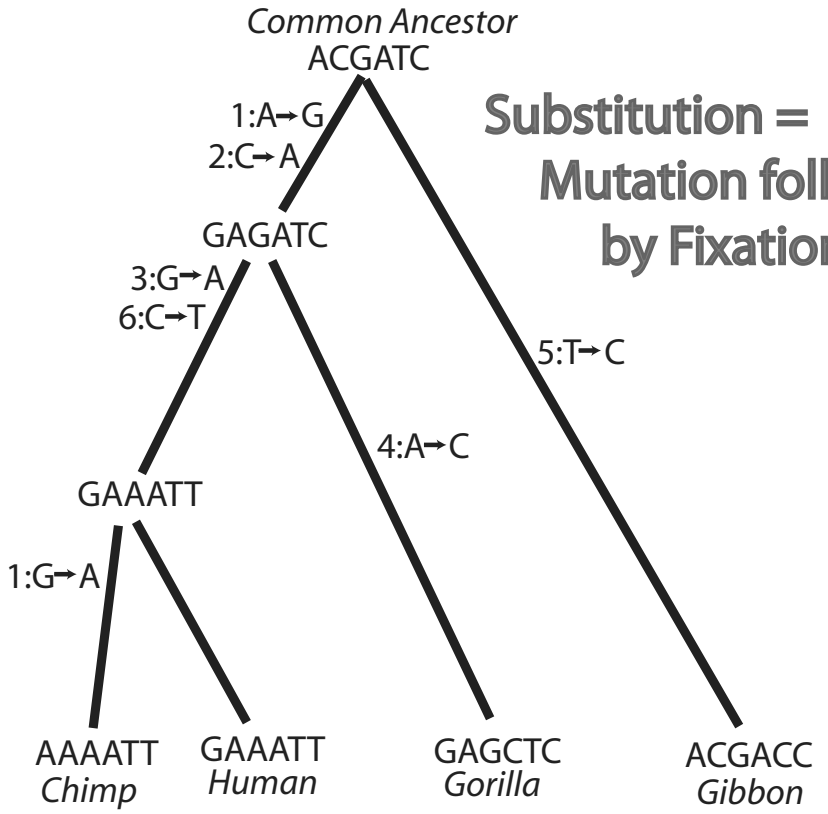


Substitution = Mutation followed by Fixation

AAAATT *Chimp*

GAAATT *Human*

GAGCTC *Gorilla*

ACGACC *Gibbon*

Likelihood (Prob. of data given model & parameter values) =

AAAATT *Chimp* GAAATT *Human* GAGCTC *Gorilla* ACGACC *Gibbon*

Likelihood for Site 1 X

AAAATT *Chimp* GAAATT *Human* GAGCTC *Gorilla* ACGACC *Gibbon*

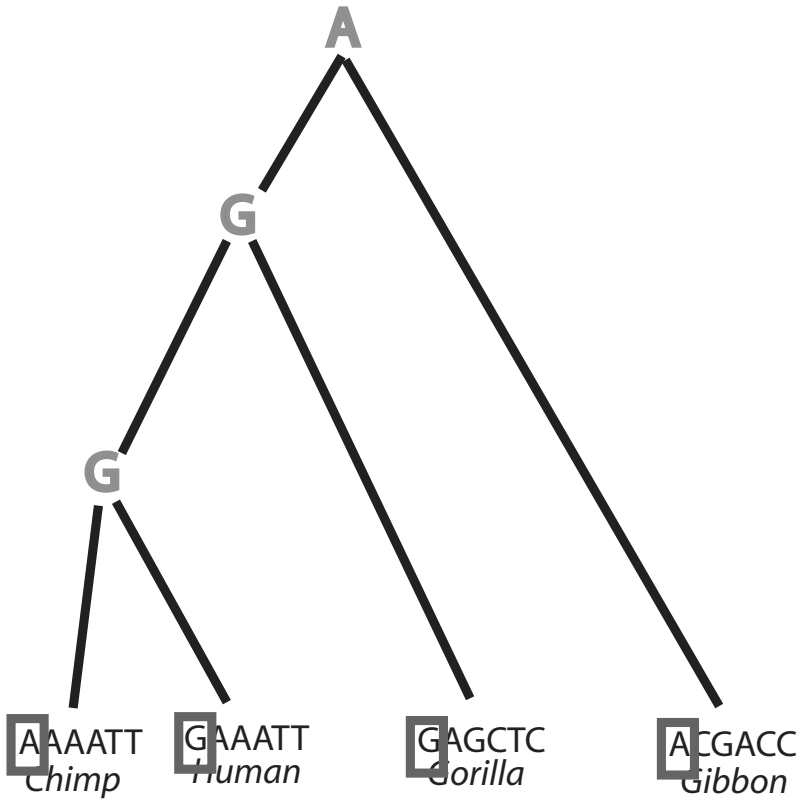
Likelihood for Site 2 X

AAAATT *Chimp* GAAATT *Human* GAGCTC *Gorilla* ACGACC *Gibbon*

... X ... X ... X

Likelihood for Site 6

AAAATT *Chimp* GAAATT *Human* GAGCTC *Gorilla* ACGACC *Gibbon*



Probabilistic models of nucleotide change (independently and identically evolving sites)

Let q_{ij} be the instantaneous rate of change at a site from nucleotide type i to j

Q is matrix of instantaneous rates (Q will have 4 rows and 4 columns because i and j can each by any of 4 nucleotide types)

For nucleotide starting as type i at time 0, probability nucleotide is type j at time t is denoted $p_{ij}(t)$.

$p_{ij}(t)$ is referred to as a *transition probability*.

Consider a **very very** small amount of evolutionary time Δt. When $i \neq j$,

$$p_{ij}(\Delta t) \doteq q_{ij}\Delta t$$

$$p_{ii}(\Delta t) \doteq 1 - \sum_{j,j\neq i} q_{ij}\Delta t$$

$$p_{ii}(\Delta t) \doteq 1 + q_{ii}\Delta t$$

where

$$q_{ii} = - \sum_{j,j\neq i} q_{ij}$$

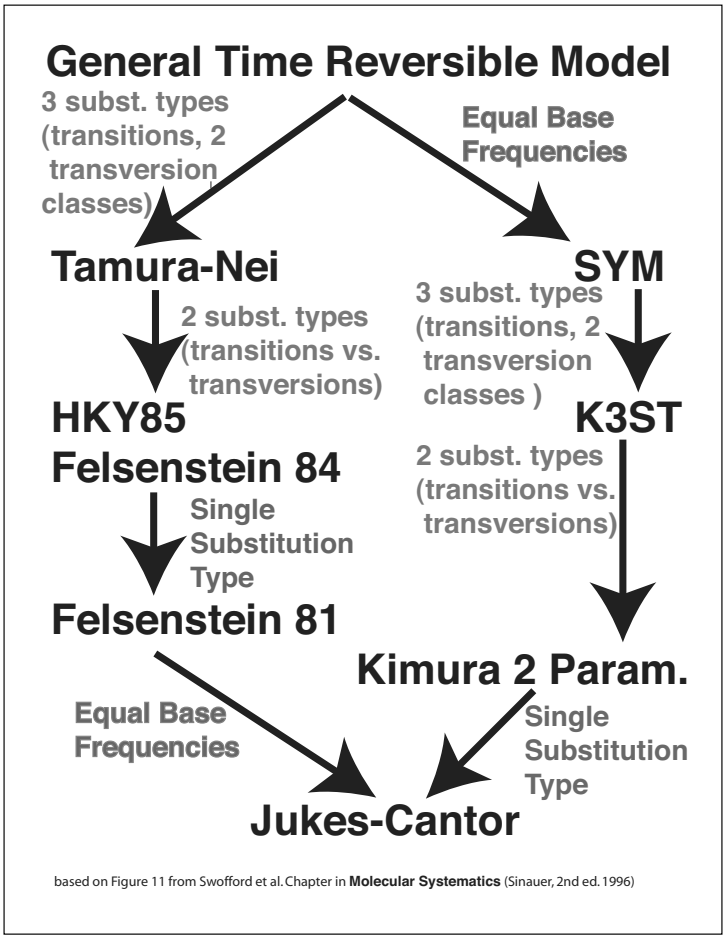
(in preceding equations, $\doteq$ can be replaced by $=$ when the limit as Δt approaches 0 is taken)

Jukes-Cantor model is simplest model of nucleotide substitution.
It assumes sequence positions evolve independently and it assumes that all possible changes at a position are equally likely.

Let π_j be probability a residue is type j. π_j is called the equilibrium probability of type j.

$$p_{ij}(\infty) = \pi_j$$

For Jukes-Cantor model, $\pi_j = 1/4$ for all 4 nucleotide types j.



Rate Matrix for Jukes-Cantor Model

FROM	**To** A	C	G	T
A	-3μ	μ	μ	μ
C	μ	-3μ	μ	μ
G	μ	μ	-3μ	μ
T	μ	μ	μ	-3μ

Note 1: Diagonal matrix elements multiplied by -1 are **rate away** from nucleotide type of that row.

Note 2: In later slide on Jukes-Cantor model, we write $s/3$ rather than μ.

Rate Matrix for Kimura 2-Parameter Model

FROM	**To** A	C	G	T
A	$-\alpha-2\beta$	β	α	β
C	β	$-\alpha-2\beta$	β	α
G	α	β	$-\alpha-2\beta$	β
T	β	α	β	$-\alpha-2\beta$

Changes involving only purines (i.e., A and G) or only pyrimidines (i.e., C and T) are transitions. Changes involving one purine and one pyrimidine are transversions.

Rate Matrix for Felsenstein 1981 Model

FROM \ **To**	A	C	G	T
A	$-\mu(\pi_C + \pi_G + \pi_T)$	$\mu\pi_C$	$\mu\pi_G$	$\mu\pi_T$
C	$\mu\pi_A$	$-\mu(\pi_A + \pi_G + \pi_T)$	$\mu\pi_G$	$\mu\pi_T$
G	$\mu\pi_A$	$\mu\pi_C$	$-\mu(\pi_A + \pi_C + \pi_T)$	$\mu\pi_T$
T	$\mu\pi_A$	$\mu\pi_C$	$\mu\pi_G$	$-\mu(\pi_A + \pi_C + \pi_G)$

Rate Matrix for Hasegawa-Kishino-Yano (a.k.a. HKY or HKY85) Model

FROM \ **To**	A	C	G	T
A	$-\mu(\pi_C + \kappa\pi_G + \pi_T)$	$\mu\pi_C$	$\mu\kappa\pi_G$	$\mu\pi_T$
C	$\mu\pi_A$	$-\mu(\pi_A + \pi_G + \kappa\pi_T)$	$\mu\pi_G$	$\mu\kappa\pi_T$
G	$\mu\kappa\pi_A$	$\mu\pi_C$	$-\mu(\kappa\pi_A + \pi_C + \pi_T)$	$\mu\pi_T$
T	$\mu\pi_A$	$\mu\kappa\pi_C$	$\mu\pi_G$	$-\mu(\pi_A + \kappa\pi_C + \pi_G)$

Rate Matrix for General Time Reversible Model

FROM \ **To**	A	C	G	T
A	$-\mu(a\pi_C + b\pi_G + c\pi_T)$	$\mu a\pi_C$	$\mu b\pi_G$	$\mu c\pi_T$
C	$\mu a\pi_A$	$-\mu(a\pi_A + d\pi_G + e\pi_T)$	$\mu d\pi_G$	$\mu e\pi_T$
G	$\mu b\pi_A$	$\mu d\pi_C$	$-\mu(b\pi_A + d\pi_C + f\pi_T)$	$\mu f\pi_T$
T	$\mu c\pi_A$	$\mu e\pi_C$	$\mu f\pi_G$	$-\mu(c\pi_A + e\pi_C + f\pi_G)$

Time Reversibility is a common property of models of sequence evolution.

Time reversibility means that $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$ for all i, j, and t.

$\pi_i q_{ij} = \pi_j q_{ji}$ for all i and j.

For phylogeny reconstruction, time reversibility means that we cannot (on the basis of sequence data alone) hope to distinguish which of two sequence is ancestral and which is the descendant.

The practical implication of time reversibility for phylogeny reconstruction is that maximum likelihood cannot infer the position of the root of the tree unless additional information information exists (e.g., which taxa are the outgroups) or additional assumptions are made (e.g., a molecular clock).

Q will represent matrix of instantaneous rates of change. For general time reversible model, entries of Q are:

		To		
From	A	C	G	T
A	$-(a\pi_C + b\pi_G + c\pi_T)$	$a\pi_c$	$b\pi_G$	$c\pi_T$
C	$a\pi_A$	$-(a\pi_A + d\pi_G + e\pi_T)$	$d\pi_G$	$e\pi_T$
G	$b\pi_A$	$d\pi_C$	$-(b\pi_A + d\pi_C + f\pi_T)$	$f\pi_T$
T	$c\pi_A$	$e\pi_C$	$f\pi_G$	$-(c\pi_A + e\pi_C + f\pi_G)$

In above matrix: a, b, c, d, e, and f cannot be negative. With any rate matrix (including above), the transition probabilities $P(t)$ can be determined from the rate matrix Q and the amount of evolution t via

$$P(t) = e^{Qt} = I + \frac{(Qt)}{1!} + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \ldots,$$

where I is the identity matrix.

Computing $p_{ij}(t)$ for the Jukes-Cantor model

The Jukes–Cantor model assumes that this is how nucleotide substitution occurs:

0. $\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$.

1. For each site in the sequence, an "event" will occur with probability $\frac{4}{3}s$ per unit evolutionary time.

2. If no event occurs, the residue at the site does not change.

3. If an event occurs, the probability that a residue is type i after the event is π_i.

What is the probability that no event occurs in t units of evolutionary time?

$$(1-\frac{4}{3}s) \times (1-\frac{4}{3}s) \times (1-\frac{4}{3}s) \ldots (1-\frac{4}{3}s) = (1-\frac{4}{3}s)^t.$$

When $\frac{4}{3}s$ is close to 0,

$$1-\frac{4}{3}s \doteq e^{-\frac{4}{3}s}.$$

$$\Pr(\text{no event}) = (1 - \frac{4}{3}s)^t \doteq e^{-\frac{4}{3}st}.$$

When s is redefined as an instantaneous rate per unit evolutionary time, the approximation becomes an equality:

$$\Pr(\text{no event}) = e^{-\frac{4}{3}st}.$$

$$\Pr(\text{at least one event}) = 1 - \Pr(\text{no event}) = 1 - e^{-\frac{4}{3}st}.$$

If there have been no "events", then the residue cannot possibly have changed after an amount of evolution t.

If there has been at least one event, then the residue is type j with probability π_j.

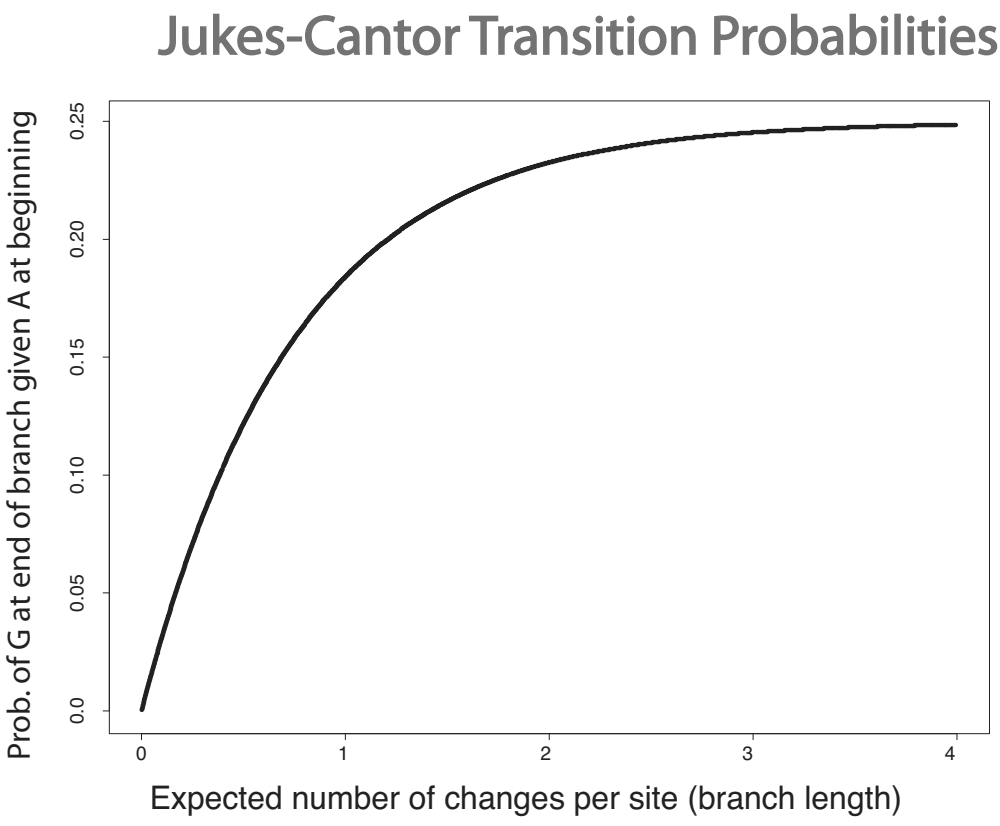
$$\begin{aligned} p_{ii}(t) &= \Pr(\text{no events}) + \Pr(\text{at least one event})\pi_j \\ &= e^{-\frac{4}{3}st} + (1 - e^{-\frac{4}{3}st})\pi_j. \end{aligned}$$

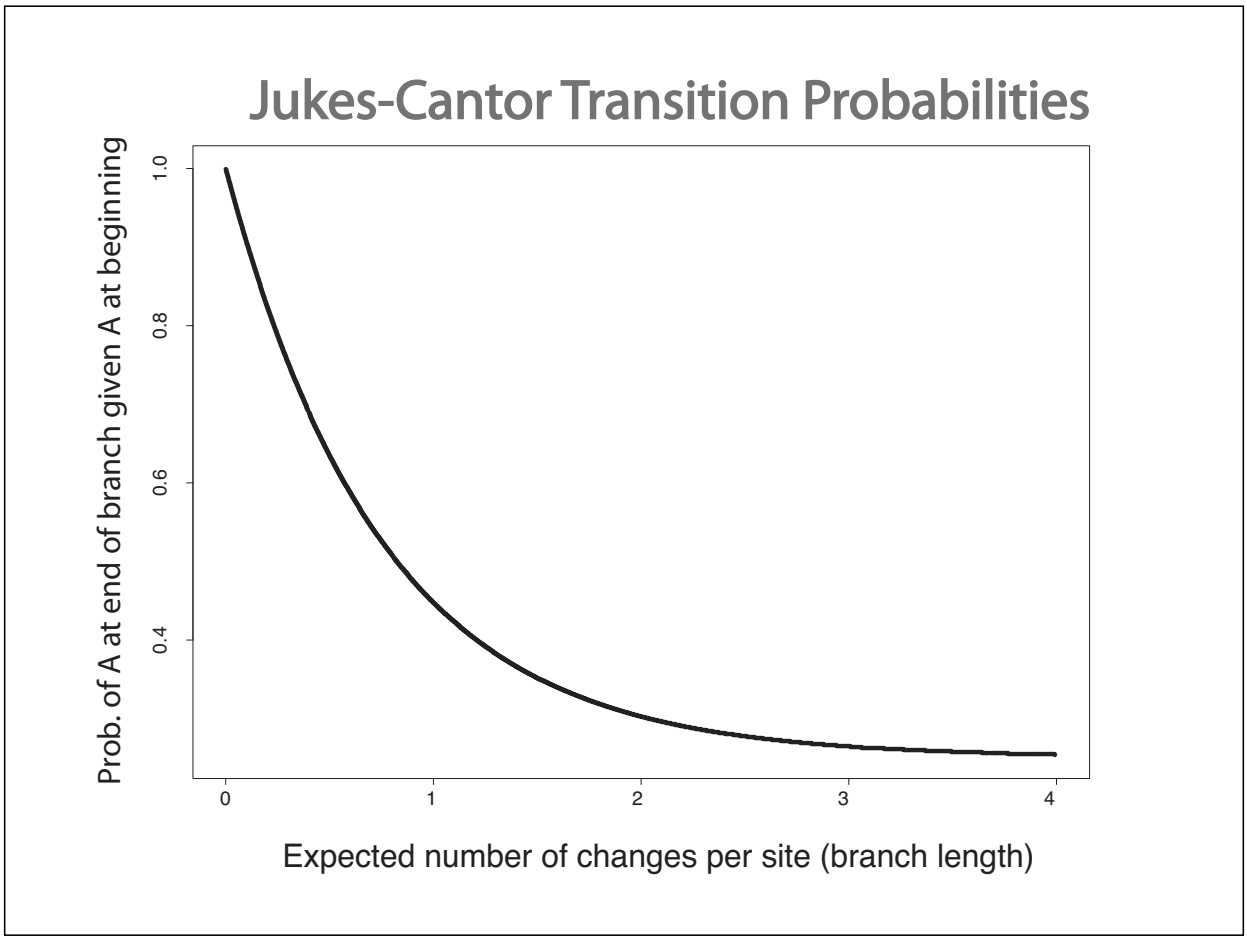
For $i \neq j$,

$$\begin{aligned} p_{ij}(t) &= \Pr(\text{at least one event})\pi_j \\ &= (1 - e^{-\frac{4}{3}st})\pi_j. \end{aligned}$$

Notice that $\frac{4}{3}s$ and t appear only as a product. $\frac{4}{3}s$ and t cannot be separately estimated. Only their product can be estimated.

Note: A generalization of the Jukes–Cantor model, the "Felsenstein 1981" model does not require $\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$.




Jukes-Cantor Transition Probabilities
Prob. of A at end of branch given A at beginning
1.0
0.8
0.6
0.4
0
1
2
3
4
Expected number of changes per site (branch length)