

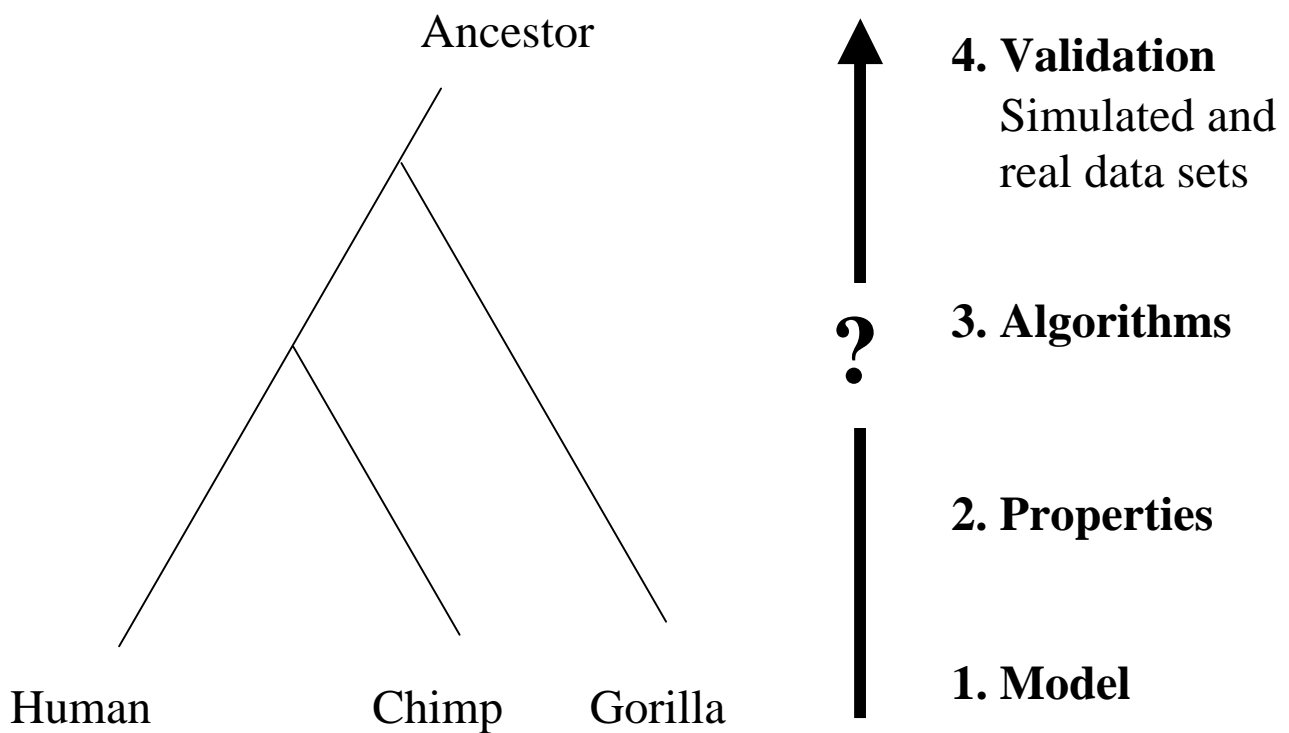
# Phylogeny reconstruction from sequence data

## Basic principles and methods

Olivier Gascuel

Montpellier – FRANCE, [gascuel@lirmm.fr](mailto:gascuel@lirmm.fr)

---



Human: ACTGAAGA . . . .

Chimp: ACAAATGA . . . . .

Gorilla: AAAAATCGA . . . . .

**Observation**

# Outline

---

## **1. Introduction**

1.1 Multiple alignment

1.2 An ideal case

1.3 A bad case

1.4 Basic facts and principles

## **2. Parsimony**

## **3. Stochastic models of sequence evolution**

## **4. Distance methods**

## **5. Maximum-Likelihood approaches**

## **6. Discussion**

## 1.1 Multiple alignment

---

- We use homologous sequences descending through Evolution from a unique ancestral sequence.
- We assume that the evolution of sequences results from simple mutational mechanisms such as substitution, insertion and deletion.
- The first step is to align the sequences in order to find the homologous sites. We search for “conserved blocks”:

$S_1$  = AGAATAGCCA  
 $S_2$  = AGGATAGGA  
 $S_3$  = AGTATGGA

A possible alignment is:

		0	1	2	3	4	5	6	7	8	9
$S_1$	:	A	G	A	A	T	A	G	C	C	A
$S_2$	:	A	G	G	A	T	A	G	G	.	A
$S_3$	:	A	G	T	A	T	.	G	G	.	A

The sites 0,1,2,3,4 are assumed to be homologous (descending from a unique ancestral site and only subject to substitution) and retained. The other sites are eliminated.

## 1.1 Multiple alignment

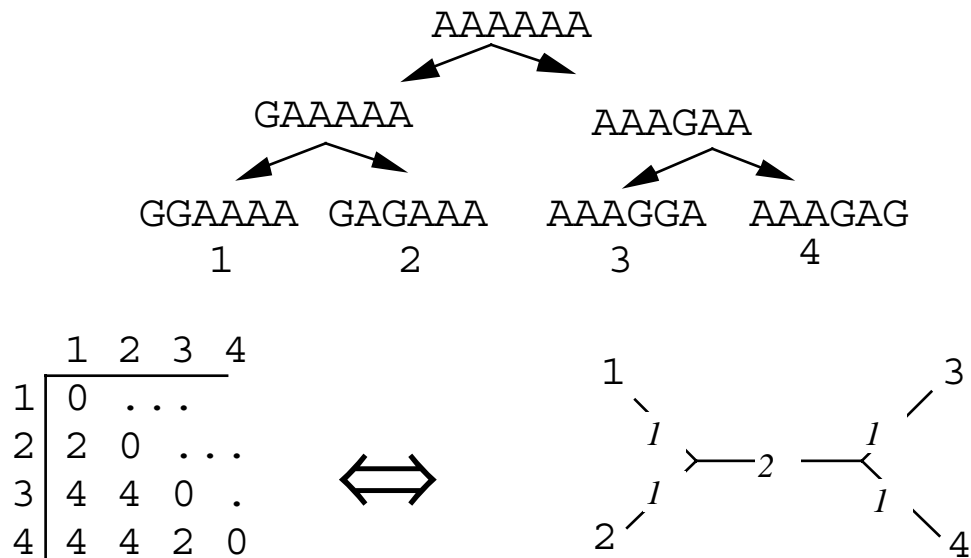
---

- Once the sequences have been aligned, we dispose of a `taxa×sites` array.
- Alignment is a crucial and difficult step, especially with proteins.
- Sequences must be sufficiently close, otherwise any alignment is doubtful.

**From now, we assume that sequences have been aligned and that the sites have only been subject to substitution events**

## 1.1 An ideal case (distance analysis)

---



- This equivalence comes from the four-point condition (Zaretskii 1965, Buneman 1971) which establishes that when the following (additive) inequality holds, the distance matrix is represented by a unique positively valued tree:

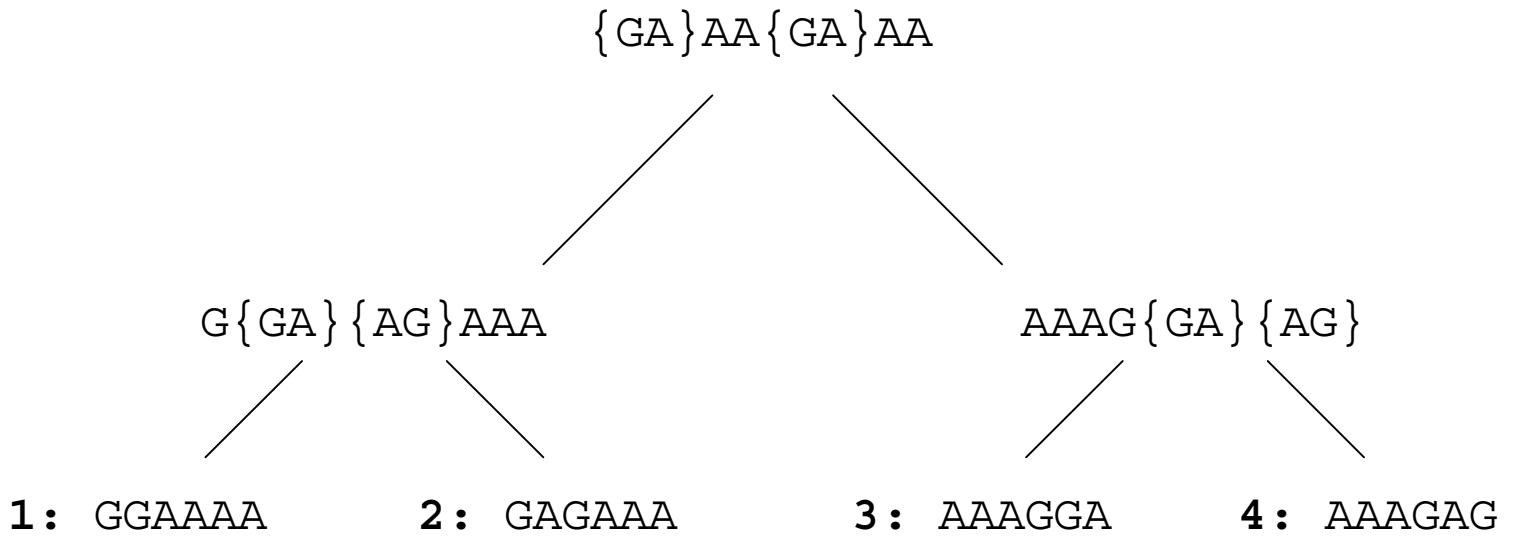
$$\forall i, j, k, l, \text{ the two largest of } (d_{ij} + d_{kl}), (d_{il} + d_{jk})$$

$$\text{and } (d_{ik} + d_{jl}) \text{ are equal.}$$

- We have lost the position of the root (unless molecular clock or outgroup) and the values of ancestral sequences.

## 1.1 An ideal case (parsimony analysis)

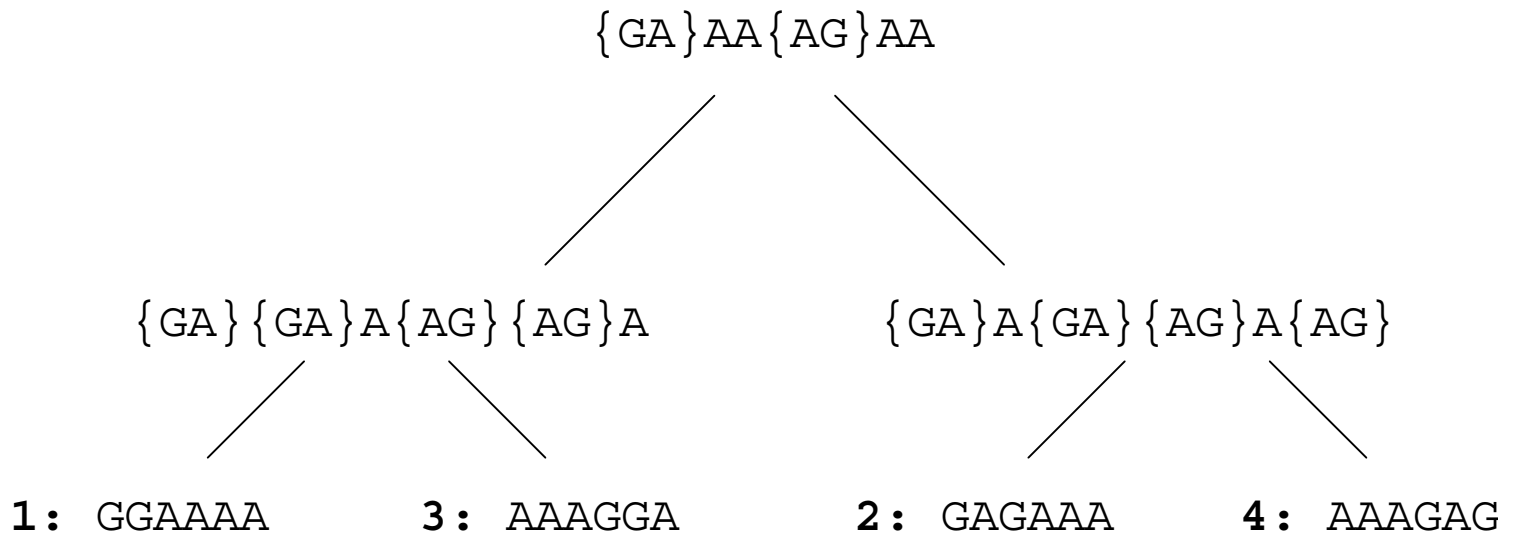
---



**6 substitutions are needed**

## 1.1 An ideal case (parsimony analysis)

---

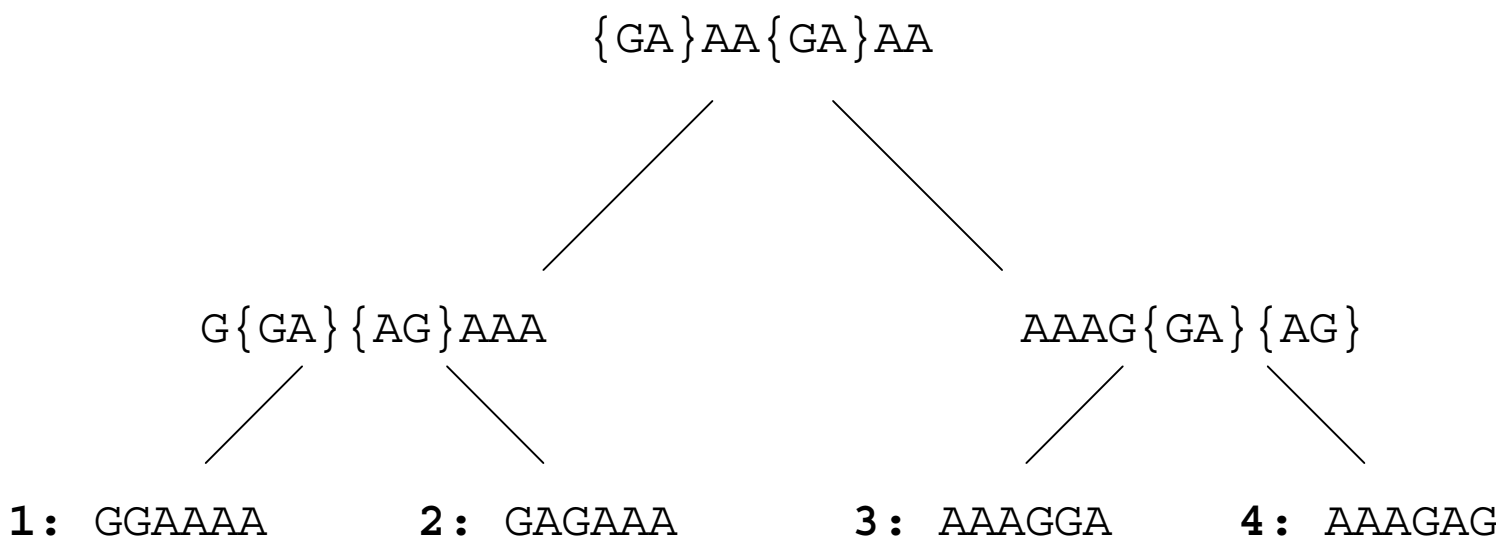


**8 substitutions are needed**

## 1.1 An ideal case (parsimony analysis)

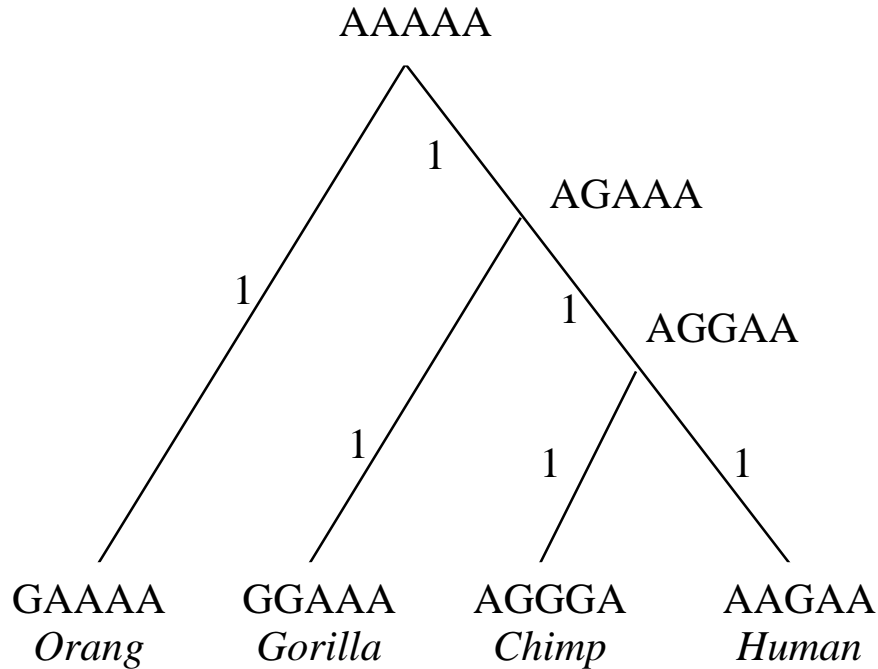
---

- The tree with best parsimony is the true tree:

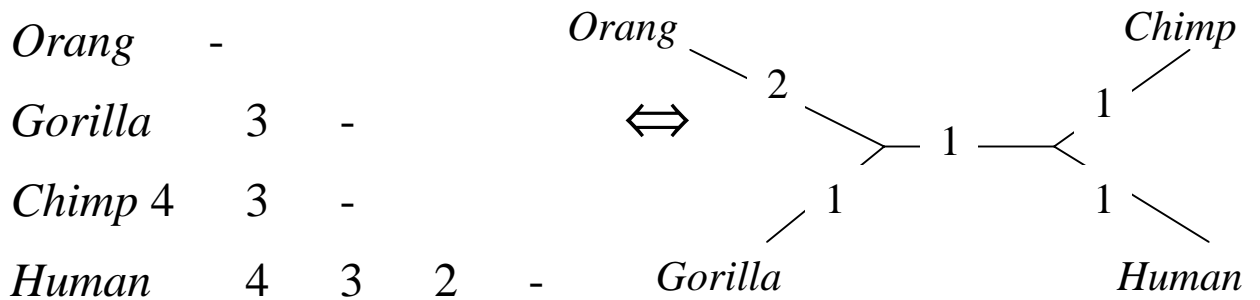


- We have reconstructed most of “ancestral” sequences.
- But we have still lost the root position (unless outgroup).

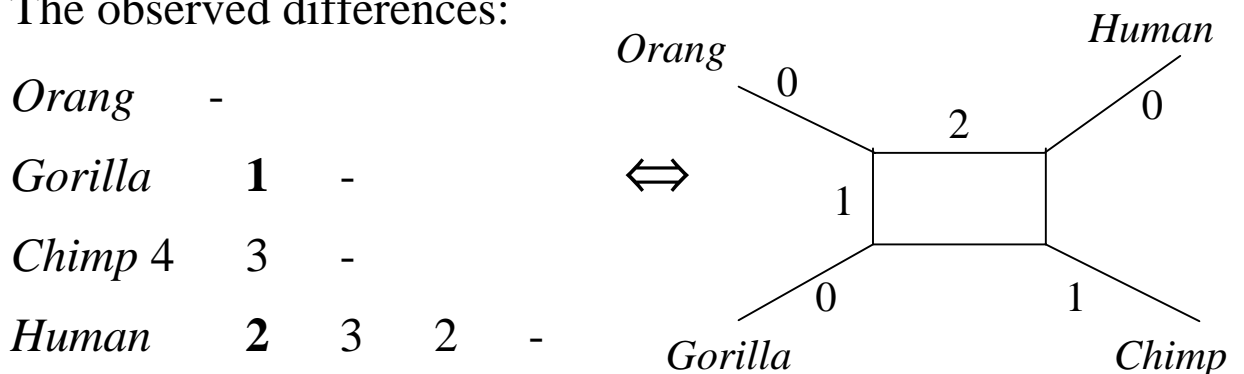
### 1.3 A (very) bad case



- The real number of substitutions:



- The observed differences:



## 1.4 Basic facts and principles

---

- **The real number of substitutions defines a tree distance that is represented by the true tree  $T$ .**
  
- **But (unfortunately) the number of observed changes under-estimates the number of substitutions that effectively occurred.**

## 1.4 Basic facts and principles

---

**Then, we can:**

- **Parsimony:** Assume that multiple substitutions are relatively rare and uniformly distributed among the sites and the tree branches; we then search for the most parsimonious tree.
- **Distance:** Assume a stochastic model of sequence evolution, use this model to estimate the true number of substitutions from the observed differences, build a tree that fits as well as possible these estimated evolutionary distances.
- **Maximum-Likelihood:** Assume a stochastic model of sequence evolution, and search for the tree with maximum-likelihood according to this model.

## 2. Parsimony methods

---

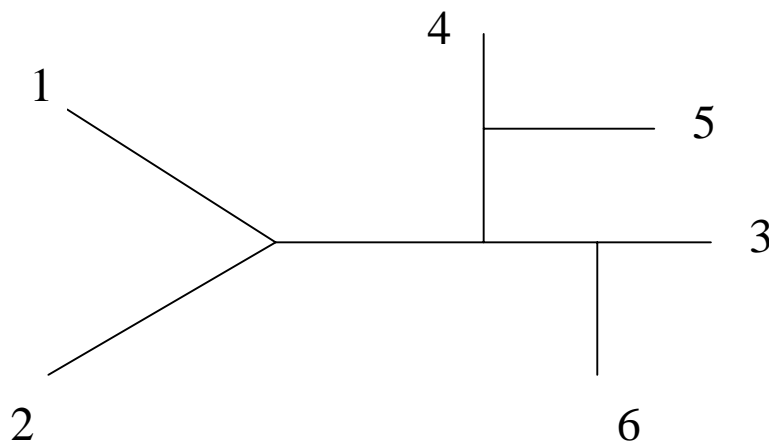
- Computing the parsimony value of any given tree is easy and in  $O(ns)$ , i.e. proportional to the number  $n$  of taxa and  $s$  the sequence length.
- But the number of trees is exponential in  $n$ , and equal to  $(2n - 5)(2n - 7)(2n - 9) \dots 1$ , e.g.  $\approx 2 \times 10^6$  when  $n = 10$ .
- Computing the best tree is NP-Hard, i.e. can require an exponential computing time in the bad cases.
- So we most often use heuristic algorithms, which only provide near-optimal trees.
- The Hendy-Penny branch-and-bound algorithm performs a careful search of the space of all trees, provides all optimal trees, but is limited to about  $n = 20$ .

## 2. Parsimony methods

---

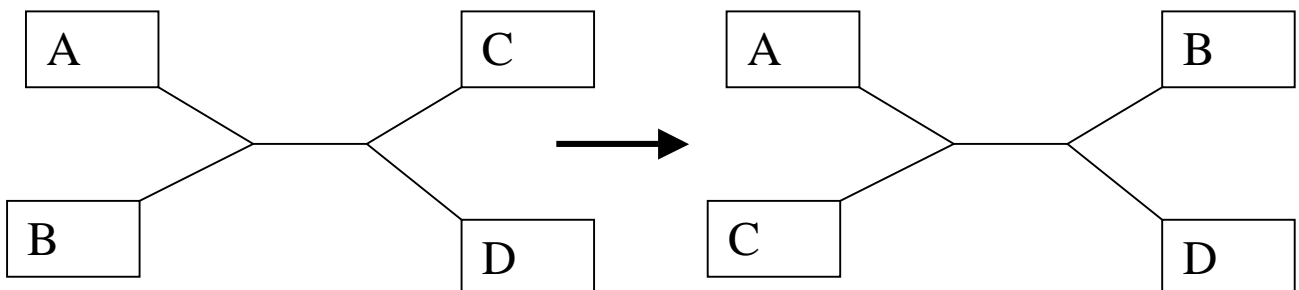
### A simple heuristic approach (DNAPARS):

**1<sup>st</sup> step:** Sequences are ordered and inserted one after the other into a growing tree; the selected insertion branch is that with minimum parsimony:



The time complexity is  $\sum_{j=4}^n sj(2j-5) = O(sn^3)$ .

**2<sup>nd</sup> step:** This initial tree is improved by tree swapping:



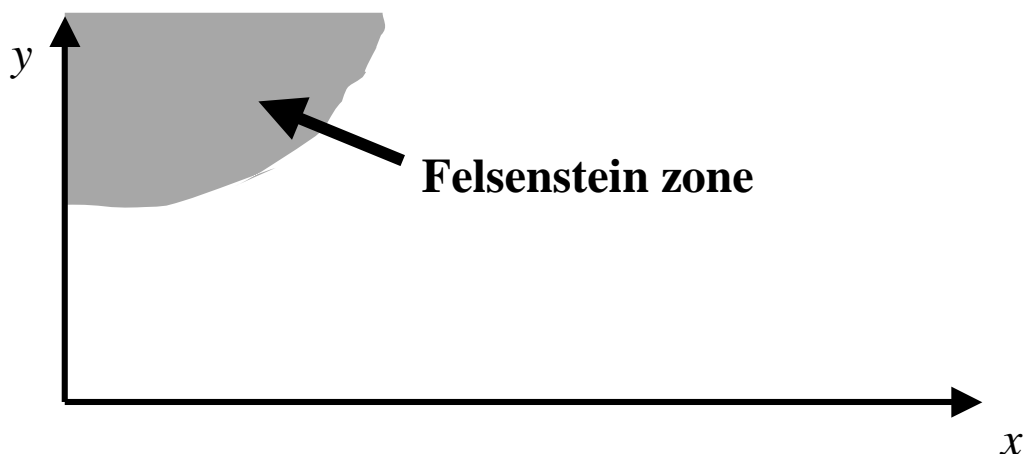
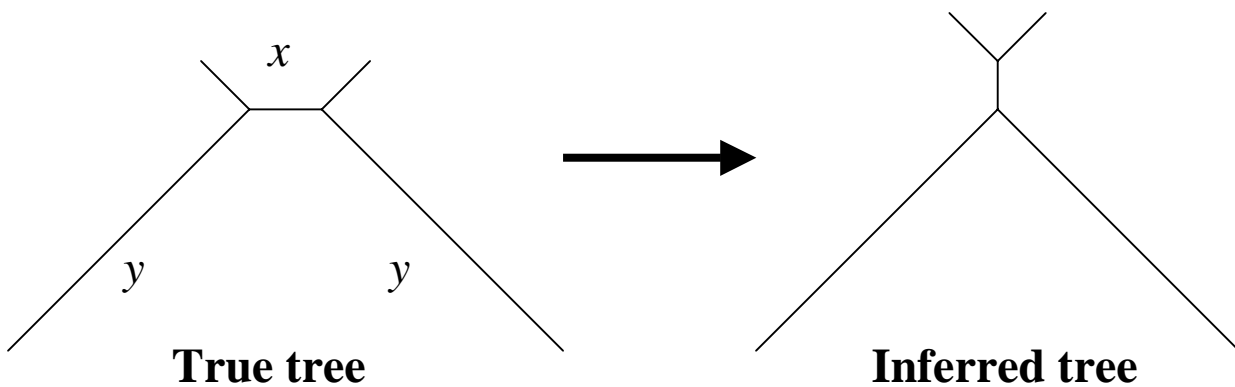
The time complexity is  $O(\#swaps \times n \times sn) \approx O(sn^3)$ .

## 2. Parsimony methods

---

### Comments:

- Relatively fast, i.e. between  $O(sn^2)$  or  $O(sn^3)$ , which allows to deal with few hundreds taxa.
- Accurate when the basic assumptions are fulfilled, i.e. when multiple substitutions are relatively rare and uniformly distributed among the sites and the branches.
- But subject to long branch attraction:



### 3. Stochastic models of sequence evolution

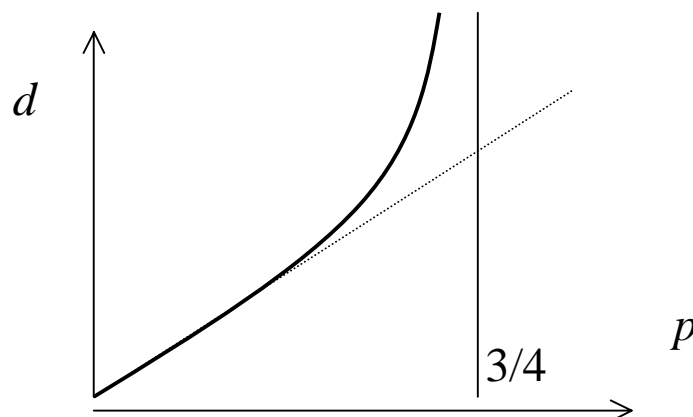
---

- To deal with multiple substitutions, we assume that mutations obey to a given stochastic process. For example, in the Jukes and Cantor model (1969) we assume that this process is markovian, i.i.d. among the sites, and that it has for shape:

$\mapsto$	A	G	C	T
A	$1-3\alpha$	$\alpha$	$\alpha$	$\alpha$
G	$\alpha$	$1-3\alpha$	$\alpha$	$\alpha$
C	$\alpha$	$\alpha$	$1-3\alpha$	$\alpha$
T	$\alpha$	$\alpha$	$\alpha$	$1-3\alpha$

- In this model, the evolutionary distance  $d$  (*i.e.*, the mutation rate) is given by ( $p$  is the probability for observing a change):

$$d = -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right),$$



### 3. Stochastic models of sequence evolution

---

- This yields the evolutionary distance estimate:

$$\delta = -\frac{3}{4} \ln\left(1 - \frac{4}{3}f\right)$$

where  $f$  is the frequency of observed changes.

### **3. Stochastic models of sequence evolution**

---

#### **Recent models are much more realistic:**

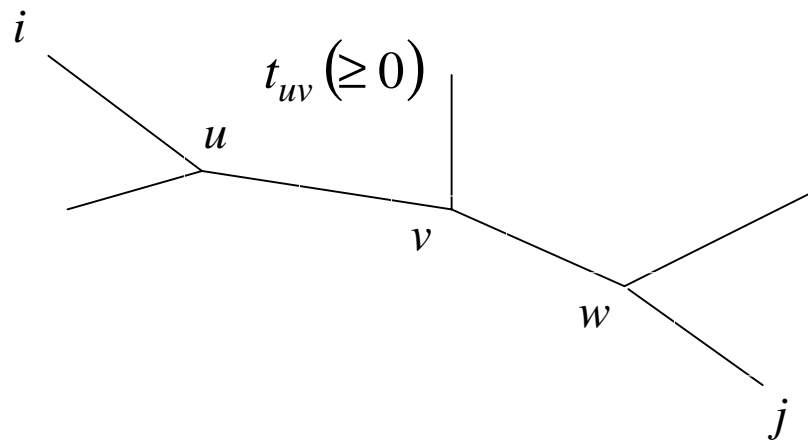
- Much more parameters, e.g., transition/transversion, different nucleotide frequencies, ...
- ... up to  $20 \times 19$  in the case of proteins (PAM, Jones-Taylor-Thornton, ...)
- Variable rates across sites (RAS)
- Variable rates across sites and subtrees (Covarion).
- Non stationarity (various GC contents within species)

**Evolutionary distances are then estimated by maximum likelihood via numerical optimization**

## 4. Distance methods

---

- We dispose of a matrix of pairwise evolutionary distance estimates  $\Delta = (\delta_{ij})$ .
- We search for the positively valued tree  $T$ , and the corresponding matrix  $(t_{ij})$ , which best fits  $\Delta$ .



- Two main approaches: least-squares or agglomerative.

## 4. Least-squares distance methods

---

- We search for the tree  $T$  that minimizes:

$$\sum_{i,j} \frac{1}{\text{Var}(\delta_{ij})} (\delta_{ij} - t_{ij})^2$$

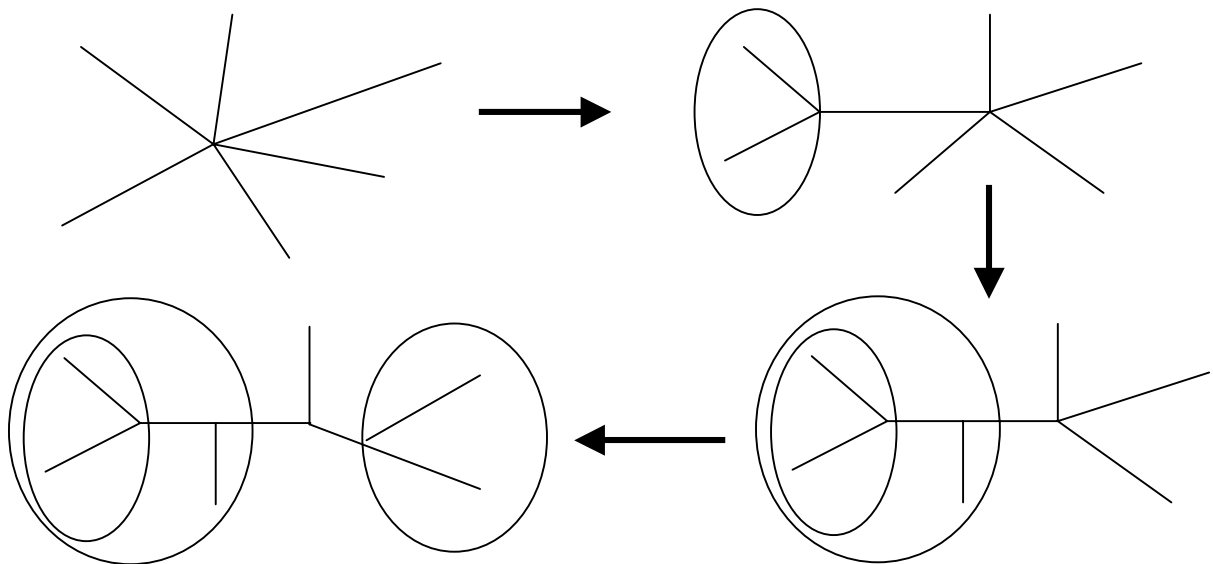
- And (Fitch and Margoliash, 1967), the variance of  $\delta_{ij}$  is estimated by  $\delta_{ij}^2$ , which is consistent with the fact that long distance estimates are much more noisy than short ones.
- This optimization problem is NP-Hard. So, practical algorithms are heuristic and use an algorithmic strategy analogous to that of DNAPARS, e.g. FITCH (Felsenstein 1997) that is in  $O(n^4)$ .
- An alternative is to use the same strategy, but to minimize the least-squares tree length estimate, e.g. BME (Desper and Gascuel, 2002) that is in  $O(n^2 \log(n))$ .

## 4. Agglomerative distance methods

---

### Iteratively:

1. **select** two taxa according to a criterion  $Q$ ;
2. **estimate** the two corresponding branch lengths and store the obtained structure;
3. **reduce** the distance matrix and replace both taxa by a unique taxon.

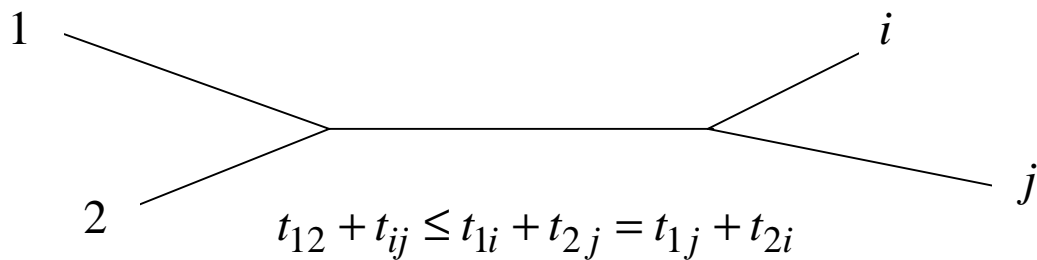


## 4. Agglomerative distance methods

---

### ADDTREE algorithm (Sattath and Tversky 1977):

- The four-point condition:



- When:  $\delta_{12} + \delta_{ij} \leq \text{Min}(\delta_{1i} + \delta_{2j}, \delta_{1j} + \delta_{2i})$ , 1 and 2 are “neighbors” for  $i$  and  $j$ .
- Agglomeration criterion (to be maximized):

$$N_{12} = \sum_{2 < i < j \leq r} H(\delta_{1i} + \delta_{2j} - \delta_{12} - \delta_{ij}) H(\delta_{1j} + \delta_{2i} - \delta_{12} - \delta_{ij})$$

(if  $x < 0$  then  $H(x) = 0$ , else  $H(x) = 1$ )

- Reduce the dissimilarity matrix using :

$$\delta_{ui} = \frac{1}{2}\delta_{1i} + \frac{1}{2}\delta_{2i}$$

- Time complexity in  $O(n^4)$ .

## 4. Agglomerative distance methods

---

- NJ (Saitou and Nei, 1987) is very close to ADDTREE, but in  $O(n^3)$ .
- BioNJ (Gascuel, 1997) incorporates the variances of the distance estimates into the reduction step, and requires the same amount of computation as NJ.
- Weighbor (Bruno et al. 2000) also incorporates these variances into the selection process; it is still in  $O(n^3)$ , but much slower than NJ and BioNJ.

## 5. Maximum Likelihood approaches

---

### Principle:

- We chose a sequence evolution model.
- Within this model, we are able to compute the likelihood of any given tree (with branch lengths), regarding the contemporary sequences at hand.
- For any given tree shape, we are then able to compute (by numerical optimization) the branch lengths and the model parameters which maximize the tree likelihood.
- Algorithmic strategy ( $\approx$  DNAPARS):
  1. We iteratively insert taxa into a growing tree
  2. The insertion branch is that with maximum likelihood
  3. For each taxon and insertion point, we have to compute the likelihood of the corresponding tree, which involves re-estimating all branch lengths (and model parameters).
  4. We apply branch-swapping at each step to improve the current tree.

## 5. Maximum Likelihood approaches

---

### Computing the likelihood ( $L$ )

- The tree  $T$  is fixed, as well as all model parameters
- The sites evolve independently. Then:

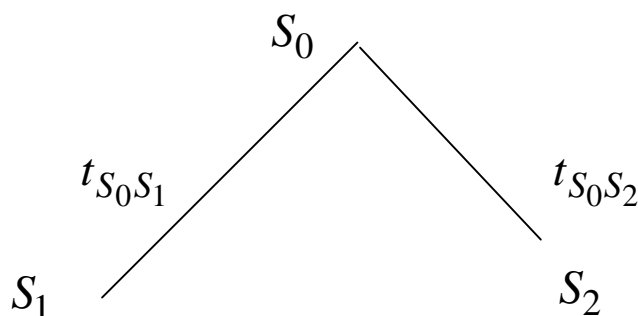
$$L(T) = \prod_{i=1}^s \sum_{c \in \{A, C, G, T\}} \pi_c L(S_{0i} = c)$$

where  $\pi_c$  is the probability of nucleotide  $c$ , and  $S_{0i}$  corresponds to the  $i$ th site of the root sequence.

- We have the recurrence formula:

$$L(S_{0i} = c) = \left[ \sum_{k \in \{A, C, \dots\}} p_{kc}(t_{S_0 S_1}) L(S_{1i} = k) \right] \left[ \sum p_{kc}(t_{S_0 S_2}) L(S_{2i} = k) \right]$$

where  $p_{kc}(t_{S_0 S_1})$  is the probability of  $k \rightarrow c$ , for an evolutionary distance (branch length) of  $t_{S_0 S_1}$ .



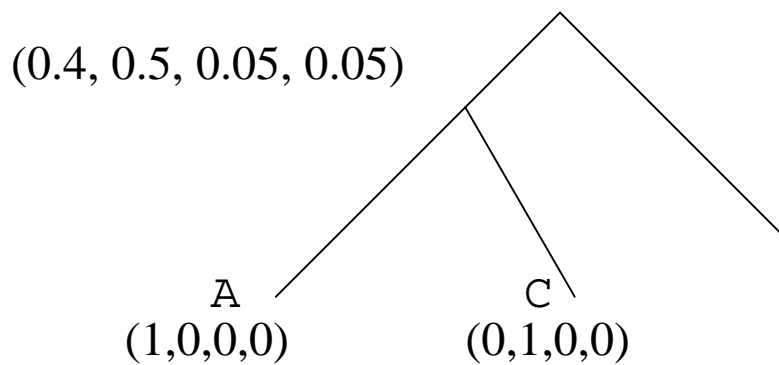
## 5. Maximum Likelihood approaches

---

- When  $S_x$  is a contemporary sequence, then:

$$L(S_{xi} = k) = 1 \text{ when } S_{xi} = k, \text{ and } 0 \text{ otherwise.}$$

- The computation of tree likelihood is analogous to that of parsimony, and requires  $O(sn)$



- Due to reversibility, the result does not depend on the root position.

## 5. Maximum Likelihood approaches

---

### Comments:

- Much slower than distance and parsimony methods, due to numerical optimization (and local minima).
- But perform best in all simulation comparisons.

## 6. Simulation comparison

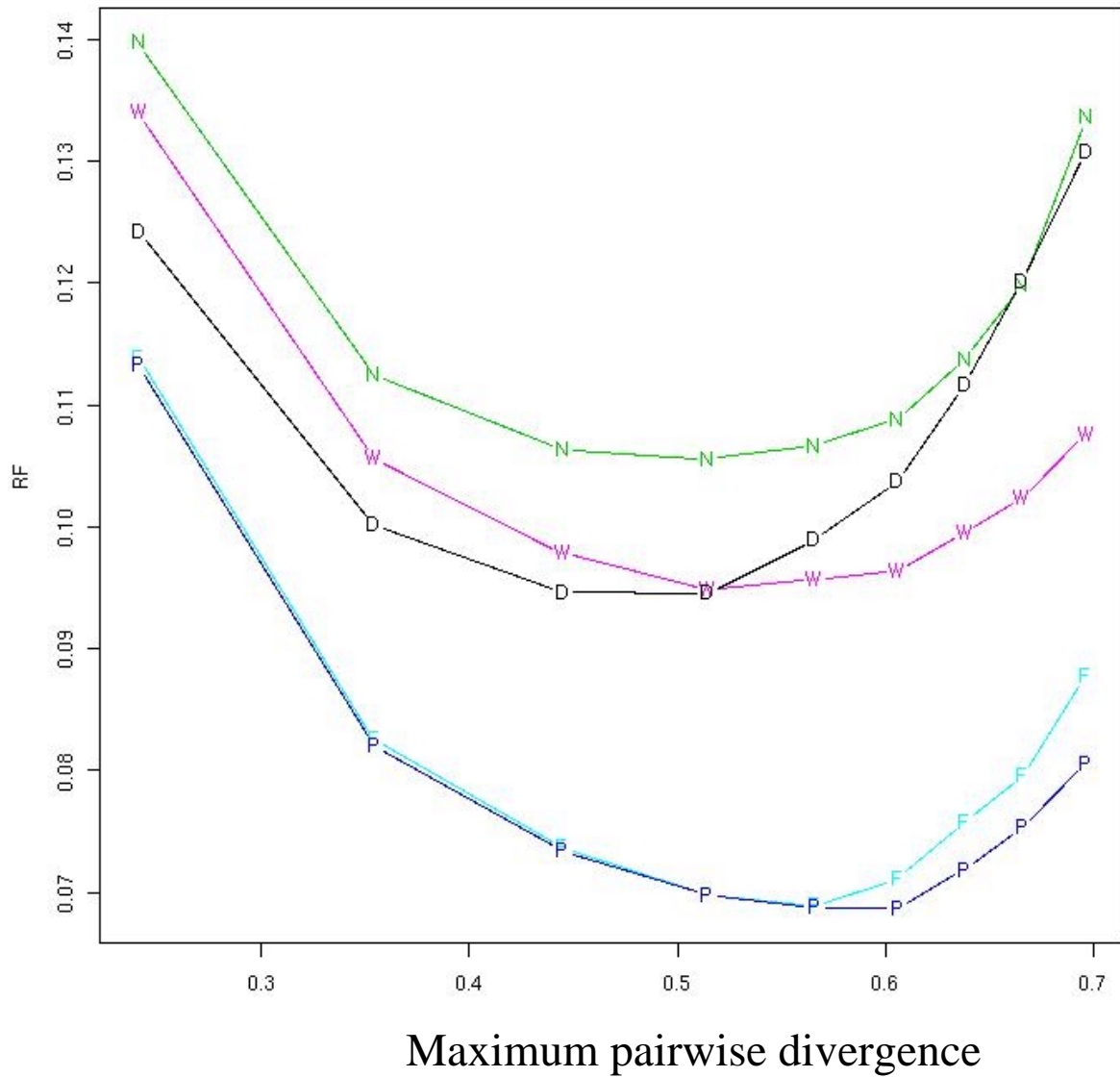
---

### Principle:

- A true tree is generated by a random speciation process (40 taxa).
- An ancestral sequence (500 sites) is generated; it evolves along the true tree according to some sequence evolution model (Kimura's).
- We reconstruct the tree from the leaf (contemporary) sequences, and compare this inferred tree with the true tree, using the topological bipartition distance.
- This process is repeated a number of times (5000).

## 6. Simulation comparison (Stéphane Guindon 2002)

### Topological accuracy



N: NJ

W: Weighbor

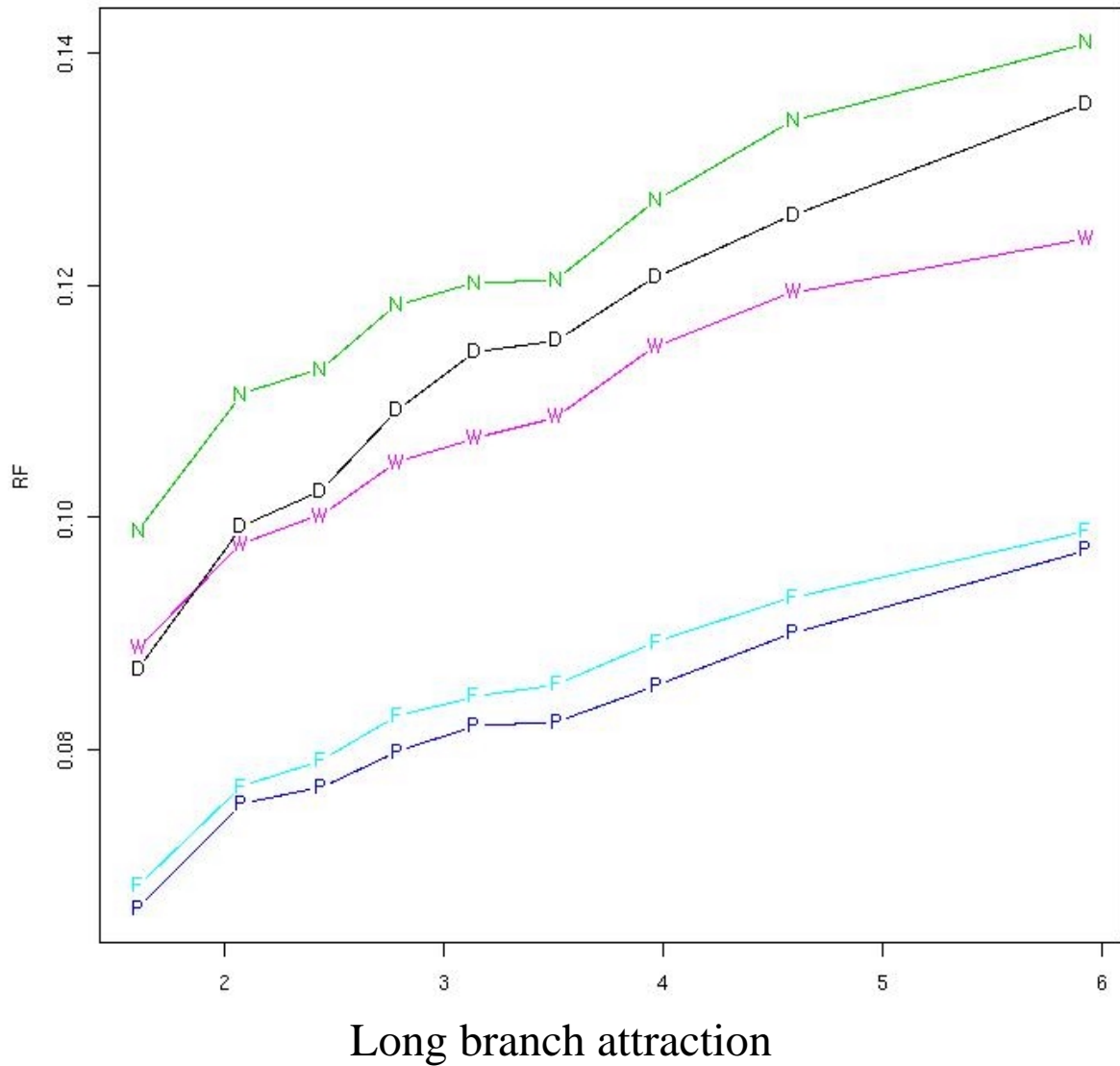
D: DNAPARS

F: FastDNAML

P: PHYML

## 6. Simulation comparison (Stéphane Guindon 2002)

### Topological accuracy



N: NJ

W: Weighbor

D: DNAPARS

F: FastDNAML

P: PHYML

## 6. Run times (40 taxa, Pentium 4 – 1.7 Ghz)

---

### For one tree:

NJ (and BME): 0.005 seconds

DNAPARS: 0.5 seconds

Weighbor: 2.0 seconds

FastDNAML: 230.0 seconds

### For 1000 trees (bootstrap):

...

FASTDNAML:  $\approx 2.5$  days.

## 6. Discussion

---

### Some crucial points:

- Gene/species trees
- Horizontal transfers
- Duplication/speciation, paralogous/orthologous sequences
- Choice of the sequence evolution model (RAS parameter?)
- Bootstrapping the data is highly recommended

SWOFFORD, D.L., G.L. OLSEN, P.J. WADDELL, and D.M. HILLIS. 1996. Phylogenetic inference. Pages 407-514 *in* Molecular Systematics (D.M. Hillis, C. Moritz and B.K. Mable, eds.). Sinauer, Sunderland, MA.

CARAUX G., GASCUEL O., ANDRIEU G. et LEVY D., "Méthodes informatiques pour la reconstruction phylogenetique", *Technique et Science Informatiques* 14(2), 113-139, 1995.

---

GUINDON, S., GASCUEL, O., "Efficient biased estimation of evolutionary distances when substitution rates vary across sites", *Molecular Biology and Evolution*, 19(4), 534-543, 2002.

ELEMENTO, O., GASCUEL, O., LEFRANC M.-P., "Reconstructing the duplication history of tandemly repeated genes", *Molecular Biology and Evolution*, 19(3), 278-288, 2002.

GASCUEL, O, BRYANT, D. and DENIS, F., "Strengths and limitations of the minimum evolution principle", *Systematic Biology*, 50(5), 621-627, 2001.

RANWEZ, V., GASCUEL, O., "Quartet based phylogenetic inference: improvements and limits," *Molecular Biology and Evolution*, 18(6), 1103-1116, 2001.