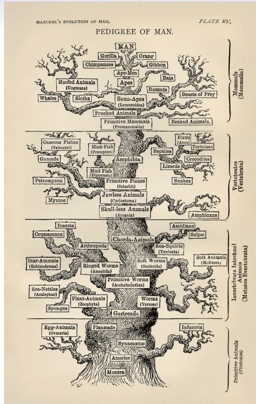
# Phylogenetic Trees

Course “Discrete Biological Models” (Modelli Biologici Discreti)

**Zsuzsanna Lipták**

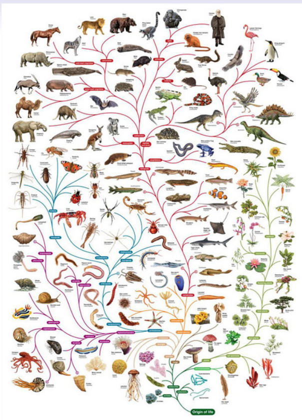
Laurea Triennale in Bioinformatica
a.a. 2014/15, fall term

These slides are partially based on the lecture notes *Algorithms for Phylogenetic Reconstruction*, by Jens Stoye and others, Bielefeld University, 2009/2010.




Tree of Life, by Ernst Haeckel, 1874

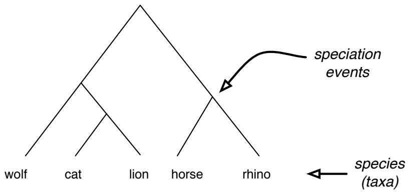




source: ETC Montessori



## What is a phylogenetic tree?




Phylogenetic trees display the evolutionary relationships among a set of objects (species). Contemporary species are represented by the leaves. Internal nodes of the tree represent speciation events (≈ common ancestors, usually extinct).



## Different types of phylogenetic trees

- rooted vs. unrooted
- binary (fully resolved) vs. multifurcating (polytomy)
- are edge lengths significant?



## Phylogenetic reconstruction

**Goal**
Given $n$ objects and data on these objects, find a phylogenetic tree with these objects at the leaves which best reflects the input data.

**Ex.**

| | $a$ | $b$ | $c$ |
|---|---|---|---|
| $a$ | 0 | 5 | 2 |
| $b$ | 5 | 0 | 4 |
| $c$ | 2 | 4 | 0 |

Can we find a tree with $a, b, c$ at the leaves s.t. the distance in the tree between $a$ and $b$ is 5, between $a$ and $c$ is 2, etc.?

## Phylogenetic reconstruction

Note:
We need to define more precisely

- what kind of input data we have,
- what kind of tree we want (e.g. rooted or unrooted), and
- what we mean by "reflect the data."

But first, ...

Say we have answered these questions, then: Could we just list all possible trees and then choose the/a best one?

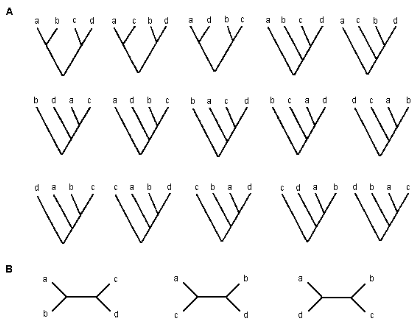


## Number of phylogenetic trees

| #*taxa* $n$ | # unrooted trees $(2n-5)!!$ | # rooted trees $(2n-3)!!$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 1 |
| 3 | 1 | 3 |
| 4 | 3 | 15 |



## Number of phylogenetic trees




All phylogenetic trees (rooted and unrooted) on 4 taxa.



## Number of phylogenetic trees

Theorem
There are $U_n = (2n-5)!! = \prod_{i=3}^{n}(2i-5)$ unrooted binary phylogenetic trees on $n$ objects, and $R_n = (2n-3)!! = \prod_{i=2}^{n}(2i-3)$ rooted binary phylogenetic trees on $n$ objects.

Proof
By induction on $n$, using that (1) we can get every unrooted tree on $n+1$ objects in a unique way by adding a new leaf to an unrooted tree on $n$ objects; (2) an unrooted binary tree with $n$ leaves has $2n-3$ edges, (3) every unrooted tree on $n$ objects can be rooted in (number of edges) ways, yielding a rooted tree on $n$ objects.



## Number of phylogenetic trees

| #*taxa* $n$ | # unrooted trees $(2n-5)!!$ | # rooted trees $(2n-3)!!$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 1 |
| 3 | 1 | 3 |
| 4 | 3 | 15 |
| 5 | 15 | 105 |
| 6 | 105 | 945 |
| 7 | 945 | 10,395 |
| 8 | 10,395 | 135,135 |
| 9 | 135,135 | 2,027,025 |
| 10 | 2,027,025 | 34,459,425 |



## Number of phylogenetic trees

So there are super-exponentially many trees:
We cannot check all of them!

## Distance data

We can have two kinds of input data:

- distance data, or
- character data (later)

Distance data is given as an $(n \times n)$ matrix $M$ with the pairwise distances between the taxa.

Ex.

| | $a$ | $b$ | $c$ |
|---|---|---|---|
| $a$ | 0 | 5 | 2 |
| $b$ | 5 | 0 | 4 |
| $c$ | 2 | 4 | 0 |

E.g., $M(a, b) = 5$ means that the distance between $a$ and $b$ is 5. Often, this is the edit distance (between two genomic sequences, or between homologous proteins, ...).

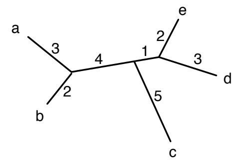


## Distance data

### Path metric of a tree

Given a tree $T$, the path-metric of $T$ is $dist_T$, defined as: $dist_T(u, v) =$ length of the (unique) path between $u$ and $v$. (In our trees edge weights are positive, so now: length of a path = sum of edge weights on path.)

### Example




$dist_T(a, b) = 5,$
$dist_T(a, d) = 11,$
$dist_T(c, d) = 9, \ldots$

### Question

Is it always possible to find a tree s.t. its path-metric equals the input distances? I.e. does such a tree exist for any input matrix $M$?



## Distance data

First of all, the input matrix $M$ has to define a metric (= a distance function), i.e. for all $x, y, z$,

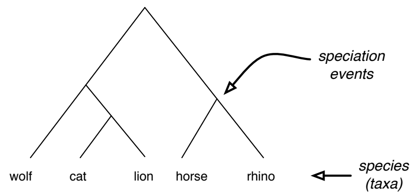
- $M(x, y) \geq 0$ and $(M(x, y) = 0$ iff $x = y)$ (positive definite)
- $M(x, y) = M(y, x)$ (symmetry)
- $M(x, y) + M(y, z) \geq M(x, z)$ (triangle inequality)

For example, the edit distance is a metric, the Hamming distance (on strings of the same length), the Euclidean distance (on $\mathbb{R}^2$).

But is this enough?



## Rooted trees and the molecular clock




In a rooted phylogenetic tree, the molecular clock assumption holds: that the speed of evolution is the same along all branches, i.e. the path distance from each leaf to the root is the same.



## Ultrametrics and the three-point condition

### Three point condition

Let $d$ be a metric on a set of objects $O$, then $d$ is an ultrametric if $\forall\, x, y, z \in O$:

$$d(x, y) \leq \max\{d(x, z), d(z, y)\}$$

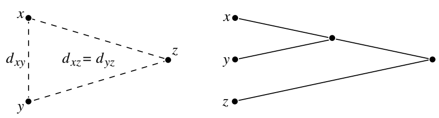



Figure : Three point condition. It implies that the path metric of a tree is an ultrametric.

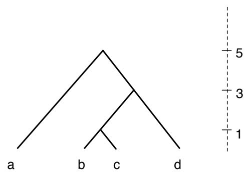
In other words, among the three distances, there is no unique maximum.



## Example

Ex. 2

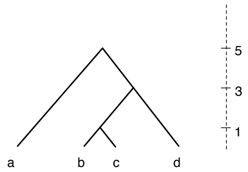
| | $a$ | $b$ | $c$ | $d$ |
|---|---|---|---|---|
| $a$ | 0 | 10 | 10 | 10 |
| $b$ | 10 | 0 | 2 | 6 |
| $c$ | 10 | 2 | 0 | 6 |
| $d$ | 10 | 6 | 6 | 0 |



## Example

Ex. 2

| | a | b | c | d |
|---|---|---|---|---|
| a | 0 | 10 | 10 | 10 |
| b | 10 | 0 | 2 | 6 |
| c | 10 | 2 | 0 | 6 |
| d | 10 | 6 | 6 | 0 |




Checking the ultrametric condition, we see that:

- for $a, b, c$ we get $2, 10, 10$ — okay
- for $a, b, d$ we get $6, 10, 10$ — okay
- for $a, c, d$ we get $6, 10, 10$ — okay
- for $b, c, d$ we get $2, 6, 6$ — okay



## Example

Compare this to our earlier example. There the matrix $M$ does not define an ultrametric!

Ex. 1 (from before)

| | a | b | c |
|---|---|---|---|
| a | 0 | 5 | 2 |
| b | 5 | 0 | 4 |
| c | 2 | 4 | 0 |

For the triple $a, b, c$ (the only triple), we get: $2, 4, 5$, and there is a unique maximum: 5.



## Example

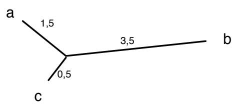
Compare this to our earlier example. There the matrix $M$ does not define an ultrametric!

Ex. 1 (from before)

| | a | b | c |
|---|---|---|---|
| a | 0 | 5 | 2 |
| b | 5 | 0 | 4 |
| c | 2 | 4 | 0 |

For the triple $a, b, c$ (the only triple), we get: $2, 4, 5$, and there is a unique maximum: 5.

Indeed, the only tree we found was not rooted:






## Ultrametrics and the three-point condition

Theorem

Given an $(n \times n)$ distance matrix $M$. There is a rooted tree whose path metric agrees with $M$ if and only if $M$ defines an ultrametric (i.e. if and only if the 3-point-condition holds). This tree is unique.



## Ultrametrics and the three-point condition

Theorem

Given an $(n \times n)$ distance matrix $M$. There is a rooted tree whose path metric agrees with $M$ if and only if $M$ defines an ultrametric (i.e. if and only if the 3-point-condition holds). This tree is unique.

Algorithm

There are algorithms which, given $M$, compute this rooted tree in $O(n^2)$ time (e.g. UPGMA).



## Additive metrics and the four-point condition

So what is the condition on the matrix $M$ for unrooted trees?

Four point condition.

Let $d$ be a metric on a set of objects $O$, then $d$ is an additive metric if $\forall\, x, y, u, v \in O$:

$$d(x, y) + d(u, v) \leq \max\{d(x, u) + d(y, u), d(x, v) + d(y, u)\}$$

In other words, among the three sums of two distances, there is no unique maximum.

## Additive metrics and the four-point condition

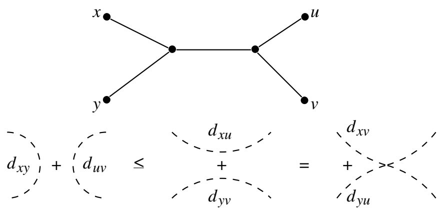
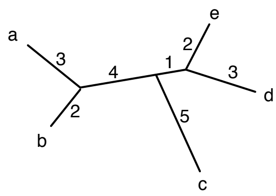



Figure : The four point condition. It implies that the path metric of a tree is an additive metric.



## Example




For ex., choose these 4 points: $a, b, c, e$. Then we get the three sums: $d(a, b) + d(c, e) = 5 + 8 = 13$, $d(a, c) + d(b, e) = 12 + 9 = 21$, and $d(a, e) + d(b, c) = 10 + 11 = 21$. Among 13, 21, 21, there is no unique maximum—okay. (Careful, this has to hold for all quadruples; how many are there?)



## Additive metrics and the four-point condition

**Theorem**
Given an $(n \times n)$ distance matrix $M$. There is an unrooted tree whose path metric agrees with $M$ if and only if $M$ defines an additive metric (i.e. if and only if the 4-point-condition holds). This tree is unique.



## Additive metrics and the four-point condition

**Theorem**
Given an $(n \times n)$ distance matrix $M$. There is an unrooted tree whose path metric agrees with $M$ if and only if $M$ defines an additive metric (i.e. if and only if the 4-point-condition holds). This tree is unique.

**Algorithm**
There are algorithms which, given $M$, compute this unrooted tree in $O(n^3)$ time (e.g. Neighbor Joining).

In fact, it is even possible to compute a "good" tree if the matrix is not additive but "almost" *(all this needs to be defined precisely, of course).*



## Summary for distance data

- When the input is a distance matrix, then we are looking for a tree whose path metric agrees with $M$.
- There are super-exponentially many trees on $n$ taxa (both rooted and unrooted).
- If the distance matrix $M$ defines an ultrametric, then a rooted tree agreeing with $M$ exists, and can be computed efficiently (i.e. in polynomial time).
- If the distance matrix $M$ defines an additive metric, then an unrooted tree agreeing with $M$ exists, and can be computed efficiently (i.e. in polynomial time).