Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

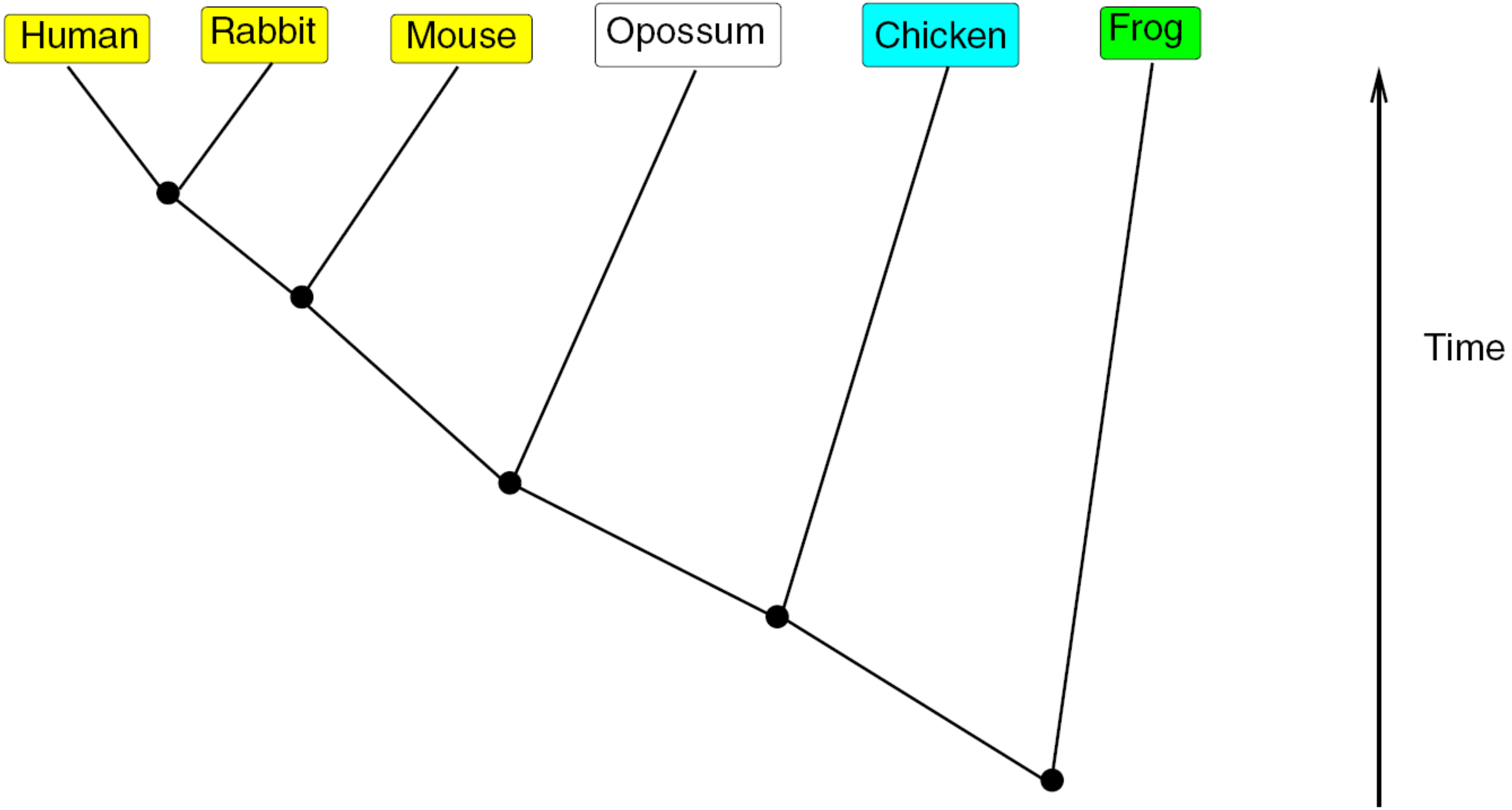
Statistical phylogenetics

Niko Beerenwinkel

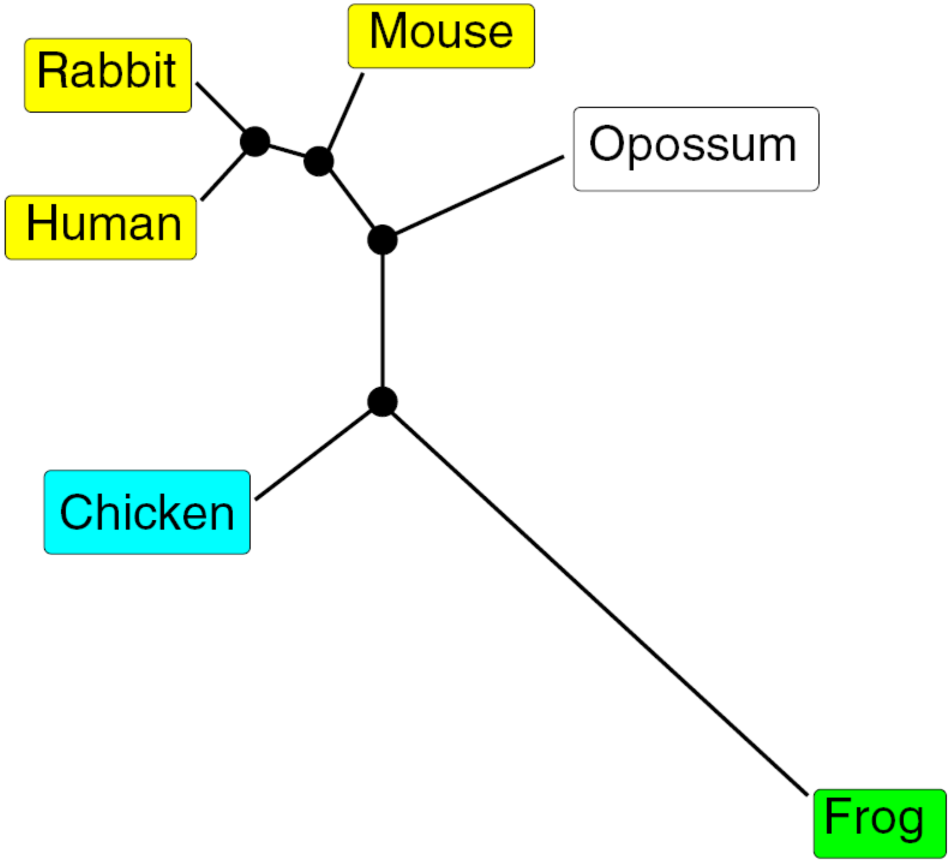
Outline

- Phylogenetic trees
- Nucleotide substitution models
- Likelihood
- ML estimation and bootstrapping
- Bayesian inference
- Rate heterogeneity
- Phylo-HMMs

Rooted phylogenetic tree



Unrooted phylogenetic tree

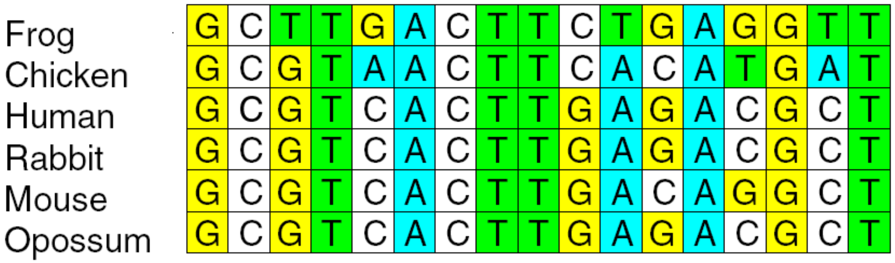


Phylogenetic trees

- Leaves = contemporary species
- Interior vertices = common ancestors
- Topology (graph structure) defines branching order (subtrees)
- Branch lengths (parameters) define time
  - absolute time t, or
  - phylogenetic time $w = \lambda t$ (λ = nucleotide substitution rate)

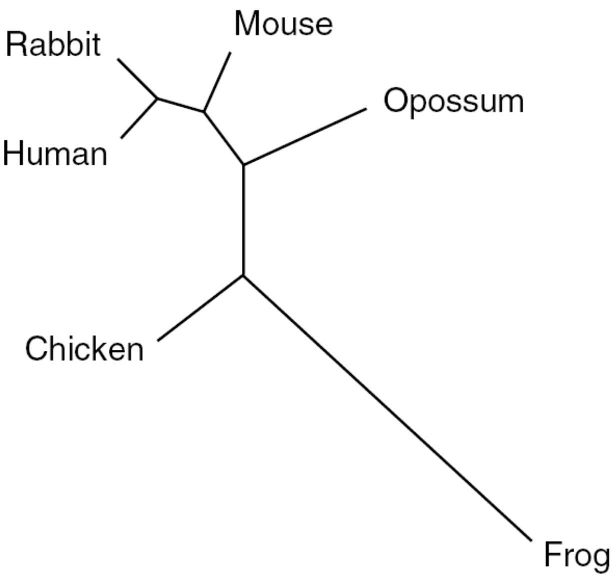
Phylogenetic inference

- Given a multiple alignment



Frog	G	C	T	T	G	A	C	T	T	C	T	G	A	G	G	T	T
Chicken	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Human	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Rabbit	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Mouse	G	C	G	T	C	A	C	T	T	G	A	C	A	G	G	C	T
Opossum	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T

- Find the best tree explaining the data



Phylogeny reconstruction methods

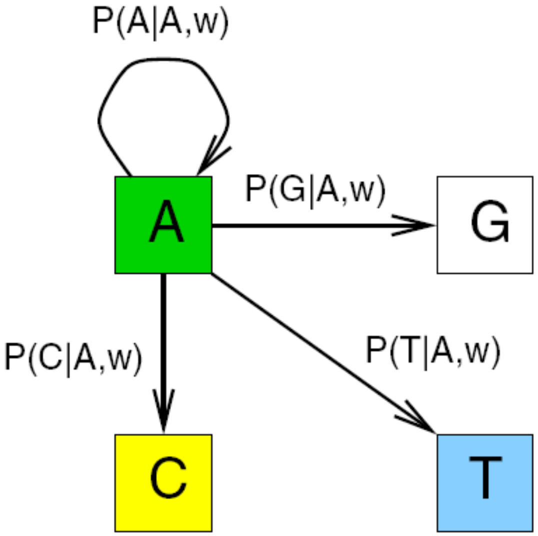
- Distance-based clustering methods
  - Define an evolutionary distance
  - Use hierarchical clustering
  - UPGMA, Neighbor joining
  - Shortcomings: information loss, not robust against violations of tree assumption

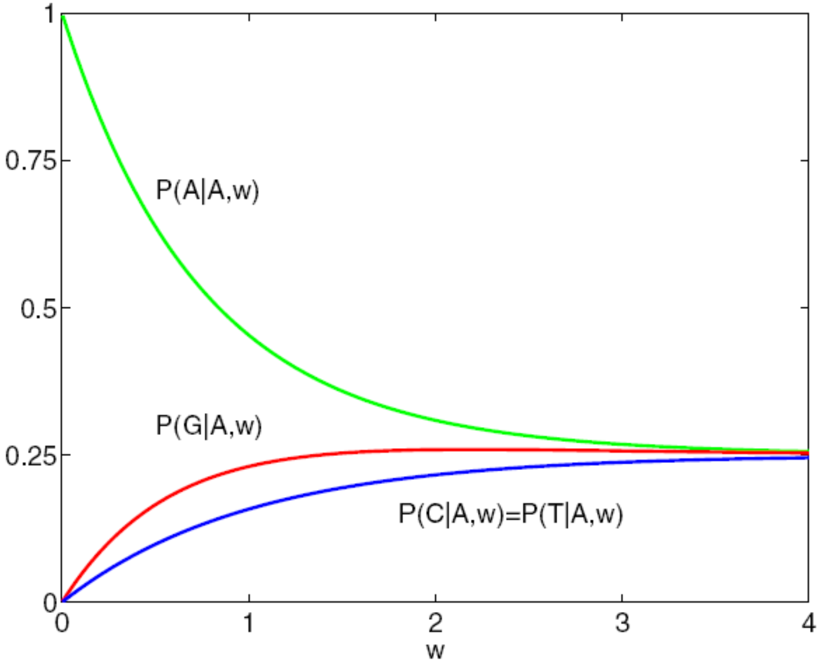
- Parsimony
  - “Minimum evolution” principle: Find the tree explaining the data by the minimum number of mutations.
  - Shortcomings: “model-free”, not consistent

- Likelihood methods

Nucleotide substitution models

- For characters x, y ∈ {A,C,G,T}, we define P(y | x, w) as the probability of observing y after w time units given that the same site was originally occupied by x.





Continuous-time homogeneous Markov chain

- Let $y_i(t) \in \{A,C,G,T\}$ be the nucleotide at position i at time t.
- We assume for all s, t > 0 and i, k $\in$ {1, ..., N},

1) a Markov process

$$P\,[y_i(t+\Delta t) \mid y_i(t), y_i(t-\Delta t), \ldots\,] = P[y_i(t{+}\Delta t) \mid y_i(t)]$$

2) homogeneous in time

$$P[y_i(s+t) \mid y_i(s)] = P[y_i(t) \mid y_i(0)]$$

3) identical across sites

$$P[y_i(t) \mid y_i(0)] = P[y_k(t) \mid y_k(0)]$$

4) independent among sites

$$P[y_1(t), \ldots, y_N(t) \mid y_1(0), \ldots, y_N(0)] = \prod_{i=1}^{N} P[y_i(t) \mid y_i(0)]$$

Transition matrix

- The nucleotide substitution process is defined by the 4 × 4 transition matrix

$$\mathbf{P}(t) = \big(P[y(t) = a \mid y(0) = b]\big)_{a,b \in \{A,C,G,T\}}$$

- $\mathbf{P}(0) = \mathbf{I}$

- Chapman-Kolmogorov equation for continuous-time homogeneous Markov chains:

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$$

for all s, t > 0.

Rate matrix

- Ansatz: $\mathbf{P}(dt) = \mathbf{P}(0) + \mathbf{R}dt$, where **R** is the *rate matrix*.
- Then

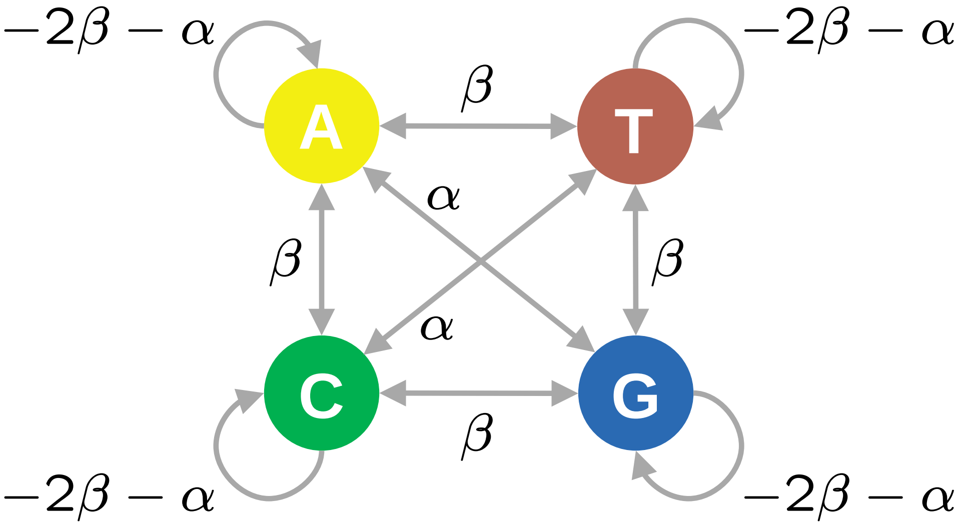
$$\mathbf{P}(t + dt) = \mathbf{P}(dt)\mathbf{P}(t) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t)$$

$$\Rightarrow \quad \frac{d\mathbf{P}(t)}{dt} = \mathbf{R}\mathbf{P}(t)$$

$$\Rightarrow \quad \mathbf{P}(t) = \exp(\mathbf{R}t) = \sum_{k=0}^{\infty} \frac{1}{k!}(\mathbf{R}t)^k$$

Example: Kimura model

- Purines = {A, G}
- Pyrimidines = {C, T}
- Transitions (rate α)
  - purine ↔ purine
  - pyrimidine ↔ pyrimidine
- Transversions (rate β)
  - purine ↔ pyrimidine



$$\mathbf{R} = \begin{pmatrix} -2\beta-\alpha & \beta & \alpha & \beta \\ \beta & -2\beta-\alpha & \beta & \alpha \\ \alpha & \beta & -2\beta-\alpha & \beta \\ \beta & \alpha & \beta & -2\beta-\alpha \end{pmatrix}$$

Equilibrium base distribution

- The marginal distribution of nucleotides

$$\mathbf{u}(w) = (P[y(w) = a])_{a \in \{A,C,G,T\}}$$

defines a homogeneous Markov chain,

$$\mathbf{u}(v + w) = \mathbf{P}(w)\mathbf{u}(v)$$

- An ergodic Markov chain converges to a unique stationary distribution $\lim_{w \to \infty} \mathbf{u}(w) = \pi = (\Pi_A, \Pi_C, \Pi_G, \Pi_T)$

characterized by $\mathbf{P}(w)\pi = \pi$.

- For the Kimura model, we find $\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$.

HKY85 model

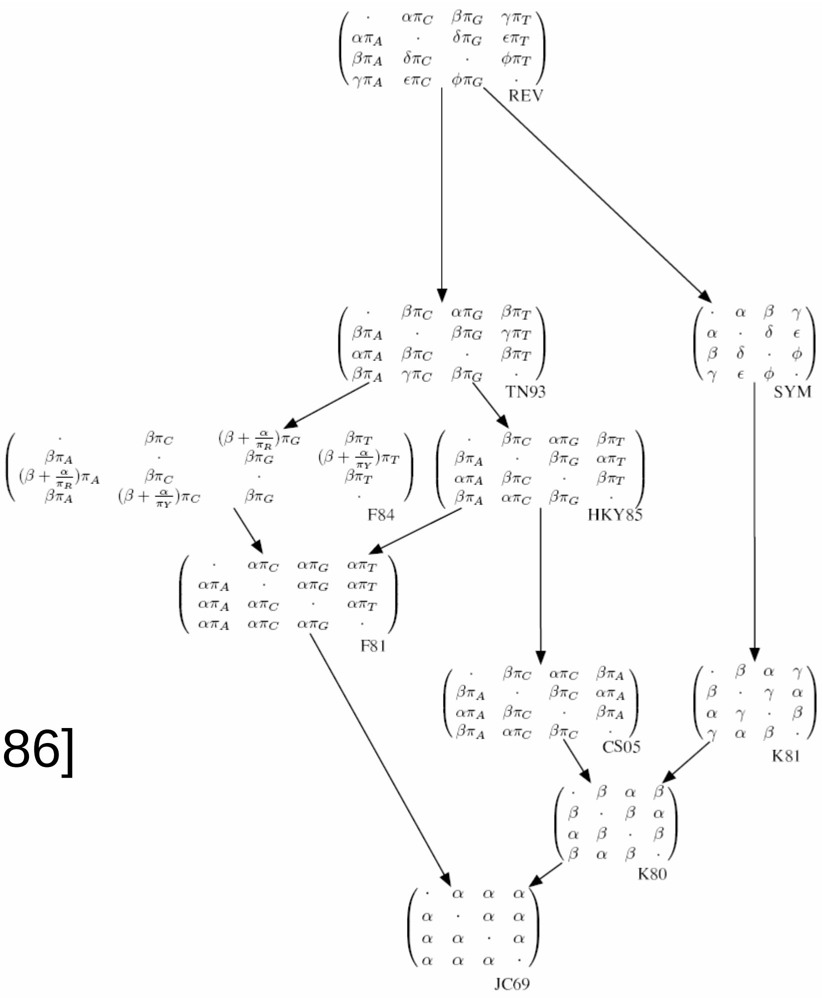
- The following modified rate matrix has the stationary distribution $\pi = (\Pi_A, \Pi_C, \Pi_G, \Pi_T)$

$$\mathbf{R} = \begin{pmatrix} * & \sqcap_A\,\beta & \sqcap_A\,\alpha & \sqcap_A\,\beta \\ \sqcap_C\,\beta & * & \sqcap_C\,\beta & \sqcap_C\,\alpha \\ \sqcap_G\,\alpha & \sqcap_G\,\beta & * & \sqcap_G\,\beta \\ \sqcap_T\,\beta & \sqcap_T\,\alpha & \sqcap_T\,\beta & * \end{pmatrix}$$

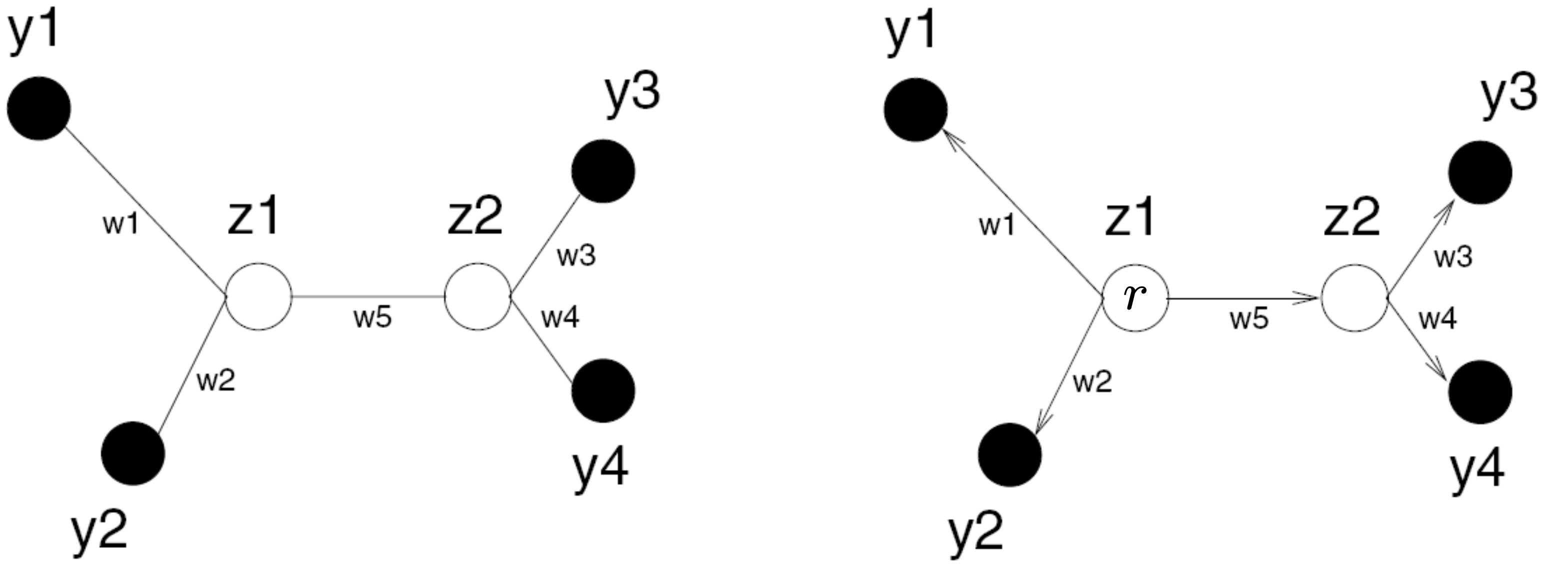
- We assume *stationarity* of the Markov chain: The nucleotide distribution is equal to π over the whole tree.

Felsenstein hierarchy

- JC69 [Jukes and Cantor, 1969]
- K80 [Kimura, 1980]
- K81 [Kimura, 1981]
- CS05 [Yap and Pachter, 2004]
- F81 [Felsenstein, 1981]
- HKY85 [Hasegawa et al., 1985]
- F84 [Felsenstein, 1989]
- TN93 [Tamura and Nei, 1993]
- SYM [Zharkikh, 1994]
- REV [Lanave et al., 1984, Tavare, 1986]



A phylogenetic tree is a Bayesian network



$$
\begin{aligned}
P(y_1, y_2, y_3, y_4, z_1, z_2 \mid \mathbf{w}) \;=\; & \Pi(z_1) P(y_1 \mid z_1, w_1) P(y_2 \mid z_1, w_2) \cdot \\
& \cdot P(z_2 \mid z_1, w_5) \cdot \\
& \cdot P(y_3 \mid z_2, w_3) P(y_4 \mid z_2, w_4)
\end{aligned}
$$

The nucleotide substitution model defines the LPDs

- In general, $x_i \in \{A,C,G,T\}$ is the random variable indicating the nucleotide at vertex i.

$$P(x_1, \ldots, x_M) = \Pi(x_r) \prod_{i \in V \setminus \{r\}} P\left(x_i \mid x_{\mathrm{pa}(i)}, w_i\right)$$

$$= \Pi(x_r) \prod_{i \in V \setminus \{r\}} \mathbf{P}(w_i)_{x_i, x_{\mathrm{pa}(i)}}$$

- Choice of the root vertex does not matter as long as the Markov chain is *reversible*, i.e.,

$$P(y \mid x, w)\Pi(x) = P(x \mid y, w)\Pi(y)$$

Marginalization over extinct species

- **y** = extant (contemporary) species
- **z** = extinct common ancestors

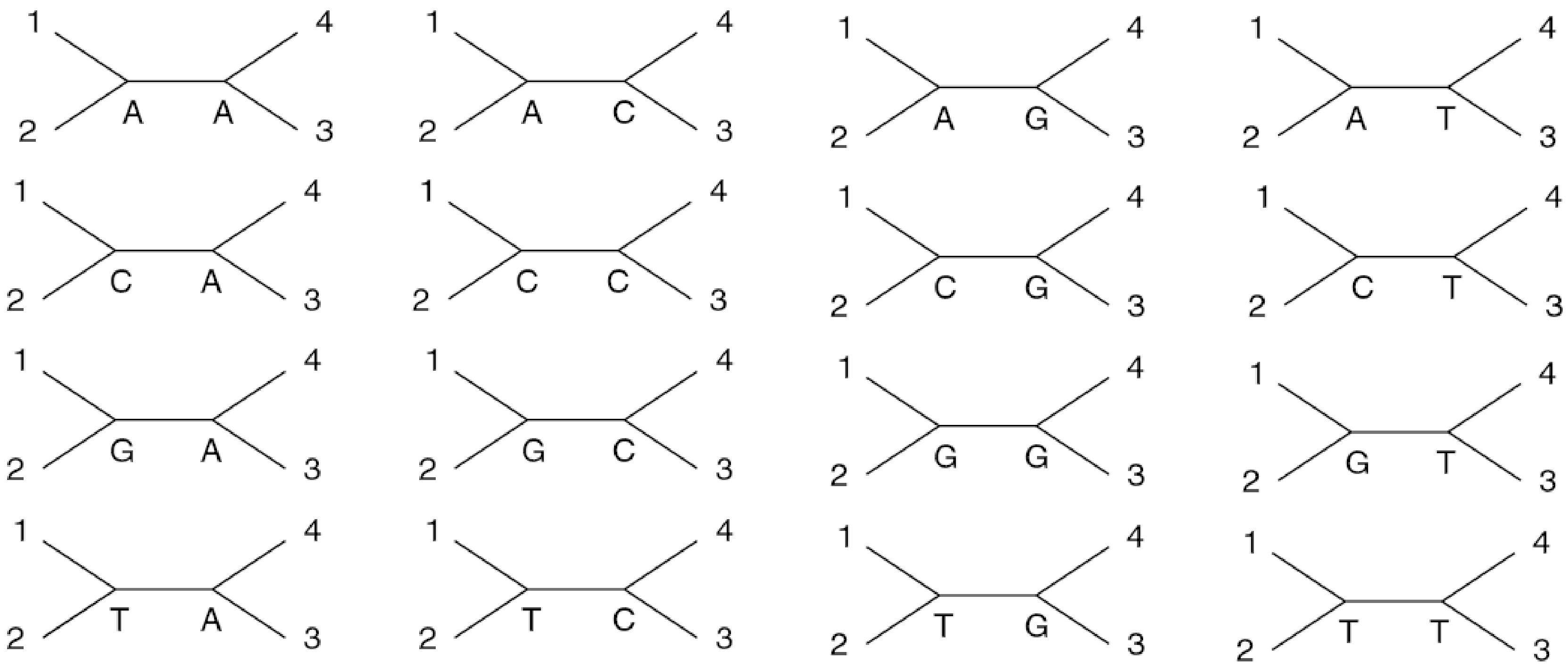
$$P(\mathbf{y} \mid \mathbf{w},\, S) = \sum_{\mathbf{z}} P(\mathbf{y}, \mathbf{z} \mid \mathbf{w},\, S)$$

where S indicates the tree topology.

- This marginal distribution can be computed efficiently with the sum-product algorithm (→ “peeling algorithm”, “Felsenstein algorithm”), a generalization of the forward algorithm.

Example

- Two extinct species (= hidden variables)



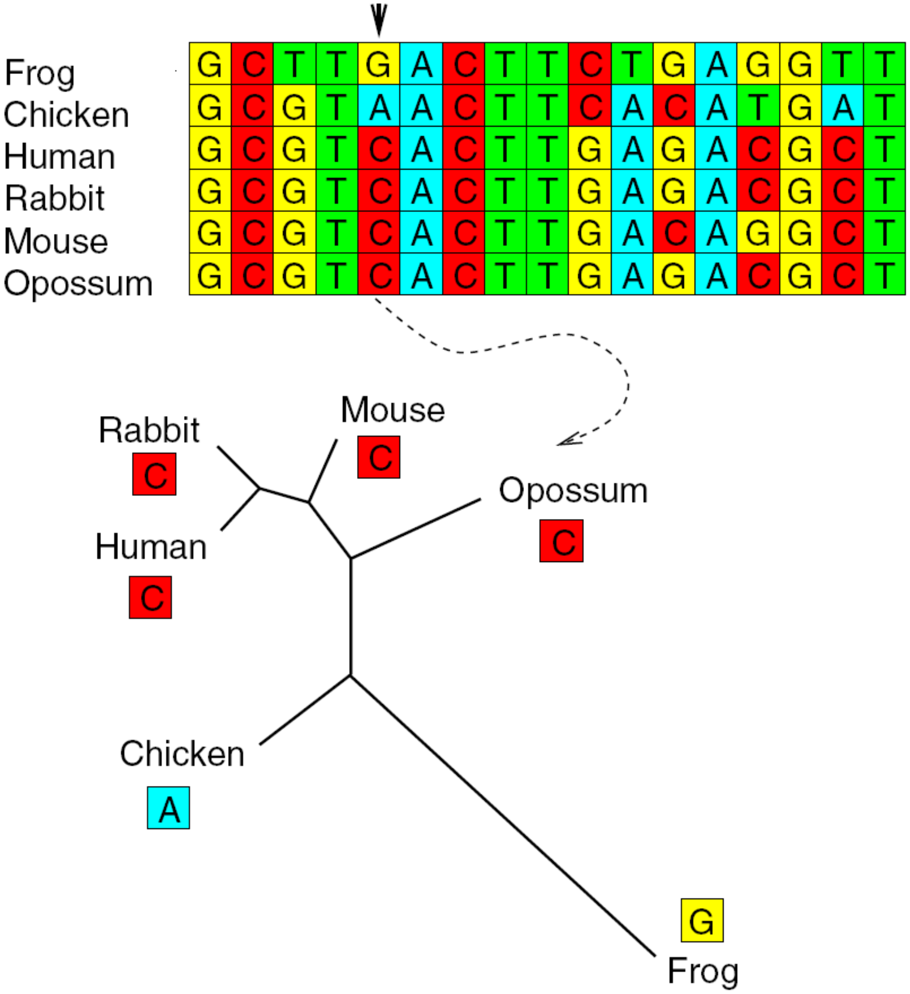
Likelihood of a phylogenetic tree

- Given a multiple alignment $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, where $\mathbf{y}_t$ is the alignment column at position t, the likelihood of tree topology S and branch lengths **w** is

$$
\begin{aligned}
P(\mathcal{D} \mid \mathbf{w},\, S) &= \prod_{t=1}^{N} P(\mathbf{y}_t \mid \mathbf{w},\, S) \\
&= \prod_{t=1}^{N} \sum_{\mathbf{z}} P(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{w},\, S)
\end{aligned}
$$

- We have omitted here and will continue to omit the parameters of the nucleotide substitution model (the rate matrix).

Likelihood of a phylogenetic tree



Frog	G	C	T	T	G	A	C	T	T	C	T	G	A	G	G	T	T
Chicken	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Human	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Rabbit	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Mouse	G	C	G	T	C	A	C	T	T	G	A	C	A	G	G	C	T
Opossum	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T

Maximum likelihood

- No analytical solution exists for the MLE problem

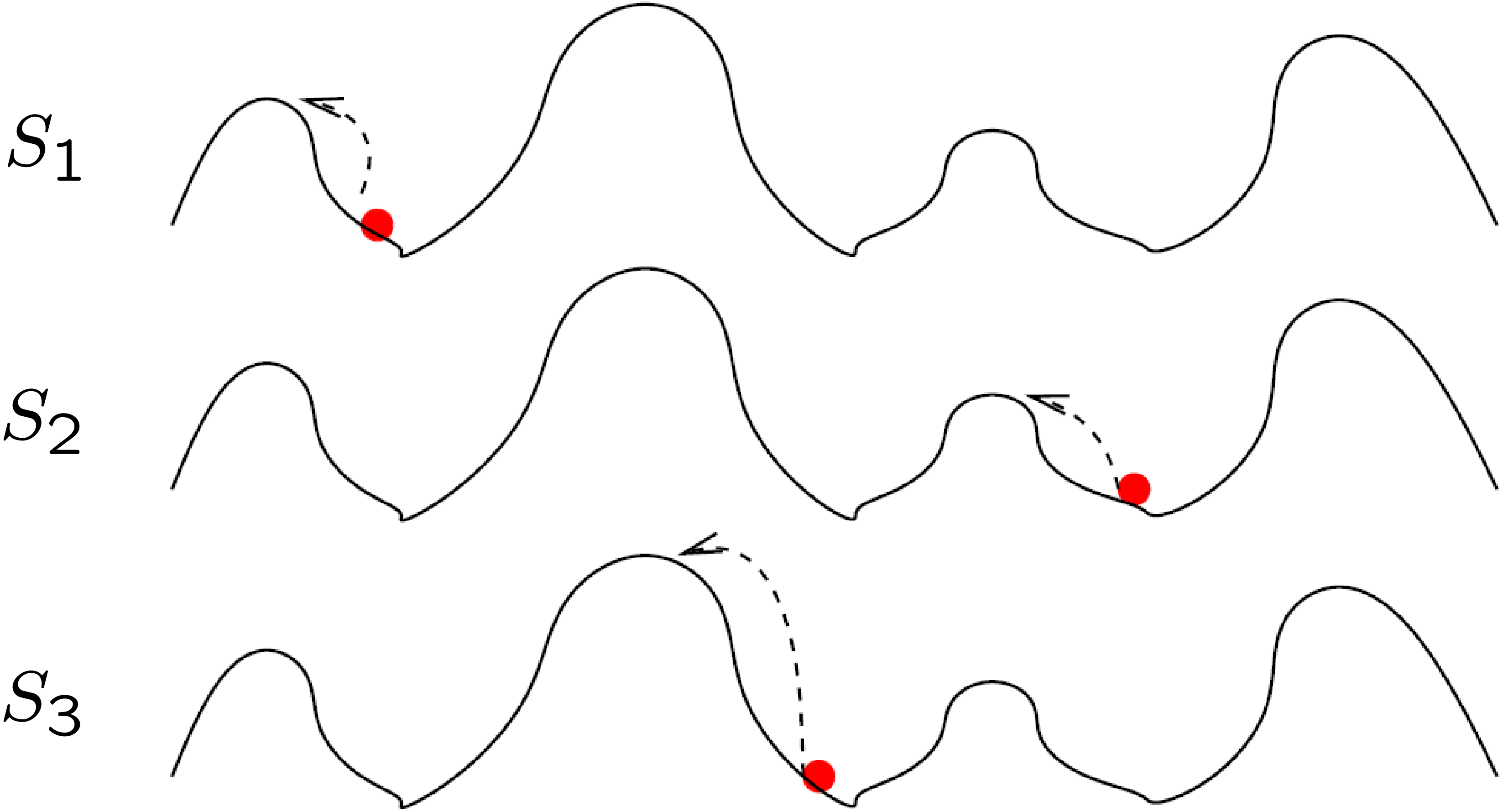
$$\max_{S,\,\mathbf{w}} \; P(\mathcal{D} \mid \mathbf{w},\, S)$$

  and numerical optimization is NP-hard.
- Branch lengths are optimized by a gradient ascent scheme:

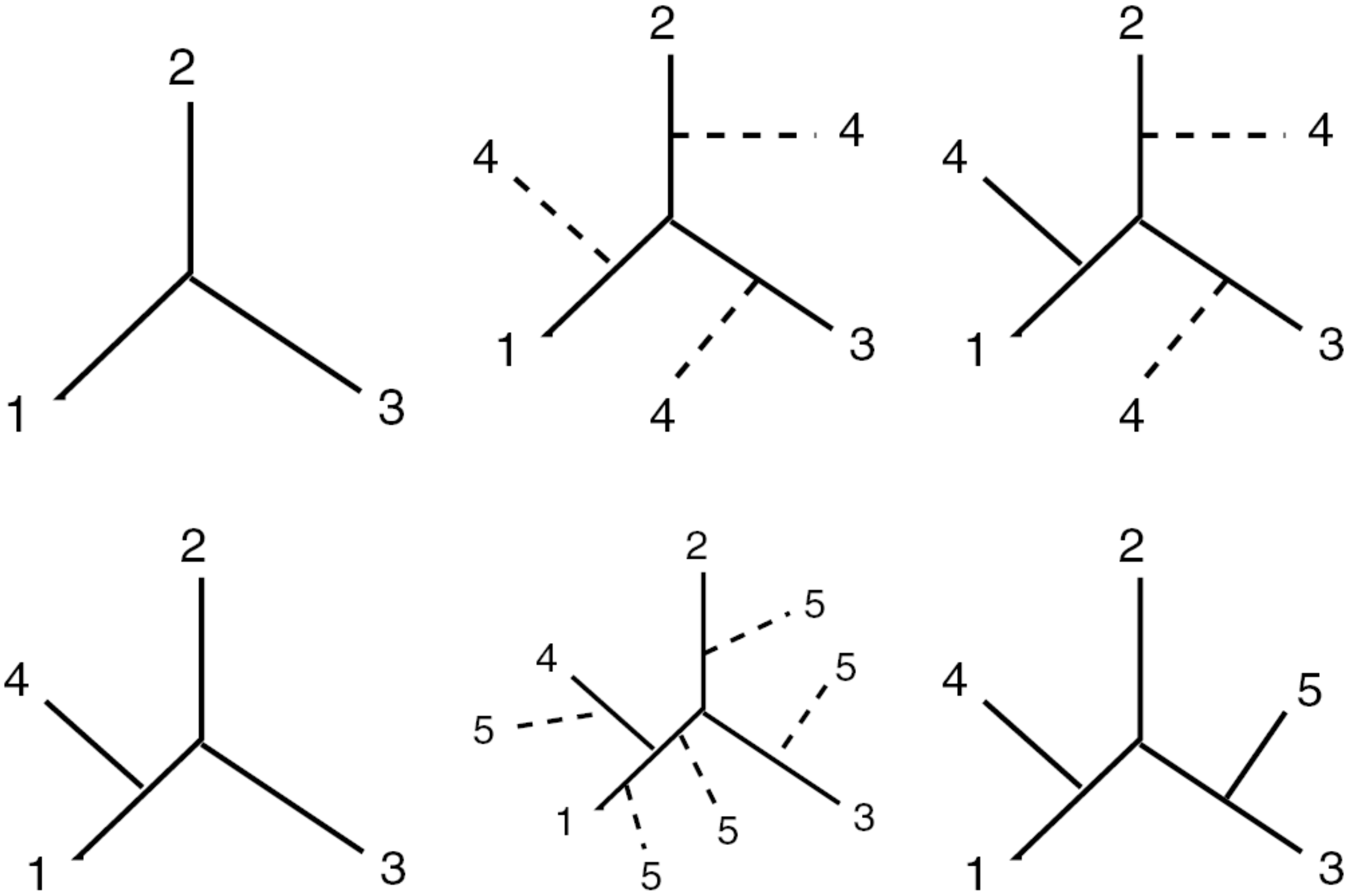
$$\mathbf{w} \to \mathbf{w} + \mathbf{A}\nabla_{\mathbf{w}} \log P(\mathcal{D} \mid \mathbf{w},\, S)$$

- There are (2n − 5)!! unrooted tree topologies for n taxa.
  $\Rightarrow$ heuristic search procedures:
  - DNAML
  - Quartet puzzling

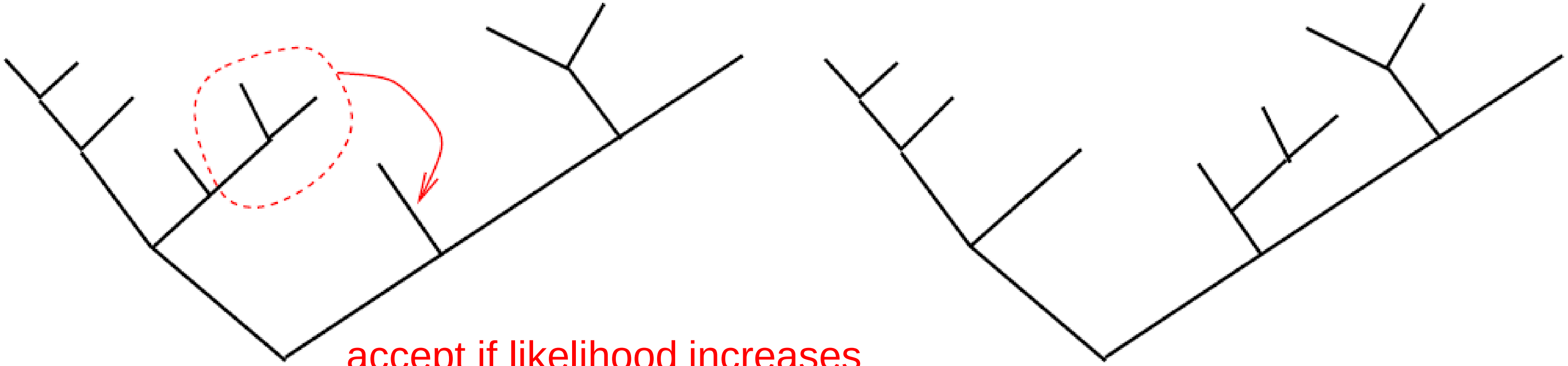
Branch lengths are optimized for each topology



DNAML: Iterative attachment of branches



DNAML: Branch regrafting

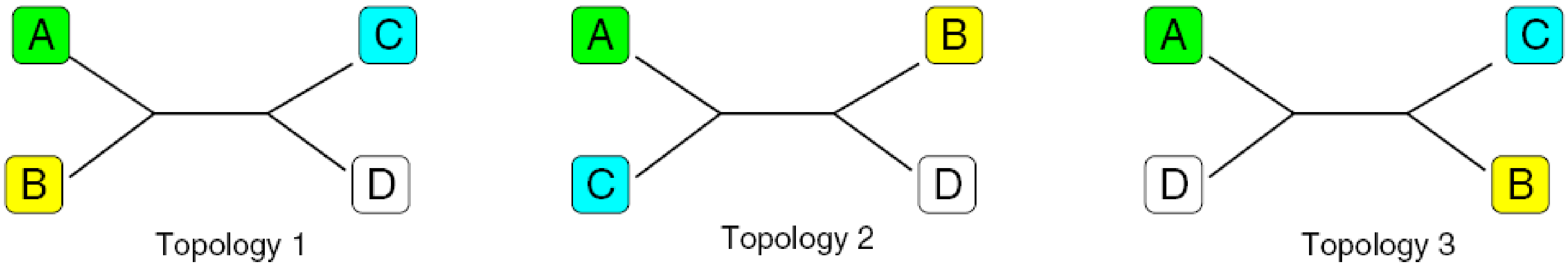


accept if likelihood increases

- DNAML employs greedy search strategies
- Results depend on the order in which alignment columns are considered
- Only few branch manipulations can be computationally afforded

Quartet puzzling

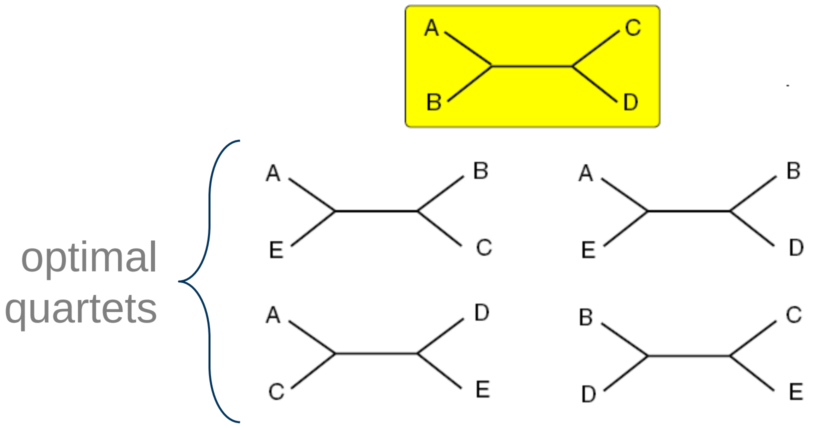
1. Construct all (n choose 4) quartet trees on four of the n given taxa using maximum likelihood

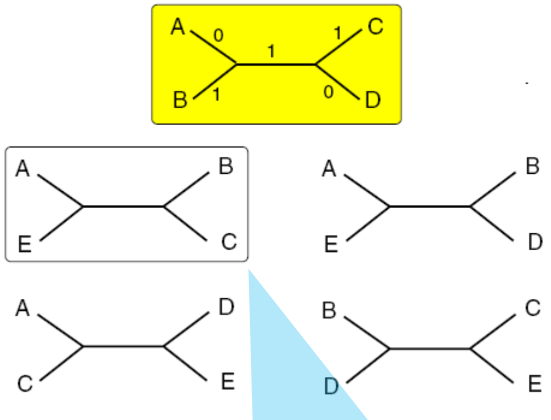


2. Combine the quartet trees into a global tree by adding taxa iteratively minimizing conflicts (puzzling step).

Example

- Consider five taxa A, B, C, D, and E.
- There are (5 choose 4) = 5 quartets.

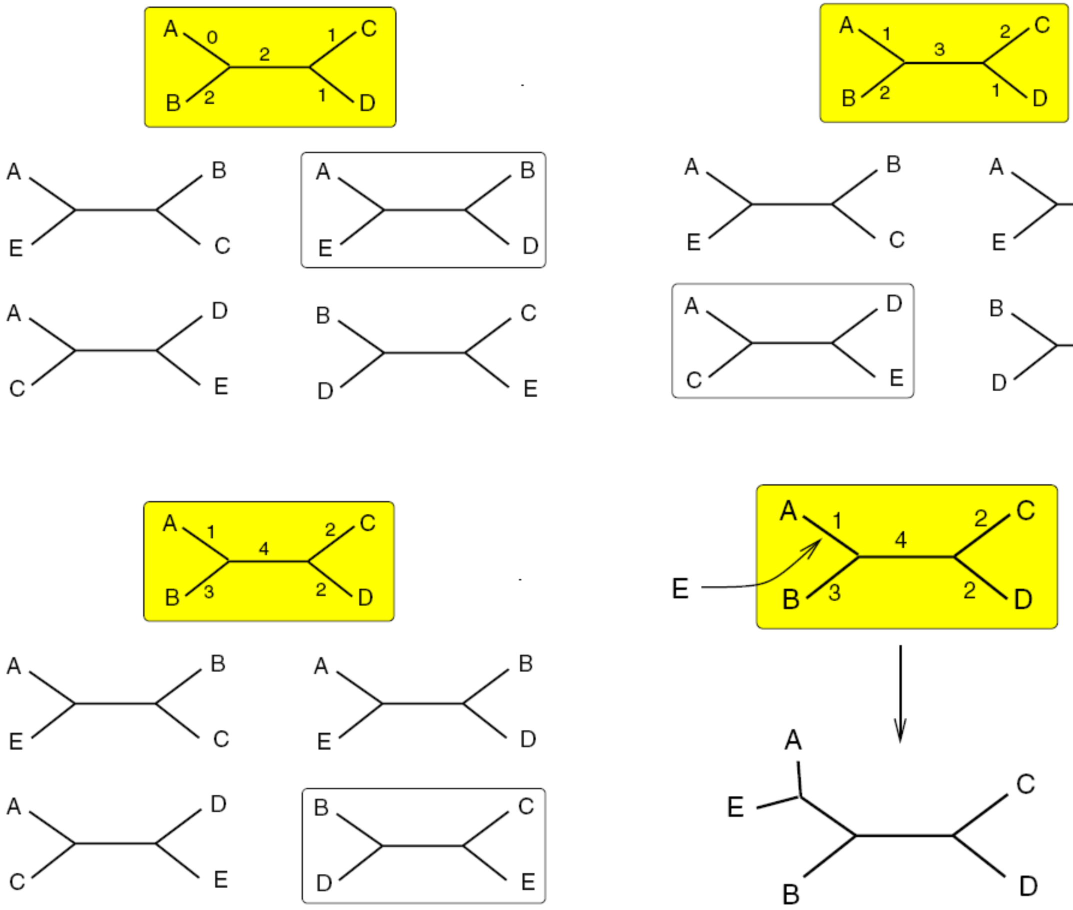




Where to place E?

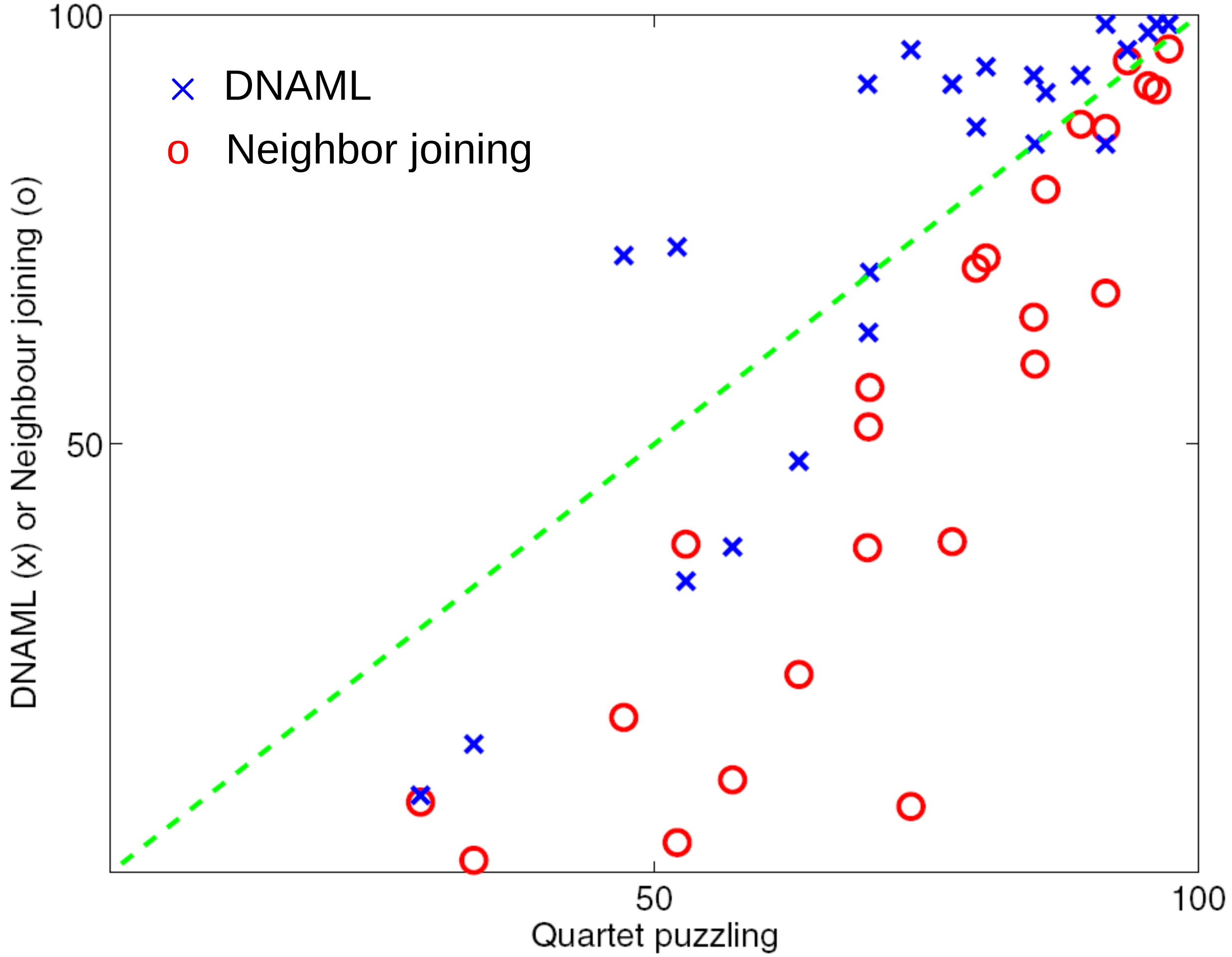
E should not be placed anywhere on the path between B and C.

Example



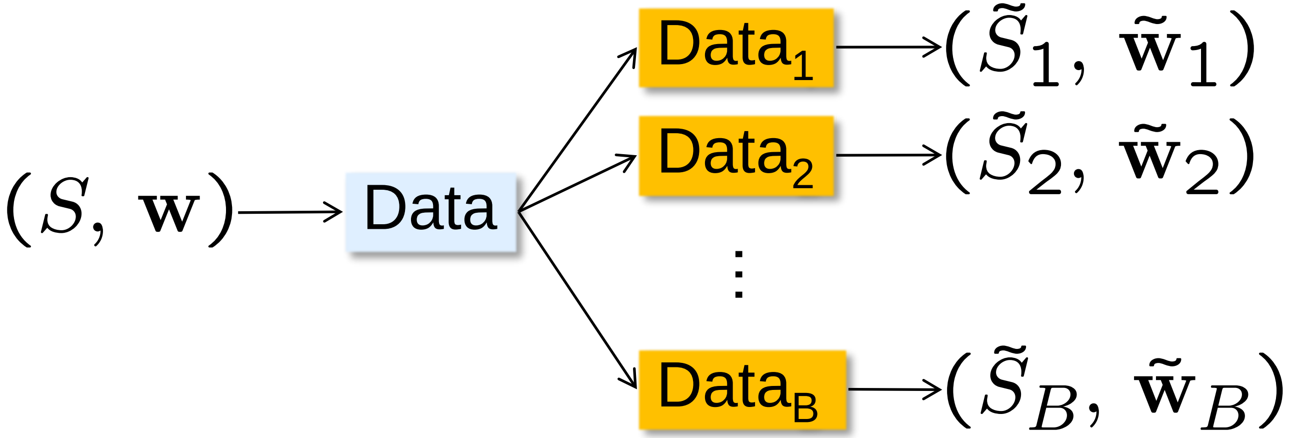
select least penalized branch

Performance of quartet puzzling



Bootstrapping phylogenetic trees

- How sure can we be that the estimated ML phylogenetic tree is correct?



$(S, \mathbf{w}) \longrightarrow$ Data

$\text{Data}_1 \longrightarrow (\tilde{S}_1, \tilde{\mathbf{w}}_1)$

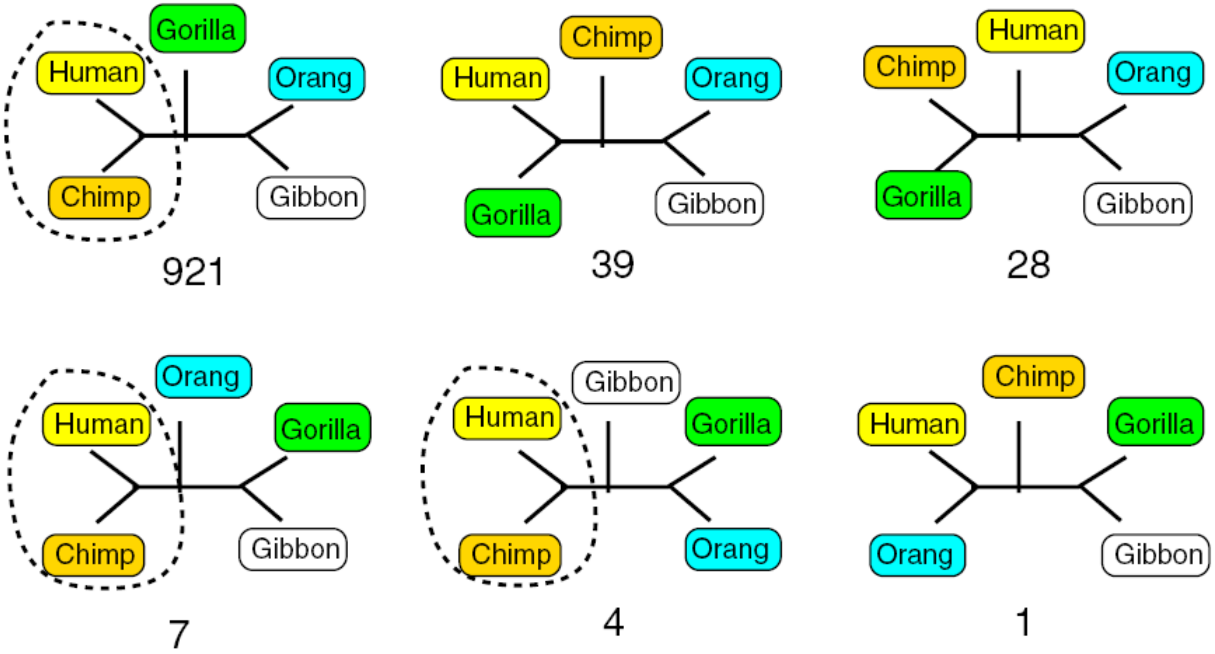
$\text{Data}_2 \longrightarrow (\tilde{S}_2, \tilde{\mathbf{w}}_2)$

$\vdots$

$\text{Data}_B \longrightarrow (\tilde{S}_B, \tilde{\mathbf{w}}_B)$

Example: (Human, Chimp)

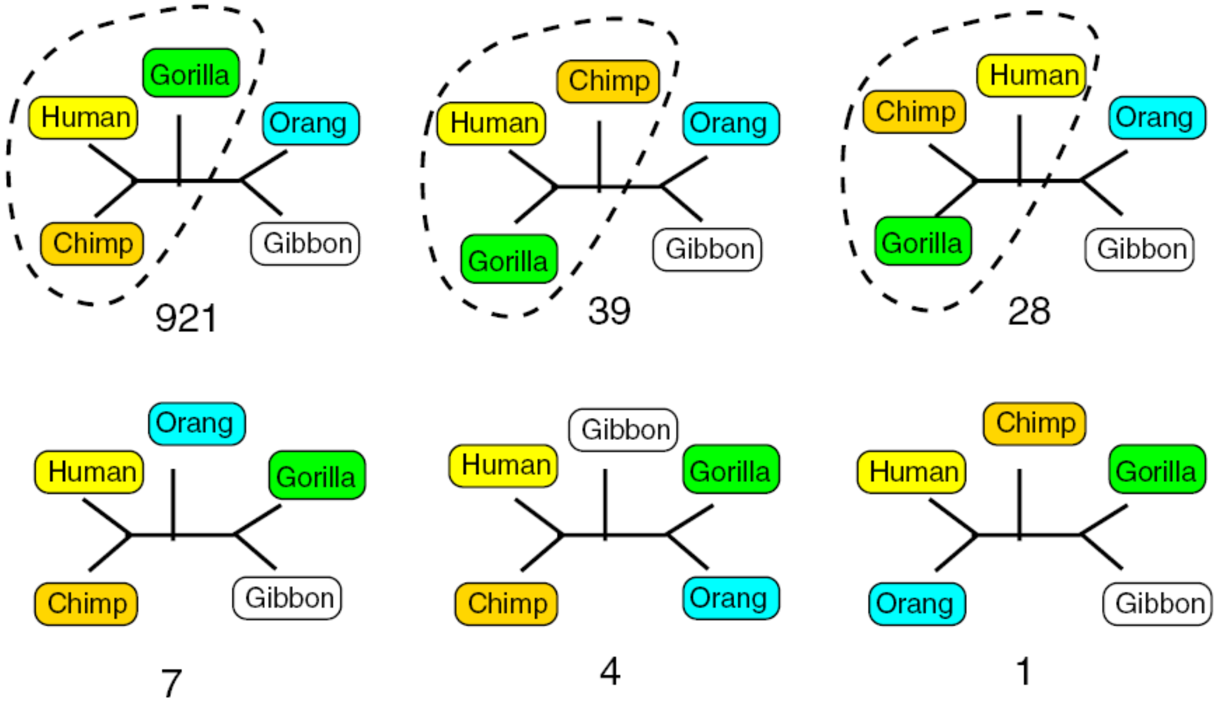
- B = 1000 bootstrap samples



$$P[(\mathrm{Human}, \mathrm{Chimp})] = \frac{921 + 7 + 4}{1000} = 0.932$$

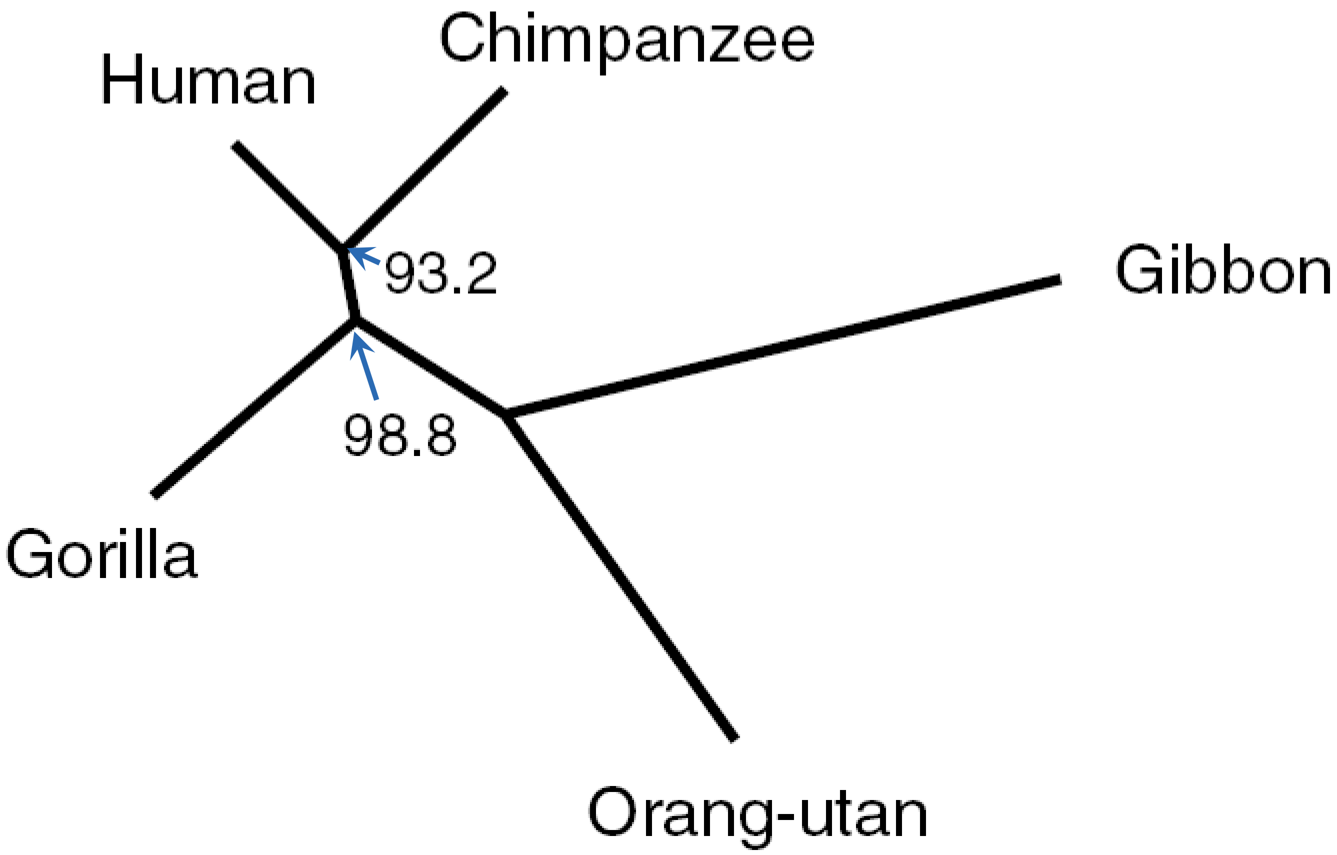
Example: (Human, Chimp, Gorilla)

- Same B = 1000 bootstrap samples



$$P[((\text{Human},\text{Chimp}),\text{Gorilla})] = \frac{921 + 39 + 28}{1000} = 0.988$$

Bootstrapped tree

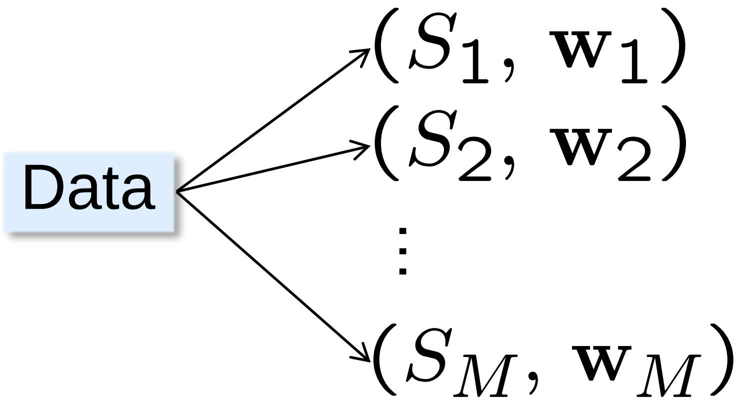


Bayesian inference

$$
\begin{aligned}
P(S \mid \mathcal{D}) \quad &\propto \quad P(S)\, P(\mathcal{D} \mid S) \\
&= \quad P(S) \int P(\mathcal{D} \mid \mathbf{w}, S) P(\mathbf{w} \mid S) d\mathbf{w}
\end{aligned}
$$

but the marginal likelihood P($\mathcal{D}$ | S) is analytically intractable

Sample from (S, w) and marginalize



Data

$(S_1, \mathbf{w}_1)$

$(S_2, \mathbf{w}_2)$

$\vdots$

$(S_M, \mathbf{w}_M)$

$$
\begin{aligned}
P(S \mid \mathcal{D}) &= \int P(S, \mathbf{w} \mid \mathcal{D}) d\mathbf{w} \\
&\approx (\#\text{ trees with topology } S)/M
\end{aligned}
$$

Rate heterogeneity

- Nucleotide substitution rates may vary across sites because of varying selective pressures. For example,
  - between coding and non-coding regions
  - among different regions of a protein (loops, catalytic residues)
  - among the three bases of a triplet coding for an amino acid
- Let us assume site-specific substitution rates r_t such that the local probabilities become $P(\mathbf{y}_t \mid r_t \mathbf{w}, S)$ and

$$P(\mathcal{D} \mid \mathbf{w}, S) = \int \prod_{t=1}^{N} P(\mathbf{y}_t \mid r_t \mathbf{w}, S) \, \mathrm{d}P(r_1, \ldots, r_N)$$

Independent substitution rates

- Let us assume that, for a hyperparameter v,

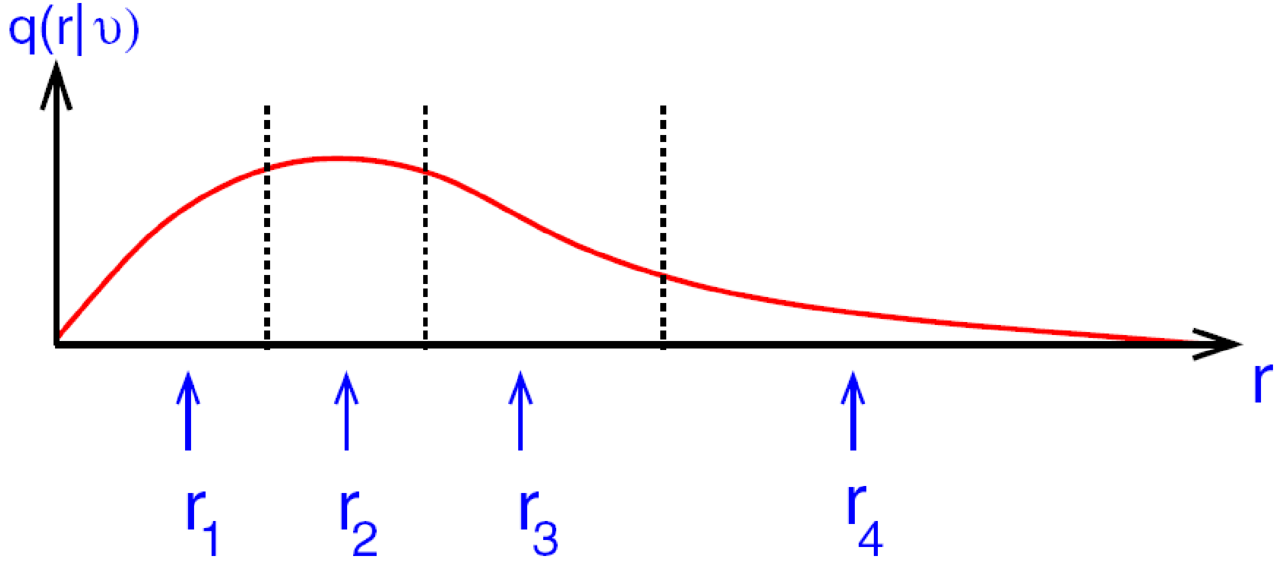
$$P(r_1, \ldots, r_N) = \prod_{t=1}^{N} q(r_t \mid v)$$

- Then the likelihood simplifies to

$$P(\mathcal{D} \mid \mathbf{w}, v, S) = \prod_{t=1}^{N} \int_0^\infty P(\mathbf{y}_t \mid r_t \mathbf{w}, S)\, q(r_t \mid v)\, dr_t$$

- q(r_t | v) = Γ(v, 1/v), gamma distribution with mean 1 and variance 1/v.

Discrete gamma distribution



- Divide the positive real line into K intervals such that each interval contains an equal area of the gamma distribution.

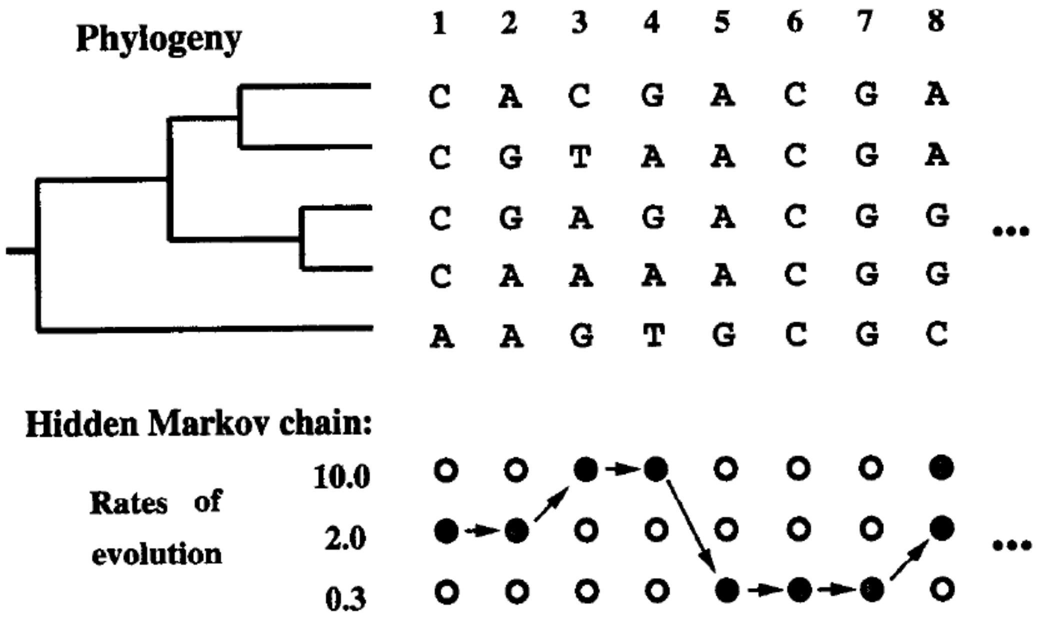
$$P(\mathcal{D} \mid \mathbf{w}, v, S) = \frac{1}{K} \prod_{t=1}^{N} \sum_{k=1}^{K} P(\mathbf{y}_t \mid r_k^v \mathbf{w}, S)$$

Spatial correlation among substitution rates

- Substitution rates will tend to be similar at neighboring sites.
- Setting

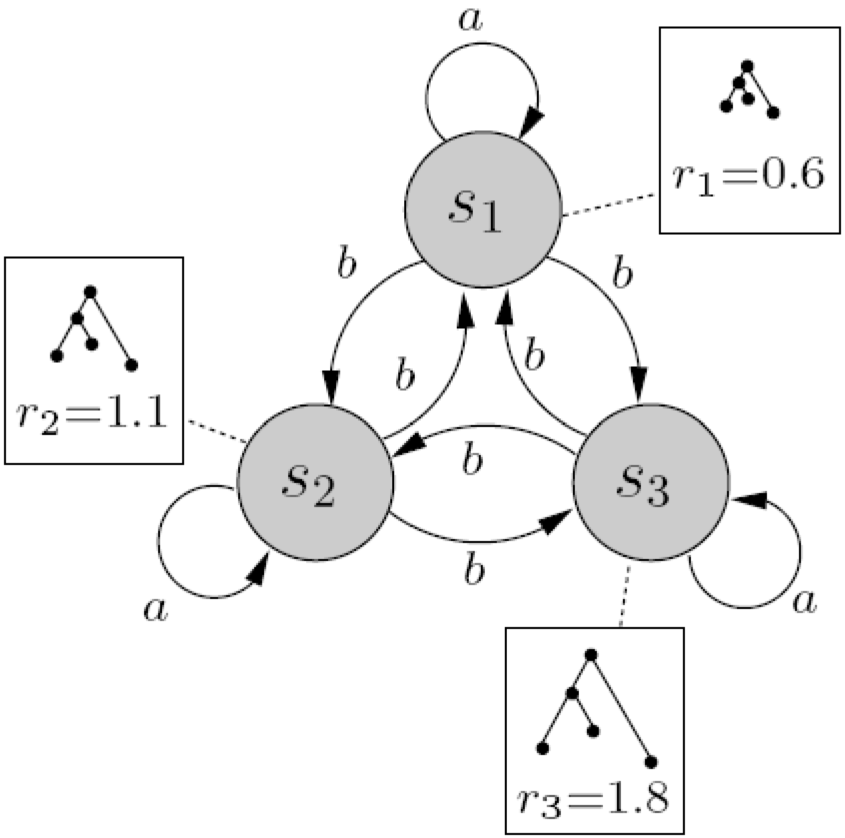
$$P(r_1, \ldots, r_N) = q(r_1) \prod_{t=2}^{N} q(r_t \mid r_{t-1}, v)$$

gives rise to a hidden Markov model:



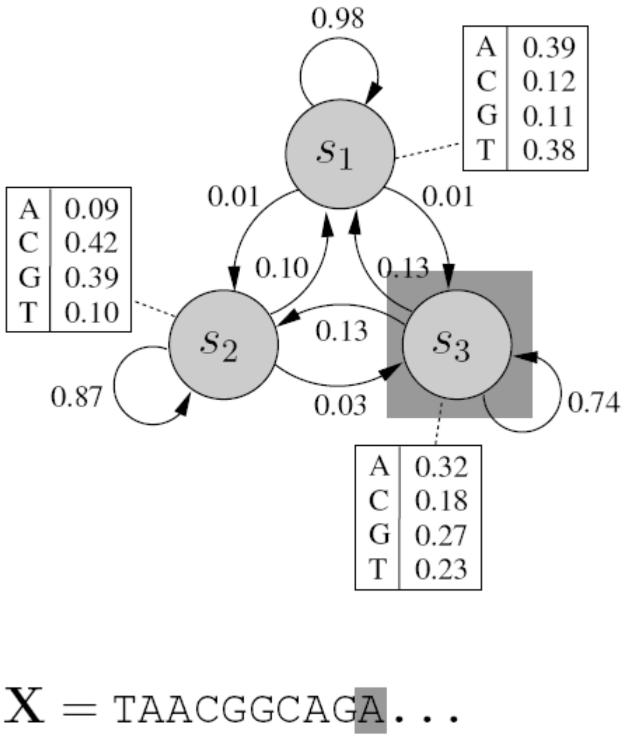
1	2	3	4	5	6	7	8
C	A	C	G	A	C	G	A
C	G	T	A	A	C	G	A
C	G	A	G	A	C	G	G
C	A	A	A	A	C	G	G
A	A	G	T	G	C	G	C

Markov chain state space



Phylogenetic hidden Markov models

HMM

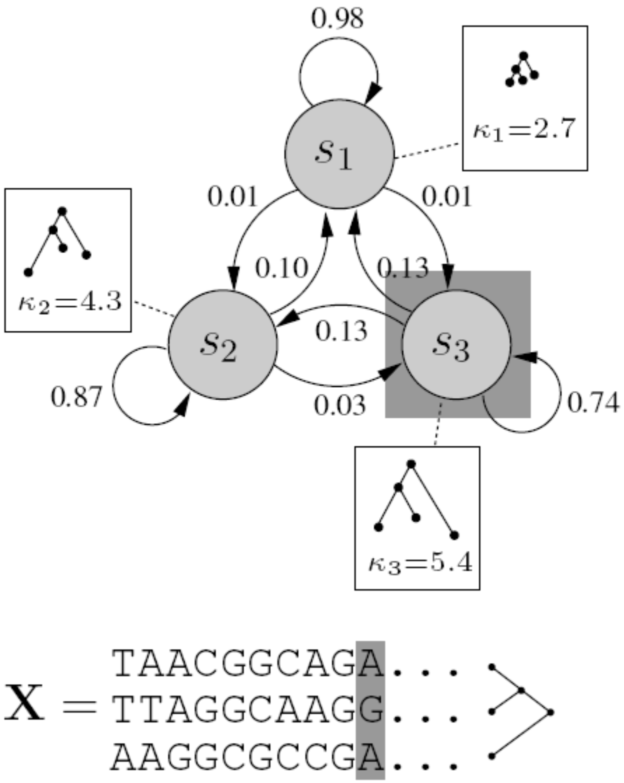


A	0.39
C	0.12
G	0.11
T	0.38

A	0.09
C	0.42
G	0.39
T	0.10

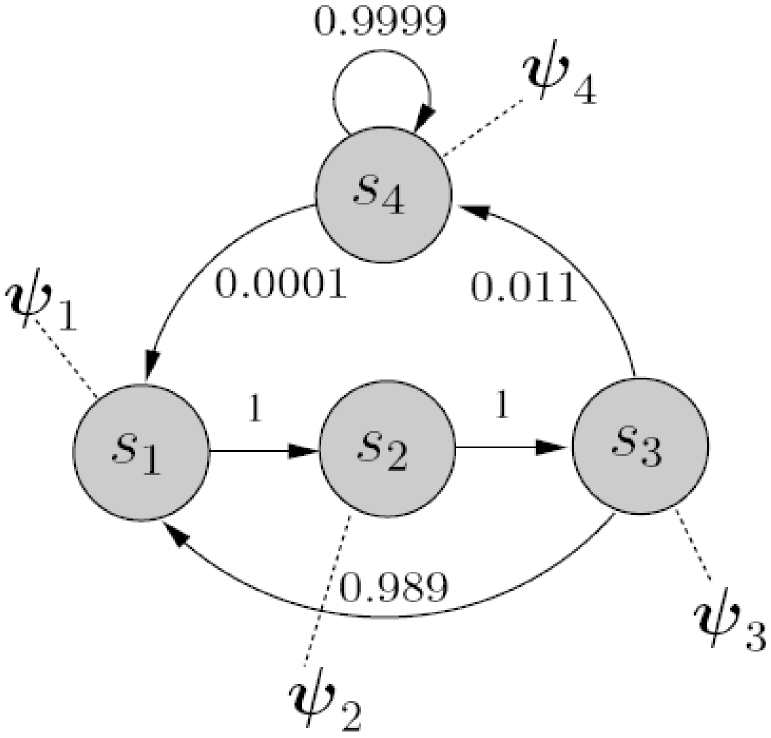
A	0.32
C	0.18
G	0.27
T	0.23

Phylo-HMM

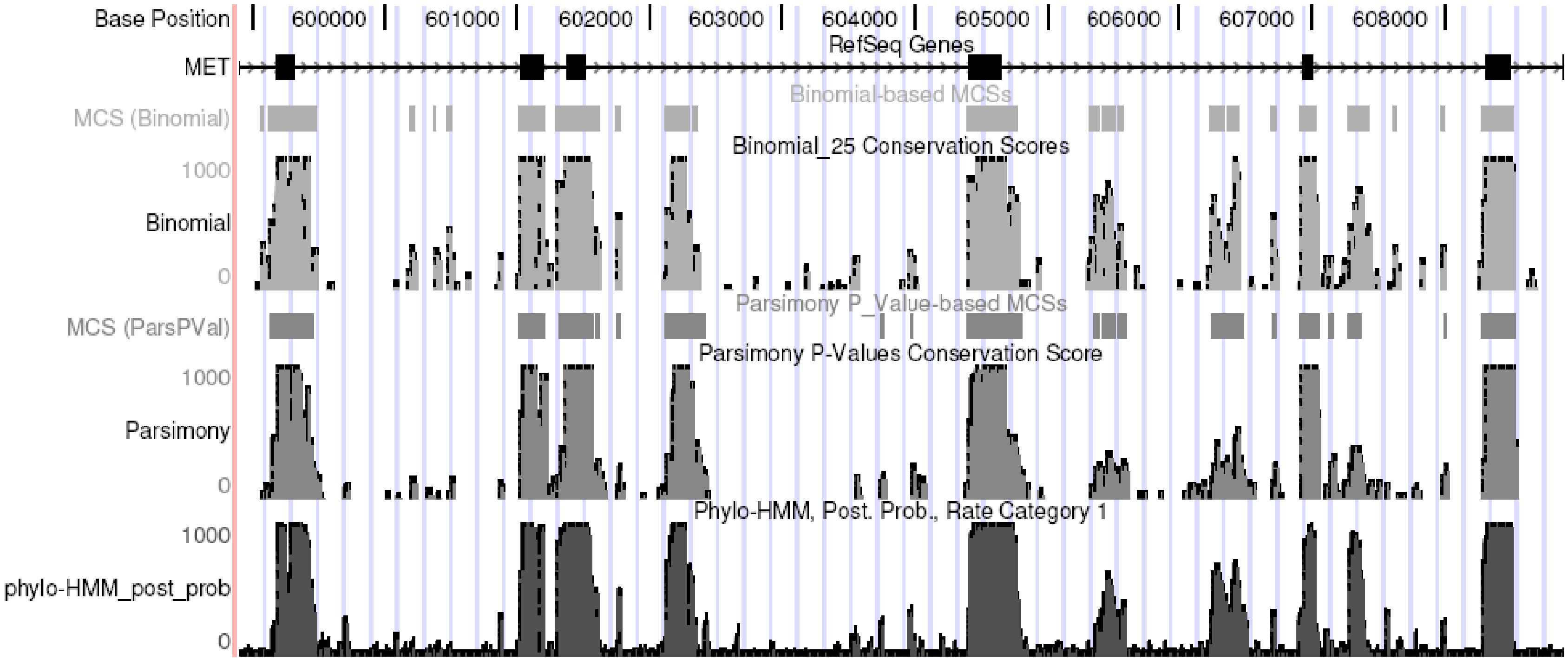


Phylo-HMM for gene finding

- Non-coding regions tend to have a higher substitution rate and a higher transition-transversion ratio
  - s_1, s_2, s_3 model codons
  - s_4 models non-coding sites
- ψ_1, ..., ψ_4 capture the parameters of the phylogenetic models



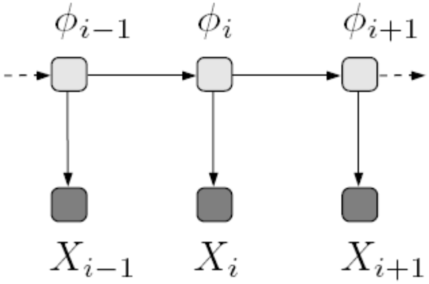
Detecting conserved genomic regions

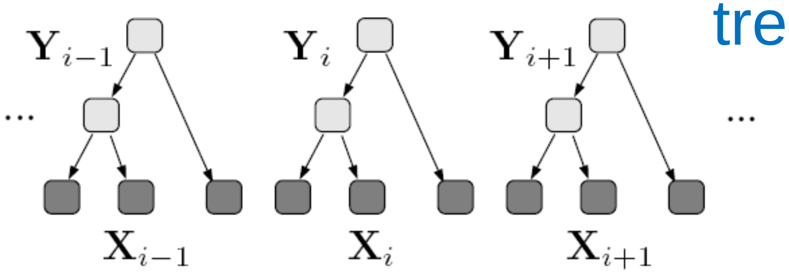


http://genome.ucsc.edu

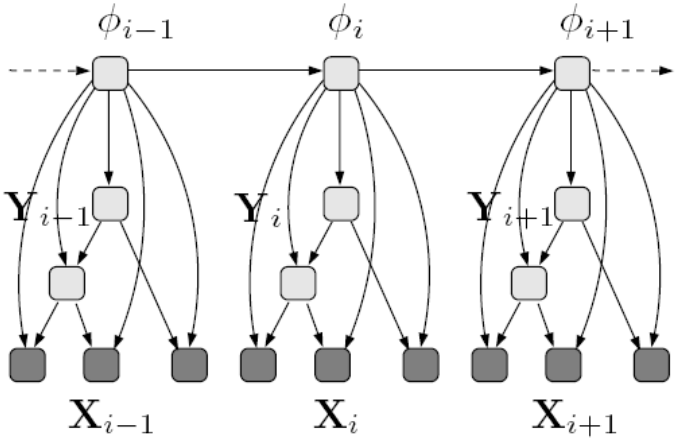
Phylo-HMMs are Bayesian networks

HMM





Phylogenetic
tree model



Phylo-HMM

Varying tree topologies are also possible (detection of recombination).

Summary

- Probabilistic phylogenetic tree models are Bayesian networks with observed and hidden random variables.
- The LPDs are defined by nucleotide substitution models, which are examples of continuous-time Markov chains.
- Phylogenetic trees can be learned using ML or Bayes.
- Rate heterogeneity across sites can be modeled using the Gamma distribution or by a HMM.
- Combining HMMs and phylogenetic trees gives rise to phylo-HMMs, a powerful model for sequence data with many applications.

References

- Husmeier D, Dybowski R, Roberts S (eds.). Probabilistic Modeling in Bioinformatics and Medical Informatics. Chapter 4.
- Beerenwinkel N and Siebourg J. Statistics, probability, and computational science. In Maria Anisimova, editor, *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*, chapter 3, pages 77–110. Springer, New York, 2012. DOI: 10.1007/978-1-61779-582-4_3. Sections 5, 7.
- Siepel A, Haussler D (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol* 11(2–3).

- Further reading:
  - Durbin R, Eddy S, Krogh A, Mitchinson G. Biological Sequence Analysis. Chapter 8.
  - Felsenstein J. Inferring Phylogenies. Sinauer, 2003.