Bridging the gaps in Bioinformatics

Phylogenetic analysis

Raphael N. Sieber, March 2025

Objectives

- Understand basic concepts for phylogenetic reconstruction
- Get an overview of available methods for phylogenetic inference, and know how to choose the best method
- Get an overview over available tools and software packages for the different methods
- Describe methods to evaluate the quality and trustworthiness of a phylogeny
- Use some tools to create phylogenetic trees
- Visualize these trees using different software

Outline

Lecture (ca. 1 h)

- Introduction to phylogenies
- Phylogentic trees
- Phylogenetic methods:
    - Substitution models
    - Maximum parsimony
    - Distance based methods (UPGMA, Minimum evolution, Neighbour Joining)
    - Probabilistic methods (Maximum likelihood, Bayesian)
- Advanced phylogenetic analyses
- Assessing the reliability of a phylogeny
- Application of phylogenetic analysis
- Phylogenetic analysis and visualization software

Practical (ca. 2h)

Bridging the gaps in Bioinformatics

Phylogenetic analysis

Raphael N. Sieber, March 2025

Introduction to phylogenies

Pylogenies aim to describe the evolutionary relationship between different taxa

Phylogenies can be based on

- phenotypic traits (binary, multi-level)
- genotypic markers (restriction patterns, SNPs, nucleotide/amino acid sequences)

There are different methods for obtaining a pylogeny

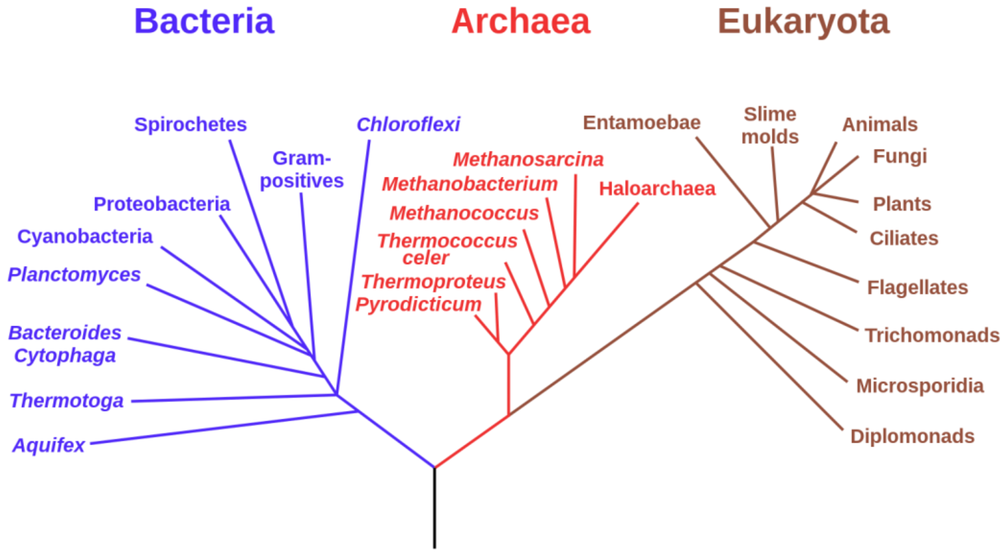
Important: model of substitution between one state and the other

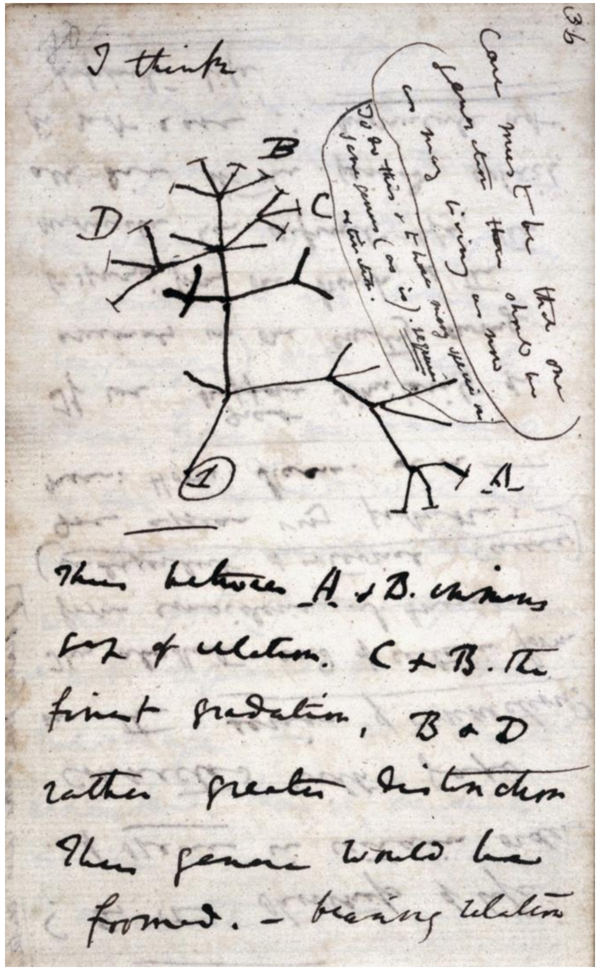
Phylogenetic trees

Introduction to phylogenies

Phylogenetic trees aim to describe the relationship between taxa

They are always based on the principle of common ancestry proposed by Charles Darwin in 1837



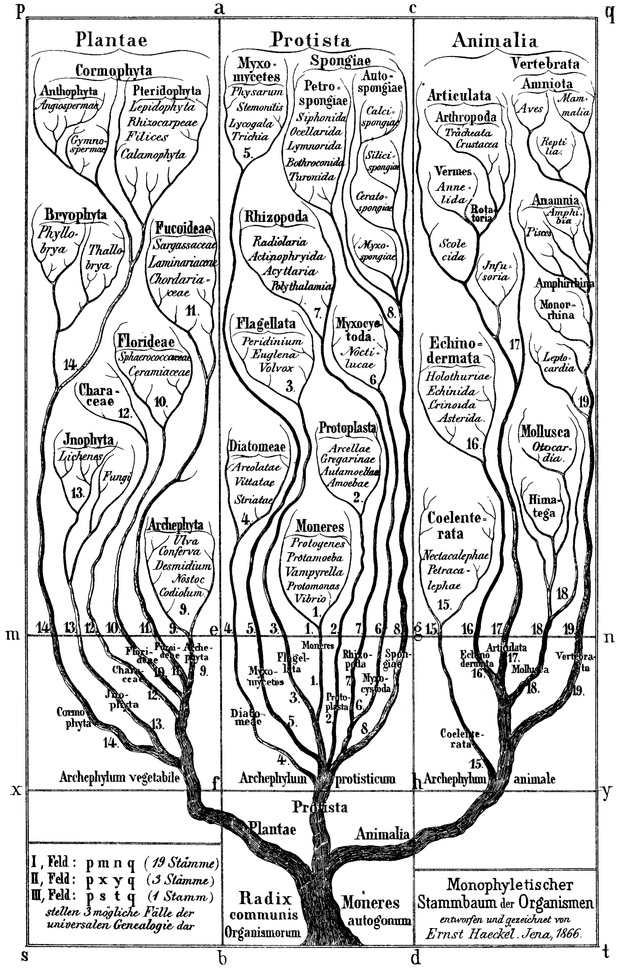


I think

Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction.

Thus between A & B immens gap of relation. C & B the finest gradation, B & D rather greater distinction. Thus genera would be formed. — bearing relation

Charles Darwin's first diagram of an evolutionary tree (Transmutation of Species, 1837) (wikipedia.org)

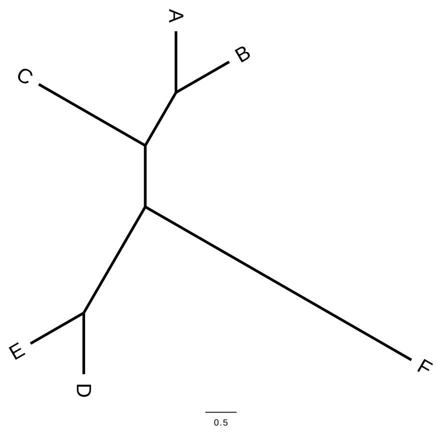


Phylogenetic tree suggested by Haeckel (Generelle Morphologie der Organismen, 1866) (wikipedia.org)

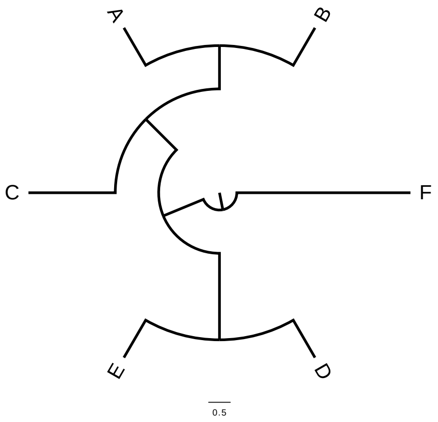
Phylogenetic trees

Different visualizations of the same tree:

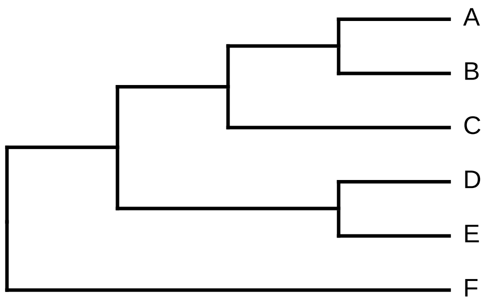
unrooted / polar



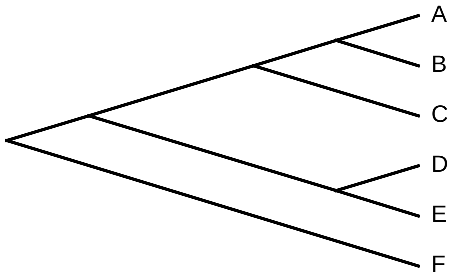
circular / radial



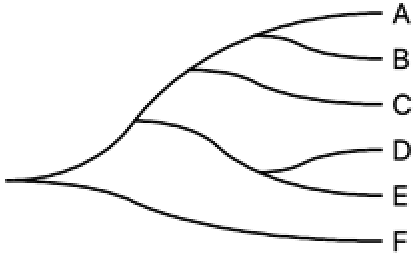
rectangular



(straight)



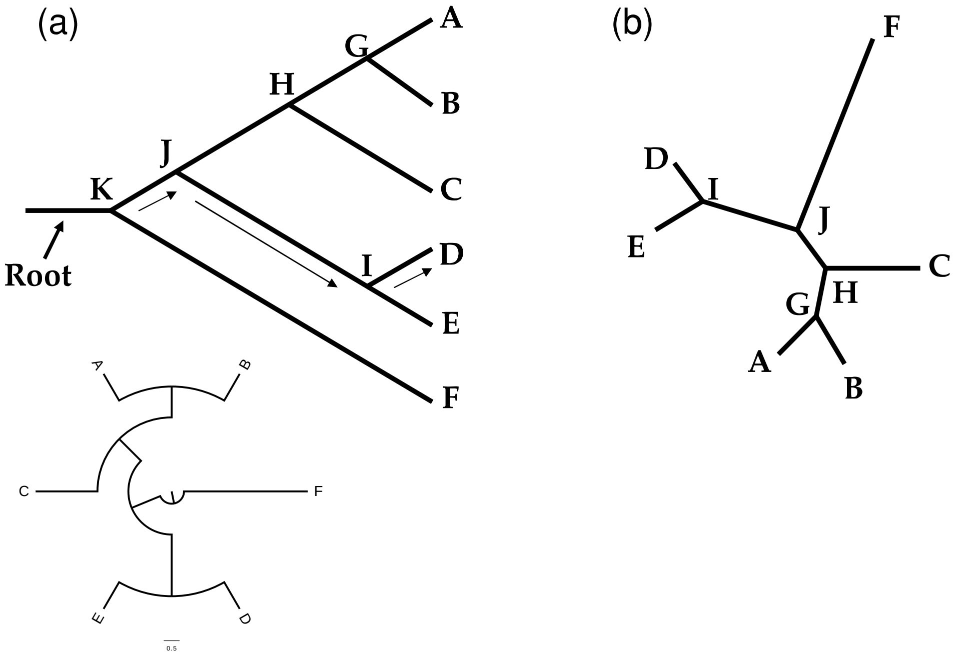
(curved)

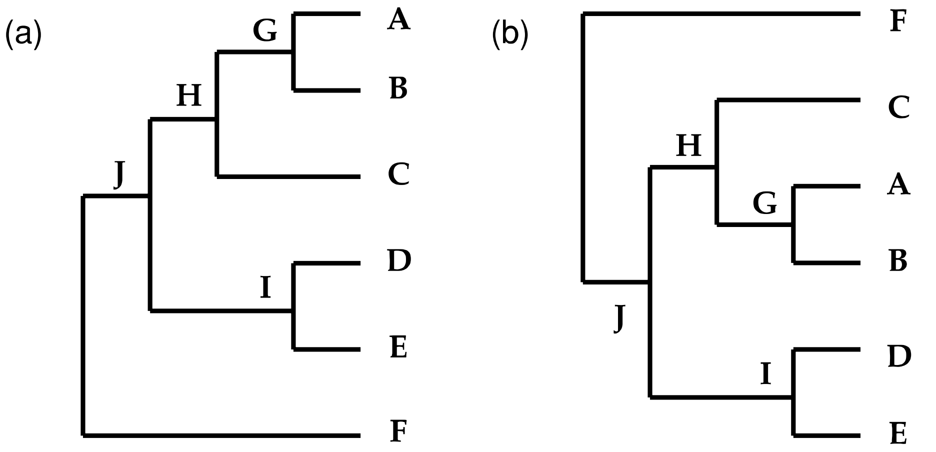


Phylogenetic trees

How to read a phylogenetic tree

- The (horizontal) distance bewteen taxa counts (center to distal in circ.)
- Rooted vs. unrooted
- Rotations are arbitrary and can be confusing

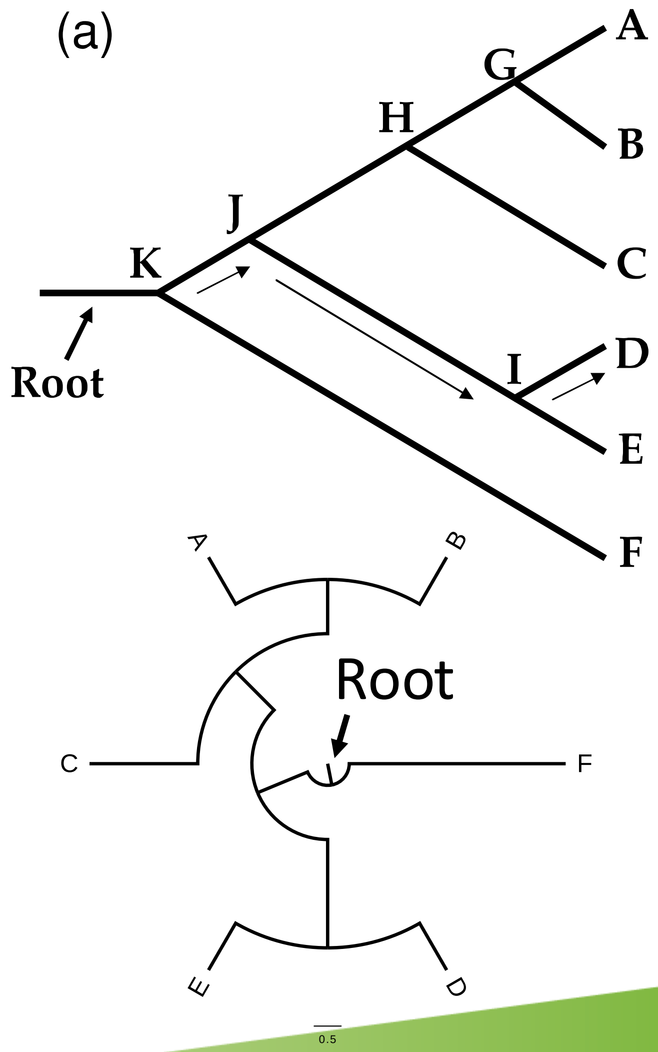




Phylogenetic trees

Methods for tree rooting

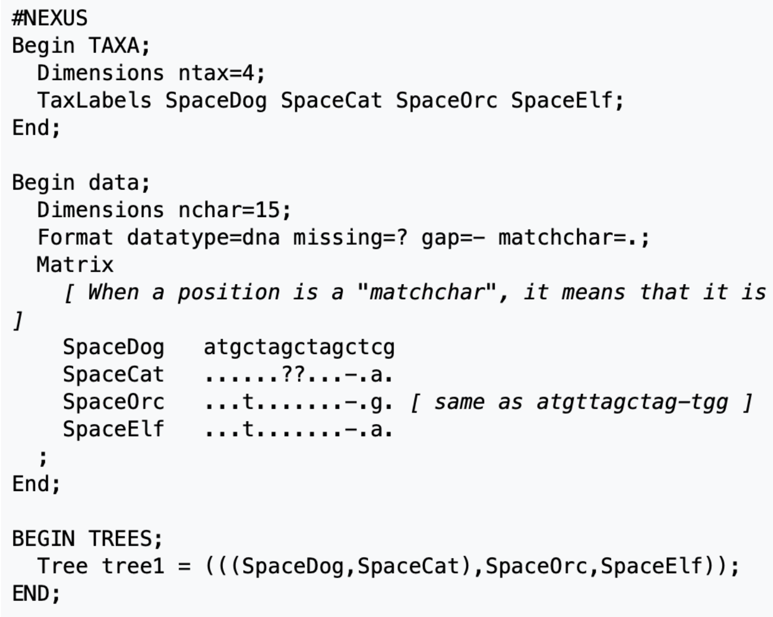
- Outgroup rooting
  - The most natural rooting, but requires knowledge of a biologically meaningful outgroup
- Midpoint rooting
  - places the root halfway between the two tips with the longest distances
- Molecular Clock Rooting
  - Assumes a constant rate of evolution



Phylogenetic trees

Tree storing formats

- Newick
  - The simplest and most common format for storing a tree structure
  - ((<taxa>:<branch_length>),(<taxa>:<branch_length>));
  - e.g.: ((((A:1,B:1),C:2),(D:1,E:1):2),F:4);
  - .nwk
  - Does only allow for limited information
- Nexus
  - Allows for more information,
  - contains a newick tree
  - E.g. FigTree uses nexus format for storing trees with properties
- (Ne)Xml
  - Most flexible and more robust than nexus



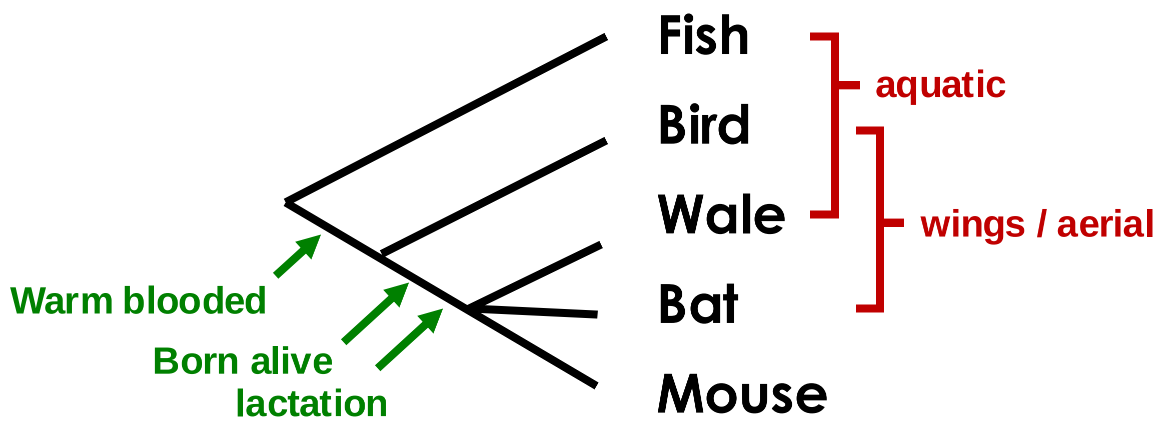
```
#NEXUS
Begin TAXA;
  Dimensions ntax=4;
  TaxLabels SpaceDog SpaceCat SpaceOrc SpaceElf;
End;

Begin data;
  Dimensions nchar=15;
  Format datatype=dna missing=? gap=- matchchar=.;
  Matrix
    [ When a position is a "matchchar", it means that it is
]
    SpaceDog   atgctagctagctcg
    SpaceCat   ......??...-.a.
    SpaceOrc   ...t.......-.g. [ same as atgttagctag-tgg ]
    SpaceElf   ...t.......-.a.
  ;
End;

BEGIN TREES;
  Tree tree1 = (((SpaceDog,SpaceCat),SpaceOrc,SpaceElf));
END;
```

Reconstructing phylogenies

Concept of constructing a phylogenetic tree

Mapping traits

	habitat	birth	lactation	warm-blooded	wings
Mouse	terrestrial	alive	yes	yes	no
Bat	aerial	alive	yes	yes	yes
Wale	aquatic	alive	yes	yes	no
Fish	aquatic	egg	no	no	no
Bird	aerial	egg	no	yes	yes



derived trait /
(syn-) apomorphy

Parallel evolution /
homoplasy

Models of substitution

How one character (trait, nucleotide or amino acid) is replaced by the other, is crucial.

E.g. different nucleotides may have different probabilities to mutate. Transitions may be more or less likely than transversions, and also changes between and within pyrimidines and purines may have different probabilites.

This is defined in different models of nucleotide substitution.

Further models exist for protein sequences, and the logic can also be used for other characters like phenotypic traits.

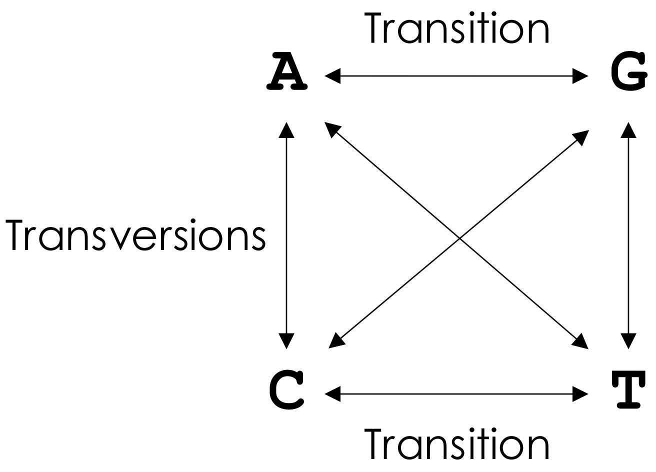
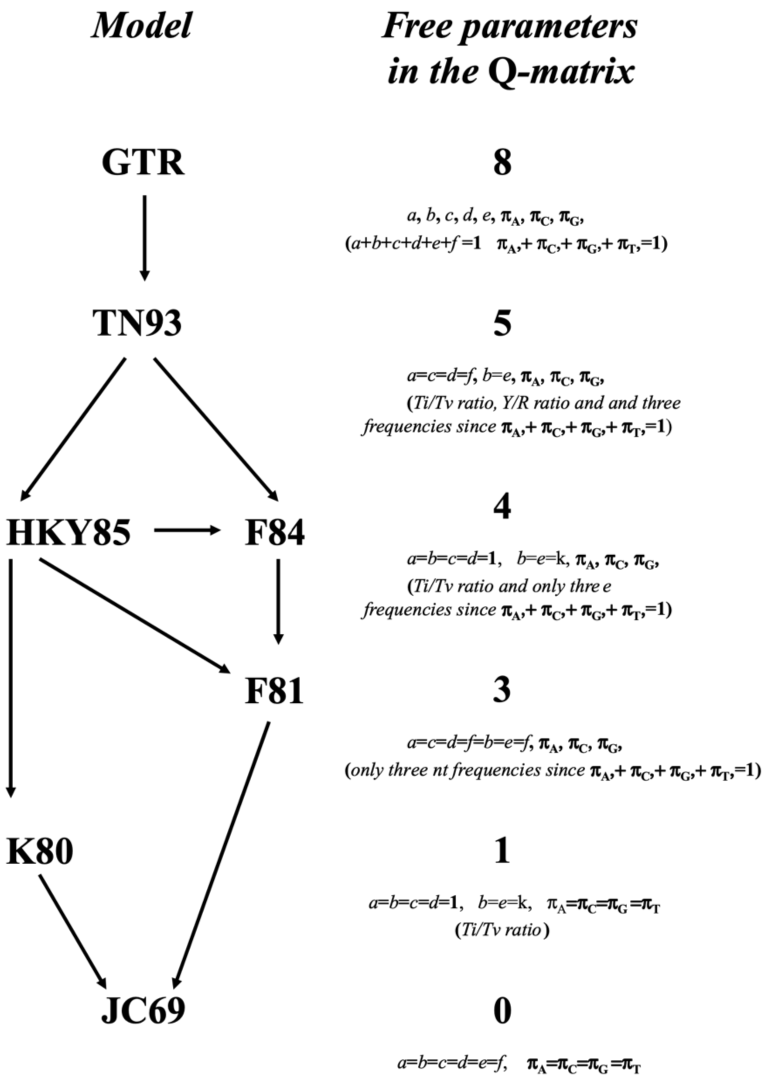


Figure from Lemeyetal et al. (2009) "Phylogenetic handbook"

Models of nucleotide substitution

Different nucleotide substitution models assume different rates of substitutions and different base frequencies:

- JC69 (Jukes-Cantor): all rates and frequencies are fixed and equal (most simple model)
- HKY85 (Hasegawa-Kishino-Yano, 85): frequencies estimated, transversions equally likely and transitions equally likely.
- GTR (general time reversible): all rates and frequencies are estimated (most free model)



Model

Free parameters in the Q-matrix

GTR **8**

$a, b, c, d, e, \pi_A, \pi_C, \pi_G,$
$(a+b+c+d+e+f=1 \quad \pi_A,+\pi_C,+\pi_G,+\pi_T,=1)$

TN93 **5**

$a=c=d=f, b=e, \pi_A, \pi_C, \pi_G,$
(*Ti/Tv ratio, Y/R ratio and and three frequencies since* $\pi_A,+\pi_C,+\pi_G,+\pi_T,=1$)

HKY85 **F84** **4**

$a=b=c=d=1, \quad b=e=k, \pi_A, \pi_C, \pi_G,$
(*Ti/Tv ratio and only thre e frequencies since* $\pi_A,+\pi_C,+\pi_G,+\pi_T,=1$)

F81 **3**

$a=c=d=f=b=e=f, \pi_A, \pi_C, \pi_G,$
(*only three nt frequencies since* $\pi_A,+\pi_C,+\pi_G,+\pi_T,=1$)

K80 **1**

$a=b=c=d=1, \quad b=e=k, \quad \pi_A=\pi_C=\pi_G=\pi_T$
(*Ti/Tv ratio*)

JC69 **0**

$a=b=c=d=e=f, \quad \pi_A=\pi_C=\pi_G=\pi_T$

Figure from Lemeyetal et al. (2009) “Phylogenetic handbook”

Phylogenetic methods overview

There are different ways of obtaining a phylogeny

- Maximum parsimony
  - Tries to find the tree with the fewest evolutionary changes (least homoplasy)

- Distance based (UPGMA / ME / NJ etc.)
  - Creates a tree based on a matrix of differences between taxa

- Probabilistic methods (Maximum likelihood and Bayesian)
  - Uses a starting tree and optimizes topology and branch lengths

Maximum parsimony

- Tries to find the tree with the fewest changes (shortest tree)
- For n taxa, an unrooted binary tree contains:
    - n terminal nodes (leaves)
    - n – 2 internal nodes
    - 2n – 3 branches
- Tree length of tree τ with N sites: $L(\tau) = \sum_{j=1}^{N} l_j$ where l_j is the length for a single site defined as:

$$l_j = \sum_{k=1}^{2N-3} c_{a(k),b(k)}$$

where $c_{a(k),b(k)}$ is the cost of change from state *a* to state *b* along branch *k*.

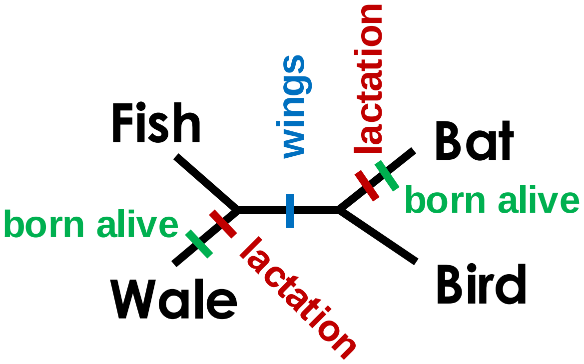
Maximum parsimony

- The cost c can differ depending on the character in question
- In DNA/protein sequence the different changes can be weighted differently (defined in the substitution model)
- Characters can be ordered (e.g. habitat: aquatic, terrestrial, aerial), so that a change from one state to the other has different costs depending on the states (e.g. higher cost for aquatic to aerial than for terrestrial to aerial).
- Computationally intense:
  - ≤~10 taxa: exhaustive search (all trees are calculated)
  - 12 to 25 taxa: branch-and-bound method (Hendy & Penny, 1982)
  - >25 taxa: Approximate methods

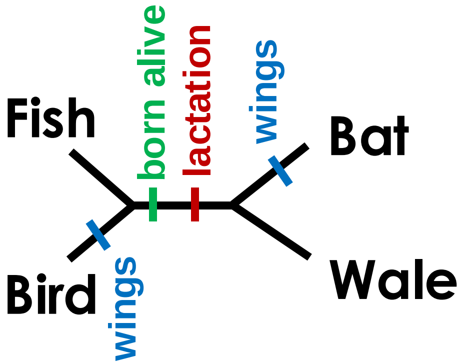
Maximum parsimony

- Example

	habitat	birth	lactation	warm-blooded	wings
Mouse	terrestrial	alive	yes	yes	no
Bat	aerial	alive	yes	yes	yes
Wale	aquatic	alive	yes	yes	no
Fish	aquatic	egg	no	no	no
Bird	aerial	egg	no	yes	yes



$L_{(\tau)}=5$



$L_{(\tau)}=4$

Maximum parsimony: Summary

- Tries to minimize evolutionary change
- Widely used in the 1970-1990's
- Intuitive and logical, especially for discrete characters
- Computationally intense for >10 taxa
- Some critical problems like long branch attraction

Distance based methods

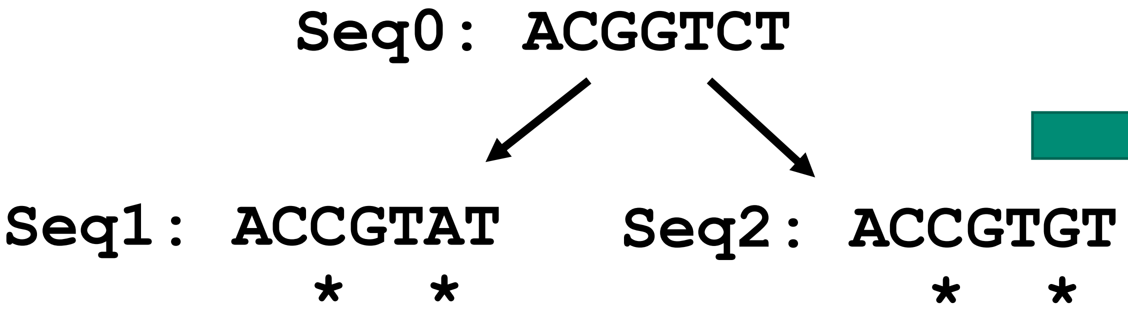
A pairwise (p-) distance can be calculated between taxa:

- from morphological characters:

	habitat	lactation
Mouse	terrestrial	yes
Fish	aquatic	no
Wale	aquatic	yes

	Mouse	Fish	Wale
Mouse	0	2	1
Fish	2	0	1
Wale	1	1	0

- from genetic sequences:



```
         Seq0: ACGGTCT
Seq1: ACCGTAT   Seq2: ACCGTGT
        *  *            *  *
```

	Seq0	Seq1	Seq2
Seq0	0	2	2
Seq1	2	0	1
Seq2	2	1	0

Distance based methods: Tree reconstruction

From a (p-) distance matrix, a tree can be reconstructed by different methods:

- unweighted-pair group method with arithmetic means (**UPGMA**) / weighted-pair group method with arithmetic means (WPGMA)
- Minimum evolution (**ME**)
- Neighbour Joining (**NJ**)

- **Note:** The p-distance is an underestimation of the true genetic distance because some of the nucleotide positions may have experienced multiple substitution events.

UPGMA / WPGMA

UPGMA = unweighted-pair group method with arithmetic means

WPGMA = weighted-pair group method with arithmetic means

Both:

- Cluster the smallest distances, group these, and cluster with the next smallest distances
- Result in rooted trees.
- When the data is ultrametric, UPGMA = WPGMA
- Very fast and deterministic method
- Limitation: Very sensitive to unequal evolutionary rates

Minimum Evolution (ME)

- Reconstructs additive distances ($dAB + dCD \leq \max(dAC + dBD, dAD + dBC)$)
- Searches for the shortest tree, meaning the tree with the lowest sum of the lengths of the branches:

$$S = \sum_{i=1}^{2n-3} v_i$$

where n = number of taxa, v_i = i^{th} branch

- Reminds of maximum parsimony, but using distances instead of traits directly

Neihgbour-joining (NJ)

- Reconstructs additive distances ($dAB + dCD \leq \max(dAC + dBD, dAD + dBC)$)
- A heuristic method
- conceptually related to clustering, but without assuming a clock-like behaviour
- Corrects for the net divergence of every leaf
- Minimizes *S* on pairs of distances to find clusters
- Very fast and efficient, with very similar output as ME

- Note: There are also additional derrived methods of NJ optimizing some aspects. These include BIONJ, generalized neighbour-joining, neighbour-joining maximum-likelihood (NJML), etc.

Maximum likelihood (ML)

Maximum likelihood is a mathematical concept to calculate the likelihood of an outcome with a given model.

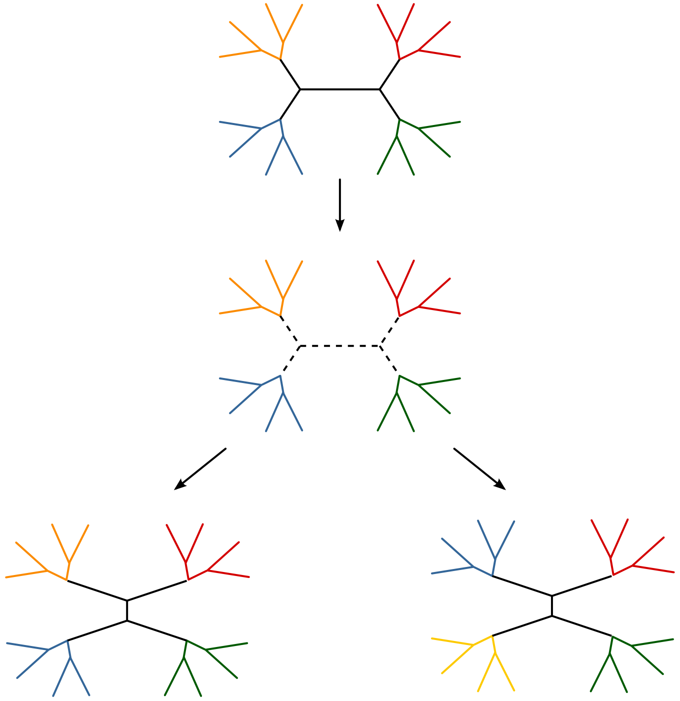
In phylogenetic analysis, the likelihood of a tree can be calculated given a **tree structure** (topology), the **branch lengths** and the **model of sequence evolution** (the substitution model).

The ML method uses different strategies (methods for tree rearrangements) to find the **tree with the highest likelihood** for the given data and model.

Maximum likelihood (ML)

Tree rearrangement algorithms:

- Nearest neighbour interchange (NNI)
  - Simplest and most used algorithm
  - exchanges the connectivity of four subtrees within the main tree
- Subtree pruning and regrafting (SPR)
  - selects and removes a subtree from the main tree and reinserts it elsewhere
- Tree bisection and reconnection (TBR)
  - detaches a subtree from the main tree and then attempts all possible connections between edges of the two resulting trees.



Nearest neighbour interchange (NNI)

Image source: Wikipedia.org

Maximum likelihood (ML) in practice

Important: choice of substitution model!

In practice, the best substitution model can be estimated using ***model finder*** in e.g. *IQTREE* or as stand alone software. However, the GTR model is usually a good choice because it allows all parameters to be estimated.

For core genome SNP data, the **ascertainment bias correction** should be used (e.g. `–m GTR+ASC` or `–m TEST+ASC`). Without `+ASC`, the branch lengths might be overestimated.

Bayesian

Bayesian approaches date back to Thomas Bayes (c. 1702–1761), a British mathematician and Presbyterian minister

Bayesian approaches calculate / estimate posterior probabilities given some prior information

Prior information can be any parameter including parameters in substitution models, sampling dates etc.

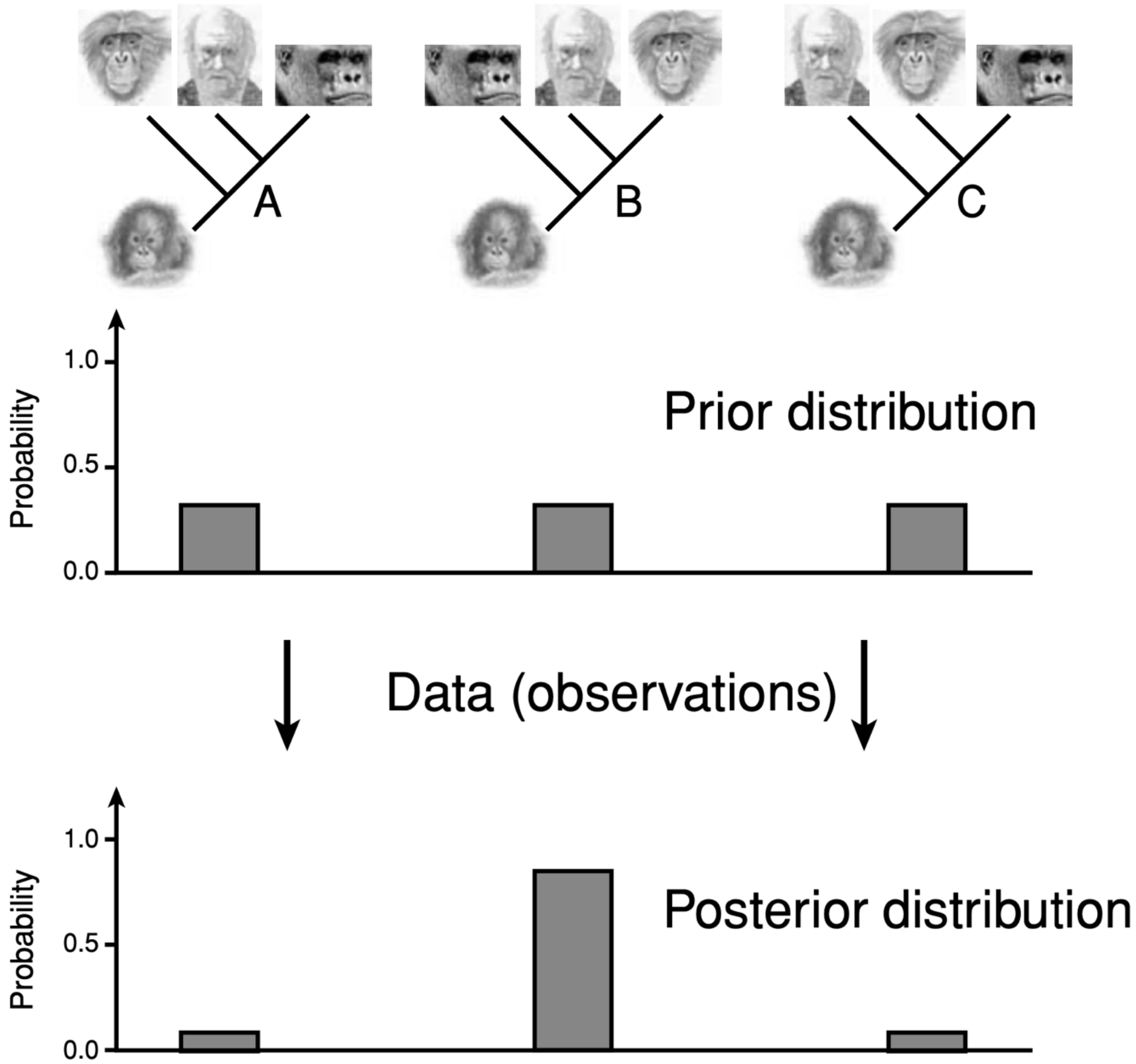


Figure from Lemeyetal et al. (2009) "Phylogenetic handbook"

Markov chain Monte Carlo (MCMC) sampling

The Bayesian approach searches for posterior probabilities in the complete parameter space, and it is therefore impossible to infer them analytically already with a handfull of taxa. Therefore, we need a search strategy.

Markov chains have the property that they converge towards an equilibrium state regardless of starting point.

Here we want a chain that converges towards our posterior probability

Markov chain Monte Carlo steps

1. Start at an arbitrary point (θ)
2. Make a small random move (to θ^*)
3. Calculate height ratio (r) of new state (to θ^*) to old state (θ)
    (a) $r > 1$: new state accepted
    (b) $r < 1$: new state accepted with probability r
        if new state rejected, stay in old state
4. Go to step 2

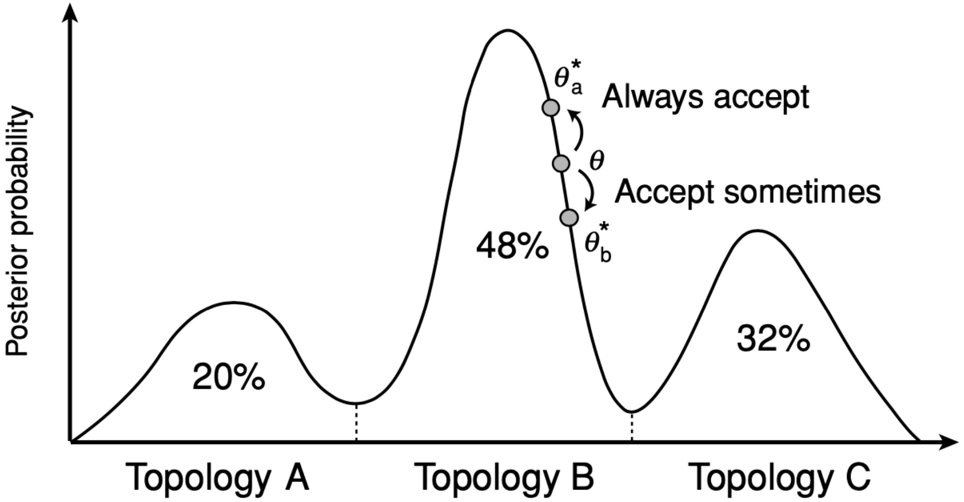


Figure from Lemeyetal et al. (2009) “Phylogenetic handbook”

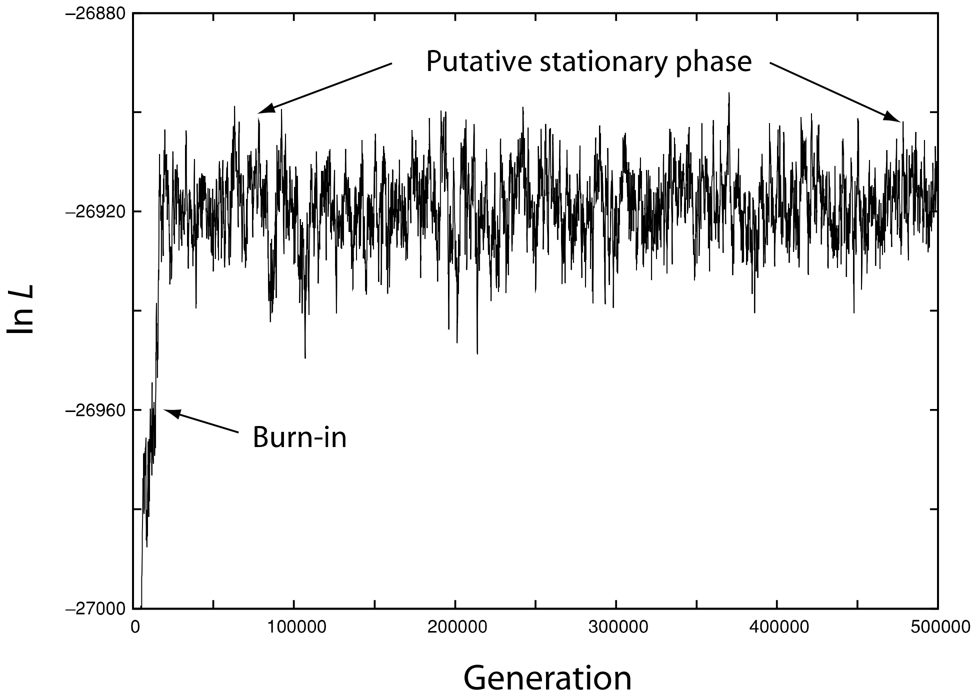
Bayesian in practice

Typically, Bayesian phylogenies are estimated in BEAST or MrBayes.

A large number of parameters can be provided (so called priors)

The Markov chain initially quickly finds towards a parameter space with high likelihood (the burn-in phase) and then circulates around these values. In a successful run, the posterior probabilities converge towards a stable value.

Bayesian phylogenies are a collection of thousands of trees, which allows to calculate a consensus tree and uncertainty for all parameters.



A typical output of the Log-likelihood of a BEAST run

Figure from Lemeyetal et al. (2009) “Phylogenetic handbook”

Advanced phylogenetic analyses

Advanced phylogenetic analyses

From a phylogeny and the given models and assumptions of evolution, a series of more advanced analyses can be done. Here are only 3 examples:

1. Time scaled trees
   - From the assumption of a molecular clock (fixed or relaxed), internal nodes and the root of a tree can be dated given sampling dates for the leaves or other internal nodes
   - Particularly useful in Bayesian tools, but also in ML trees
2. Ancestral state reconstruction
   - As with dates, known the state of some characteristics in the leaves, its state in ancestral nodes and root can be reconstructed
3. Mapping natural selection
   - From the ratio between synonymous and non-synonymous substitutions in protein-encoding sequences, the pressure of natural selection can be estimated

Assessing the reliability of a phylogeny

Assessing the reliability of a phylogeny

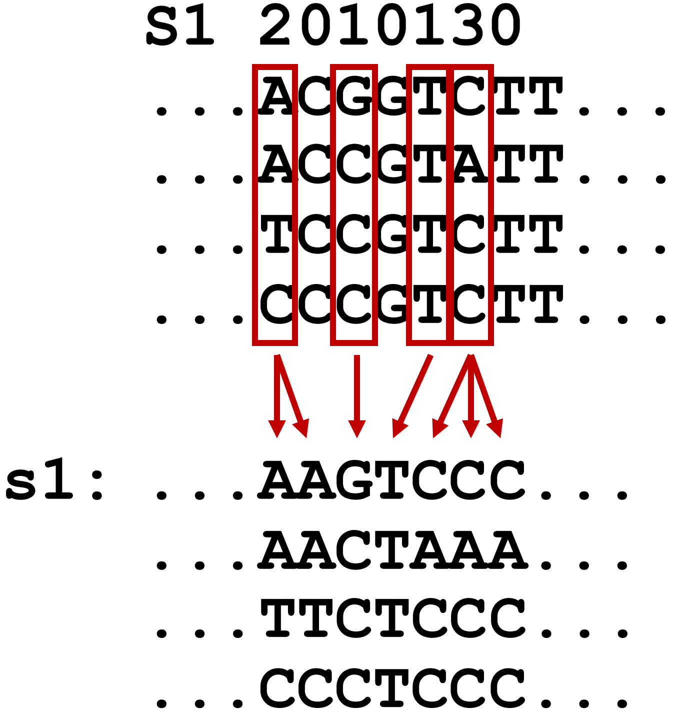
Several techniques are used to assess the reliability of an inferred tree:

- Bootstrap analysis
  - Sampling columns with replacement (same alignment length)
- Jackknifing
  - Randomly removes halv of the columns in the alignmnet
- The likelihood ratio test (LRT) (Branch test for all trees)
- Ultrafast bootstrap (UFBoot) in IQTREE
- Posterior probabilities for each split or clade (Bayesian trees)

Bootstrap

1. Alignment columns are randomly sampled with replacement until an alignment of the same length as the original is obtained
2. Create a tree with the same methods and parametrs as the original tree
3. Repeat this for n=100-2000 times
4. The proportion (%) of each clade among all the bootstrap replicates is computed on a consensus tree or the original tree as a statistical confidence for a branch / node

NB: each bootstrap replicate takes as much time to compute as the original tree.

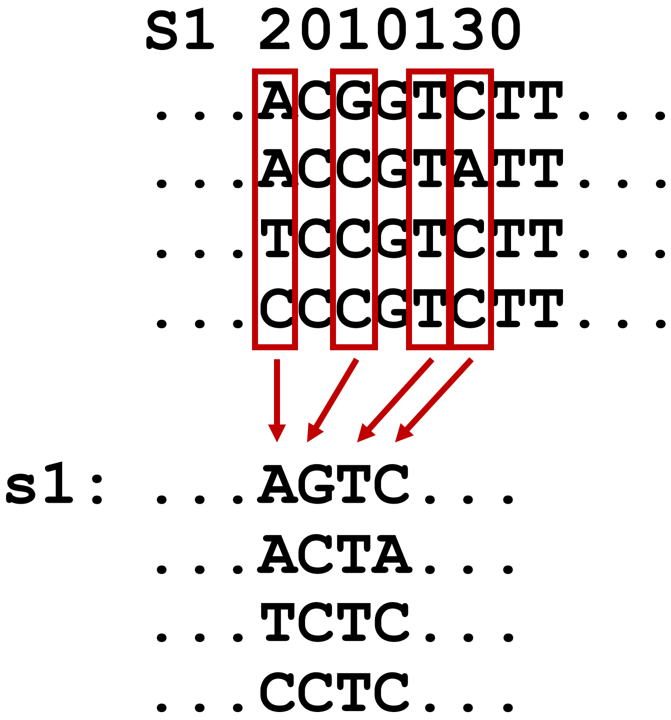


```
S1 2010130
...ACGGTCTT...
...ACCGTATT...
...TCCGTCTT...
...CCCGTCTT...

s1: ...AAGTCCC...
    ...AACTAAA...
    ...TTCTCCC...
    ...CCCTCCC...
```

Jackknife

1. Alignment columns are randomly sampled without replacement until an alignment of ½ of the original length is obtained
2. Create a tree with the same methods and parametrs as the original tree
3. Repeat this for n=100-2000 times
4. The proportion (%) of each clade among all the jackknife replicates is computed on a consensus tree or the original tree as a statistical confidence for a branch / node

NB: This is faster than bootstrap, but still requires to produce many trees.



```
S1 2010130
    ...ACGGTCTT...
    ...ACCGTATT...
    ...TCCGTCTT...
    ...CCCGTCTT...

s1: ...AGTC...
    ...ACTA...
    ...TCTC...
    ...CCTC...
```

A quick note on alignments

Core-genome SNP analysis

Core genome SNP analysis is most commonly made by mapping the reads or fasta files from one or more isolates to a reference genome

Instead of looking at just a single gene (like in the case of 16s rRNA) or a few genes (like we do in MLST), we look at mutations in the entire core genome, i.e. the part of the reference genome that is present in all isolates used in the analysis.

Tools for core-genome SNP's: **NASP**, **snippy**

From alignment to phylogeny

The quality of the alignment is crucial for the result of the phylogenetic analysis!

When doing phylogenetic analysis on alignments be aware of:

- Poorly aligned sequences
- Low-quality terminal regions in both ends (all differences may count in the phylogenetic analysis!)
- The gap-penalty can have major influence on the alignment
- Recombination in SNP matrices can introduce many SNP's (→ use software tool **gubbins** or similar to remove most of it)

Applications of phylogenetic analyses

Application of phylogenies

Phylogenies are used very commonly in biology, but also many other disciplines where biology plays a role.

Applications:

- Reconstruction of evolutionary relationships
- Reconstruction of population dynamics (over time / space)
- Outbreak analysis
- Typing of strains/variants

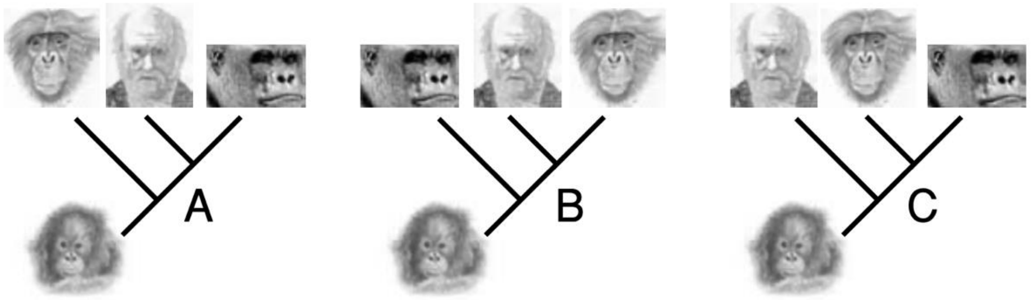
Reconstruction of evolutionary relationships

Question: How is the evolutionary relationship between species / taxa? Or "Where do we / a pathogen come from?"

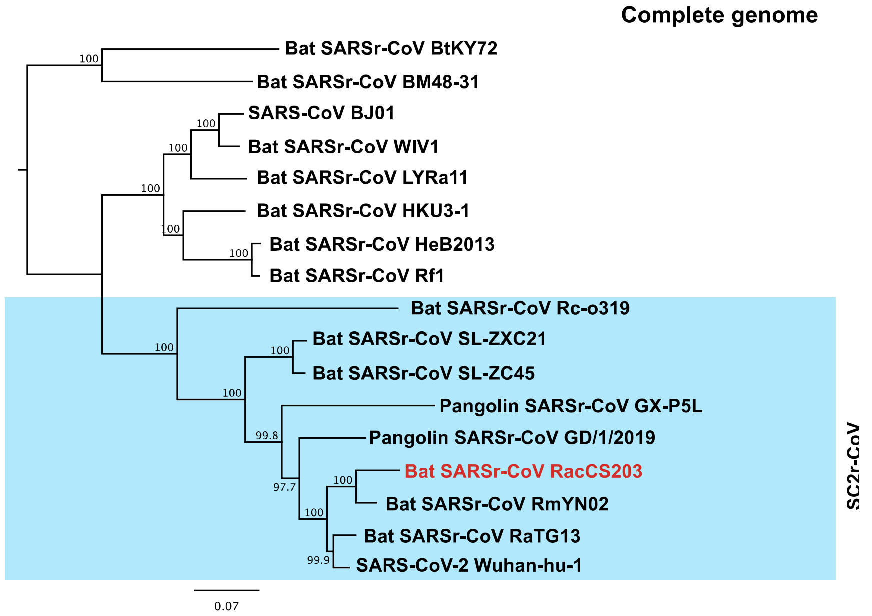
Examples:

- Human – great apes relationship
- SARS-CoV-2 origin

This allows to draw conclustions from the known relatives



Scenarios homo's closest relatives



Origin of SARS-CoV-2 virus

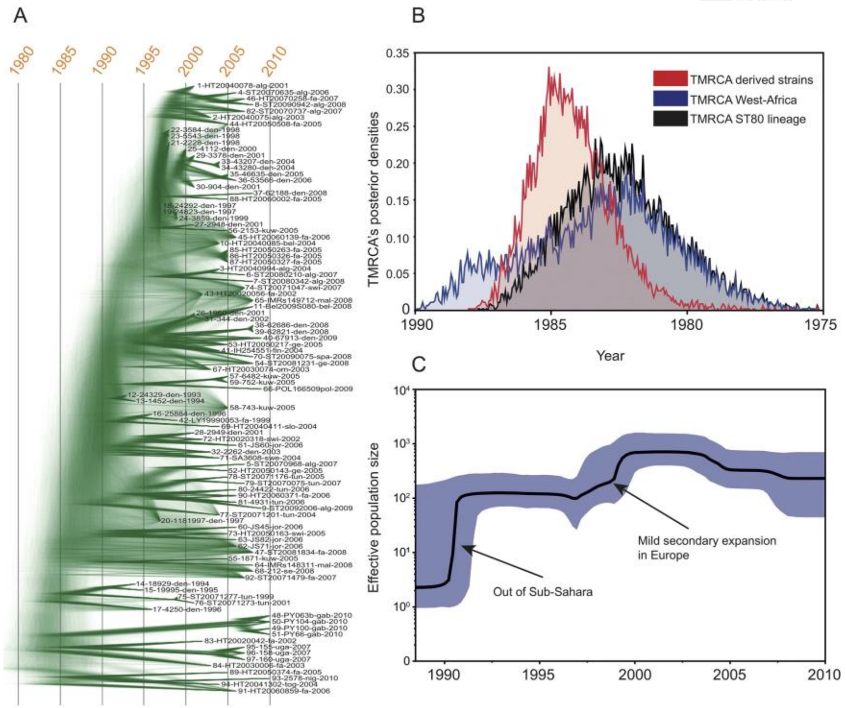
Wacharapluesadee et al. 2021, Nat. Comm

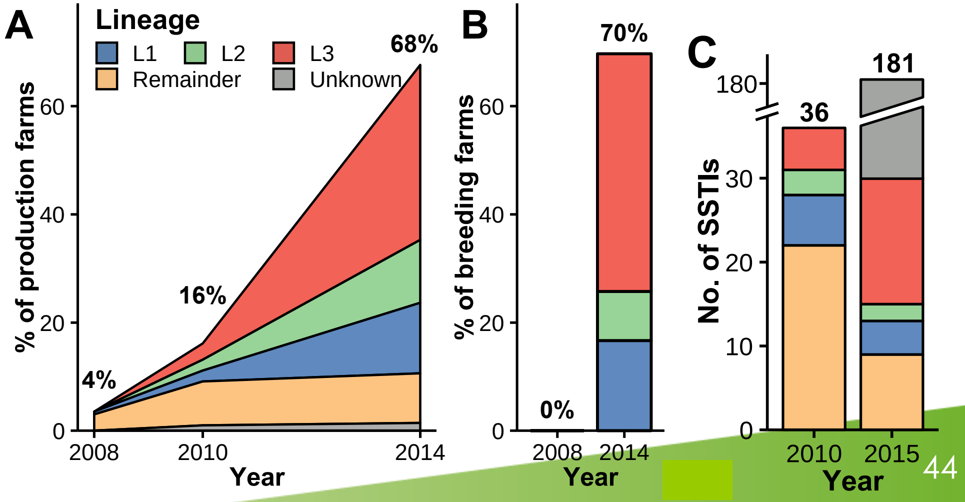
Reconstruction of population dynamics

Question: How did a population of an organism develop over time and space?

Examples:

- Bayesian reconstruction of the Staphylococcus aureus CC80 complex (top)
- Spread of MRSA in Danish pigs and humans





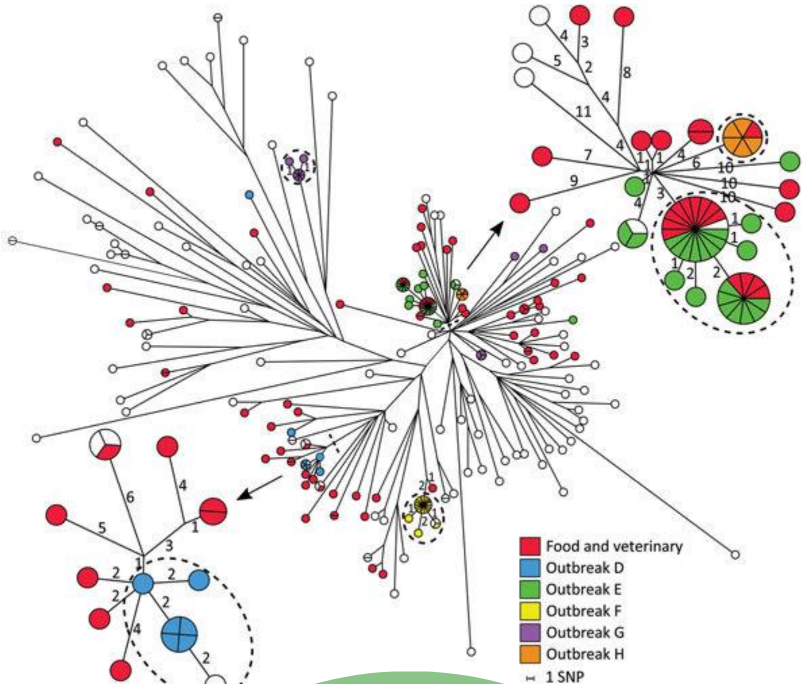
Stegger et al. (2014), mBio; Sieber et al. 2018, mBio.

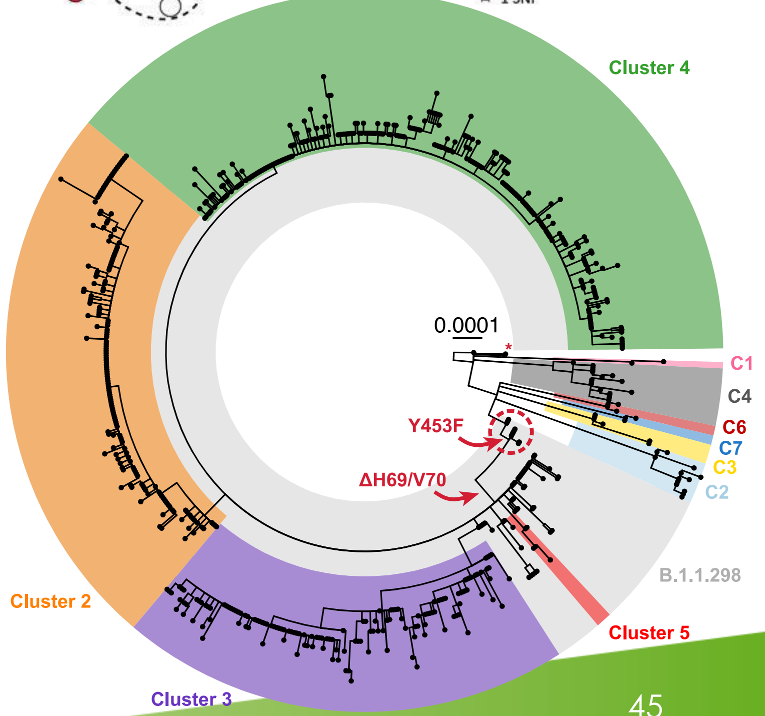
Outbreak analysis

Question: Do we have an outbreak? Which isolates are part of it and which are not?

Examples:

- Outbreaks of *Salmonella enterica* Serovar Typhimurium in Denmark (top)
- SARS-CoV-2 in Danish mink (bottom)





Gymoese et al. (2017), EID; Rasmussen et al. (2024), bioRxiv

Phylogenies for typing bacteria

Various typing methods use phylogenies for grouping types:

	Description	Typing class	Discriminatory power	Phylogenetic method?
Gram typing	Staining of cells	Phenotypic	Very low	No
MALDI-TOF	Species identification	Phenotypic	Species level	No
Serotyping	Immunological typing	Phenotypic	Within species	No
PFGE	Pulsed-field gel electrophoresis	Molecular	(High)	Yes
MLST	Multi-locus sequence typing	Molecular	High	No
wgMLST	Whole-genome MLST	Molecular	Very high	Yes
Species specific like *spa*-typing	Typing based on one variable gene	Molecular	High	No
Core genome SNPs	Typing based on sinlge nucleotide polymorphisms	Molecular	Very high	Yes

Software

Tree reconstruction software

	Distance Based	Maximum Parsimony	Maximum likelihood	Bayesian	Platform	Interface
MEGA	X	X	X		Mac, PC, Linux	GUI
PAUP	X (bionj)	X			Mac, PC, Linux (CL only)	GUI, CL
PhyML			X		Mac, PC, Linux	CL
PHYLIP	X	X	X		Mac, PC, Linux	CL
RAxML			X		Mac, PC, Linux	CL
BIONJ	X				Mac, PC, Linux	CL
IQTREE			X		Mac, PC, Linux	CL
BEAST				X	Mac, PC, Linux	GUI
MrBayes				X	Mac, PC, Linux	CL
FastTree			Approximate ML		Mac, PC, Linux	CL
VeryFastTree			Approximate ML		Mac, PC, Linux	CL

GUI: Graphical User Interface; CL: Command line

For more software and information visit: https://en.wikipedia.org/wiki/List_of_phylogenetics_software

Tree visualization software

All software is freely available

- Stand-alone software:
  - FigTree: fast and efficient tree visualization and annotation (PC, Mac and Linux)
  - MEGA: The phylo package has a great GUI and can also visualize trees
  - Treeview: Very basic tree visualization (open source, PC, Mac, Linux)
- Online tools:
  - iTOL: Nice tree visualization and annotation. Payed account needed for saving trees.
  - Microreact: Tree visualization and link to metadata incl. geographic data
  - Nextstrain: Visualization of trees, metadata and mutations (developed for virus)
- Software packages:
  - Python: BioPython, ETE3
  - R: phytools, ggtree, ape

In summary

- Phylogenies aim to reconstruct the evolutionary relationship among taxa given the provided data.
- There are different methods to estimate this relationship including distance based methods, maximum parsimony, maximum likelihood and Bayesian approaches.
- The reliability of a group can be assessed using different methods including bootstrap, the likelihood ratio test or posterior probabilities.
- A diversity of software for phylogenetic reconstruction, analysis and visualization is available, most of them for free.

Further reading

- Phylogenetic tree building in the genomic age. Paschalia Kapli, Ziheng Yang, Maximilian J. Telford. Nature Reviews Genetics 2020 https://dx.doi.org/10.1038/s41576-020-0233-0
- The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. N.p.: Cambridge University Press, 2009.
- Bioinformatics and Molecular Evolution. Higgs, Paul G., and Attwood, Teresa K.. Germany, Wiley, 2013.
- Computational Molecular Evolution. Yang, Ziheng. United Kingdom: Oxford University Press, 2006.
- Inferring phylogenies. Felsenstein, Joseph. United Kingdom: Oxford University Press, Incorporated, 2004.
- Molecular Evolution: A Statistical Approach. Yang, Ziheng. United Kingdom: Oxford University Press, 2014.

References

- Kapli P, Yang Z, Telford MJ. **Phylogenetic tree building in the genomic age.** Nat Rev Genet. 2020 Jul;21(7):428-444. doi: 10.1038/s41576-020-0233-0. Epub 2020 May 18. PMID: 32424311.
- **The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing**. Vandamme, Salemi & Lemey (eds.). Cambridge University Press, 2009.
- Wacharapluesadee S, et al.. **Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia**. Nat Commun. 2021 Feb 9;12(1):972. doi: 10.1038/s41467-021-21240-1. Erratum in: Nat Commun. 2021 Feb 25;12(1):1430. PMID: 33563978; PMCID: PMC7873279.
- Stegger M, et al.. **Origin and evolution of European community-acquired methicillin-resistant Staphylococcus aureus.** mBio. 2014 Aug 26;5(5):e01044-14. doi: 10.1128/mBio.01044-14. PMID: 25161186; PMCID: PMC4173770.
- Sieber RN, et al.. **Drivers and Dynamics of Methicillin-Resistant Livestock-Associated Staphylococcus aureus CC398 in Pigs and Humans in Denmark**. mBio. 2018:9(6):e02142-18. Doi: 10.1128/mbio.02142-18. PMID: 0425152; PMCID: PMC6234867.
- Gymoese P, et al.. **Investigation of Outbreaks of Salmonella enterica Serovar Typhimurium and Its Monophasic Variants Using Whole-Genome Sequencing, Denmark**. Emerg Infect Dis. 2017 Oct;23(10):1631-1639. doi: 10.3201/eid2310.161248. PMID: 28930002; PMCID: PMC5621559.
- Rasmussen, T. B. et al. **Emergence and spread of SARS-CoV-2 variants from farmed mink to humans and back during the epidemic in Denmark, June-November 2020**. bioRxiv 2024.02.13.580053 (2024) doi:10.1101/2024.02.13.580053.

Acknowledgements

The creation of this training material was commissioned by ECDC to Statens Serum Institut, Denmark, with the direct involvement of Raphael N. Sieber