What have we learned so far?

Birds $\mathcal{L}(\beta ; X) = P(X \mid \beta)$

model my choice

Phylogenetic tree

$\mathcal{L}(\varphi ; \text{Sites}, P) = P(\text{Sites} \mid \varphi, P)$

Inferring a phylogeny using the likelihood function

$P \approx$

Model for molecular evolution

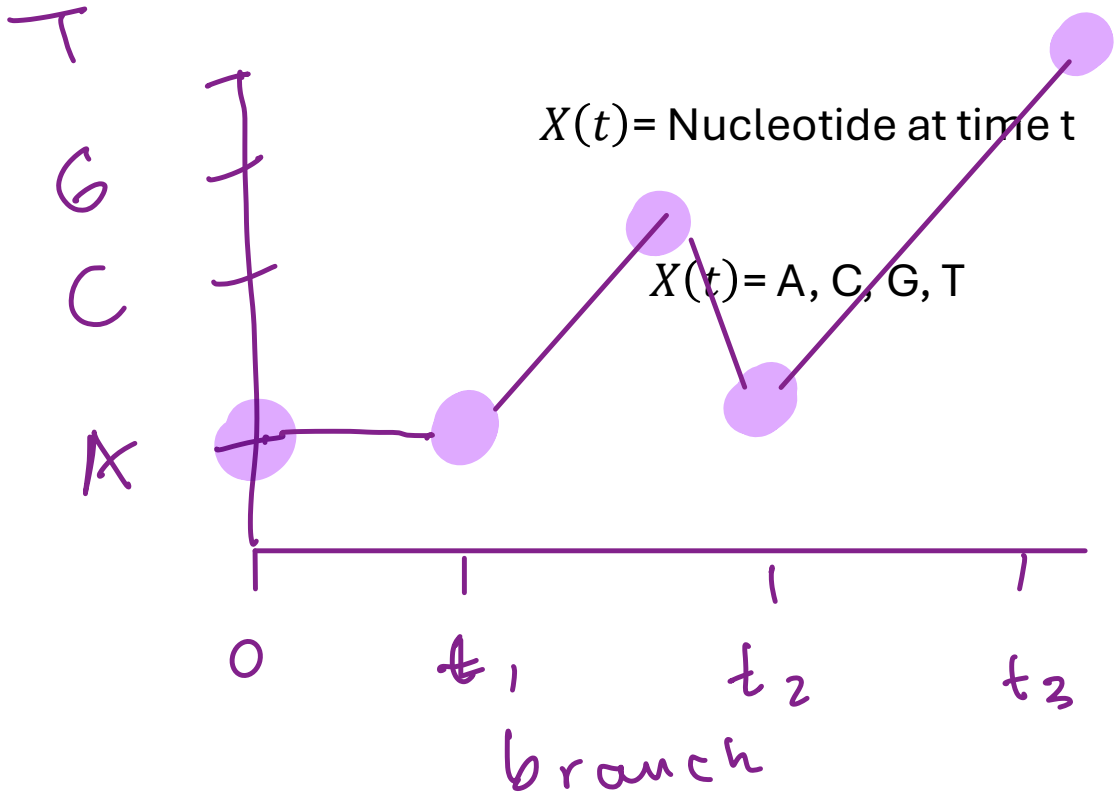
A mechanism to propose branch lengths

A mechanism to propose trees

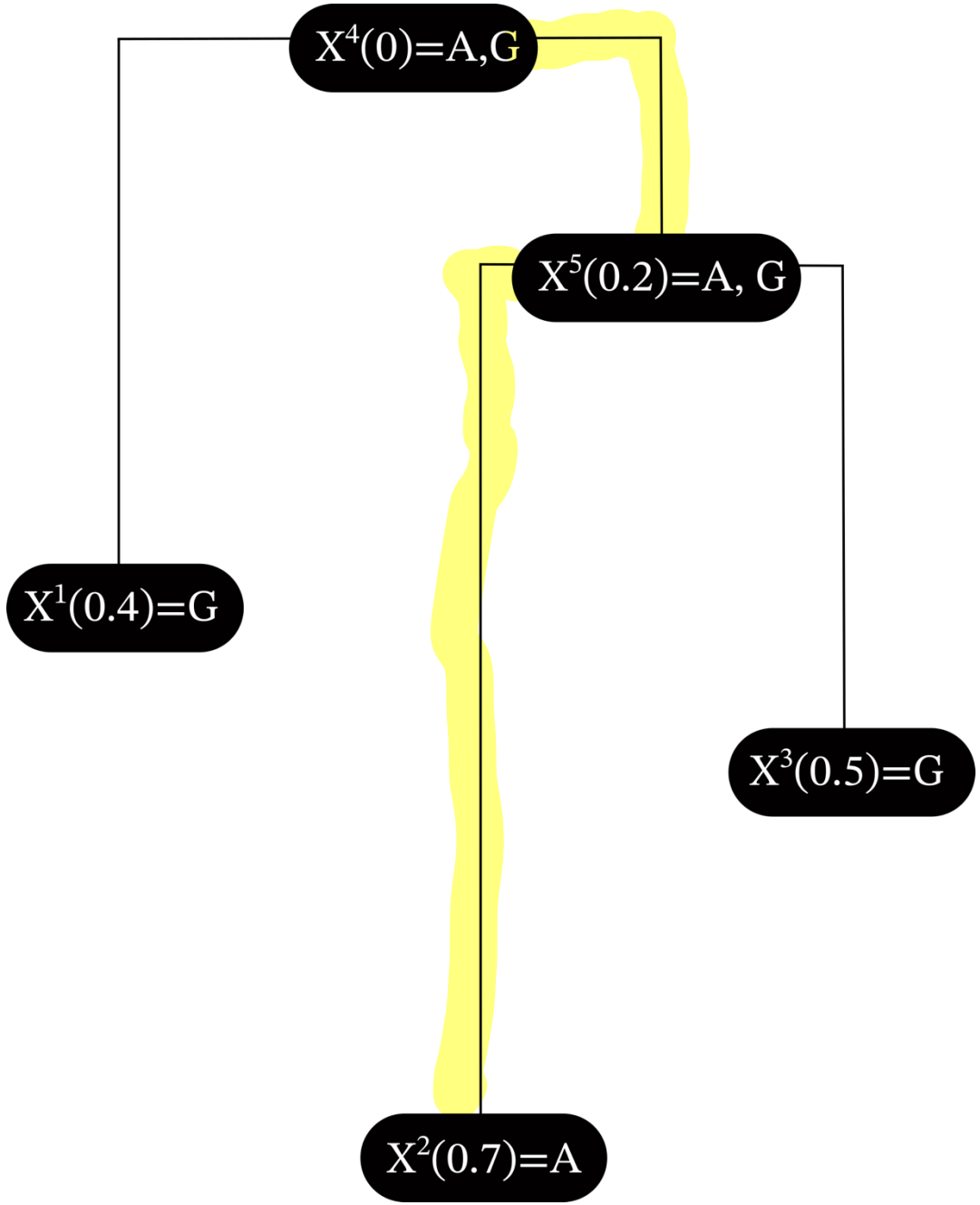
Continuous-Time Markov Chains (CTMC)

$$\{X(t), t \geq 0\}$$

Stochastic models that follow change in time with an associated **probability**



$(X(t), P(t))$



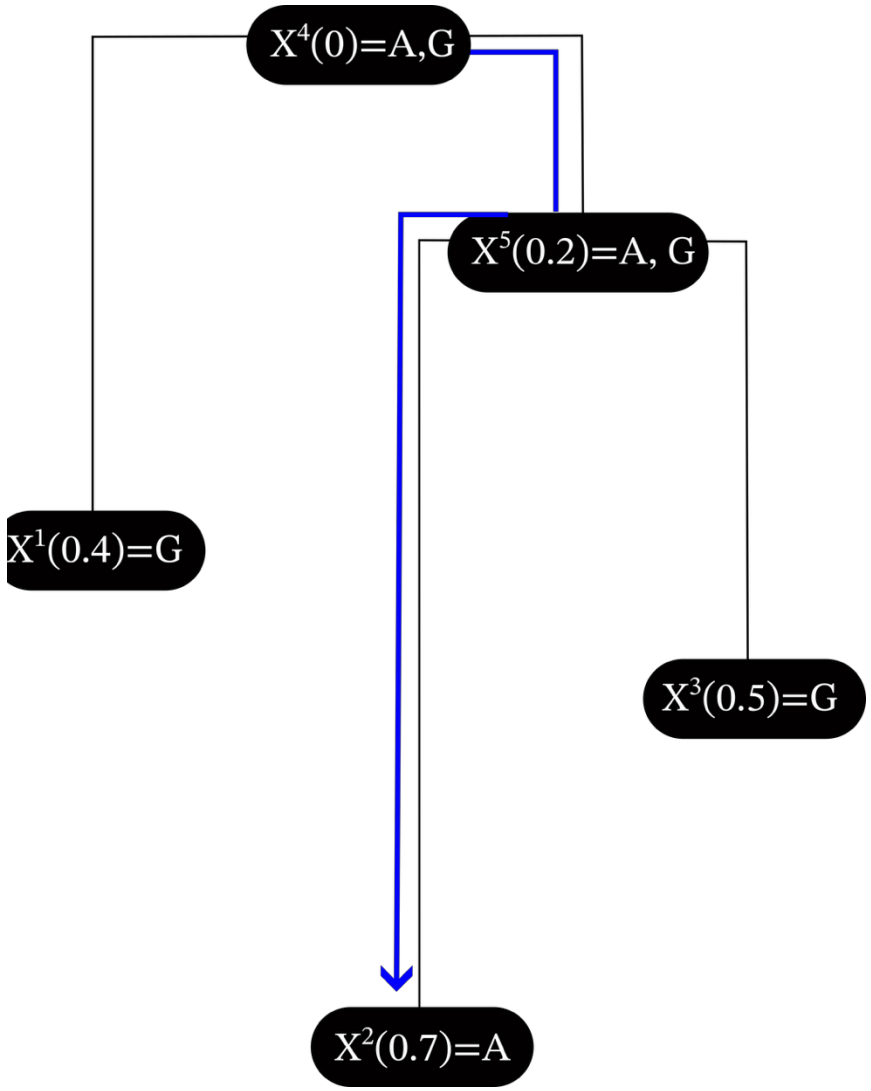
Notation

2 node

$X(0.7) = A$

r.v. that nucleotide over time

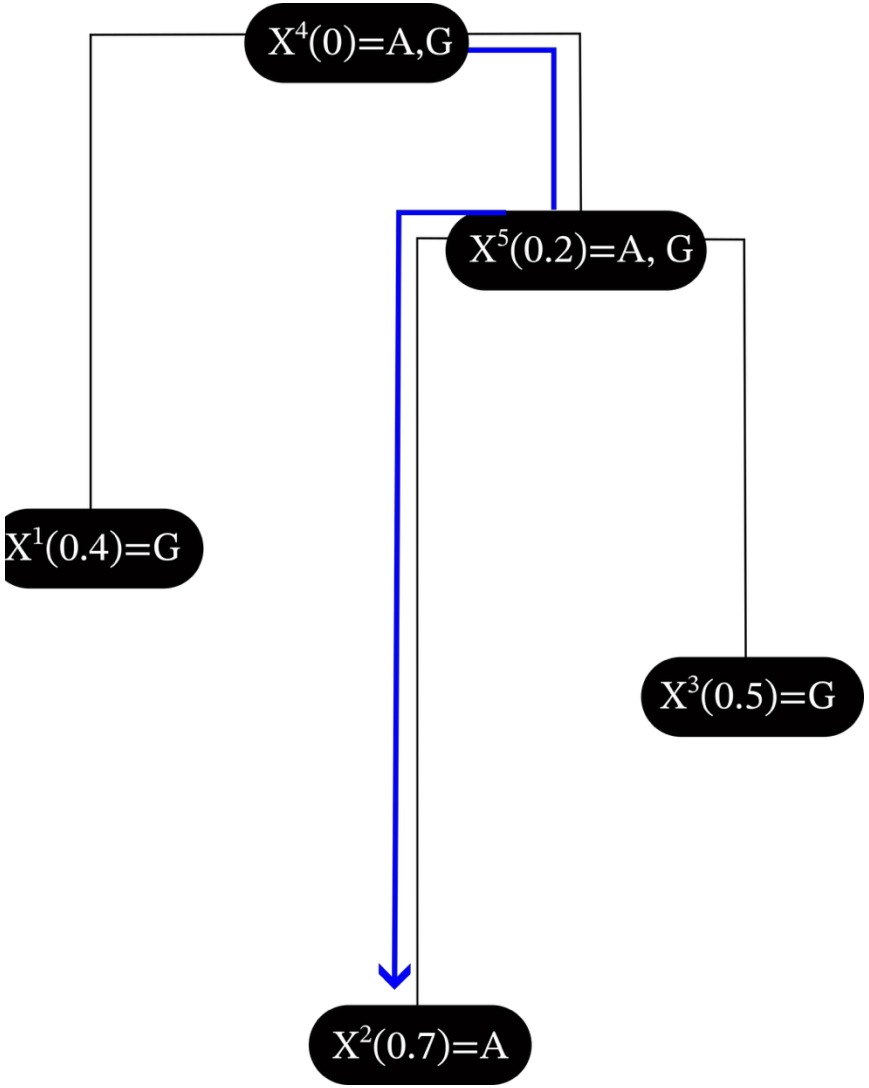
Getting a nucleotide A in Site 1 for taxon 2



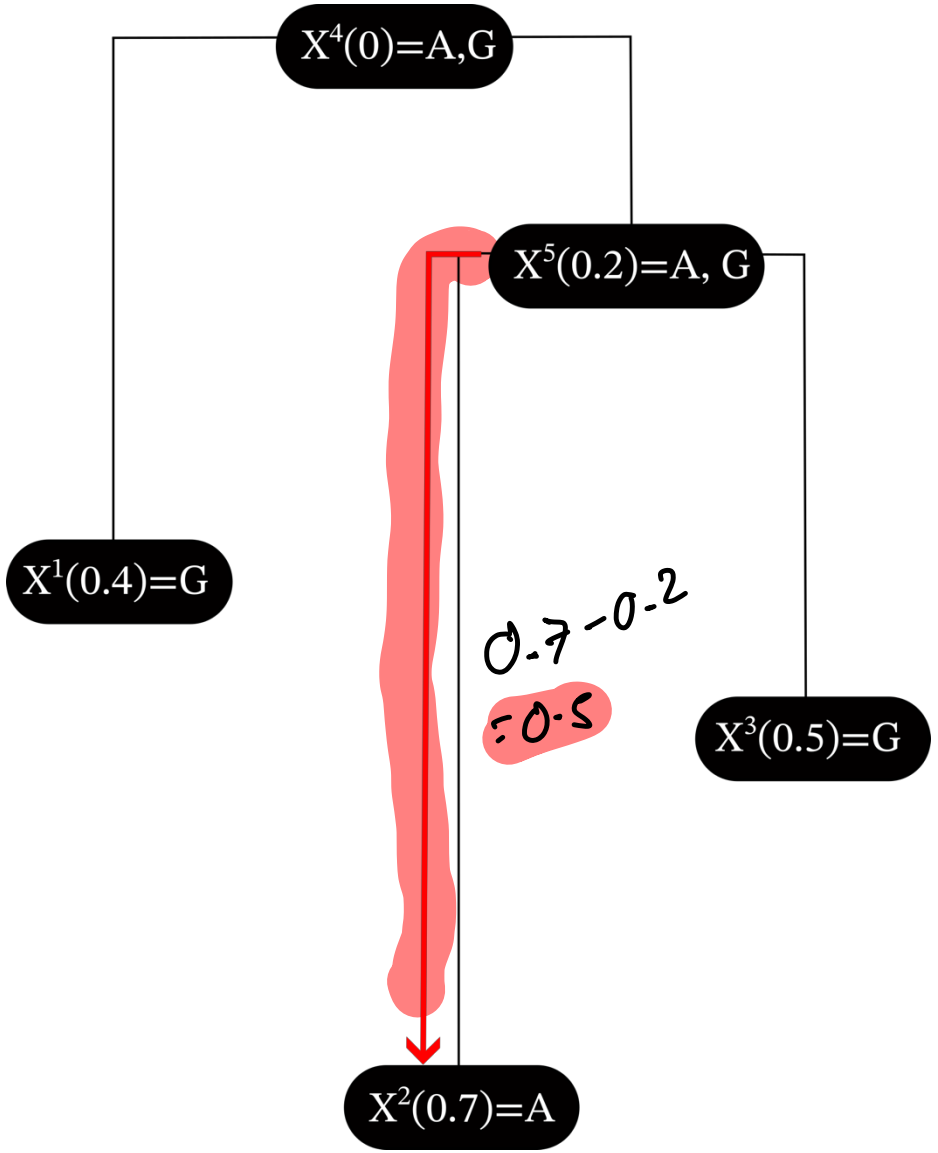
$$P(X^2(0.7) = A \mid X^1(0) = A) \quad ?$$

The Markov property

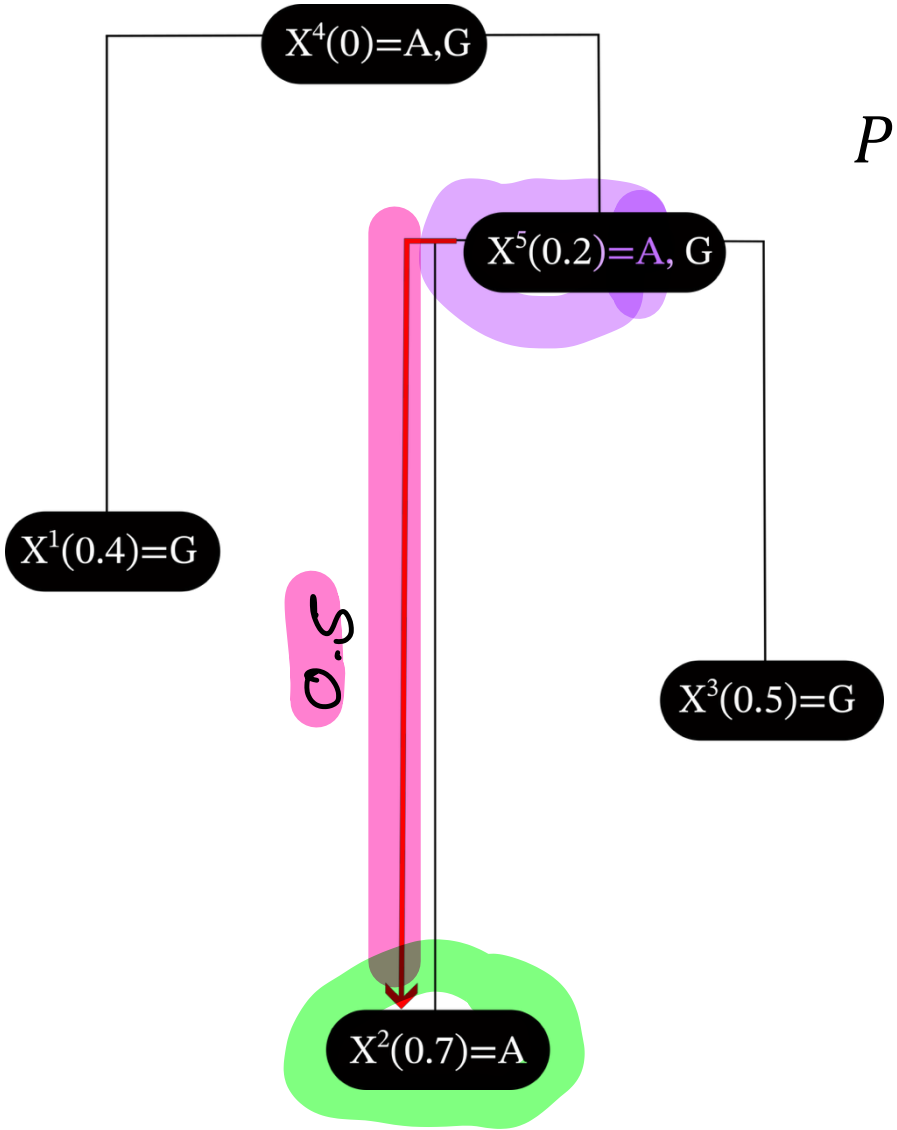
$$P(X^2(0.7)=A \mid X^5(0.2)=A, X^4(0)=A)$$
$$= P[X^2(0.7)=A \mid X^5(0.2)=A)$$



=



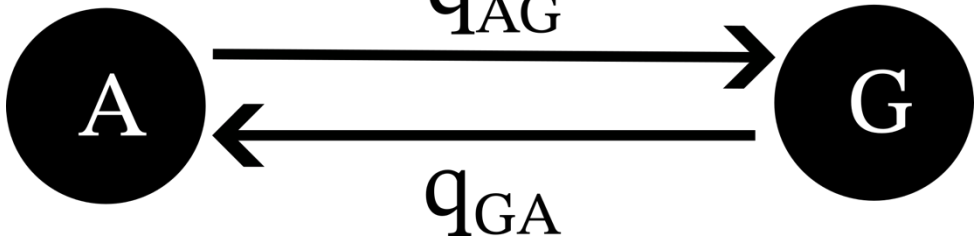
Conditional probabilities



$$P(X^2(0.7)=A|X^5(0.2)=A)=p_{AA}(0.7-0.2)=p_{AA}(0.5))$$

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{AG}(t) \\ p_{GA}(t) & p_{GG}(t) \end{pmatrix}$$

Transition rates define CTMC



rate of nucleotide substitution

$$\lim_{h \to 0} \frac{p_{ij}(h)}{h} = q_{ij}$$

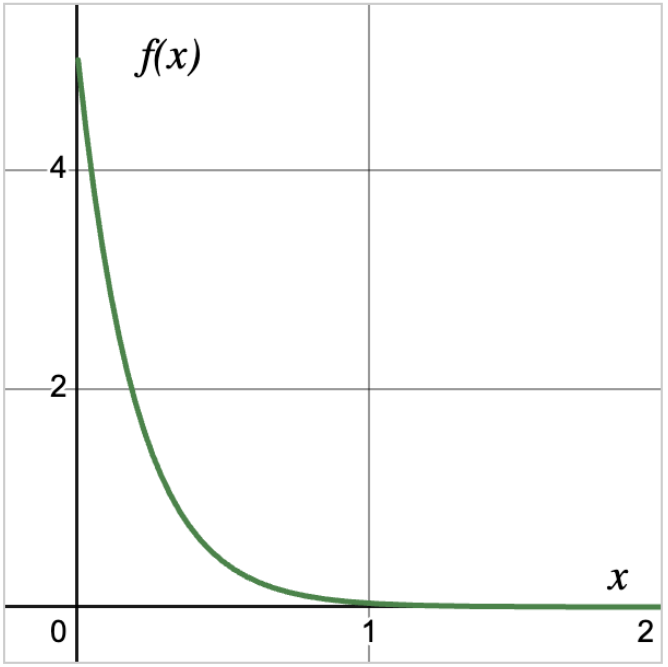
$$P(t) = \begin{pmatrix} p_{AA}(t) + p_{AG}(t) \\ p_{GA}(t) \quad p_{GG}(t) \end{pmatrix} = 1$$

$$Q = \begin{pmatrix} -q_{AG} + q_{AG} \\ q_{GA} \quad -q_{GA} \end{pmatrix} = 0$$

Relationship between P and Q

$$\frac{dP(t)}{dt} = Q \quad \text{or} \quad P(t) = e^{Qt}$$

Waiting times = Exponential distribution



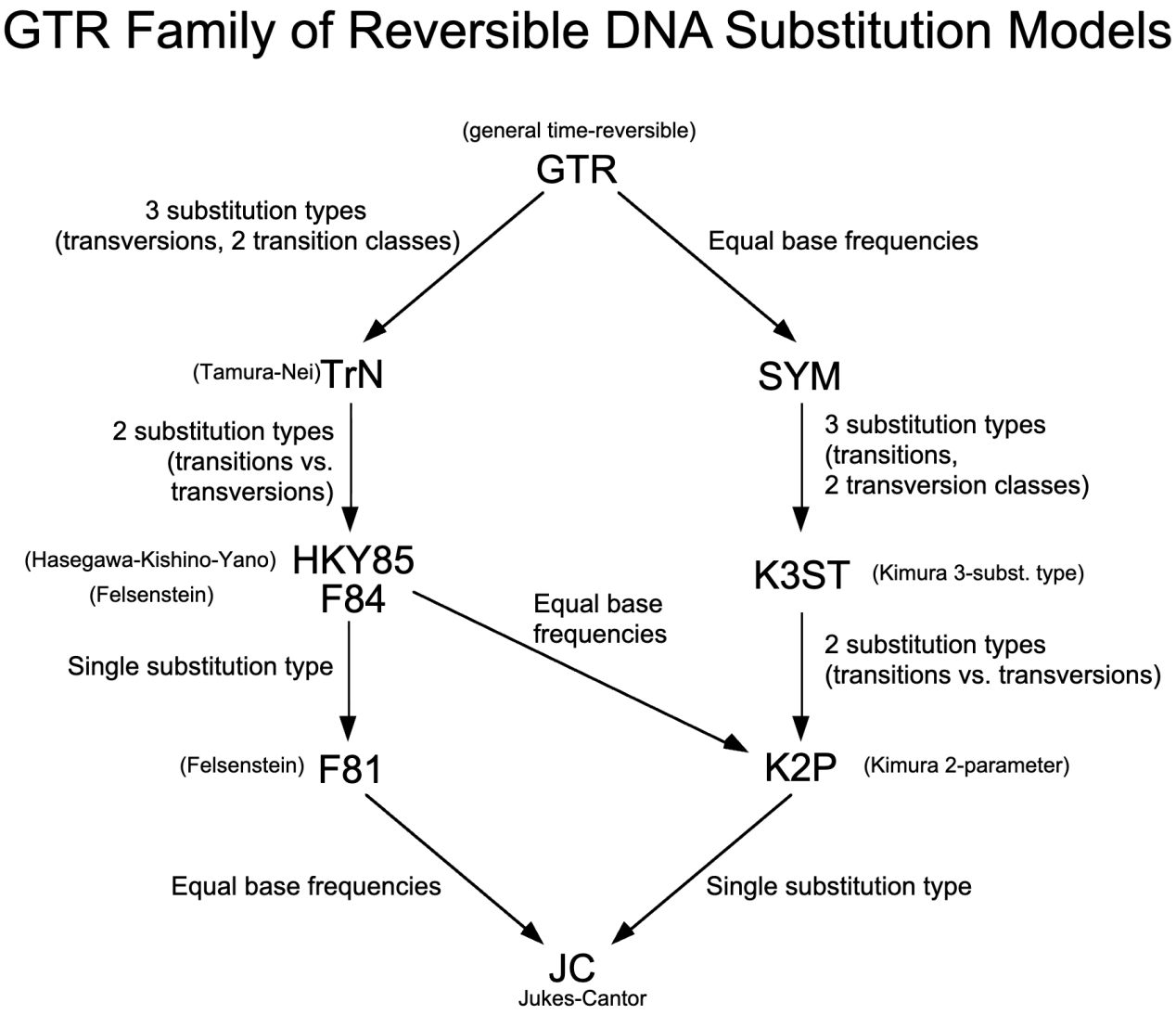
Expected waiting time for G to substitute A is

$$\frac{1}{q_{AG}}$$

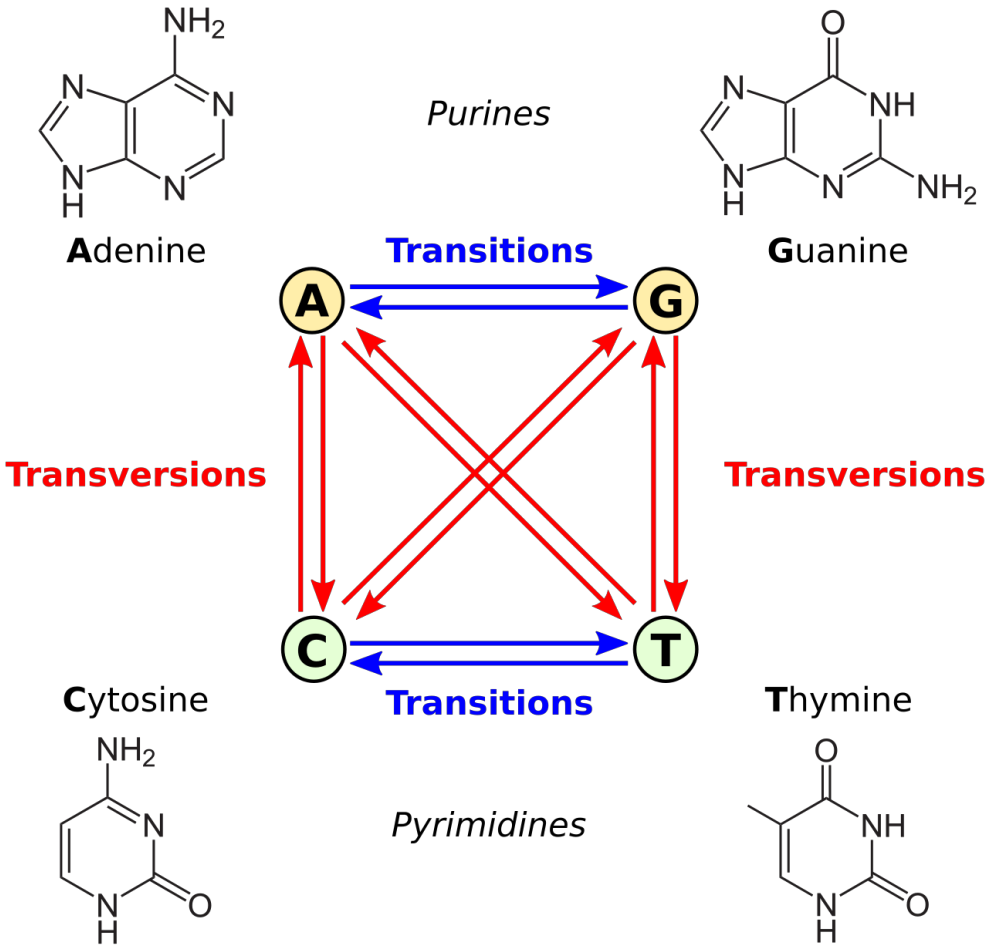
q_{AG} > large

$\Rightarrow \frac{1}{q_{AG}}$ is small

Models for molecular evolution



Substitution rate models



A C G T

$$Q = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{pmatrix} & ? & \\ & & \end{pmatrix}$$

General Time Reversible model (Tavaré, 1986)

$$Q = \begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

a, b, c, d, e, f : are exchangeability parameters

$\pi_A, \pi_C, \pi_G, \pi_T$ = are the equilibrium frequencies

JC (69)- The simplest nucleotide substitution model

Jukes Cantor 1969

$$a = b = c = d = e = f = \mu$$

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

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{pmatrix} -3\mu/4 & \mu/4 & \mu/4 & \mu/4 \\ \mu/4 & -3\mu/4 & \mu/4 & \mu/4 \\ & & & \\ & & & \end{pmatrix} \begin{array}{c} \\ A \\ C \\ G \\ T \end{array}$$

alternative parameterization

$\beta = \mu/4$

$$Q = \begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

Jukes-Cantor (1969)
The simplest nucleotide substitution model model

Transition probabilities (conditional)

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \text{if } i \neq j \end{cases}$$

Same but in matrix shape

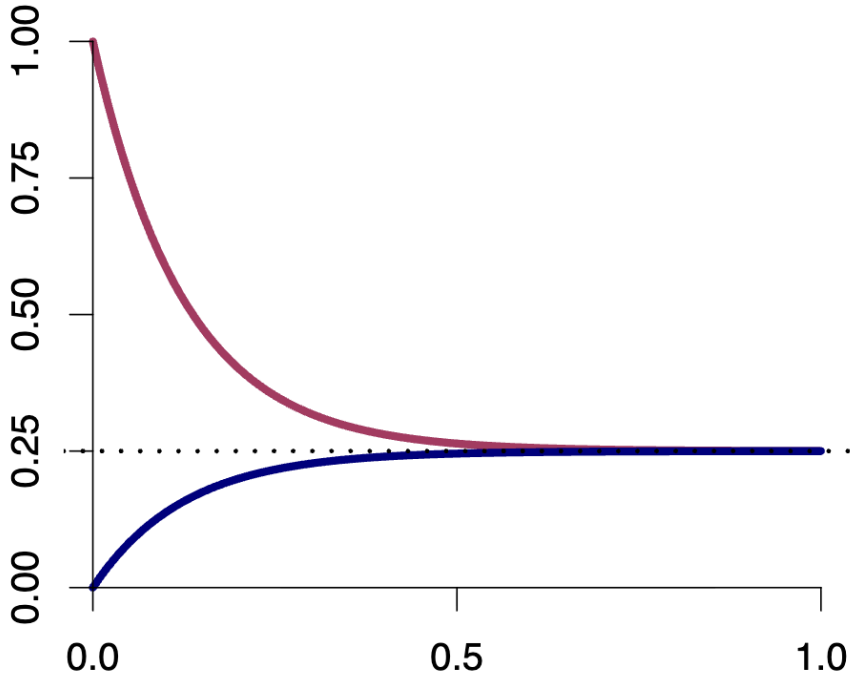
$$P(t) = \begin{bmatrix} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} \\ & & & \end{bmatrix}$$

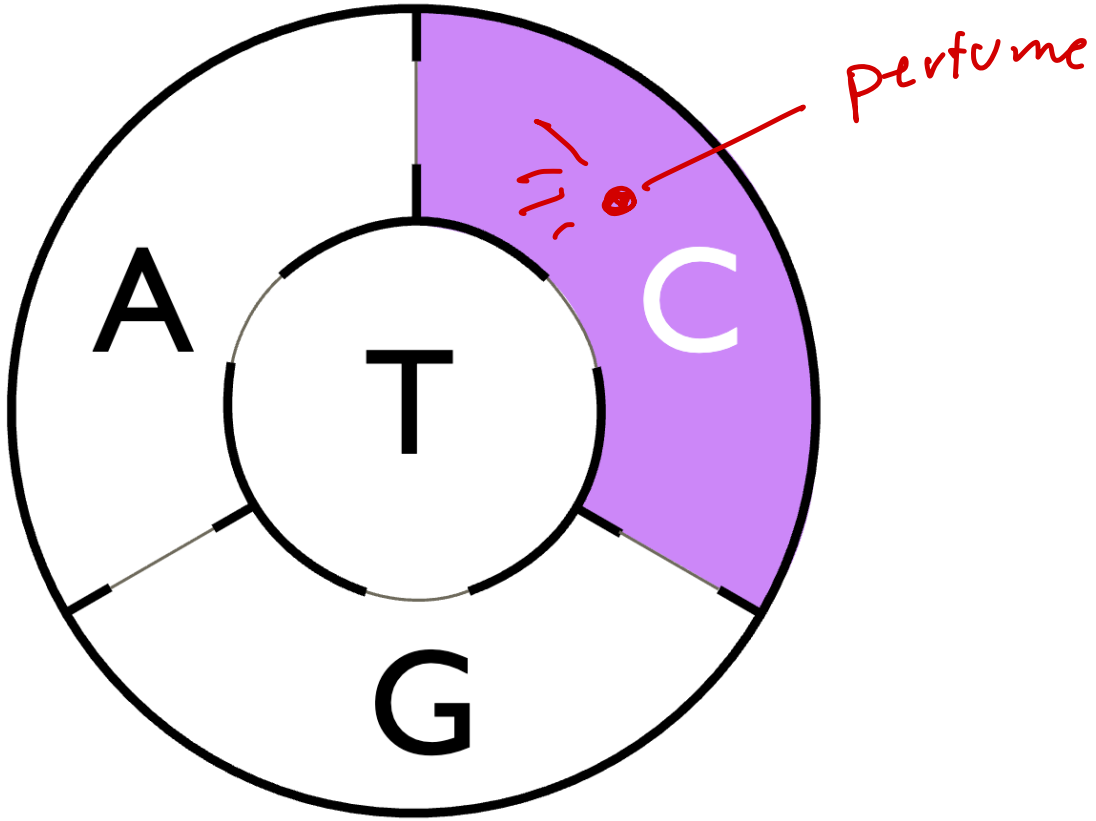
Stationary distribution

$$\pi \times P = \pi$$

$$\pi = \begin{bmatrix} \pi_A \\ \pi_C \\ \pi_G \\ \pi_T \end{bmatrix} \times P(t) = \begin{bmatrix} \pi_A \\ \pi_C \\ \pi_G \\ \pi_T \end{bmatrix}$$

Equilibrium (stationary) probabilities

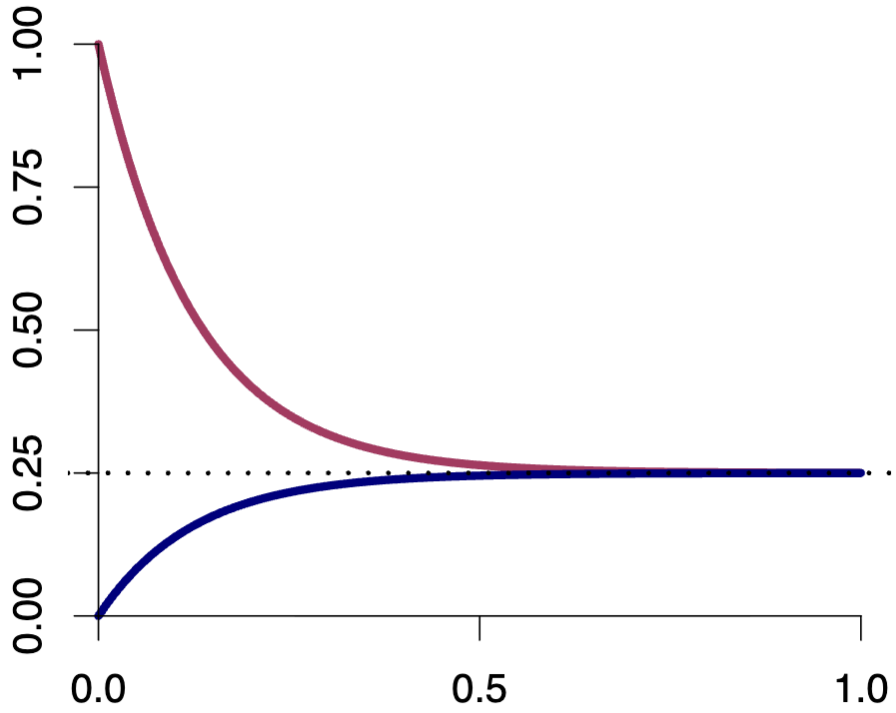


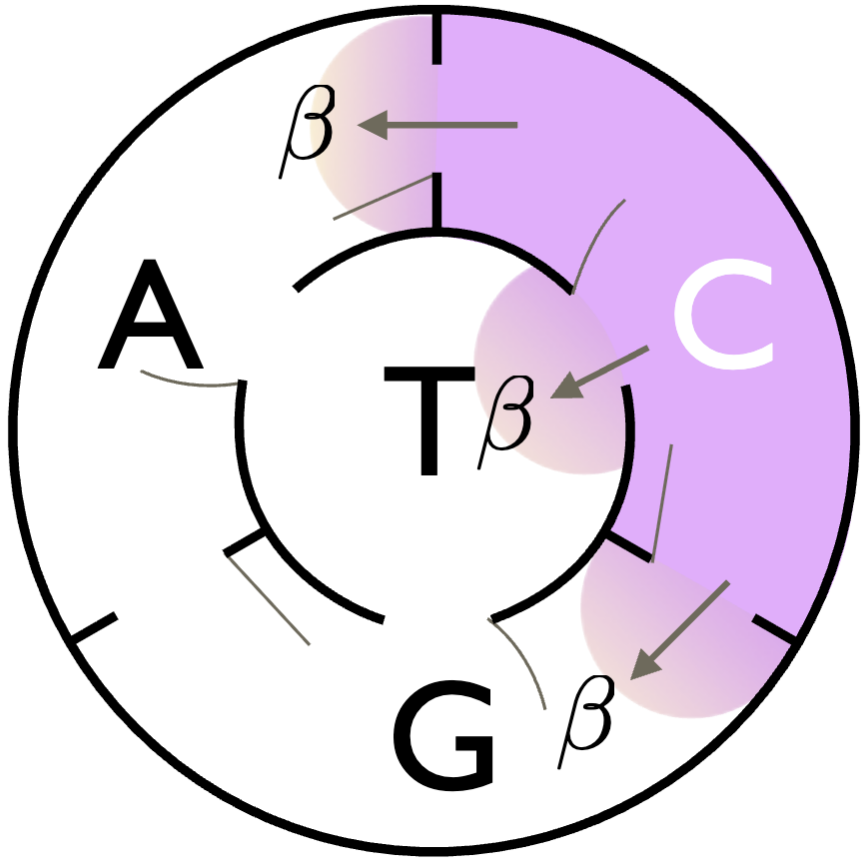


(architect: Joe Bielawski)

Slides from Paul Lewis

Equilibrium (stationary) probabilities

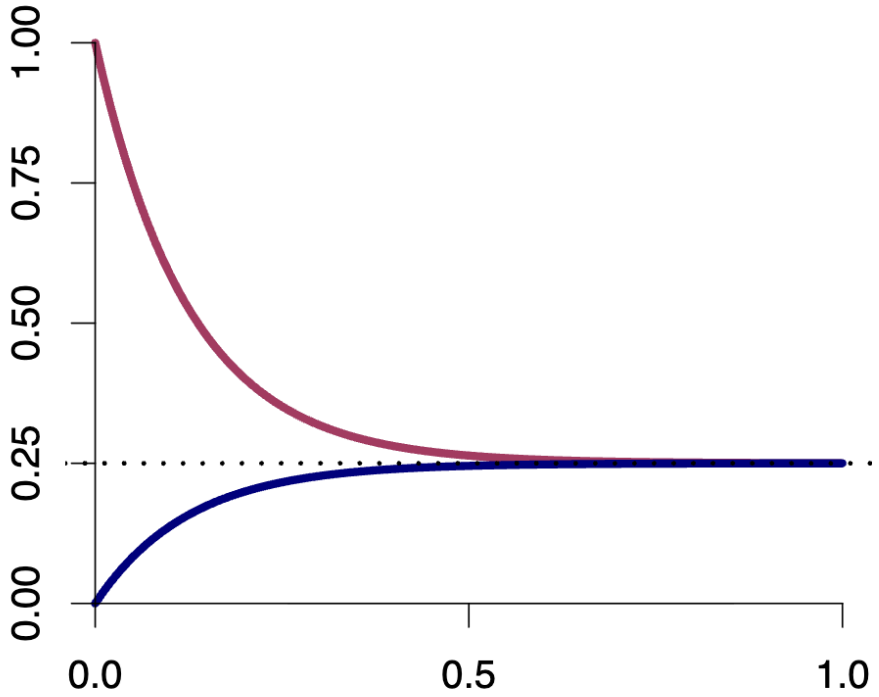


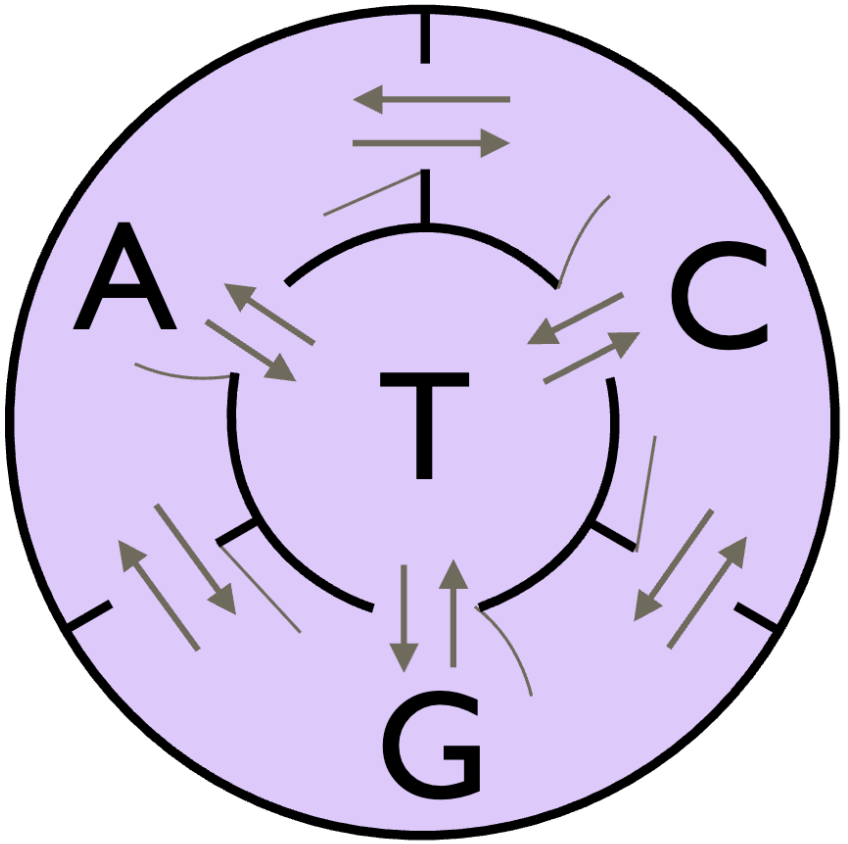


Slides from Paul Lewis

https://plewis.github.io/applets/jc-transition-probabilities/

Equilibrium (stationary) probabilities





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

Slides from Paul Lewis

Branches represent expected number of substitutions

$$Q = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}$$

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \text{if } i \neq j \end{cases}$$

Branch length $v = \mu t$

$v = 10$ $\begin{cases} \mu = 5 \\ t = 2 \end{cases}$ or $\begin{cases} \mu = 2 \\ t = 5 \end{cases}$

Exercise: Building new models of substitution rates.
After you write the model tell me, what do you think the rates represent?

$$Q = \begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

Group 1

$a = b = c = d = e = f = \boldsymbol{\mu}$

(you can make $\mu = 1$)

$(\pi_A, \pi_C, \pi_G, \pi_T)$ all free

Group 2

$a = c = d = f = \boldsymbol{\mu}$

$b = e = \mu\kappa$

(you can make $\mu = 1$)

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

Group 3

$a = c = d = f = \boldsymbol{\mu}$

$b = e = \mu\kappa$

(you can make $\mu = 1$)

$(\pi_A, \pi_C, \pi_G, \pi_T)$ all free

Group 4

$a = c = d = f = \boldsymbol{\mu}$

$b = \mu\kappa_1$

$\mathrm{e} = \mu\kappa_2$

$(\pi_A, \pi_C, \pi_G, \pi_T)$ all free

Exercise: Writing your own model of molecular evolution

Group 1

$F81$

Felsenstein 1981

Group 2

K2P

Kimura 1980

Group 3

HKY85

Hasegawa et al. 1985

Group 4

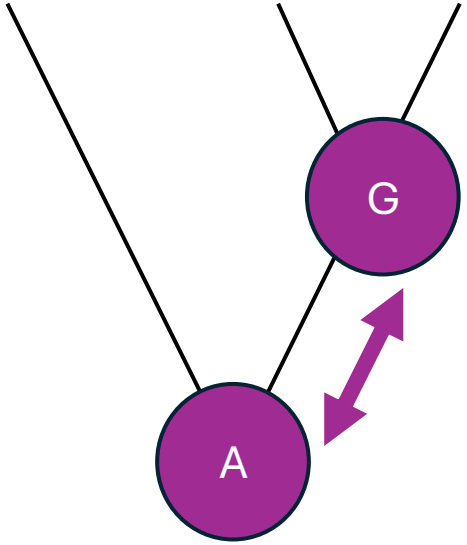
TN93

Tamura and Nei (1993)

A note on the reversible adjective

The reversibility property

$$\pi_A p_{AG}(t) = \pi_G p_{GA}(t)$$



- Any point of the tree can serve as the root

- Time reversibility is not necessary for substitution
- models

$$r_1 P_1(t) + r_2 P_2(t) + \cdots + r_k P_{k(\mathrm{t})}$$

- Invariant sites (I)
- Gamma (G)

Analogy from Paul Lewis about a model with varying substitution rates (illustrated by ChatGPT)

You buy a bunch of different shirt sizes (substitution rates) for a group of different women (sites).

It will be more expensive than buying a one-size fits all. But women will be happier with their own size rather than one-size fits all.