Evolutionary Trees

R.C.T. Lee and Chin Lung Lu

CS 5313 Algorithms for Molecular Biology

Evolutionary tree

To describe the evolutionary relationship among species

- **Root**
- **Extinct Ancestor**
- **Bifurcating Speciation**
- **Internal Node**
- **Extinct Ancestor**

- **Leaf**
- **Extant Species**
Construction of Evolutionary Tree

Distance Matrix $M$

<table>
<thead>
<tr>
<th></th>
<th>Man</th>
<th>Monkey</th>
<th>Dog</th>
<th>Horse</th>
<th>Donkey</th>
<th>Pig</th>
<th>Rabbit</th>
<th>Kangaroo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Monkey</td>
<td>0</td>
<td>12</td>
<td>16</td>
<td>15</td>
<td>12</td>
<td>11</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Dog</td>
<td>0</td>
<td>10</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>11</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Donkey</td>
<td>0</td>
<td>4</td>
<td>10</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pig</td>
<td>0</td>
<td>6</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rabbit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Kangaroo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Evolutionary Tree $T$

Rooted Evolutionary Tree

The degree of each internal node is 3, except the root node.
Unrooted Evolutionary Tree

- The degree of each internal node is 3.

Number of unrooted trees

<table>
<thead>
<tr>
<th>No of Trees</th>
<th>Structure of Trees</th>
<th>No of Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 2$</td>
<td>$s_1 \rightarrow s_2$</td>
<td>1</td>
</tr>
<tr>
<td>$n = 3$</td>
<td>$s_1 \rightarrow s_2$</td>
<td>3</td>
</tr>
<tr>
<td>$n = 4$</td>
<td>$s_1 \rightarrow s_2$</td>
<td>5</td>
</tr>
</tbody>
</table>
Number of unrooted trees

How to add a new species into the tree?

If a new species is added to an unrooted tree, the number of edges is increased by 2.

$NE(n)$: number of edges of an unrooted tree with $n$ species

By induction, we have $NE(n) = 2n - 3$.

Number of unrooted trees

$TU(n)$: number of unrooted trees for $n$ species

Since $NE(n) = 2n - 3$, we have

$$TU(n + 1) = (2n - 3)TU(n)$$

or

$$TU(n) = (2n - 5)TU(n - 1)$$

That is,

$$TU(n) = (2n - 5)(2n - 7) \cdots 1$$
### Number of unrooted trees

<table>
<thead>
<tr>
<th>No of Species</th>
<th>No of Unrooted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>:</td>
</tr>
<tr>
<td>17</td>
<td>6,190,283,353,629,375</td>
</tr>
<tr>
<td>18</td>
<td>191,898,783,962,510,625</td>
</tr>
<tr>
<td>19</td>
<td>6,332,659,870,762,850,625</td>
</tr>
<tr>
<td>20</td>
<td>221,643,095,476,699,771,875</td>
</tr>
</tbody>
</table>

### Change unrooted into rooted

1. **s₁** [root] **s₂** → **s₁** **s₂** [root]
2. **s₁** [root] **s₂** → **s₁** **s₂** [root]
3. **s₁** [root] **s₂** → **s₁** **s₂** [root]
Number of rooted trees

- $TR(n)$: number of rooted trees for $n$ species
- Since there are $2n - 3$ edges in every unrooted tree for $n$ species, we have

$$TR(n) = (2n - 3)TU(n)$$

$$= (2n - 3)(2n - 5)(2n - 7) \cdots 1$$

$$= TU(n + 1)$$

No of Species | No of Rooted Trees
---|---
2 | 1
3 | 3
4 | 15
\vdots | \vdots
17 | 191,898,783,962,510,625
18 | 6,332,659,870,762,850,626
19 | 221,643,095,476,699,771,875
20 | 8,200,794,532,637,891,559,375
**Evolution Tree Problem**

- **Input**: A distance matrix \( M \) of \( n \) species
- **Output**: Find an evolution tree \( T \) under some criterion

**Basic criterion**: \( d_T(s_i, s_j) \geq d_M(s_i, s_j) \)
- \( d_M(s_i, s_j) \): the distance between species \( s_i \) and \( s_j \) in \( M \)
- \( d_T(s_i, s_j) \): the distance between \( s_i \) and \( s_j \) in \( T \)

**Further criterion**: when two species are close to each other in the distance matrix, they should be close in the evolution tree

**Criteria of Evolution Tree**

- **MINIMAX** evolution tree: an evolutionary tree with \( d_T(s_i, s_j) \geq d_M(s_i, s_j) \) such that
  \[
  \max_{1 \leq i < j \leq n} \left[ d_T(s_i, s_j) - d_M(s_i, s_j) \right]
  \]
  is minimized

- **MINISUM** evolution tree: an evolutionary tree with \( d_T(s_i, s_j) \geq d_M(s_i, s_j) \) such that
  \[
  \sum_{1 \leq i < j \leq n} d_T(s_i, s_j)
  \]
  is minimized

- **MINISIZE** evolution tree: an evolutionary tree with \( d_T(s_i, s_j) \geq d_M(s_i, s_j) \) such that the total length of the tree is minimized
Complexities of Evolutionary Tree

<table>
<thead>
<tr>
<th></th>
<th>MINIMAX</th>
<th>MINISUM</th>
<th>MINISIZE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rooted</td>
<td>$O(n^2)$</td>
<td>NP-C</td>
<td>NP-C</td>
</tr>
<tr>
<td>Unrooted</td>
<td>NP-C</td>
<td>NP-C</td>
<td>?</td>
</tr>
</tbody>
</table>

It is still unknown whether the unrooted MINISIZE evolutionary tree problem is NP-complete.

Minimax rooted tree algorithm

**Input:** a distance matrix of a set $S$ of $n$ species

1. If $S$ contains only one species, return it as the tree.
2. Find a minimal spanning tree $T$ of $S$.
3. Find the longest $d(s_i, s_j)$ in the distance matrix. Find the longest edge $e$ in the path linking $s_i$ and $s_j$ in $T$. Let $S_i$ and $S_j$ be the two sets of species obtained by breaking edge $e$.
4. Use this algorithm recursively to find subtrees $T_i$ and $T_j$ for $S_i$ and $S_j$ respectively.
5. Construct a rooted tree with $T_i$ and $T_j$ as subtrees such that $dt(s_i, s_j) = d(s_i, s_j)$. 

By R.C.T. Lee and C.L. Lu
Basic principle of Minimax

Let \( s_i \) and \( s_j \) be the two species which have the longest distance in the distance matrix.

The longest distance is exactly preserved.

Example of Minimax algorithm

1. Consider the distance matrix as follows:

   \[
   \begin{array}{c|cccc}
   & s_1 & s_2 & s_3 & s_4 \\
   \hline
   s_1 & 0 & 2 & 3 & 3.1 \\
   s_2 & 0 & 3.6 & 5 & \\
   s_3 & 0 & 1 & & \\
   s_4 & 0 & & & \\
   \end{array}
   \]

2. Construct a minimal spanning tree \( T \) as follows:
Example of Minimax algorithm ²

3. The distance between \( s_2 \) and \( s_4 \) is the longest.

\[
\begin{array}{c|cccc}
\text{s} & s_1 & s_2 & s_3 & s_4 \\
\hline
s_1 & 0 & 2 & 3 & 3.1 \\
s_2 & 0 & 3.6 & 5 & \\
s_3 & & 0 & 1 & \\
s_4 & & & 0 & \\
\end{array}
\]

4. The path linking \( s_2 \) and \( s_4 \) in \( T \) in which \((s_1, s_3)\) is the longest edge.

Example of Minimax algorithm ³

5. Breaking \((s_1, s_3)\) obtains two subsets of species

\[
\begin{array}{c}
s_2 \quad 2 \quad s_1 \\
S_1 \\
\end{array} \quad \begin{array}{c}
s_3 \quad 1 \quad s_4 \\
S_2 \\
\end{array}
\]

6. Construct two subtrees \( T_1 \) and \( T_2 \) for \( S_1 \) and \( S_2 \) respectively.
Example of Minimax algorithm

7. Combine $T_1$ and $T_2$ by making sure that $dt(s_2, s_4) = d(s_2, s_4) = 5$.

Determination of edge weights

How to determine the edge weights if the evolutionary tree structure is given?

Example: Given a distance matrix and an unrooted evolutionary tree, determine the edge weights such that the tree size is minimized.
Determination of edge weights

Determine $x_i$ by linear programming

Minimize $x_1 + x_2 + x_3 + x_4 + x_5$

Subject to

\[
\begin{align*}
    x_1 + x_2 & \geq d_{12} \\
    x_1 + x_3 + x_4 & \geq d_{13} \\
    x_1 + x_3 + x_5 & \geq d_{14} \\
    x_2 + x_3 + x_4 & \geq d_{23} \\
    x_2 + x_3 + x_5 & \geq d_{24} \\
    x_4 + x_5 & \geq d_{34}
\end{align*}
\]

Evolutionary Trees p.23

---

Determination of edge weights

Minimize $x_1 + x_2 + x_3 + x_4 + x_5 + x_6$

Subject to

\[
\begin{align*}
    x_1 + x_2 & \geq d_{12} \\
    x_1 + x_5 + x_6 + x_3 & \geq d_{13} \\
    x_1 + x_5 + x_6 + x_4 & \geq d_{14} \\
    x_2 + x_5 + x_6 + x_3 & \geq d_{23} \\
    x_2 + x_5 + x_6 + x_4 & \geq d_{24} \\
    x_3 + x_4 & \geq d_{34} \\
    x_5 + x_1 = x_5 + x_2 = \\
    x_6 + x_3 = x_6 + x_4
\end{align*}
\]

Evolutionary Trees p.24
UPGMA: create a rooted tree

1. **UPGMA**: Unweighted Pair Group Method with Arithmetic Mean

2. **Example**: consider the distance matrix as follows

<table>
<thead>
<tr>
<th></th>
<th>$s_1$</th>
<th>$s_2$</th>
<th>$s_3$</th>
<th>$s_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>$s_2$</td>
<td></td>
<td>0</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>$s_3$</td>
<td></td>
<td></td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$s_4$</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

*Output*: a rooted evolutionary tree

1. Select the pair of species $(s_3, s_4)$ with the smallest distance and construct a rooted tree with $s_3$ and $s_4$ as leaf nodes

$$
\frac{1}{2}d(s_3, s_4) = 1
$$
2. Consider \((s_3, s_4)\) as a new species and the distances are updated as follows:

\[
\begin{align*}
    d(s_1, (s_3, s_4)) &= \frac{1}{2} (d(s_1, s_3) + d(s_1, s_4)) \\
                      &= \frac{1}{2} (4 + 3) = 3.5 \\
    d(s_2, (s_3, s_4)) &= \frac{1}{2} (d(s_2, s_3) + d(s_2, s_4)) \\
                      &= \frac{1}{2} (6 + 5) = 5.5
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>(s_1)</th>
<th>(s_2)</th>
<th>((s_3, s_4))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s_1)</td>
<td>0</td>
<td>4</td>
<td>3.5</td>
</tr>
<tr>
<td>(s_2)</td>
<td></td>
<td>0</td>
<td>5.5</td>
</tr>
<tr>
<td>((s_3, s_4))</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Select the pair of species \((s_1, (s_3, s_4))\) with the smallest distance and construct a rooted as follows:
**UPGMA: create a rooted tree**

4. Since $s_2$ is the only specie left, the final tree will look like as follows:

$$
\frac{1}{2}d(s_2, (s_1, (s_3, s_4))) = \frac{4+5+6}{3} = 2.5
$$

**Neighbor joining: unrooted tree**

Construct a starlike tree as follows:

$$
S_0 = \sum_{i=1}^{n} L_{iX} = \frac{1}{n-1} \sum_{1 \leq i < j \leq n} d(s_i, s_j)
$$
Consider the tree $T_{12}$ as follows:

$$L_{XY} = \frac{1}{2(n-2)} \left[ \sum_{k=3}^{n} (d(s_1, s_k) + d(s_2, s_k)) - (n-2)(L_{1Y} + L_{2Y}) - 2 \sum_{i=3}^{n} L_{iX} \right]$$

The tree length $S_{12}$ of $T_{12}$ is:

$$S_{12} = L_{XY} + (L_{1Y} + L_{2Y}) + \sum_{i=3}^{n} L_{iX}$$

$$= \frac{1}{2(n-2)} \sum_{k=3}^{n} (d(s_1, s_k) + d(s_2, s_k)) + \frac{1}{2} d(s_1, s_2) + \frac{1}{n-2} \sum_{3 \leq i < j \leq n} d(s_i, s_j)$$
Consider \((s_1, s_2)\) as a new species \(s_{(12)}\) and the distances are updated as follows: for \(3 \leq i \leq n\)

\[
d(s_{(12)}, s_i) = \frac{d(s_1, s_i) + d(s_2, s_i)}{2}
\]

1. For all pairs of \(s_i\) and \(s_j\), we compute the tree size \(S_{ij}\) of tree \(T_{ij}\). \((\frac{n(n-1)}{2}\) such trees)
2. Choose the pair with the smallest tree size. Suppose \(S_{12}\) is such a pair. Then consider \((s_1, s_2)\) as a new species \(s_{(12)}\) and the distances are updated as follows: \(d(s_{(12)}, s_i) = \frac{d(s_1, s_i) + d(s_2, s_i)}{2}\) for \(3 \leq i \leq n\)
3. The number of species is reduced by one and for the new distance matrix, the above procedure is again applied to find the next pair of neighbors until the number of species becomes 3.
Neighbor joining: example

Example: consider the distance matrix as follows

\[
\begin{array}{c|cccc}
   & s_1 & s_2 & s_3 & s_4 \\
\hline
s_1 & 0 & 4 & 4 & 3 \\
s_2 & & 0 & 6 & 5 \\
s_3 & & & 0 & 2 \\
s_4 & & & & 0 \\
\end{array}
\]

 Neighbor joining: example

\[
\begin{array}{c|cccc}
S_{ij} & s_1 & s_2 & s_3 & s_4 \\
\hline
s_1 & 7.5 & 8.25 & 8.25 & \\
s_2 & & 7.75 & 8.25 & \\
s_3 & & & 7.5 & \\
\end{array}
\]
Neighborhood joining: example

<table>
<thead>
<tr>
<th>(d(s_i, s_j))</th>
<th>(s_{(12)})</th>
<th>(s_3)</th>
<th>(s_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s_{(12)})</td>
<td>0</td>
<td>(\frac{4+6}{2} = 5)</td>
<td>(\frac{3+5}{2} = 4)</td>
</tr>
<tr>
<td>(s_3)</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>(s_4)</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Evolutionary Trees  p.37

Estimate edge lengths

- Fitch and Margoliash’s (1967) method:

  \[ d(s_1, s_2) = x + y, \]
  \[ d(s_1, s_3) = x + z, \]
  \[ d(s_2, s_3) = y + z \]

- \(x = \frac{d(s_1, s_2) + d(s_1, s_3) - d(s_2, s_3)}{2} \)
- \(y = \frac{d(s_2, s_1) + d(s_2, s_3) - d(s_1, s_3)}{2} \)
- \(z = \frac{d(s_3, s_1) + d(s_3, s_2) - d(s_1, s_2)}{2} \)

Evolutionary Trees  p.38
Estimate edge lengths

\[ L_{1Y_1} = \frac{d(s_1, s_2) + d(s_1, s_{345}) - d(s_2, s_{345})}{2} \]

\[ d(s_1, s_{345}) = \frac{d(s_1, s_3) + d(s_1, s_4) + d(s_1, s_5)}{3} \]

\[ d(s_2, s_{345}) = \frac{d(s_2, s_3) + d(s_2, s_4) + d(s_2, s_5)}{3} \]

\[ L_{2Y_1} = \frac{d(s_2, s_1) + d(s_2, s_{345}) - d(s_1, s_{345})}{2} \]

Compute \( L_{(12)Y_2} \) and \( L_{3Y_2} \) using the same way.

\[ \because L_{(12)Y_2} = \frac{L_{1Y_1} + L_{Y_1Y_2} + L_{2Y_1} + L_{Y_1Y_2}}{2} \]

\[ = L_{Y_1Y_2} + \frac{L_{1Y_1} + L_{2Y_1}}{2} = L_{Y_1Y_2} + \frac{d(s_1, s_2)}{2} \]

\[ \therefore L_{Y_1Y_2} = L_{(12)Y_2} - \frac{d(s_1, s_2)}{2} \]
**Unrooted minisize: approximation**

How to design a 2-approximation algorithm for the unrooted minisize evolution tree problem?

<table>
<thead>
<tr>
<th>$d(\cdot, \cdot)$</th>
<th>$s_1$</th>
<th>$s_2$</th>
<th>$s_3$</th>
<th>$s_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>$s_2$</td>
<td></td>
<td>0</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>$s_3$</td>
<td></td>
<td></td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$s_4$</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Construct a minimal spanning tree out of the distance matrix

<table>
<thead>
<tr>
<th>$d(\cdot, \cdot)$</th>
<th>$s_1$</th>
<th>$s_2$</th>
<th>$s_3$</th>
<th>$s_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>$s_2$</td>
<td></td>
<td>0</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>$s_3$</td>
<td></td>
<td></td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$s_4$</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Unrooted minisize: approximation

Conduct a breadth first search on the minimal spanning tree by choosing an arbitrary node as root.

Start from the root and visit all of the first level descendants first, then the second level descendants, and so on.

Transform the minimal spanning tree into an evolutionary tree by adding nodes one by one.
Unrooted minisize: approximation

The created tree above is indeed an evolution tree:

- Each species is at leaf.
- The degree of each internal node is three.
- For any two species $s_i$ and $s_j$, we have

\[ d_T(s_i, s_j) \geq d(s_i, s_j) \]

because the distance $d_T(s_i, s_j)$ between any two species $s_i$ and $s_j$ on the evolution tree is exactly the same as that on the minimal spanning tree, which is then $\geq d(s_i, s_j)$ by triangular inequality.

How to prove $APP < 2 \cdot |OPT|$?

- $APP$: the length of our approximate solution
- $OPT$: an optimal solution

1. $APP = |MST| < |TSP|
   - $MST$: the minimal spanning tree of distance matrix
   - $TSP$: the optimal solution of the travelling salesman problem for distance matrix

2. $|TSP| \leq 2 \cdot |OPT|$
Unrooted minisize: approximation

- \( APP = |MST| < |TSP| \)
- By construction, \( APP = |MST| \)
- TSP: find a Hamiltonian cycle (visiting all of the nodes exactly once) with minimum length
- Deleting any edge from TSP obtains a spanning tree

\[
\begin{array}{cccc}
\text{s}_1 & 6 & 5 & \text{s}_2 \\
2 & 4 & 3 & 4 \\
\text{s}_3 & 4 & \text{s}_4 & \text{s}_1
\end{array}
\]

Unrooted minisize: approximation

- \(|TSP| \leq 2 \cdot |OPT|\)
- \(ET\): an Euler tour (visiting all of edges exactly once) of \(OPT\) \(\left( |ET| = 2 \cdot |OPT| \right)\)
- \(CET\): the cycle of species corresponding to \(ET\) \(\left( |TSP| \leq |CET| \Rightarrow CET\ is a Hamilt. cycle \right)\)
- \(|CET| \leq |ET|\) since \(d_T(s_i, s_j) \geq d(s_i, s_j)\)

\[
\begin{array}{cccc}
\text{s}_3 & \text{s}_{23} & \text{s}_2 & \text{s}_{12} \\
\text{s}_{34} & \text{s}_{41} & \text{s}_1 & \text{s}_4
\end{array}
\]