

Approximate genealogies under Hitchhiking

Peter Pfaffelhuber

Joint work with Alison Etheridge (Oxford), Anton Wakolbinger (Frankfurt) and Bernhard Haubold (Weihenstephan)

March, 29th, 2006

What does $\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$ mean in practise?

or

How does the Yule process approximation perform **compared to other approximations** of selective sweeps?

Hitchhiking in the Wright-Fisher model

A finite version of the diffusion model Anton described is:

- ▶ The population of size N evolves according to a Wright-Fisher model with selection
- ▶ A beneficial allele B enters the population
- ▶ B has selective advantage $s = \frac{\alpha}{N}$ over b
- ▶ The initial frequency of B is $\frac{1}{N}$
- ▶ Assume B eventually fixates

How can the genealogy of a sample of n individuals at a neutral site linked to the selected one be approximated?

The logistic model

- ▶ Kaplan, Hudson, Langley (1989) studied a structured coalescent conditioned on the frequency path

$$dX = \alpha X(1 - X)dt, \quad X_0 = \varepsilon$$

instead of the diffusion in Anton's talk

- ▶ there is no genetic drift (might be ok for strong selection)
- ▶ Values of $\varepsilon = \frac{1}{N}, \frac{5}{\alpha}, \dots$ appear in the literature
- ▶ The same process was used in Stephan, Wiehe, Lenz (1992) and several simulation studies (Kim and Stephan (2002), Przeworski (2000), Li and Stephan (2005), ...)

Is the Yule process approximation more accurate than the logistic model?

The implementation

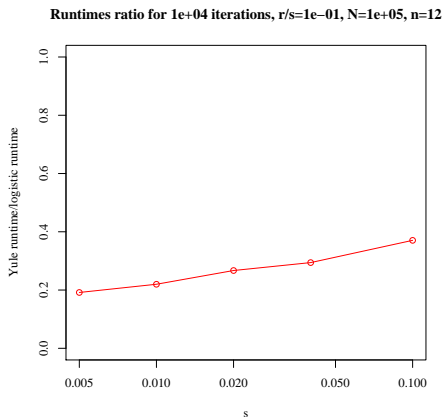
Take a sample of n at the end of a selective sweep. We compare *identity by descent* of neutral loci at the beginning of the sweep in

- ▶ the exact Wright-Fisher model
- ▶ the logistic approximation
- ▶ the Yule process approximation

Examples: What is the probability

- ▶ that $n, n - 1, \dots$ lines trace back to the founder of the selective sweep?
- ▶ the largest family in the sample has size k ?

With 1000 time steps during the sweep in the logistic model, the Yule process approximation is faster



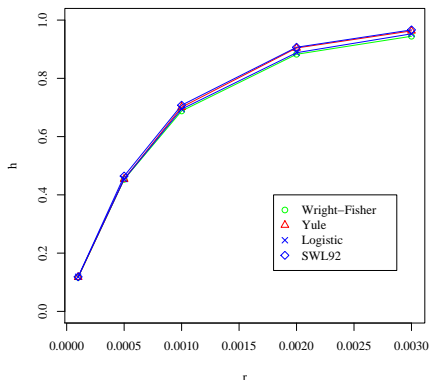
Heterozygosity

Take $n = 2$. The probability that two neutral loci do not coalesce during the sweep is well approximated.

Probability of
no coalescence
in the sweep

$$N = 10^5$$
$$s = 0.01$$

Factor of reduction in heterozygosity, $N=1e+05$, $s=1.0e-02$



Events in the non-beneficial background

The number of events in the non-beneficial background is well approximated by the logistic model;
As predicted by the *key Lemma*, only few events happen.

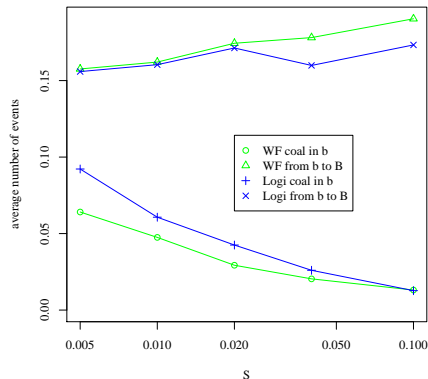
Events in the non-beneficial background

$$N = 10^5$$

$$n = 12$$

$$r/s = 0.1$$

Events in background b, $N=1e+05$, $r/s=1e-01$, $n=12$



The number of families

It was predicted that there is - up to an error of $\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$ - at most one early recombinant family.

However, for larger sample sizes the probability of more than one early recombinant family is not negligible.

Number of ≥ 2 -families

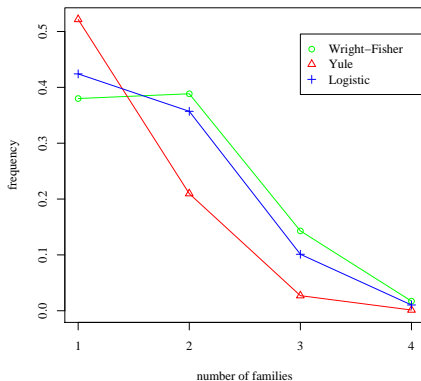
$$N = 10^5$$

$$n = 12$$

$$s = 0.01$$

$$r = 0.003$$

Number of 2-Families, $N=100000$, $s=0.01$, $r=0.003$, $n=12$



The error

To check which approximation is good, we compare the statistics

F_1 := size of the biggest family,

F_2 := size of the second biggest family.

To quantify the error of the approximations we use

$$Error = \sum_{i,j} |\mathbf{P}_{WF}[F_1 = i, F_2 = j] - P_{\square}[F_1 = i, F_2 = j]|$$

for

\square = logistic model or Yule process approximation.

The error for the Yule process approximation is $\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$.

The error

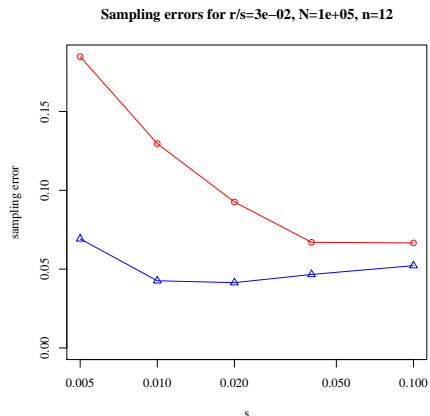
Under reasonable parameter combinations the logistic model is a better approximation for the sizes of the biggest families

Error in the approximations

$$N = 10^5$$

$$n = 12$$

$$r/s = 0.03$$



The error

Next, compare errors for

F_1 := number of lines that descend from the founder of the sweep,

F_2 := size of the largest early recombinant family

These statistics take full account of events happening in the **early phase** of a selective sweep

The error

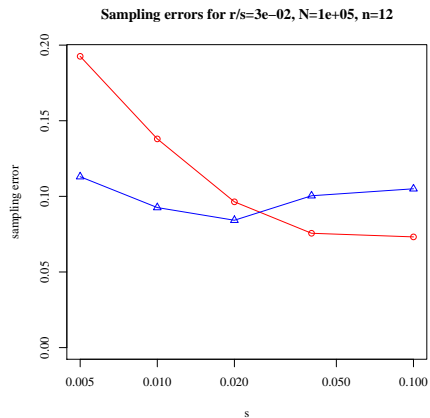
Tracing back the number of lines that go back to the founder of the sweep produces the same error in both models
As predicted the Yule process becomes better the larger α .

Error in the approximations

$$N = 10^5$$

$$n = 12$$

$$r/s = 0.03$$



Application: Tajima's D

- ▶ A commonly used statistic to test the neutral model of evolution is **Tajima's D**
- ▶ It relies on comparing the

number of **segregating sites**

with the

average pairwise difference between two lines.

To obtain the distribution of Tajima's D we

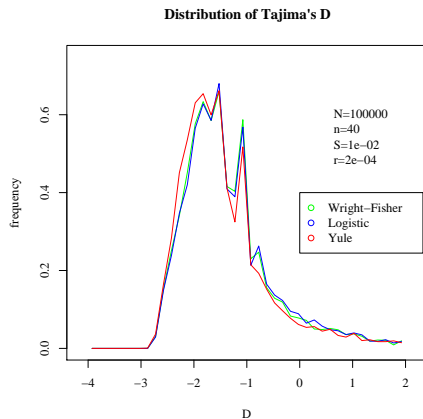
- ▶ add a neutral phase of evolution before the selective sweep
- ▶ put mutations on the tree

Application: Tajima's D

Tajima's D tends to be negative under genetic hitchhiking
All approximations are accurate.

Distribution of Tajima's
 D

$N = 10^5$
 $n = 40$
 $s = 0.01$
 $r = 0.0002$



Conclusions

logistic model	gives best approximations of family sizes
sampling formula	only works for small samples
Yule approximation	works well for strong selection serves as an analytical tool for hitchhiking

- ▶ Can the Yule process approximation be **improved**?
- ▶ Genetic structure: Is there a **multi-locus extension** of the Yule process approximation (several neutral loci)?
- ▶ Spatial structure: Does the Yule process help to understand selective sweeps in **structured populations**?