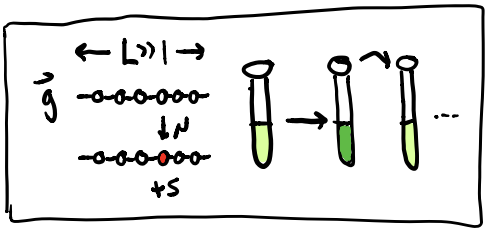
Announcements: ① PSET 3 DUE 3/9/21 (only 3 problems this time, ⇒ but #1 is long!)

② weekly comments

Last time: Multi-locus models of evolution ($L \gg 1$)



$\Rightarrow$

$$\frac{\partial f(\vec{g})}{\partial t} = \underbrace{\left[X(\vec{g}) - \bar{X}(t)\right] f(\vec{g})}_{\text{selection (nonlinear)}} + \underbrace{\sum_{\vec{g}'} M(\vec{g}' \to \vec{g}) f(\vec{g}') - M(\vec{g} \to \vec{g}') f(\vec{g})}_{\text{mutation (linear, "local")}}$$

$$+ \underbrace{\rho \sum_{\vec{g}_a, \vec{g}_b} T(\vec{g}_a, \vec{g}_b \to \vec{g}) f(\vec{g}) - \rho f(\vec{g})}_{\text{recombination (nonlinear, non-local)}}$$

$$+ \underbrace{\sqrt{\frac{f(\vec{g})}{N}}\,\eta(\vec{g}) - f(\vec{g}) \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g})}{N}}\,\eta(\vec{g}')}_{\text{genetic drift (stochastic)}}$$

No closed-form sol'n in the general case... $\Leftarrow$

$\Rightarrow$ Need asymptotic approximations (i.e., limits of N, L, ρ, M, etc...)

① Large N / "mean field" (genetic drift is negligible)

↳ ✗ inconsistent when $L \gg 1$! (e.g. $\frac{\#\text{individuals}}{\#\text{genotypes}} \ll 1$)

Today:

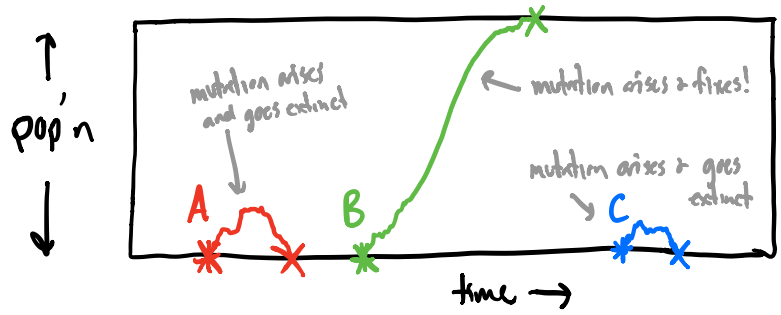
② Successive mutations regime

(i.e., mutation is small correction)

$$\frac{\partial f(\vec{g})}{\partial t} = \sim(x - \bar{x}) + \sim L\mu + \sim \rho + \sim \frac{\pi}{\sqrt{N}}$$

ϵ

⇒ i.e. new mutations fix or go extinct before next one occurs...



⇒ @ any given time, only 2 genotypes present:

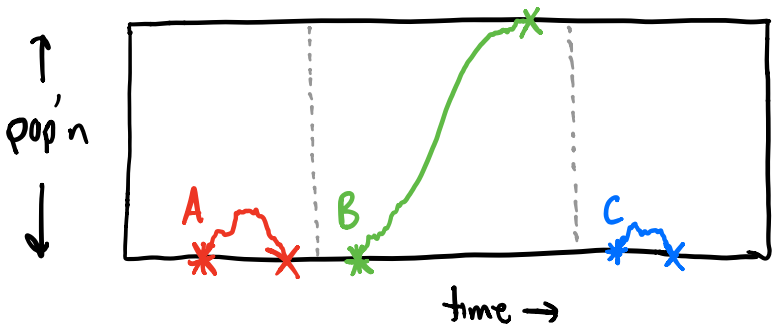
"current wildtype" $\vec{g}_0 = (1,0,1,1,0,0,0)$

mutation @ site ℓ

"single mutant" $\vec{g}_m = (1,0,1,1,0,1,0)$

⇒ what can recombination do? Nothing! (on average)

⇒ then each mutation looks like L=1 model w/



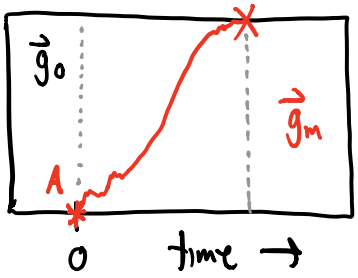
$s_{eff} \equiv X(\vec{g}_m) - X(\vec{g}_0)$

$\Rightarrow$ in this case, know exactly what happens:

(i) w/ prob $p_{fix} = \frac{2s}{1-e^{-2Nes}}$,

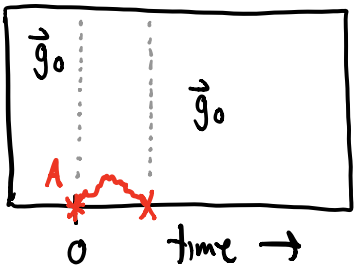
$\Rightarrow$ mutation fixes ("sweeps")

$\Rightarrow$ $\vec{g}_0 \to \vec{g}_m$; repeat!

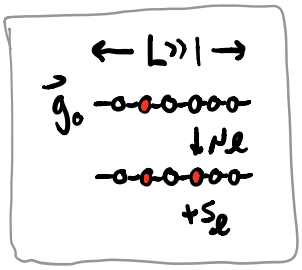


(ii) otherwise, mutation goes extinct

$\Rightarrow$ $\vec{g}_0$ stays put.

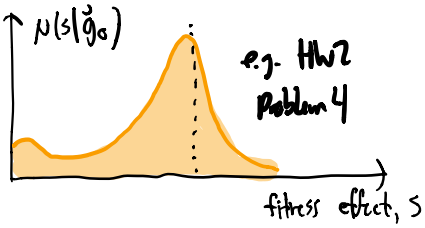


$\Rightarrow$ <u>New</u>: when $L>1$, multiple different mutations are possible...



$\Rightarrow$ useful to define fitness effects

$$s_\ell \equiv X\left(\vec{g}_0 + \begin{matrix}\text{mut'n} \\ \text{@ site } \ell\end{matrix}\right) - X(\vec{g}_0)$$



along w/ distribution of fitness effects ("DFE")

$$\nu(s|\vec{g}_0) \equiv \sum_{\ell=1}^{L} \nu_\ell \, \delta(s - s_\ell)$$

technically depends on $\vec{g}_0$

prob. of drawing a mutation w/ effect s.

$\Rightarrow$ mutations w/ fitness effect s are produced as Poisson process w/ rate $N\mu(s)$.

$\Rightarrow$ if each successful w/ prob $P_{fix}(s)$,

then successful mutations also Poisson Process

w/ total rate $R \equiv \int_0^\infty N\mu(s) \cdot P_{fix}(s)\, ds = \sum_\ell N\mu_\ell P_{fix}(s_\ell)$

$\Rightarrow$ (1) time until next successful mutation is born is:

$T_{est} \sim \text{Exponential}\left(\frac{1}{R}\right)$ [similar to $N\mu \to 0$ case in lecture 10]

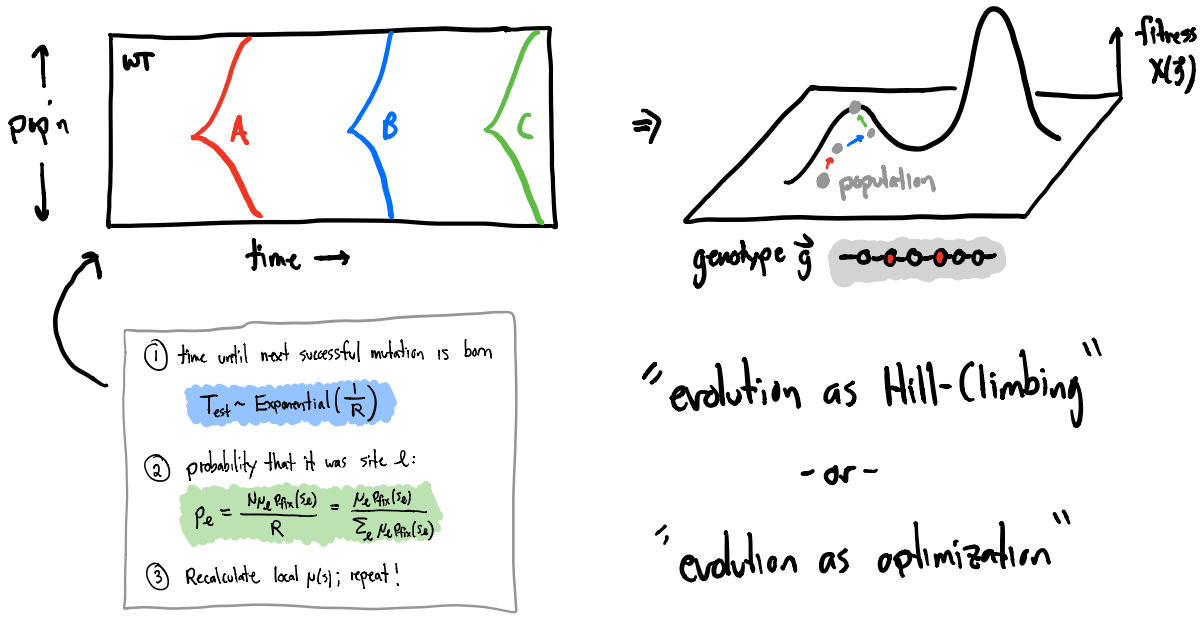
$\Rightarrow$ (2) probability that it was site ℓ:

$$p_\ell = \frac{N\mu_\ell P_{fix}(s_\ell)}{R} = \frac{\mu_\ell P_{fix}(s_\ell)}{\sum_\ell \mu_\ell P_{fix}(s_\ell)}$$

⇒ (3) $\vec{g}_0 \to \vec{g}_m$ ⇒ recalculate $\mu(s|\vec{g}_0)$ ⇒ repeat!

When approx is valid: will check carefully below...

⇒ simple algorithm for modeling evolution (not just pop gen)



① time until next successful mutation is born

$T_{est} \sim \text{Exponential}\left(\frac{1}{R}\right)$

② probability that it was site ℓ:

$p_\ell = \frac{N\mu_\ell \rho_{fix}(s_\ell)}{R} = \frac{\mu_\ell \rho_{fix}(s_\ell)}{\sum_\ell \mu_\ell \rho_{fix}(s_\ell)}$

③ Recalculate local $\mu(s)$; repeat!

"evolution as Hill-Climbing"

-or-

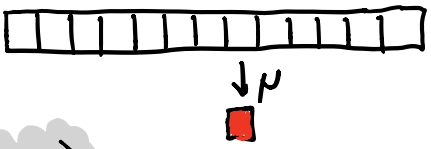
"evolution as optimization"

Note: even in these simplified settings,

⇒ fundamental limits to optimization picture...

Example: maintaining a useful function

Gene/Pathway:

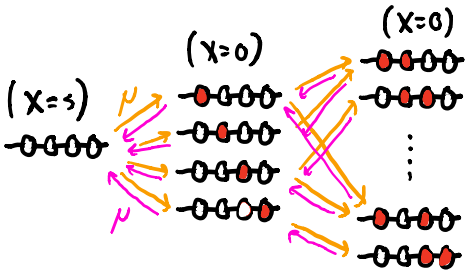


L ways to "break" (each w/ rate μ)

(e.g. for eating some low-level nutrient)

$X(\vec{0}) \equiv s;\ X(\text{else}) \equiv 0$

Genotype network:

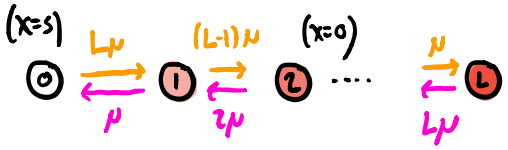


+ L levels

Key simplification:

"equivalence class"

$$f_k \equiv \sum_{|\vec{g}|=k} f(\vec{g})$$



↳ can track:

$$P_k(t) \equiv \Pr\left[f_k(t)=1\right] = \left\langle f_k(t) \right\rangle$$

How can $P_k(t)$ change?

$$\underline{k=0}: \quad \partial_t P_0 = \overbrace{N\mu\, \rho_{fix}(s) P_1}^{\text{incoming}} - \overbrace{NL\mu \cdot \rho_{fix}(-s) \cdot P_0}^{\text{outgoing}} \xrightarrow{t\to\infty} 0$$

$$\Rightarrow \frac{P_0}{P_1} = \frac{N\mu\, \rho_{fix}(s)}{NL\mu\, \rho_{fix}(-s)} = \frac{1}{L} e^{2Ns}$$

$$\underline{k=1}: \quad \partial_t P_1 = NL\mu \rho_{fix}(-s) P_0 - N\mu \rho_{fix} P_1 \qquad \text{(from 0 class)}$$
$$+ N \cdot 2\mu \cdot \left(\tfrac{1}{N}\right) P_2 - N(L-1)\mu \left(\tfrac{1}{N}\right) P_1 \xrightarrow{t\to\infty} 0$$

$$\Rightarrow P_2 = \frac{(L-1)}{2} P_1$$

$$\underline{k=2}: \quad \Rightarrow P_3 = \frac{L-2}{3} \cdot P_2 = \frac{(L-1)(L-2)}{3 \cdot 2} \cdot P_1$$

$$\Rightarrow P_k = \frac{1}{L} \frac{L!}{k!(L-k)!} P_1$$

$$\Rightarrow 1 - P_0 = \sum_{k=1}^{L} P_k = \frac{1}{L}(2^L - 1) P_1$$

$$\Rightarrow \boxed{f_1 = (1-f_0)\frac{L}{(2^L-1)}}$$

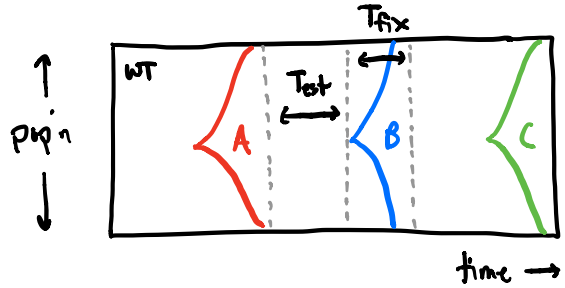
$$\Rightarrow \boxed{\frac{P_0}{1-P_0} = \exp\left[\underbrace{2Ns}_{\text{"drift barrier"}} - \underbrace{\log(2^L-1)}_{\text{"entropy of genotype space"}}\right]} = \frac{\text{prob having function}}{\text{prob broken}}$$

$$\geq \frac{0.9}{0.1} = \exp[2] \leq \exp\left(2Ns - \overbrace{\log(2^L-1)}^{\alpha L}\right)$$

$$\boxed{s \geq \frac{1}{N}} \quad \text{"drift barrier"}$$

↳ e.g. compare to deterministic case: $\bar{f}_0 = 1 - \frac{L\mu}{s}$

When is successive mutations regime a good approx?



$\Rightarrow$ need $T_{est} \sim \frac{1}{R} \ll T_{fix}$

E.g. Neutral mutations $(N(s) = U_n \delta(s))$

$$P_{fix}(0) = \frac{1}{N} \Rightarrow R = NU_n\left(\frac{1}{N}\right) = U_n \; ; \; T_{fix} \sim \mathcal{O}(N)$$

$$\Rightarrow \text{need } \frac{1}{U_n} \gg N \Rightarrow \boxed{NU_n \ll 1}$$ "weak mutation weak selection"

E.g. Strongly beneficial mutations $(N(s) = U_b \delta(s - s_b) \; ; \; Ns_b \gg 1)$

$$\Rightarrow P_{fix}(s) \approx 2s \Rightarrow R = 2NU_b s \; ; \; T_{fix} = \frac{2}{s}\log(Ns)$$

$$\Rightarrow \text{need } \frac{1}{2NU_b s} \gg \frac{2}{s}\log(Ns) \Rightarrow \boxed{NU_b \ll \frac{1}{\log(Ns_b)}} \ll 1$$

$\Rightarrow$ What does this look like for some "real" parameter values?

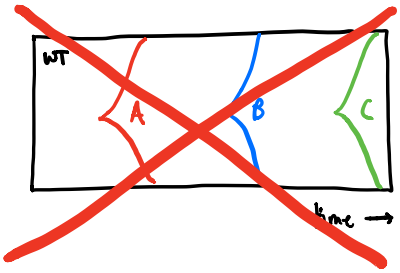
$\Rightarrow$ e.g. HW2 problem 4

$U_b \sim 5\times10^{-6}$, $s_b \sim 0.02$

just for L.O.F. muts.

$\Rightarrow$ e.g. if $N \sim 10^5$ $\Rightarrow$ $2NU_b \log(Ns) = 3$

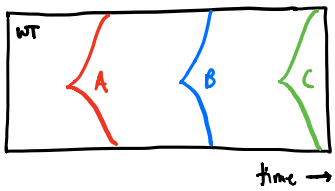
$\Rightarrow$ successive mutations picture does not apply!



$\Rightarrow$ what do things look like instead?

Neutral theory & the Coalescent

Successive mutations:



$$\frac{ds(\vec{y})}{dt} = \sim(x-\bar{x}) + \sim L\cdot\mu + \epsilon + \sim\rho + \sim\frac{\pi}{\sqrt{N}}$$

⇒ ~1 variant present @ high freqs ⇒ solved by reducing to L=1 model

* But genomes in data separated by multiple mut'ns
(e.g. humans, 2 individuals differ by ~1 mut / 1000 bp)

⇒ need to understand what's going on in these cases...

$$\frac{ds(\vec{y})}{dt} = \sim(x-\bar{x}) + 0 + \sim L\cdot\mu + \sim\rho + 0 + \sim\frac{\pi}{\sqrt{N}}$$

⇒ one other limit that's well understood:

neutral evolution in nonrecombining genome

when $X(\vec{g}) = 0$ + $\varepsilon = 0$, left with: $(\mu_\ell = \nu_\ell)$

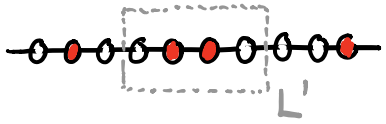
$$\frac{df(\vec{g})}{dt} = \sum_{|\vec{g}'-\vec{g}|=1} \sum_\ell \mu_\ell f(\vec{g}') \left[g_\ell (1-g_\ell') + (1-g_\ell) g_\ell' \right] - \sum_\ell \mu_\ell f(\vec{g})$$

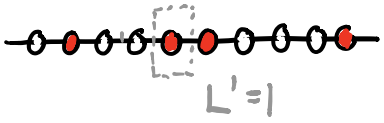
incoming mutations — outgoing mutations

$$+ \sqrt{\frac{f(\vec{g})}{N}} \eta(\vec{g}) - f(\vec{g}) \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g}')}{N}} \eta(\vec{g}')$$

genetic drift

Key insight: sites don't actually influence each other (because neutral)

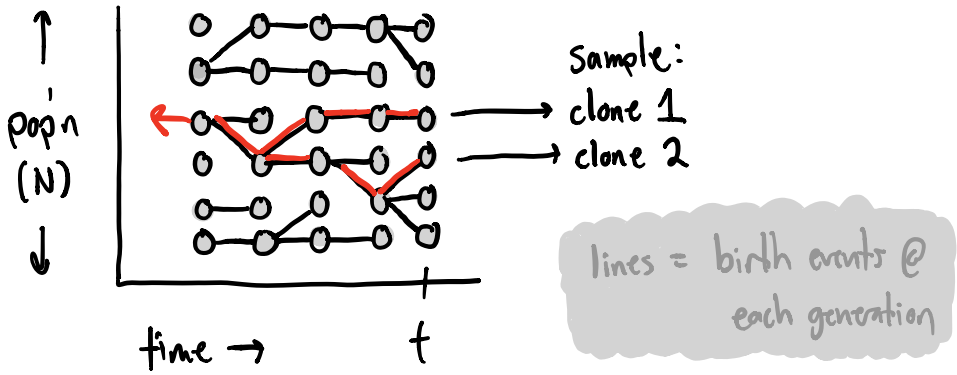
e.g. 



$p(f) = \frac{2N\mu}{f}$

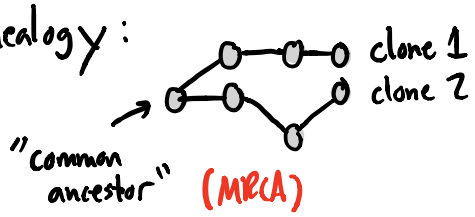
$\Rightarrow$ 2nd key insight: can take $L' = 0$ —

E.g. simulation of neutral pop'n in Wright-Fisher model:

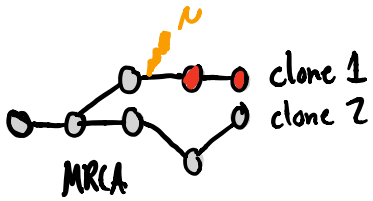


⇒ key insight: lines also = genealogical relationships backward in time!

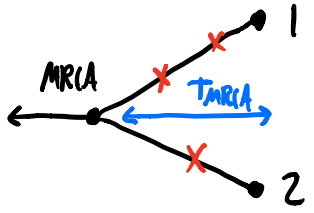
Genealogy:



⇓ differences between sampled individuals = mutations on genealogy



⇒ Mutations occur as Poisson Process w/ rate μ_ℓ on each branch



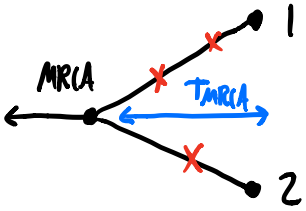
↳ total # of muts Poisson$(2\mu_\ell T_{MRCA})$

⇒ 2 extreme limits:

(1) $\mu_\ell T_{MRCA} \ll 1$ ⇒ 0 or 1 mutations on whole tree

⇒ $Pr[\text{genetic diff @ site } \ell] =$

(2) $\mu_\ell T_{MRCA} \gg 1$ ⇒ lots of forward & backward mutations along each branch.

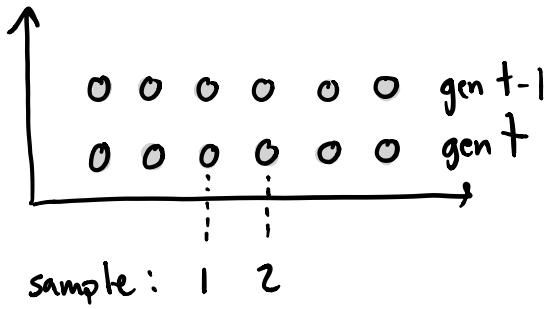


⇒ $Pr[\text{genetic diff @ site } L] =$

⇒ Key Question: what sets T_{MRCA}?

Note: T_{MRCA} is random quantity (genealogy will differ from simulation to simulation)

$\Rightarrow$ key insight: start from present & work backward in time:



Process repeats itself w/ next gen: