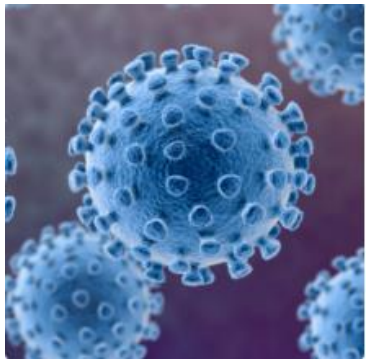


Viral Genomics and Bioinformatics (Latin America and the Caribbean)

10–14 October 2022 - Virtual course

wellcome connecting science

Introduction to phylogenetic methods

Carolina Torres

Universidad de Buenos Aires, Facultad de Farmacia y Bioquímica, Instituto de Investigaciones en Bacteriología y Virología Molecular (IBaViM), Buenos Aires, Argentina.

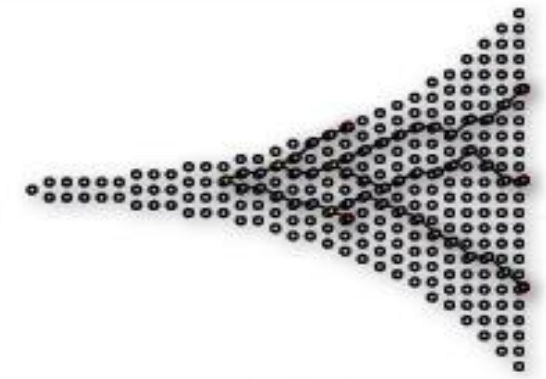
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina.

caro.torr@gmail.com | ctorres@ffyb.uba.ar

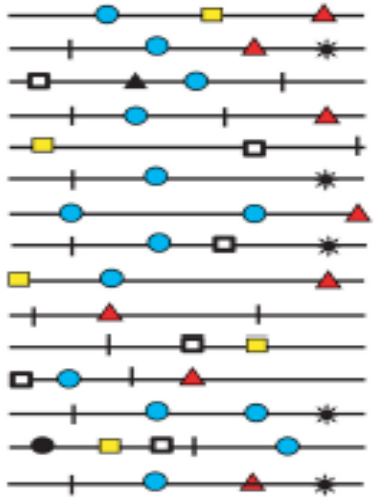
@CaroTorr

Phylogenetic analysis

Evolutionary history of the population

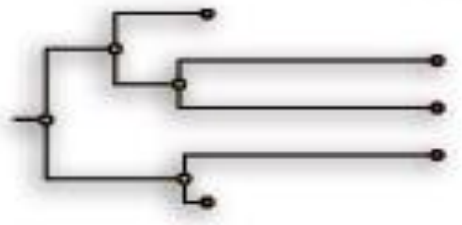


Sequences
(from some individuals of the population)



Genomes contain information about their origin

Models and methods
are used to retrieve this information

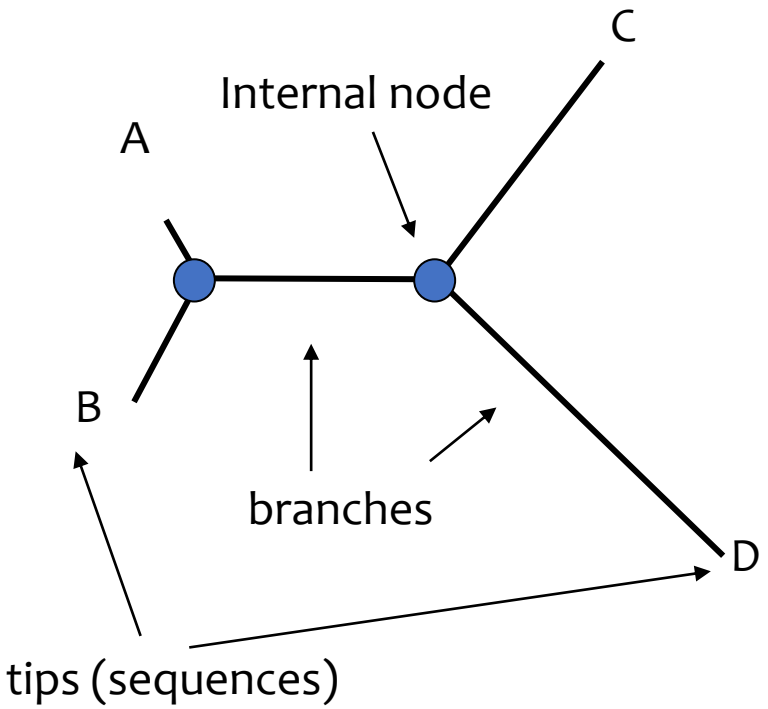


EVOLUTIONARY HYPOTHESIS

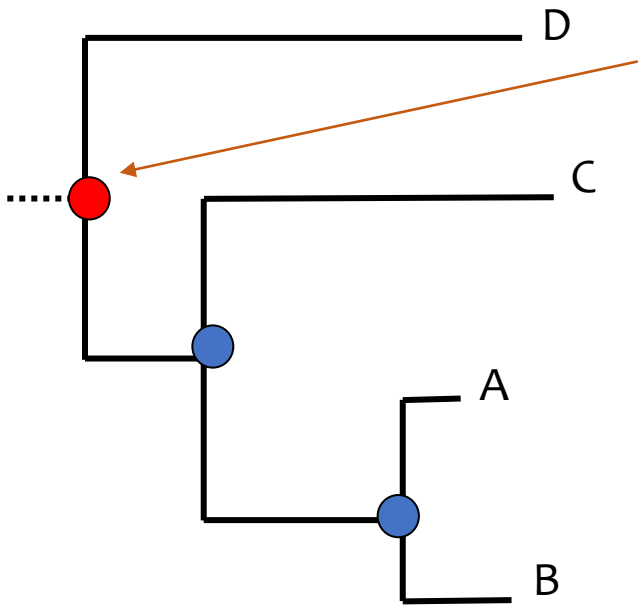
Phylogenetic analysis allows us to reconstruct the evolutionary history and study the processes that gave rise to it.

Phylogenetic trees

Unrooted tree



Rooted tree

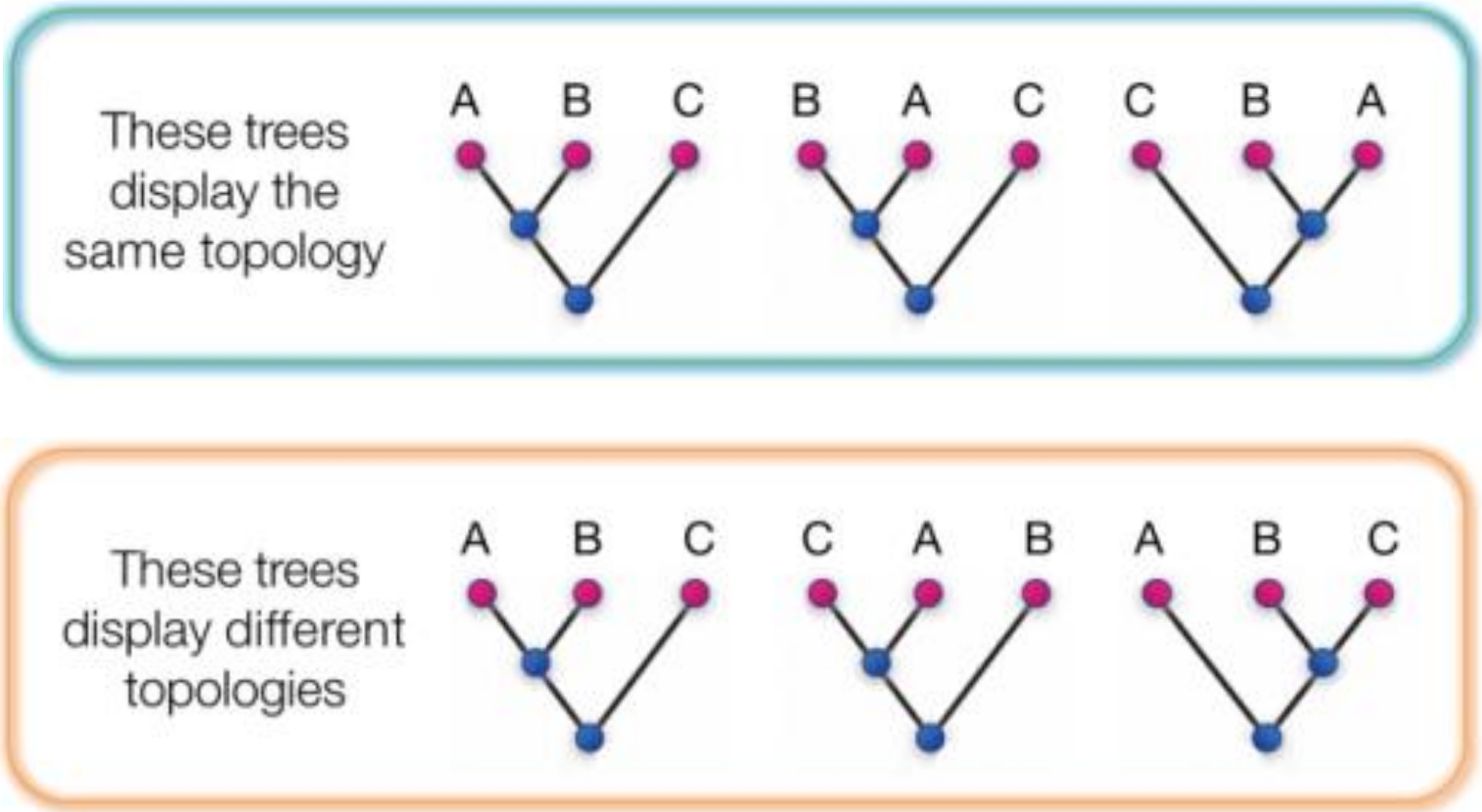


Root *(represents the most recent common ancestor of all sequences in the phylogeny)*

Direction of evolution (from past to the present)

Phylogenetic trees

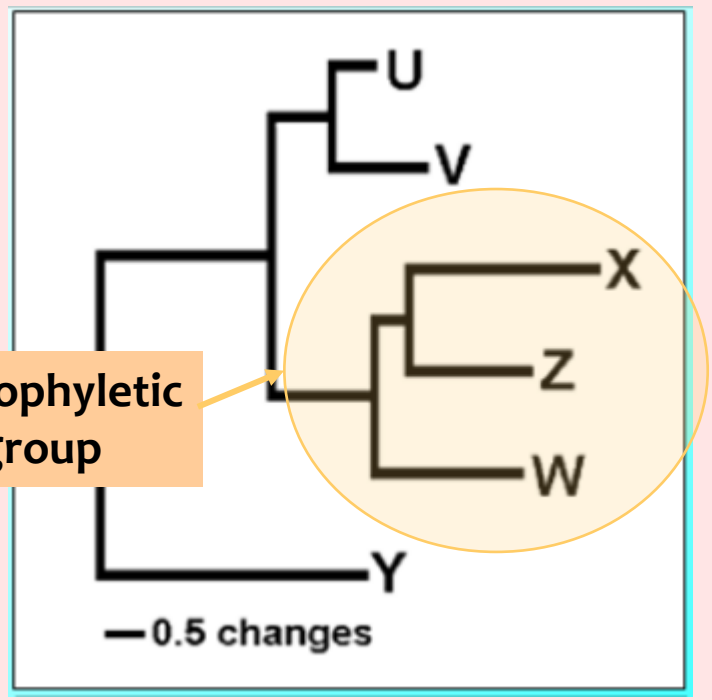
- The **topology** is the branching structure of the tree.



Branches can rotate freely over nodes

Phylogenetic trees

PHYLOGRAM



Displays information about the clustering of the sequences and their genetic distance

Monophyletic group:
group in which all sequences of interest share a common ancestor (without any other sequences within the group).

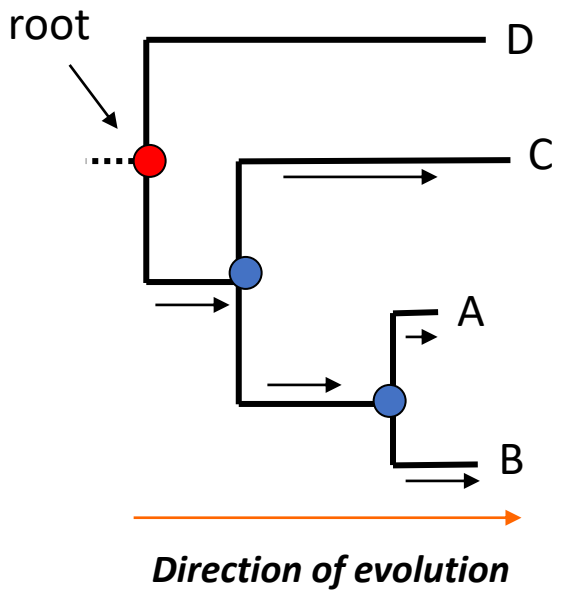
A common source of infection or transmission chain.

Phylogenetic trees

Methods for rooting phylogenetic trees

- **Outgroup Rooting**

  - ✓ The outgroup is a group of taxa whose diversification of the group of interest occurred before the internal diversification of the taxa under study (ingroup).
  - ✓ The outgroup must belong to a clearly different lineage from the ingroup, but it does not have to be so divergent that it cannot be aligned unambiguously.
  - ✓ The root is set on the branch that connects the outgroup to the ingroup.



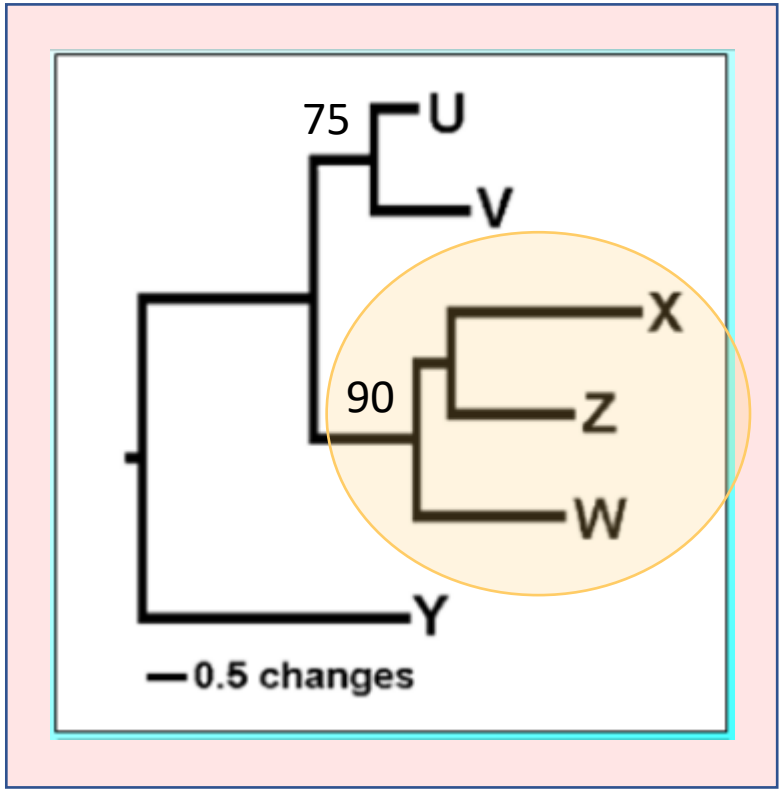
- **Midpoint Rooting**

  - ✓ The root is established at the midpoint distance between the two most divergent taxa.

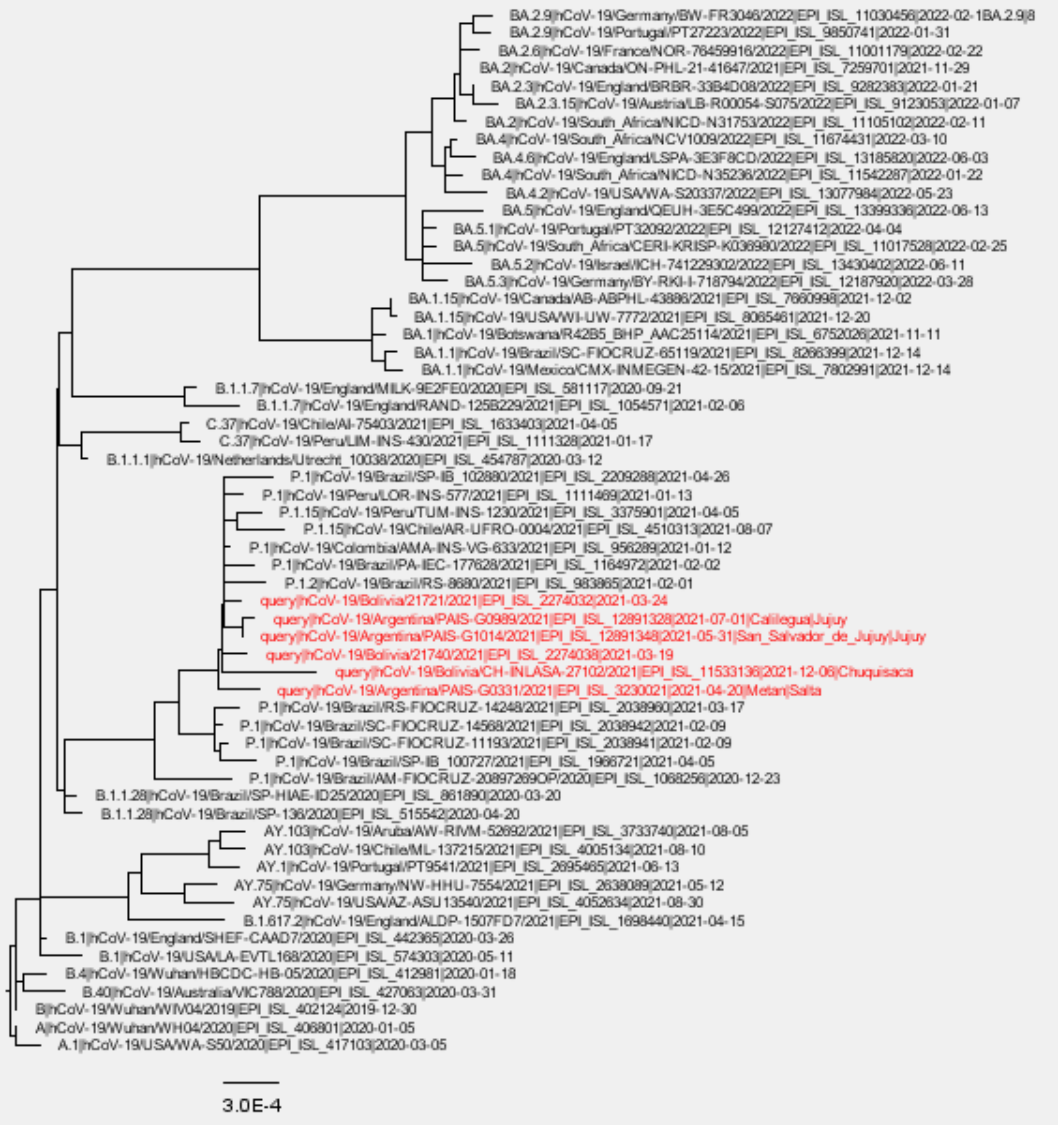
The root allows to establish the direction of the evolutionary process.

How are phylogenetic trees analysed and interpreted?

- **Root the tree** to know the direction of evolution (branch connecting the *ingroup* with the *outgroup* or midpoint).
- Analyse the groupings **from the root to the terminals**.
- Identify the **common ancestor of the sequences** of interest and analyse **with which sequences they are related** (with whom they share ancestors).
- Assess the **reliability of the clusters of interest** (how much evidence is there in the sequences for the formation of those clusters).
- Analyse the **evolutionary relationships together with the rest of the information** available (metadata): epidemiological, biological, temporal, geographical information, etc.



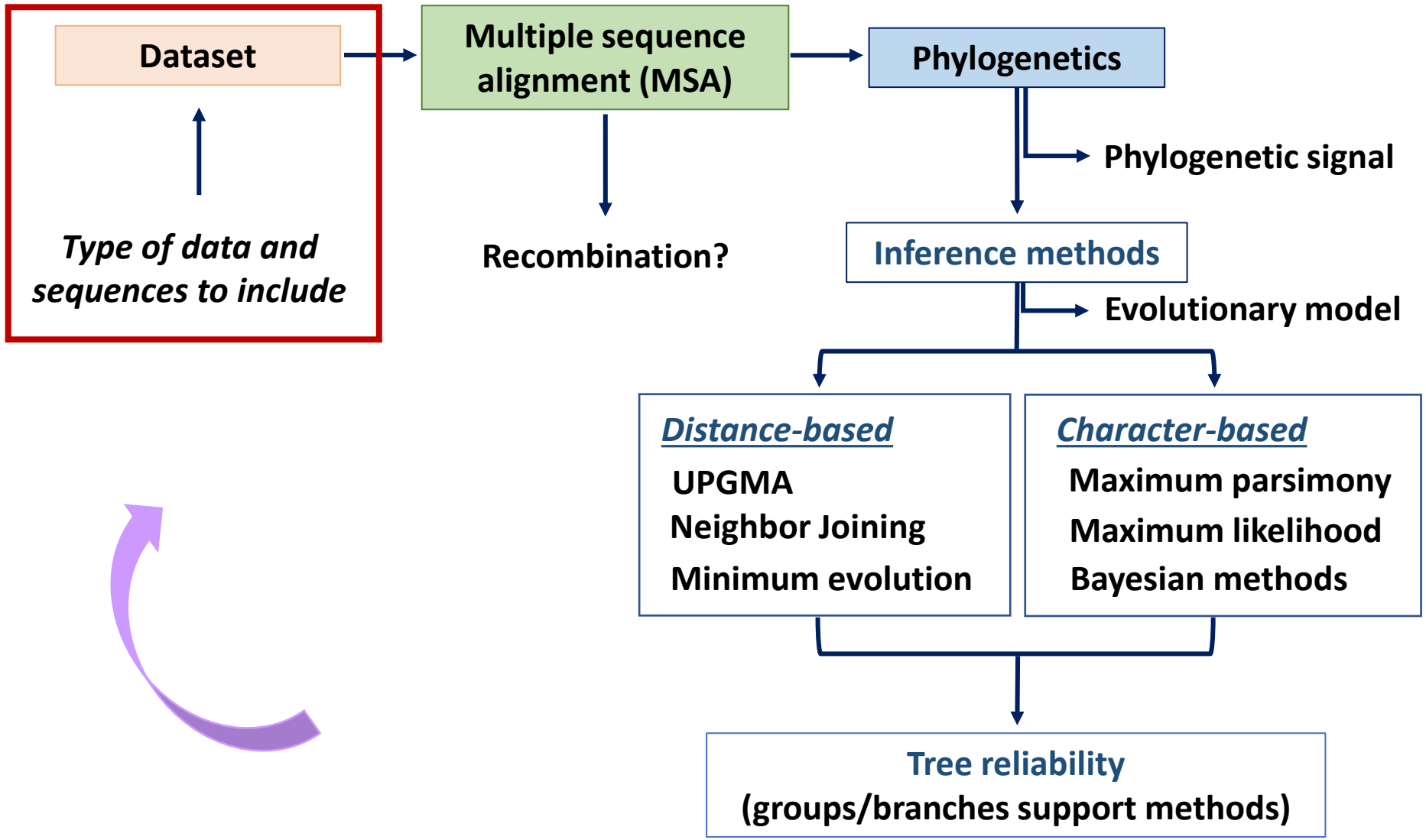
Phylogenetic analysis



- **Are the sequences in red related?**
- **Do they belong to the same transmission chain?**

…we must prepare a dataset that allows us to address this objective!

Steps for phylogenetic analysis



Sequence selection for molecular phylogenetics

- **What sequences to include?**
  - ✓ **Sequences under study**
  - ✓ **Reference sequences**
  - ✓ **Outgroups**
  - ✓ **Other sequences** (according to the objective): sequences of the organism under study **that circulates in the same geographical region** or **during the same period**, sequences with a **high % identity** or that show **specific characteristics of interest** (host, virulence, clinical behavior, etc).

The entire dataset has an impact on the results and the conclusions reached

- **Type of data**
  - ✓ **Nucleotide sequences**
  - ✓ **Amino acid sequences**
  - ✓ **SNPs**

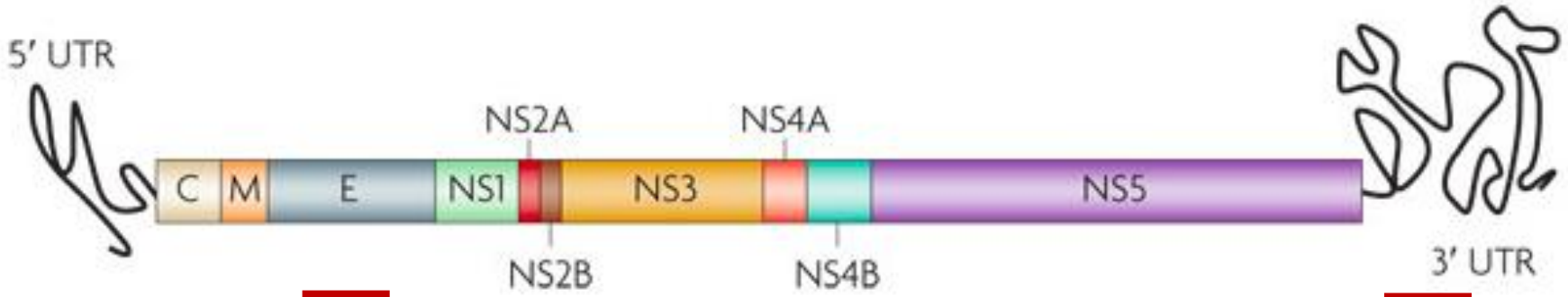
The sequences to include and the type of data to be analyzed always depend on the objective.

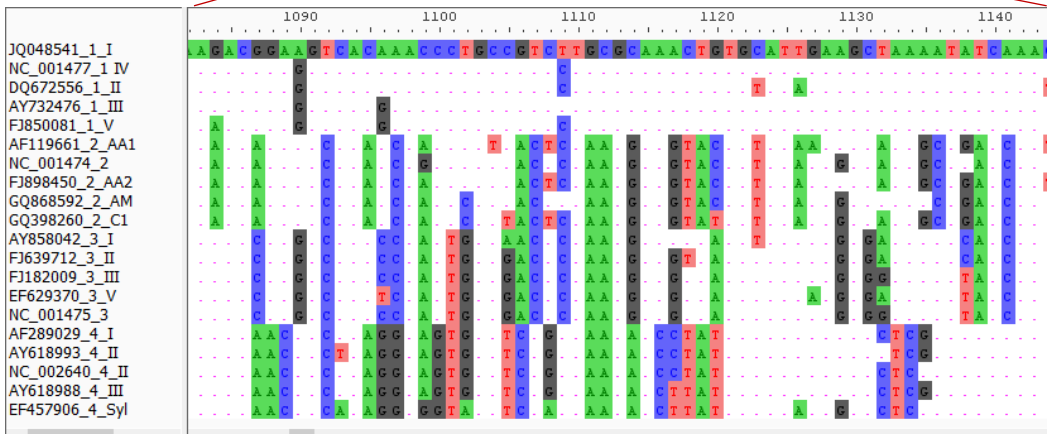
Type of data: Sequences

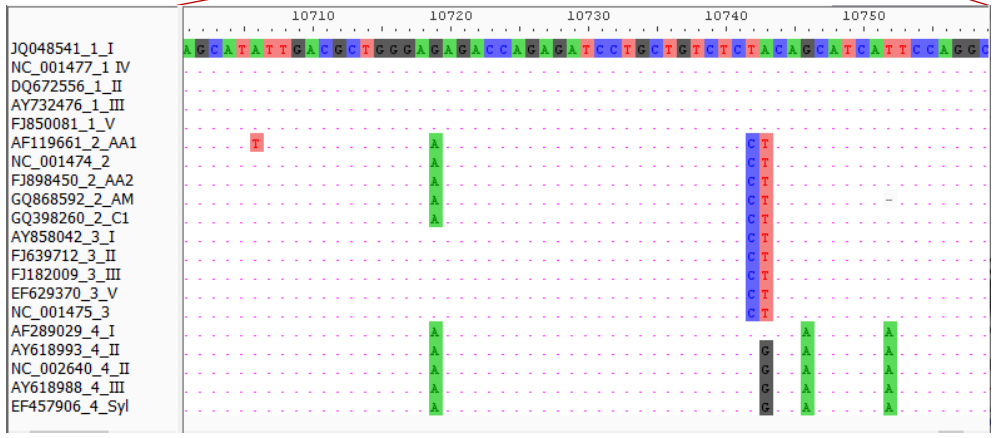
- ✓ **Nucleotides or amino acids**.
- ✓ **Complete genomes or partial sequences (which genetic region?).**

Adequate diversity for the taxonomic range under study (family, species, genotype, subgenotype, etc.).

Dengue virus genome scheme



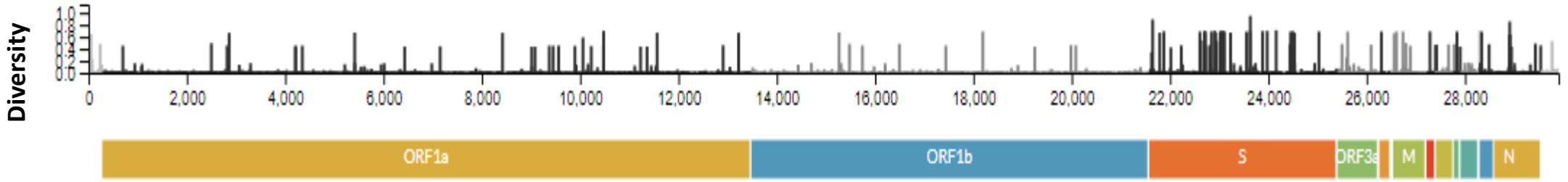




Type of data: Sequences

- ✓ **Nucleotides or amino acids**.
- ✓ **Complete genomes or partial sequences (which genetic region?).**

Adequate diversity for the taxonomic range under study (family, species, genotype, subgenotype, etc.).



In the case of SARS-CoV-2, the complete genome is usually used

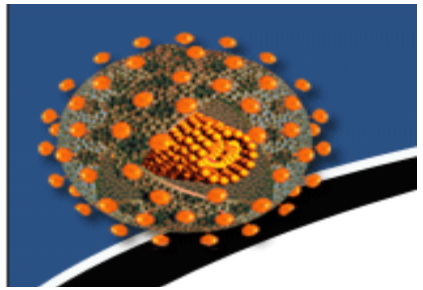
Databases

GenBank
National Center for Biotechnology Information
http://www.ncbi.nlm.nih.gov

International Nucleotide Sequence Database Colaboration

European Nucleotide Archive
https://www.ebi.ac.uk/ena

http://www.ddbj.nig.ac.jp/Welcome.html.en



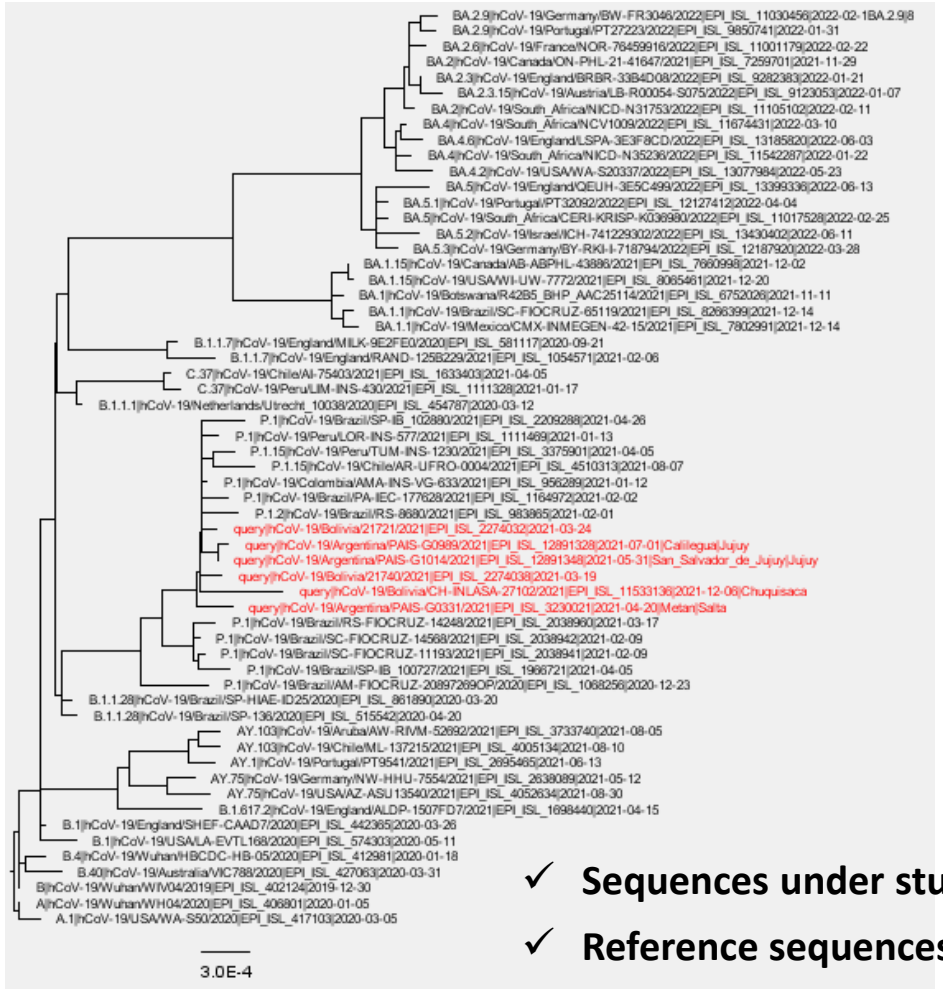
HIV DATABASES

Global Initiative on Sharing All Influenza Data

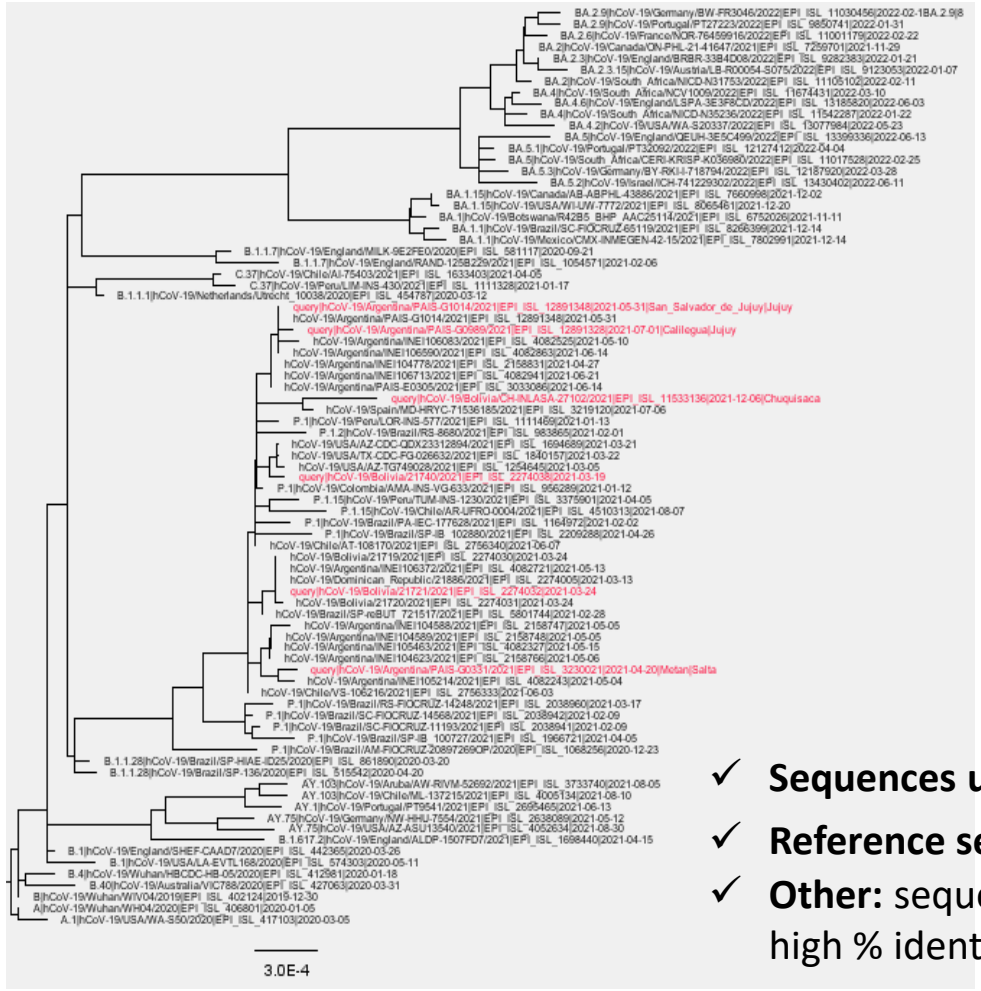
ViPR Virus Pathogen Resource

Phylogenetic analysis

- **Are the sequences in red related?**
- **Do they belong to the same transmission chain?**

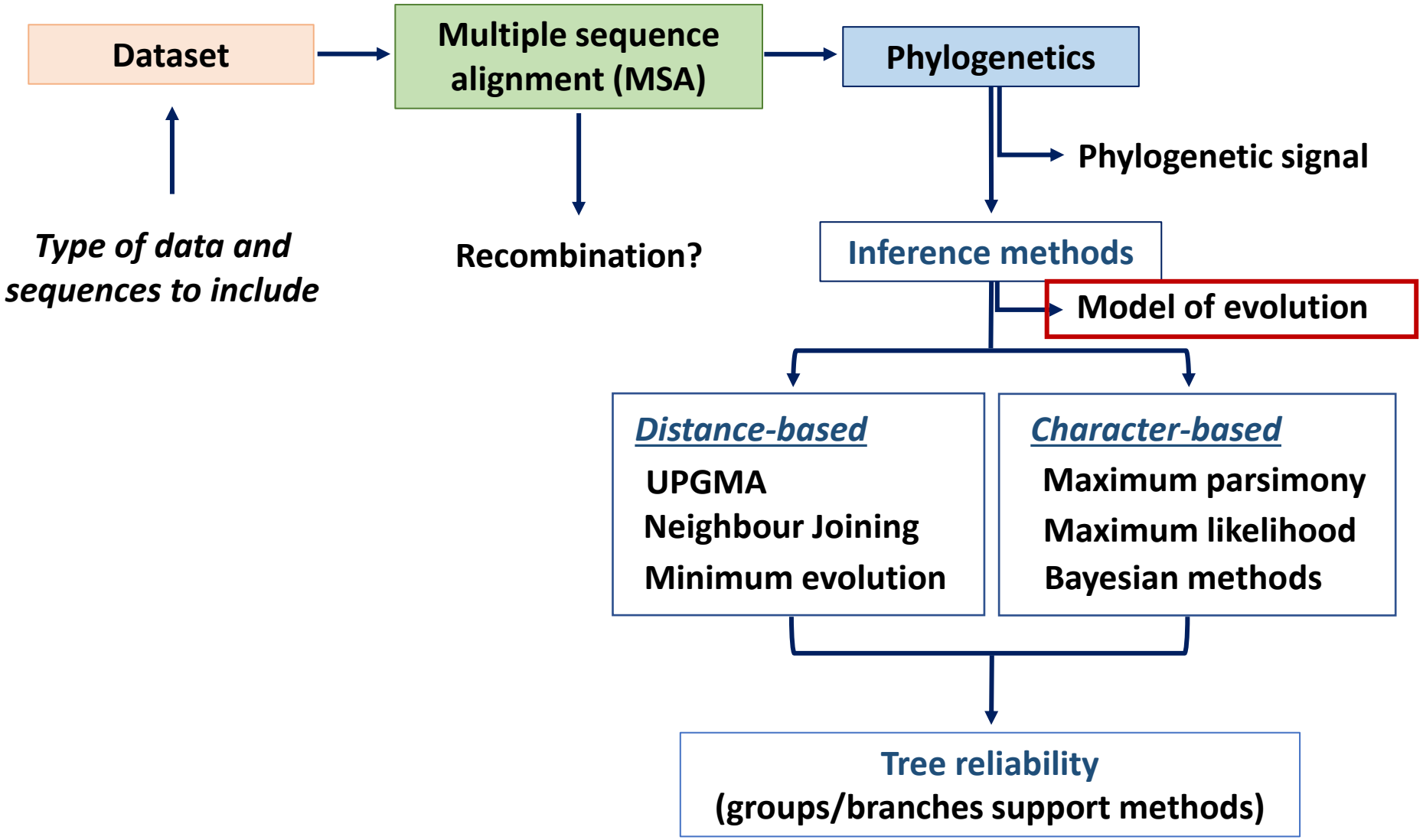


- ✓ **Sequences under study**
- ✓ **Reference sequences**



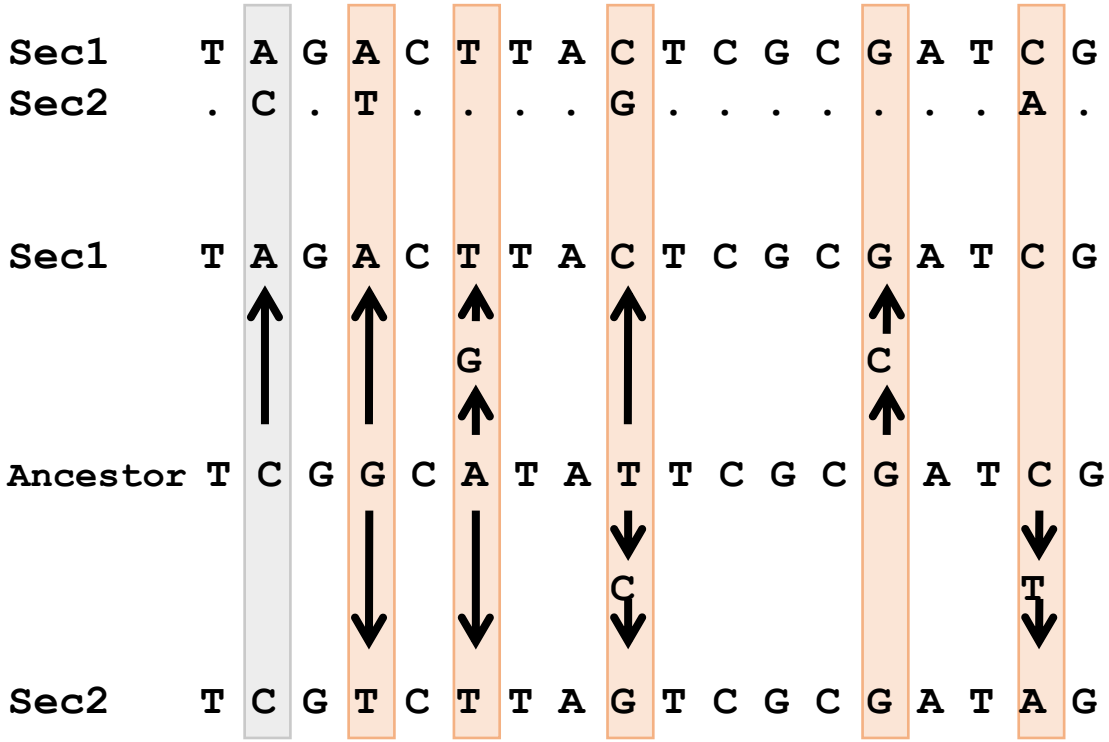
- ✓ **Sequences under study**
- ✓ **Reference sequences**
- ✓ **Other:** sequences with high % identity.

Steps for phylogenetic analysis

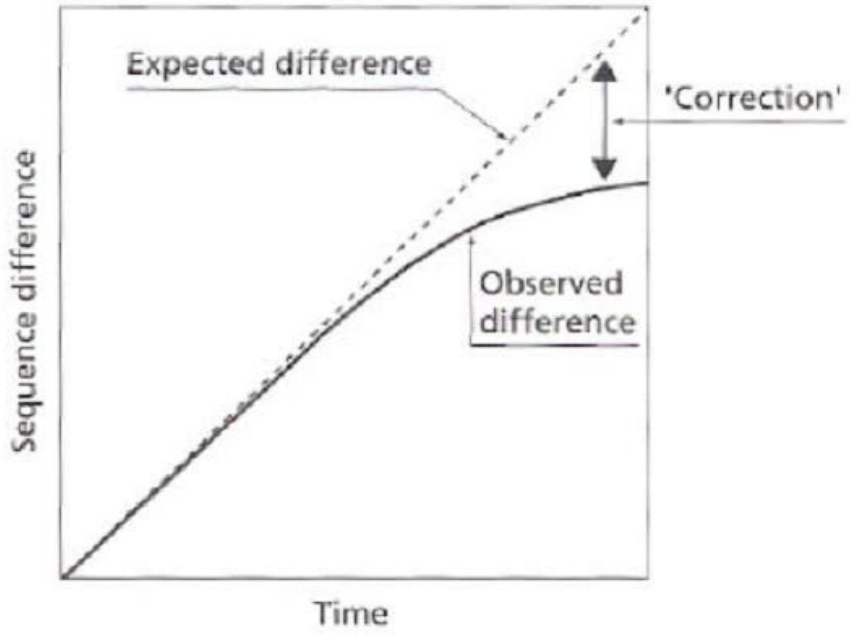


Models of molecular evolution

- They are mathematical formulations that model <u>the type and rate of change per site.</u>



Throughout the evolutionary history of sequences, some positions may have changed multiple times.



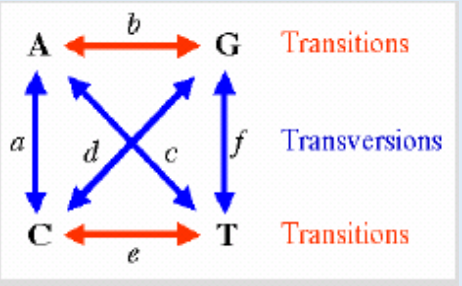
They allow distances to be "corrected" for mutations or unobserved changes

Models of molecular evolution

- **They are mathematical formulations that model <u>the type and rate of change per site.</u>**

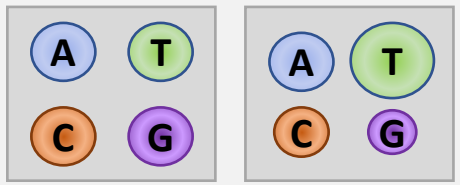
Model Parameters:

-Relative substitution rates (nts , AAs)



$$R = \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \quad \begin{matrix} A \\ C \\ G \\ T \end{matrix} \quad \begin{matrix} A & C & G & T \end{matrix}$$

-Frequencies of nts or AAs



$$\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$$

Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like `HKY` but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like `TN` but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like `K81` but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like `TPM2` but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like `TPM3` but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIMe	3	Like `TIM` but equal base freq.	012230
TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like `TIM2` but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like `TIM3` but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like `TVM` but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

- **Each site is assumed to evolve independently (nts, AAs, codons, etc.)**

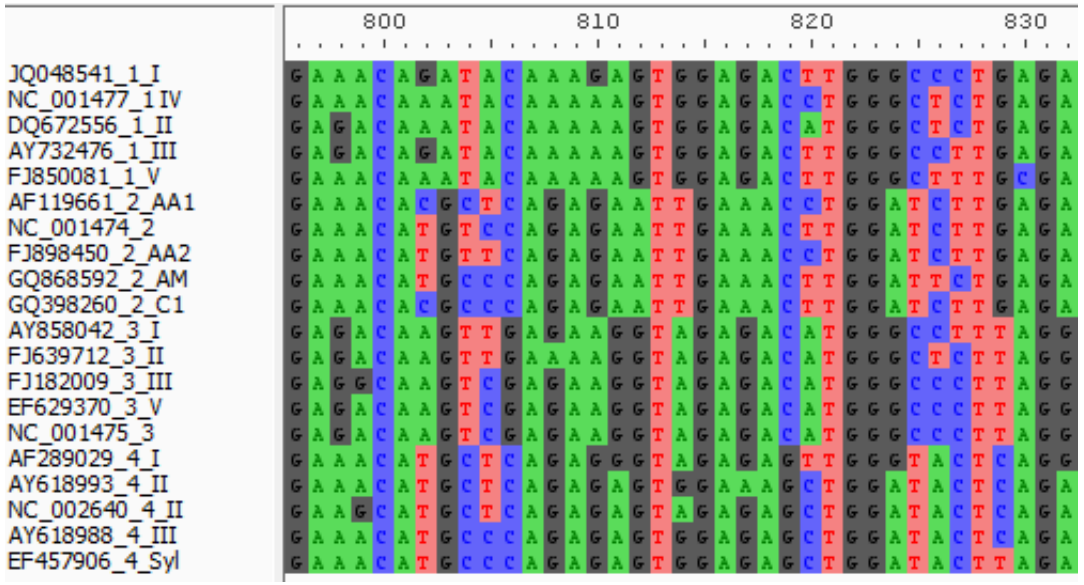
http://www.iqtree.org/doc/Substitution-Models#rate-heterogeneity-across-sites

Models of molecular evolution

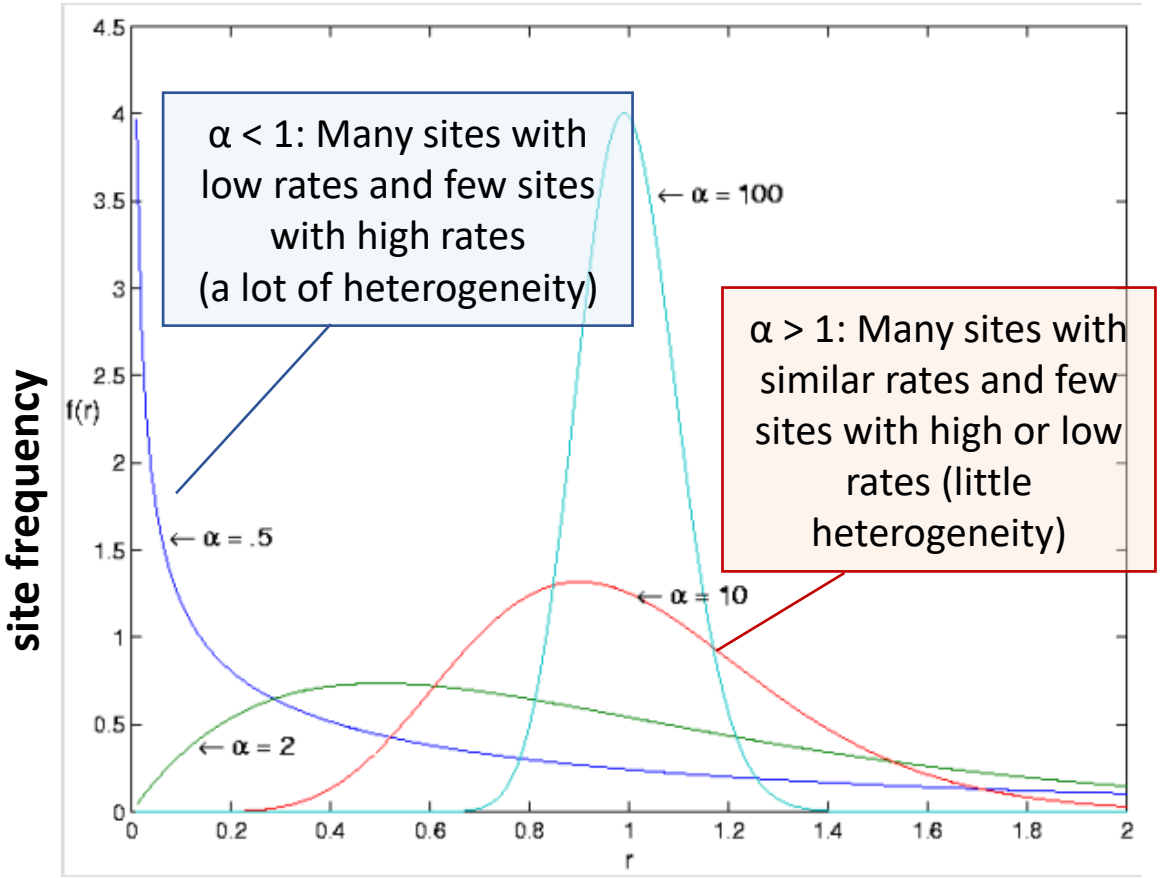
- **They are mathematical formulations that model <u>the type and rate of change per site.</u>**

Model Parameters:

-Heterogeneity of substitution rates per site



Gamma Distribution



Models of molecular evolution

- **They are mathematical formulations that model <u>the type and rate of change per site.</u>**

Model Parameters:

-Heterogeneity of substitution rates per site

+G: Gamma distribution (parameter α).

+I: proportion of Invariant sites (p-inv).

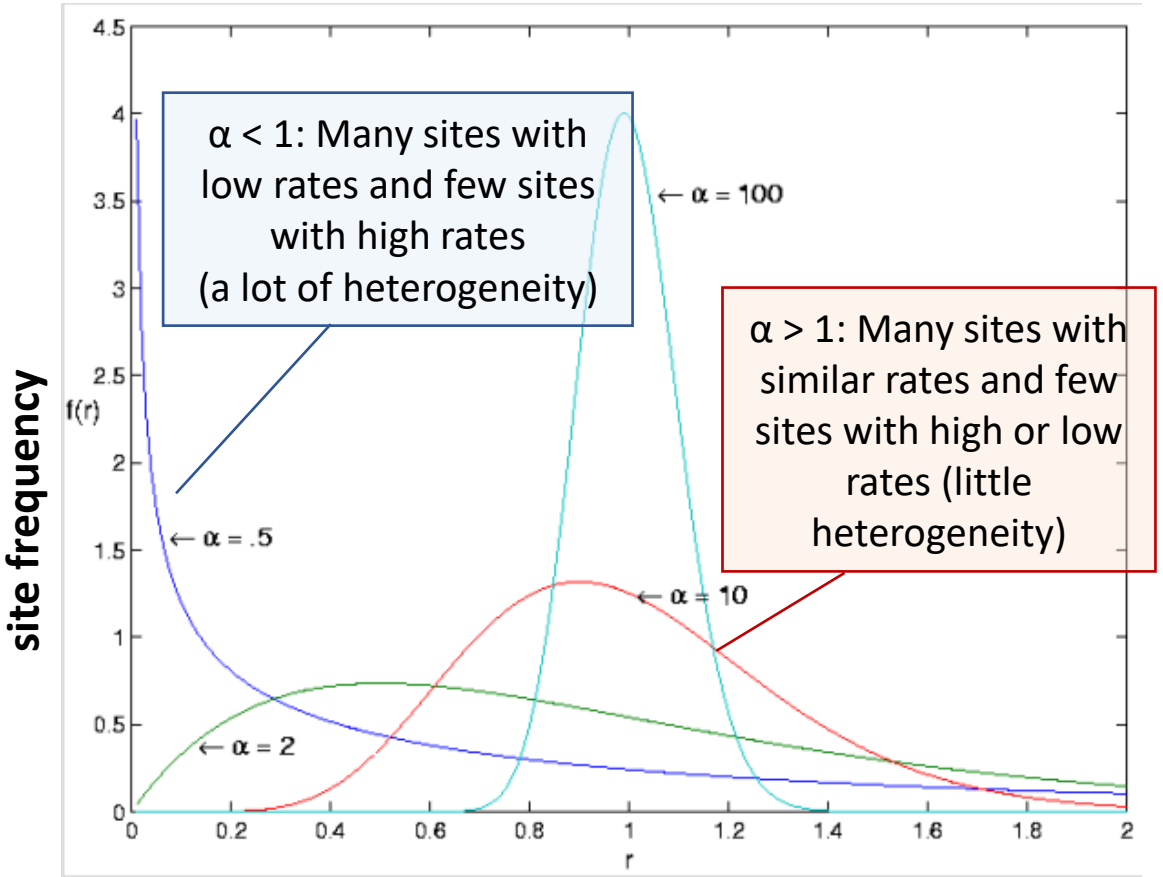
+G+I: Gamma distribution plus invariant sites.

+R: FreeRate model (generalization of the Gamma distribution assumption).

+R+I: FreeRate model plus invariant sites.

- **Programs:** ModelFinder (IQ-TREE), jModelTest , ProtTest , others.

Gamma Distribution



Models of molecular evolution

Example: Selection of the evolutionary model with IQ-TREE (ModelFinder)

```
ModelFinder will test 286 DNA models (sample size: 29796) ...
 No. Model          -LnL         df  AIC          AICc         BIC
  1  JC             46744.121    149 93786.242    93787.749    95023.259
  2  JC+I           46642.012    150 93584.023    93585.552    94829.343
  3  JC+G4          46675.504    150 93651.008    93652.536    94896.327
  4  JC+I+G4        46650.941    151 93603.882    93605.430    94857.503
  5  JC+R2          46719.714    151 93741.428    93742.977    94995.050
  6  JC+R3          46669.794    153 93645.587    93647.177    94915.813
  7  JC+R4          46639.294    155 93588.587    93590.219    94875.417
  8  JC+R5          46634.671    157 93583.342    93585.016    94886.777
 14  F81+F          45807.542    152 91919.084    91920.653    93181.008
 15  F81+F+I        45710.258    153 91726.516    91728.106    92996.742
 16  F81+F+G4       45740.064    153 91786.129    91787.719    93056.355
 17  F81+F+I+G4     45716.981    154 91741.961    91743.572    93020.489
```

(...)

```
274  GTR+F          45588.826    157 91491.652    91493.326    92795.086
275  GTR+F+I        45499.127    158 91314.253    91315.948    92625.989
276  GTR+F+G4       45523.194    158 91362.388    91364.083    92674.124
277  GTR+F+I+G4     45502.417    159 91322.834    91324.551    92642.873
278  GTR+F+R2       45492.606    159 91303.212    91304.929    92623.251
279  GTR+F+R3       45486.942    161 91295.885    91297.645    92632.528
Akaike Information Criterion:           GTR+F+R3
Corrected Akaike Information Criterion: GTR+F+R3
Bayesian Information Criterion:         GTR+F+R2
Best-fit model: GTR+F+R2 chosen according to BIC
```

selected model

```
Rate parameters:  A-C: 0.43193  A-G: 1.36320  A-T: 0.18530  C-G: 0.34865  C-T: 3.18238  G-T: 1.00000
Base frequencies:  A: 0.299  C: 0.183  G: 0.196  T: 0.322
Site proportion and rates:  (0.509,0.266) (0.491,1.762)
```

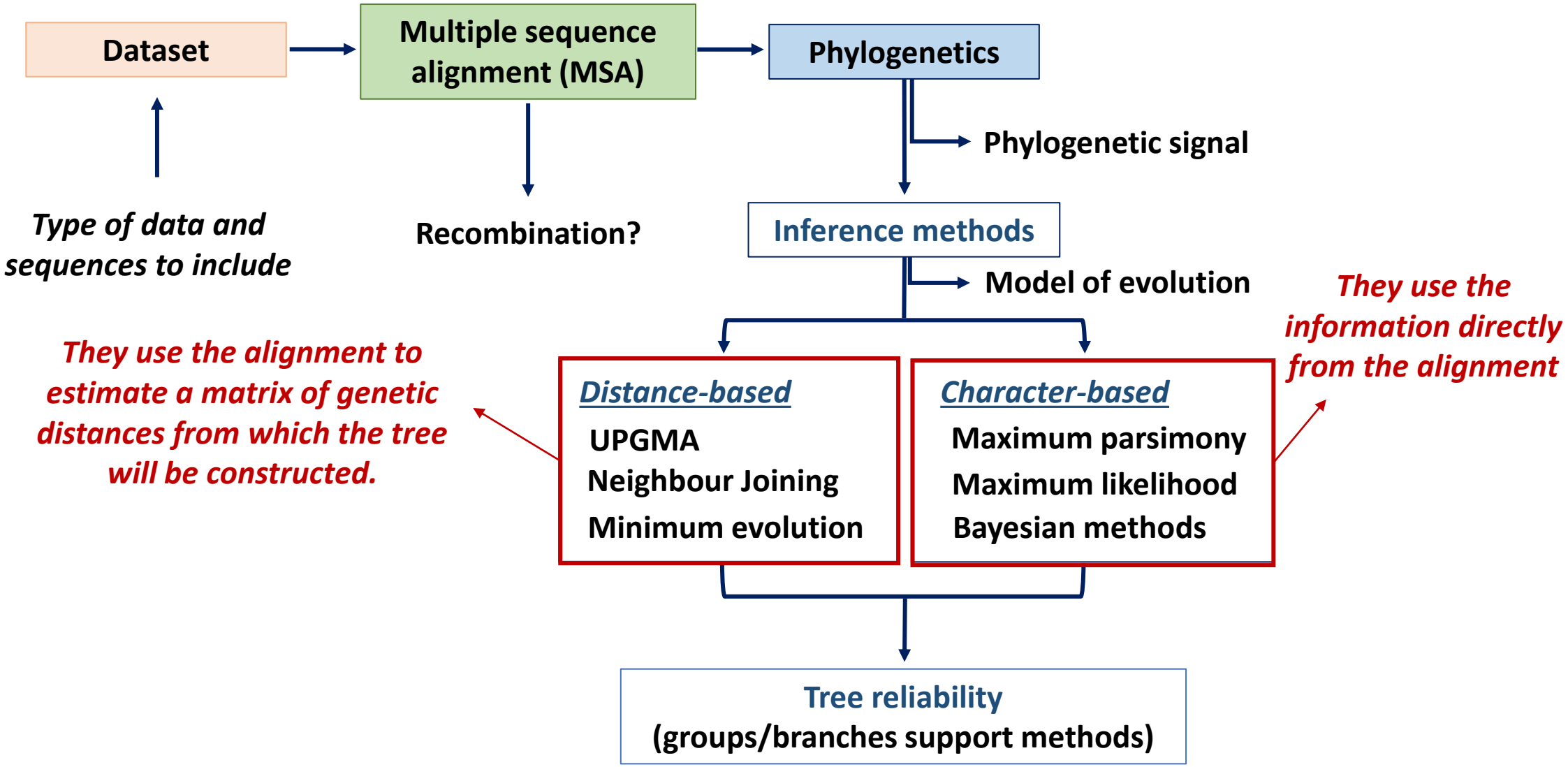
Parameters of the selected model

- **<u>The "fit" of the data (alignment) to several possible models is evaluated</u>, with parameter optimization.**
- **<u>A criterion (AIC or BIC) is applied to select the best model</u>: good fit (awarded) vs number of parameters (penalized).**
- **Complex models (more parameters) are selected only if their fit to the data is substantially better than simpler models.**

The model must be estimated for <u>each dataset</u>

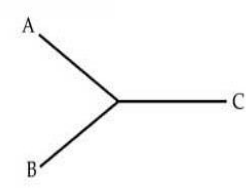
(There is no model that is specific to a gene, genome or organism...)

Steps for phylogenetic analysis

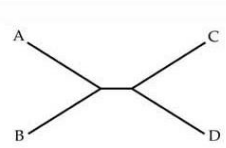


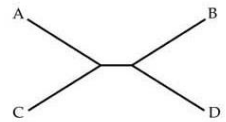
Tree space

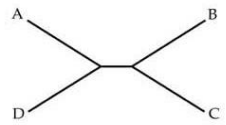
3 taxa



4 taxa







Number of taxa	Number of unrooted trees
2	1
3	1
4	3
5	15
6	105
7	954
8	10 395
9	135 135
10	2 027 025
20	**$2{,}22 \times 10^{20}$**
30	**$8{,}69 \times 10^{36}$**

Tree space

*In general, **a single tree** is obtained from the space of possible trees.*

Phylogenetic methods select one tree from the several millions/billions of possibilities.

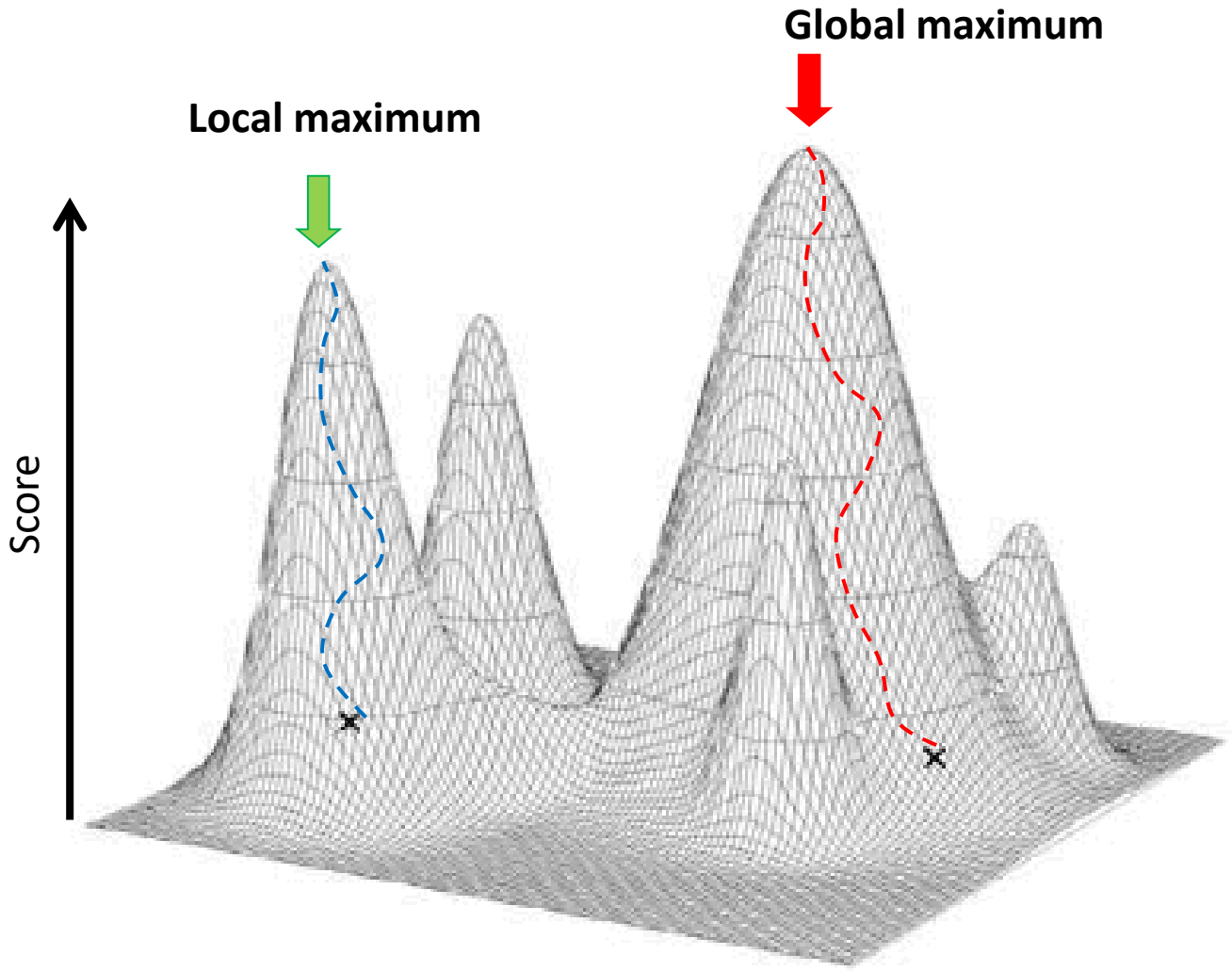
Methods of phylogenetic analysis

- **Distance-based: Tree building from genetic distance.**
  - Resume the data of the sequences into numbers: the “distance” to the other taxa.
  - The clustering methods group the taxa from the nearest to the farthest giving only one result (Neighbour Joining).
  - The methods that follow a criterion, evaluate alternative hypotheses and select the one that minimizes the distances between taxa.

- **Character-based: Tree building from character information.**
  - They evaluate alternative hypotheses (tree searching) and choose which ones (trees) meet the optimality criterion (Maximum likelihood, Maximum parsimony).
  - They explore the tree space and save some alternative hypotheses that will be summarized in one tree (Bayesian methods).

Methods of phylogenetic analysis

EXPLORATION OF TREE SPACE BY HEURISTIC SEARCH

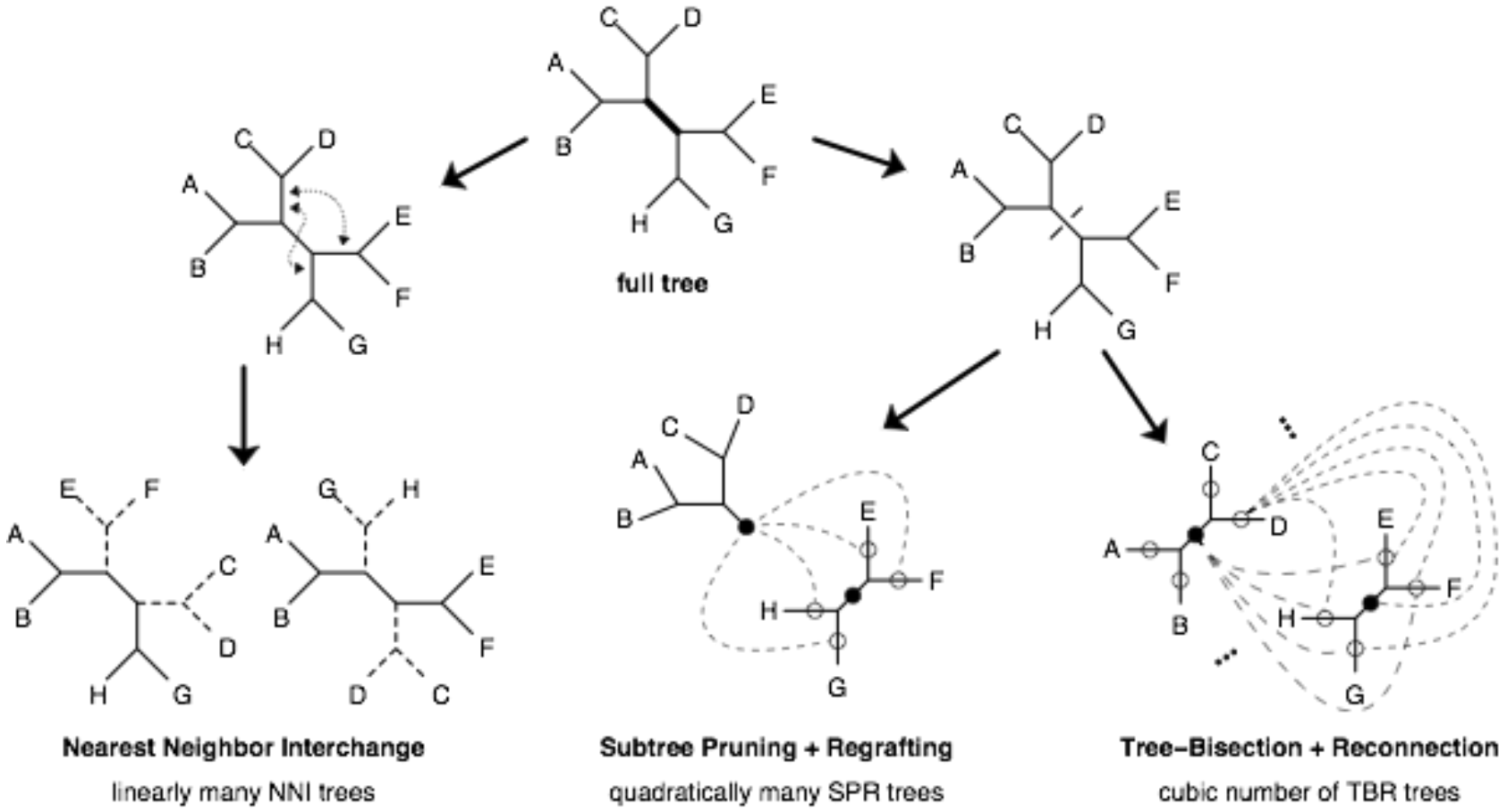


- First, the initial tree (NJ or other) is built.
- The tree is evaluated under one criterion (e.g., Maximum Likelihood).
- The initial tree is modified (e.g., by rearrangement of branches) giving new trees that are also evaluated under the criterion.
- If any of the new trees is better than the initial, the first is discarded and the best one is retained.
- The process is repeated until no better trees are found under the criterion.

Methods of phylogenetic analysis

EXPLORATION OF TREE SPACE BY HEURISTIC SEARCH

- ***Branch rearrangements (changes in topology):***



- **Nearest Neighbour Interchange (NNI)**
- **Subtree Pruning and Regrafting (SPR)**
- **Tree Bisection Reconnection (TBR)**

Methods of phylogenetic analysis

MAXIMUM LIKELIHOOD

Criterion: Tree that maximizes the probability of alignment (Likelihood= P(D|H)).

Alignment (data)

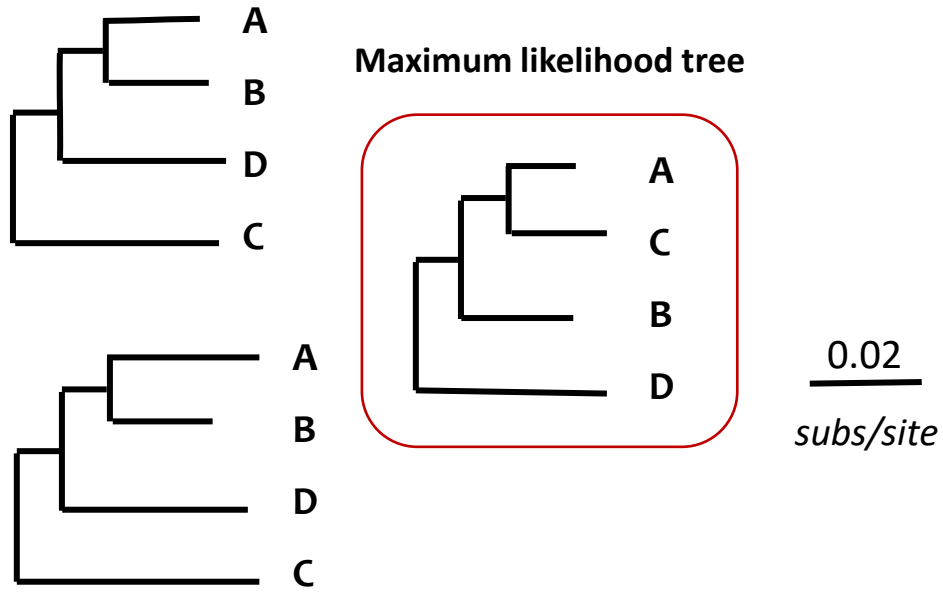
```
A  ACTTGACCCTTACGAT
B  .GC..G....G..T.C
C  .G......A...C...
D  .GC..GT...G.T..C
```

Model of molecular evolution

$$P(v_k) = \begin{pmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{pmatrix}$$

Probability of change for a branch lenght *k*

$$f = (\pi_A, \pi_C, \pi_G, \pi_T)$$



- ✓ Each HYPOTHESIS has an associated likelihood value that depends on the tree (topology and the branch lengths) and the parameters of the model of evolution.
- ✓ The hypothesis that meets the criterion of the maximum likelihood is selected.

L= L1 x L2 x L3 x … Ln (site)

- **Programs:** IQ-TREE, RAxML, PhyML, others…

Methods of phylogenetic analysis

MAXIMUM LIKELIHOOD

✓ **We look for the topology and the set of parameters (hypothesis) that maximize the probability of the data.**

L= p (data| hypothesis)

Alignment

```
A ACTTGACCCTTACGAT
B .GC..G....G..T.C
C .G.....A...C...
D .GC..GT...G.T..C
```

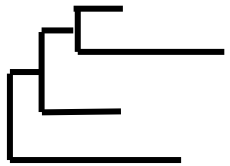
Model of evolution and its parameters

$$R = \begin{pmatrix} A & C & G & T \\ - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix}$$

$$\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$$

Phylogenetic tree

- **Topology**
- **Branch lengths**



Multidimensional optimization!

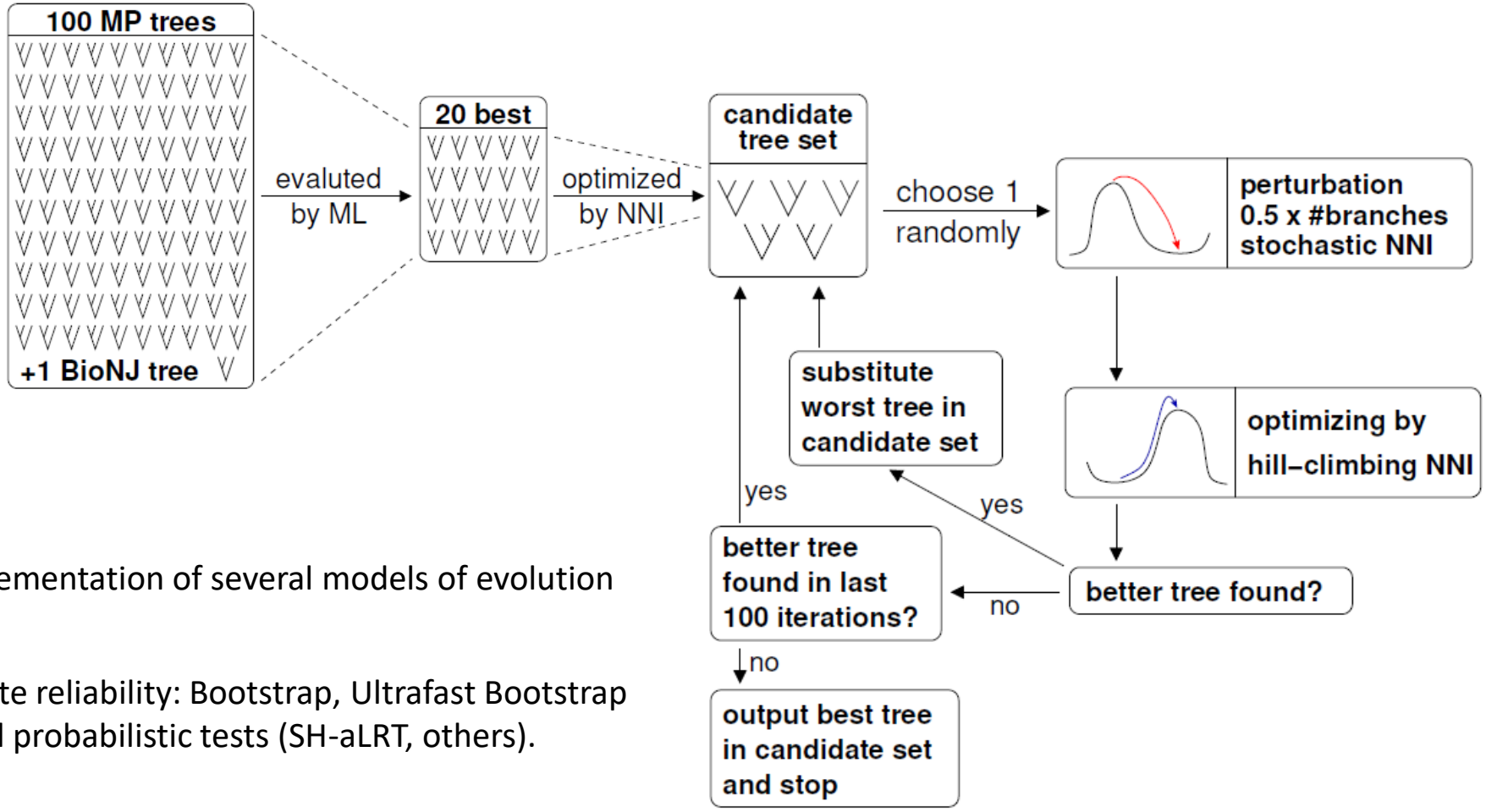
(cannot be solved in one go)

Methods of phylogenetic analysis

MAXIMUM LIKELIHOOD

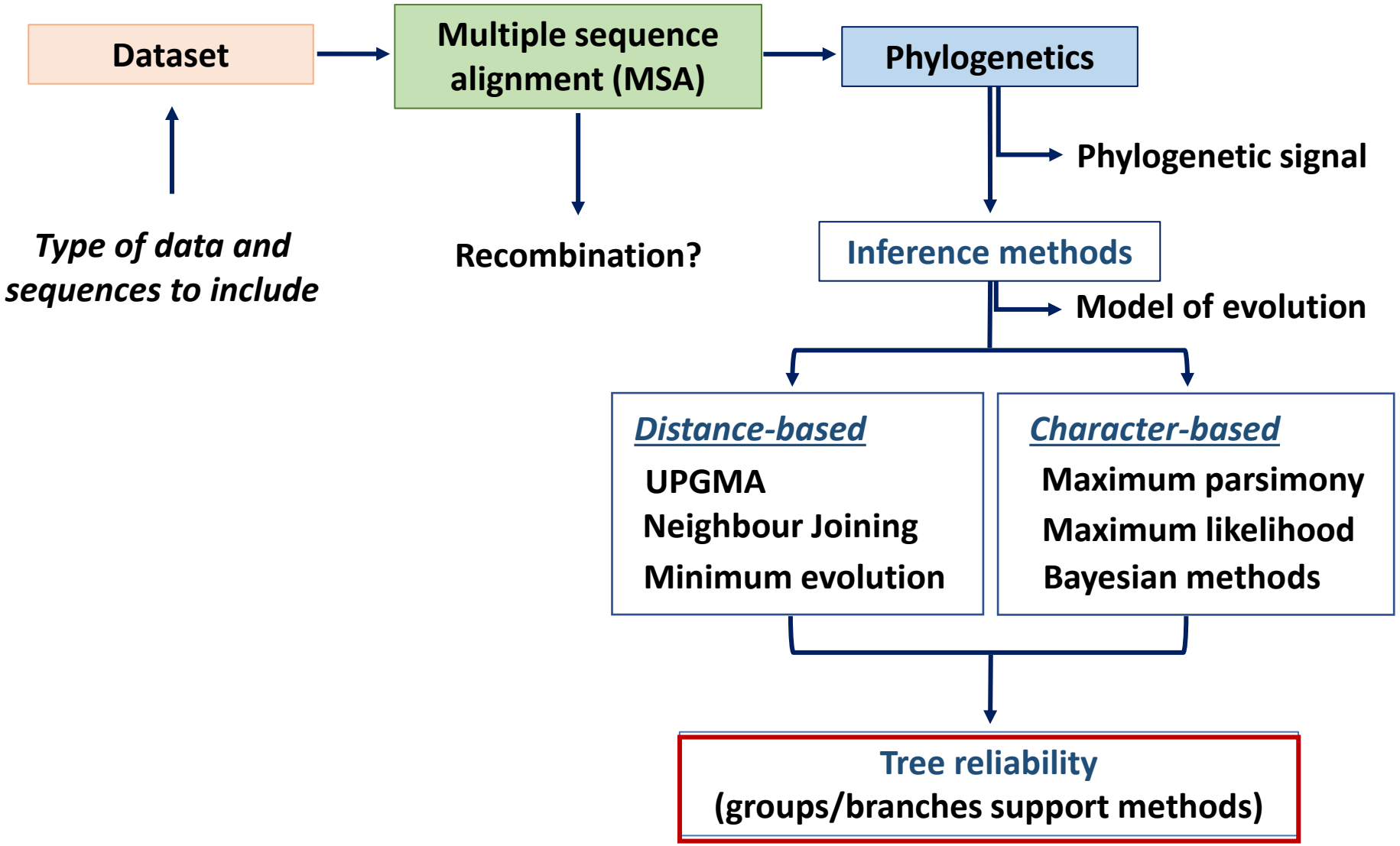
IQ-TREE

MP/BioNJ + randomization / permutation + fastNNI



- ✓ Very fast.
- ✓ Selection and implementation of several models of evolution (ModelFinder).
- ✓ Methods to evaluate reliability: Bootstrap, Ultrafast Bootstrap Approximation, and probabilistic tests (SH-aLRT, others).

Steps for phylogenetic analysis



Tree reliability

- The methods to evaluate the reliability allow us to know **how much evidence** exists in the dataset about the observed clusters, but they do NOT allow to establish if the groups exist in nature.
- Phylogenetic trees show reliable clusters in some areas and unreliable in others (higher and lower resolution spaces).
- **Methods:**

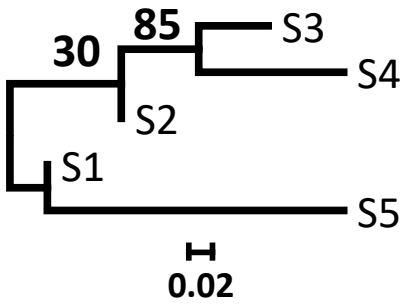
  Resampling methods: Bootstrap (Standard), Rapid Bootstrap, Ultrafast Bootstrap approximation (UFB)**.**

  Probabilistic approaches: Approximate Likelihood Ratio Test (aLRT), Shimodaira-Hasegawa-aLRT (SH-aLRT).

- The support values are usually indicated on the branches of the trees and there are recommendations on values considered "reliable" for each method (Bootstrap > 70-80%; UFB > 90-95%; SH-aLRT > 80%).

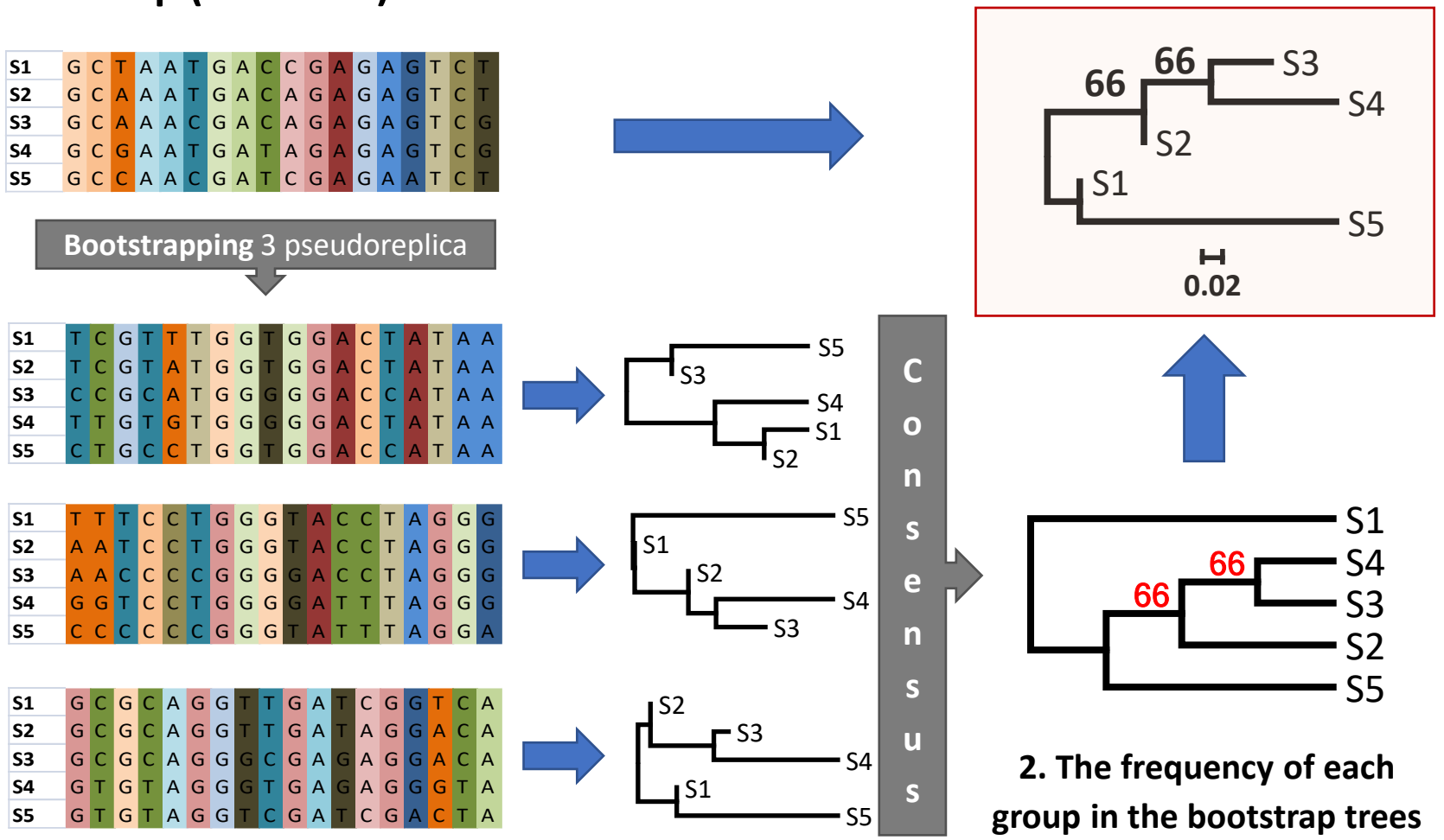
Example:

S1	G	C	T	A	A	T	G	A	C	C	G	A	G	A	G	T	C	T
S2	G	C	A	A	A	T	G	A	C	A	G	A	G	A	G	T	C	T
S3	G	C	A	A	A	C	G	A	C	A	G	A	G	A	G	T	C	G
S4	G	C	G	A	A	T	G	A	T	A	G	A	G	A	G	T	C	G
S5	G	C	C	A	A	C	G	A	T	C	G	A	G	A	A	T	C	T



Tree reliability

Bootstrap (Standard)



1. The alignment is sampled repeatedly (100-10000) and for each replica, a phylogenetic tree is obtained.

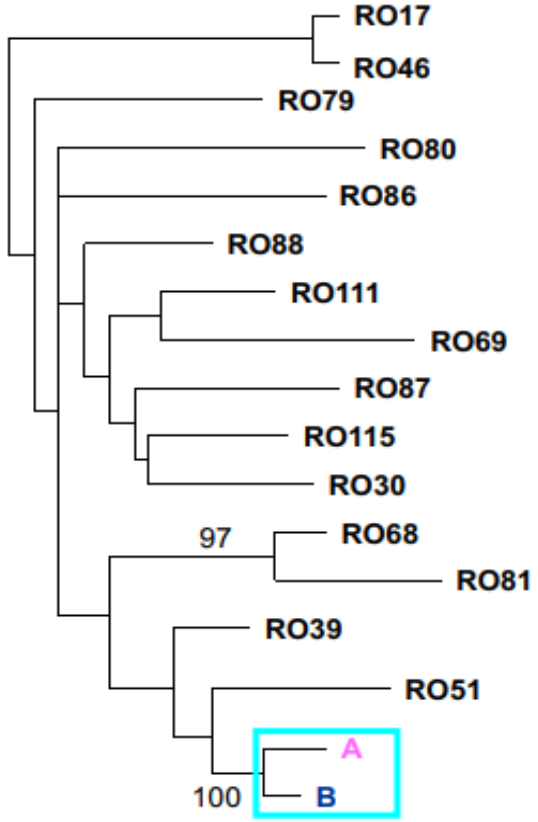
2. The frequency of each group in the bootstrap trees is estimated.

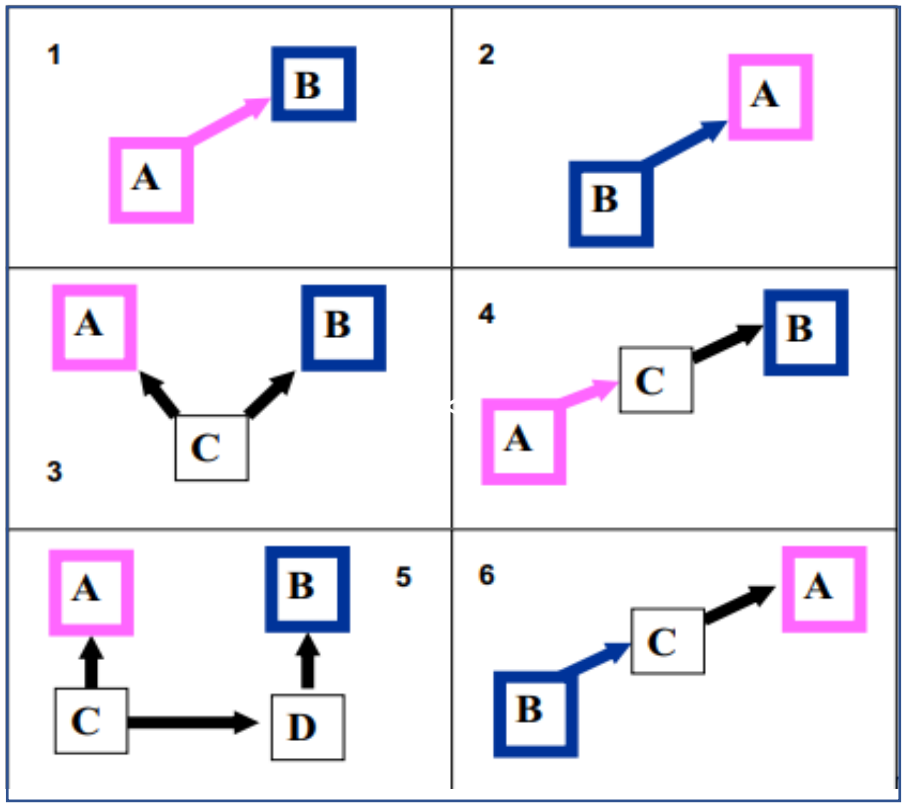
3. The bootstrap values are translated into the phylogenetic tree obtained from the actual data.

- ✓ Groups showing Bootstrap values > 70-80% are considered supported.
- ✓ It can be applied to different inference methods (MP, NJ, ML) and is implemented in several programs.
- ✓ It is very slow (demanding) and conservative.

Final considerations

A common ancestor is sought, however, there are several possible scenarios for the same topology:





Phylogenetic trees should always be interpreted in the context of all available information.

Avoid overinterpreting trees!

Other analyses using phylogenetic inference

- **Phylogenetic signal.**
- **Formal hypothesis testing on topologies.**
- **Selective pressure studies.**
- **Sequence ancestral reconstruction.**
- **Exploration of the temporal signal → Phylodynamic analyses**

References

- **The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny.** Marco Salemi, Anne-Mieke Vandamme (Eds). Cambridge University Press. (2009).
- **S. Kalyaanamoorthy, B.Q. Minh, T.K.F. Wong, A. von Haeseler, L.S. Jermiin** (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods*, 14:587-589. https://doi.org/10.1038/nmeth.4285
- **L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh** (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies.. *Mol. Biol. Evol.*, 32:268-274. https://doi.org/10.1093/molbev/msu300
- **B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. von Haeseler, R. Lanfear** (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, 37:1530-1534. https://doi.org/10.1093/molbev/msaa015.
- **D.T. Hoang, O. Chernomor, A. von Haeseler, B.Q. Minh, L.S. Vinh** (2018) UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.*, 35:518–522. https://doi.org/10.1093/molbev/msx281
- **Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.** New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010 May;59(3):307-21.