Fitting Tree Metrics with Minimum Disagreements

Evangelos Kipouridis
Any guess?

I think

The tree between A and B, various type of relation. C and B, the first predation, B and D, rather greater distance. Then some kind of formation - binary relation.
Any guess?
Darwin’s notes
Tree of life

-3.5 Billion

Single Cell

Today
Tree of life

-3.5 Billion

Single Cell

Today
Tree of life

-3.5 Billion

Single Cell

Today
Tree of life

-3.5 Billion

Single Cell

Today
Tree of life

-3.5 Billion

Today

Single Cell
Tree of life

-3.5 Billion

Today
First to discuss tree reconstruction
They're the same picture…
They're the same picture…
They're the same picture…

Hmm…

15 million years?
# Reconstruct from distance matrix

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>15</th>
<th>435</th>
<th>432</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>435</td>
<td>436</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>432</td>
<td>433</td>
<td>63</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Reconstruct from distance matrix

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>15</th>
<th>435</th>
<th>432</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>435</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>432</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If, due to noise, no matching tree?
Reconstruct from distance matrix

If, due to noise, no matching tree?

Minimize disagreements!
Reconstruct from distance matrix

If, due to noise, no matching tree?

Minimize disagreements!

Can also minimize total error, max error, L2 error…
A Battle Is Raging in the Tree of Life

Which came first, the sponge or the comb jelly?

Scientists Have Found the First Branch on the Tree of Life

Something had to diverge from the trunk eventually.
What we know...

APX - Hard
What we know…

APX - Hard

O(1) approximation for ultrametrics (structured trees) - even under mild constraints
What we know…

APX - Hard

O(1) approximation for ultrametrics (structured trees) - even under mild constraints

- What about unstructured trees?
Structuring the unstructured
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.

Input(\(\alpha, u\)) = 12, but OPT(\(\alpha, u\)) = 8
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.

Input(\(\alpha, u\)) = 12, but OPT(\(\alpha, u\)) = 8
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.

Input(α,v) = 6, but OPT(α,v) = 7
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.

Input(α,v) = 6, but OPT(α,v) = 7
Structuring the unstructured

1) Find a root.

2) Find depths of leaves.

Now for all u we know depth(u).

depth(u) = OPT'(α,u) = Input(α,u)
1) Find a root.
2) Find depths of leaves.

Now for all $u$ we know $\text{depth}(u)$.  
$\text{depth}(u) = \text{OPT}(\alpha,u) = \text{Input}(\alpha,u)$

How much did we pay?
- We moved exactly $D(\alpha)$ nodes, each introduced at most $(n-1)$ disagreements.
- $D(\text{OPT}') \leq D(\text{OPT}) + D(\alpha) (n-1)$

D() denotes disagreements in OPT
Structuring the unstructured

1) **Find a root.**
2) **Find depths of leaves.**

\[ D(OPT) = \frac{1}{2} \sum_{u} D(u) \]

\[ D(OPT') \leq D(OPT) + D(\alpha) (n-1) \]
Structuring the unstructured

1) **Find a root.**
2) **Find depths of leaves.**

$$D(OPT) = \frac{1}{2} \sum_u D(u)$$

D() denotes disagreements in OPT
$$D(OPT') \leq D(OPT) + D(\alpha) (n-1)$$
\(\alpha\) minimizes disagreements
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.

\[
D(OPT) = \frac{1}{2} \sum_{u} D(u) \geq \frac{1}{2} \cdot n \cdot D(\alpha)
\]

D() denotes disagreements in OPT
\[D(OPT') \leq D(OPT) + D(\alpha) \cdot (n-1)\]
\(\alpha\) minimizes disagreements
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.

\[ D(OPT) = \frac{1}{2} \sum_{u} D(u) \]

\[ \geq \frac{1}{2} \cdot n \cdot D(\alpha) \]

\[ D(\alpha) \leq \frac{2D(OPT)}{n} \]

D(\alpha) denotes disagreements in OPT
\[ D(OPT') \leq D(OPT) + D(\alpha) (n-1) \]
\(\alpha\) minimizes disagreements
Structuring the unstructured

1) Find a root.
2) Find depths of leaves.

\[ D(\alpha) \leq \frac{2D(OPT)}{n} \]

\[ D(OPT') \leq D(OPT) + D(\alpha)(n-1) \]

\[ D(OPT') \leq 3D(OPT) \]
Where do we stand?
Where do we stand?
Where do we stand?
Where do we stand?
Reduce to ultrametric (all leaves same depth)
Reduce to ultrametric (all leaves same depth)
Reduce to ultrametric (all leaves same depth)
So... who is the "Oldest Sister"?

Was it the sponge or the comb jelly that diverged first?
So… who is the “Oldest Sister”?

Was it the sponge or the comb jelly that diverged first?