Stochastic processes and inference in population genetics

explosion of sequencing data

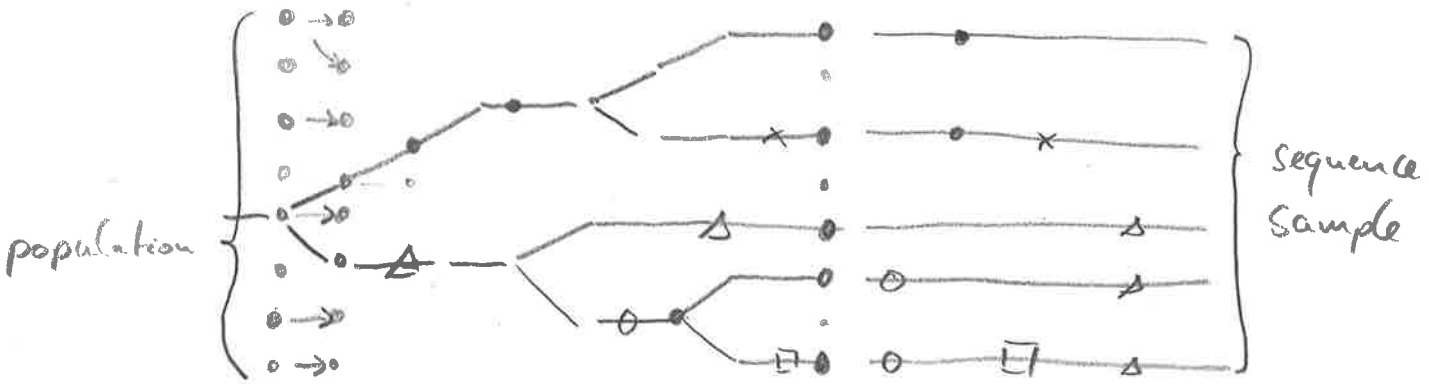
human genome:	$\sim 3\times 10^9$ bases	$\sim \$1000$
bacterial genome:	$\sim 3\times 10^6$ bases	$\sim \$10$
virus genome:	$\sim 10^4$ bases	

→ 1000s of human genomes are sequenced every day

→ what do we learn from all those data?

Population genetics

what determines and shapes genetic diversity?



- locally, sequences are related by a tree
- genetic diversity depends on the tree
- the tree depends on how one generation seeds the next
- several simple models are used as null models for inference

Coalescent model

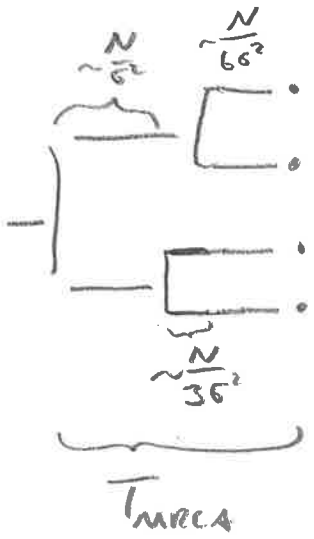
- backwards in time
- explicit model of trees
- offspring distribution O_m (const pop $\Leftrightarrow$ $\langle m \rangle = 1$)
- a randomly chosen pair has the same parent with prob:

$$N \sum_m \frac{m(m-1)}{N^2} O_m = \frac{\sigma^2}{N}$$

- sample of k $\rightarrow$ $\frac{k(k-1)}{2}$ possible pairs
- $\rightarrow$ merges with rate $\frac{\sigma^2 k(k-1)}{2N}$

  $\rightarrow$ $k \rightarrow k-1$

  $\rightarrow \rightarrow 1$



- $\langle T_{MRCA} \rangle = \frac{2N}{\sigma^2} \sum_{k=2}^{n} \frac{1}{k(k-1)}$

$$= \frac{N}{\sigma^2}\left[1 + \frac{1}{3} + \frac{1}{6} \dots\right]$$

$$= \frac{2N}{\sigma^2} \frac{n-1}{n} \qquad n = \text{sample size}$$

- average pairwise difference: $2\mu \langle T_2 \rangle = 2\mu N$
- # of mutations in sample: $\mu \langle T_{total} \rangle = \frac{N}{\sigma^2} \sum_k k T_k$

$$= \frac{N}{\sigma^2} \sum \frac{1}{k-1}$$

$$\simeq \frac{N}{\sigma^2} \log n$$

Diffusion models

- a dual process - forward in time
- considers the frequency of mutations
- if a mutation has freq. p_t at time t, what is p_{t+1}
  → depends on offspring distribution
  → here, we assume the mutation has no effect

$$\langle p_{t+1} \rangle = \langle p_t \rangle \qquad \langle \Delta p_{t+1} \rangle = \frac{\sigma^2}{N} p_t (1-p_t)$$

→ essentially binomial resampling of the mutation.

- $\frac{\partial}{\partial t} P(p,t) = \frac{\sigma^2}{2N} \frac{\partial}{\partial p^2} p(1-p) P(p,t)$
- separation of variables: $P(p,t) = a(p)\, b(t)$

$$\rightarrow \frac{1}{b(t)} \frac{\partial}{\partial t} b(t) = -\lambda = \frac{\sigma^2}{2N a(p)} \frac{\partial}{\partial p^2} p(1-p)\, a(p)$$

→ hypergeometric equation
→ BC $a(0)$ & $a(1)$ finite → discrete spectrum

$$\lambda = \frac{\sigma^2 k(k+1)}{N}$$ Same as coalescent rate !

$$P(p,t) = \sum_k c_k T_{k+1}(p)\, e^{-\frac{\sigma^2 k(k+1)}{N} t}$$

↳ Gegenbauer polynomial.

Branching processes

- stochastic dynamics of rare mutations

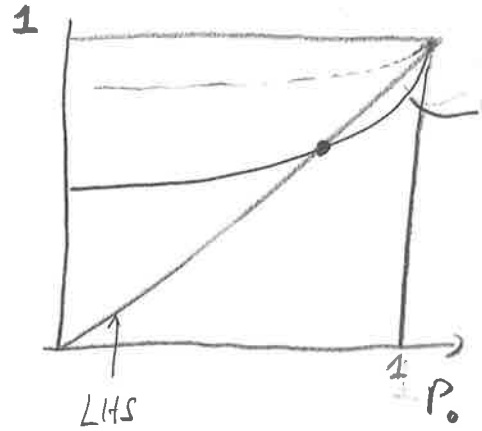
Q: how frequently is a beneficial mutation lost?

assume offspring distribution is Poisson with mean $1+s$

$$P_0 = \sum_{m=0}^{\infty} \frac{e^{-(1+s)}}{m!} (1+s)^m P_0^m = e^{-(1+s)(1-P_0)}$$

P_0: loss if all offspring are lost

graphic solution



- slope at $P_0 = 1$: $1+s$
- for $s>0$, there is a non trivial solution
- $s \ll 1$: fixation $p_{fix} = 1 - P_0$

$$p_{fix} \simeq 2s$$

- more generally: what is the distribution of the copy numbers n at time t of a clone with birth rate b & death rate d?

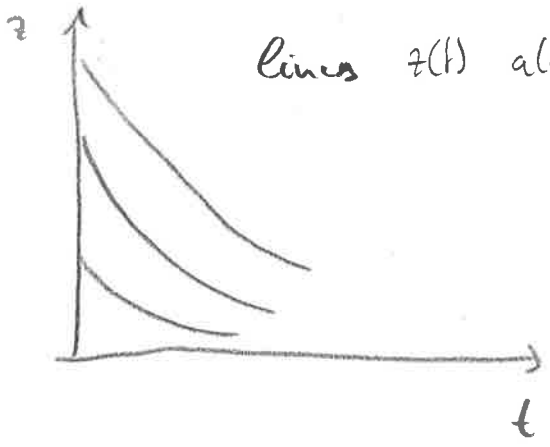
$$P(n, t+\delta t) = (1 - \delta t (b+d) n) P(n,t) + (n-1) b \delta t P(n-1,t) + (n+1) d \delta t P(n+1,t)$$

$$\Rightarrow \frac{\partial}{\partial t} P(n,t) = -(b+d) n P(n,t) + b(n-1) P(n-1,t) + d(n+1) P(n+1,t) \quad \Big| \times \sum_n z^n$$

$$\frac{\partial}{\partial t} \underbrace{\sum_{n=0}^{\infty} z^n P(n,t)}_{\text{generating function}} = \frac{\partial}{\partial t} \phi(z,t) = -(b+d) z \sum_{n=0}^{\infty} \partial_z z^n P_n + b z^2 \sum_{n=0} \partial_z z^{n-1} P_{n-1} + d \sum_n \partial_z z^{n+1} P_{n+1}$$

$$= \left[-(b+d) z + b z^2 + d \right] \partial_z \phi$$

- PDE that can be solved by characteristics



lines $z(t)$ along which $\phi(z,t)$ is constant

$$\frac{d}{dt}\phi(z(t),t) = \frac{dz}{dt}\frac{\partial\phi}{\partial z} + \frac{\partial\phi}{\partial t}$$

$$= \frac{\partial\phi}{\partial z}\left[\frac{dz}{dt} - (b+d)z + bz^2 + d\right]$$

$$= 0$$

the characteristic equation simplifies when substituting

$x = 1 - z$

$$\frac{dx}{dt} = -(b+d)(1-x) + b(1-x)^2 + d$$

$$= -(b-d)x + bx^2$$

$$= -sx + bx^2$$

$$\Rightarrow x(t) = \frac{e^{-st}}{C - b\int_0^t e^{-st'}dt'} = \frac{e^{-st}}{C + \frac{b}{s}\left[e^{-st} - 1\right]}$$

We want: $\phi(z,t)$

we have: $x(t)$ along which $\phi(1-x(t),t)$ is constant

$\phi(z,0) = z$

we need: $C(z,t)$ such that $x(t)$ goes through $(1-x,t)$

$$z = 1 - \frac{e^{-st}}{C(z,t) + \frac{b}{s}(e^{-st} - 1)}$$

$$\Rightarrow \phi(z,t) = 1 - \frac{1}{C(z,t)} = 1 - \frac{(1-z)e^{st}}{1 + (1-z)\frac{b}{s}\left[e^{st} - 1\right]}$$

Sanity checks

$$\phi(1,t) = \sum_{n=0}^{\infty} P(n,t) = 1 \quad \checkmark$$

$$\langle n \rangle = \sum_{n=0}^{\infty} n P(n,t) = \partial_z \phi(z,t)\Big|_{z=1} = e^{st} \quad \checkmark$$

Full distribution

$\phi(z,t)$ can be expressed as an geometric series

$$\frac{a+bz}{c+dz} = \frac{a}{c} + \frac{ad-bc}{dc} \sum_{n=1}^{\infty} \left(\frac{d}{c}\right)^n z^n$$

$$\phi(z,t) = \frac{(b-s)(e^{st}-1)}{be^{st}-b+s} + \frac{s^2 e^{st}}{(be^{st}-b+s)\,b(e^{st}-1)} \sum_{n=1}^{\infty} \left(\frac{b(e^{st}-1)}{be^{st}-b+s}\right)^n z^n$$

Probability of loss

$$P(0,t) = \phi(0,t) = 1 - \frac{e^{st}}{1+\frac{b}{s}(e^{st}-1)} \longrightarrow \begin{cases} 1-\frac{s}{b} & \text{if } s>0 \\ 1-\frac{1}{t} & \text{if } s=0 \\ 1-\frac{e^{st}}{1-\frac{b}{s}} & \text{if } s<0 \end{cases}$$