Introduction to phylogenetics

Seong-Hwan Jun

What is phylogenetics?

- Reconstructing evolutionary history of **homologous** organisms (taxa) on **heritable** traits.
- What does it mean to reconstruct evolutionary histories?
  - Given a set of measurements from organisms, we want to reason over possible evolutionary tree that generated the observations.
  - For example, COVID is an RNA virus. It mutates to adjust to the environment, resulting in new variants that may differ significantly from the original virus. Knowing exactly what mutations took place can yield insight into the development of vaccine or treatment.

Application: Viral infection

- Nextstrain:
  https://nextstrain.org/ncov/gisaid/global/6m?l=clock
  - Compare clock trees for ncov vs flu.
  - Run the animation.

Application: Cancer phylogenetics.

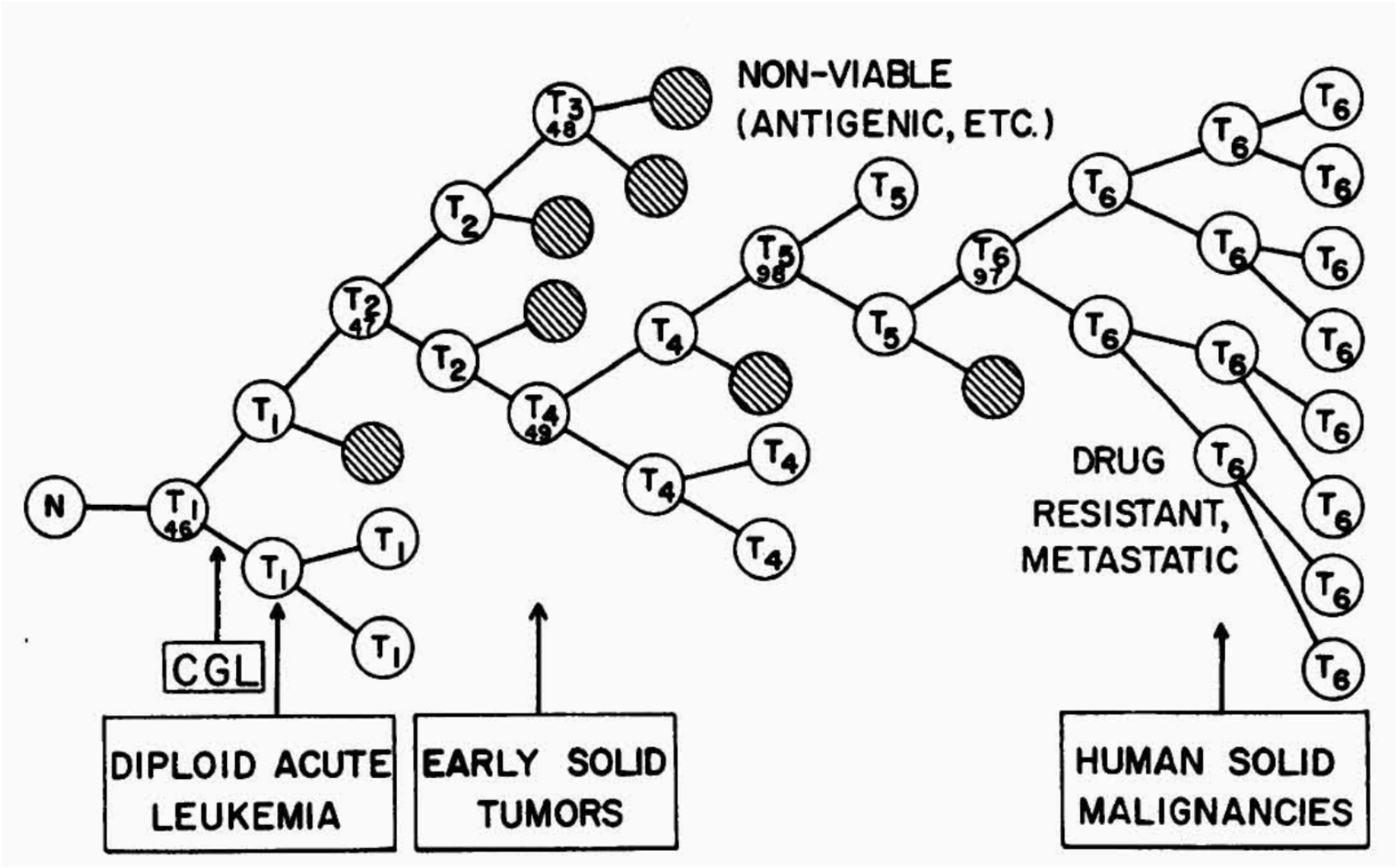


Figure 1: Nowell (1976)

Models of tumor evolution

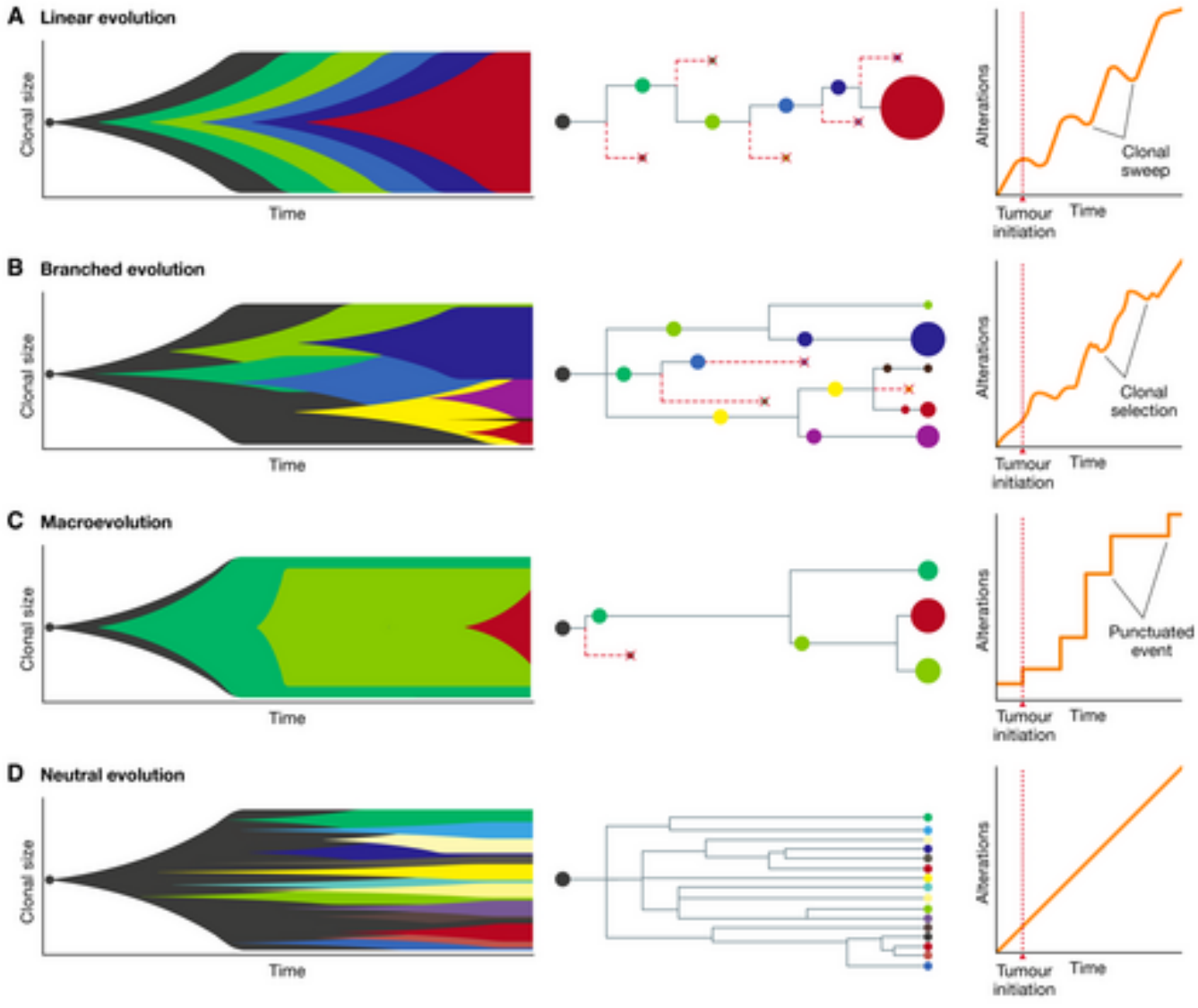
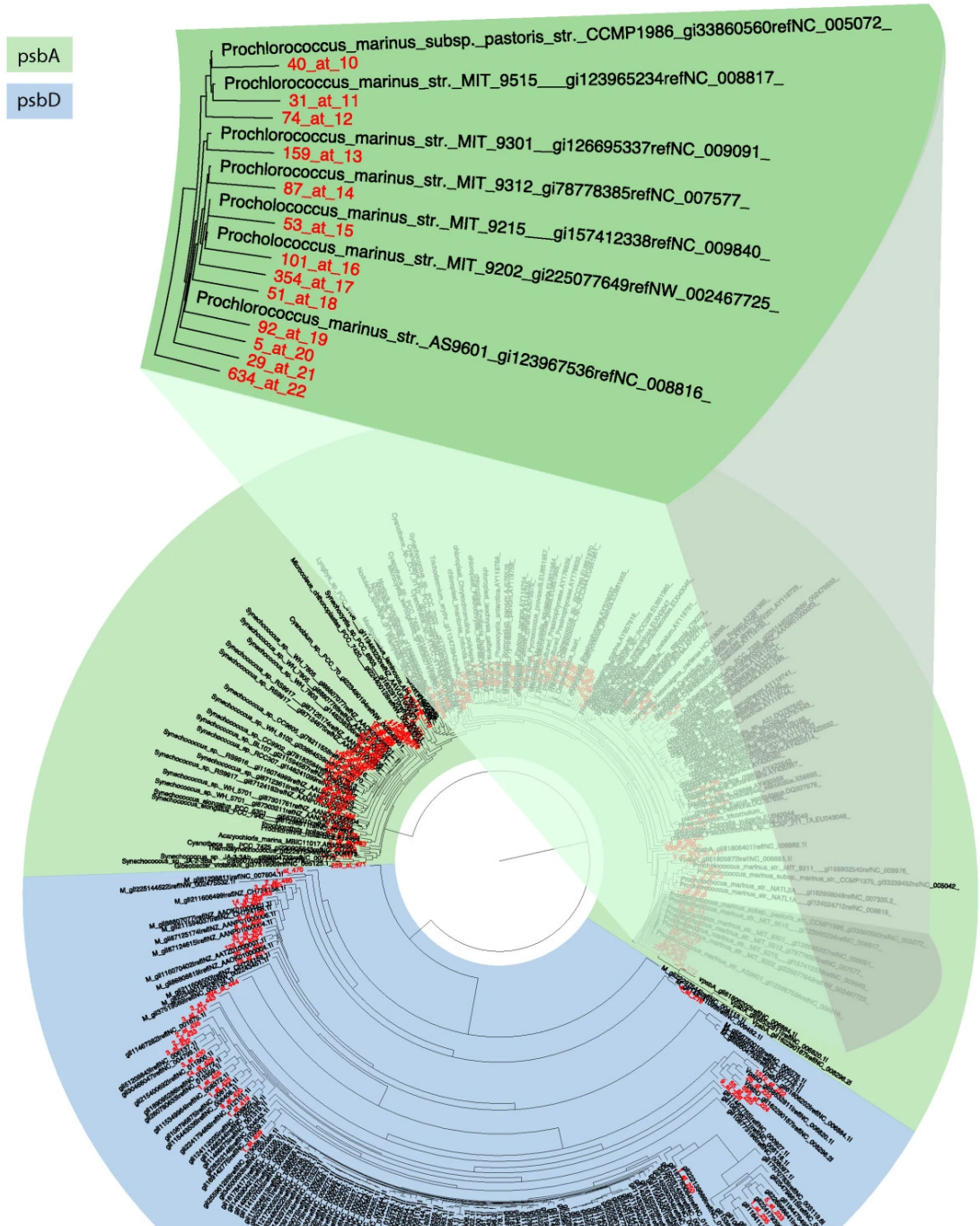


Figure 2: Vendramin, Litchfield, and Swanton (2021)

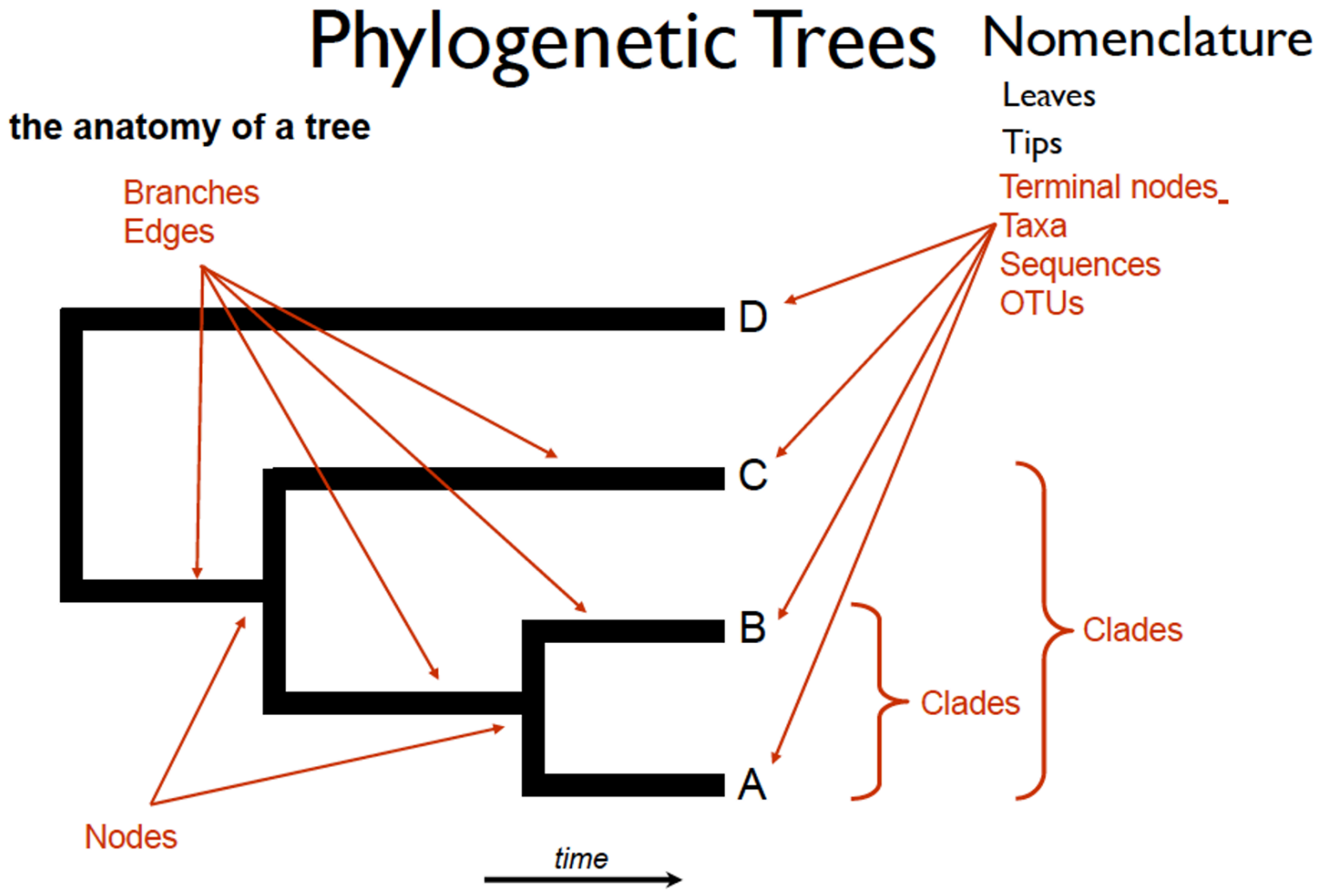
Application: Microbiome analysis

- Reconstruct phylogenetic tree from amplicon sequencing data (16S rRNA).
- Map observed microbiome sequences to reference phylogenetic tree (Matsen, Kodner, and Armbrust 2010).

Application: Microbiome analysis



Trees


Phylogenetic Trees
Nomenclature
Leaves
Tips
Terminal nodes
Taxa
Sequences
OTUs
the anatomy of a tree
Branches
Edges
D
C
B
A
Clades
Clades
Nodes
time

Trees

- $T = (V, E)$.
- V: Set of nodes.
- $X \subset V$: Set of observed taxa (leaf nodes).
- y: Set of sequence for observed data X.
- E: Set of edges.
- $b : E \to [0, \infty)$: branch lengths.
- The branch lengths typically express expected number of substitutions per site, i.e., amount of evolution that took place from parent to child.

Trees

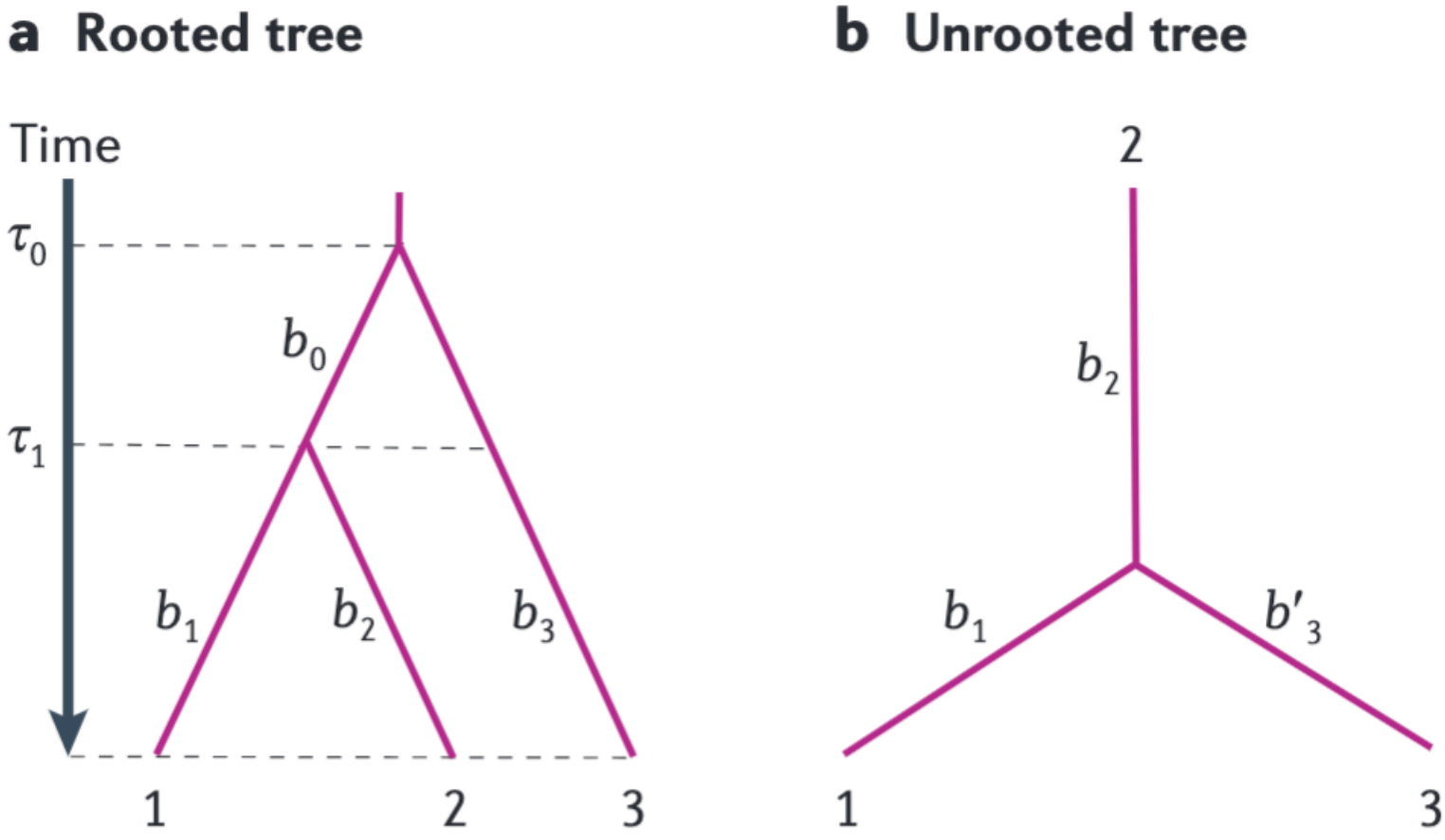


Figure 4: Yang and Rannala (2012)

What makes phylogenetic analysis challenging?

- The number of **rooted** (resp. **unrooted**) binary tree topologies with N leaves is given by $(2N-3)!!$ (resp. $(2N-5)!!$).

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

- The number of possible unrooted (resp. rooted) topologies for $N = 10$ is $2,027,025$ (resp. $34,459,425$). The growth is super-exponential (grows faster than a^N for all $a > 1$).

What makes phylogenetic analysis challenging?

- How to find the best tree? Enumeration is difficult when the number of taxa is large.
- How do you define the best tree?
- How to quantify uncertainty and perform statistical tests of hypotheses?

Distance based tree reconstruction

- Align sequences from all taxa under consideration.
- Compute the distance between pairs of sequences.
- Reconstruct the tree.

How do we calculate the distance between a pair of sequences?

Taxon1: CTTAGGTTTAAG
Taxon2: CTTAGGTTTTAG
Taxon3: ATTAGGTTTAAG
Taxon4: ATTAGGTTGAAG
Taxon5: ATTAAGGTTAAG
Taxon6: ATTAGGGTTAAG

Hamming distance?

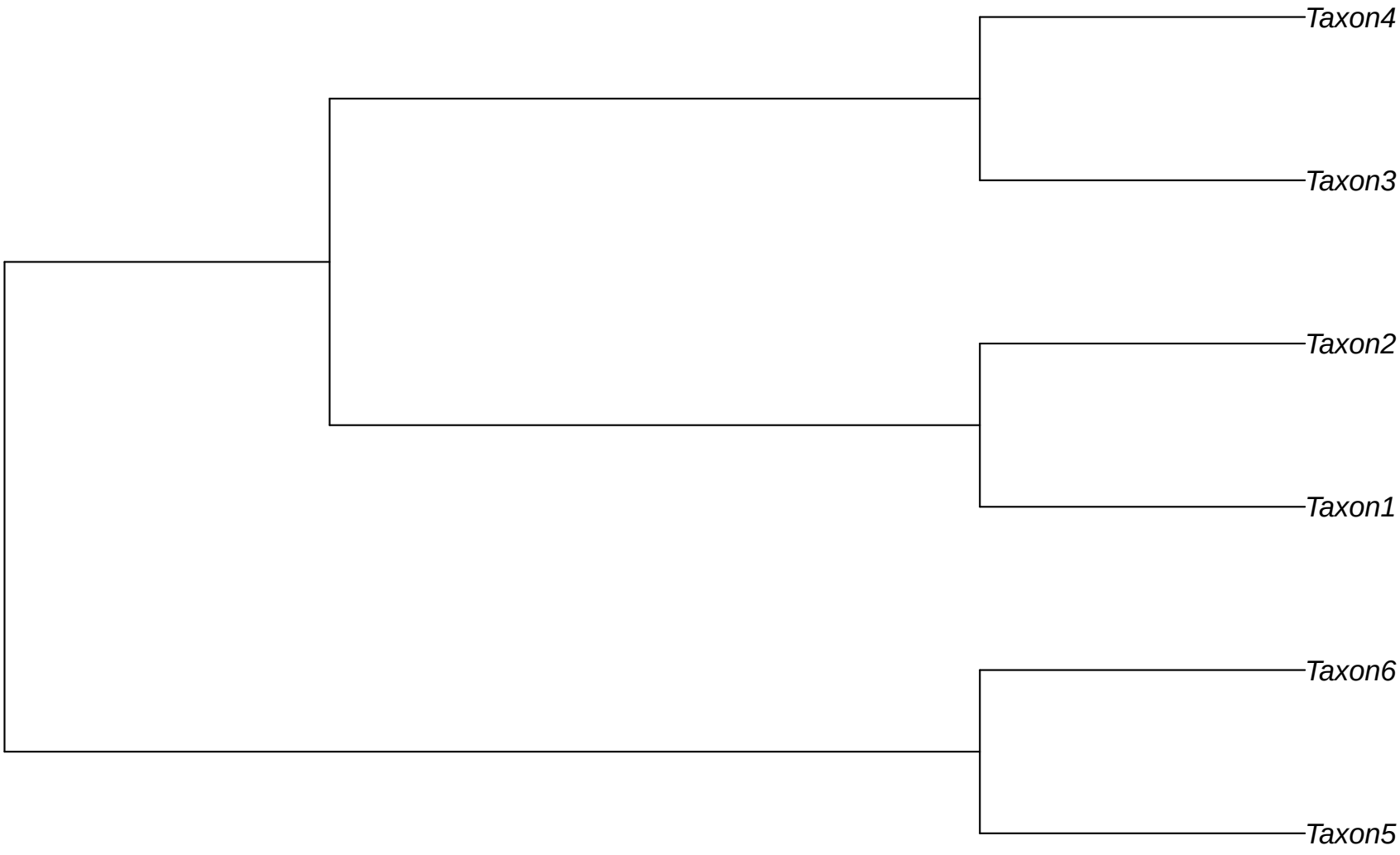
Taxon1: CTTAGGTTTAAG
Taxon2: CTTAGGTTTTAG
Taxon3: ATTAGGTTTAAG
Taxon4: ATTAGGTTGAAG
Taxon5: ATTAAGGTTAAG
Taxon6: ATTAGGGTTAAG

	$T1$	$T2$	$T3$	$T4$	$T5$	$T6$
$T1$	0	1	1	2	3	2
$T2$	–	0	2	3	4	3
$T3$	–	–	0	1	2	1
$T4$	–	–	–	0	3	2
$T5$	–	–	–	–	0	1

(1)

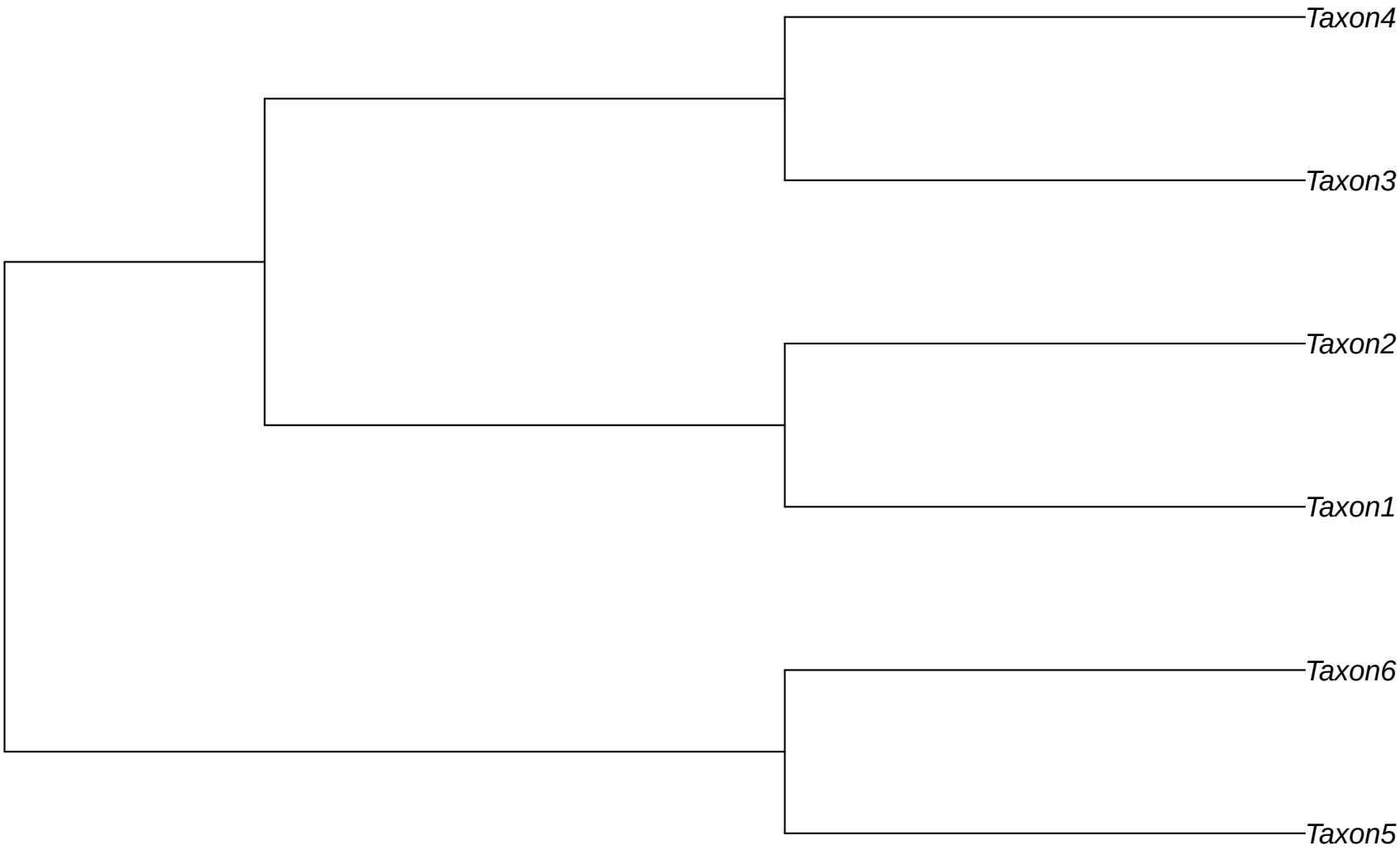
Complete linkage

Taxon1: CTTAGGTTTAAG
Taxon2: CTTAGGTTTTAG
Taxon3: ATTAGGTTTAAG
Taxon4: ATTAGGTTGAAG
Taxon5: ATTAAGGTTAAG
Taxon6: ATTAGGGTTAAG



Average linkage

Taxon1: CTTAGGTTTAAG
Taxon2: CTTAGGTTTTAG
Taxon3: ATTAGGTTTAAG
Taxon4: ATTAGGTTGAAG
Taxon5: ATTAAGGTTAAG
Taxon6: ATTAGGGTTAAG



Neighbor joining

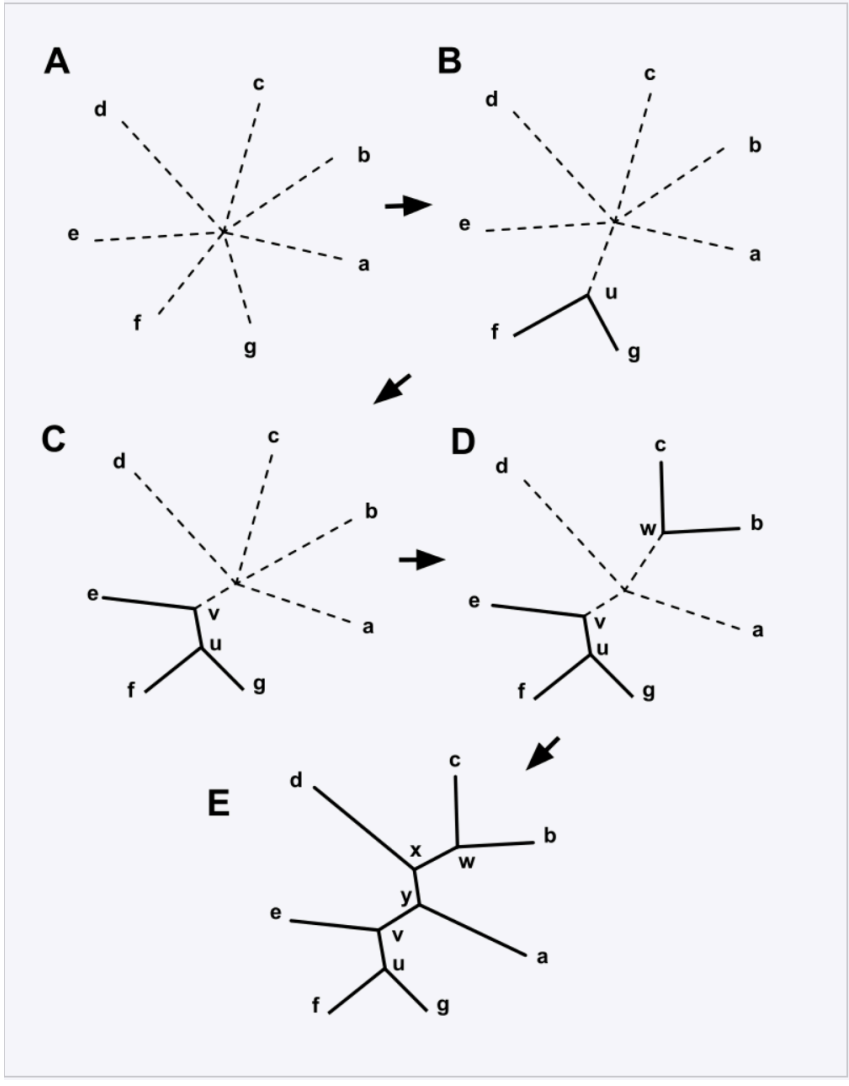


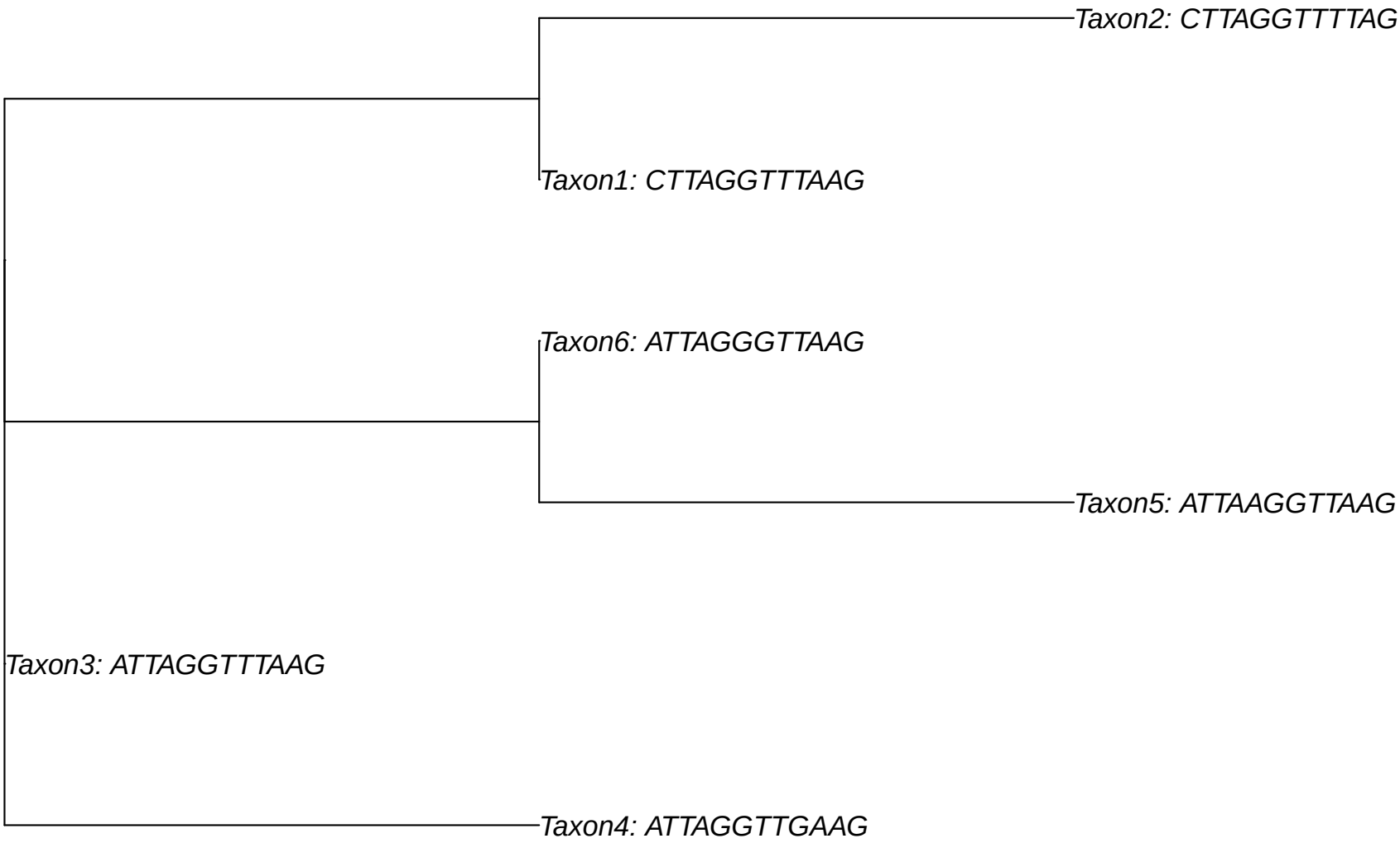
Figure 5: https://en.wikipedia.org/wiki/Neighbor_joining

Neighbor joining

- Greedily select pair of nodes to merge to minimize the branch lengths and hence tree length (sum of all branch lengths).
- A greedy algorithm so optimality is not guaranteed.
- But it's fast: $O(n^3)$ where n is the number of taxa.

Neighbor joining

Taxon1: CTTAGGTTTAAG
Taxon2: CTTAGGTTTTAG
Taxon3: ATTAGGTTTAAG
Taxon4: ATTAGGTTGAAG
Taxon5: ATTAAGGTTAAG
Taxon6: ATTAGGGTTAAG



Limitations of Hamming distance?

- Assumes constant rate of substitution across sites.
- Assumes rate of substitution is independent of the nucleotide bases.
  - The rate of of transversion and transition may not be the same (e.g., $P(A \to C)$ vs $P(A \to G)$).
- We want to be able to incorporate our knowledge of biology/evolution into the analysis.
  - e.g., multiple hits, parallel evolution.

Evolutionary model on substitution

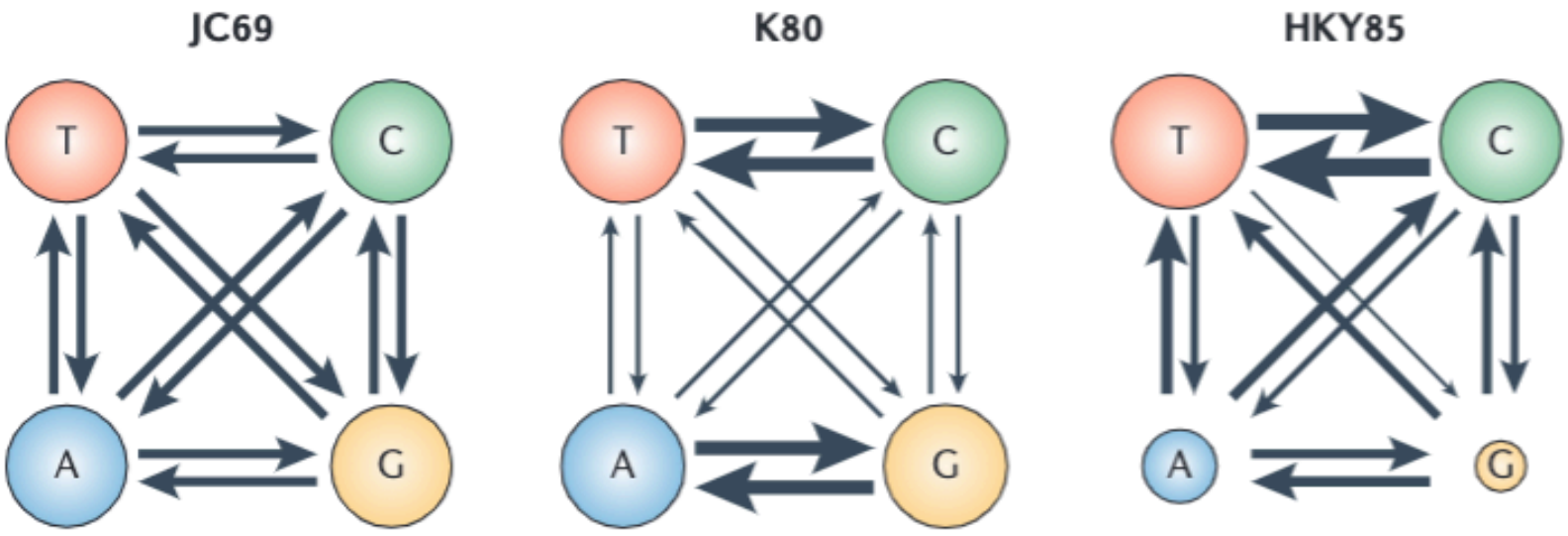


Figure 1 | **Markov models of nucleotide substitution.** The thickness of the arrows indicates the substitution rates of the four nucleotides (T, C, A and G), and the sizes of the circles represent the nucleotide frequencies when the substitution process is in equilibrium. Note that both JC69 and K80 predict equal proportions of the four nucleotides.

Figure 6: Yang2012-jd

Site specific rate of variation

- Rate of evolution may be faster or slower depending on the region of the DNA.
- Multiply the rate of substitution by $\gamma_i > 0$ for sites i.
  - Highly variable regions would have higher γ_i.
  - $\gamma_i \sim \text{Gamma}(a, b)$
  - Bayesian hierarchical model to specify common prior distribution over these parameters and share statistical strength.

Need for evolutionary modeling

- Convergent evolution.
- Gene-transfer.
- Insertion and deletion.
- Molecular clock assumption.
- In general, we need a flexible framework to be able to capture various characteristics of the underlying evolutionary biology.

Uncertainty quantification

- The distance based approaches do not allow to quantify uncertainty: we get one tree as an output.
- There may be multiple trees that can explain the observed sequences similarly well.
- Will the tree look different if we use different model or parameters? How do we test/compare two trees?

Likelihood of the data given the tree

- To quantify uncertainty we need to perform statistical analysis.
- Pinnacle of modern statistical analysis is the likelihood principle.
- Suppose we have model parameters θ and tree T, how can we compute the likelihood of the sequence data:

$$P(\mathcal{Y}|T,\theta)$$

Felsenstein pruning algorithm

- Developed by Felsenstein (1981).
- Marginalize over the sequences at the internal nodes using dynamic programming algorithm.
- Can perform the computation in $O(S|V|)$, where S is the length of the sequence and $|V|$ is the number of nodes in the tree.

Maximum likelihood approach

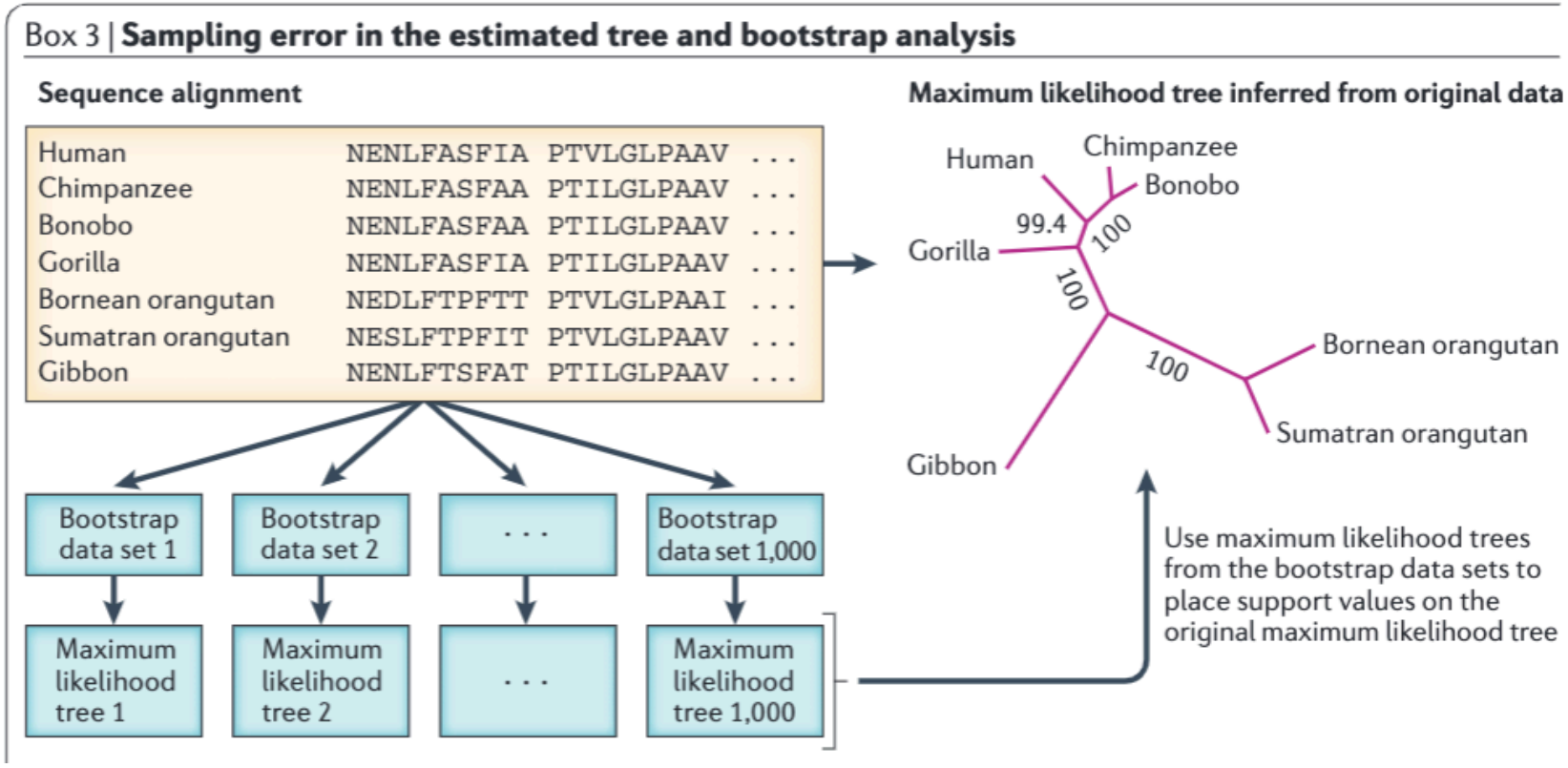


Figure 7: Yang2012-jd

Maximum likelihood approach

- We still need to search the space of trees and find the maximum tree.

Bayesian approach

The posterior distribution is given by,

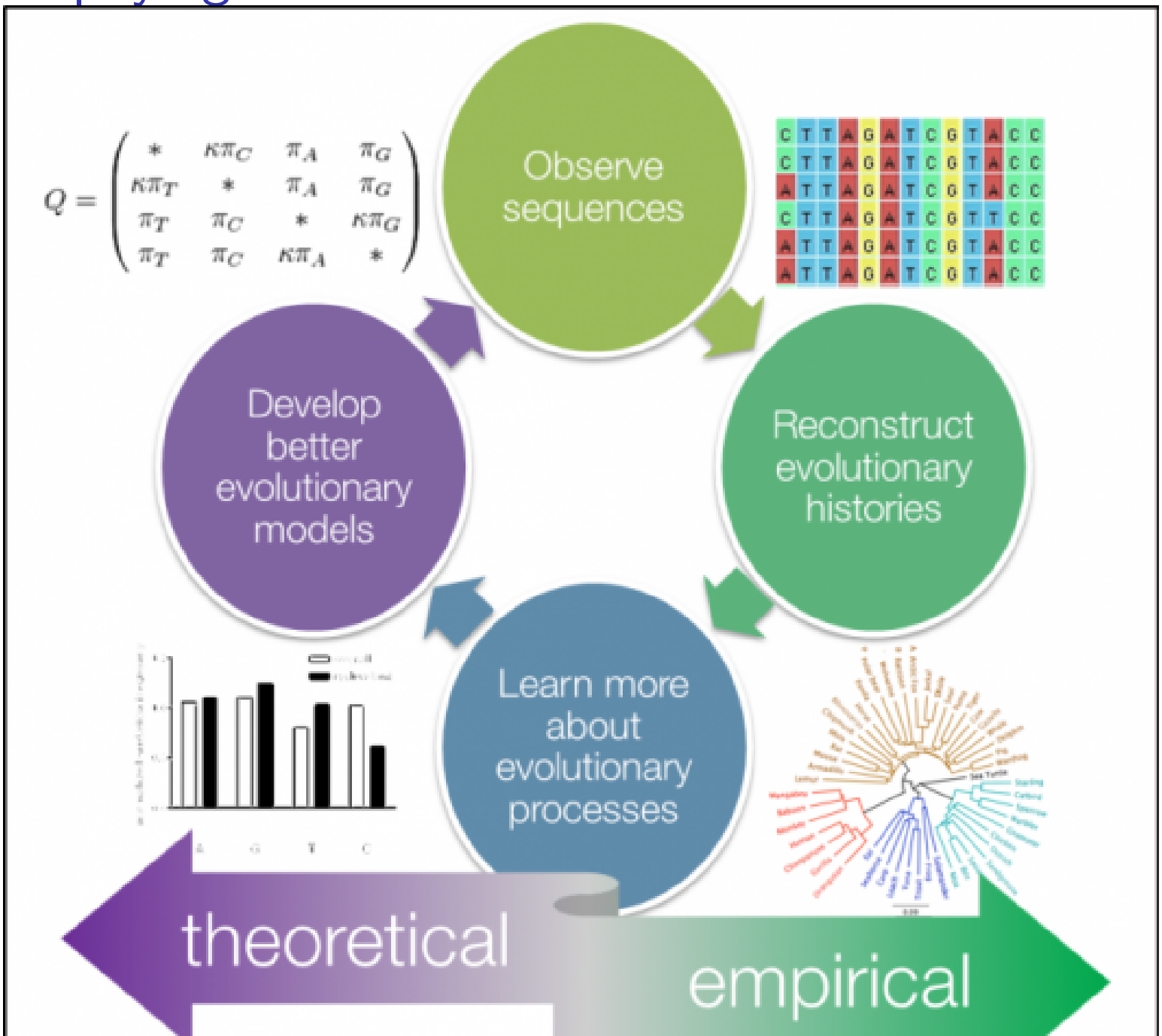
$$P(T, \theta|\mathcal{Y}) = \frac{P(\mathcal{Y}|T, \theta)P(T, \theta)}{\sum_{T'} \int_{\Theta} P(\mathcal{Y}|T', \theta)P(T, \theta)}.$$

- The summation in the numerator is finite but large, making the posterior intractable.
- Sample from the posterior using Monte Carlo methods (MCMC, Sequential Monte Carlo, Hamiltonian Monte Carlo, and Variation inference) -> active area of research.

Bayesian approach

- Does not rely on the asymptotics as for the MLE.
- Intuitive probabilistic statements about the model parameters and the tree can be made.
- Tree search is still a challenge as for the maximum likelihood approach.
- Sensitive to prior specification.

What is phylogenetics?



References

Felsenstein, J. 1981. "Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach." *J. Mol. Evol.* 17 (6): 368–76.

Matsen, Frederick A, Robin B Kodner, and E Virginia Armbrust. 2010. "Pplacer: Linear Time Maximum-Likelihood and Bayesian Phylogenetic Placement of Sequences onto a Fixed Reference Tree." *BMC Bioinformatics* 11 (October): 538.

Nowell, P C. 1976. "The Clonal Evolution of Tumor Cell Populations." *Science* 194 (4260): 23–28.

Vendramin, Roberto, Kevin Litchfield, and Charles Swanton. 2021. "Cancer Evolution: Darwin and Beyond." *EMBO J.* 40 (18): e108389.

Yang, Ziheng, and Bruce Rannala. 2012. "Molecular Phylogenetics: Principles and Practice." *Nat. Rev. Genet.* 13 (5): 303–14.