American Museum of Natural History

American Museum of Natural History
Richard Gilder Graduate School

RGGS Comparative Genomics 2 – Computational Methods (Session 13)

Jose Barba

Gerstner Scholar in Bioinformatics & Computational Biology

AMNH Arachnology Lab
Institute for Comparative Genomics
jbarba@amnh.org
josebarbamontoya.github.io

December 2, 2024
ICG conference room

Session 13 outline

- **Phylogenetic inference from NGS (part 2)**
- **Final project description /reminder**

Final project description (from the syllabus)

- This course is project-based, allowing students to develop and refine skills essential for analyzing data from their own PhD research. Whenever possible, the final project and homework assignments should incorporate the personal data of students to enhance relevance and practical application. However, if such data are not available or are insufficient, the instructors will guide the students in selecting an appropriate alternative dataset. The main goal of the final project is to provide students with experience in developing well-formulated research questions that can be addressed with -omics data, appropriate computational methods to address these questions, and fully documented, reproducible computational pipelines. Working with instructors, students will craft a research question, select the methods to use, and create a detailed tutorial for the necessary computational steps. A short description (~350 words) of the proposed research question should be submitted by November 14.

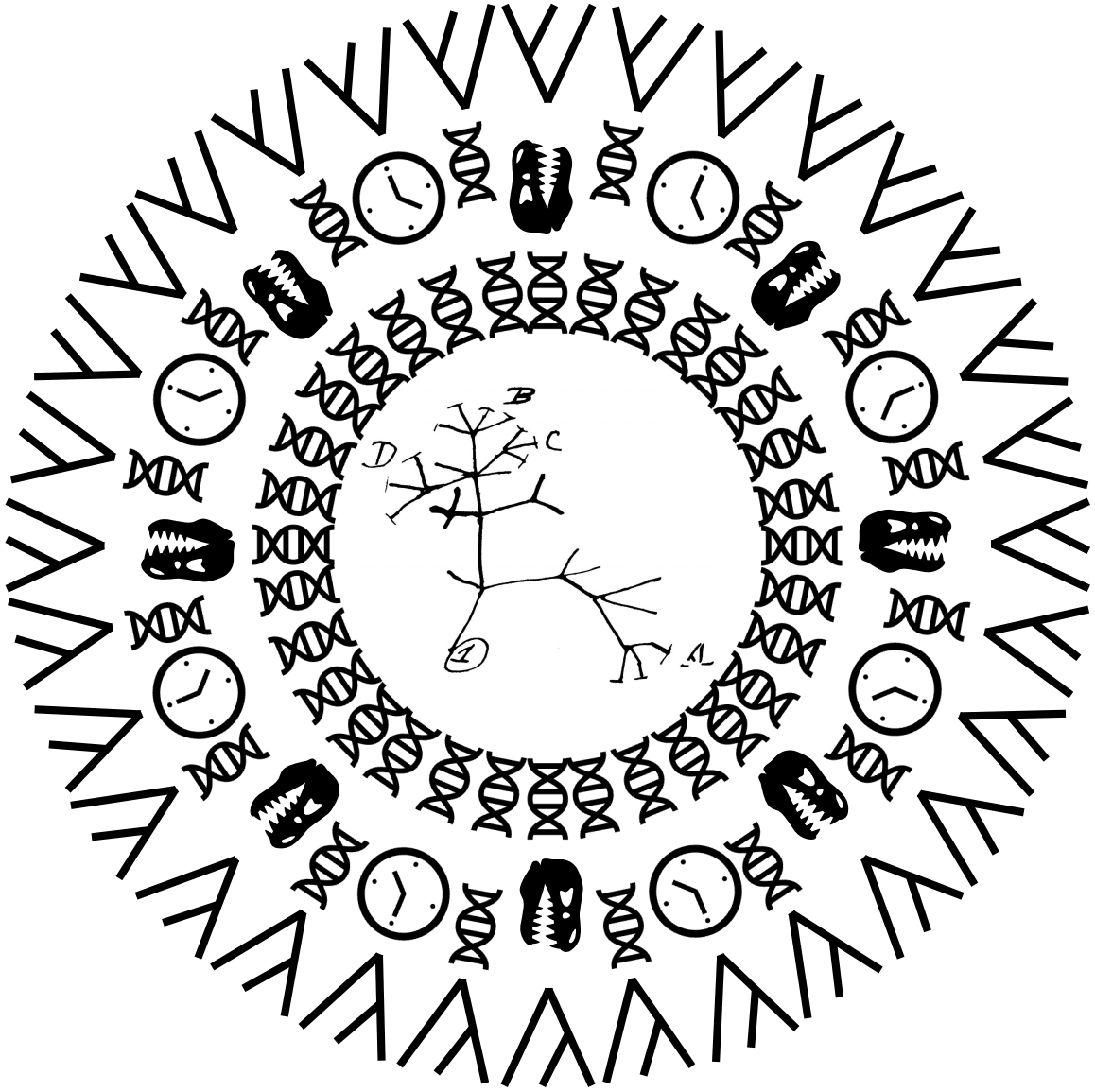
COMPLETED

Final project description (from the syllabus)

- **The final project will consist of a detailed tutorial, which will be shared on the student's GitHub account. Throughout the course, students will work incrementally to complete the project, adding and testing new steps to the pipeline as they are introduced in class. Each student will deliver a 10-minute presentation on their final project. The submitted GitHub tutorial must meet the following requirements:**
  1) An "about" section explaining the research goal, why the specific approach presented was chosen, and any alternatives that should be considered. If alternative approaches were tested and rejected, this should be described.
  2) For every step of the analysis, provide a description of the tool being used, the available options, any caveats to using the tool, and the exact command used in the project. Even trivial steps like text/data manipulations should be documented.
  3) Where relevant, elaborate on any challenges encountered, whether trivial or significant.
  4) A brief bibliography of citations for the computational tools used.
  5) There is no length requirement for the resulting document, but it must be detailed enough for anyone to fully reproduce your research using the tutorial. It must clearly describe all necessary steps.

Reconstructing the Tree of Life from NGS

- **Inference of accurate phylogenies is crucial for understanding major transitions in evolution of genomes and species**
- **Despite abundant data and powerful analysis methods, challenges persist in constructing reliable phylogenies**
- **Implementation and development of methods for phylogenomic analysis and molecular dating that mitigate the factors contributing to error**



Phylogenomics

- A phylogenomic analysis combines molecular evolution and computational techniques to infer evolutionary relationships based on genomic data
- Comparative phylogenomics offers an opportunity to identify candidate genes associated with complex traits
- Phylogenomics is a promising way to resolve the Tree of Life, as demonstrated in several cases
- Resolving power does not increase linearly with the number of characters considered

Applications of phylogenetics

- **Before NGS, phylogenetic trees were used almost exclusively to describe relationships among species in systematics and taxonomy. Today, phylogenies are used in almost every branch of biology.**
- **Besides this phylogenies are used to:**
  - Describe relationships between paralogues in a gene family
  - Histories of populations
  - Evolutionary and epidemiological dynamics of pathogens
  - Genealogical relationship of somatic cells during differentiation and cancer development
  - Evolution of language
  - Forensics

Applications of phylogenetics

- **Recently, molecular phylogenetics has become an indispensable tool for genome comparisons**
  - Classify metagenomic sequences
  - Identify genes
  - Regulatory elements and non-coding RNAs in newly sequenced genomes
  - Interpret modern and ancient individual genomes
  - Reconstruct ancestral genomes

Applications of phylogenetics

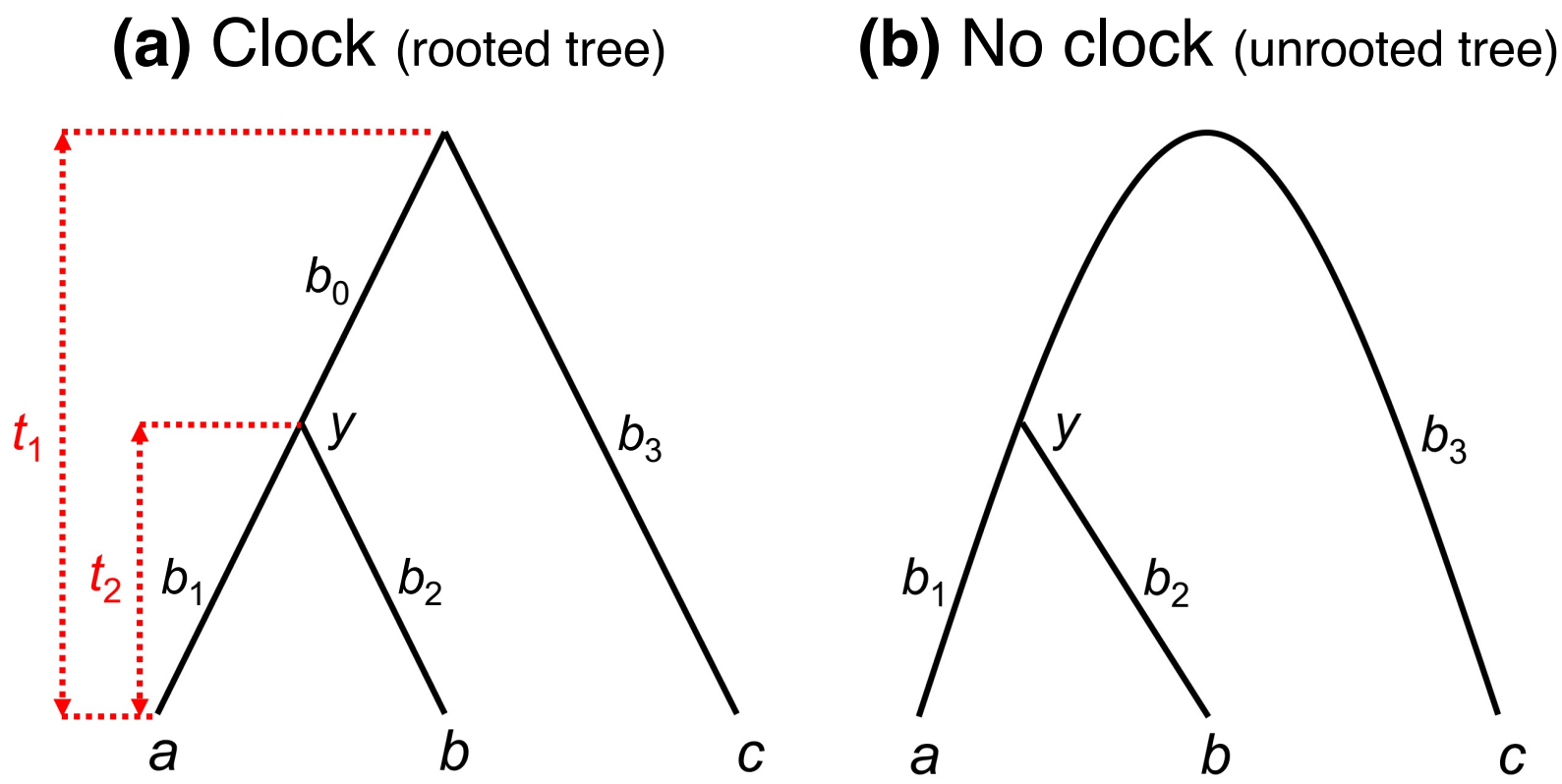
- **In other applications, the phylogeny itself may not be of direct interest but must nevertheless be accounted for in the analysis**
- **In population genetics, the development of the coalescent theory and the widespread availability of gene sequences for multiple individuals from the same species have prompted the development of genealogy-based inference methods**
  - In this case, the gene trees that describe the genealogy of sequences in a sample are highly uncertain; they are not of direct interest but nevertheless contain valuable information about parameters in the model

Applications of phylogenetics

- In species tree methods, the gene trees at individual loci may not be of direct interest and may be in conflict with the species tree
- By averaging over the unobserved gene trees under the multi-species coalescent model, those methods infer the species tree despite uncertainty in the gene trees
- In the inference of adaptive protein evolution, the phylogeny is used to trace the synonymous and nonsynonymous substitutions along branches to identify cases of accelerated amino acid change, even though the phylogeny is not of direct interest
- Nowadays, every biologist needs to know something about phylogenetic inference

Phylogenetic tree concept

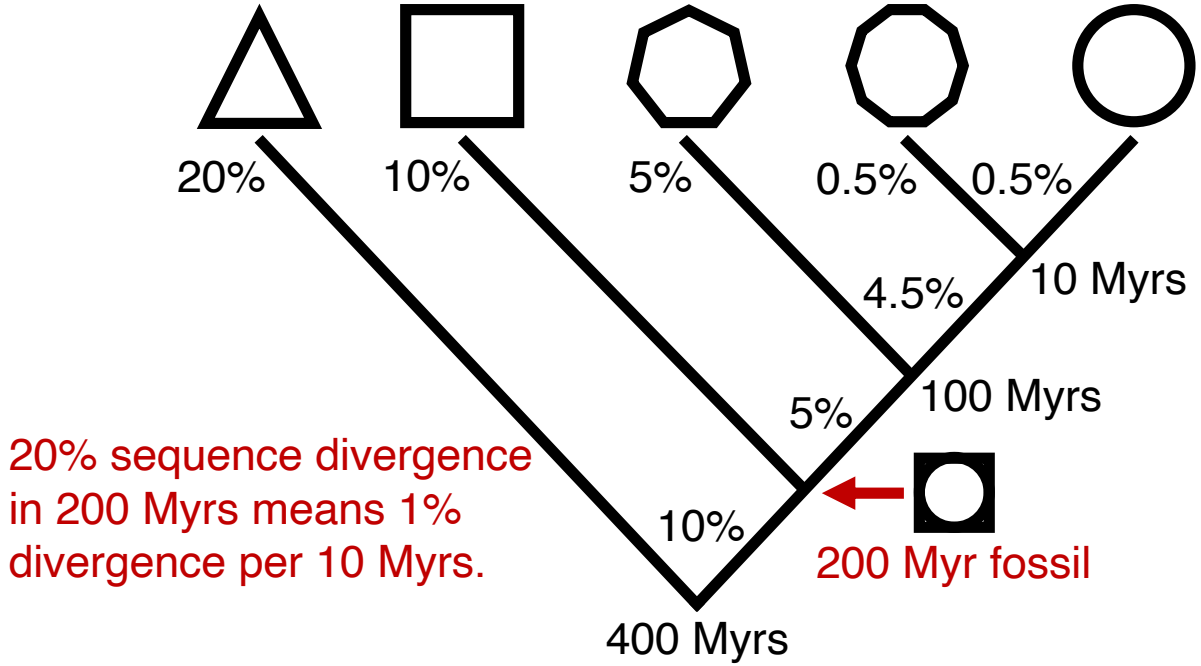
- **If substitution rate is constant over time or among lineages, molecular clock holds. For distantly related species this hypothesis should not be assumed**



Sarich and Wilson (*Science* 1973)

Molecular clock hypothesis

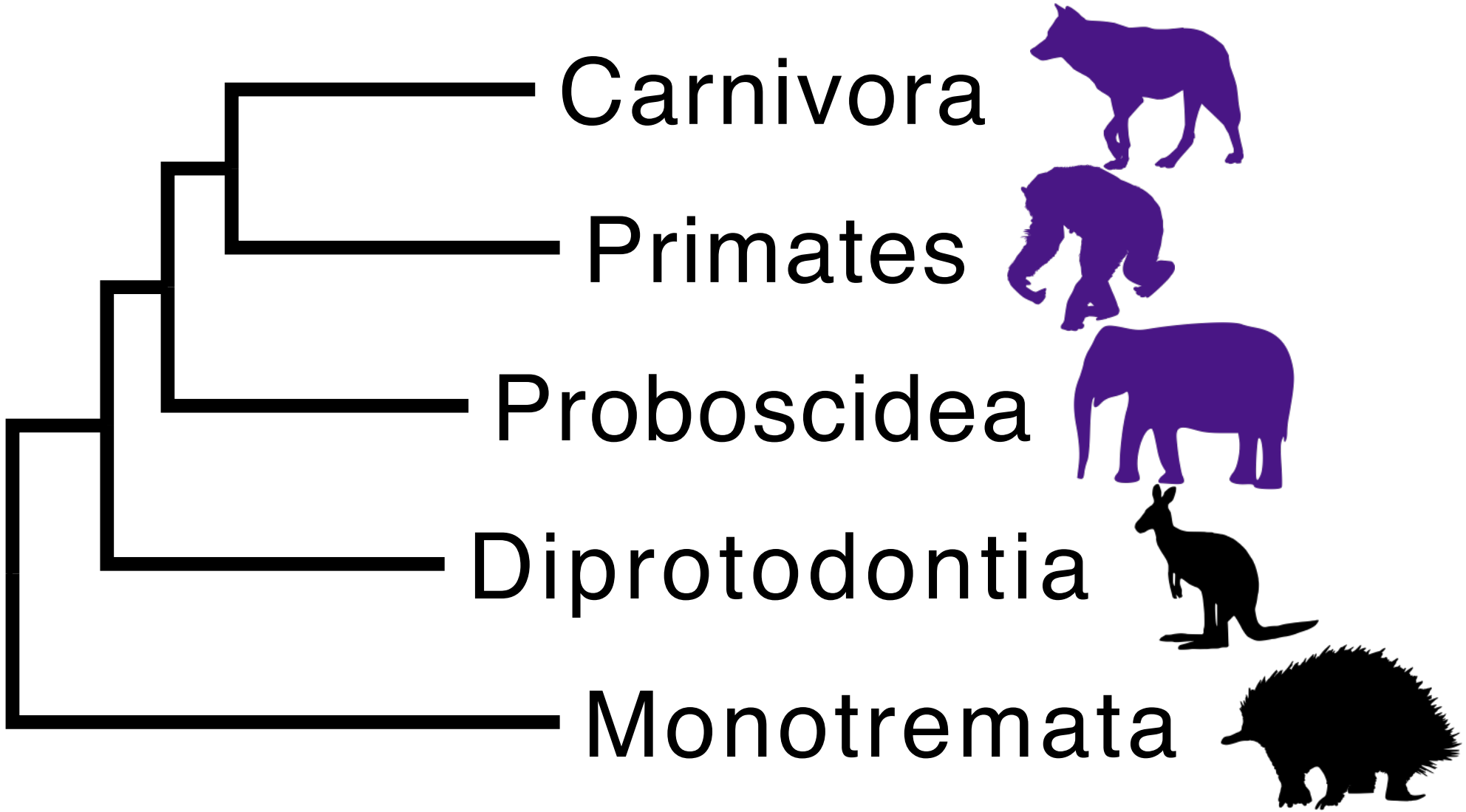
- **Simple but powerful approach measuring timescale of evolutionary divergences. Expected distance between sequences grows linearly with time of divergence**



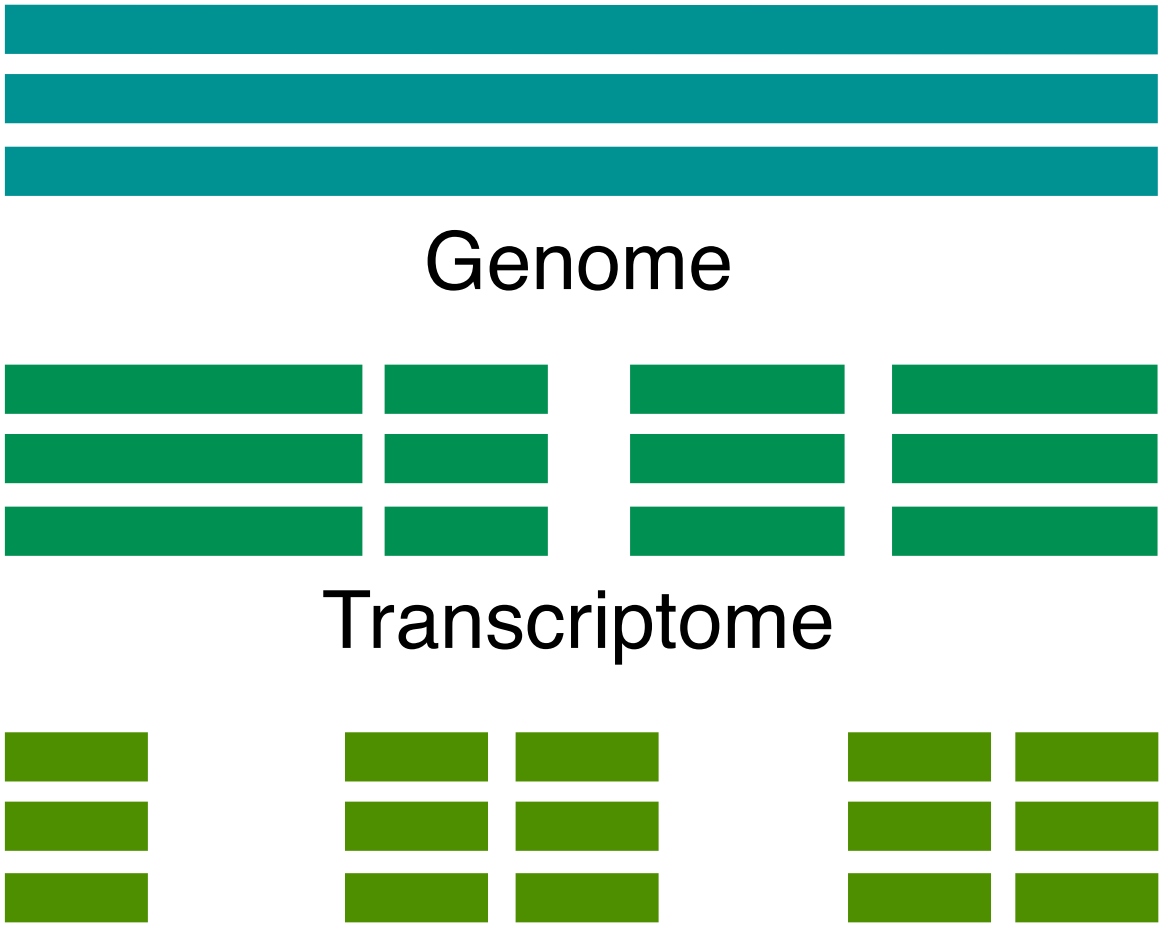
- **Ages from fossil record or geological events, can be used to translate distances between sequences or tree branch lengths into absolute geological times**

Zuckerkandl and Pauling (1962)

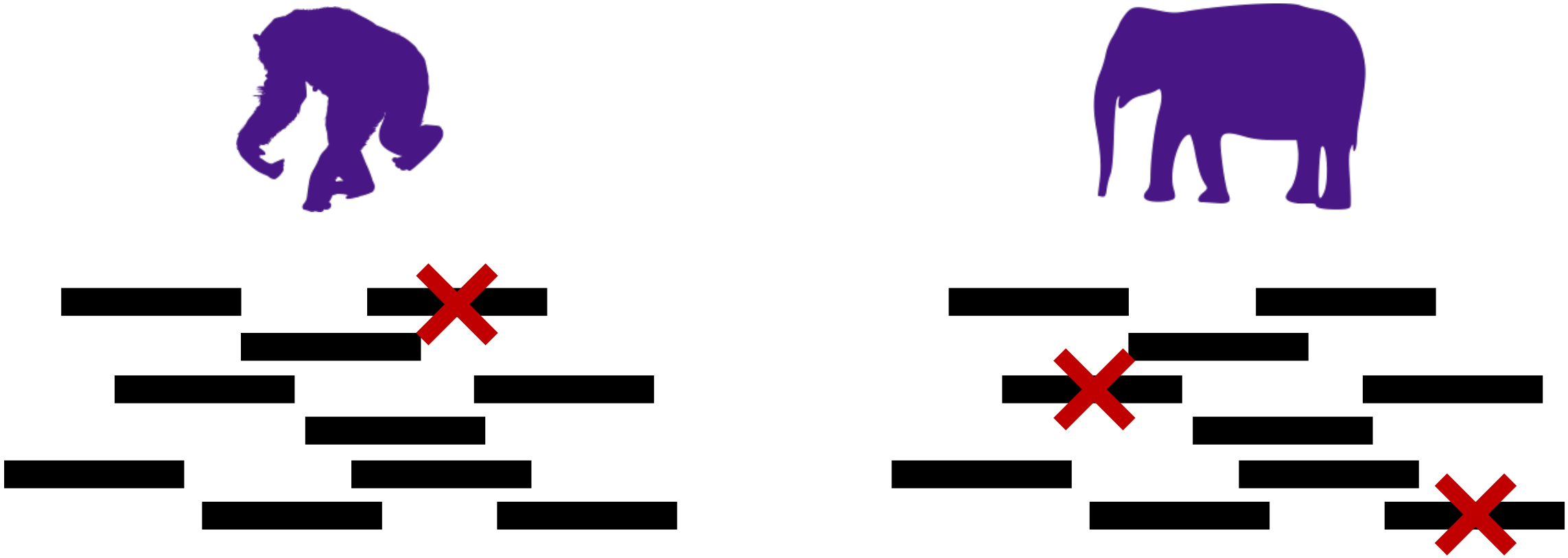
Taxon sampling



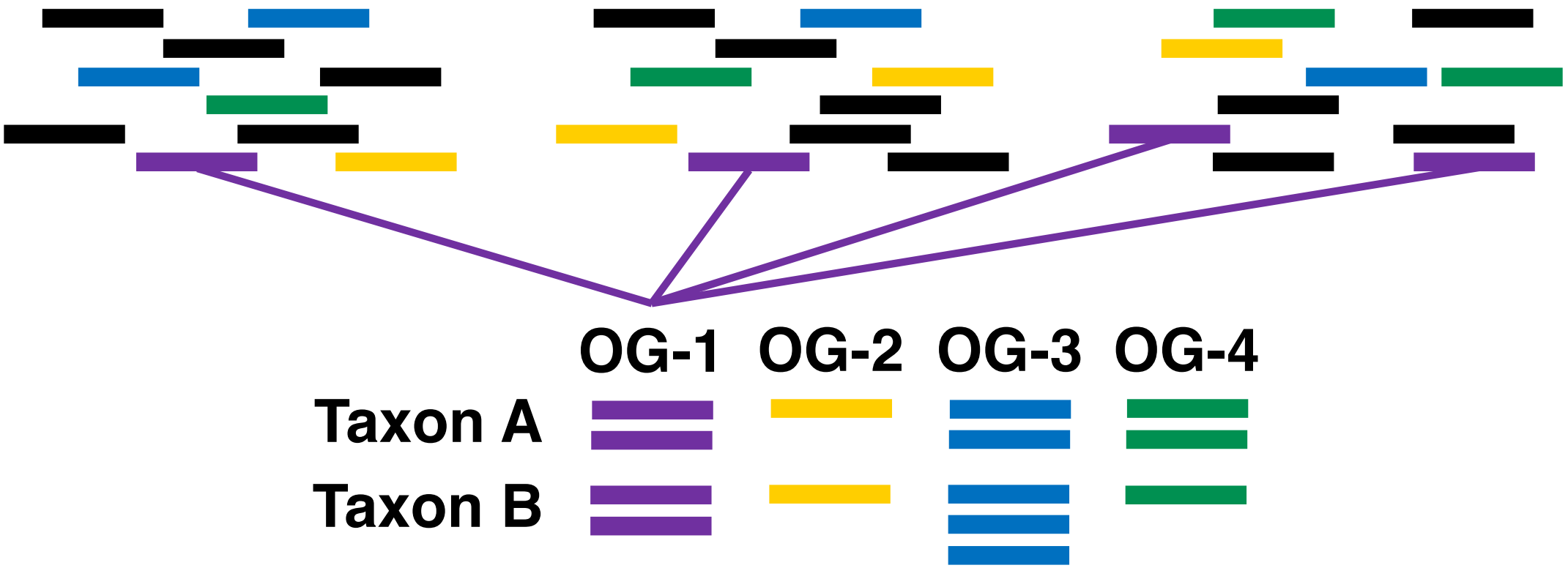
Data type



Contaminants and sequencing errors

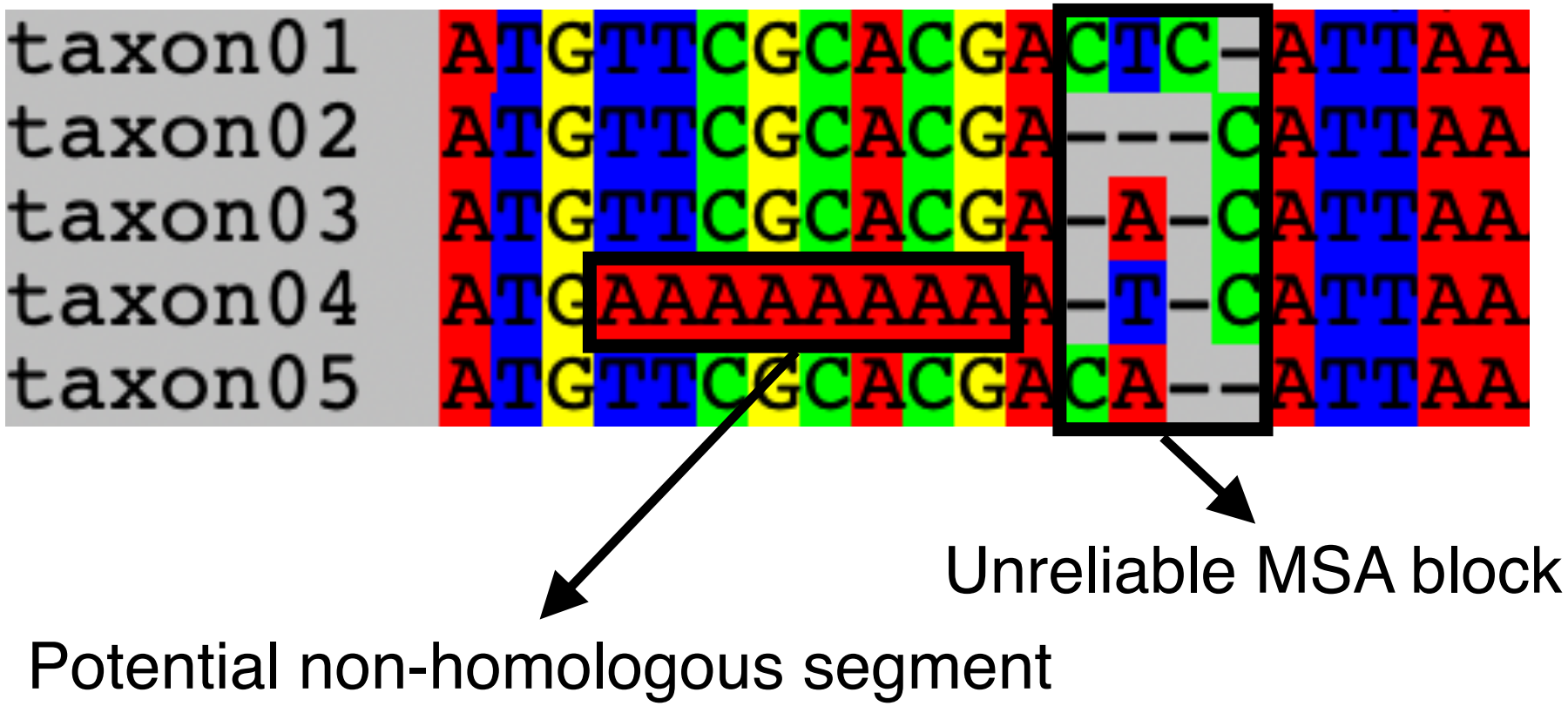


Identification of orthologous sequences

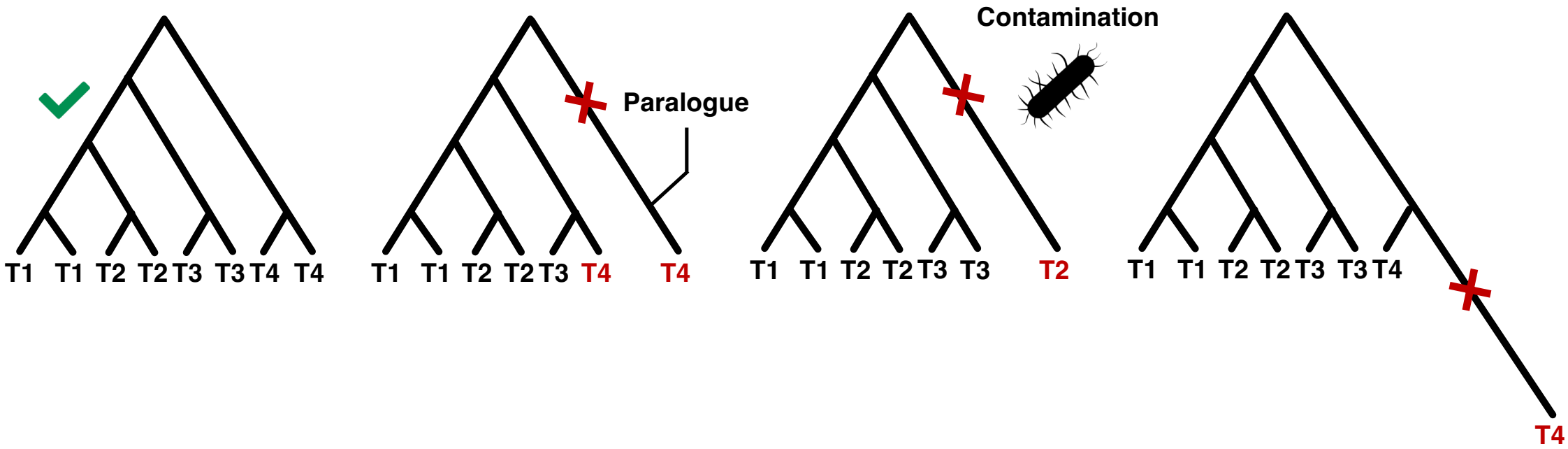


Multiple sequence alignment (MSA)

- **Assess reliability of MSA**

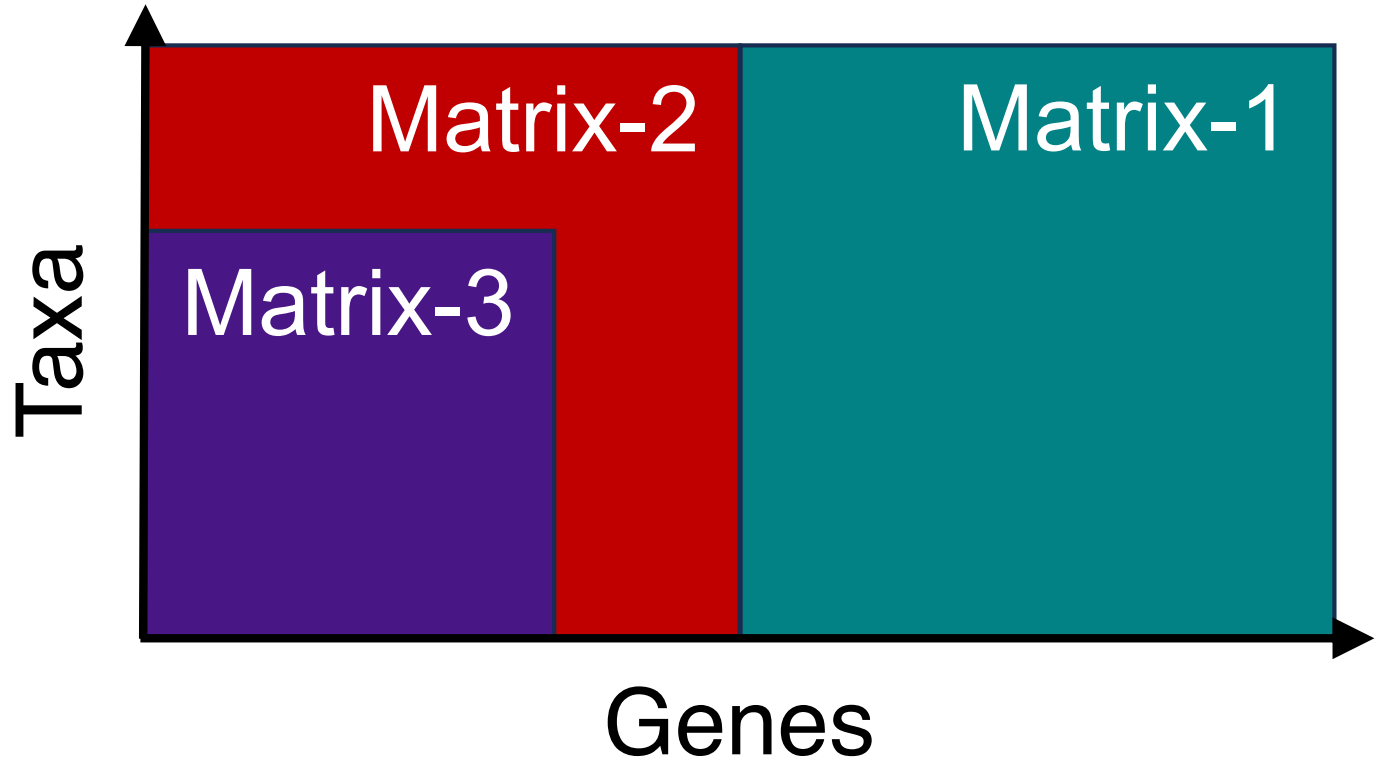


Removal of outliers



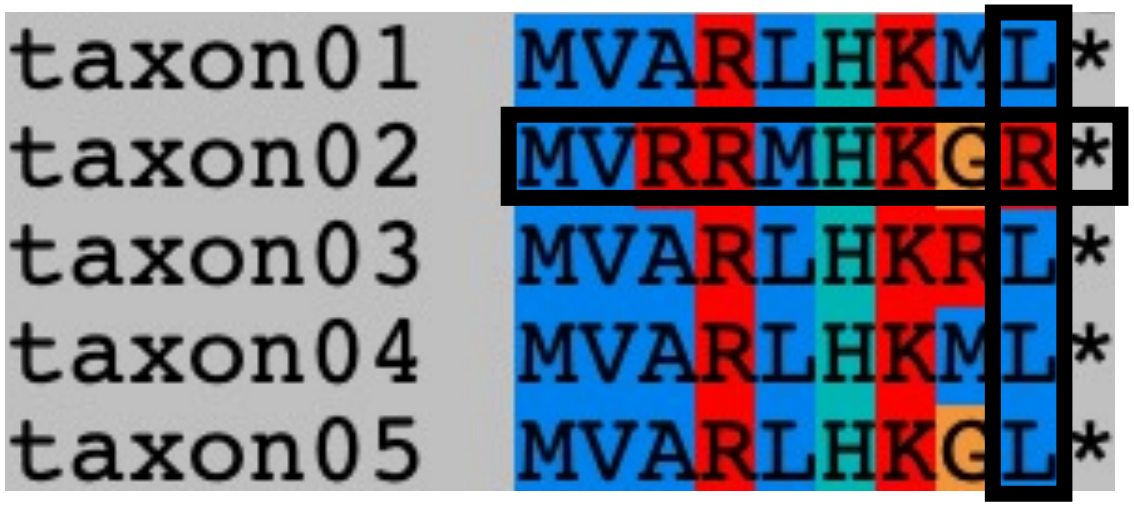
Sensitivity test

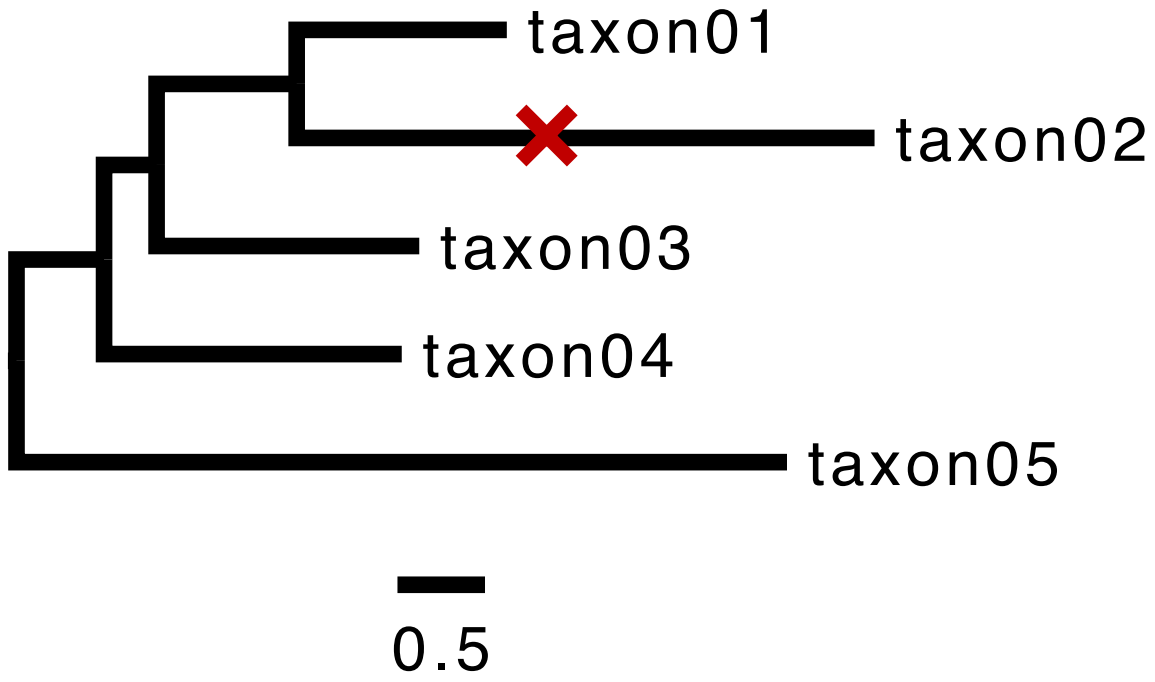
- **Analyze subsampled dataset on different properties and models**



Base composition heterogeneity

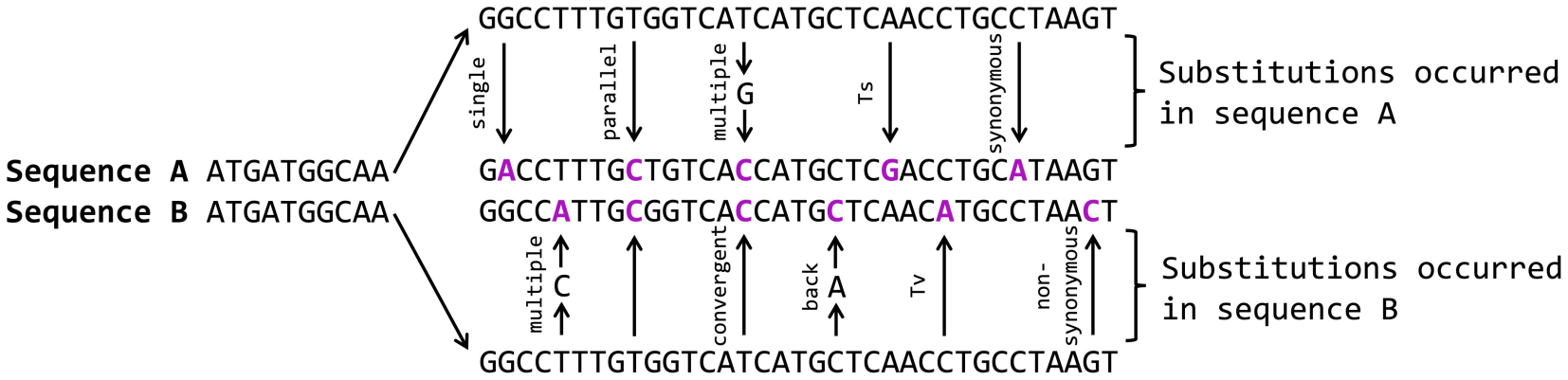
- **Compositionally heterogeneous sites/taxa**
  - Has data been recoded or excluded to reduce rate and compositional bias



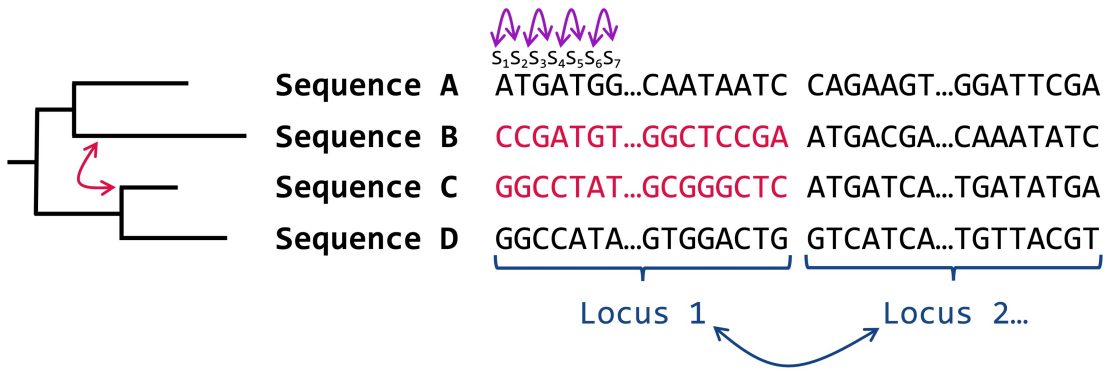


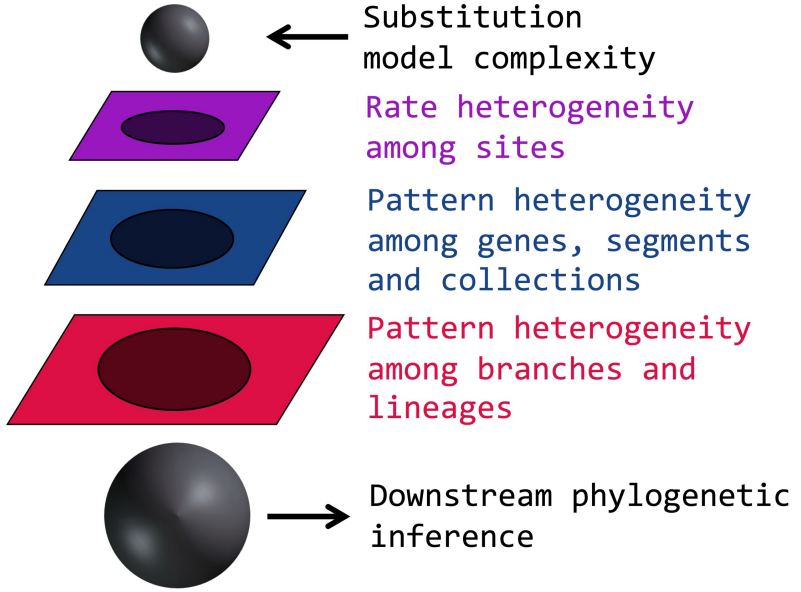
Substitution model

Types of substitutions observed between two sequences

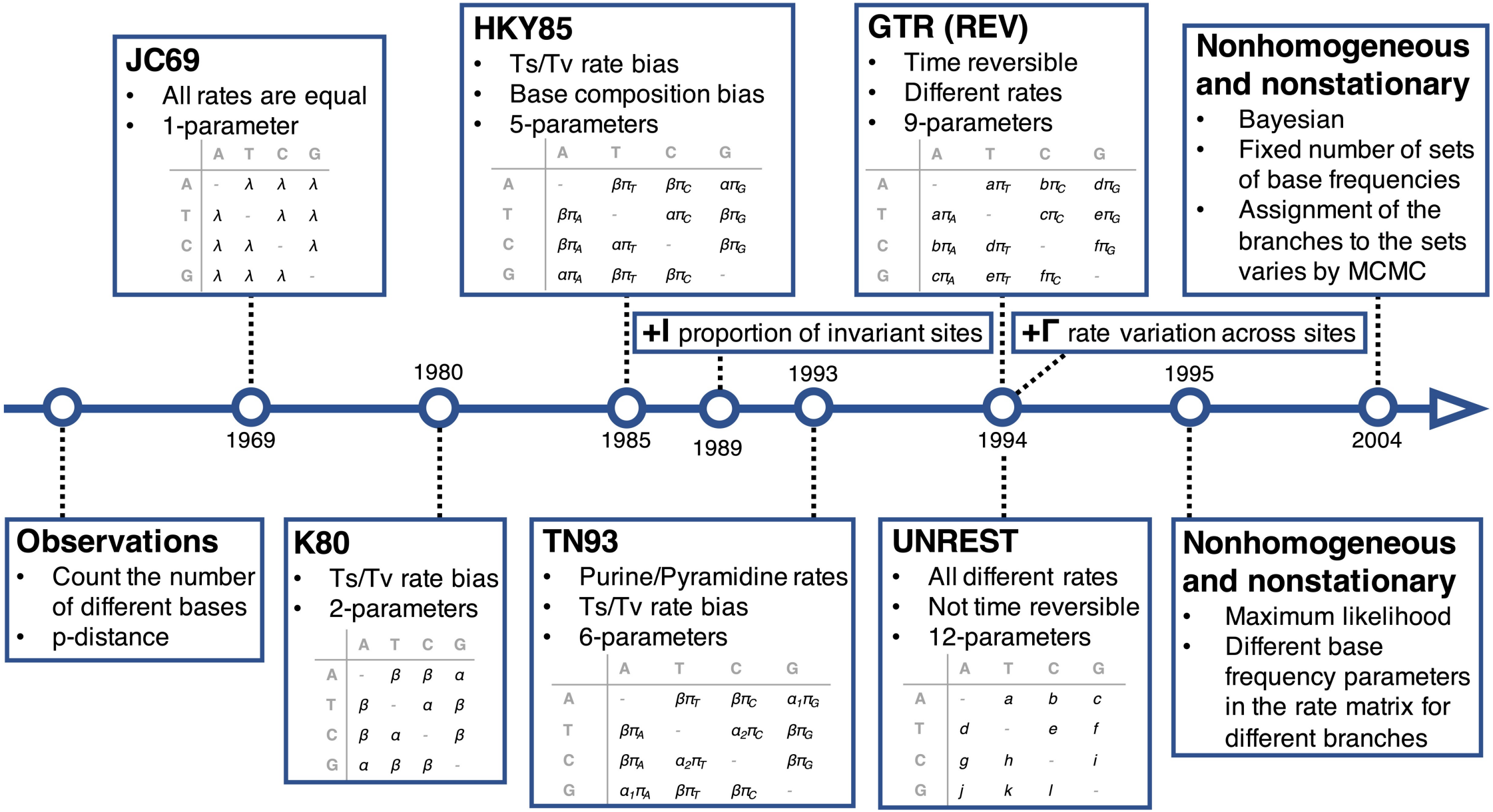


Heterogeneity of the substitution processes across sites and lineages



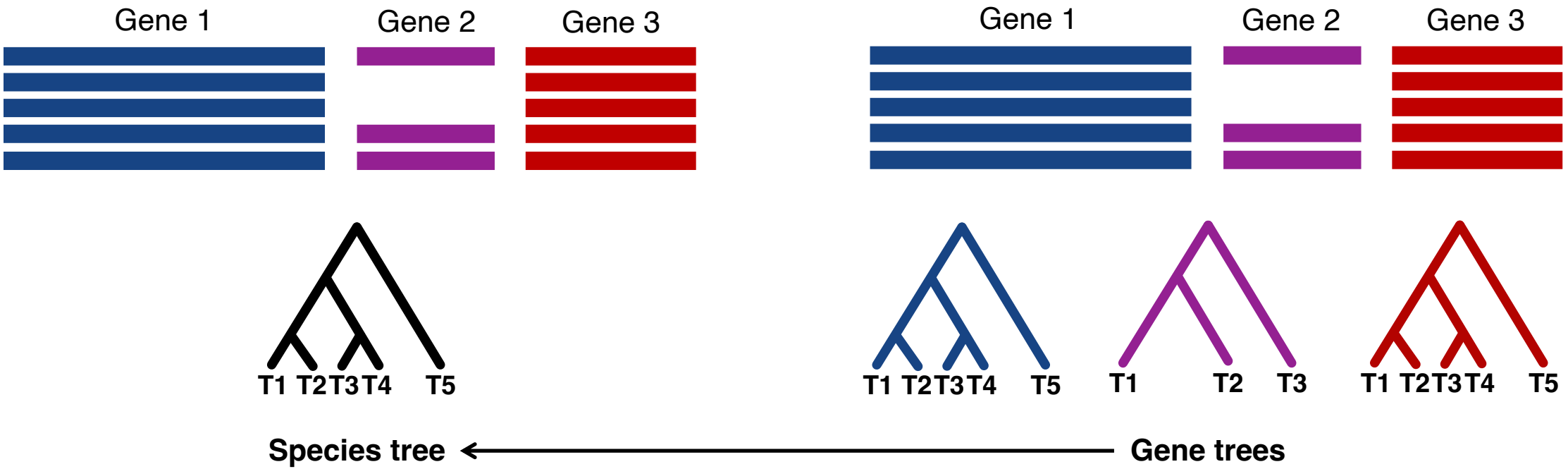


Timeline of models of DNA evolution



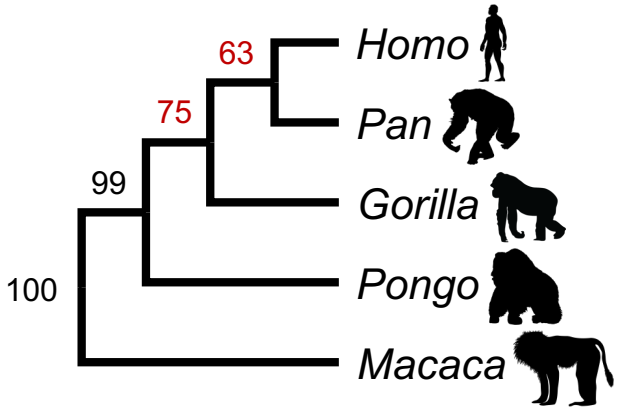
Phylogenetic inference

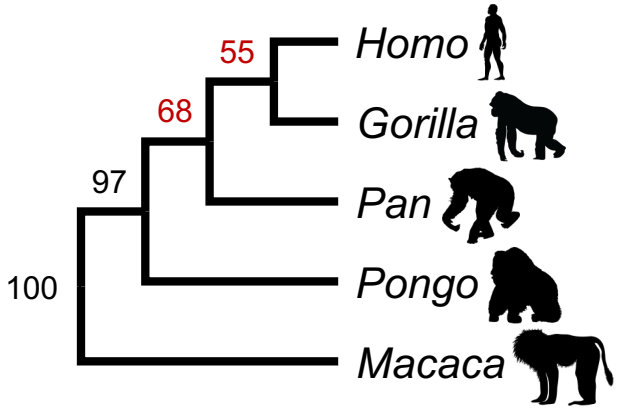
- **Concatenation and coalescence analyses**

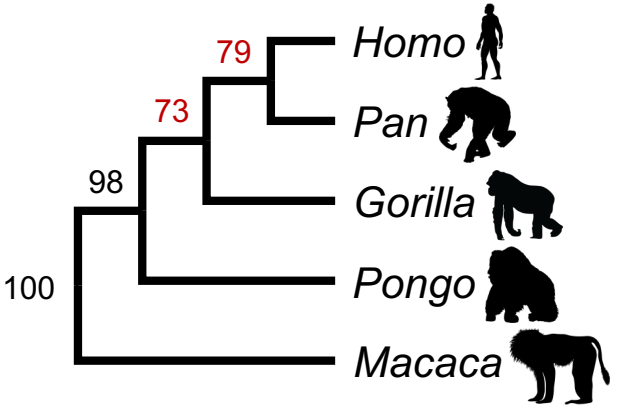


Branch support and probability

- **Do evolutionary hypothesis rely in robust?**
- **Bootstrap, posterior probability**

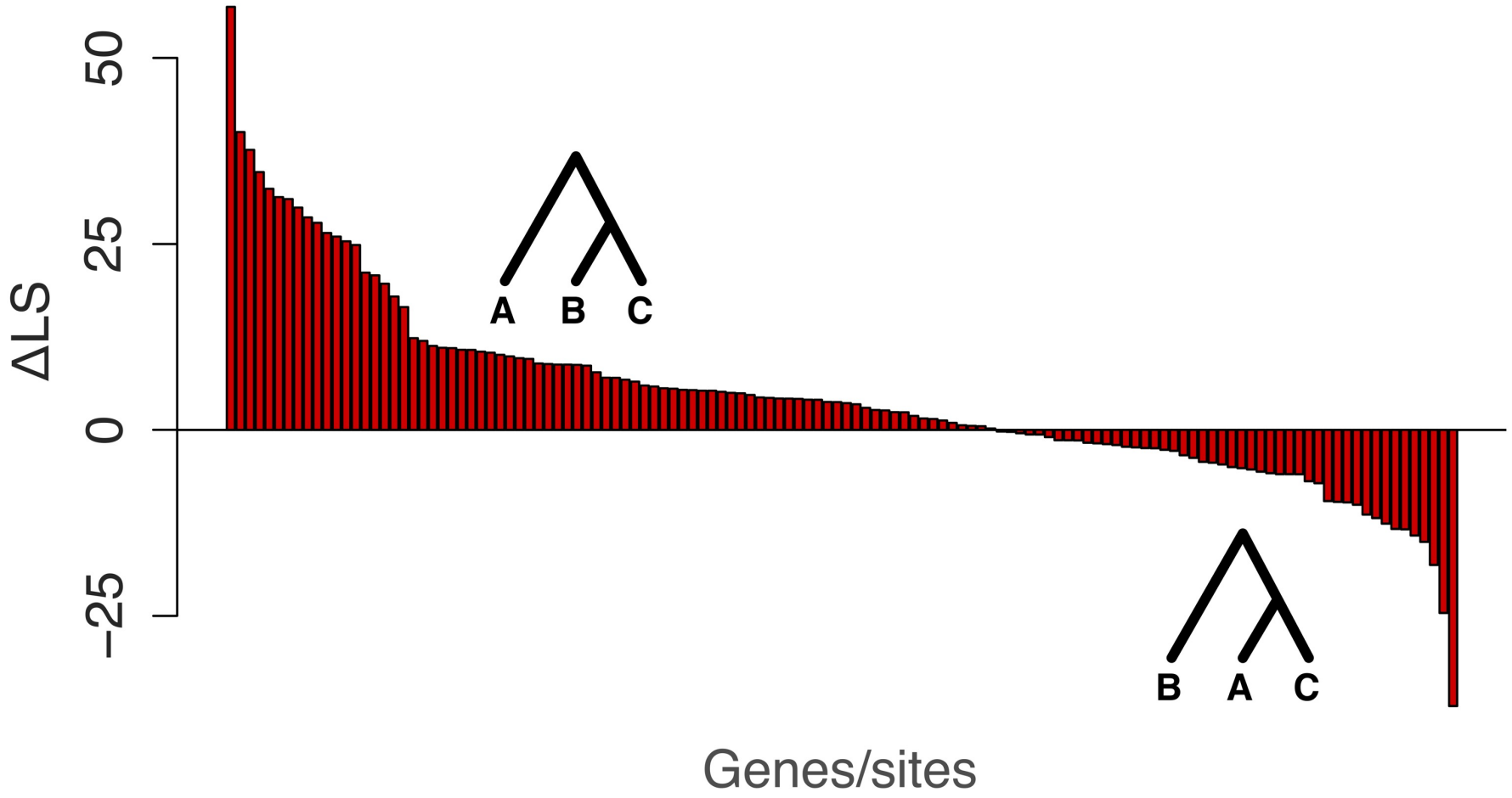






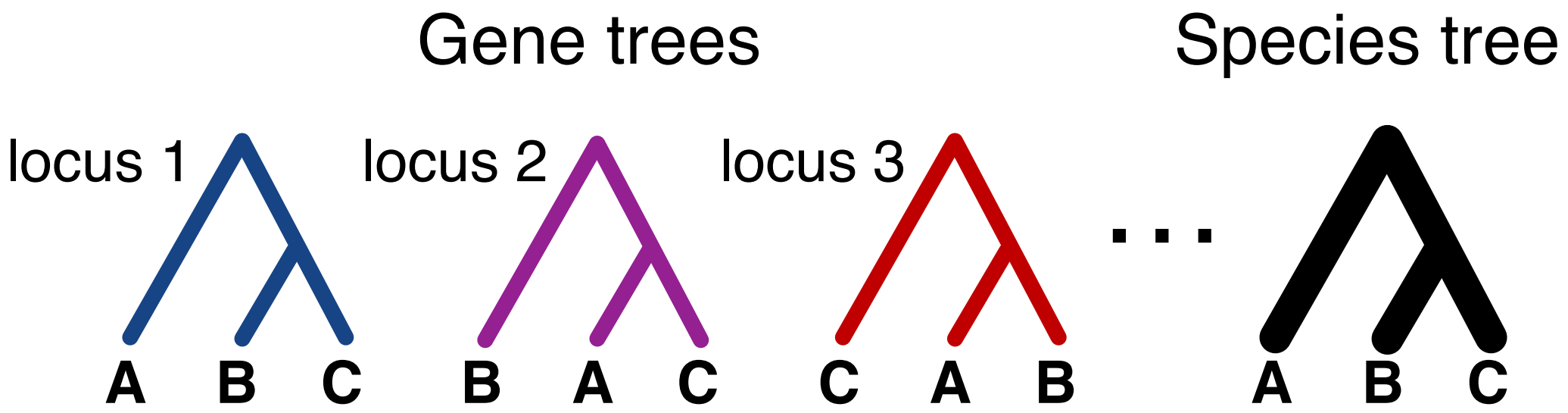
Phylogenetic signal

- **Inaccuracies or noise present in phylogenetic data can lead to erroneous conclusions about evolutionary relationships**
  - ILS, horizontal gene transfer, introgression?



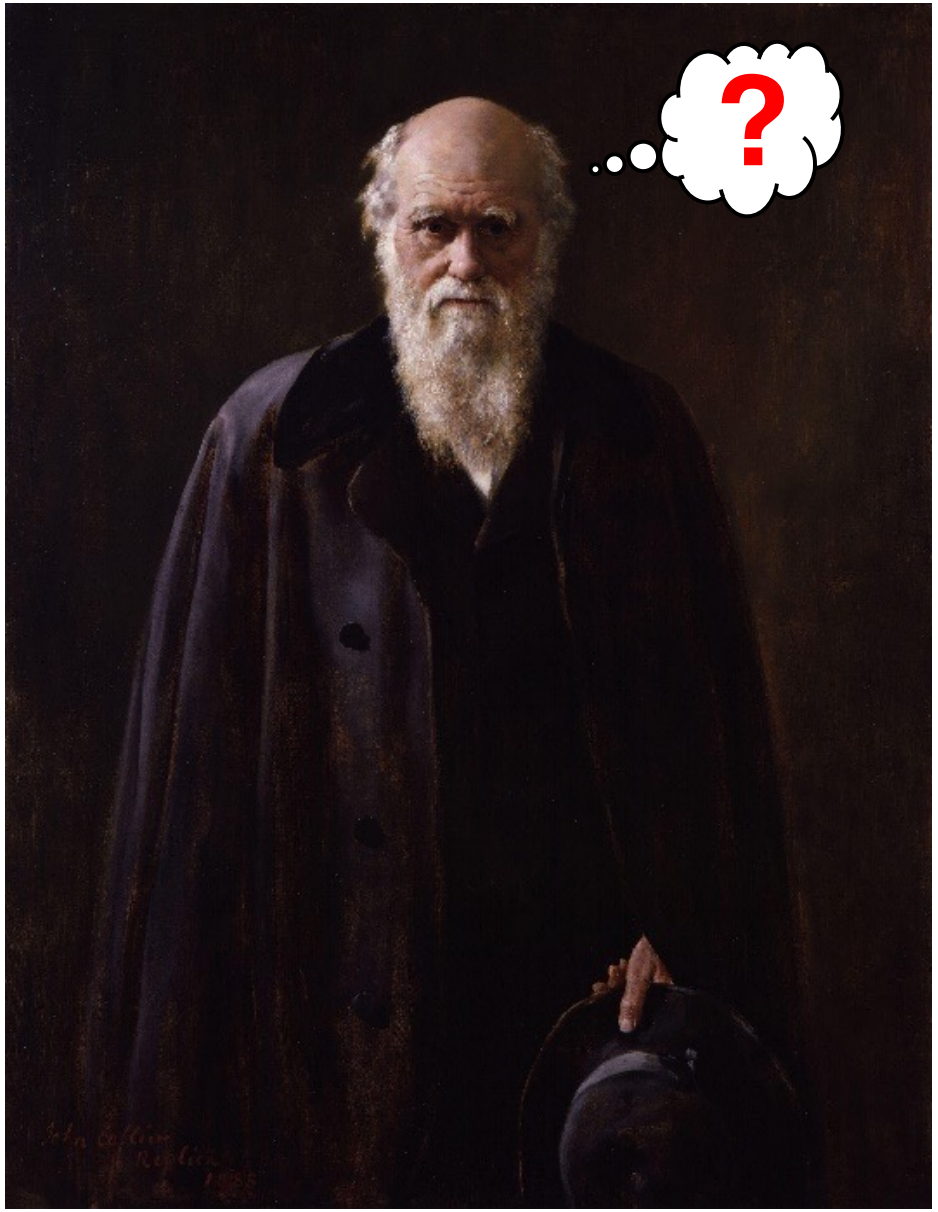
Gene-tree–species-tree incongruence

- **Loci are considered independent in terms of evolutionary history**
- **Trees inferred from concatenated genes and consensus of gene trees may conflict due to different histories of genes and/or systematic errors**



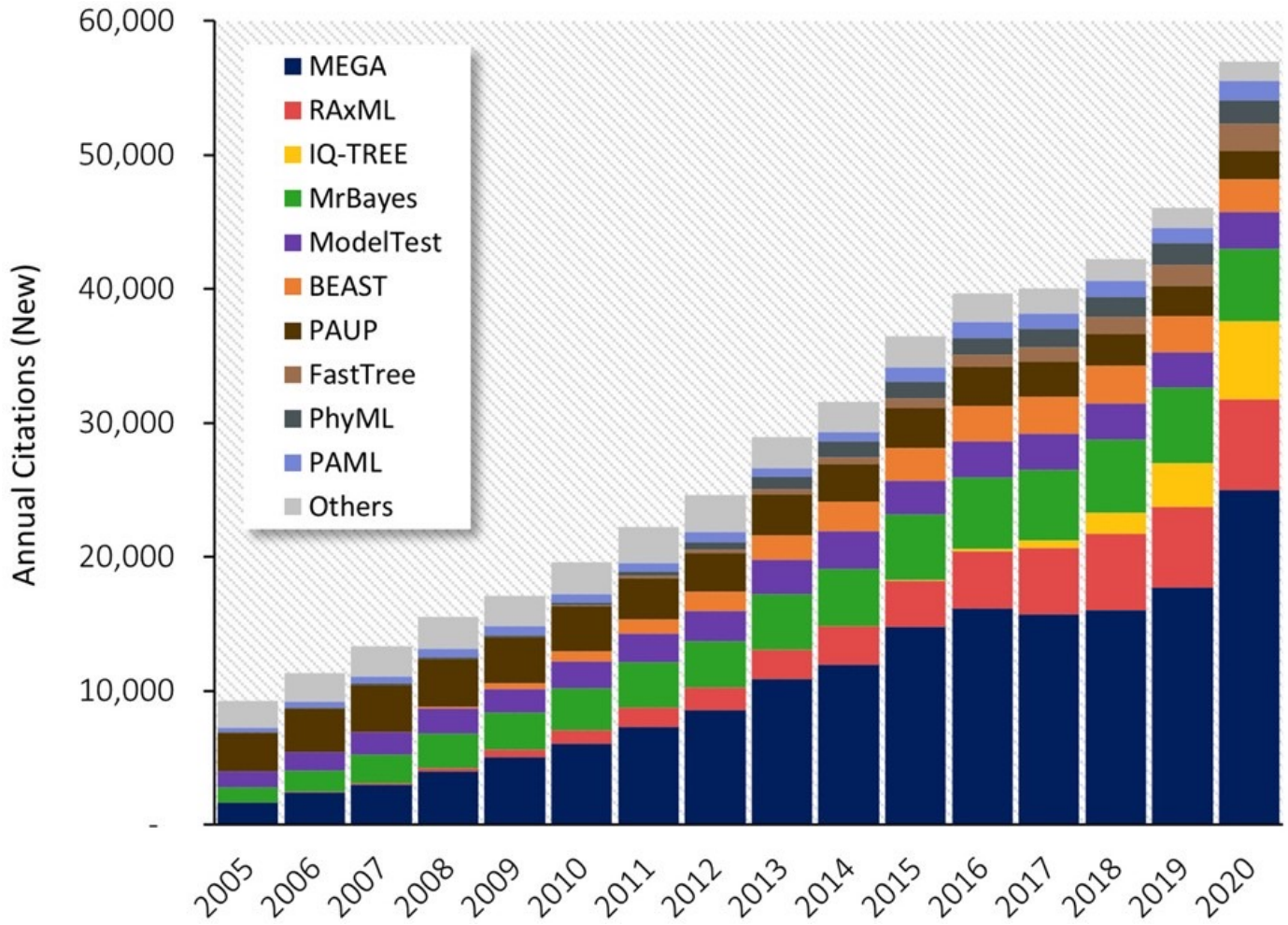
Tree reconstruction methods

- Distance matrix
- Maximum parsimony
- Maximum likelihood
- Bayesian inference



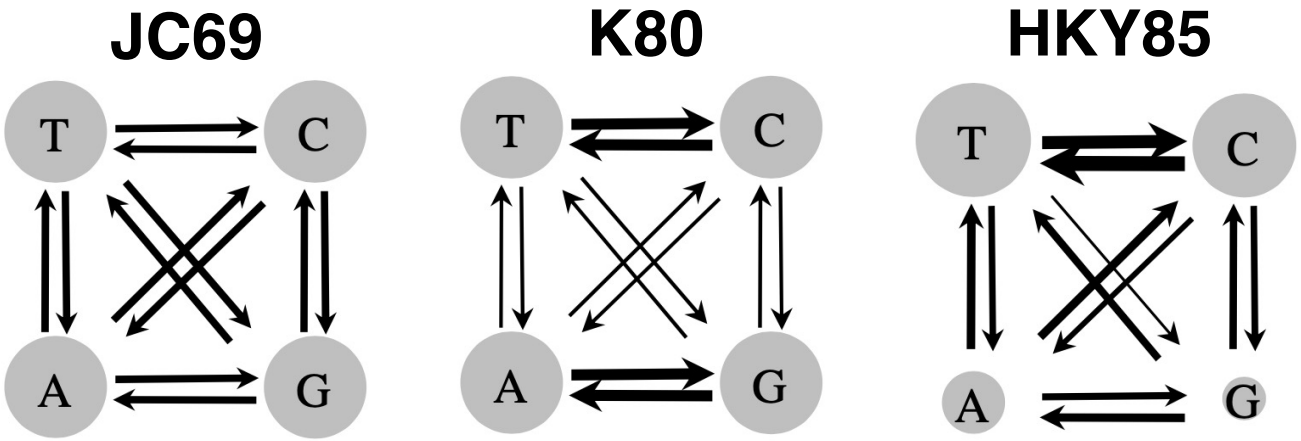
Phylogenetic inference methods

- **Annual counts of new research articles citing major software packages for molecular evolutionary and phylogenetic analyses**



Distance matrix methods

- **Distance calculation**
  - Pairwise sequence distances are calculated assuming a Markov chain model of nucleotide substitution (e.g., JC69, K80, HKY85)
  - Different sites in a DNA or protein sequence often evolve at different rates, such rate variation is accommodated by assuming a gamma (Γ) distribution of rates for sites
  - After the distances have been calculated, the sequence alignment is no longer used in distance matrix methods



Distance matrix methods

- **Least squares method**
  - Minimizes a measure of the differences between the calculated distances (dij) in the distance matrix and the expected distances (dˆ ij) on the tree (that is, the sum of branch lengths on the tree linking the two species i and j)
  - This is the same least squares method used in statistics for fitting a straight line y = a + bx to a scatterplot
  - Optimizing branch lengths (or dˆ ij) leads to the score Q for the given tree, and the tree with the smallest score is the least squares estimate of the true tree

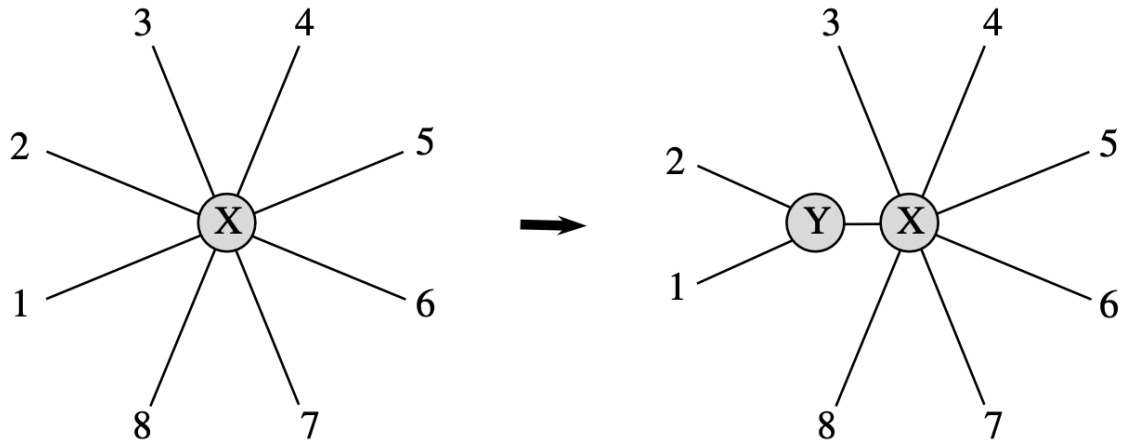
$$Q = \sum_{i=1}^{s} \sum_{i=1}^{s} (\hat{d}_{ij} - d_{ij})^2$$

Distance matrix methods

- **Minimum evolution method**
  - Uses the tree length (which is the sum of branch lengths) instead of Q for tree selection, even though the branch lengths can still be estimated using the least squares criterion
  - Under the minimum evolution criterion, shorter trees are more likely to be correct than longer trees are

Distance matrix methods

- **Neighbor joining (NJ)**
  - This is a cluster algorithm and operates by starting with a star tree and successively choosing a pair of taxa to join together (based on the taxon distances), until a fully resolved tree is obtained. The taxa to be joined are chosen to minimize an estimate of tree length
  - The two joined taxa (species 1 and 2) are then represented by their ancestor (node y), and the number of taxa that are connected to the root (node x is reduced by one). The distance matrix is then updated with the joined taxa replacing the two original taxa.



Pros of distance matrix methods

- One advantage of distance methods (especially of NJ) is their computational efficiency. The cluster algorithm is fast because it does not need to compare as many trees under an optimality criterion as maximum parsimony and maximum likelihood do.
- For this reason, NJ is useful for analyzing large data sets that have low levels of sequence divergence.

Cons of distance matrix methods

- A realistic substitution model should be used to calculate the pairwise distances
- Distance methods can perform poorly for very divergent sequences because large distances involve large sampling errors, and most distance methods do not account for the high variances of large distance estimates
- Distance methods are also sensitive to gaps in the sequence alignment

Maximum parsimony

- **Parsimony tree score**
  - The maximum parsimony method minimizes the number of changes on a phylogenetic tree by assigning character states to interior nodes on the tree
  - The character (or site) length is the minimum number of changes required for that site, whereas the tree score is the sum of character lengths over all sites.
  - The maximum parsimony tree is the tree that minimizes the tree score. Some sites are not useful for tree comparison by parsimony
  - For example, constant sites, for which the same nucleotide occurs in all species, have a character length of zero on any tree.
  - Singleton sites, at which only one of the species has a distinct nucleotide, whereas all others are the same, can also be ignored, as the character length is always one
  - The parsimony-informative sites are those at which at least two distinct characters are observed, each at least twice

Maximum parsimony

- During the late 1970s, it began to be applied to molecular data. A controversy arose concerning whether parsimony (without explicit assumptions) or likelihood (with an explicit evolutionary model) was a better method for phylogenetic analysis
- The controversy has subsided, and the importance of model-based inference methods is broadly recognized
- The use of parsimony is still common: not because it is believed to be assumption-free, but because it often produces reasonable results and is computationally efficient

Pros of maximum parsimony

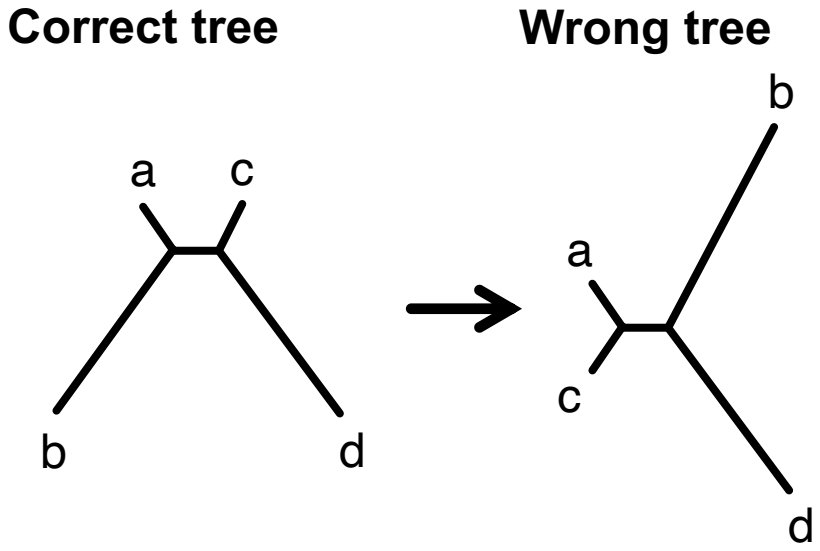
- A strength of parsimony is its simplicity
- It is easy to describe and to understand
- Well-suited to rigorous mathematical analysis
- This simplicity also helps in the development of efficient computer algorithms

Cons of maximum parsimony

- **A major weakness of parsimony is its lack of explicit assumptions, which makes it nearly impossible to incorporate any knowledge of the process of sequence evolution in tree reconstruction**
- **The failure of parsimony to correct for multiple substitutions at the same site makes it suffer from a problem known as long-branch attraction (LBA)**
- **The phenomenon of inferring an incorrect tree with long branches grouped together by parsimony or by model-based methods under simplistic models**

Long-branch attraction

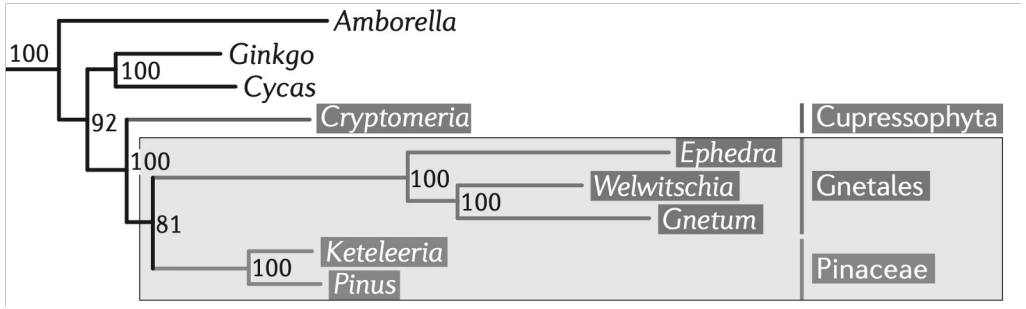
- **LBA is a phenomenon in phylogenetic reconstruction where distantly related taxa with long branches (i.e., taxa that have accumulated many evolutionary changes) are incorrectly inferred to be closely related due to their high rate of evolution**
- **This misplacement occurs because phylogenetic methods, may mistake convergent or parallel substitutions as evidence of shared ancestry**



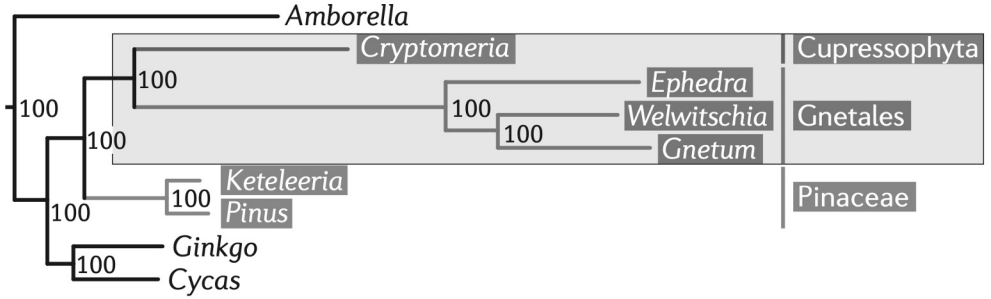
Long-branch attraction

- **LBA is a critical consideration in molecular phylogenetics, especially in studies involving ancient lineages or fast-evolving organisms**
- **If two taxa independently evolve rapidly, their sequences may appear similar due to coincidental substitutions. A phylogenetic analysis might then group these taxa together, even if they are not closely related**
- **Model-based methods (i.e., distance, ML and Bayesian) also suffer from LBA if the assumed model ignores among-site rate variation**

Correct tree



Wrong tree



Long-branch attraction

- LBA and unequal nucleotide or amino acid frequencies among species are an important source of systematic error in the reconstruction of deep phylogenies
- In such analyses, realistic substitution models and ML or Bayesian methods should be used
- Dense taxon sampling to break long branches and removing fast-evolving proteins or sites can also help

We will stop here and continue in Session 14

- **ML and Bayesian phylogenetic inference**
- **Molecular clock dating using phylogenomic data**
- **A tutorial for conducting a phylogenomic analysis using simulated data will be completed**
  - A comparison of phylogenetic methods will be conducted
  - A Bayesian timetree will be inferred

Maximum likelihood

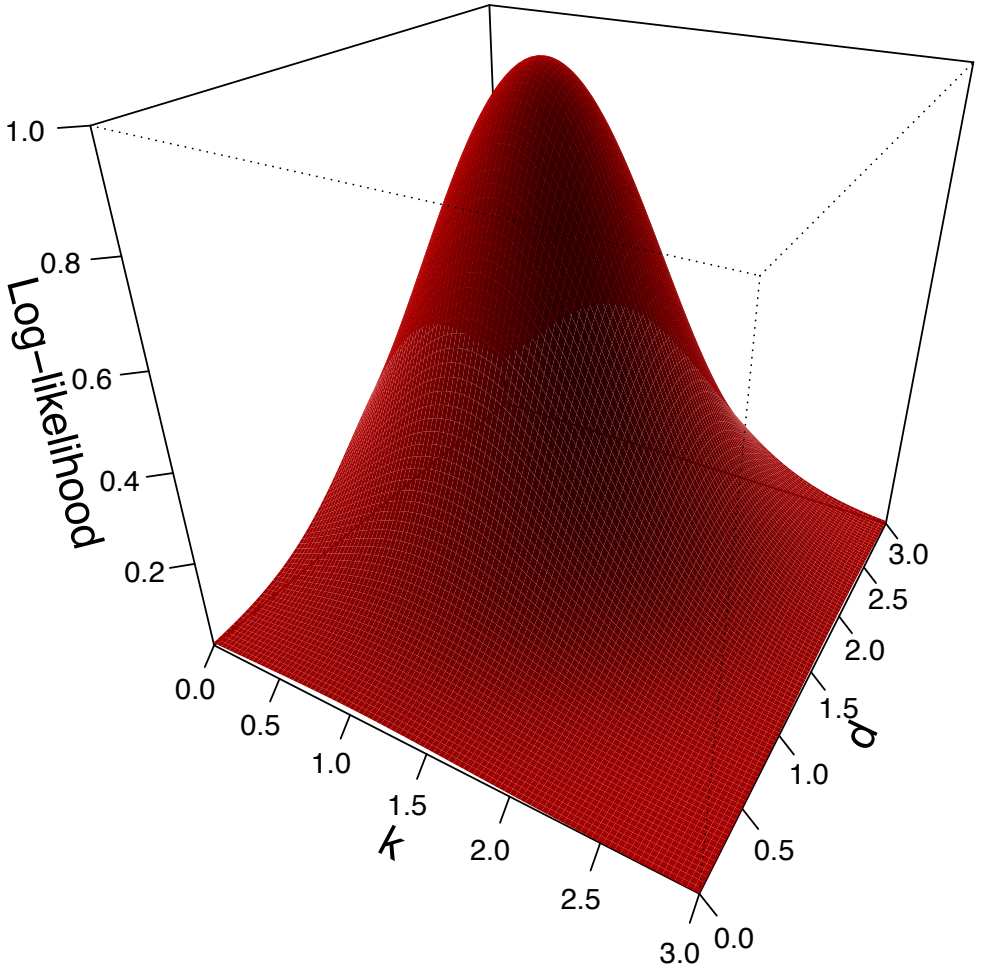
- **Basis of maximum likelihood (ML)**
  - ML was developed by R. A. Fisher in the 1920s as a statistical methodology for estimating unknown parameters in a model
  - The likelihood function is defined as the probability of the data given the parameters but is viewed as a function of the parameters with the data observed and fixed. It represents all information in the data about the parameters

$$f(D|\theta)$$

A likelihood function $f(D|\theta)$ describes the probability of data D given the values of parameters θ.

Maximum likelihood

- **The maximum MLEs of parameters are the parameter values that maximize the likelihood**
- **MLEs are found numerically using iterative optimization algorithms**
- **The MLEs have desirable asymptotic (large-sample) properties:**
  - Unbiased
  - Consistent (they approach the true values)
  - Efficient (they have the smallest variance among unbiased estimates)



Maximum likelihood

- **ML for tree reconstruction**
  - The first algorithm for ML analysis of DNA sequence data was developed by J. Felsenstein (1981)
  - Widely used due to the increased computing power and software implementations and to the development of increasingly realistic models of sequence evolution
  - Two optimization steps are involved in ML tree estimation:
    1. Optimization of branch lengths to calculate the tree score for each candidate tree
    2. Search in the tree space for the ML tree.

Maximum likelihood

- **ML for tree reconstruction**
  - The more probable the sequences given the tree, the more the tree is preferred
  - All possible trees are considered; computationally intense
  - Because the user can choose a model of evolution, the method can be useful for widely divergent groups or other difficult situations

$$P_M(D|T)$$

Likelihood provides probabilities of the sequences (***D***) given a model of their evolution (***M***) on a particular tree (***T***).

Data (*D*):
- Nucleotide sequences
- Amino acid sequences
- Codon sequences
- Morphological characters

Tree (*T*):
- Topology
- Branch lengths

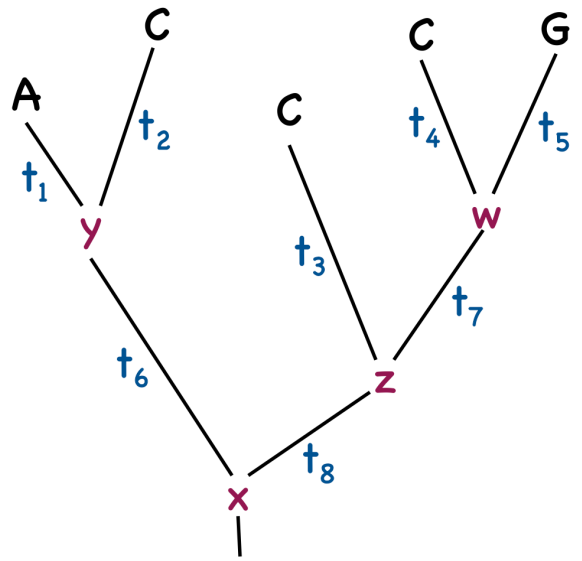
Model (*M*):
- Transition probability $P_{ij}(t)$

Maximum likelihood

- **How to find the ML tree?**
  - Objective: To find the combination of branch lengths that results in the highest likelihood (L) given a topology, and to evaluate the space of topologies, to obtain the one in which the branch lengths make the highest L, this one represents the ML tree.
    1. For each topology, find the combination of branch lengths that overall confer the maximum likelihood
    2. Explore the space of topologies
  - Assumptions:
    - To facilitate the computation of the tree likelihood, Felsenstein assumed that sites change independently from each other and across different branches. These assumptions, although unrealistic, remain foundational in most phylogenetic methods today

Maximum likelihood

- **How to calculate the L of a tree?**
  - Likelihood of a site = the sum of the probability of each possible combination of nucleotides on all internal nodes



$$\text{Prob}\,(D^{(i)} \mid T) = \sum_x \sum_y \sum_z \sum_w \text{Prob}\,(A, C, C, C, G, x, y, z, w \mid T)$$

$$\text{Prob}\,(D^{(i)} \mid T) = \text{Prob}(x)\ \text{Prob}(y \mid x, t_6)\ \text{Prob}(A \mid y, t_1)\ \text{Prob}(C \mid y, t_2)\ \text{Prob}(z \mid x, t_8)$$
$$\text{Prob}(C \mid z, t_3)\ \text{Prob}(w \mid z, t_7)\ \text{Prob}(C \mid w, t_4)\ \text{Prob}(G \mid w, t_5)$$

Maximum likelihood

- **How to calculate the L of a tree?**
  - Likelihood of a tree = product of the likelihood of all sites

$$L = \text{Prob}(D|T) = \prod \text{Prob}(D^{(i)}|T)$$

Total L
(L of the whole tree)

Product of the L of each site
(from 1 to n)

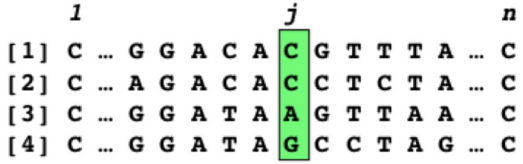
$D^{(i)}$ = given site i

Maximum likelihood

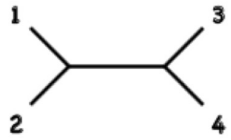
- **How to calculate the L of a tree?**
  - Likelihood of a site = the sum of the probability of each possible combination of nucleotides on all internal nodes

Exhaustive procedure

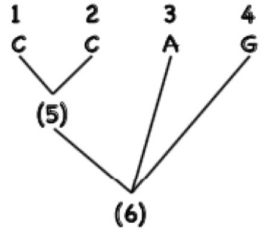
(1) Alignment



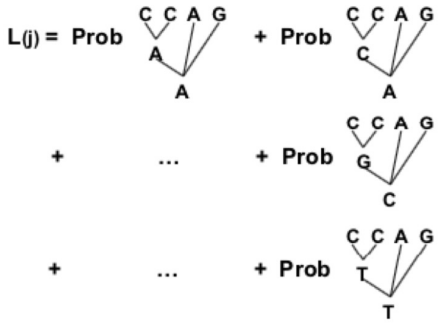
(2) Unrooted tree



(3) Arbitrary rooted tree



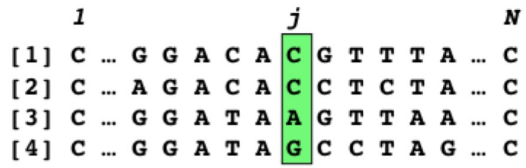
(4) L estimation at site j



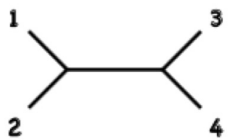
***Probability of change from one nucleotide to another. Determined by transition matrix P(t)**

$$P(t) = e^{Qt} = \begin{pmatrix} P_{A\text{-}A}(t) & P_{A\text{-}C}(t) & P_{A\text{-}G}(t) & P_{A\text{-}T}(t) \\ P_{C\text{-}A}(t) & P_{C\text{-}C}(t) & P_{C\text{-}G}(t) & P_{C\text{-}T}(t) \\ P_{G\text{-}A}(t) & P_{G\text{-}C}(t) & P_{G\text{-}G}(t) & P_{G\text{-}T}(t) \\ P_{T\text{-}A}(t) & P_{T\text{-}C}(t) & P_{T\text{-}G}(t) & P_{T\text{-}T}(t) \end{pmatrix}$$

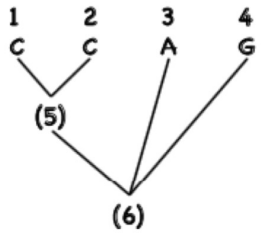
(1) Alignment



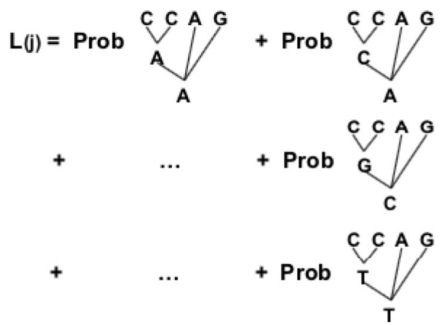
(2) Unrooted tree



(3) Arbitrary rooted tree



(4) L estimation at site j



(5) Total L = product of L from each site

$$L = L(1) \cdot L(2) \cdot \ldots L(n) = \prod_{j=1}^{N} L(j)$$

(6) L from each site and total L are expressed as lnL

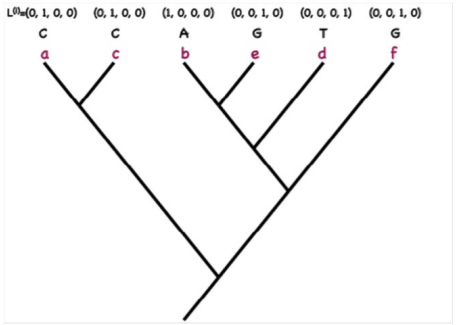
$$\ln L = \ln L(1) + \ln L(2) + \ldots + L(N) = \sum_{j=1}^{n} \ln L(j)$$

Maximum likelihood

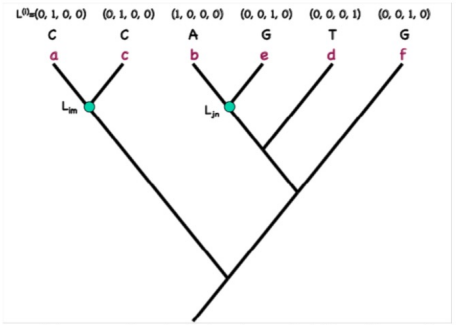
- **How to calculate the L of a tree?**
  - Likelihood of a site = the sum of the probability of each possible combination of nucleotides on all internal nodes
- **Exhaustive procedure:**
  - Number of calculations needed to obtain the likelihood for each site: (n = number of terminal taxa; n - 1 = number of internal nodes)
    - Nucleotide substitution: 4^{n-1}
    - Amino acid substitution: 20^{n-1}
    - Codon substitution: 61^{n-1}

Maximum likelihood

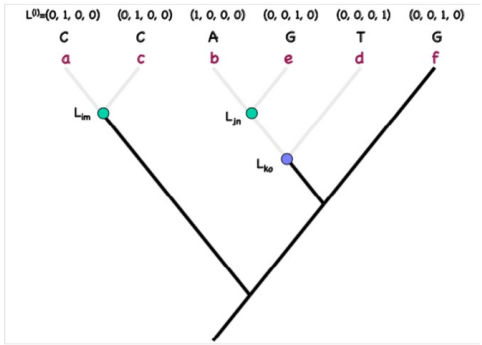
- **How to calculate the L of a tree?**
  - Likelihood of a site = the sum of the probability of each possible combination of nucleotides on all internal nodes
- **Pruning algorithm**



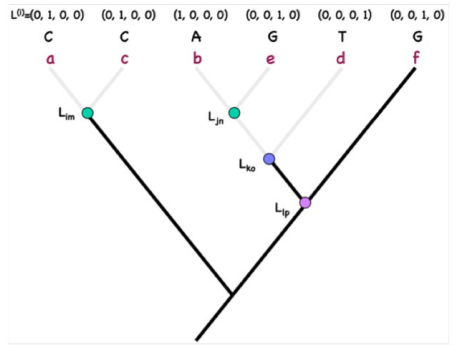
1) Obtain values of $L^{(i)}$ at the tips. Observed data. Assign prob =1 to the observed state, and prob= 0 to non observed states.



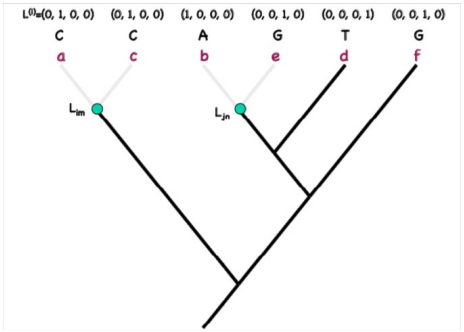
2) Proceed downwards calculating the probability of internal nodes, in which all descendants are terminal nodes. Use the method previously described.



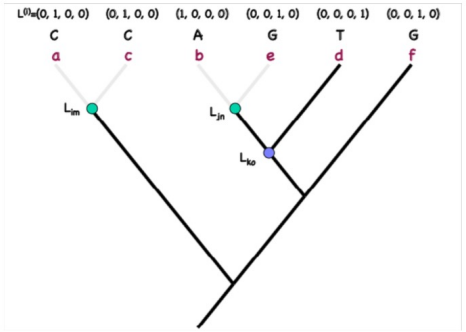
5) Continue descending through the tree.



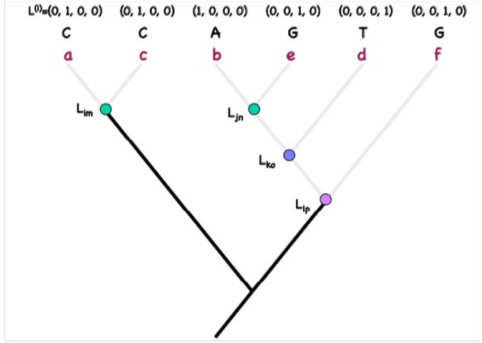
6) Continue descending through the tree.



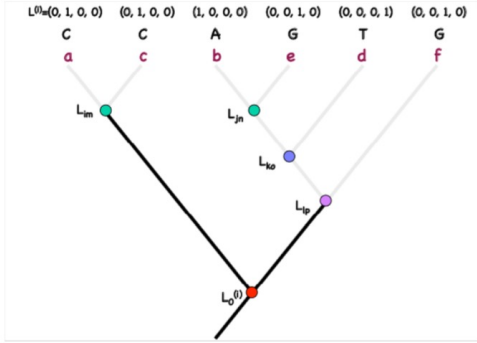
3) Once the conditional likelihoods (L_{im} and L_{jn}) are obtained, ignore their descendant terminal nodes. Proceed toward the base, using y the conditional likelihoods of the internal nodes to calculate those from deeper nodes.



4) Continue descending through the tree.



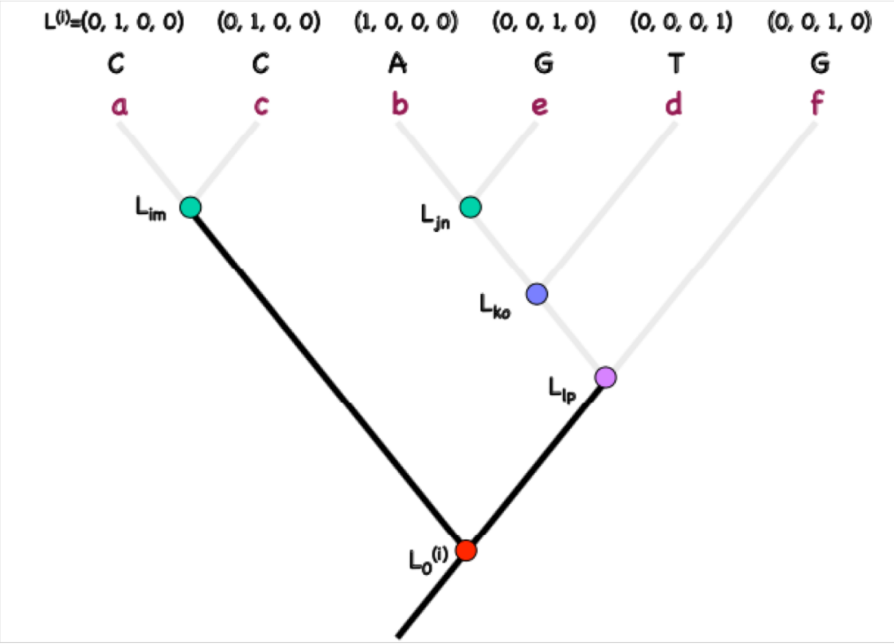
7) Continue descending through the tree.



8) Continue descending through the tree, until reaching to the root. The conditional likelihood obtained for this node ($L_0^{(i)}$) belongs to the likelihood of the tree.

Maximum likelihood

- **How to calculate the L of a tree?**
  - Likelihood of a site = the sum of the probability of each possible combination of nucleotides on all internal nodes
- **Pruning algorithm proceeding**



Proceeding

- For each site, it is calculated n-1 times (number of internal nodes)
- For each internal node, there are four calculation (number of characters)
- Each calculation is the product of two terms.
- Each term is the sum of four products.

$p(n-1)^{b2}$
p = number of sites
n = number of taxa
b = number of bases

Maximum likelihood

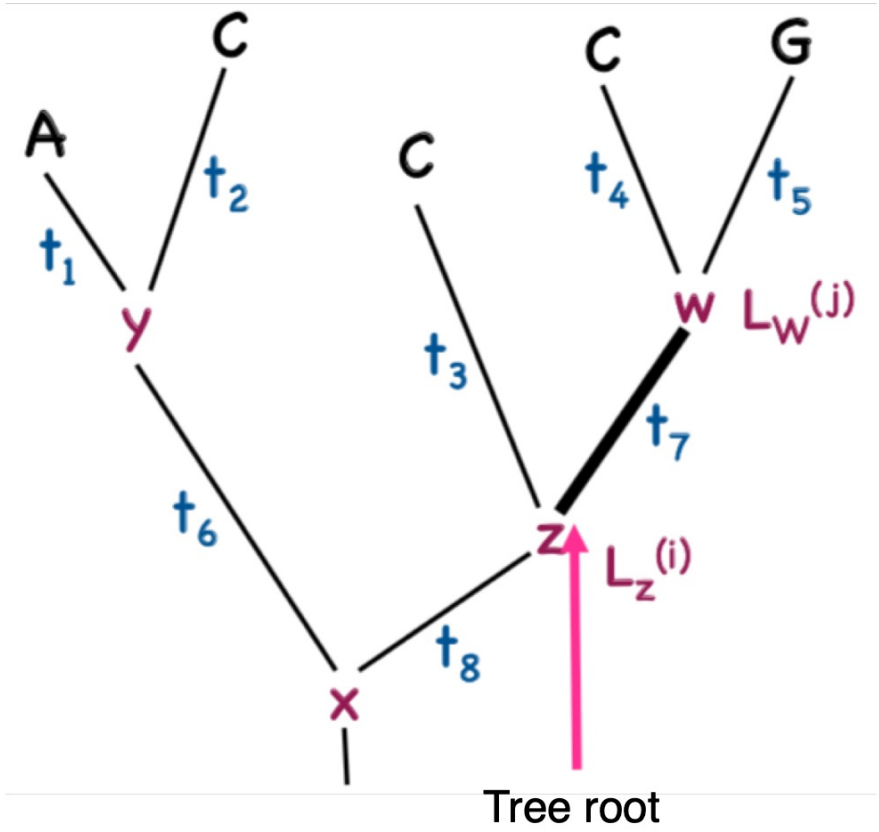
- **How to find the ML tree?**
  - Exhaustive proceeding:

Giving a substitution model

1. Start with a topology with branch lengths and values for the parameters of the model.
2. Calculate the L of that combination of topology, branch lengths and model parameter values.
3. Make a small modification to the branches, keeping everything else constant. Obtain the the value of L
4. Continue making small changes on the branch lengths and measuring the L, until reaching a combination that results on the highest L for that topology and the parameter values.
5. Make a small modification to the parameter values and modify the the branch lengths until obtaining a combination with the highest L
6. Make changes to the topology and modify the branch lengths and parameter values until identifying a combination of topology, branch lengths and parameter values that have the ML

Maximum likelihood

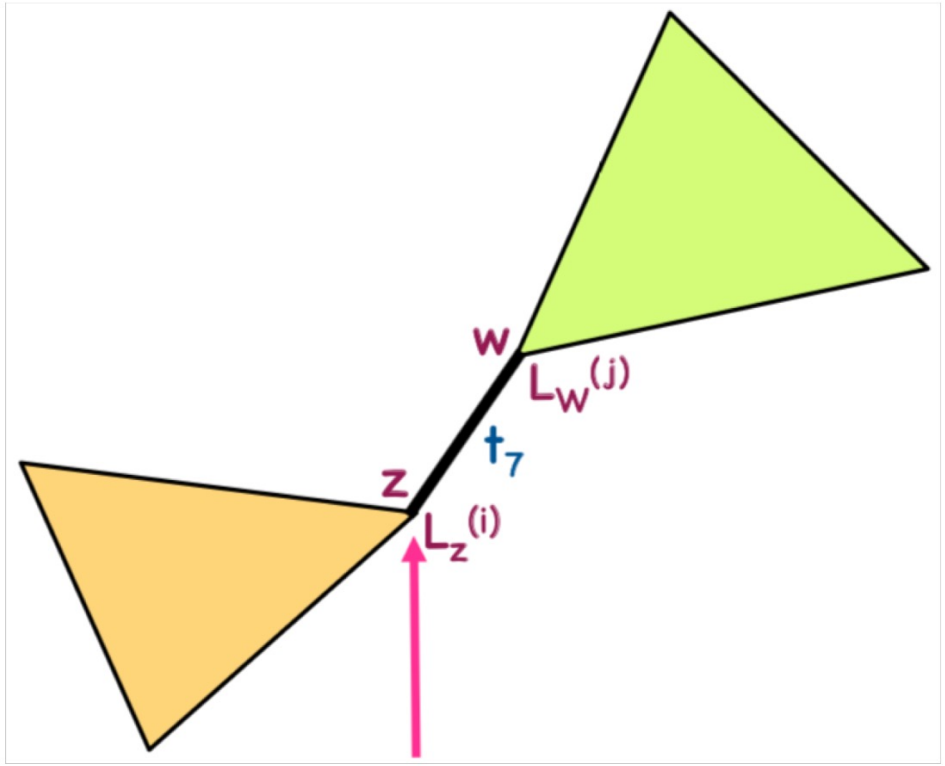
- **How to find the ML tree?**
  1. **<u>For each topology, find the combination of branch lengths that overall confer the maximum likelihood</u>**
  2. Explore the space of topologies



- Example: find the branch length t_7 that maximizes the L.
- Thanks to the reversibility of the model the tree root can be located on any point, without changing the L.
- Find the root between **z** and **w**, separated from **z** by a new branch with length cero.
- Using the pruning algorithm, obtain the conditional likelihood from node **z** and **w**.

Maximum likelihood

- **How to find the ML tree?**
  1. **<u>For each topology, find the combination of branch lengths that overall confer the maximum likelihood</u>**
  2. Explore the space of topologies



- Maximize the L of branch t_7 in a two terminals tree. (e.g. through numeric analysis of the successive approximations, or algorithms EM [expectation-maximization].
- Re-optimize the lengths within the subtrees, and optimize again the branch $\mathbf{t_7}$.
- In each step the tree L increases. Convergence is reached after a few cycles.

Maximum likelihood

- **How to find the ML tree?**
  1. **<u>For each topology, find the combination of branch lengths that overall confer the maximum likelihood</u>**
  2. Explore the space of topologies
  – <u>The pulley principle</u> allows to construct an algorithm which alters one of the branch (vi) at a time, each one being altered to that value which results in the highest likelihood. This process continues until none of the vi can be altered in a way which substantially improves the likelihood.

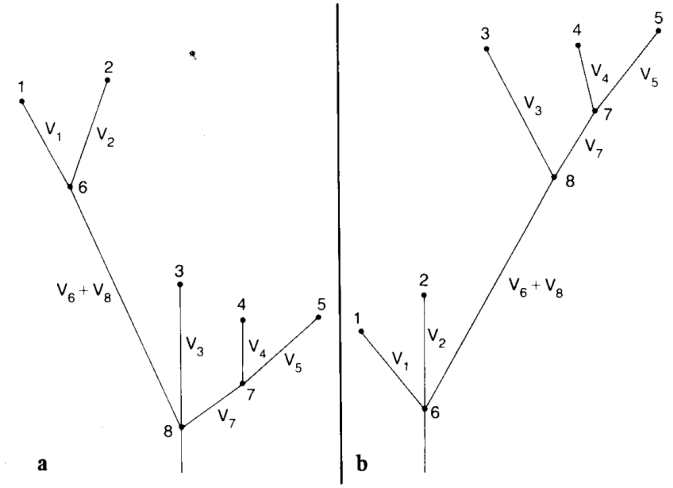


Fig. 2. Two trees whose likelihood will be equivalent to that in Fig 1 under the assumption of this paper

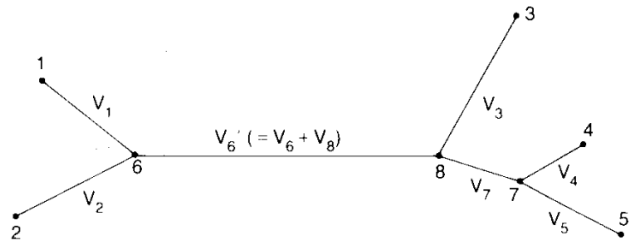


Fig. 3. The unrooted tree whose likelihood is equivalent to those in Fig 1 and 2

Maximum likelihood

- **How to find the ML tree?**
  1. For each topology, find the combination of branch lengths that overall confer the maximum likelihood
  2. **Explore the space of topologies**
     - Exhaustive methods:
       - Searching among tree space (topology): The number of unrooted bifurcating trees with n labelled tips is (2n-5)!/ [(n-3)! 2n'3], which for 10 tips is more than 2 million topologies
     - Heuristic methods:
       - NNI
       - SPR
       - TBR

Maximum likelihood

- From a statistical point of view, the tree (topology) is a model instead of a parameter, whereas branch lengths on the given tree and substitution parameters are parameters in the model
- ML tree inference is thus equivalent to comparing many statistical models, each with the same number of parameters
- The attractive asymptotic properties of MLEs apply to parameter estimation when the true tree is given but not to the ML tree

Pros and cons of ML methods

- **One advantage of the maximum likelihood method is that all of its model assumptions are explicit, so that they can be evaluated and improved**
- **The availability of a many evolutionary models in the likelihood and Bayesian method is one of its major advantages over MP**
- **ML has a clear advantage over distance or parsimony methods if the aim is to understand the process of sequence evolution**
- **The likelihood ratio test can be used to examine the fit of evolutionary models and to test interesting biological hypotheses**
  - E.g., the molecular clock and Darwinian selection affecting protein evolution

Cons of ML methods

- **The main drawback of maximum likelihood is that the likelihood calculation and tree search under the likelihood criterion is computationally demanding**
- **Another drawback is that the method has potentially poor statistical properties if the model is misspecified. This is also true for Bayesian analysis**

Bayesian inference

- **Basis of Bayesian inference**
  - It differs from maximum likelihood in that parameters in the model are considered to be random variables with statistical distributions, whereas in maximum likelihood they are unknown fixed constants
  - Before the analysis of the data, parameters are assigned a prior distribution, which is combined with the data (or likelihood) to generate the posterior distribution. All inferences concerning the parameters are then based on the posterior distribution
  - Bayesian inference has gained popularity thanks to advances in computational methods, especially Markov chain Monte Carlo algorithms (MCMC algorithms)

Bayesian inference

- **Bayesian inference relies on Bayes's theorem**
  - States that where P(T,θ) is the prior probability for tree T and parameter θ, P(DlT,θ) is the likelihood or probability of the data given the tree and parameter, and P(T,θlD) is the posterior probability.
  - The denominator P(D) is a normalizing constant, as its role is to ensure that P(T,θlD) sums over the trees and integrates over the parameters to one.
  - The theorem states that the posterior is proportional to the prior times the likelihood, or the posterior information is the prior information plus the data information

$$\mathrm{P}(T,\theta|D) = \frac{\mathrm{P}(T,\theta)\mathrm{P}(D|T,\theta)}{\mathrm{P}(D)}$$

Bayesian inference

- **Bayesian inference relies on Bayes's theorem**
  - In general, the posterior probabilities of trees cannot be directly calculated. Normalizing constant P(D) involves high-dimensional integrals (over all possible parameter θ values) and summation over all possible trees. Instead,
  - Bayesian phylogenetic inference relies on MCMC algorithms to generate a sample from the posterior distribution

$$\mathrm{P}(T,\theta|D) = \frac{\mathrm{P}(T,\theta)\mathrm{P}(D|T,\theta)}{\boxed{\mathrm{P}(D)}}$$

Pros and cons of Bayesian inference

- **Both ML and Bayesian methods use the likelihood function and thus share many statistical properties, such as consistency and efficiency**
- **However, ML and Bayesian inference represent opposing philosophies of statistical inference.**
- **The same feature of Bayesian inference may thus be viewed as either a strength or a weakness**

Pros and cons of Bayesian inference

- **Bayesian statistics is known to answer the biological questions directly and yields results that are easy to interpret:**
  - The posterior probability of a tree is just the probability that the tree is correct, given the data and model
- **In phylogenetics, it has not been possible to define a confidence interval for the tree. The widely used bootstrap method in ML has been difficult to interpret despite numerous efforts**

Pros and cons of Bayesian inference

- **However, posterior probabilities for trees and clades that have been calculated from real data sets often appear to be too high.**
  - In many analyses, nearly all nodes had posterior probabilities of ~100%.
- **Posterior tree probabilities are also sensitive to model violations, and use of simplistic models may lead to inflated posterior probabilities**

Pros and cons of Bayesian inference

- The prior probability allows incorporation of a priori information about the trees or parameters
- However, such information is rarely available, and specification of the prior is most often a burden on the user. Almost all data analyses are conducted using the default priors in the program
- High-dimensional priors are hard to specify, and an innocent-looking prior can have an undue and unexpected influence on the posterior
- It is therefore important to conduct Bayesian robustness analysis to assess the impact of the prior on the posterior estimates

Supertree and supermatrix approaches

- **Two approaches have been advocated for the phylogenetic analysis of hundreds or thousands of genes or proteins, especially when some loci are missing in some species**
- **The supertree approach separately analyses each gene and then uses heuristic algorithms to assemble the subtrees for individual genes into a supertree for all species**
- **In the supermatrix approach, sequences for multiple genes are concatenated to generate a data supermatrix, in which missing data are replaced by question marks, and the supermatrix is then used for tree reconstruction**
  - Most supermatrix analyses ignore differences in evolutionary dynamics among genes

Supertree and supermatrix approaches

- **A supermatrix analysis that assumes different evolutionary models and different trees and branch lengths for the genes is equivalent to a supertree analysis**
- **When a common tree underlies all genes, the ideal approach should be a combined (supermatrix) analysis of all genes, using the likelihood to accommodate the among-gene heterogeneity in the evolutionary process**

Missing data

- **Many genomic data sets are highly incomplete, and so most cells in the species by gene matrix will be empty**
- **Although, in theory, the likelihood function (in the ML and Bayesian methods) can properly accommodate missing data, the impact of such large-scale missing data and alignment gaps remains unclear**
- **Simulations suggest that ML and Bayesian inference generally perform better than NJ or MP in dealing with missing data, and Bayesian inference was found to perform the best**

Systemic errors

- **In the analysis of very large data sets, almost all bootstrap support values or Bayesian posterior probabilities are calculated to be 100%, even though the inferred phylogenies might be conflicting across genes or might depend on the method and model used**
- **Systematic biases are thus much more important than random sampling errors in such analyses, and methods that are robust to violations of model assumptions, even if they are less efficient, should be preferable**

Data-partitioning strategies

- **The purpose of data partitioning is to group genes or sites with similar evolutionary characteristics into the same partition so that all sites in the same partition are described using the same model, and different partitions use different models**
- **Partitioning too finely increases computation time and can cause overfitting, but partitioning too coarsely may lead to underfitting or model violation**

Data-partitioning strategies

- **<u>Mixture models</u>: Some models allow random variation among sites in substitution rate, in amino acid frequencies or in the pattern of substitution**
  - Use a statistical distribution to accommodate the among-site heterogeneity without data partitioning
  - The choice between using partition or mixture models is often philosophical

Data-partitioning strategies

- Current strategies for data partitioning include partitioning genes according to their relative substitution rates and separating the three codon positions in coding genes into different partitions
- The likelihood ratio test has also been used to decide whether two genes should be in the same or different partitions
- In summary, data partitioning should rely on our knowledge of the biological system: for example, on whether it is reasonable to assume that the same phylogeny underlies all genes

Statistical evaluation of phylogenetic methods

- **The aim of phylogenetic inference is to estimate tree topology and branch lengths**
- **Four criteria are used to evaluate tree reconstruction methods:**
  - Consistency: An estimation method is said to be consistent if the estimate converges to the true parameter value when the amount of data approaches infinity
  - Efficiency: In the statistical estimation of a parameter, an unbiased estimate with a smaller variance is more efficient than one with a larger variance
  - Robustness: A method is robust if it gives correct answers even when its assumptions are violated
  - Computational speed: Cluster algorithms are faster that criterion algorithms