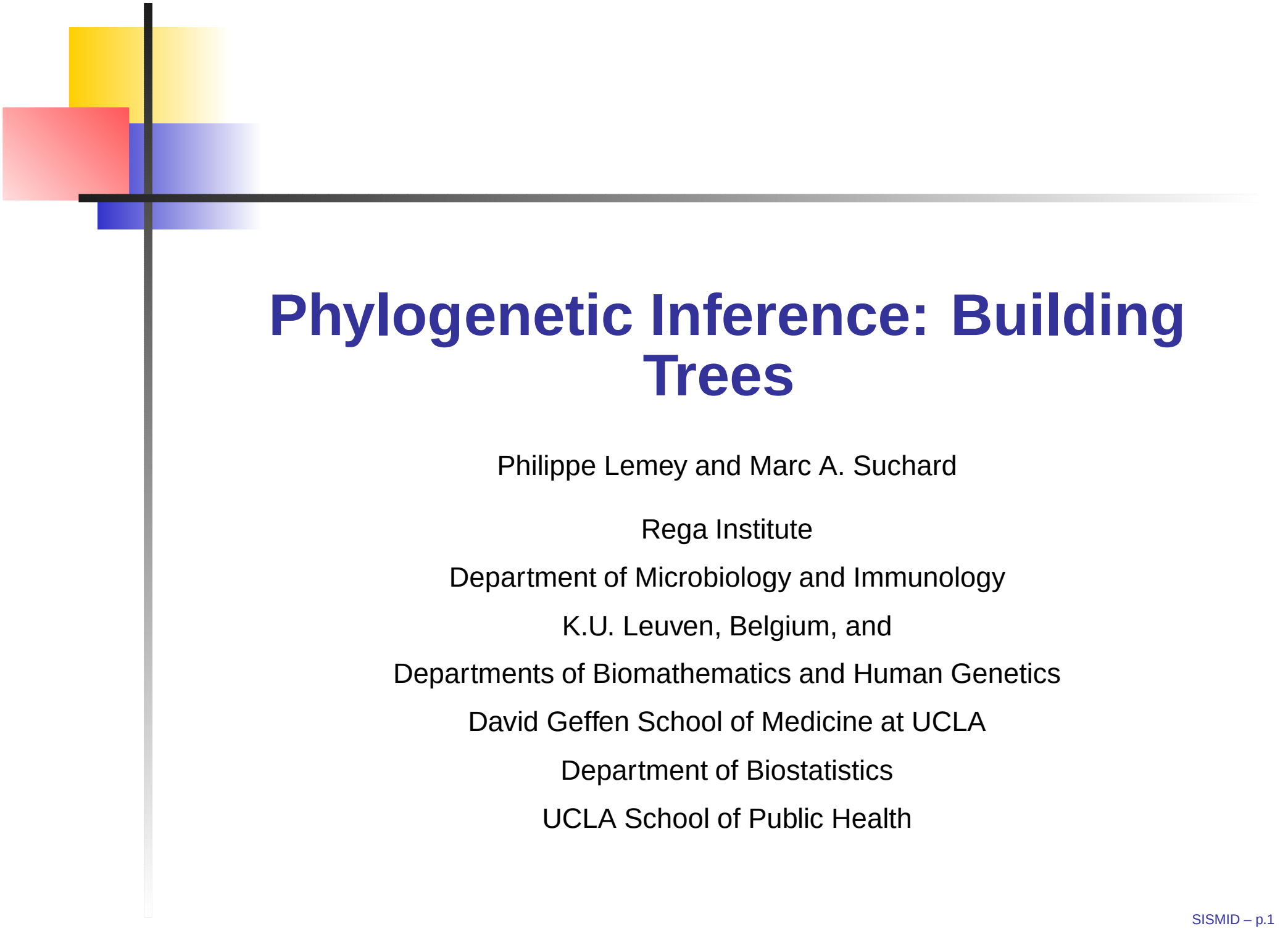


Phylogenetic Inference: Building Trees

Philippe Lemey and Marc A. Suchard

Rega Institute

Department of Microbiology and Immunology

K.U. Leuven, Belgium, and

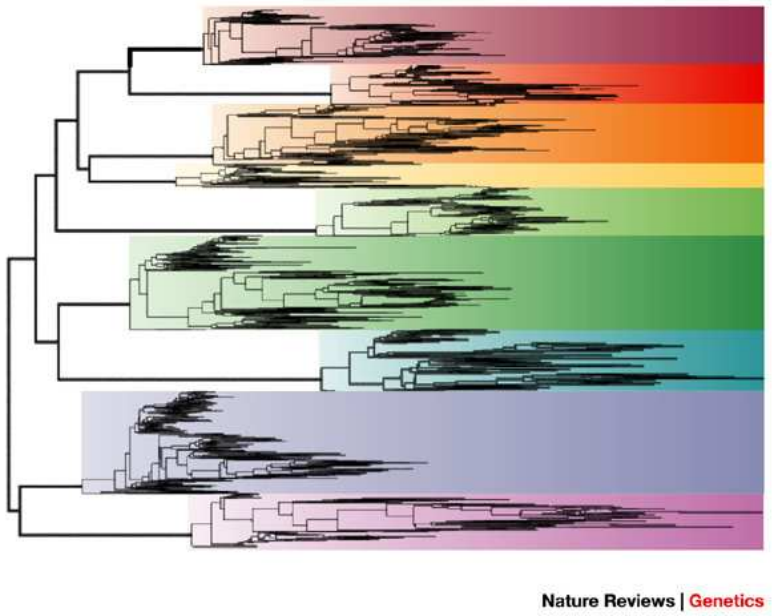
Departments of Biomathematics and Human Genetics

David Geffen School of Medicine at UCLA

Department of Biostatistics

UCLA School of Public Health

Intra-Host Viral Evolution



1195 *env* sequences from 9 HIV+ patients [taken from Rambaut et al. (2004)]

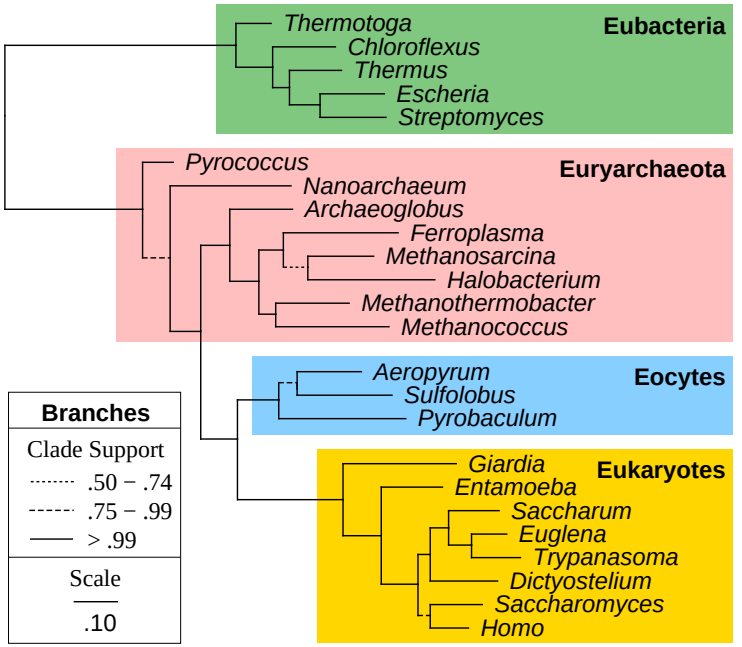
Retroviruses (and HBV) exist as a **quasi-species** within infected patients:

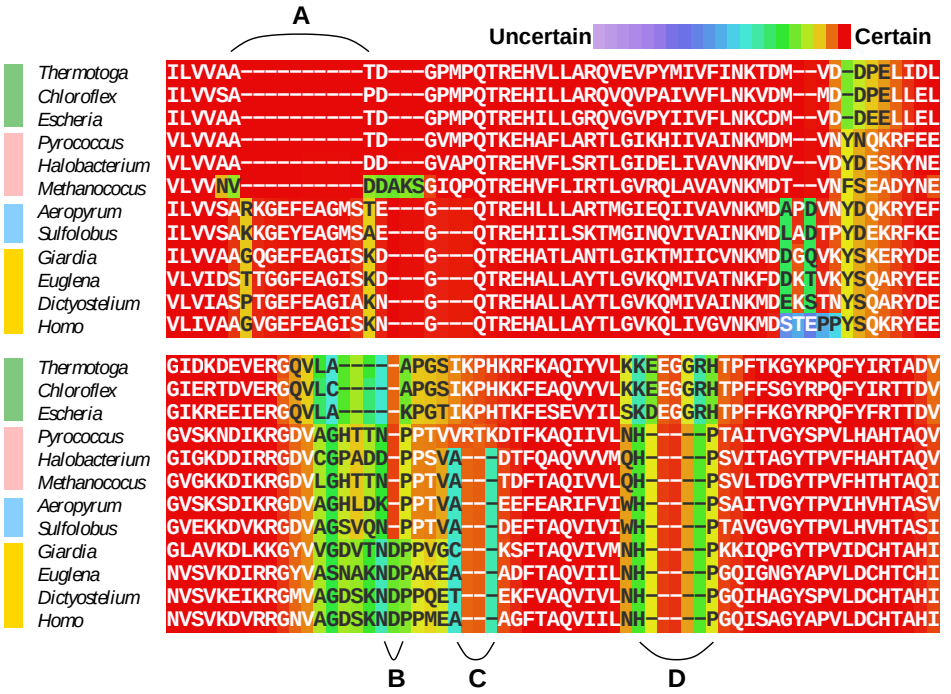
- Shared substitutions may be insufficient to resolve intra-host phylogenies

Improve resolution using joint model:

- Indel rates $\geq$ substitution rates
- Opportunity to detect intra-host recombination

Reconstructing the Tree of Life: Are Humans Just Big Slime Molds?



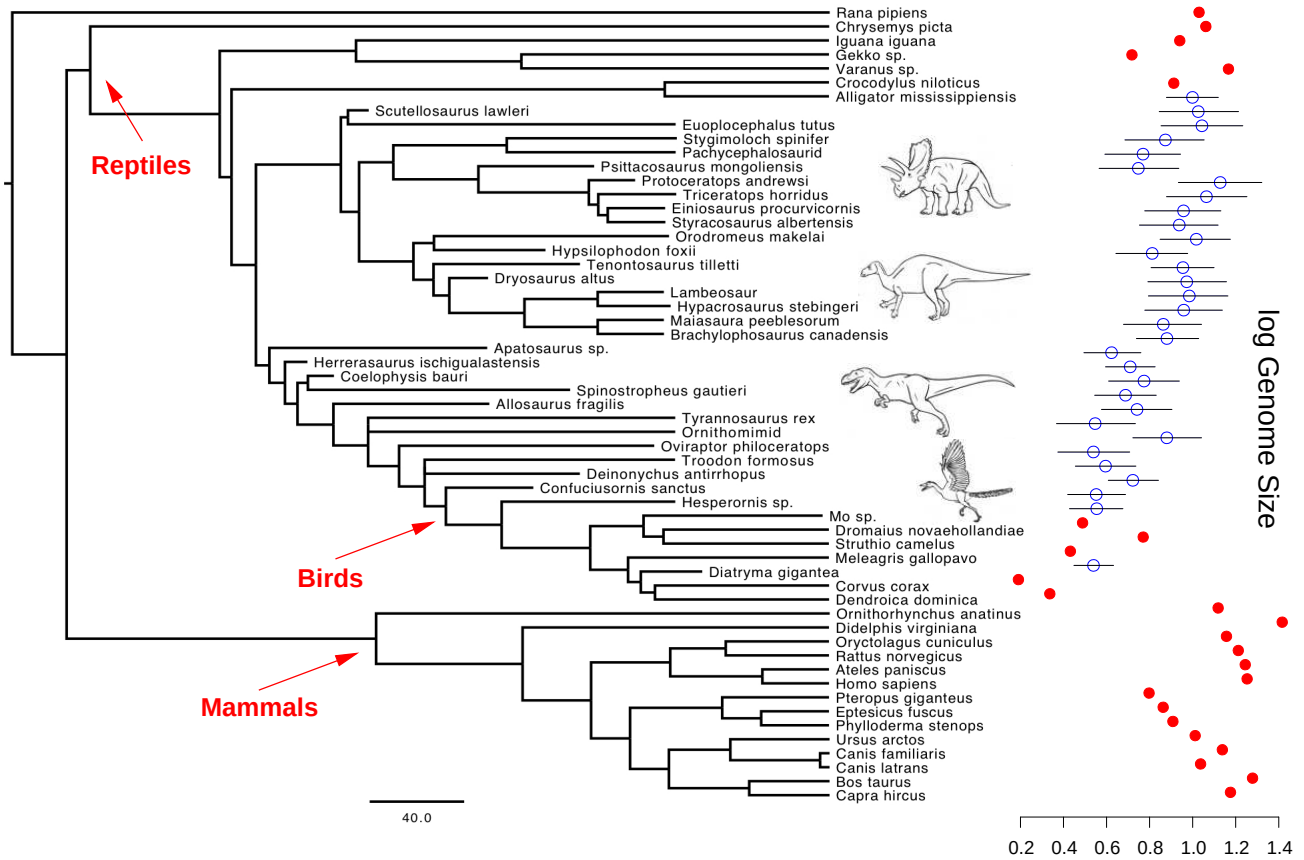


MAP Tree

Partial Au Plot

- Contentious issue among paleobiologists: Do **Archaea** (Euryarchaeota/Eocytes) form one or two domains? *Weekly World News* calls humans slime molds.

The Chicken or the (Small) Genome: Which Came First?



Evolutionary History and Genome Sizes of Reptiles, Dinosaurs, Birds and Mammals

Issue: Bird genomes are markedly smaller than those from other vertebrates.

Question: Did small genomes precede flight or co-evolve?

Maximum Parsimony (MP)

Most often used $\neq$ "best", not even statistically consistent, but **fast, fast, fast** . . . if you know the tree

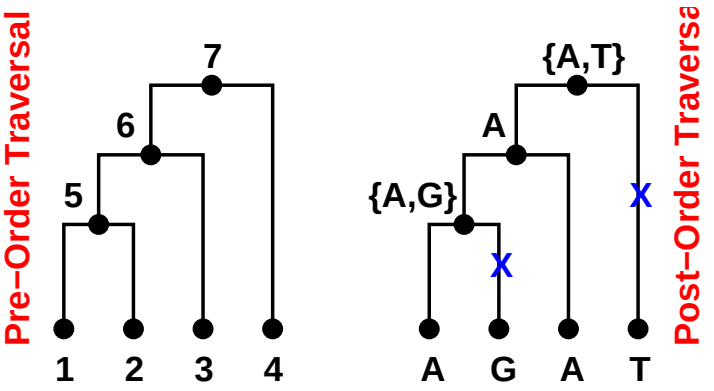
Key: Find tree with minimal # of "suspected" substitutions (internal states are not observed, 0/1 model process)

- Counting minimum # of substitutions is **easy**
- Enumerating (searching through) all possible trees is **hard**

Human	T	T	C	C	T	G	G	A	A	T
Chimp	T	A	C	C	T	G	G	A	A	T
Mouse	A	A	C	C	T	–	–	T	A	T
Fly	A	A	G	A	T	C	G	T	A	T
Site:	1	2	3	4	5	6	7	8	9	10

Along Molecular Sequence

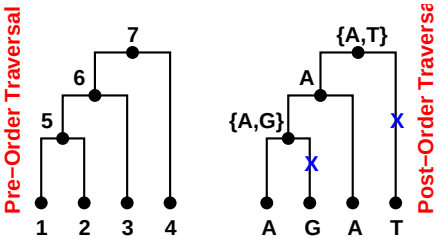
Sites are **independent**



Maximum Parsimony (MP)

A little history:

- Anthony Edwards/Luca Cavalli-Sforza (1963,1964)
  - Both students of R.A Fisher
  - Introduced both **parsimony** and **likelihood** methods (for continuous quantities, e.g. gene frequencies) in one paper
- Camin and Sokal (1965) provide first program for molecular sequences
- Fitch and Margoliash (1967) provide efficient algorithm

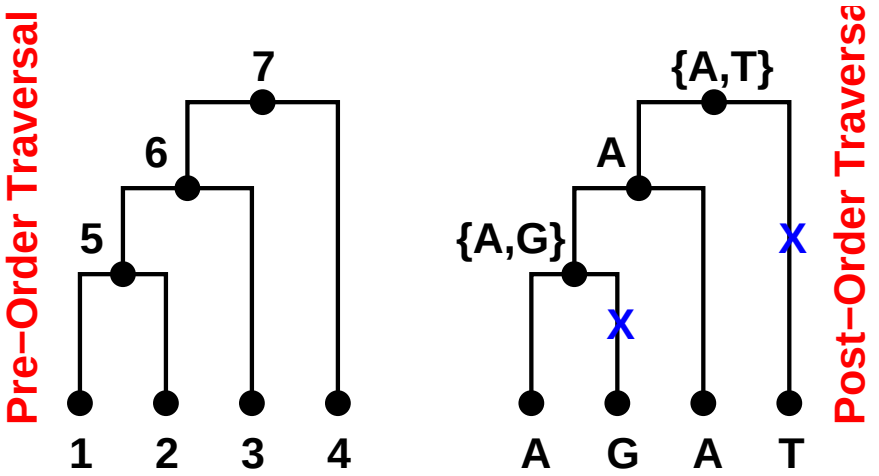


Maximum Parsimony Algorithm

procedure Fitch and Margoliash (1967) Algorithm

cost $C \leftarrow 0$ {**Initialization**}
pointer $k \leftarrow 2N - 1$ {at the root node}
To obtain the set R_k of possible states at node k {**Recursion**}
if k is leaf **then**
  $R_k \leftarrow$ observed character for taxon k
else
  Compute R_i, R_j for daughters i, j of k
  if $R_i \cap R_j \neq 0$ **then**
    $R_k \leftarrow R_i \cap R_j$
  else
    $R_k \leftarrow R_i \cup R_j$
    $C \leftarrow C + 1$
  end if
end if

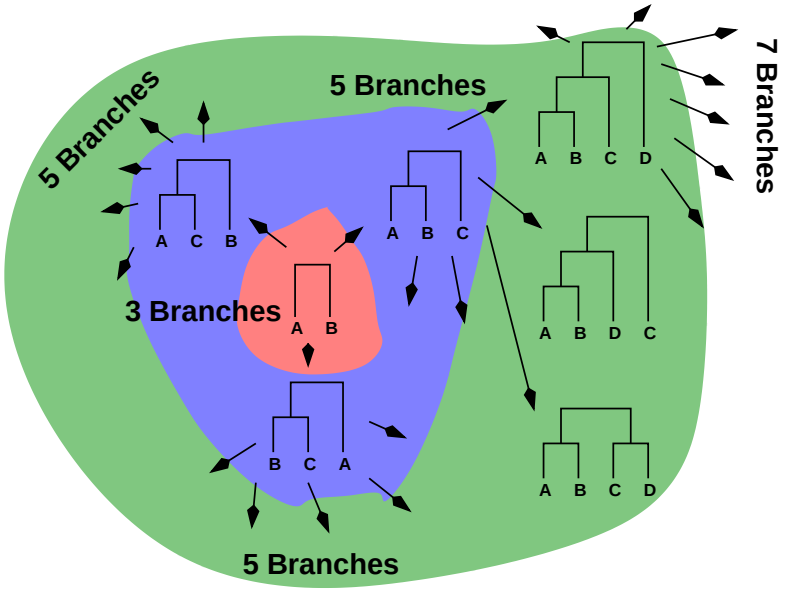
minimum cost is C {**Termination**}



Searching for the MP Tree

Complexity:

- Find MP score is **NP-complete**
- Find MP tree is **NP-hard**



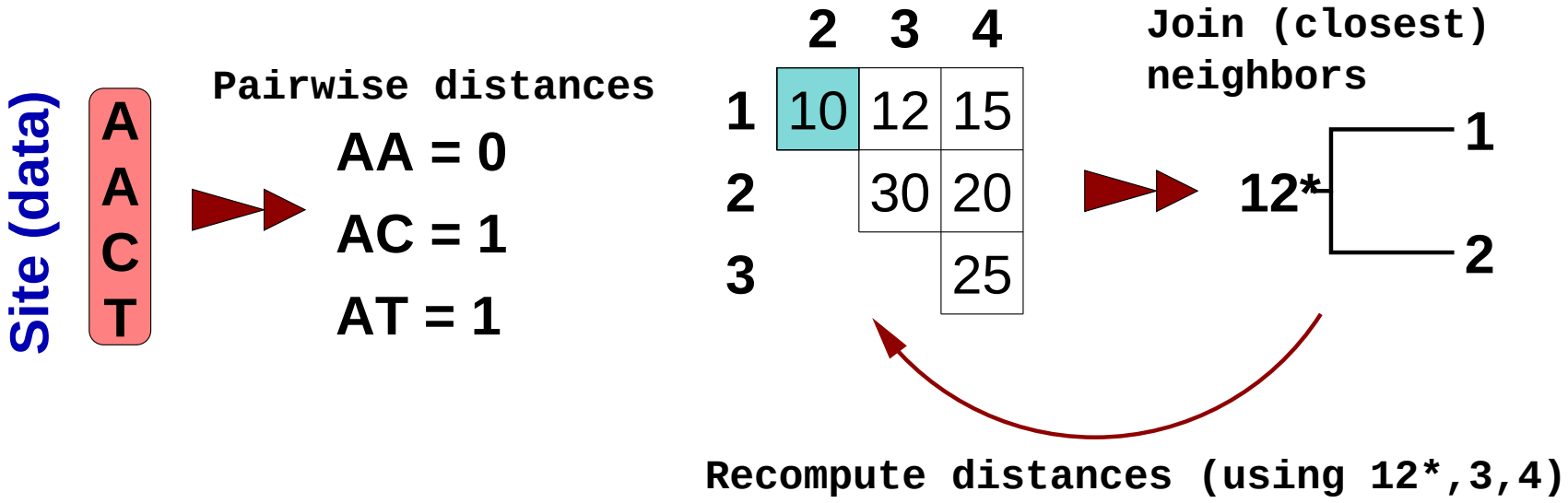
Recall that # of N-taxon rooted trees is $3 \times 5 \times \cdots \times 2N - 3$

Attack exponential-order space **Branch-and-Bound**:

- Monotonic order: min $\text{PS}_2 \leq$ min $\text{PS}_3 \leq \ldots$
- Bound if min $\text{PS}_k >$ best n-taxon PS found so far.

Neighbor-Joining (Saitou and Nei, 1987)

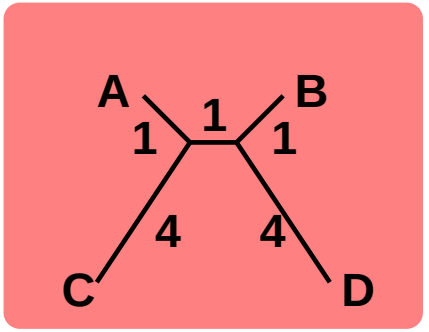
Computational algorithm: **alignment** → **single tree**



- **Advantages**: very fast, great for 1000s of sequences
- **Disadvantages**: no site-to-site rate variation, no **natural** ways to compare trees/measure data support

Neighbor-Joining

Caveat: Pairs i, j with min d_{ij} are **not** necessarily nearest neighbors. E.g., $d_{\text{AB}} = 3 < d_{\text{AC}} = 5$



Solution: Subtract off the average distances to all other leaves via

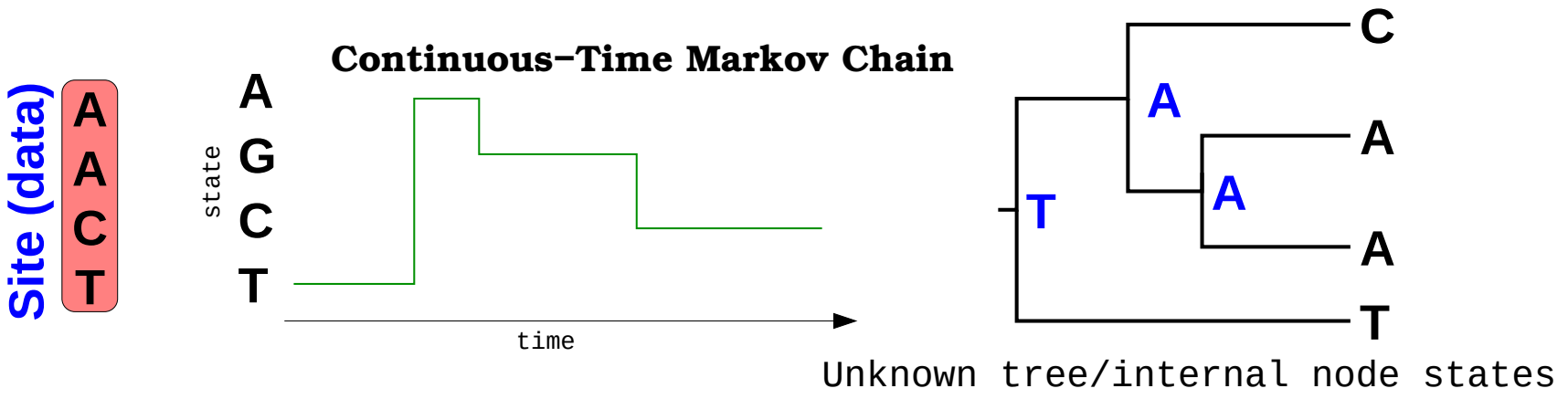
$$D_{ij} = d_{ij} - (r_i + r_j), \quad r_i = \frac{1}{|L| - 2} \sum_{k \in L} d_{ik},$$

where L is the current set of leaves. Proof in Studier and Keppler (1988).

Computational: $O(N^3)$

Likelihood-based Methods (Felsenstein, 1973)

Statistical technique: assumes an **unknown** tree and a stochastic model for character change along the tree

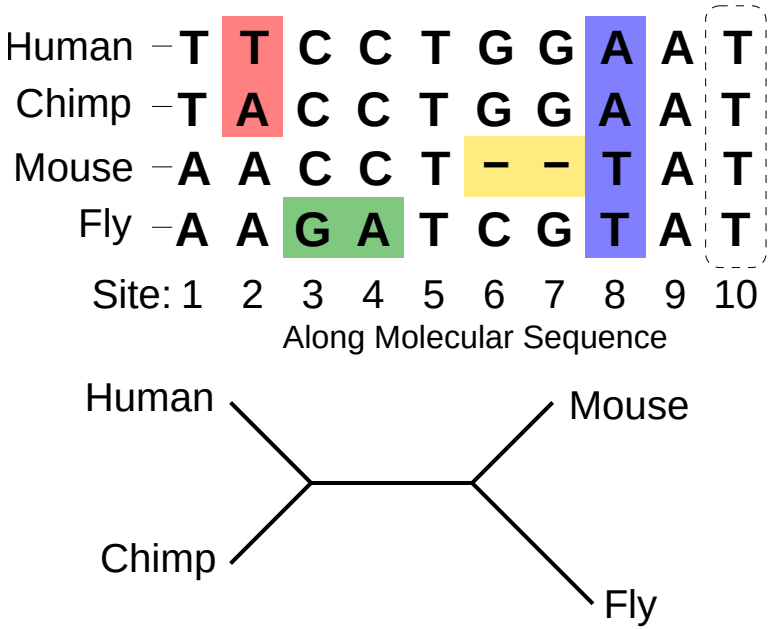


- **Advantages**: site-to-site rate/tree variation is easy, can formulate probability statements
- **Disadvantages**: must "search" tree-space $\rightarrow$ slow

Foundation of Bayesian Phylogenetics

Traditional Phylogenetic Reconstruction

Reconstruction Example



- **Substitution**: single residue replaces another
- **Insertion/deletion**: residues are inserted or deleted

Statistical Model

Assume: Homologous sites are iid and site patterns (e.g. dotted box)

$$XY \ldots Z \sim \text{Multinomial}(p_{XY \ldots Z})$$

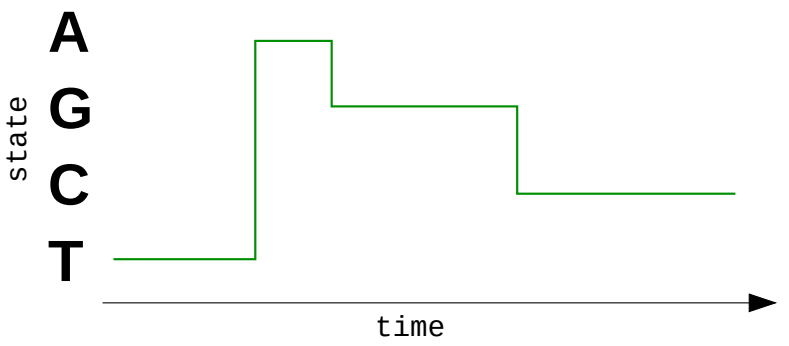
where $p_{XY \ldots Z}$ is determined by an unknown tree τ, branch lengths $t \in \mathbf{T}$ and continuous-time Markov chain model (for residue substitution) given by infinitesimal rate matrix $\mathbf{Q}$

$$\text{P}(X \to Y \text{ in time } t) = e^{t\mathbf{Q}}$$

CTMC(Q) = $\epsilon \sim$ Normal(μ, σ^2) of Phylogenetics

Continuous in elapsed time t, discrete in starting/ending state!

Memory-less process in which the probability that state b replaces state a during $(t, t+s) = sq_{ab} + o(s)$



- Infinitesimal generator matrix $\boldsymbol{Q}$ has off-diagonal entries q_{ab} and row sums $= 0$

Think: Exponential waiting time with rate $R_a = \sum_b q_{ab}$ until chain leaves a. Then the new state b is independently chosen with probabilities q_{ab}/R_a

From Infinitesimal to Finite Time

Let $p_{ab}(t) =$ the finite-time probability of the chain moving from state a at time 0 to state b at time t, then matrix $\boldsymbol{P}(t) = \{p_{ab}(t)\}$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}(t) = \boldsymbol{P}(t)\boldsymbol{Q} \text{ where } \boldsymbol{P}(0) = \boldsymbol{I}$$

with solution

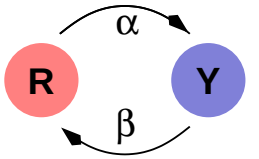
$$\boldsymbol{P}(t) = e^{t\boldsymbol{Q}} = \boldsymbol{I} + t\boldsymbol{Q} + \frac{1}{2}(t\boldsymbol{Q})^2 + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!}(t\boldsymbol{Q})^k$$

as

$$\frac{\mathrm{d}}{\mathrm{d}t}e^{t\boldsymbol{Q}} = \boldsymbol{Q}e^{t\boldsymbol{Q}} = e^{t\boldsymbol{Q}}\boldsymbol{Q} \text{ for } t \text{ real}$$

Example: Two-State Model

Consider purines (R) $\leftrightarrow$ pyrimidines (Y). Kolmogorov forward equation:



$$p_{\mathsf{RY}}(t+s) = p_{\mathsf{RR}}(t)\alpha s + p_{\mathsf{RY}}(t)(1-\beta s) + o(s)$$

yielding

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{\mathsf{RY}}(t) = \alpha p_{\mathsf{RR}}(t) - \beta p_{\mathsf{RY}}(t)$$

$$\boldsymbol{Q} = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$$

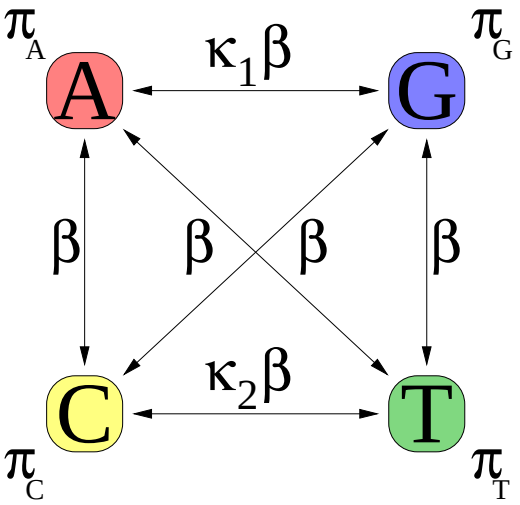
Solutions of $\boldsymbol{P}(t) = e^{t\boldsymbol{Q}}$ have the form

$$c + de^{-(\alpha+\beta)t}$$

with eigenvalues 0 and $-(\alpha+\beta)$

Standard CTMCs for Phylogenetics

- Jukes and Cantor (JC69), $\pi_a = \frac{1}{4}, \kappa_1 = \kappa_2 = 1$
- Kimura (K80), $\pi_a = \frac{1}{4}, \kappa_1 = \kappa_2$
- Hasegawa, Kishino and Yano (HKY85), $\kappa_1 = \kappa_2$ (most common)
- Tamura and Nei (TN93), right
- General Time Reversible (GTR)



Note identifiability concern in e^{tQ}. Common solution is to fix 1 d.f. such that

$$\sum_a q_{aa}\pi_a = -1$$

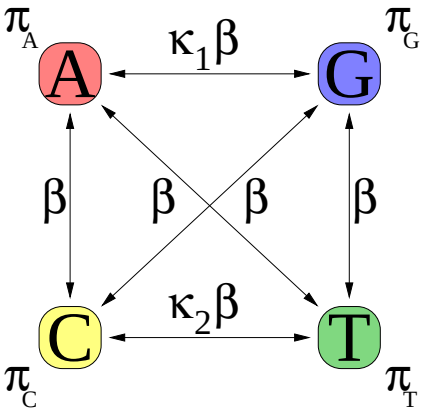
Scaling: $t = 1 \Rightarrow 1$ expected substitution per site

Explicit Parameterization of TN93

Nucleotides mutate according to a **Markovian** process

$$\Pr(X \to Y \text{ in time } t) = e^{tQ_{\text{Nuc}}}$$

where Q_{Nuc} is a 4x4 infinitesimal rate matrix and t is a branch length.



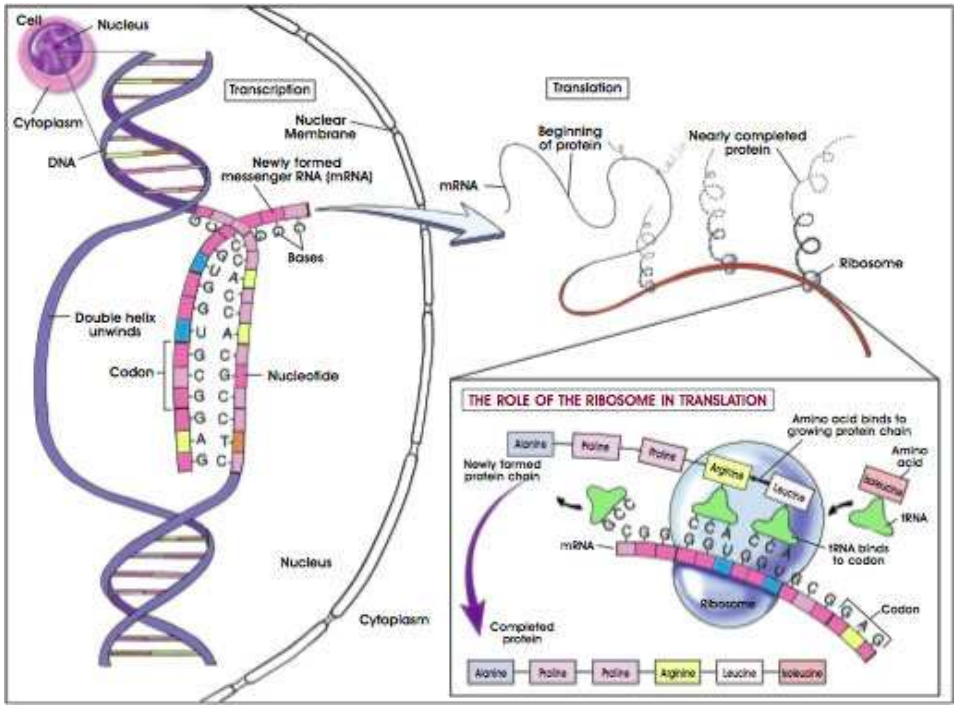
$$Q_{\text{Nuc}} = \beta \times \begin{pmatrix} - & \kappa_1\pi_G & \pi_C & \pi_T \\ \kappa_1\pi_A & - & \pi_C & \pi_T \\ \pi_A & \pi_G & - & \kappa_2\pi_T \\ \pi_A & \pi_G & \kappa_2\pi_C & - \end{pmatrix}, \text{ where}$$

κ_1, κ_2 are transition:transversion rate ratios and π is the stationary distribution of {A,G,C,T}. β controls the overall rate and can vary from site-to-site.

Site-to-Site Rate Variation

Variation occurs quite naturally and is also **an important inference**

- short range: codon phase (slow-slow-fast)
- long range: enzymatic active sites, protein folding, immunological pressures/selection



Assume: infinitesimal rates for site k are $r_k \times t \times q_{ab}$. Various priors on r_k with $\mathrm{E}(r_k) = 1$. Implicitly Bayesian

- Yang (1994) – discretized Gamma distribution

General Time Reversible CTMC

Let

$$\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{D}_{\pi}$$

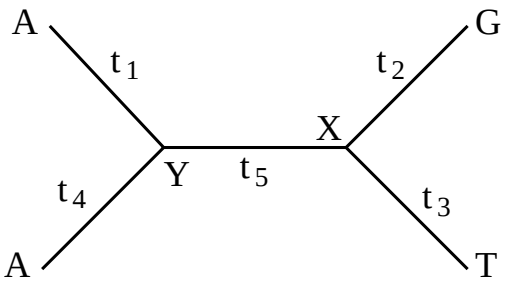
where $\boldsymbol{R}$ is symmetric and $\boldsymbol{D}_{\boldsymbol{\pi}}$ is a diagonal matrix composed of the stationary distribution $\boldsymbol{\pi}$.

- Detailed balance $\Leftrightarrow \pi_a q_{ab} = \pi_b q_{ba}$. Balance + irreducibility $\Leftrightarrow$ reversible
- **Note** $\boldsymbol{Q}$ is similar to $\boldsymbol{R}$, as $\boldsymbol{D}^{1/2}\boldsymbol{Q}\boldsymbol{D}^{-1/2} = \boldsymbol{R}$
- Hence, $\boldsymbol{Q}$ must have real eigenvalues and real eigenvectors

The properties speed up computation of the finite-time transition matrix $\boldsymbol{P}(t) = e^{t\boldsymbol{Q}}$

Calculating the Probability of a Single Site Pattern Y_i

Given the tree and **unobserved** internal node states, the probability is the product of the finite time mutation probabilities over all branches:



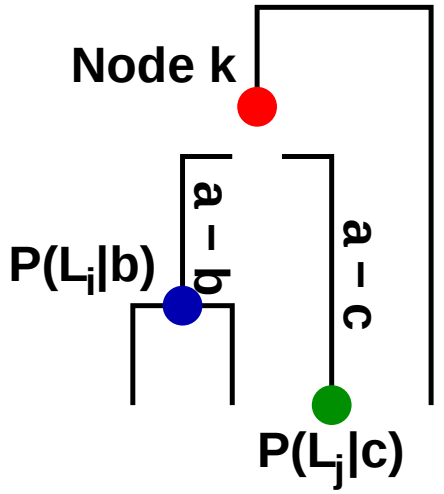
$$L(\boldsymbol{Y}_i) \propto p_{\mathsf{AAGT}} = \sum_X \sum_Y \Pr(Y \to \mathsf{A}, t_1) \Pr(X \to \mathsf{G}, t_2) * \Pr(X \to \mathsf{T}, t_3) \Pr(Y \to \mathsf{A}, t_4) \Pr(X \to Y, t_5) \pi_X \quad (1)$$

- Number of sumants grow rapidly in $N \to$ sum-product/peeling algorithm to distribute sums across the product

Pruning Algorithm Felsenstein (1981)

Let $P(L_k|a)$ = likelihood of leaves below node k given k is in state a. Then, recursively compute $P(L_k|a)$ given $P(L_i|b)$ and $P(L_j|c)$ for daughters i, j of k:

Set pointer $k \leftarrow 2N - 1$ {the root, **initialization**}
Compute $P(L_k|a)\ \forall\ a$ as follows: {**recursion**}
if k is a leaf node **then**
  if a is observed **then**
    $P(L_k|a) = 1$
  else
    $P(L_k|a) = 0$
  end if
else
  Compute $P(L_i|a)$ and $P(L_j|a)\ \forall a$ for daughters i, j of k {**post-order traversal**}
  $P(L_k|a) = \sum_b \sum_c \Pr(a \rightarrow b, t_i) P(L_i|b) \times \Pr(a \rightarrow c, t_j) P(L_j|c)$
end if
$L(\mathbf{Y}_i) \leftarrow \sum_a P(L_{2n-1}|a)\pi_a$ {**termination**}



ML Tree or MAP Tree?

Reporting uncertainty on tree estimates:

- The Bootstrap
  - Most common
  - Assumes evolutionary events are reproducible. “If I went back out to the field and recollected exchangeable data . . . ”
- Bayesian inference
  - Returns the probability of a tree given the observed data and model
  - Requires MCMC (e.g., **MrBayes** or **BEAST**)
  - **Advantages**
    - Does not rely on asymptotics (hypothesis testing)
    - Naturally incorporates uncertainty in all parameters (including discrete quantities: trees, site-classifications, etc.)
    - Arguably faster algorithms
  - **Disadvantages**
    - Must specify (justifiable) prior distributions