$$= -P_{out} + P_{in}$$

P_{out}: chem. output of incorporating R + W into chain

P_{in}: chem. input of hydrolyzing RTP WTP

Thus $P_{diss} = P_{in} - P_{out} + k_B V S_{info} \geq 0$

S_{info}: new term corresponding to writing into + propagating errors

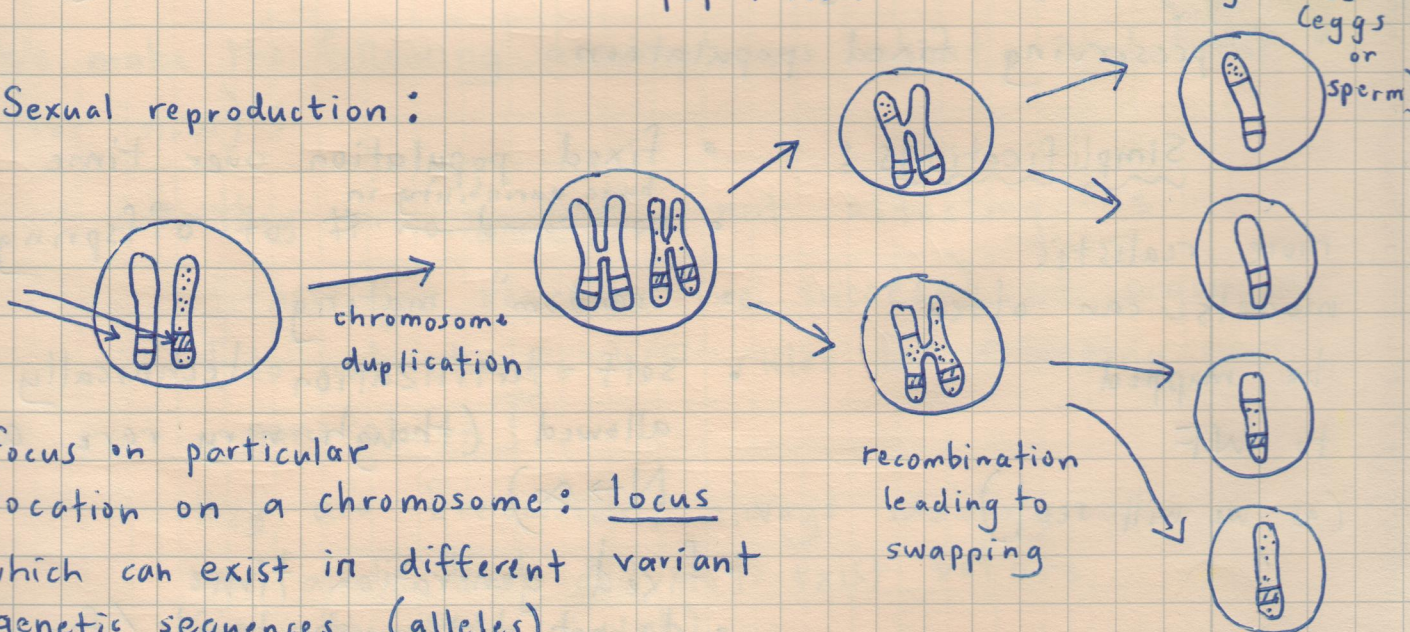
A positive error rate allows us to use less P_{in} to get same P_{out}!

In fact $\varepsilon_R, \varepsilon_W$ could be zero or slightly negative + $V > 0$ is still possible (at high error)

$\Rightarrow$ driving by disorder!

Population genetics: the evolution of genetic changes in a population over time

Sexual reproduction:

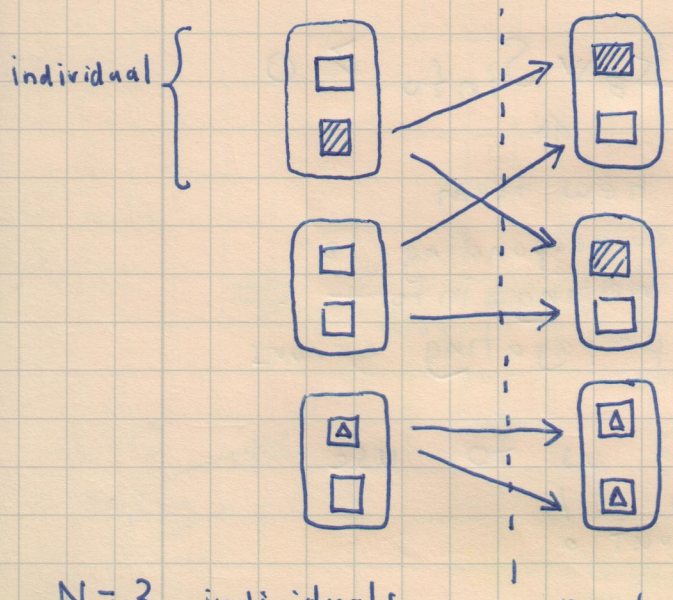


focus on particular location on a chromosome: locus

which can exist in different genetic sequences (alleles) variant

some gametes randomly merge ⇒ create new organism in next generation

abstract away whole process into simple model:
focus on the set of all of one type of locus across the entire population of species



Reproduction:
for each locus in new generation, choose a "parent" randomly in previous generation, draw arrow, copy allele from parent

⇒ Wright-Fisher model (W
highly simplified, but widely used

finite population sampling changing allele frequency randomly ⇒ genetic drift

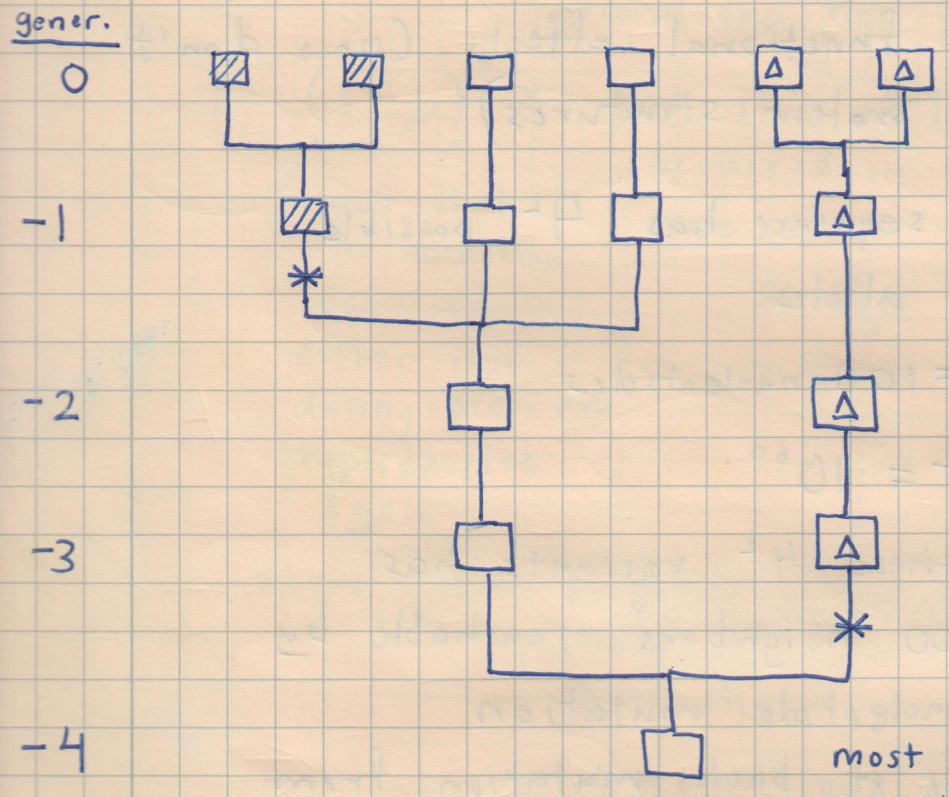
Next generation formed after fixed time t_g,
by randomly choosing 2N alleles (with replacement)
from previous generation as "parents" to copy,
preserving fixed population

Simplifications:

- fixed population over time
- ~~no bound on~~ huge variability in # of offspring (0 to N
- "random" mating
- self-fertilization technically allowed (though very rare as $N \to \infty$)
- fixed generation time
- do not distinguish male/female

more realistic models can often be mapped to WF (as we will see!)

we can draw a geneology tree for every allele in current generation (a "coalescent"), ~~we~~ with mergers ("coalescences") representing a shared parent allele ⇝ each allele acts like an indep. organism in the model, cloning itself anywhere from 0 to N times



* = mutations changing the genetic sequence (making a new allele)

most recent common ancestor (MRCA) allele of all alleles in current generation 0

We make the following assumptions about mutations (for now):

- they are <u>neutral</u>, not affecting the fitness (or reproductive success) of organism [do not change WF rules for selecting new generation]
- they are irreversible, always making a new allele that did not exist before (infinite alleles model)

We will relax these assumptions later, but they can be biologically realistic when:

1) you consider a stretch of DNA of length $L \gg 1$ in a non-coding region (98% of human genome does not code for proteins), where mutations often have no functional effects (since don't change protein structures)

or "silent" sites where mutation does not change amino acid result, i.e. GC x where x = G, C, T, A always codes for arginine

2) such a sequence has 4^L possible variant alleles

i.e. $L = 100$ nucleotides

$$4^L = 10^{60}$$

each of these 4^L variants has $3L = 300$ "neighbors" reachable by single nucleotide mutation

$\Rightarrow$ chance of back mutation from neighbor to original is very small

$$\sim \frac{1}{3L}$$

define a measure of genetic diversity in current generation:

H = prob. that any two alleles chosen at random from current generation are different = "heterozygosity"

= equivalent to looking at two lineages of the two alleles, asking "did mutation occur before coalescence"?

• do not [illegible]

going back each generation,

$\frac{1}{2N}$ = prob. that two lineages coalesced in previous generation

(choose two alleles; first one has parent X; the prob. that second one has parent X is $\frac{1}{2N}$)

$1-\underbrace{(1-\mu)}^{2} \approx$ prob. that at least one mutation occurred in the two lines since previous generation

$\approx 2\mu$ when $\mu \ll 1$

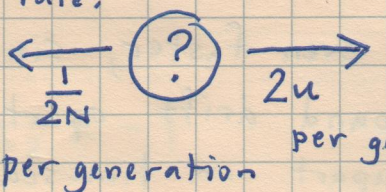
prob. that no mutation occurred in either line during previous reproductive cycle

where μ = mutation probability ~~rate~~ per generation → base pair

$\sim 5 \times 10^{-10}$ per bp per generation (prokaryotes)

to 5×10^{-8} /bp/generation (vertebrates)

We can think of the answer to the question of same or different being a transition out of ? state, allowed to occur each generation going back:



Hence H = fraction to escape to "different" $= \frac{2u}{\frac{1}{2N} + 2u}$

Rewriting $\mathcal{H} = \frac{4N\mu}{1+4N\mu} = \frac{\theta}{1+\theta}$

$\theta \equiv 4N\mu$

~~How to estimate θ?~~

If we measure μ directly, & estimate ~~$\theta = ?$~~ $\mathcal{H}$ from population sample, then we know $\theta = 4N\mu$ & can infer N.

using "silent" sites where mutations are neutral

$\theta \approx 4 \times 10^{-3}$ for single base-pair region ($L=1$) in vertebrates

$\rightarrow 10^{-1}$ for basepair in prokaryotes

humans: $\mu \sim 2.6 \times 10^{-8}$ /bp/generation

$\theta \sim 1.1 - 2 \times 10^{-3}$

$\Rightarrow N \sim 1 - 2 \times 10^{4}$!!

but actual population $N_{actual} \sim 7.6$ billion

What is going on?

N in WF model is an effective population size, b/c real populations often diverge from WF assumptions

example N fixed (idealized WF)

$N_{actual}(t) \Rightarrow$ time-varying real population

How to map from $N_{actual}(t)$ to N effective?

$\rightsquigarrow$ various choices, leading to similar answers

for example: probability that two lineages from a pair of alleles in the current generation have not coalesced going n generations into past:

$$U_c(n) = \left(1 - \frac{1}{2N}\right)^n \quad \text{in idealized WF}$$

$\Rightarrow$ or in time variable

$$U_c(t) = \left(1 - \frac{1}{2N}\right)^{\frac{t}{t_g}}$$

$t = n t_g$ in the past

$$\approx \left(e^{-\frac{1}{2N}}\right)^{t/t_g} \quad \text{for } N >> 1$$

$$= e^{-\frac{t}{2N t_g}}$$

for actual population:

$$U_c(t) = \left(1 - \frac{1}{2N_{actual}(-t_g)}\right)\left(1 - \frac{1}{2N_{actual}(-2t_g)}\right)$$

with $t = n t_g$

$$\cdots \left(1 - \frac{1}{2N_{actual}(-n t_g)}\right)$$

$$\approx \exp\left[-\sum_{i=1}^{n} \frac{1}{2N_{actual}(-i t_g)}\right]$$

$$= \exp\left[-\frac{1}{t_g}\sum_{i=1}^{n} t_g \frac{1}{2N_{actual}(-i t_g)}\right]$$

$$\approx \exp\left[-\frac{1}{t_g}\int_0^t dt' \frac{1}{2N_{actual}(t')}\right]$$

comparing models:

$$N = \left[\frac{1}{t}\int_0^t dt' \frac{1}{N_{actual}(t')}\right]^{-1}$$

$\lesssim N_{actual}(0)$ since integral dominated by smallest population in past

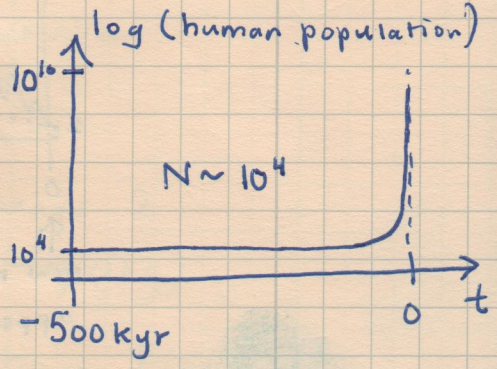
During the previous generations going back to human coalescence, "bottlenecks" may have occurred, reducing human populations to $\sim 10^4$ (or smaller) for long periods of time

How long ago was human coalescence?

$P_c(t)\,dt$ = prob. coalescence occurred between t & $t+dt$

$$= U_c(t) - U_c(t+dt)$$

$$\approx -dt\,\frac{dU_c}{dt}$$



mean coalescence time

$$t_c = \int_0^\infty dt\; t\; P_c(t)$$

$$= -\int_0^\infty dt\; t\; \frac{dU_c}{dt}$$

$$= -\int_0^\infty dt\; t \cdot \left(-\frac{1}{2Nt_g}\right) e^{-\frac{t}{2Nt_g}}$$

$$= 2Nt_g$$

if $N \sim 10^4$ $t_g \sim 20$ yrs $\Rightarrow$ $t_c \sim 400{,}000$ years

rough agreement in timescale w/ better estimates

---

mitochondrial "Eve" ~ 99–152 kyr

Y-chromosome "Adam" ~ 120–581 kyr

note for mtDNA (passed only from mother) + Y-chromosome (only from father)

$\Theta = 2N\mu$ + $N = \begin{cases} N_{males} & \text{Y-DNA} \\ N_{females} & \text{mtDNA} \end{cases}$

for native Americans, $N_e \sim 7$ to 10^4 (estimated)! at beginning of crossing over