BIO-463
Genomics and bioinformatics

Lecture 4:
Molecular evolution, phylogeny and homology

Prof. Anne-Florence Bitbol

EPFL

Schedule of this class

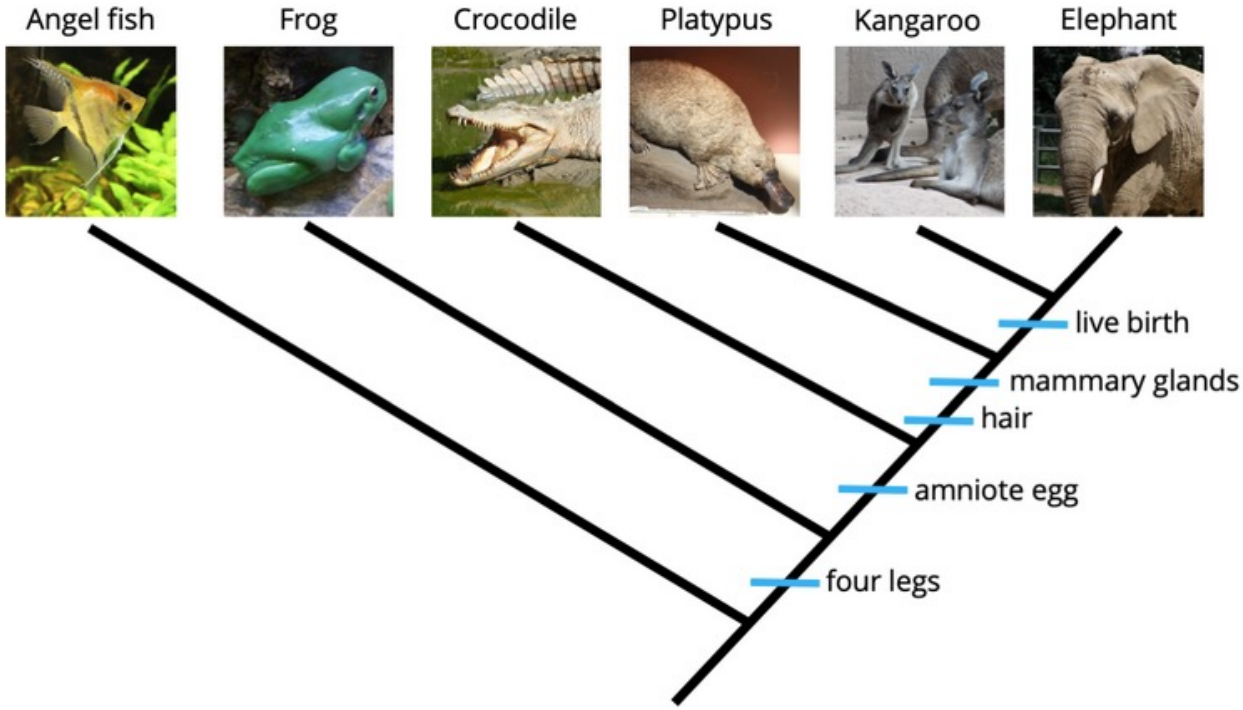
date	week	lectures	exercises	teacher
18 feb	1	Structural genomics	R exercises	**Jacques Rougemont**
25 feb	2			
4 mar	3			
11 mar	4	Population genetics	R exercises	**Anne-Florence Bitbol**
18 mar	5			
25 mar	6		Assignment 1: 25%	
1 apr	7	Gene expression	R exercises	**Raphaëlle Luisier**
8 apr	8			
15 apr	9			
22 apr		holidays		
29 apr	10	Gene expression	Assignment 2: 25%	**Raphaëlle Luisier**
6 may	11	Regulation, chromatin	Mini-projects	**Jacques Rougemont**
13 may	12			
20 may	13			
27 may	14			
			Report: 50%	

Schedule of this class

Lecture 1: Feb 18
Lecture 2: Feb 25
Lecture 3: March 4
Lecture 4: March 11
Lecture 5: March 18 – Assignment 1 available on March 20
Lecture 6: March 25 – Problem class devoted to **assignment 1**; deadline on **March 28**
Lecture 7: April 1
Lecture 8: April 8
Lecture 9: April 15 – Assignment 2 available on April 18
Lecture 10: April 29 – Problem class devoted to **assignment 2**; deadline on **May 2**
Lecture 11: May 6 – Mini-projects available on April 28; choose yours by May 6
Lecture 12: May 13
Lecture 13: May 20
Lecture 14: May 27 – Mini-project deadline on **May 30**

Motivation

- **Studying the evolution of species and genes**
  - Traditionally: based on traits such as physical or morphological features



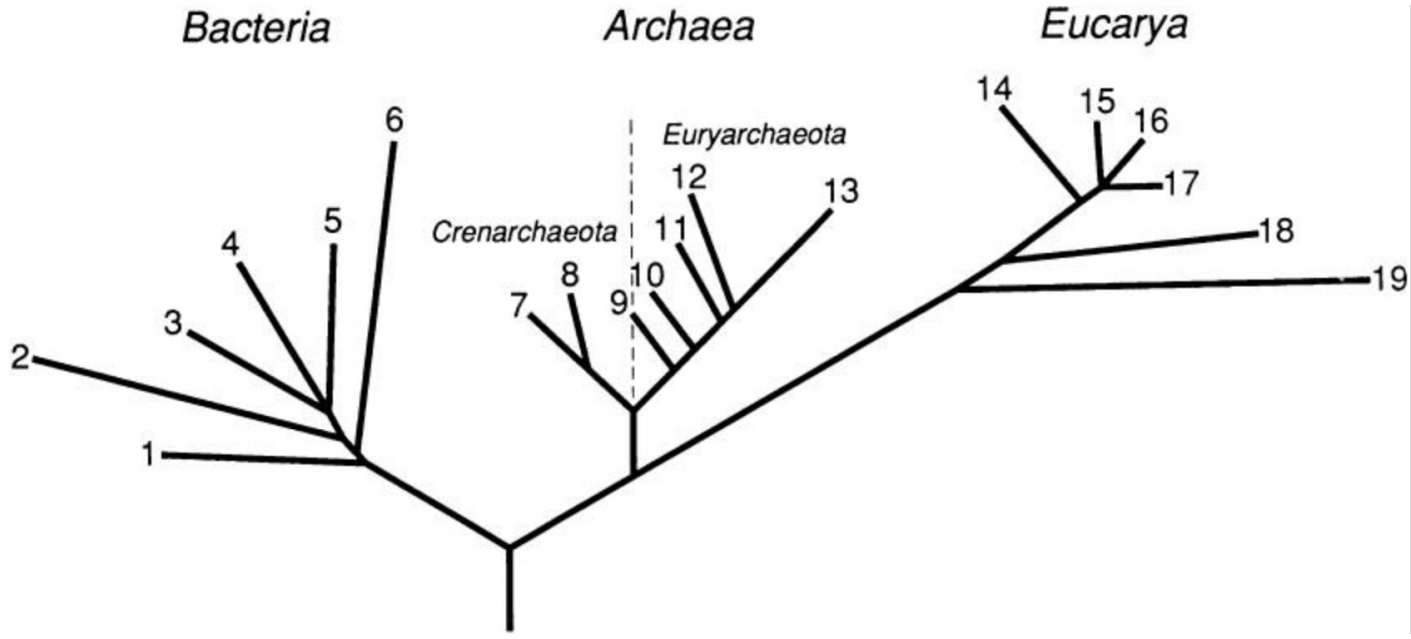
Group animals using shared characters

(assume that traits only appeared once)

Digital atlas of ancient life

Motivation

- **Studying the evolution of species and genes**
  - Traditionally: based on traits such as physical or morphological features
  - More modern: based on molecular sequence data



3 major domains of life

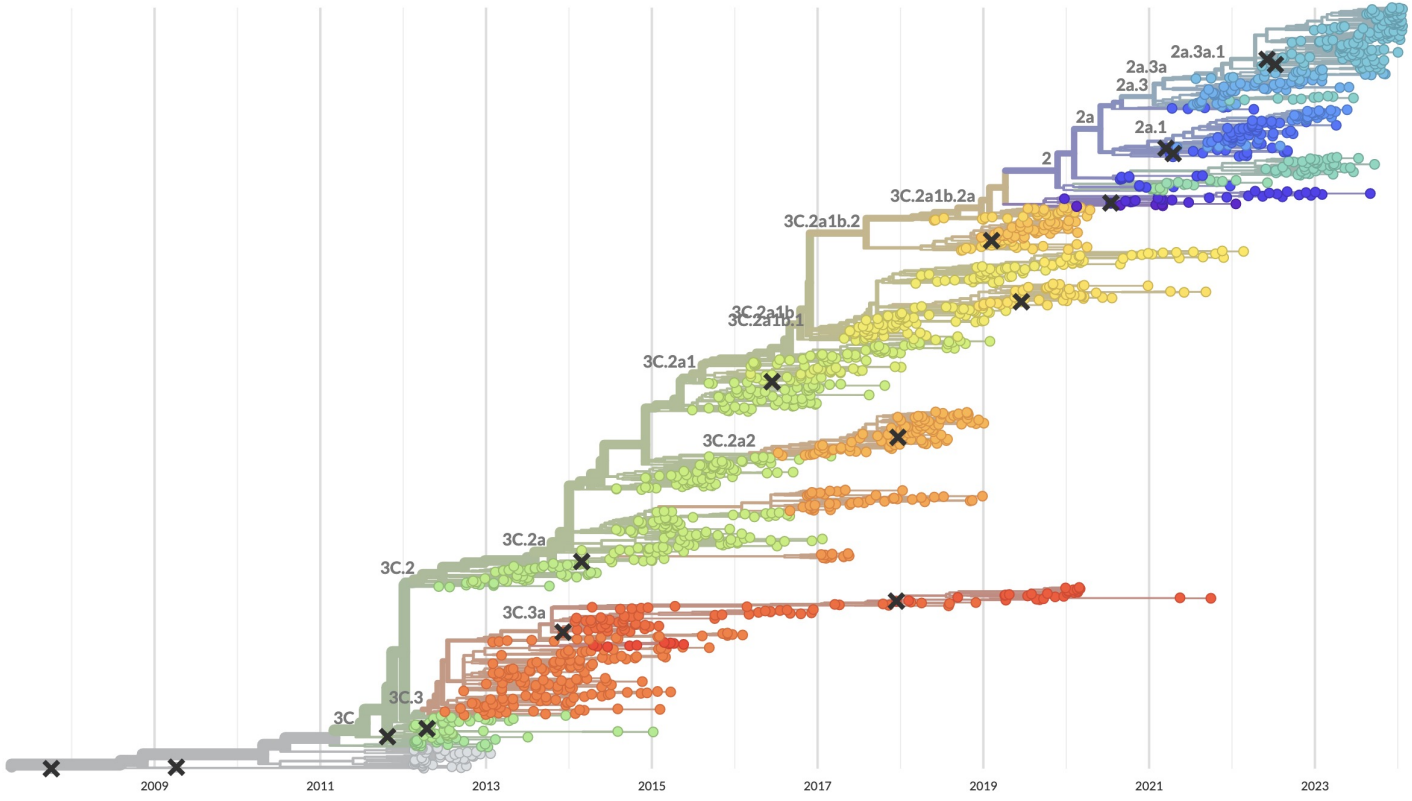
Inferred from analyzing (rRNA) sequences

Woese et al 1990 – redrawn in Pace et al 2012

Motivation

- **Studying the evolution of species and genes**

At shorter timescales: understanding the phylogeny of a virus can help design vaccines



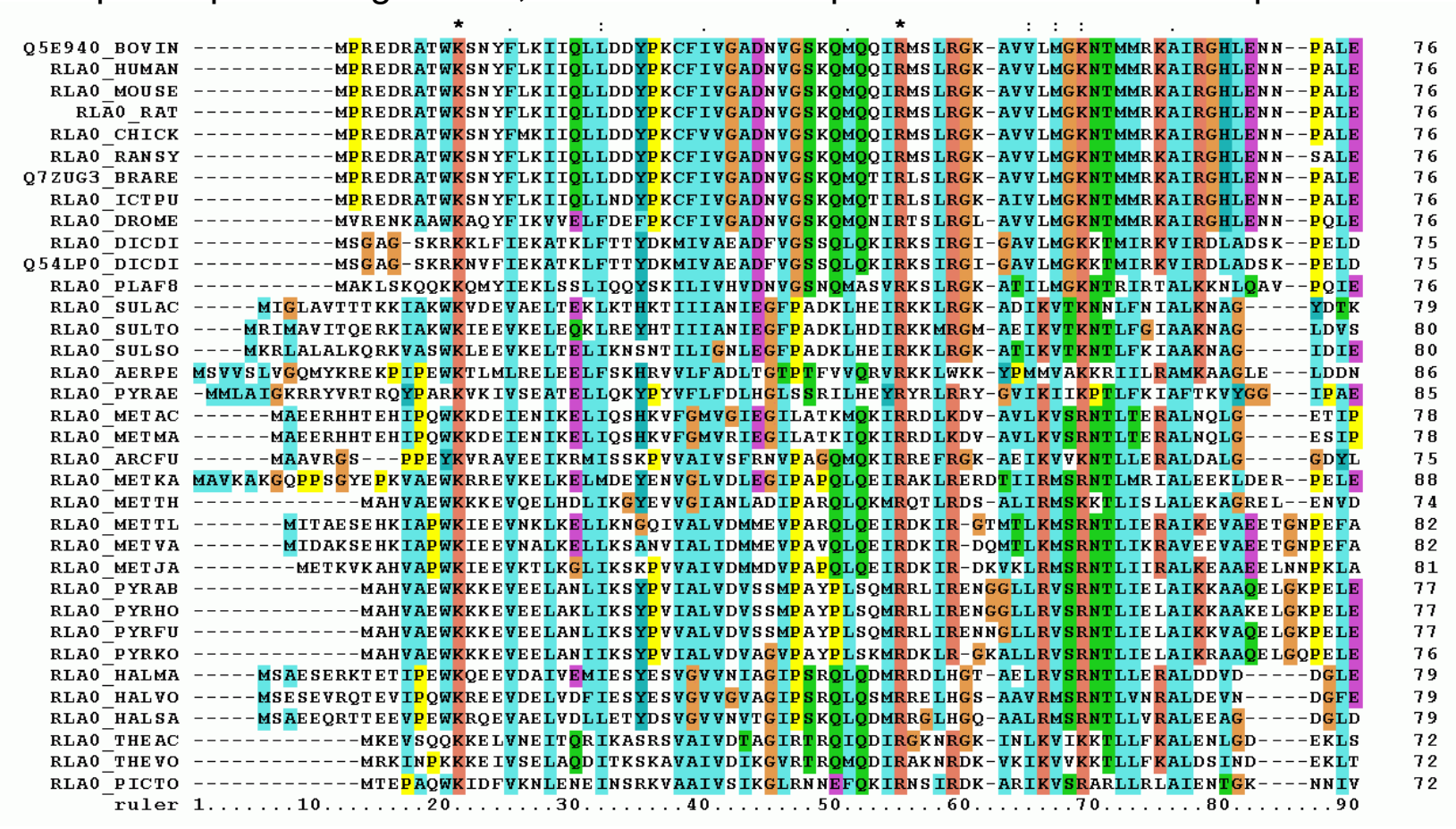
Influenza A/H3N2 evolution (sequence coding for hemagglutinin protein)

https://nextstrain.org/flu/seasonal/h3n2/ha/12y

Sequence data and phylogeny

- **Starting point: biological sequence data**

Multiple sequence alignments, of amino-acid sequences or nucleotide sequences



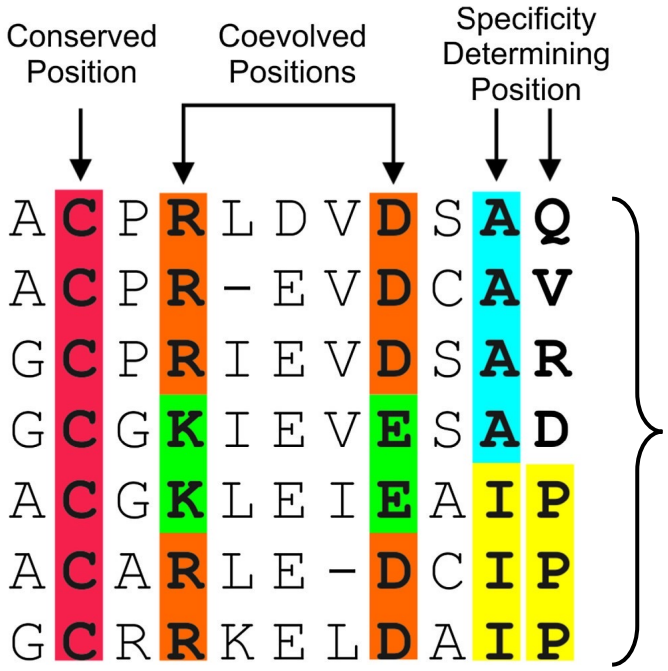
Acidic ribosomal protein P0 (first 90 positions) from several organisms

Row = sequence
Column = site (given position in 3D structure)

Colors = level of conservation

Sequence data and phylogeny

- **Starting point: biological sequence data**

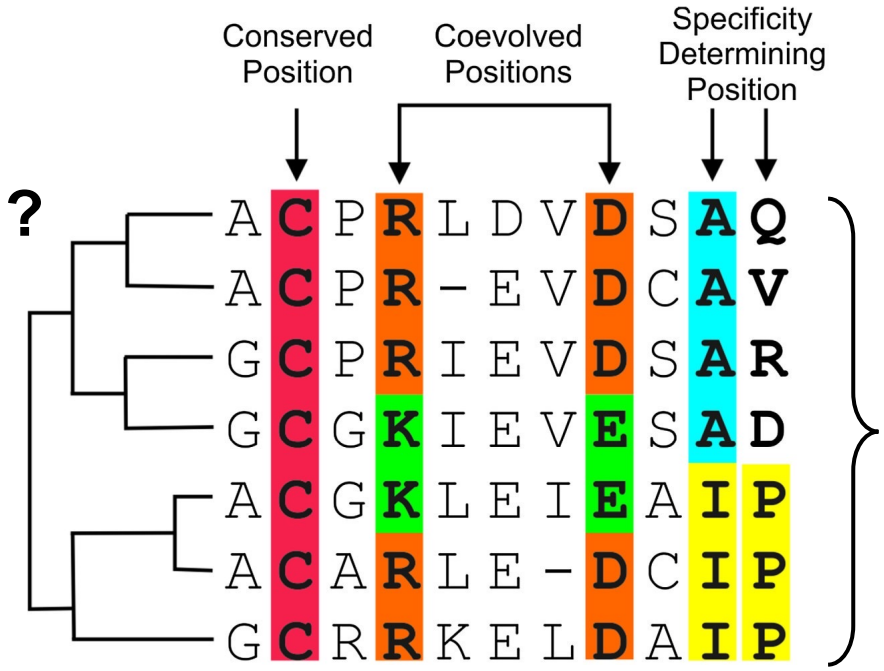


Multiple sequence alignment of homologous sequences:
same ancestry,
similar function,
similar 3D structure

Teppa et al 2012

Sequence data and phylogeny

- **Starting point: biological sequence data**



Multiple sequence alignment of homologous sequences:
same ancestry,
similar function,
similar 3D structure

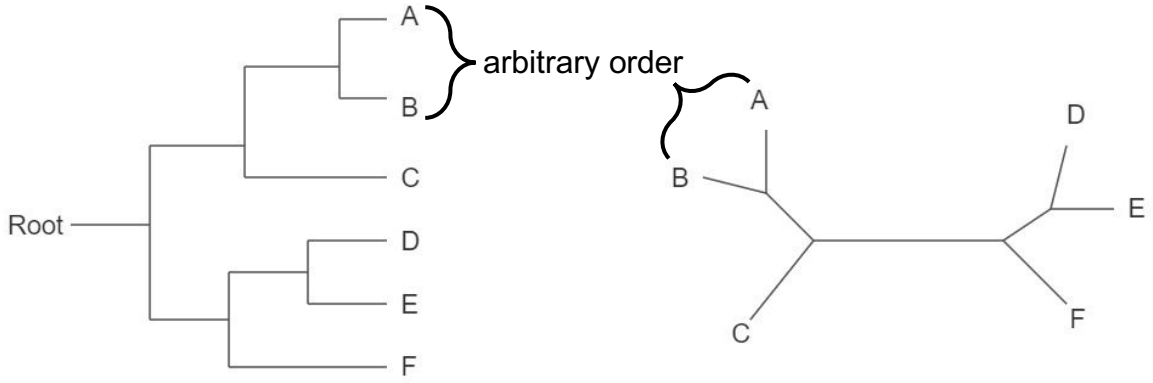
Teppa et al 2012

Goal: infer evolutionary tree (phylogenetic tree) from sequence data
Simplification: ignore coevolution – assume each site (column) evolves independently

Sequence data and phylogeny

Structure of a phylogenetic tree

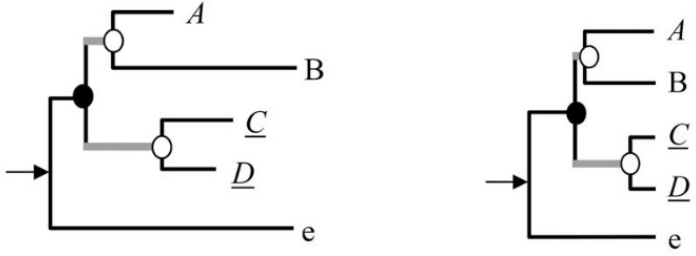
- Assume only bifurcations (2 branches, not more, from 1 node)
- Rooted versus unrooted trees:



Leaf nodes = observed species/sequences
Internal nodes = hypothetical ancestors

Root = hypothetical common ancestor of all leaves (difficult to know where it is)

- Important features of a tree:
  - Branching events → tree topology
  - Branch length may represent nothing or evolutionary distances (phylogram) or time (chronogram)



Sequence data and phylogeny

Inferring a phylogenetic tree from sequence data

- **Distance-based methods:** start from evolutionary distances between sequences, and construct a tree based on them, using clustering algorithms
  - Unweighted Paired Group Mean Arithmetic (UPGMA)
  - Neighbor Joining (NJ)

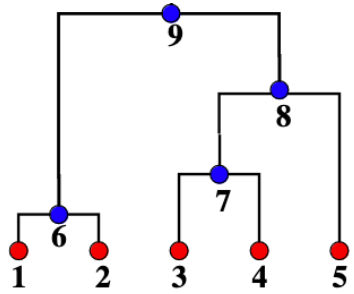
- **Character-based methods:** use a score that quantifies how well a tree describes the raw data, and find the tree with the best score. These methods directly aim to fit the states (characters, i.e. amino acids or nucleotides) observed at each site in each sequence to a tree
  - Maximum parsimony
  - Maximum likelihood
  - Bayesian (maximum a posteriori)

- Difficulty: **many** possible trees!

- Phylogenetic tree construction has been shown to be NP-complete (no solution in polynomial time) for many models

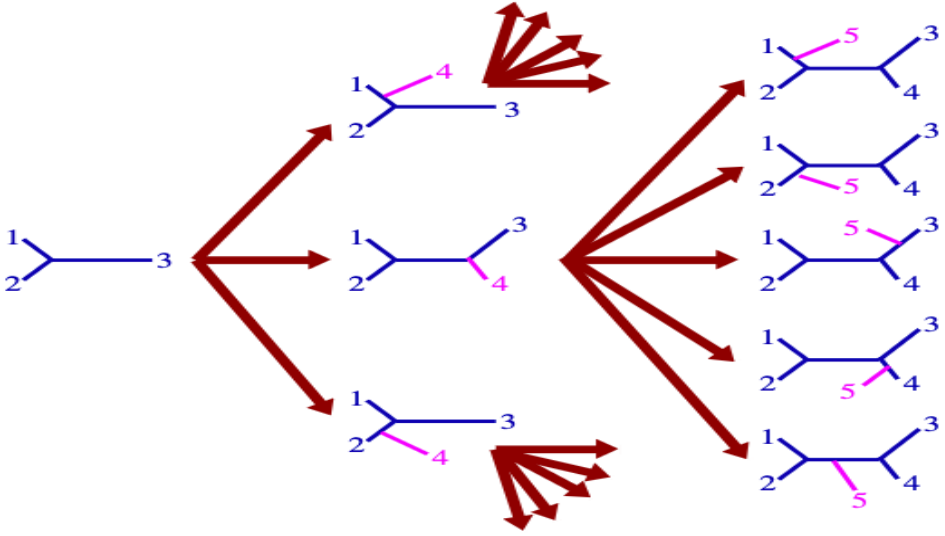
Sequence data and phylogeny

Counting possible trees with n sequences (n leaf nodes)



- Move from leaves to root → 2 edges join at each internal node (bifurcating tree)
  → at each internal node, the number of edges decreases by one
  → there are n-1 internal nodes
- Thus, there are 2n-1 total nodes (and 2n-2 total edges) in the rooted tree
- And there are 2n-2 nodes and 2n-3 edges in the unrooted tree

- Add an extra sequence: extra edge with new (n+1)th leaf can be added at any edge
  → 2n-3 times more unrooted trees with n+1 leaves than with n leaves
- Thus, 1x3x5x…x(2n-5) unrooted trees with n leaves [1 with 3, 3 with 4…] – notation: (2n-5)!!



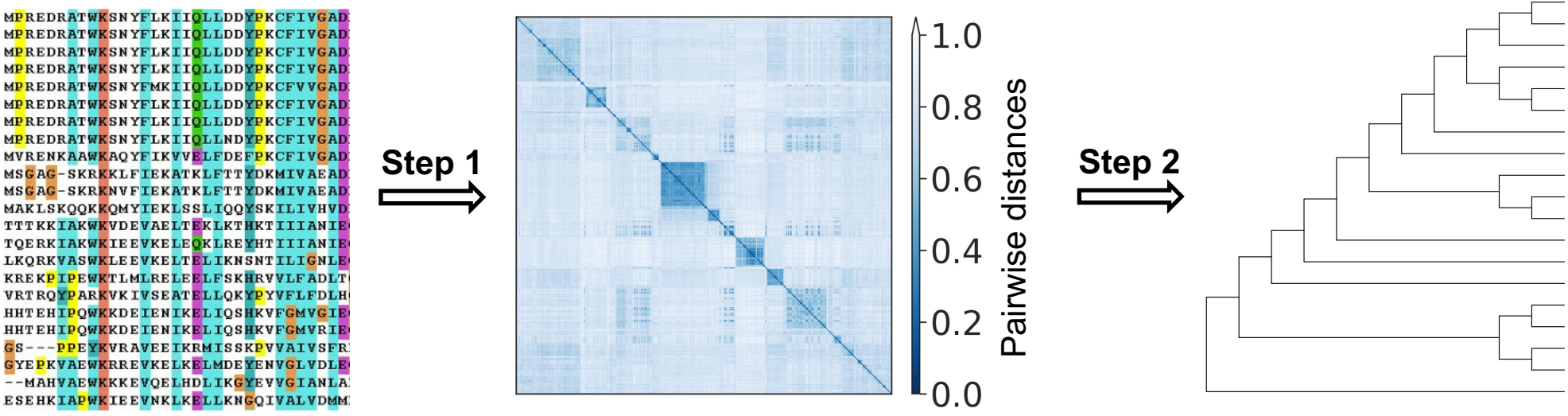
(2n-5)!! = 1x3x5x…x(2n-5) unrooted trees
This is already $>10^{20}$ for n=20…

Volker Roth, U. of Basel

Phylogeny inference: distance-based methods

General method

- **Step 1:** MSA → pairwise distances between sequences
- **Step 2:** pairwise distances between sequences → tree matching observed distances



- **Limitation:** restricting to pairwise distances leaves out some information contained in the raw data (sequences). Two different pairs of sequences can have the same distance
- However, a lot of evolutionary information is contained in distances

Phylogeny inference: distance-based methods

Step 1: Determining pairwise distances between sequences

- **Ideally:** evolutionary distance – number of mutations that actually occurred between two sequences (= sum of the number of mutations that occurred from their last common ancestor to each of them)

- **Hamming distance:** count sites that differ between two sequences (can then divide by number of sites to obtain number between 0 and 1)

AGATC
AGGCA

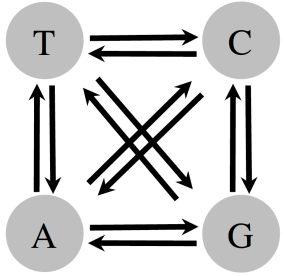
Between these two nucleotide sequences, Hamming distance=3/5

Limitation: evolutionary distance is underestimated due to multiple substitutions at the same site
Example: if A → T → A, we observe no difference but 2 mutations occurred

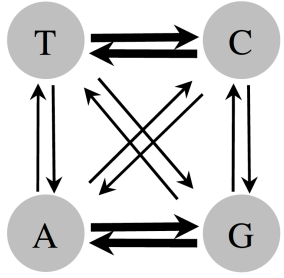
- **Jukes-Cantor distance:** simplest correction that takes into account multiple substitutions

Evolutionary model where:
- each site evolves independently of others
- all substitutions are equally likely: rate λ

(Jukes-Cantor 1969)



Remark: more sophisticated models exist, e.g. transitions more likely than transversions (Kimura 1980)



Yang 2006

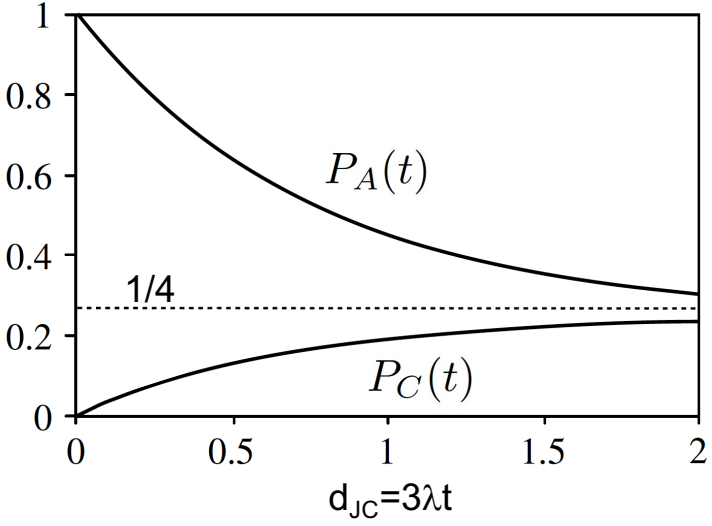
Phylogeny inference: distance-based methods

- **Step 1: Determining pairwise distances between sequences**
  - **Jukes-Cantor distance (typed notes):** simplest correction that takes into account multiple substitutions

Evolutionary model where:
- each site evolves independently of others
- all substitutions are equally likely: rate λ

$$\frac{dP_A}{dt}(t) = \lambda\left[1 - 4P_A(t)\right]$$

If the initial state at t=0 is A, then: $P_A(t) = \frac{3}{4}e^{-4\lambda t} + \frac{1}{4}$ and $P_C(t) = P_G(t) = P_T(t) = -\frac{1}{4}e^{-4\lambda t} + \frac{1}{4}$



Yang 2006

Under the Jukes-Cantor model, the evolutionary distance d_{JC} between two proteins can be estimated from their Hamming distance d_H as:

$$d_{JC} = 3\lambda t = -\frac{3}{4}\log\left[1 - \frac{4}{3}d_H\right]$$ (natural logarithm)

Examples: d_H : 0.1, 0.3, 0.5, 0.7
d_{JC} : 0.11, 0.38, 0.82, 2.03

Remarks: For small d_H, $d_{JC} \approx d_H$. If $d_H \to 3/4$, $d_{JC} \to \infty$

Phylogeny inference: distance-based methods

- **Step 2: Building a tree from a matrix of pairwise distances between sequences**
  - **Goal:** Find a phylogenetic tree that agrees with the empirical pairwise distances
    Distance d_{ij}^T between leaves i and j along the tree T should match the empirical d_{ij}, as well as possible
  - **Least-square approach:** find the tree T that minimizes $\sum_{i=1}^{n}\sum_{j\neq i}(d_{ij}-d_{ij}^T)^2$ (n: number of leaves)
  - Difficult because there are many trees; NP-complete. But efficient (polynomial) approximate algorithms:
    - Unweighted Paired Group Mean Arithmetic (UPGMA)
    - Neighbor Joining (NJ)
  - If all leaves have the same distance from the root (all species evolve at the same rate – constant molecular clock – "ultrametric tree"), then UPGMA will find the correct topology
  - If distances are additive (less strong than ultrametric, rates can differ across species), there exists a tree T such that $d_{ij}^T=d_{ij}$, and NJ works well

Four point condition: For every set of four leaves i, j, k and l, two of the distances $d_{ij}+d_{kl}, d_{ik}+d_{jl}$ and $d_{il}+d_{jk}$ must be equal and larger than the third. For instance
$d_{ij}+d_{kl} \leq d_{ik}+d_{jl} = d_{il}+d_{jk}$

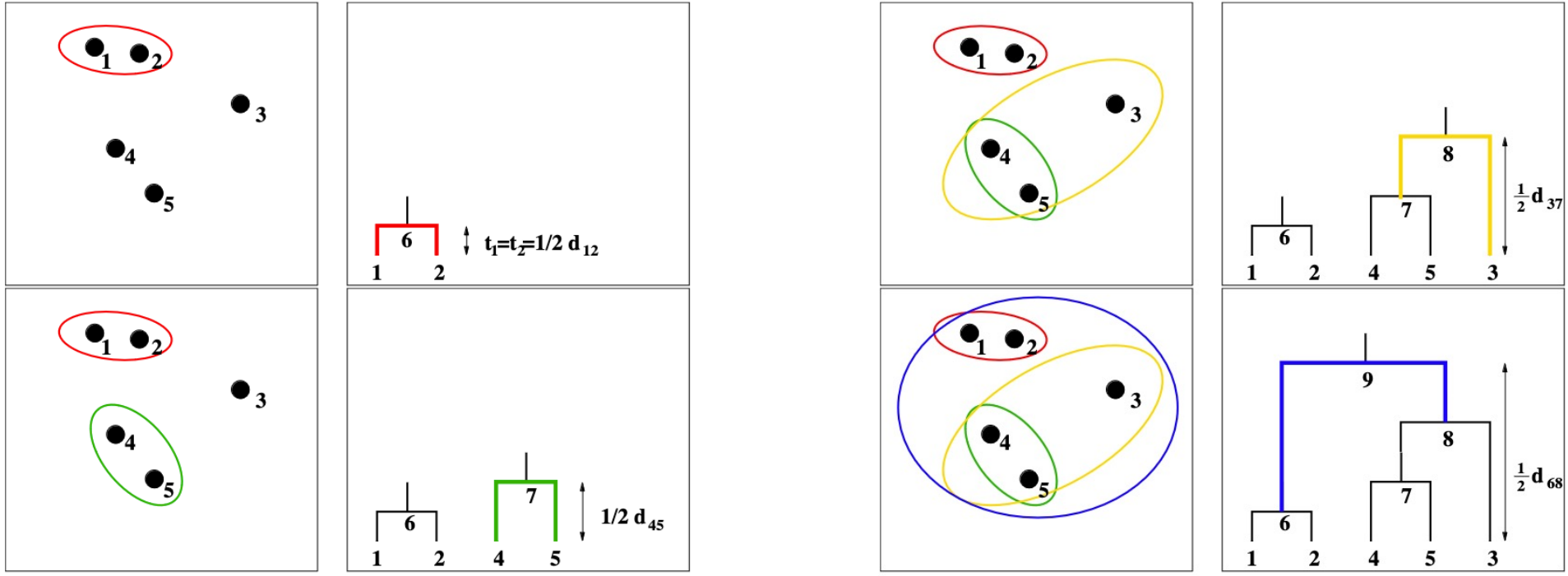
Volker Roth, U. of Basel

  - Generally, data is neither ultrametric nor additive, but NJ can often give reasonable approximations

Phylogeny inference: distance-based methods

- **Step 2: Building a tree from a matrix of pairwise distances between sequences**

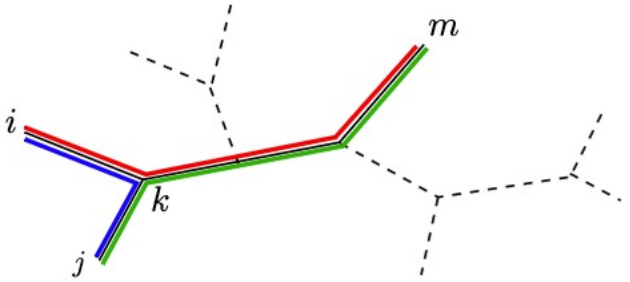
  - **UPGMA:** it is a form of hierarchical clustering – iteratively joins two nearest clusters.
    Initially, each leaf is a cluster.
    Find the 2 clusters i,j with the smallest distance d_{ij}. Group them into new cluster, and compute distance from it to all other ones as a weighted average: $d_{kl} = (\frac{n_i}{n_i + n_j})d_{il} + (\frac{n_j}{n_i + n_j})d_{jl}$
    Connect i,j to new node k, at height $d_{ij}/2$. Iterate.



Phylogeny inference: distance-based methods

- **Step 2: Building a tree from a matrix of pairwise distances between sequences**

  - **NJ:** The idea is to find direct ancestor of 2 species, join them, iterate.
    Initially, each leaf is a cluster.
    For each node i, compute $u_i = \sum_{k \neq i} \frac{d_{ik}}{(n-2)}$: average distance to all other leaves k
    Choose i and j with smallest value of d_{ij}-u_i-u_j (thus i and j are close together and far from the rest)
    Join i and j with ancestor k. The distances between k and other leaves m is defined as follows:



$$\left.\begin{aligned} d_{jm} &= d_{jk} + d_{km} \\ d_{im} &= d_{ik} + d_{km} \\ d_{ij} &= d_{ik} + d_{kj} \end{aligned}\right] \Rightarrow \boxed{d_{km} = \tfrac{1}{2}(d_{im} + d_{jm} - d_{ij})}$$

Volker Roth, U. of Basel

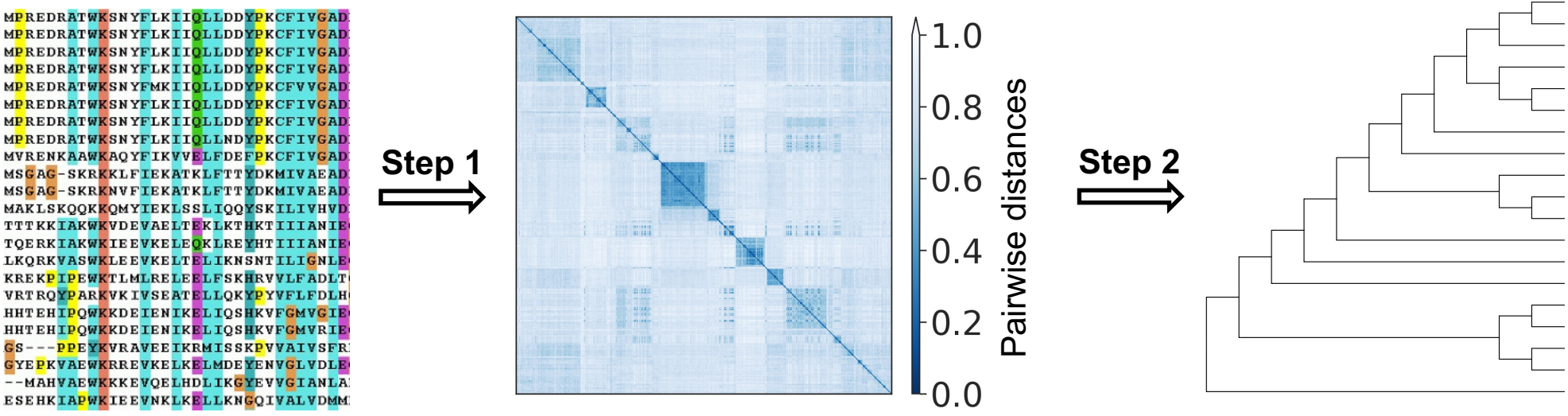
Branch lengths from i and j to the new node k are calculated as: $d_{ik} = \frac{1}{2}(d_{ij} + u_i - u_j)$, $d_{jk} = \frac{1}{2}(d_{ij} + u_j - u_i)$
Iterate.

- **Reminder: limitations:**
  - Starting from distances → we lose information from data (+ distance calculation is approximate)
  - Generally, data is not ultrametric or additive, but NJ can often give reasonable approximations

Phylogeny inference: distance-based methods

General method

- **Step 1:** MSA → pairwise distances between sequences
- **Step 2:** pairwise distances between sequences → tree matching observed distances

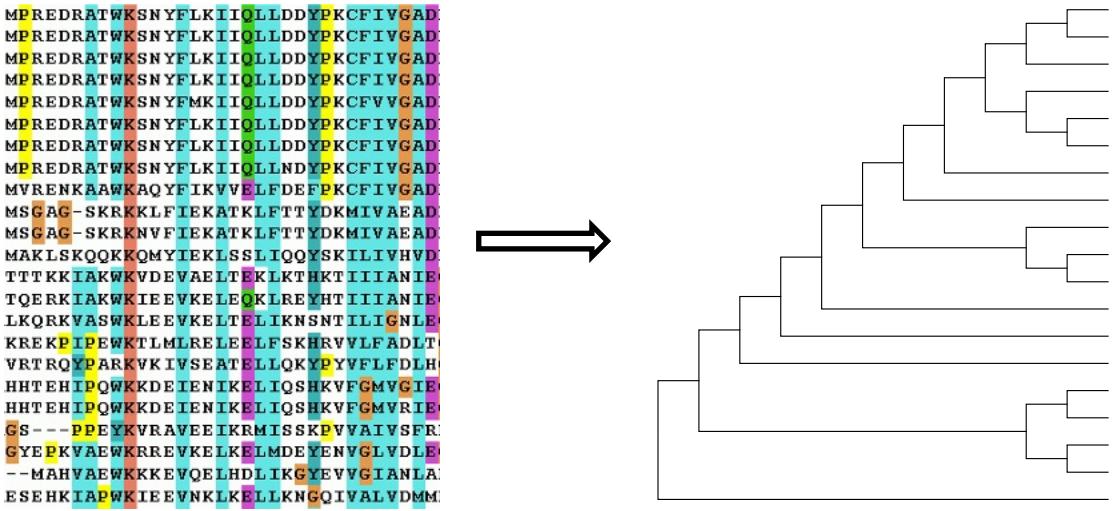


- **Limitation:** restricting to pairwise distances leaves out some information contained in the raw data (sequences). Two different pairs of sequences can have the same distance
- However, a lot of evolutionary information is contained in distances

Phylogeny inference: character-based methods

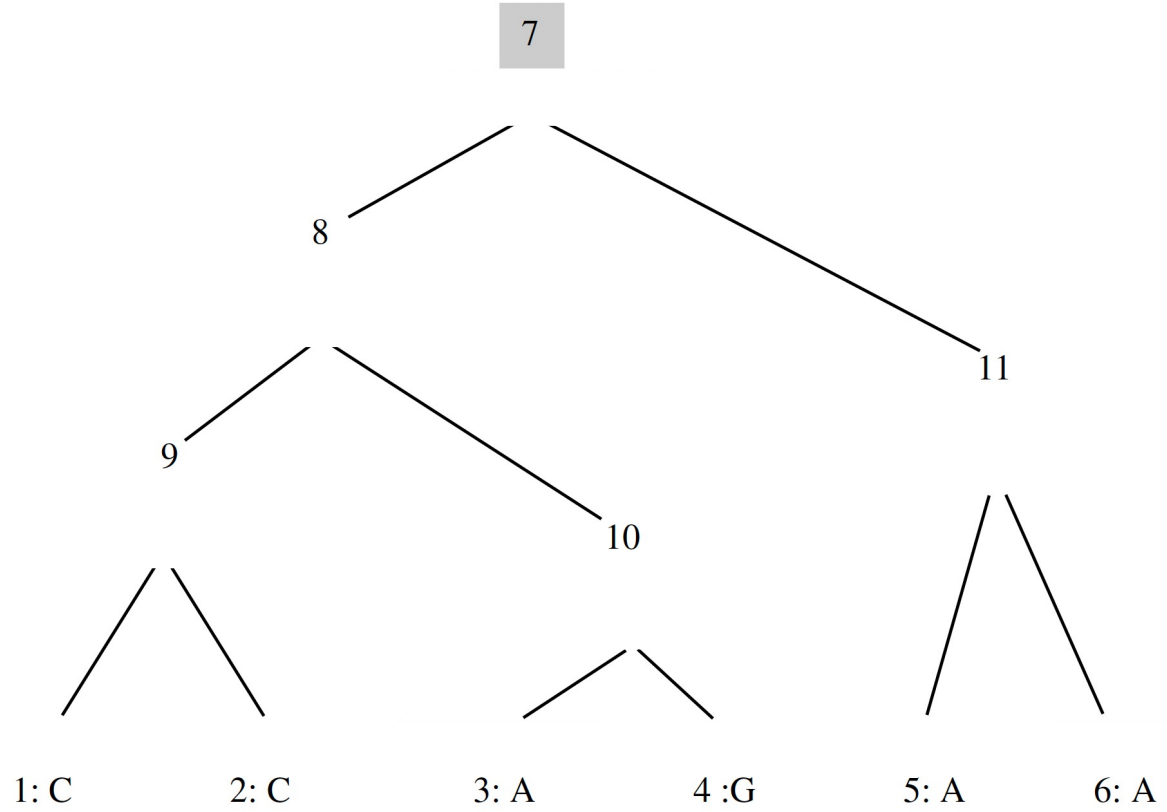
General method

- MSA → tree explaining the MSA
- Use a score that quantifies how well a tree describes the raw data, and find the tree with the best score. These methods directly aim to fit the states (characters, i.e. amino acids or nucleotides) observed at each site site in each sequence to a tree
- **Simplifying assumption:** each site evolves independently from all others



Phylogeny inference: character-based methods

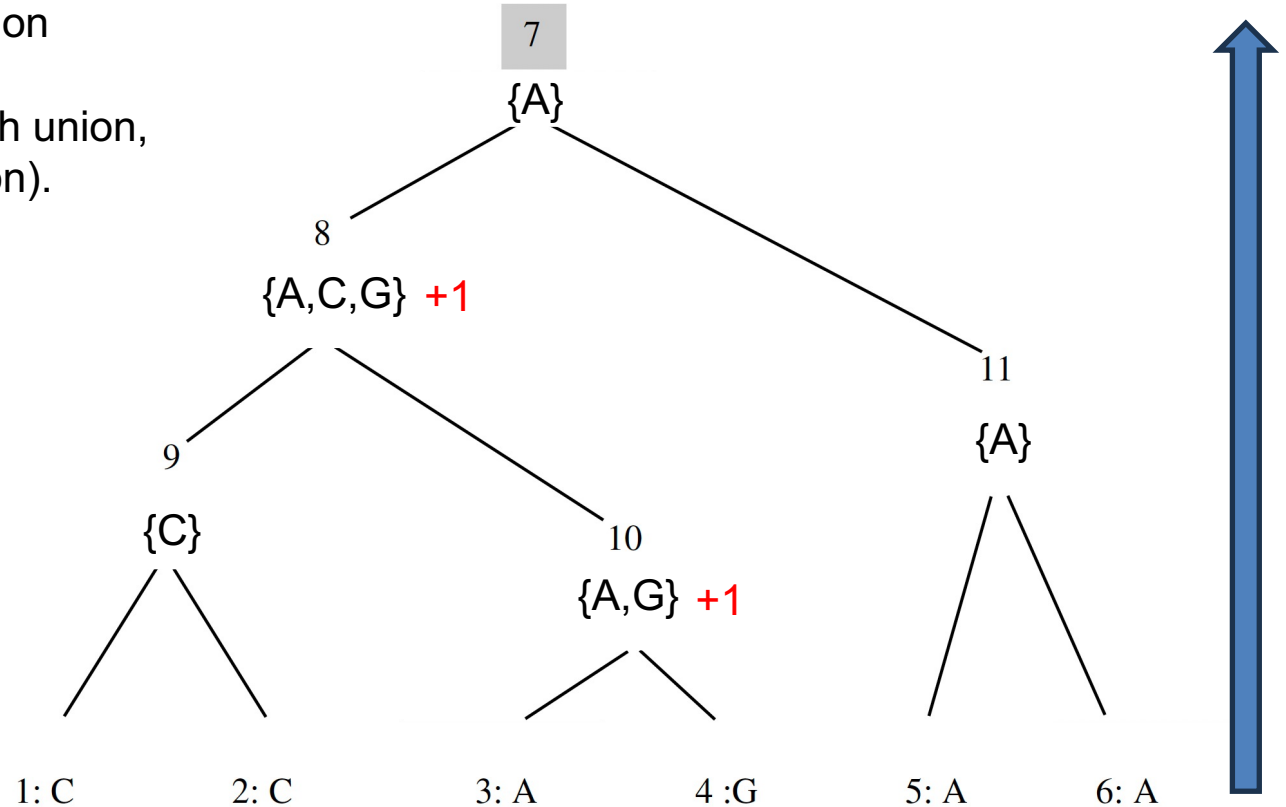
- **Parsimony**
  - **Score:** total number of substitutions (mutations) along all edges of the tree
    Minimizing this score → Occam's razor – simplest way to explain the data
  - **Scoring a given tree:**



Phylogeny inference: character-based methods

Parsimony

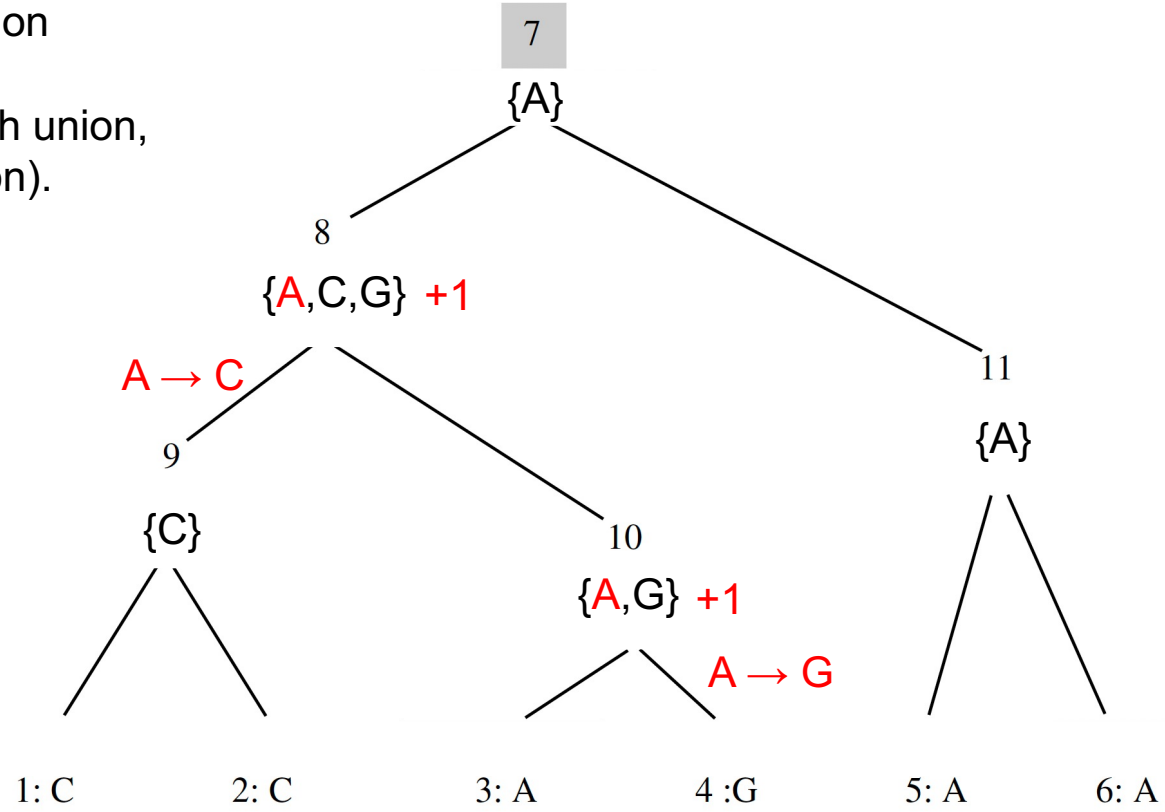
- **Score:** total number of substitutions (mutations) along all edges of the tree
  Minimizing this score → Occam's razor – simplest way to explain the data
- **Scoring a given tree: Fitch algorithm:**
  Label internal nodes with intersection
  of states of their leaves.
  If intersection is empty, replace with union,
  and increase score by 1 (1 mutation).



Phylogeny inference: character-based methods

Parsimony

- **Score:** total number of substitutions (mutations) along all edges of the tree
  Minimizing this score → Occam's razor – simplest way to explain the data
- **Scoring a given tree: Fitch algorithm:**
  Label internal nodes with intersection
  of states of their leaves.
  If intersection is empty, replace with union,
  and increase score by 1 (1 mutation).

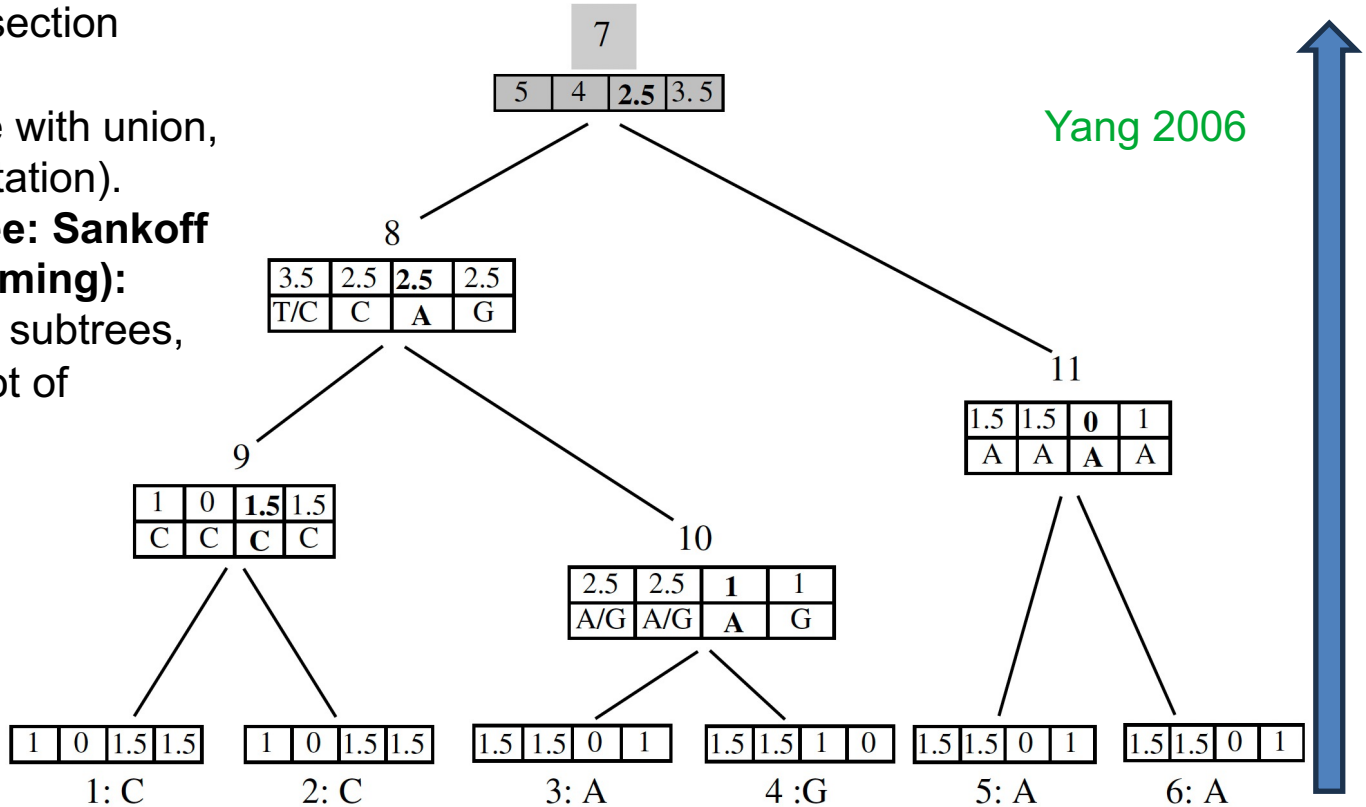


Phylogeny inference: character-based methods

Parsimony

- **Score:** total number of substitutions (mutations) along all edges of the tree
  Minimizing this score → Occam's razor – simplest way to explain the data
- **Scoring a given tree: Fitch algorithm:**
  Label internal nodes with intersection of states of their leaves.
  If intersection is empty, replace with union, and increase score by 1 (1 mutation).
- **Remark – scoring a given tree: Sankoff algorithm (dynamic programming):**
  Mutation score matrix → score subtrees, for all possible states of the root of the subtree: T,C,A,G.

From	To T	C	A	G
T	0	1	1.5	1.5
C	1	0	1.5	1.5
A	1.5	1.5	0	1
G	1.5	1.5	1	0

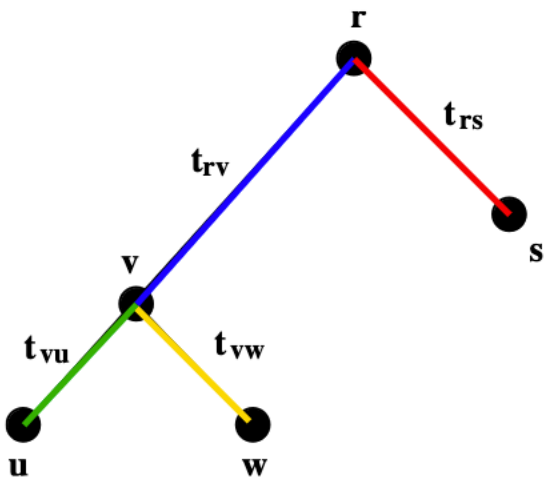


Phylogeny inference: character-based methods

Maximum likelihood

- **Score:** likelihood of the data (MSA) given the model (the tree, with topology and branch lengths, under a certain nucleotide evolution model)
- **Scoring a given tree:**
  (1) Assume that each site (each MSA column) evolves independently of others
  (2) Assume that the probability of a node having a certain state only depends on the state of its parent node and on the branch length (genetic distance) t between them
  (3) Assume that nucleotide frequencies P(x) are fixed through the phylogeny

Using (1), the likelihood is $\mathcal{L} = \prod_{i=1}^{L} P\left(x_1^{(i)}, \ldots, x_D^{(i)} \middle| T\right)$ with L: number of sites; D: number of sequences



Observed data at one given site i, for D=3: u, w, s (= x_1, x_2, x_3)

$$P(u,w,s|T) = \sum_{r,v} P(u,w,s,v,r|T)$$ Sum over ancestral states r, v

$$= \sum_{r,v} P(r) P(s|r,t_{rs}) P(v|r,t_{rv}) P(w|v,t_{vw}) P(u|v,t_{vu})$$

(2) & (3), using Bayes' theorem $P(y|x) = P(x|y)\frac{P(y)}{P(x)}$

For each branch, use a nucleotide evolution model, e.g. Jukes-Cantor

Phylogeny inference: character-based methods

Maximum likelihood

- **Reminder: Jukes-Cantor distance (see typed notes):**

  Evolutionary model where:
  - each site evolves independently of others
  - all substitutions are equally likely: rate λ

  $$\frac{dP_A}{dt}(t) = \lambda\left[1 - 4P_A(t)\right]$$

  If the initial state at t=0 is A, then: $P_A(t) = \frac{3}{4}e^{-4\lambda t} + \frac{1}{4}$ and $P_C(t) = P_G(t) = P_T(t) = -\frac{1}{4}e^{-4\lambda t} + \frac{1}{4}$

- Thus, for one given site i, the score of a branch is $P(s|r, d_{rs}) = -\frac{1}{4}e^{-4d_{rs}/3} + \frac{1}{4}$ if $r \neq s$

  $$P(s|r, d_{rs}) = \frac{3}{4}e^{-4d_{rs}/3} + \frac{1}{4} \quad \text{if } r = s$$

  where d_{rs} is the evolutionary distance between the two nodes (recall that $d=3\lambda t$)

  → express $\mathcal{L} = \prod_{i=1}^{L} P\left(x_1^{(i)}, \ldots, x_D^{(i)} \middle| T\right)$

- For a given tree topology, branch lengths can be chosen to maximize this likelihood
- These results can be compared across trees topologies, to find the tree with the highest likelihood

Phylogeny inference: character-based methods

Bayesian approach

- **Score:** posterior probability of the model (the tree, with topology and branch lengths, under a certain nucleotide evolution model) given the data (the MSA)
- **Scoring a given tree:**
  Reminder: Bayes' theorem: $P(\text{model}|\text{data}) = P(\text{data}|\text{model}) \times \frac{P(\text{model})}{P(\text{data})}$
  It is difficult to access the posterior in analytic form
  Strategy: draw samples (generate tree) from the posterior, using the Metropolis method

Comparison of approaches

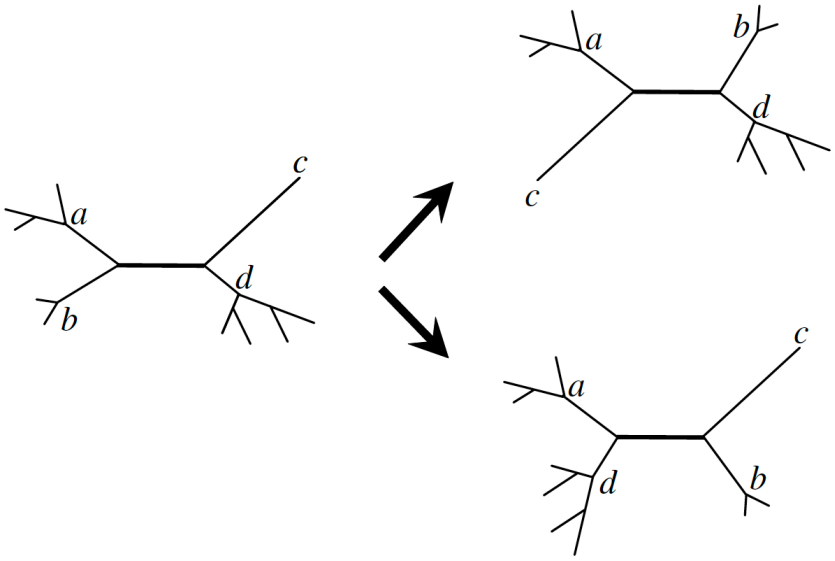
- Currently, maximum likelihood and Bayesian methods are considered the most accurate ones
- But they are computationally intensive
- Parsimony is intuitive; distance-based methods are efficient and can be used as starting points
- Distance-based methods can be good enough, e.g. for relatively close sequences

Remark: search over possible trees

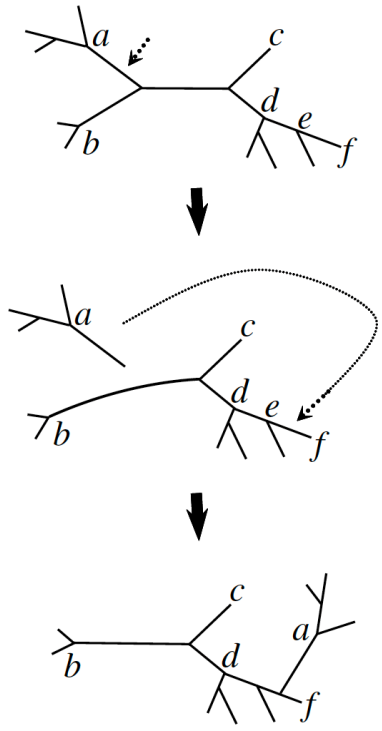
- So far, we mainly looked at how to score a given tree
- Next, need (in principle) to score all possible trees. But there are many trees!
- Heuristic search strategies exist

Phylogeny inference: character-based methods

- **Search over possible trees: heuristic strategies**
  - **Idea:** starting from a tree, construct neighboring trees by elementary operations
  - Nearest-neighbor interchange (NNI)



Subtree pruning and regrafting (SPR)

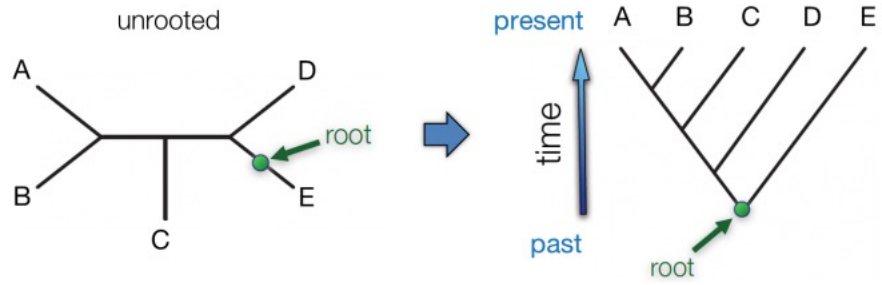


Yang 2006

Phylogeny inference: rooting a tree

Rooting a tree

- So far, we mainly saw how to select an unrooted tree
- Root = position of the most common recent ancestor; tells us the direction of evolution
- Position of the root affects interpretation about relatedness of sequences



EMBL-EBI online training

- **Where to place the root?** (Remark: there are even more rooted trees than unrooted trees)

  A common strategy is **outgroup rooting:**
  - Include a sequence that is known to be more distant from the sequences considered than they are among them (e.g. from species information)
  - Place the root where this outgroup joins the rest of the tree
  - In the illustration above, E would be the outgroup

Phylogeny inference: estimating confidence

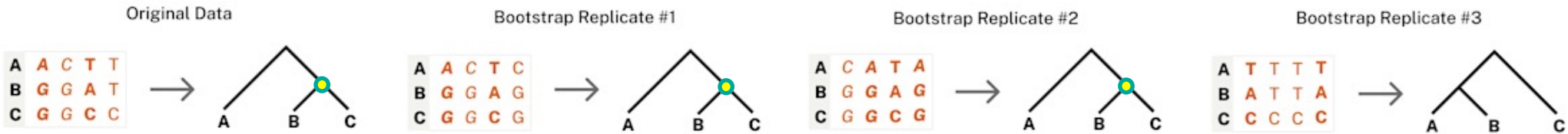
Bootstrapping

- **Reminder:** usually, in phylogeny inference methods, each site (each column of the MSA) is assumed to evolve independently

→ This can be exploited to estimate confidence

- **Method:**
  - Resample sites (columns) from the MSA: sample sites uniformly at random with replacement, to form a new MSA with the same number of columns. This new MSA is called a bootstrap replicate.
  - Infer trees for multiple bootstrap replicates
  - Bootstrap support value = percentage of bootstrapped trees that contain a particular node
  - Higher value (close to 100) means better confidence

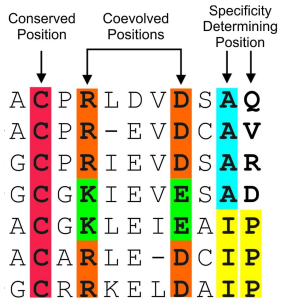
- **Example:**



Highlighted node present in 67% of bootstrap replicate trees → score 67

Geneious

Phylogeny inference: limitations



Assumptions

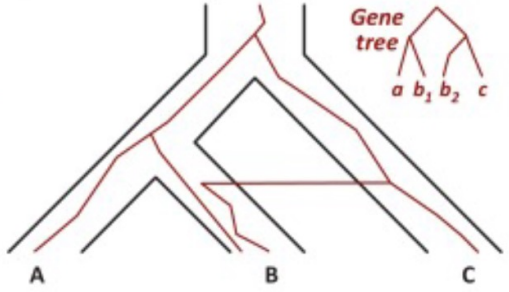
- Independently evolving sites
- Nucleotide evolution model
- Same mutation rate at each site, or limited number of classes of mutation rates

Approximations

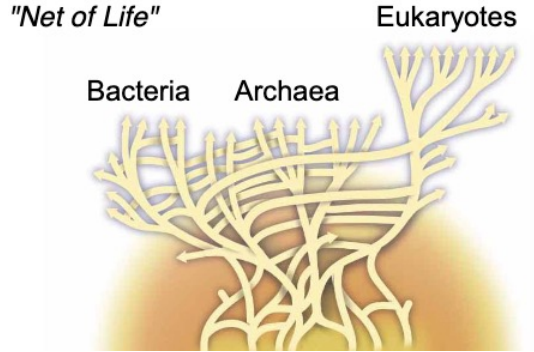
- Starting from distances in distance-based methods
- Search over all trees can't be done exhaustively → heuristic search strategies

More fundamentally

- Horizontal gene transfer between species → trees of different genes are inconsistent with each other
- No fundamental species notion in prokaryotes



Nakhleh 2013

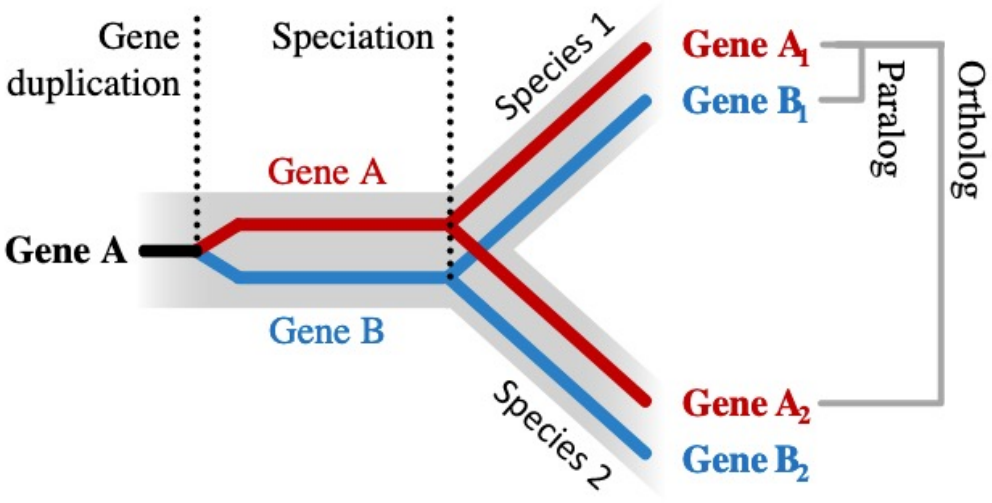


Doolittle 2000

Homologs, orthologs and paralogs

- **Definitions**

Reminder: Gene A_1 in species 1 and gene A_2 in species 2 are homologous if they share a common ancestor
Two homologous genes are:
- Orthologous if they diverged at a speciation event
- Paralogous if they diverged at a duplication event



Generally, orthologs preserve the same function, while paralogs do not and become more different

- **Practical approximate way to find orthologs**

Reciprocal best hits: pairs (**g,h**) of genes from genomes (**G,H**) such that **g** is the gene in **G** most similar to **h**, and **h** is the gene in **H** most similar to **g**