Last time:

Office hours today 12:30pm-2pm
Same zoom channel

* Approximations & self consistency

$$\underset{\substack{\uparrow\\ \text{small?}}}{\epsilon x^2} + \underset{\substack{\uparrow\\ \text{dominant?}}}{x - 1} = 0 \quad \Rightarrow \quad x \approx 1 \quad (\epsilon x^2 \approx \epsilon \ll 1)$$

$$\Rightarrow \epsilon \gg 1$$

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

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

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

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

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

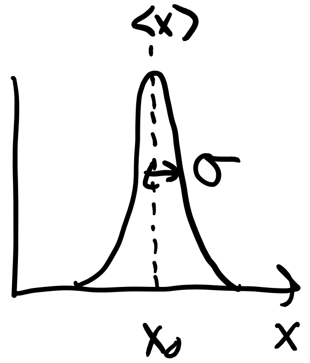
Today:

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

"Average" vs "typical"

2 main classes of behavior:

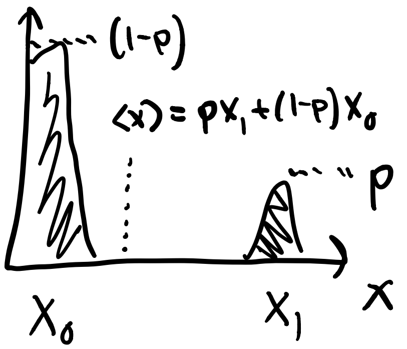
case 1:



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

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

case 2:



e.g. Binomial(N, p) whn $Np \ll 1$

e.g. did a mutation occur?

$\Rightarrow$ becomes important if we do something w/ x:

e.g. $y = F(x)$ = future growth of x # of mutations @ time 0.

$\Rightarrow$ in case 1: can use Taylor expansion:

$$y = F(x) = F\left(x_0 + \underbrace{(x-x_0)}\right) \approx F(x_0) + F'(x_0)(x-x_0)$$

$$x_0 \gg \sigma$$

$$\approx \boxed{F(x_0)} + F'(x_0)\,\sigma$$

"spread"

"deterministic"

case 2:

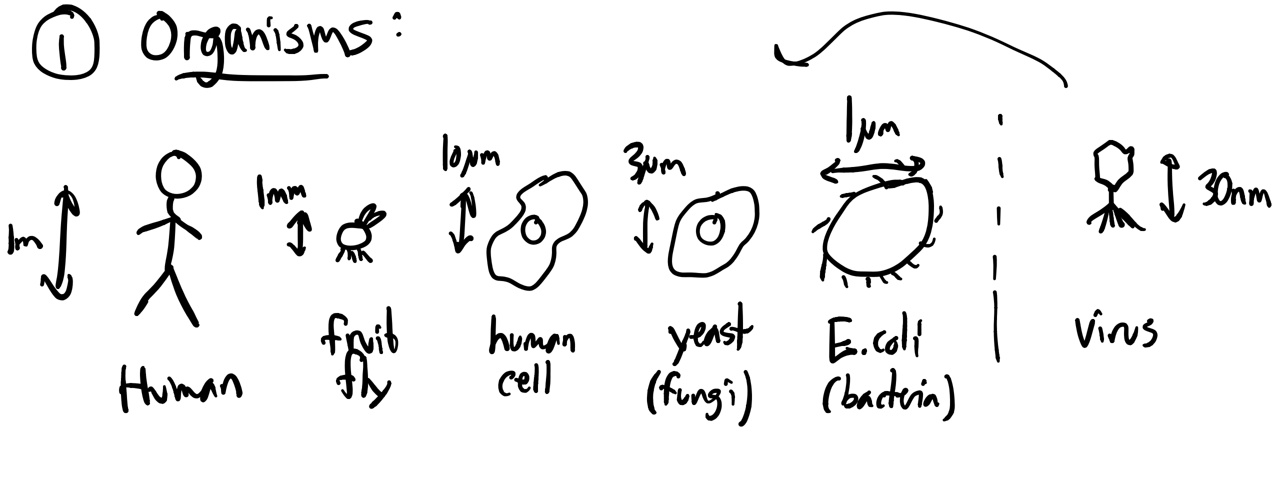
$$y = \begin{cases} F(x_0) & \text{w/ prob } 1-p \\ F(x_1) & \text{w/ prob } p \end{cases}$$

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

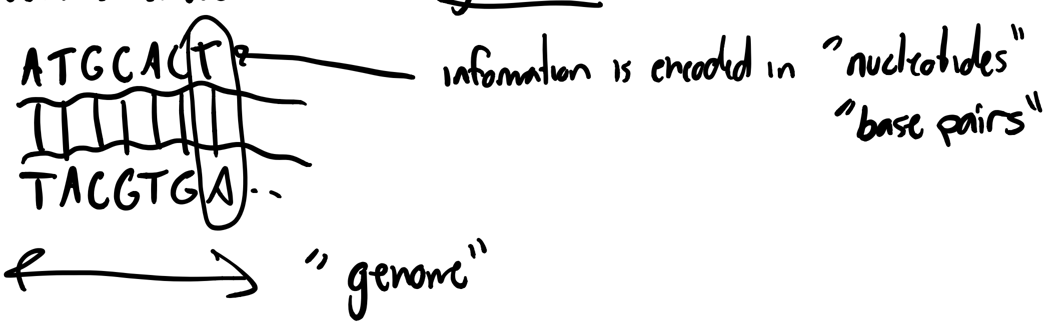
"rare event" happens separately.

Biological Background (#s / scales)

① Organisms:



instructions encoded in a single molecule of DNA:

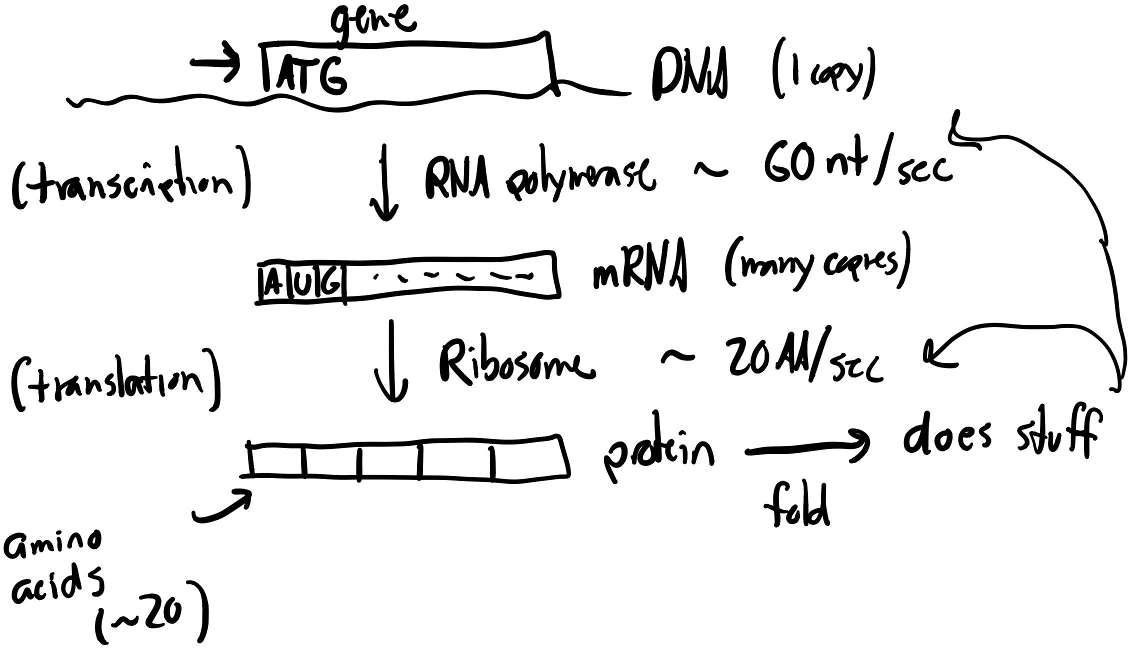


Lengths of genomes vary widely

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

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

information often encoded in genes (make proteins)



How does ribosome do it?

$\boxed{A}\boxed{T}\boxed{T}$ = "codon"

$\Downarrow$

1 amino acid
(isoleucine)

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

"genetic code"

⇒ has degeneracy

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

⇒ # of genes varies widely across organisms:

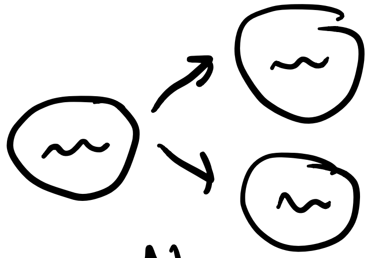
humans: 20,000 genes   yeast: 6,000 genes

E-coli ~ 4,000 genes   viruses ~ 10 genes.

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

⇒ rest of genome is "noncoding" → regulation
→ "junk"

("coding" = genes)



Δt

$\longleftrightarrow$

1 doubling time

1 generation

cell makes a copy of itself!

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

② needs to copy its DNA

(DNA polymerase)

(not usually limiting factor in growth)

humans: ~20yrs   E. coli ~ 20 mins – 1hr

humans: ~1 day (HeLa)   Prochlorococcus ~ 1 day (ocean bacteria)

Virus: → HIV ~ 13hrs

SARS-CoV-2 ~ 10hrs

⇒ when genome is copied, could introduce error ("mutations")

....ATGCCA.... parent

↓

....ATGTCA.... offspring

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

"single nucleotide mut", "substitions"

⇒ also "insertion" ATGTTTCA

↓

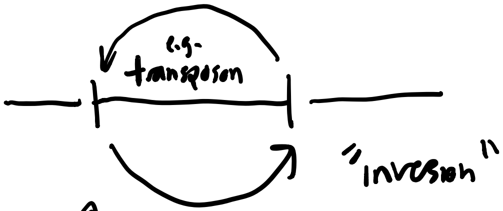
ATGTTTTTTCA "slippage of DNA pol"

⇒ or "deletion" ATGT~~T~~CA

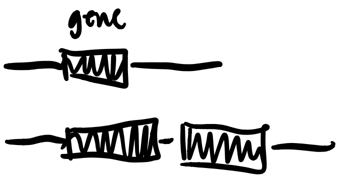
↓

ATGTCA

⇒ or "rearrangements"



⇒



"duplication"

⇒ often via special genes "transposons"

⇒ cells have machinery for fixing errors!

⇒ mutation rates vary across organisms! → (μ)

e.g. Humans : $\mu \sim 10^{-8}$/bp/gen. E.coli ~ $\mu \sim 10^{-10}$/bp/gen

human cells : $\mu \sim 10^{-10}$/bp/division "gen" Viruses ~ $\mu \sim 10^{-5}$/bp/gen

(SARS-CoV-2 10^{-6}/bp/gen)

Humans: genome is $L = 3\times10^{9}$ bp long + $\mu \sim 10^{-8}$/bp/gen

⇒ so 30 mutations per genome/gen.

⇒ ~ 10^{10} humans on earth + $\mu \times 10^{-8}$/bp/gen/individual

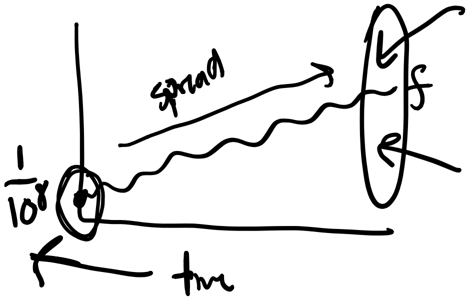
⇒ ~ 100 mutations produced @ every site in human genome every generation (in some individual)

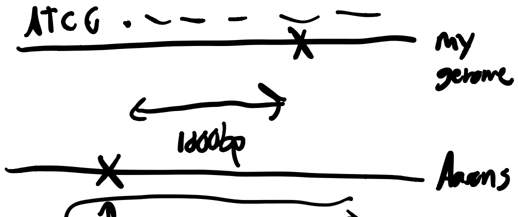
⇒ but if we pick 2 random people ⇒ differ @ 0.1% of genome.

why? why not 10^{-4}? 10^{-2}?

$$\boxed{\Rightarrow f = \frac{100}{10^{10}} \sim 10^{-8} \Rightarrow \text{differ @ } 10^{-8}?}$$

wrong!





$$\boxed{Pr[\text{diff}] = 10^{-3}} = f(1-f)$$

⇒ but not all double mutants: 10^{10} humans × 10^{-8} × 10^{-8} = 10^{-6}

⇒ sequence space is big!

E.coli: genome is ~ 4×10^{6} bp + $\mu \sim 10^{-10}$/bp/gen.

⇒ 4×10^{-4} mutations / genome / gen.

⇒ >1000 replications before single error!

in gut 10^{9} E.coli cells ⇒ so almost bp mutated every day w/in us.
(10^{10})

⇒ $\times 10^{10}$ guts ⇒ almost ~~a~~ all double mutants present in worldwide pop: $(10^{10}\times10^{10}\times10^{-10}\times10^{-10}\sim1)$

⇒ not triple: $\sim 10^{20}\times(10^{-10})^{3} \ll 1$.

$\ll 10^{82}$ atoms in universe!

⇒ more generally $L=1000$ bp ⇒ $4^{L}=4^{1000}=10^{602}$ possible gene sequences

---

what do mutations do? (genotype ⇒ phenotype map)

⇒ in general, we don't know (even for E.coli)

⇒ but in special cases, can guess based on genetic code.

e.g. if mutation occurs in a gene:

⇒ changes a codon ATC → ATT

① because of degeneracy, codons could code for same AA

⇒ doesn't change protein "synonymous mutations"

② could change to somethg else "nonsynonymous"

↳ other AA ("small change") ("missense mutations")

↳ stop codon → truncates the gene (big change)
("nonsense") ↓ loss-of-function