NTU, December 2014

Thinking about selection and genetic drift... in terms of trajectories

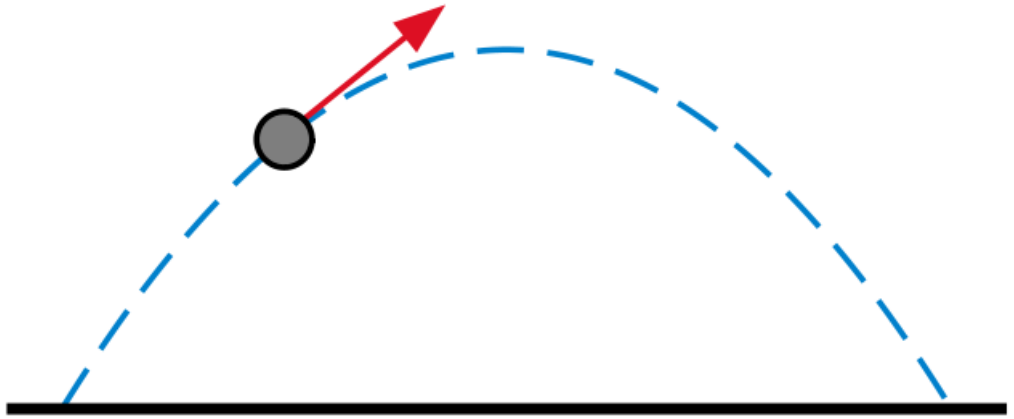
David Waxman
Centre for Computational Systems Biology
Fudan University, Shanghai PRC

Guanghua Tower at Fudan University

Content

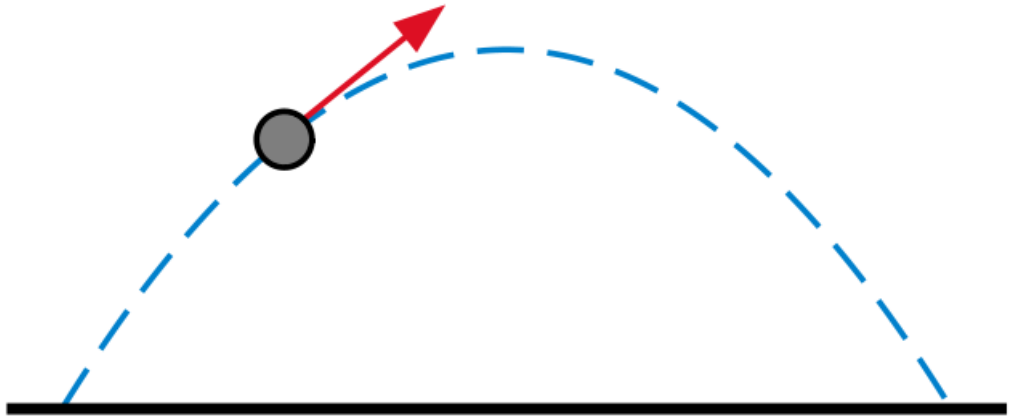
- Introduce trajectories
- Introduce selection
- Introduce random genetic drift
- Talk about conditioned trajectories and some implications

Physics deals with trajectories



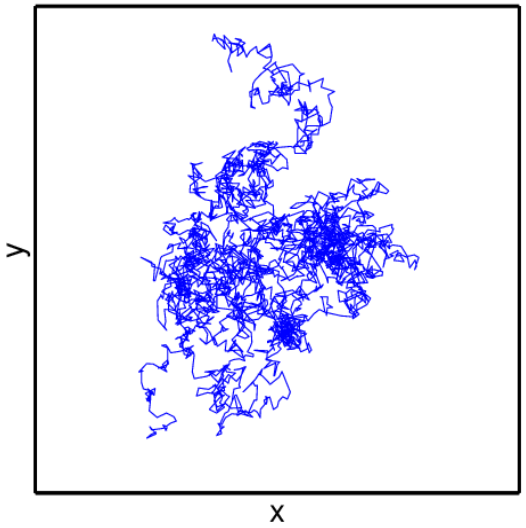
These obey well established laws (Newton,...)

Physics deals with trajectories



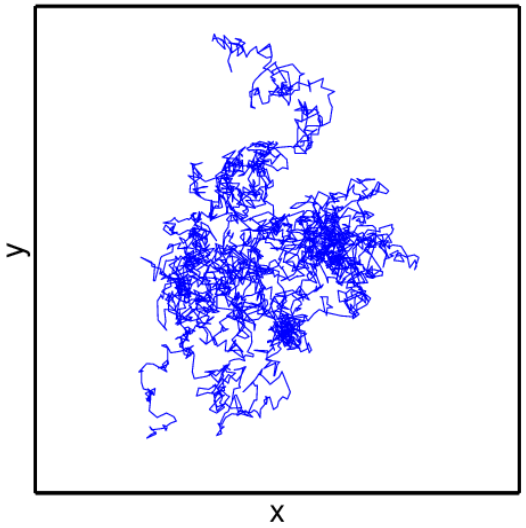
These are ‘deterministic’ trajectories

Physics deals with trajectories



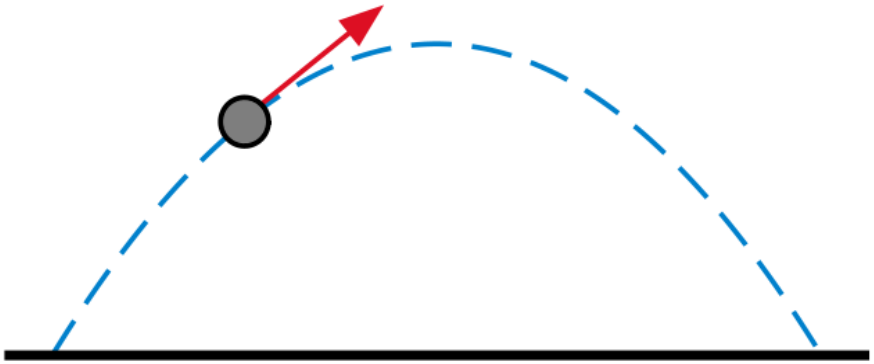
There are natural generalisations that include Brownian motion - randomness
(Langevin, Einstein, Bachelier, ...)

Physics deals with trajectories

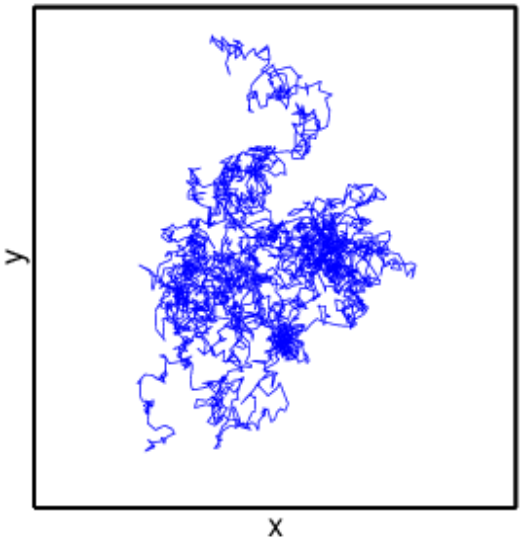


These are ‘stochastic’ trajectories

In biology there are also trajectories that are analogous to



Deterministic



Stochastic

But ... trajectories of biology are in a different space

- Simplest case: a single genetic locus with two alternative genes - i.e., two *alleles*

- The trajectory is in the space of allele frequencies

What is allele frequency - the variable in a biological trajectory?

Answer:

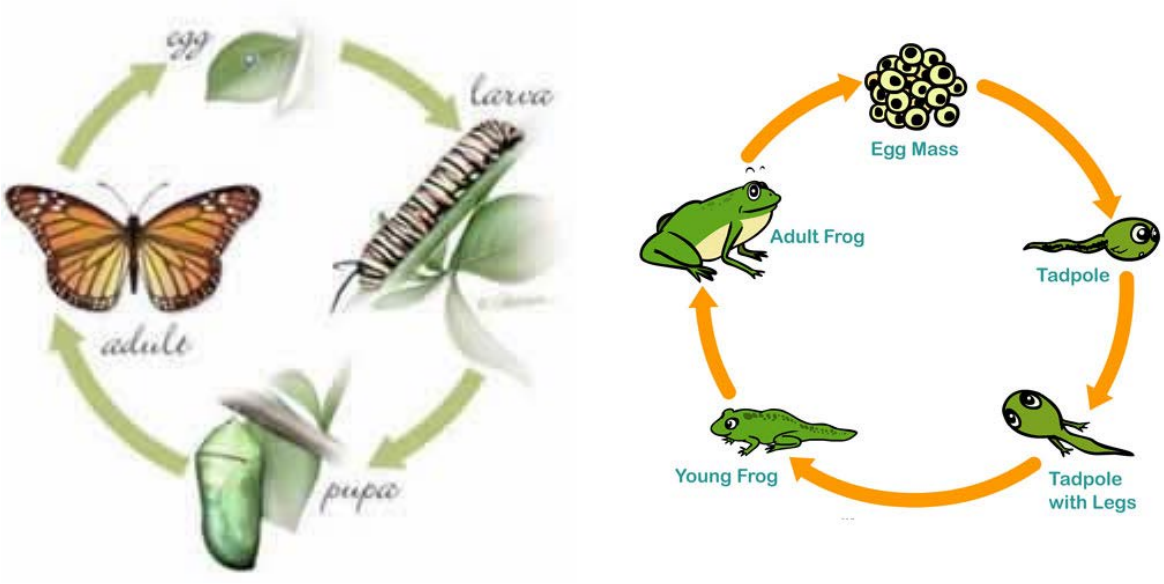
With the two alleles A and B, the variable is

$$\begin{aligned} X_t &= \text{frequency (proportion) of } A \text{ alleles in generation } t \\ &= \frac{\text{number of } A \text{ alleles in population}}{\text{number of } A \text{ alleles } + \text{ number of } B \text{ alleles}} \end{aligned}$$

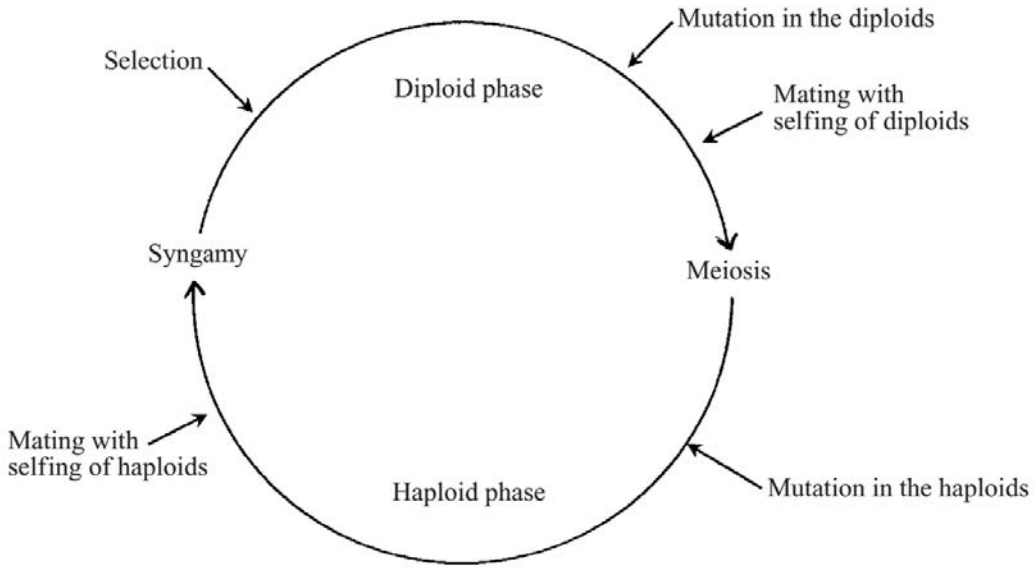
And because it is a proportion

$$0 \leq X_t \leq 1.$$

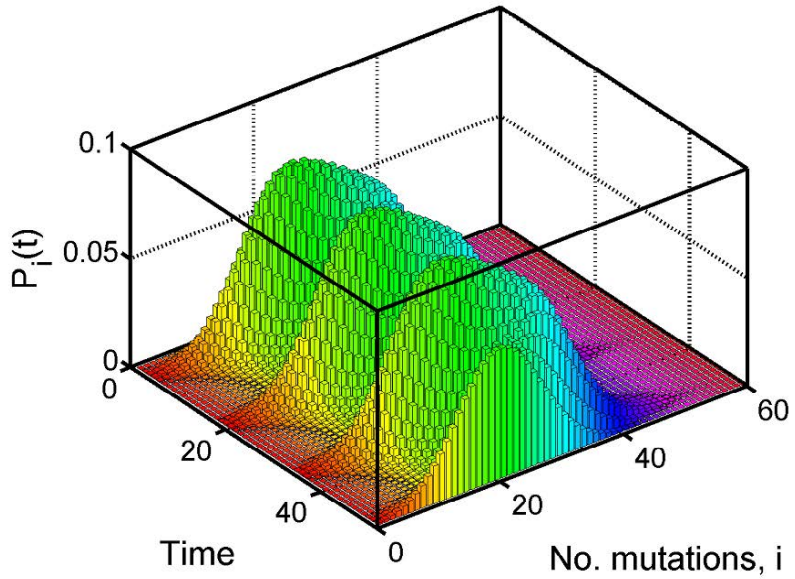
Behaviour of X_t comes from the processes of a lifecycle



Lifecycles can be complicated (ferns and fungi)

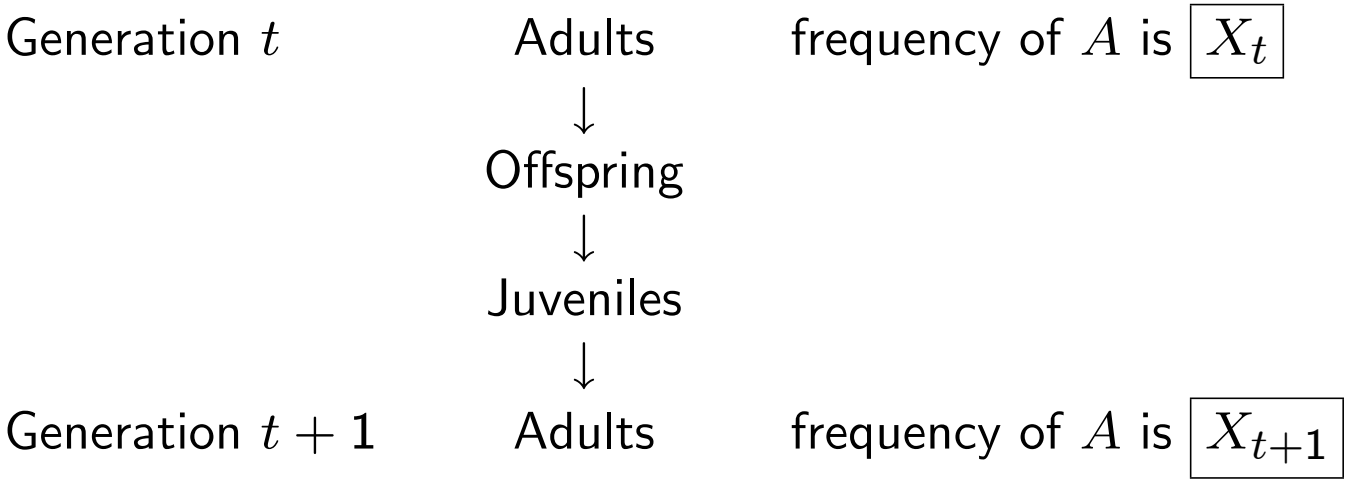


Life cycle with haploid and diploid selfing



Distribution of mutations in ferns and fungi - oscillates

Lifecycle


Generation t
Adults
frequency of A is X_t
Offspring
Juveniles
Generation $t+1$
Adults
frequency of A is X_{t+1}

Change of allele frequency when population size is infinite $(N = \infty)$

$$X_{t+1} = X_t + M(X_t)$$

Deterministic

$M(x)$ comes from evolutionary forces (selection, mutation, migration,...)

Example: deterministic dynamics with selection

Each individual carries 2 genes (diploid).

Fitnesses of AA, AB and BB individuals are $1+2s$, $1+s$ and 1

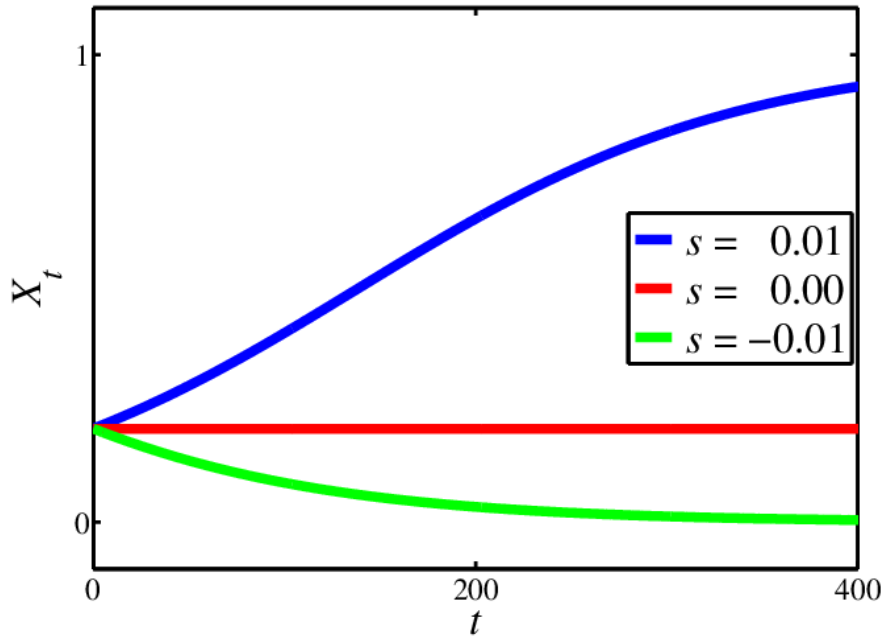
$$\frac{\text{mean number of offspring of } AA \text{ individuals}}{\text{mean number of offspring of } BB \text{ individuals}} = 1+2s$$

$$\frac{\text{mean number of offspring of } AB \text{ individuals}}{\text{mean number of offspring of } BB \text{ individuals}} = 1+s$$

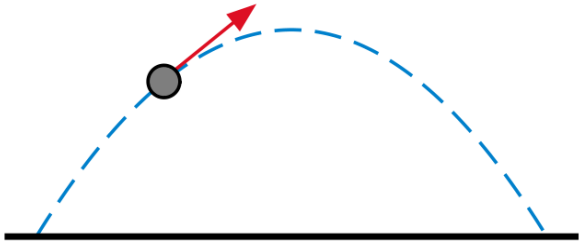
Leads to

$$X_{t+1} = X_t + sX_t(1-X_t)$$

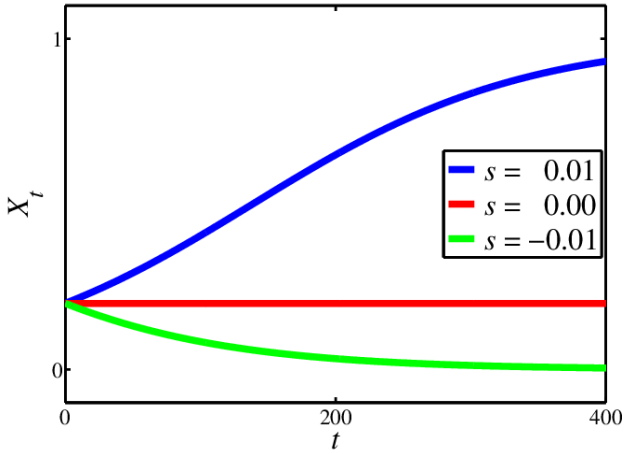
Consequences of $X_{t+1} = X_t + sX_t(1 - X_t)$



Deterministic trajectories in physics and biology

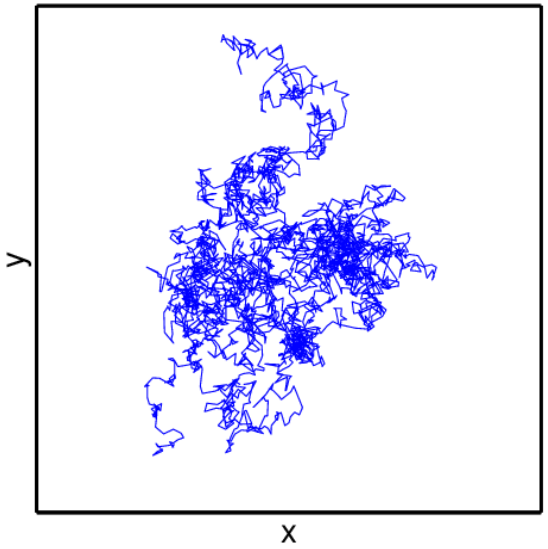


Physics

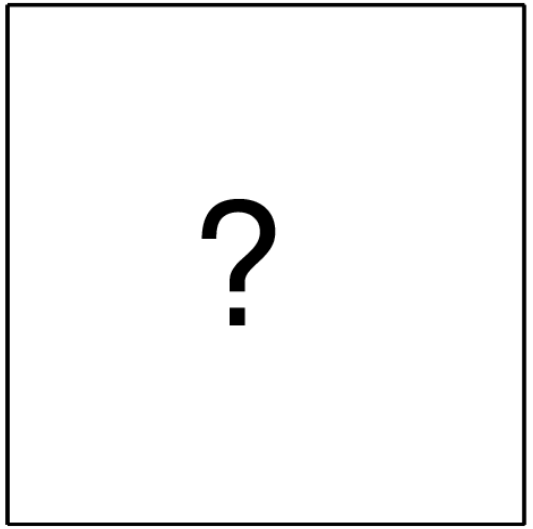


Biology

Where do the stochastic trajectories of biology come from?



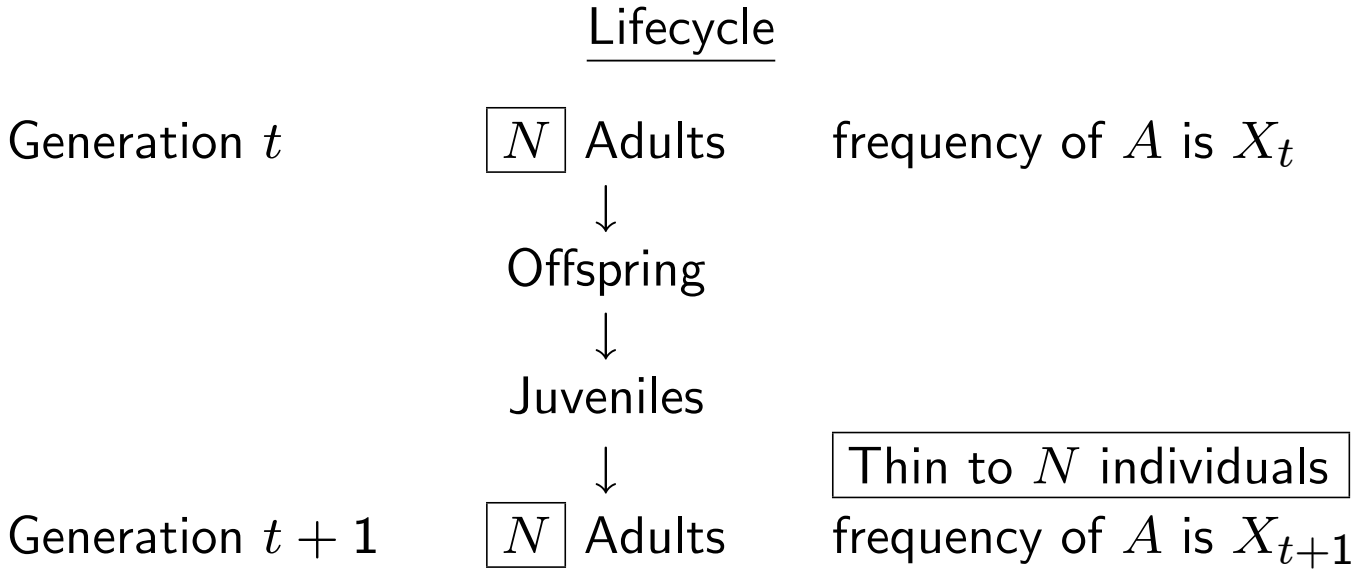
Physics



Biology

Where do the stochastic trajectories of biology come from?

Answer: **Random genetic drift, which occurs in a finite population ($N < \infty$)**



Consequence of drift

In a finite population ($N < \infty$) the deterministic equation

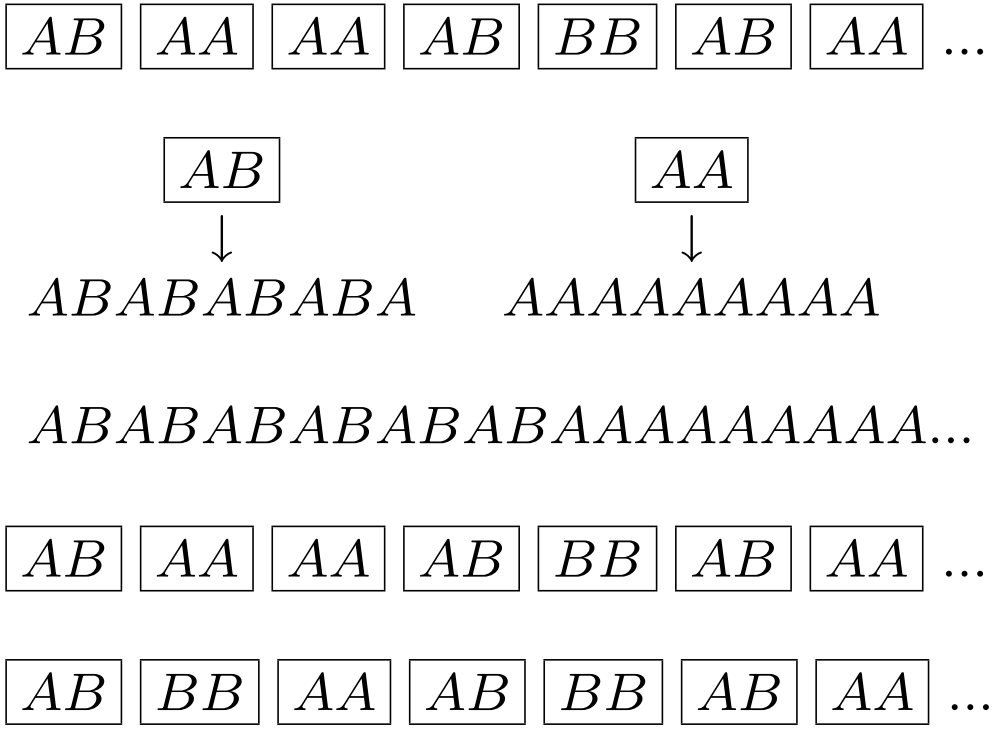
$$X_{t+1} = X_t + sX_t(1 - X_t)$$

does NOT apply - because of random genetic drift.

What is random genetic drift?

Random genetic drift - over 1 generation

N diploid sexual adults



Random genetic drift - over 1 generation

N adults

$\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Adults produce gametes

$\boxed{AB} \rightarrow ABABABABA \qquad \boxed{AA} \rightarrow AAAAAAAAA$

$ABABABABABABAAAAAAAAA\ldots$

$\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

$\boxed{AB}\ \boxed{BB}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Random genetic drift - over 1 generation

N diploid sexual adults

$\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Adults produce gametes

$\boxed{AB} \rightarrow ABABABABA \qquad \boxed{AA} \rightarrow AAAAAAAAA$

Adults die, gametes remain

$ABABABABABABAAAAAAAAA\ldots$

$\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

$\boxed{AB}\ \boxed{BB}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Random genetic drift - over 1 generation

N diploid sexual adults

$\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Adults produce gametes

$\boxed{AB} \rightarrow ABABABABA \qquad \boxed{AA} \rightarrow AAAAAAAAAA$

Adults die, gametes remain

$ABABABABABABAAAAAAAAAA\ldots$

Gametes pair up randomly

$\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

$\boxed{AB}\ \boxed{BB}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Random genetic drift - over 1 generation

N diploid sexual adults $\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Adults produce gametes

$$\boxed{AB} \downarrow ABABABABA \qquad \boxed{AA} \downarrow AAAAAAAAA$$

Adults die, gametes remain $ABABABABABABAAAAAAAAA\ldots$

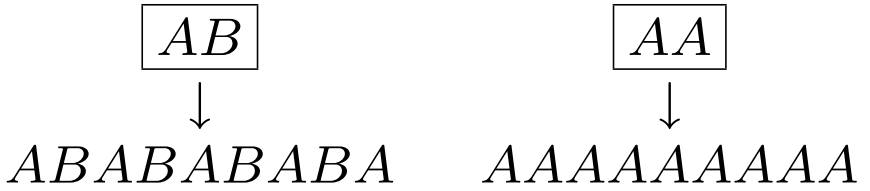
Gametes pair up randomly $\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

A <u>random</u> set of N survive $\boxed{AB}\ \boxed{BB}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Random genetic drift - over 1 generation

N diploid sexual adults $\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Adults produce gametes



Adults die, gametes remain $ABABABABABABAAAAAAAAAA\ldots$

Gametes pair up randomly $\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

N adults of next generation $\boxed{AB}\ \boxed{BB}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ \ldots$

Random genetic drift - causes trajectory to be stochastic

Generation t $\boxed{\mathbf{AB}}\ \boxed{\mathbf{AA}}\ \boxed{\mathbf{AA}}\ \boxed{\mathbf{AB}}\ \boxed{\mathbf{BB}}\ \boxed{\mathbf{AB}}\ \boxed{\mathbf{AA}}$ $X_t = \frac{9}{14}$

Adults produce gametes

$\boxed{AB}$ $\boxed{AA}$

$\downarrow$ $\downarrow$

$ABABABABA$ $AAAAAAAAA$

Adults die, gametes remain $ABABABABABABAAAAAAAAAA...$

Gametes pair up randomly $\boxed{AB}\ \boxed{AA}\ \boxed{AA}\ \boxed{AB}\ \boxed{BB}\ \boxed{AB}\ \boxed{AA}\ ...$

Generation $t+1$ $\boxed{\mathbf{AB}}\ \boxed{\mathbf{BB}}\ \boxed{\mathbf{AA}}\ \boxed{\mathbf{AB}}\ \boxed{\mathbf{BB}}\ \boxed{\mathbf{AB}}\ \boxed{\mathbf{AA}}$ $X_{t+1} = \frac{7}{14}$

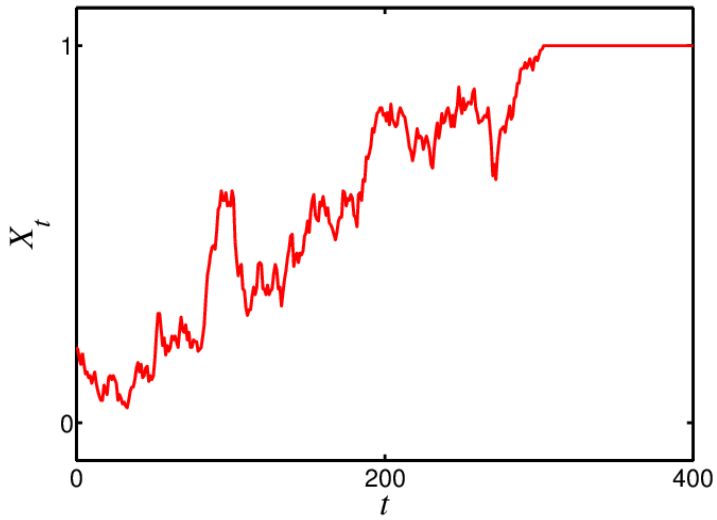
Random genetic drift - leads to stochastic trajectories

$$X_t = \frac{\text{Number of } A \text{ alleles}}{\text{Total Number of alleles}}$$

randomly takes values on

$$\left[\frac{0}{2N}, \frac{1}{2N}, \frac{2}{2N}, ..., \frac{2N}{2N}\right]$$

Drift dynamics

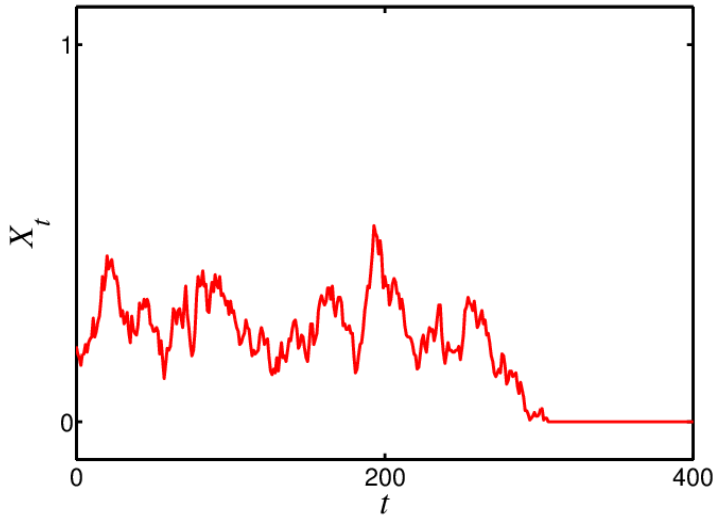


If $X_t = 1$, the N adults in the population are all

$\boxed{AA}$ $\boxed{AA}$ $\boxed{AA}$ $\boxed{AA}$ $\boxed{AA}$ $\boxed{AA}$ $\boxed{AA}$ $\boxed{AA}$ ...

and no further change occurs... FIXATION

Drift dynamics

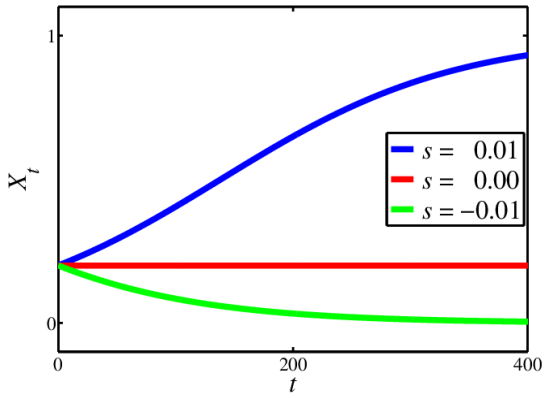


If $X_t = 0$, the N adults in the population are all

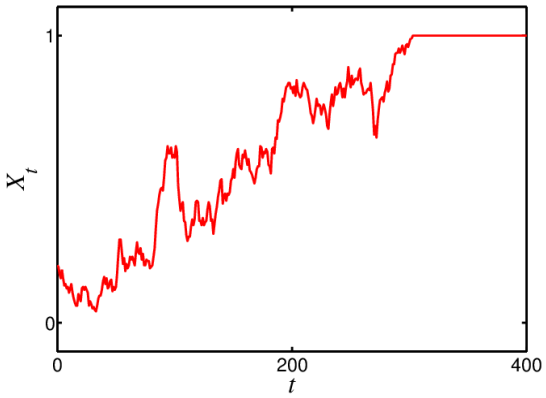
$\boxed{BB}$ $\boxed{BB}$ $\boxed{BB}$ $\boxed{BB}$ $\boxed{BB}$ $\boxed{BB}$ $\boxed{BB}$ $\boxed{BB}$ ...

and no further change occurs... LOSS

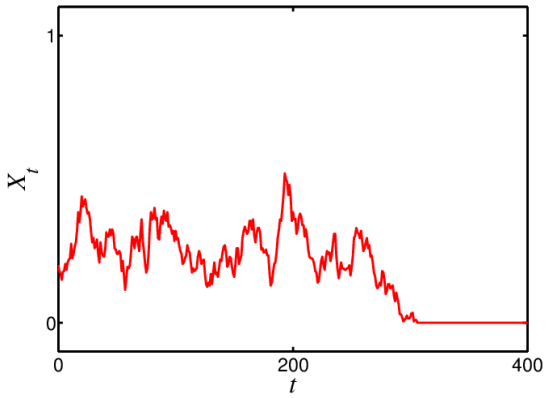
Comparison



Infinite population

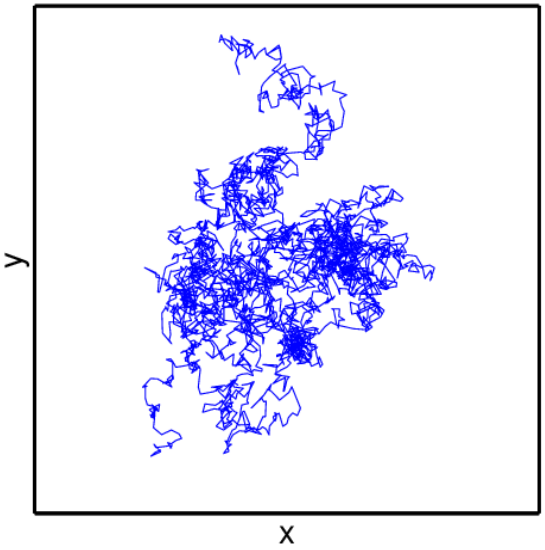


Finite population

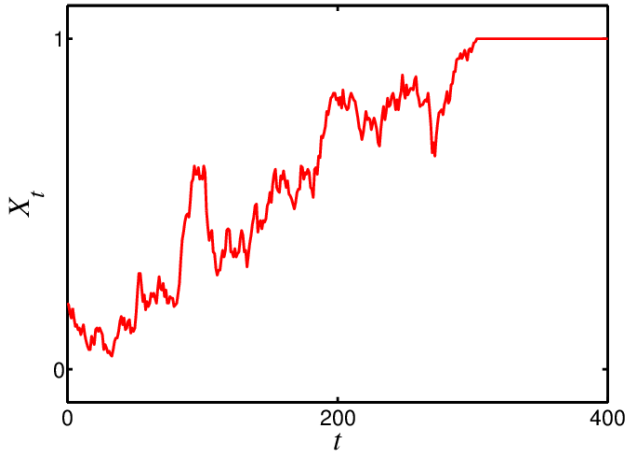


Finite population

Stochastic trajectories in physics and biology



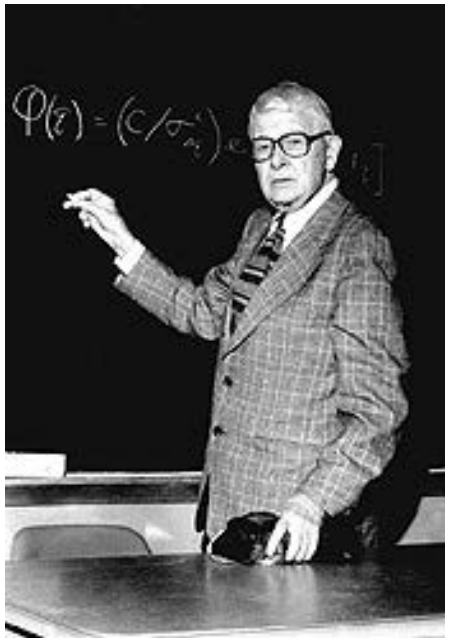
Physics



Biology (finite population)

Two scientists mathematically formulated the problem of genetic drift

R. A. Fisher



Sewall Wright

Statistical formulation of genetic drift

$\{N\}$ = a population with N individuals

$$= \underbrace{\boxed{\mathbf{AB}}\,\boxed{\mathbf{AA}}\,\boxed{\mathbf{AA}}\,\boxed{\mathbf{AB}}\,\boxed{\mathbf{BB}}\,\boxed{\mathbf{AB}}\,\boxed{\mathbf{AA}}\ldots}_{\text{N individuals}}$$

Statistical formulation of genetic drift of Wright and Fisher

$$\begin{array}{ccccc} \{N\} & \{N\} & \{N\} & \{N\} & \{N\} \\ \{N\} & \{N\} & \{N\} & \{N\} & \{N\} \\ \{N\} & \{N\} & \{N\} & \{N\} & \{N\} \end{array}$$

Imagine many **copies of a population**, each with N individuals

Wright and **Fisher** followed the fates of these copies of a population.

All initially **identical**

Yields statistical information

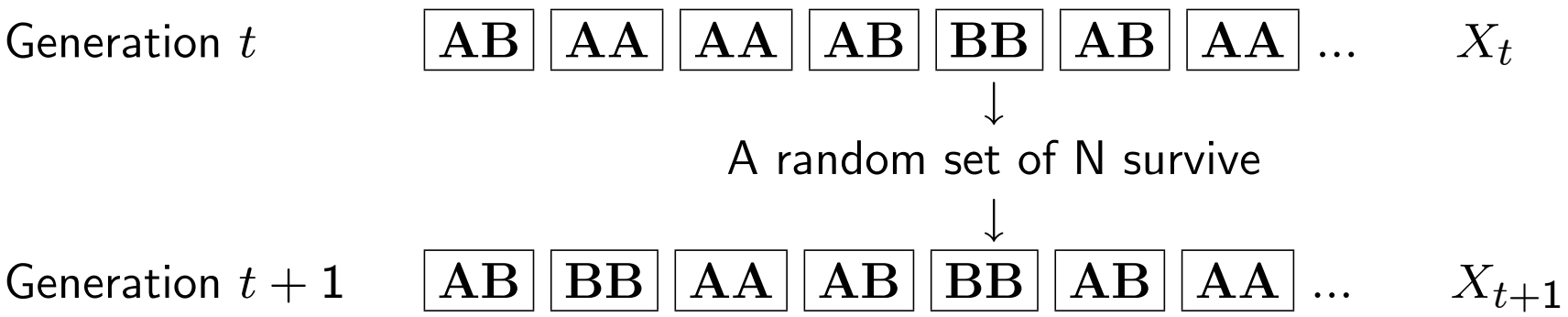
For example, with just selection, all populations eventually fix A or B

$\{N\}_A$ **fixed**	$\{N\}_B$ fixed	$\{N\}_B$ fixed	$\{N\}_B$ fixed	$\{N\}_B$ fixed
$\{N\}_B$ fixed	$\{N\}_A$ **fixed**	$\{N\}_A$ **fixed**	$\{N\}_B$ fixed	$\{N\}_A$ **fixed**
$\{N\}_B$ fixed	$\{N\}_B$ fixed	$\{N\}_B$ fixed	$\{N\}_B$ fixed	$\{N\}_A$ **fixed**

$\{N\}_{A \text{ fixed}}$ = a population with A allele fixed
estimate of fixation probability of $A = 5/15 = 1/3$

$\{N\}_{B \text{ fixed}}$ = a population with B allele fixed
estimate of fixation probability of $B = 10/15 = 2/3$

Wright Fisher model for genetic drift



Wright-Fisher model gives the rule

$$X_{t+1} = \frac{\text{Binom}(2N, X_t + sX_t(1 - X_t))}{2N}$$

$\text{Binom}(n, p)$ = binomial random number

= No. successes on n trials, each with probability p of success

Wright Fisher model for genetic drift

Define

$$f_{t,n} = \text{probability of } n \text{ copies of the } A \text{ allele in generation } t$$

$$= \text{probability that } X_t \text{ has the value } \frac{n}{2N}$$

$n = 0, 1, 2, ..., 2N.$

$f_{t,n}$ obeys

$$f_{t+1,n} = \sum_{m=0}^{2N} W_{n,m} f_{t,m} \qquad \text{Markov chain}$$

Wright Fisher model

Probability distribution obeys $f_{t+1,n} = \sum_{m=0}^{2N} W_{n,m} f_{t,m}$

$W_{n,m} =$ transition probabilities, contains information about probabilities of trajectories. A trajectory with

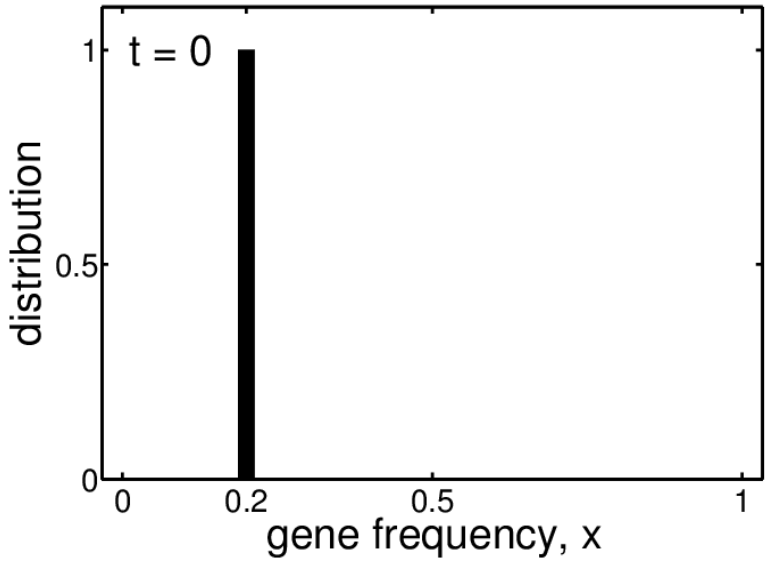
Number of A alleles	time
a	0
b	1
c	2
$\vdots$	$\vdots$

has

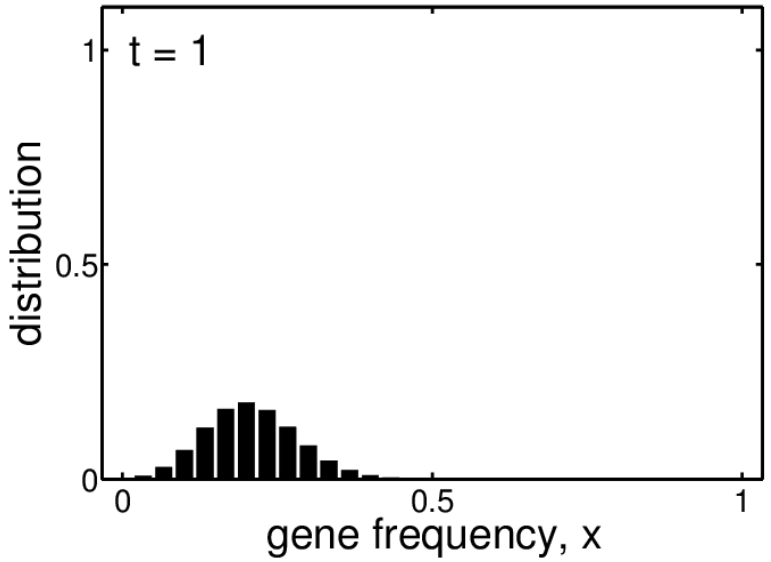
$$\text{probability } = \cdots W_{d,c} W_{c,b} W_{b,a}$$

Can look at where trajectories have reached, after time t

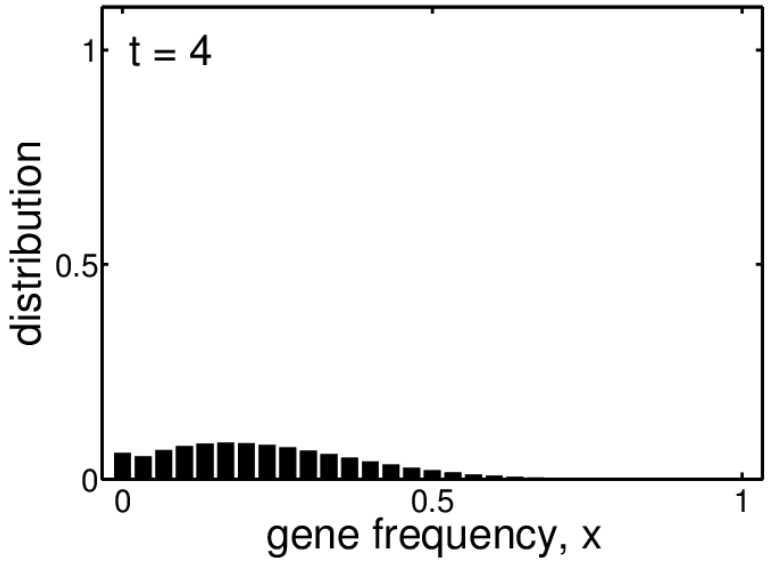
Distribution of position after time t



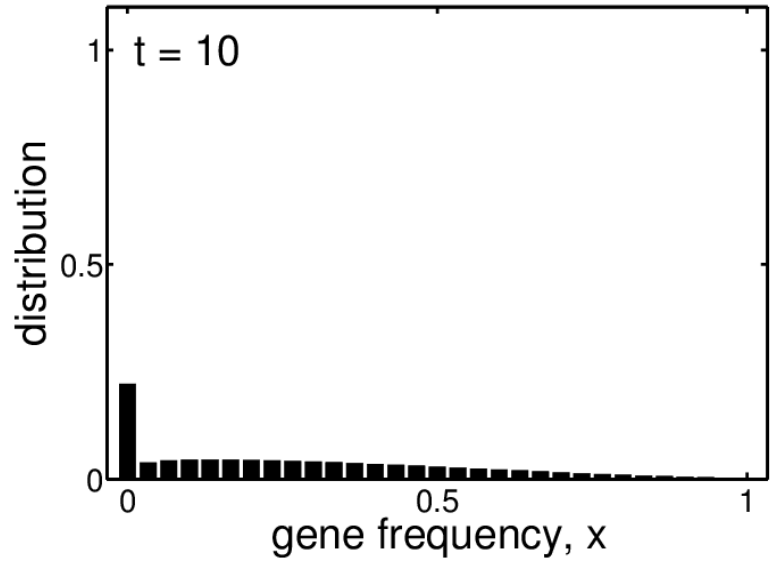
Distribution of position after time t



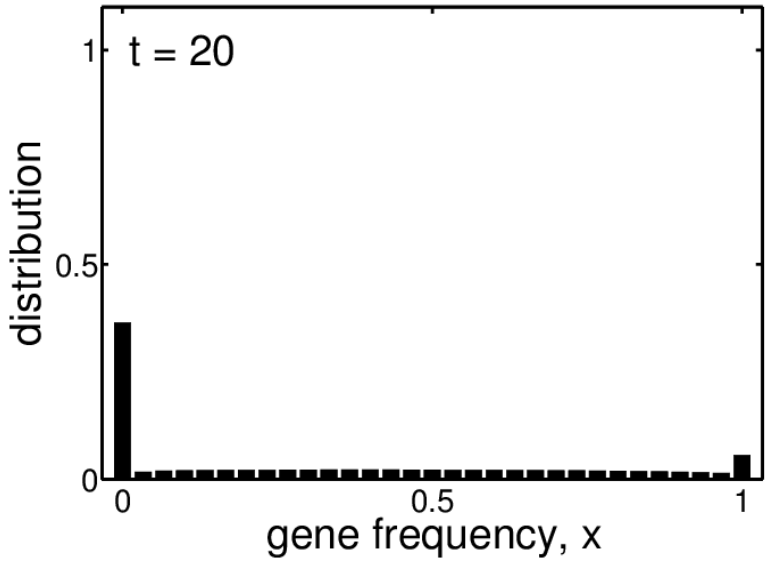
Distribution of position after time t



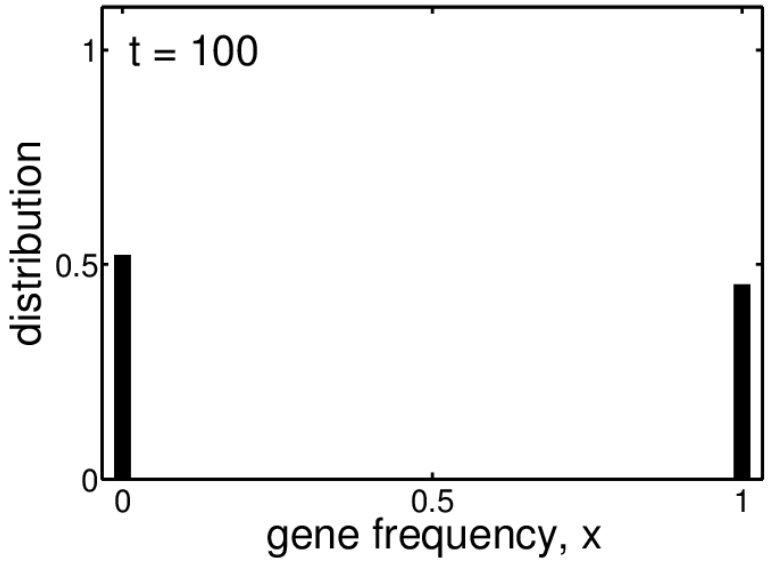
Distribution of position after time t



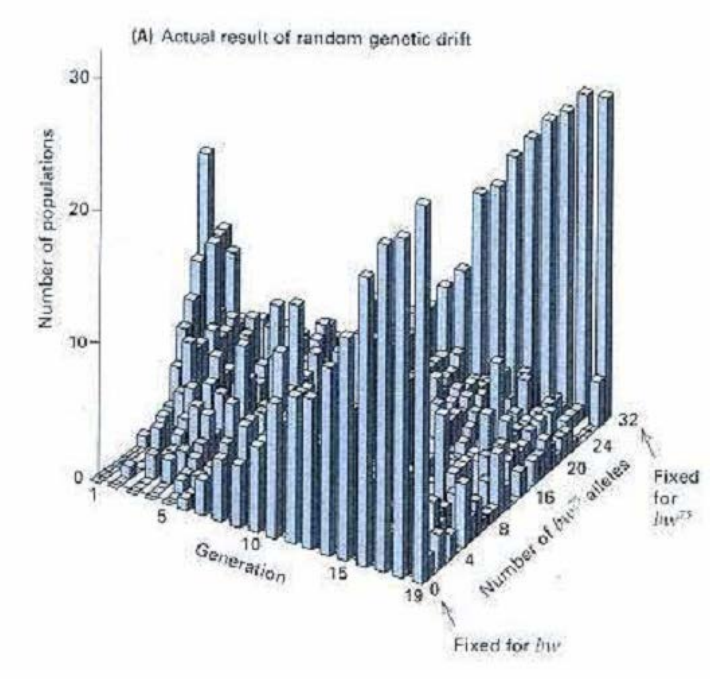
Distribution of position after time t



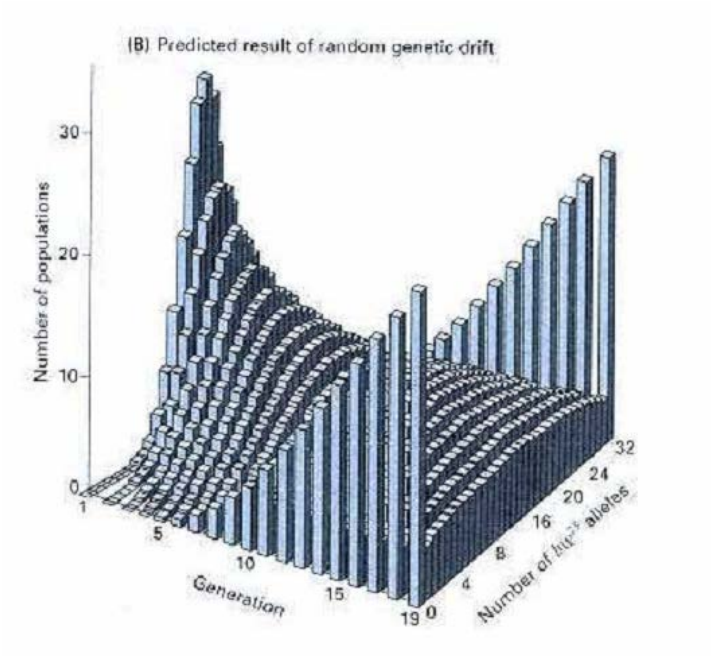
Distribution of position after time t



Experimental basis



From experiments on fly populations



From theory

The above behaviour suggests a sort of diffusion

To make analytical progress, use the

"diffusion approximation"

- replaces $f_{t+1,n} = \sum_{m=0}^{2N} W_{n,m} f_{t,m}$ by a diffusion equation which can sometimes be solved.

Diffusion approximation of $f_{t+1,n} = \sum_{m=0}^{2N} W_{n,m} f_{t,m}$

- Approximate time t as continuous
- Approximate frequency X_t as continuous: $X(t)$
- $f_{t,n} \to f(x,t) =$ probability density of $X(t)$
- $f_{t+1,n} = \sum_{m=0}^{2N} W_{n,m} f_{t,m} \to$ diffusion equation.

$$\frac{\partial}{\partial t} K(x,t|y,u) = \frac{1}{4N} \frac{\partial^2}{\partial x^2} \left[x(1-x) K(x,t|y,u) \right] - \frac{\partial}{\partial x} \left[sx(1-x) K(x,t|y,u) \right]$$

Diffusion approximation

$f_{t+1,n} = \sum_{m=0}^{2N} W_{n,m} f_{t,m}$ is approximated by

$$\frac{\partial}{\partial t} K(x,t|y,u) = \frac{1}{4N} \frac{\partial^2}{\partial x^2} \left[x(1-x) K(x,t|y,u) \right] - \frac{\partial}{\partial x} \left[sx(1-x) K(x,t|y,u) \right]$$

or equivalently

$X_{t+1} = \dfrac{\mathsf{Binom}(2N, X_t + sX_t(1-X_t))}{2N}$ is approximated by

$$dX_t = sX_t(1-X_t)dt + \sqrt{\frac{X_t(1-X_t)}{2N}} dB_t$$

Motoo Kimura - famous population geneticist

Extensively developed and applied the diffusion approximation

Conditioning

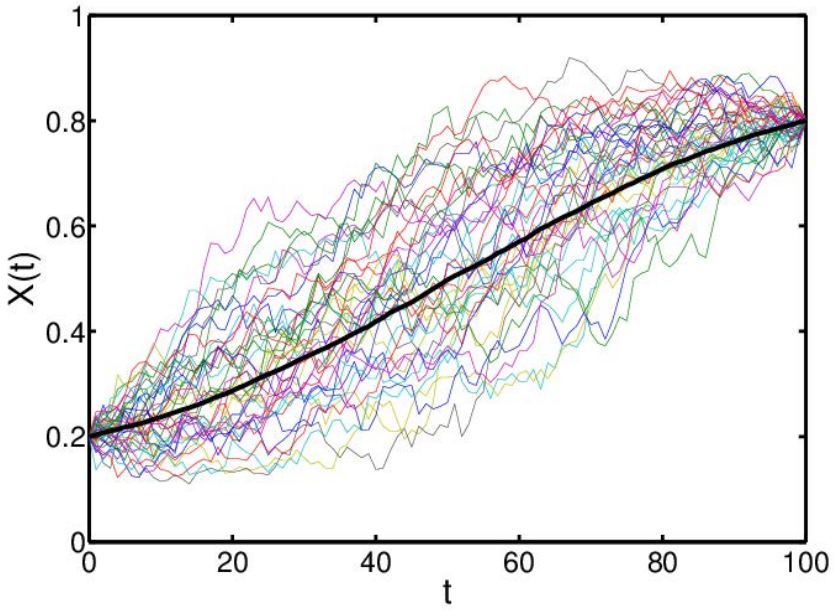
It is very natural to have observational data in the form of

an initial frequency, y, at an initial time, 0

an final frequency, z, at a final time, T

In this case we look only at trajectories going through these points.

For example from ancient DNA that is $\sim 100,000$ years old and DNA in modern humans.



Trajectories starting at frequency 0.2 at $t = 0$
and reaching frequency 0.8 at $t = 100$

What can we say about the conditioned problem?

Conditioning:

initial frequency is y at time $t = 0$

final frequency is z at time $t = T$

From a diffusion analysis: a conditioned problem has an identical description of an unconditioned one except selection gets modified:

selection strength $s \longrightarrow s_{\text{fict}}(x, t)$ (depends on y, z and T)

How does this come about?

The unconditioned problem, has fundamental solution (probability density of $X(t)$ at value x)

$$K(x,t|y,u) = E\left[\delta(x - X(t))|X(u) = y\right]$$

The conditioned problem has a fundamental solution, that achieves z at final time T, given by

$$K^{[z,T]}(x,t|y,u) = \frac{K(z,T|x,t)K(x,t|y,u)}{K(z,T|y,u)}$$

What can we say about the conditioned problem?

Unconditioned

$$\frac{\partial}{\partial t}K(x,t|y,u) = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left[x(1-x)K(x,t|y,u)\right] - \frac{\partial}{\partial x}\left[sx(1-x)K(x,t|y,u)\right]$$

Conditioned

$$\frac{\partial}{\partial t}K(x,t|y,u) = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}\left[x(1-x)K(x,t|y,u)\right] - \underbrace{\frac{\partial}{\partial x}\left[s_{\mathsf{fict}}(x,t)x(1-x)K(x,t|y,u)\right]}$$

What is the form of $s_{\text{fict}}(x,t)$?

For trajectories that fix by a specific time T

$$\begin{aligned} s_{\text{fict}}(x,t) &= s + \frac{1}{2N_e}\frac{\partial}{\partial x}P_{fix}(T|x,t) \\ &= s\coth\left(2N_e s x\right) \text{ for } T \to \infty \end{aligned}$$

where generally

$$P_{fix}(T|x,t) = \begin{array}{l}\text{probability of fixing by time } T\text{, given} \\ \text{an initial frequency of } x \text{ at time } t\end{array}$$

What is the form of $s_{\text{fict}}(x,t)$?

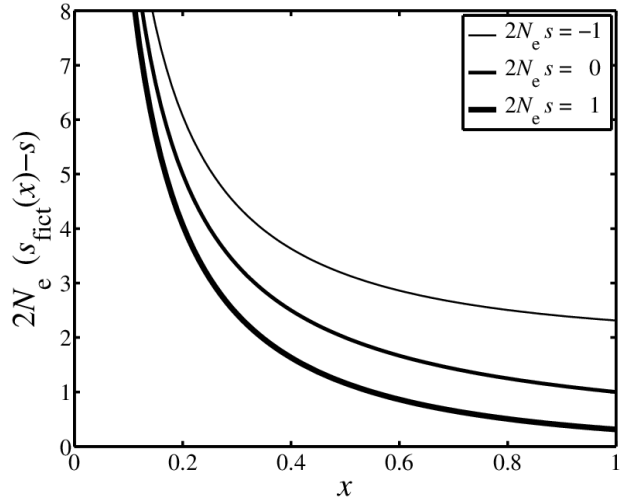
For trajectories that achieve frequency z by a specific time T

$$s_{\text{fict}}(x,t) = s + \frac{1}{2N_e}\frac{\partial}{\partial x}K(z,T|x,t)$$

where

$$K(z,T|x,t) = \begin{array}{l}\text{probability density of the frequency at time } T, \\ \text{frequency } z\text{, given a frequency of } x \text{ at time } t\end{array}$$

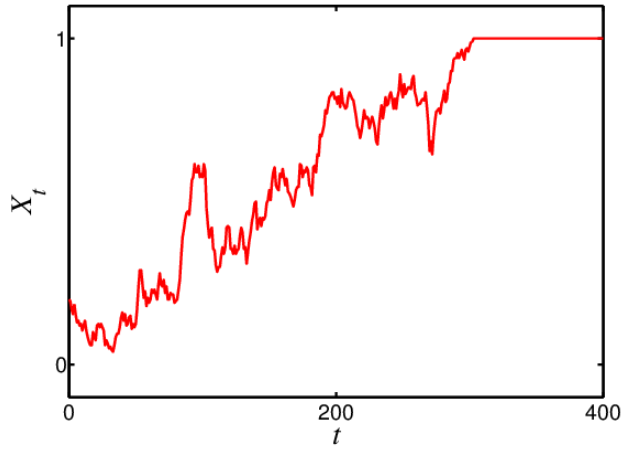
Conditioning leads to: selection strength $s \longrightarrow s_{\mathsf{fict}}(x,t)$



Much of the selection in a conditioned problem is fictitious: $|s| \ll s_{\mathsf{fict}}(x,t)$

Example

Suppose we are given a single trajectory, where the A allele fixes



Example

There are two alternatives

1. Assume this is a consequence of deterministic dynamics of a very large population, so $X_{t+1} = X_t + sX_t(1 - X_t)$

2. Assume this just is a chance fixation in a finite population - then it is a conditioned trajectory that arises from a problem with $s \longrightarrow s_{\text{fict}}(x, t)$.

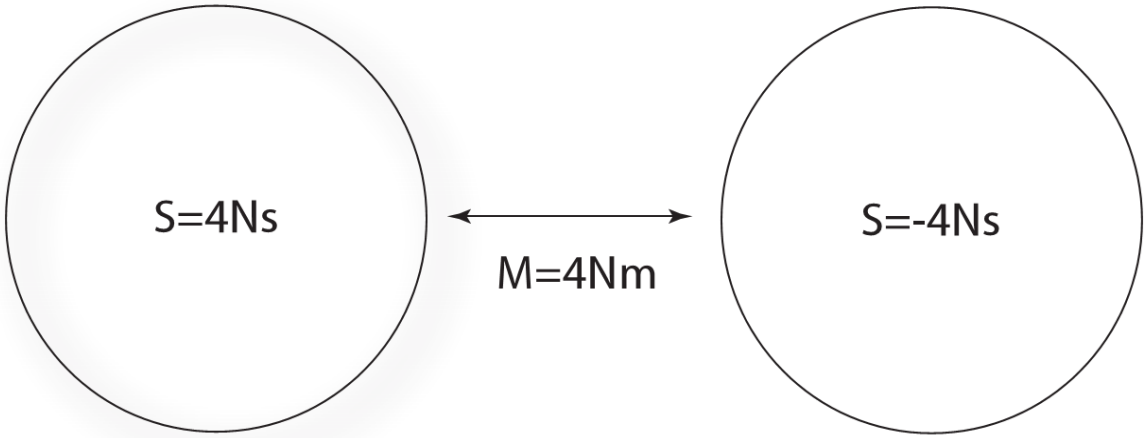
If (2) is correct, then (1) can drastically overestimate the true value of s (can be off by more than 1000% or even of the wrong sign).

- Strong evolutionary forces may be invoked in problems where effectively, conditioning has been carried out

- These strong forces may largely be an outcome of the conditioning

- They do not have a real existence and can strongly distort estimates of selection strength and other parameters

Other problems

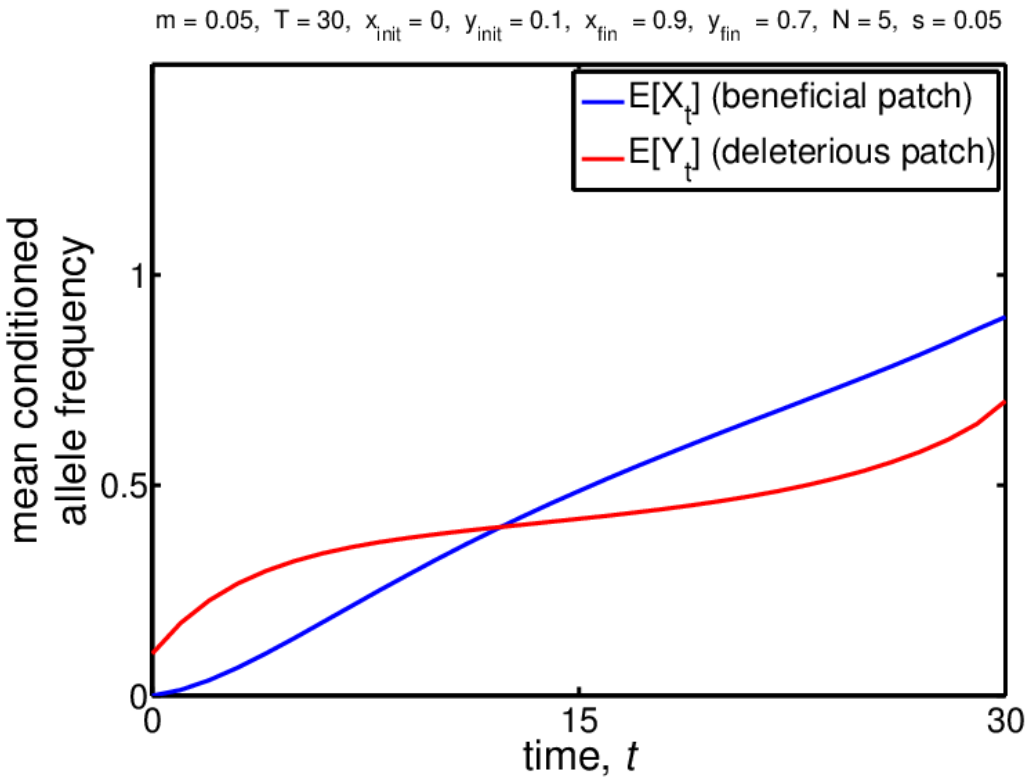


Problems of drift, involving spatial structure

$$X(t+1) = \frac{\mathsf{Bin}\left(2N, X(t) + D_1(X(t), Y(t))\right)}{2N}$$

$$Y(t+1) = \frac{\mathsf{Bin}\left(2N, Y(t) + D_2(X(t), Y(t))\right)}{2N}$$

Problems with spatial structure involve coupled diffusion equations...



Mean trajectories in a diffusion problem involving spatial structure

Summary

- Problems in genetics and evolution can usefully be looked at in terms of trajectories

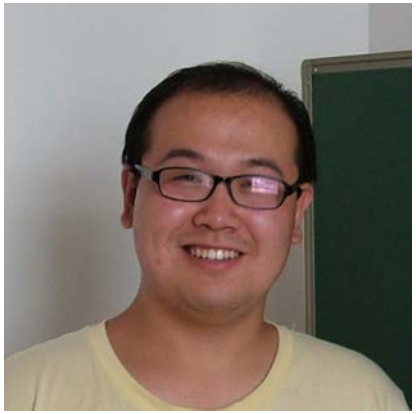
- Random genetic drift can lead to very different trajectories - compared with deterministic dynamics of an infinite population

- The act of conditioning - restricting trajectories - because of observations, can lead to fictitious forces in the problem which can greatly distort parameter estimates

Acknowledgements

Martin Lascoux

Andy Overall



Lei Zhao

Thank you for listening