Master Course
Sequence Alignment

Lecture 13

Evolution/Phylogeny
“Nothing in Biology makes sense except in the light of evolution” (Theodosius Dobzhansky (1900-1975))

“Nothing in bioinformatics makes sense except in the light of Biology (and thus evolution)”
Content

• Evolution
  – requirements
  – negative/positive selection on genes (e.g. Ka/Ks)
  – homology/paralogy/orthology (operational definition ‘bi-directional best hit’)

• Clustering
  – single linkage
  – complete linkage

• Phylogenetic trees
  – ultrametric distance (uniform molecular clock)
  – additive trees (4-point condition)
  – UPGMA algorithm
  – NJ algorithm
  – bootstrapping
Darwinian Evolution

What is needed:
1. Template (DNA)
2. Copying mechanism (meiosis/fertilisation)
3. Variation (e.g. resulting from copying errors, gene conversion, crossing over, genetic drift, etc.)
4. Selection
DNA evolution

- Gene nucleotide substitutions can be synonymous (i.e. not changing the encoded amino acid) or nonsynonymous (i.e. changing the a.a.).
- Rates of evolution vary tremendously among protein-coding genes. Molecular evolutionary studies have revealed an ~1000-fold range of nonsynonymous substitution rates (Li and Graur 1991).
- The strength of negative (purifying) selection is thought to be the most important factor in determining the rate of evolution for the protein-coding regions of a gene (Kimura 1983; Ohta 1992; Li 1997).
DNA evolution

• “Essential” and “nonessential” are classic molecular genetic designations relating to organismal fitness.
  – A gene is considered to be essential if a knock-out results in (conditional) lethality or infertility.
  – Nonessential genes are those for which knock-outs yield viable and fertile individuals.

• Given the role of purifying selection in determining evolutionary rates, the greater levels of purifying selection on essential genes leads to a lower rate of evolution relative to that of nonessential genes.
Ka/Ks Ratios

- **Ks** is defined as the number of synonymous nucleotide substitutions per synonymous site
- **Ka** is defined as the number of nonsynonymous nucleotide substitutions per nonsynonymous site
- The **Ka/Ks** ratio is used to estimate the type of selection exerted on a given gene or DNA fragment
- Need orthologous nucleotide sequence alignments
- Observe nucleotide substitution patterns at given sites and correct numbers using, for example, the Pamilo-Bianchi-Li method (Li 1993; Pamilo and Bianchi 1993).
- Correction is needed because of the following:
  
  Consider the codons specifying aspartic acid and lysine: both start AA, lysine ends A or G, and aspartic acid ends T or C. So, if the rate at which C changes to T is higher from the rate that C changes to G or A (as is often the case), then more of the changes at the third position will be synonymous than might be expected. Many of the methods to calculate Ka and Ks differ in the way they make the correction needed to take account of this type of bias.
The frequency of different values of \( \frac{Ka}{Ks} \) for 835 mouse–rat orthologous genes. Figures on the \( x \) axis represent the middle figure of each bin; that is, the 0.05 bin collects data from 0 to 0.1.
Ka/Ks ratios

Three types of selection:

1. Negative (purifying) selection $\rightarrow$ $\text{Ka/Ks} < 1$
2. Neutral selection (Kimura) $\rightarrow$ $\text{Ka/Ks} \sim 1$
3. Positive selection $\rightarrow$ $\text{Ka/Ks} > 1$
Orthologous genes are homologous (corresponding) genes in different species.

Paralogous genes are homologous genes within the same species (genome).
Orthology/paralogy

Operational definition of orthology:

**Bi-directional best hit:**
- Blast gene A in genome 1 against genome 2: gene B is best hit
- Blast gene B against genome 1: if gene A is best hit
  → A and B are orthologous

A number of other criteria is also in use (part of which is based on phylogeny)
Xenology

- **Xenologs** are homologs resulting from the horizontal transfer of a gene between two organisms.
- The function of xenologs can be variable, depending on how significant the change in context was for the horizontally moving gene. In general, though, the function tends to be similar (before and after horizontal transfer).
Multivariate statistics – Cluster analysis

Raw table

Similarity criterion

Scores

Cluster criterion

Dendrogram

Any set of numbers per column

5×5

C1  C2  C3  C4  C5  C6  ..

1  2  3  4  5
Multivariate statistics – Cluster analysis

Why do it?

• Finding a true typology
• Model fitting
• Prediction based on groups
• Hypothesis testing
• Data exploration
• Data reduction
• Hypothesis generation

But you can never prove a classification/typology!
Cluster analysis – data normalisation/weighting

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>..</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Raw table

Normalisation criterion

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>..</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Normalised table

Column normalisation: \( x/max \)

Column range normalise: \( (x-min)/(max-min) \)
Cluster analysis – (dis)similarity matrix

\[ D_{i,j} = (\sum_k |x_{ik} - x_{jk}|^r)^{1/r} \quad \text{Minkowski metrics} \]

- \( r = 2 \) *Euclidean distance*
- \( r = 1 \) *City block distance*
Cluster analysis – Clustering criteria

Scores

5x5

Similarity matrix

Cluster criterion

Dendrogram (tree)

Single linkage - Nearest neighbour

Complete linkage – Furthest neighbour

Group averaging – UPGMA

Ward

Neighbour joining – global measure
Comparing sequences
- Similarity Score -

Many properties can be used:

- Nucleotide or amino acid composition
- Isoelectric point
- Molecular weight
- Morphological characters

- But: molecular evolution through sequence alignment
Multivariate statistics
Producing a Phylogenetic tree from sequences

Multiple sequence alignment

Scores

Similarity criterion

Distance matrix

Cluster criterion

Phylogenetic tree

1
2
3
4
5

5x5
Evolution

• Most of bioinformatics is comparative biology
• Comparative biology is based upon evolutionary relationships between compared entities
• Evolutionary relationships are normally depicted in a phylogenetic tree
Where can phylogeny be used

- For example, finding out about orthology versus paralogy
- Predicting secondary structure of RNA
- Studying host-parasite relationships
- Mapping cell-bound receptors onto their binding ligands
- Multiple sequence alignment (e.g. Clustal)
Similarity criterion for phylogeny

• ClustalW: uses sequence identity with Kimura (1983) correction:
  Corrected K = - ln(1.0-K-K^2/5.0), where K is percentage divergence corresponding to two aligned sequences

• There are various models to correct for the fact that the true rate of evolution cannot be observed through nucleotide (or amino acid) exchange patterns (e.g. back mutations)

• Saturation level is ~94%, higher real mutations are no longer observable
### Lactate dehydrogenase multiple alignment

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>KITVGVGAVGMAISILMKDLADEDELAVDVEDKLKGMEMDLQHGSFLRLRTPKIVSGKDYNTANSKLVIITAGARQ</td>
</tr>
<tr>
<td>Chicken</td>
<td>KISVGVGAVGMAISILMKDLADELETLDVDVEDKLKGMEMDLQHGSFLKLTPKITSGKDYVTASHKLVIITAGARQ</td>
</tr>
<tr>
<td>Dogfish</td>
<td>KITVGVGAVGMAISILMKDLADEVALVDMEDKLKGMEMDLQHGSFLHTAKIVSGKDYVSAGSLKLVITAGARQ</td>
</tr>
<tr>
<td>Lamprey</td>
<td>SKVTIVGVGVMAAISVLLRLDADELAVDVEDRLKGMEMDLHLHGSFLKTAIVADKYSVTAHSLKLVITAGARQ</td>
</tr>
<tr>
<td>Barley</td>
<td>TKISVGDAGAVSSYAYAMLQIAEQEGSIDFEDKTKGDAIDLSNAPFTSKPFIYSAYA-EYSDAKDADLVITAGARQ</td>
</tr>
<tr>
<td>Maizey_casei</td>
<td>KVILVGDGAVGSSYAYAMLQIAQEGSIDFEDKTKGDAIDLSNAPFTSKPFIYSAYA-EYSDAKDADLVITAGARQ</td>
</tr>
<tr>
<td>Bacillus</td>
<td>TKVSVGVGVGMAISILMKDLADEAVDVEDKLKGMEMDLQHGSFLKTAIVADKYSVTAGRSLKLVIITAGARQ</td>
</tr>
<tr>
<td>Lacto__ste</td>
<td>RVVVIGAVGAGFVGSFALMNQGIAAEIVLIDANESKAIGDAMFNGKVFAPKVDIFWGDYDDCRDADLVICAGANQ</td>
</tr>
<tr>
<td>Lacto_plant</td>
<td>QKVVLGAVGSSYAFAMAQQGIAEEFVIVDDVKKRTGDAIDLAQQTAPKIKYSG-EYSDCKDADLVITAGARQ</td>
</tr>
<tr>
<td>Therma_mari</td>
<td>MKIGIVGLGRVGSFALLMKGFAREMLIDVKKRAEDGALDLHITFTRIANIYAG-DYADLKGSDVIVAAGVPQ</td>
</tr>
<tr>
<td>Bifido</td>
<td>KLVAGAGVSTLFAAAQRGIAEIVLEDIAKERVEAEVLMQHGSFYPTVSIDGDDEIECRDADMVITAGPRQ</td>
</tr>
<tr>
<td>Thermus_aqua</td>
<td>MKVIGVGSFVGSATAYALVLQGVAREVVLDLDRKLAQAHAEDILHATPFHAPVWVRSGW-YEDLEGRVVIVAAGVAQ</td>
</tr>
<tr>
<td>Mycoplasma</td>
<td>KIALIGAGNVNSFLYAMNQGLASEYGIIDINPDAGNADFEDASLFPISVSYEYKDLKADIVITAGRPQ</td>
</tr>
</tbody>
</table>

### Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.112</td>
<td>0.128</td>
<td>0.202</td>
<td>0.378</td>
<td>0.346</td>
<td>0.530</td>
<td>0.551</td>
<td>0.512</td>
<td>0.524</td>
<td>0.528</td>
<td>0.635</td>
<td>0.637</td>
</tr>
<tr>
<td>2</td>
<td>0.112</td>
<td>0.000</td>
<td>0.155</td>
<td>0.214</td>
<td>0.382</td>
<td>0.348</td>
<td>0.538</td>
<td>0.569</td>
<td>0.516</td>
<td>0.524</td>
<td>0.524</td>
<td>0.631</td>
<td>0.651</td>
</tr>
<tr>
<td>3</td>
<td>0.128</td>
<td>0.155</td>
<td>0.000</td>
<td>0.196</td>
<td>0.389</td>
<td>0.337</td>
<td>0.522</td>
<td>0.567</td>
<td>0.516</td>
<td>0.512</td>
<td>0.524</td>
<td>0.600</td>
<td>0.655</td>
</tr>
<tr>
<td>4</td>
<td>0.202</td>
<td>0.214</td>
<td>0.196</td>
<td>0.000</td>
<td>0.426</td>
<td>0.356</td>
<td>0.553</td>
<td>0.589</td>
<td>0.544</td>
<td>0.503</td>
<td>0.544</td>
<td>0.616</td>
<td>0.669</td>
</tr>
<tr>
<td>5</td>
<td>0.378</td>
<td>0.382</td>
<td>0.389</td>
<td>0.426</td>
<td>0.000</td>
<td>0.171</td>
<td>0.536</td>
<td>0.565</td>
<td>0.526</td>
<td>0.547</td>
<td>0.516</td>
<td>0.629</td>
<td>0.575</td>
</tr>
<tr>
<td>6</td>
<td>0.346</td>
<td>0.348</td>
<td>0.337</td>
<td>0.356</td>
<td>0.171</td>
<td>0.000</td>
<td>0.557</td>
<td>0.563</td>
<td>0.538</td>
<td>0.555</td>
<td>0.518</td>
<td>0.643</td>
<td>0.587</td>
</tr>
<tr>
<td>7</td>
<td>0.530</td>
<td>0.538</td>
<td>0.522</td>
<td>0.533</td>
<td>0.536</td>
<td>0.557</td>
<td>0.000</td>
<td>0.518</td>
<td>0.208</td>
<td>0.445</td>
<td>0.561</td>
<td>0.526</td>
<td>0.501</td>
</tr>
<tr>
<td>8</td>
<td>0.551</td>
<td>0.569</td>
<td>0.567</td>
<td>0.589</td>
<td>0.565</td>
<td>0.563</td>
<td>0.518</td>
<td>0.000</td>
<td>0.477</td>
<td>0.536</td>
<td>0.536</td>
<td>0.598</td>
<td>0.495</td>
</tr>
<tr>
<td>9</td>
<td>0.512</td>
<td>0.516</td>
<td>0.516</td>
<td>0.544</td>
<td>0.526</td>
<td>0.538</td>
<td>0.208</td>
<td>0.477</td>
<td>0.000</td>
<td>0.433</td>
<td>0.489</td>
<td>0.563</td>
<td>0.485</td>
</tr>
<tr>
<td>10</td>
<td>0.524</td>
<td>0.524</td>
<td>0.512</td>
<td>0.503</td>
<td>0.547</td>
<td>0.555</td>
<td>0.445</td>
<td>0.536</td>
<td>0.000</td>
<td>0.532</td>
<td>0.405</td>
<td>0.598</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.528</td>
<td>0.524</td>
<td>0.524</td>
<td>0.544</td>
<td>0.516</td>
<td>0.518</td>
<td>0.561</td>
<td>0.536</td>
<td>0.489</td>
<td>0.532</td>
<td>0.000</td>
<td>0.604</td>
<td>0.614</td>
</tr>
<tr>
<td>12</td>
<td>0.635</td>
<td>0.631</td>
<td>0.600</td>
<td>0.616</td>
<td>0.629</td>
<td>0.643</td>
<td>0.526</td>
<td>0.598</td>
<td>0.563</td>
<td>0.405</td>
<td>0.604</td>
<td>0.000</td>
<td>0.641</td>
</tr>
<tr>
<td>13</td>
<td>0.637</td>
<td>0.651</td>
<td>0.655</td>
<td>0.669</td>
<td>0.575</td>
<td>0.587</td>
<td>0.501</td>
<td>0.495</td>
<td>0.485</td>
<td>0.598</td>
<td>0.614</td>
<td>0.641</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**How can you see that this is a distance matrix?**
Cluster analysis – Clustering criteria

Scores  

Similarity matrix  

Cluster criterion  

Dendrogram (tree)  

4×5

Four different clustering criteria:
Single linkage - Nearest neighbour
Complete linkage – Furthest neighbour
Group averaging – UPGMA
Neighbour joining (global measure)

Note: these are all *agglomerative* cluster techniques; i.e. they proceed by merging clusters as opposed to techniques that are *divisive* and proceed by cutting clusters
General agglomerative cluster protocol

1. Start with $N$ clusters of 1 object each

2. Apply clustering distance criterion and merge clusters iteratively until you have 1 cluster of $N$ objects

3. Most interesting clustering somewhere in between

Dendrogram (tree)

Note: a dendrogram can be rotated along branch points (like mobile in baby room) -- distances between objects are defined along branches
Single linkage clustering (nearest neighbour)
Single linkage clustering (nearest neighbour)
Single linkage clustering (nearest neighbour)
Single linkage clustering (nearest neighbour)
Single linkage clustering (nearest neighbour)
Single linkage clustering (nearest neighbour)

Distance from point to cluster is defined as the smallest distance between that point and any point in the cluster.
Single linkage clustering (nearest neighbour)

Distance from point to cluster is defined as the smallest distance between that point and any point in the cluster.
Single linkage clustering (nearest neighbour)

Let $C_i$ and $C_j$ be two disjoint clusters:

$$d_{i,j} = \min(d_{p,q}), \text{ where } p \in C_i \text{ and } q \in C_j$$

Single linkage dendrograms typically show **chaining** behaviour (i.e., all the time a single object is added to existing cluster)
Complete linkage clustering
(furthest neighbour)
Complete linkage clustering (furthest neighbour)
Complete linkage clustering (furthest neighbour)
Complete linkage clustering (furthest neighbour)
Complete linkage clustering
(furthest neighbour)
Complete linkage clustering (furthest neighbour)
Complete linkage clustering (furthest neighbour)
Complete linkage clustering (furthest neighbour)

Distance from point to cluster is defined as the largest distance between that point and any point in the cluster
Complete linkage clustering
(furthest neighbour)

Distance from point to cluster is defined as the largest distance between that point and any point in the cluster
Complete linkage clustering (furthest neighbour)

Let $C_i$ and $C_j$ be two disjoint clusters:

$$d_{i,j} = \text{Max}(d_{p,q}), \text{ where } p \in C_i \text{ and } q \in C_j$$

More ‘structured’ clusters than with single linkage clustering
Clustering algorithm

1. Initialise (dis)similarity matrix
2. Take two points with smallest distance as first cluster
3. Merge corresponding rows/columns in (dis)similarity matrix
4. Repeat steps 2. and 3. using appropriate cluster measure until last two clusters are merged
Phylogenetic trees

Multiple sequence alignment (MSA)

Scores

Similarity/Distance matrix

Similarity criterion

Cluster criterion

Phylogenetic tree

MSA quality is crucial for obtaining correct phylogenetic tree

1 2 3 4 5

5×5
Phylogenetic tree (unrooted)

Drosophila

fugu

human

mouse

internal node

edge

leaf

OTU – Observed taxonomic unit
Phylogenetic tree (unrooted)

- Root
- Internal node
- Edge
- OTU – Observed taxonomic unit

Species:
- Drosophila
- Fugu
- Human
- Mouse
Phylogenetic tree (rooted)

Clade - group of two or more taxa that includes both their common ancestor and all of their descendents.
How to root a tree

- Outgroup – place root between distant sequence and rest group

- Midpoint – place root at midpoint of longest path (sum of branches between any two OTUs)

- Gene duplication – place root between paralogous gene copies (see earlier globin example)
# Combinatorial explosion

<table>
<thead>
<tr>
<th># sequences</th>
<th># unrooted trees</th>
<th># rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
<td>10,395</td>
</tr>
<tr>
<td>8</td>
<td>10,395</td>
<td>135,135</td>
</tr>
<tr>
<td>9</td>
<td>135,135</td>
<td>2,027,025</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
<td>34,459,425</td>
</tr>
</tbody>
</table>
A simple clustering method for building phylogenetic trees

Unweighted Pair Group Method using Arithmetic Averages

(UPGMA)

Sneath and Sokal (1973)
Let \( C_i \) and \( C_j \) be two disjoint clusters:

\[
d_{i,j} = \frac{1}{|C_i| \times |C_j|} \sum_p \sum_q d_{p,q}, \text{ where } p \in C_i \text{ and } q \in C_j
\]
Clustering algorithm: UPGMA

Initialization:
- Fill distance matrix with pairwise distances
- Start with N clusters of 1 element each

Iteration:
1. Merge cluster \( C_i \) and \( C_j \) for which \( d_{ij} \) is minimal
2. Place internal node connecting \( C_i \) and \( C_j \) at height \( d_{ij}/2 \)
3. Delete \( C_i \) and \( C_j \) (keep internal node)

Termination:
- When two clusters i, j remain, place root of tree at height \( d_{ij}/2 \)
Ultrametric Distances

• A tree $T$ in a metric space $(M,d)$ where $d$ is ultrametric has the following property: there is a way to place a root on $T$ so that for all nodes in $M$, their distance to the root is the same. Such $T$ is referred to as a uniform molecular clock tree.

• $(M,d)$ is ultrametric if for every set of three elements $i,j,k \in M$, two of the distances coincide and are greater than or equal to the third one (see next slide).

• UPGMA is guaranteed to build correct tree if distances are ultrametric (single molecular clock). But it fails if not!
Ultrametric Distances

Given three leaves, two distances are equal while a third is smaller:

\[ d(i, j) \leq d(i, k) = d(j, k) \]

\[ a + a \leq a + b = a + b \]

Nodes i and j are at the same evolutionary distance from k – dendrogram will therefore have ‘aligned’ leaves; i.e. they are all at the same distance from root.
Evolutionary clock speeds

Uniform clock: **Ultrametric distances** lead to identical distances from root to leaves.

Non-uniform evolutionary clock: leaves have different distances to the root -- an important property is that of **additive trees**. These are trees where the distance between any pair of leaves is the sum of the lengths of edges connecting them. Such trees obey the so-called **4-point condition** (next slide).
Additive trees

All distances satisfy 4-point condition:

For all leaves $i, j, k, l$:

$$d(i,j) + d(k,l) \leq d(i,k) + d(j,l) = d(i,l) + d(j,k)$$

$$(a+b)+(c+d) \leq (a+m+c)+(b+m+d) = (a+m+d)+(b+m+c)$$

Result: all pairwise distances obtained by traversing the tree
Additive trees

In additive trees, the distance between any pair of leaves is the sum of lengths of edges connecting them.

Given a set of additive distances: a unique tree $T$ can be constructed:

- For two neighbouring leaves $i,j$ with common parent $k$, place parent node $k$ at a distance from any node $m$ with

$$d(k,m) = \frac{1}{2} (d(i,m) + d(j,m) - d(i,j))$$

$$c = \frac{1}{2} ((a+c) + (b+c) - (a+b))$$

$d$ is ultrametric $\implies d$ additive
Neighbour-Joining (Saitou and Nei, 1987)

- Guaranteed to produce correct tree if distances are additive
- May even produce good tree if distances are not additive
- Global measure – keeps total branch length minimal
- At each step, join two nodes such that distances are minimal (criterion of minimal evolution)
- Agglomerative algorithm
- Leads to \textit{unrooted} tree
Neighbour joining

At each step all possible ‘neighbour joinings’ are checked and the one corresponding to the minimal total tree length (calculated by adding all branch lengths) is taken.
Neighbour joining

Finding neighbouring leaves:

Define

\[ D_{ij} = d_{ij} - (r_i + r_j) \]

Where

\[ r_i = \frac{1}{|L| - 2} \sum_k d_{ik} \]

Total tree length \( D_{ij} \) is minimal iff \( i \) and \( j \) are neighbours

Proof in Durbin book, p. 189
Algorithm: Neighbour joining

Initialisation:

• Define T to be set of leaf nodes, one per sequence
• Let L = T

Iteration:

• Pick i, j (neighbours) such that $D_{i,j}$ is minimal (minimal total tree length) *[this does not mean that the OTU-pair with smallest distance is selected!]*

• Define new node k, and set $d_{km} = \frac{1}{2} (d_{im} + d_{jm} - d_{ij})$ for all $m \in L$
• Add k to T, with edges of length $d_{ik} = \frac{1}{2} (d_{ij} + r_i - r_j)$
• Remove i, j from L; Add k to L

Termination:

• When L consists of two nodes i, j and the edge between them of length $d_{ij}$
Tree distances

Evolutionary (sequence distance) = sequence dissimilarity

- human
- mouse
- fugu
- *Drosophila*

<table>
<thead>
<tr>
<th></th>
<th>human</th>
<th>mouse</th>
<th>fugu</th>
<th><em>Drosophila</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>x</td>
<td>6</td>
<td>x</td>
<td>14</td>
</tr>
<tr>
<td>mouse</td>
<td>x</td>
<td>7</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>fugu</td>
<td>x</td>
<td>x</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td><em>Drosophila</em></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
Three main classes of phylogenetic methods

• Distance based
  – uses pairwise distances (see earlier slides)
  – fastest approach

• Parsimony
  – fewest number of evolutionary events (mutations)
  – attempts to construct maximum parsimony tree

• Maximum likelihood
  – \( L = \text{Pr}[\text{Data} | \text{Tree}] \)
  – can use more elaborate and detailed evolutionary models
**Parsimony & Distance**

**Sequences**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Drosophila</strong></td>
<td>t</td>
<td>t</td>
<td>a</td>
<td>t</td>
<td>t</td>
<td>t</td>
<td>a</td>
</tr>
<tr>
<td><strong>fugu</strong></td>
<td>a</td>
<td>a</td>
<td>t</td>
<td>t</td>
<td>t</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td><strong>mouse</strong></td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>t</td>
<td>a</td>
</tr>
<tr>
<td><strong>human</strong></td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>t</td>
</tr>
</tbody>
</table>

**parsimony**

```
Drosophila
    /   \     |
   /     \    |
mouse   fugu |
```

**distance**

```
Drosophila
    /   \     |
   /     \    |
human   fugu  \\
```

**Diagram:**

- **Parsimony:** Branch lengths represent the number of mutations.
- **Distance:** Branch lengths are distances between taxa.

**Table:**

<table>
<thead>
<tr>
<th></th>
<th>human</th>
<th>mouse</th>
<th>fugu</th>
<th>Drosophila</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td><strong>x</strong></td>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td><strong>x</strong></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

**Note:** The table and diagram illustrate the concept of parsimony and distance in molecular biology, showing how sequence data can be used to infer evolutionary relationships.
Problem: Long Branch Attraction (LBA)

- Particular problem associated with parsimony methods
- Rapidly evolving taxa are placed together in a tree regardless of their true position
- Partly due to assumption in parsimony that all lineages evolve at the same rate
- This means that also UPGMA suffers from LBA
- Some evidence exists that also implicates NJ
Maximum likelihood
Pioneered by Joe Felsenstein

• If data=alignment, hypothesis = tree, and under a given evolutionary model, maximum likelihood selects the hypothesis (tree) that maximises the observed data

• A statistical (Bayesian) way of looking at this is that the tree with the largest posterior probability is calculated based on the prior probabilities; i.e. the evolutionary model (or observations).

• Extremely time consuming method

• We also can test the relative fit to the tree of different models (Huelsenbeck & Rannala, 1997)
Bayesian methods for maximum likelihood

- Calculate the posterior probability of a tree (Huelsenbeck et al., 2001) — probability that tree is true tree given evolutionary model

\[
\text{Pr}[\text{Tree}|\text{Data}] = \text{Pr}[\text{Tree}|\text{Data}] \times \frac{\text{Pr}[\text{Tree}]}{\text{Pr}[\text{Data}]}
\]

- Most computer intensive technique
- Feasible thanks to *Markov chain Monte Carlo* (MCMC) numerical technique for integrating over probability distributions
- Gives confidence number (posterior probability) per node
Maximum likelihood

Methods to calculate ML tree:
- **Phylip** (http://evolution.genetics.washington.edu/phylip.html)
- **Paup** (http://paup.csit.fsu.edu/index.html)
- **MrBayes** (http://mrbayes.csit.fsu.edu/index.php)

Method to analyse phylogenetic tree with ML:
- **PAML** (http://abacus.gene.ucl.ac.uk/software/paml.htm)

  The strength of PAML is its collection of sophisticated substitution models to analyse trees. Tree search algorithms implemented in Paml’s baseml and codeml are rather primitive, so except for very small data sets with say, <10 species, you are better off to use another package.

- Programs such as PAML can test the relative fit to the tree of different models (Huelsenbeck & Rannala, 1997)
Maximum likelihood

- A number of ML tree packages (e.g. Phylip, PAML) contain tree algorithms that include the assumption of a uniform molecular clock as well as algorithms that don’t
- These can both be run on a given tree, after which the results can be used to estimate the probability of a uniform clock.
How to assess confidence in tree
How to assess confidence in tree

• Distance method – *bootstrap*:
  – Select multiple alignment columns *with replacement*
  – Recalculate tree
  – Compare branches with original (target) tree
  – Repeat 100-1000 times, so calculate 100-1000 different trees
  – How often is branching (point between 3 nodes) preserved for each internal node?
  – Uses samples of the data
The Bootstrap -- example

Original

1 2 3 4 5 6 7 8
- C V K V I Y S M A V R - I F S M C L R L L F T

Scrambled

3 4 3 8 6 6 8 6
V K V S I I S I
L R L T L L T L

Used multiple times in resampling example

Only boxed alignment columns are randomly selected in this example

Non-supportive
Assessing confidence in tree using Maximum Likelihood

• Bayesian method – time consuming
  – The Bayesian posterior probabilities (BPP) are assigned to internal branches in consensus tree
  – Bayesian Markov chain Monte Carlo (MCMC) analytical software such as MrBayes (Huelsenbeck and Ronquist, 2001) and BAMBE (Simon and Larget, 1998) is now commonly used
  – Uses all the data
Some versatile phylogeny software packages

- MrBayes
- Paup
- Phylip
MrBayes: Bayesian Inference of Phylogeny

• MrBayes is a program for the Bayesian estimation of phylogeny.

• Bayesian inference of phylogeny is based upon a quantity called the posterior probability distribution of trees, which is the probability of a tree conditioned on the observations.

• The conditioning is accomplished using Bayes's theorem. The posterior probability distribution of trees is impossible to calculate analytically; instead, MrBayes uses a simulation technique called Markov chain Monte Carlo (or MCMC) to approximate the posterior probabilities of trees.

• The program takes as input a character matrix in a NEXUS file format. The output is several files with the parameters that were sampled by the MCMC algorithm. MrBayes can summarize the information in these files for the user.
MrBayes: Bayesian Inference of Phylogeny

MrBayes program features include:

• A common command-line interface for Macintosh, Windows, and UNIX operating systems;
• Extensive help available via the command line;
• Ability to analyze nucleotide, amino acid, restriction site, and morphological data;
• Mixing of data types, such as molecular and morphological characters, in a single analysis;
• A general method for assigning parameters across data partitions;
• An abundance of evolutionary models, including 4 X 4, doublet, and codon models for nucleotide data and many of the standard rate matrices for amino acid data;
• Estimation of positively selected sites in a fully hierarchical Bayes framework;
• The ability to spread jobs over a cluster of computers using MPI (for Macintosh and UNIX environments only).
PAUP*

Version 4

tools for inferring and interpreting phylogenetic trees

Analyze
- Molecular sequences
- Morphological data
- Other data types

Using
- Maximum likelihood
- Parsimony
- Distance methods
Phylip programs by type of data

- DNA sequences
- Protein sequences
- Restriction sites
- Distance matrices
- Gene frequencies
- Quantitative characters
- Discrete characters
- tree plotting, consensus trees, tree distances and tree manipulation

Phylip – by Joe Felsenstein

Phylip programs by type of algorithm
• Heuristic tree search
• Branch-and-bound tree search
• Interactive tree manipulation
• Plotting trees, consensus trees, tree distances
• Converting data, making distances or bootstrap replicates

The Newick tree format

(B,(A,C,E),D);  -- tree topology

(B:6.0,(A:5.0,C:3.0,E:4.0):5.0,D:11.0); -- with branch lengths

(B:6.0,(A:5.0,C:3.0,E:4.0)Ancestor1:5.0,D:11.0)Root;  -- with branch lengths and ancestral node names