BIOINFORMATICS 1

or why all biologists need computers

Wojciech Makałowski
Institute of Bioinformatics
Faculty of Medicine

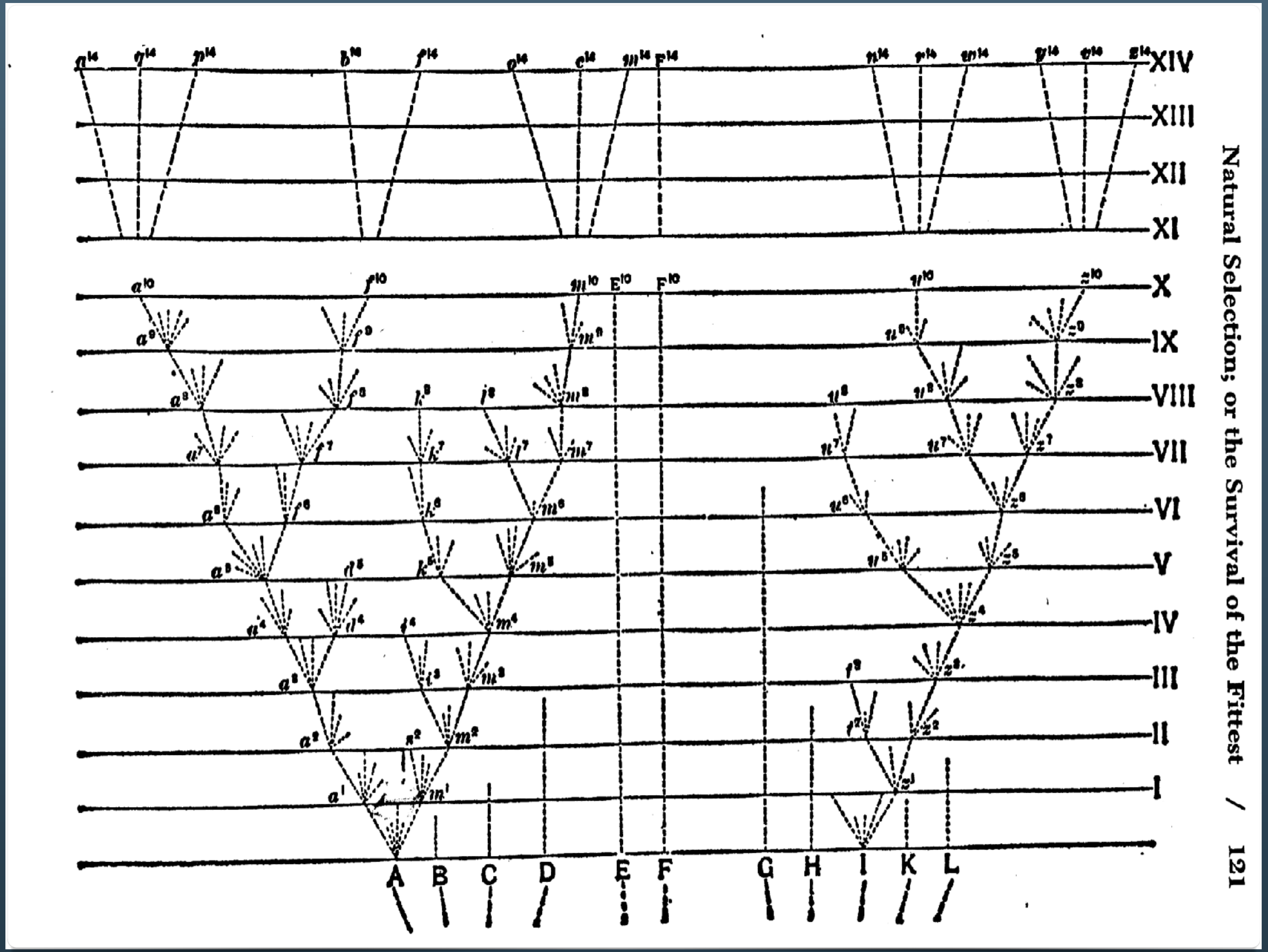
Phylogenetic Inference

What is phylogenetic analysis

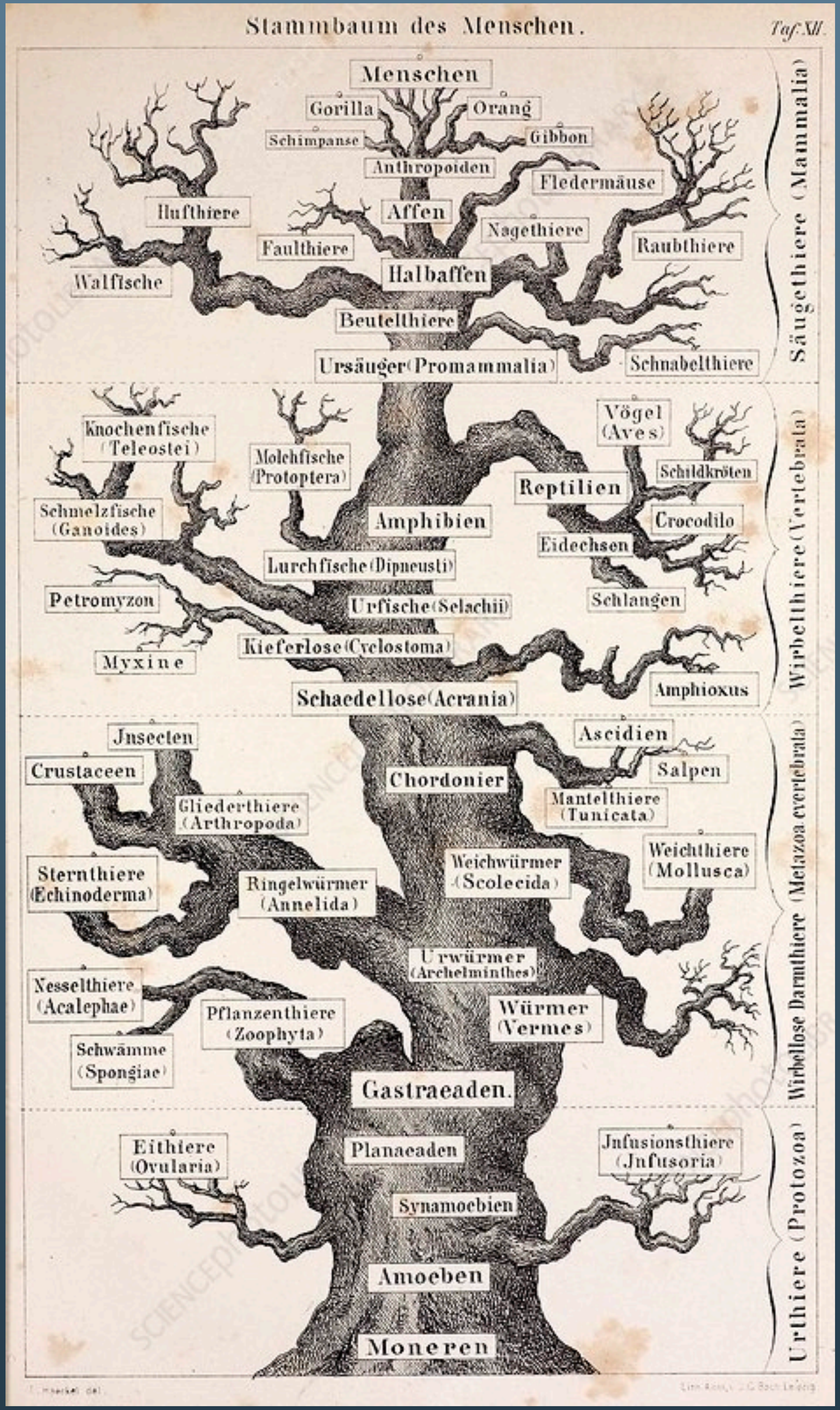
Phylogenetics is the study of evolutionary relationships

Can include morphology and physiology, paleontological, geological and molecular evidence

“…the great Tree of Life….covers the earth with ever-branching and beautiful ramifications…” Charles Darwin, 1859



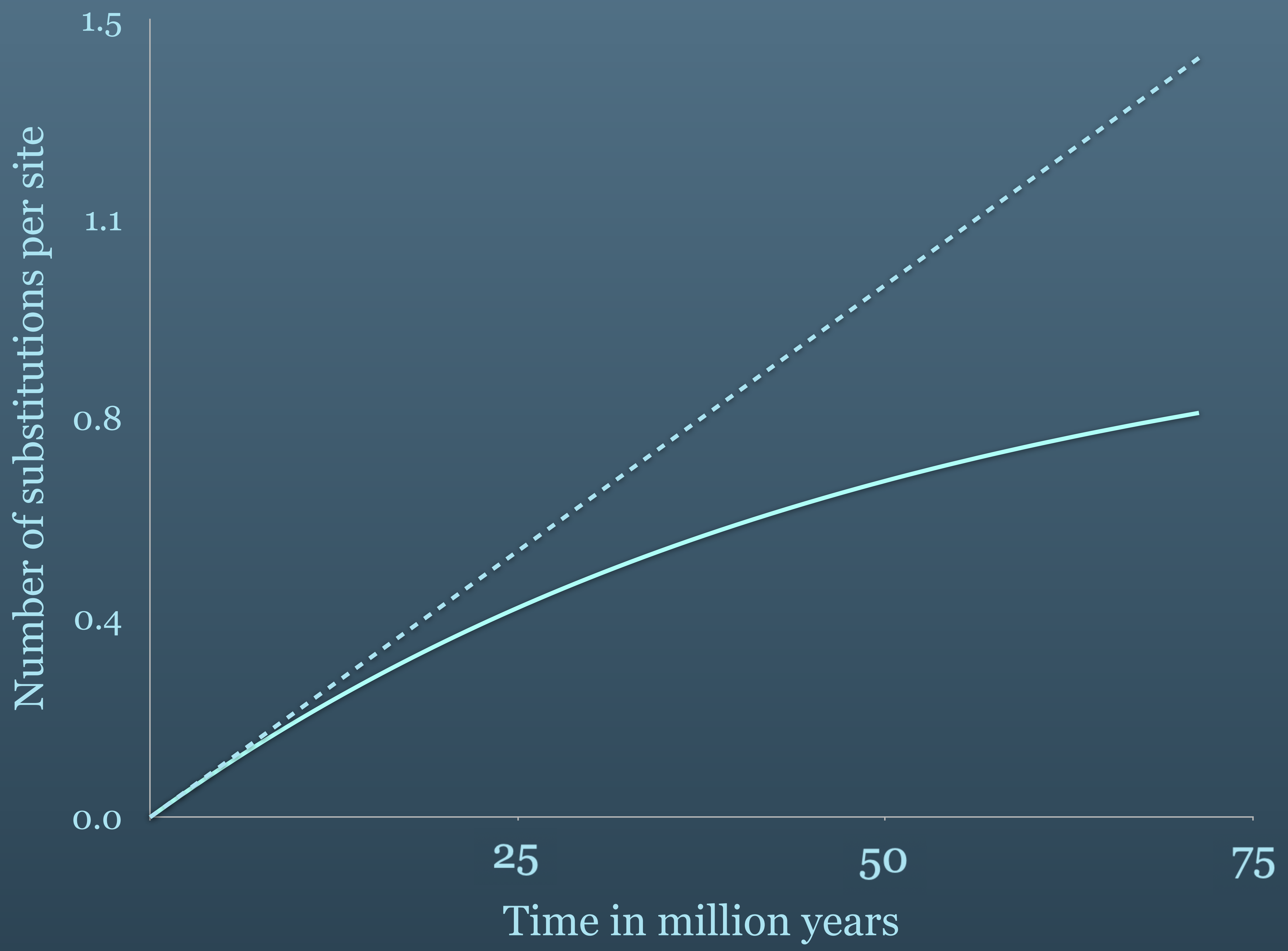
The use of trees as metaphors was promoted by Ernst Haeckel



Stem-tree of lineal progenitors of man.
(From Haeckel *Anthropogenie*, 1874)

Molecular phylogenetics

- Accumulated mutational changes in DNA and protein sequence over time constitute evidence
- Sequence-based phylogenetic analysis is performed using computers



Things to remember

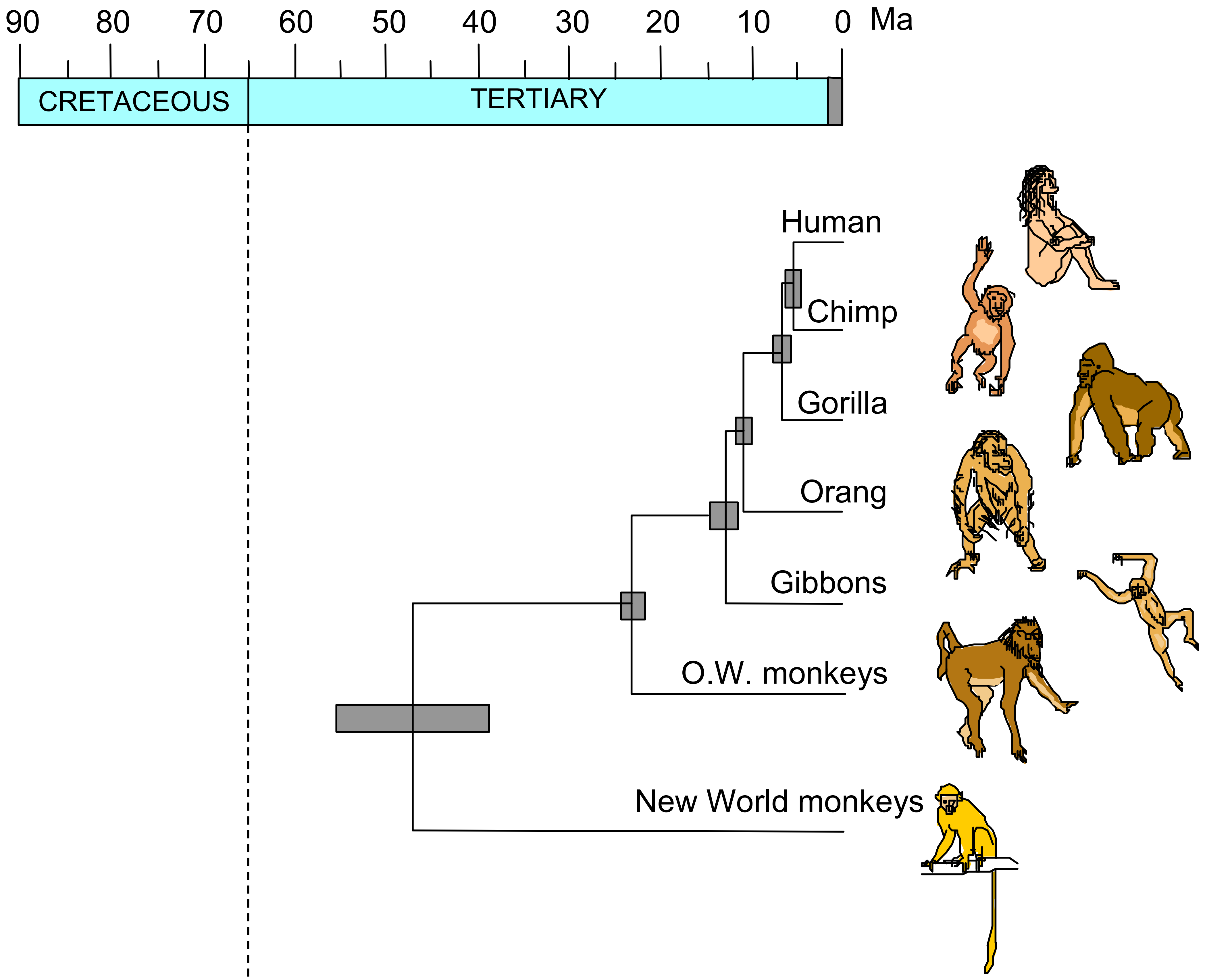
- The events that determine a phylogeny happened in the past
- They cannot be known empirically, they can only be inferred from their "end products,” whether these are morphological or molecular
- The tree is the model of evolutionary events that best explains the end product (diverged group of sequences)
- Phylogenetic analysis is modeling or estimation, and the quality or certainty of the analysis should be presented along with the result

Why do we need phylogenetics

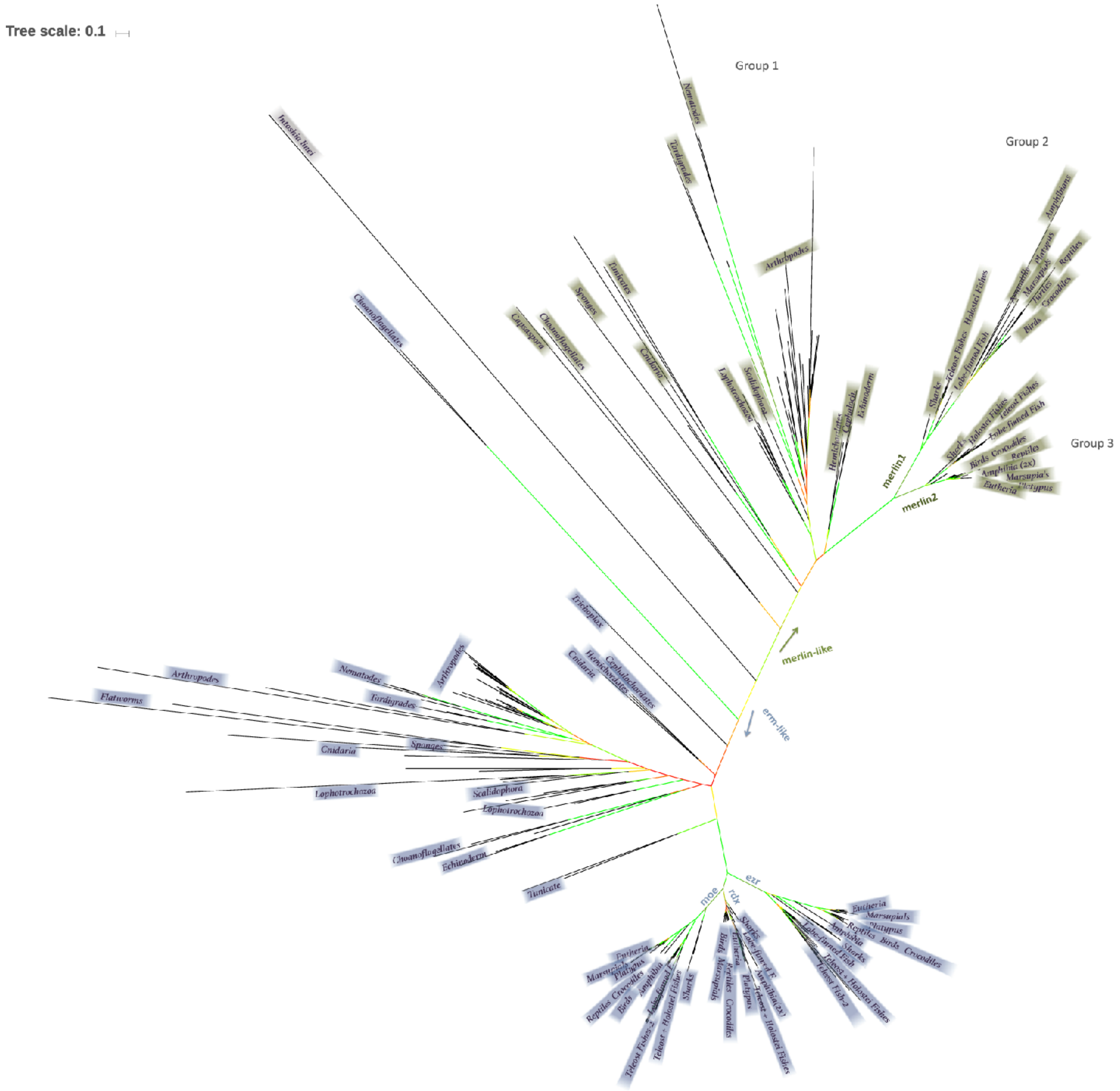
- Evolutionary studies
- Inferring functions of “new” proteins
- Medicine, pharmacology, agriculture
- Forensic science
- Ecology, nature preservation

Examples

molecular taxonomy



gene family evolution

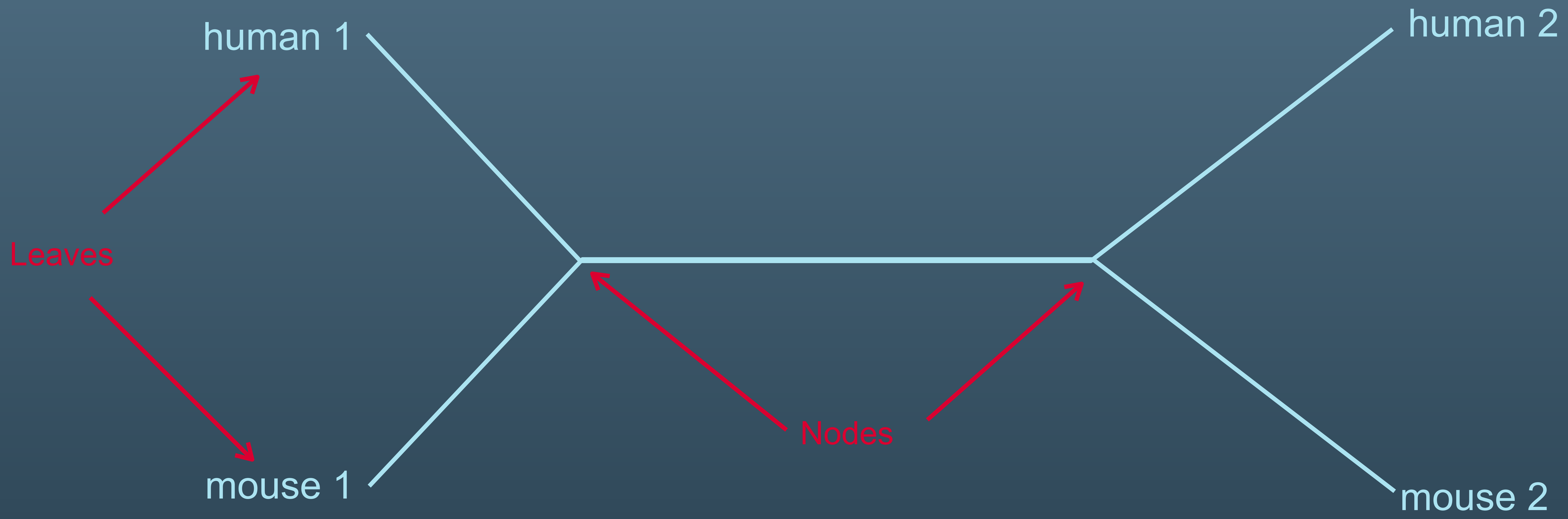


Nomenclature

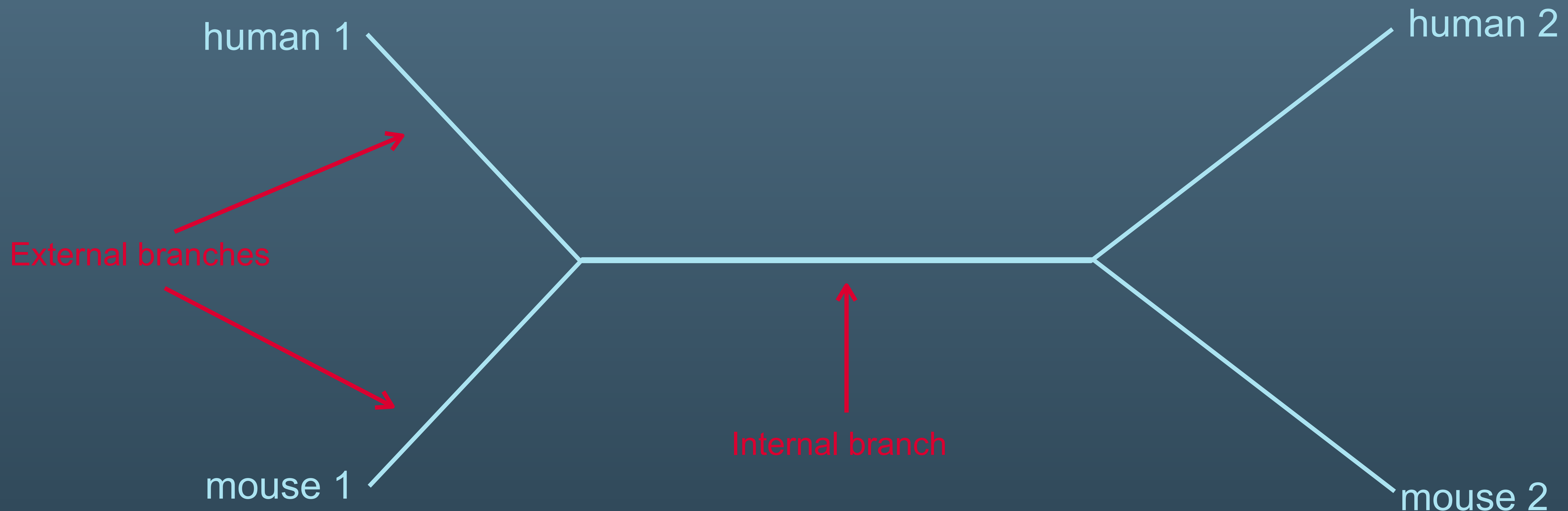
A phylogenetic tree is characterized by ”leaves”, ”nodes” and ”branches.”

- Leaves (vertices) represent species or sequences compared. They are often called Operational Taxonomy Units (OTUs)
- Nodes (vertices) are usually bifurcations and represent gene duplication or speciation events, hypothetical ancestor sequences.
- Branches (edges) are always linear and represent sequence diversity but can also be of unit length.
- The root (vertex) is optional and represents the hypothetical ancestor.

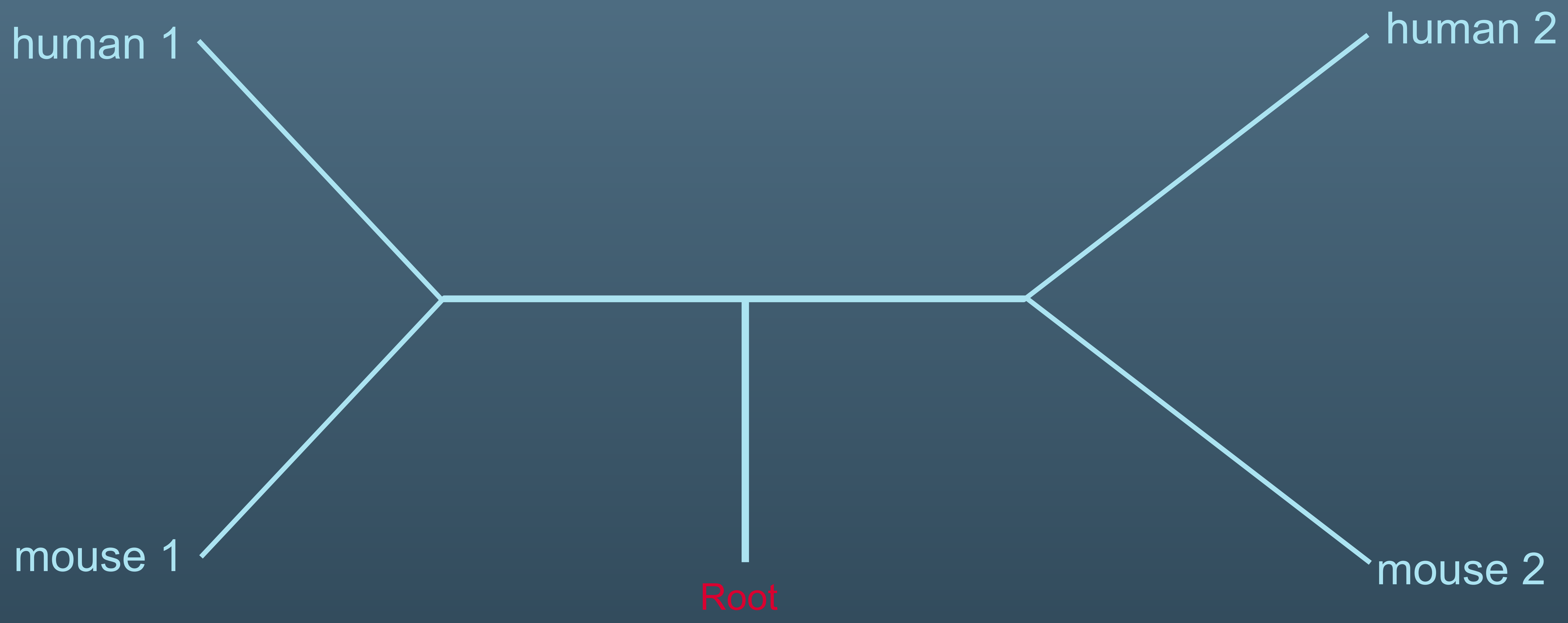
Nomenclature



Nomenclature

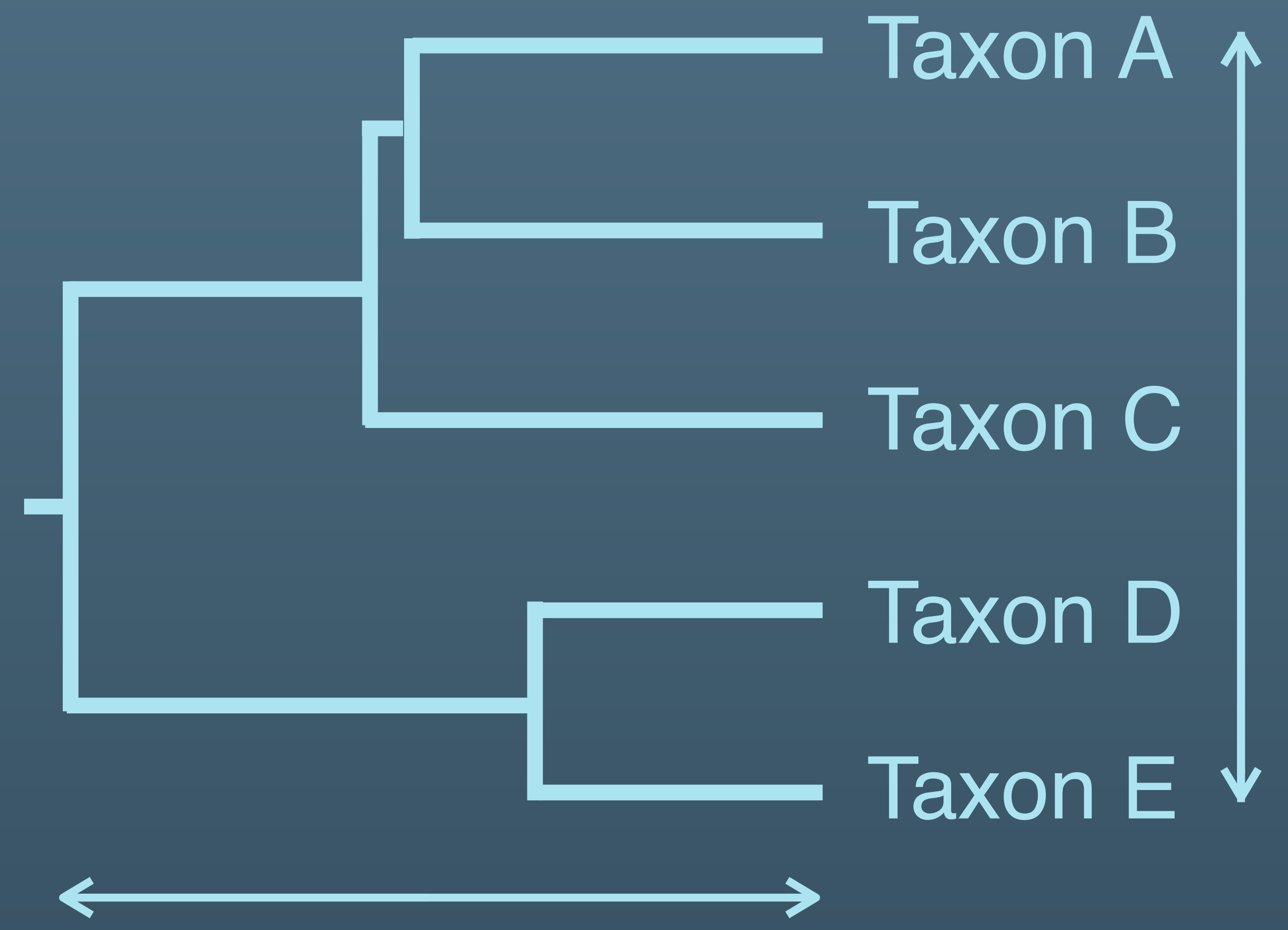


Nomenclature



The root is optional and represents the hypothetical ancestor

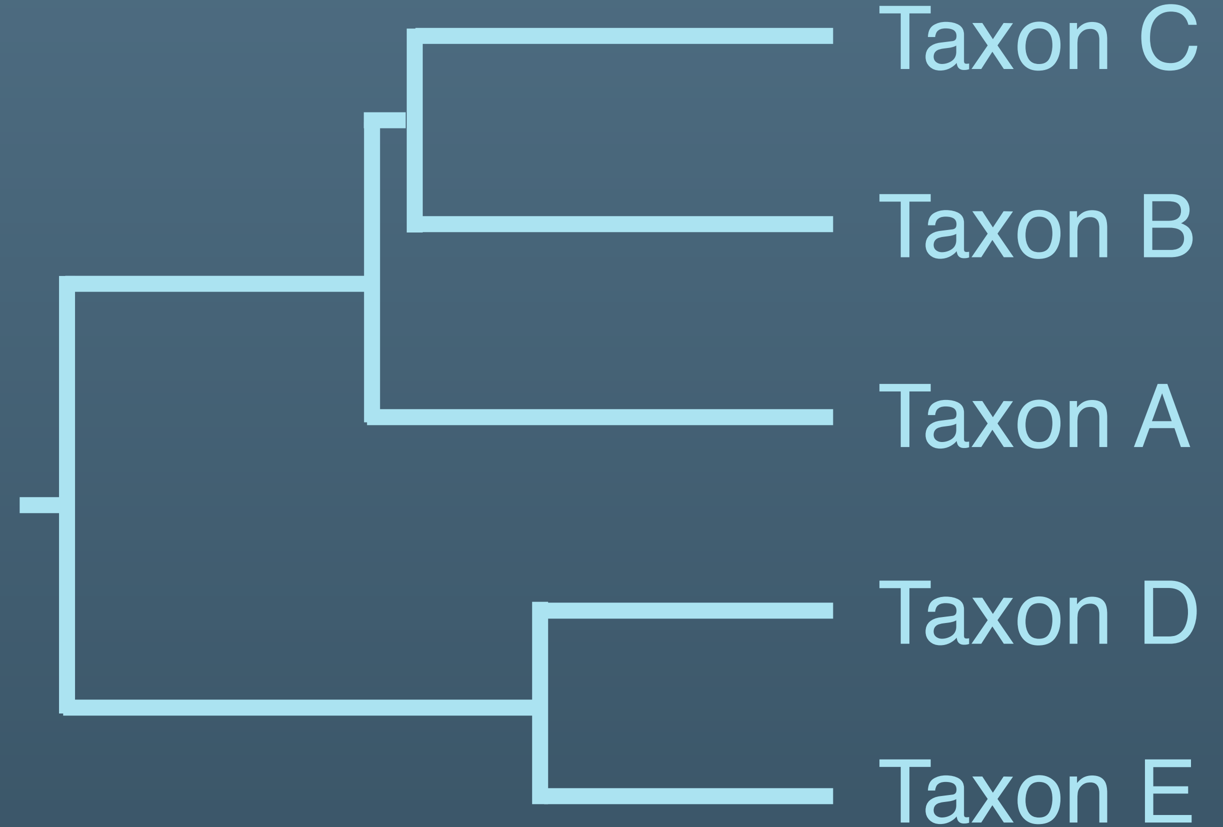
Tree interpretation



There is no meaning to the spacing between the taxa, or the order in which they appear from top to bottom

This dimension either can have no scale (for so called cladograms) or can be proportional to genetic distance (phylograms) or can be proportional to time (ultrameric trees)

Tree interpretation



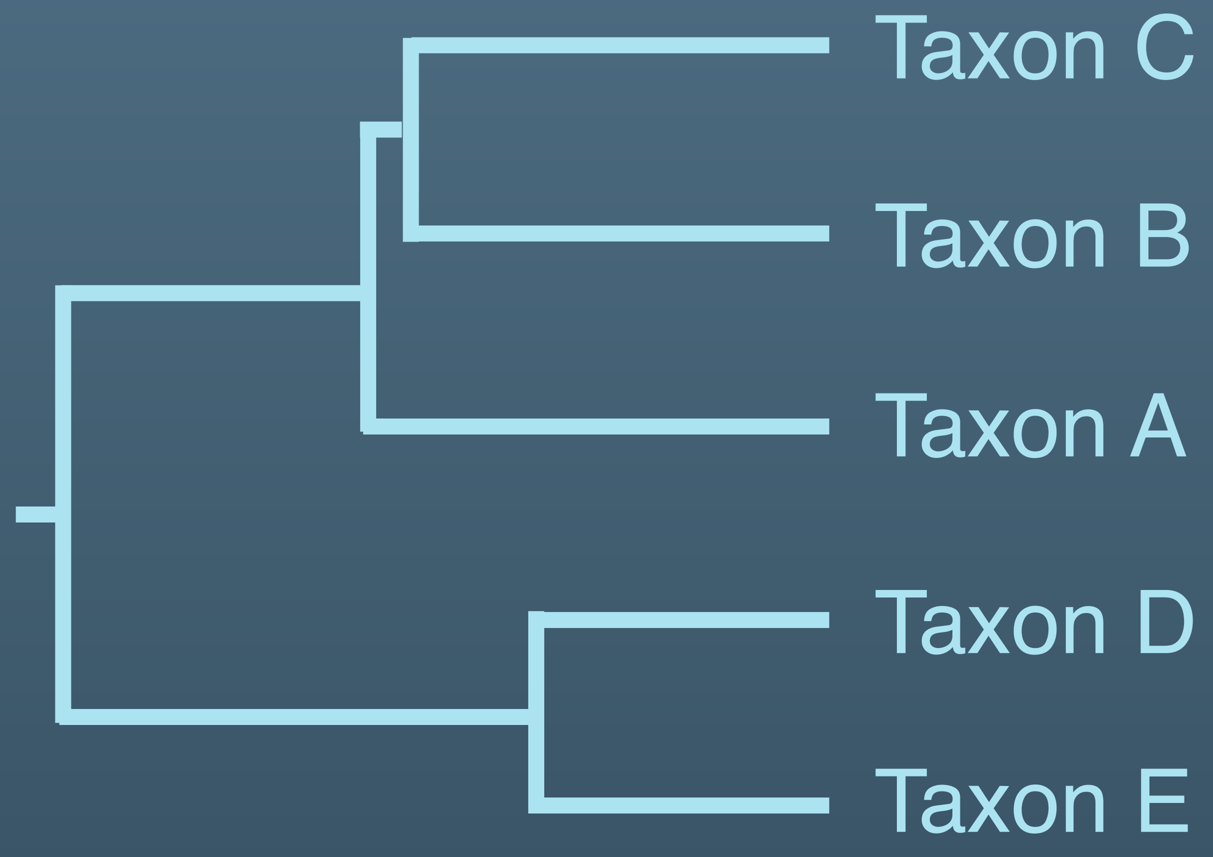
The same tree coded as a set of nested parentheses in so called the Newick tree format:

((A,(B,C)),(D,E))

This tree suggests that B and C are more closely related to each other than either to A. Moreover, A, B and C form a clade (cluster) that is a sister group to the clade consisting of D and E. If the tree has a time scale, then D and E are the most closely related.

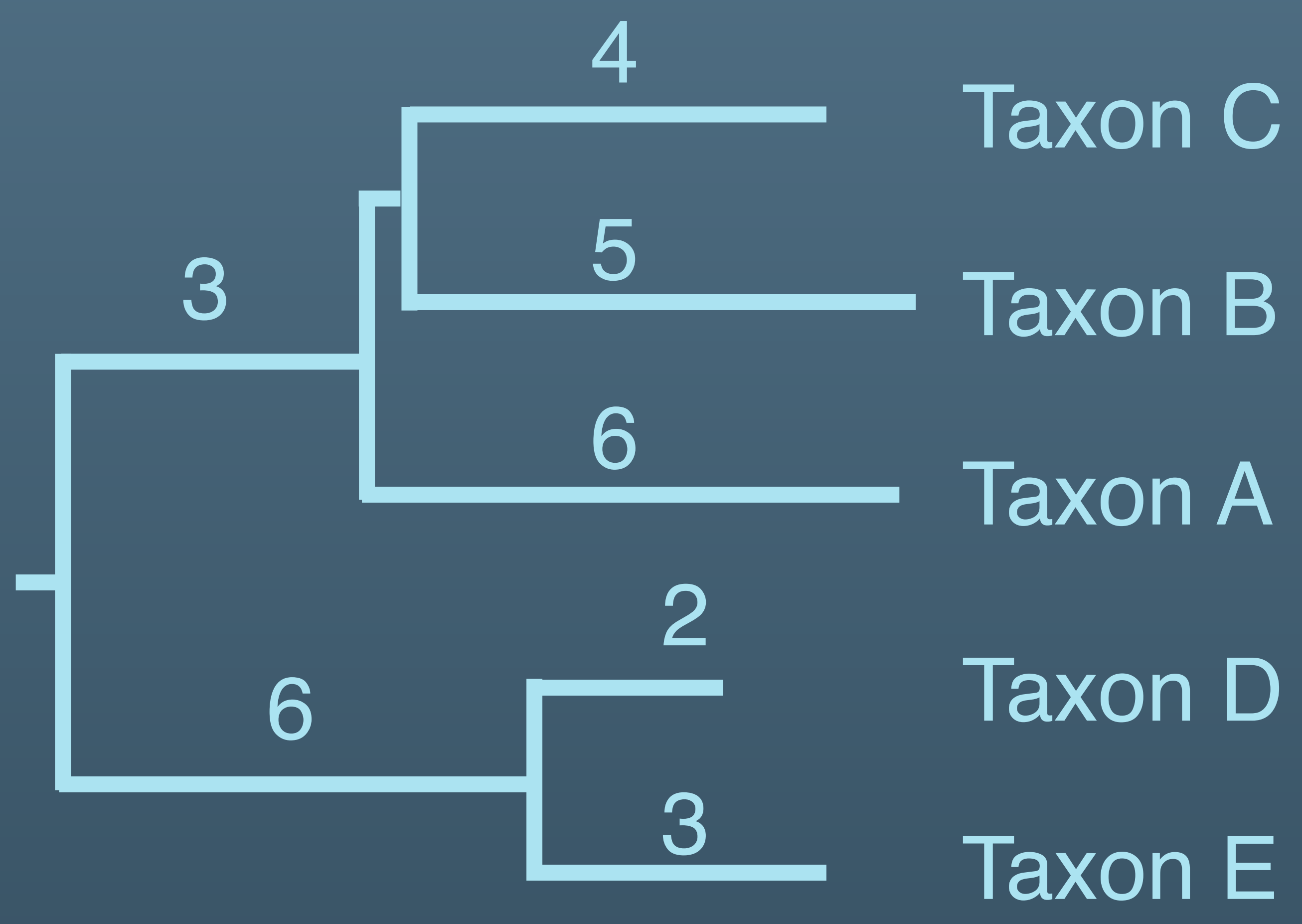
Tree interpretation

Cladogram



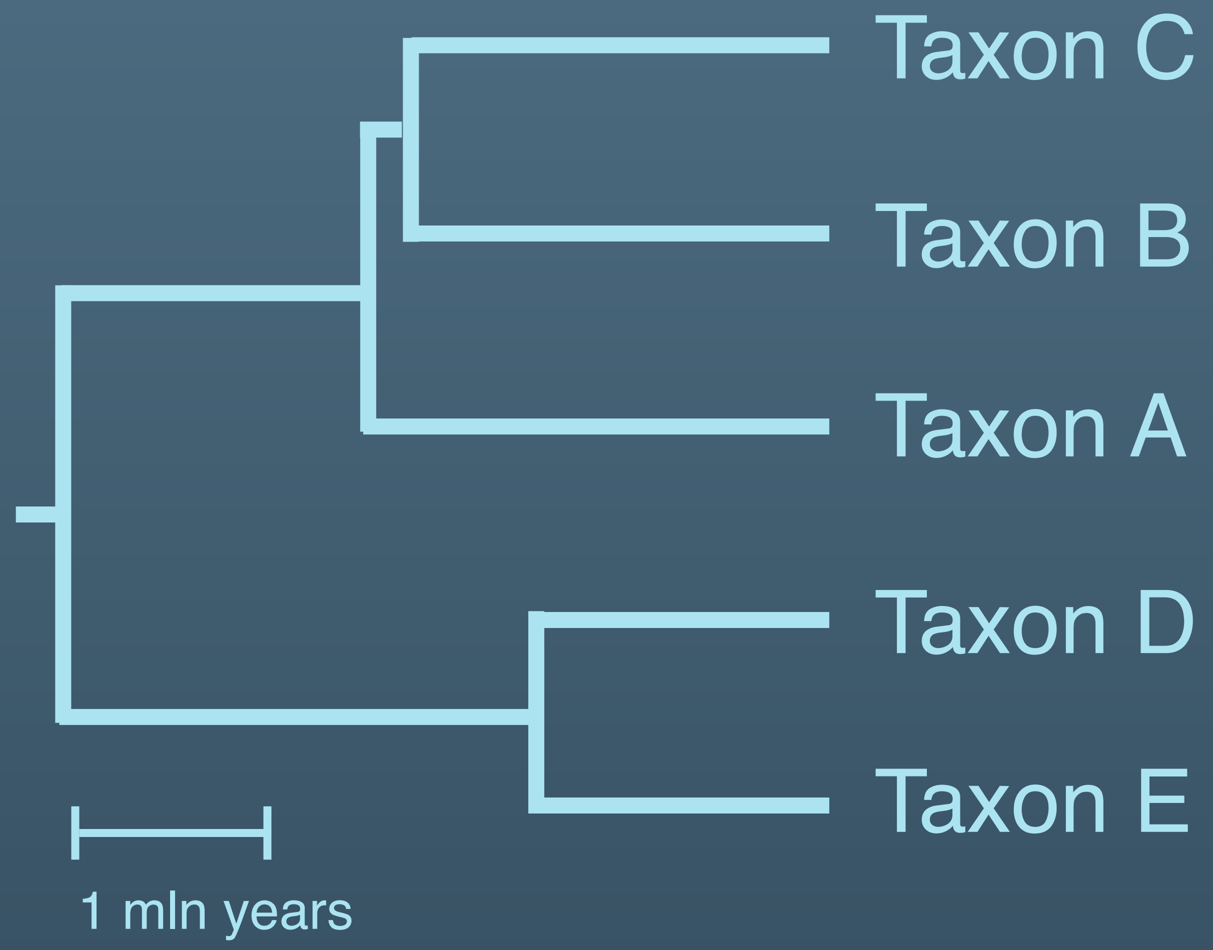
Branch length has no meaning

Phylogram



Numbers represent genetic changes

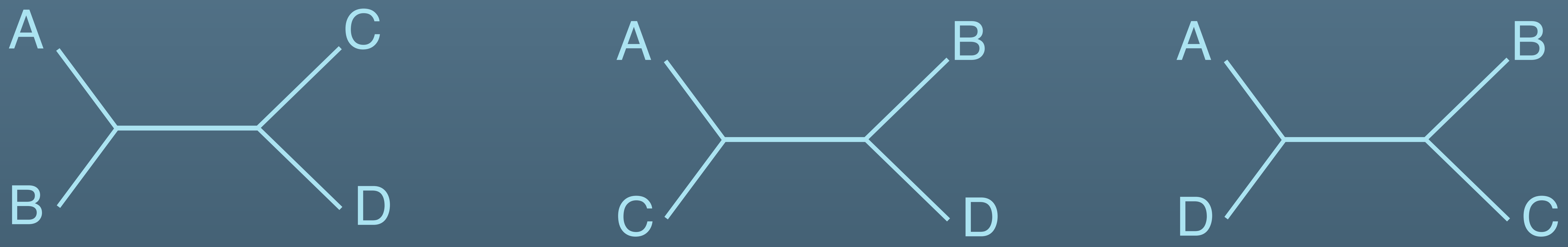
Ultrameric tree



Scale represents time

All these trees show the same evolutionary relationships between the taxa

There are three possible unrooted trees for four taxa



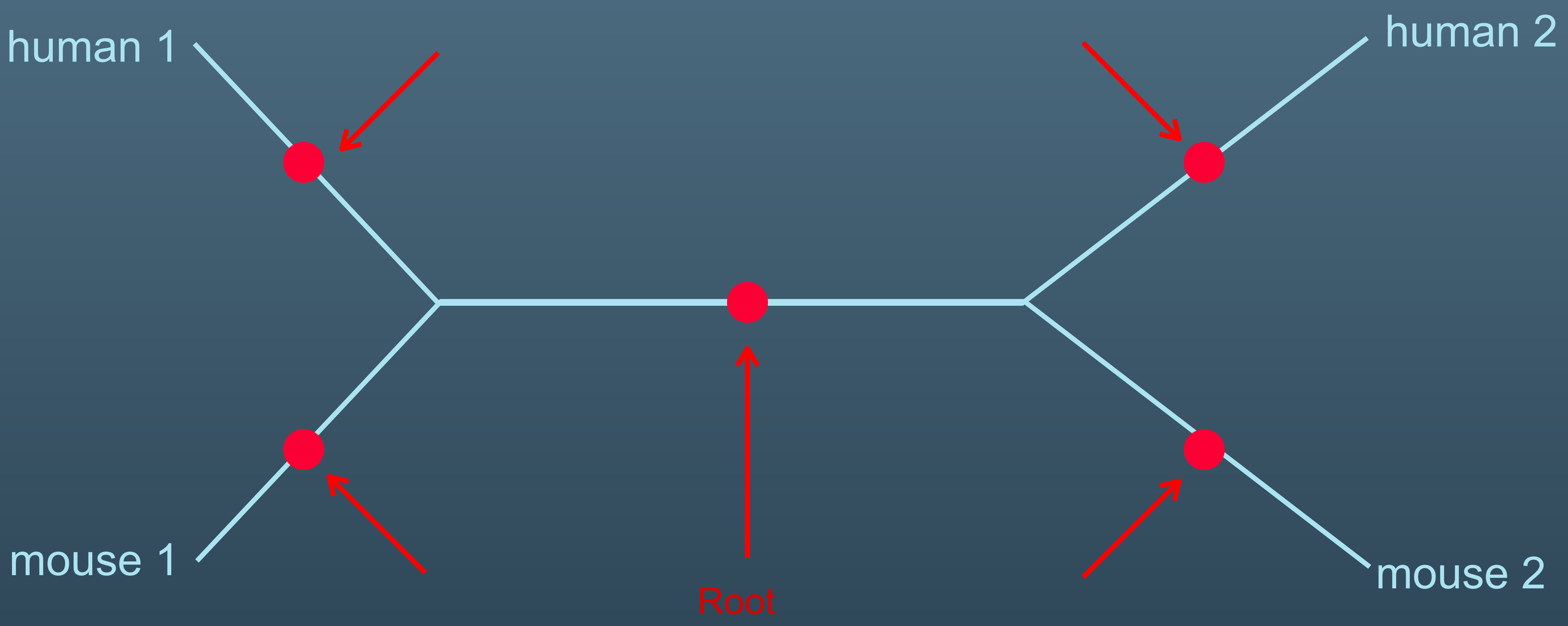
Phylogenetic tree building (or inference) methods are aimed at discovering which of the possible unrooted trees is “correct". We would like this to be the “true” biological tree — that is, one that accurately represents the evolutionary history of the taxa. However, we must settle for discovering the computationally correct or optimal tree for the phylogenetic method of choice.

The number of unrooted trees increases in a greater than exponential manner with number of taxa

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Number of OTUs	Number of unrooted trees
2	1
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
15	7,905,853,580,625
20	221,643,095,476,699,771,875

An unrooted, four-taxon tree can be rooted in five different places to produce five different rooted trees



The number of rooted trees is even higher

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

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

Number of OTUs	Number of unrooted trees	Number of rooted trees
2	1	1
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
15	7,905,853,580,625	213,458,046,676,875
20	221,643,095,476,699,771,875	8,200,794,532,637,891,559,375

Five steps in phylogenetic analysis

1. Finding all homologs
2. Multiple sequence alignment
3. Building a tree
4. Statistical assessment of a tree
5. Viewing a tree and drawing conclusions

Amino acid or nucleotide data

- Amino acid sequences are more conserved (redundant codons).
- Thus, for closely related species, DNA data bring more info.
- For long scale studies nucleotide signal is blurred by multiple substitutions at the same site.
- Some methods use codons features for the analysis.

Step 1: Finding all homologs

- Start with amino acid sequence as a seed
- Sequence similarity search is the most popular approach:

  use BLASTp and PSI-BLAST or delta-BLAST to find distant homologs
- Text search in protein databases is often useful in finding distant, very diverged homologs
- Search protein domains database, e.g. *Pfam* to find even more distant homologs

Step 2: Multiple sequence alignment

Approaches to Multiple Sequence Alignment

- Dynamic Programming
- Progressive Alignment
- Iterative Alignment
- Statistical Modeling

Dynamic programming approach

Dynamic programming with two sequences

- Relatively easy to code
- Guarantee to obtain optimal alignment

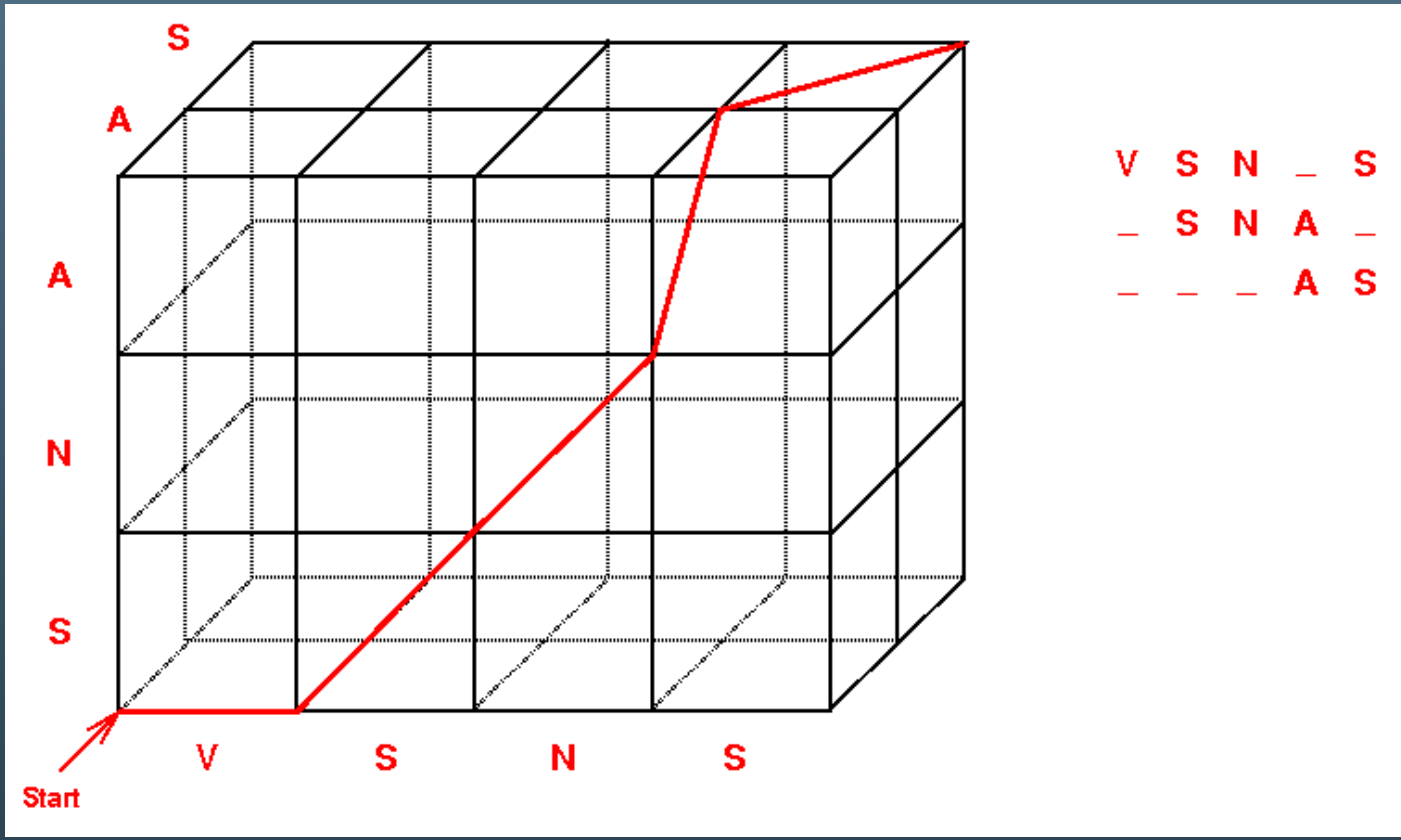
Can this be extended to multiple sequences?

Dynamic programming with three sequences

Dynamic programming with two sequences

- Relatively easy to code
- Guarantee to obtain optimal alignment

Can this be extended to multiple sequences?



```
V S N _ S
_ S N A _
_ _ _ A S
```

Dynamic programming complexity

Memory requirements if each sequence has length of n

- 2 sequences: $O(n^2)$
- 3 sequences: $O(n^3)$
- k sequences: $O(n^k)$

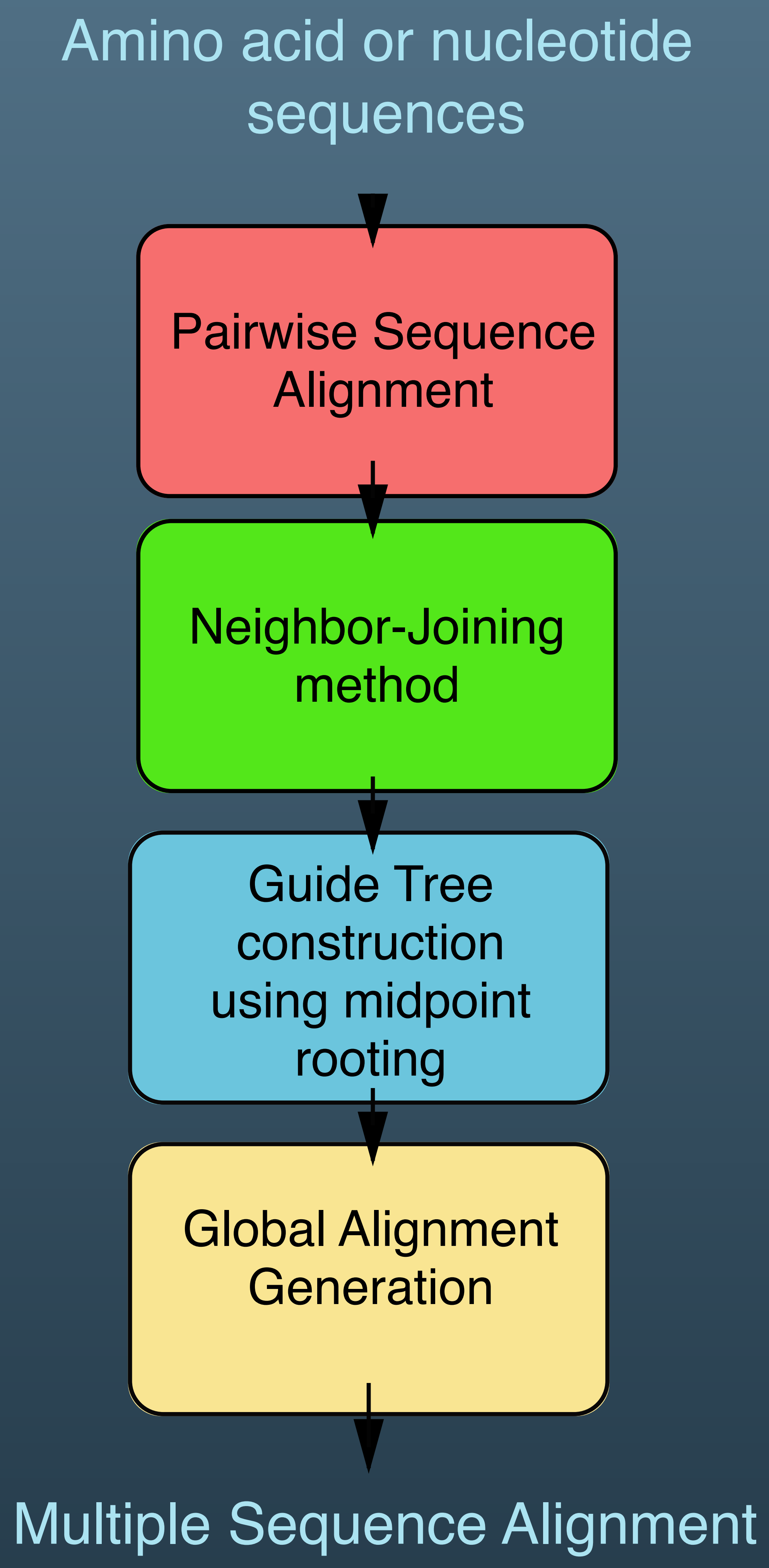
Time problem:

$$O(2^k \prod_{i=1,...,k} |s_i|)$$

If the calculation factor is one nanosecond, then for six sequences of length 100, we'll have a running time of 2^6 x 100^6 x 10^{-9}, that's roughly 64,000 seconds (almost 18 hours). Just add two more sequences, and the running time increases to 2.56 x 10^9 seconds (over 81 years)!

Solution: progressive alignment

- Devised by Feng and Doolittle in 1987 (J Mol Evol. 25(4):351-60)
- A heuristic method and as such is not guaranteed to find the 'optimal' alignment
- Requires n-1+n-2+n-3...n-n+1 pairwise alignments as a starting point
- Align most related sequences
- Add on less related sequences to initial alignment
- Most successful implementation is Clustal


Amino acid or nucleotide sequences
Pairwise Sequence Alignment
Neighbor-Joining method
Guide Tree construction using midpoint rooting
Global Alignment Generation
Multiple Sequence Alignment

Advice on progressive alignment

- Progressive alignment is a mathematical process that is completely independent of biological reality
- Can be a very good estimate
- Can be an impossibly poor estimate
- Requires user input and skills
- Treat cautiously
- Can be improved by eye (usually)
- Often helps to have color-coding
- Depending on the use, the user should be able to make a judgement on those regions that are reliable or not
- For phylogeny reconstruction, only use those positions whose hypothesis of positional homology is certain

Five steps in phylogenetic analysis

1. Finding all homologs
2. Multiple sequence alignment
3. Building a tree
4. Statistical assessment of a tree
5. Viewing a tree and drawing conclusions

Tree building methods

		Computational method	
		Optimality criterion	Clustering algorithm
Data type	Characters	Parsimony	
		Maximum likelihood	
		Bayesian inference	
	Distances	Minimum evolution	UPGMA
		Least squares	Neighbor-joining

Types of data used in phylogenetic inference

Character-based methods: Use the aligned characters, such as DNA or protein sequences, directly during tree inference.

Taxa	Characters
Species A	**ATGGCTATTCTTATAGTACG**
Species B	**ATCGCTAGTCTTATATTACA**
Species C	**TTCACTAGACCTGTGGTCCA**
Species D	**TTGACCAGACCTGTGGTCCG**
Species E	**TTGACCAGTTCTCTAGTTCG**

Distance-based methods: Transform the sequence data into pairwise distances (dissimilarities), and then use the matrix during tree building.

	Taxon A	**Taxon B**	**Taxon C**	**Taxon D**	**Taxon E**
Taxon A		0.20	0.50	0.45	0.40
Taxon B	0.23		0.40	0.55	0.50
Taxon C	0.87	0.59		0.15	0.40
Taxon D	0.73	1.12	0.17		0.25
Taxon E	0.59	0.89	0.61	0.31	

p-distances - the average difference per site (observed sequence difference)

Kimura 2-parameter distance (estimate of the true number of substitutions between taxa)

Types of computational methods

Clustering algorithms: Use pairwise distances.

- These are purely algorithmic methods, in which the algorithm itself defines the tree selection criterion. Tend to be very fast programs that produce singular trees rooted by distance. No objective function to compare to other trees, even if numerous other trees could explain the data equally well.
- Warning: finding a singular tree is not necessarily the same as finding the "true” evolutionary tree.

Optimality approaches

- Use either character or distance data. First, define an optimality criterion (minimum branch lengths, fewest number of events, highest likelihood), and then use a specific algorithm for finding trees with the best value for the objective function. Can identify many equally optimal trees, if such exist.
- Warning: finding an optimal tree is not necessarily the same as finding the "true” tree.

Computational methods for finding optimal trees

Exact algorithms

- Guarantee to find the optimal or "best" tree for the method of choice. Two types used in tree building:
  - Exhaustive search: Evaluates all possible unrooted trees, choosing the one with the best score for the method.
  - Branch-and-bound search: Eliminates the parts of the search tree that only contain suboptimal solutions

Heuristic algorithms

- Approximate or “quick-and-dirty” methods that attempt to find the optimal tree for the method of choice, but cannot guarantee to do so. Heuristic searches often operate by “hill-climbing” methods.

Parsimony methods

Optimality criterion:

- The ‘most-parsimonious’ tree is the one that requires the fewest number of evolutionary events (e.g., nucleotide substitutions, amino acid replacements) to explain the sequences

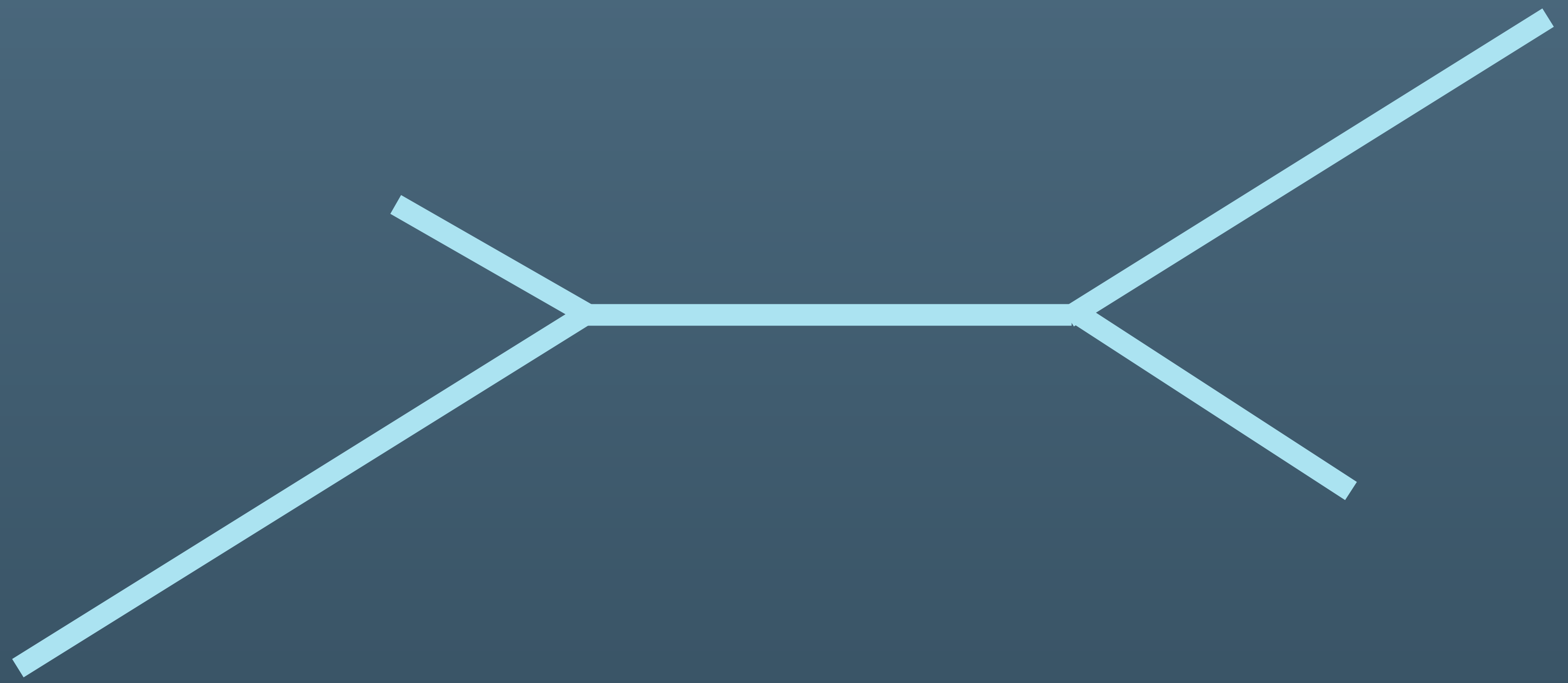
Advantages:

- Are simple, intuitive, and logical (many possible by ‘pencil-and-paper’).
- Can be used on molecular and non-molecular (e.g., morphological) data.
- Can be used for character (can infer the exact substitutions) and rate analysis.
- Can be used to infer the sequences of the extinct (hypothetical) ancestors.

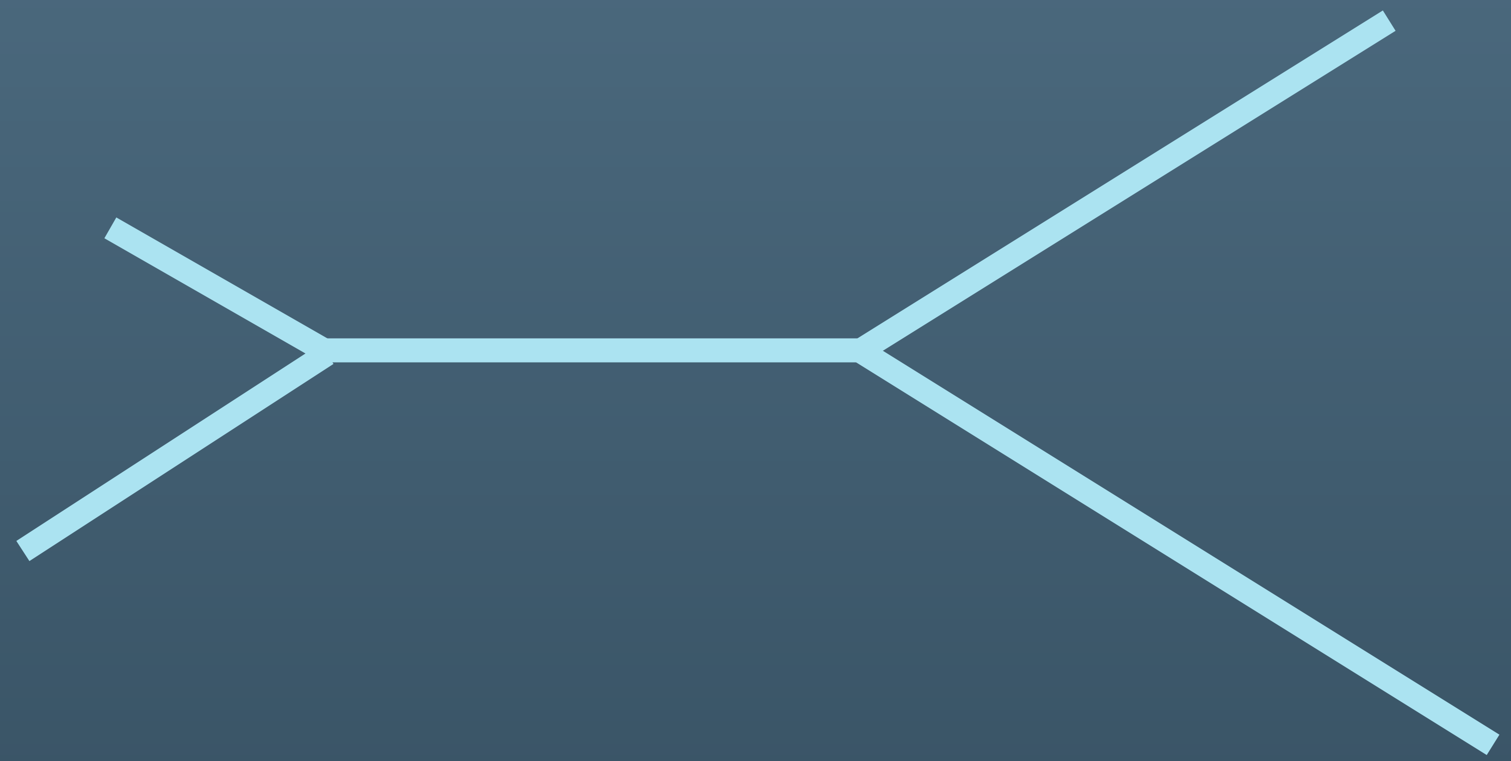
Disadvantages:

- Can be fooled by high levels of homoplasy (‘same’ events).
- Can become positively misleading in the “Felsenstein Zone” (long branch attraction)

Long branch attraction



True tree



Inferred tree

Maximum likelihood (ML) methods

ML methods evaluate phylogenetic hypotheses in terms of the probability that a proposed model of the evolutionary process and the proposed unrooted tree would give rise to the observed data. The tree found to have the highest ML value is considered to be the preferred tree.

Advantages:

- Are inherently statistical and evolutionary model-based.
- Usually the most consistent of the methods available.
- Can be used for character (can infer the exact substitutions) and rate analysis.
- Can be used to infer the sequences of the extinct (hypothetical) ancestors.
- Can help account for branch-length effects in unbalanced trees.
- Can be applied to nucleotide or amino acid sequences, and other types of data.

Disadvantages:

- Are not as simple and intuitive as many other methods.
- Are computationally very intense.
- Like parsimony, can be fooled by high levels of homoplasy.
- Violations of the assumed model can lead to incorrect trees.
- If model is wrong the inferred tree will be likely incorrect

Bayesian inference of phylogeny

Bayes' theorem, named after Reverend Thomas Bayes, describes the probability of an event, based on prior knowledge of conditions that might be related to the event

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Metropolis-Hastings algorithm

1. An initial tree, Ti, is randomly selected
2. A neighbor tree, Tj, is selected from the collection of trees.
3. The ratio, R, of the probabilities (or probability density functions) of Tj and Ti is computed as follows: R = f(Tj)/f(Ti)
4. If R ≥ 1, Tj is accepted as the current tree
5. If R < 1, Tj is accepted as the current tree with probability R, otherwise Ti is kept
6. At this point the process is repeated from Step 2 N times.

Bayesian inference of phylogeny

Bayesian inference of phylogeny uses a likelihood function to create a quantity called the posterior probability of trees using a model of evolution, based on some prior probabilities, producing the most likely phylogenetic tree for the given data.

- Start with best guess of a tree (prior probability)
- Simulation of trees using Markov Chain Monte Carlo (MCMC,)
- Keep all the best trees
- Posterior tree with probabilities

- Pitfalls and controversies
  - posterior probabilities lead to overconfidence in the results
  - controversy of using prior probabilities
  - model choice - an oversimplified model might give higher posterior probabilities

Minimum evolution (ME) methods

The tree(s) with the shortest sum of the branch lengths (or overall tree length) is chosen as the best tree

Advantages:

- Can be used on indirectly-measured distances (immunological, hybridization).
- Distances can be ‘corrected’ for unseen events.
- Usually faster than character-based methods.
- Can be used for some rate analyses.
- Has an objective function (as compared to clustering methods).

Disadvantages:

- Information lost when characters transformed to distances.
- Cannot be used for character analysis.
- Slower than clustering methods.

Clustering methods

Advantages:

- Can be used on indirectly-measured distances (immunological, hybridization).
- Distances can be ‘corrected’ for unseen events.
- The fastest of the methods available.
- Can therefore analyze very large datasets quickly.

Disadvantages:

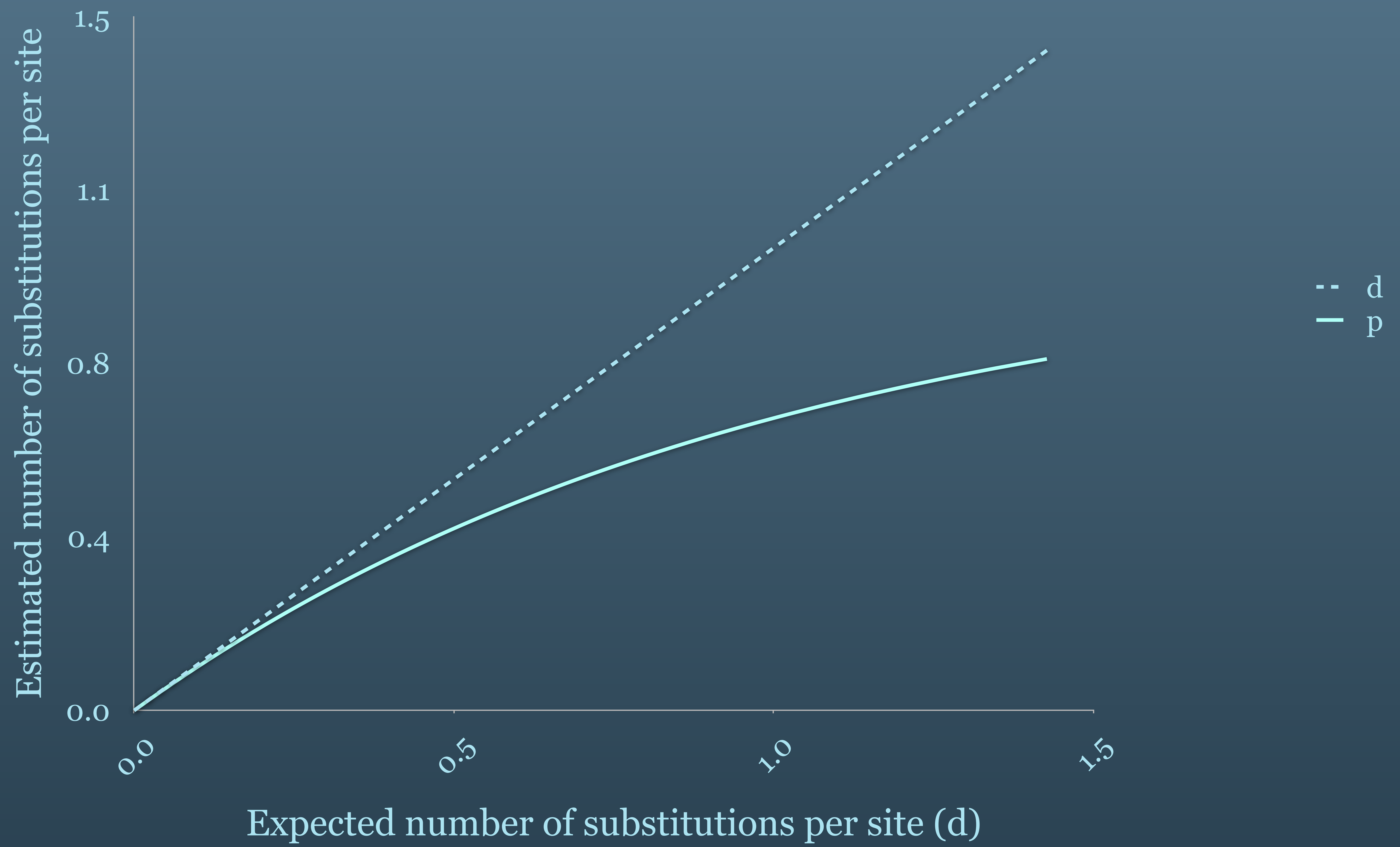
- Similarity and relationship are not necessarily the same thing, so clustering by similarity does not necessarily give an evolutionary tree.
- Cannot be used for character analysis.
- Have no explicit optimization criteria, so one cannot even know if the program worked properly to find the correct tree for the method.
- No optimality criterion

Distance methods

- Based on precomputed pairwise distances between sequences according to the scoring scheme; the actual sequence is discarded once a distance matrix is computed
- Distance score is based on number of observed differences between two aligned sequences
- Pairwise alignment identity scores can be converted directly to distance scores; more sophisticated models contain heuristics to adjust for predicted number of multiple events at each site

- Simplest distance measure = Hamming distance, number of changes (n) per unit sequence (N) = n/N; gaps can be ignored or treated as substitutions
- Assumes every change occurs only once, there are no duplicate changes at each site
- Can result in a zero or even negative branch length if that assumption is incorrect
- Alternate distance models, e.g. probabilistic models like Jukes-Cantor, Kimura, can be used to estimate probabilities that multiple changes have occurred at a site

Distance methods



Unweighted pair group method with arithmetic mean

- UPGMA (Unweighted pair group method with arithmetic mean) is a hierarchical clustering method that assumes a constant molecular clock (rate of evolution) along all branches of the tree.
- Two closest sequences are clustered first, then next two closest, etc. A rooted tree is produced.
- UPGMA assumes a molecular clock and results in a fixed (and error-prone) rooted tree topology. UPGMA methods are not recommended unless evolutionary rates can be assumed to be consistent in all branches in an entire protein group.

Algorithm

- Given a matrix of pairwise distances, find the clusters (taxa) i and j such that d_{ij} is the minimum value in the table
- Define the depth of the branching between i and j (l_{ij}) to be $d_{ij}/2$
- If i and j were the last two clusters, the tree is complete. Otherwise, create a new cluster called u.
- Define the distance from u to each other cluster to be an average of the distances d_{ki} and d_{kj}.
- Go back to step 1 with one less cluster; cluster i and j have been eliminated, and cluster u has been added.

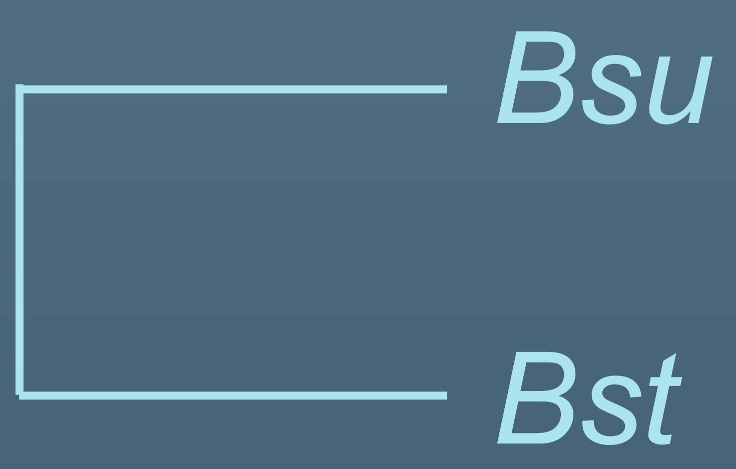
UPGMA example

	Bsu	Bst	Lvi	Amo	Mlu
Bacillus subtilis	x	0.1715	0.2147	0.3091	0.2326
Bacillus stearothermophilus		x	0.2991	0.3399	0.2058
Lactobacillus viridescens			x	0.2795	0.3943
Acholeplasma modicum				x	0.4289
Micrococcus luteus					x

Create a cluster between two taxa with the minimum distance - Bsu and Bst in the example above. Recalculate distances with Bsu-Bst cluster as a new operational unit.

Data from Olsen (1988) Phylogenetic analysis using ribosomal RNA. Meth. Enzymol. 164: 793-838.

UPGMA example

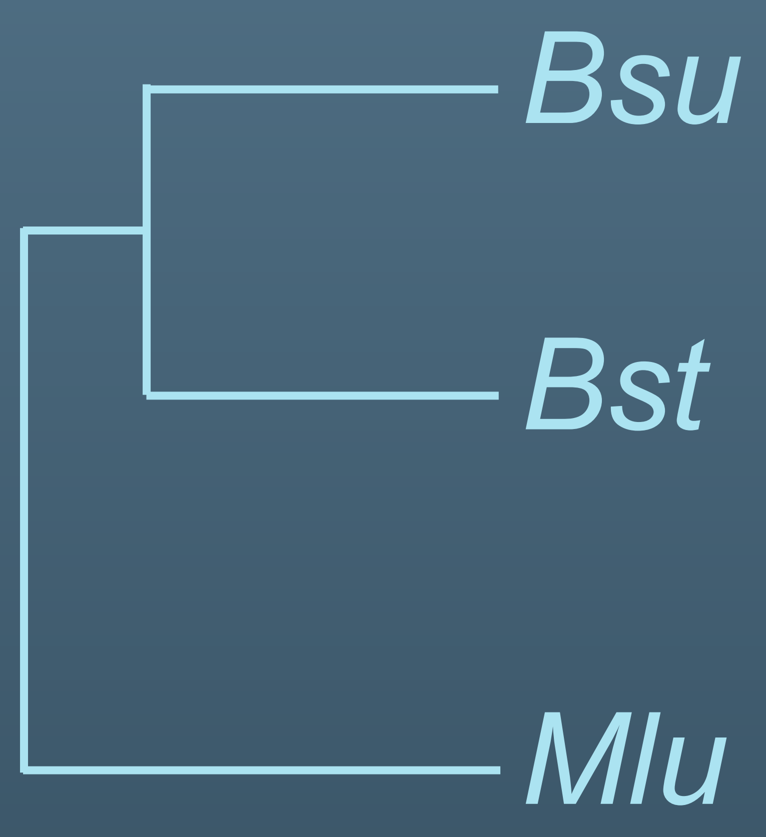


	Bsu	Bst	Lvi	Amo	Mlu
Bacillus subtilis	x	0.1715	0.2147	0.3091	0.2326
Bacillus stearothermophilus		x	0.2991	0.3399	0.2058
Lactobacillus viridescens			x	0.2795	0.3943
Acholeplasma modicum				x	0.4289
Micrococcus luteus					x

Create a cluster between two taxa with the minimum distance - Bsu and Bst in the example above. Recalculate distances with Bsu-Bst cluster as a new operational unit.

Data from Olsen (1988) Phylogenetic analysis using ribosomal RNA. Meth. Enzymol. 164: 793-838.

UPGMA example

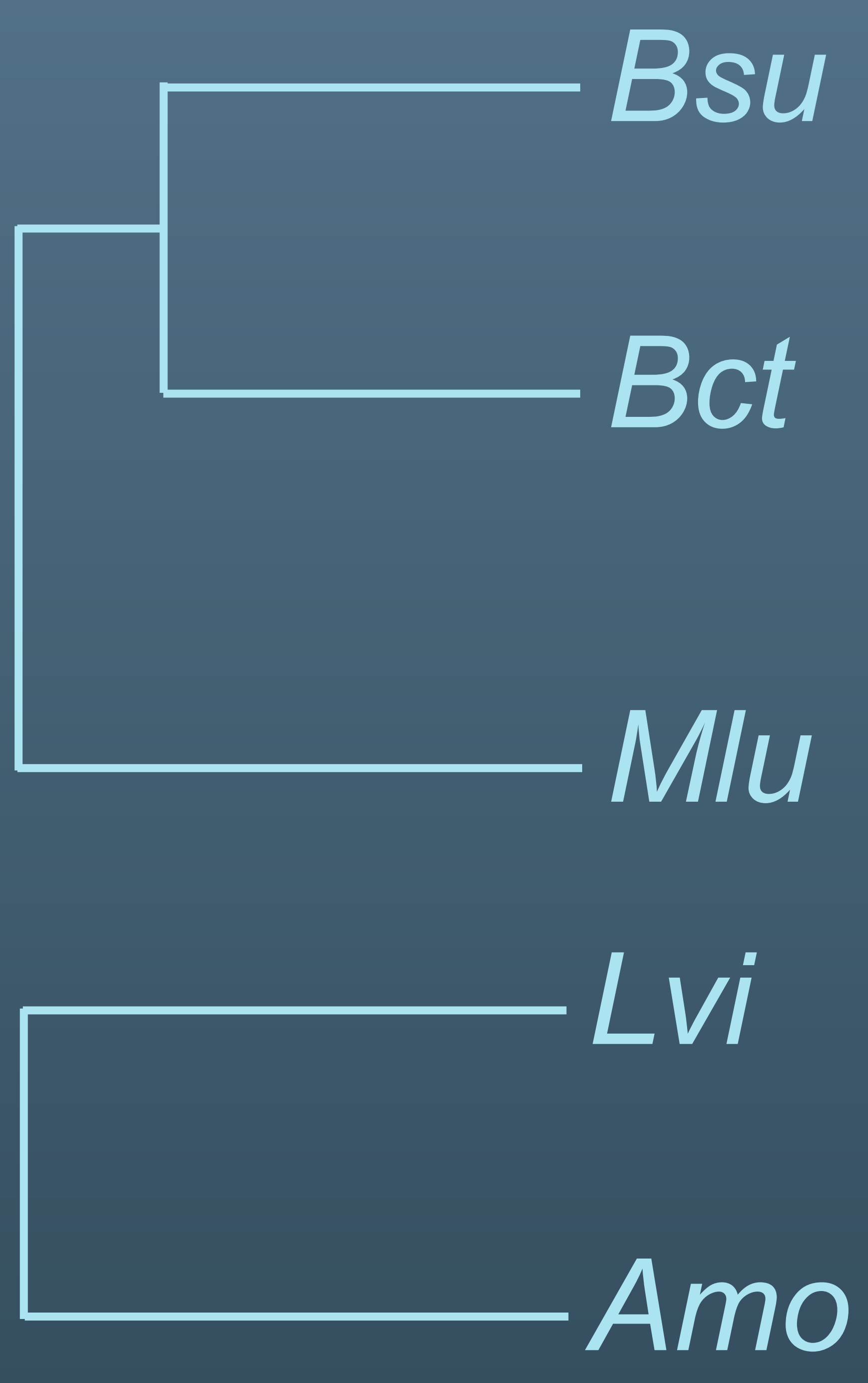


	Bsu-Bst	Lvi	Amo	Mlu
Bsu-Bst	x	0.2569	0.3245	0.2192
Lactobacillus viridescens		x	0.2795	0.3943
Acholeplasma modicum			x	0.4289
Micrococcus luteus				x

Create a cluster between two taxa with the minimum distance - Bsu-Bst and Mlu in the example above. Recalculate distances with Bsu-Bst-Mlu cluster as a new operational unit.

Data from Olsen (1988) Phylogenetic analysis using ribosomal RNA. Meth. Enzymol. 164: 793-838.

UPGMA example



	Bsu-Bst-Mlu	Lvi	Amo
Bsu-Bst-Mlu	x	0.3027	0.3593
Lactobacillus viridescens		x	0.2795
Acholeplasma modicum			x

Create a cluster between two taxa with the minimum distance - Lvi and Amo in the example above. Recalculate distances with Lvi-Amo cluster as a new operational unit.

Data from Olsen (1988) Phylogenetic analysis using ribosomal RNA. Meth. Enzymol. 164: 793-838.

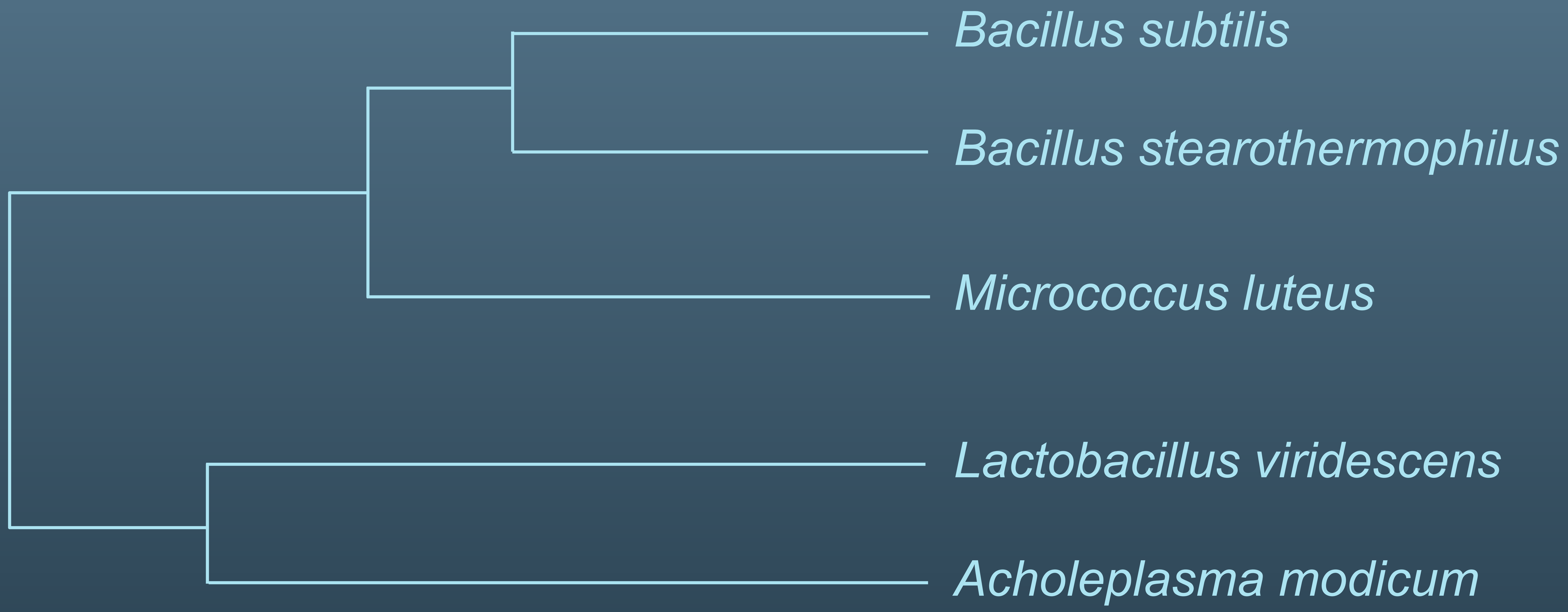
UPGMA example

	Bsu-Bst-Mlu	Lvi
Bsu-Bst-Mlu	x	0.3310
Lvi-Amo		x

Create the last cluster. Draw the tree

Data from Olsen (1988) Phylogenetic analysis using ribosomal RNA. Meth. Enzymol. 164: 793-838.

UPGMA example



Five steps in phylogenetic analysis

1. Finding all homologs
2. Multiple sequence alignment
3. Building a tree
4. Statistical assessment of a tree
5. Viewing a tree and drawing conclusions

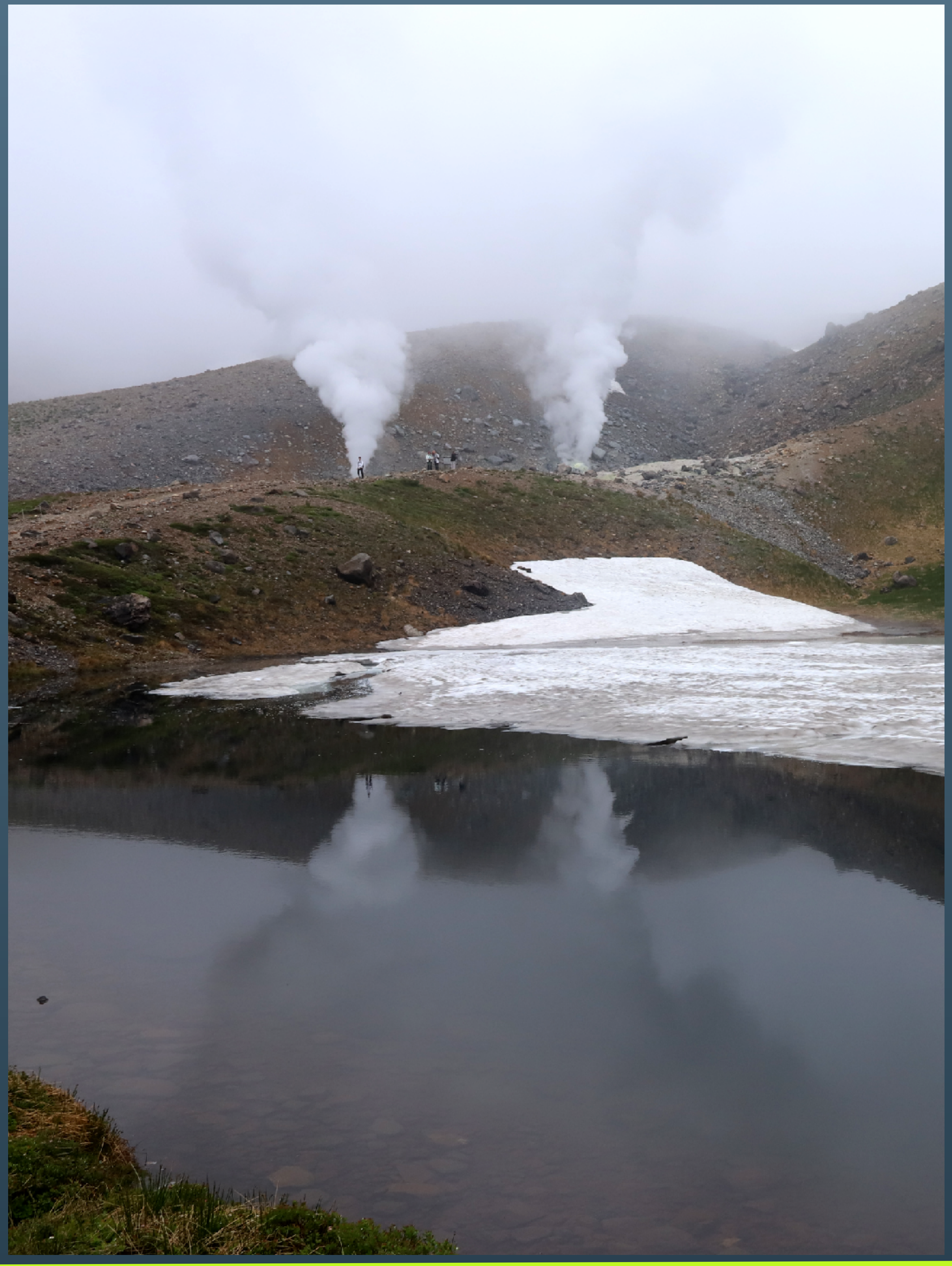
Statistical assessment of a tree

Tests of one overall hypothesis (tree) against other hypotheses

- Wilson's "winning sites" test
- Templeton's test
- Kishino-Hasegawa ML test

Tests of strength of support for lineages within trees

- Bootstrap
- Jack-knife
- Decay index



Bootstrapping

- Random sampling of columns in the original alignment to create a new alignment
- Building a tree based on the new alignment
- Repeat step 1 and 2 many times (usually 1000 times)
- Calculate how many times a given topology appears in all replicas

```
ATGGCTATTCTTATAGTACG
ATCGCTAGTCTTATATTACA
TTCACTAGACCTGTGGTCCA
TTGACCAGACCTGTGGTCCG
TTGACCAGTTCTCTAGTTCG
```

Original alignment

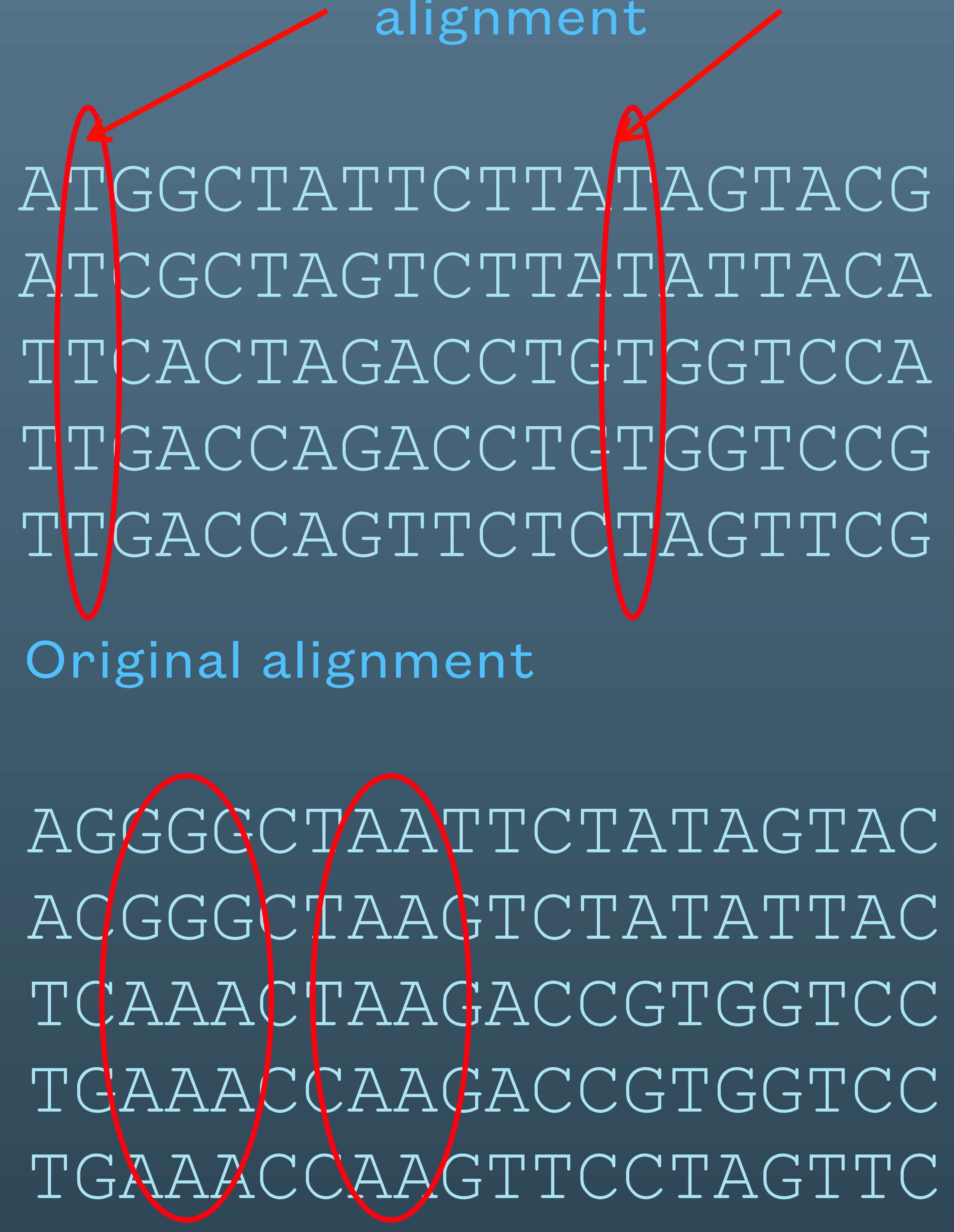
```
AGGGGCTAATTCTATAGTAC
ACGGGCTAAGTCTATATTAC
TCAAACTAAGACCGTGGTCC
TGAAACCAAGACCGTGGTCC
TGAAACCAAGTTCCTAGTTC
```

Resampled alignment

Bootstrapping

- Random sampling of columns in the original alignment to create a new alignment
- Building a tree based on the new alignment
- Repeat step 1 and 2 many times (usually 1000 times)
- Calculate how many times a given topology appears in all replicas

Some columns don’t get into the new alignment



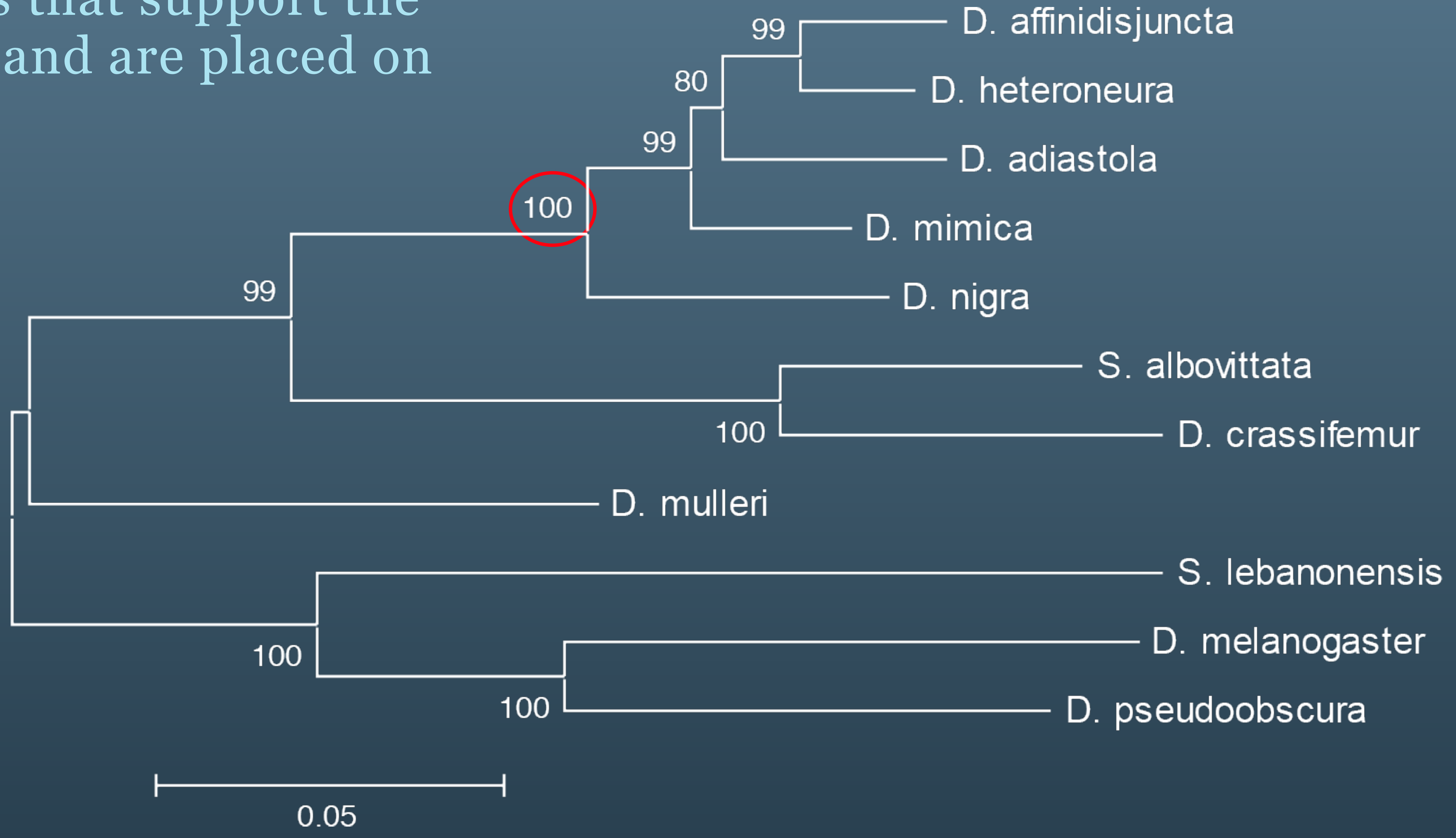
```
ATGGCTATTCTTATAGTACG
ATCGCTAGTCTTATATTACA
TTCACTAGACCTGTGGTCCA
TTGACCAGACCTGTGGTCCG
TTGACCAGTTCTCTAGTTCG
```

Original alignment

```
AGGGGCTAATTCTATAGTAC
ACGGGCTAAGTCTATATTAC
TCAAACTAAGACCGTGGTCC
TGAAACCAAGACCGTGGTCC
TGAAACCAAGTTCCTAGTTC
```

Some columns appear more than ones

Bootstrap values are usually presented as a fraction or percentage of resampled trees that support the particular branch and are placed on that branch



Difficulties with phylogenetic inference

- Horizontal or lateral transfer of genetic material (for instance through viruses) makes it difficult to determine phylogenetic origin of some evolutionary events
- Rearrangements of genetic material can lead to false conclusions
- Duplicated genes can evolve along separate pathways, leading to different functions

True versus inferred tree

- The sequence of speciation events that has led to formation of any groups of OTUs is historically unique. Consequently, only one of all possible phylogenetic trees represents the true evolutionary history, which is called true tree.
- A tree obtained from the certain data using a certain method is called an inferred tree. An inferred tree ma or may not be identical to the true tree.

Some practical advice

- Each method has its own strengths
- Use multiple methods for cross-validation
- In some cases, none of the method gives the correct phylogeny
- Selecting a high-quality input data set is the most critical step in developing a phylogeny
- The order of the input set can affect results. Good phylogenetics software provides tools for randomizing input sets
- Check for consistency by applying more than one method (NJ, MP, ML) to the same data set
- If you obtain an unreliable tree
    - GET MORE DATA!

Selected software

- MEGA: Molecular Evolutionary Genetics Analysis
  - https://www.megasoftware.net/
- RAxML (Randomized Axelerated Maximum Likelihood)
  - https://sco.h-its.org/exelixis/web/software/raxml
- MrBayes (Bayesian inference of phylogeny)
  - http://nbisweden.github.io/MrBayes/manual.html
- PHYLIP (the PHYLogeny Inference Package)
  - https://evolution.genetics.washington.edu/phylip/phylipweb.html
- Analysis of Phylogenetics and Evolution in R
  - http://ape-package.ird.fr

BIOINFORMATICS CREED

Remember about biology

Do not trust the data

Use comparative approach

Use statistics

Know the limits

Remember about biology!!!

UKM Blutspende
In Münster für Münster

Das liegt dir im Blut!

In der UKM Blutspende kommen alle Blutspenden unmittelbar den Patientinnen und Patienten am UKM zugute – in Münster für Münster!

Für deine Spende erhältst du eine Aufwandsentschädigung!

Domagkstraße 11, 48149 Münster
+49 251 83-58000
www.ukm-blutspende.de