Last time:

* **Approximations & self consistency**

$$\epsilon x^5 + x - 1 = 0 \quad \Rightarrow \quad x \approx 1 \quad (\epsilon x^2 \approx \epsilon \ll 1)$$

(ϵ: small? x, 1: dominant?)

(vs $x \approx \epsilon^{-1/5}$ when $\epsilon \gg 1$)

* **Probability** $(x \sim p(x))$

$\Rightarrow$ **Generating functions**: $H_x(z) \equiv \langle e^{-z \cdot x} \rangle = \int e^{-zx} p(x) dx$

e.g. Poisson $p(n) = \frac{\lambda^n}{n!} e^{-\lambda} \iff H_n(z) = e^{-\lambda(1-e^{-z})}$

$\Rightarrow$ **Central limit theorem**: $X_1, X_2, \ldots, X_n \sim p(x)$

as $n \to \infty \Rightarrow \frac{1}{n} \sum_{i=1}^{n} X_i \to \text{Gaussian}\left(\langle x \rangle, \frac{\text{Var}(x)}{n}\right)$

Today:

1. Intuition about probability
2. Biological background (#s & scales)
3. Simple model of evolution (if time permits)

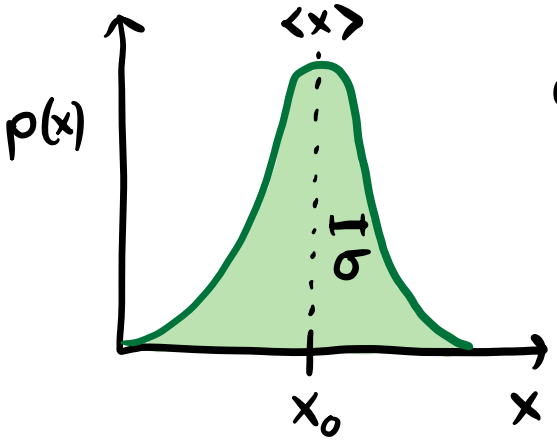
Intuition about random variables ("average" vs "typical")

Probability can be hard b.c. it forces us to think about many outcomes all @ once...

⇒ often want some way of summarizing "typical" behavior

we will frequently encounter 2 broad classes:

Case 1 ("fuzzy noise"):

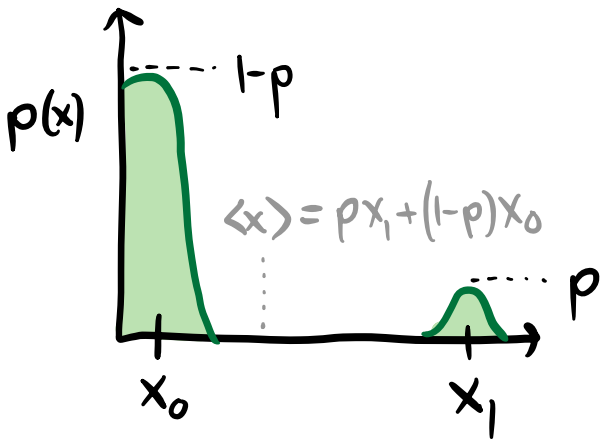


e.g. Binomial(N, p)
when $Np \gg 1$ & $N(1-p) \ll 1$

$\Rightarrow \quad x \approx x_0 \pm \sigma$

⇒ average is good summary of "typical"

Case 2 ("jagged noise"):



e.g. Binomial(N, p)
when $Np \ll 1$
e.g. did a mutation occur or not?

$\Rightarrow$ no realization of x has $x \approx \langle x \rangle$

$\Rightarrow$ average is <u>poor</u> summary of typical!

(better off guessing $x \approx x_0$, + rare exceptions)

$\Rightarrow$ distinction becomes important if we <u>do something</u> w/ x:

e.g. $y = F(x)$ = "future growth of x mutations"

$\Rightarrow$ in case 1: can treat noise as <u>small perturbation</u>
& use approx. methods above.

e.g. using Taylor expansion around $x \approx x_0$:

$$F(x) = F\big(x_0 + (x - x_0)\big) \approx F(x_0) + F'(x_0)(x - x_0)$$

w/ $x = x_0 \pm \sigma$ ↓

$$y \approx F(x_0) \pm F'(x_0)\,\sigma$$

↑ deterministic guess

↖ small spread ("fuzziness") due to noise.

case 2:

$$y = \begin{cases} F(x_0) & \text{w/ prob } 1-p \\ F(x_1) & \text{w/ prob } p \end{cases}$$

↙ this can be "typical" case (most of time)

↖ "rare event" happens separately.

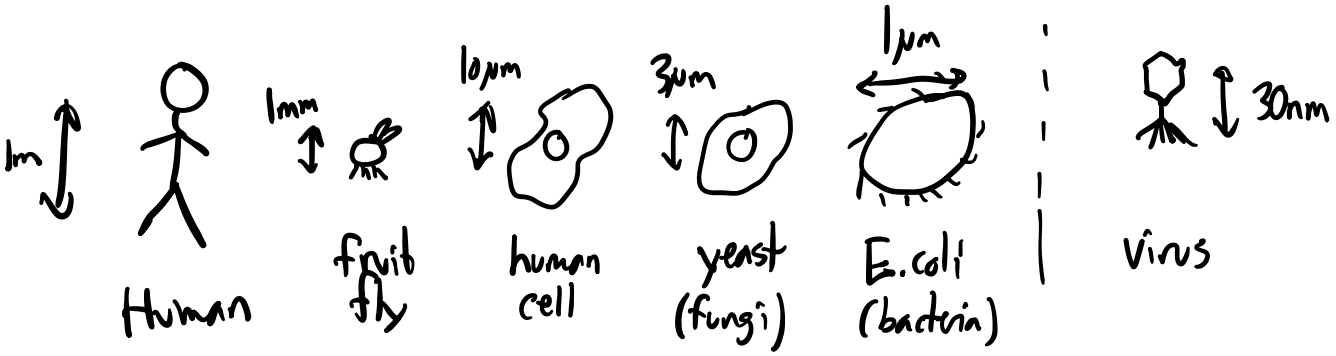
⇒ often used to thinking about case 1 dist'ns.

⇒ evolution will have many examples of case 2!

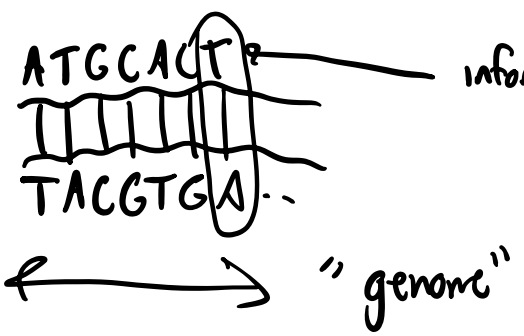
Biological background (key #'s and scales)

(1) Organisms come in huge range of shapes & sizes:

"Model organisms" we will encounter in this course:



(2) Despite diffs, these organisms are similar in that instructions to create them are encoded in a single* long molecule of DNA:



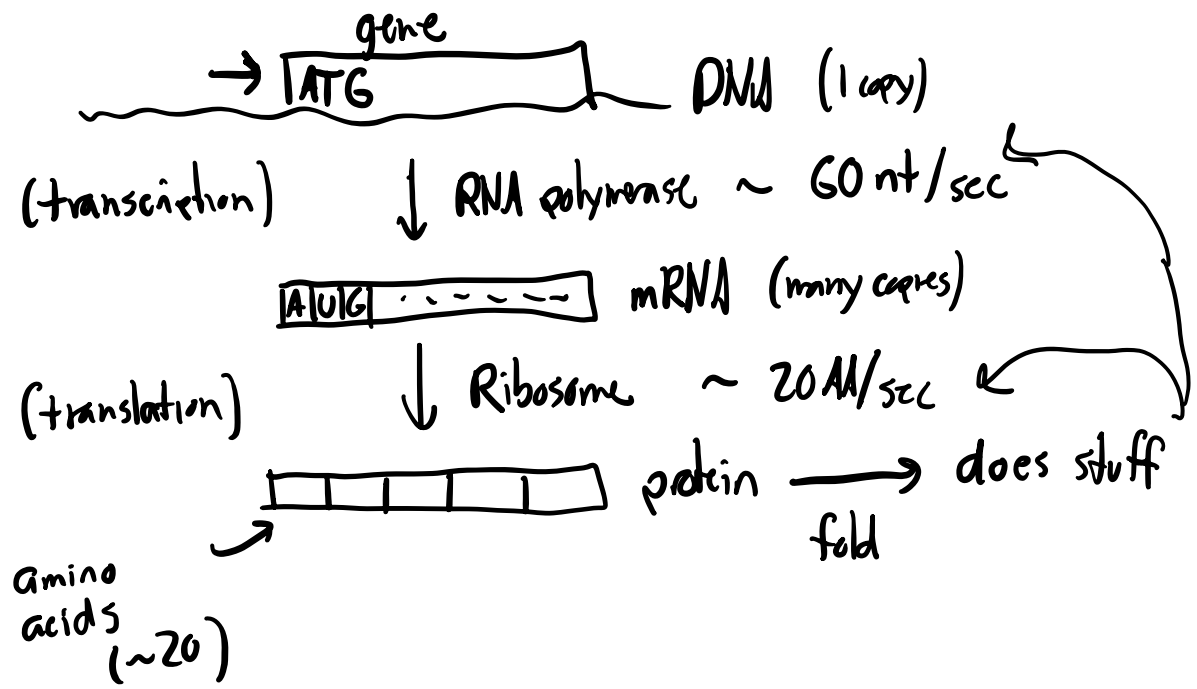
We take this for granted now, but pretty crazy from physics perspective! (important info in 1 molecule vs many)

Lengths of genomes vary widely across species:

human: ~ 10^9 bp yeast: 10^7 bp virus: 10^4 ~ 10^5 bp

fruit fly: ~ 10^8 bp bacteria: 10^6 bp (1Gbp, 1Mb, 1kb
= 10^9 bp, 10^6 bp, 1000bp.

information often encoded in genes (make proteins)



How does ribosome do it? ATT = "codon"

⇒ 1 amino acid (isoleucine)

$4^3 = 64$ different codons ⟶ 20 amino acids
+ "start codon"
+ "stop" codon

"genetic code"

⇒ has degeneracy

⇒ typical protein ~ 300 AA (1000 bp of DNA)

⇒ # of genes varies widely across organisms:

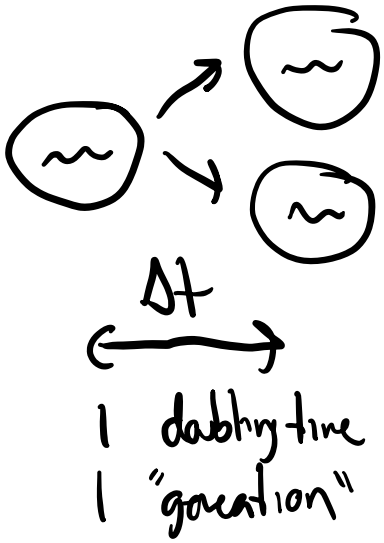
Humans: 20,000 genes Yeast: 6,000 genes

E. coli ~ 4,000 genes Viruses ~ 10 genes.

→ 1000x bigger genome ⇒ but 5x as many genes.

⇒ rest of genome is "noncoding" → regulation
("coding" = genes) → "junk"

⇒ net effect of doing all these things
is that the organism makes a copy of itself!



1 doubling time
1 "generation"

① new cell wall, all other proteins (including ribosomes!)

② needs to copy its DNA (DNA polymerase) (not usually limiting factor in growth)

Some characteristic generation times:

humans: ~20 yrs

human cells (HeLa): ~1 day

Virus: HIV ~ 13 hrs
SARS-CoV-2 ~ 10 hrs

E. coli ~ 20 mins - 1 hr (lab)
1 hr - 1 day? (in gut)

Prochlorococcus ~ 1 day
(ocean bacterium, one of the most abundant photosynthetic organisms on earth, $N \sim 10^{27}$)

⇒ Since n=1 genome, can make errors during copying

....ATGCCA.... parent
....ATGTCA.... offspring

"mutations"

⇒ simplest mutations are "point mutations" (A→T, T→C, ...)

aka "single nucleotide mutations" / "substitutions" / "SNPs"

⇒ can also have "insertions":

...ATGTTCA...
↓
...ATGTTTTTCA...
(+3T)

⇒ or "deletions"

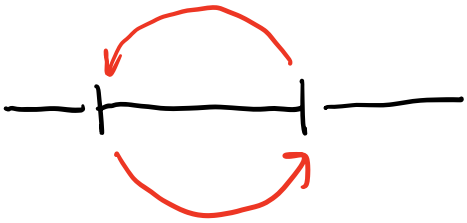
...ATGTTTCA...
↓
...ATGTCA...
(-2)

e.g. slippage of DNA pol

⇒ can also have larger "structural rearrangements":

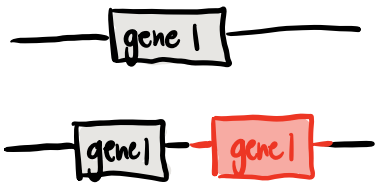
(eg. >1kb)

e.g. "inversions"

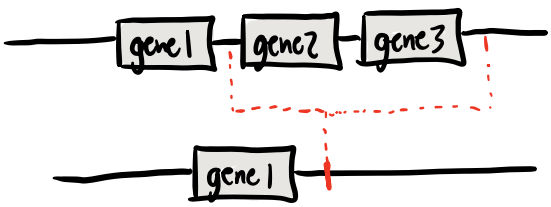


often mediated by special genes known as "transposons"

e.g. "duplications"



e.g. "deletions"



⇒ upshot: can get pretty complicated

⇒ cells have sophisticated machinery for fixing errors that occur in genome...

⇒ net mutation rates (μ) vary across organisms!

e.g. Humans: $\mu \sim 10^{-8}$ single nucleotide muts / bp / gen

Human cells: $\mu \sim 10^{-10}$/bp/division E. coli: $\mu \sim 10^{-10}$/bp/gen

viruses: up to $\mu \sim 10^{-5}$/bp/gen (SARS-CoV2 10^{-6}/bp/gen)

$\Rightarrow$ Using these #'s, can already make some interesting predictions...

Evolutionary "Fermi Problems"

e.g. in Humans genome is $L = 3 \times 10^9$ bp

+ mutation rate $\mu \sim 10^{-8}$/bp/gen

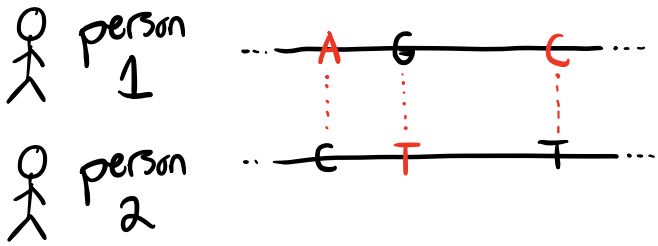
(actually ×2, since there are 2 copies of each chromosome "diploid")

$$\Rightarrow L \cdot \mu = 3 \times 10^9 \frac{\text{bp}}{\text{genome}} \times 10^{-8} \frac{\text{mutations}}{\text{bp} \cdot \text{gen}} \approx 30 \text{ mutations per genome per gen.}$$

$\Rightarrow$ there are $N \sim 10^{10}$ humans on earth, so

$\Rightarrow N \times \mu \sim 10^{10} \times 10^{-8} \sim 100$ mutations produced @ every site in human genome per generation (in some individual)

⇒ but, if we pick 2 random people & compare genomes:



Empirically: differ @ ~0.1% of genome

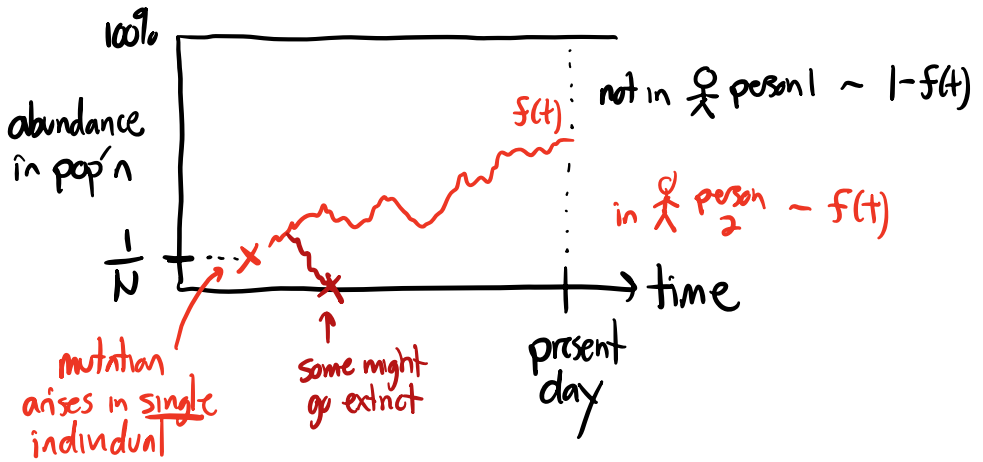
Question: what sets this scale? why not 10^{-4} or 10^{-2}?

⇒ one factor: dynamics over time

⇒ for mutations produced in previous generation,

$$\Pr\left[\begin{array}{c}\text{mut in person 1}\\ \text{or person 2}\end{array}\right] \approx 100\ \text{muts/bp} \times \frac{2}{10^{10}} \sim 10^{-8}/\text{bp}$$

⇒ many differences observed today occurred in past



$\Rightarrow$ must understand mut'n trajectories over time ("dynamics")

(goal of next several lectures...)

Another Fermi calculation:

* all single mutations produced every gen in humans

but all pairs of mutations are not:

$$\Pr\left[\begin{array}{c}\text{site 1 \& site 2}\\ \text{mutated in same}\\ \text{newborn}\end{array}\right] \sim N \times \mu \times \mu \sim 10^{10} \times 10^{-8} \times 10^{-8} \sim 10^{-6}$$

$\Rightarrow$ must wait $\sim 10^{6}$ gens (20 million yrs!) for a given pair of sites to mutate @ same time

$\Rightarrow$ Upshot: past dynamics even more important for combinations of mut'ns.

$\Rightarrow$ can also repeat same calculations for E. coli...

$\Rightarrow$ genome is $L = 4 \times 10^{6}$ bp + $\mu \sim 10^{-10}$/bp/gen

$\Rightarrow L \times \mu \sim 4 \times 10^{-4}$ mutations/genome/gen

$\Rightarrow$ >1000 replications before a single error!

$\Rightarrow N_g \sim 10^9 - 10^{10}$ E. coli cells in single person's gut

$\Rightarrow N_g \times \mu \sim 0.1 - 1$ (almost every bp mutated w/in us each day)

$\Rightarrow N_h \sim 10^{10}$ guts in human pop'n,

$\Rightarrow N_h \times N_g \times \mu \times \mu \sim 0.1 - 1 \Rightarrow$ almost all double mutations produced in worldwide E. coli pop'n each day

$\Rightarrow$ but <u>not</u> triple mutants $\left(10^{10} \times 10^{10} \times (10^{-10})^3 \ll 1\right)$

$\Rightarrow$ more generally, for single gene of $L \sim 1000$ bp

$\Rightarrow 4^L \simeq 10^{600}$ possible DNA sequences!

compare to $\sim 10^{82}$ atoms in universe

$\Rightarrow$ sequence space is <u>very</u> big (& sparsely populated)

What do mutations do?

"genotype ⇒ phenotype map"

⇒ in general, we don't know a priori (even for model organisms like E.coli!)

⇒ but in special cases, can make some guesses based on structure of the genetic code...

e.g. if mutation occurs in a gene:

⇒ changes a codon (e.g. ATC → ATT)

① due to degeneracy, codon could code for same AA

⇒ doesn't change protein — "synonymous mutation"

② could change to something else — "nonsynonymous mut'n"

↳ e.g. other AA (small change?) — "missense mut'n"

↳ e.g. stop codon ⇒ truncates gene (big change)

"loss-of-function" / "nonsense" mut'n