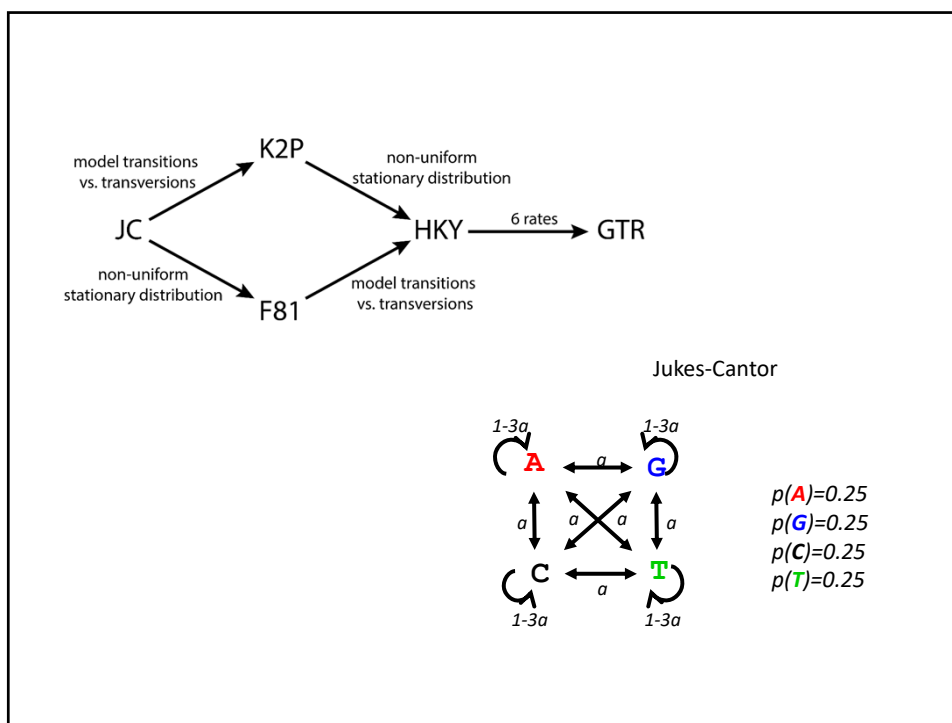


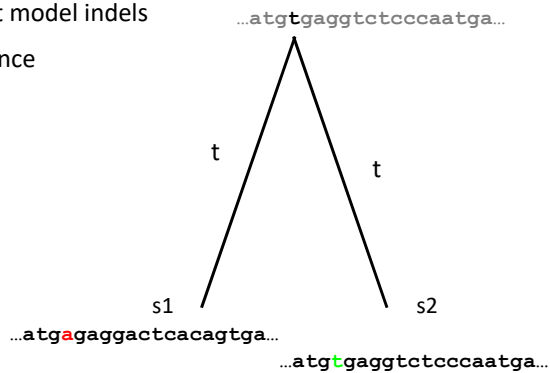


We apply these models to the following scenario

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:

Last week: Given a site evolving according to Jukes Cantor with parameter α , 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 occurred since the divergence of s_1 and s_2 ?

```

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

```

← n nucleotides with μ mismatches →

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

```

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

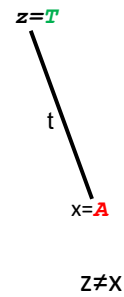
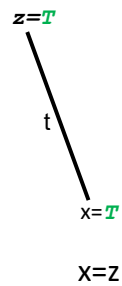
```

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

- $x=y$
- $x \neq y$

Subproblem:

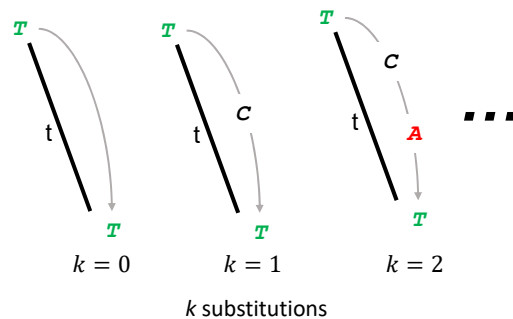
Given a site evolving according to Jukes Cantor with parameter α , what is the probability of observing z at time 0 and x at time t ?



5

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

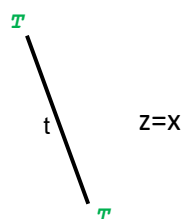
$$\sum_{k=0} P(\varphi_T(0) = 1, \varphi_T(t) = 1 | JC(\alpha), k)$$



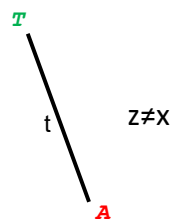
Last week, we used a differential equation to integrate over all values of k

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 ?



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



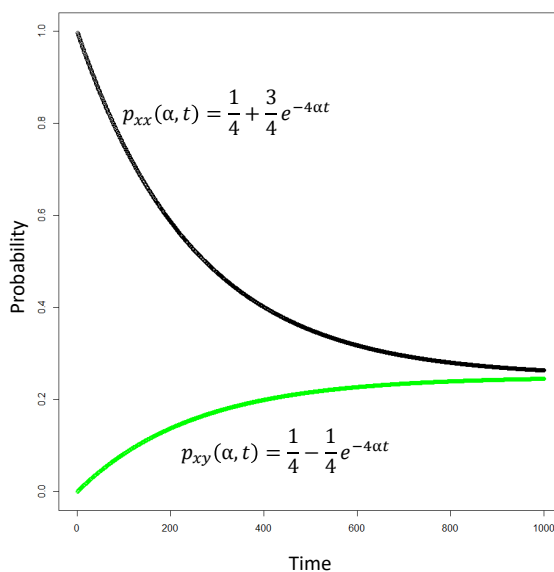
$$p_{xy}(\alpha, t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

7

When $t = 0$, the probability of observing

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

When $t \rightarrow \infty$, the probability of observing any of the four nucleotides is $\frac{1}{4}$.



8

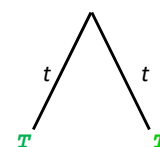
Given a site evolving according to Jukes Cantor with parameter α , what is the probability of observing x aligned with x at time t ?

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

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

$$= \frac{1}{16}(1 + 3e^{-8\alpha t})$$

$z = \{C, G, A, T\}$



...ACCTGTCCG**T**AAC**T**TT...
...ACTTATCTG**T**TACGAT...

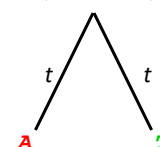
Given a site evolving according to Jukes Cantor with parameter α , what is the probability of observing x aligned with y at time t ?

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

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

$$= \frac{1}{16}(1 - e^{-8\alpha t})$$

$z = \{C, G, A, T\}$



...ACCTGTCCG**A**AAC**T**TT...
...ACTTATCTG**T**TACGAT...

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...

```

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

```

...CACATACGAAGATACGAACGAGC...
...CAGATAGGAAGAGACGATCTAGC...
<----->
n nucleotides with  $\mu$  mismatches

```

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 α

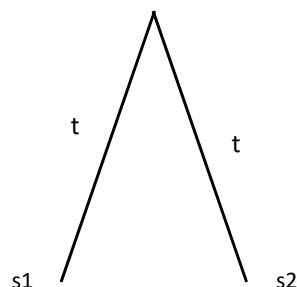
Given an alignment of n nucleotides with m observed mismatches...

```

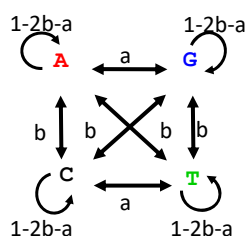
...CACATACGAAGATACGAACGAGC...
...CAGATAGGAAGAGACGATCTAGC...
<----->
n nucleotides with  $\mu$  mismatches

```

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



Correcting for distances for multiple substitutions with sequence evolution models



Given an alignment of n nucleotides with m observed mismatches...

...CACATACGAAGATACGAACGAGC...
 ...CAGATAGGAAGAGACGATCTAGC...
 $\longleftrightarrow$
 n nucleotides with μ 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

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

...CACATACGAAGATACGAACGAGC...
 ...CAGATAGGAAGAGACGATCTAGC...
 $\longleftrightarrow$
 n nucleotides with μ mismatches

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

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

Correcting for multiple substitutions with Jukes-Cantor

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

