Evolutionary Distances
Inferring Phylogenies

Luca Bortolussi[1]

[1]Dipartimento di Matematica ed Informatica
Università degli studi di Trieste
luca@dmi.units.it

Trieste, 28th November 2007

Outline

Outline

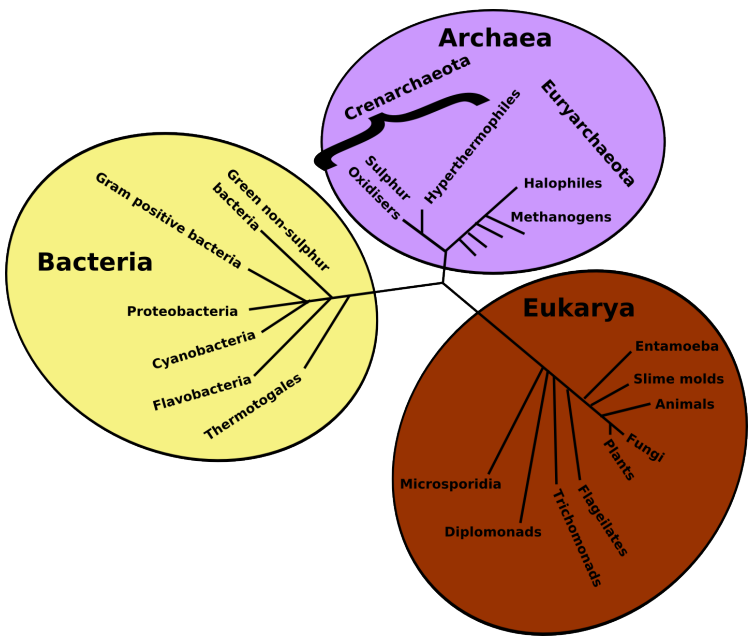
RECONSTRUCTING HISTORY OF LIFE

WHAT MEANS "PHYLOGENETIC INFERENCE"?

All species on Earth come from a common ancestor.
If we have data from a pool of species, we wish to reconstruct the history of speciation events that lead to their emergence:

We want to find the phylogenetic tree giving this information!

This is an hard task, because data is often incomplete (we lack information about most of the ancestor species) and noisy.



METHODS TO INFER PHYLOGENY

APPROACHES TO PHYLOGENY

- Distance-based methods
- Parsimony methods
- Likelihood methods
- Bayesian inference methods

DISTANCE-BASED METHODS

Given a matrix of pairwise distances, find the tree that explains it better. Several algorithms:

- UPGMA (clustering methods)
- Neighbor Joining
- Fitch-Margolias (sum of squares methods)

AN EXAMPLE: PRIMATES

DNA FROM PRIMATES

Tarsius	AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCT...
Lemur	AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCAT...
Homo Sapiens	AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCT...
Chimp	AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCT...
Gorilla	AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCAT...
Pongo	AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCT...
Hylobates	AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTT...
Macaco Fuscata	AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTT...

DISTANCE MATRIX

$$\begin{pmatrix} 0.00 & 0.29 & 0.40 & 0.39 & 0.38 & 0.34 & 0.38 & 0.37 \\ 0.29 & 0.00 & 0.37 & 0.38 & 0.35 & 0.33 & 0.36 & 0.34 \\ 0.40 & 0.37 & 0.00 & 0.10 & 0.11 & 0.15 & 0.21 & 0.24 \\ 0.39 & 0.38 & 0.10 & 0.00 & 0.12 & 0.17 & 0.21 & 0.24 \\ 0.38 & 0.35 & 0.11 & 0.12 & 0.00 & 0.16 & 0.21 & 0.26 \\ 0.34 & 0.33 & 0.15 & 0.17 & 0.16 & 0.00 & 0.22 & 0.24 \\ 0.38 & 0.36 & 0.21 & 0.21 & 0.21 & 0.22 & 0.00 & 0.26 \\ 0.37 & 0.34 & 0.24 & 0.24 & 0.26 & 0.24 & 0.26 & 0.00 \end{pmatrix} \begin{matrix} \text{Tarsius} \\ \text{Lemur} \\ \text{Homo Sapiens} \\ \text{Chimp} \\ \text{Gorilla} \\ \text{Pongo} \\ \text{Hylobates} \\ \text{Macaco Fuscata} \end{matrix}$$

Outline

What is an optimization problem?

Two ingredients

1. A search space S (possibly constrained)
2. A function $f : S \to \mathbb{R}$ to optimize:
   find $\bar{x}$ such that $f(\bar{x}) = \max_{x \in S} f(x)$.

- If S is discrete (integers, graphs, trees), then we talk of combinatorial optimization.
- If S is continuous ($\mathbb{R}$), then we talk of continuous optimization.

Bad news... (or good ones?)

"Interesting" combinatorial optimization problems are usually $\mathcal{NP}$-hard.

The space of trees

The search space is the space of trees, usually with branch lengths. Branch lengths can be optimized for a given tree topology in an easy way. The bottleneck is the identification of the correct tree topology!

Tree topologies

- Rooted vz Unrooted
- Labeled vz Unlabeled (leaves)
- Bifurcating vz Multifurcating

HOW MANY TOPOLOGIES ARE THERE?

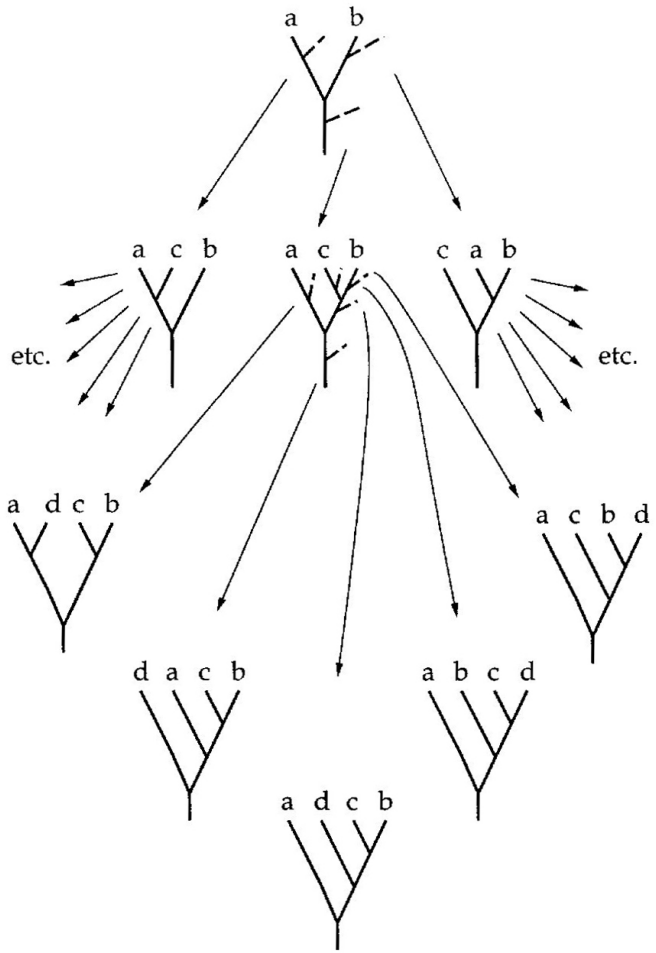
We count the rooted labeled bifurcating trees.

KEY PROPERTY

Each tree with n tips (leaves) can be obtained in a unique way from a tree with k leaves adding in sequence the remaining $n - k$ leaves.
In fact, reversing the sequence and removing the $n - k$ leaves, one obtains a unique tree with k leaves.

HOW MANY TOPOLOGIES ARE THERE?

We count the rooted labeled bifurcating trees.



TREES WITH n LEAVES

$$NT_{r,l,b}(n) = 3 \cdot 5 \cdots (2n-3) = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

How many topologies are there?

We count the rooted labeled bifurcating trees.

Species	Number of Trees
3	3
4	15
5	105
6	945
7	10.395
8	135.135
9	2.027.025
10	34.459.425
15	213.458.046.676.875
20	8.200.794.532.637.891.559.375
30	$4,9518 \times 10^{38}$
40	$1,00985 \times 10^{57}$
50	2.75292×10^{76}

Trees with n leaves

$$NT_{r,l,b}(n) = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Searching the tree space: heuristic methods

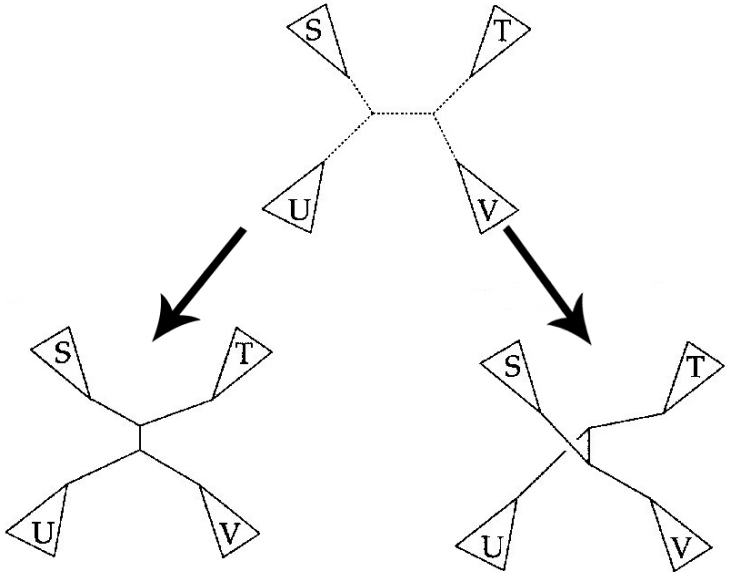
- The $\mathcal{NP}$-hardness of optimization problems on the tree space calls for heuristic solutions.
- Heuristic optimization algorithms usually search the space by following a trajectory that hopefully will hit good solutions, even the optimal one (trajectory-based heuristics).
- The next point of the trajectory is chosen in the neighbor of the current one.

Neighborhoods for the tree space

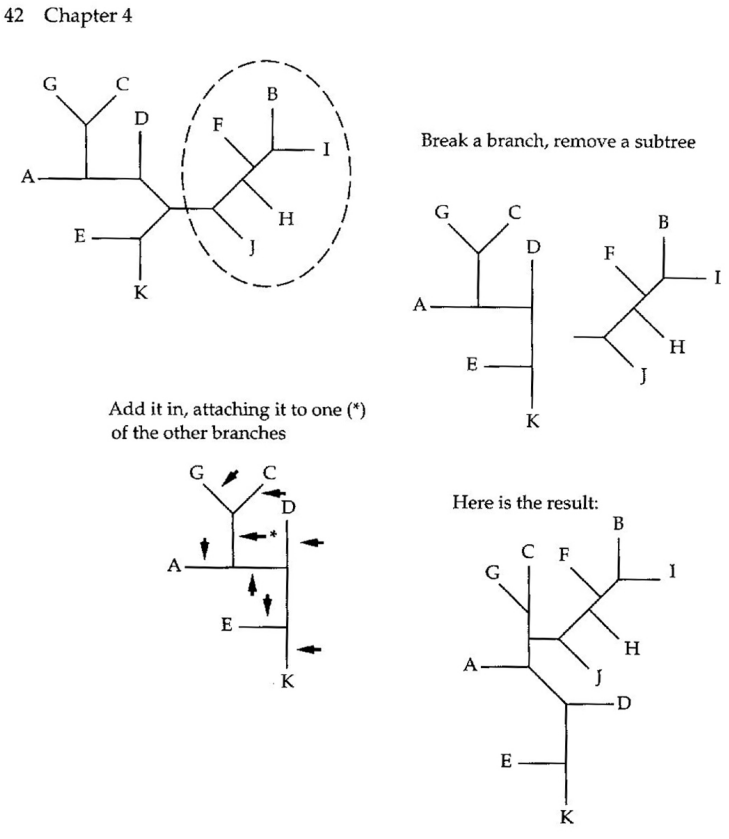
- Nearest-neighbor interchanges
- Subtree pruning and regrafting

NEIGHBORHOODS FOR THE TREE SPACE

Nearest-neighbor interchanges

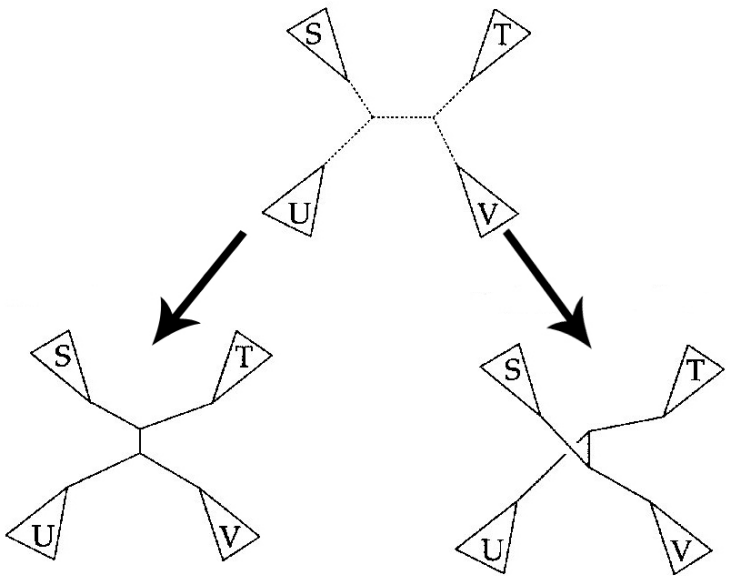


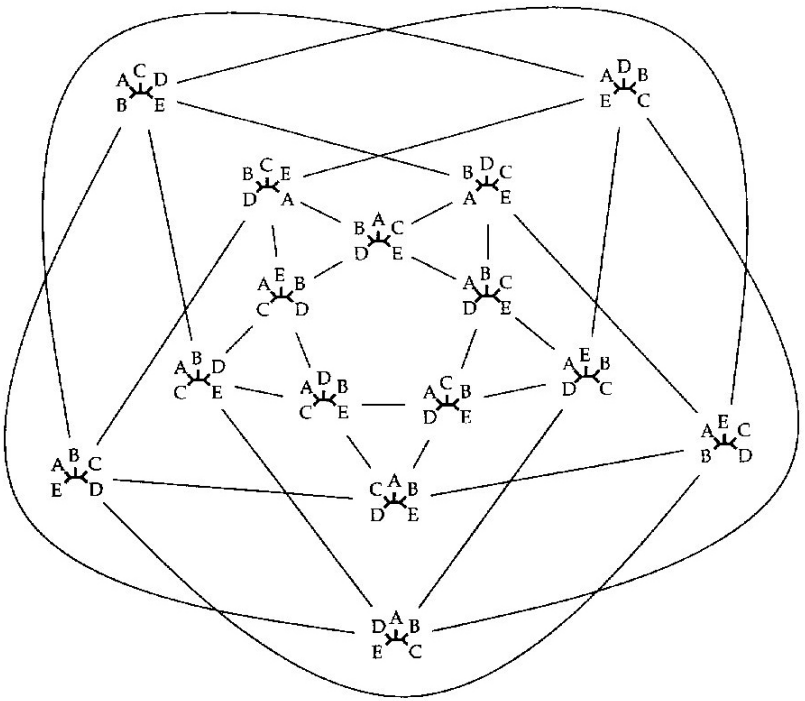
Subtree pruning and regrafting



NEIGHBORHOODS FOR THE TREE SPACE

Nearest-neighbor interchanges

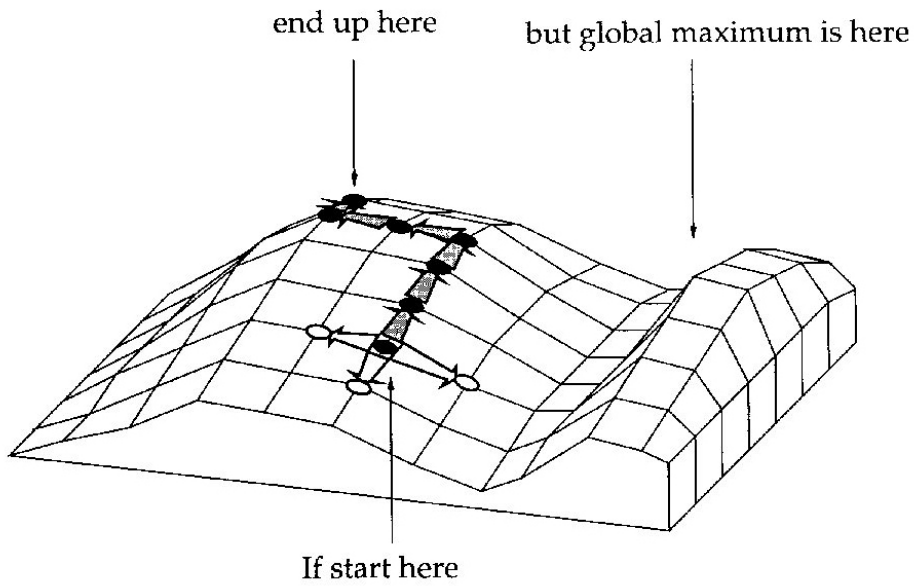




HEURISTIC ALGORITHMS

TRAJECTORY METHODS

- Local Optimization
- Greedy Search
- Simulated Annealing
- GRASP
- Taboo Search



Heuristic algorithms

Trajectory methods

- Local Optimization
- Greedy Search
- Simulated Annealing
- GRASP
- Taboo Search

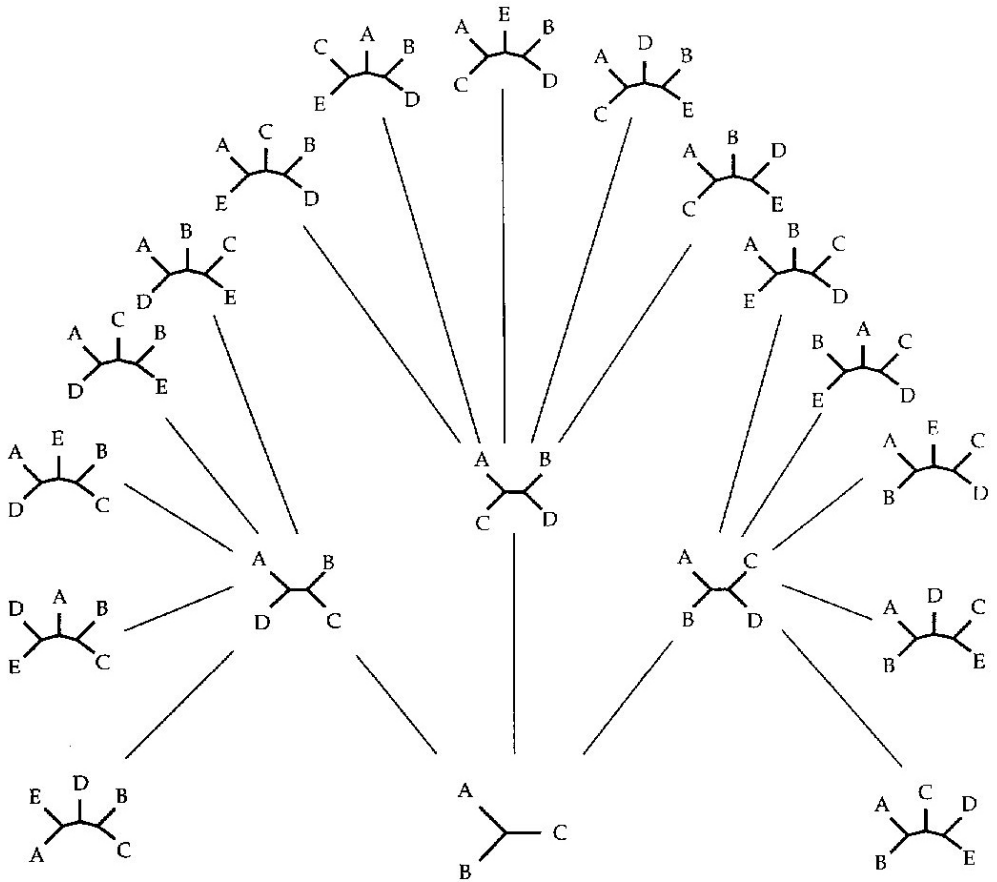
Population Methods

- Genetic Algorithms
- Ant Colony Optimization
- Particle swarm optimization

Branch and Bound Search

- Branch and Bound is a technique that explores a portion of the search space that is guaranteed to contain the optimal solution. It is an intelligent exhaustive search.
- Trees are constructed incrementally, adding one species at time. If the current tree is known to lead to a worse solution to the best one found so far, the algorithm backtracks and reconsider previous choices.
- The algorithm needs to estimate upper and lower bounds on the function to be optimized from partial solutions.
- It worst complexity is exponential, but it works well in practice.

BRANCH AND BOUND - SEARCH TREE



Branch and Bound can be seen as a visit of a search tree, pruning some subtrees (according to bounds) and backtracking before reaching leaves.

Outline

Least Square Method

We have our observed distance matrix D_{ij} and a tree T with branch lengths predicting an additive distance matrix d_{ij}.

Target

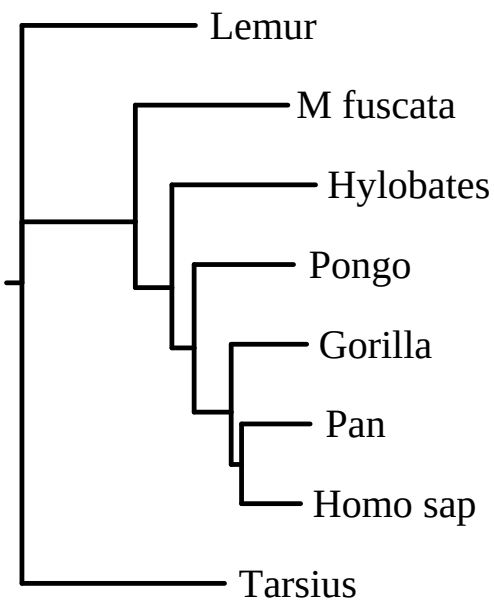
Find the tree T minimizing the error between d_{ij} and D_{ij}, i.e. the tree minimizing the weighted least square sum

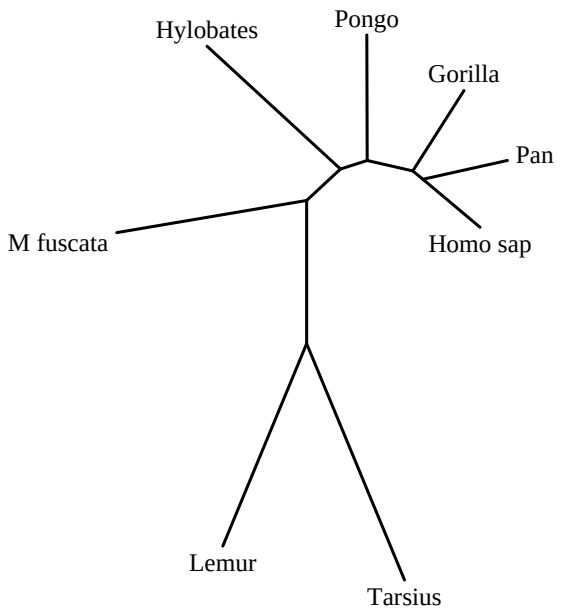
$$S(T) = \sum_{i,j} w_{ij}(D_{ij} - d_{ij})^2$$

- Given a tree topology, the best branch lengths for S can be computed by solving a linear system.
- A least square algorithms needs to search the tree space for the best tree T: this is an $\mathcal{NP}$-hard problem.
- The search for the best tree can use branch and bound methods or heuristic state space explorations.
- This method gives the best explanation of the data

FITCH-MARGOLIAS ALGORITHM

- Letting $w_{ij} = \frac{1}{D_{ij}^2}$ in $S(T) = \sum_{i,j} w_{ij}(D_{ij} - d_{ij})^2$, we obtain the method of Fitch-Margoliash.
- The choice of $\frac{1}{D_{ij}^2}$ has statistical reasons: it takes into account the variance from the expected additive distances.





Outline

Heuristic methods: UPGMA

Hierarchical clustering

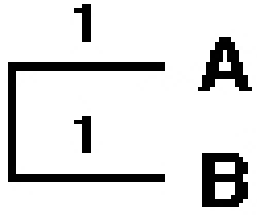
- Hierarchical clustering works by iteratively merging the two closest clusters (sets of elements) in the current collection of clusters.
- It requires a matrix of distances among singletons.
- Different ways of computing intercluster distances give rise to different HC-algorithms.

UPGMA

- UPGMA (Unweighted Pair Group Method with Arithmetic mean) computes the distance between two clusters as $d(A, B) = \frac{1}{|A||B|} \sum_{i \in A, j \in B} d_{ij}$.
- When two clusters A and B are merged, their union is represented by their ancestor node in the tree.
- The distance between A and B is evenly split between the two branches entering in A and B

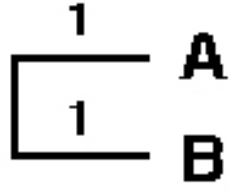
UPGMA - AN EXAMPLE

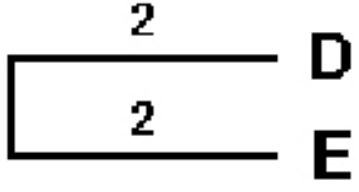
	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8



$$d([A,B],C) = \frac{1}{2}(d(A,C) + d(B,C)) = 4$$

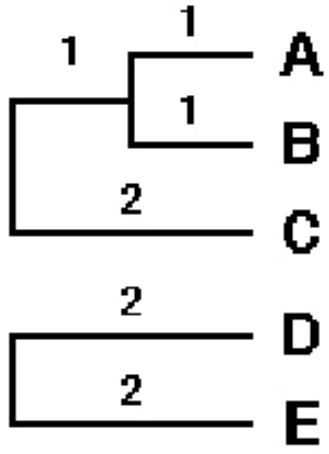
	A,B	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8



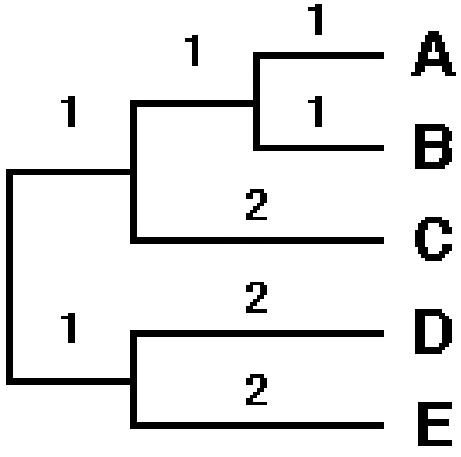


UPGMA - AN EXAMPLE

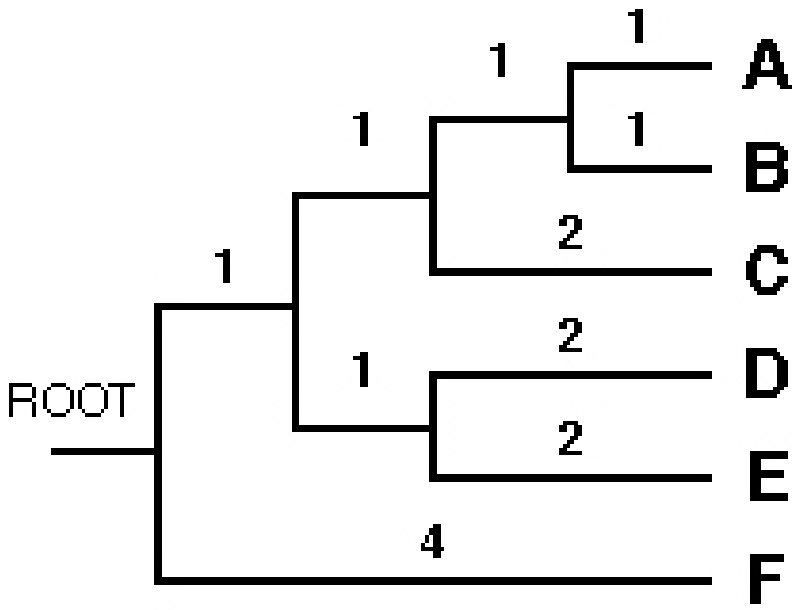
	A, B	C	D, E
C	4		
D, E	6	6	
F	8	8	8



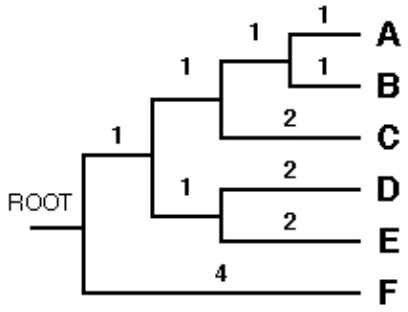
	A, B, C	D, E
D, E	6	
F	8	8

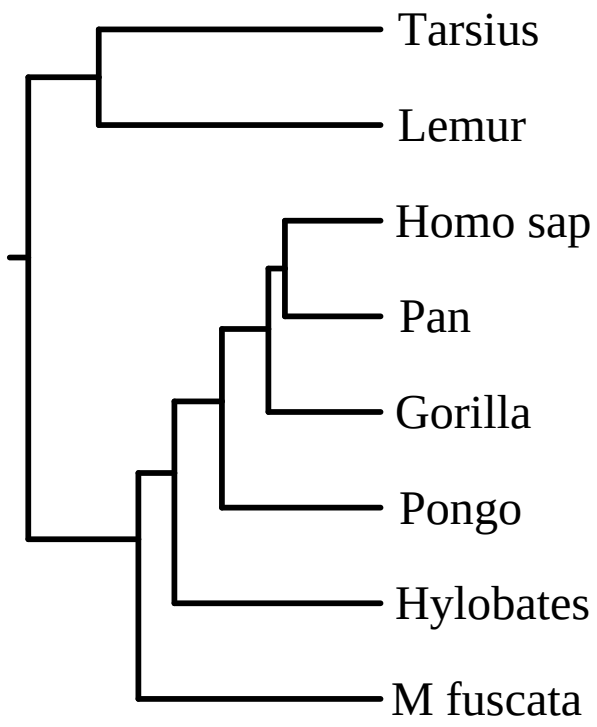


UPGMA - AN EXAMPLE



UPGMA - CONSIDERATIONS





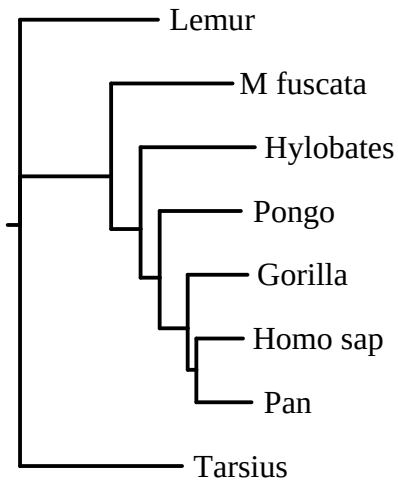
HYPOTHESIS

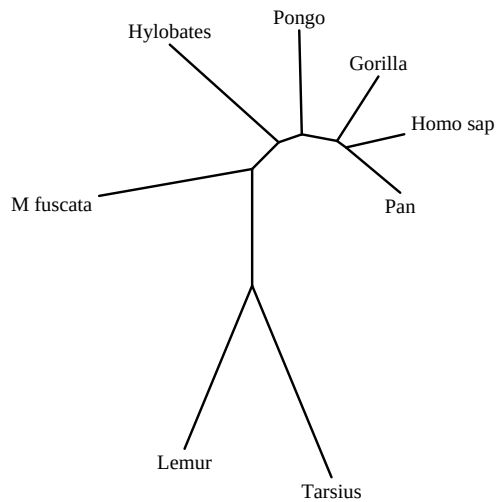
UPGMA reconstructs correctly the tree if the input distance is an ultrametric (molecular clock).

Heuristic Methods: Neighbor-Joining

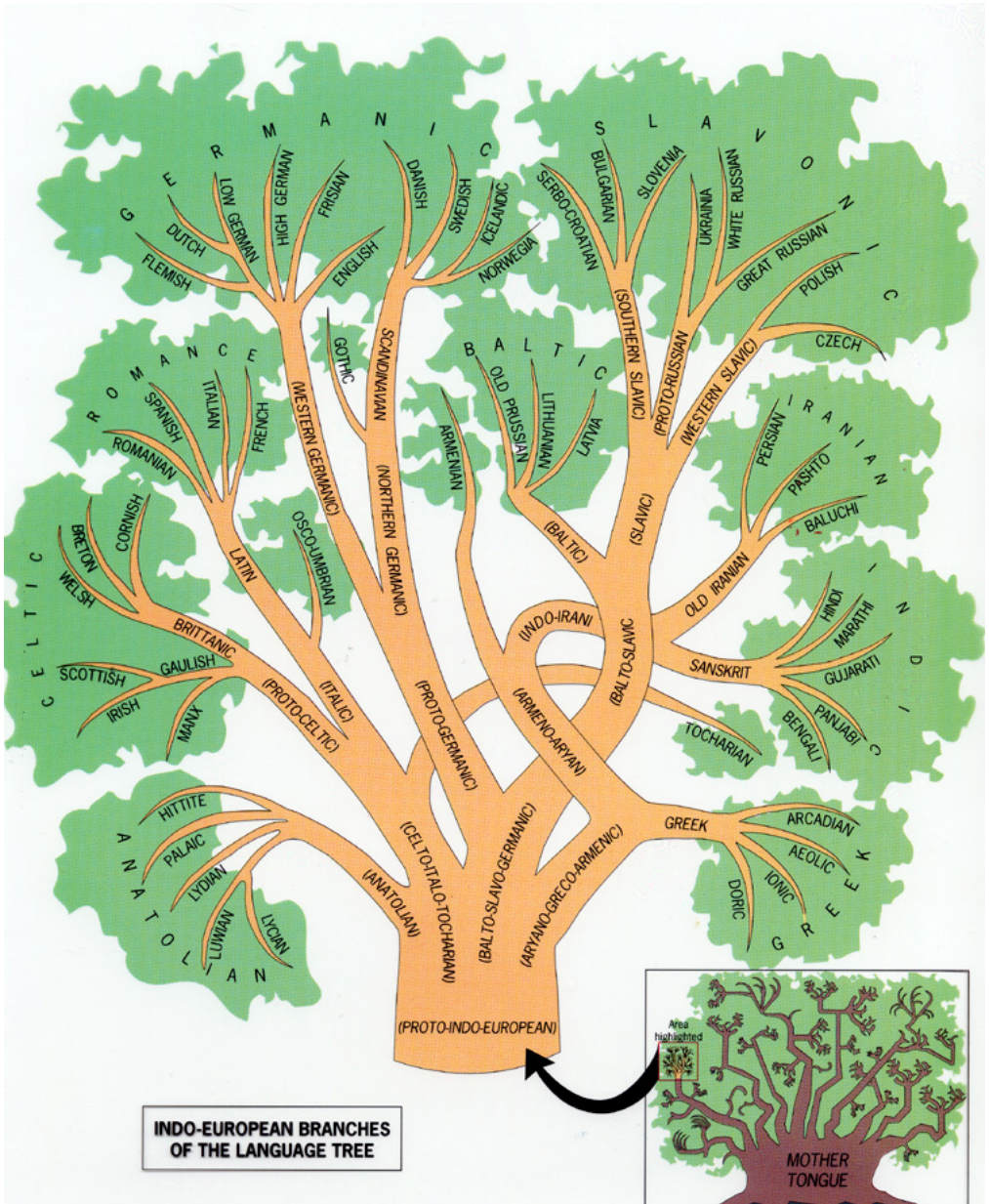
Neighbor-Joining

- Neighbor-Joining works similarly to UPGMA, but it merges together the two clusters minimizing $D_{ij} = d_{ij} - r_i - r_j$, where $r_i = \frac{1}{C-2} \sum_k d_{ik}$ is the average distance of i from all other nodes.
- When i and j are merged, their new ancestor x has distances from another node k equal to $d_{xk} = \frac{1}{2}(d_{ik} + d_{jk} - d_{ij})$
- The branch lengths are $d_{ix} = \frac{1}{2}(d_{ij} + r_i - r_j)$ and $d_{jx} = \frac{1}{2}(d_{ij} + r_j - r_i)$.
- NJ reconstructs the correct tree if the input distance is additive.





NOT ONLY DNA EVOLVE...



The end

Thanks for the attention!