Step 3. Construction of the phylogenetic tree

Distance methods

Character methods
  Maximum parsimony
  Maximum likelihood
Distance methods

Simplest distance measure:

Consider every pair of sequences in the multiple alignment and count the number of differences.

Degree of divergence = Hamming distance (D)

\[
D = \frac{n}{N}
\]

where \( N \) = alignment length

\( n \) = number of sites with differences

Example:

\[
\begin{align*}
\text{AGGCTTTTCA} \\
\text{AGCCTTCTCA}
\end{align*}
\]

\( D = \frac{2}{10} = 0.2 \)
Problem with distance measure:

As the distance between two sequences increases, the probability increases that more than one mutation has occurred at any one site.

<table>
<thead>
<tr>
<th>time point</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>scenario 1</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>scenario 2</td>
<td>A</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>scenario 3</td>
<td>A</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>scenario 4</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
</tbody>
</table>

Therefore, methods have been developed to compensate for this.
Corrected distances

1. Jukes and Cantor

\[ d_{AB} = \frac{3}{4} \ln \left( 1 - \frac{4}{3} f_{AB} \right) \]

2. Kimura two parameter model

rate of transitions is different from rate of transversions

\[ d_{AB} = -\frac{1}{2} \ln \left[ (1 - 2P - Q) \sqrt{1 - 2Q} \right] \]

P = the fraction of sequence positions differing by a transition
Q = the fraction of sequence positions differing by a transversion.
In general, transitions more common than transversions
Distance methods

UPGMA (unweighted pair group method with arithmetic mean)

Neighbor-joining
UPGMA and the effect of unequal rates of evolution
Errors in tree topology may be remedied

* Transformed distance matrix.

This is used in *neighbor joining*. 
Neighbor joining is also different from UPGMA in that it uses the *star decomposition* method.
Procedure of neighbor joining

1. Given a matrix of pairwise distances \((d)\), for each terminal node \(i\) calculate its net divergence \((r_i)\) from all other taxa using the formula

\[
    r_i = \sum_{k=1}^{N} d_{ik}
\]

where \(N\) is the number of terminal nodes in the current matrix. Note the assumption that \(d_{ii} = 0\), otherwise the summation would need to skip over \(k = i\).

2. Create a rate-corrected distance matrix \((M)\) in which the elements are defined by

\[
    M_{ij} = d_{ij} - (r_i + r_j) / (N - 2)
\]

for all \(i\) and with \(j > i\) (the matrix is symmetrical, and the case of \(i = j\) is not interesting). Only the values \(i\) and \(j\) for which \(M_{ij}\) is minimum need be recorded; saving the entire matrix is unnecessary.

3. Define a new node \(u\) whose three branches join nodes \(i\), \(j\), and the rest of the tree. Define the lengths of the tree branches from \(u\) to \(i\) and \(j\):

\[
    v_{iu} = d_{ij} / 2 + (r_i - r_j) / [2(N - 2)]
\]

\[
    v_{ju} = d_{ij} - v_{iu}
\]

4. Define the distance from \(u\) to each other terminal node (for all \(k \neq i\) or \(j\))

\[
    d_{ku} = (d_{ik} + d_{jk} - d_{ij}) / 2
\]

5. Remove distances to nodes \(i\) and \(j\) from the data matrix, and decrease \(N\) by 1.

6. If more than two nodes remain, go back to step 1. Otherwise, the tree is fully defined except for the length of the branch joining the two remaining nodes \((i\) and \(j))\). Let this remaining branch be

\[
    v_{ij} = d_{ij}
\]

Each step has generated one internal node and has estimated the lengths of two of the branches connected to that node. The tree can be drawn from these data.
<table>
<thead>
<tr>
<th></th>
<th>Bsu</th>
<th>Bst</th>
<th>Lvi</th>
<th>Amo</th>
<th>Mlu</th>
<th>R</th>
<th>R/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bsu</td>
<td>—</td>
<td>0.1715</td>
<td>0.2147</td>
<td>0.3091</td>
<td>0.2326</td>
<td>0.9279</td>
<td>0.3093</td>
</tr>
<tr>
<td>Bst</td>
<td>—</td>
<td>—</td>
<td>0.2991</td>
<td>0.3399</td>
<td>0.2058</td>
<td>1.0163</td>
<td>0.3388</td>
</tr>
<tr>
<td>Lvi</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.2795</td>
<td>0.3943</td>
<td>1.1876</td>
<td>0.3959</td>
</tr>
<tr>
<td>Amo</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.4289</td>
<td>1.3574</td>
<td>0.4525</td>
</tr>
<tr>
<td>Mlu</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.2616</td>
<td>0.4205</td>
</tr>
</tbody>
</table>

Lvi to node 1 distance = 0.2795/2 + (0.3959 - 0.4525)/2 = 0.1114
Amo to node 1 distance = 0.2795 - 0.1114 = 0.1681

<table>
<thead>
<tr>
<th></th>
<th>Bsu</th>
<th>Bst</th>
<th>Mlu</th>
<th>Node 1</th>
<th>R</th>
<th>R/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bsu</td>
<td>—</td>
<td>0.1715</td>
<td>0.2326</td>
<td>0.1222</td>
<td>0.5263</td>
<td>0.2631</td>
</tr>
<tr>
<td>Bst</td>
<td>—</td>
<td>—</td>
<td>0.2058</td>
<td>0.1798</td>
<td>0.5571</td>
<td>0.2785</td>
</tr>
<tr>
<td>Mlu</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.2719</td>
<td>0.7103</td>
<td>0.3551</td>
</tr>
<tr>
<td>Node 1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.5739</td>
<td>0.2869</td>
<td></td>
</tr>
</tbody>
</table>

Bsu to node 2 distance = 0.1222/2 + (0.2631 - 0.2869)/2 = 0.0492
node 1 to node 2 distance = 0.1222 - 0.0492 = 0.0730

<table>
<thead>
<tr>
<th></th>
<th>Bst</th>
<th>Mlu</th>
<th>Node 2</th>
<th>R</th>
<th>R/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bst</td>
<td>—</td>
<td>0.2058</td>
<td>0.1146</td>
<td>0.3204</td>
<td>0.3204</td>
</tr>
<tr>
<td>Mlu</td>
<td>—</td>
<td>—</td>
<td>0.1912</td>
<td>0.3970</td>
<td>0.3970</td>
</tr>
<tr>
<td>Node 2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.3058</td>
<td>0.3058</td>
</tr>
</tbody>
</table>

Bst to node 3 distance = 0.1146/2 + (0.3204 - 0.3058)/2 = 0.0646
node 2 to node 3 distance = 0.1146 - 0.0646 = 0.0500

<table>
<thead>
<tr>
<th></th>
<th>Mlu</th>
<th>Node 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mlu</td>
<td>—</td>
<td>0.1412</td>
</tr>
<tr>
<td>Node 3</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Mlu to node 3 distance = 0.1412

**Figure 30** Neighbor joining of 5S rRNA evolutionary distance estimates. The data and abbreviations are as in Figure 29. Each table presents the pairwise distance values input to the round of analysis (upper right half of the matrix). The rightmost two columns present the row totals for the uncorrected distances (the row being defined based on the full symmetrical matrix; see equation 41) and the total divided by the number of terminal nodes minus two. The rate-corrected pairwise distances as defined by equation (42) are given in the lower left half of the matrix. The minimum corrected distance value in each table and the corresponding uncorrected pairwise distance are shown in bold. The corresponding pair of taxa (or clusters) are removed from the matrix and replaced by their common ancestral node in the next table and distances based on equation (43). The inferred tree is diagrammed in Figure 15A.
Neighbor joining creates an *unrooted* tree
Character-based methods

Maximum parsimony
Maximum likelihood
Maximum parsimony

\textit{parsimony} - principle in science where the simplest answer is the preferred.

In phylogeny: The preferred phylogenetic tree is the one that requires the fewest evolutionary steps.
Maximum parsimony

1. Identify all informative sites in the multiple alignment

2. For each possible tree, calculate the number of changes at each informative site.

3. Sum the number of changes for each possible tree.

4. Tree with the smallest number of changes is selected as the most likely tree.
Maximum parsimony
Identify informative sites

<table>
<thead>
<tr>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A A G A G T G C A</td>
</tr>
<tr>
<td>2</td>
<td>A G C C G T G C G</td>
</tr>
<tr>
<td>3</td>
<td>A G A T A T C C A</td>
</tr>
<tr>
<td>4</td>
<td>A G A G A T C C G</td>
</tr>
</tbody>
</table>

* * *
Site 3 - non informative

1-G  
\[ \text{G} \quad \text{A} \quad \text{A-3} \]

2-C  
\[ \text{*} \quad \text{A} \quad \text{A-4} \]

1-G  
\[ \text{A} \quad \text{A} \quad \text{C-2} \]

3-A  
\[ \text{A} \quad \text{A} \quad \text{A-4} \]

4-A  
\[ \text{A} \quad \text{A} \quad \text{A-3} \]
Site 5 - informative

1-G  A-3
   G   *
2-G  A-4

1-G  G-2
   *   *
A   A
3-A  A-4

1-G  G-2
   *   *
A   A
4-A  A-3
Summing changes:

<table>
<thead>
<tr>
<th></th>
<th>site 5</th>
<th>site 7</th>
<th>site 9</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree I</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Tree II</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Tree III</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

=> Tree I most likely.
Step 4. Evaluate the tree - Bootstrapping
(from www.icp.ucl.ac.be/~opperd/private/bootstrap.html)

Bootstrapping is a way of testing the reliability of the dataset and the tree, allows you to assess whether the distribution of characters has been influenced by stochastic effects.

**Bootstrapping in practice**

Take a dataset consisting of in total $n$ sequences with $m$ sites each. A number of resampled datasets of the same size ($n \times m$) as the original dataset is produced. However, each site is sampled at random and no more sites are sampled than there were original sites.
Sample 1 0 1 2 0 3 0 1 2 0 1 (← number of times each site is sampled)

A  A G G C U C C A A A  A  G G G U U U C A A A
B  A G G U U C G A A A  B  G G G U U U G A A A
C  A G C C C C G A A A  C  G C C C C C G A A A
D  A U U U C C G A A C  D  U U U C C C G A A C

Sample 2 1 0 0 0 2 2 2 0 0 3

A  A G G C U C C A A A  A  A U U C C C C C A A A
B  A G G U U C G A A A  B  A U U C C G G A A A
C  A G C C C C G A A A  C  A C C C C C G A A A
D  A U U U C C G A A C  D  A C C C C C G G C C C

---

A B C
B 1
C 6 5
D 8 7 4
Consensus tree.

The number of times each branch point or node occurred
(bootstrap proportion) is indicated at each node.
Bootstrapping typically involves 100-1000 datasets. Bootstrap values > 70% are generally considered to provide support for the clade designation.
Software for phylogenetic analysis

**PHYLIP** (Phylogenetic Inference Package)
*Joe Felsenstein*
Command line
WebPhylip

Examples in practical
- **DNADIST** = create a distance matrix
- **NEIGHBOR** = neighbor joining / UPGMA
- **DNAPARS** = maximum parsimony

**PAUP** (Phylogenetic Analysis Using Parsimony)
Exercises in molecular phylogeny

* What animal is most closely related to the extinct quagga?

Clustalw alignment. Neighbor joining is used in the clustalw progressive alignment method.
Exercises in molecular phylogeny

Is the south american opossum evolutionarily related to the australian ‘marsupial wolf’?
DNA analysis of Neanderthal individuals
neighbor joining

maximum parsimony
AIDS epidemic and the evolution of HIV

$SIV_{cpz} \rightarrow HIV-1 \rightarrow HIV-1 \text{ M group}$

$SIV_{sm} \rightarrow HIV-2$
Phylogenetic relationships between HIV and SIV viruses.
Sharp et al, Phil. Trans. R. Soc. Lond (2001) 356, 867-876
From: Worobey M, et al

Origin of AIDS: contaminated polio vaccine theory refuted.


Phylogenetic tree reconstruction using *maximum likelihood*
Phylogenetic analysis may be used to identify horizontal gene transfer.
Some Chlamydia (eubacterium) proteins cluster with plant homologs.
Phylogenetic analysis may be used to identify horizontal gene transfer.

Aquifex aeolicus (of \textit{Eubacteria}) has a large number of genes that seem to originate from \textit{Archeae}.

\textbf{Figure 1.} Genes of apparent archaeal origin in the genome of \textit{Aquifex aeolicus}. Yellow circles represent genes encoding proteins with reliable best hits to archaeal homologs. Gene clusters conserved in \textit{Aquifex} and archaean are boxed. The largest cluster contains genes for a predicted RNA helicase, a nuclease and a zinc-finger-containing nucleic acid-binding protein; the remaining genes encode uncharacterized proteins, most of which are conserved in archaean and \textit{Aquifex} only.