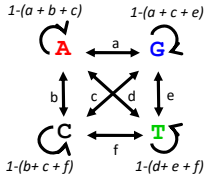
Review: Models of DNA sequence evolution

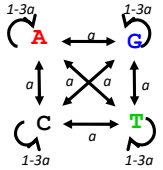
Properties of DNA substitution models



- State space: {E_1 = A, E_2 = G, E_3 = C, E_4 = T}
- States are fully connected
- Transition probabilities: substitution frequencies (*a, b, c, d* ...)
- Implicitly also specifies stationary base frequencies: $\varphi^* = (p_A, p_G, p_C, p_T)$

GACTAGCTAGACATAGCTAGACAGATACGAAGATACGAACTAGCTAGACATATTACATATAC

Jukes-Cantor model (1969)

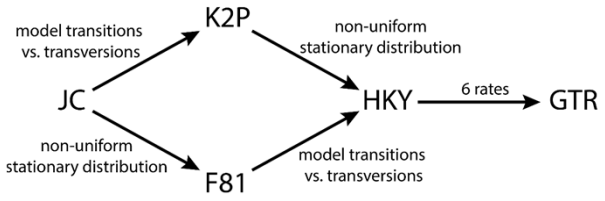


p(A)=0.25
p(G)=0.25
p(C)=0.25
p(T)=0.25

Assumptions:

- All substitutions have equal probability
- Base frequencies are equal

Models of DNA sequence evolution



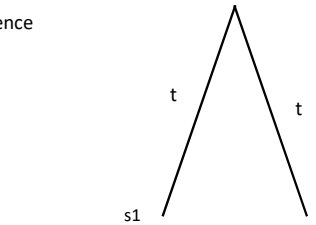
We apply these models to the following scenario

Sequences s_1 and s_2 are DNA sequences of length *n*

s_1 and s_2 have been diverging from a common ancestor for t million years (MY) according to the Jukes Cantor (JC) model with parameter *a*

- All substitutions are equally probable
- This framework does not model indels
- Assumes site independence

…atgcgaggactcacagtga…
…atgtgaggtctcccaatga…



…atgcgaggactcacagtga…
…atgtgaggtctcccaatga…

Questions to ask:

Thursday: Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing *x* aligned with *y*?

…ATGCGAGGACTCXCAGTGA…
…ATGTGAGGTCTCYCAATGA…

Today: Given an alignment of s_1 and s_2 with *m* observed mismatches, how many substitutions occured since the divergence of s_1 and s_2?

…CACATACGAAGATACGAACGAGC…
…CAGATAGGAAGAGACGATCTAGC…
n nucleotides with *m* mismatches

Questions to ask:

***Thursday:* Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing *x* aligned with *y*?**

…ATGCGAGGACTCXCAGTGA…
…ATGTGAGGTCTCYCAATGA…

Today: Given an alignment of s_1 and s_2 with *m* observed mismatches, how many substitutions occured since the divergence of s_1 and s_2?

…CACATACGAAGATACGAACGAGC…
…CAGATAGGAAGAGACGATCTAGC…
n nucleotides with *m* mismatches

Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing *x* aligned with *y*?

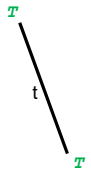
…ATGCGAGGACTCXCAGTGA…
…ATGTGAGGTCTCYCAATGA…

For the Jukes Cantor model, there are 2 cases of interest

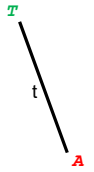
- x=y
- x≠y

Subproblem:

Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing *z* at time 0 and *x* at time *t*?



z=x

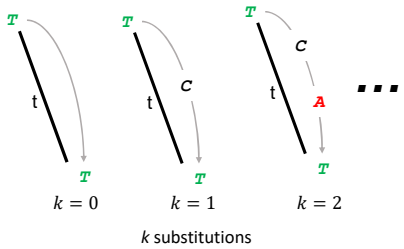


z≠x

8

Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing ***T*** at time 0 and ***T*** at time t?

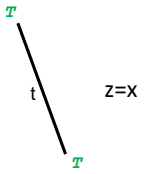
$$\sum_{k=0} P(\varphi_T(0) = 1, \varphi_T(t) = 1 \mid JC(\alpha), k)$$



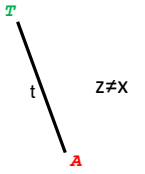
On Thursday, we used a differential equation to integrate over all values of k

Subproblem:

Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing *z* at time 0 and *x* at time *t*?



$$p_{xx}(\alpha, t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$



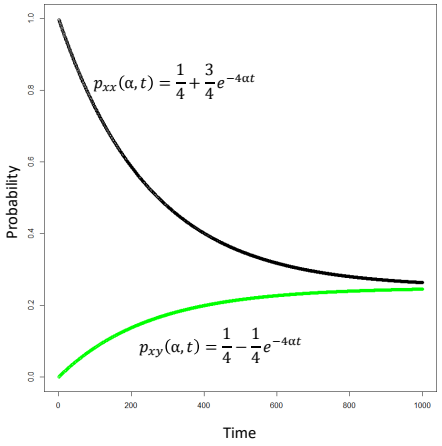
$$p_{xy}(\alpha, t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

10

When ***t = 0***, the probability of observing

- the same nucleotide is one
- a different nucleotide is zero

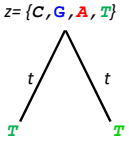
When ***t*** → ∞, the probability of observing any of the four nucleotides is $\frac{1}{4}$.



11

Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing *x* aligned with *x*?

...ACCTGTCCGTAACTTT...
...ACTTATCTGTTACGAT...

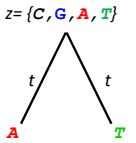


$$P\binom{x}{x}\alpha t) = p_x p_{xx}^2 + 3p_z p_{zx}^2$$

$$= p_z\left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^2 + 3p_z\left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^2$$

Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing *x* aligned with *y*?

...ACCTGTCCGAAACTTT...
...ACTTATCTGTTACGAT...



$$P\left(\begin{matrix}x\\y\end{matrix}\middle|\alpha t\right) = 2p_x p_{xx} p_{zx} + 2p_z p_{zx}^2$$
$$= 2p_x\left(\tfrac{1}{4}+\tfrac{3}{4}e^{-4\alpha t}\right)\left(\tfrac{1}{4}-\tfrac{1}{4}e^{-4\alpha t}\right)$$
$$+2p_z\left(\frac{1}{4}-\frac{1}{4}e^{-4\alpha t}\right)^2$$

Questions to ask:

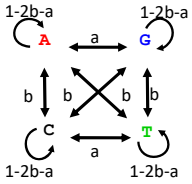
Thursday: Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing *x* aligned with *y*?

...ATGCGAGGACTCXCAGTGA...
...ATGTGAGGTCTCYCAATGA...

Today: **Given an alignment of s_1 and s_2 with *m* observed mismatches, how many substitutions occured since the divergence of s_1 and s_2?**

...CACATACGAAGATACGAACGAGC...
...CAGATAGGAAGAGACGATCTAGC...
n nucleotides with *m* mismatches

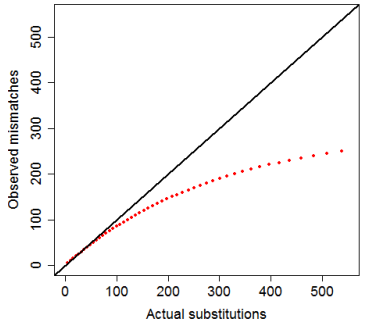
Correcting for distances for multiple substitutions with sequence evolution models



Given an alignment of *n* nucleotides with *m* observed mismatches…

...CACATACGAAGATACGAACGAGC...
...CAGATAGGAAGAGACGATCTAGC...
n nucleotides with *m* mismatches

…estimate the expected number of substitutions since the divergence of the two sequences



Correcting for multiple substitutions with Jukes-Cantor

Given an alignment of *n* nucleotides that differs at *m* positions, the expected number of substitutions since the divergence of the two sequences is given by

$$D = \frac{-3}{4}\ln\left(1 - \frac{4}{3}\frac{m}{n}\right).$$

...CACATACGAAGATACGAACGAGC...
...CAGATAGGAAGAGACGATCTAGC...

n nucleotides with *m* mismatches

For example, if we observe 200 mismatches in an alignment of 1000 nucleotides, then the number of actual substitutions is

$$\frac{-3}{4}\ln\left(1 - \frac{800}{3000}\right) = 233 \text{ substitutions}$$

18