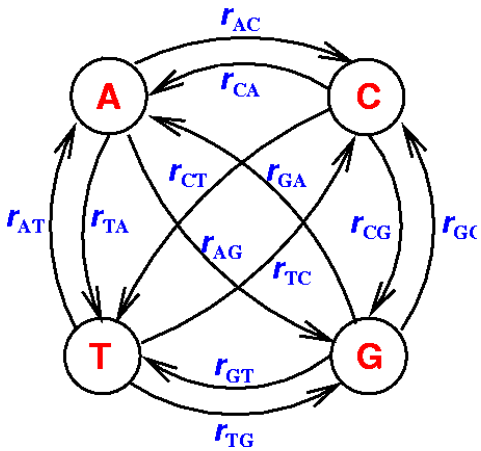
iMed. ULisboa
Research Institute for Medicines

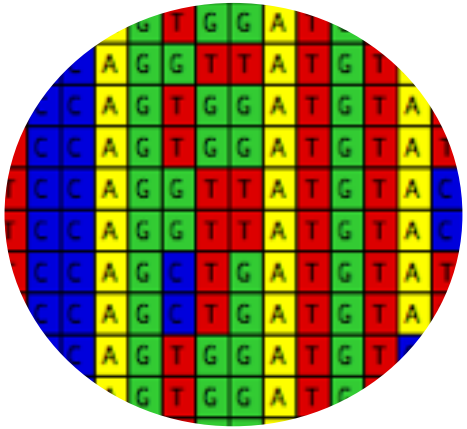
LISBOA
UNIVERSIDADE DE LISBOA

FACULDADE DE FARMÁCIA
Universidade de Lisboa

PATHOGEN MULTIOMICS AND BIOINFORMATICS

Module 5: Introduction to Phylogenetics and Public Health I





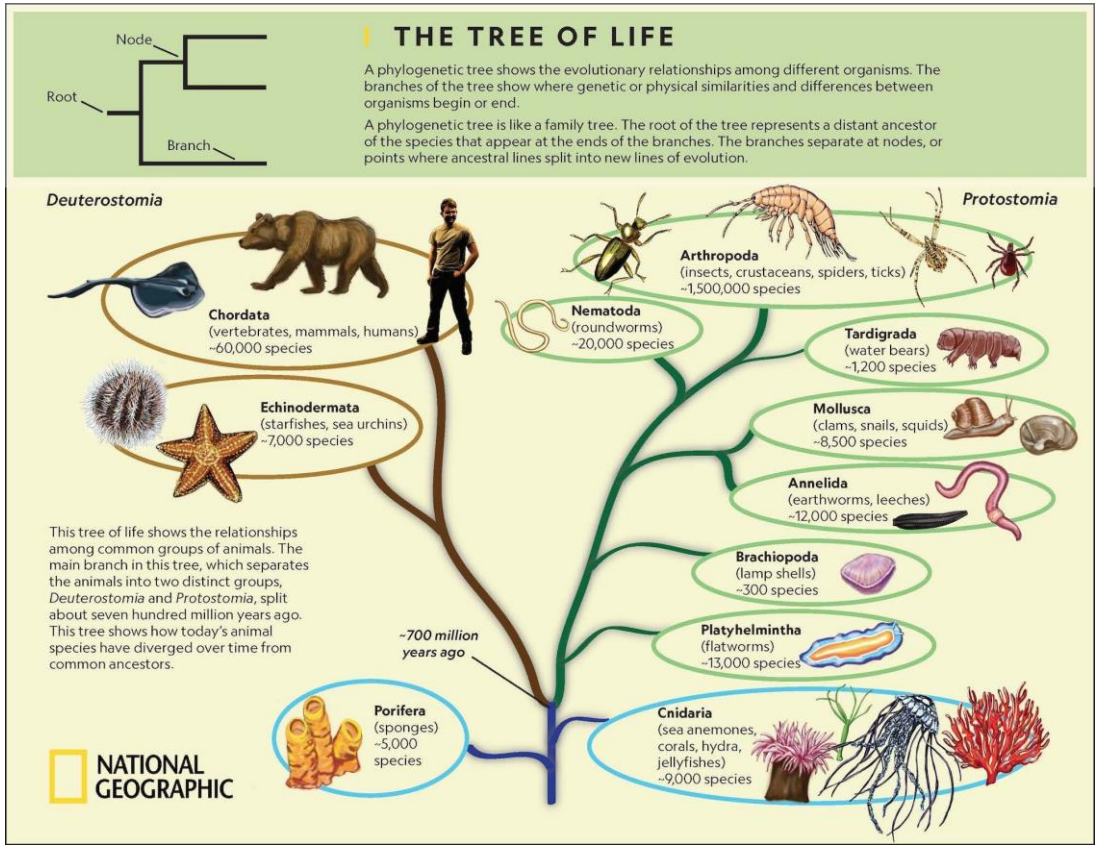
Ifeanyi Ezeonwumelu
&
João Perdigão

25.06.2021

iMed. ULisboa
Research Institute for Medicines

Phylogenetics

*Phylogenetics pertains to the **study of the evolutionary relationships***



Between what?

- *organisms, e.g., species or strains*
- *genes*
- *genomes*
- *Etc.*

Phylogenetics should refer to how closely the taxa are

and…

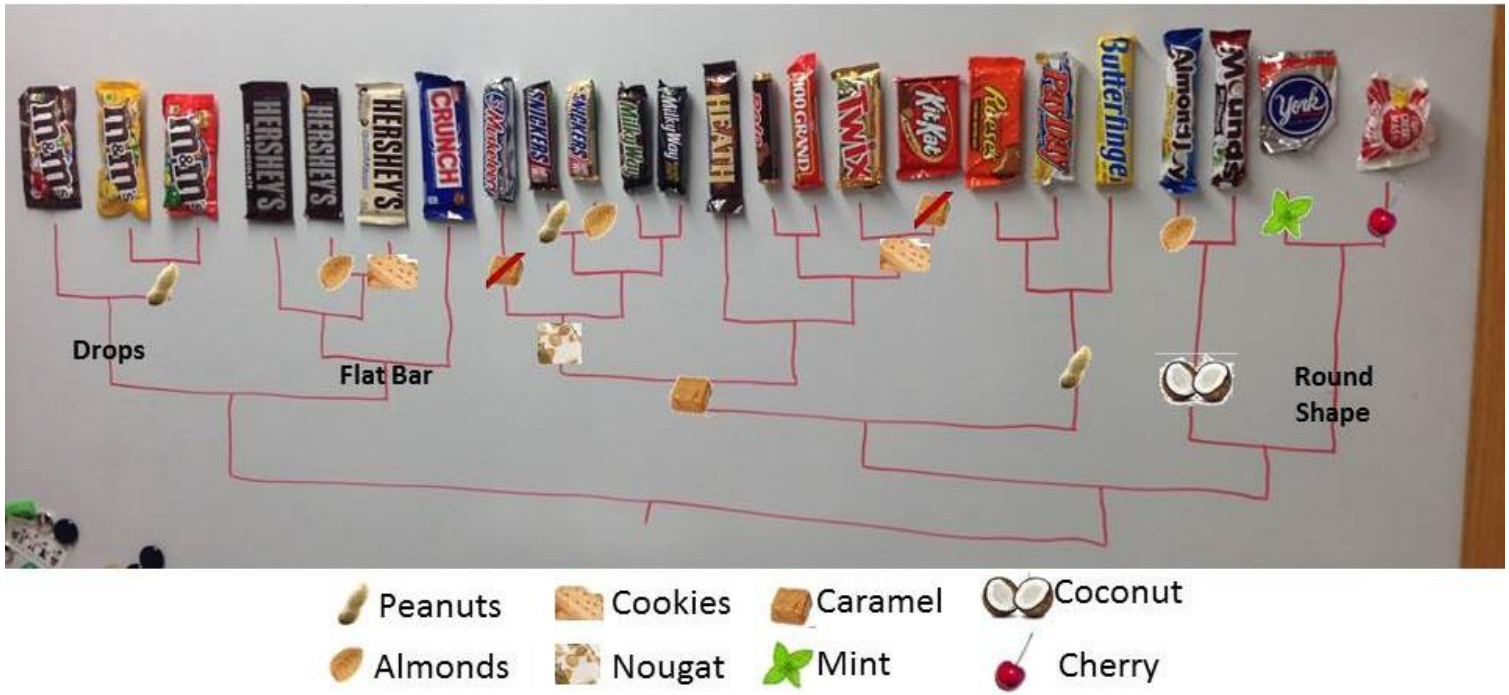
its evolutionary history

iMed. ULisboa
Research Institute for Medicines

Phylogenetics

What characters to use to construct a tree?

Early characters for phylogenetic reconstruction relied upon morphological and physiological characteristics



Example: ability to grow at different temperatures, drug resistance, sugar fermentation

Problems?

iMed. ULisboa
Research Institute for Medicines

Molecular Phylogenetics

Morphological/Physiological characters have two main problems associated:

Proneness to convergent evolution

Limited number of characters – poorly informative

However … molecular data (DNA or Protein sequence),

are less prone to convergent evolution

can provide an increasing number of characters

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

But… which characters are these?

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

Each homologous position between sequences comprise a character?

An alignment provides a way to:

- Identify homology – regions of common ancestry
- Contrast regions

iMed. ULisboa
Research Institute for Medicines

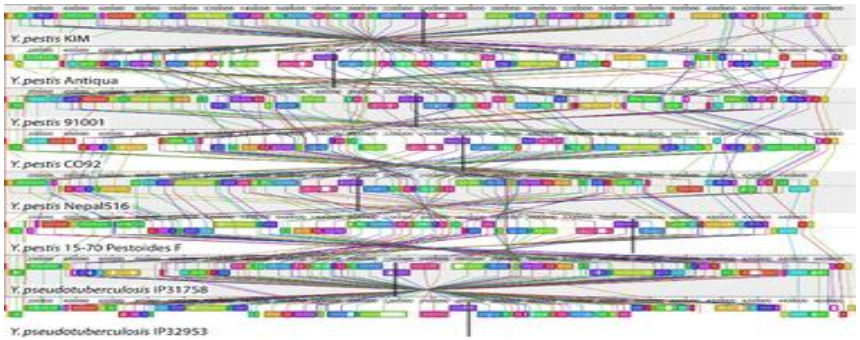
Molecular Phylogenetics

In summary, molecular phylogenetics require a Multiple Sequence Alignment

High amount of information – each column is an evolutionary marker

Possibility of going genome-wide – compare entire genomes

Problem: *alignment of entire genomes consumes large amounts of memory, even more upon tree bulding*



Research Questions & Trade-offs:

- *Sequence length (full length vs gene fragments)*
- *Genetic variation (conserved vs variable regions)*

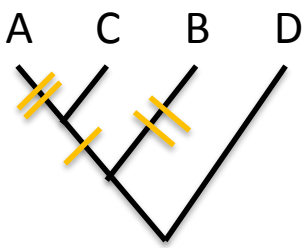
iMed. ULisboa
Research Institute for Medicines

Methods for Phylogenetic Reconstruction

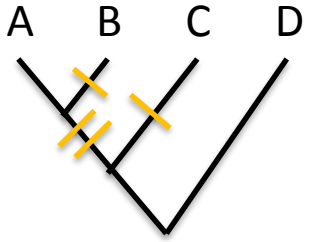
Maximum Parsimony

- *Simplest possible evolution scenario – the best tree is the shortest tree*
- *The rationale is to have the tree with the least homoplasy*

A: GGG
B: GTG
C: TGT
D: TTT



Length: 5

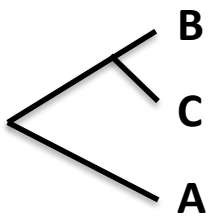


Length: 4

Distance Matrix Methods

- *Start by calculating pairwise distances from sequence data*
- *Tree is constructed from pairwise distances through clustering algorithms (e.g. Neighbour-Joining)*

	A	**B**	**C**
A	-	4	4
B	4	-	2
C	4	2	-



iMed. ULisboa
Research Institute for Medicines

Methods for Phylogenetic Reconstruction

Maximum Likelihood

More robust approach with parameter estimation using a probabilistic model:

- *Tree topology and branch lengths; nucleotide frequencies & substitution rates*
- *Measure how well the model fits the data*

Calculates the likelihood for every column first and for the entire alignment using different permutations, random starting values.

Likelihood (Model) = Probability (Data | Model)
Maximum Likelihood – Best set of parameter values yielding the highest possible likelihood

Software: PhyML [MEGA, Seaview, IQ-TREE, etc]

Bayesian Inference

- Based on the calculation of posterior probabilities given a set of prior parameter values
- Requires starting prior value
- Involves millions of iterations and mesasures the convergence to parameter values over the iterations

*Bayes Rule: P (Model|Data) = P (Data | M1) * P (M1)/P (Data)*

Software: MrBayes, BEAST

iMed. ULisboa
Research Institute for Medicines

Nucleotide Substitution Models

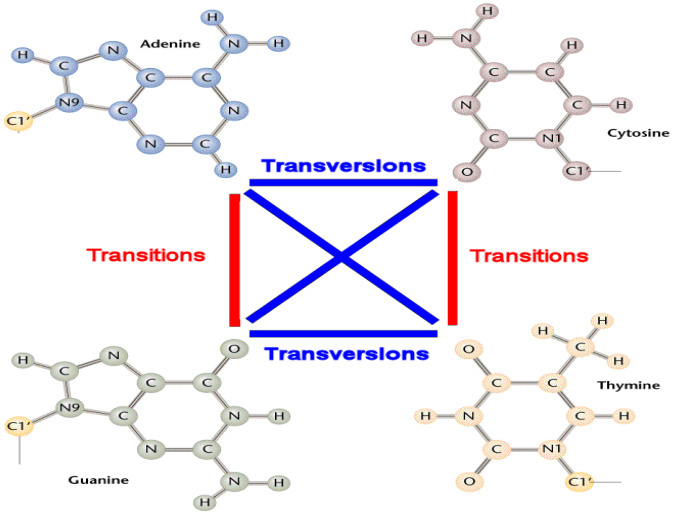
Models of Evolution

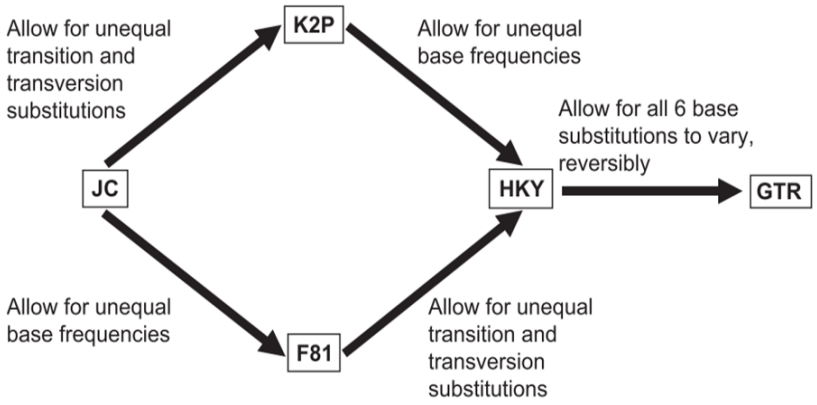
Aim to infer the number of real evolutionary events based on observed events!

Need to know how sequences are evolving – requires a model!!

Possibility of the existence of different susbstitution rates across site (gamma distribution).

The six possible substitution patterns for nucleotide data





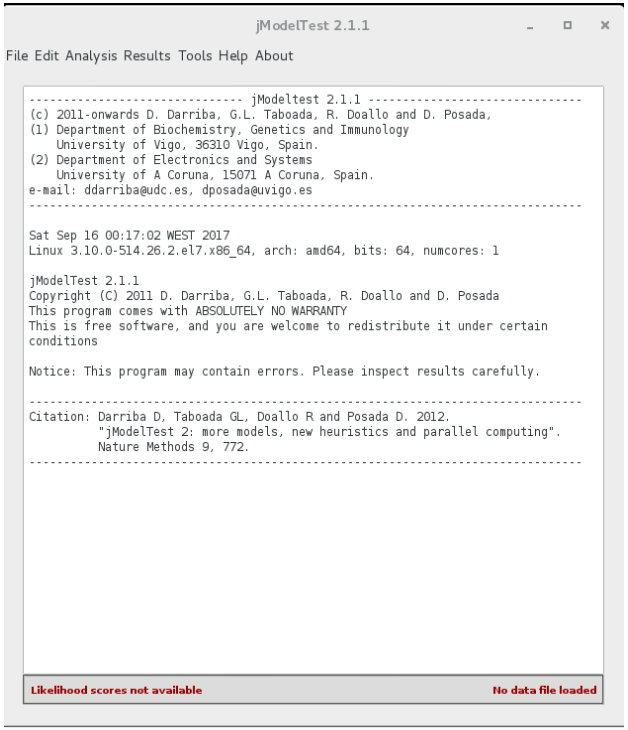
Transitions occur at higher frequency than Transversions

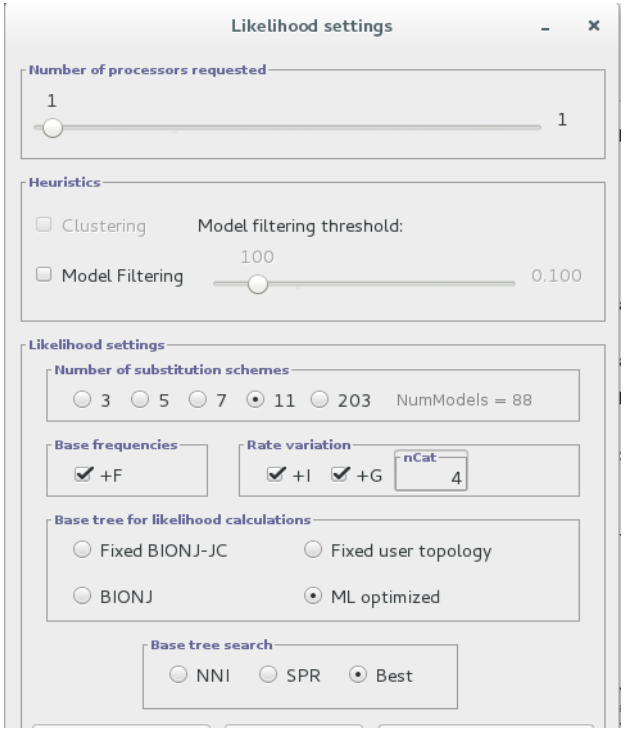
iMed. ULisboa
Research Institute for Medicines

How to choose the right parameters?

Answer: *test for a set of parameter permutations and choose on the combination that yields the combination that best fits your data*

jModelTest





Also possible: R, phangorn package (model.test function); ModelTest online server
Increasingly integrated in popular Tree building programs: RAxML, IQ-Tree

iMed. ULisboa
Research Institute for Medicines

Tree Searching

Tree space – the number of all possible trees for a given dataset

How many branches are present in a tree with 3 tips? And with 4 tips?

Number of branches in a tree with x tips = 2x-3

3 tips ->	3 branches
4 tips ->	5 branches
5 tips ->	7 branches
10 tips ->	17 branches
20 tips ->	37 branches
40 tips ->	77 branches
100 tips ->	197 branches

But, how many possible trees:

$$\prod_{i=2}^{n-1}(2i-3)$$

For every tree with x tips it is possible to construct 2x-3 derive trees by adding na extra tip

3 tips ->	1 tree
4 tips ->	3 trees
5 tips ->	15 trees
6 tips ->	105 trees
7 tips ->	945 trees
8 tips ->	10395 trees
9 tips ->	135135 trees
10 tips ->	2027025 trees
100 tips ->	1.7×10^{182} trees

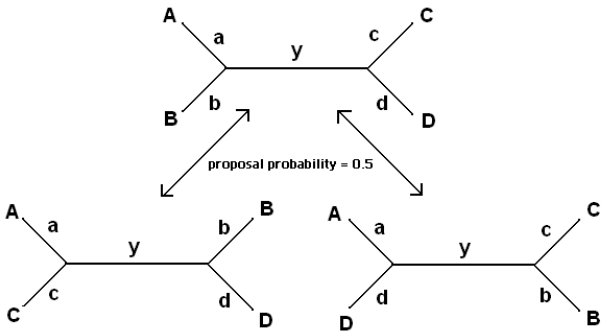
It is not possible to exhaustively screen and search all possible trees in present day datasets.

Answer: start with random tree(s) and progress by making alterations and removing less likely pathways

iMed. ULisboa
Research Institute for Medicines

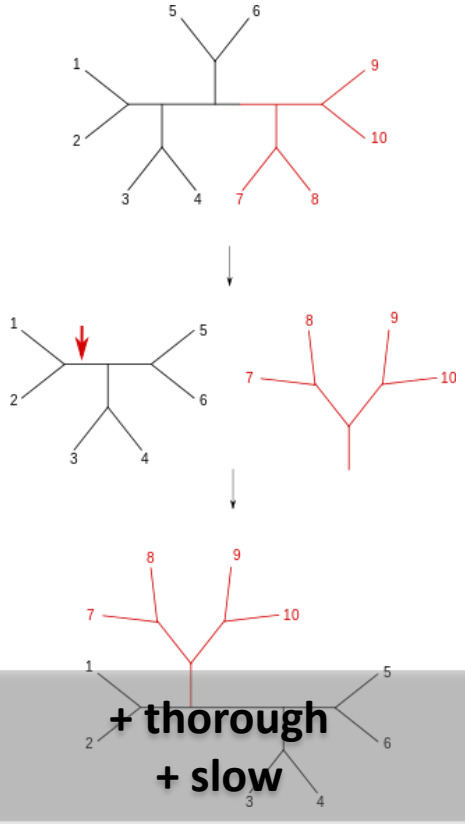
Tree Searching

Nearest Neighbour Interchange (NNI)



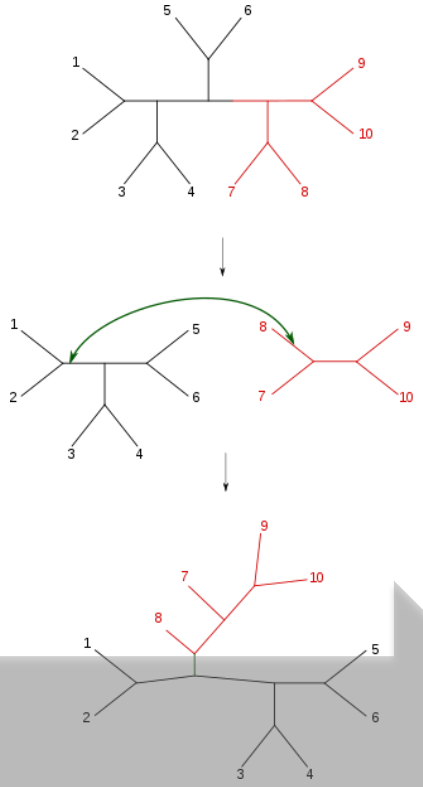
Swap neighbors at every internal branch

Subtree Prunning and Regrafting (SPR)



Cut a subtree at every possible point and regrafts at multiple points

Tree Bisection and Reconnection (TBR)



Divides tree in two parts, regrafts using every possible branch of the detached tree at all possible branches of the other tree

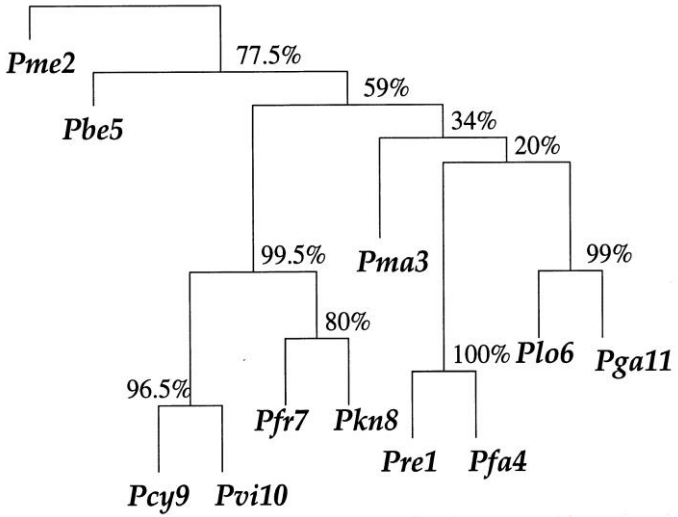
+ thorough
+ slow

iMed. ULisboa
Research Institute for Medicines

Tree Reliability

Bootstrap

- Widely used;
- Random sampling from the alignment (with replacement) until achieving the original length;
- Tree reconstruction n times – for each new alignment;
- Calculation for each branch the occurrence of that same clade in each tree;
- Expressed as a percentage/fraction (0-100%).
- Can be extremely time consuming



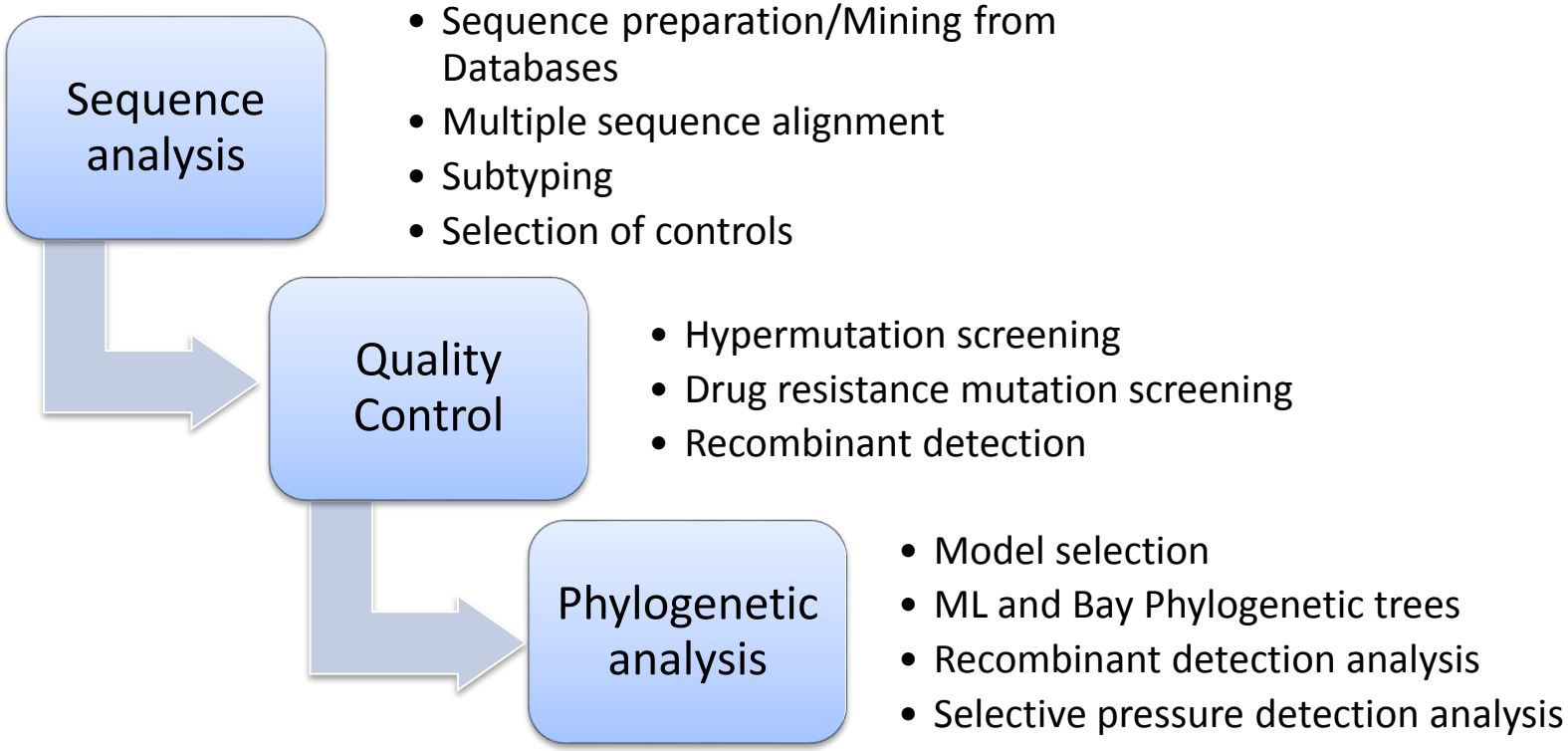
Alternatives:

- Jacknife – removes a position from the alignment (leave one out)

- aLRT – aproximate Likelihood Ratio Test – provides a *p*-value – likelihood gains between having a branch and not having a branch

iMed. ULisboa
Research Institute for Medicines

The Phylogenetic reconstruction approach



- Sequence preparation/Mining from Databases
- Multiple sequence alignment
- Subtyping
- Selection of controls

- Hypermutation screening
- Drug resistance mutation screening
- Recombinant detection

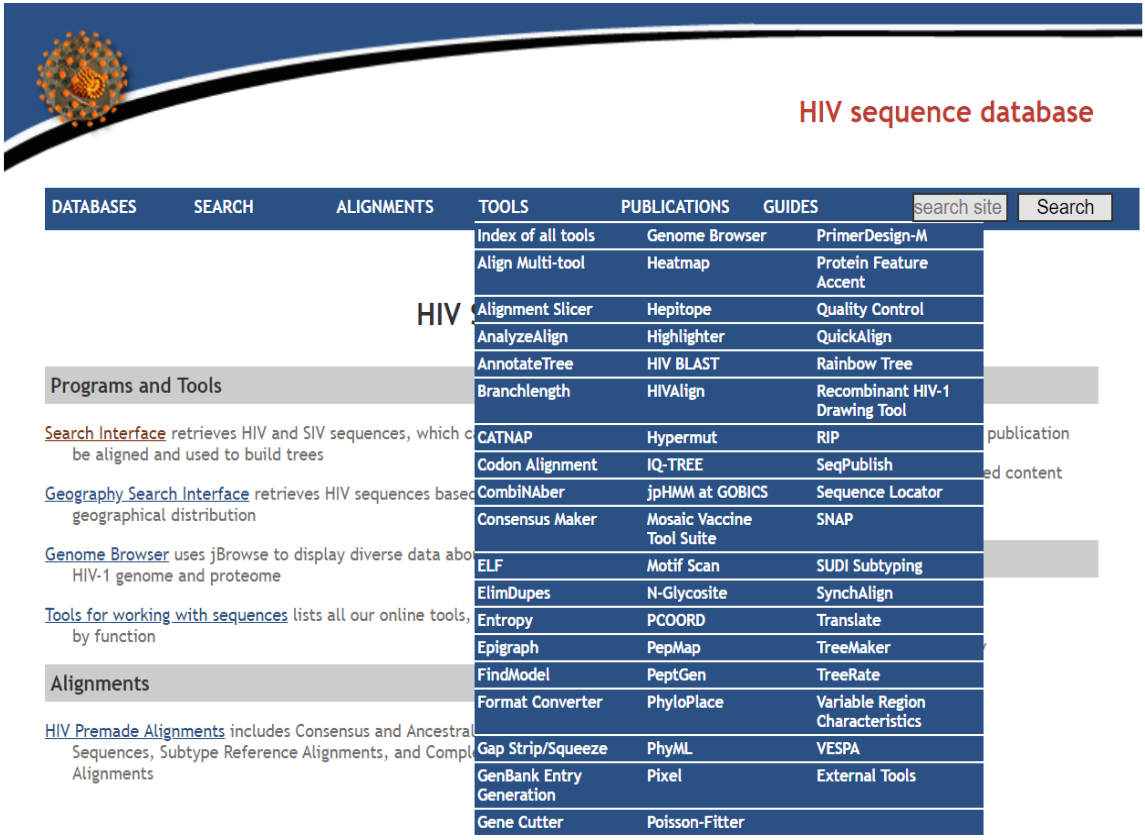
- Model selection
- ML and Bay Phylogenetic trees
- Recombinant detection analysis
- Selective pressure detection analysis
- Bayesian evolutionary analysis

iMed. ULisboa
Research Institute for Medicines

Quality Control

LANL (Los Alamos National Laboratory, New Mexico, USA) databases: **HIV, SIV** and **HCV**
https://www.hiv.lanl.gov/content/index

A One-Stop Platform for Phylogenetic reconstruction & QC !!!



EMBL (EuropeanMolecularBiology Laboratory) database, maintained at the EMBLEBI (European Bioinformatics Institute, Hinxton, England, UK).

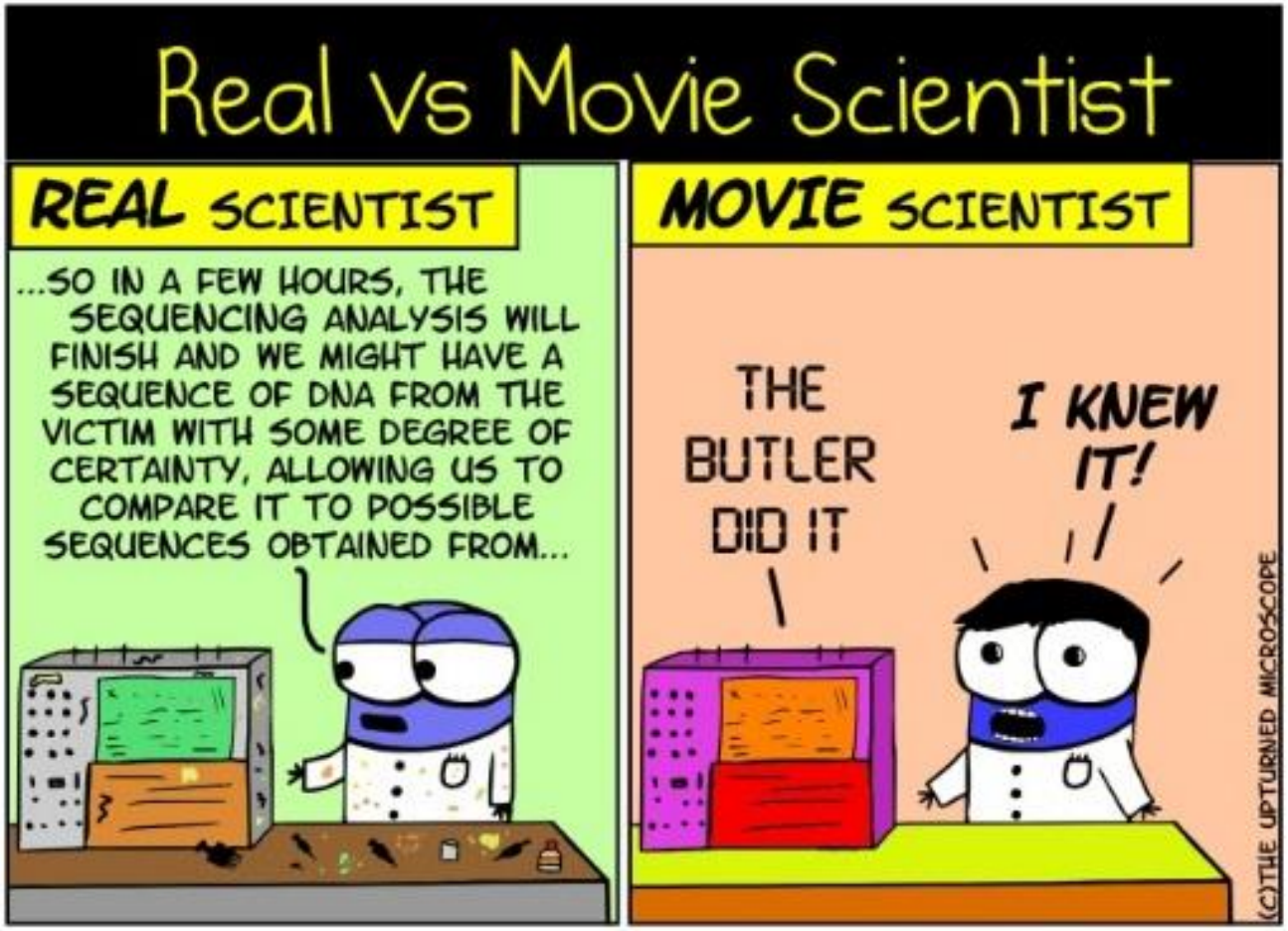
GenBank, maintained at the NCBI (National Center for Biotechnology Information, Bethesda, Maryland, USA).
https://www.ncbi.nlm.nih.gov/genbank/

DDBJ (DNA Data Bank of Japan), maintained at the NIG/CIB (National Institute of Genetics, Center for Information Biology, Mishima, Japan).

iMed. ULisboa
Research Institute for Medicines

Applications

DNA sequence analysis



ses;

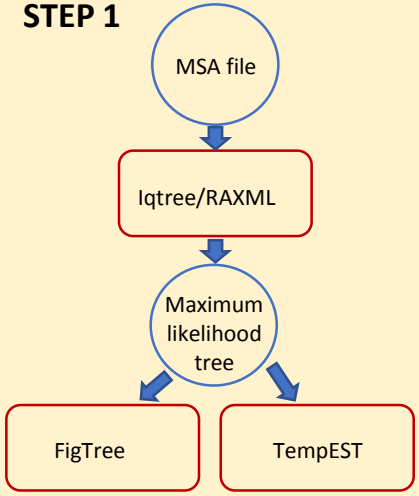
iMed. ULisboa
Research Institute for Medicines

Phylogeography Practice Session: OUTLINE

TempEST v1.5.3; FigTree v1.4.3; IQ-Tree v2; MEGA X

Session I

STEP 1



IqTree:

- ML tree

FigTree:

- Tree visualization

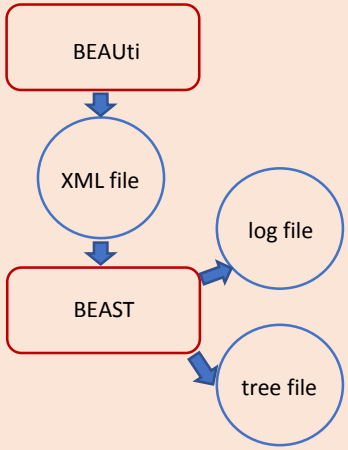
ClusterPicker:

- Transmission Clusters

TempEST:

- Remove outliers
- Determine Molecular clock likelihood

STEP 2



BEAUTi:

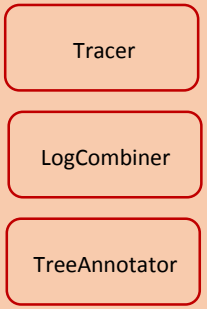
- Define data partitions, select models and define priors
- Decision on best Model/prior selection: specify MLE

BEAST:

- Bayesian Inference (MCMC)

Session II

STEP 3



Tracer:

- Check for convergence of MCMC, visualize log files, summarize data.

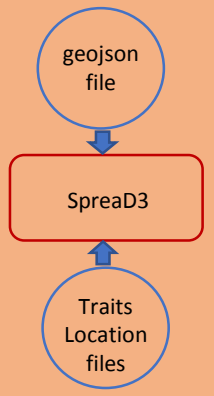
LogCombiner:

- Combine output files

TreeAnnotator:

- Summarize tree files to Maximum clade credibility trees

STEP 4



spreaD3:

- Spatiotemporal visualization
- BF for location transitions

SpreaD3 v0.9.6

BEAST v1.10 package

iMed. ULisboa
Research Institute for Medicines

Phylogeography Practice Session: Case study

SCIENTIFIC REPORTS
nature research

Check for updates

Epidemic history and baseline resistance to NS5A-specific direct acting drugs of hepatitis C virus in Spain

Claudia Palladino[1,4], Ifeanyi Jude Ezeonwumelu[1,4], Irene Mate-Cano[2], Pedro Borrego[1], Paula Martínez-Román[2], Sonia Arca-Lafuente[2], Salvador Resino[2], Nuno Taveira[1,3] & Verónica Briz[2]