Selection, mutation and drift in the Moran model

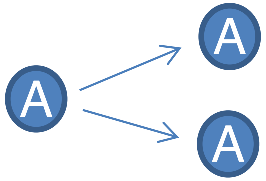
So far we looked at what happens when a mutant is introduced in a population.

What if mutations occur continuously during the evolution of the population?

Mutants don’t simply “take over” because they also mutate away.
Relevant would be to look at the dynamics of the proportion of mutants.

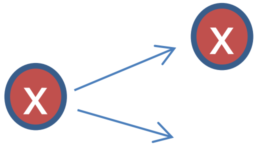
Selection, mutation and drift in the Moran model

An ‘A’ individual duplicates



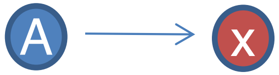
$\sigma(N-n)dt$

Another type duplicates



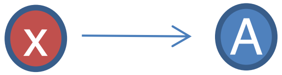
ndt

An A-type mutates



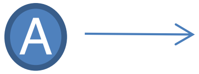
$\mu(N-n)dt$

Another type mutates

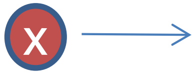


$\frac{\mu n}{3}dt$

At each duplication a randomly chosen individual is removed:



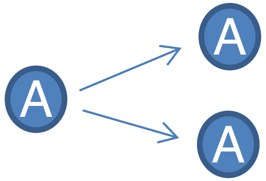
$\left(1-\frac{n}{N}\right)$



$\frac{n}{N}$

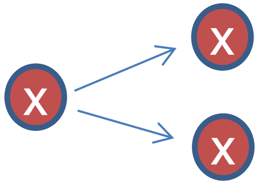
Or, in terms of $f=\frac{n}{N}$

An ‘A’ individual duplicates



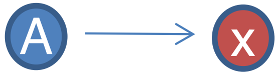
$N\sigma(1-f)dt$

Another type duplicates



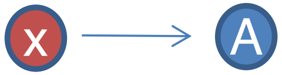
$Nfdt$

An A-type mutates



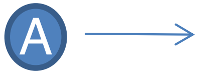
$N\mu(1-f)dt$

Another type mutates

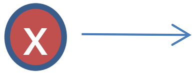


$\frac{\mu}{3}Nfdt$

At each duplication a randomly chosen individual is removed:



$(1-f)$

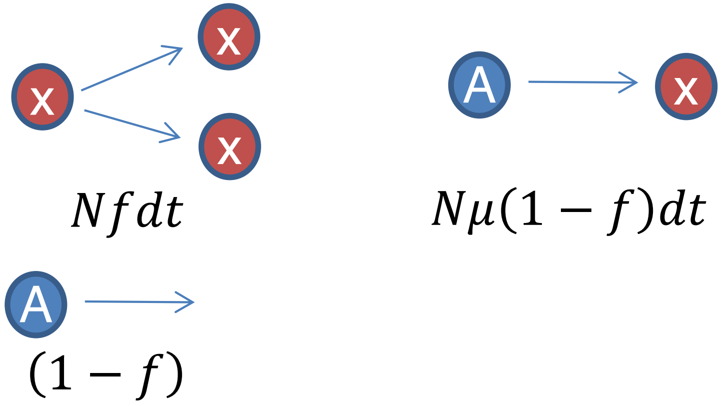


f

Mutants don’t simply “take over” because they also mutate away.
Relevant would be to look at the dynamics of the proportion of mutants.

Selection, mutation and drift in the Moran model

Scenarios that increase the mutant frequency by 1/*N*

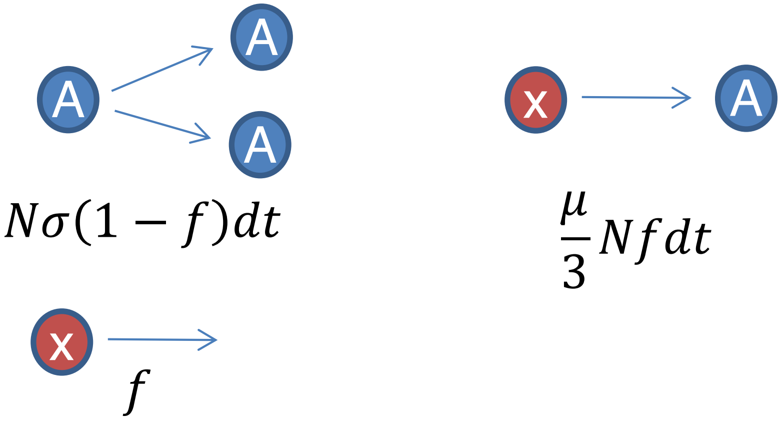


$Nf\,dt$

$N\mu(1-f)dt$

$(1-f)$

Scenarios that decrease the mutant frequency by 1/*N*



$N\sigma(1-f)dt$

$\frac{\mu}{3}Nf\,dt$

f

Probability of changing by 1 individual during time dt:

$$T\left(f,\delta f=-\frac{1}{N},dt\right)=N\left[\sigma f(1-f)+\frac{\mu}{3}f\right]dt$$

$$T\left(f,\delta f=+\frac{1}{N},dt\right)=N[f(1-f)+\mu(1-f)]dt$$

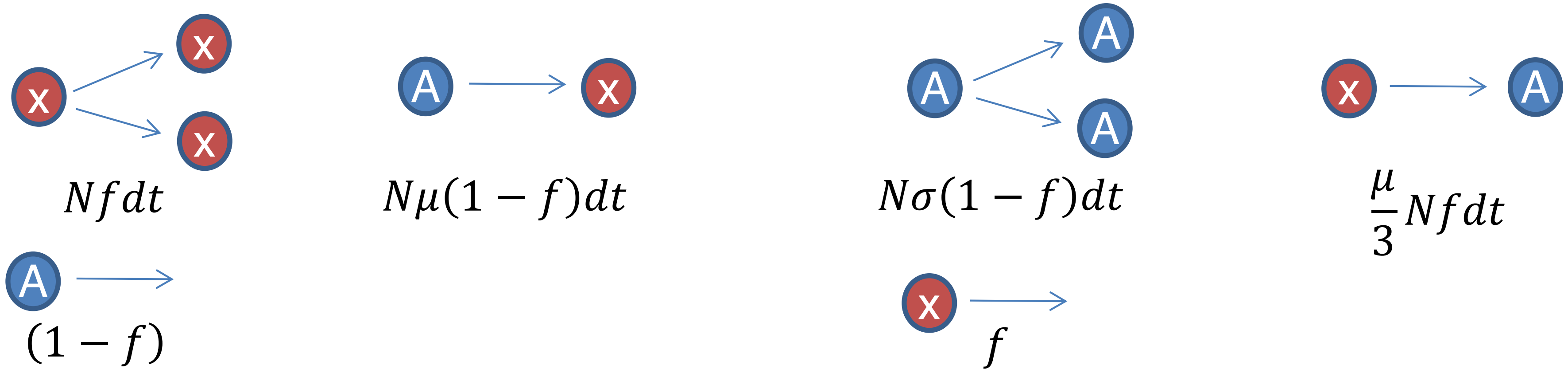
Thus $\langle\delta f\rangle_f=\left[f(1-f)+\mu(1-f)-\sigma f(1-f)-\frac{\mu}{3}f\right]dt$

$$=\left[(1-\sigma)f(1-f)+\mu\left(1-\frac{4f}{3}\right)\right]dt$$

Selection, mutation and drift in the Moran model

Scenarios that increase the mutant frequency by 1/*N*

Scenarios that decrease the mutant frequency by 1/*N*



$Nfdt$

$N\mu(1-f)dt$

$N\sigma(1-f)dt$

$\frac{\mu}{3}Nfdt$

$(1-f)$

f

Probability of changing by 1 individual during time dt:

$$T\left(f, \delta f = -\frac{1}{N}, dt\right) = N\left[\sigma f(1-f) + \frac{\mu}{3}f\right]dt$$

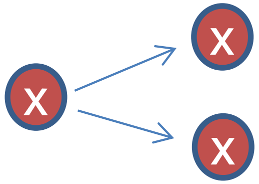
$$T\left(f, \delta f = +\frac{1}{N}, dt\right) = N[f(1-f) + \mu(1-f)]dt$$

Thus $\langle(\delta f)^2\rangle_f \; = \frac{1}{N}\left[f(1-f) + \mu(1-f) + \sigma f(1-f) + \frac{\mu}{3}f\right]dt$

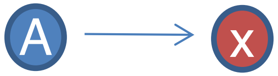
$$= \left[(1+\sigma)f(1-f) + \mu\left(1 - \frac{2f}{3}\right)\right]dt$$

Selection, mutation and drift in the Moran model

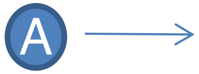
Scenarios that increase the mutant frequency by 1/*N*



$Nf\,dt$

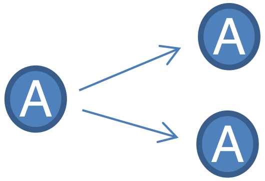


$N\mu(1-f)dt$

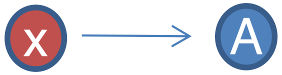


$(1-f)$

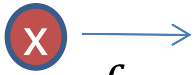
Scenarios that decrease the mutant frequency by 1/*N*



$N\sigma(1-f)dt$



$\frac{\mu}{3}Nf\,dt$



f

$$T\left(f,\delta f=-\frac{1}{N},dt\right)=N\left[\sigma f(1-f)+\frac{\mu}{3}f\right]dt$$

$$T\left(f,\delta f=+\frac{1}{N},dt\right)=N[f(1-f)+\mu(1-f)]dt$$

Let's study the *stochastic* dynamics of f using again a diffusion approximation.

$$\begin{aligned}\langle\delta f\rangle_f &=\left[f(1-f)+\mu(1-f)-\sigma f(1-f)-\tfrac{\mu}{3}f\right]dt\\ &=\left[(1-\sigma)f(1-f)+\mu\left(1-\frac{4f}{3}\right)\right]dt\end{aligned}$$

$$\begin{aligned}\langle(\delta f)^2\rangle_f &=\frac{1}{N}\left[f(1-f)+\mu(1-f)+\sigma f(1-f)+\tfrac{\mu}{3}f\right]dt\\ &=\left[(1+\sigma)f(1-f)+\mu\left(1-\frac{2f}{3}\right)\right]dt\end{aligned}$$

Diffusion model

Definitions: f = Fraction of the population with the mutant genotype

$P(f,t)$ = Probability that the mutant represents fraction f population at time t

$T(f,\delta f,dt)$ = Probability that the fraction changes from f to $f+\delta f$ in a small time interval dt

Diffusion model

Definitions: f = Fraction of the population with the mutant genotype

$P(f,t)$ = Probability that the mutant represents fraction f population at time t

$T(f,\delta f,dt)$ = Probability that the fraction changes from f to $f+\delta f$ in a small time interval dt

Master equation: $$P(f,t+dt) = \int T(f-\delta f,\delta f,dt)P(f-\delta f,t)d(\delta f)$$

Let's call $f-\delta f = y$ and $\delta f = x$, and rewrite $P(f,t+dt) = \int T(y,x,dt)P(y,t)dx$

Which says that the probability to have a mutant fraction *f* at the time *t*+*dt* is given by the integral, over all possible changes *x* of the mutant fraction, of the probability that the fraction changes from *y* = *f* – *x* to *f* in time *dt* times the probability that the mutant fraction was *y* at time *t*.

Diffusion model

Definitions: f = Fraction of the population with the mutant genotype

$P(f,t)$ = Probability that the mutant represents fraction f
population at time t

$T(f,\delta f,dt)$ = Probability that the fraction changes from
f to $f+\delta f$ in a small time interval dt

$$P(f,t+dt)=\int T(y,x,dt)P(y,t)dx$$

We then expand the right-hand side around $y=f$:

$$\int\left[T(f,x,dt)P(f,t)+(y-f)\frac{\partial}{\partial y}\big(T(y,x,dt)P(y,dt)\big)\Big|_{y=f}+\frac{(y-f)^2}{2}\frac{\partial^2}{\partial y^2}\big(T(y,x,dt)P(y,dt)\big)\Big|_{y=f}\right]dx$$

$$\int T(f,x,dt)P(f,t)dx=P(f,t)\int T(f,x,dt)dx=P(f,t)$$

$$\int(y-f)\frac{\partial}{\partial y}\big(T(y,x,dt)P(y,dt)\big)\Big|_{y=f}dx=-\frac{\partial}{\partial y}\left(\int xT(y,x,dt)dx\right)P(y,dt)\Bigg|_{y=f}$$

$$=-\frac{\partial}{\partial y}\left[\langle\delta f\rangle_y P(y,dt)\right]\Big|_{y=f}$$

$$\int\frac{(y-f)^2}{2}\frac{\partial^2}{\partial y^2}\big(T(y,x,dt)P(y,dt)\big)\Big|_{y=f}dx=\frac{1}{2}\frac{\partial^2}{\partial y^2}\left[\langle(\delta f)^2\rangle_y P(y,dt)\right]\Big|_{y=f}$$

Diffusion model

Putting it all together

$$P(f, t + dt) = \int T(y, x, dt) P(y, t) dx =$$

$$P(f, t) - \frac{\partial}{\partial y}\left[\langle \delta f \rangle_y P(y, dt)\right]\Big|_{y=f} + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left[\langle (\delta f)^2 \rangle_y P(y, dt)\right]\Big|_{y=f}$$

And rearranging

$$\frac{\partial P(f,t)}{\partial t} = -\frac{\partial}{\partial y}\left[M_{\delta f}(y) P(y, dt)\right]\Big|_{y=f} + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left[V_{\delta f}(y) P(y, dt)\right]\Big|_{y=f}$$

With

$$M_{\delta f}(y) = \frac{\langle \delta f \rangle_y}{dt} \qquad V_{\delta f}(y) = \frac{\langle (\delta f)^2 \rangle_y}{dt}$$

This is the diffusion equation of gene-frequency change in population genetics.

We now formally solve for its *steady-state* solution.

Selection, mutation and drift at a single site

At steady-state we have: $\frac{\partial P(f,t)}{\partial t} = -\frac{\partial}{\partial y}\left[M_{\delta f}(y)P(y,t)\right]_{y=f} + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left[V_{\delta f}(y)P(y,t)\right]_{y=f} = 0$

If we define: $U(f) = -M_{\delta f}(f)P(f,t) + \frac{1}{2}\frac{d}{df}\left[V_{\delta f}(f)P(f,t)\right]$

The steady-state equation becomes: $\frac{dU(f)}{df} = 0 \Rightarrow U(f) =$ constant

It can be shown that the solution we want has: $U(f) = 0$

or $M_{\delta f}(f)P(f,t) = \frac{1}{2}\frac{d}{df}\left[V_{\delta f}(f)P(f,t)\right]$ Defining: $V_{\delta f}(f)P(f,t) = X(f)$

we have: $2\frac{M_{\delta f}(f)}{V_{\delta f}(f)}X(f) = \frac{d}{df}\left[X(f)\right] \Leftrightarrow \log\left[X(f)\right] = C + 2\int \frac{M_{\delta f}(f)}{V_{\delta f}(f)}df$

and we find that the steady-state solution is:

$$P(f) = \frac{C}{V_{\delta f}(f)}\exp\left[2\int \frac{M_{\delta f}(f)}{V_{\delta f}(f)}df\right]$$

Selection, mutation and drift at a single site

For the Moran model we had:

$$M_{\delta f} = \frac{\left\langle \delta f \right\rangle_f}{dt} = \left[(1-\sigma)f(1-f)+\mu\left(1-\frac{4f}{3}\right)\right], \quad V_{\delta f} = \frac{\left\langle (\delta f)^2 \right\rangle_f}{dt} = \frac{1}{N}\left[(1+\sigma)f(1-f)+\mu\left(1-\frac{2f}{3}\right)\right]$$

$$\frac{M_{\delta f}}{V_{\delta f}} = N\frac{(1-\sigma)f(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(1+\sigma)f(1-f)+\mu\left(1-\frac{2f}{3}\right)} = N\frac{-sf(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)}$$

Although we could directly substitute this into: $P(f) = \frac{C}{V_{\delta f}(f)}\exp\left[2\int\frac{M_{\delta f}(f)}{V_{\delta f}(f)}df\right]$

it is pedagogically more instructive to take the limit in which both μ and $s = \sigma\text{-}1$ are small.

Selection, mutation and drift at a single site

$$\frac{M_{\delta f}}{V_{\delta f}} = N\frac{(1-\sigma)f(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(1+\sigma)f(1-f)+\mu\left(1-\frac{2f}{3}\right)} = N\frac{-sf(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)}$$

Expanding to first order in s and μ we get:

$$F(s,\mu) = F(0,0) + s\left.\frac{\partial F(s,\mu)}{\partial s}\right|_{(0,0)} + \mu\left.\frac{\partial F(s,\mu)}{\partial \mu}\right|_{(0,0)}$$

$$\frac{\partial}{\partial s}\left[N\frac{-sf(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)}\right]$$

$$\text{At } (0,0)\quad \frac{-Nf(1-f)\big(2f(1-f)\big)}{\big(2f(1-f)\big)^2} = -\frac{N}{2}$$

$$= \frac{-Nf(1-f)\left((2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)\right) - f(1-f)N\left(-sf(1-f)+\mu\left(1-\frac{4f}{3}\right)\right)}{\left((2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)\right)^2}$$

Selection, mutation and drift at a single site

$$\frac{M_{\delta f}}{V_{\delta f}} = N\frac{(1-\sigma)f(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(1+\sigma)f(1-f)+\mu\left(1-\frac{2f}{3}\right)} = N\frac{-sf(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)}$$

Expanding to first order in s and μ we get:

$$F(s,\mu) = F(0,0) + s\left.\frac{\partial F(s,\mu)}{\partial s}\right|_{(0,0)} + \mu\left.\frac{\partial F(s,\mu)}{\partial \mu}\right|_{(0,0)}$$

$$\frac{\partial}{\partial \mu}\left[N\frac{-sf(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)}\right]$$

$$\text{At } (0,0) \quad \frac{N\left(1-\frac{4f}{3}\right)\left(2f(1-f)\right)}{\left(2f(1-f)\right)^2} = \frac{N\left(1-\frac{4f}{3}\right)}{2f(1-f)}$$

$$= \frac{N\left(1-\frac{4f}{3}\right)\left((2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)\right)-\left(1-\frac{2f}{3}\right)N\left(-sf(1-f)+\mu\left(1-\frac{4f}{3}\right)\right)}{\left((2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)\right)^2}$$

Selection, mutation and drift at a single site

$$\frac{M_{\delta f}}{V_{\delta f}} = N\frac{(1-\sigma)f(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(1+\sigma)f(1-f)+\mu\left(1-\frac{2f}{3}\right)} = N\frac{-sf(1-f)+\mu\left(1-\frac{4f}{3}\right)}{(2+s)f(1-f)+\mu\left(1-\frac{2f}{3}\right)}$$

Expanding to first order in s and μ we got:

$$\frac{M_{\delta f}(f)}{V_{\delta f}(f)} \approx -\frac{Ns}{2}+\frac{N\mu}{2}\left[\frac{1}{f}-\frac{1}{3\left(1-f\right)}\right]$$

We now substitute this into the steady-state equation

$$P(f)=\frac{C}{V_{\delta f}(f)}\exp\left[2\int\frac{M_{\delta f}(f)}{V_{\delta f}(f)}df\right]$$

Selection, mutation and drift at a single site

$$\frac{M_{\delta f}(f)}{V_{\delta f}(f)} \approx -\frac{Ns}{2} + \frac{N\mu}{2}\left[\frac{1}{f} - \frac{1}{3(1-f)}\right] \qquad P(f) = \frac{C}{V_{\delta f}(f)}\exp\left[2\int \frac{M_{\delta f}(f)}{V_{\delta f}(f)}df\right]$$

$$2\int \frac{M_{\delta f}(f)}{V_{\delta f}(f)}df = \int\left[-Ns + N\mu\left(\frac{1}{f} - \frac{1}{3(1-f)}\right)\right]df = -Nsf + N\mu\log(f) + \frac{N\mu}{3}\log(1-f)$$

$$P(f) = \frac{C}{V_{\delta f}(f)}\exp\left[2\int \frac{M_{\delta f}(f)}{V_{\delta f}(f)}df\right] = \frac{C}{V_{\delta f}(f)}\exp\left[-Nsf + N\mu\log(f) + \frac{N\mu}{3}\log(1-f)\right]$$

We obtain $$P(f) \approx \frac{C}{V_{\delta f}}e^{-Nsf}f^{N\mu}\left(1-f\right)^{\frac{N\mu}{3}}$$

This is a function only of the products *Ns* and *Nμ,* the former quantifying the balance between selection and drift, and the latter the balance between mutation and drift.

Evolution with finite populations

Recap

Finite populations: selection-substitution dynamics

All members of a population in which individuals simply replicate have a common ancestor on average 2*N* generations in the past (genetic drift).

Probability of fixation of a mutant that starts at a frequency *f* in the population, given that its frequency changes by *δf* in a small interval *dt*:

$$\pi\left(\frac{1}{N}\right) = \frac{1 - e^{\frac{2s}{2+s}}}{1 - e^{\frac{2Ns}{2+s}}}, \text{ with } s = \sigma - 1$$

The total probability that within a generation a mutant will arise that will take over the population is*:*

$$N\mu\pi \approx N\mu \frac{1 - e^{\frac{2s}{2+s}}}{1 - e^{\frac{2Ns}{2+s}}}$$

In the limit of s->0 (neutral mutations), $N\mu\pi \approx \mu$*.* That is, the probability that a neutral mutation will arise and will take over the population is *μ*.

Adaptive mutations fix at a rate that depends strongly on population size, i.e. the bigger the population, the more adaptive mutations fix. Conversely, the smaller the population size, the higher the chance of disadvantageous mutants to take over.

Selection, mutation and drift at a single site

Steady-state distribution of the frequency of mutant:

$$P(f) \approx \frac{C}{V_{\delta f}} e^{-Nsf} f^{N\mu} \left(1 - f\right)^{\frac{N\mu}{3}}$$

Selection, mutation and drift at a single site

$P(f) \approx Ce^{-Nsf} f^{N\mu - 1} (1-f)^{\frac{N\mu}{3} - 1}$ $s = \sigma - 1$ is the fitness advantage of the wildtype

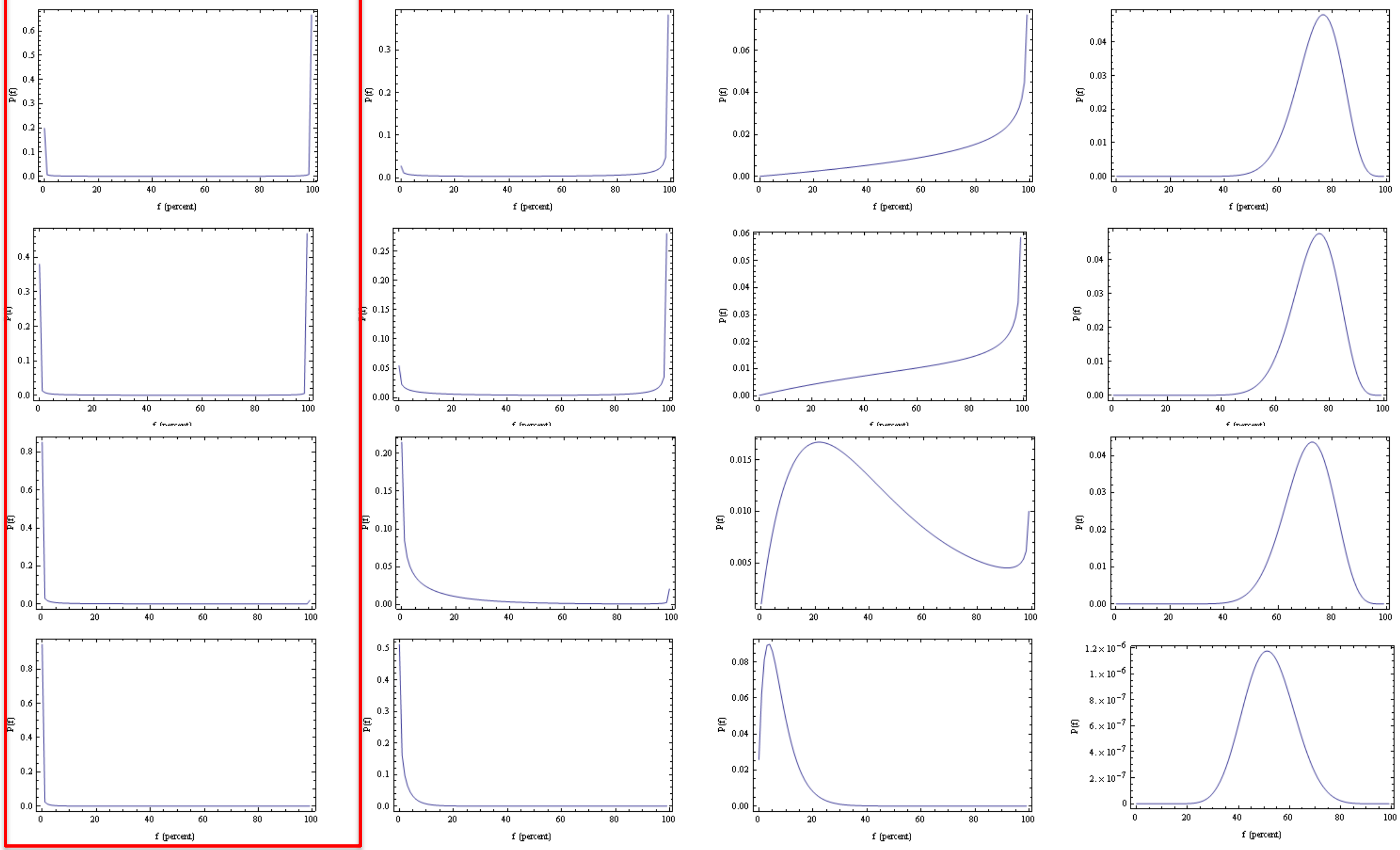
$N\mu = 0.05$ $N\mu = 0.5$ $N\mu = 2$ $N\mu = 20$

$Ns = 0$

$Ns = 1$

$Ns = 5$

$Ns = 25$



Qualitative Summary:
Drift-mutation and Drift-selection balance

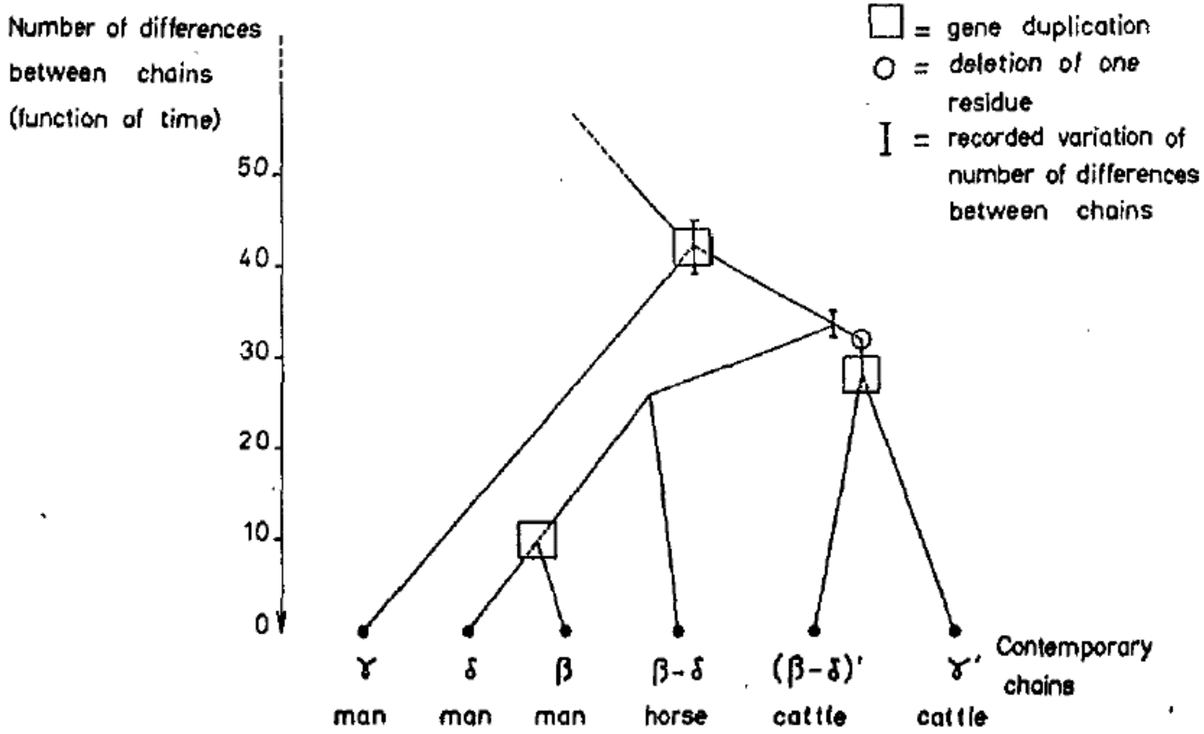
Population consists of copies of the same high fitness type at any point in time.	Population consists of a mixture of high fitness types and a small fraction of low fitness mutants.
Population consists of copies of the same type at any point in time. This type can be either high or low fitness (mutant or wildtype).	Population consists of a mixture of both lower and higher fitness types.

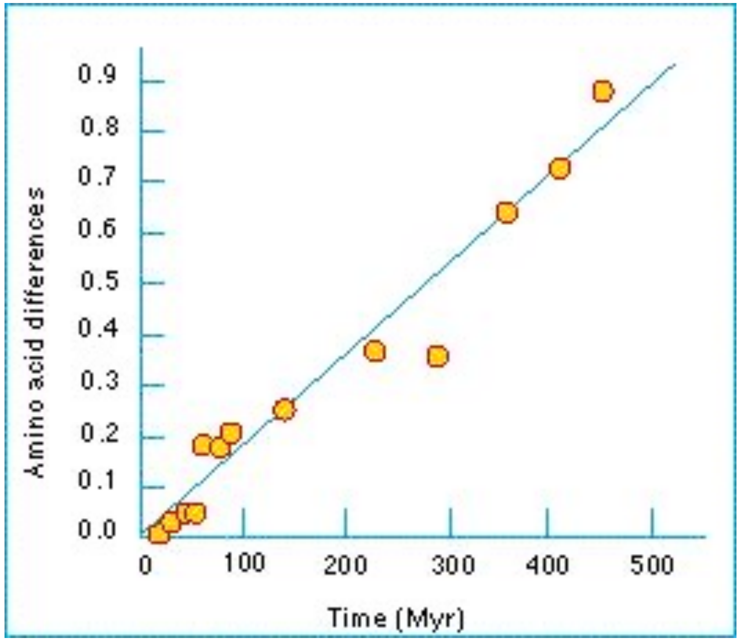
Ns 1

1

Nμ

The First Molecular Data


Number of differences between chains (function of time)
50
40
30
20
10
0
□ = gene duplication
○ = deletion of one residue
I = recorded variation of number of differences between chains
γ man
δ man
β man
β-δ horse
(β-δ)' cattle
γ' cattle
Contemporary chains



Pauling and Zuckerkandl (1965)

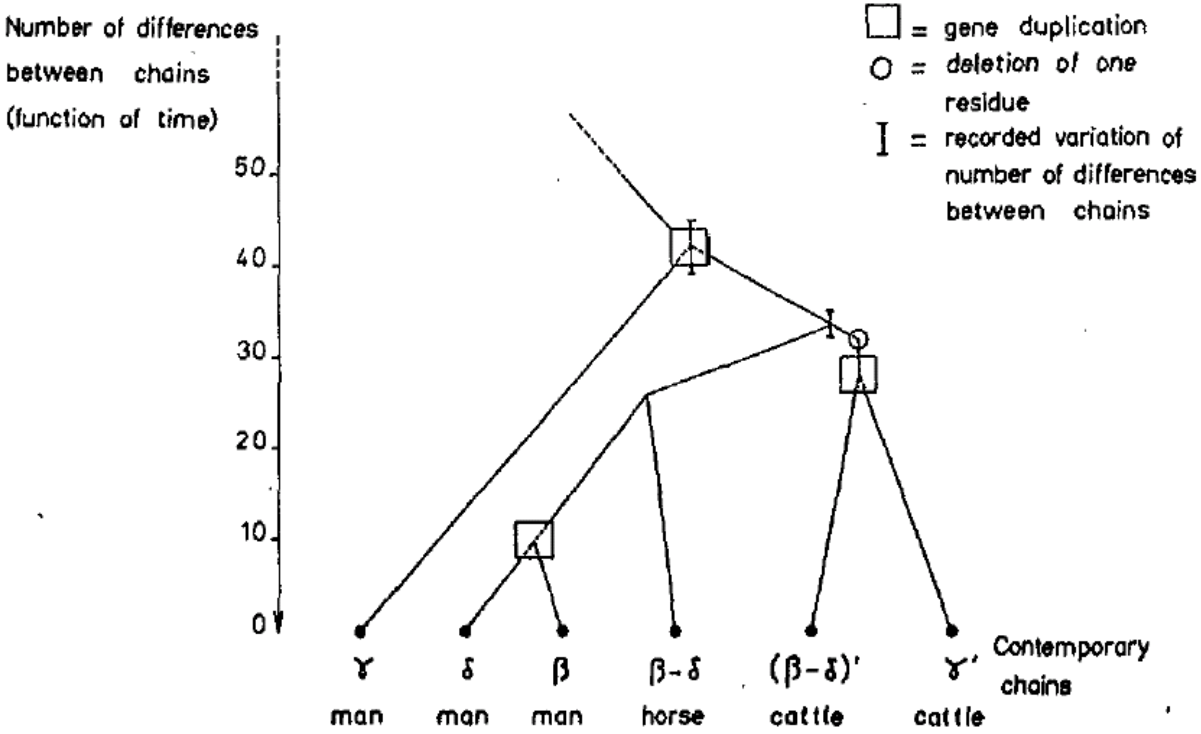
- Substantial numbers of amino acid differences between hemoglobin proteins in different mammals.
- The number of differences roughly proportional to the evolutionary distance estimated from the fossil record (molecular clock).

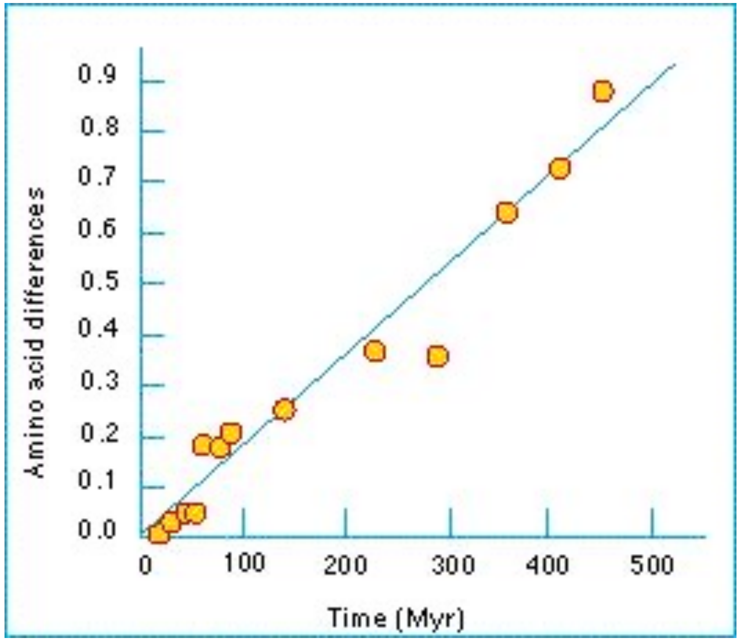
Substitution dynamics

The total probability that within a generation a (neutral, s ≈ 0) mutant will arise that will take over the population is*:*

$$N\mu\pi \approx N\mu \frac{1-e^{\frac{2s}{2+s}}}{1-e^{\frac{2Ns}{2+s}}} \approx N\mu \frac{-2s}{-2Ns} \approx \mu$$

The First Molecular Data

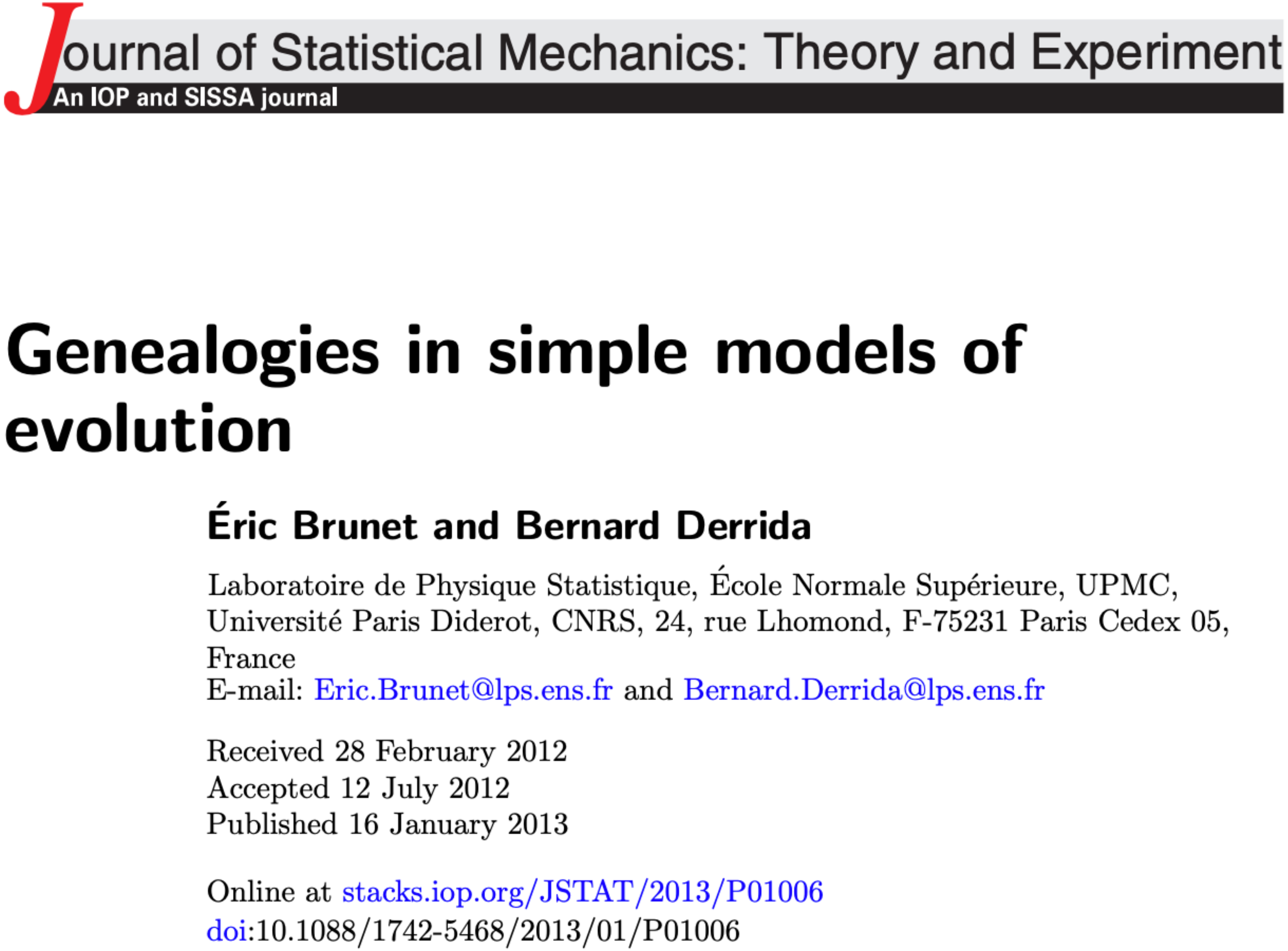

Number of differences between chains (function of time)
50
40
30
20
10
0
□ = gene duplication
○ = deletion of one residue
I = recorded variation of number of differences between chains
γ man
δ man
β man
β-δ horse
(β-δ)' cattle
γ' cattle
Contemporary chains



To explain the molecular data Kimura proposed in 1968 that the vast majority of single nucleotide changes are selectively *neutral.*

The molecular clock: most mutations are neutral, and will fix at a roughly constant rate (μ). The differences between species (substitutions) are the mutations that fixated (spread in the population), not all the mutations that took place.

How about sexually reproducing populations?



Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Genealogies in simple models of evolution

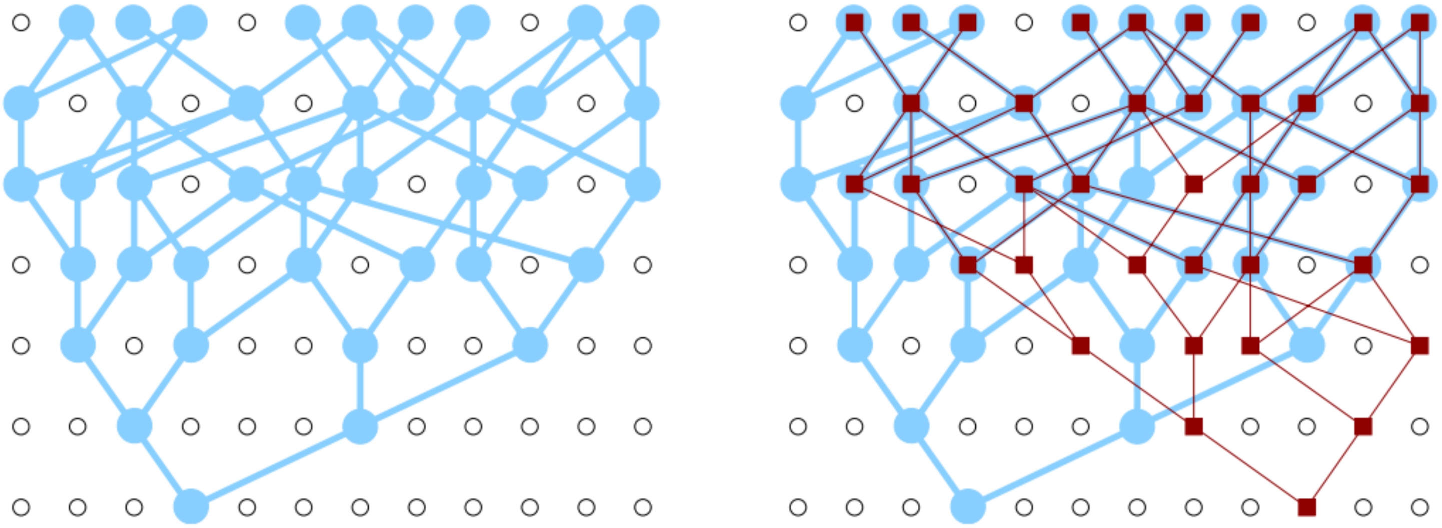
Éric Brunet and Bernard Derrida

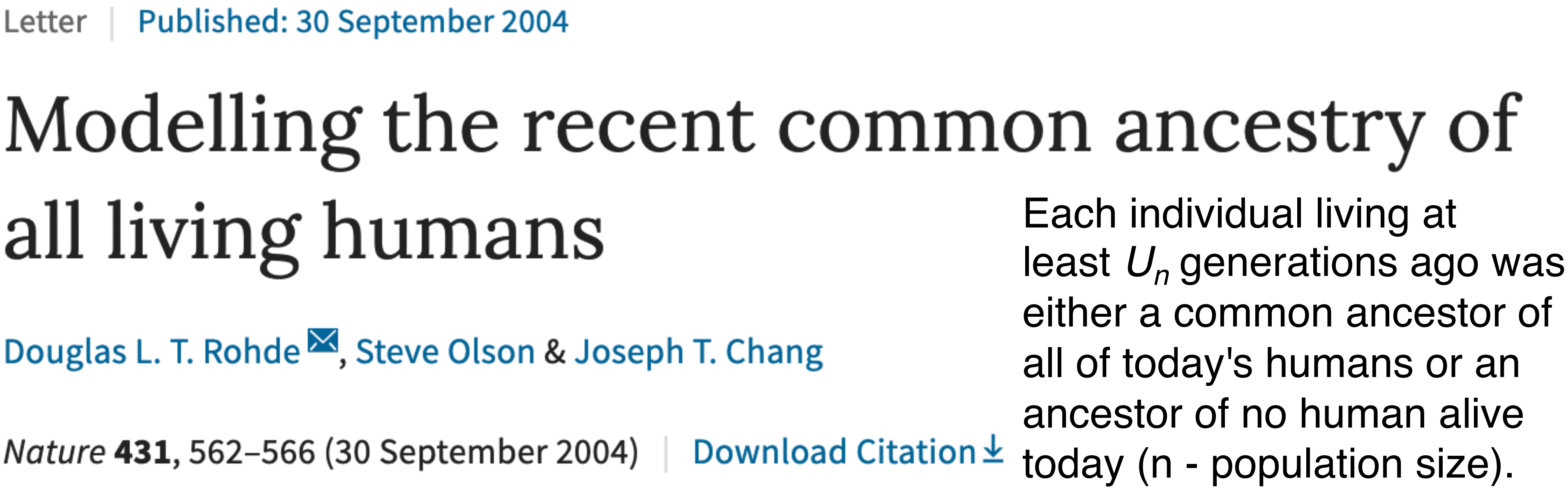
Laboratoire de Physique Statistique, École Normale Supérieure, UPMC, Université Paris Diderot, CNRS, 24, rue Lhomond, F-75231 Paris Cedex 05, France
E-mail: Eric.Brunet@lps.ens.fr and Bernard.Derrida@lps.ens.fr

Received 28 February 2012
Accepted 12 July 2012
Published 16 January 2013

Online at stacks.iop.org/JSTAT/2013/P01006
doi:10.1088/1742-5468/2013/01/P01006

How about sexually reproducing populations?





Letter | Published: 30 September 2004

Modelling the recent common ancestry of all living humans

Douglas L. T. Rohde, Steve Olson & Joseph T. Chang

Nature **431**, 562–566 (30 September 2004) | Download Citation

Each individual living at least U_n generations ago was either a common ancestor of all of today's humans or an ancestor of no human alive today (n - population size).

Up to now we discussed “forward” evolution of finite populations.

However, the problem that we usually want to solve is to reconstruct the evolutionary scenario from current day sequence data.

Reconstructing phylogenies

- Substitution models

- Pairwise distances

- Likelihood of a phylogenetic tree: Felsenstein’s algorithm

- Reconstructing phylogenetic trees:
    - Maximum likelihood
    - Neighbor-joining

Substitution models

Neutral evolution of a letter in the DNA: Jukes-Cantor model

- Under neutral evolution, a single base is substituted with another base at a rate that simply equals the mutation rate μ.

- Let $P(\alpha|\beta,t)$ denote the probability to evolve from letter β to letter α in a time t

- We want to compute this probability as a function of time t and mutation rate.

Neutral evolution of a letter in the DNA: Jukes-Cantor model

$P(\alpha|\beta,t)$ - probability to evolve from letter β to letter α in time t

$$P(\alpha|\beta,t+dt) = (1-\mu\, dt)P(\alpha|\beta,t) + \textstyle\sum_{\gamma\neq\alpha}\frac{\mu}{3}\, dt\; P(\gamma|\beta,t)$$

$$= P(\alpha|\beta,t) - \mu\, dt\; P(\alpha|\beta,t) + \textstyle\sum_{\gamma\neq\alpha}\frac{\mu}{3}\, dt\; P(\gamma|\beta,t)$$

$$P(\alpha|\beta,t+dt) - P(\alpha|\beta,t) = -\mu\, dt\; P(\alpha|\beta,t) + \textstyle\sum_{\gamma\neq\alpha}\frac{\mu}{3}\, dt\; P(\gamma|\beta,t)$$

$$\frac{\partial P(\alpha|\beta,t)}{\partial t} = -\mu\, P(\alpha|\beta,t) + \sum_{\gamma\neq\alpha}\frac{\mu}{3}\; P(\gamma|\beta,t) = -\mu\, P(\alpha|\beta,t) + \frac{\mu}{3}\big(1 - P(\alpha|\beta,t)\big)$$

$$\frac{\partial P(\alpha|\beta,t)}{\partial t} = \frac{\mu}{3} - \frac{4\mu}{3}\, P(\alpha|\beta,t)$$

Neutral evolution of a letter in the DNA: Jukes-Cantor model

$P(\alpha|\beta,t)$ - probability to evolve from letter β to letter α in time t

$$\frac{\partial P(\alpha|\beta,t)}{\partial t} = \frac{\mu}{3} - \frac{4\mu}{3} P(\alpha|\beta,t)$$

$$\frac{\partial P(\alpha|\beta,t)}{\partial t} e^{\frac{4\mu t}{3}} = \frac{\mu}{3} e^{\frac{4\mu t}{3}} - \frac{4\mu}{3} P(\alpha|\beta,t) e^{\frac{4\mu t}{3}}$$

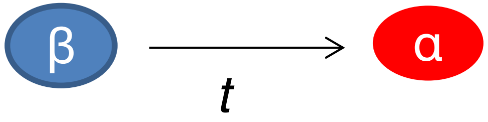
$$\frac{\partial P(\alpha|\beta,t)}{\partial t} e^{\frac{4\mu t}{3}} + \frac{4\mu}{3} P(\alpha|\beta,t) e^{\frac{4\mu t}{3}} = \frac{\mu}{3} e^{\frac{4\mu t}{3}}$$

$$\frac{\partial}{\partial t}\left(P(\alpha|\beta,t) e^{\frac{4\mu t}{3}}\right) = \frac{\mu}{3} e^{\frac{4\mu t}{3}}$$

$$P(\alpha|\beta,t) e^{\frac{4\mu t}{3}} = \frac{\mu}{3}\frac{3}{4\mu} e^{\frac{4\mu t}{3}} + C$$

$$P(\alpha|\beta,t) = \frac{1}{4} + Ce^{-\frac{4\mu t}{3}}$$

Jukes-Cantor model (1969)



$$P(\alpha|\beta,t) = \frac{1}{4} + Ce^{-\frac{4\mu t}{3}}$$

We have the boundary condition: $P(\alpha|\beta,0) = \delta_{\alpha\beta}$ where $\delta_{\alpha\beta} = \begin{cases} 0 \; if \; \alpha \neq \beta \\ 1 \; if \; \alpha = \beta \end{cases}$

from which we can determine *C* and obtain:

$$C = \delta_{\alpha\beta} - \frac{1}{4} \text{ and } P(\alpha|\beta,t) = \frac{1}{4} + \left(\delta_{\alpha\beta} - \frac{1}{4}\right) e^{-\frac{4\mu t}{3}}$$

The probability that base β has changed into another (has been substituted) as a function of time is

$$P(\bar{\beta}|\beta,t) = \frac{3}{4}\left(1 - e^{-\frac{4\mu t}{3}}\right)$$

It takes in average $\frac{3}{4\mu}$ generations before a base is substituted.

In Lenski’s experiment there are about 7 generations per day, and about 2000 per year.

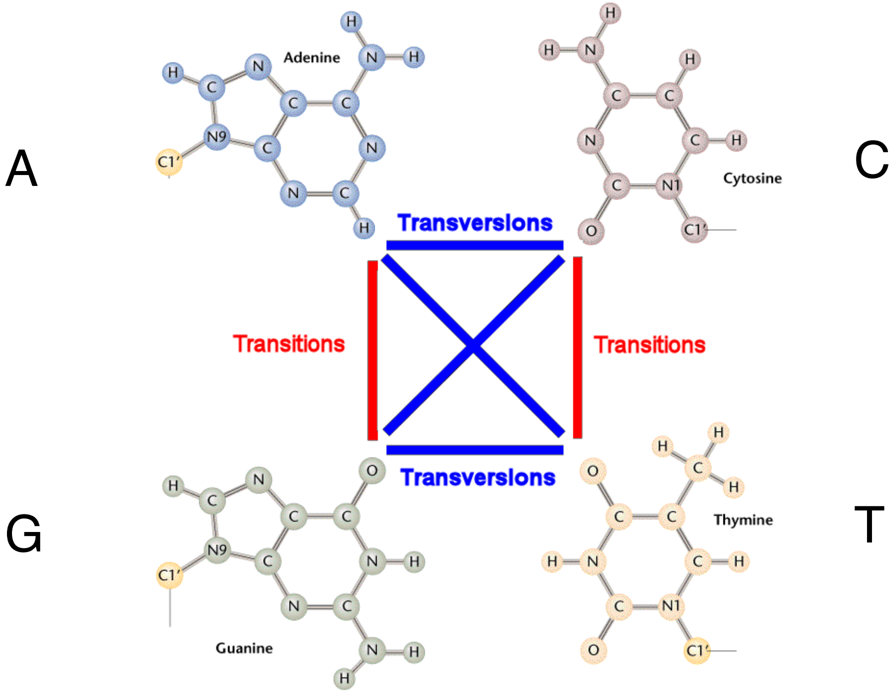
With a substitution rate of 10^{-9} it would take 375’000 years on average before a neutral position is substituted.

More general substitution models

- The Jukes-Cantor model assumes *all* mutations occur equally often but this is not true in reality.

For example:

- Transitions are more common than transversions.
- Mutations from C,G pairs to A,T pairs occur more often than vice-versa.



More general substitution models

We can more generally define a substitution-rate matrix: $P(\alpha|\beta,dt) = R_{\alpha\beta}dt$ for $\alpha \neq \beta$

Recall that $P(\alpha|\beta,t+dt) = (1-\mu\, dt)P(\alpha|\beta,t) + \sum_{\gamma\neq\alpha}\frac{\mu}{3}dt\, P(\gamma|\beta,t)$

$$\frac{\partial P(\alpha|\beta,t)}{\partial t} = -\mu\, P(\alpha|\beta,t) + \sum_{\gamma\neq\alpha}\frac{\mu}{3}\, P(\gamma|\beta,t)$$

Here $\frac{\mu}{3} = R_{\alpha\beta}$ for all $\alpha \neq \beta$ and $-\mu = -\sum_{\alpha\neq\beta} R_{\alpha\beta}$

Thus, we can generalize to a process in which mutation rates are not equal for all types of mutations, but are given by $R_{\alpha\beta}$. Further define $R_{\beta\beta} = -\sum_{\beta\neq\alpha} R_{\beta\alpha}$.

Then we can generally write $\frac{\partial P(\alpha|\beta,t)}{\partial t} = \sum_{\gamma} R_{\alpha\gamma}\; P(\gamma|\beta,t)$

which gives $P(\alpha|\beta,t) = (e^{Rt})_{\alpha\beta}$

Note: this solution takes into account the possibility of multiple mutations at the same position!

More general substitution models

We can more generally define a substitution-rate matrix: $P(\alpha|\beta, dt) = R_{\alpha\beta}dt$ for $\alpha \neq \beta$

Jukes & Cantor 1969

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & X & \alpha & \alpha & \alpha \\ C & \alpha & X & \alpha & \alpha \\ G & \alpha & \alpha & X & \alpha \\ T & \alpha & \alpha & \alpha & X \end{array}$$

1 parameter
equiprobable changes

Kimura 1980

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & X & \alpha & \kappa.\alpha & \alpha \\ C & \alpha & X & \alpha & \kappa.\alpha \\ G & \kappa.\alpha & \alpha & X & \alpha \\ T & \alpha & \kappa.\alpha & \alpha & X \end{array}$$

2 parameters
transition rate ≠
transversion rate

Tamura 1992

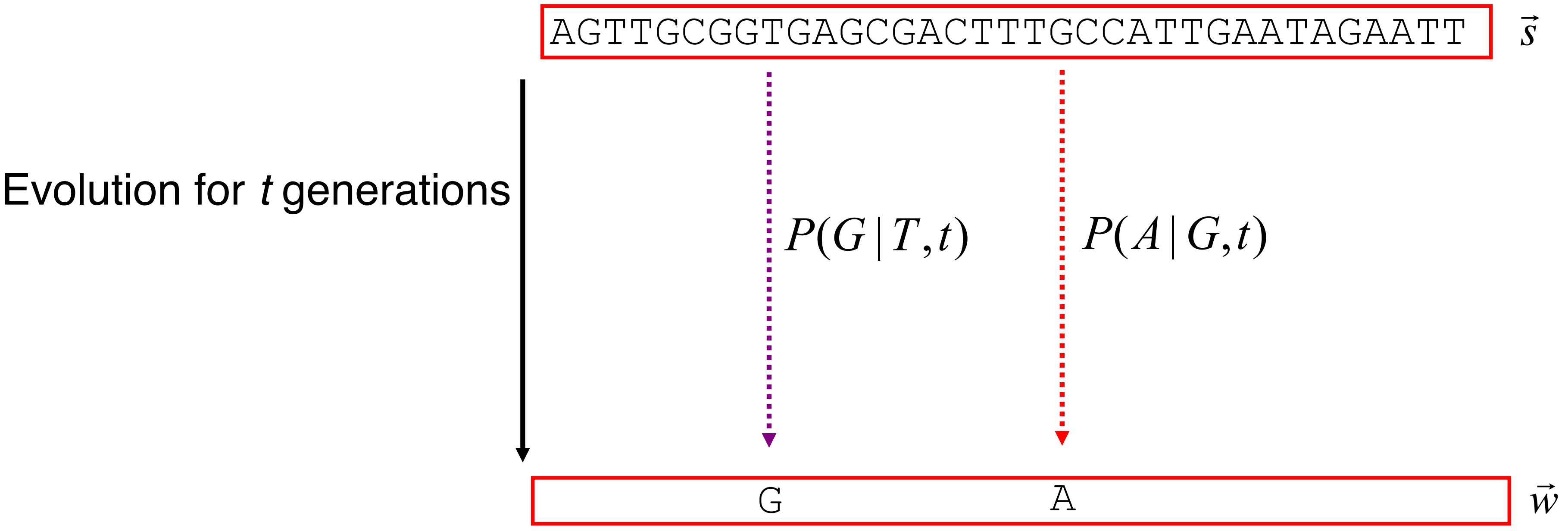
$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & X & \alpha\frac{1-\theta}{2} & \kappa\alpha\frac{1-\theta}{2} & \alpha\frac{1-\theta}{2} \\ C & \alpha\frac{\theta}{2} & X & \alpha\frac{\theta}{2} & \kappa\alpha\frac{\theta}{2} \\ G & \kappa\alpha\frac{\theta}{2} & \alpha\frac{\theta}{2} & X & \alpha\frac{\theta}{2} \\ T & \alpha\frac{1-\theta}{2} & \kappa\alpha\frac{1-\theta}{2} & \alpha\frac{1-\theta}{2} & X \end{array}$$

3 parameters
stationary GC% = $\theta \neq$ 50%

Pairwise distances

Neutral evolution of a sequence

Let us generalize now to an entire sequence.


AGTTGCGGTGAGCGACTTTGCCATTGAATAGAATT
$\vec{s}$
Evolution for *t* generations
$P(G|T,t)$
$P(A|G,t)$
G
A
$\vec{w}$

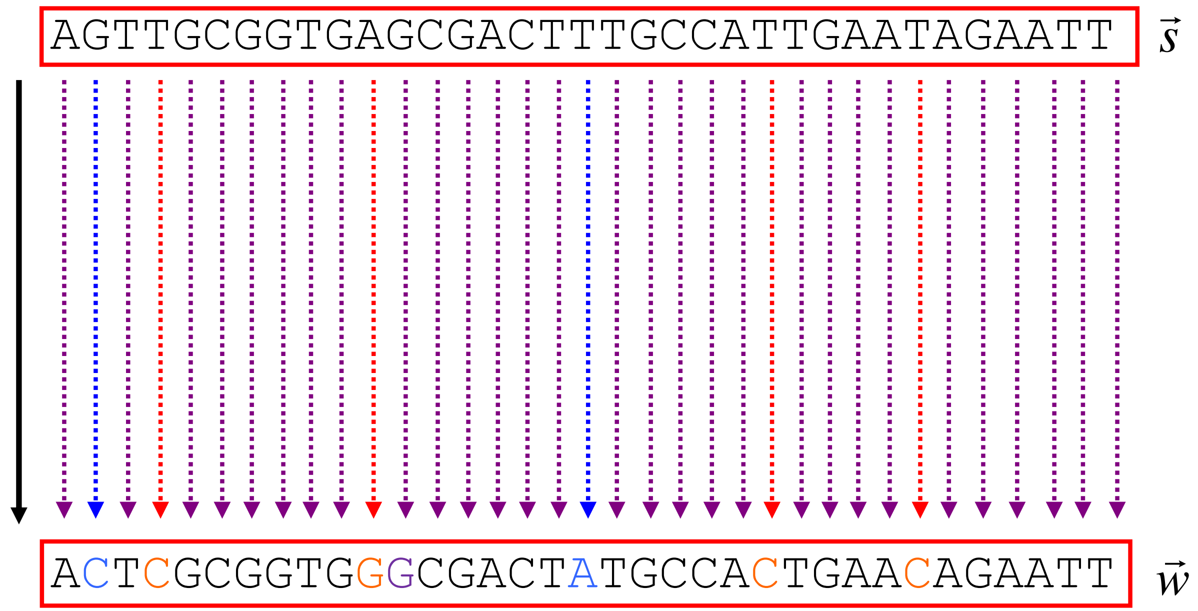
- At each position *i* in the sequence, there is a probability *P(w_i|s_i,t)* to evolve from ancestral base s_i to the descendant's base w_i

- The crucial second ingredient is that the probabilities for the evolution at different positions are *independent,* i.e. $P\left(w_i, w_j \mid s_i, s_j, t\right) = P\left(w_i \mid s_i, t\right) P\left(w_j \mid s_j, t\right)$

- Therefore, the entire evolutionary scenario has probability: $P\left(\vec{w} \mid \vec{s}, t\right) = \prod_i P\left(w_i \mid s_i, t\right)$

Neutral evolution of a sequence under Jukes-Cantor model

Evolution for *t* generations



AGTTGCGGTGAGCGACTTTGCCATTGAATAGAATT $\vec{s}$

ACTCGCGGTGGGCGACTATGCCACTGAACAGAATT $\vec{w}$

Under the simply Jukes-Cantor model, each position has probability $c = \frac{1}{4}\left(1 + 3e^{-\frac{4\mu t}{3}}\right)$ to show the same base, and $1 - c$ probability to show a different base. The probability that $\vec{s}$ evolves to $\vec{w}$ will be given by

$$P(\vec{w}|\vec{s}, t) = c^{L-d}(1-c)^d = \left(\frac{1}{4}\right)^{L-d}\left(1 + 3e^{-\frac{4\mu t}{3}}\right)^{L-d}\left(\frac{3}{4}\right)^{d}\left(1 - e^{-\frac{4\mu t}{3}}\right)^{d}$$

Where L is the length of the sequences and d is the number of observed differences.

Thus, for a given position in a sequence, we have inferred the probabilities to observe each of the 4 bases after an evolutionary time *t*, given that the ancestral sequence had any of the 4 bases.

What we want to get however is the evolutionary distance, i.e. the *time* that separates two current-day sequences.

Let’s first find the evolutionary distance between a sequence and its ancestral sequence.

Neutral evolution of a sequence under Jukes-Cantor model

AGTTGCGGTGAGCGACTTTGCCATTGAATAGAATT $\vec{s}$

t generations

ACTCGCGGTGGGCGACTATGCCAGTGAACAGAATT $\vec{w}$

In this example: d=6,L=35

$$P(\vec{w}|\vec{s},t) = c^{L-d}(1-c)^{d} = \left(\frac{1}{4}\right)^{L-d}\left(1+3e^{-\frac{4\mu t}{3}}\right)^{L-d}\left(\frac{3}{4}\right)^{d}\left(1-e^{-\frac{4\mu t}{3}}\right)^{d}$$

Finding the maximum likelihood distance to the common ancestor means finding t for which $P(\vec{w}|\vec{s},t)$ is maximal. Typically it is easier to maximize $log[P(\vec{w}|\vec{s},t)]$, which boils down to finding the value of $c = \frac{1}{4}\left(1+3e^{-\frac{4\mu t}{3}}\right)$ at which the maximum of $log[P(\vec{w}|\vec{s},t)]$ occurs.

Neutral evolution of a sequence under Jukes-Cantor model

AGTTGCGGTGAGCGACTTTGCCATTGAATAGAATT $\vec{s}$

t generations

ACTCGCGGTGGGCGACTATGCCAGTGAACAGAATT $\vec{w}$

In this example: d=6,L=35

$$P(\vec{w}|\vec{s},t) = c^{L-d}(1-c)^d = \left(\frac{1}{4}\right)^{L-d}\left(1+3e^{-\frac{4\mu t}{3}}\right)^{L-d}\left(\frac{3}{4}\right)^{d}\left(1-e^{-\frac{4\mu t}{3}}\right)^{d}$$

$$\frac{\partial log[P(\vec{w}|\vec{s},t)]}{\partial c} = \frac{\partial}{\partial c} log[c^{L-d}(1-c)^d] = \frac{\partial}{\partial c}[(L-d)\,log(c) + d\,log(1-c)]$$

$$= \frac{L-d}{c} - \frac{d}{1-c}$$

Setting this derivative to 0, we obtain $c = \frac{L-d}{L}$

Because $c = \frac{1}{4}\left(1+3e^{-\frac{4\mu t}{3}}\right)$, we get $\mu t = -\frac{3}{4} log\left[1-\frac{4d}{3L}\right]$

We call μt the mutational *distance* between the two sequences, i.e. the expected number of times each position has been mutated.

Neutral evolution of a sequence under Jukes-Cantor model

$\vec{s}$

AGTTGCGGTGAGCGACTTTGCCATTGAATAGAATT

t generations

$\vec{w}$

ACTCGCGGTGGGCGACTATGCCAGTGAACAGAATT

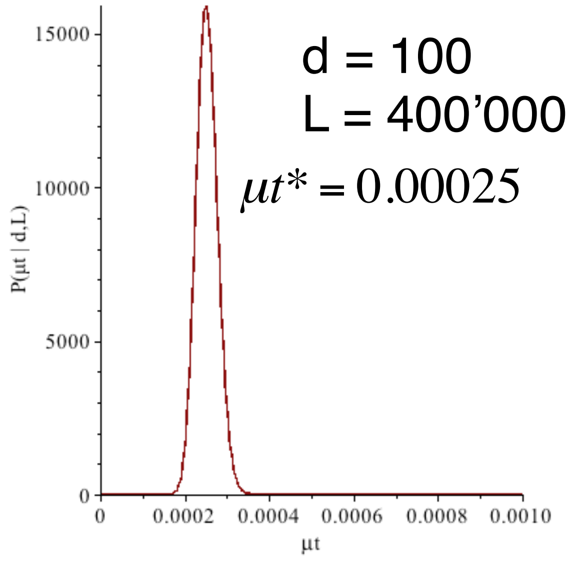
In this example: d=6,L=35

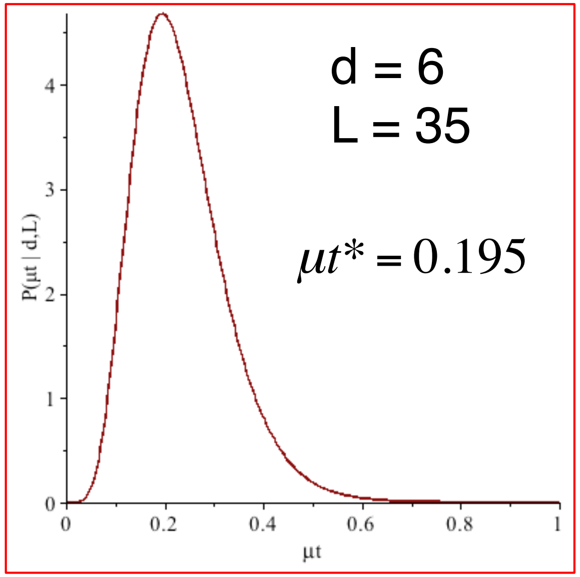
Posterior probability (Bayes theorem)

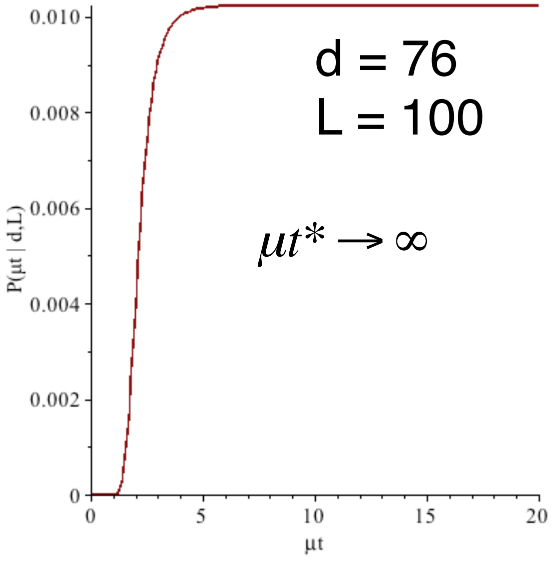
$$P(\mu t \mid d,L) = \frac{P(d \mid \mu t,L)\,P(\mu t)}{\int P(d \mid \mu t,L)\,P(\mu t)\,d(\mu t)}$$

Assuming uniform prior for μt

$$P(d|\mu t,L) = c^{L-d}(1-c)^{d} = \left(\frac{1}{4}\right)^{L-d}\left(1+3e^{-\frac{4\mu}{3}t}\right)^{L-d}\left(\frac{3}{4}\right)^{d}\left(1-e^{-\frac{4\mu}{3}t}\right)^{d}$$







Distances in Kimura’s 1980 model

Rate matrix:

$$R = \begin{array}{c} \\ \\ \end{array}\begin{array}{cc} \begin{array}{cccc} \textbf{A} & \textbf{C} & \textbf{G} & \textbf{T} \end{array} & \\ \begin{pmatrix} -\mu(2+k) & \mu & k\mu & \mu \\ \mu & -\mu(2+k) & \mu & k\mu \\ k\mu & \mu & -\mu(2+k) & \mu \\ \mu & k\mu & \mu & -\mu(2+k) \end{pmatrix} & \begin{array}{c} \textbf{A} \\ \textbf{C} \\ \textbf{G} \\ \textbf{T} \end{array} \end{array}$$

Recall that

in this model rate of transitions is different than rate of transversions

the diagonal elements are set to – (sum of the rest of the elements in a row) by the requirement that this is a rate matrix.

Distances in Kimura's 1980 model

Rate matrix:

$$R = \begin{array}{c} \\ \end{array}\begin{array}{cc} \begin{array}{cccc} \textbf{A} & \textbf{C} & \textbf{G} & \textbf{T} \end{array} & \\ \begin{pmatrix} -\mu(2+k) & \mu & k\mu & \mu \\ \mu & -\mu(2+k) & \mu & k\mu \\ k\mu & \mu & -\mu(2+k) & \mu \\ \mu & k\mu & \mu & -\mu(2+k) \end{pmatrix} & \begin{array}{c} \textbf{A} \\ \textbf{C} \\ \textbf{G} \\ \textbf{T} \end{array} \end{array}$$

Solving $\frac{\partial P(\alpha \mid \beta, t)}{\partial t} = \sum_{\gamma} R_{\alpha\gamma} P(\gamma \mid \beta, t)$ to get $P(\alpha \mid \beta, t) = \left(e^{Rt}\right)_{\alpha\beta}$

Recall that $e^{Rt} = Me^{Dt}M^{-1}$ where M is the matrix with the eigenvectors of R in its columns and D is the diagonal matrix of eigenvalues of R.

$$M = \begin{pmatrix} -1 & 1 & 0 & -1 \\ 1 & 1 & -1 & 0 \\ -1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad D = \begin{pmatrix} -4\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -2(k+1)\mu & 0 \\ 0 & 0 & 0 & -2(k+1)\mu \end{pmatrix}$$

Distances in Kimura's 1980 model

Rate matrix:

$$R = \begin{array}{c} \\ \\ \\ \\ \\ \end{array}\begin{array}{cccc|c} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} & \\ -\mu(2+k) & \mu & k\mu & \mu & \mathbf{A} \\ \mu & -\mu(2+k) & \mu & k\mu & \mathbf{C} \\ k\mu & \mu & -\mu(2+k) & \mu & \mathbf{G} \\ \mu & k\mu & \mu & -\mu(2+k) & \mathbf{T} \end{array}$$

Solving $\frac{\partial P(\alpha\,|\,\beta,t)}{\partial t} = \sum_{\gamma} R_{\alpha\gamma} P(\gamma\,|\,\beta,t)$ to get $P(\alpha\,|\,\beta,t) = \left(e^{Rt}\right)_{\alpha\beta}$

Recall that $e^{Rt} = Me^{Dt}M^{-1}$ where M is the matrix with the eigenvectors of R in its columns and D is the diagonal matrix of eigenvalues of R.

Solution: $P(\alpha\,|\,\beta,t) = \begin{array}{cccc|c} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} & \\ r(t) & v(t) & s(t) & v(t) & \mathbf{A} \\ v(t) & r(t) & v(t) & s(t) & \mathbf{C} \\ s(t) & v(t) & r(t) & v(t) & \mathbf{G} \\ v(t) & s(t) & v(t) & r(t) & \mathbf{T} \end{array}$

with: $v(t) = \frac{1}{4}\left(1 - e^{-4\mu t}\right)$

$s(t) = \frac{1}{4}\left(1 + e^{-4\mu t} - 2e^{-2(k+1)\mu t}\right)$

$r(t) = 1 - 2v(t) - s(t)$

Distances in Kimura's 1980 model

$$P(\alpha \mid \beta, t) = \begin{array}{c} \\ \\ \\ \\ \end{array}\begin{array}{cccccc} & \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} & \\ \Bigg(& r(t) & v(t) & s(t) & v(t) & \Bigg) \; \mathbf{A} \\ & v(t) & r(t) & v(t) & s(t) & \mathbf{C} \\ & s(t) & v(t) & r(t) & v(t) & \mathbf{G} \\ & v(t) & s(t) & v(t) & r(t) & \mathbf{T} \end{array}$$

with:

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

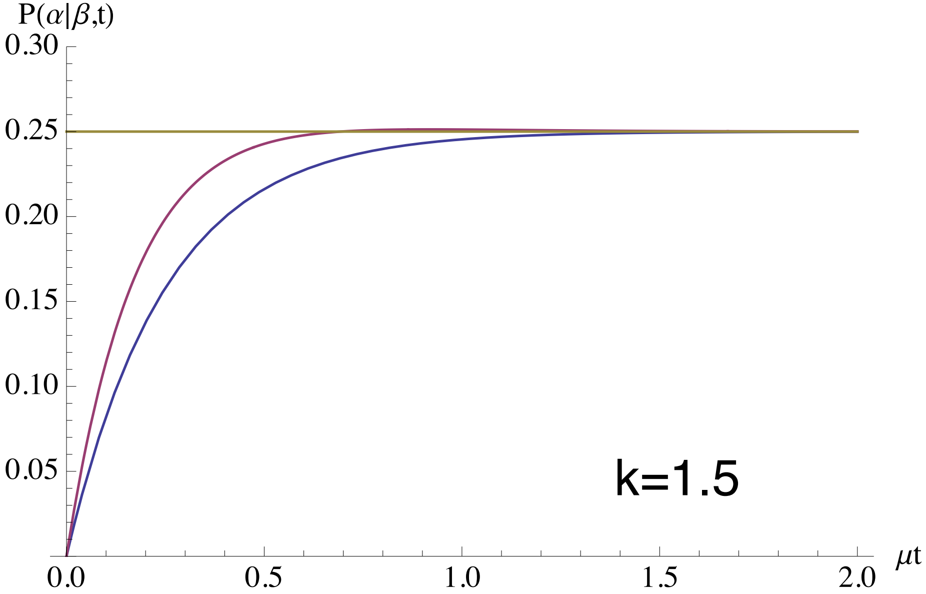
$$s(t) = \frac{1}{4}\left(1 + e^{-4\mu t} - 2e^{-2(k+1)\mu t}\right)$$

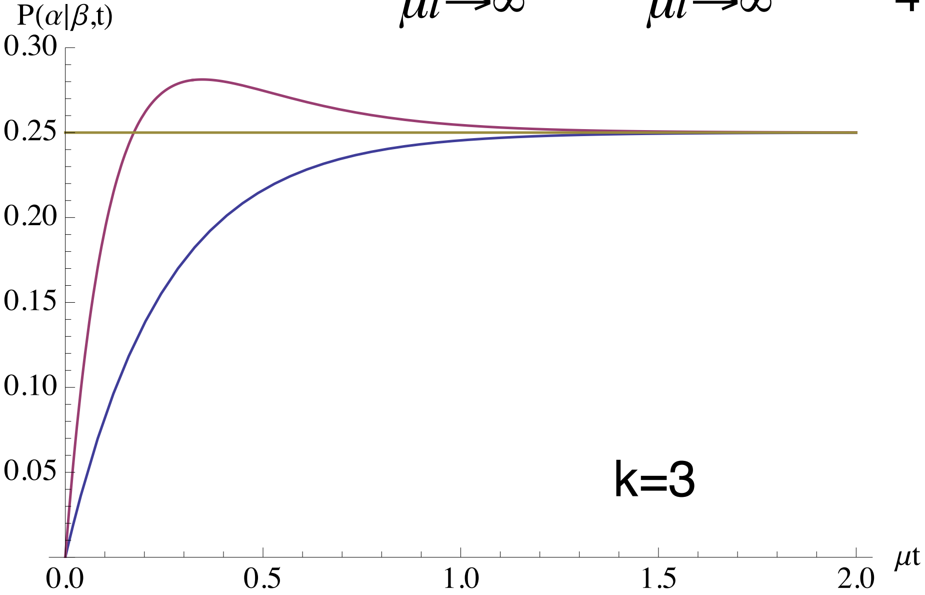
$$r(t) = 1 - 2v(t) - s(t)$$

Examples:

v(t) in blue, s(t) in red

Observe that $\lim_{\mu t \to \infty} v(t) = \lim_{\mu t \to \infty} s(t) = \frac{1}{4}.$





Distances in Kimura's 1980 model

$$P(\alpha \mid \beta, t) = \begin{array}{cccc|c} \textbf{A} & \textbf{C} & \textbf{G} & \textbf{T} & \\ r(t) & v(t) & s(t) & v(t) & \textbf{A} \\ v(t) & r(t) & v(t) & s(t) & \textbf{C} \\ s(t) & v(t) & r(t) & v(t) & \textbf{G} \\ v(t) & s(t) & v(t) & r(t) & \textbf{T} \end{array}$$

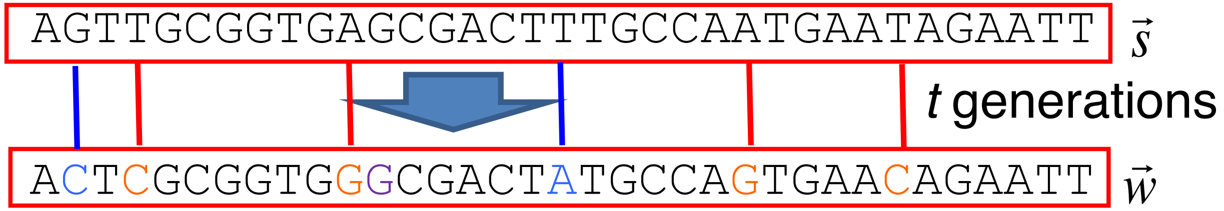
with:

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

$$s(t) = \frac{1}{4}\left(1 + e^{-4\mu t} - 2e^{-2(k+1)\mu t}\right)$$

$$r(t) = 1 - 2v(t) - s(t)$$

Maximum likelihood distance: Count the fraction of transitions *S* and transversions V.



AGTTGCGGTGAGCGACTTTGCCAATGAATAGAATT $\vec{s}$

t generations

ACTCGCGGTGGGCGACTATGCCAGTGAACAGAATT $\vec{w}$

In this example:
S = 4/35, *V* = 2/35

Knowing that $P\left(\vec{w} \mid \vec{s}, t\right) = s(t)^{LS} v(t)^{LV} r(t)^{L(1-S-V)}$

we can compute the maximum likelihood values of the parameters μt and k.

Distances in Kimura's 1980 model

$$P\left(\vec{w} \mid \vec{s},t\right) = s(t)^{LS}\, v(t)^{LV}\, r(t)^{L(1-S-V)}$$

$$s(t)^{LS}\, v(t)^{LV}\left(1-2v(t)-s(t)\right)^{L(1-S-V)}$$

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

$$s(t) = \frac{1}{4}\left(1+e^{-4\mu t} - 2e^{-2(k+1)\mu t}\right)$$

$$r(t) = 1-2v(t)-s(t)$$

and

$$\log\left(P\left(\vec{w} \mid \vec{s},t\right)\right) = LS\log(s(t)) + LV\log(v(t)) + L(1-S-V)\log(1-2v(t)-s(t))$$

To obtain maximum likelihood values of the parameters we solve

$$\frac{\partial \log\left[P\left(\vec{w} \mid \vec{s},t\right)\right]}{\partial s(t)} = 0 \text{ and } \frac{\partial \log\left[P\left(\vec{w} \mid \vec{s},t\right)\right]}{\partial v(t)} = 0$$

Distances in Kimura's 1980 model

$$\log\left(P\left(\vec{w} \mid \vec{s},t\right)\right) = LS\log(s(t)) + LV\log(v(t)) + L(1-S-V)\log(1-2v(t)-s(t))$$

$$\frac{\partial \log\left[P\left(\vec{w} \mid \vec{s},t\right)\right]}{\partial s(t)} = \frac{LS}{s(t)} - \frac{L(1-S-V)}{1-2v(t)-s(t)} \qquad \frac{\partial \log\left[P\left(\vec{w} \mid \vec{s},t\right)\right]}{\partial v(t)} = \frac{LV}{v(t)} - \frac{2L(1-S-V)}{1-2v(t)-s(t)}$$

In simplified notation $\frac{LS}{s} - \frac{L(1-S-V)}{1-2v-s} = 0$ and $\frac{LV}{v} - \frac{2L(1-S-V)}{1-2v-s} = 0$

which lead to $s = S$ and $v = \frac{V}{2}$

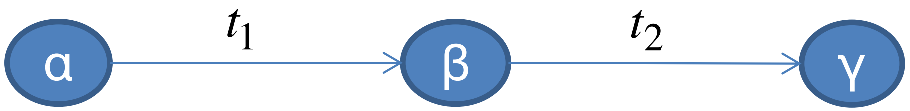
$$v(t) = \frac{1}{4}\left(1 - e^{-4\mu t}\right)$$
$$s(t) = \frac{1}{4}\left(1 + e^{-4\mu t} - 2e^{-2(k+1)\mu t}\right)$$

imply

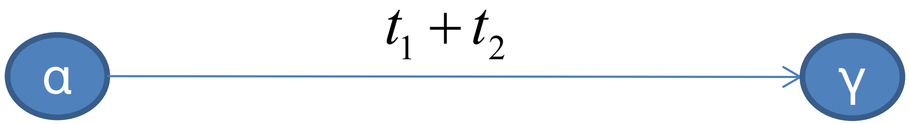
$$\mu t_* = -\frac{1}{4}\log\left[1-2V\right]$$
$$k_* = \frac{2\log\left[1-V-2S\right]}{\log(1-2V)} - 1$$

Additivity and Reversibility

The substitution process is *additive* in time:



$$\sum_{\beta} P(\beta \mid \alpha, t_2) P(\gamma \mid \beta, t_1) = \sum_{\beta} \left(e^{Rt_2}\right)_{\gamma\beta} \left(e^{Rt_1}\right)_{\beta\alpha} = \left(e^{R(t_1+t_2)}\right)_{\gamma\alpha} = P(\gamma \mid \alpha, t_1 + t_2)$$



For most rate matrices, the substitution process is also reversible.

Denote the limit distribution q_α of the substitution process $\dfrac{\partial P\left(\alpha \mid \beta, t\right)}{\partial t} = 0 \Rightarrow \sum_{\beta} R_{\alpha\beta} q_{\beta} = 0$

For the Jukes-Cantor and Kimura 1980 models $\vec{q} = (0.25, 0.25, 0.25, 0.25)$.

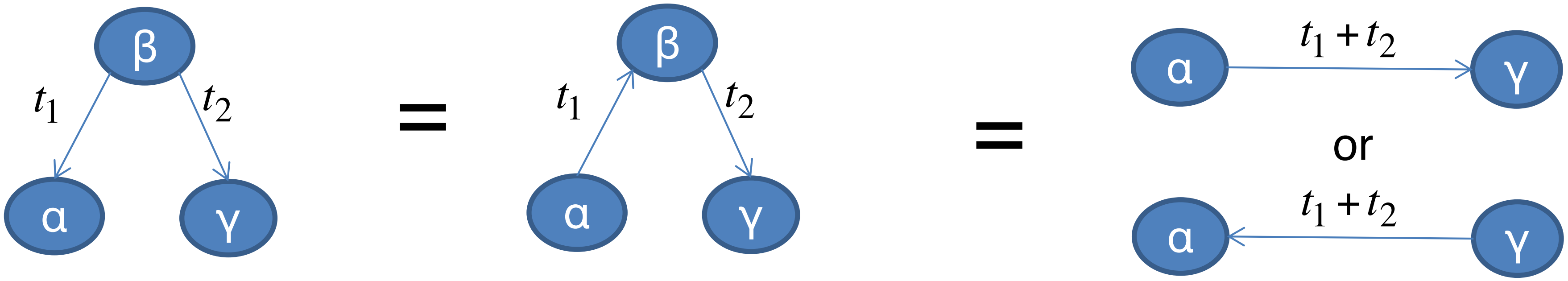
In general, a reversible model obeys detailed balance:

$$R_{\alpha\beta} q_{\beta} = R_{\beta\alpha} q_{\alpha} \Rightarrow P\left(\alpha \mid \beta, t\right) q_{\beta} = P\left(\beta \mid \alpha, t\right) q_{\alpha}$$

No dependence on the root node

Using $P(\alpha \mid \beta, t) q_\beta = P(\beta \mid \alpha, t) q_\alpha$ we can derive the following:

$$\sum_\beta P(\alpha \mid \beta, t_1) P(\gamma \mid \beta, t_2) q_\beta \qquad \sum_\beta P(\beta \mid \alpha, t_1) P(\gamma \mid \beta, t_2) q_\alpha$$



- The probability to start from a base *β* and evolve to letter *α* over *t_1* generations and to letter *γ* over *t_2* generations is the *same* as:
- To start from letter *α* and evolve to letter *γ* over *t_1+t_2* generations.

or

- To start from letter *γ* and evolve to letter *α* over *t_1+t_2* generations.

Reconstructing phylogenies

- Substitution models
- Pairwise distances
- Likelihood of a phylogenetic tree: Felsenstein’s algorithm
- Reconstructing phylogenetic trees:
    - Maximum likelihood
    - Neighbor-joining