How can a single bad apple spoil the bunch? Using polyhedral geometry to understand phylogenetic inference on non tree metrics

M. A. “Angie” Cueto¹  F. A. “Erick” Matsen²

¹Department of Mathematics
University of California, Berkeley
(advisor: Bernd Sturmfels)

²Miller Institute Fellow
Departments of Integrative Biology, Mathematics, and Statistics
University of California, Berkeley
(going to Fred Hutchinson Cancer Research Center in 2010)

September 10, 2009

Phylogenetics is the inference of evolutionary history

A bit of phylogenetics terminology

unrooted tree

rooted tree

What happens when we add an extra taxon to a phylogenetic data set?

Motivation 2: can unacknowledged recombination destroy the “base” tree?

Motivation 3: understand geometry of phylogenetic inference

Balanced minimum evolution (BME) phylogenetics

The BME length of an $n$ taxon tree $T$ with respect to an $n \times n$ distance matrix $D$ is (Semple and Steel, 04):

$$\ell(T, D) = \sum_{ij} w_T(i, j)D(i, j).$$

Let $p(i, j)$ denote the path from $i$ to $j$ in the unrooted tree $T$, and define the weight of a path from leaf $i$ to leaf $j$ as

$$w_T(i, j) = \prod_{v \in p(i, j)} \frac{1}{\deg(v) - 1}.$$

Given a distance matrix $D$, the BME criterion simply picks the tree $T$ minimizing $\ell(T, D)$.

When $D$ is a distance matrix which comes from a tree $T_0$, $\ell(T, D)$ will be minimized at $T = T_0$.

Neighbor-joining is a greedy BME length minimizer (Desper, Gascuel, Pauplin, Semple, Steel, etc.)
Terminology in polyhedral geometry

\[
\text{conv}(x_1, \ldots, x_n) = \left\{ \lambda_1 x_1 + \cdots + \lambda_n x_n : \lambda_i \geq 0, \sum_{i=1}^{n} \lambda_i = 1 \right\}
\]

\[
\text{cone}(x_1, \ldots, x_n) = \left\{ \lambda_1 x_1 + \cdots + \lambda_n x_n : \lambda_i \geq 0 \right\}
\]
Normal fan construction: separating facets

Normal fan construction: that's it

"F-vector" for BME normal fans

<table>
<thead>
<tr>
<th>dimension</th>
<th>number of taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 1 1</td>
</tr>
<tr>
<td>1</td>
<td>3 52 90,262</td>
</tr>
<tr>
<td>2</td>
<td>3 210 640,140</td>
</tr>
<tr>
<td>3</td>
<td>250 1,742,445</td>
</tr>
<tr>
<td>4</td>
<td>105 2,373,345</td>
</tr>
<tr>
<td>5</td>
<td>15 1,715,455</td>
</tr>
<tr>
<td>6</td>
<td>635,265</td>
</tr>
<tr>
<td>7</td>
<td>105,945</td>
</tr>
<tr>
<td>8</td>
<td>5,460</td>
</tr>
<tr>
<td>9</td>
<td>105</td>
</tr>
</tbody>
</table>

Note that the BME normal fan for 4 and 5 taxa was constructed by Eickmeyer et. al., *Alg. Mol. Biol.* 3:1, 2008

What our distance matrices will look like

What does BME do with these sorts of distance matrices?

Question 1:
Which of the following trees is the BME optimal tree after adding a taxon to a tree metric with the topology...

- of course
- no
- yes (1)

etc...
Question 2: when do we get these interesting examples? Sufficient conditions

Answers:

1. for any quartet branch lengths there exists an “interesting” extra column
   ▶ but that column must contain entries of size at least three times the quartet internal branch length
2. for certain quartet branch lengths there is a weak constraint on “interesting” extra columns

Answer:

1. for any quartet branch lengths there exists an “interesting” extra column
   ▶ but that column must contain entries of size at least three times the quartet internal branch length
2. for certain quartet branch lengths there is a weak constraint on “interesting” extra columns
Q2, A1: for any quartet branch lengths, there exists an interesting column

starting with \( \frac{a}{b \leftarrow c} \), interesting columns are

\[ [0, 3, 3, 0] \]

to get a tree of topology

\[
\begin{array}{c}
\text{a} \\
\text{e} \\
\text{d} \\
\text{c} \\
\text{b} \\
\end{array}
\]

(distance to scale)

There is an analogous picture for any set of pendant branch lengths, but there must always be an entry of size at least three times the internal branch length of the quartet.

Q2, A2: for certain quartet branch lengths, can add a column of any norm

start with \( \frac{a}{b \leftarrow c} \) or \( \frac{a}{b \leftarrow d} \)

plus any positive linear combination of

\[
\begin{array}{cccc}
\text{a} & \text{b} & \text{c} & \text{d} \\
\text{c} & \text{a} & \text{b} & \text{d} \\
\text{b} & \text{a} & \text{c} & \text{d} \\
\text{d} & \text{a} & \text{b} & \text{c} \\ 
\end{array}
\]

The entries of the extra column must satisfy

\[ 0 \leq D_{ae} \leq D_{de} \leq \min(D_{be}, D_{ce}) \]

Future directions

- Can we get “interesting” examples from unacknowledged recombination?
- What else can we say about the geometric structure of the BME partition?
- What about likelihood and parsimony?