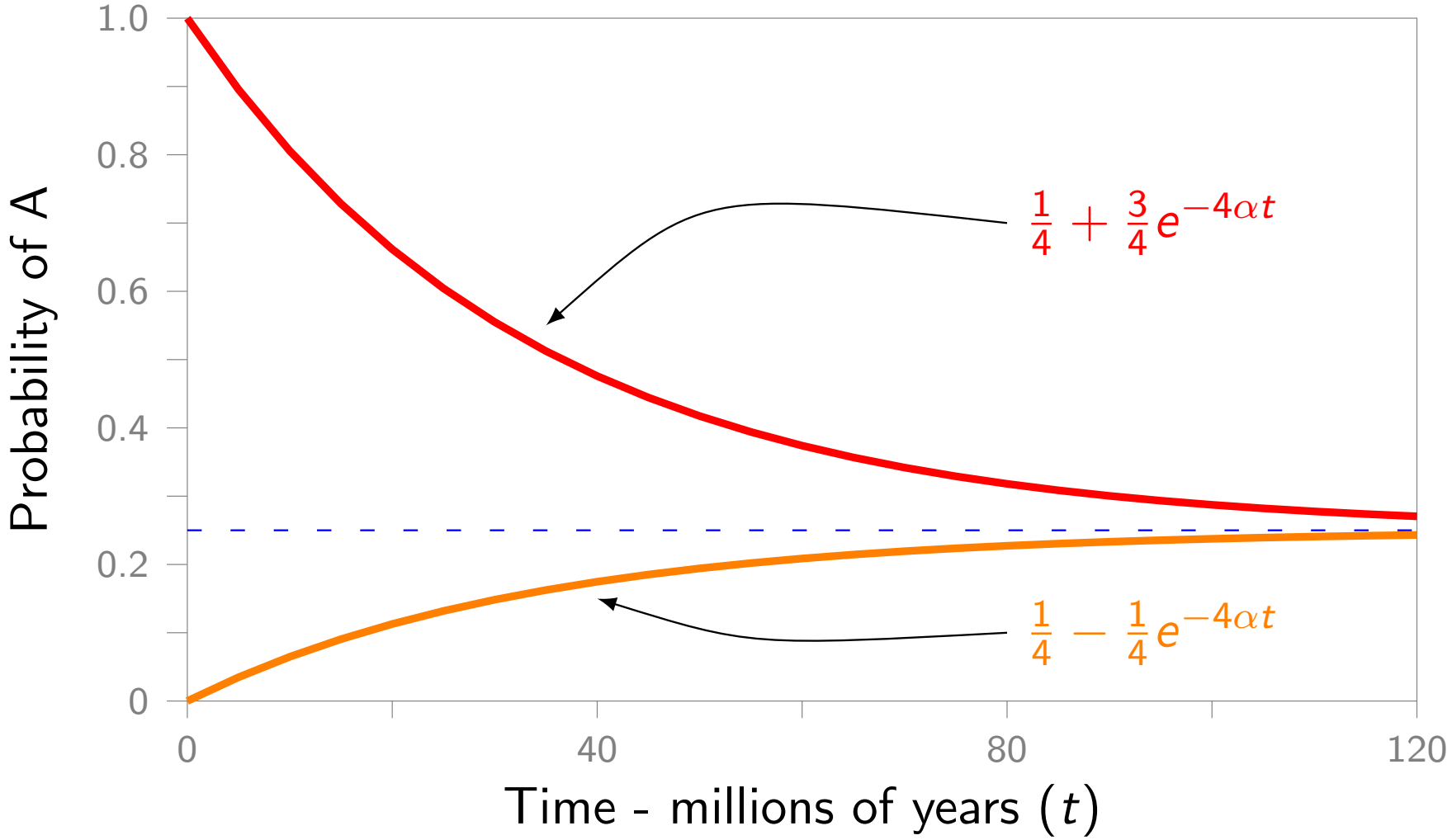
Change over time



Change over time

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

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

while $P_0 = 0$ then

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

where $P_{i \to j,t}$ 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 ...

$$Q_t = 1 - [P^2_{i \to i,t} + 3 \times P^2_{i \to j,t}]$$

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 ...

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

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

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 $\mu = 3\alpha$,

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

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

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

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

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

It has variance

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

Change over time

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

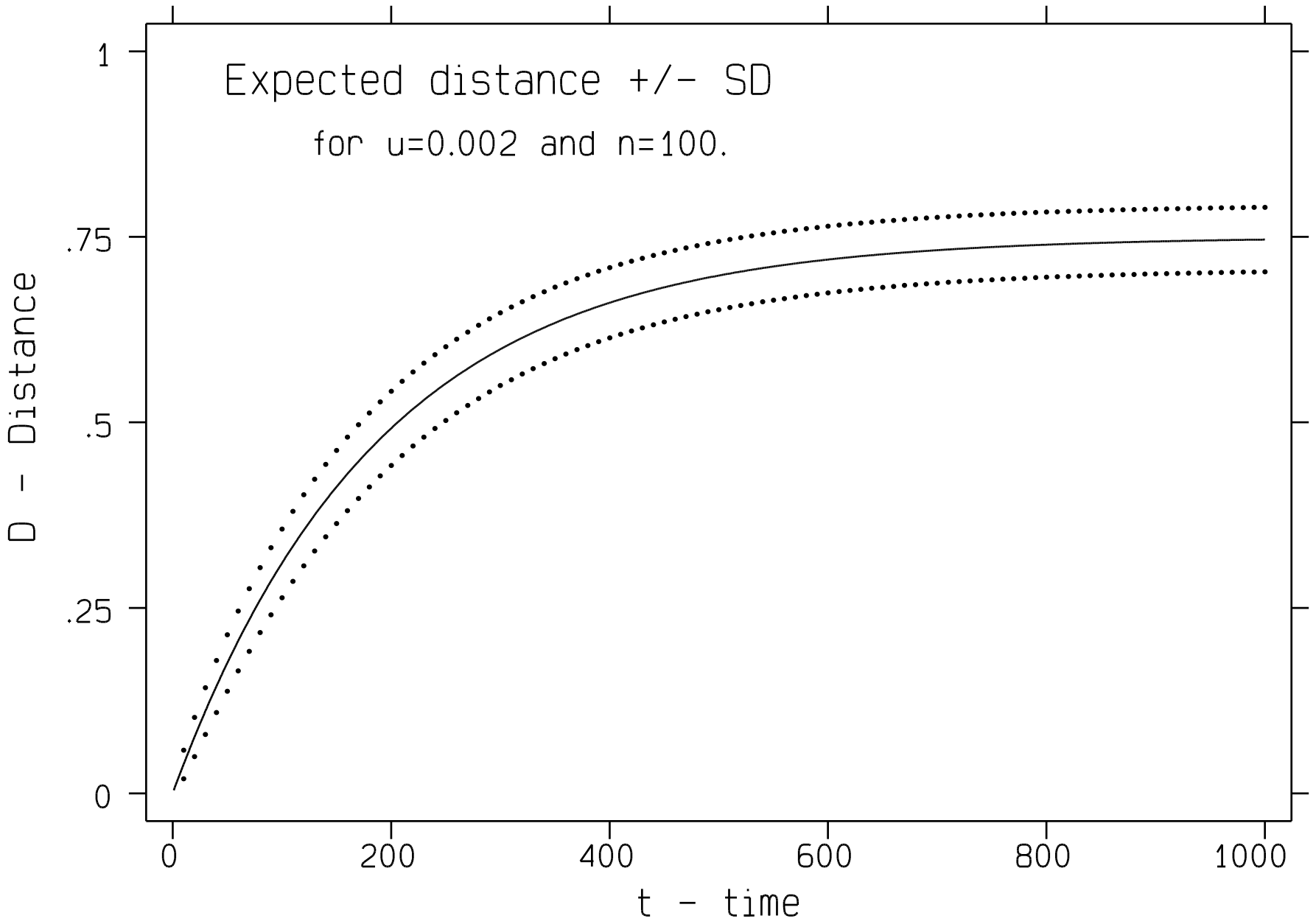
$$D_{JC} = -\frac{3}{4} ln(1 - \frac{4}{3} 0.22)$$

$$D_{JC} = 0.2604$$

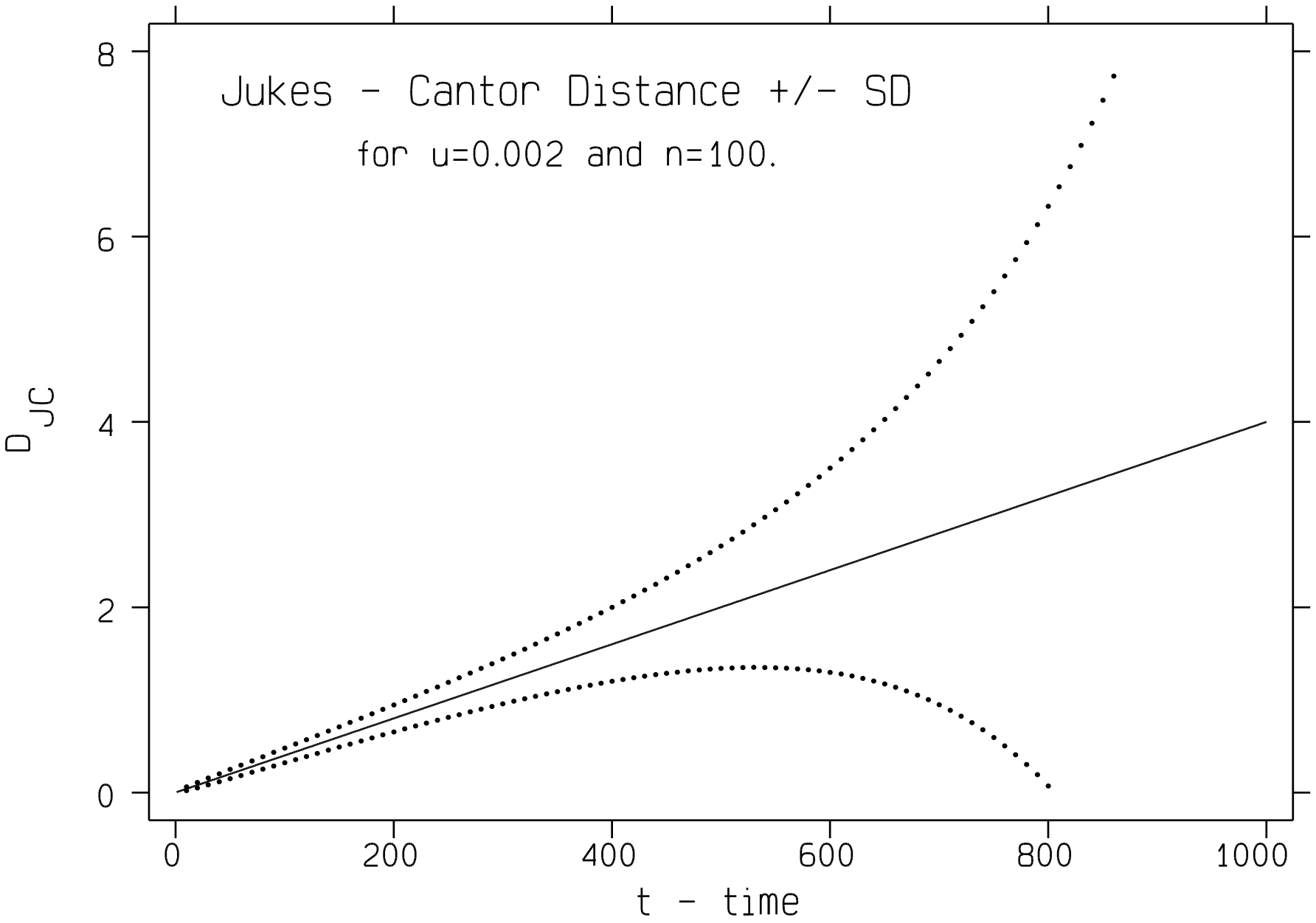
So this suggests that in your sequence of 100bp although you observed 22 differences between the sequences, on average you should expect these 22 differences to have been caused by 26.04 (0.2604 ± 0.0586) actual changes.

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.

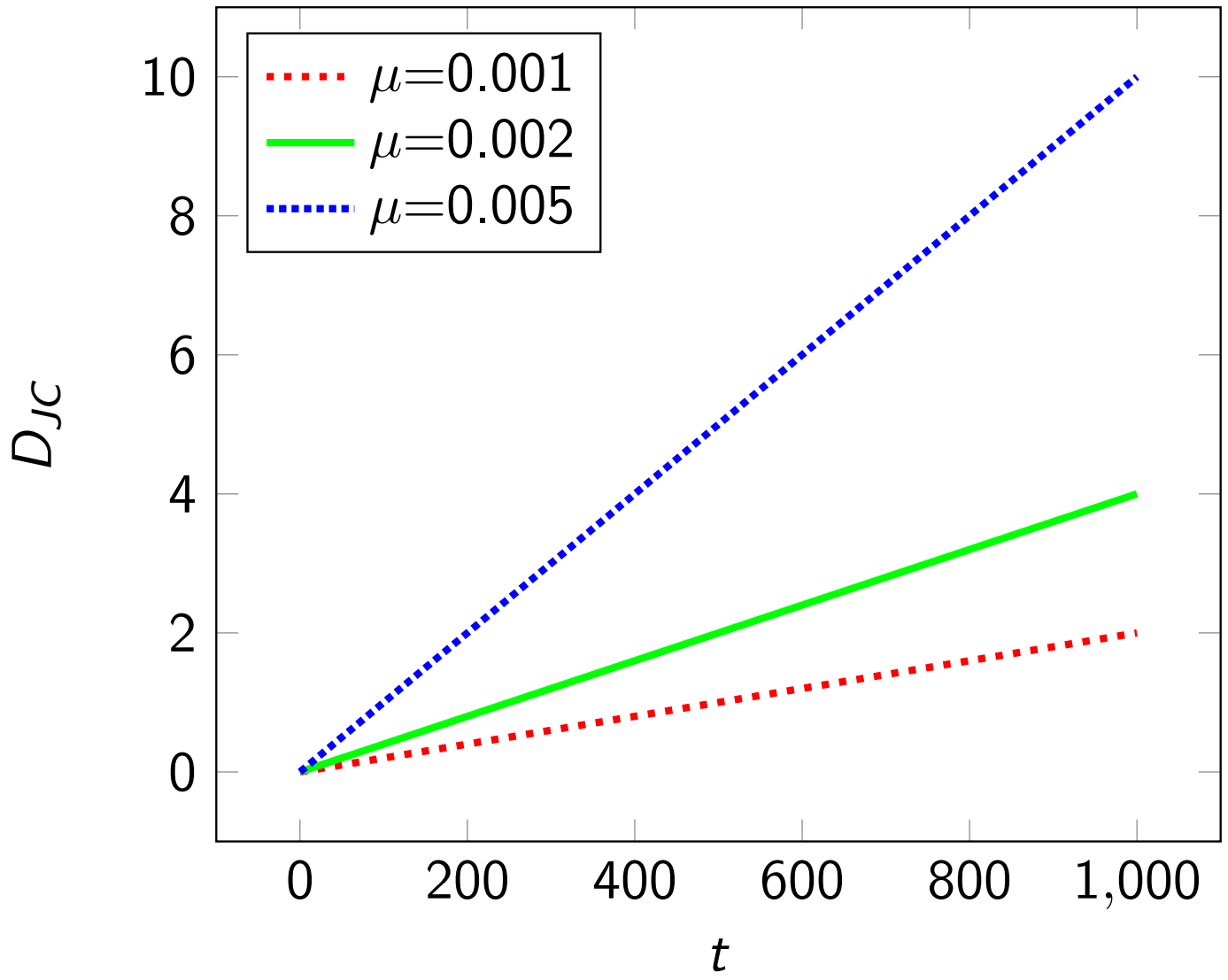
Back to the Saturation problem



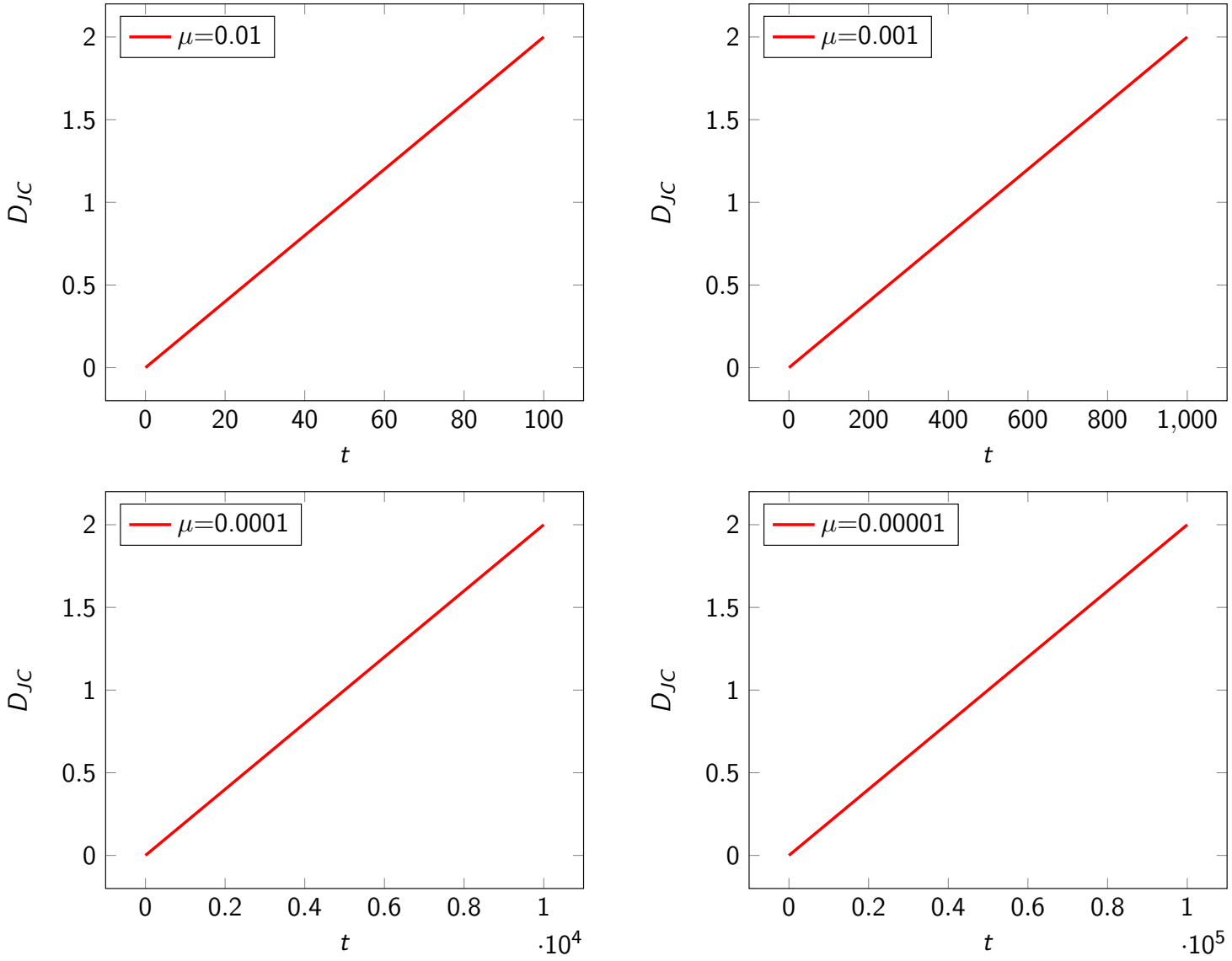
Change over time



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.

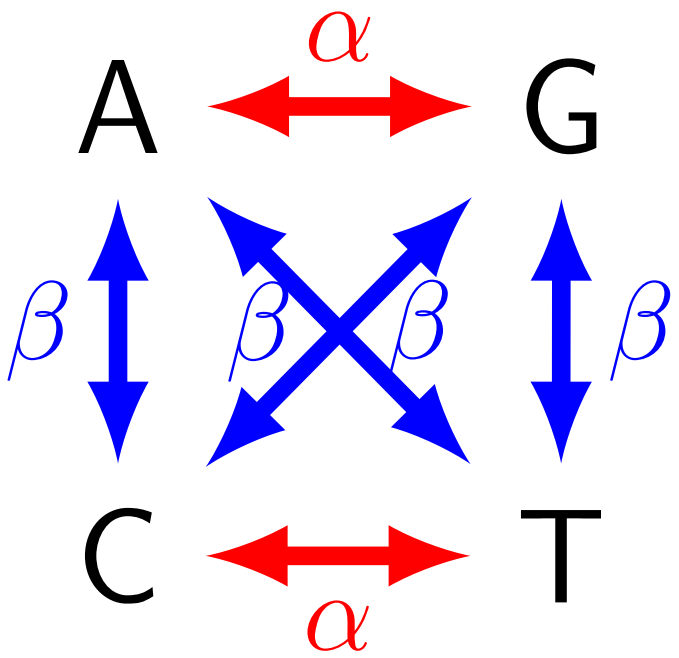
$2\mu t$, often labelled k, is a measure of the number of substitutions per site.

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

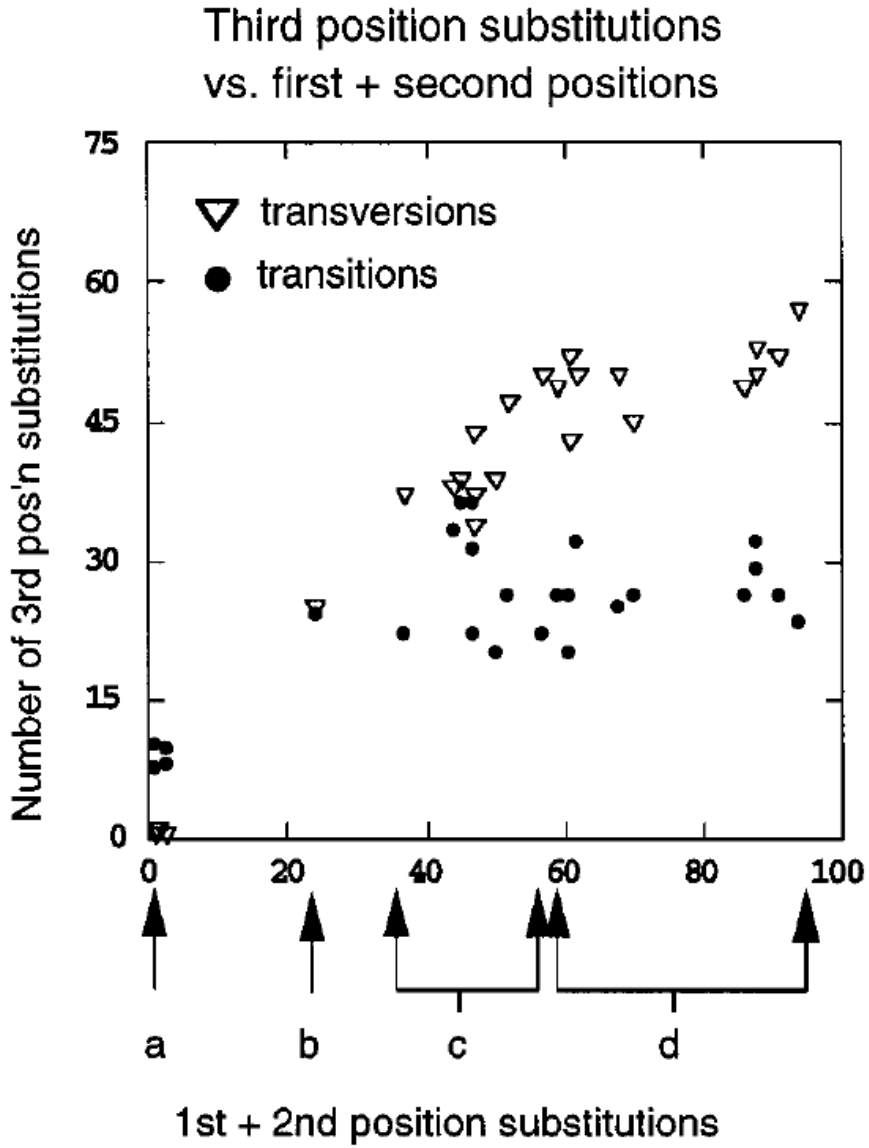
Two parameter models

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

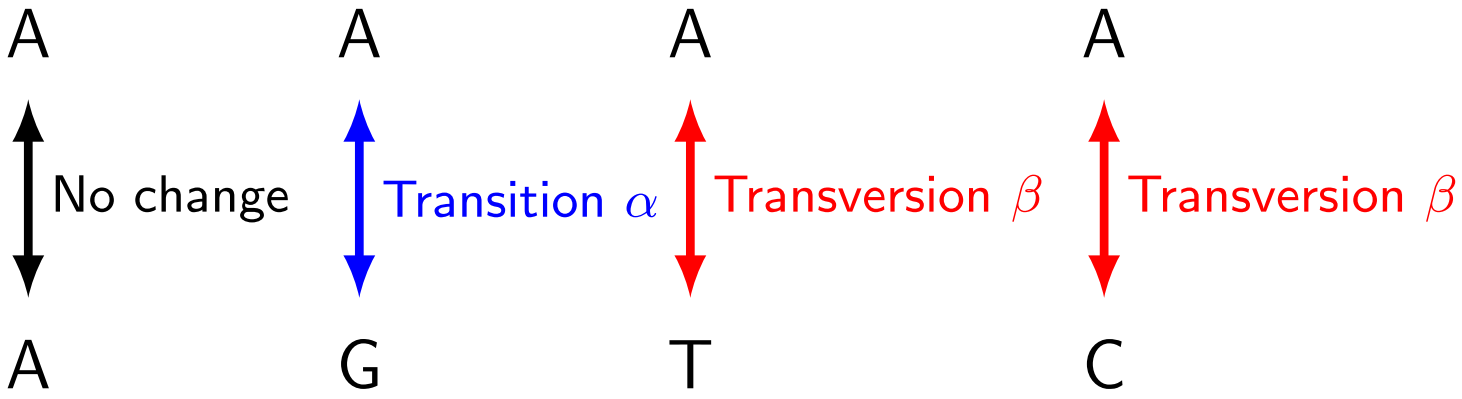
Kimura Two Parameters model

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	–

Two parameters



Prob nucleotides change over time

Let P_t^{X-Y} represent the probability of a path from X to Y at time t.
Then at time 0 if there is an A then $P_0^A = 1$. What happens over time?

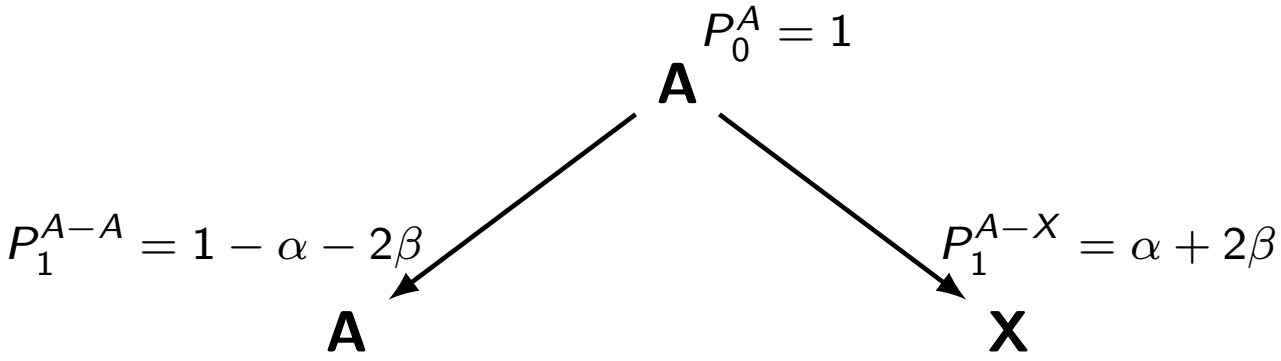
Prob nucleotides change over time

Let P_t^{X-Y} represent the probability of a path from X to Y at time t. Then at time 0 if there is an A then $P_0^A = 1$. What happens over time?

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

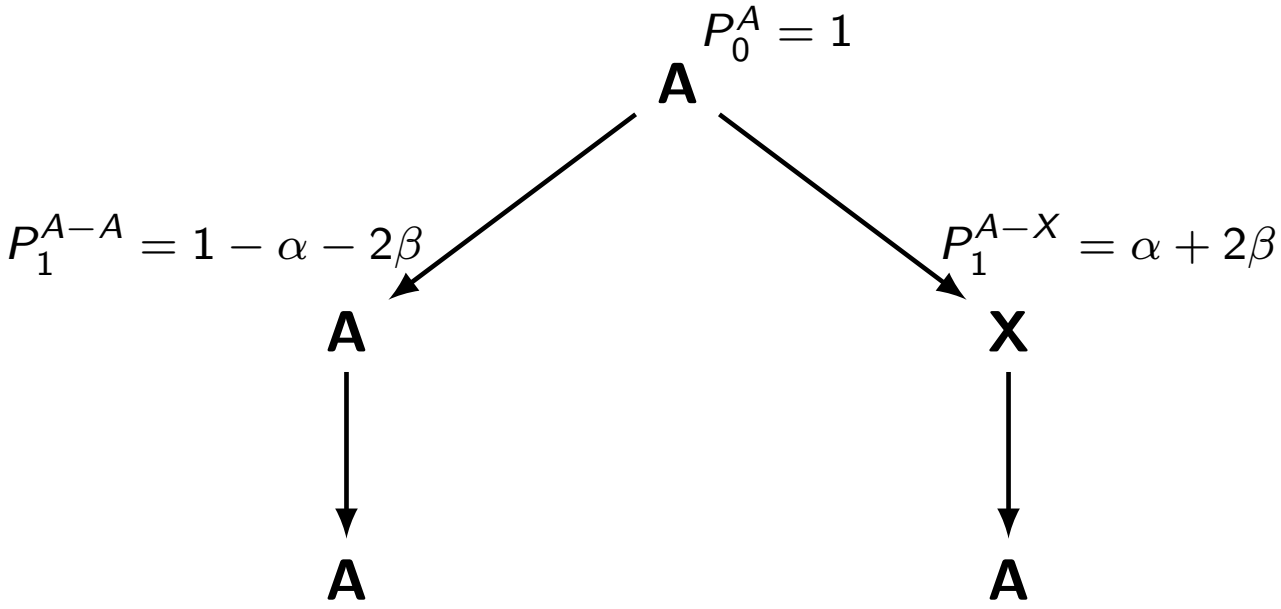
Prob nucleotides change over time

Let P_t^{X-Y} represent the probability of a path from X to Y at time t. Then at time 0 if there is an A then $P_0^A = 1$. What happens over time?



Prob nucleotides change over time

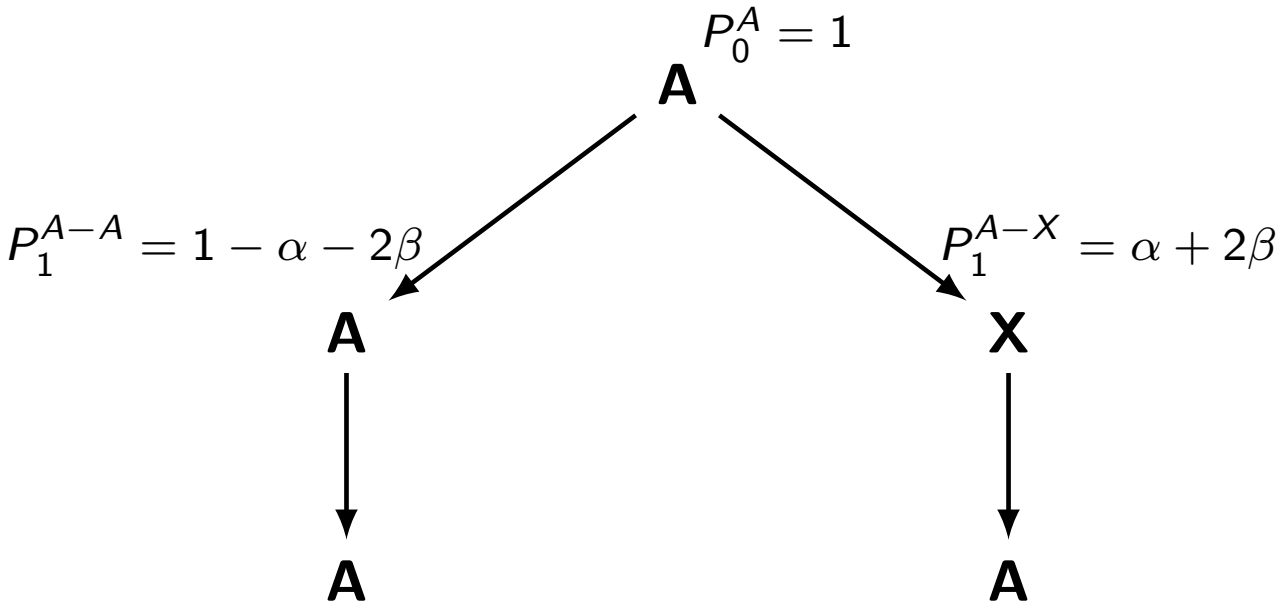
Let P_t^{X-Y} represent the probability of a path from X to Y at time t. Then at time 0 if there is an A then $P_0^A = 1$. What happens over time?



$$P_2^{A-A-A} = (1 - \alpha - 2\beta)P_1^{A-A} \qquad P_2^{A-X-A} = \alpha P_1^{A-G} + \beta P_1^{A-T} + \beta P_1^{A-C}$$

Prob nucleotides change over time

Let P_t^{X-Y} represent the probability of a path from X to Y at time t. Then at time 0 if there is an A then $P_0^A = 1$. What happens over time?



$$P_2^{A-A-A} = (1 - \alpha - 2\beta)P_1^{A-A} \qquad P_2^{A-X-A} = \alpha P_1^{A-G} + \beta P_1^{A-T} + \beta P_1^{A-C}$$

$$P_2^A = (1 - \alpha - 2\beta)P_1^A + \alpha P_1^G + \beta P_1^T + \beta P_1^C$$

Prob nucleotides change over time

So it is easier (though not necessary) to follow each individual nucleotide over time. So if we let P_t^X represent the probability of nucleotide X at time t.

Then ...

$$P_{t+1}^A = (1-\alpha-2\beta)P_t^A + \beta P_t^T + \beta P_t^C + \alpha P_t^G$$

Prob nucleotides change over time

So it is easier (though not necessary) to follow each individual nucleotide over time. So if we let P_t^X represent the probability of nucleotide X at time t.

Then ...

$$P_{t+1}^A = (1-\alpha-2\beta)P_t^A + \beta P_t^T + \beta P_t^C + \alpha P_t^G$$

$$P_{t+1}^T = \beta P_t^A + (1-\alpha-2\beta)P_t^T + \alpha P_t^C + \beta P_t^G$$

Prob nucleotides change over time

So it is easier (though not necessary) to follow each individual nucleotide over time. So if we let P_t^X represent the probability of nucleotide X at time t.

Then ...

$$
\begin{aligned}
P_{t+1}^A &= (1-\alpha-2\beta)P_t^A + \beta P_t^T + \beta P_t^C + \alpha P_t^G \\
P_{t+1}^T &= \beta P_t^A + (1-\alpha-2\beta)P_t^T + \alpha P_t^C + \beta P_t^G \\
P_{t+1}^C &= \beta P_t^A + \alpha P_t^T + (1-\alpha-2\beta)P_t^C + \beta P_t^G
\end{aligned}
$$

Prob nucleotides change over time

So it is easier (though not necessary) to follow each individual nucleotide over time. So if we let P_t^X represent the probability of nucleotide X at time t.

Then ...

$$
\begin{aligned}
P_{t+1}^A &= (1-\alpha-2\beta)P_t^A + \beta P_t^T + \beta P_t^C + \alpha P_t^G \\
P_{t+1}^T &= \beta P_t^A + (1-\alpha-2\beta)P_t^T + \alpha P_t^C + \beta P_t^G \\
P_{t+1}^C &= \beta P_t^A + \alpha P_t^T + (1-\alpha-2\beta)P_t^C + \beta P_t^G \\
P_{t+1}^G &= \alpha P_t^A + \beta P_t^T + \beta P_t^C + (1-\alpha-2\beta)P_t^G
\end{aligned}
$$

Prob nucleotides different between two sequences

This can be solved (simple differential equations) to get the following.

By transition (A/G, T/C)

$$P_t = \frac{1}{4} + \frac{1}{4}e^{-8\beta t} - \frac{1}{2}e^{-4(\alpha+\beta)t}$$

(where P_t are the transition differences)

By transversion (A/T, A/C, T/G, C/G)

$$Q_t = \frac{1}{2} - \frac{1}{2}e^{-8\beta t}$$

(and Q_t are the transversion differences)

So distance is the number of nucleotide substitutions per site:

$$D = 2\alpha t + 4\beta t = -\frac{1}{2} ln(1 - 2P - Q) - \frac{1}{4} ln(1 - 2Q)$$

$$Var(D) = [c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2]/n$$

with
$c_1 = 1/(1 - 2P - Q)$
$c_2 = 1/(1 - 2Q)$
$c_3 = (c_1 + c_2)/2$

Kimura two parameter

As an example, if you have a sequence, as before, of 100bp with 22 differences and 11 of these are transition differences and 11 are transversion differences. Then $P = 11/100, Q = 11/100$,

$$D_{2p} = -\frac{1}{2} ln(1 - 2P - Q) - \frac{1}{4} ln(1 - 2Q)$$

so $D_{2p} = 0.2624$ even though the Hamming distance is $D = 22/100 = 0.22$.

On the other hand, if 2 are transition differences and 20 are transversions, then $P = 2/100, Q = 20/100$, and so $D_{2p} = 0.2649$ even though the Hamming distance is $D = 22/100 = 0.22$ and the Jukes Cantor distance is $D_{JC} = 0.2604$.

Other models

Tamura's (1992) model:

- In Kimura's two parameter model the frequencies of each nucleotide are assumed equal (0.25).
- Tamura's method relaxes this assumption and lets some genomes have a biased GC content. Let the proportions of GC be $\theta_1 = \pi_G + \pi_C$ and of AT be $\theta_2 = \pi_A + \pi_T$.
- Then $D = -h \ln(1 - P/h - Q) - \frac{1}{2}(1 - h) \ln(1 - 2Q)$ with $h = 2\theta_1(1 - \theta_1)$

Other models

Tamura and Nei's (1993) model:

They noted that in mitochondria there were often different rates between $A \leftrightarrow G$ transitions and $T \leftrightarrow C$ transitions. In addition, they let the equilibrium frequency of each nucleotide be unique. So their model is

	A	T	C	G
A	–	$\beta\pi_T$	$\beta\pi_C$	$\alpha_1\pi_G$
T	$\beta\pi_A$	–	$\alpha_2\pi_C$	$\beta\pi_G$
C	$\beta\pi_A$	$\alpha_2\pi_T$	–	$\beta\pi_G$
G	$\alpha_1\pi_A$	$\beta\pi_T$	$\beta\pi_C$	–

Other models

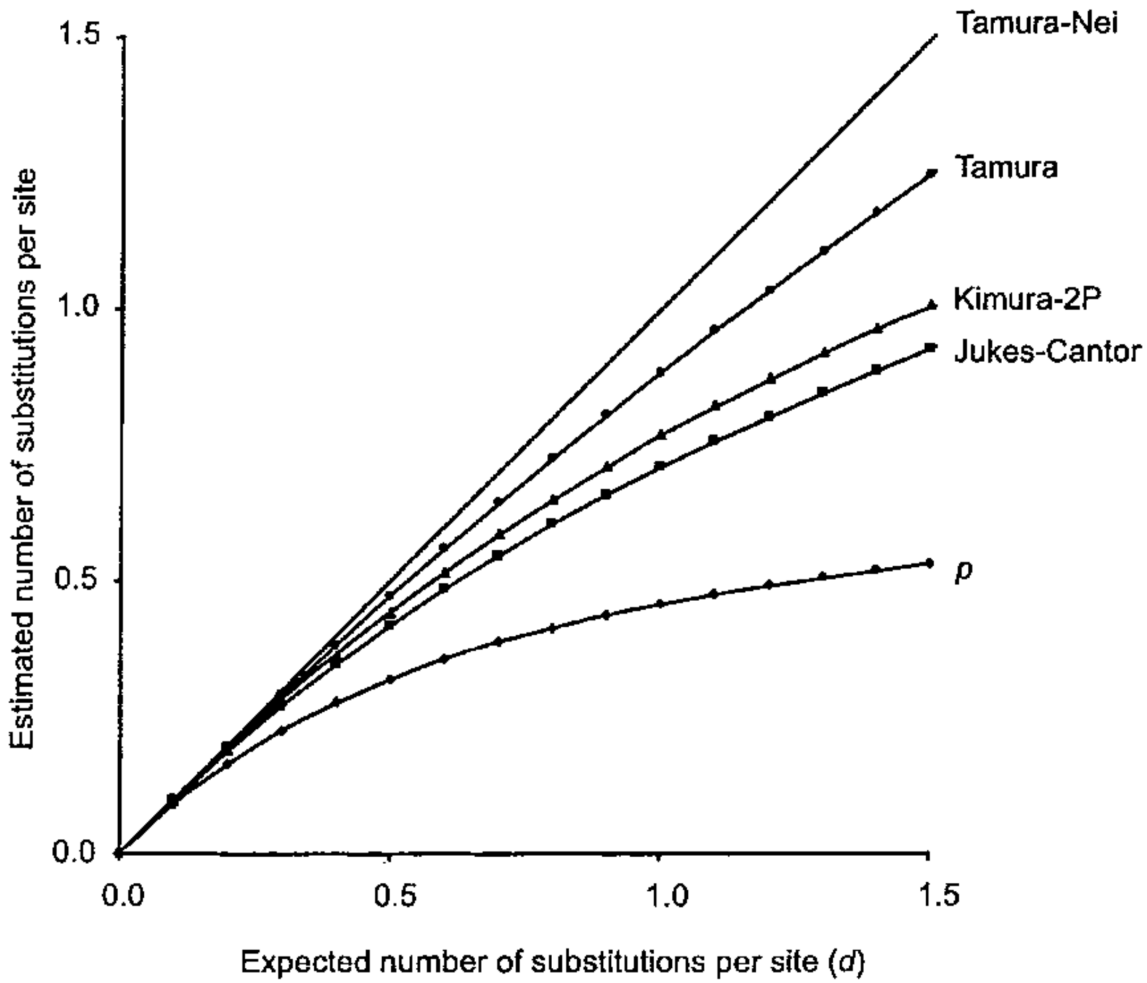
So Tamura and Nei's (1993) model gives:

$$
\begin{aligned}
D_{TN} = & -\frac{2\pi_A\pi_G}{\pi_R}\ln(1-\frac{\pi_R}{2\pi_A\pi_G}P_1-\frac{1}{2\pi_R}Q) \\
& -\frac{2\pi_T\pi_C}{\pi_Y}\ln(1-\frac{\pi_Y}{2\pi_T\pi_C}P_2-\frac{1}{2\pi_Y}Q) \\
& -2(\pi_R\pi_Y-\frac{\pi_A\pi_G\pi_Y}{\pi_R}-\frac{\pi_T\pi_C\pi_R}{\pi_Y})\ln(1-\frac{1}{2\pi_R\pi_Y}Q)
\end{aligned}
$$

with

P_1 the proportion of transitions between A and G
P_2 the proportion of transitions between T and C
Q the proportion of transversions.

Other models



Where $\pi_A = 0.3$, $\pi_T = 0.4$, $\pi_C = 0.2$, $\pi_G = 0.1$, $\alpha_1/\beta = 4$ and $\alpha_2/\beta = 8$.

From Nei and Kumar 2000.

Other models

Jukes Cantor model

	A	T	C	G
A	–	α	α	α
T	α	–	α	α
C	α	α	–	α
G	α	α	α	–

Kimura 2p model

	A	T	C	G
A	–	β	β	α
T	β	–	α	β
C	β	α	–	β
G	α	β	β	–

Tamura model

	A	T	C	G
A	–	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$
T	$\beta\theta_2$	–	$\alpha\theta_1$	$\beta\theta_1$
C	$\beta\theta_2$	$\alpha\theta_2$	–	$\beta\theta_1$
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	–

Tamura Nei model

	A	T	C	G
A	–	$\beta\pi_T$	$\beta\pi_C$	$\alpha_1\pi_G$
T	$\beta\pi_A$	–	$\alpha_2\pi_C$	$\beta\pi_G$
C	$\beta\pi_A$	$\alpha_2\pi_T$	–	$\beta\pi_G$
G	$\alpha_1\pi_A$	$\beta\pi_T$	$\beta\pi_C$	–

Other models

And the models get more complicated up to and including the GTR (General time reversible) that permits all rates and all frequencies to be different.

And the solutions become more complicated up to and including MLE (Maximum likelihood estimates) and Bayesian methods.

Other models

General Time Reversible - GTR

	A	T	C	G
A	–	$a\,\pi_T$	$b\,\pi_C$	$c\,\pi_G$
T	$a\,\pi_A$	–	$d\,\pi_C$	$e\,\pi_G$
C	$b\,\pi_A$	$d\,\pi_T$	–	$f\,\pi_G$
G	$c\,\pi_A$	$e\,\pi_T$	$f\,\pi_C$	–