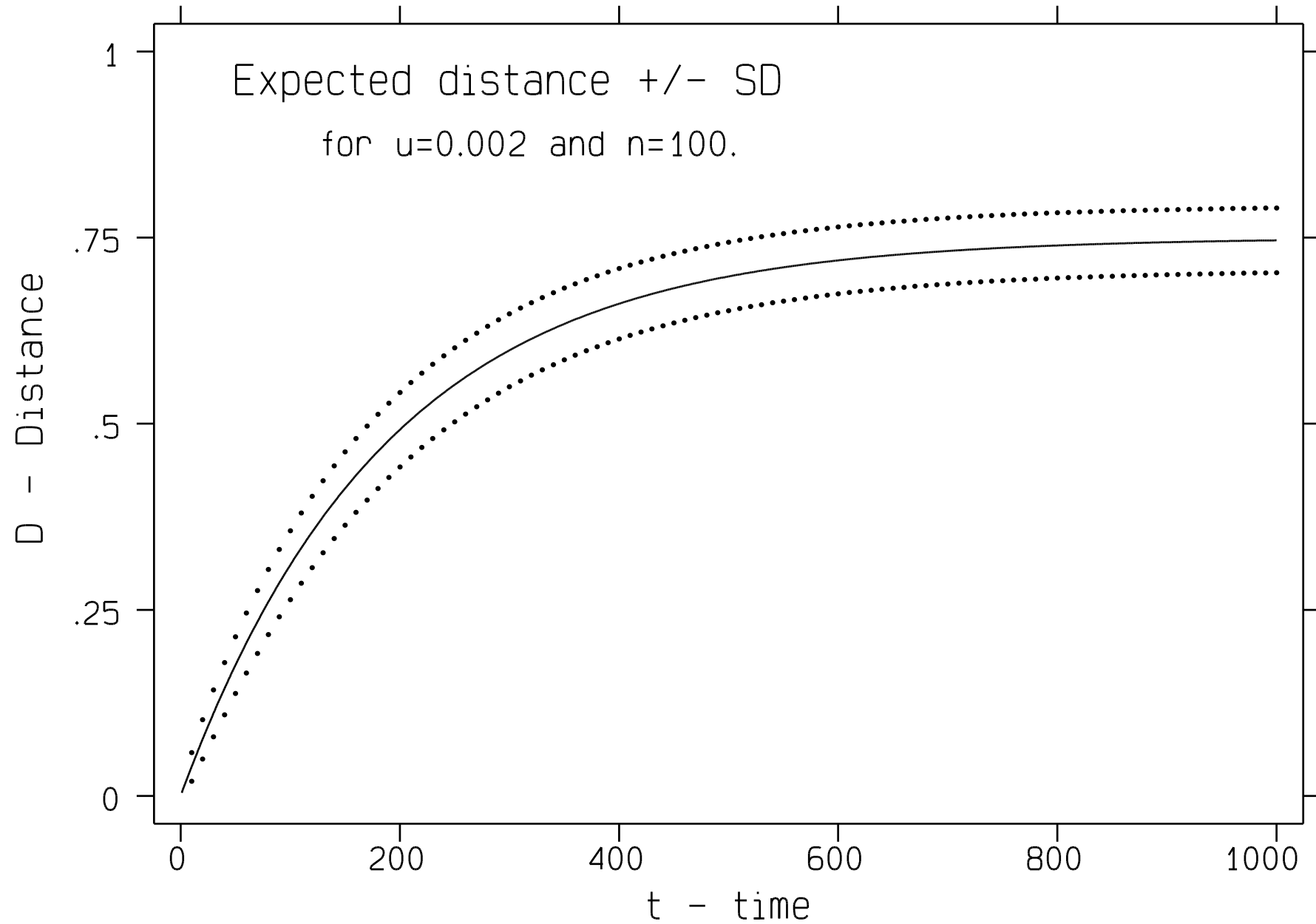
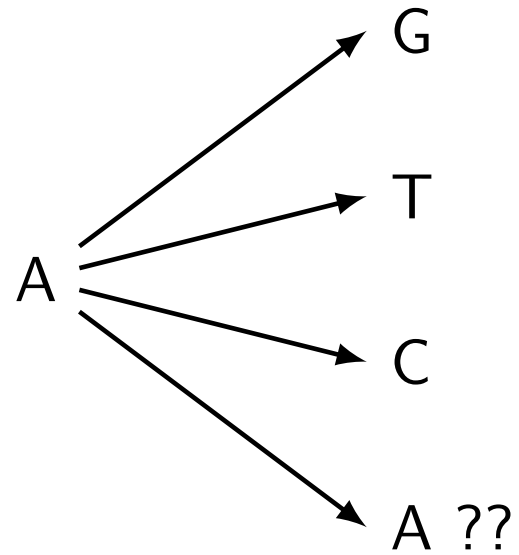


Saturation problem

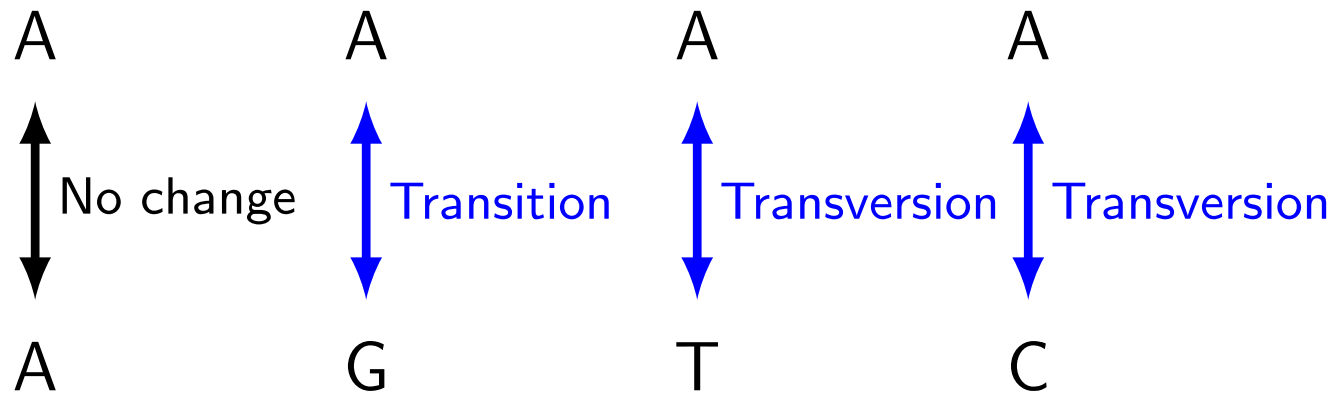


Saturation problem

A T G A T T A C G G G G T A C G C T T A G
A T G A C T A T G G A G T T C G C A T A G



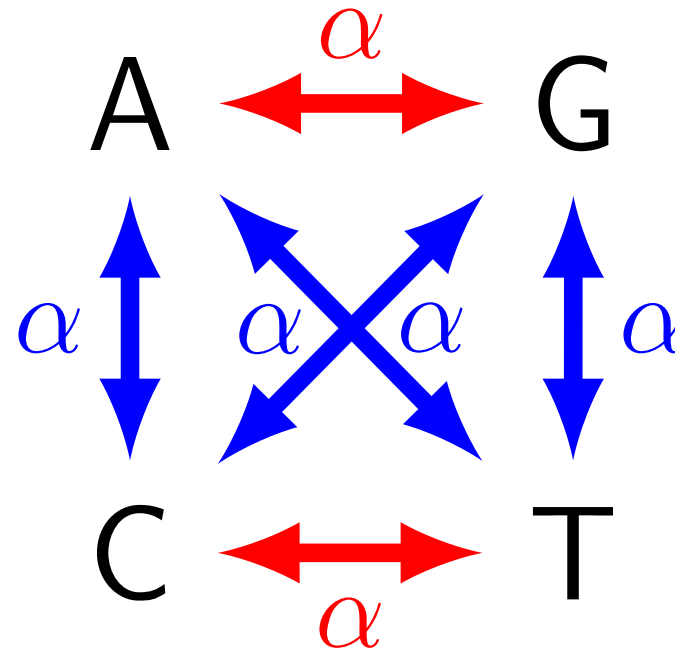
Saturation problem



One parameter models

purines

pyrimidines



Jukes-Cantor model

Nucleotide substitutions arise at any site with equal frequency.

Each nucleotide changes to one of the three others with the same probability

	A	T	C	G
A	—	α	α	α
T	α	—	α	α
C	α	α	—	α
G	α	α	α	—

Prob nucleotides stay the same time over time

Let P_t represent the probability that a nucleotide is A at time t .

Then at time 0 if there is an A then $P_0 = 1$. What happens over time?

Prob nucleotides stay the same time over time

Let P_t represent the probability that a nucleotide is A at time t .

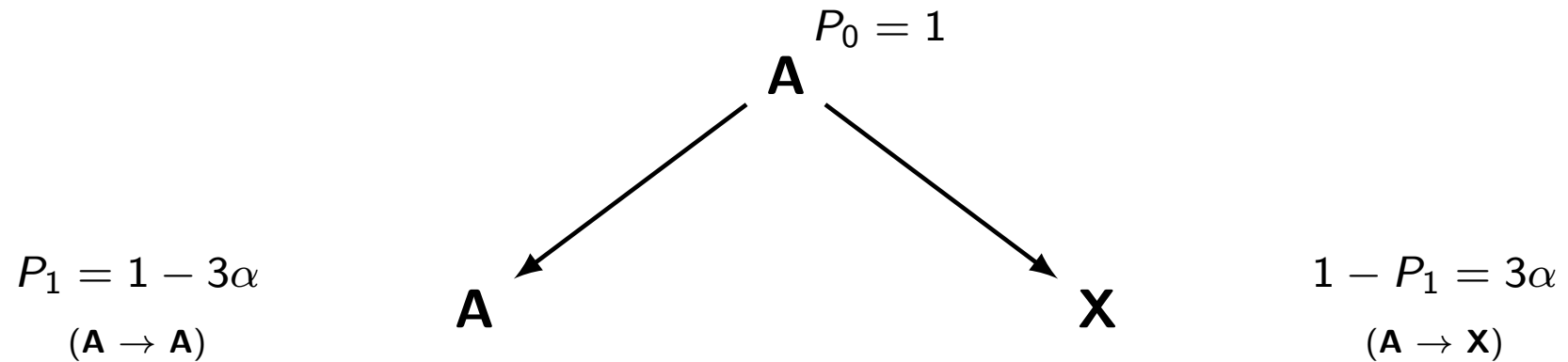
Then at time 0 if there is an A then $P_0 = 1$. What happens over time?

$$\mathbf{A} \quad P_0 = 1$$

Prob nucleotides stay the same time over time

Let P_t represent the probability that a nucleotide is A at time t .

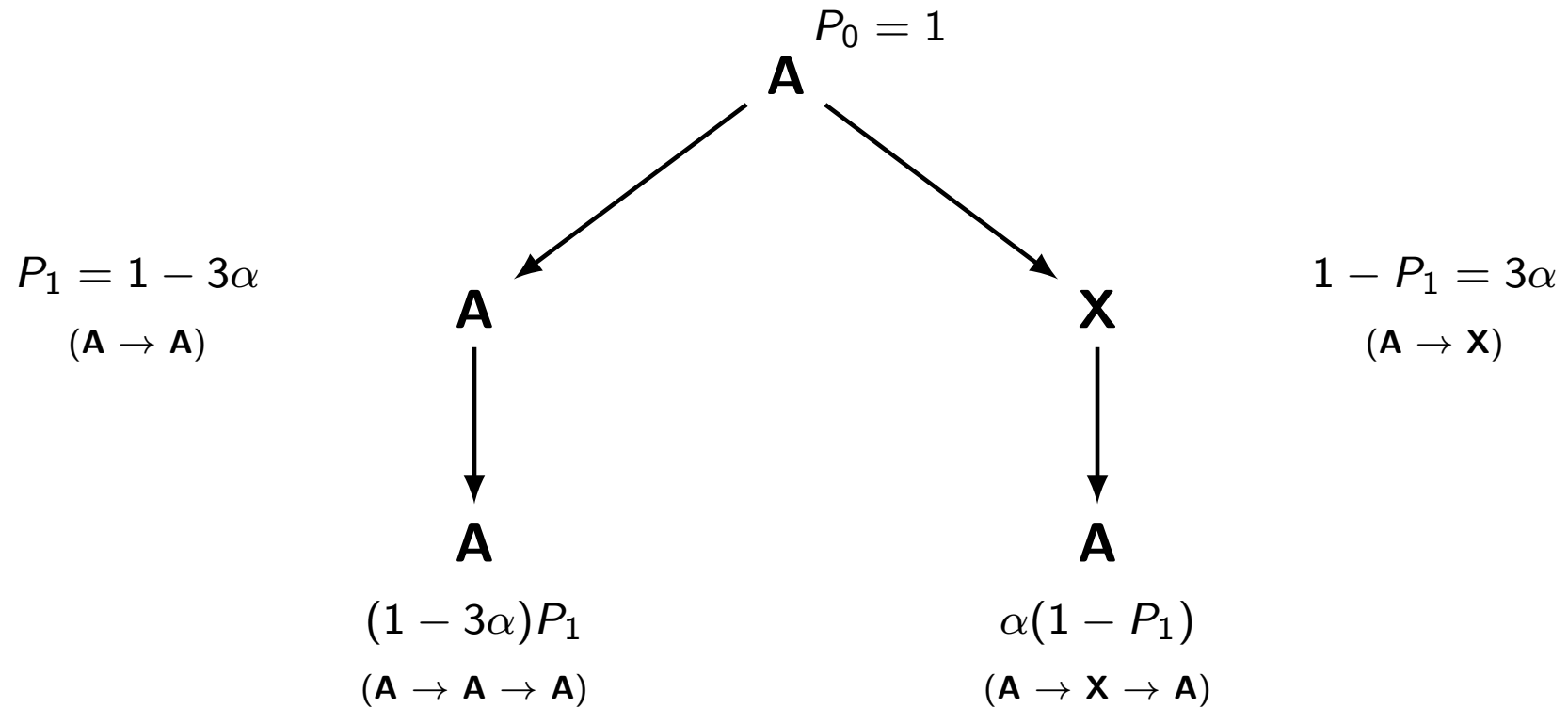
Then at time 0 if there is an A then $P_0 = 1$. What happens over time?



Prob nucleotides stay the same time over time

Let P_t represent the probability that a nucleotide is A at time t .

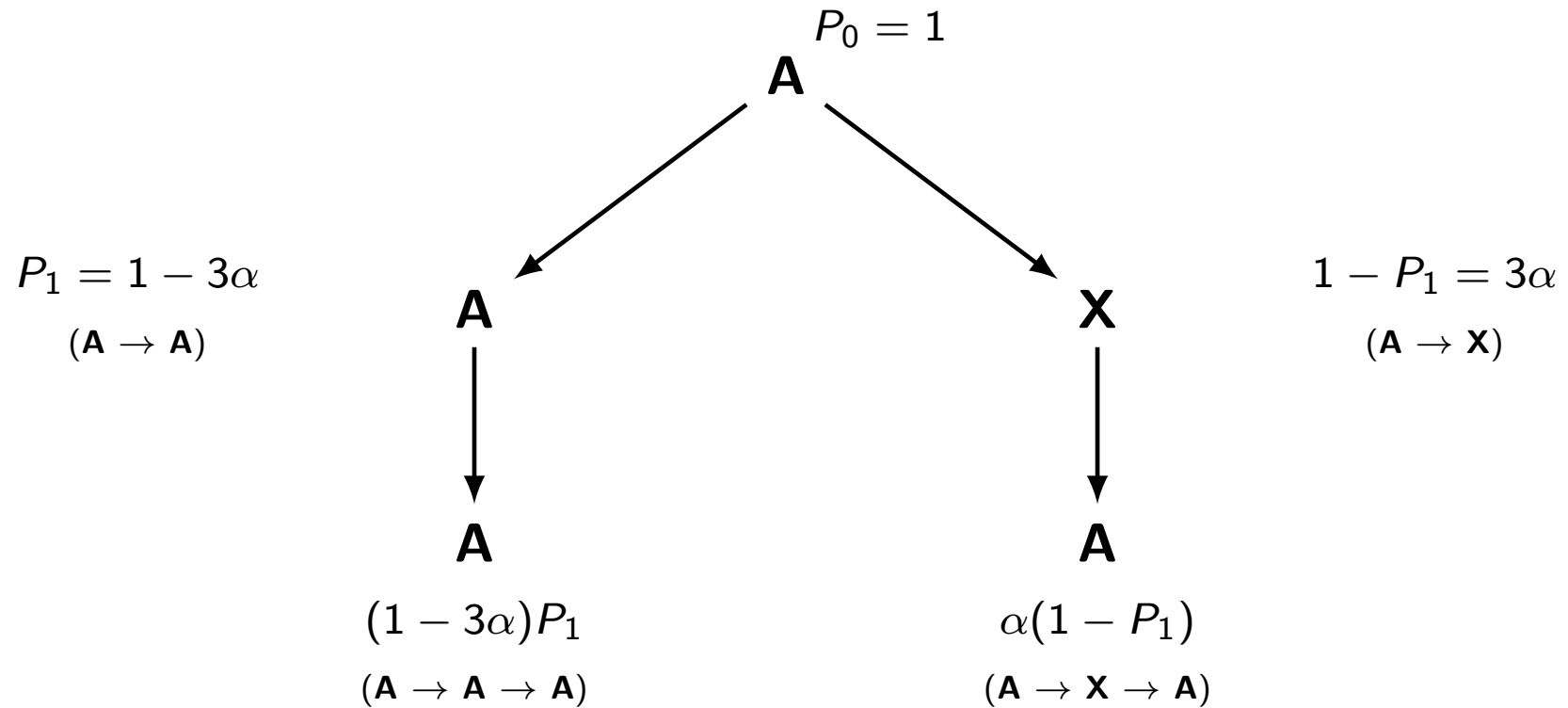
Then at time 0 if there is an A then $P_0 = 1$. What happens over time?



Prob nucleotides stay the same time over time

Let P_t represent the probability that a nucleotide is A at time t .

Then at time 0 if there is an A then $P_0 = 1$. What happens over time?

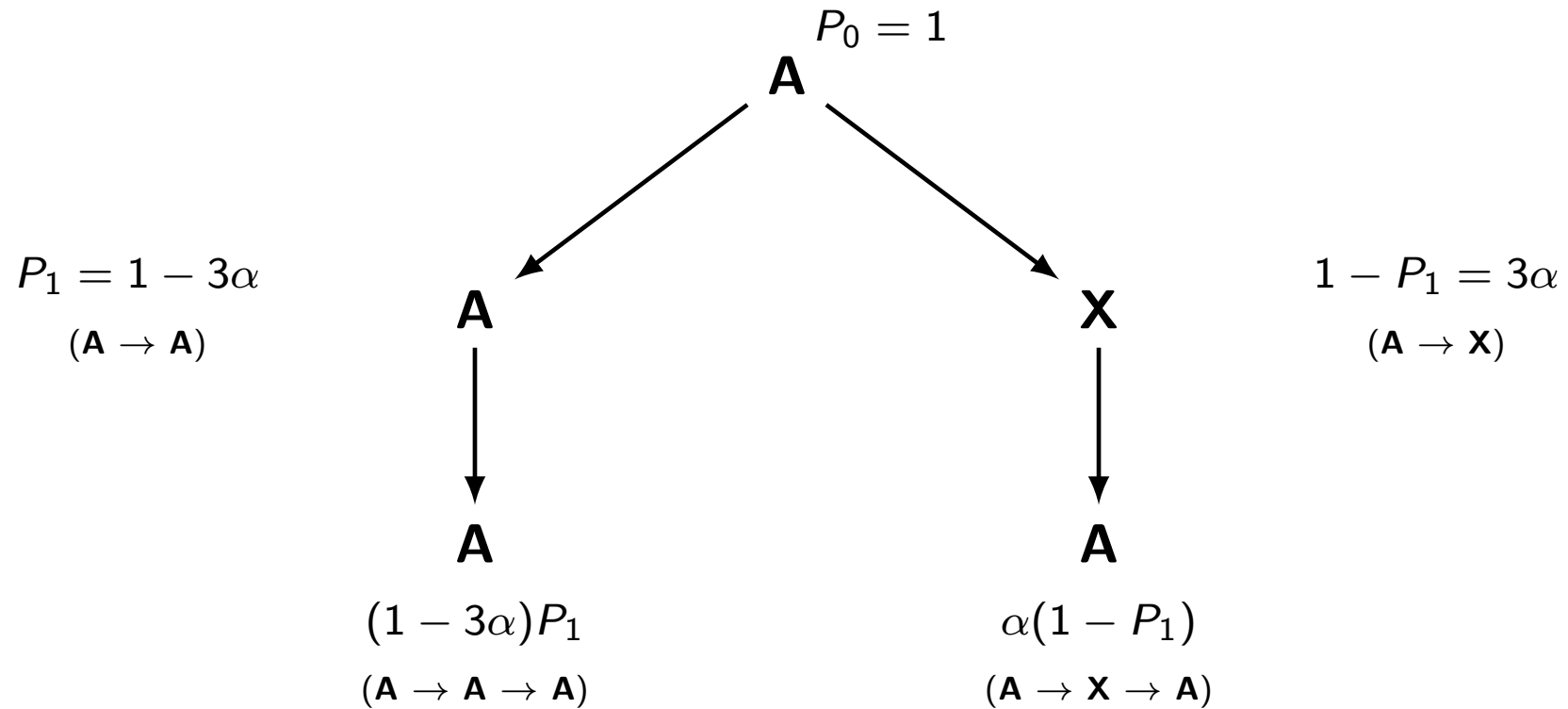


$$P_2 = (1 - 3\alpha)P_1 + \alpha(1 - P_1)$$

Prob nucleotides stay the same time over time

Let P_t represent the probability that a nucleotide is A at time t .

Then at time 0 if there is an A then $P_0 = 1$. What happens over time?



$$P_2 = (1 - 3\alpha)P_1 + \alpha(1 - P_1)$$

$$P_{t+1} = (1 - 3\alpha)P_t + \alpha(1 - P_t)$$

Change over time

So

$$P_{t+1} = (1 - 3\alpha)P_t + \alpha(1 - P_t)$$

or if you prefer,

$$\frac{dP_t}{dt} = P_{t+1} - P_t$$

$$\frac{dP_t}{dt} = -4\alpha P_t + \alpha$$

Either way this has a solution of ...

$$P_t = \frac{1}{4} + \left(P_0 - \frac{1}{4}\right)e^{-4\alpha t}$$

Change over time

From

$$P_t = \frac{1}{4} + (P_0 - \frac{1}{4})e^{-4\alpha t}$$

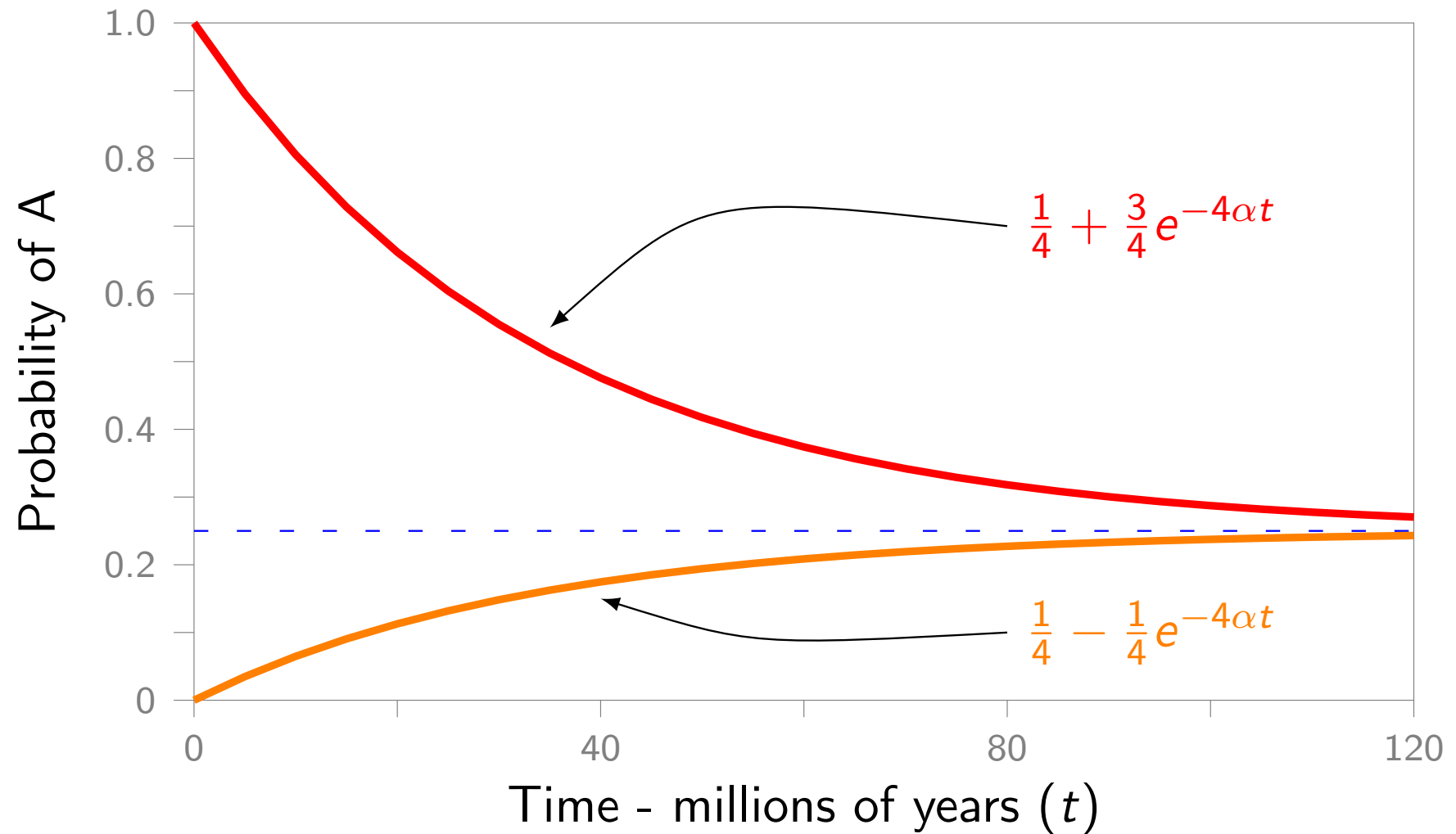
if $P_0 = 1$ then

$$P_A = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

or if $P_0 = 0$ then

$$P_A = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

Change over time



Change over time

This holds whether the nucleotide is an A, T, C, or G. From $P_0 = 1$ then

$$P_{i \rightarrow i, t} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

while $P_0 = 0$ then

$$P_{i \rightarrow j, t} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

where $P_{i \rightarrow j, t}$ is the probability of a change from nucleotide i to j in time t .

Change over time

So the overall probability, Q_t , that two sequences differ at a site is

$$Q_t = 1 - \text{“probability they stayed the same”}$$

Change over time

So the overall probability, Q_t , that two sequences differ at a site is

$$Q_t = 1 - \text{“probability they stayed the same”}$$

so ...

$$Q_t = 1 - [P_{i \rightarrow i,t}^2 + 3 \times P_{i \rightarrow j,t}^2]$$

The latter is multiplied by 3 because nucleotide i could change to anyone of the 3 other nucleotides and each term is squared because you must have both sequences either stay the same or change identically in parallel (both species must have the same change since their last common ancestor else they will be different).

Stick everything in and you get ...

$$Q_t = 1 - \left(\frac{1}{4} + \frac{3}{4} e^{-8\alpha t} \right)$$

$$Q_t = \frac{3}{4} (1 - e^{-8\alpha t})$$

Change over time

To be lazy, I will use P_t as Q_t . So ...

$$P_t = \frac{3}{4}(1 - e^{-8\alpha t})$$

This tells you how two (homologous, ie: aligned) nucleotides change over time as a function of t and α .

Change over time

The number of changes in one sequence is just

$$D_1 = 3\alpha t$$

The number of changes between two sequences is thus,

$$\begin{aligned} D_2 &= 2(3\alpha t) &= 6\alpha t \\ & &= 2\mu t \end{aligned}$$

where $\mu = 3\alpha$.

Change over time

If P is the prob of a difference and I am examining a 100bp sequence then I should see about $100 \times P$ differences. e.g. if $P = 0.22$ then in 100bp I should see 22 differences. The Hamming distance is $D = 22/100$. But I can invert this to see how many changes actually occurred.

Change over time

Sticking in $\mu = 3\alpha$,

$$P = \frac{3}{4} \left(1 - e^{-\frac{4}{3}(2\mu t)} \right)$$

We now wish to solve for $2\mu t$ (which is our best guess of divergence) in terms of P (an observable quantity).

Change over time

Sticking in $\mu = 3\alpha$,

$$P = \frac{3}{4}(1 - e^{-\frac{4}{3}(2\mu t)})$$

We now wish to solve for $2\mu t$ (which is our best guess of divergence) in terms of P (an observable quantity).

Our best guess of P is just D – the Hamming distance.

Change over time

$$P = \frac{3}{4}(1 - e^{-\frac{4}{3}(2\mu t)})$$

$$1 - \frac{4}{3}P = e^{-\frac{4}{3}(2\mu t)}$$

$$\ln(1 - \frac{4}{3}P) = -\frac{4}{3}(2\mu t)$$

$$2\mu t = -\frac{3}{4}\ln(1 - \frac{4}{3}P)$$

$$D_{JC} = -\frac{3}{4}\ln(1 - \frac{4}{3}D)$$

So here we are estimating P , the probability of a difference at one nucleotide, from the Hamming distance D , the average percent difference.

Change over time

So the genetic distance between two sequences, a best estimate of the number of changes that have occurred, is

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

This is known as the Jukes-Cantor (1969) distance.

It has variance

$$\text{Var}(D_{JC}) = \frac{D(1 - D)}{n\left(1 - \frac{4}{3}D\right)^2}$$

Change over time

If the Hamming distance is $D = 22/100 = 0.22$ then the Jukes-Cantor distance is

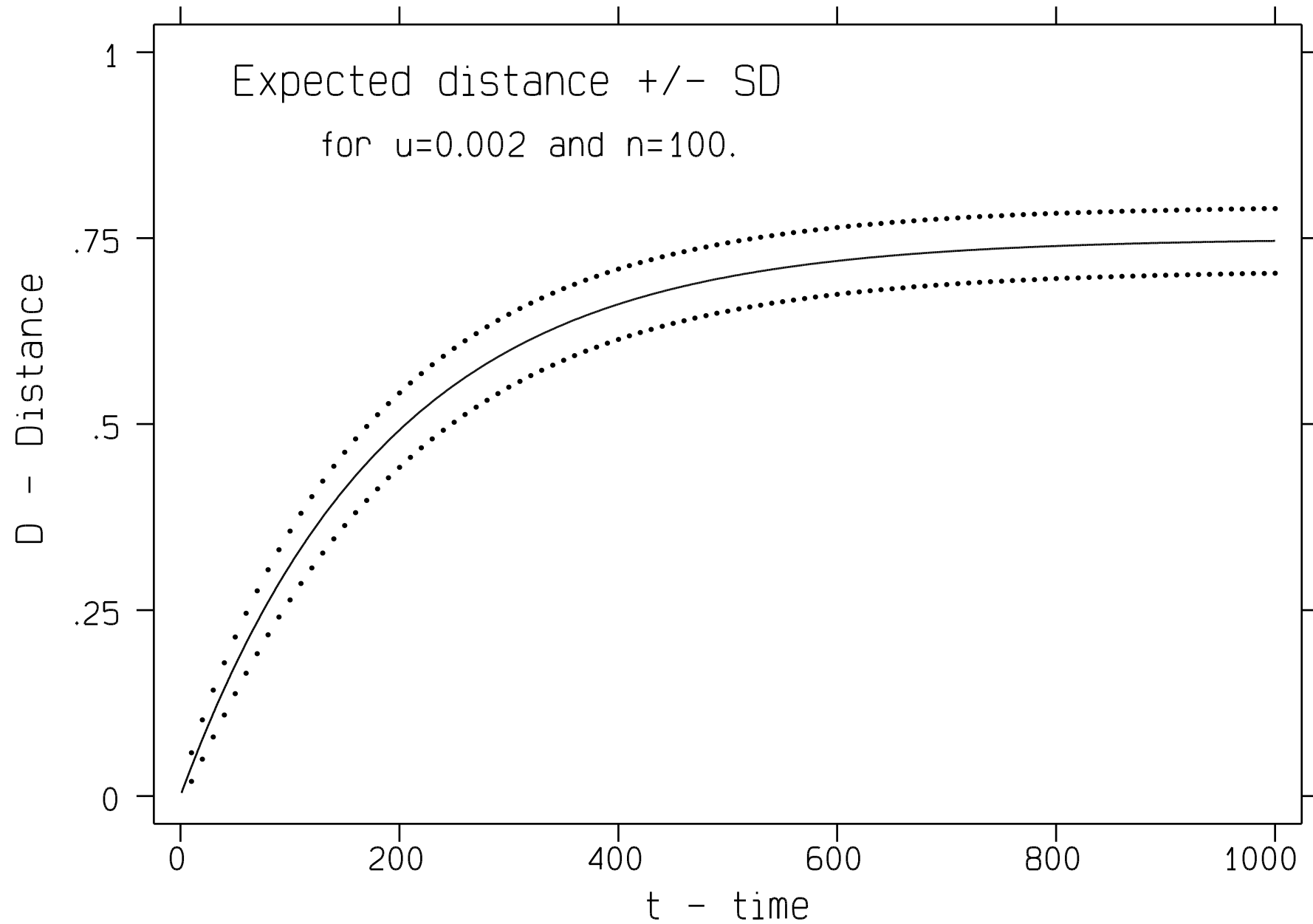
$$D_{JC} = -\frac{3}{4} \ln\left(1 - \frac{4}{3}0.22\right)$$

$$D_{JC} = 0.2604$$

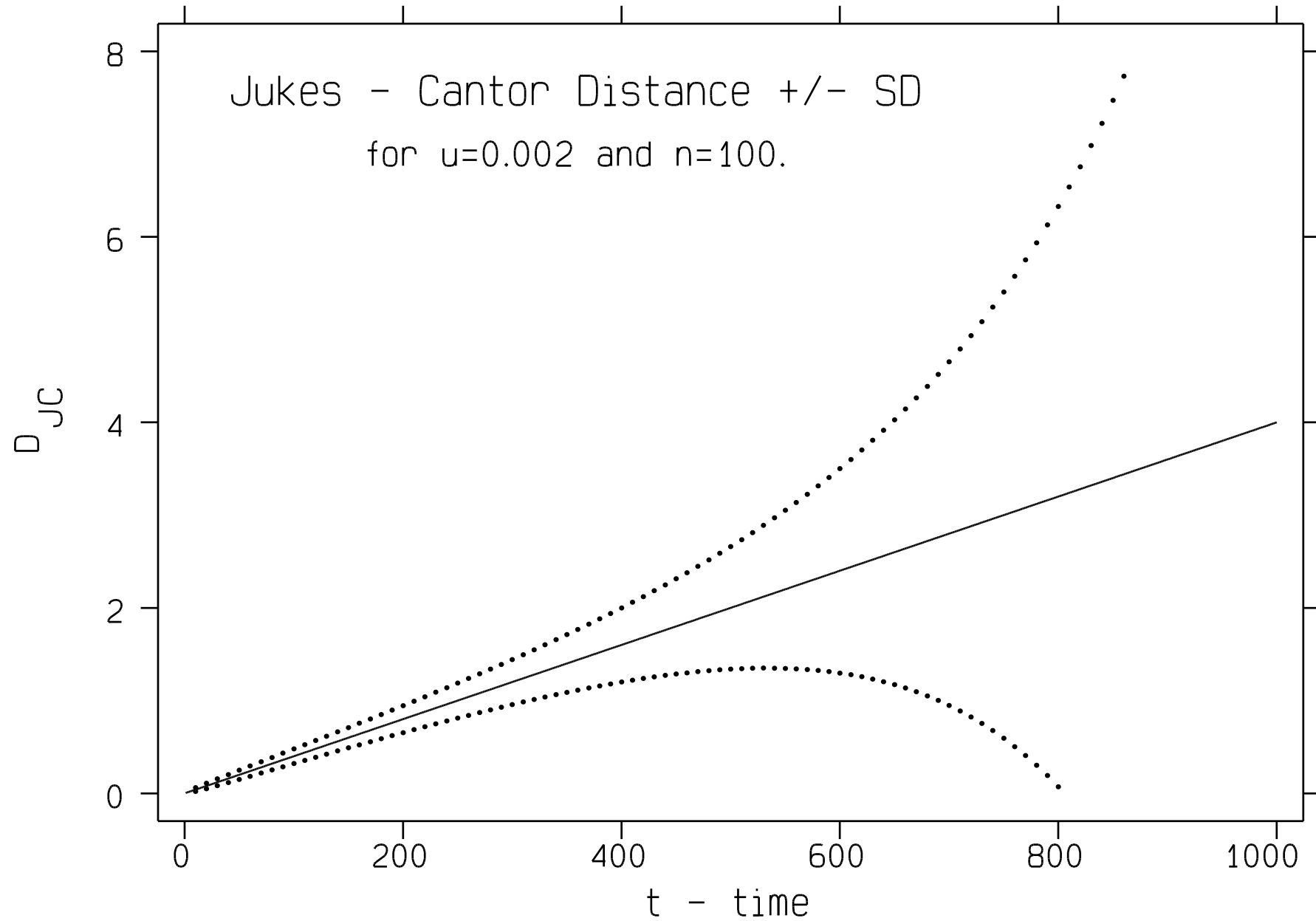
So this suggests that in your sequence of 100bp although you observed 22 differences between the sequences, on average you should expect these 22 differences to have been caused by 26.04 (0.2604 ± 0.0586) actual changes.

The extra changes are changes that might have caused a site to change and then to change back again; or to change in parallel within each species.

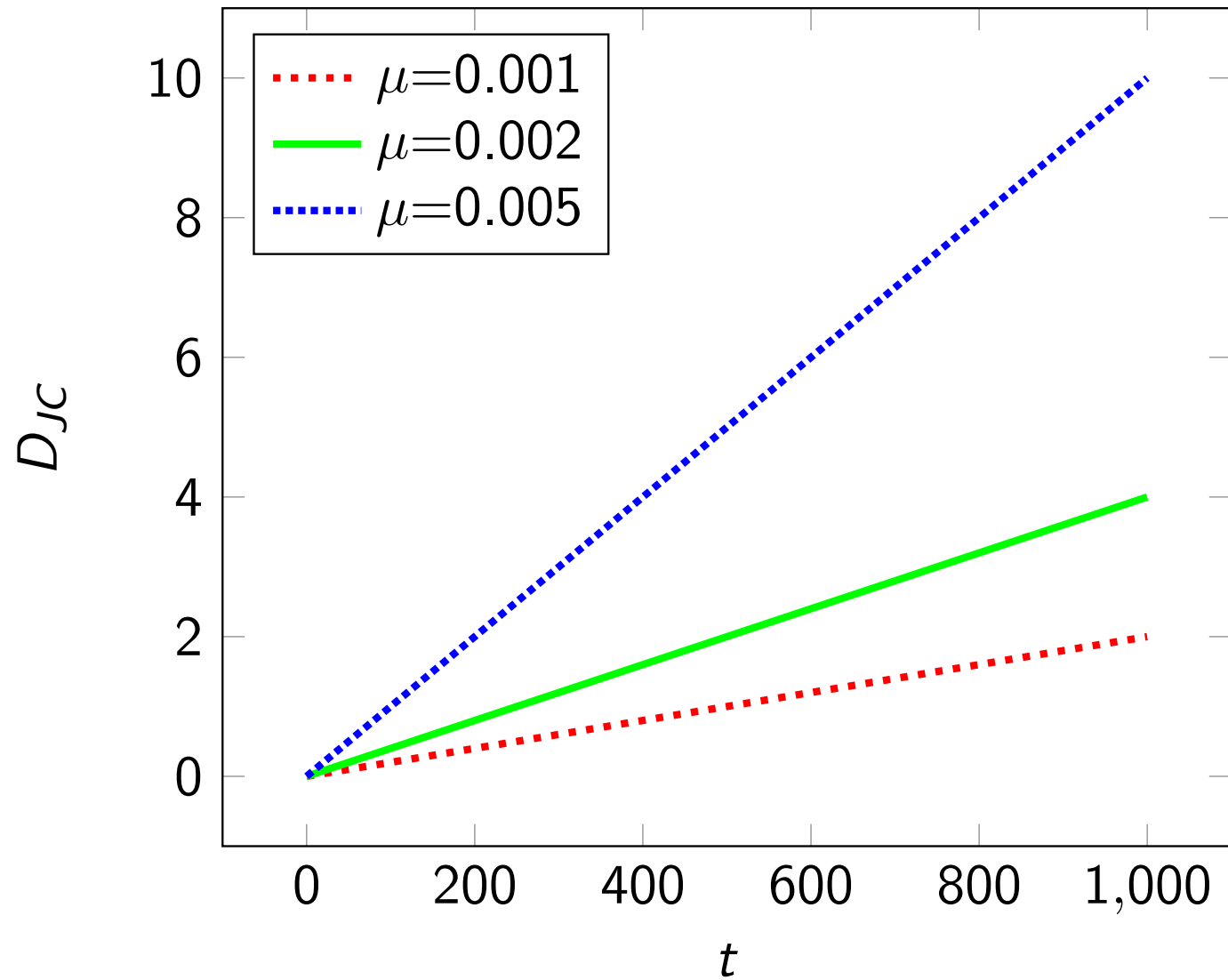
Back to the Saturation problem



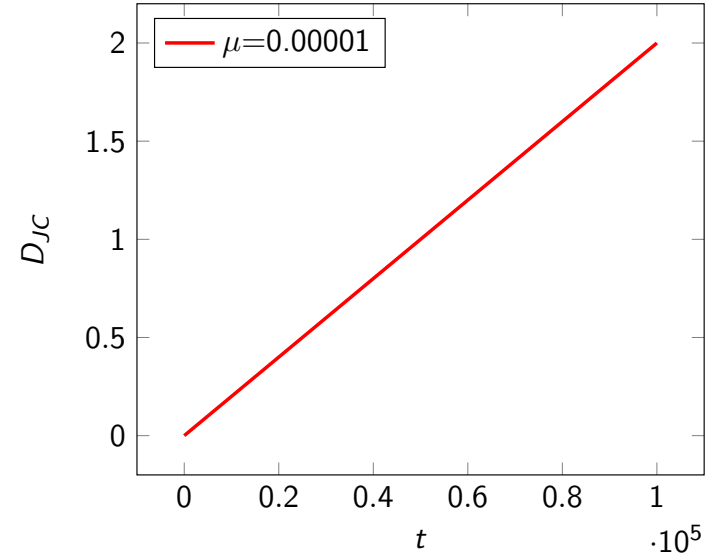
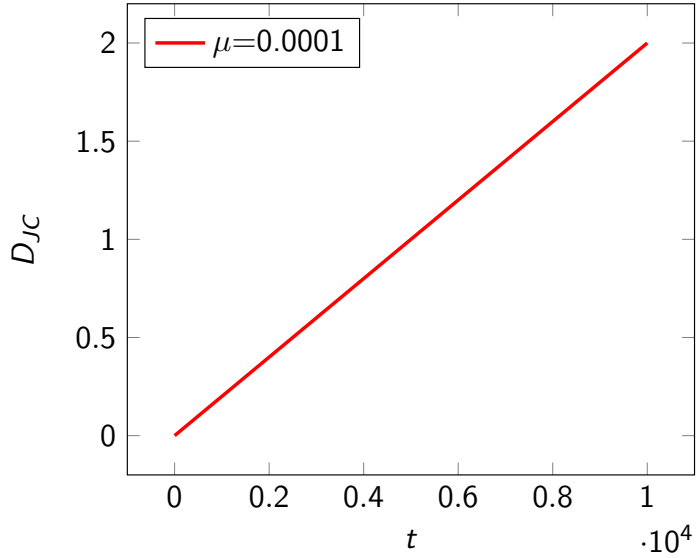
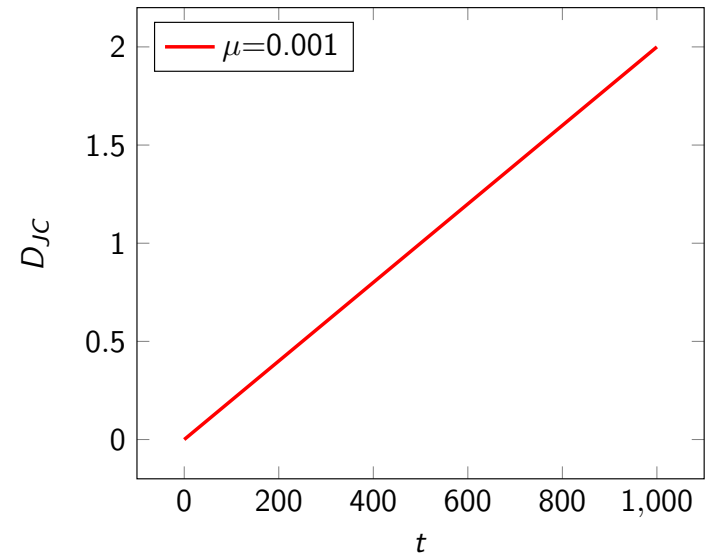
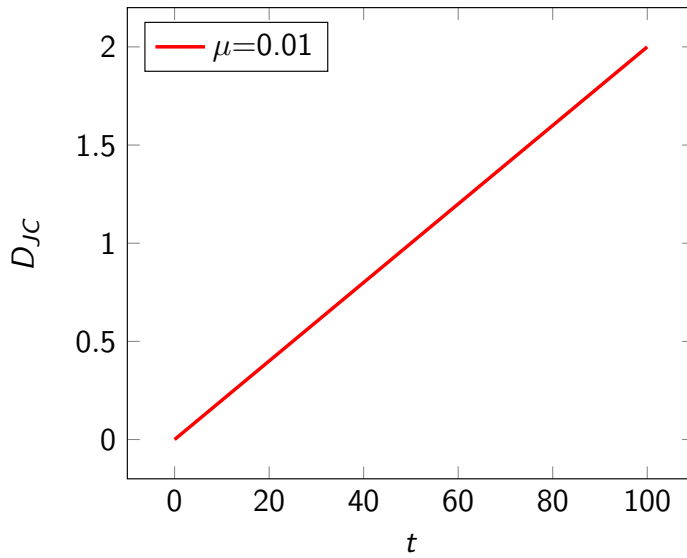
Change over time



Change over time – A faster rate = faster divergence



Change over time – it's all the same



Change over time

The value of μ doesn't matter, the value of t doesn't matter.

It is the product μt that matters.

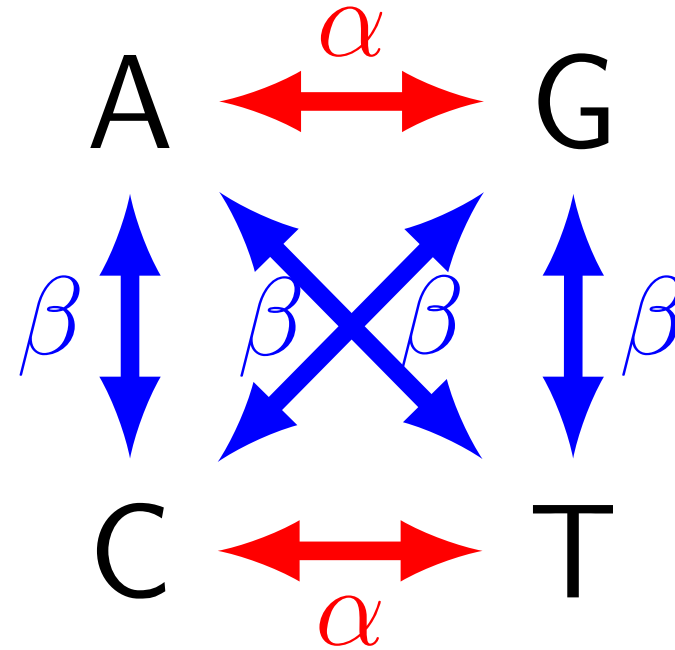
$2\mu t$, often labelled k , is a measure of the number of substitutions per site.

The distinction between time and substitution rate is lost without extraneous information.

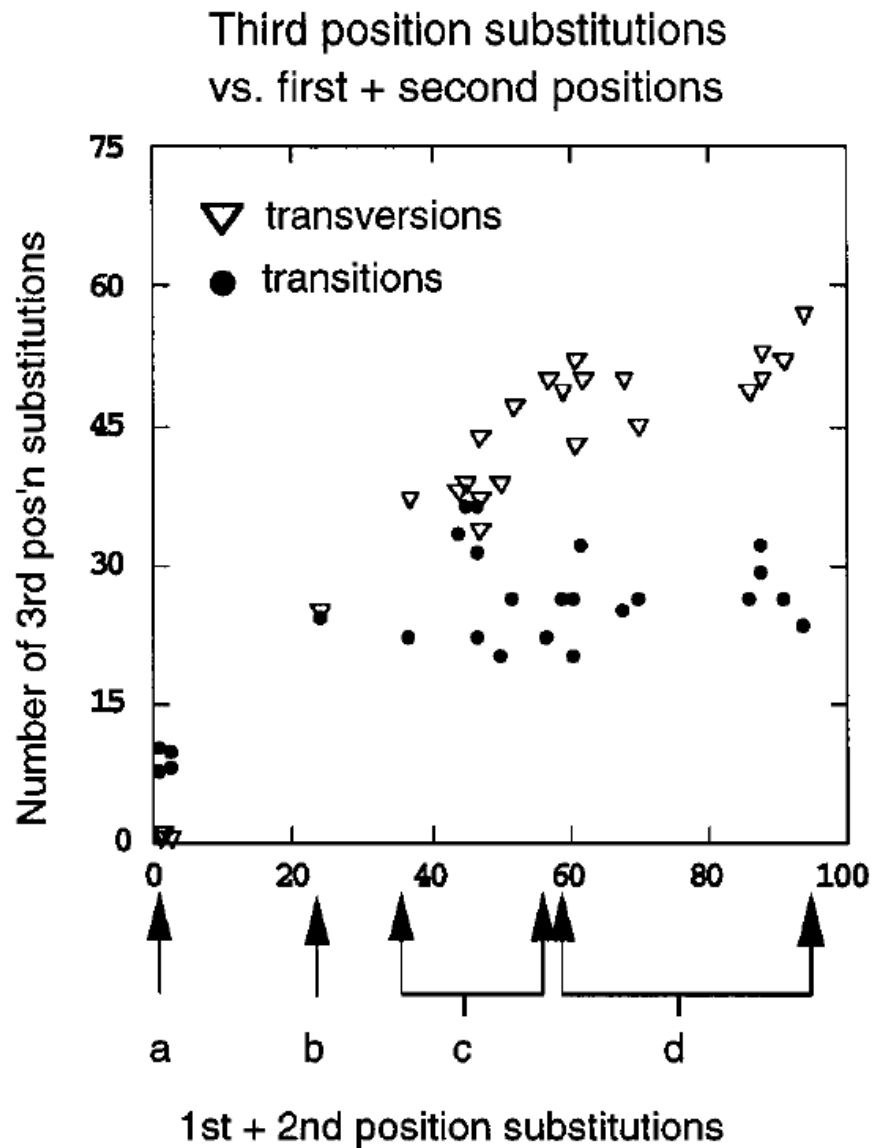
Two parameter models

purines

pyrimidines



Saturation of mtDNA transitions



Observed numbers of transitions and transversions at third-position sites plotted against total substitutions at first- plus second- position sites for pairwise comparisons within and between species. Small letters roughly denote taxonomic levels of comparisons: a 5 intraspecific; b 5 intrageneric (*H. placei* vs. *H. contortus*); c 5 intrafamilial (between trichostrongylid genera); d 5 interordinal (*C. elegans* vs. *A. suum* vs. the trichostrongylids). Note how rapidly the ts/tv ratio changes in going from intraspecific to interspecific comparisons.

Nematode mtDNA ND4 genes
From: Blouin et al. 1998 MBE 15:1719

Kimura Two Parameters model

Nucleotide substitutions arise at any site with equal frequency.

But each nucleotide changes to/from purine/pyrimidine at a different rate from changes within a grouping

	A	T	C	G
A	—	β	β	α
T	β	—	α	β
C	β	α	—	β
G	α	β	β	—