Approximating genealogies under genetic hitchhiking with recurrent mutation

Peter Pfaffelhuber
(joint with Joachim Hermisson)

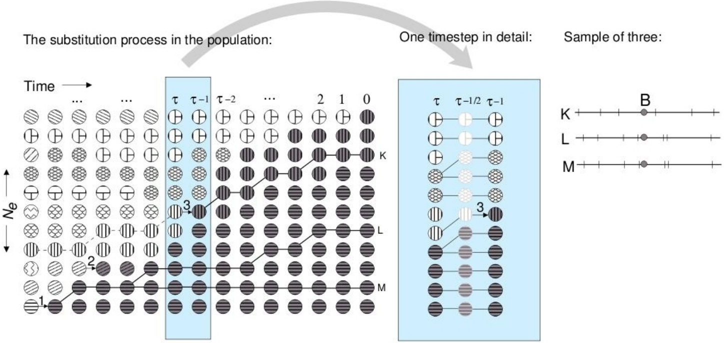
La Londe, September 2008

Goal

- Goal: detect selection in a genome
- Use sample variation data to find candidate genes
- Needed: prediction of sequence diversity under various forms of selection
- (Classical) selective Sweep: Variation around a stronlgy beneficial allele is strongly reduced
- Here: selection starts acting at $t = 0$
  beneficial allele arises recurrently during fixation
- Soft sweep: beneficial allele has multiple origins

Soft Sweep Patterns

- Recurrent mutation in a Wright-Fisher model (Pennings, Hermisson, 2006)



Soft Sweep Patterns

- Classical selective sweep: neutral variation dragged to high frequency together with beneficial allele
- Soft sweeps: Multiple mutants introduce different patterns of neutral variation
- Consequence: Different haplotype blocks around the selected site

Lactose gene (from Tishkoff et al (2007))

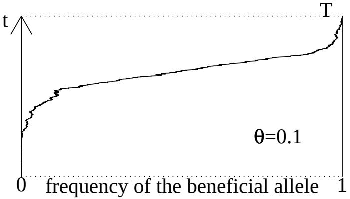
- Not all adults can digest milk (→ lactase persistence LP)
- Probably connection to cattle domestication
- Europe: Swedes 90% LP, Spanish 50% LP;
  SNP C/T-13910 associated with LP
- Asia: Chinese 1% LP
- Africa: West-African agriculturalists 5-20% LP;
  G/C-14010 most significantly associated SNP with LP
- ⇒: **Different origins of LP**

The Wright-Fisher diffusion

- Frequency path of beneficial allele is

$$dX = \Big(\tfrac{\theta}{2}(1-X)+\alpha X(1-X)\Big)dt+\sqrt{X(1-X)}dW, \qquad X_0 = 0$$

s	selective advantage
u	mutation rate
N	population size
α	$:= sN \gg 1$
θ	$:= 2uN$
dt	$\equiv Ndt$ generations
T	fixation time

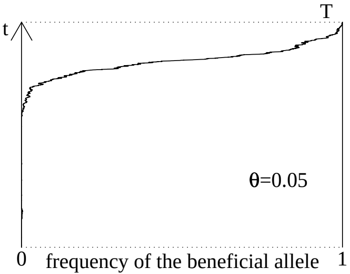


The Wright-Fisher diffusion

- Frequency path of beneficial allele is

$$dX = \Big(\tfrac{\theta}{2}(1-X)+\alpha X(1-X)\Big)dt+\sqrt{X(1-X)}dW, \qquad X_0 = 0$$

s	selective advantage
u	mutation rate
N	population size
α	$:= sN \gg 1$
θ	$:= 2uN$
dt	$\equiv N dt$ generations
T	fixation time

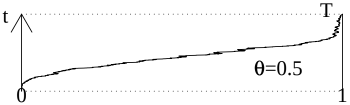


The Wright-Fisher diffusion

- Frequency path of beneficial allele is

$$dX = \Big(\tfrac{\theta}{2}(1-X)+\alpha X(1-X)\Big)dt+\sqrt{X(1-X)}dW, \qquad X_0 = 0$$

s	selective advantage
u	mutation rate
N	population size
α	$:= sN \gg 1$
θ	$:= 2uN$
dt	$\equiv$ Ndt generations
T	fixation time



Fixation times

- Let
$$T_0 := \sup\{t \geq 0 : X_t = 0\}, \qquad T^* := T - T_0.$$
- Fixation times
  - For $\theta > 0$,
$$\begin{aligned}\mathbb{E}[T] &= \frac{1}{\alpha\theta} + \frac{2\log\alpha}{\alpha} + \mathcal{O}\Big(\frac{1}{\alpha}\Big) + \frac{1}{\theta}\mathcal{O}\big(\alpha e^{-\alpha}\big),\\ \mathbb{E}[T^*] &= \frac{2\log\alpha}{\alpha} + \mathcal{O}\Big(\frac{1}{\alpha}\Big),\\ \mathbb{V}[T^*] &= \mathcal{O}\Big(\frac{1}{\alpha^2}\Big).\end{aligned}$$
  - For $\theta \geq 1$, almost surely, $T = T^*$.

The structured coalescent

- Sample n individuals at time T
- Genealogy at selected/linked neutral site given by structured coalescent
- Kaplan, Hudson, Langley (1989); extension by Barton, Etheridge, Sturm (2004)
- time T_0: random partition ξ of $\{1, ..., n\}$.
- **Goal:** describe/approximate ξ

The structured coalescent

- Discrete model: given $X_t = x$, birth events of beneficial alleles:

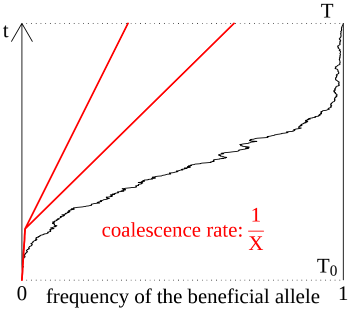
$$\text{rate } \frac{Nx}{2}$$

common ancestry of a given pair

$$\text{probability } \frac{1}{\binom{Nx}{2}}$$

$\Rightarrow$ **unscaled coalescence rate**

$$\frac{1}{Nx}$$


t
T
T$_0$
0
frequency of the beneficial allele
1
coalescence rate: $\frac{1}{X}$

The structured coalescent

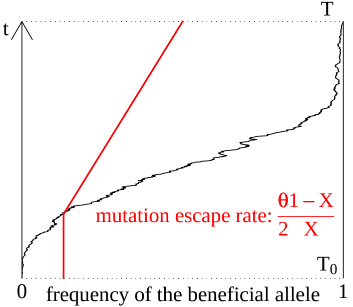
- Discrete model: given $X_t = x$, probability of following a mutant is

$$u(1-x).$$

Probability of picking a beneficial allele is x.

$\Rightarrow$ **unscaled mutation rate**

$$\frac{u(1-x)}{x}$$


T
t
mutation escape rate: $\frac{\theta 1 - X}{2 \;\; X}$
T_0
0
frequency of the beneficial allele
1

The structured coalescent

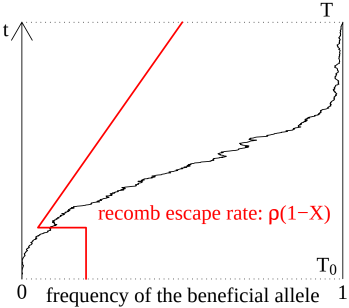
- Discrete model: given $X_t = x$, Frequency of recombinants of beneficial allele with wild-type is

$$rx(1-x)$$

Probability of picking a beneficial allele is x.

$\Rightarrow$ **unscaled recombination rate**

$$r(1-x)$$


T
t
recomb escape rate: ρ(1−X)
T_0
0
frequency of the beneficial allele
1

The structured coalescent

- Discrete model: given $X_t = x$, birth events of wild-type alleles:

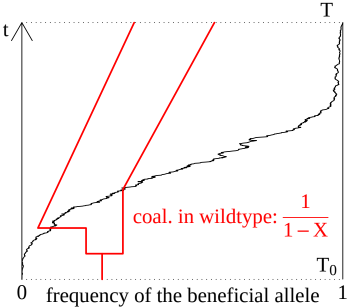
$$\text{rate } \frac{N(1-x)}{2}$$

common ancestry of a given pair

$$\text{probability } \frac{1}{\binom{N(1-x)}{2}}$$

⇒ unscaled coalescence rate

$$\frac{1}{N(1-x)}$$


T
t
coal. in wildtype: $\frac{1}{1-X}$
T_0
0
frequency of the beneficial allele
1

The structured coalescent

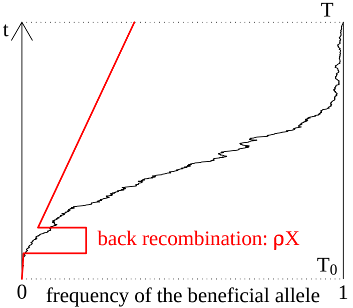
- Discrete model: given $X_t = x$, Frequency of recombinants of beneficial allele with wild-type is

$$rx(1-x)$$

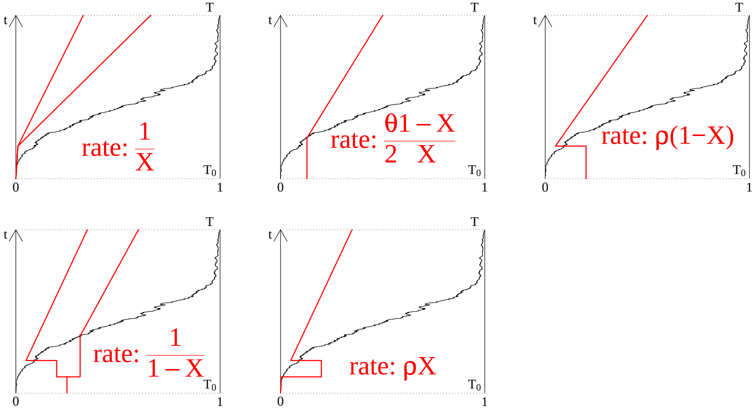
Probability of picking a wild-type allele is $1-x$.

⇒ unscaled recombination rate

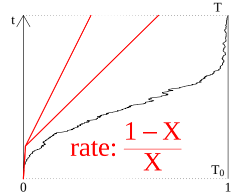
$$rx$$

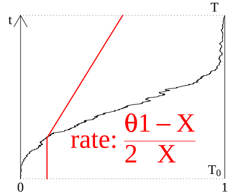


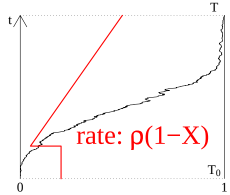
The structured coalescent

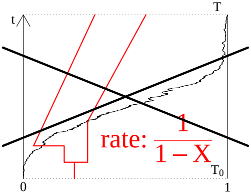


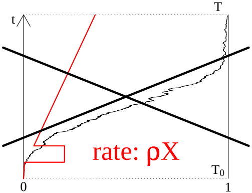
The structured coalescent











Probability $\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$ for $\rho = \mathcal{O}\left(\frac{\alpha}{\log \alpha}\right)$

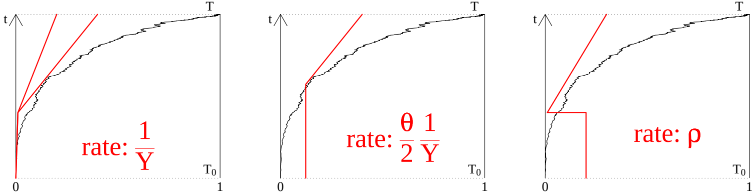
The structured coalescent

Time rescaling $d\tau = (1 - X)dt$:

$$dY = \left(\frac{\theta}{2} + \alpha Y\right)d\tau + \sqrt{Y}dW, \qquad Y_0 = 0.$$

Supercritical Feller branching process with immigration
Stop when hitting $Y = 1$

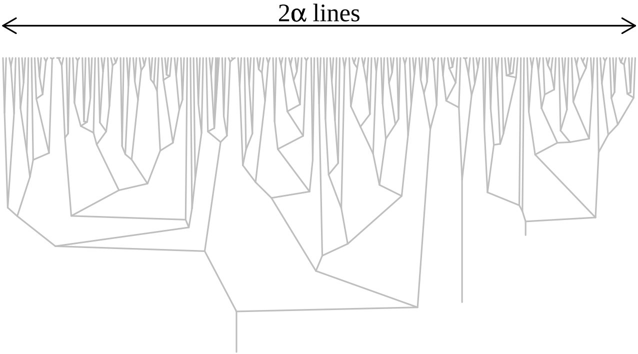
The structured coalescent



Coalescent generates a marked (rate ρ) genealogy of a supercritical Feller branching process with immigration (rate $\theta/2$)

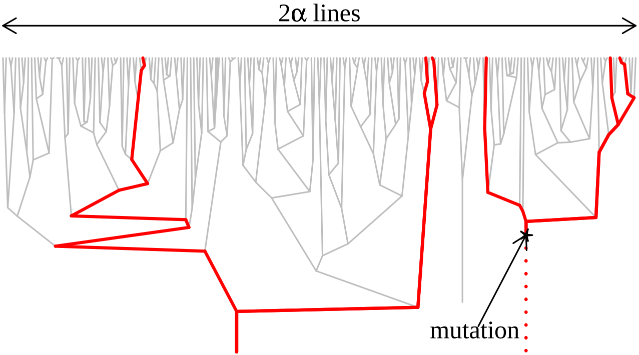
The Yule process approximation

- splitting rate α per line, immigration rate: θ



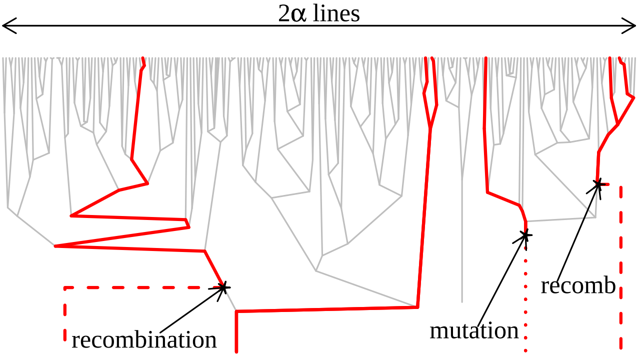
The Yule process approximation

- splitting rate α per line, immigration rate: θ



The Yule process approximation

- splitting rate α per line, immigration rate: θ
- recombinations: rate ρ along Yule tree



The Yule process approximation

- Given: sample of size n
- Yule process approximation: random partition Υ of $\{1, \dots, n\}$
- Let $\rho = \gamma \frac{\alpha}{\log \alpha}$.

- Theorem

$$\sup_A \Big| \mathbb{P}[\xi \in A] - \mathbb{P}[\Upsilon \in A] \Big| = \mathcal{O}\Big(\frac{1}{(\log \alpha)^2} \Big)$$

where the error is uniform on compacta in γ, θ.

Related work

- $\theta = 0$: Durret, Schweinsberg (2004,...), Etheridge, P, Wakolbinger (2006): Yule approximation for classical sweeps
- $\rho = 0$: Pennings, Hermisson (2006): family sizes of origins of beneficial allele follow the Ewens sampling formula

- P, Studeny (2007): Yule approximation for genealogies of two neutral loci
- Leocard (2008): Yule approximation for several neutral loci

Application: heterozygosity

- Heterozygosity H_t: probability that two randomly picked individuals carry different alleles
- Consider neutral locus linked to the selected one
- Assuming no mutations at neutral locus during the sweep,

$$H_T = \mathbb{P}[\text{no coalescence by } T_0] \cdot H_{T_0}.$$

Application: expected heterozygosity

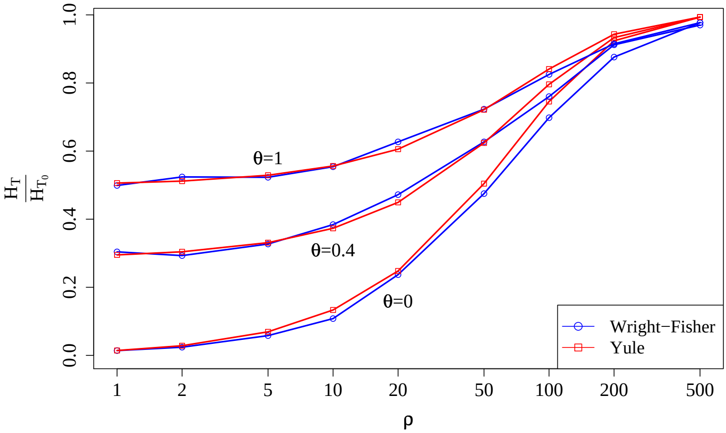
- Using Yule process approximation for $\rho = \gamma \frac{\alpha}{\log \alpha}$:

$$\frac{H_T}{H_{T_0}} = 1 - \frac{p_1^2}{\theta + 1} - \frac{2\gamma}{\log \alpha} \sum_{i=2}^{\lfloor 2\alpha \rfloor} \frac{2i + \theta}{(i+\theta)^2 (i+1+\theta)} p_i^2 + \mathcal{O}\Big(\frac{1}{(\log \alpha)^2}\Big)$$

with

$$p_i := \exp\Big(-\frac{\rho}{\alpha} \sum_{j=i+1}^{\lfloor 2\alpha \rfloor} \frac{1}{j}\Big).$$

Application: expected heterozygosity



Summary

- Soft sweeps from recurrent mutation generalize *classical* sweeps
- Ewens sampling formula gives family decomposition at selected site
- Yule process with immigration and marks approximates genealogy at linked neutral locus

Outlook

- Lactase Persistence: partial sweep, structured population
- What is a good approximation to the genealogy under sweeps in structured populations?