Invariants methods for phylogenetic inference

Phylogenetic invariants work by examining patterns in the data in order to determine whether they show relationships that are predicted by particular tree topologies.

Typically, invariants are linear or polynomial expressions that evaluate to specific expected values for each possible tree topology. By tabulating site pattern frequencies, we can calculate the values of the invariants and ask whether they approximate these expected values. This information allows selection of a preferred tree, and rejection of the other possibilities.
Early work on phylogenetic invariants

Cavender and Felsenstein (1987)
Quadratic invariants for 2-state data

Lake (1987)
Linear invariants for 4-state nucleotide data
Much subsequent work by mathematicians...
Much subsequent work by mathematicians

$$\mathbb{E} \langle Y_1, \theta \rangle \mathbb{E} \langle Y_2, \phi \rangle = \mathbb{E} \langle Y_1, \phi \rangle \mathbb{E} \langle Y_2, \theta \rangle.$$ 

Before closing this discussion of the invariants of the Kimura two-parameter model for the two-leaf tree, it is both of independent interest and convenient for later examples to relate our notation and results to those of Cavender (1989, 1991). Following Cavender, we let $A$ (resp. $G, C, T$) double as the function on $\mathbb{G} = \{A, G, C, T\}$ which takes the value 1 on $A$ (resp., $G, C, T$) and 0 elsewhere. We then see that $A - G = (1/2)(\phi + \theta)$ and $C - T = (1/2)(\phi - \theta)$; and, letting $\otimes$ denote the tensor product of functions on $\mathbb{G}$, we see that

\[
(A - G) \otimes (C - T) = \frac{1}{4}(\phi + \theta) \otimes (\phi - \theta)
\]

\[
= \frac{1}{4}(\phi \otimes \phi - \phi \otimes \theta + \theta \otimes \phi - \theta \otimes \theta)
\]

and

\[
(C - T) \otimes (A - G) = \frac{1}{4}(\phi - \theta) \otimes (\phi + \theta)
\]

\[
= \frac{1}{4}(\phi \otimes \phi - \theta \otimes \phi + \phi \otimes \theta - \theta \otimes \theta).
\]

Despite providing a great deal of recreational opportunities for mathematicians/algebraic statisticians, these methods were almost completely ignored by empirical evolutionary biologists!
Lake’s linear invariants
(“Evolutionary Parsimony”)

For each set of 4 sequences (quartet):

For the tree ((1,2),(3,4)), tabulate

\[ E = f_{AA|CC} + f_{AA|TT} + f_{CC|AA} + f_{CC|GG} + f_{GG|CC} + f_{GG|TT} + f_{TT|AA} + f_{TT|GG} \]
\[ u = f_{AG|CT} + f_{AG|TC} + f_{CT|AG} + f_{CT|GA} + f_{GA|CT} + f_{GA|TC} + f_{TC|AG} + f_{TC|GA} \]
\[ H = f_{AG|CC} + f_{AG|TT} + f_{CT|AA} + f_{CT|GG} + f_{GA|CC} + f_{GA|TT} + f_{TC|AA} + f_{TC|GG} \]
\[ J = f_{AA|CT} + f_{AA|TC} + f_{CC|AG} + f_{CC|GA} + f_{GG|CT} + f_{GG|TC} + f_{TT|AG} + f_{TT|GA} \]

Then:

\[ X = E + u - H - J \]

Calculate similar terms \( Y \) and \( Z \) for the trees ((1,3),(2,4)) and ((1,4),(2,3)).

\[ Y = F + v - L - N \]
\[ Z = G + w - Q - S \]

(patterns not involving 2 pyrimidines and 2 purines are ignored)

The true tree is expected to take a positive value for one of \( X \), \( Y \), and \( Z \); the other two are expected to be zero (i.e., the invariants).
Inferring evolutionary trees using matrix rank and the Singular Value Decomposition (SVD)

- Eriksson (2005)
- Fernández-Sánchez and Casanellas (2016)

Exploit the fact that there are linear dependencies in site-pattern frequencies that are tree-topology specific (the “invariants”)

Site pattern frequencies

A tree for 4 taxa, which may be a subtree of a larger tree
Site pattern frequencies

\[
\begin{array}{c}
\text{A} \\
\text{G} \\
\text{A} \\
\end{array}
\quad \begin{array}{c}
\text{G} \\
\text{G} \\
\end{array}
\]
Site pattern frequencies

```
      C
     /|
    /  
   C   G
  /    |
 A   A   G
   /    |
  A    C   G
```
Site pattern frequencies

A

C A
C C
A A
A
G
C
G
G

.
Allman-Rhodes-Erikkson invariants from Singular Value Decomposition

For each set of 4 sequences (quartet), we can count the relative frequencies of the 256 possible site patterns

<table>
<thead>
<tr>
<th>$p_{ijkl}$</th>
<th>Taxon A</th>
<th>Taxon B</th>
<th>Taxon C</th>
<th>Taxon D</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>$p_{AAAA}$</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>$p_{AAAC}$</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>$p_{AAAG}$</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>$p_{AAAT}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>129</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>$p_{GGGA}$</td>
</tr>
<tr>
<td>130</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>$p_{GGGC}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>255</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>$p_{TTTG}$</td>
</tr>
<tr>
<td>256</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>$p_{TTTT}$</td>
</tr>
</tbody>
</table>
For each set of 4 sequences (quartet):

Represent the pattern frequencies by three "flattening matrices" (one for each resolution of the quartet):

\[
\text{Flat}_{\{1,3\},\{2,4\}}(P) = \begin{pmatrix}
AA & AC & AG & AT & CA & CC & \ldots \\
AA & p_{\text{AAA}} & p_{\text{AAC}} & p_{\text{AAG}} & p_{\text{AAT}} & p_{\text{ACA}} & p_{\text{CAC}} & \ldots \\
AC & p_{\text{ACA}} & p_{\text{ACC}} & p_{\text{AAG}} & p_{\text{ACT}} & p_{\text{ACA}} & p_{\text{ACCC}} & \ldots \\
AG & p_{\text{AGA}} & p_{\text{AGC}} & p_{\text{AAG}} & p_{\text{AGT}} & p_{\text{ACG}} & p_{\text{ACGC}} & \ldots \\
AT & p_{\text{ATA}} & p_{\text{ATC}} & p_{\text{ATG}} & p_{\text{ATT}} & p_{\text{ACT}} & p_{\text{ACTC}} & \ldots \\
CA & p_{\text{CAA}} & p_{\text{CAC}} & p_{\text{CAAG}} & p_{\text{CAAT}} & p_{\text{CCAA}} & p_{\text{CCAC}} & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ldots \\
\end{pmatrix}
\]

Allman-Rhodes and Eriksson main result:

Under very general Markov assumptions, the flattening matrices are full rank (16) for the two incorrect trees, but the rank of the matrix corresponding to the true is tree expected to be 4.

("rank" = number of linearly independent rows and columns)
Intuition on reduced rank/linear dependencies

E.g., all 4 of these site patterns have the same expected frequency

\[ f(AC|AC) = f(AC|CA) = f(CA|AC) = f(CA|CA) \]

These patterns are **not** all expected to have the same expected frequency *if they evolved on the other tree*

\[ f(AA|CC) \neq f(AC|CA) \neq f(CA|AC) \neq f(CC|AA) \]
Flattenings for a 2-State Jukes-Cantor model

\[
\begin{array}{c}
\text{A} \\
\text{A} \\
\text{A} \\
\text{A} \\
\text{G} \\
\text{G} \\
\text{G} \\
\text{G} \\
\text{A} \\
\text{A} \\
\text{A} \\
\text{A} \\
\text{G} \\
\text{G} \\
\text{A} \\
\text{G} \\
\text{G} \\
\text{G} \\
\text{A} \\
\text{G} \\
\text{A} \\
\text{G} \\
\text{G} \end{array}
\]

\[
= \begin{array}{c}
\text{G} \\
\text{G} \\
\text{G} \\
\text{G} \\
\text{A} \\
\text{A} \\
\text{G} \\
\text{G} \\
\text{A} \\
\text{G} \\
\text{A} \\
\text{G} \\
\text{G} \\
\text{A} \\
\text{G} \\
\text{A} \\
\text{G} \\
\text{G} \\
\text{A} \\
\text{G} \\
\text{A} \\
\text{G} \\
\text{G} \end{array}
\]

Flattening matrix for 1,2 | 3,4

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td></td>
<td></td>
<td></td>
<td>a</td>
</tr>
</tbody>
</table>
Flattenings for a 2-State Jukes-Cantor model

Flattening matrix for 1,2 | 3,4

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>AG</td>
<td></td>
<td>b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td></td>
<td>b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td></td>
<td></td>
<td></td>
<td>a</td>
</tr>
</tbody>
</table>
Flattenings for a 2-State Jukes-Cantor model

\[
\begin{align*}
\text{Flattening matrix for } 1,2 \mid 3,4 \\
\begin{array}{|c|cccc|}
\hline
& AA & AG & GA & GG \\
\hline
AA & a & b & b & \\
AG & b & & b & \\
GA & b & & b & \\
GG & b & b & & a \\
\hline
\end{array}
\end{align*}
\]
Flattenings for a 2-State Jukes-Cantor model

Flattening matrix for 1,2 | 3,4

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>AG</td>
<td>b</td>
<td></td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td>b</td>
<td></td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td>c</td>
<td>b</td>
<td>b</td>
<td>a</td>
</tr>
</tbody>
</table>
Flattenings for a 2-State Jukes-Cantor model

Flattening matrix for 1,2 | 3,4

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>AG</td>
<td>b</td>
<td>d</td>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td>GA</td>
<td>b</td>
<td>d</td>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td>GG</td>
<td>c</td>
<td>b</td>
<td>b</td>
<td>a</td>
</tr>
</tbody>
</table>
Some numbers

```
 2
3
 0.1
0.1

“True” branch lengths in expected substitutions/site

```

```
Expected flattening matrix for 1,2|3,4

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0.093008</td>
<td>0.061355</td>
<td>0.061355</td>
<td>0.068115</td>
</tr>
<tr>
<td>AG</td>
<td>0.061355</td>
<td>0.046728</td>
<td>0.046728</td>
<td>0.061355</td>
</tr>
<tr>
<td>GA</td>
<td>0.061355</td>
<td>0.046728</td>
<td>0.046728</td>
<td>0.061355</td>
</tr>
<tr>
<td>GG</td>
<td>0.068115</td>
<td>0.061355</td>
<td>0.061355</td>
<td>0.093008</td>
</tr>
</tbody>
</table>
```

Expected site-pattern frequencies

- $p_{AAAA} = 0.09300841$
- $p_{AAAG} = 0.06135527$
- $p_{AAGA} = 0.06135527$
- $p_{AAGG} = 0.06811487$
- $p_{AGAA} = 0.06135527$
- $p_{AGAG} = 0.04672782$
- $p_{AGGA} = 0.04672782$
- $p_{AGGG} = 0.06135527$
- $p_{GAAA} = 0.06135527$
- $p_{GAAG} = 0.04672782$
- $p_{GAGA} = 0.04672782$
- $p_{GAGG} = 0.06135527$
- $p_{GGAA} = 0.06811487$
- $p_{GGAG} = 0.06135527$
- $p_{GGGA} = 0.06135527$
- $p_{GGGG} = 0.09300841$

etc.
Some numbers

Expected flattening matrix for $1,2|3,4$

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0.093008</td>
<td>0.061355</td>
<td>0.061355</td>
<td>0.068115</td>
</tr>
<tr>
<td>AG</td>
<td>0.061355</td>
<td>0.046728</td>
<td>0.046728</td>
<td>0.061355</td>
</tr>
<tr>
<td>GA</td>
<td>0.061355</td>
<td>0.046728</td>
<td>0.046728</td>
<td>0.061355</td>
</tr>
<tr>
<td>GG</td>
<td>0.068115</td>
<td>0.061355</td>
<td>0.061355</td>
<td>0.093008</td>
</tr>
</tbody>
</table>

Delete redundant 3rd row and column...

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0.093008</td>
<td>0.061355</td>
<td>0.068115</td>
</tr>
<tr>
<td>AG</td>
<td>0.061355</td>
<td>0.046728</td>
<td>0.061355</td>
</tr>
<tr>
<td>GG</td>
<td>0.068115</td>
<td>0.061355</td>
<td>0.093008</td>
</tr>
</tbody>
</table>

Note that we can now obtain the last column of the above matrix as a linear combination of the first two columns:

\[
f_{AA,GG} = -f_{AA,AA} + 2.62617 f_{AA,AG} = 0.068115
\]

\[
f_{AG,GG} = -f_{AG,AA} + 2.62617 f_{AG,AG} = 0.061355
\]

\[
f_{GG,GG} = -f_{GG,AA} + 2.62617 f_{GG,AG} = 0.093008
\]

\[\therefore\text{matrix has only two linearly independent rows and columns; rank is 2}\]
Estimating the rank (2-state data)

To estimate the rank, compute the singular value decomposition (SVD) of each matrix, obtaining singular values

\[ s = [s_1, s_2, s_3, s_4] \]

For the true tree, \[ \sum_{i=3}^{4} s_i \approx 0 \]. Otherwise, \[ \sum_{i=3}^{4} s_i > 0 \]

Thus, for each of the three trees for four taxa, we can compute the Frobenius distance from each to the nearest rank 2 matrix:

\[
\text{score} = \sqrt{\sum_{i=3}^{4} s_i^2}
\]

where the \( s_i \) are the 4 singular values resulting from the SVD

Then choose the tree with the lowest score.
The Singular Value Decomposition (SVD)

Decompose an initial matrix into 3 new ones, such that multiplying the new matrices as shown below returns the original matrix exactly

\[ M = U \Sigma V^T \]
The Singular Value Decomposition (SVD)

\[ f_{12,34} = \begin{bmatrix} 0.093008 & 0.061355 & 0.061355 & 0.068115 \\ 0.061355 & 0.046728 & 0.046728 & 0.061355 \\ 0.061355 & 0.046728 & 0.046728 & 0.061355 \\ 0.068115 & 0.061355 & 0.061355 & 0.093008 \end{bmatrix} = U \Sigma V^T \]

\[ U = \begin{bmatrix} -0.562539 & 0.707107 & 0.428427 & 0 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.562539 & 0.707107 & 0.428427 & 0 \end{bmatrix}, \quad V = \begin{bmatrix} -0.562539 & 0.707107 & 0.428427 & 0 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.562539 & 0.707107 & 0.428427 & 0 \end{bmatrix} \]

\[ \Sigma = \begin{bmatrix} 0.254578 & 0 & 0 & 0 \\ 0 & 0.024893 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \]

Last two singular values are zero; rank = 2

Check:

\[ f_{12,34} = U \Sigma V^T \]

\[ \begin{bmatrix} -0.562539 & 0.707107 & 0.428427 & 0 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.562539 & 0.707107 & 0.428427 & 0 \end{bmatrix} \begin{bmatrix} 0.254578 & 0 & 0 & 0 \\ 0 & 0.024893 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -0.562539 & -0.428427 & -0.428427 & -0.562539 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.562539 & 0.707107 & 0.428427 & 0 \end{bmatrix} \]

\[ \begin{bmatrix} 0.093008 & 0.061355 & 0.061355 & 0.068115 \\ 0.061355 & 0.046728 & 0.046728 & 0.061355 \\ 0.061355 & 0.046728 & 0.046728 & 0.061355 \\ 0.068115 & 0.061355 & 0.061355 & 0.093008 \end{bmatrix} \]
When the flattening corresponds to a tree that did not generate the data

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>AG</td>
<td>b</td>
<td>d</td>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td>GA</td>
<td>b</td>
<td>d</td>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td>GG</td>
<td>c</td>
<td>b</td>
<td>b</td>
<td>a</td>
</tr>
</tbody>
</table>

Flattening matrix for 1,2|3,4

Flattening matrix for 1,3|2,4
When the flattening corresponds to a tree that did not generate the data

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>AG</td>
<td>b</td>
<td>d</td>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td>GA</td>
<td>b</td>
<td>d</td>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td>GG</td>
<td>c</td>
<td>b</td>
<td>b</td>
<td>a</td>
</tr>
</tbody>
</table>

**Flattening matrix for 1,2|3,4**

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>d</td>
</tr>
<tr>
<td>AG</td>
<td></td>
<td>c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
When the flattening corresponds to a tree that did not generate the data

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$a$</td>
<td>$b$</td>
<td>$b$</td>
<td>$c$</td>
</tr>
<tr>
<td>AG</td>
<td>$b$</td>
<td>$d$</td>
<td>$d$</td>
<td>$b$</td>
</tr>
<tr>
<td>GA</td>
<td>$b$</td>
<td>$d$</td>
<td>$d$</td>
<td>$b$</td>
</tr>
<tr>
<td>GG</td>
<td>$c$</td>
<td>$b$</td>
<td>$b$</td>
<td>$a$</td>
</tr>
</tbody>
</table>

Flattening matrix for $1,2|3,4$

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GA</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$a$</td>
<td>$b$</td>
<td>$b$</td>
<td>$d$</td>
</tr>
<tr>
<td>AG</td>
<td>$b$</td>
<td>$c$</td>
<td>$d$</td>
<td>$b$</td>
</tr>
<tr>
<td>GA</td>
<td>$b$</td>
<td>$d$</td>
<td>$c$</td>
<td>$b$</td>
</tr>
<tr>
<td>GG</td>
<td>$d$</td>
<td>$b$</td>
<td>$b$</td>
<td>$a$</td>
</tr>
</tbody>
</table>

Flattening matrix for $1,3|2,4$

No redundant rows; matrix is full rank (=4)
When the flattening corresponds to a tree that did not generate the data

Demonstration that matrix is full rank:

\[
\mathbf{f}_{13,24} = \begin{bmatrix}
0.093008 & 0.061355 & 0.061355 & 0.046728 \\
0.061355 & 0.068115 & 0.046728 & 0.061355 \\
0.061355 & 0.046728 & 0.068115 & 0.061355 \\
0.046728 & 0.061355 & 0.061355 & 0.093008
\end{bmatrix}
\]

\[
= \mathbf{U}\Sigma\mathbf{V}^T
\]

\[
\mathbf{U} = \begin{bmatrix}
-0.524622 & 0.707107 & 0 & 0.474101 \\
-0.474101 & 0 & -0.707107 & -0.524622 \\
-0.474101 & 0 & 0.707107 & -0.524622 \\
-0.524622 & -0.707107 & 0 & 0.474101
\end{bmatrix},
\quad
\mathbf{V} = \begin{bmatrix}
-0.524622 & 0.707107 & 0 & 0.474101 \\
-0.474101 & 0 & -0.707107 & -0.524622 \\
-0.474101 & 0 & 0.707107 & -0.524622 \\
-0.524622 & -0.707107 & 0 & 0.474101
\end{bmatrix}
\]

\[
\mathbf{\Sigma} = \begin{bmatrix}
0.250629 & 0 & 0 & 0 \\
0 & 0.046280 & 0 & 0 \\
0 & 0 & 0.021387 & 0 \\
0 & 0 & 0 & 0.003950
\end{bmatrix}
\]

All singular values are nonzero; matrix is full rank (= 4)

Note: we just used JC model here for illustration; the method works under much more general models (more later)
Calculation of SVD Scores (4-state data)

\[ score = \sqrt{\sum_{i=5}^{16} s_i^2} \]

= “Frobenius distance” to nearest rank 4 matrix

Simulation conditions:

- tree = (((1:0.05,2:0.05):0.05,3:0.1):0.05,4:0.15)
- 1,000,000 sites
- HKY model: \( \kappa=4 \pi=(0.1, 0.2, 0.3, 0.4) \)
- all sites share same history (no incomplete lineage sorting, horizontal transfer, gene duplication and loss, etc.)

<table>
<thead>
<tr>
<th>SV (s)</th>
<th>1,2 ( \mid ) 3,4</th>
<th>1,3 ( \mid ) 2,4</th>
<th>1,4 ( \mid ) 2,3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.279686</td>
<td>0.278714</td>
<td>0.278716</td>
</tr>
<tr>
<td>2</td>
<td>0.21899</td>
<td>0.219191</td>
<td>0.219191</td>
</tr>
<tr>
<td>3</td>
<td>0.10902</td>
<td>0.110392</td>
<td>0.110389</td>
</tr>
<tr>
<td>4</td>
<td>0.056873</td>
<td>0.05709</td>
<td>0.05709</td>
</tr>
<tr>
<td>5</td>
<td>8E-05</td>
<td>0.006875</td>
<td>0.006886</td>
</tr>
<tr>
<td>6</td>
<td>6.14E-05</td>
<td>0.006315</td>
<td>0.006305</td>
</tr>
<tr>
<td>7</td>
<td>4.93E-05</td>
<td>0.003286</td>
<td>0.003286</td>
</tr>
<tr>
<td>8</td>
<td>3.8E-05</td>
<td>0.003244</td>
<td>0.003246</td>
</tr>
<tr>
<td>9</td>
<td>3.26E-05</td>
<td>0.002905</td>
<td>0.002903</td>
</tr>
<tr>
<td>10</td>
<td>3.09E-05</td>
<td>0.002499</td>
<td>0.002499</td>
</tr>
<tr>
<td>11</td>
<td>2.69E-05</td>
<td>0.001471</td>
<td>0.001472</td>
</tr>
<tr>
<td>12</td>
<td>2.23E-05</td>
<td>0.001182</td>
<td>0.001181</td>
</tr>
<tr>
<td>13</td>
<td>1.3E-05</td>
<td>0.001009</td>
<td>0.001008</td>
</tr>
<tr>
<td>14</td>
<td>1.03E-05</td>
<td>0.000937</td>
<td>0.000937</td>
</tr>
<tr>
<td>15</td>
<td>6.19E-06</td>
<td>0.000382</td>
<td>0.000384</td>
</tr>
<tr>
<td>16</td>
<td>1.56E-06</td>
<td>0.000377</td>
<td>0.000375</td>
</tr>
</tbody>
</table>

score

\[ 0.000133 \quad 0.011353 \quad 0.011354 \]
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of \textit{inconsistent quartets} (i.e., seek a solution to the Maximum Quartet Consistency problem).

Suppose we infer these quartet relationships for 5 taxa.
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of inconsistent quartets (i.e., seek a solution to the Maximum Quartet Consistency problem).
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

```
12|34
12|35
12|45
14|35
23|45
```
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

```
12|34
12|35
12|45
14|35
23|45
```
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of inconsistent quartets (i.e., seek a solution to the Maximum Quartet Consistency problem).

```
12|34  
12|35  
12|45  
14|35  
23|45  
```
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of inconsistent quartets (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

4 consistent quartets, 1 inconsistent quartet
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

```
12|34
12|35
12|45
14|35
23|45
```
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

```
12|34
12|35
12|45
14|35
23|45
```
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

```
12|34
12|35
12|45
14|35
23|45
```
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of inconsistent quartets (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of inconsistent quartets (i.e., seek a solution to the Maximum Quartet Consistency problem).

12\mid34
12\mid35
12\mid45
14\mid35
23\mid45
Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of inconsistent quartets (i.e., seek a solution to the Maximum Quartet Consistency problem).

2 consistent quartets, 3 inconsistent quartet

Now evaluate the remaining 13 trees and choose the one that maximizes the number of consistent quartets
Handling >4 taxa

While evaluation of each possible tree might work well for 5-tip trees, the number of possible trees for $n$ tips grows too quickly to make it a general strategy.

Must use a heuristic algorithm to search for the best tree:

- The default in PAUP* is a heavily modified version of “QFM” (Reaz et al., 2014)
- Other algorithms are available in PAUP* and elsewhere
- Unfortunately, the MQC problem is NP-hard (i.e., exact solution will be slow for large numbers of tips)
Allman-Rhodes-Eriksson method (ErikSVD)

Work for extremely general models:

*Does* assume that all sites in the alignment are
independently and identically distributed according to a
general Markov model.

But:

- No assumption of stationarity/time-reversibility!
- No assumption of homogeneity over the tree!
  Each branch may have its own transition matrix, or even
  multiple transition matrices along the same branch.

Enables inferences about the tree topology to be made without having to
estimate the parameters of the underlying model(s).

A recent improvement by Jesús Fernández-Sánchez and Marta Casanellas
(Syst. Biol. 2016): Do an additional row and column normalization to reduce
the error associated with low counts for certain entries of the flattening
matrices (“Erik+2”)
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees

Gene tree matches species tree
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees

New mutation
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees

A

B

C

D
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees

First coalescence
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees
Ancestral polymorphism and species trees

Final coalescence
Ancestral polymorphism and species trees

“Incomplete lineage sorting” (ILS)

Gene tree conflicts with species tree
The multispecies coalescent model

- Independence between branches—coalescent events that occur in one population are independent of what happens in other populations within the phylogeny.
- Panmixia—within a population, all pairs of lineages are equally likely to coalesce.
- Divergence is instantaneous and complete—no gene flow occurs after speciation
- ILS only—no other evolutionary processes (e.g., horizontal transfer, duplication and loss, ...) have led to incongruence between gene trees and the species tree.
- No recombination within genes; free recombination between genes
Some species tree methods that assume the multispecies coalescent process

- Full data methods
  - Fully Bayesian (integrate over gene trees within species trees, estimate posterior distribution of population sizes, branch lengths, and other model parameters in addition to the species tree)
    - BEST (Liu and Pearl, 2007; Liu, 2008)
    - *BEAST (Heled and Drummond, 2010)
    - SNAPP (Bryant et al., 2012)
    - BPP (Yang and Rannala, 2010)
  - SVDQuartets (Chifman and Kubatko, 2014, 2015)

- Summary methods (start with estimated gene trees)
  - Methods that use branch lengths:
    - STEM (Kubatko et al., 2009)
    - STEAC (Liu et al. 2009)
  - Methods that only use topology information
    - STAR (Liu et al. 2009)
    - Minimize deep coalescences ((PhyloNet; Than & Nakhleh 2009)
    - MP-EST (Liu et al. 2010)
    - ST-ABC (Fan and Kubatko 2011)
    - STELLS (Wu 2011)
    - ASTRAL (Mirarab et al., 2014; Mirarab and Warnow, 2015)
“Anomaly zones”

Probabilities of all 15 gene-tree topologies for values of x, y, and z

y=1, x=1

y=1, x=0.001

y=0.01, x=0.01

Most probable gene tree is discordant with the species tree

https://www.asc ohio-state.edu/kubatko.2/SpeciesTreeEstimation2017.pdf
Handling incomplete lineage sorting

When times between species divergences are short (or when population sizes are large), the history of individual genes may be discordant from the species tree topology in several (or many) ways:

**SVDQuartets:** Expected rank of flattening matrix is 10 for the true species tree and 16 for the other two trees, under GTR+I+G or any of its submodels!
SVDQuartets

Does not need to assume a molecular clock or constant population size


Θ=0.1
100 sites/locus
10,000,000 loci
(1,000,000,000 sites)

3X increase in mutation rate)
SVQuartets

But more data may be needed to achieve the same level of accuracy (the price of generality)

\[ \Theta = 0.1 \]

100 sites/locus
10,000,000 loci
(1,000,000,000 sites)

3X increase in mutation rate
SVDQuartets

Can also handle migration between pairs of sister lineages (IM model)

A disadvantage:

- No estimates of node ages (branch lengths) or theta parameter
- But we’re working on that...

- Can calculate (analytically) expected site pattern probabilities as a function of the species tree, node age, and theta parameters by integrating over the coalescent times and summing over all labeled histories (Chifman and Kubatko, 2015)

\[
\sum_{x \in \{k\}} p^*_x = \sum_{(G,t) \in \mathcal{G}_4(G,t) \mid (G_5,t_*) \in \mathcal{G}_5} \int_{t_*} p_{\sigma(i_1i_2i_4i_5)}(G,t) \\
\times f((G_5, t_*))(S_5, t_*) \, dt_* \\
= \sum_{(G,t) \in \mathcal{G}_4(G,t) \mid (G_5,t_*) \in \mathcal{G}_5} \int_{t} \prod_{b \in B^*} f_{P_b}(t) f_{P_{b^*}}(t, t^*) \, dt \, dt^*
\]

- Perform a ML optimization of node ages and theta, maximizing fit of observed to expected pattern frequencies under a multinomial model
- Recently, we have derived equations for computing the first and second order derivatives for node ages and theta, allowing estimation of sampling variance via the Fisher Information Matrix.
SVDQuartets

<table>
<thead>
<tr>
<th>Node</th>
<th>True age</th>
<th>Estimated Age</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.5</td>
<td>0.5206</td>
<td>0.0623</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>0.9714</td>
<td>0.0822</td>
</tr>
<tr>
<td>Z</td>
<td>1.5</td>
<td>1.498</td>
<td>0.1024</td>
</tr>
</tbody>
</table>

estimated $\Theta = 0.101 (0.0044)$

(need to test bigger trees)
• For 4 species, the dominant quartet topology is the species tree [Allman, et al. 2010]

• For >4 species, the dominant topology may be different from the species tree [Degnan and Rosenberg, 2006]

1. Break up input each gene tree into trees on 4 taxa (quartet trees)

2. Find all \(\binom{n}{4}\) dominant quartet topologies

3. Combine dominant quartet trees

\[p_1 = 30\% \quad p_2 = 30\% \quad p_3 = 40\%\]
ASTRAL with >4 species

Find the species tree with the maximum number of induced quartet trees shared with a collection of gene trees

Three example gene trees and their induced quartets:
(trees are considered to be unrooted)

Number of loci with this tree

10
(human,chimp),(gorilla,orang)
(human,chimp),(gorilla,gibbon)
(human,chimp),(orang,gibbon)
(chimp,gorilla),(orang,gibbon)

5
(chimp,gorilla),(human,orang)
(human,gibbon),(chimp,gorilla)
(human,chimp),(orang,gibbon)
(chimp,gorilla),(orang,gibbon)

1
(human,orang),(chimp,gorilla)
(human,chimp),(gorilla,gibbon)
(human,orang),(chimp,gibbon)
(human,orang),(gorilla,gibbon)
(chimp,orang),(gorilla,gibbon)
### ASTRAL with >4 species

Compute the total number of times each resolved quartet was found over all input gene trees

<table>
<thead>
<tr>
<th>Resolved quartet</th>
<th>Induced by tree(s)</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>(human, chimp), (gorilla, orang)</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>(human, chimp), (gorilla, gibbon)</td>
<td>1,3</td>
<td>10+1=11</td>
</tr>
<tr>
<td>(human, chimp), (orang, gibbon)</td>
<td>1,2</td>
<td>10+5=15</td>
</tr>
<tr>
<td>(human, gorilla), (orang, gibbon)</td>
<td>1,2</td>
<td>10+5=15</td>
</tr>
<tr>
<td>(chimp, gorilla), (orang, gibbon)</td>
<td>1,2</td>
<td>10+5=15</td>
</tr>
<tr>
<td>(chimp, gorilla), (human, orang)</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>(human, gibbon), (chimp, gorilla)</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>(human, orang), (chimp, gorilla)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>(human, orang), (chimp, gibbon)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>(human, orang), (gorilla, gibbon)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>(chimp, orang), (gorilla, gibbon)</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
ASTRAL with >4 species

Find the species tree that maximizes the number of consistent quartets (*a la* SVDQuartets)

For each species tree evaluated, we sum the weights of all satisfied quartets

Could also evaluate 11 more species trees (have to search over all unrooted species trees, including trees that never appeared as a gene tree)
Where do the gene trees come from?

That’s your problem! Typically, people run RAxML or IQ-TREE to estimate gene trees. ASTRAL is a very fast method once you have the gene trees, but the gene-tree estimation typically dominates the total run time.

ASTRAL makes a consistent estimate of the species tree, as long as the input gene trees themselves are estimated using a consistent method. If the gene trees estimates are biased, there is no guarantee of consistency.

Astral provides exact and heuristic algorithms for the MQC tree search. The exact method will be too slow if there are very many tips.

Download at: https://github.com/smirarab/ASTRAL
Selected References


