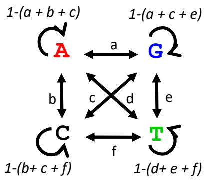
Review: Models of DNA sequence evolution

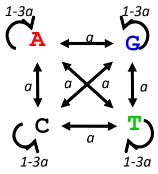
Properties of DNA substitution models



- 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

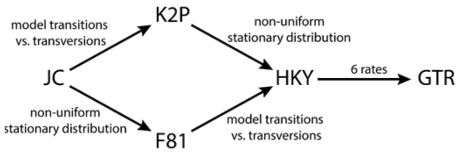
Jukes-Cantor model (1969)



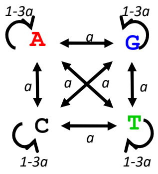
p(A)=0.25
p(G)=0.25
p(C)=0.25
p(T)=0.25

Assumptions:

- All substitutions have equal probability
- Base frequencies are equal



Jukes-Cantor



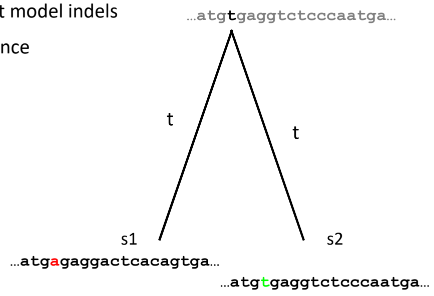
p(A)=0.25
p(G)=0.25
p(C)=0.25
p(T)=0.25

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



Questions to ask:

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

```
...ATGCGAGGACTCXCAGTGA...
...ATGTGAGGTCTCYCAATGA...
```

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 *α*, what is the probability of observing *x* aligned with *y*?

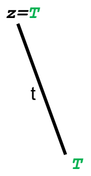
```
...ATGCGAGGACTCXCAGTGA...
...ATGTGAGGTCTCYCAATGA...
```

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

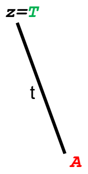
- x=y
- x≠y

Subproblem:

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



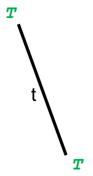
z=x



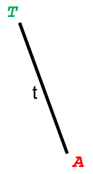
z≠x

Subproblem:

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



z=x



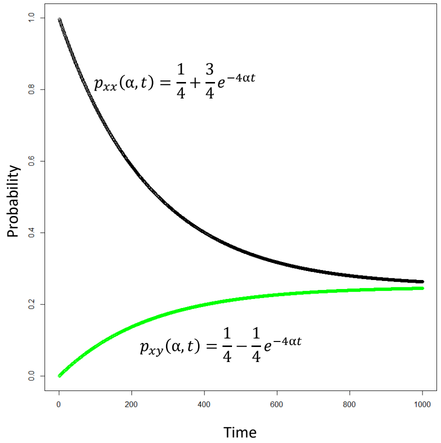
z≠x

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

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

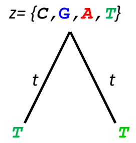
When ***t = 0***, the probability of observing
- the same nucleotide is one
- a different nucleotide is zero

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



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

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

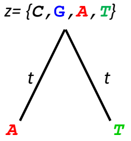


$$P\left(\begin{smallmatrix}x\\x\end{smallmatrix}\middle|\alpha t\right) = p_x p_{xx}^2 + 3p_z p_{zx}^2$$

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

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

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



$$P\left(\begin{smallmatrix}x\\y\end{smallmatrix}\middle|\alpha t\right) = 2p_x p_{xx} p_{zx} + 2p_z p_{zx}^2$$

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

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

Questions to ask:

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

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

Next: 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