# 9 Phylogenetic tree

## 9.1 Introduction to phylogenetic trees

A phylogenetic provides additional views on the analysis of multiple sequences.

### Elements of phylogenetic tree

- Terminal nodes: sequences, groups of genes, species, operational taxonomic units
- Internal nodes: hypothetical ancestral units
- Edges: often represent distances

### Types of trees

- Cladogram or phylogram
- Bifurcating or multifurcating
- Rooted or unrooted

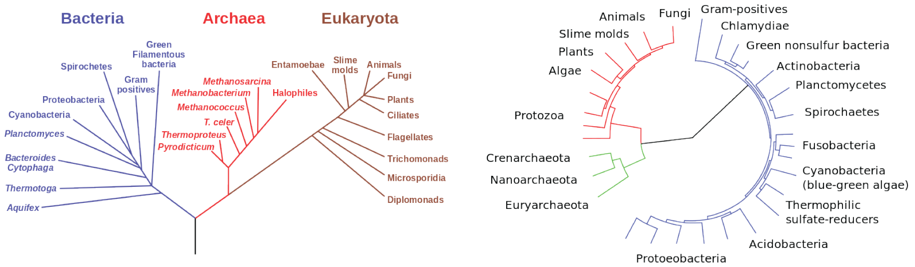


**Figure 9.1:** Phylogenetic trees (sources: TimVickers, Wikimedia Commons, NASA Astrobiology Institute, Wikimedia Commons))

### Rooted and unrooted trees

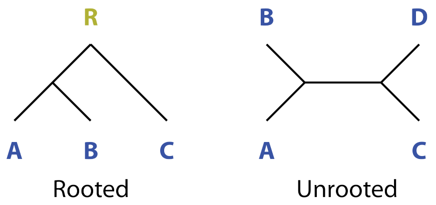
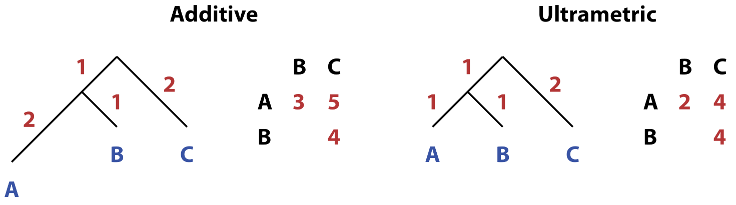


**Figure 9.2:** A rooted tree with three nodes and an unrooted tree with four nodes

### Additive and ultrametric trees



Additive

|   | B | C |
|---|---|---|
| A | 3 | 5 |
| B |   | 4 |

Ultrametric

|   | B | C |
|---|---|---|
| A | 2 | 4 |
| B |   | 4 |

**Figure 9.3:** Additive and ultrametric trees

An ultrrametic tree is a special version of additive tree. It assumes that the distances from two sequences to their common ancestor are always equal.

### Number of topologically different trees

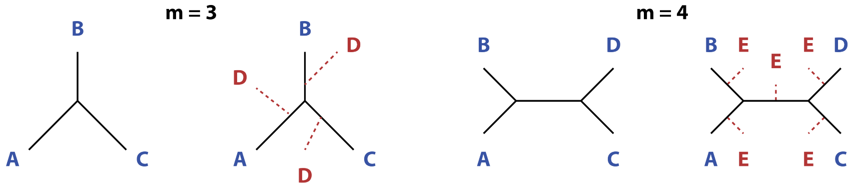


**Figure 9.4:** Adding one external node to unrooted trees

The number of all possible topologically different unrooted trees $\mathrm{T}_{\text{unroot}}(m)$ can be obtained by the double factorial of $2m - 5$.

$$\mathrm{T}_{\text{unroot}}(m) = (2m-5)!! \equiv \frac{(2m-5)!}{2^{m-3}(m-3)!}$$

$\mathrm{T}_{\text{root}}(m)$ can be calculated from $\mathrm{T}_{\text{unroot}}(m)$.

$$\mathrm{T}_{\text{root}}(m) = (m-1) \times \mathrm{T}_{\text{unroot}}(m)$$

### Example of the number of unrooted trees

What is the number of all possible topologically different unrooted trees when m = 7?

$$\mathrm{T}_{\text{unroot}}(7) = (2 \times 7 - 5)!! = 9!! = 1 \times 3 \times 5 \times 7 \times 9 = 945$$

or

$$\mathrm{T}_{\text{unroot}}(7) = \frac{(2 \times 7 - 5)!}{2^{7-3}(7-3)!} = \frac{9!}{2^4(4)!} = 945$$

### Exercise 9.1

1. Calculate the number of all possible topologically different unrooted trees when m = 5.

2. Construct an additive rooted tree for the distance matrix below. Estimate the edge values by trial and error.

|   | B | C |
|---|---|---|
| A | 4 | 7 |
| B |   | 5 |

## 9.2 Tree reconstruction methods

A number of methods have been proposed to reconstruct a phylogenetic tree.

### Two types of reconstruction methods

- Distance-based methods
- Character-based methods

### Distance-based methods

A distance is a positive value with larger values indicating that two sequences are separated further.

- PGMA (pair-group method using arithmetic mean)
- Neighbor-joining (NJ)

### Character-based methods

Character based methods rely on characters (amino acid/nucleotide letters) to reconstruct a tree.

- Maximum parsimony
- Maximum likelihood

### Evaluation of reconstructed trees

Bootstrapping is one of the methods to test the robustness of a reconstruct tree by adding noises and comparing the results.

1. Randomly generate a pseudo MAS from the original MSA
2. Reconstruct a tree
3. Repeat the process
4. Compare the trees

## 9.3 Distance-based methods

PGMA (pair-group method using arithmetic mean) and neighbor-joining are two popular distance-based methods to reconstruct a phylogenetic tree.

### UPGMA

UPGMA is an unweighted version of PGMA. It requires the evolutionary rate should be constant (ultrametric). Pairwise distances need to be pre-calculated, for instance, by DP.

- $w$: A new node
- $u$, $v$: Child nodes of $w$
- $m_A$ The number of original sequences in subtree $A$
- $D_{A,B}$: Distance between sequences/subtrees $A$ and $B$

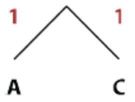
$$D_{w,x} = \frac{m_u D_{u,x} + m_v D_{v,x}}{m_u + m_v}$$

### Example of UPGMA

Reconstruct a phylogenetic tree from the pre-calculated distances below.

| | B | C | D |
|---|---|---|---|
| A | 4 | 2 | 5 |
| B | | 4 | 8 |
| C | | | 5 |

**Step 1a**. Find a pair with the closest distance



(A, C) is the closest pair and make a new internal node.

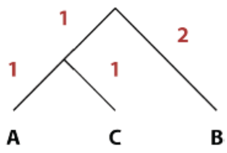
**Step 1b**. Recalculate the distances

$$d_{B,(AC)} = \frac{d_{B,A} + d_{B,C}}{2} = 4, \quad d_{D,(AC)} = \frac{d_{D,A} + d_{D,C}}{2} = 5$$

**Step 1c**. Update the distance matrix with a new node (AC)

| | B | D |
|---|---|---|
| (AC) | 4 | 5 |
| B | | 8 |

**Step 2a**. Find a pair with the closest distance



(AC, B) is the next closest pair and make a new internal node.

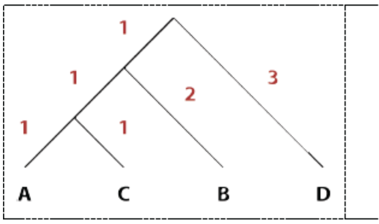
**Step 2b**. Recalculate the distance

$$d_{((AC)B),D} = \frac{2 \times d_{(AC),D} + d_{B,D}}{3} = 6$$

**Step 2c**. Update the distance matrix with a new node ((AC)B)

| | D |
|---|---|
| ((AC)B) | 6 |

**Step 3**. Complete the tree



Add the last sequence D to the tree.

### Evaluation on how well fitted to the original distances

Several criteria are available to find the best-fitted tree for a given distance matrix, such as the Cavalli-Sforza and Edwards criterion:

$$\sum_{i,j} (M_{i,j} - d_{i,j})^2$$

where $M_{i,j}$ and $d_{i,j}$ are respectively the original and the calculated pairwise distances.

### Example of the Cavalli-Sforza and Edwards criterion

Original

| | B | C | D |
|---|---|---|---|
| A | 4 | 2 | 5 |
| B | | 4 | 8 |
| C | | | 5 |

Reconstructed

| | B | C | D |
|---|---|---|---|
| A | 4 | 2 | 6 |
| B | | 4 | 6 |
| C | | | 6 |

$$\sum_{i,j} (M_{i,j} - d_{i,j})^2 = 2((5-6)^2 + (8-6)^2 + (5-6)^2) = 12$$

### WPGMA

WPGMA is a weighted version of PGMA.

$$D_{w,x} = \frac{D_{u,x} + D_{v,x}}{2}$$

### Neighbor-joining (NJ) method

It stats with the initial tree and then select two sequences which results in the smallest sum of edge lengths. It continues until there are no sequences to join. Unlike UPGMA, it does not require a constant evolutionary rate.

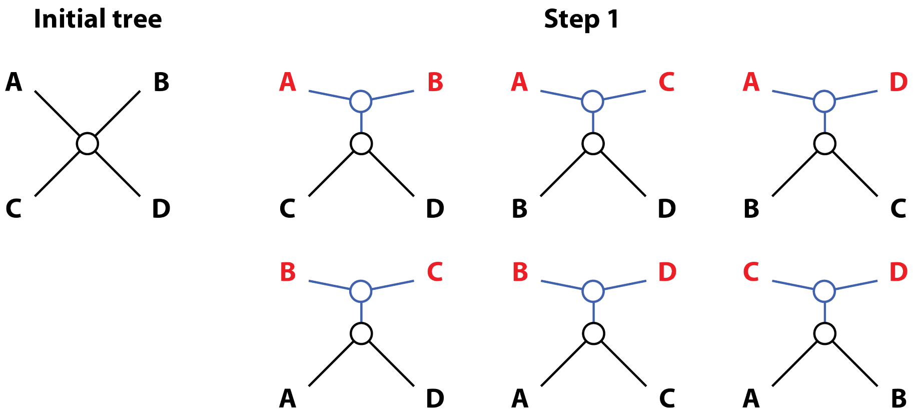


**Figure 9.5:** All possible combinations of adding one node the four sequences

### Exercise 9.2

1. Reconstruct a phylogenetic tree by using UPGMA and the following pre-calculated distances.

|   | B | C |
|---|---|---|
| A | 2 | 3 |
| B |   | 5 |

2. Create the distance matrix of the reconstructed tree.

|   | B | C |
|---|---|---|
| A |   |   |
| B |   |   |

3. Calculated the Cavalli-Sforza and Edwards criterion.

## 9.4 Maximum parsimony

Maximum parsimony is a character-based method to reconstruct a phylogenetic tree.

### Definition of parsimony

**Definition of parsimony** (source: http://www.merriam-webster.com)
*noun* | par·si·mo·ny | \ˈpär-sə-ˌmō-nē\

**a :** the quality of being careful with money or resources **:** thrift
**b :** the quality or state of being stingy

### Tree search method of maximum parsimony

The maximum parsimony method uses a tree search to find the tree with the minimum number of mutations.

**Algorithm 9.1:** Maximum parsimony with the minimum union operations

```
Construct an MSA;
foreach column $c \in MSA$ do
    foreach tree $t \in$ all possible topologically different trees do
        Count the number of union operations in $c$ for tree $t$;
    end
    Add one point to the tree with the minimum union operations;
end
Report the tree with the maximum point;
```

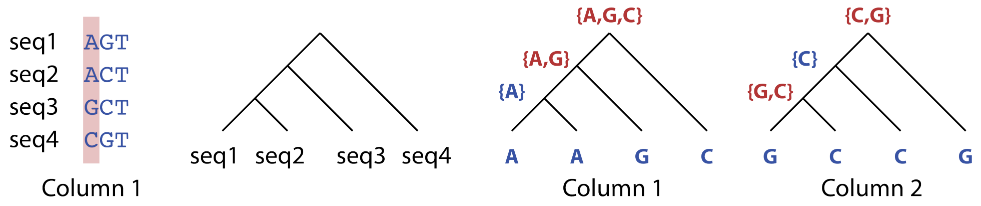
### Count the number of union operations

Either intersection or union operation is performed for each internal node.

- $s_i = s_j \cap s_k$ If there is at least one element in $s_j$ and $s_k$
- $s_i = s_j \cup s_k$ Otherwise

### Example of counting the number of union operations

Count the number of union operations for the first and the second columns.



### Column 1

- $A \cap A \rightarrow \{A\}$
- $\{A\} \cup G \rightarrow \{A, G\}$
- $\{A, G\} \cup C \rightarrow \{A, G, C\}$

# of union operations: 2

### Column 2

- $G \cup C \rightarrow \{G, C\}$
- $\{G, C\} \cap C \rightarrow \{C\}$
- $\{C\} \cup G \rightarrow \{C, G\}$

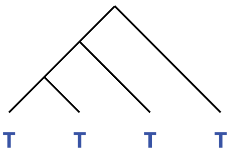
# of union operations: 2

### Exercise 9.3

1. What is the number of union operations for the third column?

```
seq1   AGT
seq2   ACT
seq3   GCT
seq4   CGT
```

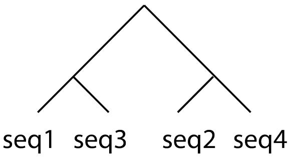
Column 3



Column 3

2. What is the number of union operations for each column?

```
seq1   AGT
seq2   ACT
seq3   GCT
seq4   CGT
```



- Column 1:
- Column 2:
- Column 3:

## 9.5 Maximum likelihood

The maximum likelihood can be used to reconstruct a phylogenetic tree.

### Conditional probabilities

- $P(H|D)$ where $D$ is observed data and $H$ is a hypothesis
- $P(M|D)$ where $D$ is observed data and $M$ is a model

Not easy to solve $P(H|D)$ or$P(M|D)$ directly

### Bayes' theorem

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

- $P(H|D)$, $P(M|D)$: conditional probabilities
- $P(H)$, $P(D)$: marginal probabilities
- $P(H|D)$: posterior probability
- $P(H)$: prior probability
- $P(D|H)$: likelihood
- $L(H|D)$: likelihood function (equivalent with $P(D|H)$)

### Maximum likelihood estimate (MLE)

We assume a uniform prior distribution for $P(H)$. Then, we can find the hypothesis that achieves the maximum likelyhood $L(H|D)$.

$$\hat{\theta} \in \underset{\theta \in \Theta}{\arg\max}\ L(\theta|D)$$

### Example of fair and unfair dice

Roll a die three times, and a 6 comes up three times in a row.

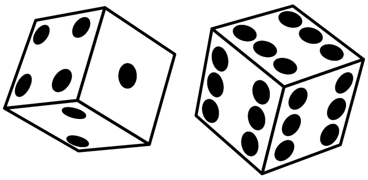


**Figure 9.6:** Two dice - a fair die and a die with three 6s

Probabilities:

- Three die roles with a fair die $P(D|H = fair) = (1/6)^3 \simeq 0.028$
- Three die roles with the unfair die: $P(D|H = unfair) = (3/6)^3 = 0.125$

Maximum likelihood estimate:

- $\underset{\theta \in (fair, unfair)}{\arg\max} \quad L(\theta|D) = unfair$

### Tree search method of maximum likelihood

The maximum likelihood method is also based on tree search. It tries to find the tree with the highest likelihood for a given MSA.

### Example of tree search method

Calculate the likelihood $L(T = tree1|D)$.

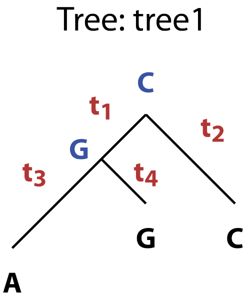


**Figure 9.7:** Two dice - a fair die and a die with three 6s

$$\text{Likelihood: } L(T = tree1|D) = P(D|T = tree1) = P_{CG}(t1)P_{CC}(t_2)P_{GA}(t_3)P_{GG}(t_4)$$

### Log-likelihood

- Logarithm is a monotonically increasing function
- $\log(ab) = \log(a) + \log(b)$

### Time complexity of tree search

Since it needs to search all possible trees, both the maximum parsimony and the maximum likelihood methods are NP hard problems.

- Exhaustive search: up to 8-10 sequences
- Branch and bound or pruning: up to 15-20 sequences
- Heuristics: 100+ sequences