Some motivations

Evolutionary Biology

- Evolutionary Biology studies the change of inheritable characters of populations over time
- Inheritable characters are called alleles
- To study these changes, an evolutionary biologist consults phenotypic and genotypic data
- Studying phenotypes leads to a morphological distinction of species
- The distinction of species according to genotypes is called phylogenetics
- Changes in allelic frequencies within a species falls into the field of population genetics

Questions that might be asked:

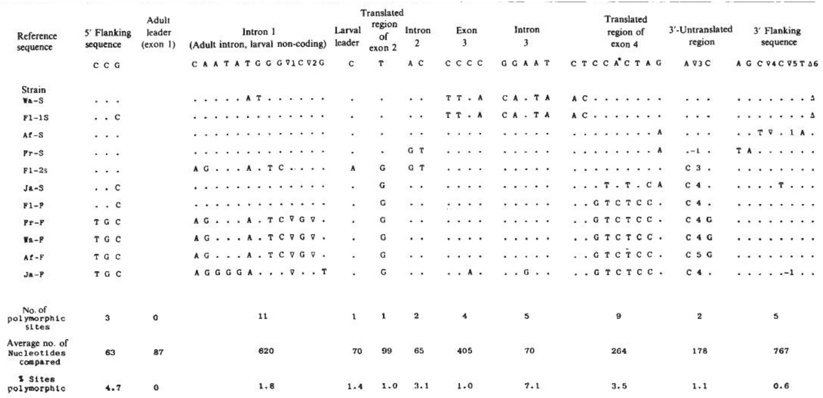
- Is there a way to predict a phenotype from a genotype?
- According to Darwin, the fittest will survive. Is this true?
- How can we measure fitness of a genotype?
- What inferences can we make on the history of a population, given its current genotype?
- How can we see on a genetical level that a population adapts to its environment?

There are tons of data:

- The human genome has $\approx 3 \cdot 10^9$ base pairs
- PCR, the basis for modern sequencing, was discovered only 15 years ago
- On a standard sequencer, it takes two hours to read 64 strains of $\approx$ 500-700 bases for a sequencer
- Latest development 454 sequencing: since 2005 it is possible to sequence $20 \cdot 10^6$ bases in 4-5 hours...

DNA data is special:

- The *AdH*-locus from Kreitman (1983)



Reference sequence	5′ Flanking sequence	Adult leader (exon 1)	Intron 1 (Adult intron, larval non-coding)	Larval leader	Translated region of exon 2	Intron 2	Exon 3	Intron 3	Translated region of exon 4	3′-Untranslated region	3′ Flanking sequence
	C C G		C A A T A T G G G ∇1 C ∇2 G	C	T	A C	C C C C	G G A A T	C T C C A* C T A G	A ∇3 C	A G C ∇4 C ∇5 T Δ6
Strain											
Wa-S	. . .		. . . . . A T	.	.	. .	T T . A	C A . T A	A C	. . .	. . . . Δ
Fl-1S	. . C		. . . .	.	.	. .	T T . A	C A . T A	A C	. . .	. . . . Δ
Af-S	. . .		. . . .	.	.	. .	. . . .	. . . . .	. . . . A	. . .	. . T ∇ . 1 A .
Fr-S	. . .		. . . .	.	.	G T	. . . .	. . . . .	. . . . A	. -1 .	T A . . . . . .
Fl-2s	. . .		A G . . . A . T C	A	G	G T	. . . .	. . . . .	. . . .	C 3 .	. . . .
Ja-S	. . C		. . . .	.	G	. .	. . . .	. . . . .	. . . T . T . C A	C 4 .	. . . . T . . .
Fl-F	. . C		. . . .	.	G	. .	. . . .	. . . . .	. . G T C T C C .	C 4 .	. . . .
Fr-F	T G C		A G . . . A . T C ∇ G ∇ .	.	G	. .	. . . .	. . . . .	. . G T C T C C .	C 4 G	. . . .
Wa-F	T G C		A G . . . A . T C ∇ G ∇ .	.	G	. .	. . . .	. . . . .	. . G T C T C C .	C 4 G	. . . .
Af-F	T G C		A G . . . A . T C ∇ G ∇ .	.	G	. .	. . . .	. . . . .	. . G T C T C C .	C 5 G	. . . .
Ja-F	T G C		A G G G G A . . . ∇ . . T	.	G	. .	. . A .	. . G . .	. . G T C T C C .	C 4 .	. . . . . -1 . .
No. of polymorphic sites	3	0	11	1	1	2	4	5	9	2	5
Average no. of Nucleotides compared	63	87	620	70	99	65	405	70	264	178	767
% Sites polymorphic	4.7	0	1.8	1.4	1.0	3.1	1.0	7.1	3.5	1.1	0.6

DNA data is special:

- The data structure is complex
- There is coding DNA, introns, regulatory regions, making every base special
- In a given population, most bases agree in all individuals
- DNA samples from the same population are not independent

Mathematical population genetics

- Mathematical population genetics is an own field
- Changes in allele frequencies are modelled by a stochastic process
- Keywords: diffusion limit, measure-valued diffusion, Markov process on general state spaces, dual process, martingale problem, super-process, particle representation, resampling model, branching process

Applied mathematical population genetics

- Quantitative predictions help to answer biological questions
- There are standard models
- Mostly, it is easy to name all mechanisms that must be modelled: reproduction, mutation, selection, recombination, structure,...
- Even for simple models there are still open questions

1: Basic models: Wright-Fisher model, Moran model, neutral theory, mutation models

2: Diffusion theory and applications

3: Applications: the Ewens sampling formula, site frequency spectrum, mismatch distribution

4: Recombination

5: Selection

6: Neutrality tests

Literature

- Ewens, W. J., Mathematical Population Genetics, 2002
- Wakeley, J., Coalescent Theory: An Introduction,Roberts & Co., 2007 (visit www.coalescenttheory.com)
- Gillespie, J., Population Genetics: A Concise Guide, Johns Hopkins University Press, Second Edition, 2004
- Hein, J., Schierupp, M. and Wiuf, C., Sequence Variation, Genealogies and Evolution: A Primer in Coalescent Theory, Oxford University Press, 2004
- Karlin, S. and Taylor, H.M., A Second Course in Stochastic Processes, Academic Press, 1981
- Etheridge, A., Diffusion Process Models in Mathematical Genetics, Lecture Notes
- Pfaffelhuber, P. and Pennings, P., Population Genetics Tutorial, companion manuscript

Basic models

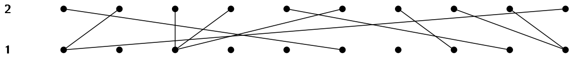
Introduction

- Assume a large population of (haploid) size N (often used: N diploids $= 2N$ haploids)
- Individuals have genotypes
- Genotypes are inherited to the next generation
- Every individual has only one parent

Wright-Fisher model

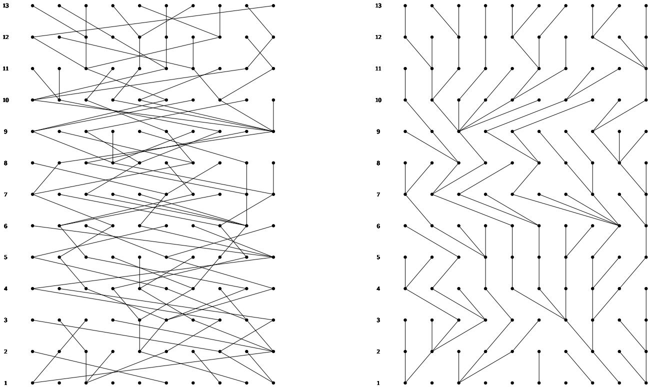
- standard population model of non-overlapping generations
- Example: Population size is 10

- Parents are picked at random
- Offspring gets genetic information from the parent.



Wright-Fisher model

The tangled and untangled versions after some generations



Wright-Fisher model

- Z_i: number of offspring of individual $i \sim B(N, \frac{1}{N}) \approx Poi(1)$
- Allele A frequency $X_t = x$ at time t
- $$\mathbb{P}[N \cdot X_{t+1} = k | X_t = x] = \binom{N}{k} x^k (1-x)^{N-k}$$
- X_{t+1} only depends on X_t, but not on $X_{t-1}, X_{t-2}, \ldots$
- The process $(X_t)_{t=0,1,\ldots}$ is a Markov chain

Exercise

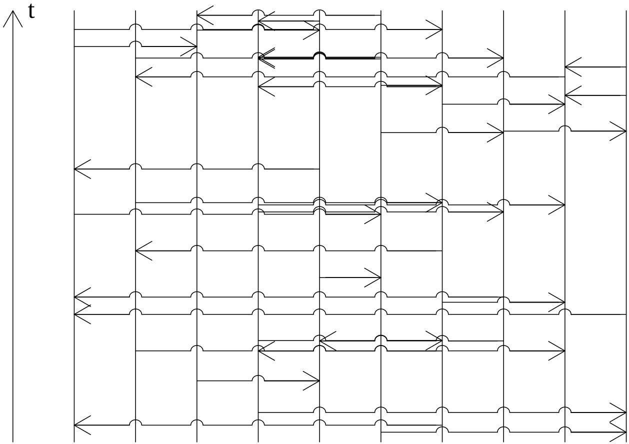
Obviously the Wright-Fisher model as we introduced it here is a model for haploid populations. (Every individual only has one parent and one set of genes.) Assume we also want to model diploids in the model. Can you draw a similar figure for the diploid model?

Moran model

- Standard population model of overlapping generations
- Every individual resamples at rate 1
- Resampling: choose second individual at random; one of them dies, the other one reproduces

Moran model

- Individual at the tip dies, the other one reproduces



Cannings models

- Allele frequency of A is $X_t = x$ at time t
- Rates

$$Q_{x,x+1/N} = Q_{x,x-1/N} = \frac{1}{2} N x (1 - x)$$

Cannings models

- Wright-Fisher model: binomial offspring distribution
- Attention: offspring distribution of different individuals are dependent!
- But: they are exchangeable: $Z_1, \ldots, Z_N$: numbers of offspring of all individuals; π: Permutation of $\{1, \ldots, N\}$

$$\mathcal{L}(Z_1, \ldots, Z_N) = \mathcal{L}(Z_{\pi(1)}, \ldots, Z_{\pi(N)})$$

- General: exchangeable offspring distribution: Cannings model

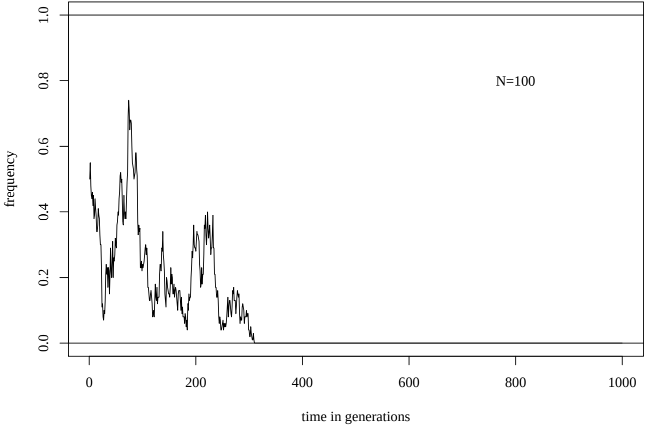
Genetic drift

- Measure for speed of loss/fixation of allelels
- Wright-Fisher model, allele frequency x.

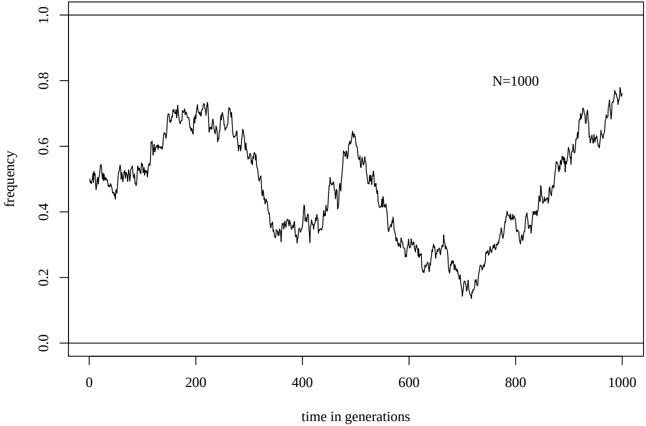
$$\begin{aligned}\mathbb{P}_x[\text{loss in one generation}] &= (1-x)^N \\ \mathbb{V}[X_{t+1}|X_t = x] &= \tfrac{1}{N^2}\mathbb{V}[NX_{t+1}|X_t = x] \\ &= \frac{Nx(1-x)}{N^2} = \frac{x(1-x)}{N}\end{aligned}$$

- Genetic drift strongest in small populations

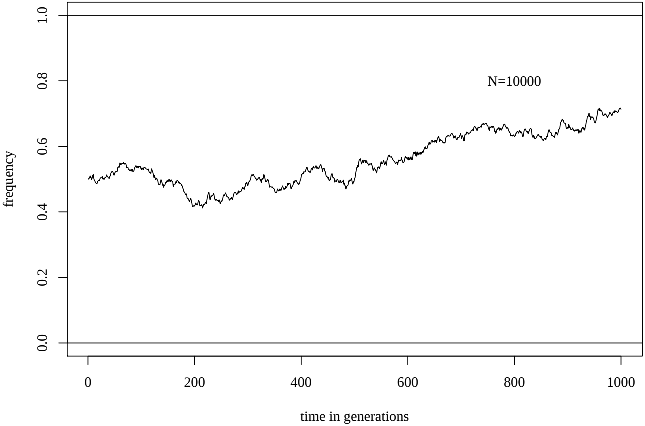
Genetic drift



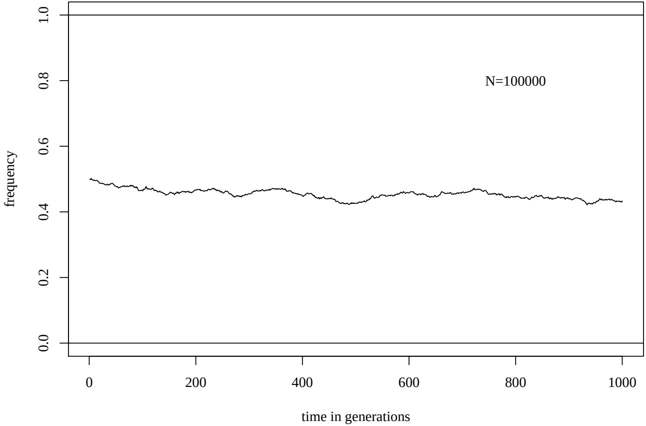
Genetic drift



Genetic drift



Genetic drift



Looking backwards in time

- Usually: data obtained from sample of size n of population of size N
- Allele A has frequency x at time 0; population evolves for time t
- Question: What is the distribution of allelic frequency of A in the sample?
- Possible calculation: compute random allelic frequency in the population; sample independent from population gives frequency in the sample

Looking backwards in time

- Question: What is the distribution of allelic frequency of A in the sample?
- Another possibility: every individual in sample has an ancestor at time 0
- $$\begin{array}{c}\text{Individual at time } t \text{ has allele } A \\ \Longleftrightarrow \\ \text{Ancestor at time 0 has allele } A\end{array}$$
- Possible: two individuals at time t have same ancestor at time 0

The coalescent in the Wright-Fisher model

- Sample of size n in big population of size N

$\mathbb{P}[n \text{ different ancestors one generation ago}]$

$$= \left(1 - \frac{1}{N}\right) \cdot \ldots \cdot \left(1 - \frac{n-1}{N}\right)$$

$$= 1 - \frac{\binom{n}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

$\mathbb{P}[\text{ less than } n-1 \text{ ancestors one generation ago}]$

$$\leq \frac{\binom{N}{n-2}(n-2)^n}{N^n} = \mathcal{O}\left(\frac{1}{N^2}\right)$$

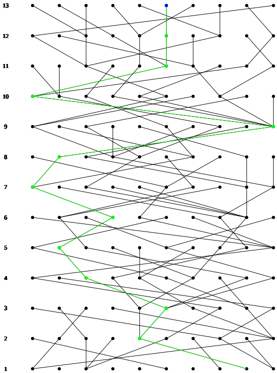
$$\mathbb{P}[n-1 \text{ different ancestors one generation ago}] = \frac{\binom{n}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

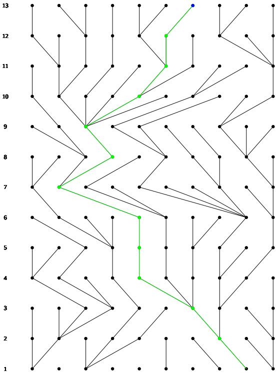
The coalescent in the Wright-Fisher model

- $\widetilde{T}_n$: waiting time until first coalescence event [generations]
- $$\mathbb{P}[\widetilde{T}_n > tN] \approx \left(1 - \frac{\binom{n}{2}}{N}\right)^{tN} \approx \exp\left(-\binom{n}{2}t\right)$$
- $T_n := \frac{\widetilde{T}_n}{N}$: waiting time until first coalescence event [N genertations]
- T_n approximately Exp$\left(\binom{n}{2}\right)$-distributed
- Restart argument:
  T_{n-1}: waiting time from T_n until second coalescence event approximately Exp$\left(\binom{n-1}{2}\right)$-distributed

The coalescent in the Wright-Fisher model

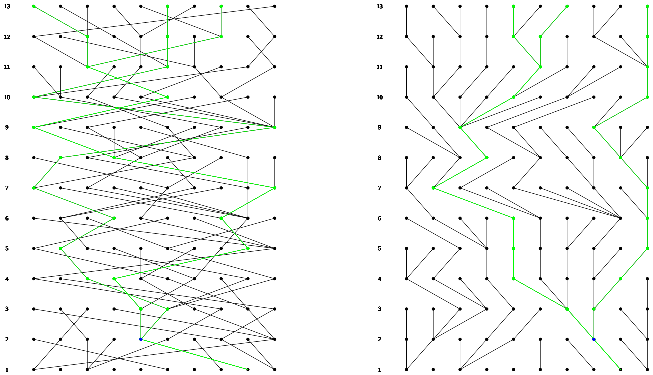
Green lines are ancestral lines of the sample





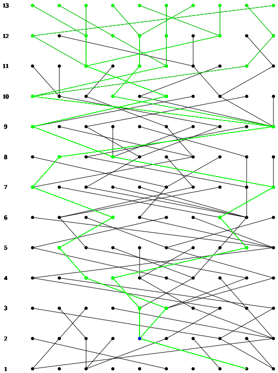
The coalescent in the Wright-Fisher model

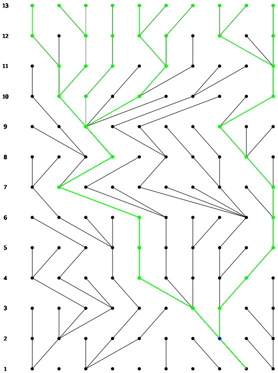
Lines in a sample share ancestry



The coalescent in the Wright-Fisher model

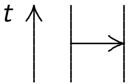
Genealogy of the whole population





The coalescent in the Moran model

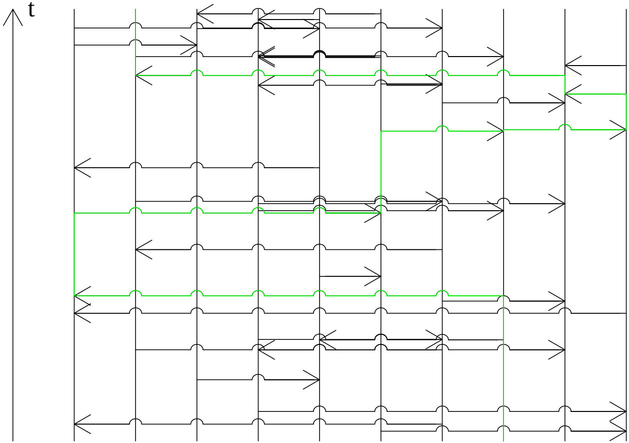
- Every pair resamples at rate $\frac{1}{N}$
- Backward in time, resampling is coalescence



- $\widetilde{T}_n$: time of first coalescence event in sample of size n
- $\widetilde{T}_n \sim \text{Exp}\left(\frac{\binom{n}{2}}{N}\right)$
- $T_n := \frac{\widetilde{T}_n}{N} \sim \text{Exp}\left(\binom{n}{2}\right)$

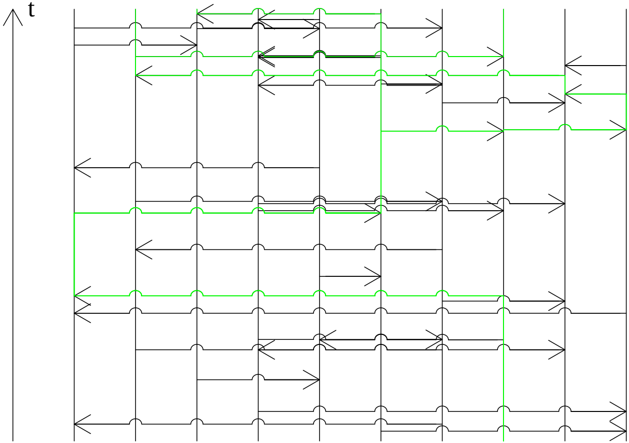
The coalescent in the Moran model

- Easy: ancestral line of one individual



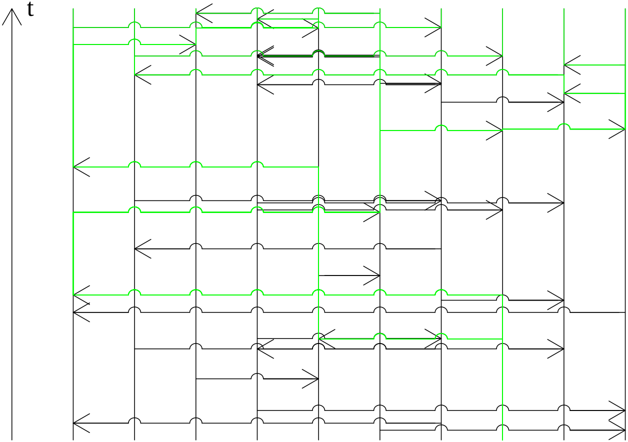
The coalescent in the Moran model

- All coalescence events at different time points



The coalescent in the Moran model

- Population MRCA different from sample MRCA

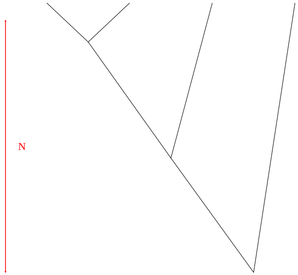


Kingman's coalescent

- Start with n lines.
- If there are k lines left, coalesce two of them at rate $\binom{k}{2}$
- Stop if only one line left
- The path of this process describes a genealogical tree
- Time is measured in units of N generations

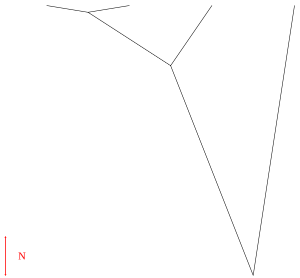
Kingman's coalescent

- The sample genealogy for $n = 4$



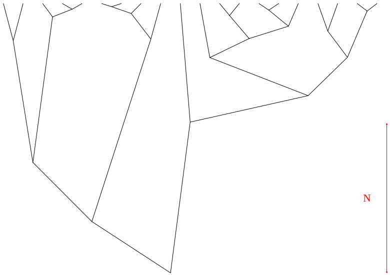
Kingman's coalescent

- The sample genealogy for $n = 4$



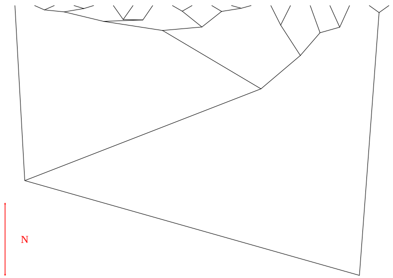
Kingman’s coalescent

- The sample genealogy for $n = 20$



Kingman's coalescent

- The sample genealogy for $n = 20$



Kingman's coalescent

- $T_k \sim \mathrm{Exp}\left(\binom{k}{2}\right)$: time the coalescent spends with k lines
- Time to the most recent common ancestor $T_{MRCA} = \sum_{k=2}^{n} T_k$

$$\mathbb{E}[T_{MRCA}] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n} \frac{1}{k-1} - \frac{1}{k} = 2\left(1 - \tfrac{1}{n}\right)$$

$$\mathbb{V}[T_{MRCA}] = \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} = 4\sum_{k=2}^{n} \left(\frac{1}{(k-1)} - \frac{1}{k}\right)^2$$

$$= 4\left[2\left(\sum_{k=2}^{n} \frac{1}{k^2}\right) + 1 - \frac{1}{n^2} - \sum_{k=2}^{n-1} \frac{2}{k(k-1)}\right]$$

$$= 8\left(\sum_{k=2}^{n} \frac{1}{k^2}\right) - 4\left(1 - \tfrac{1}{n}\right)^2$$

Kingman's coalescent

- Total tree length $L_n = \sum_{k=2}^{n} kT_k$
- $T_k \sim \text{Exp}\left(\binom{k}{2}\right)$, $kT_k \sim \text{Exp}\left(\frac{k-1}{2}\right)$. So,

$$\mathbb{E}[L_n] = \sum_{k=2}^{n} \frac{2}{k-1} = 2 \sum_{k=1}^{n-1} \frac{1}{k}$$

$$\mathbb{V}[L_n] = 4 \sum_{k=1}^{n-1} \frac{1}{k^2}$$

Mutations

- Without mutations, observed data would be extremely boring...
- Darwin: Variation shaped by natural selection
- Kimura: Neutral models can explain much of observed variation
- Empirical population genetics: what kind of variation is shaped by neutrality? What is different if neutrality does not hold?

Mutations in Wright-Fisher and Moran model

- Wright-Fisher model: offspring has different allele with probability μ
- Moran model: every line mutates at rate μ
  biologically unrealistic (mutation only during reproduction)

Mutations in the coalescent

- Recall: coalescence at rate $\binom{k}{2}$
- Mutations probability/rate μ per line [1], i.e., rate $N\mu$ [N]. Set $\theta := 2N\mu$
- Alternative description: given branch of length ℓ [N], no. of mutations is Poisson with parameter $\frac{\theta}{2}$
- Recall: If $X \sim \text{Exp}(\alpha)$, $Y \sim \text{Exp}(\beta)$ then

$$\mathbb{P}[X < Y] = \frac{\alpha}{\alpha + \beta}.$$

- Especially:

$$\mathbb{P}[\text{coalescence before mutation}] = \frac{\binom{k}{2}}{\binom{k}{2} + \frac{\theta k}{2}} = \frac{k-1}{k-1+\theta}.$$

Two alleles model

- Mutations occur between two possible states, A and B.
- Mutation probability/rate are

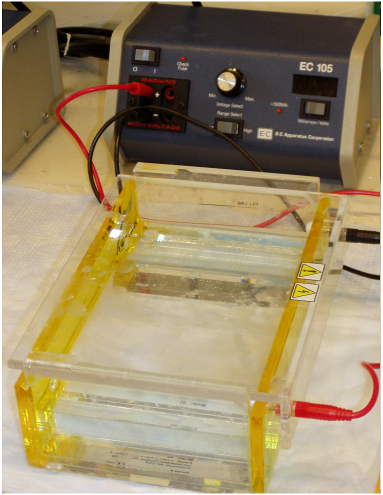
$$A \to B : \mu_A, \qquad B \to A : \mu_B.$$

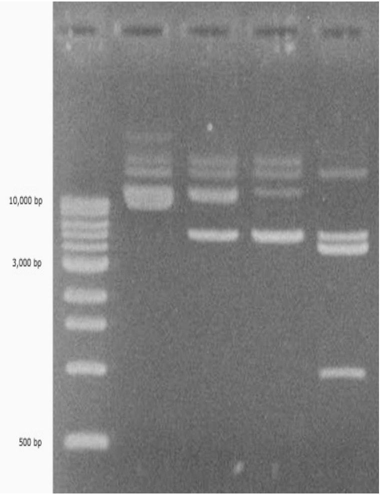
- Leads to one-dimensional models (constant population size!); most techniques known

Infinite alleles model

- Mutation probability/rate is μ
- If offspring is mutant, it carries a completely new allele
- Used for: electrophoretic data (in the old days)

Infinite alleles model





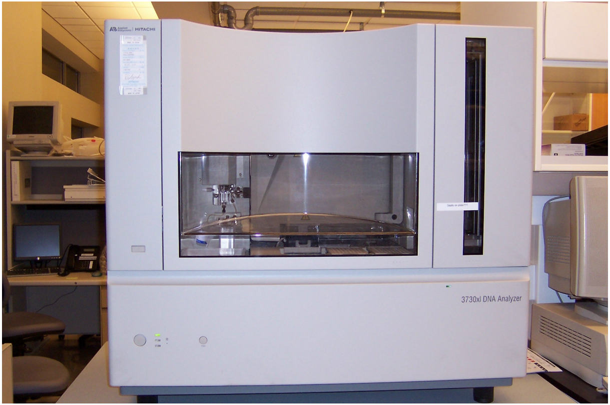
Stepwise mutation model

- Used for microsatellites
- Microsatellite: stretch of non-coding DNA, one short motif rapeated for a random number of times (`TCCTAGAGAGAGAGAGCCCGA`)
- Mutation: one repetition less or more
- Sequencing microsatellites: electrophoresis (cheap)

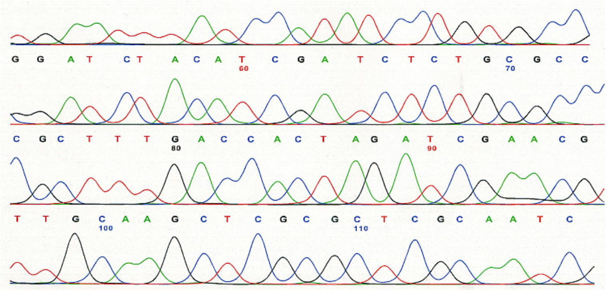
Infinite sites model

- Mutation probability/rate is μ
- If offspring is mutant: one new allele at a single site
- Every mutation hits a new site
- Used for: DNA sequence data

DNA sequencer

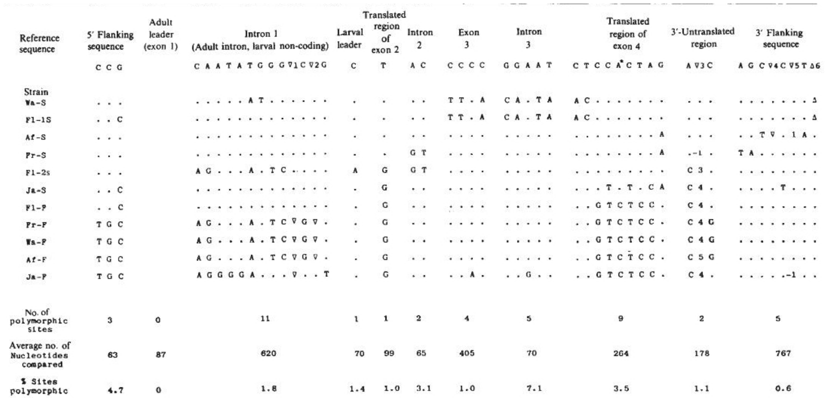


DNA raw data


G G A T C T A C A T C G A T C T C T G C G C C
60
70
C G C T T T G A C C A C T A G A T C G A A C G
80
90
T T G C A A G C T C G C G C T C G C A A T C
100
110

Infinite sites model

- Probably all mutations hit new sites in the Kreitman data



Reference sequence	5′ Flanking sequence	Adult leader (exon 1)	Intron 1 (Adult intron, larval non-coding)	Larval leader	Translated region of exon 2	Intron 2	Exon 3	Intron 3	Translated region of exon 4	3′-Untranslated region	3′ Flanking sequence
	CCG		CAATATGGG▽1C▽2G	C	T	AC	CCCC	GGAAT	CTCCA*CTAG	A▽3C	AGC▽4C▽5T△6
Strain											
Wa-S	...		.....AT.......	.	.	..	TT.A	CA.TA	AC.......	...	........△
Fl-1S	..C		....	.	.	..	TT.A	CA.TA	AC.......	...	........△
Af-S	...		....	.	.	..	....	.....	........A	...	..T▽.1A.
Pr-S	...		....	.	.	GT	....	.....	........A	.-1.	TA.......
Fl-2s	...		AG...A.TC.....	A	G	GT	....	.....	....	C3.	....
Ja-S	..C		....	.	G	..	....	.....	...T.T.CA	C4.	....T...
Fl-F	..C		....	.	G	..	....	.....	..GTCTCC.	C4.	....
Pr-F	TGC		AG...A.TC▽G▽.	.	G	..	....	.....	..GTCTCC.	C4G	....
Wa-F	TGC		AG...A.TC▽G▽.	.	G	..	....	.....	..GTCTCC.	C4G	....
Af-F	TGC		AG...A.TC▽G▽.	.	G	..	....	.....	..GTCTCC.	C5G	....
Ja-F	TGC		AGGGGA...▽..T	.	G	..	..A.	..G..	..GTCTCC.	C4.	.....-1..
No. of polymorphic sites	3	0	11	1	1	2	4	5	9	2	5
Average no. of Nucleotides compared	63	87	620	70	99	65	405	70	264	178	767
% Sites polymorphic	4.7	0	1.8	1.4	1.0	3.1	1.0	7.1	3.5	1.1	0.6

Remarks on mutation models

- infinite sites is a refinement of infinite alleles
- sequences in infinite sites models called haplotypes
- Further models: state-dependent mutation rates, finite sites model, indel mutations, ...

Heterozygosity

- Heterozygosity $h(t)$: probability of picking two different alleles at time t
- If $x_1, \dots, x_K$ are allele frequencies at time 0,

$$h = 1 - \sum_{k=1}^{K} x_i^2$$

- In a Wright-Fisher infinite allele model,

$$\begin{aligned} h(t+1) &= (1-(1-\mu)^2) + (1-\mu)^2\bigl(1-\tfrac{1}{N}\bigr)h(t) \\ &\approx 2\mu + \bigl(1-2\mu-\tfrac{1}{N}\bigr)h(t). \end{aligned}$$

In equilibrium $h(t+1) = h(t)$ and so

$$h(2\mu + \tfrac{1}{N}) = 2\mu, \qquad h = \frac{\theta}{\theta+1}$$

Heterozygosity

- The coalescent describes genealogies in equilibrium. Using

$$\mathbb{P}[\text{mutation before coalescence}] = \frac{\frac{\theta k}{2}}{\binom{k}{2} + \frac{\theta k}{2}} = \frac{\theta}{k - 1 + \theta}$$

for $k = 2$ immediately gives

$$h = \frac{\theta}{\theta + 1}$$

Segregating sites

- Mutation rate is $\frac{\theta}{2}$ per line $[N]$
- How many segregating sites do you expect for a sample of size 2? ... of size n?
- S_n: (random) number of segregating sites in sample of size n
- $$\mathbb{E}[S_n] = \mathbb{E}\big[\mathbb{E}[S_n|L_n]\big] = \mathbb{E}\big[\tfrac{\theta}{2}L_n\big] = \frac{\theta}{2}2\sum_{i=1}^{n-1}\frac{1}{i} = \theta\sum_{i=1}^{n-1}\frac{1}{i}$$
- Especially: no. of different sites in sample of size 2 is θ, in expecatation

Segregating Sites

- Moreover,

$$\begin{aligned}\mathbb{V}[S_n] &= \mathbb{E}\big[\mathbb{E}[S_n^2|L_n]\big] - \mathbb{E}[S_n]^2 \\ &= \mathbb{E}\big[\tfrac{\theta}{2}L_n + \tfrac{\theta^2}{4}L_n^2\big] - \tfrac{\theta^2}{4}\mathbb{E}[L_n]^2 \\ &= \tfrac{\theta}{2}\mathbb{E}[L_n] + \tfrac{\theta^2}{4}\mathbb{V}[L_n] = \theta\sum_{i=1}^{n-1}\tfrac{1}{i} + \theta^2\sum_{i=1}^{n-1}\tfrac{1}{i^2}\end{aligned}$$

Pairwise Differences

- $\widehat{\theta}_\pi$: Average number of pairwise differences

$$\widehat{\theta}_\pi = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} S_{ij}.$$

S_{ij}: number of sites different in sequences i and j

- $$\mathbb{E}[S_{ij}] = \theta \qquad \Rightarrow \qquad \mathbb{E}[\widehat{\theta}_\pi] = \theta$$

- Tajima (1983) has shown that

$$\mathbb{V}[\widehat{\theta}_\pi] = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2.$$

Mutation rate estimators

- Only the combined parameter $\theta = 2N\mu$ can be estimated!
- $$\widehat{\theta}_\pi := \frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} S_{ij}, \qquad \widehat{\theta}_W = \frac{S_n}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

  are unbiased estimators for θ!
- $\widehat{\theta}_W$ is consistent, but $\widehat{\theta}_\pi$ is not!

Sequence data and trees

- Consider the following sequences:

1	AATCCTTTGGAATTCCCT
2	GACCCTTTAGAATCCCAT
3	GACCCTTTAGGATTCCAT
4	GACCTTCGAGAGTCCTAT
5	GACCTCCGAGAATCCTAT

- Is there a way to put mutations on a tree which has leaves 1,2,3,4 and 5 that explains the data?
- Is the tree marked by mutations as informative as the data?

Effective population size

- Hammer (2004) sequenced the same locus (5239 bases) on 41 human X-chromosomes
- They found 16 segregating sites
- $$\widehat{\theta}_W = \frac{16}{\sum_{i=1}^{40} \frac{1}{i}} = 3.74 \text{ per locus} = 0.07\% \text{ per base}$$
- Moreover, $$\widehat{\theta}_\pi = 0.035\% \text{ per base} .$$
- Use this to estimate the human population size!
- Assume humans and chimpanzees split $T = 10^7$ years ago. Generation time is 25 years. Mutation rate is $2 \cdot 10^{-8}$ per base per generation.

Effective population size

- For N humans, there are $1.5N$ X chromosomes.
- Advantage of X-chromosomes: males only carry one allele, so there are no heterozygotes
- If N is the population size for diploids, $\theta = 3N\mu$
- $\widehat{\theta}_W, \widehat{\theta}_\pi$ are unbiased estimators for $3N\mu$

Effective population size

- Divergence D between humans and chimpanzees is 1.6% (per base)
- $D = 2T\mu$, so

$$\begin{aligned}\hat{\mu} = \frac{D}{2T} &= \frac{1.6\%}{2 \cdot 10^7}[\text{base and year}] \\ &= 25\frac{1.6\%}{2 \cdot 10^7}[\text{base and generation}] \\ &= 2 \cdot 10^{-8}[\text{base and generation}].\end{aligned}$$

- Population size, estimated using $\widehat{\theta}_W$:

$$\widehat{N}_W = \frac{\widehat{\theta}_W}{3\mu} = \frac{0.07\%}{6 \cdot 10^{-8}} \approx 1.2 \cdot 10^4.$$

- Why is N so low??

Effective population size

- Why is N so low??
- Model assumptions not met:
  - overlapping generations
  - selection
  - life-times not exponentially distributed
  - expanding population
  - not randomly mating
- Instead of census population sizes, effective population sizes are considered in practise

Effective population size

Let $\bullet$ be some property of a model in population genetics. This can be e.g. the rate of loss of heterozygosity, the offspring variance of a single individual, the speed of the coalescent or the time of fixation of a neutral allele. If there is a real population with census population size $N_{\mathcal{X}}$ and behaving as a model $\mathcal{X}$ the effective size of the population $\mathcal{X}$ is the size of an ideal (panmictic, constant-size etc.) Wright-Fisher population such that $\bullet$ is the same quantity in $\mathcal{X}$ and the Wright-Fisher model. This is denoted the $\bullet$-effective population size.

Effective population size

- $\bullet =$ loss of heterozygosity
- Assume no new mutations
- h_t: heterozygosity at time t,

$$h_1 = \frac{1}{N}0 + \left(1 - \frac{1}{N}\right)h_0 = \left(1 - \frac{1}{N}\right)h_0$$

so

$$h_t = \left(1 - \frac{1}{N}\right)^t \cdot h_0.$$

- Heterozygosity lost at rate $1 - \frac{1}{2N}$
- Assume a real population where heterozygosity is lost at rate a
- The real population size has N_e such that

$$a = 1 - \frac{1}{N_e}, \qquad \text{i.e., } N_e^{heterozygosity} = \frac{1}{1-a}.$$

Effective population size

- • = offspring variance
- Some model: Z_i number of offspring of individual i
- $\mathbb{V}[Z_i] = \sigma^2$

$$
\begin{aligned}
\mathbb{COV}[Z_i Z_j] &= \sum_{z=0}^{N} \mathbb{P}[Z_i = z]\mathbb{E}[Z_i Z_j | Z_i = z] - 1 \\
&= \sum_{z=0}^{N} \mathbb{P}[Z_i = z] z \mathbb{E}[Z_j | Z_i = z] - 1 \\
&\approx \frac{1}{N} \sum_{z=0}^{N} \mathbb{P}[Z_i = z] z (N - z + 1) - 1 = -\frac{\sigma^2}{N}
\end{aligned}
$$

Effective population size

- Some allele carried by first Nx individuals at time t, $X_t = x$.

$$\mathbb{V}[X_{t+1}] = \frac{1}{N^2}\mathbb{V}\Big[\sum_{i=1}^{Nx} Z_i\Big] = \frac{1}{N^2}\Big(\sum_{i=1}^{Nx}\mathbb{V}[Z_i] + \sum_{i=1}^{Nx}\sum_{\substack{j=1\\ j\neq i}}^{Nx}\mathbb{COV}[Z_i, Z_j]\Big)$$

$$\approx \frac{1}{N^2}(\sigma^2 Nx - N^2x^2\frac{\sigma^2}{N}) = \sigma^2\frac{x(1-x)}{N}$$

- Wright-Fisher model:

$$\mathbb{V}[X_{t+1}] = \frac{x(1-x)}{N}.$$

- $$N_e^{offspring\ variance} = \frac{N}{\sigma^2}$$

Exercise

- Assume $Z_I = N$ for a randomly chosen I each generation.
- What are the 'loss of heterozygosity' and 'offspring variance' effective population size?

Diffusion Theory

Definition

- A strong Markov process $\mathcal{X} = (X_t)_{t \geq 0}$ for which the sample paths are (almost surely) continuous is called a diffusion process.
- Diffusions we consider fulfill:
  - $\Box_k := \lim\limits_{t \to 0} \dfrac{\mathbb{E}_x[(X_t - x)^k]}{t}$ exist for $k = 1, 2, \ldots$.
  - $\Box_3, \Box_4, \ldots = 0$
- $\mu := \Box_1$: infinitesimal mean
- $\sigma^2 := \Box_2$: infinitesimal variance

Generator

- $\mathcal{X} = (X_t)_{t\geq 0}$: real-valued Markov process.
- For $f \in \mathcal{B}(\mathbb{R})$ define

$$(Gf)(x) := \lim_{t\to 0} \frac{\mathbb{E}_x[f(X_t) - f(x)]}{t}$$

  whenever the limit exists.
- The set $\mathcal{D}(G)$ for which the limits exists: domain of G
- G: (infinitesimal) generator of $\mathcal{X}$.

Example

- $\mathcal{X} = (X_t)_{t\geq 0}$: Poisson process with rate λ
- The generator is

$$\begin{aligned}
&\tfrac{1}{t}\mathbb{E}_x[f(X_t) - f(x)] \\
&\quad = \tfrac{1}{t}\big(e^{-\lambda t}\cdot f(x) + e^{-\lambda t}\lambda t \cdot f(x+1) - f(x) + \mathcal{O}(t^2)\big) \\
&\quad = \tfrac{1}{t}\big(- \lambda t f(x) + \lambda t f(x+1) + \mathcal{O}(t^2)\big) \\
&\quad \xrightarrow{t\to 0} \lambda\big(f(x+1) - f(x)\big)
\end{aligned}$$

- $\mathcal{X} = (X_t)_{t\geq 0}$: Jump process with rates $\lambda(x)$

$$(Gf)(x) = \sum_{x_{\text{new}}} \lambda(x, x_{\text{new}})\big(f(x_{\text{new}}) - f(x)\big).$$

Example: Brownian motion

- $\mathcal{X}$: standard Brownian motion
- For $f \in \mathcal{C}^2(\mathbb{R})$,

$$\mathbb{E}_x[f(X_t)] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right) f(y)dy$$

$$\overset{z=\frac{y-x}{\sqrt{t}}}{=} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) f(x+\sqrt{t}z)dz$$

$$(Gf)(x) = \lim_{t\to 0} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tfrac{1}{t}(f(x+\sqrt{t}z) - f(x))dz$$

$$= \lim_{t\to 0} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tfrac{1}{t}(f'(x)\sqrt{t}z + f''(x)\tfrac{tz^2}{2} + O(t^{3/2}))dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tfrac{1}{2} f''(x) z^2 dz = \tfrac{1}{2} f''(x).$$

Example: Diffusion

- $\mathcal{X} = (X_t)_{t\geq 0}$: Diffusion with μ and σ^2: for $f \in \mathcal{C}^2(\mathbb{R})$
- Generator given by

$$\begin{aligned}
&\tfrac{1}{t}\mathbb{E}_x[f(X_t) - f(x)] \\
&\quad = \tfrac{1}{t}\mathbb{E}_x[f'(x)(X_t - x) + \tfrac{1}{2}f''(x)(X_t - x)^2 + \ldots (X_t - x)^k] \\
&\quad \xrightarrow{t\to 0} \mu(x)f'(x) + \tfrac{1}{2}\sigma^2(x)f''(x)
\end{aligned}$$

Diffusion Approximation

- $\mathcal{X}, \mathcal{X}^1, \mathcal{X}^2, \ldots$: strong Markov processes (on compact state space) with generators $G_1, G_2, \ldots$.
- If
$$G_N f \xrightarrow{N \to \infty} Gf$$
for enough functions G then $\mathcal{X}_N \Rightarrow \mathcal{X}$.

Diffusion Approximation

- Moran model with alleles A and a of size N
- $\mathcal{X}^N$: Frequency path of allele A
- Theorem:

$$(X^N_{Nt})_{t\geq 0} \Rightarrow \mathcal{X}$$

$\mathcal{X} = (X_t)_{t\geq 0}$: Wright-Fisher diffusion with $\mu(x) = 0, \sigma^2(x) = x(1-x)$

- 'Proof':

$$\begin{aligned} G_N f(x) &= N \cdot (xN)(1-x) \cdot \\ &\qquad \left(\tfrac{1}{2} f(x + \tfrac{1}{N}) + \tfrac{1}{2} f(x - \tfrac{1}{N}) - f(x)\right) \\ &= N^2 \cdot x(1-x)\left(\tfrac{1}{2N^2} f''(x) + \mathcal{O}\left(\tfrac{1}{N^3}\right)\right) \\ &\xrightarrow{N\to\infty} \tfrac{1}{2} x(1-x) f''(x) \end{aligned}$$

Diffusion Approximation

- $\mathcal{X}^N = (X_t^N)_{t=0,1,\dots}$: Frequency of allele A in Wright-Fisher model and mutation probability μ from $a \to A$, $2N\mu \to \theta$
- Theorem:

$$(X^N_{[Nt]})_{t\geq 0} \Rightarrow \mathcal{X}$$

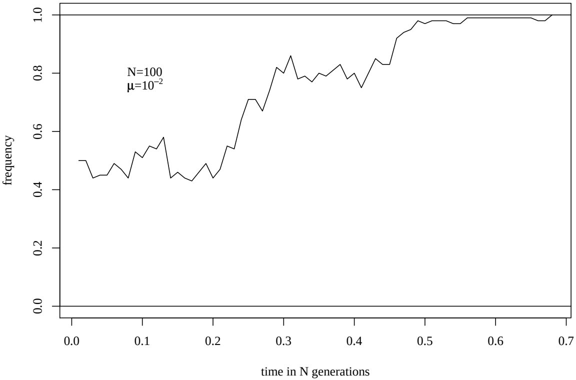
  $\mathcal{X} = (X_t)_{t\geq 0}$: Wright-Fisher diffusion with $\mu(x) = \frac{\theta}{2}, \sigma^2(x) = x(1-x)$
- 'Proof': $NX_1^N \sim B(N, x + \mu(1-x))$, so

$$N\mathbb{E}_x[X_1^N - x] = N\mu(1-x) \xrightarrow{N\to\infty} \frac{\theta}{2}(1-x),$$

$$\begin{aligned} N\mathbb{E}_x[(X_1^N - x)^2] &\approx N\cdot \mathrm{Var}[X_1^N] = \tfrac{1}{N}\cdot \mathrm{Var}[NX_1^N] \\ &= (x+\mu(1-x))(1-x-\mu(1-x)) \approx x(1-x) \end{aligned}$$

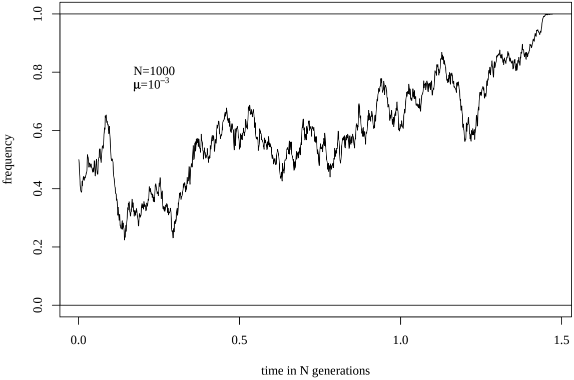
Diffusion Approximation

- No distinction possible on the timescale of N generations



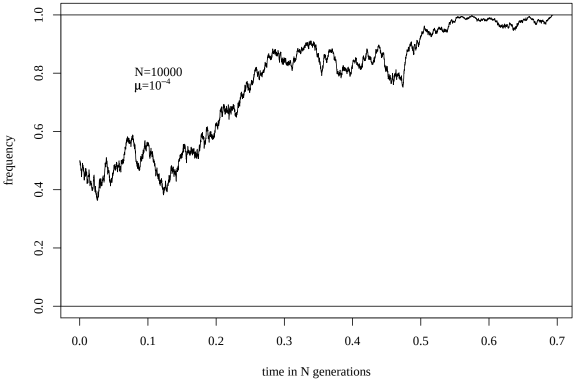
Diffusion Approximation

- No distinction possible on the timescale of N generations



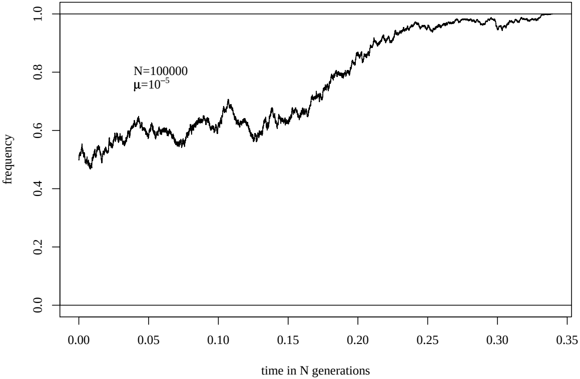
Diffusion Approximation

- No distinction possible on the timescale of N generations



Diffusion Approximation

- No distinction possible on the timescale of N generations



Chapman-Kolmogoroff equations

- The transition density function $p(.,.,.)$ of a process $\mathcal{X} = (X_t)_{t\geq 0}$ is

$$\mathbb{P}_x[X_t \in A] = \int_A p(t,x,y)dy.$$

- Chapman-Kolmogoroff equations: for any Markov process with transition function $p(.,.,.)$ and $s < t$

$$p(t,x,z) = \int p(t-s,x,y)p(s,y,z)dy.$$

The backward equation

- $\mathcal{X} = (X_t)_{t\geq 0}$: diffusion with μ and σ^2.
- g : smooth function
- What is
$$u(t,x) := \mathbb{E}_x[g(X_t)]?$$
- u is solution of the Kolmogoroff backward equation
$$\frac{\partial u}{\partial t} = \mu(x)\frac{\partial u}{\partial x} + \tfrac{1}{2}\sigma^2(x)\frac{\partial^2 u}{\partial x^2}.$$

The backward equation

Proof:

$$\begin{aligned}
\frac{\partial u(t,x)}{\partial t} &= \tfrac{1}{h}\mathbb{E}_x[g(X(t+h)) - g(X(t))] \\
&= \lim_{h\to 0} \tfrac{1}{h}\mathbb{E}_x[\mathbb{E}_{X(h)}[g(X(t))] - g(X(t))] \\
&= \lim_{h\to 0} \tfrac{1}{h}\mathbb{E}_x[u(t,X(h)) - u(t,x)] \\
&= \lim_{h\to 0} \tfrac{1}{h}\mathbb{E}_x\Big[((X(h)-x)\frac{\partial u(t,x)}{\partial x} \\
&\qquad\qquad + \tfrac{1}{2}(X(h)-x)^2\frac{\partial^2 u(t,x)}{\partial x^2} + O((X(h)-x)^3)\Big] \\
&= \mu(x)\frac{\partial u(t,x)}{\partial x} + \tfrac{1}{2}\sigma(x)\frac{\partial^2 u(t,x)}{\partial x^2}
\end{aligned}$$

Example

- $\mathcal{X} = (X_t)_{t\geq 0}$: neutral Wright-Fisher diffusion
- Consider $u(t,x) = \mathbb{E}_x[X_t(1-X_t)]$.
- $u(t,x)$ is the probability to pick one A and one a from the population at time t
- $$\frac{\partial u(t,x)}{\partial t} = -x(1-x) = -u(0,x)$$
- So, $u(t,x) = (1-e^{-t})x(1-x)$.

Example

- We found

$$\mathbb{E}_x[X_t(1-X_t)] = (1-e^{-t})x(1-x)$$

- Using the coalescent:

$$\begin{aligned}
&\mathbb{E}_x[X_t(1-X_t)] \\
&\quad = \mathbb{P}[\text{coalescence by time } t]\cdot 0 \\
&\qquad + \mathbb{P}[\text{no coalescence by time } t]\cdot x(1-x) \\
&\quad = (1-e^{-t})x(1-x).
\end{aligned}$$

The forward equation

- $\mathcal{X} = (X_t)_{t\geq 0}$: diffusion with μ and σ^2.
- Assume: transition density function exists.
- What is
$$p(t,x,y) = \mathbb{P}_x[X_t \in dy]?$$
- $p(.,.,.)$ solves the Kolmogoroff forward equation
$$\frac{\partial p(t,x,y)}{\partial t} = -\frac{\partial}{\partial y}\big(\mu(y)p(t,x,y)\big) + \tfrac{1}{2}\frac{\partial^2}{\partial y^2}\big(\sigma^2(y)p(t,x,y)\big).$$

Derivatives applied to both the infinitesimal parameters and the function p!

The forward equation

Proof:

$$\begin{aligned}\frac{\partial p(t,x,z)}{\partial t} &= \lim_{s\to 0}\frac{\partial}{\partial s}\int p(t,x,y)p(s,y,z)dy\\ &= \lim_{s\to 0}\int p(t,x,y)\frac{\partial p(s,y,z)}{\partial s}dy\\ &= \lim_{s\to 0}\int p(t,x,y)\Big(\mu(y)\frac{\partial p(s,y,z)}{\partial y}+\tfrac{1}{2}\sigma^2(y)\frac{\partial^2 p(s,y,z)}{\partial y^2}\Big)dy\\ &= \lim_{s\to 0}\int -p(s,y,z)\Big(\frac{\partial}{\partial y}\big(p(t,x,y)\mu(y)\big)\\ &\qquad\qquad -\tfrac{1}{2}\frac{\partial^2}{\partial y^2}\big(p(t,x,y)\sigma^2(y)\big)\Big)\\ &= -\frac{\partial}{\partial z}\big(\mu(z)p(t,x,z)\big)+\tfrac{1}{2}\frac{\partial^2}{\partial z^2}\big(\sigma^2(z)p(t,x,z)\big).\end{aligned}$$

Stationary distribution

- Integrate forward equation

$$\frac{\partial}{\partial t} P_x[X_t \leq y]$$
$$= \int_{-\infty}^{y} \Big(- \frac{\partial}{\partial z}\big(\mu(z)p(t,x,z)\big) + \tfrac{1}{2}\frac{\partial^2}{\partial z^2}\big(\sigma^2(z)p(t,x,z)\big)\Big) dz$$
$$= -\mu(y)p(t,x,y) + \tfrac{1}{2}\frac{\partial}{\partial y}\sigma^2(y)p(t,x,y)$$

- For $t \to \infty$, $p(t,x,y) \to \psi(y)$, $\frac{\partial}{\partial t}\mathbb{P}_x[X_t \leq y] \to 0$ and so

$$-\mu(y)\psi(y) + \tfrac{1}{2}\frac{\partial}{\partial y}\sigma^2(y)\psi(y) = 0,$$

i.e.,
$$\psi(y) = \frac{C}{\sigma^2(y)} \exp\Big(2\int_{\eta}^{y} \frac{\mu(z)}{\sigma^2(z)} dz\Big).$$

Application: Mutation-Drift balance

- Mutation $A \to a$ at rate $\frac{\theta_A}{2}$,
- Mutation $a \to A$ at rate $\frac{\theta_a}{2}$,
- $\mathcal{X} = (X_t)_{t \geq 0}$: Wright-Fisher diffusion with

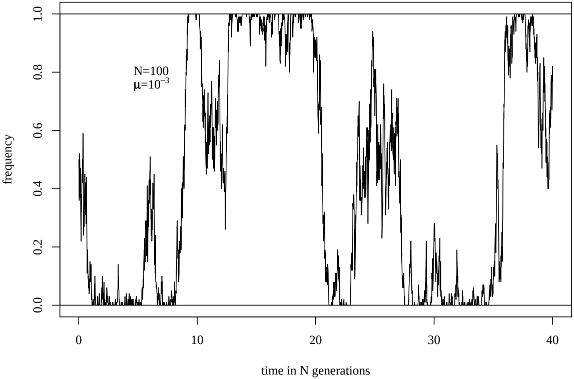
$$\mu(x) = -\frac{\theta_A}{2} x + \frac{\theta_a}{2}(1-x), \qquad \sigma^2(x) = x(1-x).$$

- Stationary distribution:

$$\psi(y) = \frac{C}{y(1-y)} \exp\left(\int_\eta^y -\frac{\theta_A}{1-z} + \frac{\theta_a}{z} dz \right)$$
$$= C y^{\theta_a - 1}(1-y)^{\theta_A - 1}$$

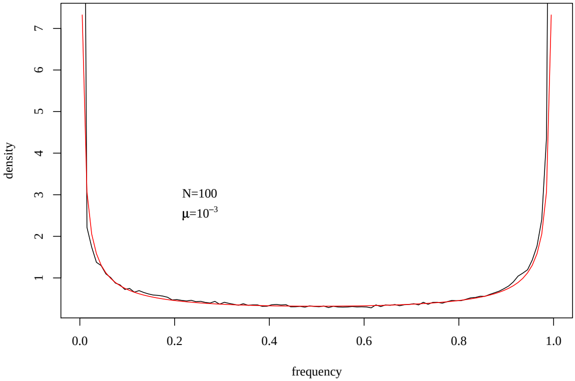
Diffusion Approximation

- For small mutation rates process frequently near boundaries



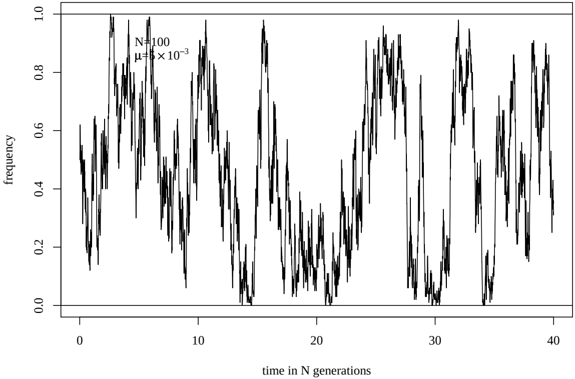
Diffusion Approximation

- Time average in simulations similar to equilibrium distribution



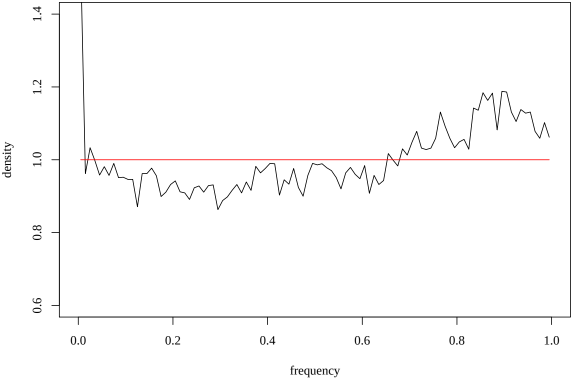
Diffusion Approximation

- For intermediate mutation rates process purely random frequencies



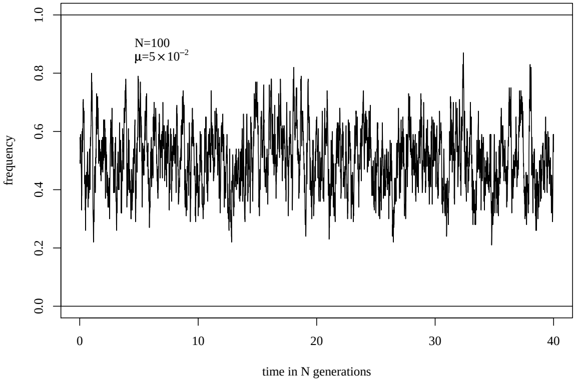
Diffusion Approximation

- For intermediate mutation rates process purely random frequencies



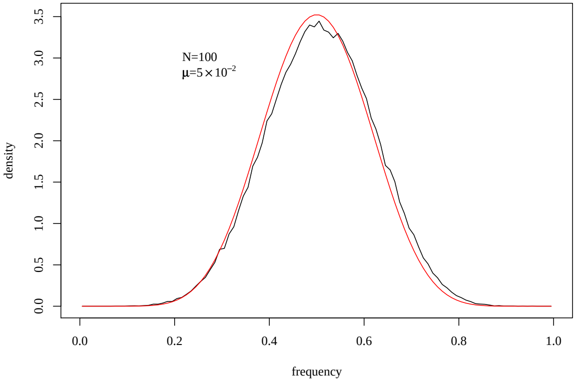
Diffusion Approximation

- For big mutation rates high heterozygosity



Diffusion Approximation

- For big mutation rates high heterozygosity



Exercise

Is heterozygosity increasing or decreasing with mutation rate?

Boundary Behavior

- Analysis did not need boundary behavior (absorbing, reflecting)
- All diffusions we consider: boundary behavior clear from finite model

Questions

- Assume $\mathcal{X}$ has absorbing states at 0 and 1
- T_0, T_1: absorption times (or ∞) at 0 and 1
- What is
$$\mathbb{P}_x[T_1 < T_0]?$$
- What is
$$\mathbb{E}_x[T_0 \wedge T_1]?$$
- What does $\mathcal{X}$, conditioned on $\{T_1 < T_0\}$ look like?

Speed and Scale

- Set $g(X_t) := 1_{X_t \leq y}$
- Backward equation for $u_y(t,x) := \mathbb{E}_x[1_{X_t \leq y}] = \mathbb{P}_x[X_t \leq y]$:

$$\frac{\partial}{\partial t}\mathbb{P}_x[X_t \leq y] = \mu(x)\frac{\partial}{\partial x}\mathbb{P}_x[X_t \leq y] + \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}\mathbb{P}_x[X_t \leq y]$$

Speed and Scale

- Assume $\mathcal{X}$ has absorbing states at 0 and 1
- Set

$$P_0(t,x) := \mathbb{P}_x[\mathcal{X} \text{ absorbed at 0 at time } t] = \mathbb{P}[X_t = 0],$$

$$P_0(t,x) \xrightarrow{t\to\infty} P_0(x) = \mathbb{P}_x[\mathcal{X} \text{ eventually absorbed at 0}]$$

- and

$$P_1(x) = \mathbb{P}_x[\mathcal{X} \text{ eventually absorbed at 1}]$$

Speed and Scale

- Letting $y \to 0$,

$$\frac{\partial}{\partial t}P_0(t,x) = \mu(x)\frac{\partial}{\partial x}P_0(t,x) + \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}P_0(t,x)$$

- After infinite time,

$$0 = \mu(x)\frac{\partial}{\partial x}P_0(x) + \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}P_0(x).$$

Speed and Scale

- After infinite time,

$$0 = \mu(x)\frac{\partial}{\partial x}P_0(x) + \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}P_0(x).$$

- Solving is easy: observe $P_0(0) = 1, P_0(1) = 0$.
  For some $\xi \in [0, 1]$,

$$\frac{\partial}{\partial x}P_0(x) = C \cdot \exp\Big(-2\int_\xi^x \frac{\mu(y)}{\sigma^2(z)}dz\Big)$$

$$P_0(x) = \frac{\int_x^1 \exp\Big(-2\int_\xi^y \frac{\mu(z)}{\sigma^2(z)}dz\Big)dy}{\int_0^1 \exp\Big(-2\int_\xi^y \frac{\mu(z)}{\sigma^2(z)}dz\Big)dy}.$$

Speed and Scale

- The scale function is

$$S(x) = \int_{x_0}^{x} \exp\left(-2 \int_{\xi}^{y} \frac{\mu(z)}{\sigma^2(z)} dz \right) dy$$

  for some $x_0 \in [0,1]$

- $$P_0(x) = \frac{S(1) - S(x)}{S(1) - S(0)}$$

- Similar:

$$P_1(x) = \frac{S(x) - S(0)}{S(1) - S(0)}$$

  Especially:

$$\mathbb{P}[\mathcal{X} \text{ eventually absorbed }] = 1.$$

Speed and Scale

- Stop the diffusion upon hitting a_0, b_0 with $0 < a_0 < x < b_0 < 1$

$$\mathbb{P}_x[\mathcal{X} \text{ hits } a_0 \text{ before } b_0] = \frac{S(b_0) - S(x)}{S(b_0) - S(a_0)}.$$

Speed and Scale

- $\mathcal{X}$ has absorbing boundaries 0 and 1
- Random time of absorption is T
- What is
$$w(x) = \mathbb{E}_x\Big[\int_0^T g(X_s)ds\Big]?$$
- $g = 1$: $w(x) =$ mean time until absorption

Speed and Scale

- Separating the integral into $[0, h]$ and $[h, T]$,

$$w(x) = \mathbb{E}_x\Big[\int_0^h g(X_s)ds\Big] + \mathbb{E}_x\big[w(X_h)\big],$$

$$\begin{aligned}\mathbb{E}_x\Big[\int_0^h g(X_s)ds\Big] &= hg(x) + \mathcal{O}(h^2),\\ \mathbb{E}_x[w(X_h)] &= \mathbb{E}_x[w(x) + (X_h - x)w'(x)\\ &\qquad + \tfrac{1}{2}(X_h - x)^2 w''(x) + \mathcal{O}(h^2)]\\ &= w(x) + h\big(\mu(x)w'(x) + \tfrac{1}{2}\sigma^2(x)w''(x) + \mathcal{O}(h)\big)\end{aligned}$$

- So,

$$\mu(x)w'(x) + \tfrac{1}{2}\sigma^2(x)w''(x) = -g(x), \qquad w(0) = w(1) = 0.$$

Speed and Scale

- So,
$$\mu(x)w'(x) + \tfrac{1}{2}\sigma^2(x)w''(x) = -g(x)$$
- Equivalently,

$$\exp\left(2\int_\xi^x \frac{\mu(z)}{\sigma^2(z)}dz\right)\frac{2\mu(x)}{\sigma^2(x)}w'(x)$$
$$+\exp\left(2\int_\xi^x \frac{\mu(z)}{\sigma^2(z)}dz\right)w''(x) = -\frac{2g(x)}{\sigma^2(x)}\exp\left(2\int_\xi^x \frac{\mu(z)}{\sigma^2(z)}dz\right),$$
$$\frac{d}{dx}\left(\exp\left(2\int_\xi^x \frac{\mu(z)}{\sigma^2(z)}dz\right)w'(x)\right) = -\frac{2g(x)}{\sigma^2(x)}\exp\left(2\int_\xi^x \frac{\mu(z)}{\sigma^2(z)}dz\right).$$

Speed and Scale

- So,

$$S(x) = \int_{x_0}^{x} \exp\Big(-2\int_{\xi}^{y} \frac{\mu(z)}{\sigma^2(z)}dz\Big)dy$$

and set

$$m(x) = \frac{1}{\sigma^2(x)S'(x)} = \frac{1}{\sigma^2(x)} \exp\Big(2\int_{\xi}^{x} \frac{\mu(z)}{\sigma^2(z)}dz\Big)$$

- So,

$$\frac{d}{dx}\Big(\frac{w'(x)}{S'(x)}\Big) = -2m(x)g(x),$$

Speed and Scale

- Recall

$$\frac{d}{dx}\Big(\frac{w'(x)}{S'(x)}\Big) = -2m(x)g(x).$$

- Integrating,

$$\frac{w'(x)}{S'(x)} = -2\int_{x_0}^{x} m(\xi)g(\xi)d\xi + \beta,$$
$$w(x) = -2\int_0^x S'(\eta)\int_0^\eta m(\xi)g(\xi)d\xi d\eta + \beta\int_0^x S'(\eta)d\eta + \alpha$$

- Since $w(0) = 0$ we find $\alpha = 0$.

Speed and Scale

- $$\begin{aligned} w(x) &= -2\int_0^x \int_\xi^x S'(\eta)d\eta m(\xi)g(\xi)d\xi + \beta\big(S(x)-S(0)\big) \\ &= -2\int_0^x \big(S(x)-S(\xi)\big)m(\xi)g(\xi)d\xi + \beta\big(S(x)-S(0)\big) \end{aligned}$$
- Since $w(1)=0$,

$$\beta = \frac{2}{S(1)-S(0)}\int_0^1 \big(S(1)-S(\xi)\big)m(\xi)g(\xi)d\xi$$

Speed and Scale

$$\begin{aligned}
w(x) &= \frac{2}{S(1)-S(0)}\Big(\big(S(x)-S(0)\big)\int_0^1 \big(S(1)-S(\xi)\big)m(\xi)g(\xi)d\xi \\
&\qquad - \big(S(1)-S(0)\big)\int_0^x \big(S(x)-S(\xi)\big)m(\xi)g(\xi)d\xi\Big) \\
&= \frac{2}{S(1)-S(0)}\Big(\big(S(x)-S(0)\big)\int_x^1 \big(S(1)-S(\xi)\big)m(\xi)g(\xi)d\xi \\
&\quad + \int_0^x \Big[\big(S(x)-S(0)\big)\big(S(1)-S(\xi)\big) - \big(S(1)-S(0)\big)\big(S(x)-S(\xi)\big)\Big] \\
&\qquad\qquad m(\xi)g(\xi)d\xi \\
&= 2\mathbb{P}[T_1 < T_0]\int_x^1 \big(S(1)-S(\xi)\big)m(\xi)g(\xi)d\xi \\
&\qquad +2\mathbb{P}[T_0 < T_1]\int_0^x \big(S(\xi)-S(0)\big)m(\xi)g(\xi)d\xi
\end{aligned}$$

Speed and Scale

- Theorem:

$$\mathbb{E}_x\Big[\int_0^T g(X_s)ds\Big] = \int_0^1 G(x,\xi)g(\xi)d\xi$$

for the Green function

$$G(x,\xi) = \begin{cases} 2\frac{S(x)-S(0)}{S(1)-S(0)} \cdot \big(S(1)-S(\xi)\big)m(\xi), & x \le \xi \le 1, \\ 2\frac{S(1)-S(x)}{S(1)-S(0)} \cdot \big(S(\xi)-S(0)\big)m(\xi), & 0 \le \xi \le x, \end{cases}$$

- Take $g(x) = 1_{[x_1,x_2]}$ to see:

$$\int_{x_1}^{x_2} G(x,\xi)d\xi = \text{ mean time spent in } [x_1,x_2].$$

Exercise

- Show
$$\mathbb{E}_x[T^2] = 2\int_0^1\int_0^1 G(x,\xi)G(\xi,\eta)d\eta d\xi.$$

Speed and Scale

- Consider diffusion $\mathcal{X} = (X_t)_{t\geq 0}$ with μ and σ^2.
- Let $\tau(t)$ be such that

$$d\tau = m(X_t)dt.$$

- Then

$$S(X_{\tau(t)})_{t\geq 0}$$

is a Brownian motion.

Speed and Scale

- Example: Mean absorption time for Wright-Fisher diffusion

$$\mu(x) = 0, \qquad \sigma^2(x) = x(1-x).$$

- $$S(x) = \int_0^x \exp\Big(-2\int_0^y \frac{\mu(z)}{\sigma^2(z)} dz\Big) dy = x,$$
$$m(x) = \frac{1}{x(1-x)}.$$

- $$\mathbb{E}_x[T] = 2\int_x^1 x(1-\xi)\frac{1}{\xi(1-\xi)} d\xi + 2\int_0^x (1-x)\xi \frac{1}{\xi(1-\xi)} d\xi$$
$$= -2\big(x\log x + (1-x)\log(1-x)\big)$$

Speed and Scale

- Alternative: use the coalescent

$$\begin{aligned}\mathbb{E}_x[T] &= \int_0^\infty \mathbb{P}_x[T > t]dt \\ &= \int_0^\infty \sum_{n=2}^\infty \mathbb{P}[K_t = n](1 - x^n - (1-x)^n)dt\end{aligned}$$

and

$$\int_0^\infty \mathbb{P}_x[K_t = n]dt = \mathbb{E}_x\Big[\int_0^\infty 1_{K_t=n}dt\Big] = \mathbb{E}[T_n] = \frac{2}{n(n-1)}$$

Speed and Scale

- $$\sum_{n=2}^{\infty}\frac{1}{n(n-1)}x^{n}=\int_0^x\int_0^y\sum_{n=2}^{\infty}z^{n-2}dz=-\int_0^x\log(1-y)dy$$
  $$=(1-y)\log(1-y)+1-y\Big|_0^x=(1-x)\log(1-x)+1-x$$

  and so

  $$\mathbb{E}_x[T]=\sum_{n=2}^{\infty}\frac{2}{n(n-1)}(1-x^n-(1-x)^n)$$
  $$=-2\big(-1+(1-x)\log(1-x)+1-x+x\log(x)+x\big)$$
  $$=-2(x\log(x)+(1-x)\log(1-x))$$

Conditioned Diffusions

- $\mathcal{X}$: Wright-Fisher diffusion, modeling frequency of allele A
- $\{\text{Fix}\}$: event of eventual fixation of the A allele and

$$h(x) := \mathbb{P}_x[\text{Fix}].$$

- $$\begin{aligned}\mathbb{E}_x[f(X_t)|\text{Fix}] &= \frac{\mathbb{E}_x[f(X_t), \text{Fix}]}{h(x)} \\ &= \frac{\mathbb{E}_x\left[f(X_t)\mathbb{P}_x[\text{Fix}|X_t]\right]}{h(x)} \\ &= \mathbb{E}_x\left[\frac{f(X_t)h(X_t)}{h(x)}\right].\end{aligned}$$

Conditioned Diffusions

- Generator of the conditioned process

$$
\begin{aligned}
(G^*f)(x) &= \lim_{t\to 0}\tfrac{1}{t}\big(\mathbb{E}_x[f(X_t)|\text{Fix}] - f(x)\big)\\
&= \lim_{t\to 0}\tfrac{1}{t}\left(\mathbb{E}_x\left[\frac{f(X_t)h(X_t)}{h(x)}\right] - f(x)\right)\\
&= \frac{(Gfh)(x)}{h(x)}.
\end{aligned}
$$

Conditioned Diffusions

- $\mathcal{X}$: diffusion with μ, σ^2, i.e.,

$$(Gf)(x) = \mu(x)f'(x) + \tfrac{1}{2}\sigma^2(x)f''(x).$$

- We computed

$$h(x) = \frac{S(x) - S(0)}{S(1) - S(0)}.$$

- Is conditioned process $\mathcal{X}^*$ again a diffusion?
- If yes, what is $\mu^*, (\sigma^2)^*$?

Conditioned Diffusions

- Assume $S(0) = 0$

$$
\begin{aligned}
(G^* f)(x) &= \frac{(Gfh)(x)}{h(x)} \\
&= \frac{\mu(x)\big(S(x)f'(x) + S'(x)f(x)\big)}{S(x)} \\
&\qquad + \frac{\frac{1}{2}\sigma^2(x)\big(S(x)f''(x) + 2S'(x)f'(x) + S''(x)f(x)\big)}{S(x)} \\
&= \big(\mu(x) + \tfrac{1}{2}\sigma^2(x)\frac{S'(x)}{S(x)}\big)f'(x) + \tfrac{1}{2}\sigma^2(x)f''(x)
\end{aligned}
$$

- Conditioned diffusion has

$$\mu^*(x) = \mu(x) + \tfrac{1}{2}\sigma^2(x)\frac{S'(x)}{S(x)}, \qquad (\sigma^2)^*(x) = \sigma^2(x).$$

Conditioned Diffusions

- Assume a new allele enters a population. If it fixes, how long does this take?
- Consider diffusion with $\mu(x) = 0, \sigma^2(x) = x(1-x)$, conditioned on fixation, i.e.

$$\mu^*(x) = \frac{\sigma^2(x)}{x} = 1 - x, \qquad (\sigma^2)^*(x) = x(1-x).$$

Conditioned Diffusions

- Thus,

$$S^*(x) = \int_1^x \exp\Big(-2\int_1^y \frac{1}{z}dz\Big)dy = \int_1^x \frac{1}{y^2}dy$$
$$= 1 - \frac{1}{x} = -\frac{1-x}{x}$$
$$m^*(x) = \frac{x^2}{x(1-x)} = \frac{x}{1-x}$$

- So,

$$G(\varepsilon,\xi)\xrightarrow{\varepsilon\to 0} 2(S(1)-S(\xi))m(\xi) = 2$$
$$\mathbb{E}_0[T] = \int_0^1 2dt = 2.$$

Applications

Ewens Sampling Formel

- Sample of size n
- $a_i := \#$ alleles that appear i **times in the sample**
- **What probability does a configuration**

  $$\mathbf{(a_1, \ldots, a_n)?}$$

  have?
- Example: $n = 2$

  $$(2,0) : 2 \text{ alleles with frequency } 1$$
  $$(0,1) : 1 \text{ alleles with frequency } 2$$

Ewens-Sampling Formel

- For $N \to \infty, \mu \to 0$, such that $\mathbf{2N}\mu \to \theta$,

$$\mathbb{P}[(\mathbf{a_1,\dots,a_n})] = \frac{\mathbf{n!}}{\theta\cdots(\theta+\mathbf{n}-1)}\frac{\theta^{\sum \mathbf{a_j}}}{\mathbf{a_1!\cdots a_n!}\cdot 1^{\mathbf{a_1}}\cdots \mathbf{n}^{\mathbf{a_n}}}$$

- Conjectured by W. J. Ewens (1972), proved by S. Karlin and J. McGregor (1972)
- Examples:

$$n = 2: \qquad \mathbb{P}[(2,0)] = \frac{2}{\theta(\theta+1)}\frac{\theta^2}{2} = \frac{\theta}{\theta+1}$$

$$\mathbb{P}[(0,...,0,1)] \xrightarrow{\theta\to 0} 1 \quad \text{(all alleles equal)}$$

$$\mathbb{P}[(n,0,...,0)] \xrightarrow{\theta\to\infty} 1 \quad \text{(all alleles different)}$$

The coalescent and the infinite alleles model

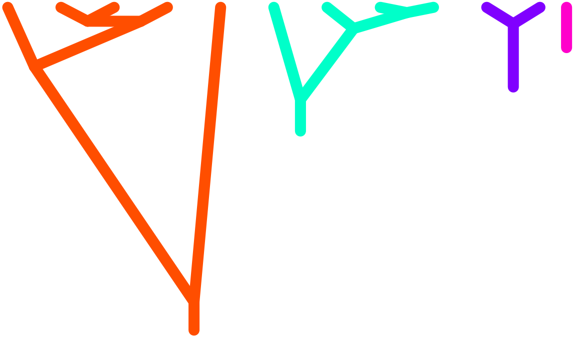
- Coalesce any two lines at rate 1
- Poisson process with rate $\frac{\theta}{2}$ on the tree gives mutations
- Two individuals carry the same allele iff they are not separated by a mutation event

$$\Leftrightarrow$$

- Coalesce any two lines at rate 1
- Every line is killed at rate $\frac{\theta}{2}$
- Two individuals carry the same allele iff they belong to the same part of the tree

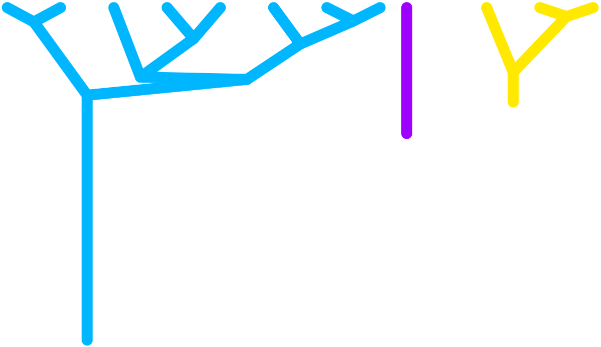
Genealogien

- Mutation (Rate $\theta/2$ pro Linie); Koaleszenz (Rate 1 pro Paar)



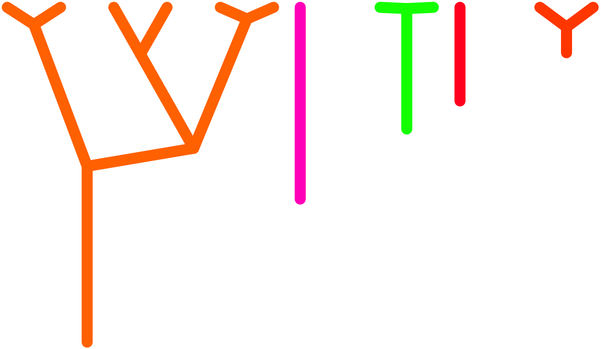
Genealogien

- Mutation (Rate $\theta/2$ pro Linie); Koaleszenz (Rate 1 pro Paar)



Genealogien

- Mutation (Rate $\theta/2$ pro Linie); Koaleszenz (Rate 1 pro Paar)



Hoppe's urn

- Families in the coalescent with killing

$$\Leftrightarrow$$

- Urn with one colored (mass 1) and one black ball (mass θ)
- Draw ball relative to its weight
- Colored ball $\rightarrow$ add new ball with same color
- Black ball $\rightarrow$ add new ball with new color
- Stop when the urn contains n colored balls
- Two balls are in same family $\iff$ they carry same color

Hoppe's urn

- Why is this the same?
- Coalescent has $k+1$ lines

$$\mathbb{P}[\text{next step is killing}] = \frac{\frac{\theta}{2}(k+1)}{\binom{k+1}{2}\frac{\theta}{2}(k+1)} = \frac{\theta}{\theta+k}.$$

- Hoppe's urn with k colored balls

$$\mathbb{P}[\text{next balls has new color}] = \frac{\theta}{\theta+k}.$$

- Hoppe's urn generates coalescent with killing forward in time

Number of allleles

- Let $\eta_k = 1$ iff kth ball in Hoppe''s urn is black (otherwise $\eta_k = 0$)
$$\mathbb{P}[\eta_k = 1] = \frac{\theta}{\theta + k - 1}.$$
- Now,
$$\mathbb{E}[\text{number of alleles}] = \mathbb{E}\Big[\sum_{k=1}^{n} \eta_k\Big] = \sum_{k=1}^{n} \frac{\theta}{\theta + k - 1}$$
$$\mathbb{V}[\text{number of alleles}] = \sum_{k=1}^{n} \mathbb{V}\big[\eta_k\big] = \sum_{k=1}^{n} \frac{\theta(k-1)}{(\theta + k - 1)^2}$$

Ewens Sampling formula: A simple proof

- Proof of Ewens Sampling formula by induction:
- $n = 1$: $\mathbb{P}[(1)] = 1$
- $n - 1 \rightarrow n$: Use Hoppe's urn and

$$\sum_{k=1}^{n} k a_k = n$$

to make the induction step:

Ewens Sampling formula: A simple proof

$$
\begin{aligned}
\mathbb{P}[(a_1,\dots,a_n)] &= \mathbb{P}[(a_1-1,a_2,\dots)]\frac{\theta}{\theta+n-1} \\
&\quad + \sum_{k=1}^{n} \mathbb{P}[(a_1,\dots,a_k+1,a_{k+1}-1,\dots)]\frac{k(a_k+1)}{\theta+n-1} \\
&= \frac{(n-1)!}{\theta\cdots(\theta+n-1)}\Big[\frac{\theta^{\sum a_j}}{(a_1-1)!a_2!a_3!\cdots 2^{a_2}3^{a_3}\cdots} \\
&\quad + \sum_{k=1}^{n} \frac{\theta^{\sum a_j} a_{k+1}k}{a_1!a_2!\cdots 1^{a_1}2^{a_2}\cdots}\frac{k+1}{k}\Big] \\
&= \frac{(n-1)!}{\theta\cdots(\theta+n-1)}\frac{\theta^{\sum a_j}}{a_1!a_2!\cdots 1^{a_1}2^{a_2}\cdots}\Big(a_1+\sum_{k=1}^{n}(k+1)a_{k+1}\Big)
\end{aligned}
$$

A fast proof

- **Loss-List**: at each coalescent or mutation event a line is lost
- coalescent has k lines: k possibilities which line is lost
- **Number of loss-lists** is $n!$
- Number of ways to put n objekts into families, such that a configuration $(a_1, a_2, \ldots)$ arises:

$$\frac{n!}{\prod_{k=1}^{n} (k!)^{a_k} a_k!}$$

A fast proof

- Fix a loss-list and a decomposition of all individuals into families, that leads to the configuration $(a_1, a_2, \dots)$
- Coalescent has k lines:

$$\mathbb{P}[\text{loss by mutation}] = \frac{k\theta/2}{k\theta/2 + \binom{k}{2}} = \frac{\theta}{\theta + k - 1},$$

$$\mathbb{P}[\text{loss by coalescence}] = \frac{\binom{k}{2}}{k\theta/2 + \binom{k}{2}} = \frac{k-1}{\theta + k - 1}$$

- $\Rightarrow$ **loss list has probability**

$$\frac{\prod_{k=1}^{n}((k-1)!)^{a_k}\theta^{a_k}}{(\theta+n-1)\cdots(\theta+1)\cdot\theta}$$

- **Multiplication gives the Ewens Sampling formula**

Number of allels

- What is
$$\mathbb{P}\Big[\sum_{i=1}^{n} a_i = k\Big]?$$
- Observe that
$$\frac{n!}{a_1!a_2!\cdots 1^{a_1}2^{a_2}\cdots}$$
is the number of permutations of $\{1,\ldots,n\}$ having a_i cycles of length i
- Recall the Stirling number of the first kind
$$S_n^k$$
is the number of permutations with k cycles

Number of alleles

- This gives

$$\mathbb{P}\Big[\sum_{j=1}^{n} a_j = k\Big] = \frac{\theta^k}{\theta \cdots (\theta + n - 1)} S_n^k$$

- As

$$\mathbb{P}\Big[(a_1, a_2, \ldots) | \sum_{j=1}^{n} a_j = k\Big] = \frac{n!}{S_n^k} \sum_{j=1}^{n} \frac{(1/j)^{a_j}}{a_j!}$$

$\sum_{j=1}^{n} a_j$ is sufficient for estimators of θ!

The site frequency spectrum

- Infinite sites model
- Polymorphic sites are called SNPs
- Size of a SNP is the number of individuals in the sample that carry the mutant allele
- What is the expected number of SNPs that have size i?

The site frequency spectrum

- S_i: number of mutations of size i
- We already computed

$$\mathbb{E}\Big[\sum_{i=1}^{n-1} S_i\Big] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

- Coalescent is in state $k \iff$ it has k lines
- A branch is of size i if exactly i of the sampled individuals are descendants of this branch

The site frequency spectrum

- We write

$$\mathbb{E}[S_i] = \sum_{k=2}^{n}\sum_{l=1}^{k}\mathbb{P}[l\text{th branch at state } k \text{ is of size } i]\cdot$$
$$\mathbb{E}[\text{number of mutations on } l\text{th branch at state } k].$$

- The easy part:

$$\mathbb{E}[\text{number of mutations on } l\text{th branch at state } k]$$
$$= \frac{\theta}{2}\cdot\mathbb{E}[\text{length of the } l\text{th branch at state } k] = \frac{\theta}{k(k-1)}$$

Polya's urn

- Urn cointains some balls with different colors
- Take out one ball, put it back and add one of the same color
- Example: start with 2 balls $'0'$ and $'1'$
- Upon adding $n-2$ balls, what is the probability that k are descendants of $'0'$?

$$\frac{1\cdots(k-1)\cdot 1\cdots(n-k-1)}{2\cdots(n-1)}\binom{n-2}{k-1}=\frac{1}{n-1}$$

Polya's urn

- Polya-urn-genealogy coincides with coalescent structure
- Reason: each line has same chance to split as each pair has the same cahnce to coalesce
- Start Polya urn with k balls and stop it with n balls: color counts give sizes of branches in the coalescent

$$
\begin{aligned}
&\mathbb{P}[l\text{th line at state } k \text{ is of size } i] \\
&\quad = \binom{n-k}{i-1} \frac{(i-1)!(k-1)\cdots(n-i-1)}{k\cdots(n-1)} \\
&\quad = \frac{k-1}{i}\binom{n-k}{i-1}\frac{i!}{(n-i)\cdots(n-1)} = \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}}\frac{k-1}{i}.
\end{aligned}
$$

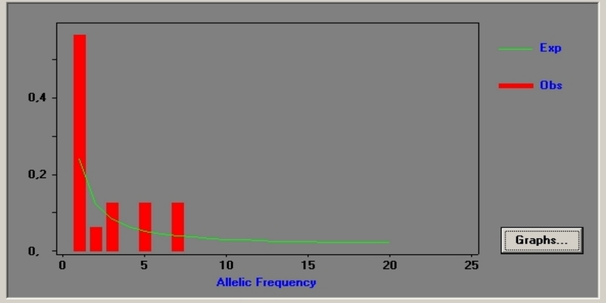
Site frequency spectrum

- Putting everything together,

$$\begin{aligned}\mathbb{E}[S_i] &= \sum_{k=2}^{n}\sum_{l=1}^{k}\frac{\binom{n-k}{i-1}}{\binom{n-1}{i}}\frac{k-1}{i}\frac{\theta}{k(k-1)}\\ &= \frac{\theta}{i}\frac{1}{\binom{n-1}{i}}\sum_{k=2}^{n}\binom{n-k}{i-1}\\ &= \frac{\theta}{i}\frac{1}{\binom{n-1}{i}}\sum_{k=2}^{n}\left(\binom{n-(k-1)}{i}-\binom{n-k}{i}\right)\\ &= \frac{\theta}{i}\frac{1}{\binom{n-1}{i}}\left(\sum_{k=1}^{n-1}\binom{n-k}{i}-\sum_{k=2}^{n}\binom{n-k}{i}\right)=\frac{\theta}{i}.\end{aligned}$$

The site frequency spectrum

- Look at the X-chromosome dataset from Hammer (2004)



The site frequency spectrum

- Why does the plot only include allele frequencies up to 20?
- Why are there so many singleton mutations in the population sample?

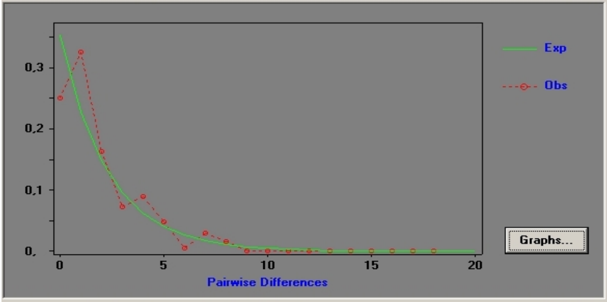
The mismatch distribution

- For a sample of size n there are $\binom{n}{2}$ pairs
- Every pair (i, j) of sequences has a number of differences S_{ij}
- The empirical distribution of $\{S_{ij} : 1 \leq i < j \leq n\}$ is the mismatch distribution.
- This may be compared to

$$\mathbb{P}[S_{ij} = k] = \left(\frac{\theta}{\theta + 1}\right)^k \frac{1}{\theta + 1}.$$

The mismatch distribution

- Look at the X-chromosome dataset from Hammer (2004)



Recombination

Some biology

- Consider two pairs of alleles A/a and B/b
- First consider a cross $AABB \times aabb$ (P-generation)
- All offspring must have $AaBb$ (F_1-generation)
- Cross one child with a homozygote $AaBb \times aabb$
- All offspring should have $AaBb$ or $aabb$ (F_2-generation)
- In fact we also find $Aabb$ and $aaBb$!
- Why?

Some biology

- The F_1-generation certainly has one set of chromosomes carrying AB and one set of chrosmosomes carrying ab
- During production of germ cells, the F_1-generation rearranges the combinations of alleles

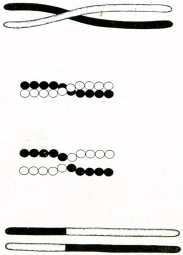
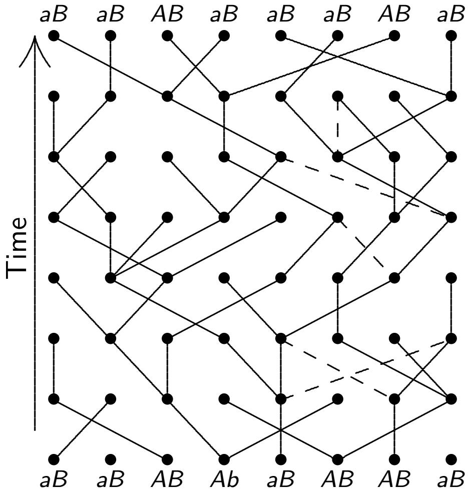


Fig. 64. Scheme to illustrate a method of crossing over of the chromosomes.

The Wright-Fisher model with recombination

- Consider the evolution of two different loci on a chromosome
- We extend the Wright-Fisher model by the rule
    - With probability r, the two loci choose two different ancestors

The Wright-Fisher model with recombination



aB *aB* *AB* *aB* *aB* *aB* *AB* *aB*
Time
aB *aB* *AB* *Ab* *aB* *AB* *AB* *aB*

- The *A*-locus is traced back using solid lines
- The *B*-locus is traced back using dashed lines

The ancestral recombination graph

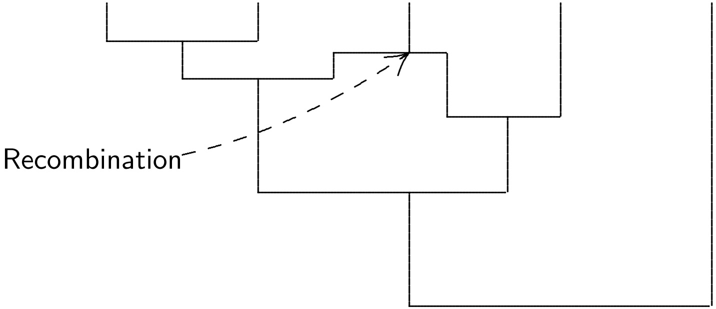
- We assume that N is big and $N \cdot r \to \rho$.
- Trace back two linked loci
- Both loci have two different ancestors after an Exp(ρ) waiting time

The ancestral recombination graph (two loci)

- Start with *n* pairs of linked loci on *n* lines
- Any pair of lines coalesces at rate 1
- A line carrying two loci splits at rate ρ
- Stop upon either
  - hitting a single line (Marjoram, Griffiths)
  - the MRCA at both loci is found (Hudson)

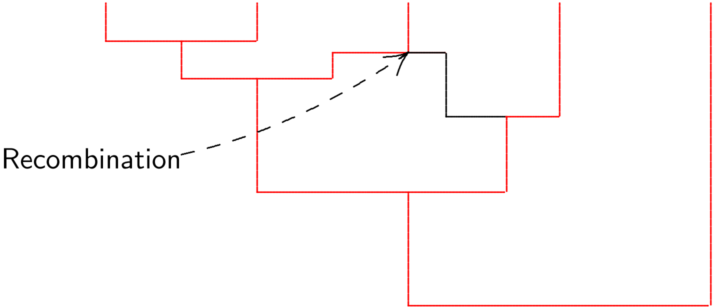
The ancestral recombination graph

- The *A* locus has the left, the *B* locus the right ancestor



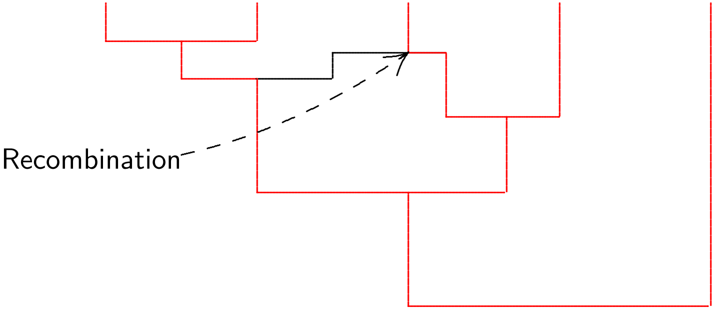
The ancestral recombination graph

- The genealogy at the *A*-locus



The ancestral recombination graph

- The genealogy at the *B*-locus

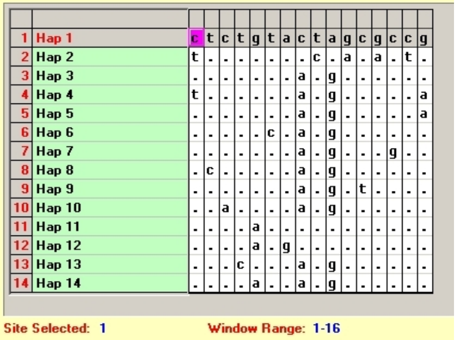


Some remarks

- It is easy to construct the ancestral recombination graph on $3, 4, \ldots$ loci
- General notion: every locus has its own genealogy
- If $\mathcal{T}_1, \ldots, \mathcal{T}_n$ are the trees at loci $1, \ldots, n$. Then $(\mathcal{T}_i)_{1 \le i \le n}$ is not a Markov chain!
- Programs like `ms` from Richard Hudson construct ancestral recombination graphs along a recombining chromosome
- Most analysis done for two loci

Recombination and data

- Look at the haplotypes of the X-chromosome dataset
- Was there recombination in the sample genealogy?



1	Hap 1	c	t	c	t	g	t	a	c	t	a	g	c	g	c	c	g
2	Hap 2	t	.	.	.	.	.	.	.	c	.	a	.	a	.	t	.
3	Hap 3	.	.	.	.	.	.	.	a	.	g	.	.	.	.	.	.
4	Hap 4	t	.	.	.	.	.	.	a	.	g	.	.	.	.	.	a
5	Hap 5	.	.	.	.	.	.	.	a	.	g	.	.	.	.	.	a
6	Hap 6	.	.	.	.	.	c	.	a	.	g	.	.	.	.	.	.
7	Hap 7	.	.	.	.	.	.	.	a	.	g	.	.	.	g	.	.
8	Hap 8	.	c	.	.	.	.	.	a	.	g	.	.	.	.	.	.
9	Hap 9	.	.	.	.	.	.	.	a	.	g	.	t	.	.	.	.
10	Hap 10	.	.	a	.	.	.	.	a	.	g	.	.	.	.	.	.
11	Hap 11	.	.	.	.	a	.	.	.	.	.	.	.	.	.	.	.
12	Hap 12	.	.	.	.	a	.	g	.	.	.	.	.	.	.	.	.
13	Hap 13	.	.	.	c	.	.	.	a	.	g	.	.	.	.	.	.
14	Hap 14	.	.	.	.	a	.	.	a	.	g	.	.	.	.	.	.

Site Selected: 1 Window Range: 1-16

The four-gamete rule

- If you can find four different gametes (which is the same as genotype or haplotype) in a sample, by considering just two segregating sites a recombination event must have taken place between the two sites.
- 'Only if' also holds

Linkage Disequilibrium

- Consider two loci with allels A, a and B, b and mutations $A \leftrightarrow a, B \leftrightarrow b$.
- Set

$$\begin{aligned} X_A &= \text{ frequency of allele } A \\ X_A &= \text{ frequency of allele } B \\ X_{AB} &= \text{ frequency of combination } AB \end{aligned}$$

Linkage Disequilibrium

- Set

$$D = X_{AB} - X_A X_B, \qquad r^2 = \frac{D^2}{X_A(1-X_A)X_B(1-X_B)}$$

- In a sample, D and r^2 can be estimated using the frequencies $\widehat{X}_A, \widehat{X}_B, \widehat{X}_{AB}$ in the sample
- Using the ancestral recombination graph, we can compute

$$\mathbb{E}[D], \qquad \sigma^2 := \frac{\mathbb{E}[D^2]}{\mathbb{E}[X_A(1-X_A)X_B(1-X_B)]} \approx \mathbb{E}[r^2]$$

  in equilibrium.
- There are two model parameters (θ, ρ)

Linkage Disequilibrium

- $\mathbb{E}[X_{AB}]$: Probability that a randomly picked individual has A at the A-locus and B at the B-locus
- $\mathbb{E}[X_A X_B]$: Probability that the A and B locus of two different individuals carry alleles A and B

Linkage Disequilibrium

- There are several explanations for

$$\mathbb{E}[D] = 0$$

  in equilibrium.

Linkage Disequilibrium

- We next compute

$$\sigma^2 = \frac{2\theta + \rho + 5}{(2\theta + \rho + 5)(2\theta + 2\rho - 3) - 4}$$

- If we consider only single sites as loci, $\theta \ll \rho$ and thus σ^2 is not much incluenced by θ
- For large ρ,

$$\sigma^2 \approx \frac{1}{\rho}.$$

Linkage Disequilibrium

- f: probability that two linked pairs (i.e. they are in the same individual) of L and R loci are heterozygous at both the L and R locus.
- g: probability that two pairs of loci, where the first pair is linked and the second pair is unlinked (i.e. comes from two different individuals), is heterozygous
- h probability that two pairs of unlinked loci are heterozygous.

Linkage Disequilibrium

- Observe

$$
\begin{aligned}
&\mathbb{E}[X_A X_a X_B X_b] = \tfrac{1}{4}h \\
&\mathbb{E}[D^2] = \tfrac{1}{2}\big(\mathbb{E}[(X_{AB} - X_A X_B)(X_{ab} - X_a X_b)] \\
&\qquad\qquad\qquad\qquad + \mathbb{E}[(X_{Ab} - X_A X_b)(X_{aB} - X_a X_B)] \\
&\ = \tfrac{1}{4}\big(\mathbb{E}[2X_{AB}X_{ab} + 2X_{Ab}X_{aB}] + 4\mathbb{E}[X_A X_B X_a X_b] \\
&\qquad\quad - 2\mathbb{E}[X_{AB}X_a X_b + X_{Ab}X_a X_B + X_{aB}X_A X_b + X_{ab}X_A X_B]\big) \\
&\ = \tfrac{1}{4}(f - 2g + h),
\end{aligned}
$$

Linkage Disequilibrium

- Set
$$\mathcal{C} = 2\theta^2 \frac{1}{1+\theta}$$
- Using the ancestral recombination graph,
$$\begin{aligned} f &= \frac{\mathcal{C}}{1+2\theta+2\rho} + \frac{2\rho}{1+2\theta+2\rho} g \\ g &= \frac{\mathcal{C}}{3+2\theta+\rho} + \frac{1}{3+2\theta+\rho} \cdot f + \frac{\rho}{3+2\theta+\rho} h \\ h &= \frac{\mathcal{C}}{6+2\theta} + \frac{4}{6+2\theta} \cdot g \end{aligned}$$
- Solving the linear system gives the assertion

Selection

Selection

- Selection = dependence of offspring distribution on genetic type

$$\begin{array}{ccccccc}
 & \textit{meiosis} & & \textit{random union} & & \textit{survival} & \\
\text{Adults} & \Longrightarrow & \text{Gametes} & \Longrightarrow & \text{Zygotes} & \Longrightarrow & \text{Adults} \\
(N) & \textit{fertility} & (\infty) & \textit{sexual} & (\infty) & \textit{viability} & (N) \\
 & \textit{selection} & & \textit{selection} & & \textit{selection} &
\end{array}$$

Some keywords

- Viability selection
- Sexual selection
- Gametic selection
- Fecundity selection
- Density and frequency dependent selection
- Pleiotropy
- Epistasis

Modeling selection

- Assume the frequency of A is x
- What is the frequency after one generation in a Wright-Fisher model?
- Allele a is less fit that A

Genotype	AA	Aa	aa
Newborns	x^2	$2x(1-x)$	$(1-x)^2$
Viability	1	$1-sh$	$1-s$
Adults	$x^2/\bar{w}$	$2x(1-x)(1-sh)/\bar{w}$	$(1-x)^2(1-s)/\bar{w}$

with

$$\bar{w} = x^2+2x(1-x)(1-sh)+(1-x)^2(1-s) = 1-s(1-x)(1-x(1-2h)).$$

Modeling selection

- h is the dominance coefficient
- $h = 0$: Selection against a recessive allele
- $h = 1$: Selection against a dominant allele
- $h = \frac{1}{2}$: gametic selection

Selection and the Wright-Fisher model

- Assume again that all individuals choose their parents independently at random
- Given $X_t = x$ in the last generation, the probability of picking an individual with allele A is

$$\begin{aligned}\tilde{x} = \frac{x^2 + (1-sh)x(1-x)}{\bar{w}} &= \frac{x(1-sh) + shx^2}{\bar{w}} \\ &= x + \frac{-x\bar{w} + x(1-sh) + shx^2}{\bar{w}} \\ &= x + \frac{-x\big(1 - s(1-x)(1-x(1-2h))\big) + x(1-sh) + shx^2}{\bar{w}} \\ &= x + \frac{sx(1-x)(1-x+2hx) - shx(1-x)}{\bar{w}} \\ &= x + \frac{sx(1-x)(1-h+x(2h-1))}{\bar{w}}\end{aligned}$$

Exercise

- Is it biologically realistic to say that all individuals pick their parent independently?

Diffusion Approximation

- $\mathcal{X}^N = (X_t^N)_{t=0,1,\dots}$: Frequency of allele A in Wright-Fisher model with selection. Selection and dominance coefficient s and h with $sN \to \alpha$
- Theorem:
$$(X_{[Nt]}^N)_{t\geq 0} \Rightarrow \mathcal{X}$$
$\mathcal{X} = (X_t)_{t\geq 0}$: Wright-Fisher diffusion with
$\mu(x) = \alpha x(1-x)(1-h+x(2h-1)), \sigma^2(x) = x(1-x)$
- 'Proof': $NX_1^N \sim B(N, \tilde{x})$, so
$$N\mathbb{E}_x[X_1^N - x] = Nsx(1-x)(1-h+x(2h-1)) + \mathcal{O}(Ns^2),$$
$$N\mathbb{E}_x[(X_1^N - x)^2] = \tilde{x}(1-\tilde{x}) = x(1-x) + \mathcal{O}(s)$$

Fixation probability

- "Kimura's formula": Probability of fixation of an allele under selection
- $(X_t)_{t\geq 0}$: frequency path of the fitter allele; this is a diffusion with
$$\mu(x) = \frac{\alpha}{2}x(1-x), \qquad \sigma^2(x) = x(1-x).$$
- The scale function is
$$\begin{aligned} S(x) &= \int_0^x \exp\Big(-2\int_0^y \frac{\mu(z)}{\sigma^2(z)}dz\Big)dy \\ &= \int_0^x e^{-\alpha y}dy = \frac{1}{\alpha}(1-e^{-\alpha x}) \end{aligned}$$

Selection in the Wright-Fisher model

- The scale function is

$$S(x) = \frac{1}{\alpha}(1 - e^{-\alpha x})$$
$$\Longrightarrow \mathbb{P}_x[\text{fix}] = P_1(x) = \frac{1 - e^{-\alpha x}}{1 - e^{-\alpha}}$$

- For a finite population, $Ns \gg 1$,

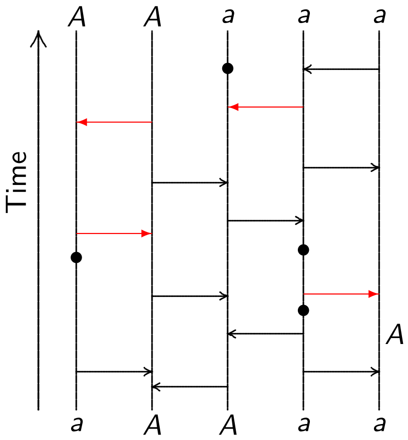
$$\mathbb{P}_x[\text{fix}] \approx \frac{1 - e^{-s}}{1 - e^{-\alpha}} \approx s.$$

- Even for highly beneficial mutations, the probability of loss is high!

Genealogies under selection

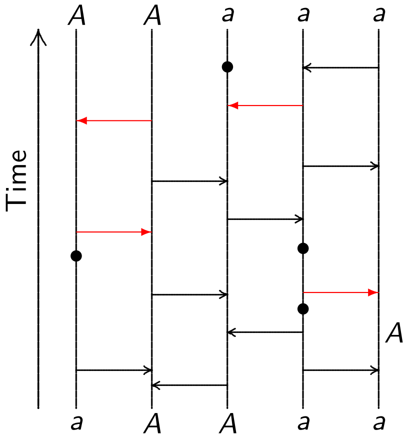
- What does the genealogy under selection look like?
- How does it differ from neutral genealogies?
- Complication: coalescence probabilities depend on allelic states, but these are unknown when looking backward in time.
- To study genealogies in equilibrium, we take a two-allele model with two-way mutation

The Moran model with selection



- The population is assumed to be in selection-mutation-drift equilibrium
- Alleles are a and A
- Each pair resamples with rate 1
- Each lines mutates with rate $\frac{\theta}{2}$
- Each line creates red arrows with rate $\frac{\alpha}{2}$

The Moran model with selection



- Black arrows can be used by any allele
- Only A alleles can use red arrows
- The state at all times can be read from this graphical representation

Generator

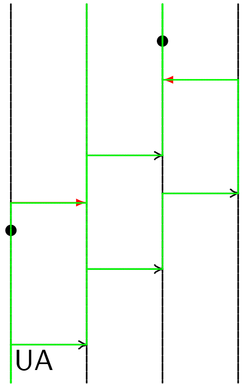
- $X = (X_t)_{t\geq 0}$: Frequency path of A
- Generator of X_t:

$$\begin{aligned} Gf(x) = \tfrac{\alpha}{2}Nx(1-x)\big(f(x+\tfrac{1}{N}) - f(x)\big) \\ + \tfrac{\theta}{2}N(1-x)(f(x+\tfrac{1}{N}) - f(x)) \\ + \tfrac{\theta}{2}Nx(f(x-\tfrac{1}{N}) - f(x)) \\ + \binom{N}{2}x(1-x)(f(x+\tfrac{1}{N}) + f(x-\tfrac{1}{N}) - 2f(x)) \end{aligned}$$

$$\xrightarrow{N\to\infty}$$

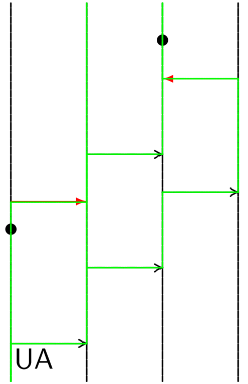
$$\big(\tfrac{\alpha}{2}x(1-x) + \tfrac{\theta}{2}(1-x) - \tfrac{\theta}{2}x\big)f'(x) + \tfrac{1}{2}x(1-x)f''(x)$$

A sample of size 2



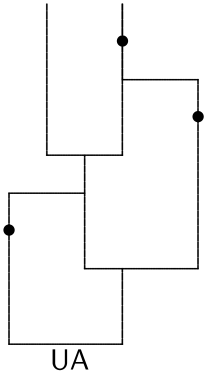
- Question: Can we trace back the ancestry of a sample?
- Again: for the sample only arrows and bullets affecting the sample are important
- Observation for a large population: when a sample is hit by a red arrow it almost always comes from outside the sample. Each line is hit at rate $\frac{1}{2}sN = \frac{\alpha}{2}$.

A sample of size 2



- From the UA forwards in time the genealogy can be found
- Determine the allele of the UA
- Red arrows can only be used by A alleleles
- At a selection event there is a continuing and an incoming branch
- When a red arrow is not used use the continuing branch
- Alleles in the sample and the ancestry can be found

The ASG

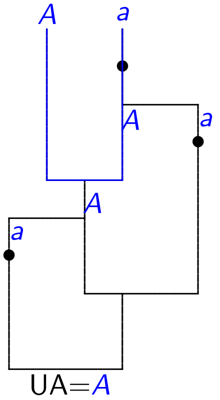


- Two lines coalesce at rate 1
- At rate $\frac{\alpha}{2}$ each line is hit by a red arrow; thus it produces a new line in the ancestry graph
- Mutations occur at rate $\frac{\theta}{2}$

The ASG

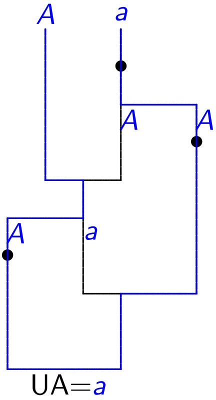
- The allele of the UA is given by the equilibrium distribution
- Finding the true genealogy can be done going from the UA forwards
- In simulations it takes a long time to reach the UA

The ASG



- Assume the UA has allele A

The ASG



- Assume the UA has allele *a*

Duality

- X_t: frequency of a neutral allele without new mutations
- K_t: number of lines in Kingman's coalescent
- 'Duality':

$$\mathbb{E}[X_t^n|X_0 = x] = \mathbb{E}[x^{K_t}|K_0 = n].$$

- Y_t: frequency of a beneficial allele without new mutations
- L_t: Number of lines in the ancestral selection graph
- 'Duality':

$$\mathbb{E}[(1 - Y_t)^n|Y_0 = y] = \mathbb{E}[(1 - y)^{L_t}|L_0 = n].$$

The structured coalescent

- Model: Two-allele (A/a) Wright-Fisher, two-way mutation (probability μ) and selection with $h = \frac{1}{2}$
- X_t: frequency of A at time t
- Assume $\mathcal{X} = (X_t)_{t<0}$ is known
- What does the genealogy of a sample at time $t = 0$ look like conditioned on $\mathcal{X}$?

The structured coalescent

- Pick one individual at time t

$$\mathbb{P}\left[\begin{array}{l}\text{ancestor is } a \text{ at}\\ \text{time } t-1\end{array}\middle|\begin{array}{l}\text{individual is } A \text{ at time } t,\\ X_{t-1}=x\end{array}\right]$$
$$=\frac{\mu(1-x)(1-\frac{s}{2})}{\mu(1-x)(1-\frac{s}{2})+x(1-\mu)}=\mu\frac{1-x}{x}+\mathcal{O}(\mu^2+\mu s)$$

- Since $X_t = X_{t-1}+\mathcal{O}(\frac{1}{\sqrt{N}})$,

$$\mathbb{P}\left[\begin{array}{l}\text{ancestor is } a \text{ at}\\ \text{time } t-1\end{array}\middle|\begin{array}{l}\text{individual is } A \text{ at}\\ \text{time } t,\ \mathcal{X}\end{array}\right]\approx\mu\frac{1-X_t}{X_t}$$

The structured coalescent

- Pick two individuals at time t

$$\begin{aligned}&\mathbb{P}\left[\begin{matrix}\text{common ancestor in } A\\ \text{at time } t-1\end{matrix}\middle|\begin{matrix}\text{both individuals } A \text{ at}\\ \text{time } t,\ X_{t-1}=x\end{matrix}\right]\\ &= \left(\frac{x(1-\mu)}{x(1-\mu)+\mu(1-x)(1-\frac{s}{2})}\right)^2 \frac{1}{Nx} = \frac{1}{Nx} + \mathcal{O}\left(\frac{\mu}{N}\right)\end{aligned}$$

$$\mathbb{P}\left[\begin{matrix}\text{common ancestor in } A\\ \text{at time } t-1\end{matrix}\middle|\begin{matrix}\text{both individuals } A \text{ at}\\ \text{time } t,\ \mathcal{X}\end{matrix}\right] \approx \frac{1}{NX_t}$$

$$\mathbb{P}\left[\begin{matrix}\text{both ancestors in } a \text{ at}\\ \text{time } t-1\end{matrix}\middle|\begin{matrix}\text{both individuals}\\ A \text{ at time } t,\ \mathcal{X}\end{matrix}\right] = \mathcal{O}(\mu^2)$$

The structured coalescent

- Rescale time by N
- Assume $(X_t)_{t\in\mathbb{R}}$ is known
- Rates in the structured coalescent from time t backwards

$$\begin{aligned}
\text{coalescence in } A\text{:} &\qquad \frac{1}{X_t} \\
\text{coalescence in } a\text{:} &\qquad \frac{1}{1-X_t} \\
\text{jump from } A \text{ to } a\text{:} &\qquad \tfrac{\theta}{2}\frac{1-X_t}{X_t} \\
\text{jump from } a \text{ to } A\text{:} &\qquad \tfrac{\theta}{2}\frac{X_t}{1-X_t}
\end{aligned}$$

Example

- These rates also apply for general h
- Under balancing selection ($h > 1$) and weak mutation the MRCA of a sample is far in the past

Recombination in the structured coalescent

- Assume the frequency path $\mathcal{X}$ of allele A is known
- Look at a linked B/b-locus
- The recombination probability between A/a and B/b locus is r per generation
- What does the genealogy at the B/b locus look like conditioned on $\mathcal{X}$?

Recombination in the structured coalescent

- Pick an individual at time t

$$\mathbb{P}\left[\begin{array}{l}\text{ancestor at } A/a \text{ and } B/b \text{ locus}\\ \text{identical at time } t-1\end{array}\right] = 1 - r$$

$$\mathbb{P}\left[\begin{array}{l}\text{ancestors different and ancestor of}\\ B/b\text{-locus linked to } a \text{ at time } t-1\end{array}\middle|\begin{array}{l}B/b\text{-locus linked to } A\\ \text{at time } t,\ \mathcal{X}\end{array}\right]$$
$$= r(1 - X_{t-1})$$

Recombination in the structured coalescent

- Rescale time by N, $\rho := Nr$
- Additional rates in the structured coalescent from time t backwards

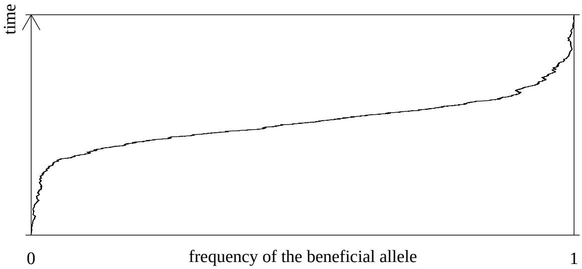
$$\begin{aligned} &\text{jump from } A \text{ to } a: \qquad \rho(1 - X_t) \\ &\text{jump from } a \text{ to } A: \qquad \rho X_t \end{aligned}$$

Example: Hitchhiking

- Assume a beneficial allele enters and fixes in a population in small time
- What does the genealogy at a linked locus look like?

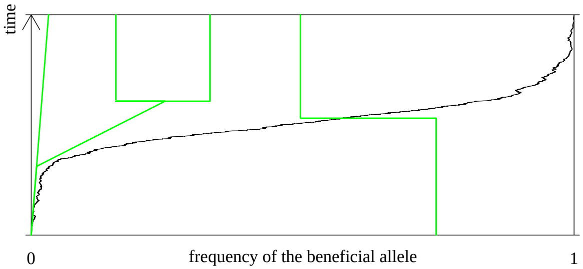
Example: Hitchhiking

- The frequency path $\mathcal{X}$ of the beneficial allele



Example: Hitchhiking

- ...and the genealogy of a linked neutral locus



Example: Hitchhiking

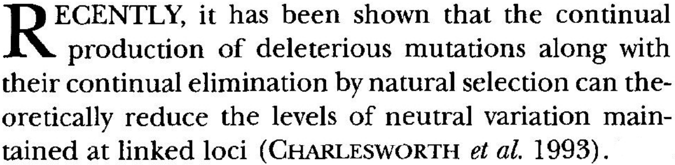
- There are several ways to detect a hitchhiking event. Can you explain why biologists expect to find

(i) reduced diversity (e.g. measured as the total number of mutations in a sample) close to a strongly beneficial locus that recently fixed?

(ii) an excess of high-frequency variants close to the selected site relative to other mutational classes?

Background selection

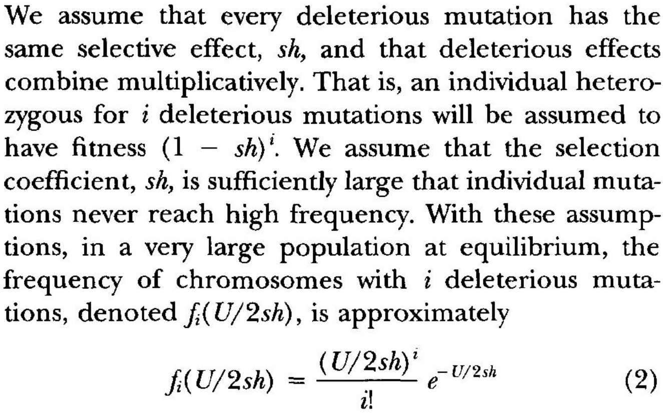
- Certainly some mutations are deleterious
- Neutral mutations on chromosomes carrying deleterious mutations are quickly lost
- Hudson and Kaplan (1995) say:



> RECENTLY, it has been shown that the continual production of deleterious mutations along with their continual elimination by natural selection can theoretically reduce the levels of neutral variation maintained at linked loci (CHARLESWORTH *et al.* 1993).

Background selection

- Assume that in each generation every individual has Pois($U/2$) new deleterious mutations and:



We assume that every deleterious mutation has the same selective effect, sh, and that deleterious effects combine multiplicatively. That is, an individual heterozygous for i deleterious mutations will be assumed to have fitness $(1 - sh)^i$. We assume that the selection coefficient, sh, is sufficiently large that individual mutations never reach high frequency. With these assumptions, in a very large population at equilibrium, the frequency of chromosomes with i deleterious mutations, denoted $f_i(U/2sh)$, is approximately

$$f_i(U/2sh) = \frac{(U/2sh)^i}{i!} e^{-U/2sh} \qquad (2)$$

Background selection

- Assume that the frequency of chromosomes carrying i deleterious mutations is f_i; set $\theta = U/2sh$
- After selection,

$$f_i' = \frac{f_i(1-hs)^i}{\sum_{j=0}^{\infty} f_j(1-hs)^j} = \frac{e^{-\theta}(\theta)^i(1-hs)^i}{i!e^{-\theta}\sum_{j=0}^{\infty}\frac{\theta^j(1-hs)^j}{j!}}$$

$$= e^{-\theta(1-hs)}\frac{\theta^i(1-hs)^i}{i!}$$

$$= \text{pois}(\theta(1-hs))(i) \approx \text{pois}(\theta - U/2)(i)$$

- After accumulating new mutations, $f'' \approx \text{pois}(\theta)$

Background selection

- A quick argument: Selection is like 'thinning' out offspring which are not fit.
- A thinned Poisson distribution is again Poisson.
- Therefore, after selection, $f' = \text{pois}(\theta(1 - sh))$
- After mutation, $f'' \approx \text{pois}(\theta)$.

Background selection

- How is variation reduced under background selection?
- Charlesworth, Morgan and Charlesworth (1993) give the result

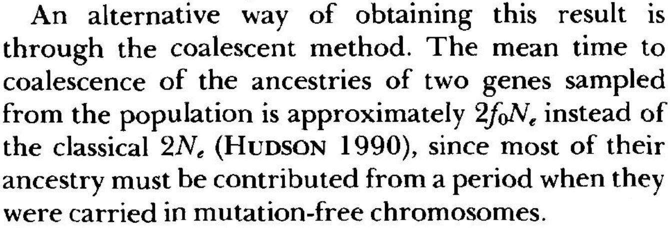
$$\pi \approx 4 f_0 N_e v.$$

where

  - $\pi := \mathbb{E}$[number of neutral mutations in a sample of size 2]
  - v: neutral mutation rate
  - N_e: number of diploid individuals
  - $f_0 := e^{-U/2sh}$ is the frequency of the class without deleterious mutations

Background selection

- They have two arguments. One is based on genealogies:



An alternative way of obtaining this result is through the coalescent method. The mean time to coalescence of the ancestries of two genes sampled from the population is approximately $2f_0N_e$ instead of the classical $2N_e$ (HUDSON 1990), since most of their ancestry must be contributed from a period when they were carried in mutation-free chromosomes.

Background selection

- Assume an individual carries $j > 0$ deleterious mutations. What is the time τ_{j-1} (in generations) in the past it has an ancestor carrying $j-1$ deleterious mutations?
- Using the structured coalescent and the frequencies f_i, τ_{j-1} is exponentially distributed with parameter

$$U\frac{f_{j-1}}{f_j} = U\frac{(U/2sh)^{j-1}j!}{(U/2sh)^j(j-1)!} = 2shj.$$

- So, the time in the past when the ancestor is in the mutation-free class has expectation

$$\frac{1}{2sh}\sum_{k=1}^{j}\frac{1}{k}.$$

Background selection

- Once two lines are in the mutation-free class they coalesce at rate $\frac{1}{2Nf_0}$. Since
$$2Nf_0 \gg \frac{1}{2sh}\sum_{k=1}^{j}\frac{1}{k}$$
as long as Nsh is large, most time is spent to coalesce both lines in the mutation-free class.
- Therefore,
$$\begin{aligned}&\mathbb{E}[\text{mutations in a sample of size 2}]\\&\quad = 2v \cdot \mathbb{E}[\text{coalescence time of two lines}] \approx 4vNf_0\end{aligned}$$

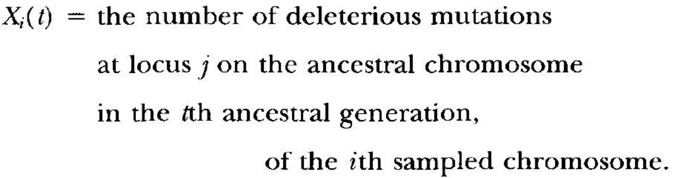
Background selection

- Can you explain why biologists

(i) expect to see patterns of a neutral evolution model with a reduced population size under background selection?

Background selection

- Look at a neutral locus linked to a locus under background selection (=locus j)
- Recombination probability is R per generation
- Pick two individuals and set



$X_i(t)$ = the number of deleterious mutations
at locus j on the ancestral chromosome
in the tth ancestral generation,
of the ith sampled chromosome.

Background selection

- Approximately, $X_1(t)$ and $X_2(t)$ are independent
- The probability of coalescence in generation t is

$$\Lambda_t = \sum_k \frac{P(X_1(t) = k)^2}{2Nf_k(u(x_j)\Delta x/2sh)}$$

where $u(x_j)\Delta x$ is the deleterious mutation rate at locus j.

Background selection

- We assume:
  - t is large and $P[X_1(t) = k] \xrightarrow{t\to\infty} P_\infty(k)$
  - $U/2sh$ is small so that we only have to worry about $k = 0, 1$
- Recall: at locus j, if $u/2$ is the deleterious mutation rate,

$$\mathbb{P}[\text{ancestor is in } k = 0 | \text{line is in } k = 1] \approx \frac{u}{2}\frac{1 - u/2sh}{u/2sh} \approx sh$$

- at the neutral locus,

$$\mathbb{P}\left[\text{ancestor is in } k = 0 \middle| \begin{array}{l}\text{line is in } k = 1,\\ \text{recombination}\end{array}\right] \approx 1 - u/2sh$$

Background selection

- In equilibrium,

$$P_\infty(1) = (1 - R - sh)P_\infty(1) + R\frac{u}{2sh},$$
$$P_\infty(1) = \frac{uR}{2sh(R - sh)}.$$

- So,

$$\Lambda_\infty \approx \frac{\left(\frac{uR}{2sh(R+sh)}\right)^2}{u/2sh} + \frac{\left(1 - \frac{uR}{2sh(R+sh)}\right)^2}{1 - u/2sh}$$
$$\approx \frac{uR^2}{2sh(r + sh)^2} + \left(1 - \frac{2uR}{2sh(R + sh)}\right)(1 + u/sh)$$
$$\approx 1 + \frac{uR^2 - 2uR(R + sh) + u(R + sh)^2}{2sh(R + sh)^2} = 1 + \frac{ush}{2(r + sh)^2}$$

Background selection

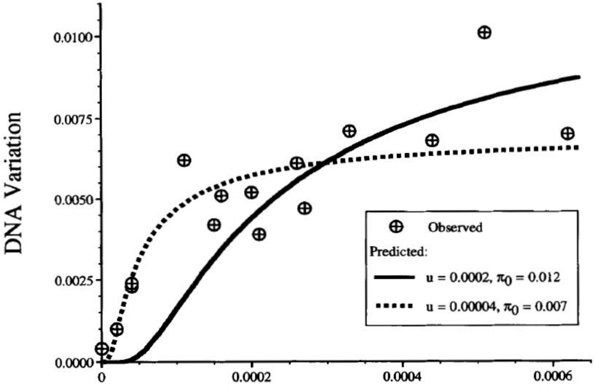
- The mean time to coalescence is approximately

$$\Lambda_\infty^{-1} \approx 1 - \frac{ush}{2(R+sh)^2} \approx 1 - \frac{u}{4R}.$$

- Especially: the coalescence time increases with distance to the selected locus.

Background selection

- low recombination rate ⇒ variation reduced
- data from third chromosome of *D. melanogaster*:



Selection

Selection

- Selection = dependence of offspring distribution on genetic type

$$\begin{matrix}\text{Adults}\\(N)\end{matrix} \xRightarrow[\textit{fertility selection}]{\textit{meiosis}} \begin{matrix}\text{Gametes}\\(\infty)\end{matrix} \xRightarrow[\textit{sexual selection}]{\textit{random union}} \begin{matrix}\text{Zygotes}\\(\infty)\end{matrix} \xRightarrow[\textit{viability selection}]{\textit{survival}} \begin{matrix}\text{Adults}\\(N)\end{matrix}$$

Some keywords

- Viability selection
- Sexual selection
- Gametic selection
- Fecundity selection
- Density and frequency dependent selection
- Pleiotropy
- Epistasis

Modeling selection

- Assume the frequency of A is x
- What is the frequency after one generation in a Wright-Fisher model?
- Allele a is less fit than A

Genotype	AA	Aa	aa
Newborns	x^2	$2x(1-x)$	$(1-x)^2$
Viability	1	$1-sh$	$1-s$
Adults	$x^2/\bar{w}$	$2x(1-x)(1-sh)/\bar{w}$	$(1-x)^2(1-s)/\bar{w}$

with

$$\bar{w} = x^2+2x(1-x)(1-sh)+(1-x)^2(1-s) = 1-s(1-x)(1-x(1-2h)).$$

Modeling selection

- h is the dominance coefficient
- $h = 0$: Selection against a recessive allele
- $h = 1$: Selection against a dominant allele
- $h = \frac{1}{2}$: gametic selection

Selection and the Wright-Fisher model

- Assume again that all individuals choose their parents independently at random
- Given $X_t = x$ in the last generation, the probability of picking an individual with allele A is

$$\begin{aligned}\tilde{x} = \frac{x^2 + (1-sh)x(1-x)}{\bar{w}} &= \frac{x(1-sh) + shx^2}{\bar{w}} \\ &= x + \frac{-x\bar{w} + x(1-sh) + shx^2}{\bar{w}} \\ &= x + \frac{-x\big(1 - s(1-x)(1-x(1-2h))\big) + x(1-sh) + shx^2}{\bar{w}} \\ &= x + \frac{sx(1-x)(1-x+2hx) - shx(1-x)}{\bar{w}} \\ &= x + \frac{sx(1-x)(1-h+x(2h-1))}{\bar{w}}\end{aligned}$$

Exercise

- Is it biologically realistic to say that all individuals pick their parent independently?

Diffusion Approximation

- $\mathcal{X}^N = (X_t^N)_{t=0,1,\dots}$: Frequency of allele A in Wright-Fisher model with selection. Selection and dominance coefficient s and h with $sN \to \alpha$
- $$(X^N_{[Nt]})_{t\geq 0} \Rightarrow \mathcal{X}$$

  $\mathcal{X} = (X_t)_{t\geq 0}$: Wright-Fisher diffusion with $\mu(x) = \alpha x(1-x)(1-h+x(2h-1)), \sigma^2(x) = x(1-x)$
- 'Proof': $NX_1^N \sim B(N, \tilde{x})$, so

  $$\begin{aligned} N\mathbb{E}_x[X_1^N - x] &= Nsx(1-x)(1-h+x(2h-1)) + \mathcal{O}(Ns^2), \\ N\mathbb{E}_x[(X_1^N - x)^2] &= \tilde{x}(1-\tilde{x}) = x(1-x) + \mathcal{O}(s) \end{aligned}$$

Fixation probability

- "Kimura's formula": Probability of fixation of an allele under selection
- $(X_t)_{t\geq 0}$: frequency path of the fitter allele; this is a diffusion with
$$\mu(x) = \frac{\alpha}{2}x(1-x), \qquad \sigma^2(x) = x(1-x).$$
- The scale function is
$$\begin{aligned} S(x) &= \int_0^x \exp\Big(-2\int_0^y \frac{\mu(z)}{\sigma^2(z)}dz\Big)dy \\ &= \int_0^x e^{-\alpha y}dy = \frac{1}{\alpha}(1-e^{-\alpha x}) \end{aligned}$$

Selection in the Wright-Fisher model

- The scale function is

$$S(x) = \frac{1}{\alpha}(1 - e^{-\alpha x})$$
$$\implies \mathbb{P}_x[\text{fix}] = P_1(x) = \frac{1 - e^{-\alpha x}}{1 - e^{-\alpha}}$$

- For a finite population, $Ns \gg 1$,

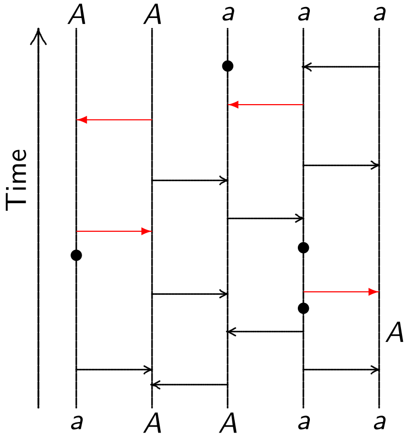
$$\mathbb{P}_x[\text{fix}] \approx \frac{1 - e^{-s}}{1 - e^{-\alpha}} \approx s.$$

- Even for highly beneficial mutations, the probability of loss is high!

Genealogies under selection

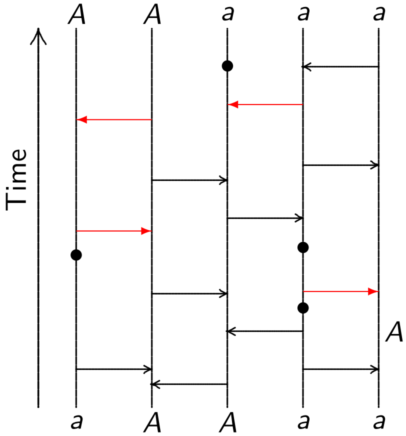
- What does the genealogy under selection look like?
- How does it differ from neutral genealogies?
- Complication: coalescence probabilities depend on allelic states, but these are unknown when looking backward in time.
- To study genealogies in equilibrium, we take a two-allele model with two-way mutation

The Moran model with selection



- The population is assumed to be in selection-mutation-drift equilibrium
- Alleles are a and A
- Each pair resamples with rate 1
- Each lines mutates with rate $\frac{\theta}{2}$
- Each line creates red arrows with rate $\frac{\alpha}{2}$

The Moran model with selection



- Black arrows can be used by any allele
- Only A alleles can use red arrows
- The state at all times can be read from this graphical representation

Generator

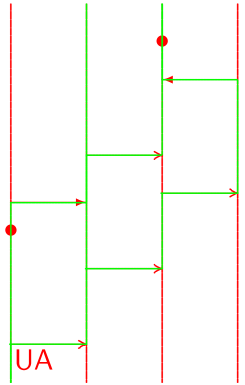
- $X = (X_t)_{t\geq 0}$: Frequency path of A
- Generator of X_t:

$$\begin{aligned} Gf(x) = &\tfrac{\alpha}{2} Nx(1-x)\big(f(x+\tfrac{1}{N}) - f(x)\big) \\ &+ \tfrac{\theta}{2} N(1-x)(f(x+\tfrac{1}{N}) - f(x)) \\ &+ \tfrac{\theta}{2} Nx(f(x-\tfrac{1}{N}) - f(x)) \\ &+ \binom{N}{2} x(1-x)(f(x+\tfrac{1}{N}) + f(x-\tfrac{1}{N}) - 2f(x)) \end{aligned}$$

$$\xrightarrow{N\to\infty}$$

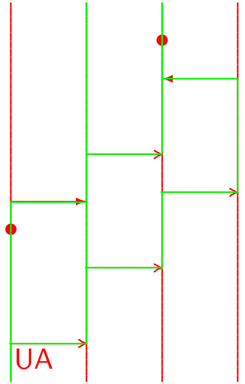
$$\big(\tfrac{\alpha}{2}x(1-x) + \tfrac{\theta}{2}(1-x) - \tfrac{\theta}{2}x\big)f'(x) + \tfrac{1}{2}x(1-x)f''(x)$$

A sample of size 2



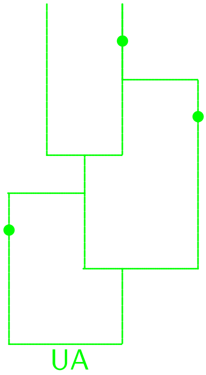
- Question: Can we trace back the ancestry of a sample?
- Again: for the sample only arrows and bullets affecting the sample are important
- Observation for a large population: when a sample is hit by a red arrow it almost always comes from outside the sample. Each line is hit at rate $\frac{1}{2}sN = \frac{\alpha}{2}$.

A sample of size 2



- From the UA forwards in time the genealogy can be found
- Determine the allele of the UA
- Red arrows can only be used by *A* alleleles
- At a selection event there is a continuing and an incoming branch
- When a red arrow is not used use the continuing branch
- Alleles in the sample and the ancestry can be found

The ASG

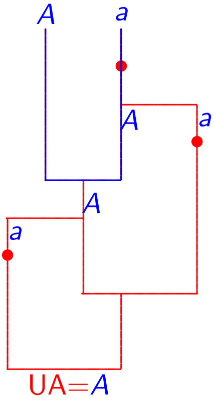


- Two lines coalesce at rate 1
- At rate $\frac{\alpha}{2}$ each line is hit by a red arrow; thus it produces a new line in the ancestry graph
- Mutations occur at rate $\frac{\theta}{2}$

The ASG

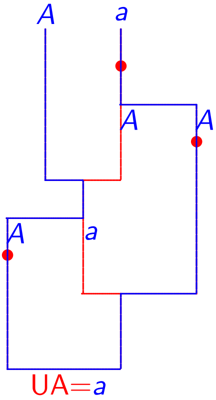
- The allele of the UA is given by the equilibrium distribution
- Finding the true genealogy can be done going from the UA forwards
- In simulations it takes a long time to reach the UA

The ASG



- Assume the UA has allele A

The ASG



- Assume the UA has allele a

Duality

- X_t: frequency of a neutral allele without new mutations
- K_t: number of lines in Kingman's coalescent
- 'Duality':
$$\mathbb{E}[X_t^n | X_0 = x] = \mathbb{E}[x^{K_t} | K_0 = n].$$
- Y_t: frequency of a beneficial allele without new mutations
- L_t: Number of lines in the ancestral selection graph
- 'Duality':
$$\mathbb{E}[(1 - Y_t)^n | Y_0 = y] = \mathbb{E}[(1 - y)^{L_t} | L_0 = n].$$

The structured coalescent

- Model: Two-allele (A/a) Wright-Fisher, two-way mutation (probability μ) and selection with $h = \frac{1}{2}$
- X_t: frequency of A at time t
- Assume $\mathcal{X} = (X_t)_{t<0}$ is known
- What does the genealogy of a sample at time $t = 0$ look like conditioned on $\mathcal{X}$?

The structured coalescent

- Pick one individual at time t

$$\mathbb{P}\left[\begin{array}{l|l}\text{ancestor is } a \text{ at} & \text{individual is } A \text{ at time } t, \\ \text{time } t-1 & X_{t-1} = x\end{array}\right] \\ = \frac{\mu(1-x)(1-\frac{s}{2})}{\mu(1-x)(1-\frac{s}{2}) + x(1-\mu)} = \mu\frac{1-x}{x} + \mathcal{O}(\mu^2 + \mu s)$$

- Since $X_t = X_{t-1} + \mathcal{O}(\frac{1}{\sqrt{N}})$,

$$\mathbb{P}\left[\begin{array}{l|l}\text{ancestor is } a \text{ at} & \text{individual is } A \text{ at} \\ \text{time } t-1 & \text{time } t,\ \mathcal{X}\end{array}\right] \approx \mu\frac{1-X_t}{X_t}$$

The structured coalescent

- Pick two individuals at time t

$$\mathbb{P}\left[\begin{array}{l}\text{common ancestor in } A\\ \text{at time } t-1\end{array}\middle|\begin{array}{l}\text{both individuals } A \text{ at}\\ \text{time } t,\ X_{t-1}=x\end{array}\right]$$

$$=\left(\frac{x(1-\mu)}{x(1-\mu)+\mu(1-x)(1-\frac{s}{2})}\right)^2\frac{1}{Nx}=\frac{1}{Nx}+\mathcal{O}\left(\frac{\mu}{N}\right)$$

$$\mathbb{P}\left[\begin{array}{l}\text{common ancestor in } A\\ \text{at time } t-1\end{array}\middle|\begin{array}{l}\text{both individuals } A \text{ at}\\ \text{time } t,\ \mathcal{X}\end{array}\right]\approx\frac{1}{NX_t}$$

$$\mathbb{P}\left[\begin{array}{l}\text{both ancestors in } a \text{ at}\\ \text{time } t-1\end{array}\middle|\begin{array}{l}\text{both individuals}\\ A \text{ at time } t,\ \mathcal{X}\end{array}\right]=\mathcal{O}(\mu^2)$$

The structured coalescent

- Rescale time by N
- Assume $(X_t)_{t\in\mathbb{R}}$ is known
- Rates in the structured coalescent from time t backwards

$$\begin{aligned}
\text{coalescence in } A\text{:} &\qquad \frac{1}{X_t} \\
\text{coalescence in } a\text{:} &\qquad \frac{1}{1-X_t} \\
\text{jump from } A \text{ to } a\text{:} &\qquad \tfrac{\theta}{2}\frac{1-X_t}{X_t} \\
\text{jump from } a \text{ to } A\text{:} &\qquad \tfrac{\theta}{2}\frac{X_t}{1-X_t}
\end{aligned}$$

Example

- These rates also apply for general h
- Under balancing selection ($h > 1$) and weak mutation the MRCA of a sample is far in the past

Recombination in the structured coalescent

- Assume the frequency path $\mathcal{X}$ of allele A is known
- Look at a linked B/b-locus
- The recombination probability between A/a and B/b locus is r per generation
- What does the genealogy at the B/b locus look like conditioned on $\mathcal{X}$?

Recombination in the structured coalescent

- Pick an individual at time t

$$\mathbb{P}\left[\begin{array}{l}\text{ancestor at } A/a \text{ and } B/b \text{ locus}\\ \text{identical at time } t-1\end{array}\right] = 1 - r$$

$$\mathbb{P}\left[\begin{array}{l}\text{ancestors different and ancestor of}\\ B/b\text{-locus linked to } a \text{ at time } t-1\end{array}\middle| \begin{array}{l}B/b\text{-locus linked to } A\\ \text{at time } t,\ \mathcal{X}\end{array}\right]$$

$$= r(1 - X_{t-1})$$

Recombination in the structured coalescent

- Rescale time by N, $\rho := Nr$
- Additional rates in the structured coalescent from time t backwards

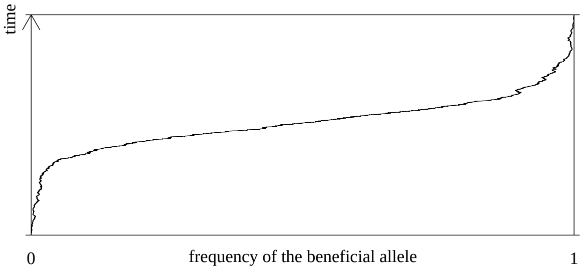
$$\begin{aligned} &\text{jump from } A \text{ to } a: && \rho(1 - X_t) \\ &\text{jump from } a \text{ to } A: && \rho X_t \end{aligned}$$

Example: Hitchhiking

- Assume a beneficial allele enters and fixes in a population in small time
- What does the genealogy at a linked locus look like?

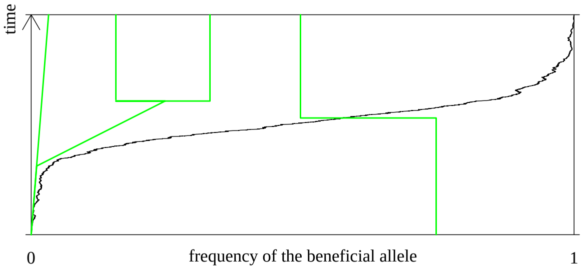
Example: Hitchhiking

- The frequency path $\mathcal{X}$ of the beneficial allele



Example: Hitchhiking

- ...and the genealogy of a linked neutral locus



Example: Hitchhiking

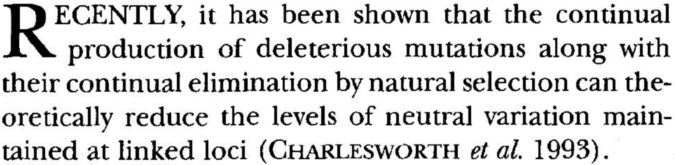
- There are several ways to detect a hitchhiking event. Can you explain why biologists expect to find

(i) reduced diversity (e.g. measured as the total number of mutations in a sample) close to a strongly beneficial locus that recently fixed?

(ii) an excess of high-frequency variants close to the selected site relative to other mutational classes?

Background selection

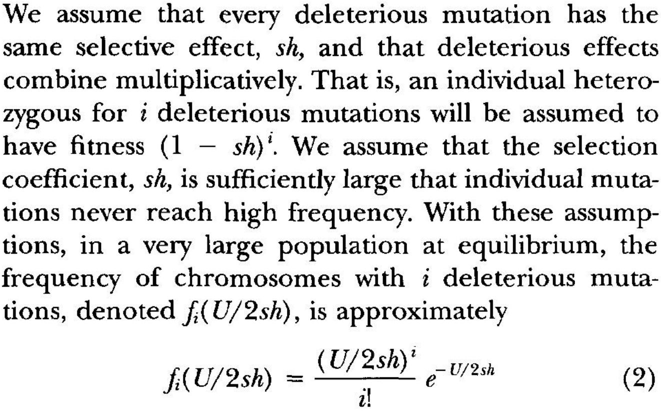
- Certainly some mutations are deleterious
- Neutral mutations on chromosomes carrying deleterious mutations are quickly lost
- Hudson and Kaplan (1995) say:



> RECENTLY, it has been shown that the continual production of deleterious mutations along with their continual elimination by natural selection can theoretically reduce the levels of neutral variation maintained at linked loci (CHARLESWORTH *et al.* 1993).

Background selection

- Assume that in each generation every individual has Pois($U/2$) new deleterious mutations and:



We assume that every deleterious mutation has the same selective effect, sh, and that deleterious effects combine multiplicatively. That is, an individual heterozygous for i deleterious mutations will be assumed to have fitness $(1 - sh)^i$. We assume that the selection coefficient, sh, is sufficiently large that individual mutations never reach high frequency. With these assumptions, in a very large population at equilibrium, the frequency of chromosomes with i deleterious mutations, denoted $f_i(U/2sh)$, is approximately

$$f_i(U/2sh) = \frac{(U/2sh)^i}{i!} e^{-U/2sh} \qquad (2)$$

Background selection

- Assume that the frequency of chromosomes carrying i deleterious mutations is f_i; set $\theta = U/2sh$
- After selection,

$$
\begin{aligned}
f_i' = \frac{f_i(1-hs)^i}{\sum_{j=0}^{\infty} f_j(1-hs)^j} &= \frac{e^{-\theta}(\theta)^i(1-hs)^i}{i!e^{-\theta}\sum_{j=0}^{\infty}\frac{\theta^j(1-hs)^j}{j!}} \\
&= e^{-\theta(1-hs)}\frac{\theta^i(1-hs)^i}{i!} \\
&= \mathrm{pois}(\theta(1-hs))(i) \approx \mathrm{pois}(\theta - U/2)(i)
\end{aligned}
$$

- After accumulating new mutations, $f'' \approx \mathrm{pois}(\theta)$

Background selection

- A quick argument: Selection is like 'thinning' out offspring which are not fit.
- A thinned Poisson distribution is again Poisson.
- Therefore, after selection, $f' = \text{pois}(\theta(1 - sh))$
- After mutation, $f'' \approx \text{pois}(\theta)$.

Background selection

- How is variation reduced under background selection?
- Charlesworth, Morgan and Charlesworth (1993) give the result

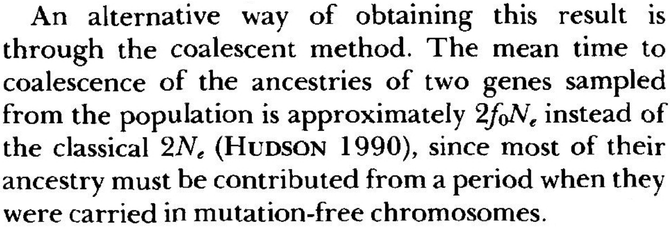
  $$\pi \approx 4 f_0 N_e v.$$

  where

  - $\pi := \mathbb{E}$[number of neutral mutations in a sample of size 2]
  - v: neutral mutation rate
  - N_e: number of diploid individuals
  - $f_0 := e^{-U/2sh}$ is the frequency of the class without deleterious mutations

Background selection

- They have two arguments. One is based on genealogies:



An alternative way of obtaining this result is through the coalescent method. The mean time to coalescence of the ancestries of two genes sampled from the population is approximately $2f_0N_e$ instead of the classical $2N_e$ (Hudson 1990), since most of their ancestry must be contributed from a period when they were carried in mutation-free chromosomes.

Background selection

- Assume an individual carries $j > 0$ deleterious mutations. What is the time τ_{j-1} (in generations) in the past it has an ancestor carrying $j-1$ deleterious mutations?
- Using the structured coalescent and the frequencies f_i, τ_{j-1} is exponentially distributed with parameter

$$U\frac{f_{j-1}}{f_j} = U\frac{(U/2sh)^{j-1}j!}{(U/2sh)^j(j-1)!} = 2shj.$$

- So, the time in the past when the ancestor is in the mutation-free class has expectation

$$\frac{1}{2sh}\sum_{k=1}^{j}\frac{1}{k}.$$

Background selection

- Once two lines are in the mutation-free class they coalesce at rate $\frac{1}{2Nf_0}$. Since

$$2Nf_0 \gg \frac{1}{2sh} \sum_{k=1}^{j} \frac{1}{k}$$

as long as Nsh is large, most time is spent to coalesce both lines in the mutation-free class.

- Therefore,

$$\begin{aligned}&\mathbb{E}[\text{mutations in a sample of size 2}]\\ &\quad = 2v \cdot \mathbb{E}[\text{coalescence time of two lines}] \approx 4vNf_0\end{aligned}$$

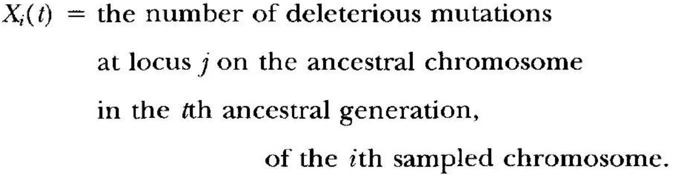
Background selection

- Can you explain why biologists

(i) expect to see patterns of a neutral evolution model with a reduced population size under background selection?

Background selection

- Look at a neutral locus linked to a locus under background selection (=locus j)
- Recombination probability is R per generation
- Pick two individuals and set



$X_i(t)$ = the number of deleterious mutations
at locus j on the ancestral chromosome
in the tth ancestral generation,
of the ith sampled chromosome.

Background selection

- Approximately, $X_1(t)$ and $X_2(t)$ are independent
- The probability of coalescence in generation t is

$$\Lambda_t = \sum_k \frac{P(X_1(t) = k)^2}{2Nf_k(u(x_j)\Delta x/2sh)}$$

where $u(x_j)\Delta x$ is the deleterious mutation rate at locus j.

Background selection

- We assume:
  - t is large and $P[X_1(t) = k] \xrightarrow{t\to\infty} P_\infty(k)$
  - $U/2sh$ is small so that we only have to worry about $k = 0, 1$
- Recall: at locus j, if $u/2$ is the deleterious mutation rate,

$$\mathbb{P}[\text{ancestor is in } k = 0 | \text{line is in } k = 1] \approx \frac{u}{2}\frac{1 - u/2sh}{u/2sh} \approx sh$$

- at the neutral locus,

$$\mathbb{P}\left[\text{ancestor is in } k = 0 \middle| \begin{matrix}\text{line is in } k = 1,\\ \text{recombination}\end{matrix}\right] \approx 1 - u/2sh$$

Background selection

- In equilibrium,

$$P_\infty(1) = (1 - R - sh)P_\infty(1) + R\frac{u}{2sh},$$
$$P_\infty(1) = \frac{uR}{2sh(R - sh)}.$$

- So,

$$\begin{aligned}\Lambda_\infty &\approx \frac{\left(\frac{uR}{2sh(R+sh)}\right)^2}{u/2sh} + \frac{\left(1 - \frac{uR}{2sh(R+sh)}\right)^2}{1 - u/2sh}\\ &\approx \frac{uR^2}{2sh(r + sh)^2} + \left(1 - \frac{2uR}{2sh(R + sh)}\right)(1 + u/sh)\\ &\approx 1 + \frac{uR^2 - 2uR(R + sh) + u(R + sh)^2}{2sh(R + sh)^2} = 1 + \frac{ush}{2(r + sh)^2}\end{aligned}$$

Background selection

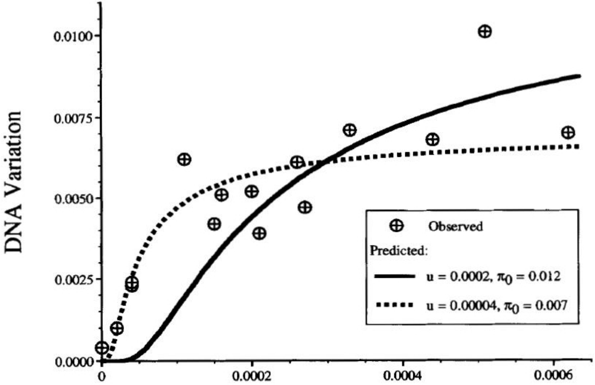
- The mean time to coalescence is approximately

$$\Lambda_{\infty}^{-1} \approx 1 - \frac{ush}{2(R+sh)^2} \approx 1 - \frac{u}{4R}.$$

- Especially: the coalescence time increases with distance to the selected locus.

Background selection

- low recombination rate ⇒ variation reduced
- data from third chromosome of *D. melanogaster*:



Neutrality tests

General task

- Assume you have gathered sequence variation data from a species. How can you decide statistically if the population evolved neutrally?
- If you are sure that the population did not evolve neutrally, which forces were responsible for the shape of the sequence variation data?

Statistical Inference

- Every statistical test consists of:
  - A null hypothesis H_0 (which is to be rejected)
  - A test statistic T (which must be computed from data)
  - The distribution of T under H_0 (which must be known from theory)
- According to these ingredients one computes

  $$p = \mathbb{P}[T \text{ more extreme than the given data}]$$

  which is the p-value.
- If $p < 0.05$ the null hypothesis is rejected. This means that one assumption of H_0 is probably not satisfied.

Tajima's D

- Under neutrality,

  $$\widehat{\theta}_\pi := \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} S_{ij} \qquad \text{and } \widehat{\theta}_W := \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

  are unbiased estimators of θ.
- Tajima's D compares these two:

  $$d := \widehat{\theta}_\pi - \widehat{\theta}_W$$

Tajima's *D*

- Tajima (1989) computed

$$\mathbb{V}[d] = c_1\theta + c_2\theta^2$$

with

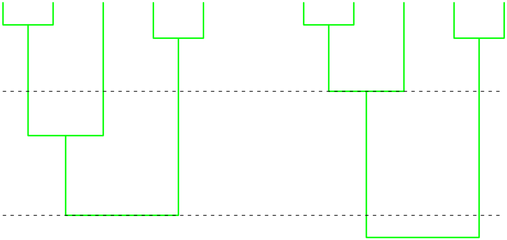
$$\begin{aligned} c_1 &= b_1 - \frac{1}{a_1}, & c_2 &= b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}, \\ b_1 &= \frac{n+1}{3(n-1)}, & b_2 &= \frac{2(n^2+n+3)}{9n(n-1)}, \\ a_1 &= \sum_{i=1}^{n-1} \frac{1}{i}, & a_2 &= \sum_{i=1}^{n-1} \frac{1}{i^2}. \end{aligned}$$

Tajima's D

- Since
$$\mathbb{E}[S(S-1)] = (a_2 + (a_1)^2)\theta^2,$$
we have the unbiased esimator
$$\widehat{\mathbb{V}}[d] = \frac{c_1}{a_1}S + \frac{c_2}{a_1^2 + a_2}S(S-1).$$
- So,
$$D := \frac{\widehat{\theta}_\pi - \widehat{\theta}_W}{\widehat{\mathbb{V}}[D]}$$
roughly has $\mathbb{E}[D] \approx 0$ and $\mathbb{V}[D] \approx 1$

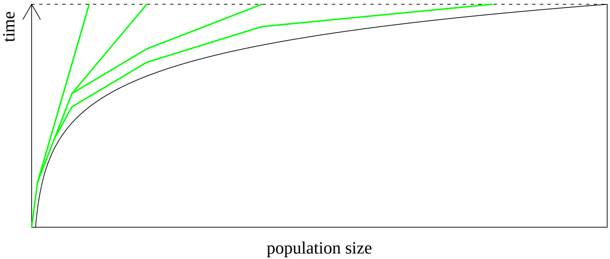
Tajima's D

- $\widehat{\theta}_W$ is equal in both trees, on average
- $\widehat{\theta}_\pi$ is higher for the right tree, on average



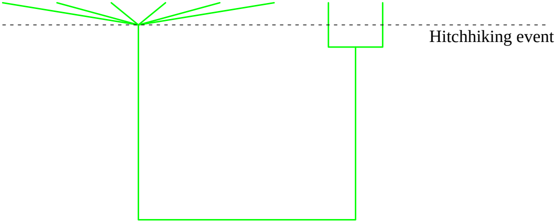
Tajima's D

- Tajima's D is expected to be negative in expanding populations



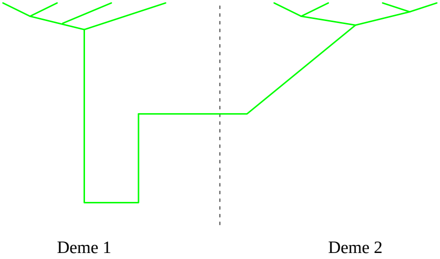
Tajima's *D*

- Tajima's D is expected to be negative after a hitchhiking event



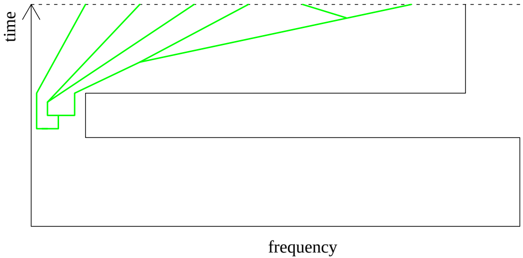
Tajima's *D*

- Tajima's *D* is expected to be positive in structured populations populations



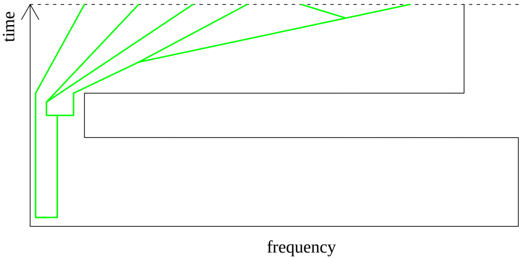
Tajima's *D*

- Tajima's *D* can be negative after a population botleneck populations



Tajima's *D*

- Tajima's *D* can also be positive after a population botleneck populations



Tajima's D

- Tajima argued that the distribution of Tajima's D might be close to a β distribution
- In practise, the distribution is found by simulation

Fu and Li's D

- Consider the result

$$\mathbb{E}[S_i] = \frac{\theta}{i}$$

- Other unbiased estimators for θ are

$$\widehat{\theta}_{S_1} = S_1, \qquad \widehat{\theta}_{S_{>1}} = \frac{\sum_{i=2}^{n-1} S_i}{\sum_{i=2}^{n-1} \frac{1}{i}}$$

Fu and Li's D

- Fu and Li (1993) use

$$d = \widehat{\theta}_{S_{>1}} - \widehat{\theta}_{S_1}, \qquad D = \frac{d}{\widehat{\mathbb{V}}[d]}$$

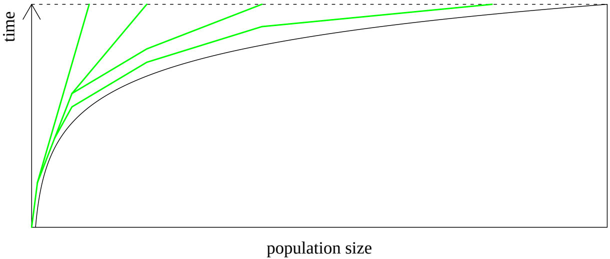
  with some expression for $\widehat{\mathbb{V}}[d]$ for a statistic testing the neutral model

- Again, approximately,

$$\mathbb{E}[D] \approx 0, \qquad \mathbb{V}[D] \approx 1$$

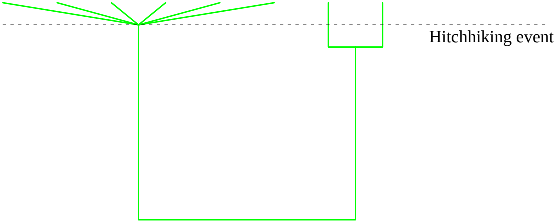
Fu and Li's *D*

- Fu and Li's *D* is expected to be negative in expanding populations



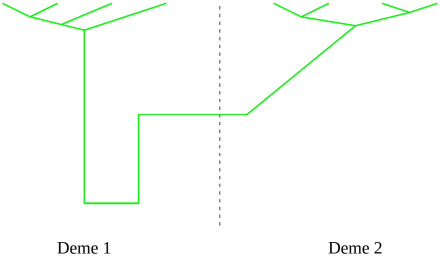
Fu and Li's *D*

- Fu and Li's *D* is expected to be negative after a hitchhiking event



Fu and Li's *D*

- Fu and Li's *D* is expected to be positive in structured populations populations



Exercise

- Can you draw a genealogical tree (with mutations on the tree) for the case that
  - Tajima's D is negative and Fu and Li's D is approximately 0?
  - Fu and Li's D is positive and Tajima's D is approximately 0?
- Why are Tajima's and Fu and Li's D said to be statistics based on the site frequency spectrum?

The McDonald-Kreitman test

- Look at coding regios on a chromosome
- The genetic code: translation table from $4^3 = 64$ possible tripels of bases (codons) to 20 different amino acids (plus `start` and `stop` states
- Example: Lysin encoded by `AAA` and `AAG`
- Some mutations in the DNA sequence do not change the amino acid sequence (synonymous mutations)
- Others change in the amino acid sequence (non-synonymuos mutations)

The McDonald-Kreitman test

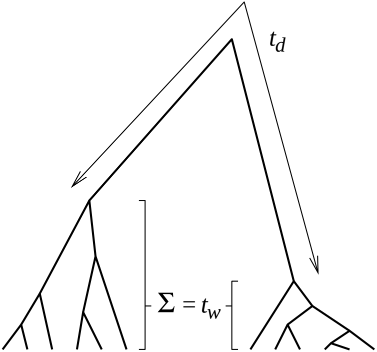
- Assume you have sequences from the same coding regions in two different species
- Some mutations are 'private' to one species
  → polymorphism within a species
- Some mutations are fixed between the species (substitutions)
  → divergence between species

The McDonald-Kreitman test

- Synonymous mutations occur at rate μ_s
- Non-synonymous mutations occur at rate μ_n

The McDonald-Kreitman test

- t_d: time in the tree for substitutions
- t_w: time in the tree for private mutations



The McDonald-Kreitman test

- Data can be arranged in a 2×2 contingency table

	diverged	polymorphic	Total
synonymous	$\mu_s t_d$	$\mu_s t_w$	$\mu_s t$
non-synonymous	$\mu_n t_d$	$\mu_n t_w$	$\mu_n t$
Total	μt_d	μt_w	μt

The McDonald-Kreitman test

- Example from McDonald, Kreitman (1991): *Adh* gene in 12 sequences from *D. melanogaster* and 6 from *D. simulans* and 12 from *D. yakuba*.

	diverged	polymorphic	Total
synonymous	17	42	59
non-synonymous	7	2	9
Total	24	44	68

The McDonald-Kreitman test

- Example: Fisher's exact test gives $p < 0.01$
- Interpretation: Excess of non-synonymous divergence
- These indicate adaptively driven mutations
- An excess of non-synonymous private mutations would indicate background selection