# Substitution models

Askar Gafurov

November 9, 2023

# Modelling the evolution of genomes

- The ultimate goal: to model the evolutionary distance between two genomes
  - Input: sequences $S_1, S_2 \in \{A, C, G, T\}^* = \Sigma^*$, evolutionary time $t$
  - Output: $\Pr[S_1 \xrightarrow{t} S_2]$ (formal way to denote: $\Pr[S_2 \mid S_1, t]$)
    - ⋆ *Probability of sequence $S_1$ to mutate into sequence $S_2$ in evolutionary time $t$*
    - ⋆ Formally: *Probability of observing sequence $S_2$, given that its evolutionary ancestor in time $t$ is sequence $S_1$*
- Requirements:
  - $\Pr[S \xrightarrow{t=0} S] = 1$ (no evolution in zero time)
  - $\forall S' \in \Sigma^* : \Pr[S' \xrightarrow{t=\infty} S] = \pi_S$ (with enough time, the starting point is irrelevant)
  - $\Pr[S_1 \xrightarrow{t_1} S_2 \wedge S_2 \xrightarrow{t_2} S_3] = \Pr[S_1 \xrightarrow{t_1} S_2] \cdot \Pr[S_2 \xrightarrow{t_2} S_3]$ (no memory)
  - $\Pr[S_1 \xrightarrow{t=t_1+t_2} S_3] = \sum_{S_2 \in \Sigma^*} \Pr[S_1 \xrightarrow{t_1} S_2] \cdot \Pr[S_2 \xrightarrow{t_2} S_3]$ (multiplicativity)
    - ⋆ we can break time $t$ into two parts $t_1$ and $t_2$, and sum over all possible intermediate states

# What can we do with such a model (in the near future)

- Given a phylogenetic tree (phylogeny) $T = (\mathbf{S} \subset \Sigma^*, E \subset \mathbf{S}^2, t : E \to \mathbf{R})$ of sequences $\mathbf{S}$ with times $\mathbf{t}(\cdot, \cdot)$ on the edges, we can compute its total probability by multiplying probabilities of each edge:

$$\Pr[\mathbf{S} \mid E, \mathbf{t}] = \Pr[S_{\text{root}}] \cdot \prod_{e:(S_a,S_s)\in E} \Pr[S_a \overset{\mathbf{t}(S_a,S_s)}{\to} S_s]$$

- This allows us to compute the likelihood $\mathcal{L}(E, \mathbf{t}; \mathbf{S})$ of a potential phylogeny $T$ structure $E$ and times $\mathbf{t}$ w.r.t. sequences $\mathbf{S}$ in the nodes
  - We can choose the best phylogeny structure by maximizing the total likelihood
- We can even maximize the likelihood using only sequences **in the leaves** (present species) by using the Felsenstein algorithm (*next week*)

# Simplifying assumptions

- No indels, only substitutions
  - $\implies |S_1| = |S_2| = n$
- All bases mutate independently
  - Compute mutation prob. for each base, and then multiply:

$$\begin{aligned}\Pr[S_1 = (a_1, \ldots, a_n) \xrightarrow{t} S_2 = (b_1, \ldots, b_n)] = \\ = \Pr[a_1 \xrightarrow{t} b_1] \cdot \Pr[a_2 \xrightarrow{t} b_2] \cdot \ldots \cdot \Pr[a_n \xrightarrow{t} b_n] = \\ = \prod_{i=1}^{n} \Pr[a_i \xrightarrow{t} b_i].\end{aligned}$$

  - Now, we only need to model the **evolution of a single base** $\Pr[a \xrightarrow{t} b]$

# Substitution model for one base

- $\Pr[a \xrightarrow{t} b]$ for a fixed time $t$ has only 16 possible input combinations $\{A, C, G, T\}^2$
- Written as a matrix: $S(t) = \begin{pmatrix} \Pr[A \xrightarrow{t} A] & \Pr[A \xrightarrow{t} C] & \Pr[A \xrightarrow{t} G] & \Pr[A \xrightarrow{t} T] \\ \Pr[C \xrightarrow{t} A] & \Pr[C \xrightarrow{t} C] & \Pr[C \xrightarrow{t} G] & \Pr[C \xrightarrow{t} T] \\ \Pr[G \xrightarrow{t} A] & \Pr[G \xrightarrow{t} C] & \Pr[G \xrightarrow{t} G] & \Pr[G \xrightarrow{t} T] \\ \Pr[T \xrightarrow{t} A] & \Pr[T \xrightarrow{t} C] & \Pr[T \xrightarrow{t} G] & \Pr[T \xrightarrow{t} T] \end{pmatrix}$
- General properties of matrix $S(t)$:
  - $\Pr[C \xrightarrow{t} G] =$

# Substitution model for one base

- $\Pr[a \xrightarrow{t} b]$ for a fixed time $t$ has only 16 possible input combinations $\{A, C, G, T\}^2$
- Written as a matrix: $S(t) = \begin{pmatrix} \Pr[A \xrightarrow{t} A] & \Pr[A \xrightarrow{t} C] & \Pr[A \xrightarrow{t} G] & \Pr[A \xrightarrow{t} T] \\ \Pr[C \xrightarrow{t} A] & \Pr[C \xrightarrow{t} C] & \Pr[C \xrightarrow{t} G] & \Pr[C \xrightarrow{t} T] \\ \Pr[G \xrightarrow{t} A] & \Pr[G \xrightarrow{t} C] & \Pr[G \xrightarrow{t} G] & \Pr[G \xrightarrow{t} T] \\ \Pr[T \xrightarrow{t} A] & \Pr[T \xrightarrow{t} C] & \Pr[T \xrightarrow{t} G] & \Pr[T \xrightarrow{t} T] \end{pmatrix}$
- General properties of matrix $S(t)$:
  - $\Pr[C \xrightarrow{t} G] = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \cdot S(t) \cdot \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}^T$
  - $S(0) =$

# Substitution model for one base

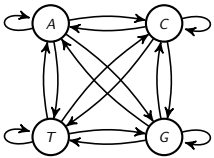
- $\Pr[a \xrightarrow{t} b]$ for a fixed time $t$ has only 16 possible input combinations $\{A, C, G, T\}^2$
- Written as a matrix: $S(t) = \begin{pmatrix} \Pr[A \xrightarrow{t} A] & \Pr[A \xrightarrow{t} C] & \Pr[A \xrightarrow{t} G] & \Pr[A \xrightarrow{t} T] \\ \Pr[C \xrightarrow{t} A] & \Pr[C \xrightarrow{t} C] & \Pr[C \xrightarrow{t} G] & \Pr[C \xrightarrow{t} T] \\ \Pr[G \xrightarrow{t} A] & \Pr[G \xrightarrow{t} C] & \Pr[G \xrightarrow{t} G] & \Pr[G \xrightarrow{t} T] \\ \Pr[T \xrightarrow{t} A] & \Pr[T \xrightarrow{t} C] & \Pr[T \xrightarrow{t} G] & \Pr[T \xrightarrow{t} T] \end{pmatrix}$
- General properties of matrix $S(t)$:
  - $\Pr[C \xrightarrow{t} G] = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \cdot S(t) \cdot \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}^T$
  - $S(0) = I_4$
  - $S(t_1) \cdot S(t_2) =$

# Substitution model for one base

- $\Pr[a \xrightarrow{t} b]$ for a fixed time $t$ has only 16 possible input combinations $\{A, C, G, T\}^2$
- Written as a matrix: $S(t) = \begin{pmatrix} \Pr[A \xrightarrow{t} A] & \Pr[A \xrightarrow{t} C] & \Pr[A \xrightarrow{t} G] & \Pr[A \xrightarrow{t} T] \\ \Pr[C \xrightarrow{t} A] & \Pr[C \xrightarrow{t} C] & \Pr[C \xrightarrow{t} G] & \Pr[C \xrightarrow{t} T] \\ \Pr[G \xrightarrow{t} A] & \Pr[G \xrightarrow{t} C] & \Pr[G \xrightarrow{t} G] & \Pr[G \xrightarrow{t} T] \\ \Pr[T \xrightarrow{t} A] & \Pr[T \xrightarrow{t} C] & \Pr[T \xrightarrow{t} G] & \Pr[T \xrightarrow{t} T] \end{pmatrix}$
- General properties of matrix $S(t)$:
  - $\Pr[C \xrightarrow{t} G] = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \cdot S(t) \cdot \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}^T$
  - $S(0) = I_4$
  - $S(t_1) \cdot S(t_2) = \left( \sum_{x \in \Sigma} \Pr[i \xrightarrow{t_1} x] \cdot \Pr[x \xrightarrow{t_2} j] \right)_{i,j \in \Sigma} \overset{\text{multiplicativity}}{=} \left( \Pr[i \xrightarrow{t_1+t_2} j] \right)_{i,j \in \Sigma} = S(t_1 + t_2)$
    - $S(k \cdot t) = S^k(t)$

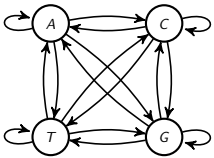
# Model with discrete time

- Assume that evolutionary time $t$ is discrete
  - **at most one mutation** occurs in time 1
- A base now has 4 possible states, and has a chance to transit between them in each time step, or stay the same $\Longrightarrow$ Markov chain
- $S(t) =$



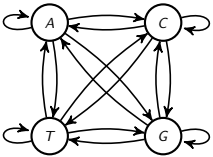
# Model with discrete time

- Assume that evolutionary time $t$ is discrete
  - **at most one mutation** occurs in time 1
- A base now has 4 possible states, and has a chance to transit between them in each time step, or stay the same $\Longrightarrow$ Markov chain
- $S(t) = S^t(1) \Longrightarrow$ only need to define $S(1)$



# Model with discrete time

- Assume that evolutionary time $t$ is discrete
  - **at most one mutation** occurs in time 1
- A base now has 4 possible states, and has a chance to transit between them in each time step, or stay the same $\Longrightarrow$ Markov chain
- $S(t) = S^t(1) \Longrightarrow$ only need to define $S(1)$
- Stationary distribution (equilibrium)




$$S(\infty) = \lim_{t\to\infty} S(t) = \lim_{t\to\infty} S^t(1) = \begin{pmatrix} \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_A & \pi_C & \pi_G & \pi_T \end{pmatrix}$$

# Quick summary so far

- Evolution model = prob. $\Pr[S_1 \xrightarrow{t} S_2] = \Pr[S_2 \mid S_1, t]$ of observing $S_2$ given that its ancestor in evolutionary time $t$ is $S_1$
- Assuming only substitutions
  - $|S_1| = |S_2| = n$
- Assuming independent evolution for each base
  - $\Pr[S_1 = (a_1, \ldots, a_n) \xrightarrow{t} S_2 = (b_1, \ldots, b_n)] = \prod_{i=1}^{n} \Pr[a_i \xrightarrow{t} b_i]$
  - Only need to define a (substitution) model for a single base
  - $\Pr[a \xrightarrow{t} b] = S(t) = \begin{pmatrix} \Pr[A \xrightarrow{t} A] & \Pr[A \xrightarrow{t} C] & \Pr[A \xrightarrow{t} G] & \Pr[A \xrightarrow{t} T] \\ \Pr[C \xrightarrow{t} A] & \Pr[C \xrightarrow{t} C] & \Pr[C \xrightarrow{t} G] & \Pr[C \xrightarrow{t} T] \\ \Pr[G \xrightarrow{t} A] & \Pr[G \xrightarrow{t} C] & \Pr[G \xrightarrow{t} G] & \Pr[G \xrightarrow{t} T] \\ \Pr[T \xrightarrow{t} A] & \Pr[T \xrightarrow{t} C] & \Pr[T \xrightarrow{t} G] & \Pr[T \xrightarrow{t} T] \end{pmatrix}$
  - $S(t_1 + t_2) = S(t_1) \cdot S(t_2)$
- For discrete time, only need to define $S(1)$
  - Classic Markov chain with states $\{A, C, G, T\}$, $S(1)$ = matrix of transition probabilities

# Jukes-Cantor JC69 model

- The plan: define Markov chains with continuous time (CTMC), where all substitutions are equally likely
  - $S(t) =$

# Jukes-Cantor JC69 model

- The plan: define Markov chains with continuous time (CTMC), where all substitutions are equally likely
  - $S(t) = \begin{pmatrix} 1-3s(t) & s(t) & s(t) & s(t) \\ s(t) & 1-3s(t) & s(t) & s(t) \\ s(t) & s(t) & 1-3s(t) & s(t) \\ s(t) & s(t) & s(t) & 1-3s(t) \end{pmatrix} = I + \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix} \cdot s(t)$

# Jukes-Cantor JC69 model

- The plan: define Markov chains with continuous time (CTMC), where all substitutions are equally likely
  - $S(t) = \begin{pmatrix} 1-3s(t) & s(t) & s(t) & s(t) \\ s(t) & 1-3s(t) & s(t) & s(t) \\ s(t) & s(t) & 1-3s(t) & s(t) \\ s(t) & s(t) & s(t) & 1-3s(t) \end{pmatrix} = I + \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix} \cdot s(t)$
- Let's look at $s(t)$ closely
  - $s(0) = 0$
  - Let's denote the first derivative of $s(t)$ at zero as $\alpha$:
    - ★ Formally, $\alpha := s'(0) \stackrel{\text{def.}}{=} \lim\limits_{\varepsilon \to 0} \dfrac{s(0+\varepsilon) - s(0)}{\varepsilon} = \lim\limits_{\varepsilon \to 0} \dfrac{s(\varepsilon)}{\varepsilon}$
    - ★ $\alpha = \left. \dfrac{\partial \Pr[a \xrightarrow{t} b]}{\partial t} \right|_{t=0}$

# Derivative of $S(t)$

$$S'(t) \stackrel{\text{def.}}{=} \lim_{\varepsilon \to 0} \frac{S(t+\varepsilon) - S(t)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{S(t)S(\varepsilon) - S(t)}{\varepsilon} =$$

$$= \lim_{\varepsilon \to 0} \frac{S(t)(S(\varepsilon) - I)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{S(t) \cdot \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix} \cdot s(\varepsilon)}{\varepsilon} =$$

$$= S(t) \cdot \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix} \cdot \lim_{\varepsilon \to 0} \frac{s(\varepsilon)}{\varepsilon} =$$

$$= S(t) \cdot \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

# Differential equation

- We've got diff. equation $S'(t) = S(t) \cdot R$, where $R = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$
- $R$ is called transition rate matrix
- It is really a system of 16 ordinary differential equations $S'(t)_{a,b} = (S(t) \cdot R)_{a,b}$
  - for $(A, A)$: $-3s'(t) = (1 - 3s(t))(-3\alpha) + 3s(t)\alpha = -3\alpha + 12\alpha s(t)$
    - ⋆ $s'(t) = \alpha - 4\alpha s(t)$
  - for $(A, C)$: $s'(t) = (1 - 3s(t))\alpha + s(t)(-3\alpha) + 2s(t)\alpha = \alpha - 4\alpha s(t)$
  - which reduces to a single ordinary differential equation $s'(t) = \alpha - 4\alpha s(t)$ with start condition $s(0) = 0$
- Solution: $s(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}; \quad 1 - 3s(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$

$$\frac{ds}{dt} = \alpha - 4\alpha s$$

$$\frac{ds}{\alpha - 4\alpha s} = dt$$

$$\frac{1}{\alpha}\int \frac{ds}{1-4s} = \int 1dt$$

$$|(1-4s) = x, -4ds = dx|$$

$$\frac{1}{-4\alpha}\int \frac{dx}{x} = \int 1dt$$

$$\frac{1}{-4\alpha}\ln(1-4s) = t + C$$

$$1 - 4s = e^{-4\alpha t + C}$$

$$s = \frac{1 - e^{-4\alpha t + C}}{4}$$

$$s(0) = 0 \Longrightarrow \frac{1-e^{C}}{4} = 0 \Longrightarrow C = 0$$

Solution: $s(t) = \frac{1-e^{-4\alpha t}}{4}; 1 - 3s(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$

PARENTAL ADVISORY HARD MATH

# Equilibrium for Jukes-Cantor model

$$\lim_{t\to\infty} \Pr[A \xrightarrow{t} A] = \lim_{t\to\infty} \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} = \frac{1}{4}$$
$$\lim_{t\to\infty} \Pr[A \xrightarrow{t} C] = \lim_{t\to\infty} \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} = \frac{1}{4}$$

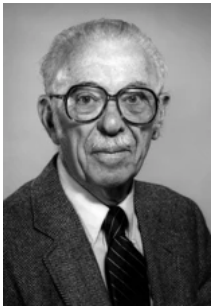
$$S(\infty) = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

# Quick summary so far

- Jukes-Cantor substitution model:
  - Continuous time $t$
  - Equal probability of substitution $\forall a \neq b : \Pr[a \xrightarrow{t} b] = s(t)$
  - Matrix form

$$S(t) = \begin{pmatrix} 1-3s(t) & s(t) & s(t) & s(t) \\ s(t) & 1-3s(t) & s(t) & s(t) \\ s(t) & s(t) & 1-3s(t) & s(t) \\ s(t) & s(t) & s(t) & 1-3s(t) \end{pmatrix}$$

- Diff. equation $s'(t) = 1 - 3s(t), s(0) = 0$
- $Pr[a \xrightarrow{t} b] = S_{JC}(t)_{a,b} = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} & a = b \\ \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} & a \neq b \end{cases}$
- Equilibrium for JC: $\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$



# Example for Jukes-Cantor

- Input: $S_1 = TAACCGT$, $S_2 = AATGCGT$, evolutionary time $t = 0.5$, $\alpha = 3$
- Result:

$$\Pr[S_1 \xrightarrow{t} S_2] = \prod_{i=1}^{n} \Pr[a_i \xrightarrow{t} b_i] = \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^{\#(a_i = b_i)} \cdot \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^{\#(a_i \neq b_i)} =$$

$$= \left(\frac{1}{4} + \frac{3}{4}e^{-6}\right)^{4} \cdot \left(\frac{1}{4} - \frac{1}{4}e^{-6}\right)^{3} \approx (0.2519)^4 \cdot (0.2493)^3 \approx 0.0000624$$

# Example for Jukes-Cantor

- Input: $S_1 = TAACCGT$, $S_2 = AATGCGT$, evolutionary time $t = 0.5$, $\alpha = 3$
- Result:

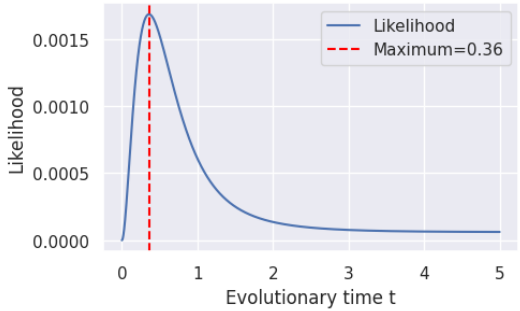
$$\Pr[S_1 \xrightarrow{t} S_2] = \prod_{i=1}^{n} \Pr[a_i \xrightarrow{t} b_i] = \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^{\#(a_i = b_i)} \cdot \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^{\#(a_i \neq b_i)} =$$

$$= \left(\frac{1}{4} + \frac{3}{4}e^{-6}\right)^{4} \cdot \left(\frac{1}{4} - \frac{1}{4}e^{-6}\right)^{3} \approx (0.2519)^4 \cdot (0.2493)^3 \approx 0.0000624$$

- Notice that parameters $t = 30, \alpha = 1/20$ would give the same result
  - Because $t$ and $\alpha$ are always in a product
  - Standard practice is to select $\alpha$ such that $E[\#\text{ mutations in time } t = 1] = 1$
    - $\#$ mutations in time $t = 1 \sim Poisson\,(\lambda = 3\alpha)$, $E = 3\alpha$, $E[\#] = 1$ when $\alpha = 1/3$

# Estimation of evolutionary time in JC model

- Input: $S_1 = TAACCGT$, $S_2 = AATGCGT$, $\alpha = 1/3$ (standard)
- Goal: find the best evolutionary time $t^*$
- Best = with highest likelihood
  - likelihood $\mathcal{L}(t; S_1, S_2, \alpha) = \Pr[S_1 \xrightarrow{t} S_2 \mid \alpha] = \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^{\#(a_i = b_i)} \cdot \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^{\#(a_i \neq b_i)}$.
  - $t^* = \arg\max_{t \geq 0} \mathcal{L}(t; S_1, S_2, \alpha) = -\frac{1}{4\alpha} \ln\left(1 - \frac{4}{3}d\right)$, where $d :=$ proportion of different positions



# Exact estimator of evolutionary time in JC model

$$t^* = \arg\max_{t \geq 0} \mathcal{L}(t; S_1, S_2, \alpha) = \arg\max_{t \geq 0} \log \mathcal{L}(t; S_1, S_2, \alpha) =$$

$$= \arg\max_{t \geq 0} \#(a_i = b_i) \log(1 - 3s(t)) + \#(a_i \neq b_i) \log s(t).$$

$$\frac{df}{ds} = -\frac{3\#(a_i = b_i)}{1 - 3s} + \frac{\#(a_i \neq b_i)}{s} = \frac{(1 - 3s)\#(\neq) - 3s\#(=)}{s(1 - 3s)}.$$

$$\frac{ds}{dt} = \alpha \cdot e^{-4\alpha t}.$$

$$\frac{df}{dt} = 0 \Longrightarrow \frac{df}{ds}\frac{ds}{dt} = 0 \Longrightarrow \frac{df}{ds} = 0 \Longrightarrow \frac{(1 - 3s)\#(\neq) - 3s\#(=)}{s(1 - 3s)} = 0 \Longrightarrow$$

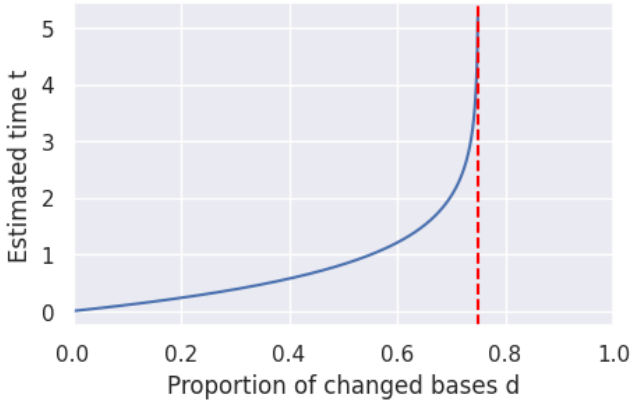
$$\Longrightarrow (1 - 3s)\#(\neq) - 3s\#(=) = 0 \Longrightarrow s = \frac{\#(\neq)}{3 \cdot (\#(\neq) + \#(=))} = \frac{\#(\neq)}{3n}.$$

$$\frac{1}{4} - \frac{1}{4}e^{-4\alpha t} = \frac{\#(\neq)}{3n} \Longrightarrow -4\alpha t = \ln\left(1 - \frac{4\#(\neq)}{3n}\right) \Longrightarrow$$

$$\Longrightarrow t = \frac{-\ln\left(1 - \frac{4}{3}\frac{\#(\neq)}{n}\right)}{4\alpha} = \frac{-\ln\left(1 - \frac{4}{3}d\right)}{4\alpha}.$$

# Behaviour of the time estimator

$$t^* = -\frac{1}{4\alpha} \ln\left(1 - \frac{4}{3} \cdot d\right)$$



# More general models

- JC69 model: rate matrix $R_{JC69} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$
- Sum in a row must equal to 0
- $R_{a,b} := \frac{\partial \Pr[a \xrightarrow{t} b]}{\partial t}$ speed of change from $a$ to $b$
- In general: $R = \begin{pmatrix} * & \mu_{A,C} & \mu_{A,G} & \mu_{A,T} \\ \mu_{C,A} & * & \mu_{C,G} & \mu_{C,T} \\ \mu_{G,A} & \mu_{G,C} & * & \mu_{G,T} \\ \mu_{T,A} & \mu_{T,C} & \mu_{T,G} & * \end{pmatrix}$
  - Diagonal is set to make row sum up to 0
  - Some regularity conditions apply

# Solution to a general model

- The differential equation $S'(t) = S(t) \cdot R$ holds for any rate matrix $R$
- The general solution is $S(t) = e^{Rt}$
- How to compute $e^{Rt}$?
  - diagonalization of matrix $R = Q \cdot \Lambda \cdot Q^{-1}$, where
    - $Q$ = orthogonal matrix (of eigenvectors)
    - $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_4)$ is a diagonal matrix (of eigenvalues)
  - $R^n = (Q \cdot \Lambda \cdot Q^{-1})^n = Q\Lambda Q^{-1} Q\Lambda Q^{-1} Q \ldots Q^{-1} Q\Lambda Q^{-1} = Q\Lambda^n Q^{-1} =$
    $Q \cdot \text{diag}(\lambda_1^n, \ldots, \lambda_4^n) \cdot Q^{-1}$

$$e^{Rt} = \sum_{i=0}^{\infty} \frac{(Rt)^n}{n!} = \sum_{i=0}^{n} \frac{Q \cdot \text{diag}((\lambda_1 t)^n, \ldots, (\lambda_4 t)^n) \cdot Q^{-1}}{n!} =$$

$$= Q \cdot \text{diag}\left( \sum_{i=0}^{\infty} \frac{(\lambda_1 t)^n}{n!}, \ldots, \sum_{i=0}^{\infty} \frac{(\lambda_1 t)^n}{n!} \right) \cdot Q^{-1} = Q \cdot \text{diag}\left( e^{\lambda_1 t}, \ldots, e^{\lambda_4 t} \right) \cdot Q^{-1}$$

# Solution in general form

$$\frac{dS}{dt} = SR \Longrightarrow \int \frac{dS}{S} = \int R dt \Longrightarrow \ln S = Rt + C \Longrightarrow S = e^{Rt+C}; S(0) = I \Longrightarrow S(t) = e^{Rt}$$

$$R_{JC69} = \begin{pmatrix} -1 & -1 & -1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \cdot \text{diag}(-4\alpha, -4\alpha, -4\alpha, 0) \cdot \begin{pmatrix} -0.25 & -0.25 & -0.25 & 0.75 \\ -0.25 & -0.25 & 0.75 & 0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

$$S_{JC69}(t) = \begin{pmatrix} -1 & -1 & -1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \cdot \text{diag}\left(e^{-4\alpha t}, e^{-4\alpha t}, e^{-4\alpha t}, 1\right) \cdot \begin{pmatrix} -0.25 & -0.25 & -0.25 & 0.75 \\ -0.25 & -0.25 & 0.75 & 0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

# Kimura's K80 model

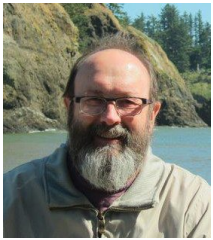
- Also called Kimura's 2 parameter model (K2P)
- A and G are **purines**, C and T are **pyrimidines**
  - Transitions: within the same group $A \longleftrightarrow G$, $C \longleftrightarrow T$
  - Transversions: between the groups
- **Transitions are more frequent** than transversions
  - $\kappa := \dfrac{\text{rate of transitions}}{\text{rate of transversions}}$, set rate of transversions to 1
- $R_{K80} = \begin{pmatrix} * & 1 & \kappa & 1 \\ 1 & * & 1 & \kappa \\ \kappa & 1 & * & 1 \\ 1 & \kappa & 1 & * \end{pmatrix}$
- Equilibrium is still $\pi_A = \pi_C = \pi_G = \pi_T = 25\%$

# Hasewaga-Kishino-Yano HKY85 model

- Transition/transversion ratio $\kappa$ & arbitrary equilibrium $(\pi_A, \pi_C, \pi_G, \pi_T)$
- $R_{HKY85} = \begin{pmatrix} * & \pi_C & \kappa \cdot \pi_G & \pi_T \\ \pi_A & * & \pi_G & \kappa \cdot \pi_T \\ \kappa \cdot \pi_A & \pi_C & * & \pi_T \\ \pi_A & \kappa \cdot \pi_C & \pi_G & * \end{pmatrix}$

# Other models

- Kimura's 3 parameter model (K3P, K81)
  - 1 transition rate + 2 transversion rates
  - admits Hadamard transformation (generalized Fourier)
- Felsenstein F81 model
  - JC69 + arbitrary equilibrium
- Tamura T92 model
  - K80 + GC content
- Tamura and Nay TN93 model
  - 2 transition rates + 1 transversion rate
- Tavaré GTR86 model (General Time Reversible)
  - everything from the above: arbitrary equilibrium + 6 rate parameters



# Summary

- Evolution model: $\Pr[S_1 \xrightarrow{t} S_2]$
  - Independent base evolution $\Longrightarrow \Pr[S_1 \xrightarrow{t} S_2] = \prod_{i=1}^{n} \Pr[a_i \xrightarrow{t} b_i]$
  - Continuous time $t$ + Only substitutions $\Longrightarrow$ Continuous time Markov chains (CTMC)
- Substitution model for one base (CTMC)
  - substitution rate matrix $R = \begin{pmatrix} * & \mu_{A,C} & \mu_{A,G} & \mu_{A,T} \\ \mu_{C,A} & * & \mu_{C,G} & \mu_{C,T} \\ \mu_{G,A} & \mu_{G,C} & * & \mu_{G,T} \\ \mu_{T,A} & \mu_{T,C} & \mu_{T,G} & * \end{pmatrix}$, rows sum up to zero
  - $S_{a,b}(t) = \Pr[a \xrightarrow{t} b]$ from $S(t) = e^{Rt}$ using diagonalization trick
- Different rate matrices $R$ give different models:
  - JC69 model: all substitutions are equally likely, equilibrium 25%
  - K80 model: transition/transversion ratio $\kappa$, equlibrium 25%
  - HKY85 model: K80 + arbitrary equilibrium