao, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. 2018. "ASTRAL-III: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees." BMC Bioinformatics 19 (S6): 153. doi:10.1186/s12859-018-2129-y.
  - Sayyari, Erfan, and Siavash Mirarab. 2016. "Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies." Molecular Biology and Evolution 33 (7): 1654-68. doi:10.1093/molbev/msw079.
  - Mirarab, Siavash, Rezwana Reaz, Md. Shamsuzzoha Bayzid, Theo Zimmermann, M. S. Swenson, and Tandy Warnow. 2014. "ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation." Bioinformatics 30 (17): i541-48. doi:10.1093/bioinformatics/btu462.
  - Mirarab, Siavash, and Tandy Warnow. 2015. "ASTRAL-II: Coalescent-Based Species Tree Estimation with Many Hundreds of Taxa and Thousands of Genes." Bioinformatics 31 (12): i44-52. doi:10.1093/bioinformatics/btv234.

References

- SVDQuartets

  - Chifman, J. and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model, Bioinformatics 30(23): 3317-3324.
  - Chifman, J. and L. Kubatko. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites, Journal of Theoretical Biology 374: 35-47.
  - Swofford, D. PAUP* (* Phylogenetic Analysis Using PAUP), Version 4.0a165, available at `https://paup.phylosolutions.com`.