Filogenia e evolução molecular
How to build Phylogenetic Trees

1. Select Sequences
2. Align Sequences
3. Choose model and method; Build tree
4. Evaluate Tree
   - Good
   - Needs Improvement
5. Interpret Phylogeny
## Sequências, evolução e relógio molecular

![ClustalX 1.4b Alignment](image)

| 1 | PEAPAL1_CD     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 2 | PEAPAL_CDS     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 3 | PEAPAL2_CDS    | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 4 | MSPAL_CDS1     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 5 | TFRPAL1X_C     | VKDAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 6 | DMPAL1_CDS     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 7 | SSNPAL_CDS     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 8 | POPPALGA_C     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 9 | PCPAL2_CDS     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 10| PCPAL3_CDS     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 11| ATHPAL2_CDS    | MKAHLKDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 12| ATHPAL_CDS     | MKAHLKDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 13| TOBTPAL1A_C    | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 14| TOMPAL5A_C     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 15| TOMPEAMLY      | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 16| NTPHEAL_CDS    | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 17| TOBPAL1_CDS    | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 18| IPBPAL_CDS     | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |
| 19| POPPALB_CDS    | VKAAKLHEDPL0K-PKQDRY-ALRTSPQWLGLPVIRQSTKSIEREINSYDNPLID |

**ruler**

```
..: *** : **** : * : *************** : * : **** : *** : *
```

```
130 140 150 160 170 180
```
Sequências, evolução e relógio molecular

ARMADILLIDIUM  
CARCINUS  
HOMARUS

ARMS TsCKGfYDrGLFAQLDLRVCEDCYNLYRKPf

QIVDTSCKGVYDRAfL-DLEHVCDDCYNLYRTSYVA
QVFDQACKGVYDrNLFKKLDRVCEDCYNLYRKPf

...

********

ARMADILLIDIUM  
CARCINUS  
HOMARUS

AECDRTDCYTFESCLKDLMMHDFINEYKEMALMV
SACRSNCYSNVLVFRQCMDDLLLMDEFDQYARKVQVMV
TTCRENCYSNWVFRQCLDDLLLNVIDEYVSNVQM-

**

**

**

**

**

****
Sequências, evolução e relógio molecular
(a) Missense mutation (transversion)
DNA:  TAC  TCC  ACC  ACG  ATA  ——
mRNA:  AUG  AGG  UGG  UGC  UAU  ——
Protein: met  arg  trp  cys  tyr  ——

(b) Missense mutation (transition)
DNA:  TAC  TCC  AGC  ACG  ATA  ——
mRNA:  AUG  AGG  UCG  UGC  UAU  ——
Protein: met  arg  ser  cys  tyr  ——

(c) Nonsense mutation
DNA:  TAC  TCC  ATC  ACG  ATA  ——
mRNA:  AUG  AGG  UAG  UGC  UAU  ——
Protein: met  arg  STOP  cys  tyr  X  X

(d) Frameshift mutation
DNA:  TAC  TCC  AAG  CAC  GAT  ——
mRNA:  AUG  AGG  UUC  GUG  CUA  ——
Protein: met  arg  phe  val  leu  ——

(e) Synonymous (silent) mutation
DNA:  TAC  TCC  GAC  ACG  ATA  ——
mRNA:  AUG  AGG  CUG  UGC  UAU  ——
Protein: met  arg  leu  cys  tyr  ——

In phase (reading frames)

Out of phase

Sequências, evolução e relógio molecular
Sequências, evolução e relógio molecular

Seq 1: A G C G A G
Seq 2: G C G G A C

Number of changes:

Seq 1:
C → A → T → G

Seq 2:
C → G

1 2 3
Sequences, evolution and molecular clock

One substitution happened - one is visible

Sequence 0

Sequence 1

Sequence 2

\[ \text{Fraction of sites that differ is } d = \frac{3}{10} \]
Sequências, evolução e relógio molecular

Two substitutions happened - only one is visible

Two substitutions happened - nothing is visible
Estimating Genetic Differences

If all nucleotides equally likely, observed difference would plateau at 0.75

Simply counting differences underestimates distances.

Fails to count for multiple hits
Models of nucleotide substitution allow for the calculation of probabilities of specific base changes along a branch. They include different parameters that aim to describe distinct aspects of the process of nucleotide substitution.
Models of evolution

- **Jukes–Cantor (JC)**
  - Equal base frequencies: \( \pi_A = \pi_C = \pi_G = \pi_T \)
  - All substitutions equally likely: \( \alpha = \beta \)

- **Kimura 2 parameter (K2P)**
  - Equal base frequencies: \( \pi_A = \pi_C = \pi_G = \pi_T \)
  - Transversions and transitions have different substitution rates: \( \alpha \neq \beta \)

- **Felsenstein (F81)**
  - Unequal base frequencies: \( \pi_A \neq \pi_C \neq \pi_G \neq \pi_T \)
  - All substitutions equally likely: \( \alpha = \beta \)

- **Hasegawa et al. (HKY85)**
  - Unequal base frequencies: \( \pi_A \neq \pi_C \neq \pi_G \neq \pi_T \)
  - Transversions and transitions have different substitution rates: \( \alpha \neq \beta \)

- **General reversible (REV)**
  - Unequal base frequencies: \( \pi_A \neq \pi_C \neq \pi_G \neq \pi_T \)
  - All six pairs of substitutions have different rates

Page RDM, Holmes EC (1998)
Molecular Evolution: a phylogenetic approach
Molecular Clock
Proposed that for any given protein, the rate of molecular evolution is approximately constant over time in all lineages.
Molecules reflect evolutionary divergence
Molecular Clock
Hypothesis

The number of genetic differences between two sequences is proportional to the divergence time between the individuals the sequences belong to.

The mutation rate is sufficiently constant to allow a reasonable time divergence estimate.
If proteins evolve at constant rates, then the number of substitutions between two sequences may be used to estimate divergence times.

This is analogous to the dating of geological times by radioactive decay.
The differences between 1 and 2 are the result of changes on the lineage leading to descendant 1 + those on the lineage leading to descendant 2.

Assumption: Life is monophyletic.
$T = \text{divergence time}$

$K = \text{number of substitutions}$

$r = \frac{k}{2T}$

$r = \text{rate of nucleotide substitution}$
Phylogeny based on nucleotide differences in the gene for cytochrome c

Values are estimates of the minimum number of nucleotide substitutions that have occurred along the lineages in the gene coding for this protein.
A molecular clock for Cytochrome c

- Yeast vs mould
- Angiosperms vs animals
- Insects vs vertebrates
- Mammals vs reptiles
- Fish vs land vertebrates
- Amphibians vs birds and mammals
- Birds vs reptiles
- Birds vs mammals

Time since divergence (millions of years)

Amino acid substitutions (per 100 residues) in cytochrome c
It has two trees with the same topology using:
"hard" data - the fossil record - for the black version
"soft" data - molecular clock estimates - for the red

Divergence times

P  Paleontological data

M  Molecular data
Árvores representam TAXA (genes, populações, espécies, etc.)
Usadas para inferência filogenética

Nó ancestral ou Raiz da árvore

Nós internos ou pontos de divergência (representam ancestrais hipotéticos dos taxa)

Ramos ou linhagens

Terminal node (leaf)

Nós terminais
• A statistical estimation of the true evolutionary history of a group of taxa

• Estimated from contemporary species and molecules
Clades, taxa e árvores filogénéticas

Phylogenies = Evolutionary relationships

This dimension can:
• be proportional to genetic distance (differences) = phylogram or additive trees;
• be proportional to time = ultrametric trees;
• have no scale what so ever.

((A, (B, C)), (D, E)) = phylogeny

B - C closer, sister clade A
A - B - C, sister clade D - E
If there was a temporal or genetic scale then D - E taxa are the closest related, and diverged more recently

Alternação de caracteres.
Não tem significado o espaçamento entre os taxa nem a ordem absoluta porque aparecem.
All of these rearrangements show the same evolutionary relationships between the taxa.
Trees can be scaled or unscaled (with or without branch lengths)
Cladogramas

shows relative recency of common ancestry
Filogramas = árvores aditivas

Provides additional information: branch lengths.

Numbers correspond to “amount” of evolutionary change on branch.
Árvores ultramétricas

Special kind of additive tree in which all of tips are equidistant from root.

Used to depict evolutionary time expressed as sequence divergence or directly as years.
Phylogenetic tree building (or inference) methods are aimed at discovering which of the possible unrooted trees is "correct".

We would like this to be the “true” biological tree — that is, one that accurately represents the evolutionary history of the taxa. However, we must settle for discovering the computationally correct or optimal tree for the phylogenetic method of choice.
The number of unrooted trees increases in a greater than exponential manner with number of taxa.

\[
(2N - 5)!! = \text{# unrooted trees for } N \text{ taxa}
\]

<table>
<thead>
<tr>
<th># Taxa (N)</th>
<th># Unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>10,935</td>
</tr>
<tr>
<td>9</td>
<td>135,135</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
</tr>
<tr>
<td>30</td>
<td>3.58 \times 10^{36}</td>
</tr>
</tbody>
</table>

This table shows the number of unrooted trees for taxa ranging from 3 to 10. The formula \((2N - 5)!!\) calculates the number of unrooted trees for any number of taxa N.
An unrooted, four-taxon tree theoretically can be rooted in five different places to produce five different rooted trees.

The unrooted tree 1:

Rooted tree 1

Rooted tree 2

Rooted tree 3

Rooted tree 4

Rooted tree 5

These trees show five different evolutionary relationships among the taxa!
Each unrooted tree theoretically can be rooted anywhere along any of its branches.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Unrooted trees X</th>
<th>roots</th>
<th>Rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>7</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>9</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
<td>11</td>
<td>10 395</td>
</tr>
<tr>
<td>8</td>
<td>10 935</td>
<td>13</td>
<td>135 125</td>
</tr>
<tr>
<td>9</td>
<td>135 135</td>
<td>15</td>
<td>2 027 025</td>
</tr>
<tr>
<td>30</td>
<td>$3.58 \times 10^{36}$</td>
<td>57</td>
<td>$2.04 \times 10^{38}$</td>
</tr>
</tbody>
</table>

For 10 sequences there are more than 34 million rooted trees.
For 20 sequences there are $8,200,794,532,637,891,559,000$ trees.
In a recent study of 135 human mtDNA sequences there were potentially $2.113 \times 10^{267}$ trees. This number is larger than number of particles known in the universe!!
There are two major ways to root trees:

**By outgroup:**
Uses taxa (the “outgroup”) that are known to fall outside of the group of interest (the “ingroup”). Requires some prior knowledge about the relationships among the taxa. The outgroup can either be species (e.g., birds to root a mammalian tree) or previous gene duplicates (e.g., a-globins to root b-globins).

**By midpoint or distance:**
Roots the tree at the midway point between the two most distant taxa in the tree, as determined by branch lengths. Assumes that the taxa are evolving in a clock-like manner. This assumption is built into some of the distance-based tree building methods.
• As a basis of comparison we need to designate an outgroup
  Which is a species or group of species that is closely related to the ingroup, the various species we are studying

• Outgroup comparison
  Is based on the assumption that homologies present in both the outgroup and ingroup must be primitive characters that predate the divergence of both groups from a common ancestor
Clades, taxa e árvores filogénéticas

**Plesiomorphies**

Ancestral character

**Synapomorphies**

Shared derived character

**Autapomorphies**

Uniquely derived character
Monophyletic. In this tree, grouping 1, consisting of the seven species B–H, is a monophyletic group, or clade. A monophyletic group is made up of an ancestral species (species B in this case) and all of its descendant species. Only monophyletic groups qualify as legitimate taxa derived from cladistics.

Paraphyletic. Grouping 2 does not meet the cladistic criterion: It is paraphyletic, which means that it consists of an ancestor (A in this case) and some, but not all, of that ancestor’s descendants. (Grouping 2 includes the descendants I, J, and K, but excludes B–H, which also descended from A.)

Polyphyletic. Grouping 3 also fails the cladistic test. It is polyphyletic, which means that it lacks the common ancestor of (A) the species in the group. Furthermore, a valid taxon that includes the extant species G, H, J, and K would necessarily also contain D and E, which are also descended from A.
Phylogenetic Methods

**Distance:**
- Tree based on pairwise distances between sequences
- Evolutionary models applied to pairwise distances to account for multiple substitutions per site and rate heterogeneity among sites

**Parsimony:**
- Minimize the number of substitutions
- Assumes sites are independent
- Assumes 1 substitution per site

**Maximum Likelihood:**
- Maximize probability of sequences given tree
- Evolutionary models applied to each position to account for multiple substitutions per site and rate heterogeneity among sites
- Gives single tree with highest likelihood
- Assumes sites are independent

**Bayesian:**
- Maximize posterior probability of tree given sequences
- Evolutionary models applied to each position to account for multiple substitutions per site and rate heterogeneity among sites
- Integrates over all trees
- Assumes sites are independent
Molecular phylogenetic tree building methods:

Are mathematical and/or statistical methods for inferring the divergence order of taxa, as well as the lengths of the branches that connect them. There are many phylogenetic methods available today, each having strengths and weaknesses. Most can be classified as follows:

<table>
<thead>
<tr>
<th>DATA TYPE</th>
<th>COMPUTATIONAL METHOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters</td>
<td>PARSIMONY</td>
</tr>
<tr>
<td></td>
<td>MAXIMUM LIKELIHOOD</td>
</tr>
<tr>
<td>Distances</td>
<td>MINIMUM EVOLUTION</td>
</tr>
<tr>
<td></td>
<td>LEAST SQUARES</td>
</tr>
<tr>
<td></td>
<td>Clustering algorithm</td>
</tr>
<tr>
<td></td>
<td>UPGMA</td>
</tr>
<tr>
<td></td>
<td>NEIGHBOR-JOINING</td>
</tr>
</tbody>
</table>
Methods of reconstructing phylogenies (evolutionary trees)

**Distance matrix methods.** Tree that best predicts the entries in a table of pairwise distances among species.

**Parsimony methods.** Tree that allows evolution of the sequences with the fewest changes is preferred.

**Maximum likelihood.** Tree that has highest probability that the observed data would evolve.

*Also Bayesian methods: tree which is most probable a posteriori given some prior distribution on trees.*
Types of data used in phylogenetic inference

Distance-based methods: Transform the sequence data into pairwise distances (dissimilarities), and then use the matrix during tree building.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species A</td>
<td>----</td>
<td>0.20</td>
<td>0.50</td>
<td>0.45</td>
<td>0.40</td>
</tr>
<tr>
<td>Species B</td>
<td>0.23</td>
<td>----</td>
<td>0.40</td>
<td>0.55</td>
<td>0.50</td>
</tr>
<tr>
<td>Species C</td>
<td>0.87</td>
<td>0.59</td>
<td>----</td>
<td>0.15</td>
<td>0.40</td>
</tr>
<tr>
<td>Species D</td>
<td>0.73</td>
<td>1.12</td>
<td>0.17</td>
<td>----</td>
<td>0.25</td>
</tr>
<tr>
<td>Species E</td>
<td>0.59</td>
<td>0.89</td>
<td>0.61</td>
<td>0.31</td>
<td>----</td>
</tr>
</tbody>
</table>

Example 1: Uncorrected “p” distance (=observed percent sequence difference)

Example 2: Kimura 2-parameter distance (estimate of the true number of substitutions between taxa)
Métodos principais de filogenia molecular
Métodos principais de filogenia molecular

**Distance**

**Parsimony**

<table>
<thead>
<tr>
<th>sequences</th>
<th>distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>5  4</td>
</tr>
<tr>
<td>4</td>
<td>5  4  2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>sites</th>
<th>sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  2  3  4  5  6  7</td>
<td>T T A T T A A</td>
</tr>
<tr>
<td>1  2  3  4  5  6  7</td>
<td>A A T T T A A</td>
</tr>
<tr>
<td>1  2  3  4  5  6  7</td>
<td>A A A A A T A</td>
</tr>
<tr>
<td>1  2  3  4  5  6  7</td>
<td>A A A A A A T</td>
</tr>
</tbody>
</table>
Construction of a distance tree using clustering with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

First, construct a distance matrix:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>GCTTGTCCGTTACGAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>ACTTGTCTGTTACGAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ACCTGTCCGAAACGAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>ACTTGACCGTTTCCTT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>AGATGACCGTTTCGAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>ACTACACCTCTTATGAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Métodos principais de filogenia molecular**

**UPGMA**

**First round**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B</strong></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>E</strong></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Choose the most similar pair, cluster them together and calculate the new distance matrix.

\[
\begin{align*}
\text{dist}(A,B),C &= \frac{\text{dist}(A,C) + \text{dist}(B,C)}{2} = 4 \\
\text{dist}(A,B),D &= \frac{\text{dist}(A,D) + \text{dist}(B,D)}{2} = 6 \\
\text{dist}(A,B),E &= \frac{\text{dist}(A,E) + \text{dist}(B,E)}{2} = 6 \\
\text{dist}(A,B),F &= \frac{\text{dist}(A,F) + \text{dist}(B,F)}{2} = 8 \\
\end{align*}
\]
## Métodos principais de filogenia molecular

**UPGMA**

### First round

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td></td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>4</td>
<td></td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Choose the most similar pair, cluster them together and calculate the new distance matrix.

- \( \text{dist}(A, B), C = \frac{(\text{dist}AC + \text{dist}BC)}{2} = 4 \)
- \( \text{dist}(A, B), D = \frac{(\text{dist}AD + \text{dist}BD)}{2} = 6 \)
- \( \text{dist}(A, B), E = \frac{(\text{dist}AE + \text{dist}BE)}{2} = 6 \)
- \( \text{dist}(A, B), F = \frac{(\text{dist}AF + \text{dist}BF)}{2} = 8 \)
**Second round**

<table>
<thead>
<tr>
<th></th>
<th>A,B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

**Third round**

<table>
<thead>
<tr>
<th></th>
<th>A,B</th>
<th>C</th>
<th>D,E</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D,E</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

\[\text{dist}(D,E),(A,B) = \frac{\text{dist}(D(A,B)) + \text{dist}(E(A,B))}{2} = 6\]
\[\text{dist}(D,E),C = \frac{\text{dist}(D(C)) + \text{dist}(E(C))}{2} = 6\]
\[\text{dist}(D,E),F = \frac{\text{dist}(D(F)) + \text{dist}(E(F))}{2} = 8\]
<table>
<thead>
<tr>
<th></th>
<th>A,B</th>
<th>C</th>
<th>D,E</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D,E</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

**Forth round**

\[
\begin{align*}
\text{dist}(A,B,C),DE &= \frac{\text{dist}(ABDE) + \text{dist}(CDE)}{2} = 6 \\
\text{dist}(A,B,C),F &= \frac{\text{dist}(ABF) + \text{dist}(CF)}{2} = 8
\end{align*}
\]
Fifth round

<table>
<thead>
<tr>
<th></th>
<th>AB,C</th>
<th>D,E</th>
</tr>
</thead>
<tbody>
<tr>
<td>D,E</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Sixth round

<table>
<thead>
<tr>
<th></th>
<th>ABC,DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>8</td>
</tr>
</tbody>
</table>

Note the this method identifies the root of the tree.
UPGMA fails when rates of evolution are not constant

A tree in which the evolutionary rates are not equal

From http://www.icp.ucl.ac.be/~opperd/private/upgma.html
Métodos principais de filogenia molecular
Making trees using neighbor-joining

The neighbor-joining method of Saitou and Nei (1987).
Is especially useful for making a tree having a large number of taxa.

Begin by placing all the taxa in a star-like structure.
Next, identify neighbors (e.g. 1 and 2) that are most closely related. Connect these neighbors to other OTUs via an internal branch, XY. At each successive stage, minimize the sum of the branch lengths.
Define the distance from X to Y by

\[ d_{XY} = \frac{1}{2}(d_{1Y} + d_{2Y} - d_{12}) \]
The neighbor joining method joins at each step, the two closest sub-trees that are not already joined. It is based on the minimum evolution principle. One of the important concepts in the NJ method is *neighbors*, which are defined as two taxa that are connected by a single node in an unrooted tree.
Métodos principais de filogenia molecular

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>
We have in total 6 OTUs (N=6).

**Step 1:** We calculate the net divergence \( r(i) \) for each OTU from all other OTUs

\[
\begin{align*}
  r(A) &= 5+4+7+6+8 = 30 \\
  r(B) &= 42 \\
  r(C) &= 32 \\
  r(D) &= 38 \\
  r(E) &= 34 \\
  r(F) &= 44
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>7</td>
<td>10</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td>6</td>
<td>9</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>
Step 2: Now we calculate a new distance matrix using for each pair of OUTs the formula:

\[ M_{ij} = d_{ij} - \frac{r_i + r_j}{N-2} \]

or in the case of the pair A,B:

\[ M_{AB} = d_{AB} - \frac{(r_A + r_B)}{(N-2)} = 5 - \frac{(30 + 42)}{(6-2)} = -13 \]
Step 2: Now we calculate a new distance matrix using for each pair of OUTs the formula:

\[ M(ij) = d(ij) - \frac{[r(i) + r(j)]}{(N-2)} \]

or in the case of the pair A,B:

\[ M(AB) = d(AB) - \frac{[(r(A) + r(B))]}{(N-2)} = \]
\[ M(AB) = 5 - \frac{[30 + 42]}{(6-2)} = -13 \]

\[ M(AC) = 4 - \frac{[30 + 32]}{(6-2)} = -11.5 \]
\[ M(AD) = 7 - \frac{[30 + 38]}{(6-2)} = -10 \]

etc........

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-11.5</td>
<td>-11.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-10</td>
<td>-10</td>
<td>-10.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-10</td>
<td>-10</td>
<td>-10.5</td>
<td>-13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-10.5</td>
<td>-10.5</td>
<td>11</td>
<td>-11.5</td>
<td>-11.5</td>
<td></td>
</tr>
</tbody>
</table>
**Step 3:** Now we choose as neighbors those two OTUs for which $M_{ij}$ is the smallest. These are A and B and D and E. Let's take A and B as neighbors and we form a new node called U (joining AB).
Now we calculate the branch length from the internal node U to the external OTUs A and B.

\[ S(AU) = \frac{d(AB)}{2} + \frac{[r(A)-r(B)]}{2(N-2)} = \]
\[ S(AU) = \frac{5}{2} + \frac{[30-42]}{2(6-2)} = 1 \]
\[ S(BU) = 5 - S(AU) = 4 \]
Step 4: Now we define new distances from U to each other terminal node:

\[
\begin{align*}
d(CU) &= d(AC) + d(BC) - d(AB) / 2 = 3 \\
d(DU) &= d(AD) + d(BD) - d(AB) / 2 = 6 \\
d(EU) &= d(AE) + d(BE) - d(AB) / 2 = 5 \\
d(FU) &= d(AF) + d(BF) - d(AB) / 2 = 7 \\
\end{align*}
\]

\[
\begin{array}{cccccc}
A & B & C & D & E \\
\hline
A & & & & & \\
B & 5 & & & & \\
C & 4 & 7 & & & \\
D & 7 & 10 & 7 & & \\
E & 6 & 9 & 6 & 5 & \\
F & 8 & 11 & 8 & 9 & 8 \\
\end{array}
\]

\[
\begin{array}{cccccc}
U & C & D & E & F \\
\hline
U & & & & & \\
C & 3 & & & & \\
D & 6 & 7 & & & \\
E & 5 & 6 & 5 & & \\
F & 7 & 8 & 9 & 8 & \\
\end{array}
\]

\[
\begin{align*}
r(A) &= 30 \\
r(B) &= 42 \\
r(C) &= 32 \\
r(D) &= 38 \\
r(E) &= 34 \\
r(F) &= 44 \\
\end{align*}
\]
Step 4: And we create a new distance matrix

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>
Step 5: Now, N is N-1 = 5, and the entire procedure is repeated starting at step 1
Métodos principais de filogenia molecular

NJ

Diagrama de filogenia com distâncias:
- A-B: 4
- A-C: 1
- A-D: 2.25
- A-E: 0.5
- A-F: 4.75
- B-C: 2
- B-D: 2.75
- C-E: 2.75
- C-F: 1
- D-E: 1
- E-F: 1
Comparison of UPGMA and NJ

UPGMA

NJ
Comparison of UPGMA and NJ

**Distance Methods:** evolutionary distances (number of substitutions) are computed for all pairs of taxa.

**UPGMA:** unweighted pairgroup method with arithmetic means.
- assumes equal rate of substitutions (therefore is always rooted, as the taxa that has accumulated more sequences is evidently older) (if the substitutions rates are different among taxa, then the tree maybe wrong)
- sequential clustering algorithms
- pairs of taxa are clustered in order of decreasing similarity

**Neighbor Joining:** finding shortest (minimum evolution) tree by finding neighbors that minimize the total length of the tree. Shortest pairs are chosen to be neighbors and then joined in distance matrix as one OTU.
- the algorithm does not assume that the molecular clock is constant for sequences in the tree. If there are unequal substitution rates, the tree is more accurate than UPGMA.
William of Ockham (or Occam) was a 14th-century English logician and Franciscan friar who's name is given to the principle that when trying to choose between multiple competing theories the simplest one is probably the best. This principle is known as Ockham's razor.
**Maximum parsimony**

- Minimize the number of substitutions
- Assumes sites are independent
- Assumes <1 substitution per site

**Species**

<table>
<thead>
<tr>
<th>Species</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
</tr>
</tbody>
</table>

**Tree 1**

1 (A) ➔ 2 (G) ➔ 3 (A) ➔ 4 (G)

- 2 changes

**Tree 2**

1 (A) ➔ 3 (A) ➔ 2 (G) ➔ 4 (G)

- 1 change

Parsimony methods

consider first site of these 4 sequences

Unrooted tree A

Two possible character reconstructions for the evolution of site 1 on unrooted tree A

1 change required (most parsimonious)

5 changes required
...Now find most parsimonious reconstructions for the other 4 sites

4 sequences

1. ATATT
2. ATCGT
3. GCAGT
4. GCCGT

Note there are 2 equally parsimonious reconstructions for site 3
In phylogenetic analysis, parsimony is used to justify the choice of a tree that represents the smallest number of evolutionary changes.

As an example, if we wanted to use the DNA sequences from seven sites to determine the most parsimonious arrangement of four species, we would begin by tabulating the sequence data.

Then, we would draw all possible phylogenies for the four species, including the three shown here.
**Maximum Parsimony**

States that, given certain rules about how DNA changes over time, a tree can be found that reflects the most likely sequence of evolutionary events.

**APPLICATION**

The most efficient way to study the various phylogenetic hypotheses is to begin by first considering the most parsimonious—that is, which hypothesis requires the fewest total evolutionary events (molecular changes) to have occurred.

**TECNIQUE**

Follow the numbered steps as we apply the principle of parsimony to a hypothetical phylogenetic problem involving four closely related bird species.

1. First, draw the possible phylogenies for the species (only 3 of the 15 possible trees relating these four species are shown here).

2. Tabulate the molecular data for the species (in this simplified example, the data represent a DNA sequence consisting of just seven nucleotide bases).

3. Now focus on site 1 in the DNA sequence. A single base-change event, marked by the crossbar in the branch leading to species I, is sufficient to account for the site 1 data.
Continuing the comparison of bases at sites 2, 3, and 4 reveals that each of these possible trees requires a total of four base-change events (marked again by crossbars). Thus, the first four sites in this DNA sequence do not help us identify the most parsimonious tree.

After analyzing sites 5 and 6, we find that the first tree requires fewer evolutionary events than the other two trees (two base changes versus four). Note that in these diagrams, we assume that the common ancestor had GG at sites 5 and 6. But even if we started with an AA ancestor, the first tree still would require only two changes, while four changes would be required to make the other hypotheses work. Keep in mind that parsimony only considers the total number of events, not the particular nature of the events (how likely the particular base changes are to occur).
At site 7, the three trees also differ in the number of evolutionary events required to explain the DNA data.

RESULTS

To identify the most parsimonious tree, we total all the base-change events noted in steps 3–6 (don’t forget to include the changes for site 1, on the facing page). We conclude that the first tree is the most parsimonious of these three possible phylogenies. (But now we must complete our search by investigating the 12 other possible trees.)
• The most widely-used method, familiar notion in science ("simplicity")

• Shared attributes among taxa are inherited from common ancestors

• When character conflicts occur, ad hoc hypotheses cannot be avoided if you want to explain all the data, and assumptions of homoplasy must be invoked
Parsimony methods

Optimality criterion: The ‘most-parsimonious’ tree is the one that requires the fewest number of evolutionary events (e.g., nucleotide substitutions, amino acid replacements) to explain the sequences.

**Advantages:**
- Are simple, intuitive, and logical (many possible by ‘pencil-and-paper’).
- Can be used on molecular and non-molecular (e.g., morphological) data.
- Can tease apart types of similarity (shared-derived, shared-ancestral, homoplasy)
- Can be used for character (can infer the exact substitutions) and rate analysis.
- Can be used to infer the sequences of the extinct (hypothetical) ancestors.

**Disadvantages:**
- Are simple, intuitive, and logical (derived from “Medieval logic”, not statistics!)
- Can be fooled by high levels of homoplasy (‘same’ events).
- Can become positively misleading in the “Felsenstein Zone”:

[See Stewart (1993) for a simple explanation of parsimony analysis, and Swofford et al. (1996) for a detailed explanation of various parsimony methods.]
• **Parsimony** seeks solutions that minimize the amount of change required to explain the data (underestimates superimposed changes)

• **ML** attempts to estimate the actual amount of change (by specifying the evolutionary model that will account for the data with the highest likelihood)

• Methods that incorporate models of evolutionary change can make more efficient use of the data
Maximum likelihood (ML) methods

**Optimality criterion:** ML methods evaluate phylogenetic hypotheses in terms of the probability that a proposed model of the evolutionary process and the proposed unrooted tree would give rise to the observed data. The tree found to have the highest ML value is considered to be the preferred tree.

**Advantages:**
- Are based on explicit model of evolution.
- Usually the most ‘consistent’ of the methods available.
- Can be used for character (can infer the exact substitutions) and rate analysis.
- Can be used to infer the sequences of the extinct (hypothetical) ancestors.
- Can help account for branch-length effects.

**Disadvantages:**
- Are based on explicit model of evolution.
- Are not as simple and intuitive as many other methods.
- **Are computationally very intense** (limits number of taxa and length of sequence).
- **Sloooooow!!!**
- Violations of the assumed model can lead to incorrect trees.
Maximum likelihood (ML) method

O exemplo dos dados...

Rolamos os dados e obtemos “14” pontos

Qual o par de dados que mais provavelmente originará esse resultado?

Equivalente a: qual a árvore que mais provavelmente terá originado essas sequências?

(Which tree is most likely to have yielded these sequences?)

Idea from Gavin Naylor
How many ways of obtaining the score “14” are there for each pair?

- 5 + 2
  - 6 + 8
- 5 + 6
  - 2 + 12
  - 3 + 11
  - 4 + 10
  - 5 + 9
  - 6 + 8
- 5 + 7
  - 2 + 12
  - 3 + 11
  - 4 + 10
  - 5 + 9
  - 6 + 8
  - 7 + 7
  - 8 + 6

Idea from Gavin Naylor
What is the probability of any single outcome for each of the three sets of pairs?

1/6 x 1/8 + 1/6 x 1/12 + 1/8 x 1/12 = 1/48 + 1/72 + 1/96

= 1/48

= 1/72

= 1/96

Ideia from Gavin Naylor
Now multiply ways of obtaining the score “14” by the probability of any single outcome to get the likelihood.

\[
\begin{align*}
\frac{1}{48} \times 1 & = 0.0208 \\
\frac{1}{72} \times 5 & = 0.0694 \\
\frac{1}{96} \times 7 & = 0.0729
\end{align*}
\]

Notice that none of the likelihoods are very “likely”, but (8+12) is more likely than the other two.
1. Calculate likelihood for each site on a specific tree.

2. Sum up the L values for all sites on the tree.

3. Compare the L value for all possible trees.

4. Choose tree with highest L value.