Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Phylogenetic inference

Hélène Chiapello, MIA-T, INRA Toulouse

Outline

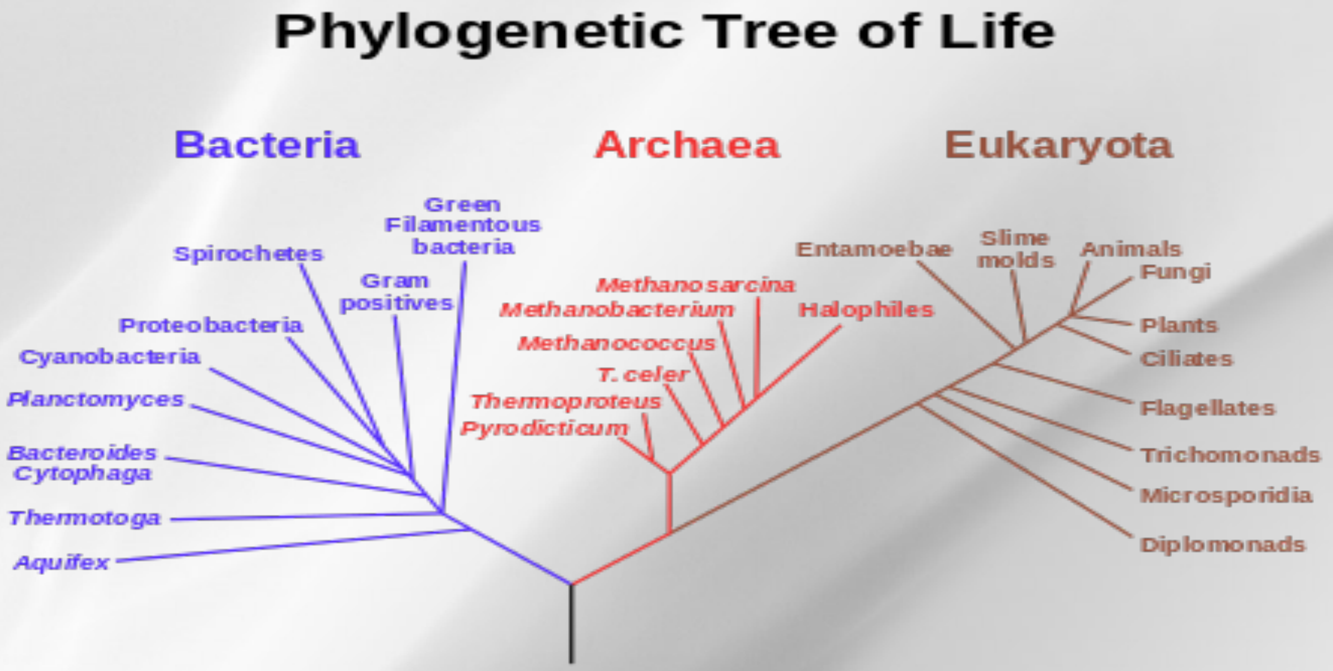
- Introduction and basic concepts in phylogeny
  - Trees
  - Genetic distances and nucleotide substitution models
- Phylogenetic inference methods
  - Distance methods
  - Parcimony methods
  - Maximum likehood methods
  - Bayesian methods
- Phylogeny in practice
  - Testing tree topologies (bootstrap)
  - How to choose a method ?

Introduction

- **Phylogenetics** is the study of evolutionary relationships among groups of organisms (e.g. species, populations)
- The result of phylogenetic studies is a hypothesis about the evolutionary history of taxonomic groups: their **phylogeny**
- **Phylogenetic methods** aims at representing similarities and differences between taxa using a **phylogenetic tree**
- Underlying asumption : taxa joined together in the tree are implied to have descended from a common ancestor **through different speciation events**

What is a phylogenetic tree ?

- In **biology**, a phylogenetic tree is a **branching diagram** for representing the inferred evolutionary relationships among various biological entities
- In **mathematics**, a tree is an **undirected graph** in which any two vertices are connected by exactly **one simple path**. In other words, any connected graph without simple cycles is a tree.


Phylogenetic Tree of Life
Bacteria
Archaea
Eukaryota
Green Filamentous bacteria
Spirochetes
Gram positives
Proteobacteria
Cyanobacteria
Planctomyces
Bacteroides Cytophaga
Thermotoga
Aquifex
Methanosarcina
Methanobacterium
Halophiles
Methanococcus
T. celer
Thermoproteus
Pyrodicticum
Entamoebae
Slime molds
Animals
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia
Diplomonads

Molecular phylogenetics

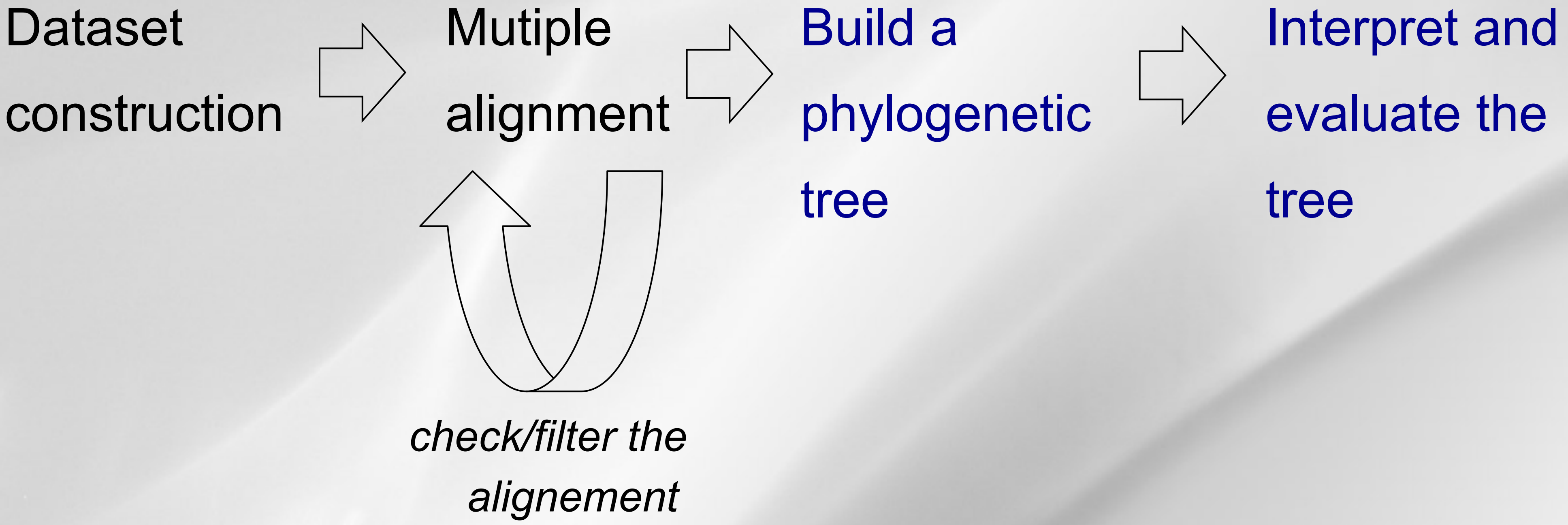
Here : we focus on **molecular phylogenetics**, based on different kind of **molecular sequence data**

Trees are infered from **heritable characters** like:

- Binary patterns : presence/absence, 0/1
- Microsattelites data, SNPs, Insertions, Deletions
- **Aligned genetic sequences** (ADN, ARN, proteins) **in most cases**

In molecular phylogenetics, we infered the evolutionary history of sequences: it is not always the same of the one of the corresponding species !!!

Usual workflow in phylogenetic analysis



Dataset construction

Mutiple alignment

Build a phylogenetic tree

Interpret and evaluate the tree

check/filter the alignement

Dataset construction

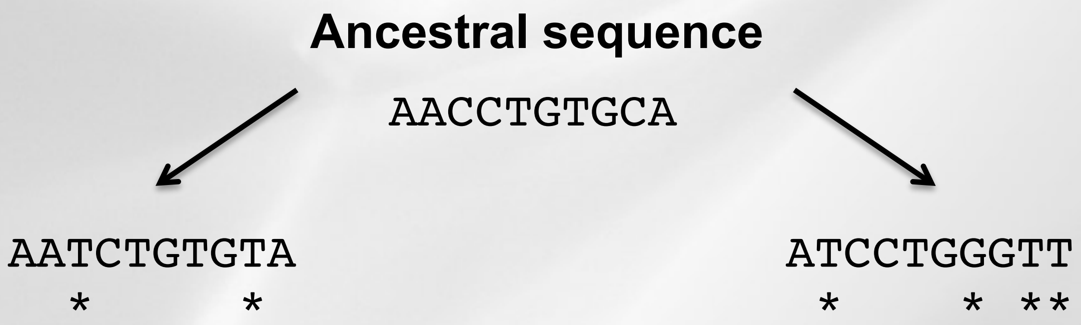
- **Criteria to choose good sequences dataset:** universality, conserved structure, no horizontal transfer, apropriate evolutionary rate.

- Some **popular genes** used in molecular phylogenetics
  - **Procaryotes**: ribosomal RNA (rRNA) 16S, betaglucosidase,…
  - **Eukaryotes**: rRNA 18S, actin, EF1, RPB1, mitochondrial genes,…

- **Protein coding genes:** nucleic alignments (if closed sequences) or proteic alignements (if distant sequences) of homolog sequences

Mutiple alignment as dataset

Hypothese: aligned sequences are **homologous, *i.e.* vertically derived from an ancestral sequence of common ancestor**



Ancestral sequence

```
AACCTGTGCA
```

```
AATCTGTGTA
  *     *
```

```
ATCCTGGGTT
 *    * **
```

```
Seq1 AATCTGTGTA
Seq2 ATCCTGGGTT
      **   *  *
```

⇦ 1 site = 1 position in the alignment

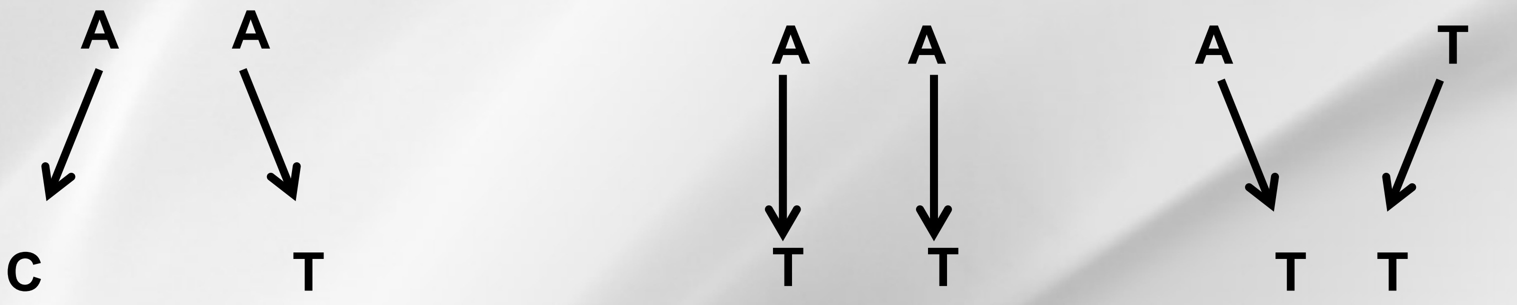
In phylogeny we will focus on **sites** of the alignment, either directly or indirectly via computation of a distance.

Homology vs Homoplasy

Homology is any similarity between shared characters that is due to their shared ancestry

Homoplasy occurs when characters are similars, but are not derived from a common ancestor

Homoplasies often result from **parallel** or **convergent** evolution



A A

C T

A A

T T

A T

T T

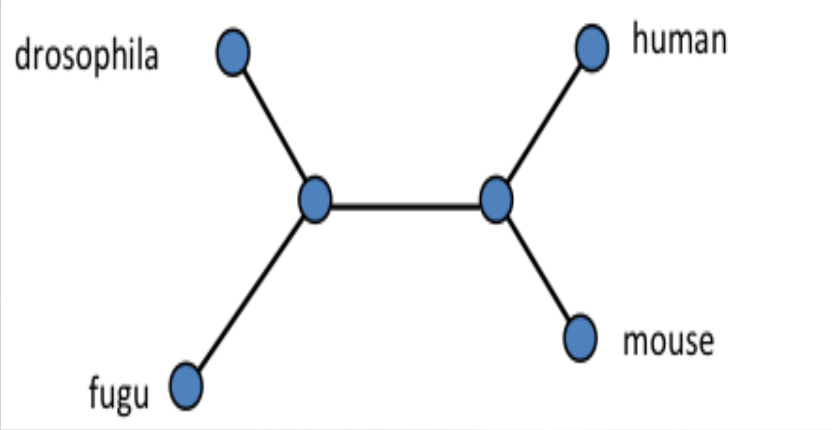
Phylogenetic inference should distinguish homoplasies from real phylogenetic signal

Quality of the genetic dataset is essential !

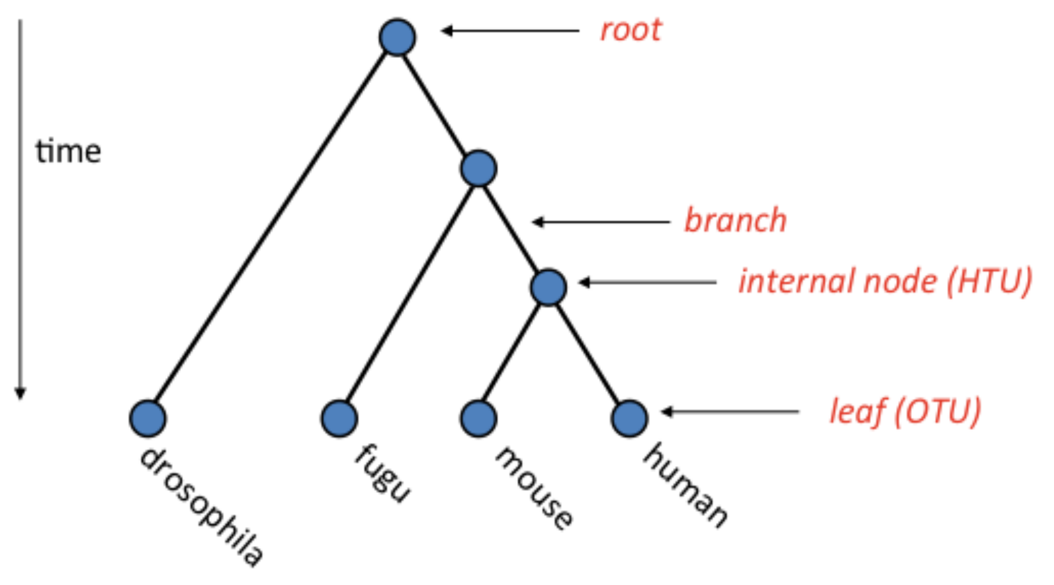
Phylogenetic tree: terminology

Structure of an **unrooted** (a) and a **rooted phylogenetic tree** (b)

(a)


drosophila
human
fugu
mouse

(b)


time
root
branch
internal node (HTU)
leaf (OTU)
drosophila
fugu
mouse
human

A tree is defined by its **topology** and its **branch lengths**.

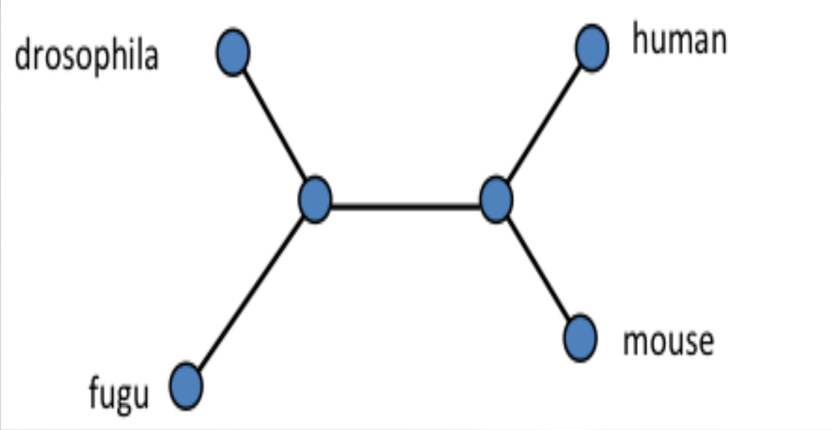
Taxa are often named

- OTU: Operational Taxonomic Units
- HTU: Hypothetical Taxonomic Units

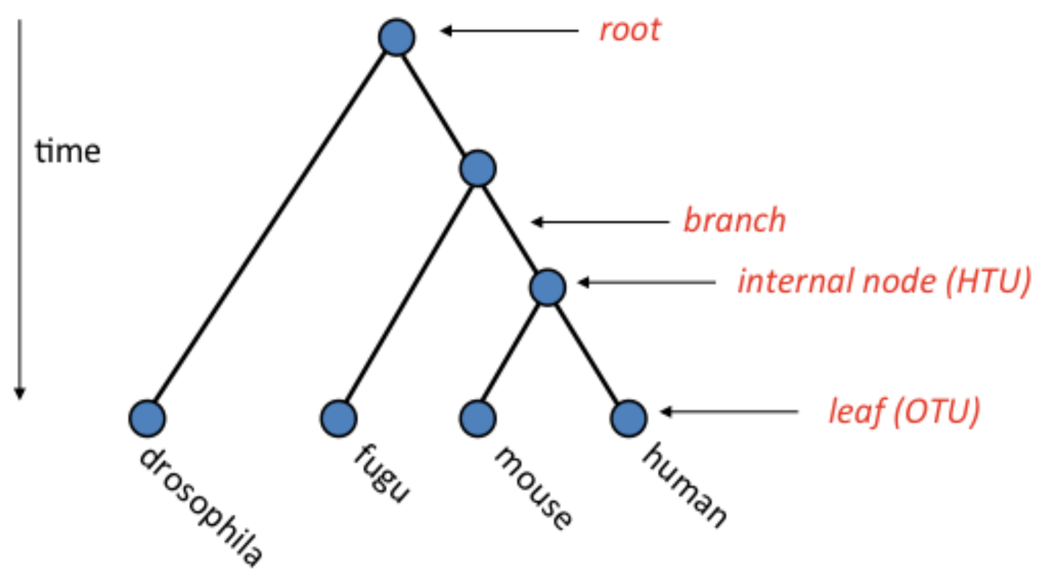
Phylogenetic tree: terminology

Structure of an **unrooted** (a) and a **rooted phylogenetic tree** (b)

(a)


drosophila
human
mouse
fugu

(b)


root
time
branch
internal node (HTU)
leaf (OTU)
drosophila
fugu
mouse
human

- Phylogeny focus on **bifurcating trees** : each internal node is of degree 3
- Most phylogenetical methods produce **unrooted trees**

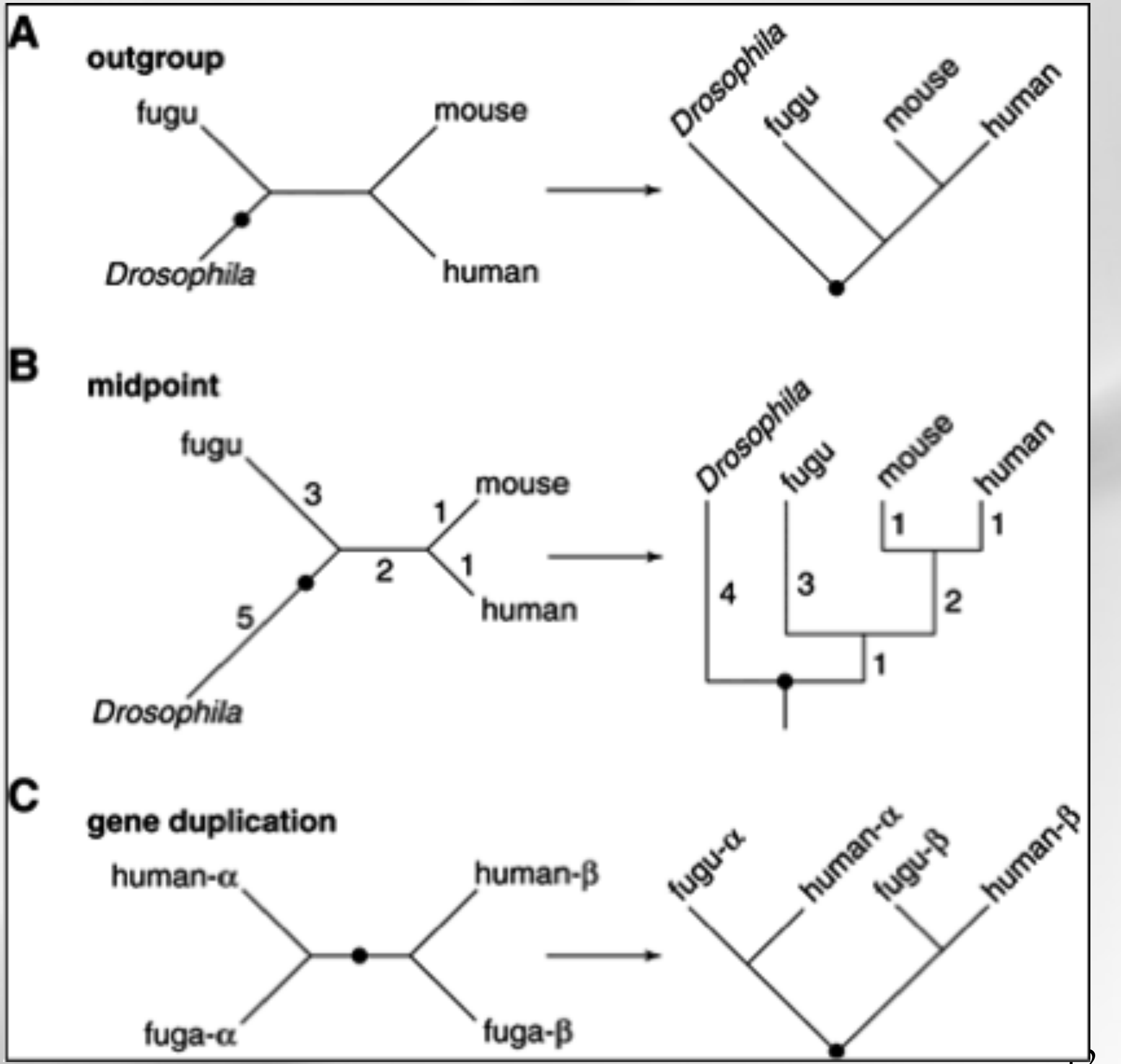
Introduction: how rooting a tree ?

Three methods exist:

A. **Outgroup** rooting

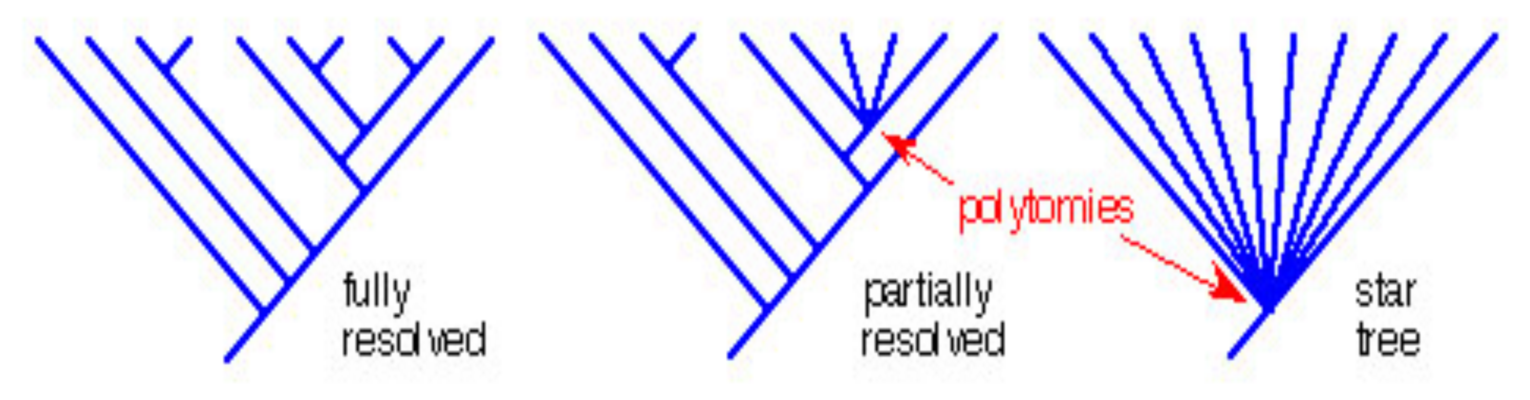
B. **Midpoint** rooting

C. Usage of external knowledge (ex. ancestral **gene duplication**)


A
outgroup
fugu
mouse
Drosophila
human
Drosophila
fugu
mouse
human
B
midpoint
fugu
3
1
mouse
2
1
human
5
Drosophila
Drosophila
fugu
mouse
human
1
1
4
3
2
1
C
gene duplication
human-α
human-β
fuga-α
fuga-β
fugu-α
human-α
fugu-β
human-β

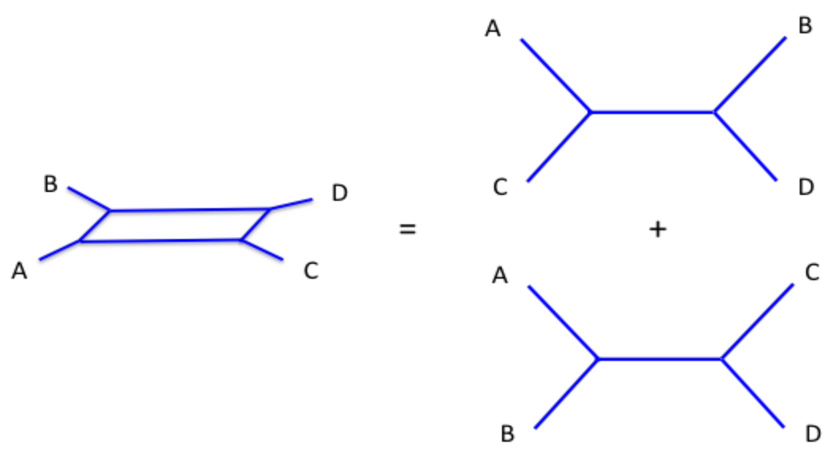
Is evolution always tree like ?

Some processes lead to **non-bifurcating trees** :


fully
resolved
partially
resolved
polytomies
star
tree

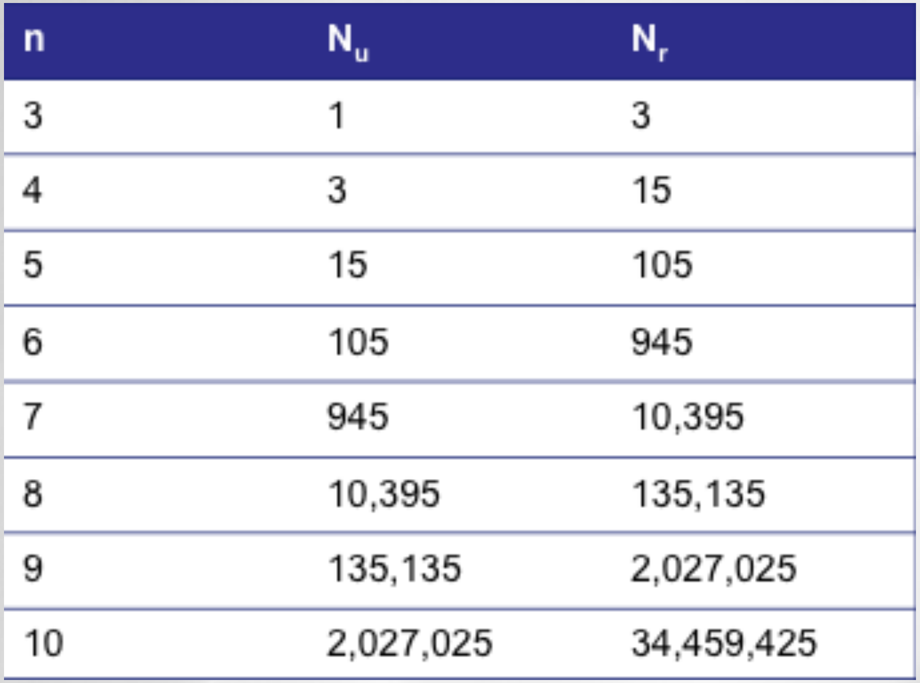
- Multifurcations on phylogenetic trees are konwn as **polytomies** an include trees with **internal polytonies** (partially unresolved tree) and **star-like trees**

- **Networks** are a way of representing two conflicting tree topologies


B
D
A
C
=
A
B
C
D
+
A
C
B
D

Number of tree topologies

- Number of possible unrooted (N_U) and rooted (N_T) trees for n=1 to 10 OTUs



n	N_u	N_r
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

$$N_u = 3x5x7x\ldots(2n-5) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

$$N_r = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

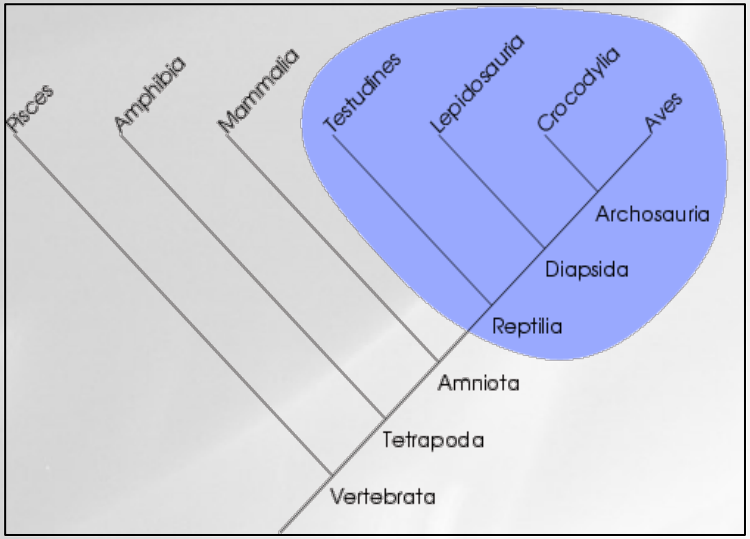
- Conclusion: an exhaustive search of all possible trees is usually impossible => **heuristic** strategies

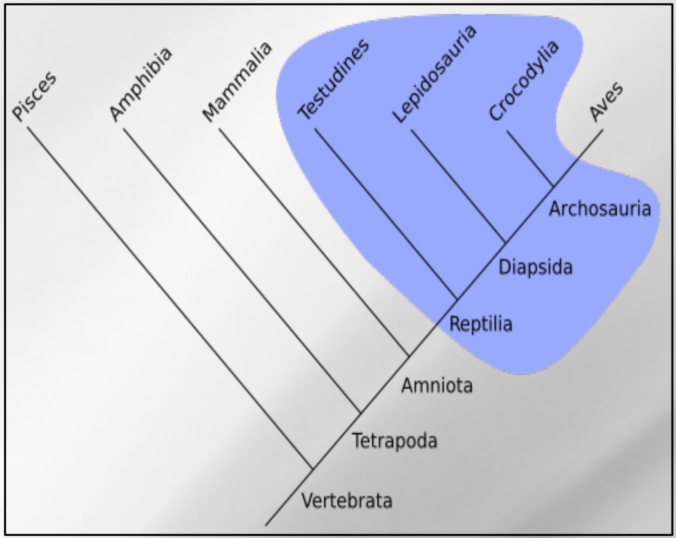
Terminology

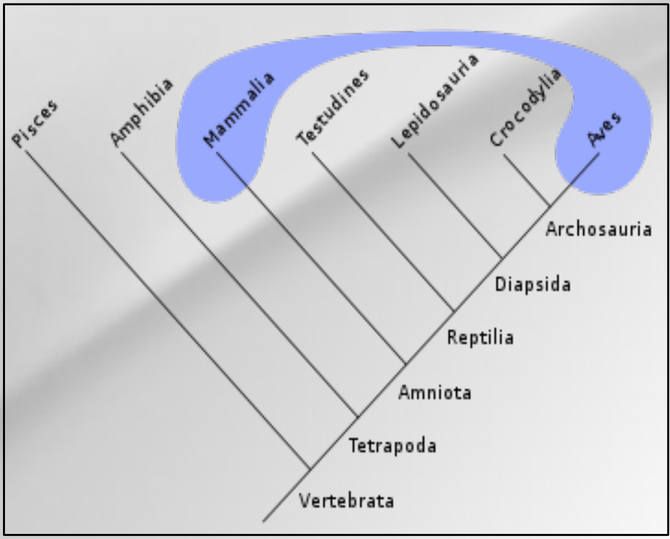
Monophyletic : a group of taxa is monophyletic if it includes all descendants from its inferred common ancestor

Paraphyletic : a group of taxa is paraphyletic if it does not include all descendants from its inferred common ancestor

Polyphyletic : a group of taxa is poliphyletic if it includes some descendants but not the inferred common ancestor

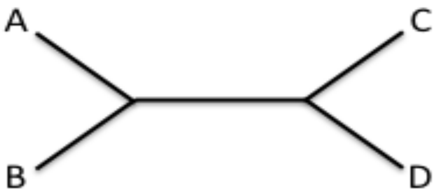

Pisces
Amphibia
Mammalia
Testudines
Lepidosauria
Crocodylia
Aves
Archosauria
Diapsida
Reptilia
Amniota
Tetrapoda
Vertebrata


Pisces
Amphibia
Mammalia
Testudines
Lepidosauria
Crocodylia
Aves
Archosauria
Diapsida
Reptilia
Amniota
Tetrapoda
Vertebrata


Pisces
Amphibia
Mammalia
Testudines
Lepidosauria
Crocodylia
Aves
Archosauria
Diapsida
Reptilia
Amniota
Tetrapoda
Vertebrata

Formats for phylogenetic trees

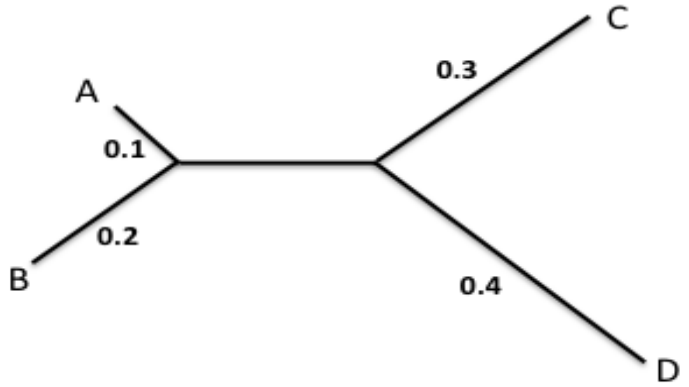
- **Two main formats: NEWICK and NEXUS**



NEWICK: ((A,B), (C,D))

```
#NEXUS
BEGIN TAXA;
 TAXLABELS A B C D;
END;

BEGIN TREES;
 TREE tree1 = ((A,B),(C,D));
END;
```



NEWICK: ((A:0.1,B:0.2):0.2, (C:0.3,D:0.4))

```
#NEXUS:
Begin trees;
Translate
1 A,
2 B,
3 C,
4 D,
;
Tree tree2= [&U] ((1:0.1,2:0.2):0.2, (3:0.3,3:0.4));
End;
```

Usual workflow in phylogenetic analysis

Dataset construction ⇨ Mutiple alignment ⇨ Build a phylogenetic tree ⇨ Interpret and evaluate the tree

Build a phylogenetic tree:

- *Evolutionary distance choice*
- *Method choice*

Genetic (evolutionary) distances

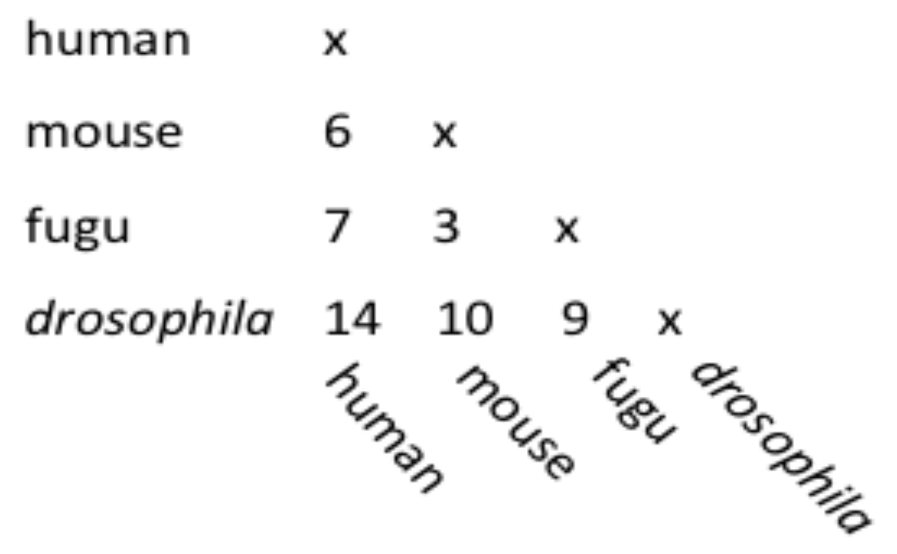
A **genetic (evolutionary) distance** is a measure of the divergence between two genetic sequences

Calculation of distance between two sequences is a central point on phylogenetic analysis

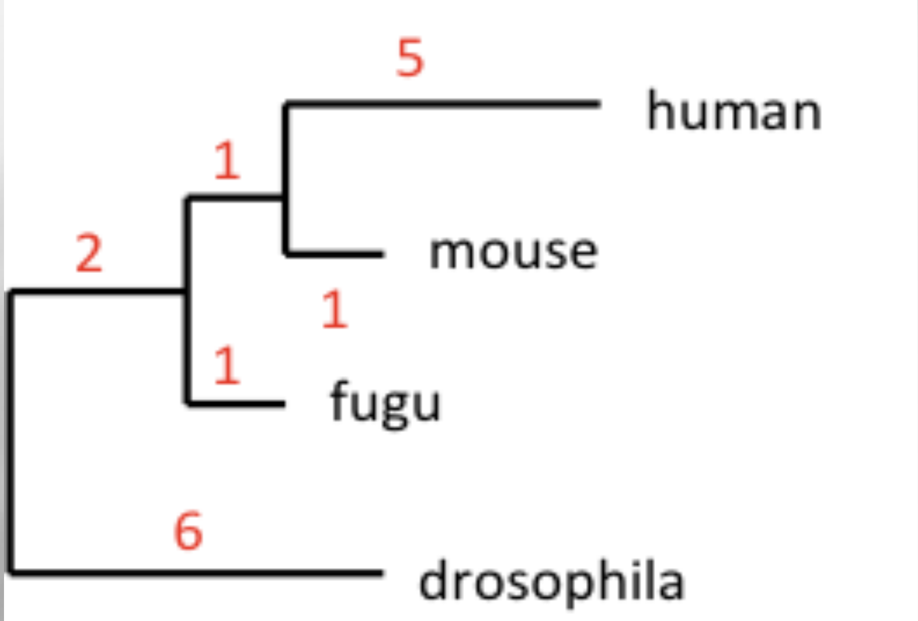
- Pairwise distance calculation is the first step of **distance matrix methods** in phylogeny (UPGMA, NJ)
- **Models of nucleotide/amino-acid sustitutions** used in distance-calculation form the basis of **likehood and Bayesian analysis methods**

Distances and trees

- For sequences related by an evolutionary tree, the **branch lengths** represent the distance between the nodes (sequences) in the tree
- If a **molecular clock hypothesis** is assumed then the genetic distance is linearly proportional to the time elapsed

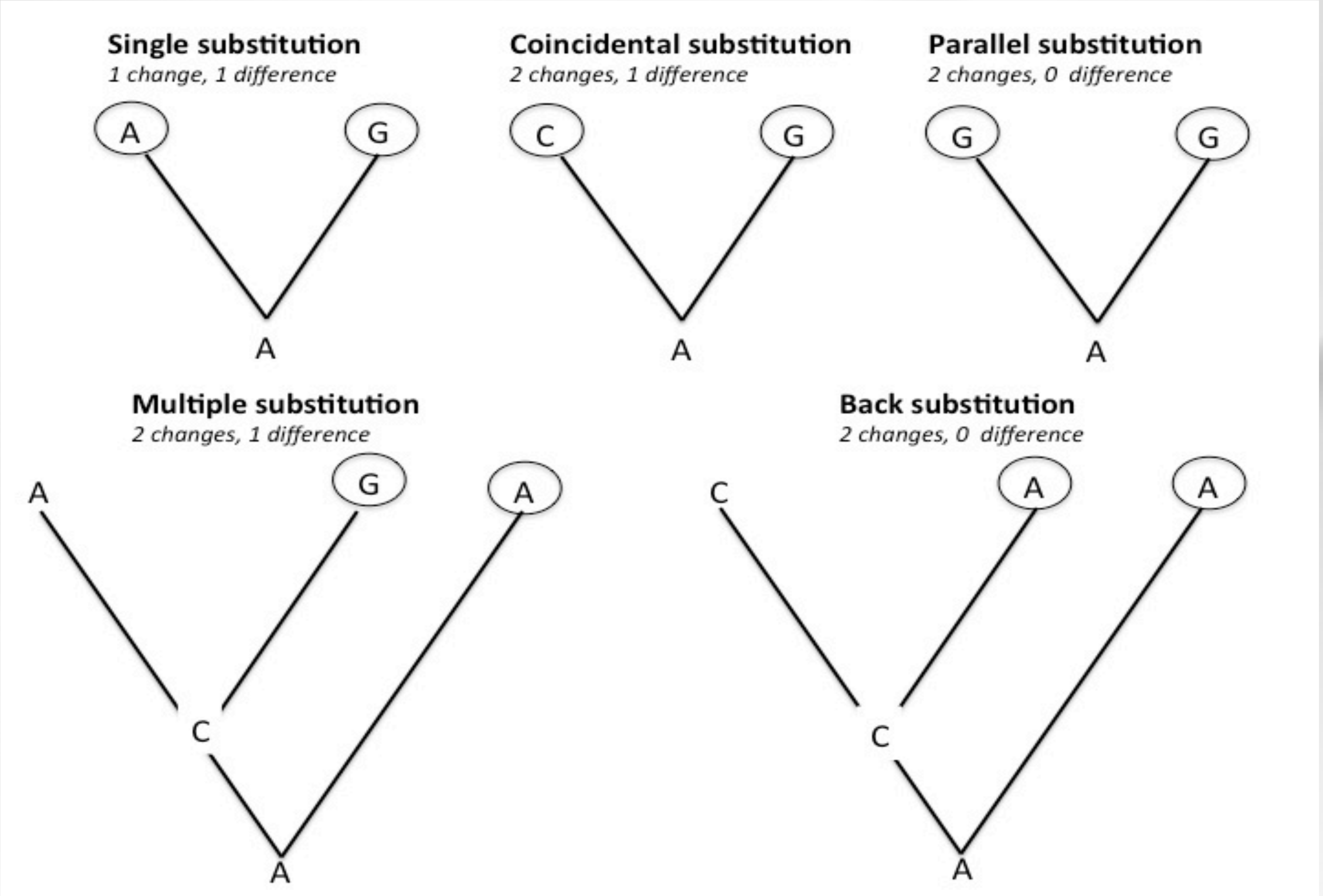


	human	mouse	fugu	*drosophila*
human	x			
mouse	6	x		
fugu	7	3	x	
drosophila	14	10	9	x



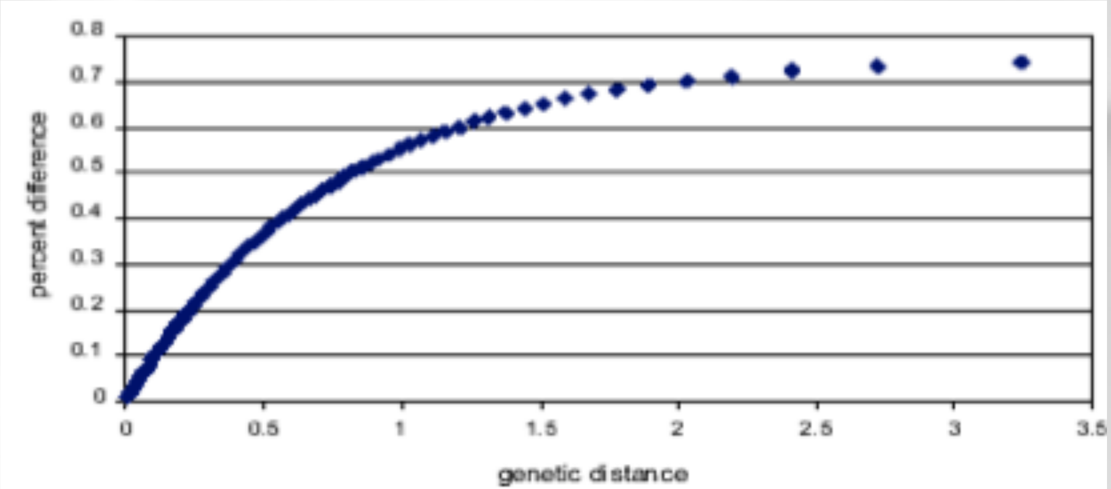
Observed and genetic distances

Observed nucleotide differences are not very informative !


Single substitution
1 change, 1 difference
A
G
A
Coincidental substitution
2 changes, 1 difference
C
G
A
Parallel substitution
2 changes, 0 difference
G
G
A
Multiple substitution
2 changes, 1 difference
A
G
A
C
A
Back substitution
2 changes, 0 difference
C
A
A
C
A

Observed and genetic distances

- **The observed distance** can be computed by counting the number of sites where two sequences differ : it is expressed as **the number of nucleotide differences per site (p-distance) ;**
- The observed distance is an under-estimation of the genetic distance due to multiple substitutions per site and saturation : **substitution models** are used.



Nucleotide substitution models

- Nucleotide substitution rate can be modeled as a stochastic process using **time continuous stationary Markov models ;**
- Underlying asumptions :
  - At any given site, the rate of change from base i to j is independant from the base that occupied that site prior i (**Markov property**) ;
  - Substitution rates do not change over time (**homogenity**) ;
  - The relative frequencies of A, C, G, and T are at equilibrium (**stationarity**)

Instantaneous rate matrix Q :

$$Q = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \left[\begin{array}{cccc} -\mu_A & \mu_{AT} & \mu_{AC} & \mu_{AG} \\ \mu_{TA} & -\mu_T & \mu_{TC} & \mu_{TG} \\ \mu_{CA} & \mu_{CT} & -\mu_C & \mu_{CG} \\ \mu_{GA} & \mu_{GT} & \mu_{GC} & -\mu_G \end{array}\right] \end{array}$$

Probability of from base i to base j :

$$P_{ij}(t) = e^{Q(t)}$$

The Jukes & Cantor model (JC, 1969)

- The simplest possible nucleotide substitution model :
  - All base frequencies are equal (0.25)
  - Only **one parameter** = the **susbtitution rate μ**
- Given the **proportion p of sites** that differ between the two sequences the Jukes-Cantor estimate of the **evolutionary distance d** is given by :

$$Q = \begin{bmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{bmatrix}$$

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

where p is the proportion of sites that show differences.

The JC model - exercise

```
Seq1 TCAAGTCAGGTTCGA
       |   |  ||
Seq2 TCCAGTTAGACTCGA
       | | |  |  |
Seq3 TTCAATCAGGCCCGA
```

	Seq1	Seq2	Seq3
Seq2			
Seq3			

Observed distances

Observed distance

$$d_{obs}(seq1 - seq2) = ?$$

	Seq1	Seq2	Seq3
Seq2			
Seq3			

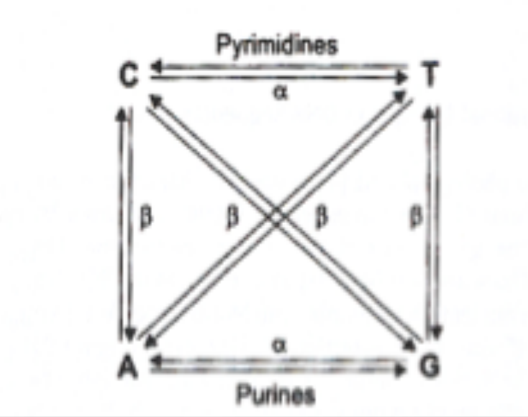
J&C distance

$$d_{JC}(seq1 - seq2) = ?$$

Evolutionary distances

The Kimura model (1980)

- The model is defined by **2 parameters**
  - all base frequencies are equal (0.25)
  - It distinguishes **the rate of transition** substitutions **α** and **the rate of transversion** substitutions **β**



- The **Kimura two-parameter distance d** is given by:

$$Q = \begin{array}{c|cccc} & A & T & C & G \\ \hline A & -\mu_A & \beta & \beta & \alpha \\ T & \beta & -\mu_T & \alpha & \beta \\ C & \beta & \alpha & -\mu_C & \beta \\ G & \alpha & \beta & \beta & -\mu_G \end{array}$$

$$d = -\frac{1}{2}\ln\left(1-2p-q\right)-\frac{1}{4}\ln(1-2q)$$

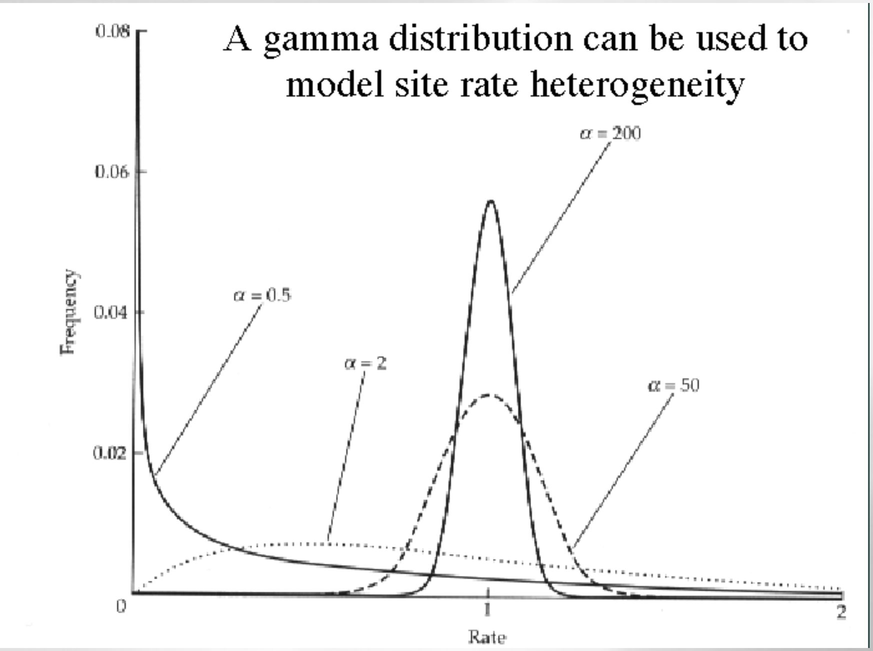
where p is the proportion of sites that show transitional differences and q is the proportion of sites that show transversional differences.

Other models

- The **Felsenstein's 1981** model is an extension of the JC69 model in which base frequencies are allowed to vary from 0.25
- The **HKY85 model** can be thought of as combining the extensions made in the Kimura80 and Felsenstein81 models: it distinguishes between the rate of transitions and transversions and it allows unequal base frequencies.
- The **GTR (Generalised time-reversible, Tavaré 1986)** model is the most general neutral, independent, finite-sites, time-reversible model possible :
  - All bases can have unequal frequencies
  - All type of mutations are distinghuished

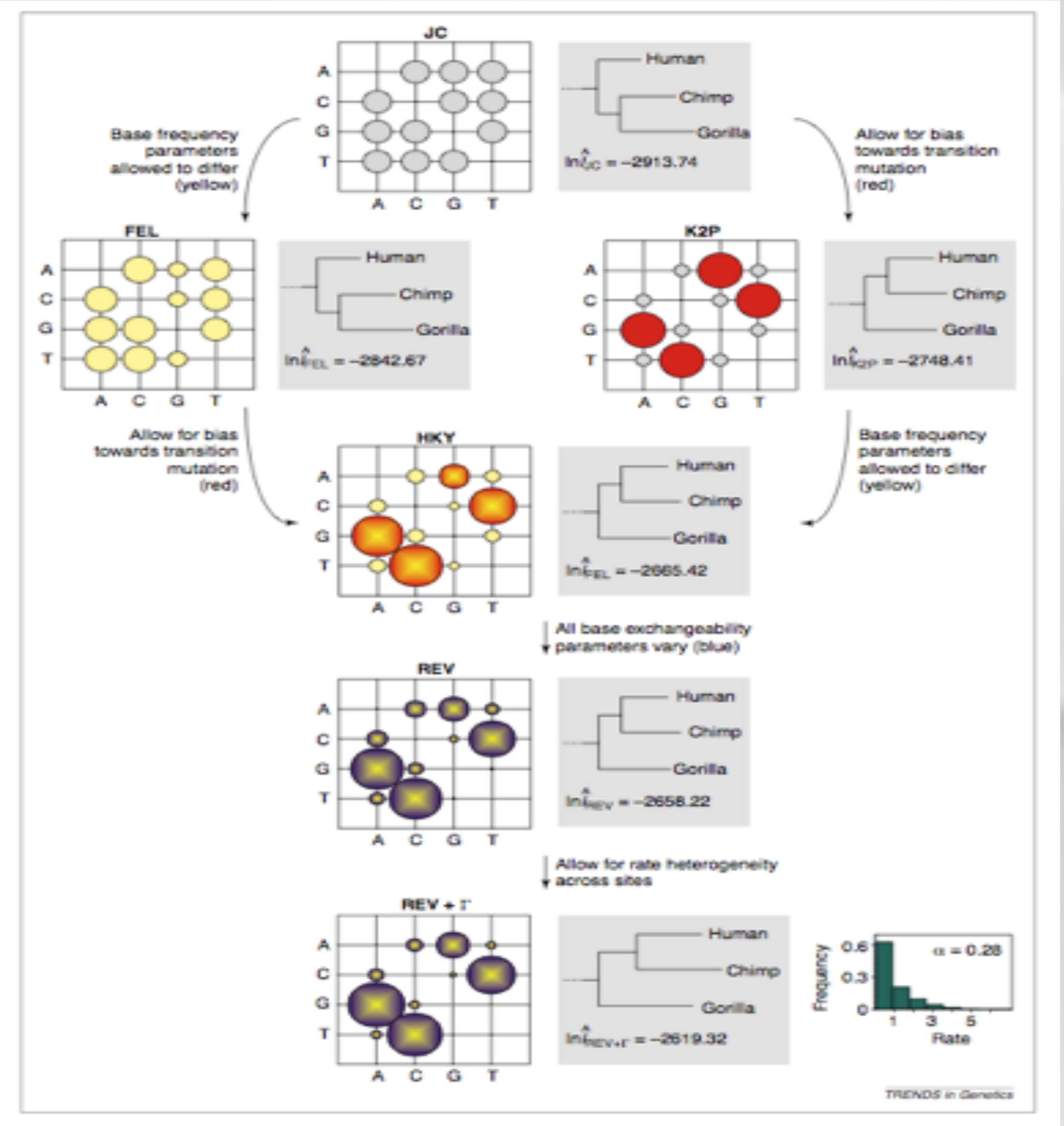
Rate heterogeneity among sites

- The **rate of substitution can vary substantially** for different position of an an alignment
- To account for the site-dependent rate variation, the common approach is to use a **Gamma distribution which model distribution rates between sites**



Usually, rather than using the continuous Gamma distribution, **discrete categories of equally probable substitution rates** are used to obtained an approximation of the function (4 to 8 site categories)

Nucleotide models : summary

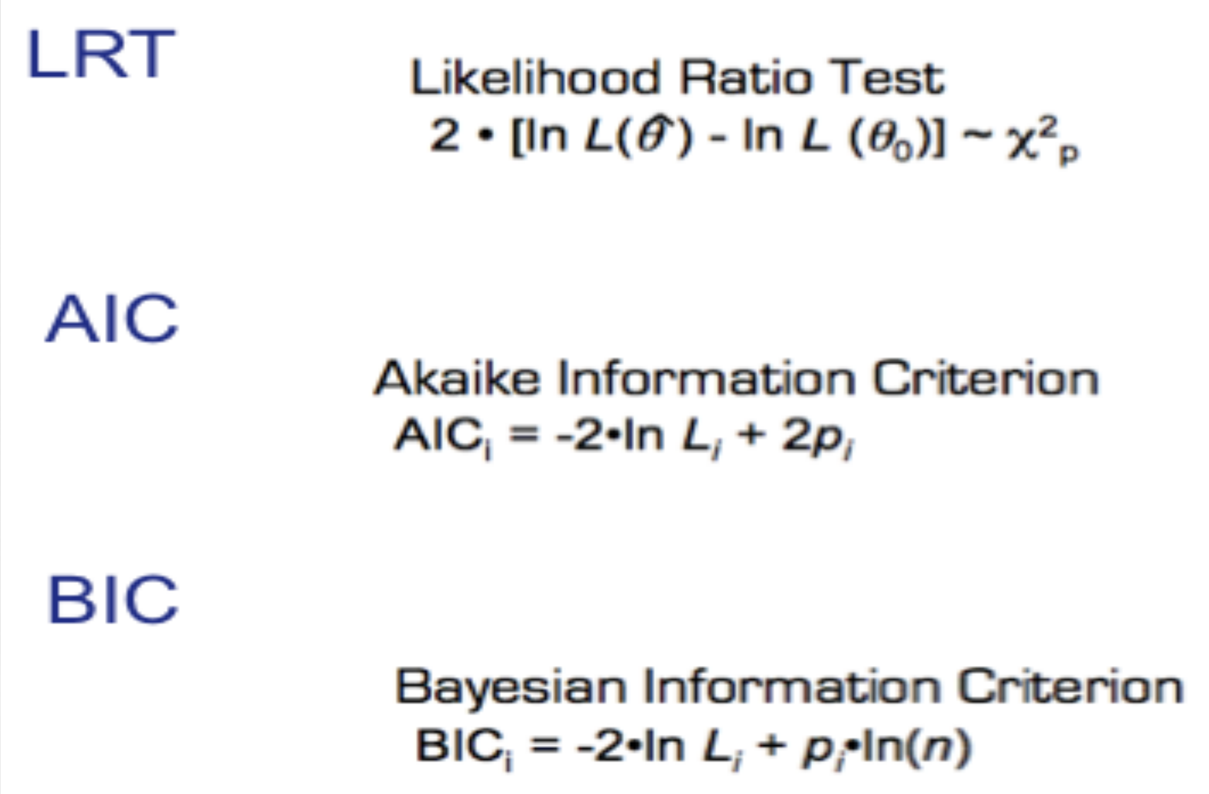

JC
A
C
G
T
A C G T
Human
Chimp
Gorilla
$\ln \hat{L}_{JC} = -2913.74$
Base frequency parameters allowed to differ (yellow)
Allow for bias towards transition mutation (red)
FEL
Human
Chimp
Gorilla
$\ln \hat{L}_{FEL} = -2842.67$
K2P
Human
Chimp
Gorilla
$\ln \hat{L}_{K2P} = -2748.41$
Allow for bias towards transition mutation (red)
Base frequency parameters allowed to differ (yellow)
HKY
Human
Chimp
Gorilla
$\ln \hat{L}_{FEL} = -2665.42$
All base exchangeability parameters vary (blue)
REV
Human
Chimp
Gorilla
$\ln \hat{L}_{REV} = -2658.22$
Allow for rate heterogeneity across sites
REV + Γ
Human
Chimp
Gorilla
$\ln \hat{L}_{REV+\Gamma} = -2619.32$
Frequency
0.6
0.3
0
α = 0.28
1
3
5
Rate
TRENDS in Genetics

Choosing among models

- **It is crucial step**
- Different evolutionary models can lead to different results : **inaccurate branch lengths**, even sometimes **wrong tree topology**
- The most complex model with the largest number of parameters is not necessarily the most appropriate, it **depends of the question and the data**
- The best-fit model of evolution for a particular dataset can be selected using **sound statistical techniques, for example :**
  - Hierarchical **Likehood Ratio Tests** (hLTRs)
  - **Information criteria** (ex : Akaike Information criterion=AIC)

Choosing among models

- In practice : **adjust the model to the analyzed dataset**
- **Use statistical methods to select the best fitted model* :**



LRT

Likelihood Ratio Test

$2 \cdot [\ln L(\hat{\theta}) - \ln L(\theta_0)] \sim \chi^2_p$

AIC

Akaike Information Criterion

$AIC_i = -2 \cdot \ln L_i + 2p_i$

BIC

Bayesian Information Criterion

$BIC_i = -2 \cdot \ln L_i + p_i \cdot \ln(n)$

LRT criterion can be used to compare models which are subsets of each other

AIC and BIC criteria compare all of the models simultaneously according to some measure of fitness

*Keane & al., *BMC Evolutionary Biology* 2006

Selection of the best fitted model

Example: Hierarchical LRT of models of molecular evolution

Ho	**Models compared**
Equal base frequencies	Ho: JC69 1 parameter H1 : F81 2 parameters
Equal ti/tv rates	Ho : F81 2 parameters H1 : HKY 5 parameters
Equal ti and equal tv rates	Ho : HKY 5 parameters H1 : GTR 9 parameters
Equal rates among sites	Ho : GTR 9 parameters H1 : GTR+ τ 9 parameters +n
Proportion of invariable sites	Ho : GTR+ τ 9 parameters +n H1 : GTR+ τ + I 9 parameters +n +1

where I means there is a significant proportion of invariable sites, and τ means a gamma distribution is being used to account for rate variation among sites

Protein models

- **Similar concept: multiple substitutions of amino acids** lead to underestimation of evolutionary distances between two homologous proteins.
- **Substitution frequency of amino acids depends of the AA :** it is higher between closed amino-acids in term of physical properties (polarity, hydrophibicity,...)
- **Too much (190) parameters to estimate parameters of probabilistic model** => **empirical models** are used
- Transition rate between amino acids are estimated once from **big reference alignments obtained by concatenation of several homologs proteins**

Main protein evolutionary models

Model	**Dataset**	**Ref**
Poisson	Poisson process	Zuckerkandl, 1965
PAM	1300 protein sequences from 71 homolog families	Dayhoff 1978
Blosum	Extension of PAM dataset	Henikoff 1992
JTT	16 300 sequences	Jones 1992
mtREV	Mitochondrial DNA	Adachi 1996
WAG & LG	**Likehood methods**	**Whelan 2001**

Model choice is based on the same tests as for nucleotide evolutionary models (LRT, AIC, BIC)

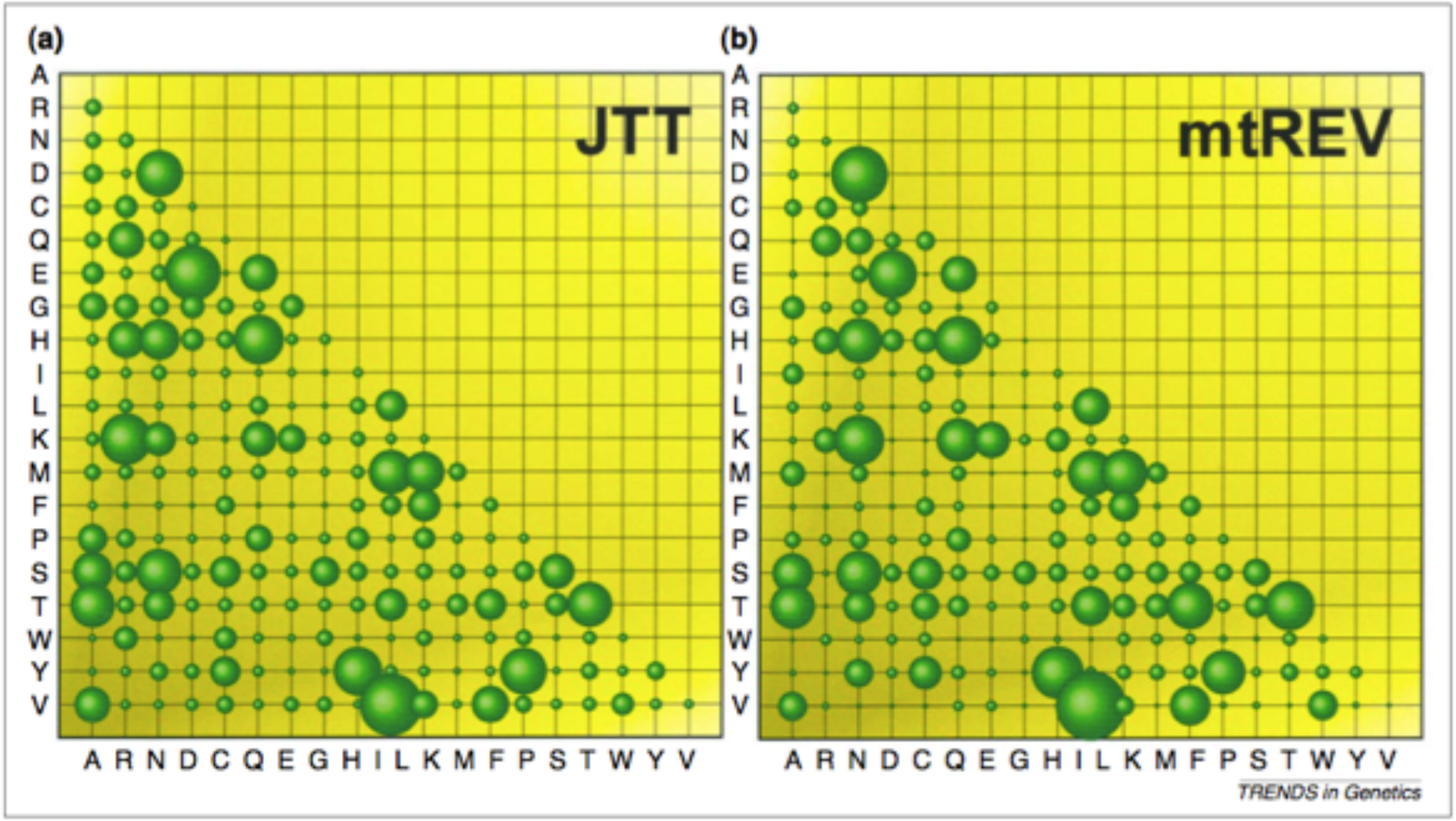
Main protein evolutionary models

Model	**Dataset**	**Ref**
Poisson	Poisson process	Zuckerkandl, 1965
PAM	1300 protein sequences from 71 homolog families	Dayhoff 1978
Blosum	Extension of PAM dataset	Henikoff 1992
JTT	16 300 sequences	Jones 1992
mtREV	Mitochondrial DNA	Adachi 1996
WAG & LG	**Likehood methods**	**Whelan 2001**

WAG and LG models are the more used models

Protein models

Example: JTT (1992, 16 300 sequences) vs mtREV (for mitochondrial proteins)



Choosing among models: training

- **Dataset 1**: **amino-acid** alignment of the phosphoglycerate kinase (pgk) protein in 21 bacterial species of the firmicutes group (adapted from [1] ; formats : clustalw, fasta, phylip).

Firmicutes are a phylum of bacteria, most of which have Gram-positive cell wall structure. It includes Bacilli, Clostridia, mollicutes…

[1] Wolf & al, 2004, *International Journal of Systematic and Evolutionary microbiology*, ***Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data***

Choosing among models: training

- **Dataset 2: nucleic** alignment of the beta glucosidase gene (bglA) in 10 bacterial species of the *Staphylococcus* and *Listeria* genus (adapted from [2], formats : clustalw, fasta, phylip).

bglA is the betaglucosidase gene used in MLST studies

Here focus on two bacterial genus : 5 listeria strains, 5 staphylococcus strains

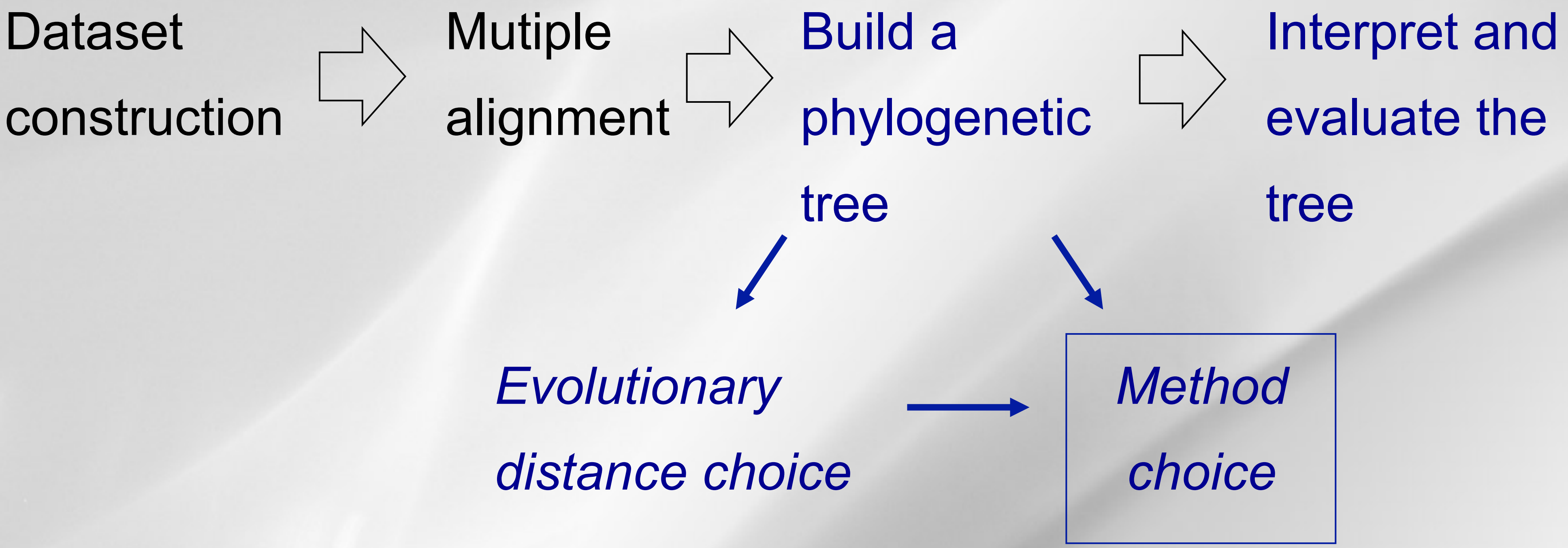
[2] Ragon *et al*. 2008. PLOS Pathogens. **A new perspective on Listeria monocytogenes evolution**

Exercise 1: solution

Exercise 1: Choose the best-fit evolutionary model for the two datasets *(use modelgenerator.jar)*

- **Dataset 1**: **amino-acid** alignment of the phosphoglycerate kinase (pgk) protein in 21 bacterial species of the firmicutes
- **Dataset 2: nucleic** alignment of the beta glucosidase gene (bglA) in 10 bacterial species of the *Staphylococcus* and *Listeria* genus
- *Modelgenerator : take as input fasta or phylip files, choose 4 categories of sites*

Usual workflow in phylogenetic analysis



Dataset construction

Mutiple alignment

Build a phylogenetic tree

Interpret and evaluate the tree

Evolutionary distance choice

Method choice

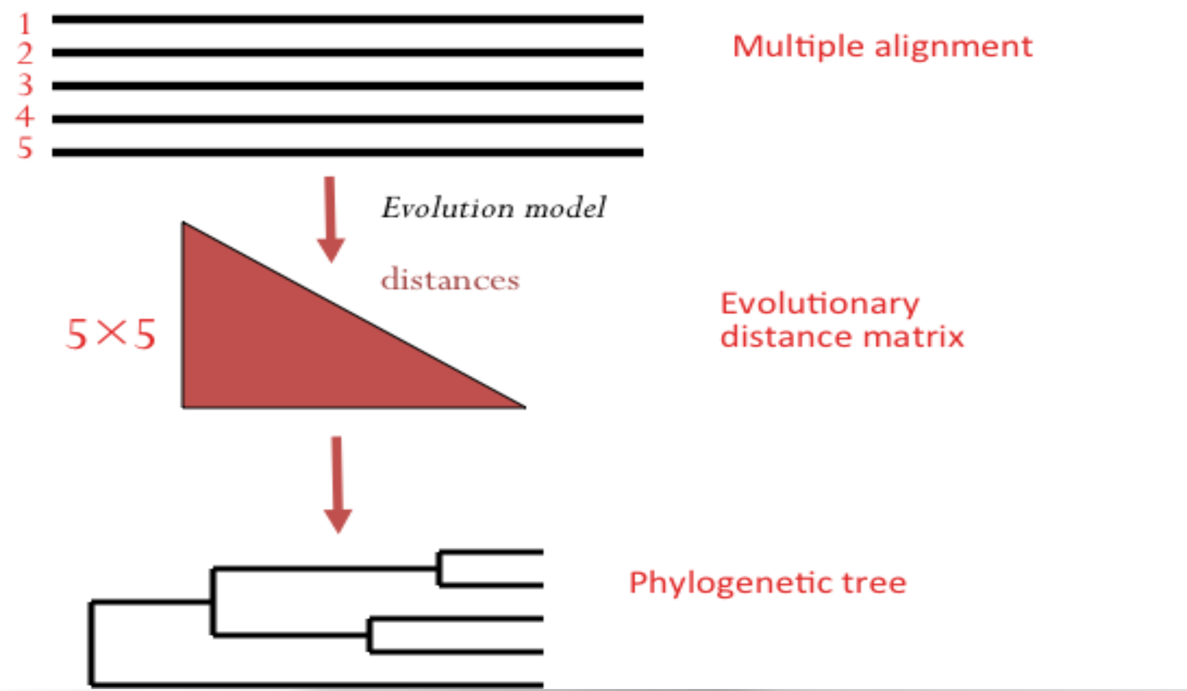
Method choice

- Main methods for inferring phylogenetic trees:

Input data	**Method**	**Principle of the algorithms**
Distance matrix	Unweighted Pair Group Method (UPGMA)	clustering
	Neighbor-Joining (NJ)	clustering
Character state	Maximum Parsimony (MP)	Search for the tree(s) of minimum character changes
	Maximum Likehood (ML)	Search for the tree(s) that maximizes the probability of observing the character states giving a tree topology and a model of evolution
	Bayesian Inference	Target a probability distribution of trees (set of possible trees for the data)

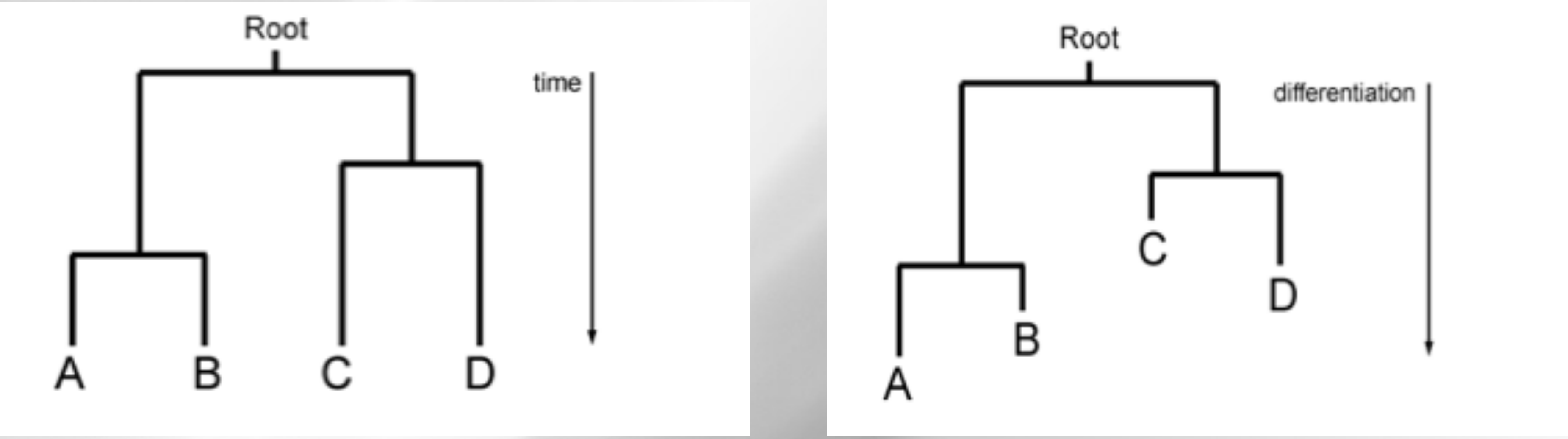
Distance methods for inferring a phylogenetic tree

- Introduced in phylogeny in 1960
- Try to fit a tree **to a matrix of pairwise genetic distances**
- Need to choose an **evolutionary model**


1
2
3
4
5
Multiple alignment
Evolution model
distances
5×5
Evolutionary distance matrix
Phylogenetic tree

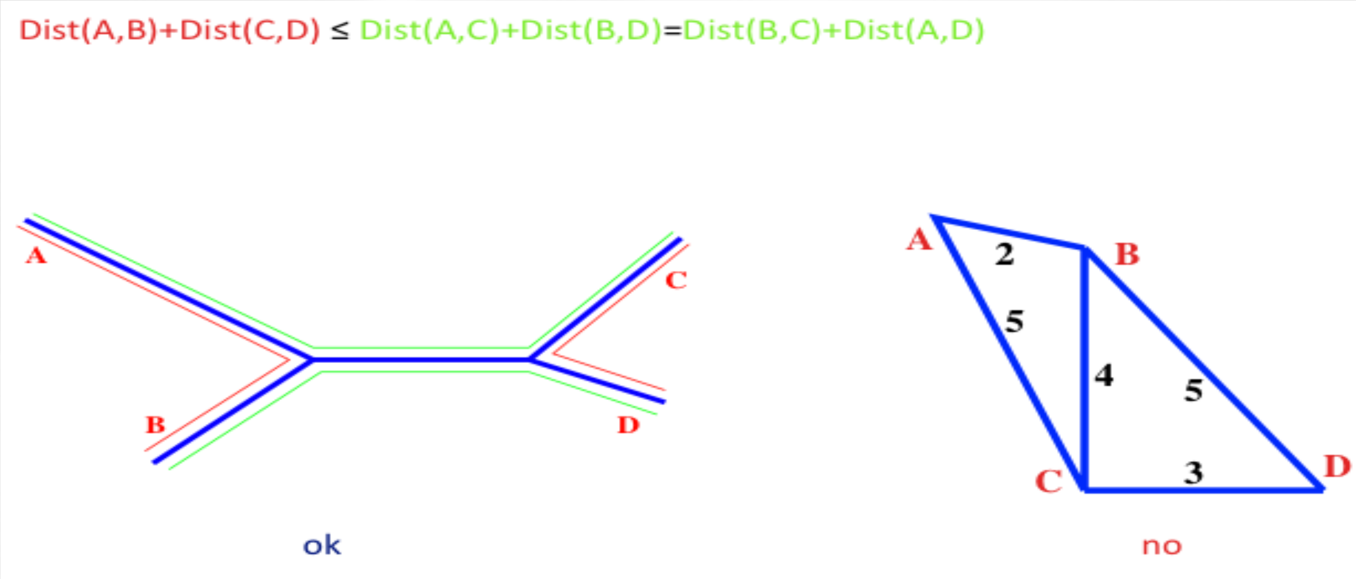
Distance methods

- Two main methods
  - **UPGMA**: a clustering method that produced ultrametric trees
  - **Neighbor-Joining**: use a greedy algorithm to compute the Minimal Evolution tree *i.e.* the optimal topology is the one which minimizes the tree length


Root
time
A
B
C
D
Root
differentiation
C
D
B
A

Neighbor-Joining

- First algorithm proposed by **Saitou & Nei** (1987)
- Very fast : polynomial-time algorithm
- Produces unrooted trees
- Produces the wright topology if matrix **distances are patristic**



$\mathrm{Dist}(A,B)+\mathrm{Dist}(C,D) \leq \mathrm{Dist}(A,C)+\mathrm{Dist}(B,D)=\mathrm{Dist}(B,C)+\mathrm{Dist}(A,D)$

Plateforme Bioinformatique Midi-Pyrénées

Neighbor-Joining (NJ)

Principle of the algorithm:

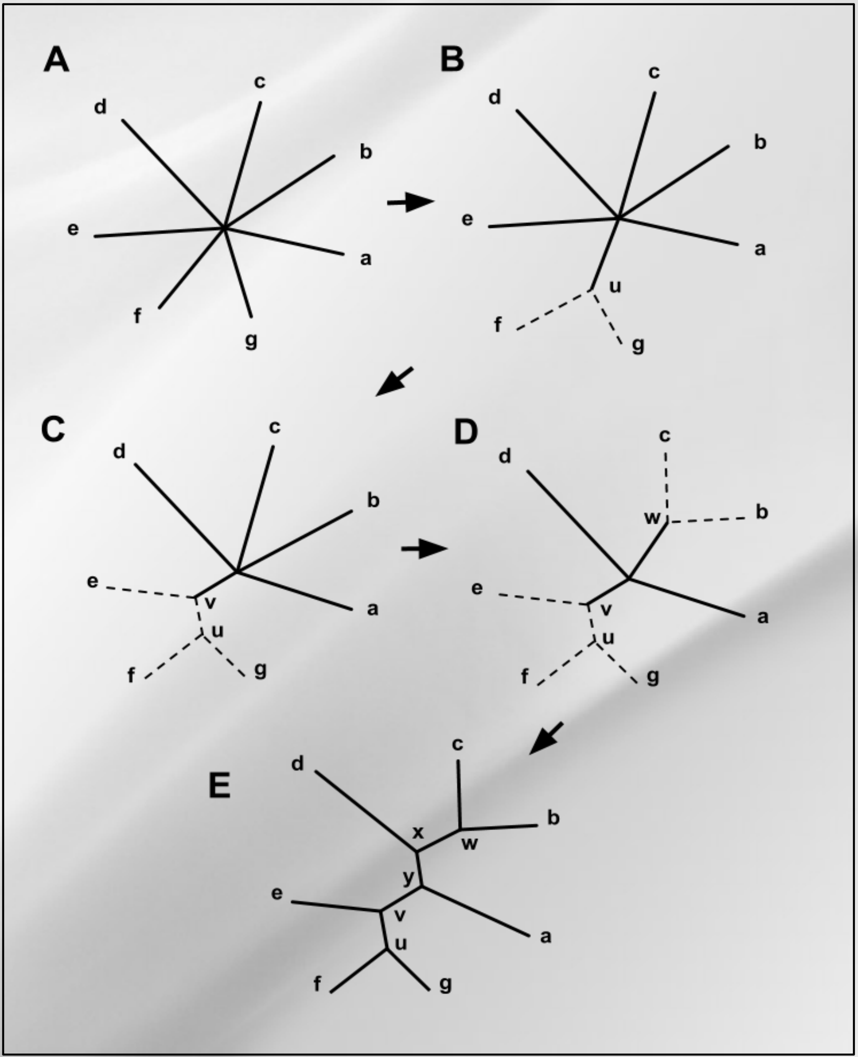
- Start with **a star tree** (A)

- Compute the **matrix Qij** and find the pair of taxa with lowest value (here f and g)

- Join f and g and **create a new internal node, u**, as shown in (B)

- Compute the distances from node u to the nodes a-e

- **Repeat the process** :
  - u and e are joined to the newly created v, as shown in (C).
  - Two more iterations lead first to (D), and then to (E).



Neighbor-Joining in practice

- **NJ: Fast** but problems may occur for **very divergent sequences** or **heterogeneous** datasets
- **BioNJ* algorithm:**
  - A variant of NJ which improves its accuracy by making use of a simple first-order model of the variances and covariances of evolutionary distance estimates.
  - When the substitution rates are low (maximum pairwise divergence ~0.1 substitutions per site) or when they are constant among lineages, BIONJ is only slightly better than NJ.
  - When the substitution rates are higher and vary among lineages, BIONJ clearly has better topological accuracy*.

**Gascuel Molecular Biology and Evolution 1997*

Neighbor-Joining in practice

- Choose **an evolutionary model** and compute **a distance matrix (see next slide)**
- **NJ/BioNJ softwares:**
  - **Neighbor** (PHYLIP, NJ) http://evolution.genetics.washington.edu/phylip.html
  - **BioNJ** http://www.atgc-montpellier.fr/bionj/ or http://phylogeny.lirmm.fr/phylo_cgi/one_task.cgi?task_type=bionj
  - **QuickTree (NJ)** http://www.sanger.ac.uk/resources/software/quicktree/
  - **Seaview** (NJ and BioNJ) http://pbil.univ-lyon1/fr/software/seaview

Evolutionary models in NJ

NJ softwares do not implement all models !

- **At small distances** (~10% of variable sites) the different evolutionary models produce very similar distance estimates => no problem
- **At intermediate distances** (20 to 30% of variable sites), different model asumptions become more important => It is recommanded to use realistic models for distance estimation, especially if the sequences are longs
- **At large distances** (40% of variable sites), the different model produces very different distance estimates. Sometimes the distance estimates become infinite. => The solution is to use realistic models for distance estimation AND to add sequences to break down the long distances

Neighbor-Joining: training

Exercise 2: *Neighbor-Joining* trees for the two alignement datasets.

- ***Question 1.*** *Use* ***seaview or Phylip*** *to build a Neighbor-Joining tree for the two datasets.*
- ***Question 2*** *Use* ***figtree*** *to root the trees using two different methods : midpoint rooting and and arbitrary outgroup (for instance Staph_epider for the nucleic dataset and Myco_geni for the proteic dataset)*

Parsimony

- Main concept (adapted from Fitch, 1971):

  Seek the tree(s) that **minimizes the net amount of evolutionary change (in term of character change) required to explain the data**
- Very used on **morphological data** (presence/absence of characters) but also relevant for **biological sequences** (a character = a site with 4 states=A,T,C,G or molecular polymorphism data like SINE)
- Produces **unrooted tress**
- Does not require any evolutionary model
- Take into account explicitly ancestral states

Parsimony

The problem of finding the parsimony tree can be separated into **three steps**:

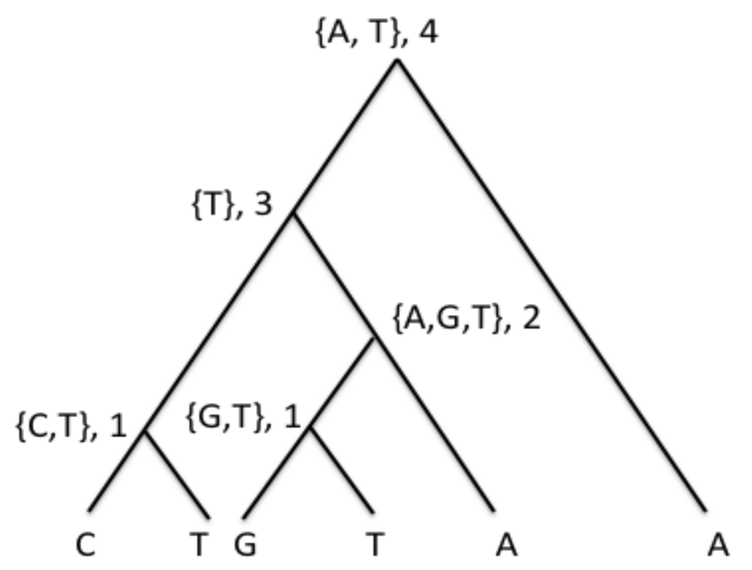
- **Step 1:** Compute the **minimal amount of character change** required in a given tree (compute changes for each character and sum up all characters)
- **Step 2:** Search for **all possible tree topologies**
- **Step 3:** Choose **the tree(s) that minimize this number of character changes**.

Parsimony

Step 1: Compute the minimal amount of character change required in a given tree (compute changes for each character and sum up all characters)

- Compute the minimum number of changes for a site in a tree (for instance with the ***Fitch algorithm***)

Seq1. ...C...
Seq2. ...T...
Seq3. ...G...
Seq4. ...T...
Seq5. ...A...
Seq6. ...A...

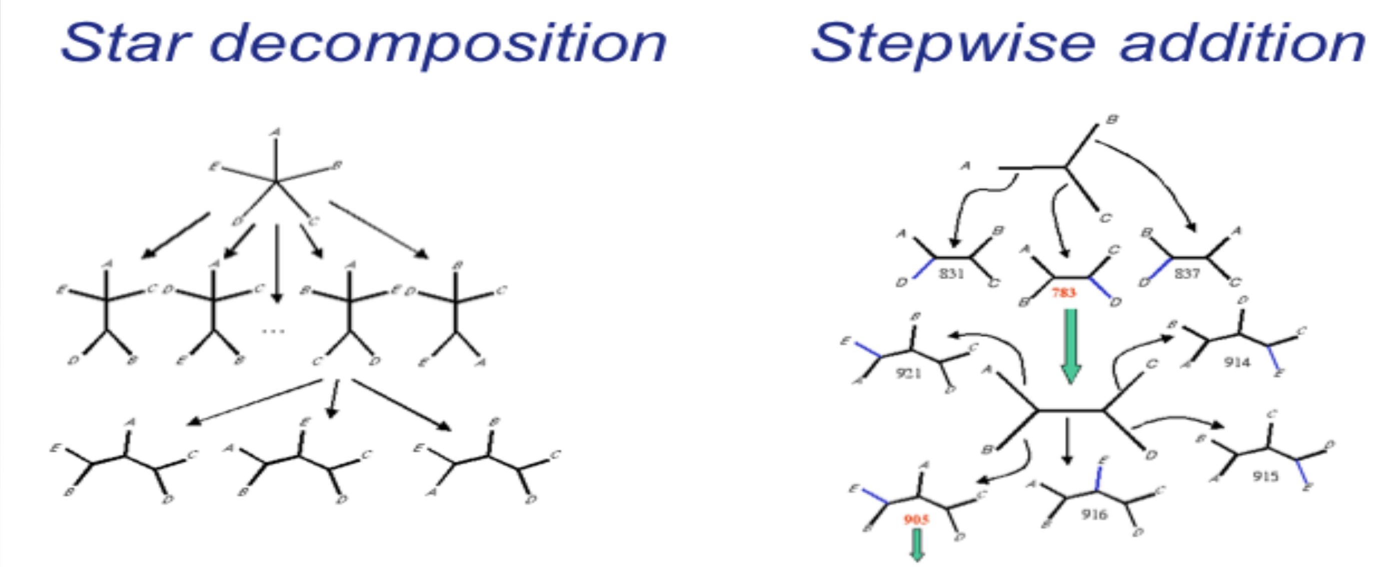


Parsimony score :
1+1+2+3+4 =11

- Sum over the number of sites to obtain the **parsimony score** of a tree

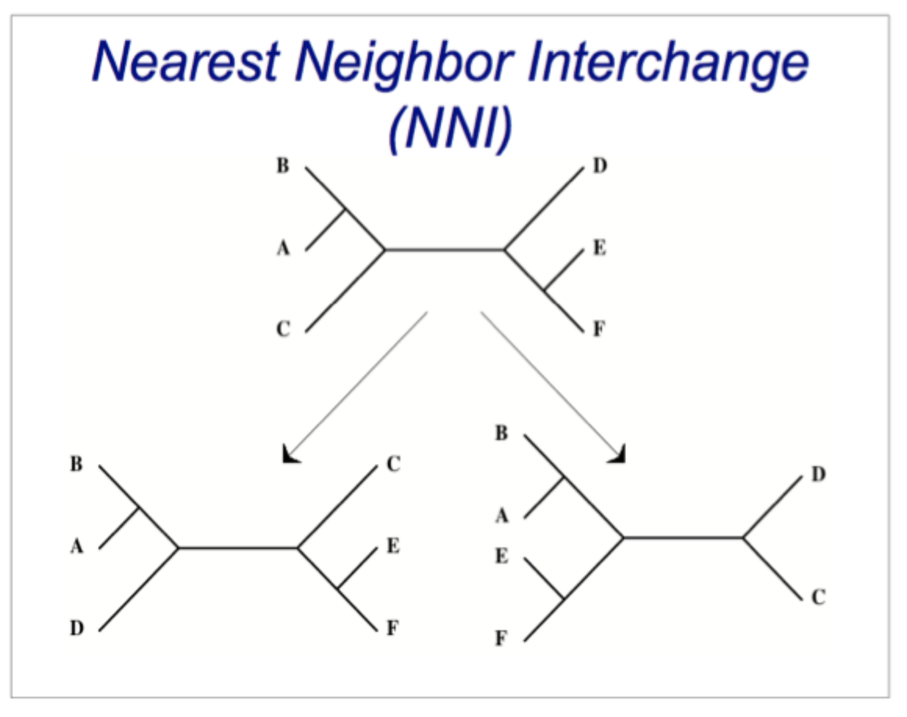
Parsimony

- **Step 2 : generate all possible tree topologies**
  - **Exact methods** (max. 20 taxa) : example=*Branch and Bound* algorithm
  - **Heuristic methods** : choose an **intitial tree topology** (star decomposition, stepwise addition, random choice) and perform **tree-rearrangement perturbations** like *Nearest Neighbor Interchange (NNI)* or *Subtree Pruning and Regrafting (SPR)*


Star decomposition
Stepwise addition
831
783
837
921
914
915
905
916

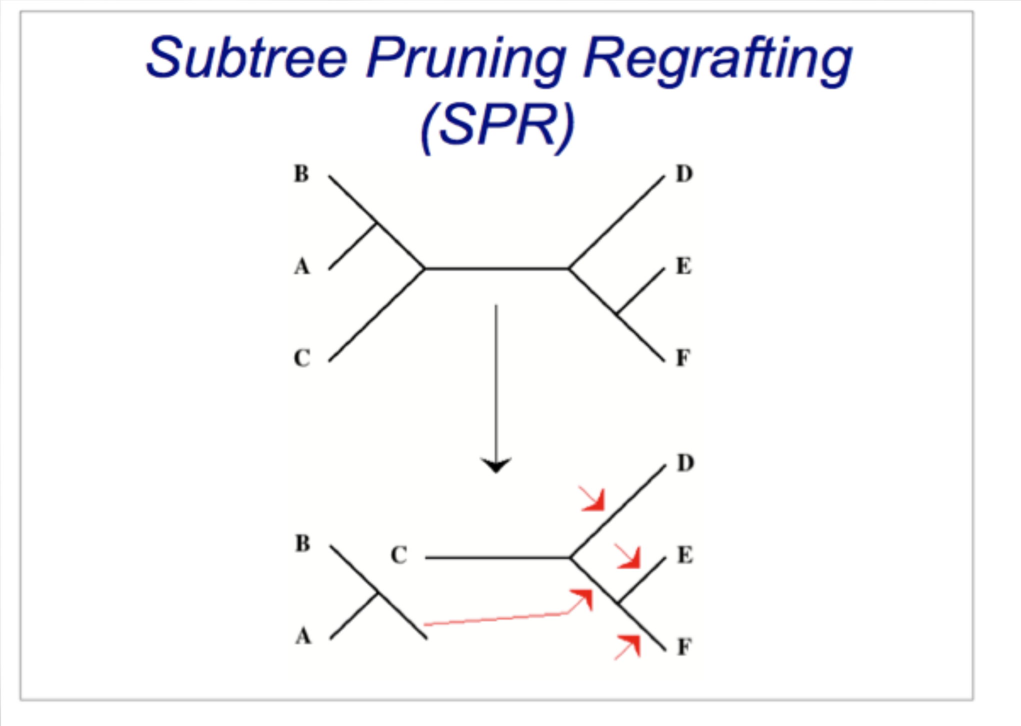
Tree rearrangments

- Exploration of tree topologies using different kind of local rearrangments:



Nearest Neighbor Interchange (NNI)

Small changes => local space exploration



Subtree Pruning Regrafting (SPR)

Medium changes => best space exploration

- Iterate and keep always the best (more parsimonious) tree
- Stop after n iterations if the swapping process do not produce a better tree

Parsimony in practice

- + : can be applied to any kind of characters, good performances if substitution events are rare
- - : no statistical justification and some sites are excluded i.e. **non informative sites** = invariant sites (AAAA) and two-states sites with one character in one occurrence (AAAT) (all the tree are equal for theses sites)
- **Sotwares for parsimony**
  - PHYLIP (dnapars, protpars)
  - Seaview
  - MacClade http://macclade.org/macclade.html
  - (PAUP)

Parsimony: training

Exercise 3: build a *parsimony* tree for the two alignement datasets *(use Seaview or Phylip)*

Maximum Likehood methods

- **The most frequently used methods**
- **Sound mathematical and statistical foundations**
- The **evolution model is central**, the method is only possible for aligned sequences
- In statistics, **maximum-likelihood estimation (MLE) is** a general method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters.

Maximum Likehood methods

Adressed question: what is the probabilty to observe the data by considering an evolutionary model with its parameters and a tree topology ?

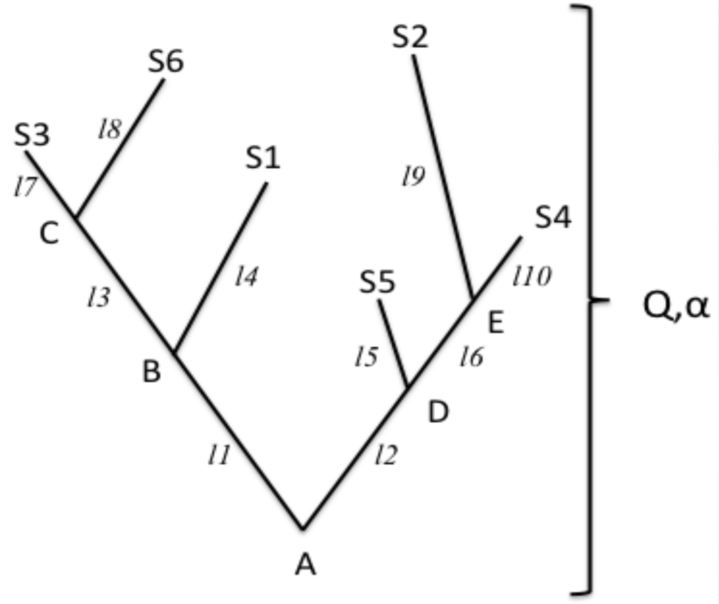
$$\Pr(D/T)$$

- **Input:** A set of observed sequences and an underlying evolutionary model.
- **Desired Output:** The weighted tree that maximizes the likelihood of the data

Maximum Likehood methods

Parameters of the probabilistic model :

- A phylogenetic tree T, with an arbitrary root and valuated branch lengths
- A normalized Q-matrix, common to all tree branches
- An α parameter which determines the variation of the evolutionary rates between sites using the Gamma distribution



- l_i are branch lengths (#subst/site)
- A, B, C, D, E are the unknown ancestral states

- Likehood computation of observed data :

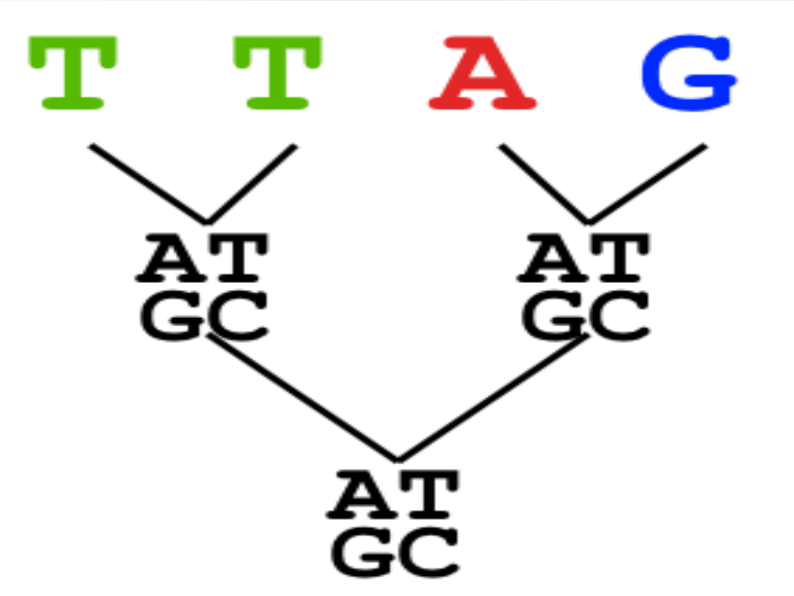
$$Log(L) = \sum_{sites} \log(L(site))$$

$$L(site) = \sum_{A}\sum_{B}\sum_{C}\sum_{D}\sum_{E} \Pr ob(S1,S2,S3,S4,S5,S6,A,B,C,D,E \mid T)$$

Maximum likehood: Example

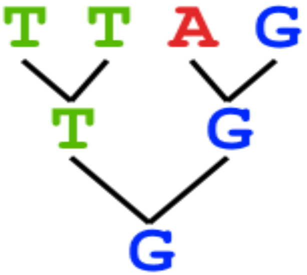
```
Sequence W: A C G C G T T G G G
Sequence X: A C G C G T T G G G
Sequence Y: A C G C A A T G A A
Sequence Z: A C A C A G G G A A
                      ▲
```

All possible evolutionary paths of a site



Likehood of a site

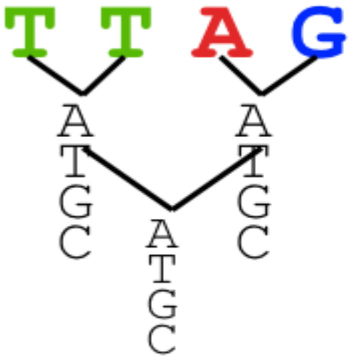
Likehood of a path



L(path) = L(root) x Π L(branches)

=P(G→T)P(G→G) P(G→A)P(G→G) P(T→T)P(T→T)

Sum over all paths


T
T
A
G
A
T
G
C
A
T
G
C
A
T
G
C

L(Column Cluster 1) = Σ L(all possible Evolutionary Paths)

= L(path1) + L(path2) + L(path3) + … + L(path64)

Felsenstein algorithm

- 5 internal nodes => $5^4 = 1024$ possible combinations

- **Pruning Felsenstein algorithm** :
  progressive computation of **the likehood of a site to have nucleotide i** (with tree T and model M fixed) **from leaves to root** by using a **recursive** strategy

- **Calculate tree Likelihood by multiplying the likehood for each position**

Maximum Likehood features

- **Branch length l are estimated using the Q matrix (of an evolutionary model).**

l=expected number of subtitutions per site = µt (mutation rate x time)

$$P_{(l)} = e^{Q_l}$$

- **Reversibility of the process** (symetry of Q matrix) : it is possible to show that if the base substitution model is reversible
- **Root position** : Likelihood remains the same regardless of where the root is. So search for the best tree only needs to be carried out on unrooted trees
- **Can take into account variation of the evolutionary rates** between sites using K possible categories of sites

Maximum likehood algorithm in practice

- **Pick an evolutionary model** (result of modelgenrator can help)
- For each site, generate **all possible tree structures** (same methods as in MP)
- Based on the e**volutionary model**, calculate **likelihood** of these trees.
- **Choose the tree with the Maximum Likelihood**

Maximum likehood in practice

- **+:** Works well for distantly related sequences and under different molecular clock theory ; Can incorporate any desirable evolutionary model ; Sound mathematical foundations

- **-:** Bad Approx. under Bad Evolutionary Models ; Computationally Intensive (=>slow)

- **Sotwares for Maximum likehood**

- PHYLIP (dnaml, protml)
- PhyML http://atgc.lirmm.fr/phyml/
- RaXML http://sco.h-its.org/exelixis/web/software/raxml/index.html

Maximum likehood: training

Exercise 4: build a *Maximum likehood* tree for the two alignement datasets *(use Seaview or phyml)*

Bayesian phylogenetic inference

- **The most recent method, now** becomes very used
- **Use probabilistic evolutionary models** (the same as in maximum likehood methods)
- The central concept of the method is **posterior probability**; a Bayesian analysis produces a posterior probability distribution of trees
- If the data are informative, most of the posterior probabilities will focus on **one tree or a small subset of trees**

Bayesian phylogenetic inference

Central question: what is the probability of the model/tree taking into account the data D ?

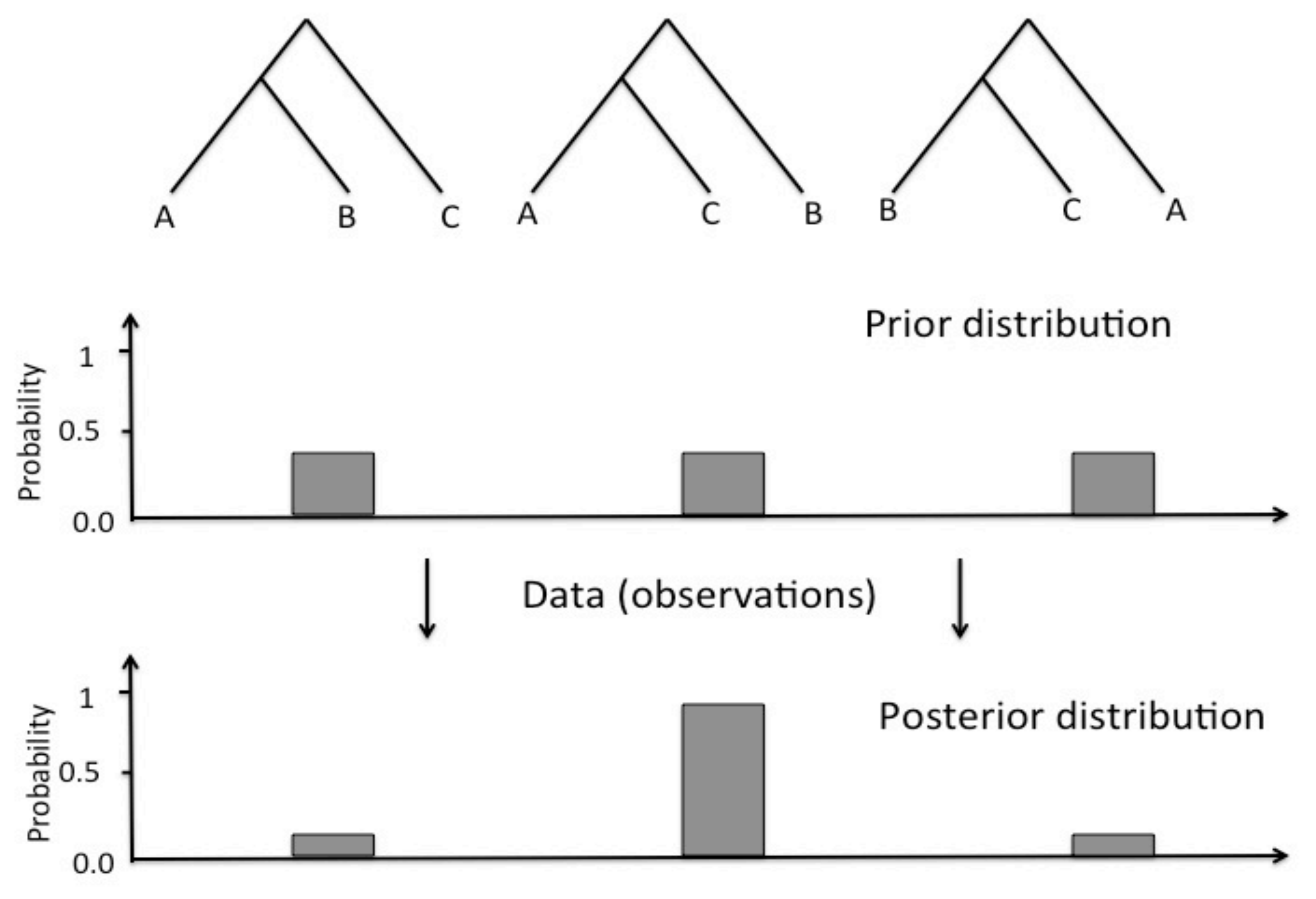
- **Start with a prior belief about trees** (prior distribution of possible trees)
- Collect data and use **an evolutionary model** and **Bayes theorem** to obtain a **posterior probability distribution of trees**

prior ***likehood***

$$\Pr(T / D) = \frac{\Pr(T)\Pr(D / T)}{\Pr(D)}$$

data probability

Bayesian phylogenetic inference



Bayesian phylogenetic inference

- Is is not possible to derive the posterior probability analytically
- The posterior probabilty is derived by using a **Markov Chain Monte-Carlo sampling (MCMC)** strategy:

1- start from an arbitrary point

2- make small random changes to the current values of the model parameters

3- accept or reject these changes according to its posterior probability

- This process is repeated during **n generations** until **convergence**.

Bayesian phylogenetic inference

- **Input:**
  - A set of aligned sequences
  - A prior distribution about trees
  - An underlying evolutionary model.
- **Desired Output:**
  - **One (or a few) valuated tree(s)** with maximal posterior probabilities

Bayesian phylogenetic inference

- **Powerful** but **complex** method
- Can produce **either one either several tree topologies** with high posterior probabilities
- Use an **a priori distribution** for parameters
- Use **heuristic to explore tree spaces**
- **Convergence problems:** for some phylogenetic problems, difficult or impossible to achieve convergence within a reasonable number of generations

Sotware for Bayesian inference

- **MrBayes** : http://mrbayes.sourceforge.net

Mr Bayes parameters

- **Set the evolutionary model,** eventually with a discrete gamma-distributed rate variation across sites (N=4) and a proportion of invariable sites (I) (or let MrBayes choose)
- **Set the MCMC parameters:**
  - **Number of chains Nc:** by default Nc=2 and MrBayes will run two simultaneous, completely independent analyses starting from different random tree
  - **Number of generations Ngen : typically** Ngen≥10000
  - **Criterion for convergence diagnostic,** typically by comparing the variance among and within tree samples MrBayes will run diagnostic every **runfreq** generations and report clades ot at least **minfrequency.**

Bayesian inference: training

Exercise 5: Build a MrBayes tree for the two datasets.

Bayesian inference: training

Exercise 5: Build a MrBayes tree on the nucleic dataset.

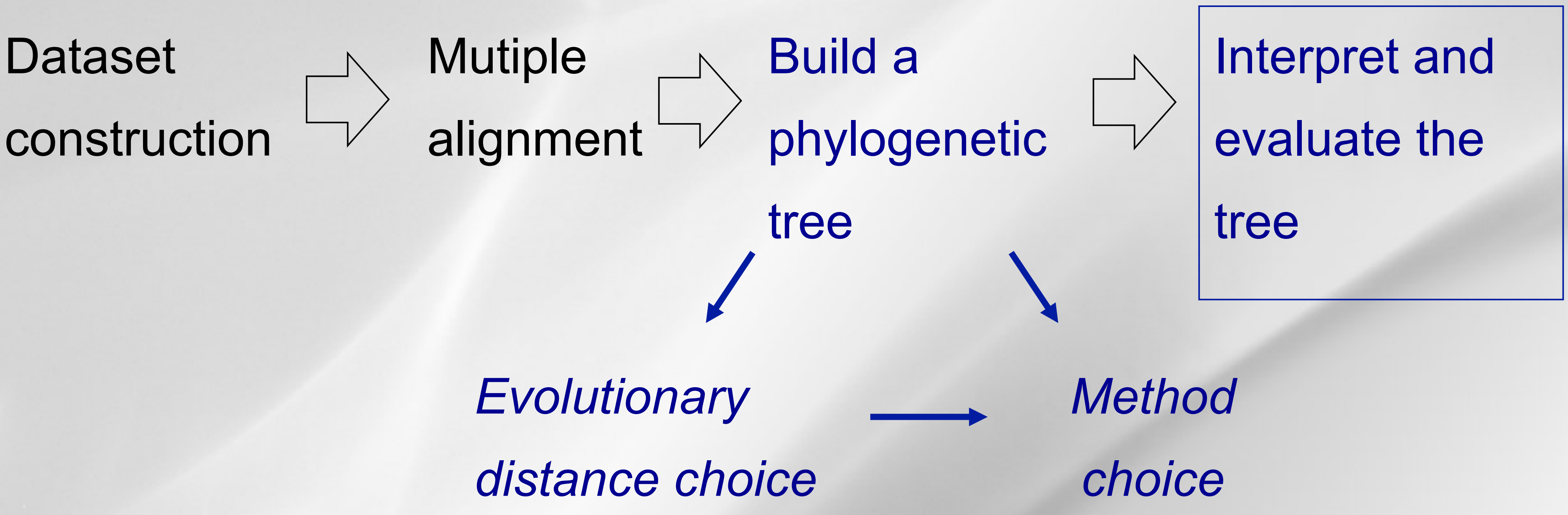
(i) Read the nexus datafile

(ii) Set the evolutionary model

(iii) Run the analysis

(iv) Summarize the samples

Usual workflow in phylogenetic analysis



Dataset construction

Mutiple alignment

Build a phylogenetic tree

Interpret and evaluate the tree

Evolutionary distance choice

Method choice

Testing tree topologies

Confidence issue

- How confident are we on the inferred tree ?
- Which parts of the tree are reliable/not reliable ?
- How can we validate the tree ?

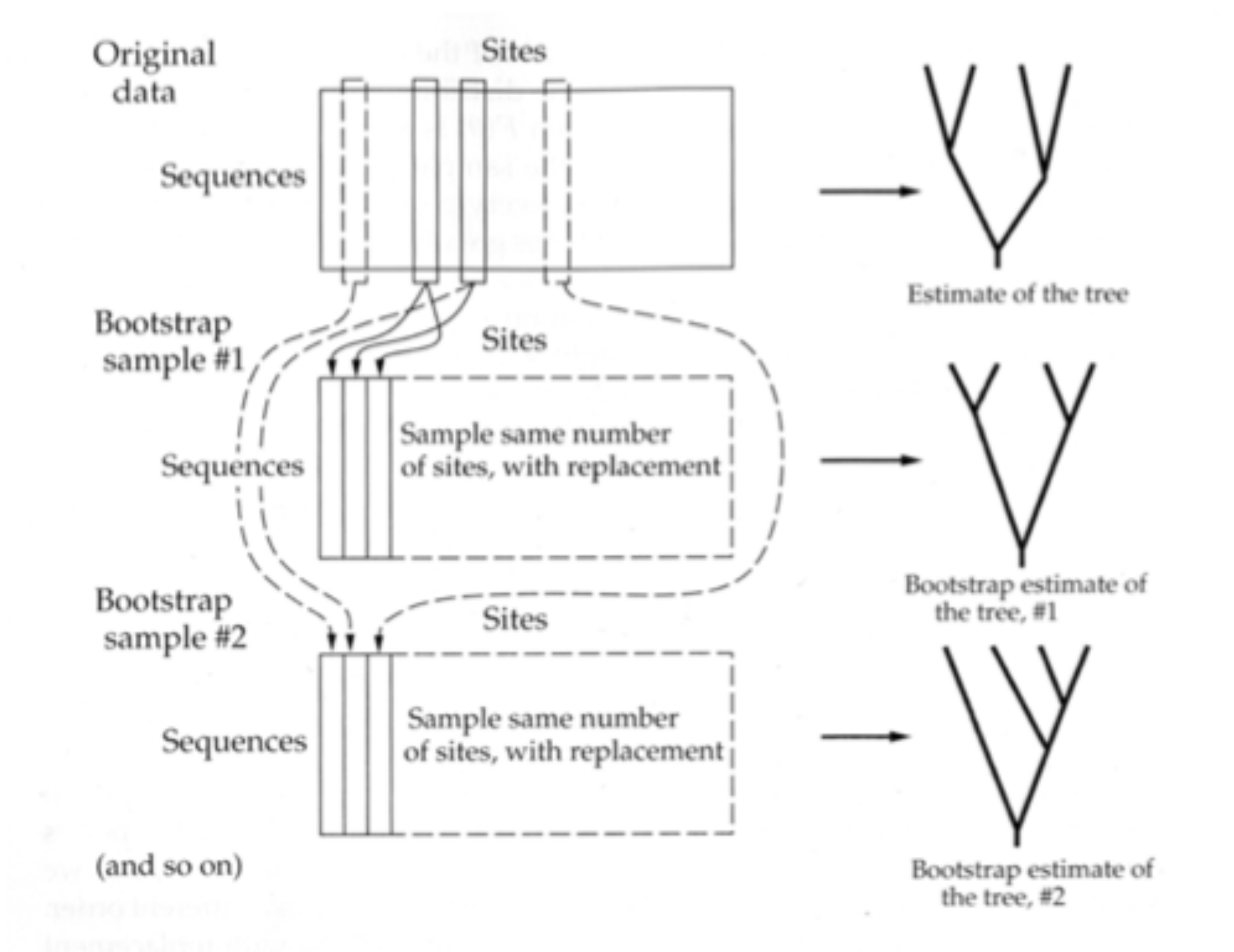
Problem: the true tree is unknown !

Solution :

- use **bootstrap** (or jacknife) to evaluate the **reliability** of the inferred tree and specific clades
- combine **subsampling** and **consensus trees** to get support values on branches

Testing tree topologies

- **Bootstrap**: resample “nucleotides” from the alignment;


Original data
Sites
Sequences
Estimate of the tree
Bootstrap sample #1
Sites
Sequences
Sample same number of sites, with replacement
Bootstrap estimate of the tree, #1
Bootstrap sample #2
Sites
Sequences
Sample same number of sites, with replacement
Bootstrap estimate of the tree, #2
(and so on)

Bootstrap process and consensus tree

Bootstrap process

- Infer **several trees** using **resampling techniques;**
- Identify and conserve only the **core information contained and repeated in many trees ;**
- Combine the several trees to produce a **consensus tree** which is compatible with all (or most) of the trees**.**
- In general, the consensus tree has **no branch lengths** and **a lower resolution** than the original tree**.**
- Superimpose boostrap values **on the original tree**

Consensus tree

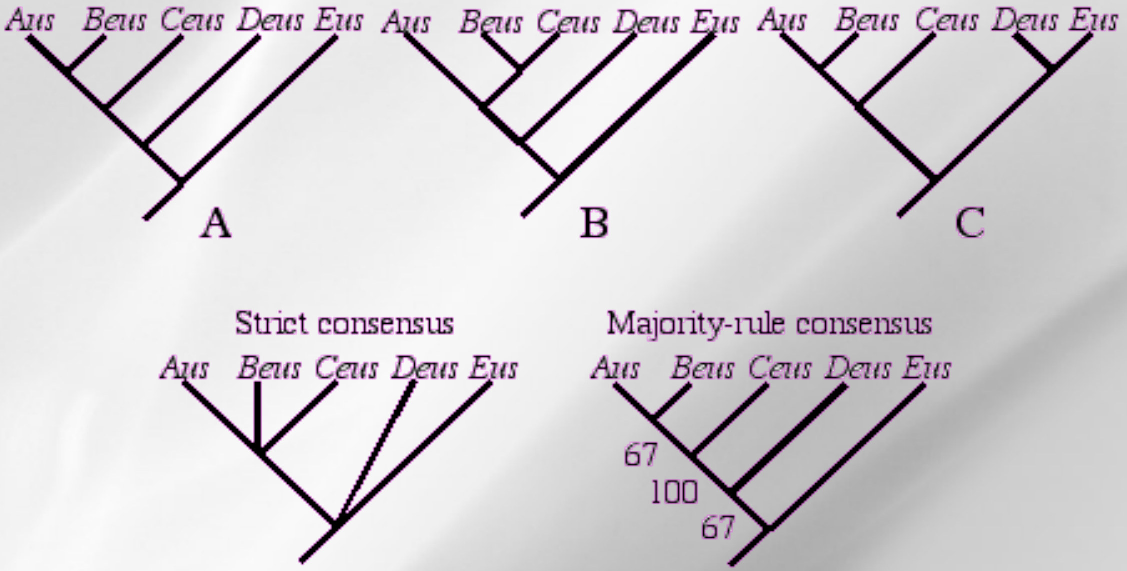
.**Consensus rules:**

- **Strict Consensus:** clades presents in all trees;
- **Majority Rule:** clades presents in at least half of the trees;
- **Extended Majority Rule:** clades presents in at least half of the trees and some more until the tree is resolved.

Consensus tree

.Consensus rules:

- **Strict Consensus:** clades presents in all trees;
- **Majority Rule:** clades presents in at least half of the trees;
- **Extended Majority Rule:** clades presents in at least half of the trees and some more until the tree is resolved.


Aus
Beus
Ceus
Deus
Eus
A
B
C
Strict consensus
Majority-rule consensus
67
100
67

Boostrap values guidelines

Be cautious with boostrap values interpretation:

- Bootstrap values have no clear-cut statistical interpretation;
- A bootstrap value of 95% doesn’t mean that the corresponding clade has 95% chance of being “true”;
- Bootstrap values are **difficult to interpret quantitavely**.

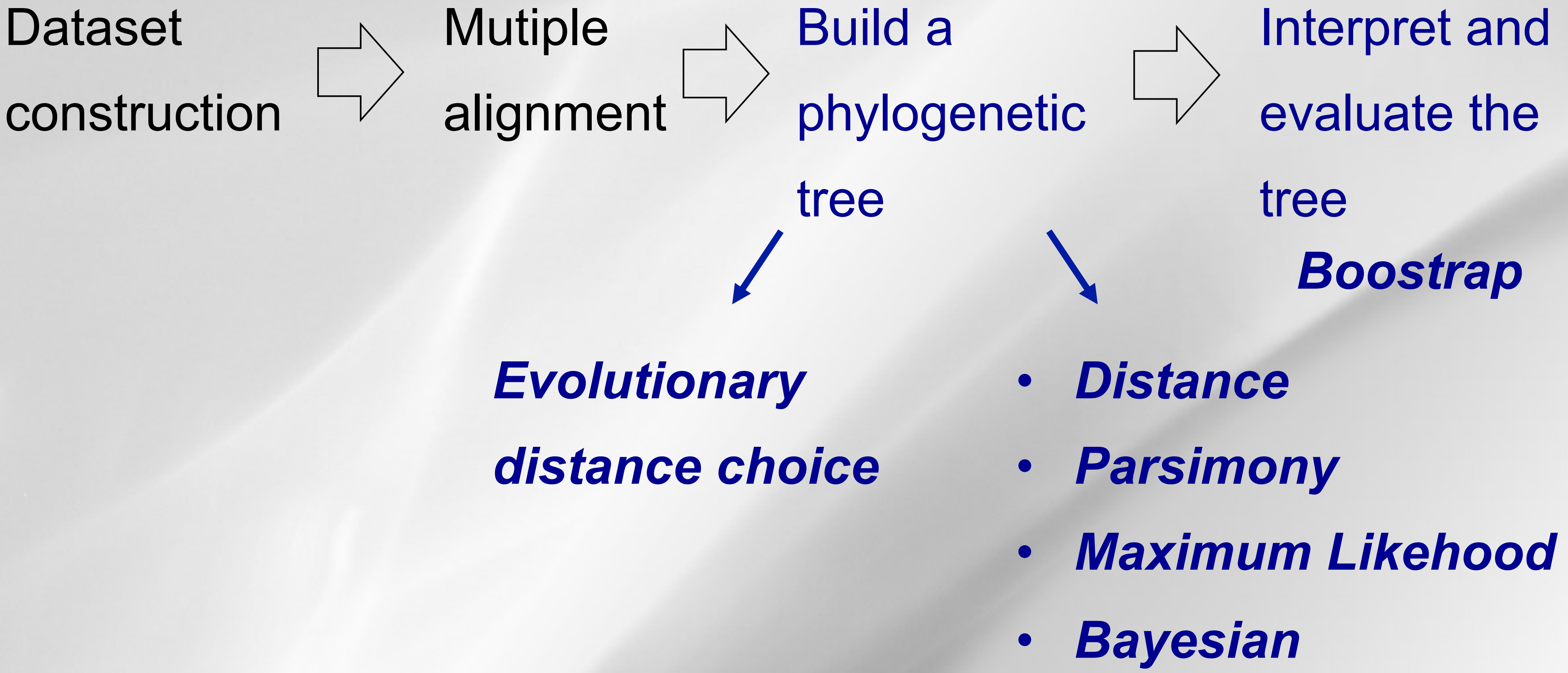
However Bootstrap values are (quite) easy to interpret **qualitatively**:

- The higher the bootstrap value, the more **confident** you can be in your clade;
- 95%, 90% and 66% consitute traditional threshold for being confident in a clade.

Bootstrap: training

Exercise 6: Compute bootstrap values for the bioNJ and parsimony trees.

Conclusion: overall view



Dataset construction

Mutiple alignment

Build a phylogenetic tree

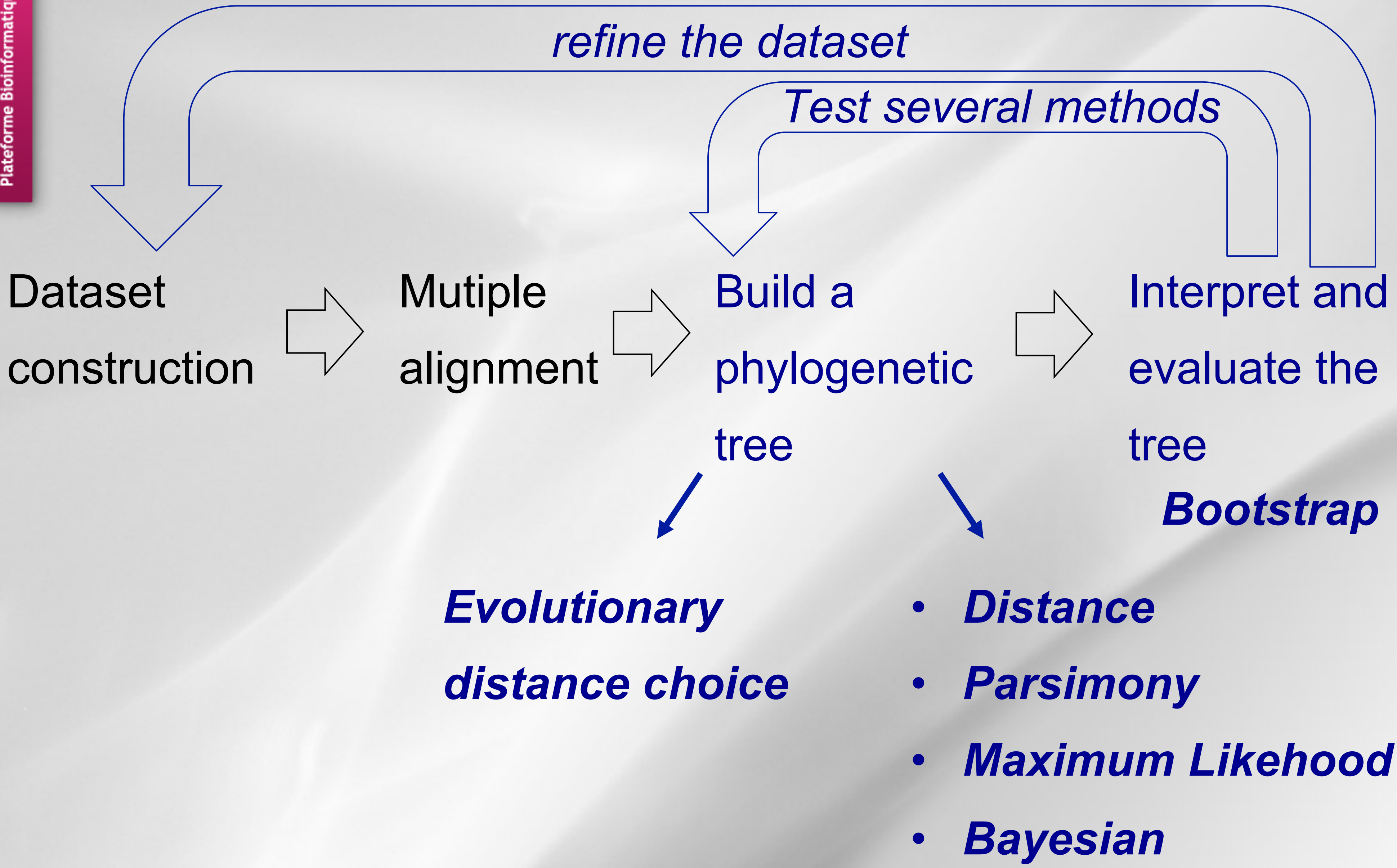
Interpret and evaluate the tree

Boostrap

Evolutionary distance choice

- ***Distance***
- ***Parsimony***
- ***Maximum Likehood***
- ***Bayesian***

Conclusion: overall view



refine the dataset

Test several methods

Dataset construction

Mutiple alignment

Build a phylogenetic tree

Interpret and evaluate the tree

Bootstrap

Evolutionary distance choice

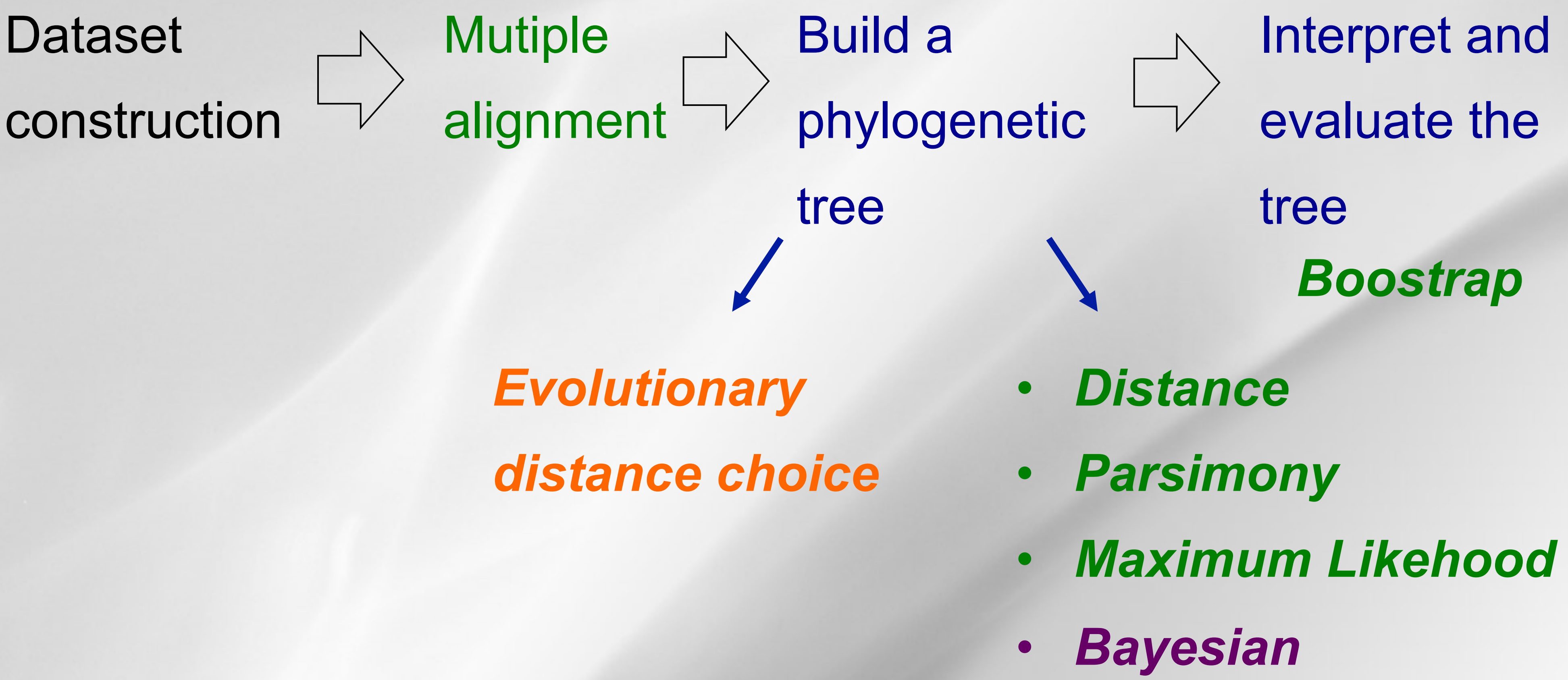
- ***Distance***
- ***Parsimony***
- ***Maximum Likehood***
- ***Bayesian***

Conclusion: overall view

Implemented in Seaview

Use modelgenerator.jar

Use Mr Bayes



Dataset construction

Mutiple alignment

Build a phylogenetic tree

Interpret and evaluate the tree

Boostrap

Evolutionary distance choice

- ***Distance***
- ***Parsimony***
- ***Maximum Likehood***
- ***Bayesian***

Conclusion: method comparison

Neighbor-joining (fast)

- Consistent: proven to construct the correct tree if distances are patristic.
- Problems with long and divergent sequences

Parsimony (medium)

- good for closely related sequences
- can be used with any kind of data
- No clear interpretation of branch length

Conclusion: method comparison

Likelihood method (slow)

- Sound statistic foundations
- Works well for distantly related sequences
- Can incorporate any desirable evolutionary model

Bayesian method (very slow)

- Powerful but complex method

Frequent problems

- **Long Branch Attraction:** Long branches tend to cluster together in the tree:

Solution: “break down” long branches by adding some taxa to the analysis;

- **Saturation:** Characters have evolved for so long that they are almost random:

Solutions: Remove saturated sites and/or taxa; When available, use proteic sequences instead of nucleic sequences;

- **Missing Data:** Some characters are missing from the alignment:

Solutions: Use methods that can handle missing values, such as ML; Use as many characters as possible.

Bibliography

- **The Phylogenetic Handbook.** A Practical Approach to Phylogenetic Analysis and Hypothesis testing. 2009, 2nd edition. Edited by Philipppe Lemey, Marco Salemi and Anne-Mieke Vandamme, eds. Cambridge University Press, Cambridge, U.K. 723 pp.
- **Concepts et méthodes en phylogénie moléculaire.** 2010. Guy Perrière and Céline Brochier-Armanet. Collection IRIS. Springer. 250 pp.
- **Computational Molecular Evolution.** 2006. Ziheng Yang. Oxford Series in Ecology and Evolution, Oxford University Press. Oxford, U.K. 357 pp.
- **Inferring Phylogenies**. 2004. Joseph Felsenstein. Sinauer Associates, Inc. Sunderland, MA, U.S.A. 664 pp.

Useful web sites

- **LIRMM web site :**

http://phylogeny.lirmm.fr

- **PHYLIP (Felsenstein lab, Univ. of Washington) web site :**
http://evolution.gs.washington.edu/phylip/software.html

Plateforme Bioinformatique Midi-Pyrénées

The End !

Questions ?