Linked selection & clonal interference

Last time, we saw that in sufficiently large populations (even recombining ones!) we will eventually reach a point where $NU_{b,eff} \cdot s \cdot \frac{1}{s}\log(Ns) \gg 1$

where sweeping beneficial mutations will interfere w/ each other. ⟹ this regime has historically been very challenging to analyze

"clonal interference" / "Hill robertson interference"

⟹ Most progress came only recently, w/ big contribution from physicists [e.g. Tsimring et al 1996, Rouzine et al 2003, Desai & Fisher 2007, ...]

⟹ analytical progress enabled by taking step back & thinking about very simple "staircase model" of asexual evolution: all beneficial mutations provide same benefit s, occur @ total rate $U_b = L\lambda_b N$ & never run out

⟹ all individuals can be charactrized by # of mutations that they have (k) ⟹ fitness $X(k) = ks$.

⟹ rather than keep track of genotypes (1,0,0,1,00,1,0) can coarse grain over all genotypes w/ same k:

$$f(k,t) \equiv \sum_{\vec{g}\,:\,|\vec{g}|=k} f(\vec{g}) \quad = \text{"fitness class } k\text{"}$$

(fraction of individuals w/ fitness sk)

$\Rightarrow$ in terms of fitness, dynamics of population can be equally well described by $\{f(k)\}$, rather than $\{f(\vec{g})\}$

$\Rightarrow$ "fitness distribution"

the fitness classes satisfy the coarse-grained SDE

$$\frac{df(k)}{dt} = s(k-\bar{k})f(k) + U_b\left[f(k-1)-f(k)\right] + \sqrt{\frac{f(k)}{N}}\eta(k) - f(k)\sum_{k'}\sqrt{\frac{f(k')}{N}}\eta(k')$$

$\Rightarrow$ big simplification: now 1 dimensional system, rather than 2^L dimensional

$\Rightarrow$ may be tempting to drop noise terms "for large pop'n s"

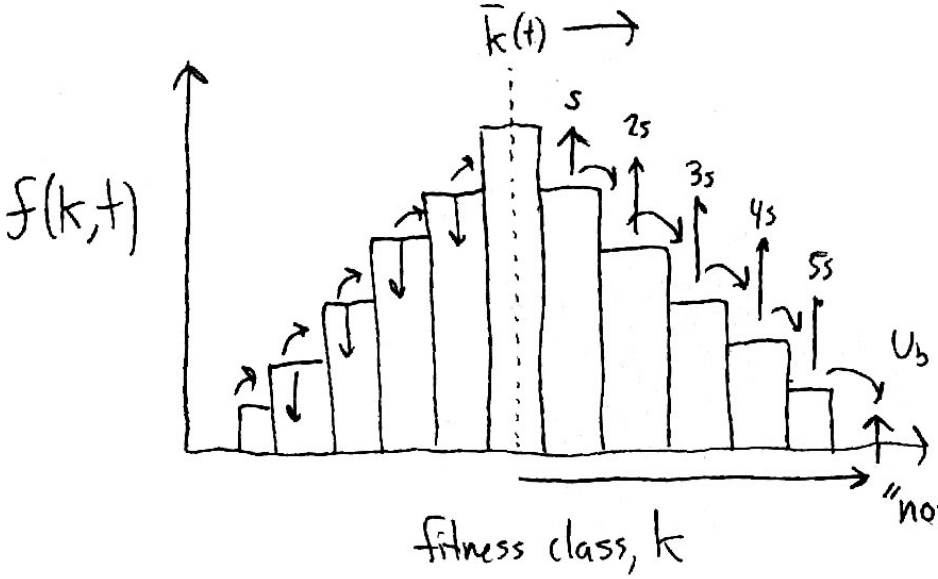
$\Rightarrow$ this leads to nonsense results:

$$f_{det}(k,t) = \frac{\left[\frac{U_b}{s}(e^{st}-1)\right]^k}{k!} e^{-\frac{U_b}{s}(e^{st}-1)}$$

$\longrightarrow$ gains fitness exponentially fast

$\Rightarrow$ problem: "fractional" individuals produced by mutations grow very fast and dominate $f_{det}(t)$, even if they don't typically occur.

Instead, if we simulate this model, we find that the population fitness distribution develops into a "travelling wave" w/ a fixed profile that translates to higher fitness @ a fixed rate $v \equiv \frac{d\langle \bar{x}(t) \rangle}{dt} \equiv s \frac{d\langle \bar{k}(t) \rangle}{dt} \equiv \frac{s}{\tau}$ ← typical time it takes $\bar{k}(t)$ to increase by 1.



"nose", "lead", q fitness classes above mean.

How can we understand this behavior?

⇒ will present heuristic analysis based on Desai & Fisher 2007

⇒ analysis will apply for asymptotic regime where:

$Ns \gg 1$, $U_b \ll s_b$, $s\tau \gg 1$ (will understand this one later)
$q \gg 1$

⇒ together imply that mutations are only important for establishing new nose class. (when $f_q(0)=0$)
after that selection much more important for growth of fitness class.

also imply that most of population is concentrated near $\bar{k}$:

$\Rightarrow$ from SDE: $\frac{d\langle \bar{k} \rangle}{dt} = \sum_k (k-\bar{k})^2 f(k,t) > s$

↳ $\frac{!}{\tau}$ $\Rightarrow$ $Var(k) = \frac{1}{s\tau} \ll 1$

$\Rightarrow$ then implies that $\bar{k}(t)$ changes very abruptly during each step:

$\Rightarrow$ @ steady state, we know that $f_{\bar{k}+1}(\tau) = f_{\bar{k}+1}(0)e^{s\tau} = f_{\bar{k}}(0)$

because mean fitness has increased by 1:

$\Rightarrow$ if $\bar{k}(t)$ is dominated by these 2 classes, then

$\bar{k}(t) = \bar{k}(0) + f_{\bar{k}+1}(t) = \bar{k}(0) + e^{s(t-\tau)} / (1 + e^{s(t-\tau)})$

$\Rightarrow$ $\bar{k}(t) \to \bar{k}(t)+1$ w/in $\mathcal{O}(\frac{1}{s})$ of τ (which is $\gg \frac{1}{s}$ by assumption)

$\Rightarrow$ thus, for most of time in $t \in (0,\tau)$, $\bar{k}(t) \approx \bar{k}(0)$

this gives us all the tools we'll need to understand the behavior of travelling wave @ steady state.

(1) Let's focus on establishment of a new class @ the nose:

(a) Mutations are produced by fittest occupied class ($f_{-1}(t)$) @ rate $NU_b f_{-1}(t)$. these mutants establish w/ prob ~ qs, and will grow as $f_k(t) \approx \frac{1}{Nqs} e^{qs(t-\tau_k)}$

where τ_k is establishment time of k^{th} successful mutant.

$\Rightarrow$ we can define an establishment time for the whole class as $f_0(t) \equiv \frac{1}{Nqs} e^{qs(t-\tau)} \equiv \sum_{k=1}^{K_{max}} \frac{1}{Nqs} e^{qs(t-\tau_k)}$

$\Rightarrow$ by construction, this is same as click time @ steady state.

(b) since fittest occupied class was created by such an establishment event in the last click, we must have $f_{-1}(t) = \frac{1}{Nqs} e^{(q-1)st}$. this allows us to solve for the time that the k^{th} successful mutation arises:

using same logic as last class:

$$k \sim \int_0^{\tau_k} N U_b f_{-1}(t) \cdot qs \, dt \approx \int_0^{\tau_k} U_b e^{(q-1)st} \, dt \simeq \frac{U_b}{(q-1)s}\left(e^{(q-1)s\tau_k} - 1\right)$$

$$\Rightarrow \tau_k \approx \frac{1}{(q-1)s} \log\left(\frac{s(q-1)k}{U_b}\right) \quad \left(\text{assuming } (q-1)s\tau_k \gg 1\right)$$

$$\Rightarrow f_k(t) = \frac{1}{Nsq} e^{qst} \left(\frac{s(q-1)k}{U_b}\right)^{-1-\frac{1}{q-1}} \approx \frac{e^{st}}{Nqs}\left(\frac{sqk}{U_b}\right)^{-1-1/q}$$

extra little bit matters a lot again.

$\Rightarrow$ we can then solve for the establishment time for the whole class:

$$f_0(t) = \frac{1}{Nqs} e^{qs(t-\tau)} \equiv \frac{1}{Nqs} e^{sqt} \sum_{k=1}^{K_{max}} \left(\frac{sqk}{U_b}\right)^{-1-1/q} \approx \frac{e^{st}}{Nqs}\left(\frac{U_b}{s}\right)$$

$$\Rightarrow \tau = \frac{1}{qs} \log\left(\frac{s}{U}\right) \Rightarrow \text{if we know } q, \text{ can solve for click time.}$$

$$\Rightarrow \text{know that } 1 \approx f_k(0) = \frac{1}{Nqs} e^{(q-1)s\tau + (q-2)s\tau + \cdots + s\tau} = \frac{1}{Nqs} e^{\frac{q^2 s\tau}{2}}$$

$\Rightarrow$ system of 2 equations for τ & q:

$$q \approx \frac{\log(Nqs)}{\log(s/U)}, \quad \tau = \frac{1}{qs}\log(s/U)$$

$$\Rightarrow \boxed{\tau = \frac{\log^2(s/U)}{s\log(Ns)}, \quad q \approx \frac{\log(Ns)}{\log(s/U)}}$$

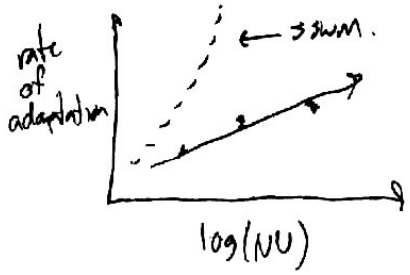
$\Rightarrow$ or in terms of rate of adaptation:

$$\frac{d\langle \bar{k} \rangle}{dt} = \frac{s\log(Ns)}{\log^2(s/U)}$$

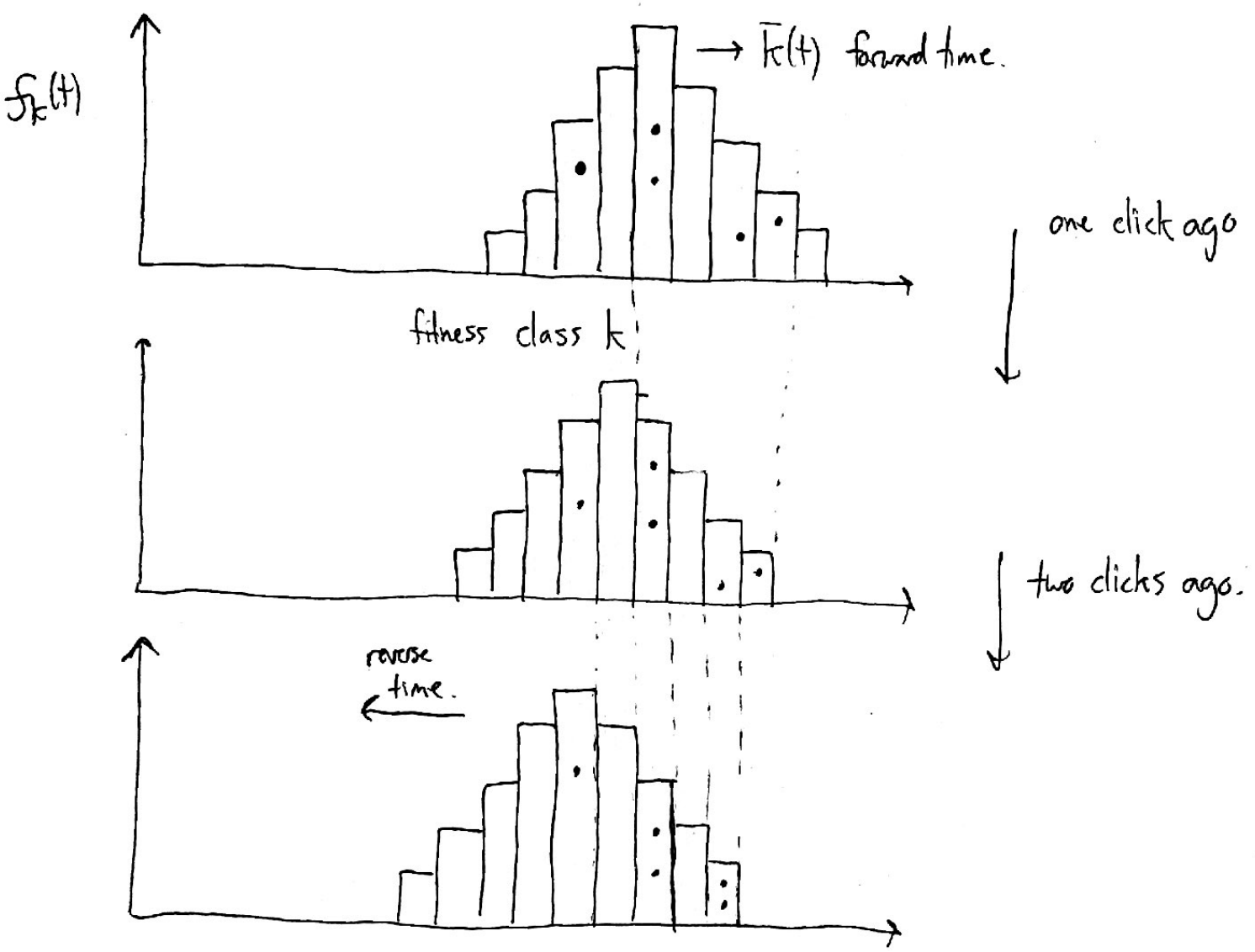
[compare to SSWM regime: $\frac{d\langle \bar{k} \rangle}{dt} = NUs$]

$\Rightarrow$ rate of adaptation linear in s, but now only logarithmic in population size (N) or supply of mutations. this was early test of whether clonal interference was relevant in lab evolution experiments:

e.g. Miralles et al 2000, de Visser et al 2003
Desai et al 2007, & others...

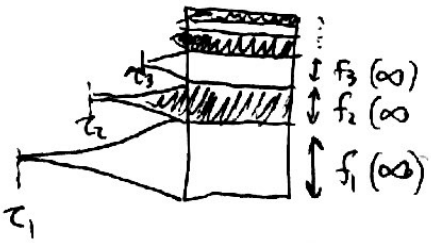


Now that we understand forward time dynamics of fitness distribution, can we understand backward in time dynamics of linked neutral genealogies? consider a sample from present day:



(1) no appreciable chance of coalescing until in f_0 class $\left(\frac{\tau}{N f_{-1}} \ll 1\right)$
→ since mostly growing deterministically, few new mutations.

(2) once in f_0 class, can coalesce if both individuals share the same founding mutant lineage:
(just like in classic sweep case)



where $f_k(\infty) = \dfrac{\frac{1}{Nqs} e^{qs(t-\tau_k)}}{\frac{1}{Nqs} e^{qs(t-\tau)}} = (qk)^{-1-1/q}$

↳ biggest mutant lineage is size $1/q$.

$\Rightarrow$ probability of 2 individuals coalescing in a single click is

$$p_c = \sum_{k=1}^{\infty} f_k(\infty)^2 \approx \int_1^{\infty} (qk)^{-2(1-1/q)} dk \approx 1/q^2$$

(note, not $1/q$ because lineage sizes are skewed)

$\Rightarrow$ however this is missing key part of puzzle.

- our argument has focused on sizes of typical first mutant lineages. But there is always a probability that the first lineage happens anomalously early.

$\Rightarrow$ since selection is strong, doesn't have to be that early to have huge effect on $f_1(\infty)$.

$\Rightarrow$ e.g. if $\tau_1 = \tau$, then first lineage will occupy >50% of new nose class. high probability of coalescing.

probability of such an event is

$$p_c^* \approx \int_0^{\tau} NU_b f_{-1}(t) \cdot qs\, dt \approx \frac{U_b}{qs} e^{(q-1)s\tau} \approx \frac{1}{q} \gg \frac{1}{q^2}$$

thus, typically have coalescence after q clicks

$$\Rightarrow T_c \sim q\tau = \frac{1}{s}\log\left(\frac{s}{U}\right) \quad \left[\text{roughly independent of } N\right]$$

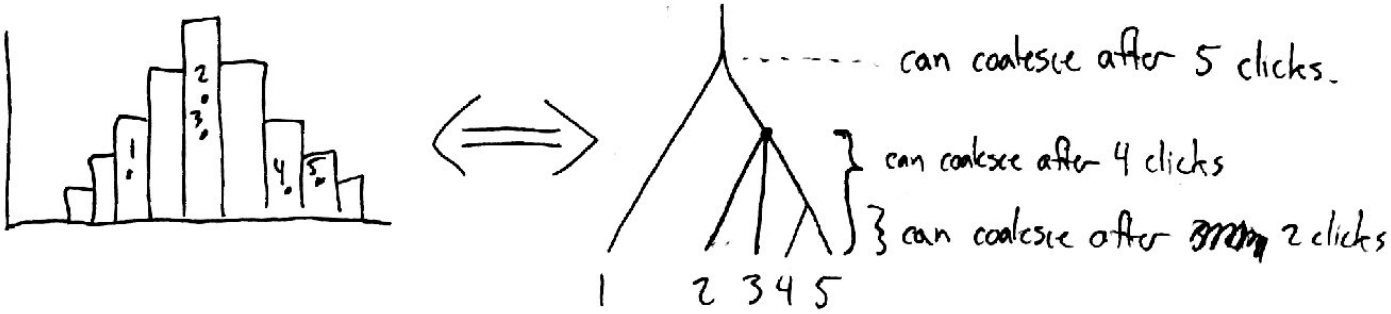
[roughly the time it takes for fitness dist'n to move forward by it's width]

$\Rightarrow$ but again, coalescence is bursty: once large fluctuation happens to get pairwise coalescence, not that much less likely to get coalescence between lots of lineages:

$$p_c(n) \approx \int_0^{\tau - \frac{1}{qs}\log(n)} NU_3 f_{-1}(t) \cdot qs\, dt \sim 1/qn \Rightarrow$$ multiple merger coalescent

(technically, Bolthausen–Sznitman)

$\Rightarrow$ another interesting feature of coalescence in travelling wave:



$\Rightarrow$ time of coalescence (+ burstiness of branches) in past provides information about relative fitnesses of lineages today

$\Rightarrow$ relative fitness today provides information about which lineages are most likely to take over in the future

$\Rightarrow$ Neher & Shraiman (Elife 2014) proposed a clever method to make this intuition precise: they use shape of genealogy of influenza strains today to predict which clades are most likely to take over in following season.

$\Rightarrow$ works pretty well $\Rightarrow$ performance comparable to or better than vaccine strains selected by hand (in terms of genetic distance)

$\Rightarrow$ coalescent properties are "universal" in that surprisingly insensitive to precise details of model: (1) variation in fitness effects $U_{be}(s)$

(2) $s \ll U_b$ (infinitesimal limit) [Neher & Shraiman 2013]

Good et al 2012
Fisher 2013

(3) addition of recombination - will examine here: (Weissman & Hallatschek 2014)

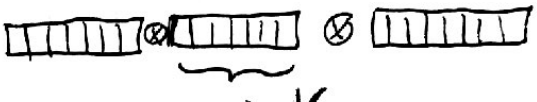
$\Rightarrow$ Idea is similar to "linkage block" ansatz from last lecture:

All mutations w/in $\ell^* \sim \frac{1}{rT_c}$ of each other will not recombine away from each other w/in a typical coalescence timescale

$\Rightarrow$ effectively asexual on this length scale.

$\Rightarrow$ need additional assumption that mutations w/ $\ell \gg \ell^*$ will recombine many times during fixation timescale (selection timescale) and will therefore act independently (this happens to be true here, but not always $\Rightarrow$ active area of research)

Under these assumptions, genome breaks up into collection of independent asexual linkage blocks, each w/ size ℓ^*



$\Rightarrow$ How to solve for ℓ^*?

$\Rightarrow$ w/in block, asexual, so $U_{eff} = N\lambda_b \frac{1}{rT_c}$, $T_c \sim \frac{1}{s}\log\left(\frac{s}{U_{b,eff}}\right)$

$\Rightarrow T_c \sim \frac{1}{s}\log\left(\frac{sr}{N\lambda_b} T_c\right) \sim \frac{1}{s}\log\left(\frac{r}{N\lambda_b}\log\left(\frac{r}{N\lambda_b}\right)\right)$

$\ell^* \sim \frac{s}{r}\frac{1}{\log\left(\frac{r}{N\lambda_b}\right)}$

T_c weakly dependent on r (unlike in selective sweeps case) no dependence on N.

$\Rightarrow$ self consistent when $NU_{eff} \log(Ns) \gg 1$ (lots of CI w/in block)

$(Ns)\cdot\frac{N}{r}\cdot\lambda_b\cdot\frac{\log(Ns)}{\log\left(\frac{r}{N\lambda_b}\right)} \gg 1 \Rightarrow \boxed{Ns\cdot\frac{N}{r}\cdot\lambda_b \gg 1}$ ✓