Lecture 23. Phylogeny methods, part 3 (Distance methods)

Joe Felsenstein

Department of Genome Sciences and Department of Biology
A phylogeny with branch lengths

and the pairwise distances it predicts
A phylogeny with branch lengths
Least squares trees

Least squares methods minimize

\[ Q = \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} (D_{ij} - d_{ij})^2 \]

over all trees, using the distances \( d_{ij} \) that they predict. Cavalli-Sforza and Edwards suggested \( w_{ij} = 1 \), Fitch and Margoliash suggested \( w_{ij} = 1/D_{ij}^2 \).
Statistical assumptions of least squares trees

Implicit assumption is that distances are (independently?) Normally distributed with expectation $d_{ij}$ and variance proportional to $1/w_{ij}^2$:

$$D_{ij} \sim \mathcal{N}(d_{ij}, K/w_{ij})$$

Thus the different weightings correspond to different assumptions about the error in the distances. Also, there is assumed to be no covariance of distances.

In fact, the distances will covary, since a change in an interior branch of the tree increases (or decreases) all distances whose paths go through that branch.
Matrix approach to fitting branch lengths

If we stack the distances up into a column vector $D$, we can solve the least squares equation (obtained by taking derivatives of the quadratic form $Q$):

$$D^T = \begin{pmatrix} D_{12}, D_{13}, D_{14}, D_{15}, D_{23}, D_{24}, D_{25}, D_{34}, D_{35}, D_{45} \end{pmatrix}$$

$$X^T D = (X^T X) v.$$

where the “design matrix" $X$ is has 1’s whenever a given branch lies on the path for the given distance.

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$
The Jukes-Cantor model for DNA
The distance for the Jukes-Cantor model

- Differences per site:
  - 0
  - 0.49
  - 0.75
  - 1

- Branch length:
  - 0.7945
If you don’t correct for “multiple hits"

Left: the true tree. Right: a tree fitting the uncorrected distances
Approximate variances for distances
under the Jukes-Cantor model

Distance as a function of fraction of nucleotide differences is

\[ \hat{t} = -\frac{3}{4} \ln \left( 1 - \frac{4}{3}D \right) \]

The “delta method" approximates the variance of one as a function of the variance of the other:

\[ \text{Var}(\hat{t}) \approx \left( \frac{\partial \hat{t}}{\partial D} \right)^2 \text{Var}(D) \]
Approximate variances, continued

The variance of fraction of nucleotide difference with \( n \) sites is the binomial variance

\[
\text{Var}(D) = \frac{D(1 - D)}{n}
\]

and since

\[
\frac{\partial \hat{t}}{\partial D} = \frac{1}{1 - \frac{4}{3}D}
\]

we get

\[
\text{Var}(D) \approx \frac{D(1 - D)/n}{(1 - \frac{4}{3}D)^2}
\]
Standard deviation of distance
as it increases with distance (given the JC model)
The UPGMA algorithm

1. Choose the smallest of the $D_{ij}$
2. make a new “tip” (ij)
3. Have i and j connected to this new tip, by a node whose “time” ago in branch length units is $D_{ij}/2$.
4. Have the weight of the new tip be $w_{(ij)} = w_i + w_j$
5. For each other tip, aside from i and j, compute

$$D_{(ij),k} = D_{k,(ij)} = \frac{w_iD_{ik} + w_jD_{jk}}{w_i + w_j}$$

6. Delete the rows and columns of the $D$ matrix for i and j.
7. If only one row left, stop, else return to step 1.

This can be done in $O(n^2)$ time if you save minimum elements of each row.
Sarich’s (1969) immunological distances

with columns and rows corresponding to the smallest distance highlighted.

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th>bear</th>
<th>raccoon</th>
<th>weasel</th>
<th>seal</th>
<th>sea lion</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0</td>
<td>32</td>
<td>48</td>
<td>51</td>
<td>50</td>
<td>48</td>
<td>98</td>
<td>148</td>
</tr>
<tr>
<td>bear</td>
<td>32</td>
<td>0</td>
<td>26</td>
<td>34</td>
<td>29</td>
<td>33</td>
<td>84</td>
<td>136</td>
</tr>
<tr>
<td>raccoon</td>
<td>48</td>
<td>26</td>
<td>0</td>
<td>42</td>
<td>44</td>
<td>44</td>
<td>92</td>
<td>152</td>
</tr>
<tr>
<td>weasel</td>
<td>51</td>
<td>34</td>
<td>42</td>
<td>0</td>
<td>44</td>
<td>38</td>
<td>86</td>
<td>142</td>
</tr>
<tr>
<td>seal</td>
<td>50</td>
<td>29</td>
<td>44</td>
<td>44</td>
<td>0</td>
<td>24</td>
<td>89</td>
<td>142</td>
</tr>
<tr>
<td>sea lion</td>
<td>48</td>
<td>33</td>
<td>44</td>
<td>38</td>
<td>24</td>
<td>0</td>
<td>90</td>
<td>142</td>
</tr>
<tr>
<td>cat</td>
<td>98</td>
<td>84</td>
<td>92</td>
<td>86</td>
<td>89</td>
<td>90</td>
<td>0</td>
<td>148</td>
</tr>
<tr>
<td>monkey</td>
<td>148</td>
<td>136</td>
<td>152</td>
<td>142</td>
<td>142</td>
<td>142</td>
<td>148</td>
<td>0</td>
</tr>
</tbody>
</table>
UPGMA tree for Sarich (1969) data

Lecture 23. Phylogeny methods, part 3 (Distance methods) – p.15/23
An unclocklike tree (left), the distances from it (center) and the UPGMA tree from those distances (right)
The distortion of the tree is due to “short-branch attraction" in which B and C, close to each other in the true tree, cluster first.
Neighbor-joining algorithm

1. For each tip, compute \( u_i = \sum_{j \neq i}^n D_{ij} / (n - 2) \)

2. Choose the \( i \) and \( j \) for which \( D_{ij} - u_i - u_j \) is smallest.

3. Join items \( i \) and \( j \). Compute the branch length from \( i \) to the new node \((v_i)\) and from \( j \) to the new node \((v_j)\) as

\[
\begin{align*}
v_i &= \frac{1}{2} D_{ij} + \frac{1}{2} (u_i - u_j) \\
v_j &= \frac{1}{2} D_{ij} + \frac{1}{2} (u_j - u_i)
\end{align*}
\]

4. compute the distance between the new node \((ij)\) and each other tip as

\[
D_{(ij),k} = \frac{D_{ik} + D_{jk} - D_{ij}}{2}
\]

5. delete tips \( i \) and \( j \) from the tables and replace them by the new node, \((ij),\) which is now treated as a tip.

6. If more than two nodes remain, go back to step 1. Otherwise connect the two remaining nodes by a branch of length \( D_{ij} \).
"Star decomposition" tree search method used in Neighbor-Joining method
Neighbor-joining tree for the Sarich (1969) immunological distance data
References, page 1


Farris, J. S. 1986. Distances and statistics. *Cladistics* **2**: 144-157. [debate was cut off after this]
References, page 1


How it was done

This projection produced

- using the \texttt{prosper} style in \LaTeX, 
- using \LaTeX\ to make a \texttt{.dvi} file, 
- using \texttt{dvips} to turn this into a Postscript file, 
- using \texttt{ps2pdf} to mill it into a PDF file, and 
- displaying the slides in Adobe Acrobat Reader.

Result: nice slides using freeware.