Population Genetics and Evolution – II

The Mechanisms of Evolution: Mutation and Drift

Luca Peliti

Dipartimento di Fisica and Sezione INFN
Università di Napoli “Federico II”

Helsinki / June 2013

Special issue JSTAT: January 2013

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Evolutionary dynamics and statistical physics

Daniel Fisher[1], Michael Lässig[2] and Boris Shraiman[3]

[1] McCullough Laboratory, Department of Applied Physics, Stanford University, CA 94305, USA
[2] Institut für theoretische Physik, University of Cologne, D-50937 Cologne, Germany
[3] Kavli Institute for Theoretical Physics and Department of Physics, University of California, Santa Barbara, CA 91306, USA
E-mail: dsfisher@stanford.edu, mlaessig@uni-koeln.de and shraiman@kitp.ucsb.edu

Received 18 January 2013
Accepted 18 January 2013
Published 12 February 2013

Online at stacks.iop.org/JSTAT/2013/N01001
doi:10.1088/1742-5468/2013/01/N01001

Abstract. This introductory article provides the background to and motivation for this special issue and the relationship between evolutionary dynamics and statistical physics.

J. Stat. Mech. (2013) N

Outline

Mutations

Hardy-Weinberg equilibrium

- Sexual reproduction, diploid genome
- Notation: A, a variant *alleles* at one *locus* (ultimately, DNA subsequences)
- Genotypes: AA & aa *homozygotes*, Aa *heterozygote* (same as aA)
- Population of size N, with genotype frequency vector $(x_{\mathrm{AA}}, x_{\mathrm{Aa}}, x_{\mathrm{aa}})$
- Then $p = 2x_{\mathrm{AA}} + x_{\mathrm{Aa}}$ is the frequency of the A allele, and $q = 2x_{\mathrm{aa}} + x_{\mathrm{Aa}}$ that of the a allele

Hardy-Weinberg equilibrium

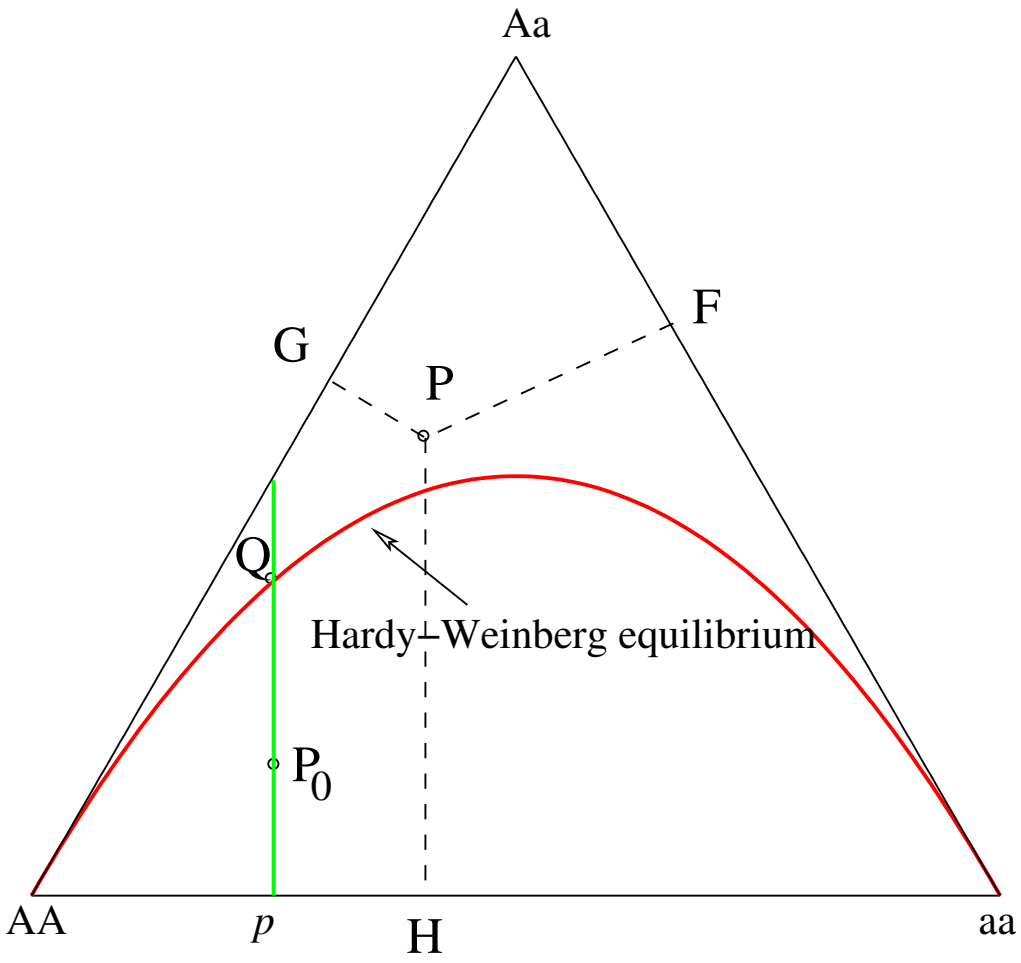
- **Hardy-Weinberg theorem:** Assume
  - Large population (fluctuations are neglected)
  - Neutral genotypes (fitness equal for everybody)
  - Mating is random (*panmictic* population)
- Then, *at the next generation*:

$$x_{\mathrm{AA}} = p^2 \qquad x_{\mathrm{Aa}} = 2pq \qquad x_{\mathrm{aa}} = q^2$$

- Allele frequencies determine the genotype frequencies!

Hardy-Weinberg equilibrium

De Finetti diagram:



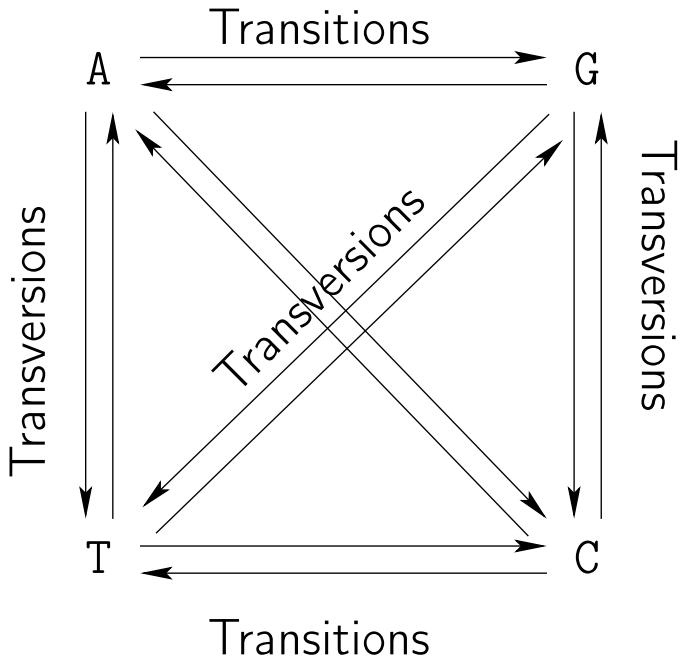
Nature of mutations

- Sequence mutations are changes in the offspring DNA wrt that of its parent(s)
- According to their *nature*, *small* (point) mutations are:

  Transitions: A ⇋ G or C ⇋ T

  Transversions: A ⇋ C, T or G ⇋ C, T

  Indels: Insertion or deletion of a short nucleotide sequence



Mutations in coding sequences

- In *coding sequences* each nucleotide triplet codes for a codon
- According to their *effects* mutations are:

  Synonymous or silent: The mutated codon corresponds to the same amino acid (weakest effect)

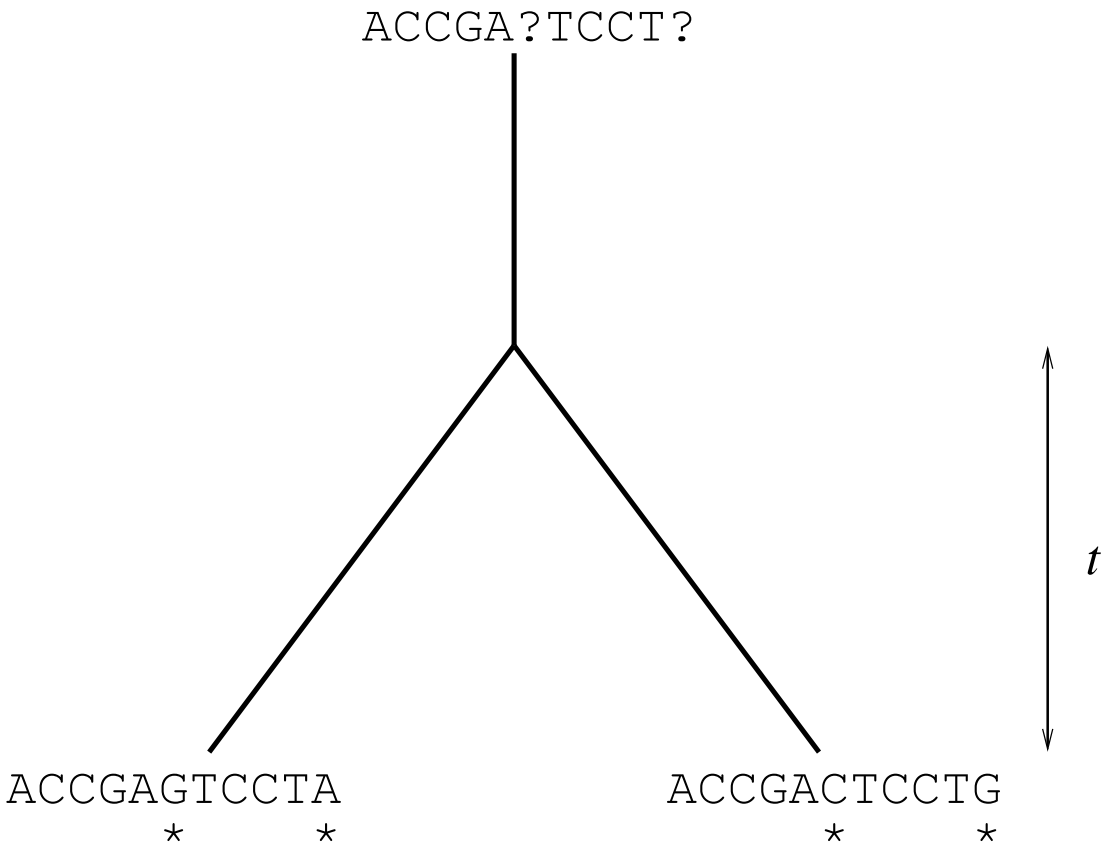
  Non-synonymous or missense: The mutated codon corresponds to a *different* amino acid (stronger effect)

  Nonsense: The replacement changes the codon into one of the stop ones (*much* stronger effect)
- Indels with a length which is *not* a multiple of 3 produce **reading frame shifts**: all codons after the indel are affected (strongest effect)

Mutation rates

- Mutations are a *stochastic process*, due both to the effect of the environment and of the organism's internal workings
- Mutation rates can be estimated by comparing *orthologous* sequences in two related life forms and counting changes
- One assumes a simple mutation model and estimates its parameters by making the comparison

Mutation rates



Jukes' rule: the time separation between the two sequence is $2t$
Assumes that backward evolution is the same as forward evolution (*reversibility*)

Mutation rates

- The comparison evaluates *substitution* rates, rather than *mutation* rates
- However, for *neutral* mutations the rates are equal (see later) (*Kimura*)
- The estimate is based on four general assumptions (all of them false!):
  1. The rates are uniform (do not depend on the position in the genome)
  2. They are constant in time
  3. They are the same for the two branches
  4. The equilibrium frequencies of the nucleotides are the same for the ancestral sequence and for the two “evolved” ones

Model for nucleotide substitution

- **Substitution matrix** $\mathrm{W} = (\mu_{ji})$: rate of substitution $j \longleftarrow i$, $i, j \in \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$
- Frequency of base i: $f_i(t)$
- Evolution equation for f_i:

$$\frac{\mathrm{d}f_i}{\mathrm{d}t} = \sum_{j\,(\neq i)}{}' \left[\mu_{ij} f_j - \mu_{ji} f_i\right]$$

- Equilibrium frequencies: f_i^{eq}: $\sum_{j\,(\neq i)} \left[\mu_{ij} f_j^{\mathrm{eq}} - \mu_{ji} f_i^{\mathrm{eq}}\right] = 0$
- **Evolution matrix** $\mathrm{P}(t) = (p_{ji}(t))$: conditional probability to find nucleotide j at time t, given that nucleotide i was in that position at $t = 0$
- Observed data: **Divergence matrix** $\mathrm{X}(t) = (x_{ji}(t))$: joint pdf to find nucleotide j in the first sequence and nucleotide i at the same position in the second sequence

Model for nucleotide substitution

- Equation for P(t):

$$\frac{\mathrm{d}p_{ij}}{\mathrm{d}t} = \sum_{k\,(\neq i)}{}' \left[\mu_{ik}p_{kj} - \mu_{ki}p_{kj}\right] \qquad p_{ij}(0) = \delta_{ij}$$

- Divergence matrix:

$$\mathrm{X}(t) = \mathrm{P}(t)\mathrm{X}(0)\mathrm{P}^{\mathrm{T}}(t) \qquad x_{ij}(0) = f_i^{\mathrm{eq}}\delta_{ij}$$

- Symmetry: $\mathrm{X}^{\mathrm{T}} = \mathrm{X}$, i.e., $x_{ji}(t) = x_{ij}(t)$ (not exactly satisfied due to sampling errors)
- Normalization constraint on the diagonal elements: $2x_{ii} = 2f_i - \sum_{i\,(\neq j)}{}' x_{ij} - \sum_{j\,(\neq i)}{}' x_{ji}$
- Thus W (16 entries) has only 6 independent parameters

Model for nucleotide substitution

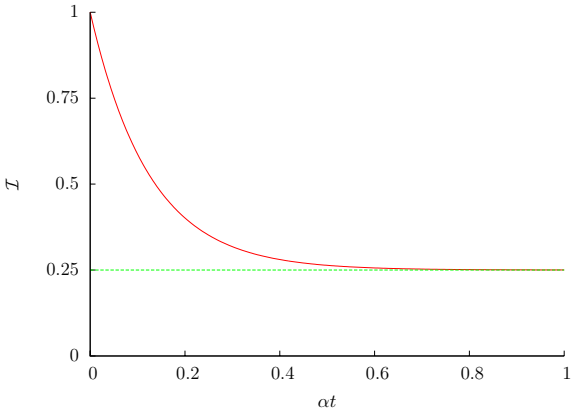
Jukes-Cantor model

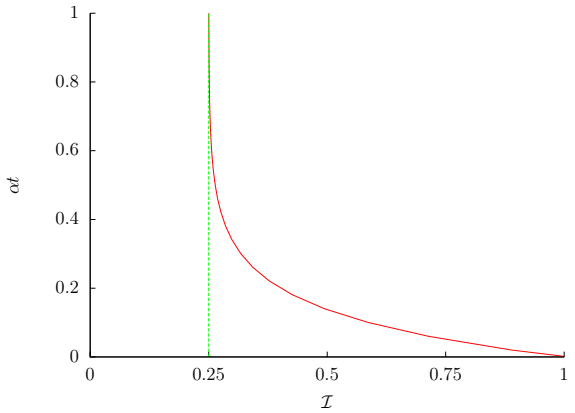
All substitutions are equally probable: $\mu_{ij} = \alpha, \forall (i \neq j)$

- $f_i^{\mathrm{eq}} = \frac{1}{4}, \forall i; p_{ij}(t) = \frac{1}{4}\left(1 - \mathrm{e}^{-4\alpha t} + 4\delta_{ij}\mathrm{e}^{-4\alpha t}\right)$
- Probability of observing the *same* nucleotide in the two sequences:

$$\mathcal{I}(t) = \frac{1}{4}\left(1 + 3\mathrm{e}^{-8\alpha t}\right)$$

- Thus $\alpha t = -\frac{1}{8}\ln\left(\frac{4\mathcal{I}-1}{3}\right)$

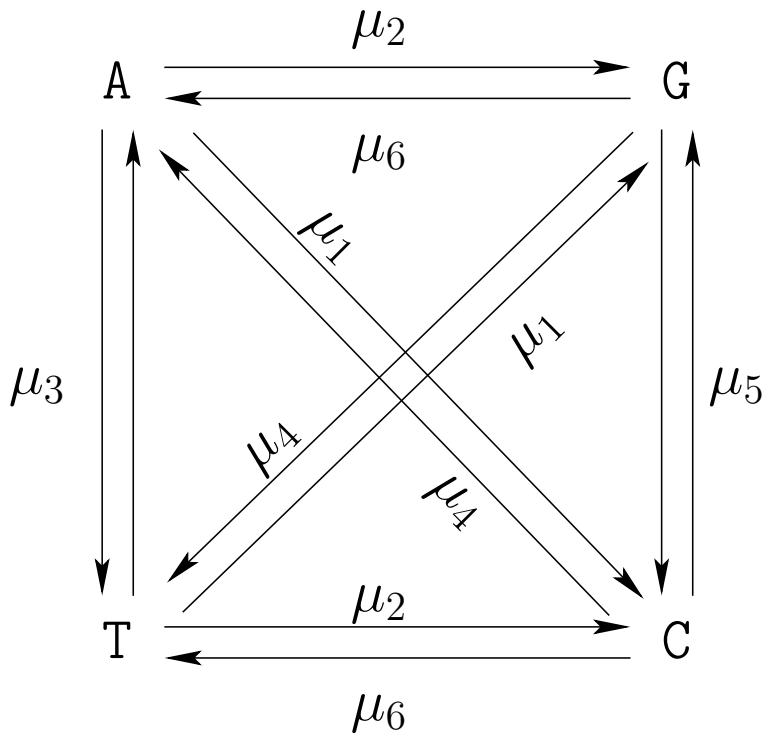




Model for nucleotide substitution

General 6-parameter model

- A substitution $\mathtt{A} \longleftarrow \mathtt{C}$ implies the corresponding substitution $\mathtt{T} \longleftarrow \mathtt{G}$ in the opposite strand
- Thus $w_{\mathtt{AC}} = w_{\mathtt{TG}}$, ecc.
- Thus we have only 6 independent rates from *stable sequences*:



Reversibility vs. detailed balance

O. ZAGORDI AND J.-L. LOBRY, 2005

- Detailed balance: $\mu_{ij} f_j^{\mathrm{ex}} = \mu_{ji} f_i^{\mathrm{ex}}, \forall i \neq j$
- Reversibility: $\mathrm{P}(-t) = \mathrm{P}(t)$ (needed by Jukes' rule)
- Theorem: Reversibility $\Leftrightarrow$ Detailed balance
- Problem: A model which fits the data is reversible?
- Answer: Chargaff rule: $f_{\mathrm{A}} = f_{\mathrm{T}}$, $f_{\mathrm{G}} = f_{\mathrm{C}}$ (*no strand bias*)
- There are only *five* independent observable quantities in X!
- One can impose an additional constraint on the model, e.g., $\mu_1 \mu_6 = \mu_2 \mu_4$ (reversibility)

Infinite allele and infinite site model

- We often want to model mutations starting from a given wild type
- **Infinite allele model:** *Each mutation produces a wholly new genotype*
- No structure in the mutants: all mutants are as different from the wild type as from each other
- **Infinite site model:** *Each mutation hits a different site*
- Mutants can be binned in k-classes: Classes with k mutations wrt wild type

Mutations and selection

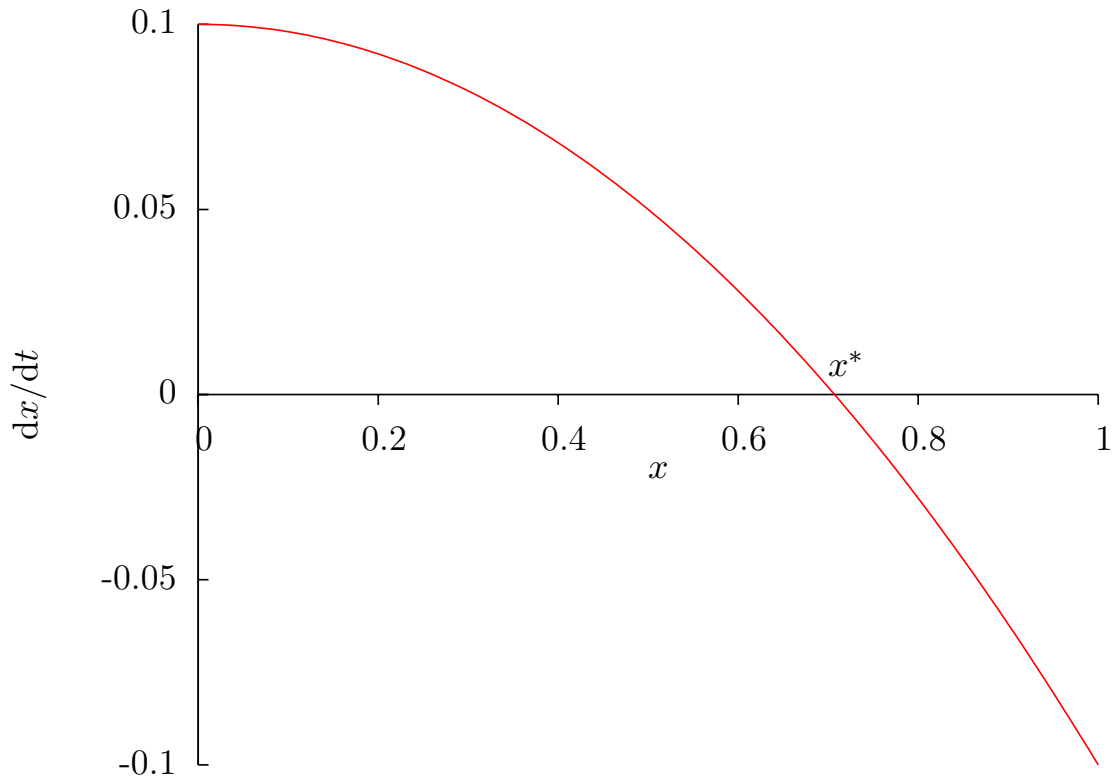
A simple model

- Population with two types: A and B
- Selection coefficient $s = f_{\mathrm{A}} - f_{\mathrm{B}}$
- **Mutation:** $\mathrm{A} \underset{\rightarrow}{\overset{\mu}{\leftarrow}} \mathrm{B}$

Evolution equation:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = sx(1-x) + \mu(1-x) - \mu x = sx(1-x) + \mu(1-2x)$$

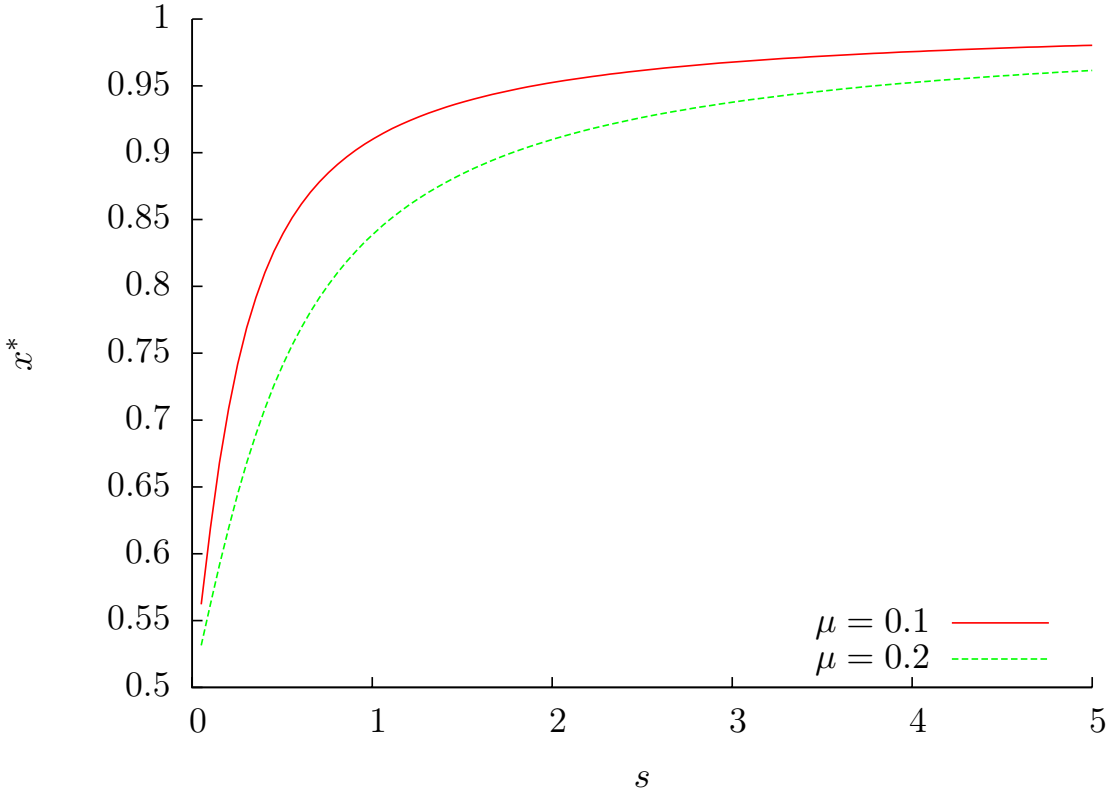
A simple model



A simple model

Fixed point x^*:

$$x^* = \frac{s - 2\mu + \sqrt{s^2 + 4\mu^2}}{2s}$$



Optimization?

- $\langle f \rangle_x = f_A x + f_B(1-x)$ is not maximal at x^*
- But define

$$\Phi(x) = \underbrace{\langle f \rangle_x}_{-\text{“energy”}} + \mu \underbrace{\log\left[x(1-x)\right]}_{\text{“entropy”}}$$

- Then

$$\begin{aligned} \frac{\mathrm{d}\Phi}{\mathrm{d}t} &= s\frac{\mathrm{d}x}{\mathrm{d}t} + \mu\frac{1-2x}{x(1-x)}\frac{\mathrm{d}x}{\mathrm{d}t} \\ &= x(1-x)\left[s + \mu\frac{1-2x}{x(1-x)}\right]^2 \geq 0 \end{aligned}$$

- Φ increases and reaches its maximum at the fixed point

Multiple alleles

- r alleles: $\alpha \overset{\mu}{\leftrightharpoons} \beta \qquad \alpha, \beta = 1, \ldots, r \qquad \mu(\alpha \longrightarrow \beta) = \mu_\beta$
- Set $x_r = 1 - \sum_{j=1}^{r-1} x_j$
- Define:

$$s_j = f_j - f_r = \frac{\partial \langle f \rangle_{\boldsymbol{x}}}{\partial x_j}, \qquad j = 1, \ldots, r-1$$

$$\Gamma_{jk}(\boldsymbol{x}) = \begin{cases} -x_j x_k, & \text{if } j \neq k \\ x_j(1 - x_j), & \text{if } j = k \end{cases} \qquad \Gamma \text{ positive definite}$$

- Evolution equation for $\boldsymbol{x} = (x_1, \ldots, x_{r-1})$:

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \sum_{k=1}^{r-1} \Gamma_{jk}(\boldsymbol{x}) s_k + \mu_j(1 - x_j) - x_j \sum_{\alpha(\neq j)}{}' \mu_k$$

Optimization II

- Define

$$M(\boldsymbol{x}) = \sum_{\alpha} \mu_\alpha \log x_\alpha$$

- Then

$$\sum_k \Gamma_{jk}(\boldsymbol{x}) \frac{\partial M}{\partial x_k} = \mu_j(1 - x_j) - x_j \sum_{\alpha(\neq j)}{}' \mu_\alpha = \mu_j - x_j \sum_\alpha \mu_\alpha$$

and

$$\begin{aligned} \frac{\mathrm{d}x_j}{\mathrm{d}t} &= \sum_k \Gamma_{jk}(\boldsymbol{x}) \frac{\partial}{\partial x_k} \left[\langle f \rangle_{\boldsymbol{x}} + M(\boldsymbol{x})\right] = \sum_k \Gamma_{jk}(\boldsymbol{x}) \frac{\partial \Phi}{\partial x_k} \\ \Phi(\boldsymbol{x}) &= \langle f \rangle_{\boldsymbol{x}} + M(\boldsymbol{x}) \end{aligned}$$

- Thus

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = \sum_{j,k} \frac{\partial \Phi}{\partial x_j} \Gamma_{jk}(\boldsymbol{x}) \frac{\partial \Phi}{\partial x_k} \geq 0$$

The quasispecies (QS) model

M. EIGEN, 1971

- Nonoverlapping generations; large number of alleles
- Mutation rate $k \overset{Q_{k\ell}}{\leftrightharpoons} \ell$ depending on "distance" of alleles
- Evolution equation for $\boldsymbol{x} = (x_1, \ldots, x_r)$:

$$x_j(t+1) = \frac{1}{\langle W \rangle_{\boldsymbol{x}}} \sum_{k=1}^{r} Q_{jk} W_k x_k(t)$$

where $\langle W \rangle_{\boldsymbol{x}} = \sum_j W_j x_j$

Asymptotic behavior of the QS model

- Define the unnormalized population vector $\boldsymbol{y}(t)$:

$$\begin{aligned} \boldsymbol{y}(0) &= \boldsymbol{x}(0) \\ y_j(t+1) &= \sum_{k=1}^{r} Q_{jk} W_k y_k(t) = \sum_{k=1}^{r} T_{jk} y_k(t) \end{aligned}$$

- Decompose $\boldsymbol{y}$ according to the right eigenvectors of $\mathrm{T} = (Q_{jk} W_k)$:

$$\begin{aligned} \boldsymbol{y} &= \sum_{\kappa} c_\kappa \boldsymbol{\xi}^{(\kappa)} \\ \mathrm{T} \cdot \boldsymbol{\xi}^{(\kappa)} &= \lambda^{(\kappa)} \boldsymbol{\xi}^{(\kappa)} \end{aligned}$$

- **Perron-Frobenius theorem:** the largest eigenvalue $\lambda^{(0)}$ is positive and has a unique right eigenvector $\boldsymbol{\xi}^{(0)}$, $\xi_i^{(0)} > 0$, $\forall i$
- Thus, for $n \gg 1$

$$\mathrm{T}^n \cdot \boldsymbol{y} = \sum_{\kappa} \left(\lambda^{(\kappa)}\right)^n c_\kappa \boldsymbol{\xi}^{(\kappa)} \simeq \left(\lambda^{(0)}\right)^n c_0 \boldsymbol{\xi}^{(0)}$$

The composition vector $\boldsymbol{x}$

Since

$$\boldsymbol{x}(t) = \frac{\boldsymbol{y}(t)}{\sum_j y_j(t)}$$

we have

$$\lim_{t\to\infty} \boldsymbol{x}(t) = \boldsymbol{\xi}^{(0)}$$

independently of the initial condition

The error threshold

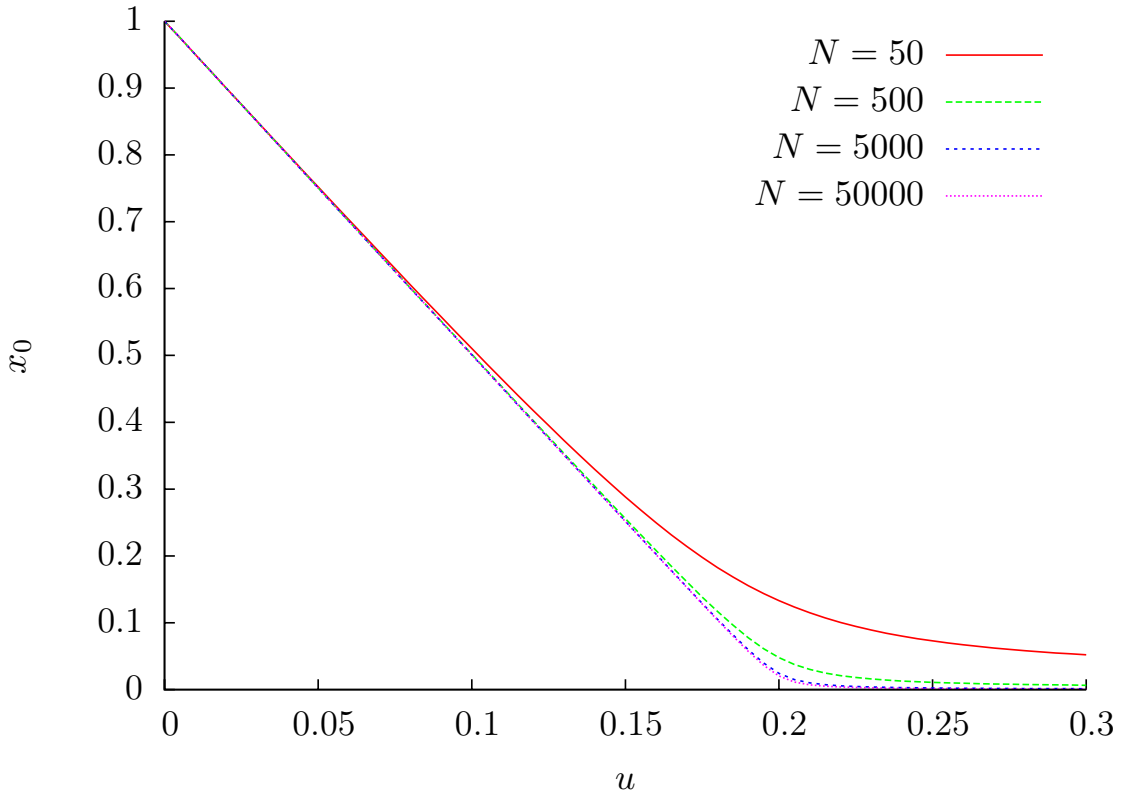
- One optimal genotype 0: $W_0 > W_k = W, \forall k \neq 0, r \gg 1$
- Mutation probability $k \longrightarrow 0$: $u/r \simeq 0$
- Define $W/W_0 = 1 - s$
- Then

$$x_0(t+1) = \frac{W_0(1-u)x_0}{W_0 x + W(1-x_0)} = \frac{(1-u)x_0}{1-s+sx_0}$$

The error threshold

Fixed point:

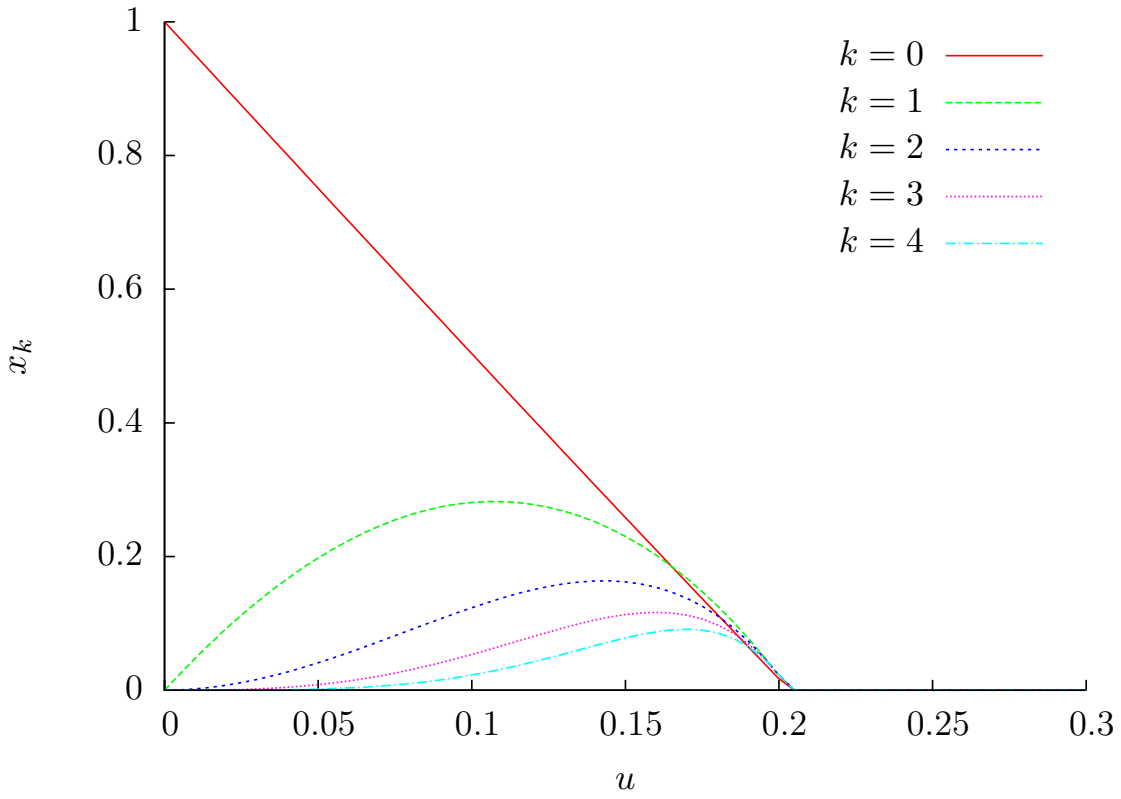
$$x_0^* = 1 - \frac{u}{s}$$



Selection factor $s = 0.2$

Error classes

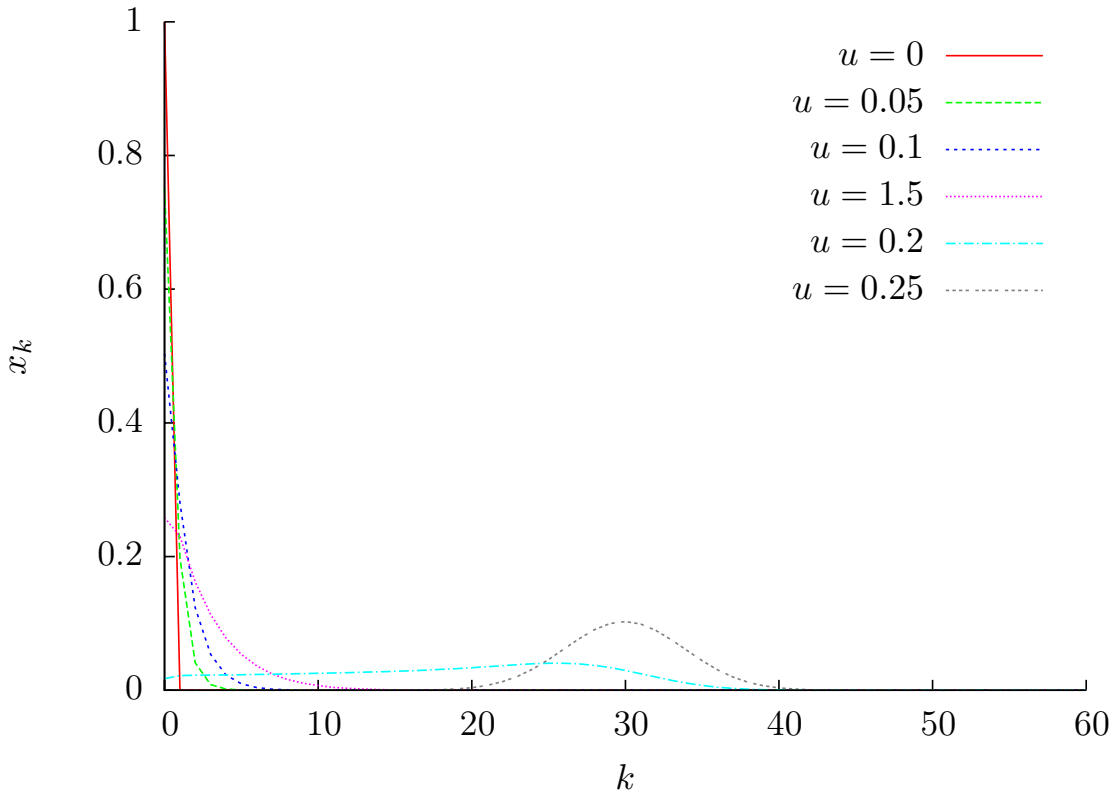
x_k: fraction of individuals with k "errors" with respect to selected type



$N = 60$ loci, two alleles, selection factor $s = 0.2$

Error classes

x_k: fraction of individuals with k "errors" with respect to selected type



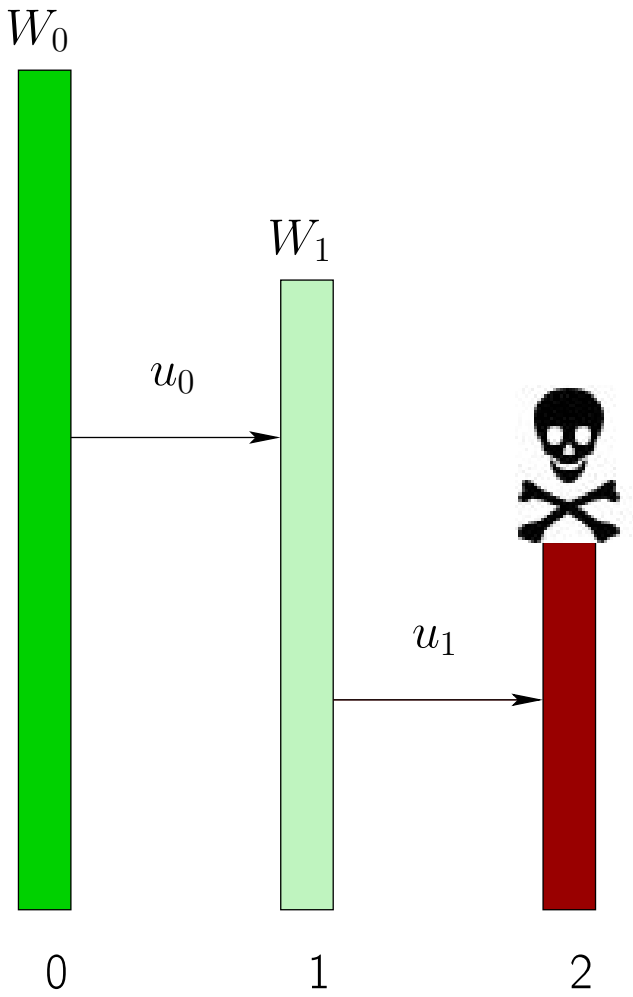
$N = 60$ loci, two alleles, selection factor $s = 0.2$

Error threshold vs. extinction

J. BULL ET AL., 2005; C. O. WILKE, 2005

- Simple model with three genotype classes:
  - Class 0: Fitness $W_0 > 1$, mutation probability u_0 to Class 1
  - Class 1: Fitness $W_1 < W_0$, mutation probability $u_1 < u_0$ to Class 2
  - Class 2: Fitness $W_2 = 0$ (does *not* reproduce)

Error threshold vs. extinction



Error threshold vs. extinction

Evolution equation for the population vector $\boldsymbol{n} = (n_0, n_1, n_2)$:

$$\begin{aligned} \boldsymbol{n}(t+1) &= \mathsf{T}\boldsymbol{n}(t) \\ \mathsf{T} &= \begin{pmatrix} (1-u_0)W_0 & 0 & 0 \\ u_0 W_0 & (1-u_1)W_1 & 0 \\ 0 & u_1 W_1 & 0 \end{pmatrix} \end{aligned}$$

The total population is given by $N(t) = \sum_j n_j$

Error threshold vs. extinction

Eigenvalues and eigenvectors:

$$
\begin{aligned}
\lambda^{(0)} &= W_0(1-u_0) \\
\boldsymbol{n}^{(0)} &= \left(\frac{(1-u_0)(W_0(1-u_0) - W_1(1-u_1))}{W_0 u_0 u_1}, \frac{(1-u_0)W_0}{W_1 u_1}, 1 \right) \\
\lambda^{(1)} &= W_1(1-u_1) \\
\boldsymbol{n}^{(1)} &= \left(0, \frac{1-u_1}{u_1}, 1 \right)
\end{aligned}
$$

$N(t) \sim (\lambda^{\max})^t$: extinction if $\lambda^{\max} < 1$

Error threshold vs. extinction

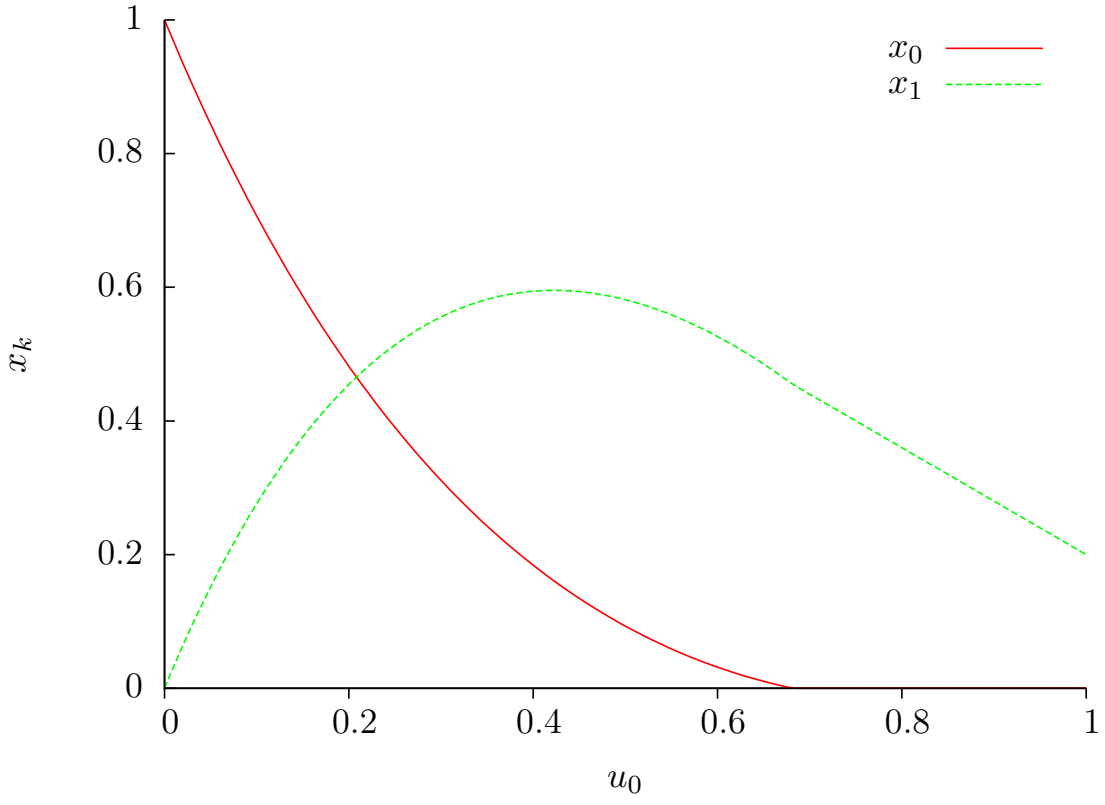
Error threshold:

$$(1 - u_0)W_0 = (1 - u_1)W_1$$

Extinction threshold:

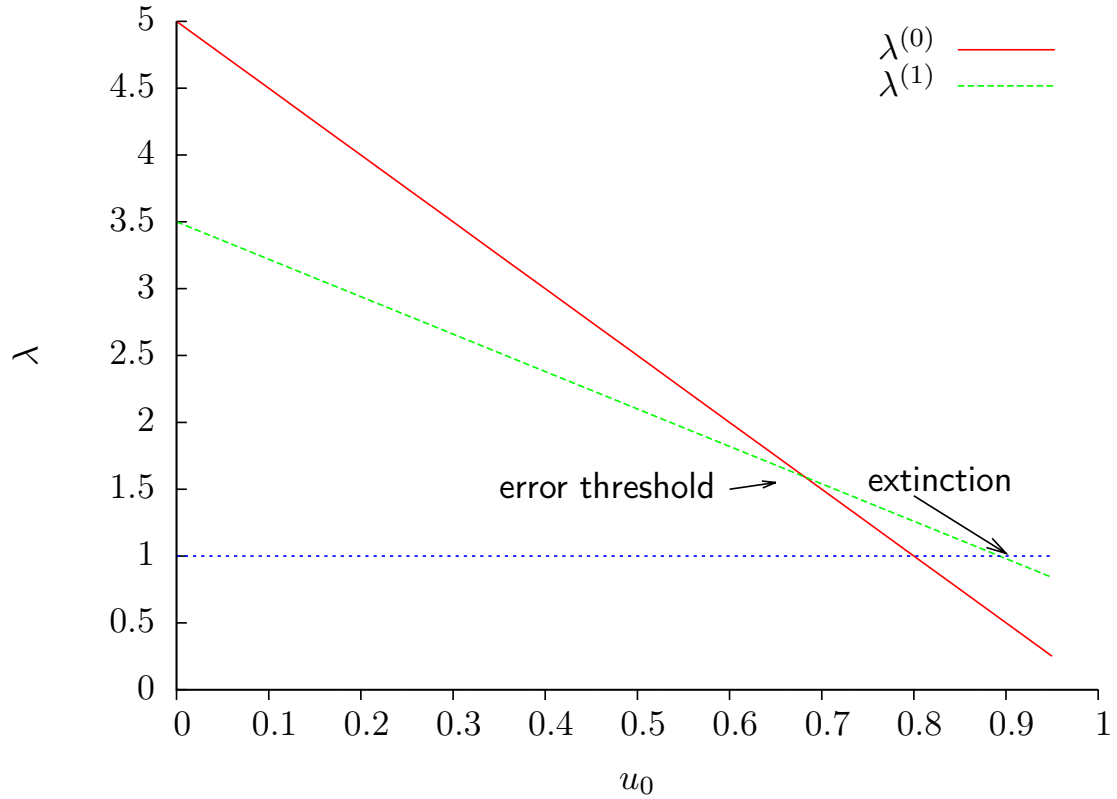
$$\lambda^{\max}(W_0, W_1, u_0, u_1) = 1$$

The transitions



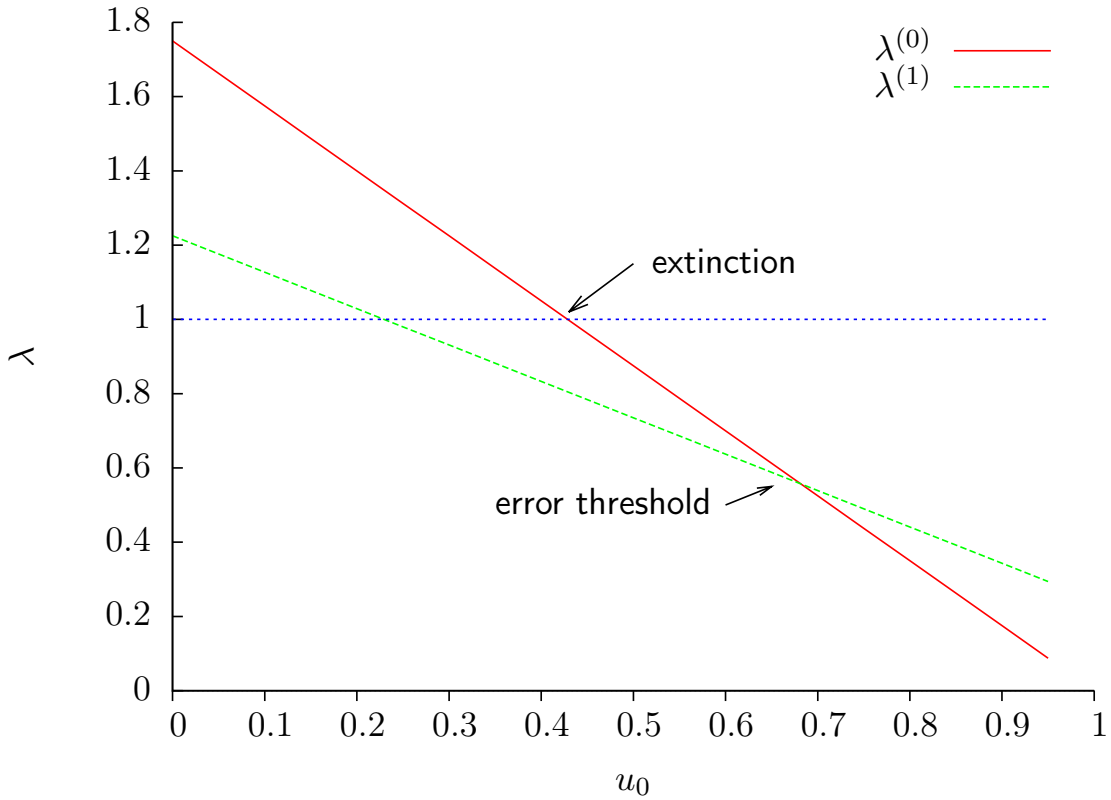
$W_1 = 0.7W_0$, $u_1 = 0.8u_0$: The error threshold is independent of W_0

The transitions



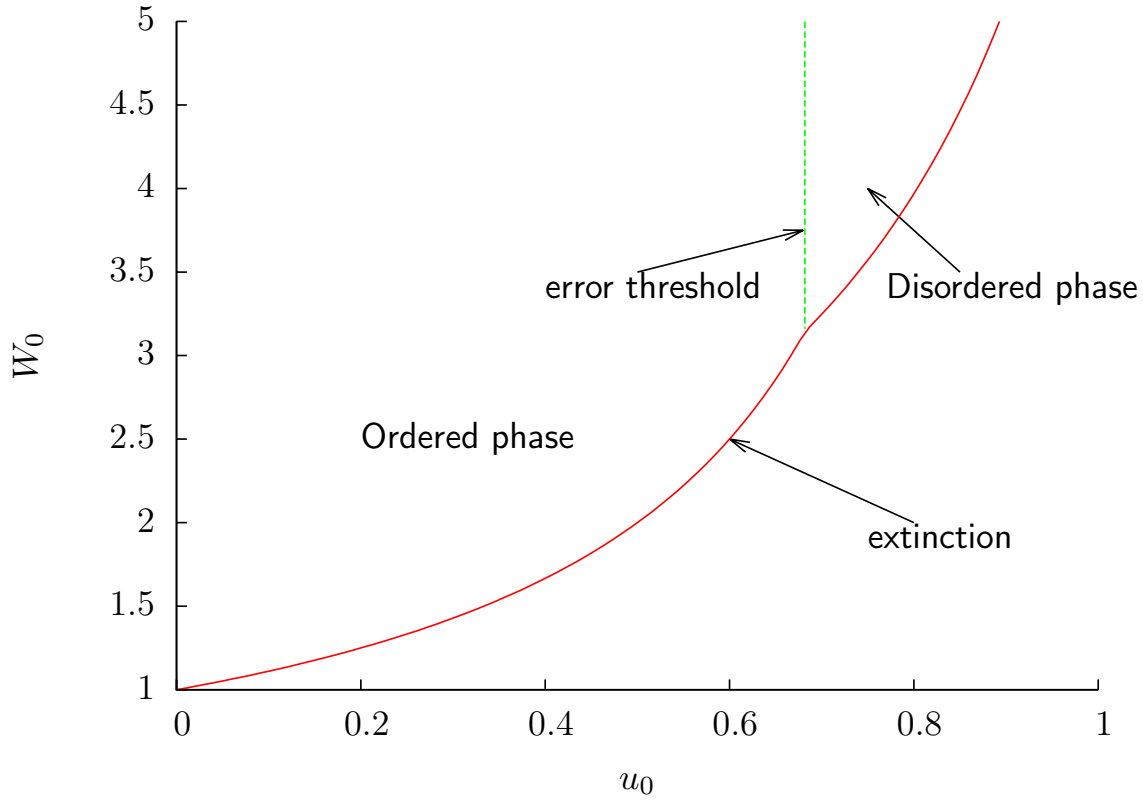
$W_0 = 5.0$: The error catastrophe *delays* the extinction

The transitions



$W_0 = 1.75$: Extinction *prevents* the error catastrophe

The transitions



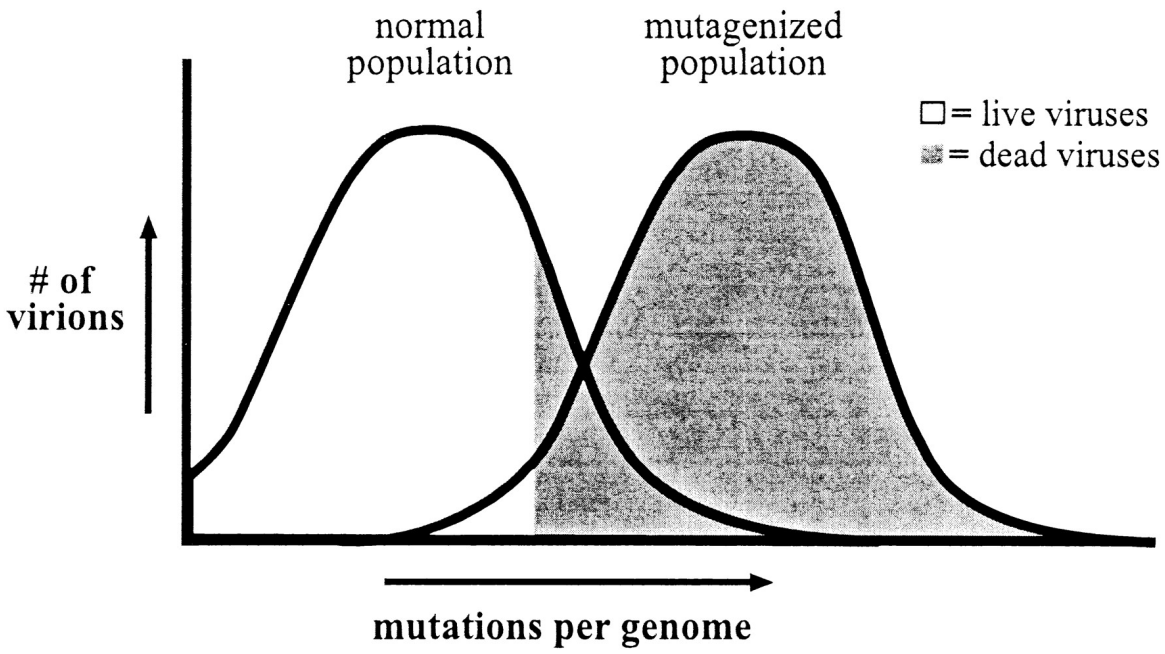
Phase diagram in the (u_0, W_0) plane

Lethal mutagenesis in RNA viruses

- RNA viruses have large mutation rates (of the order of 1 mutation per genome per replication) in spite of their *very small* genome length (a few 10^4 bases)
- It has been suggested that increasing the mutation rate can push a viral population beyond the extinction threshold
- This technique has been named **lethal mutagenesis**
- Can it be effective?

Lethal mutagenesis in RNA viruses

The idea:



Error Catastrophe in an RNA virus?

RNA virus error catastrophe: Direct molecular test by using ribavirin

Shane Crotty*, Craig E. Cameron†, and Raul Andino*‡

S. CROTTY ET AL., PNAS **98**, 6895 (2001)

Error Catastrophe in an RNA virus?

Mutation frequency per base μ in a normal RNA virus population (poliovirus in HeLa cells), expressed per 10^4 bases

	μ
RT-PCR	0.21
Normal virus population	2.12

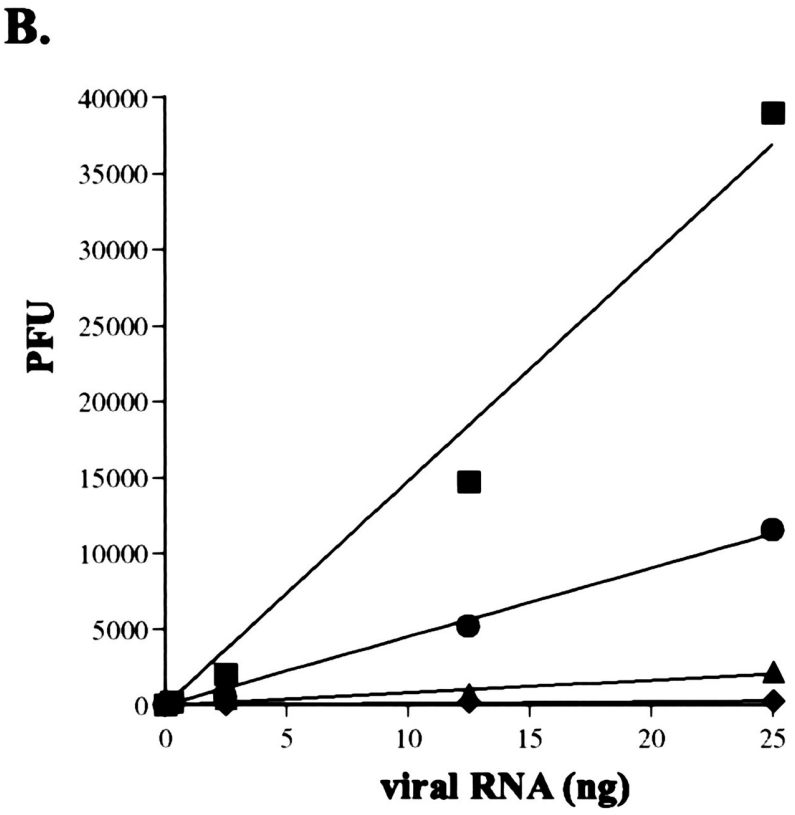
Error Catastrophe in an RNA virus?

Mutation frequency per base μ in a ribavirin-treated RNA virus population, expressed per 10^4 bases

Ribavirin conc.	G $\rightarrow$ A	C $\rightarrow$ T	Total μ
0	0.5	1.2	2.1
100 μM	—	1.3	2.5
400 μM	4.4	5.0	9.3
1000 μM	6.8	12.0	20.8

Error Catastrophe in an RNA virus?

Infectivity of ribavin-mutagenized poliovirus



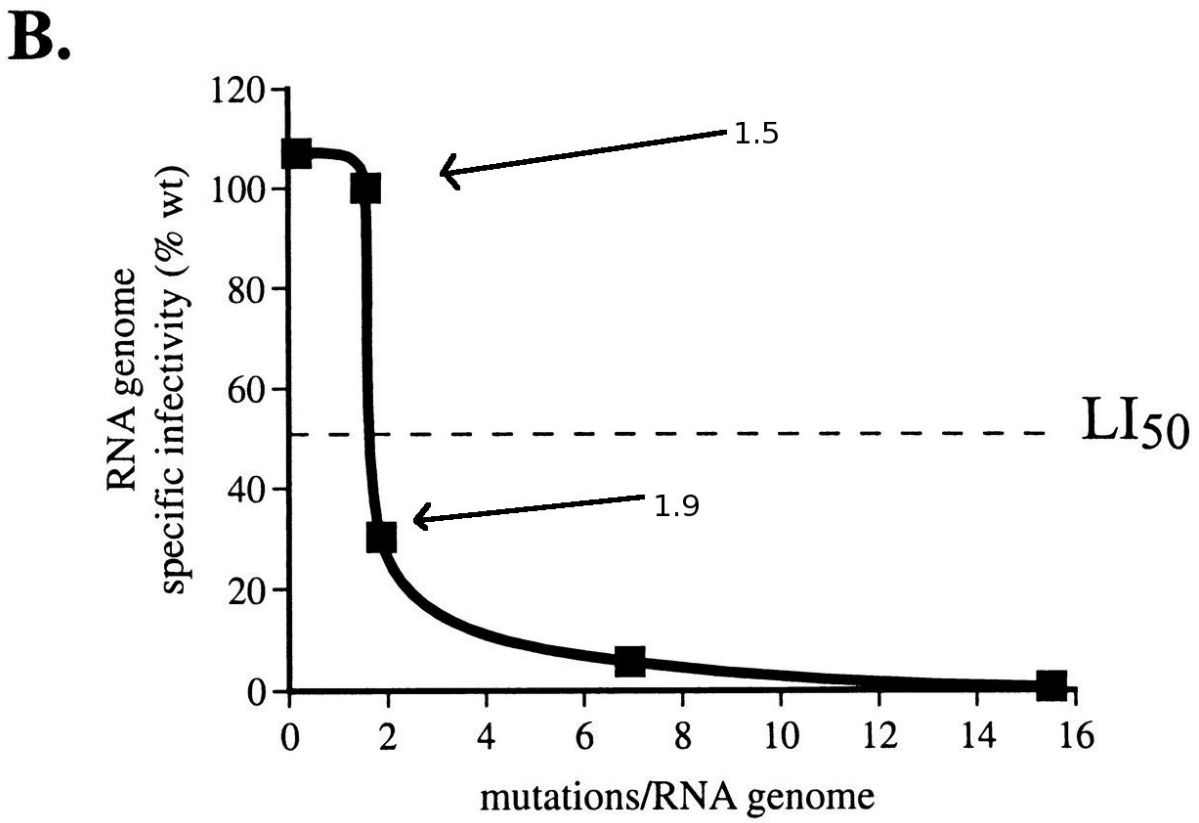
PFU: plaque-forming units

■: untreated cells; ●: 100 μM; ▲: 400 μM; ♦: 1000 μM

Measure obtained by electroporating viral RNA into HeLa cells

Error Catastrophe in an RNA virus?

Infectivity vs. # of mutations



Error Catastrophe in an RNA virus?

The data presented here demonstrate that high genetic variability, a biological property that is normally a major advantage for an RNA virus, can be turned into a weapon against the virus by increasing that mutation rate beyond tolerable levels and causing a genetic meltdown.

Unlike RNA viruses, DNA-based organisms generally have much lower mutation frequencies and do not exist near the error threshold. They appear to be able to absorb 300- to 5,000-fold higher increases in mutation frequencies before significant loss of viability is seen, although DNA viruses may be an exception.

Error Catastrophe in an RNA virus?

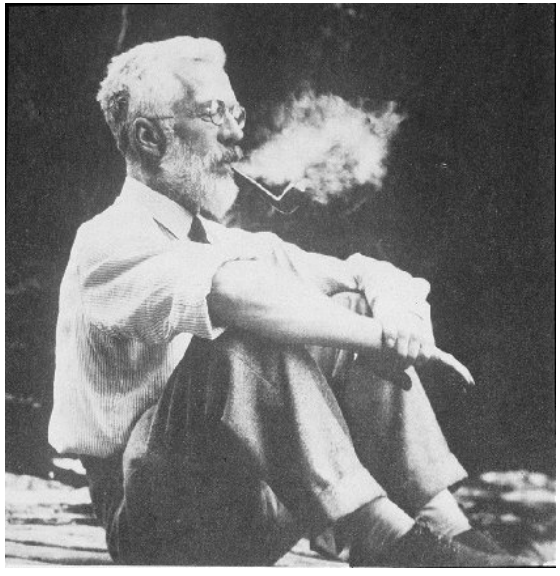
Discussion:

- Rather than error threshold, the experiments shows that RNA viruses can only sustain a limited amount of mutations, because of the several constraints on their genome (cf. *E. Holmes, 2005*)
- The reduction in infectivity seems more a path to extinction than an error catastrophe
- Difficult to conclude if data on the frequency spectrum of mutants are not available

Drift

The Population Genetics Trimurti

Sewall Wright



Ronald A. Fisher

Motoo Kimura

Finite population

The Wright-Fisher model

- Population size N, number n_k of individuals of type k, $k = 1, \ldots, r$, with fitness w_k
- Nonoverlapping generations
- Given the composition vector $\boldsymbol{x} = (x_i)$, $x_i = n_i/N$, the numbers n'_k in the next generation are distributed according to

$$\text{Prob}(n'_1, \ldots, n'_r) = \frac{N!}{n'_1! \cdots n'_r!} \xi_1^{n'_1} \cdots \xi_r^{n'_r}$$

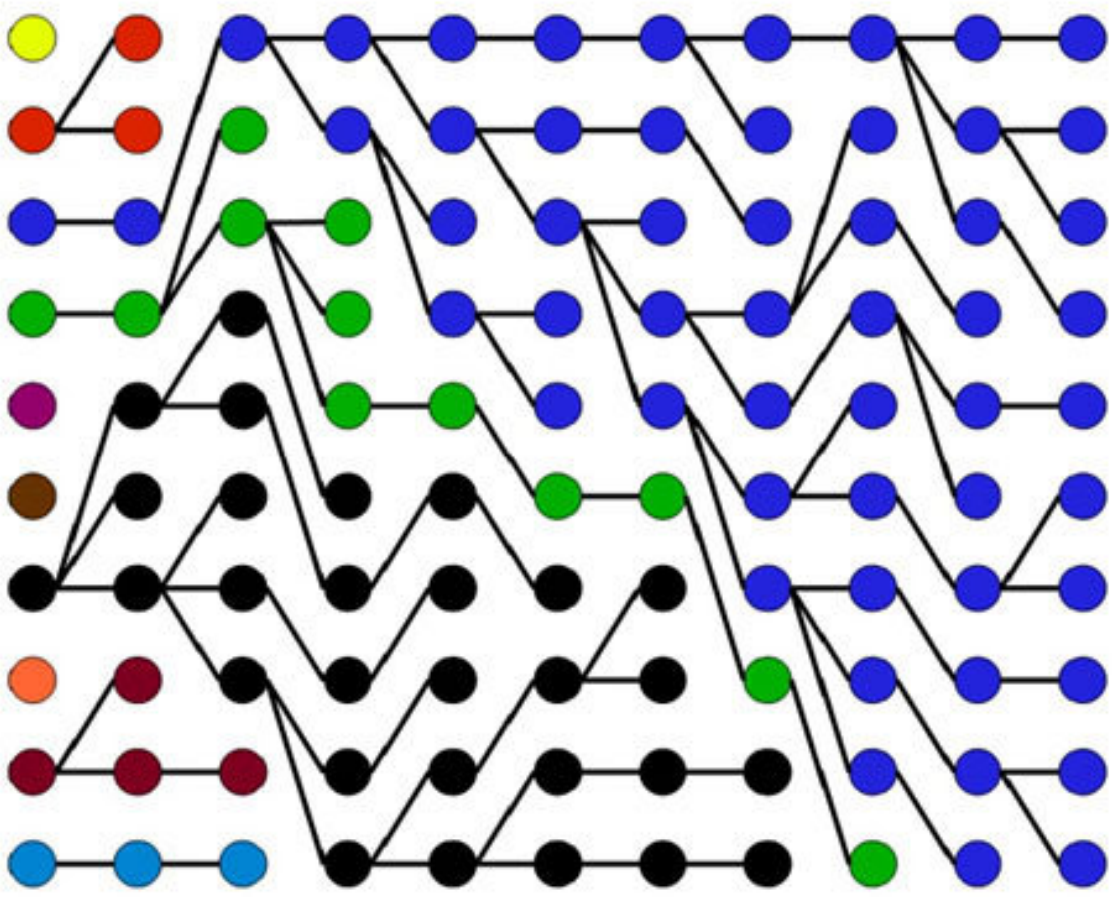
where

$$\xi_k = \frac{x_k w_k}{\sum_j x_j w_j}$$

- Thus n'_k is approximately distributed as a Gaussian with mean $N\xi_k$ and variance $N\xi_k(1 - \xi_k)$

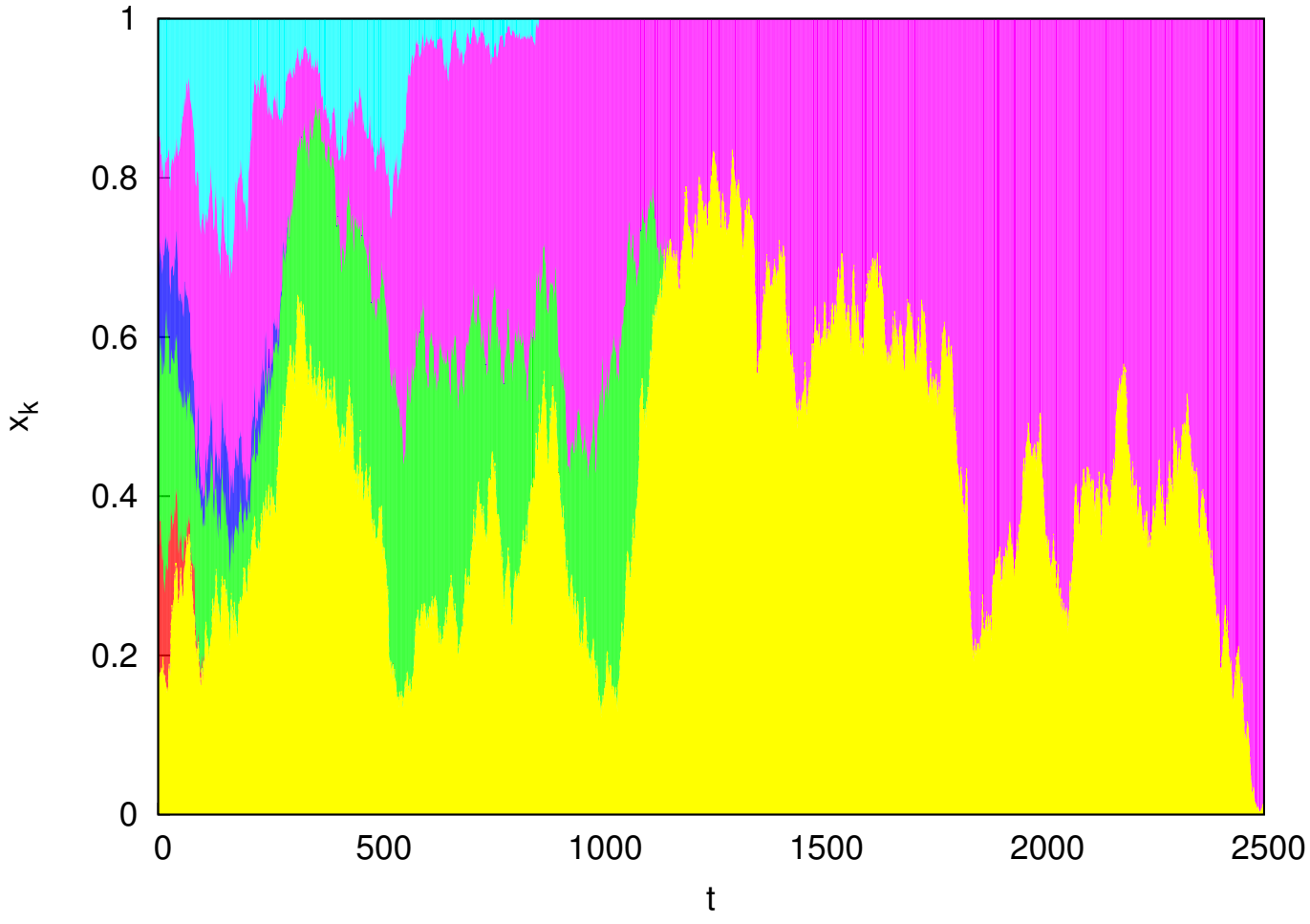
Finite population

The Wright-Fisher model



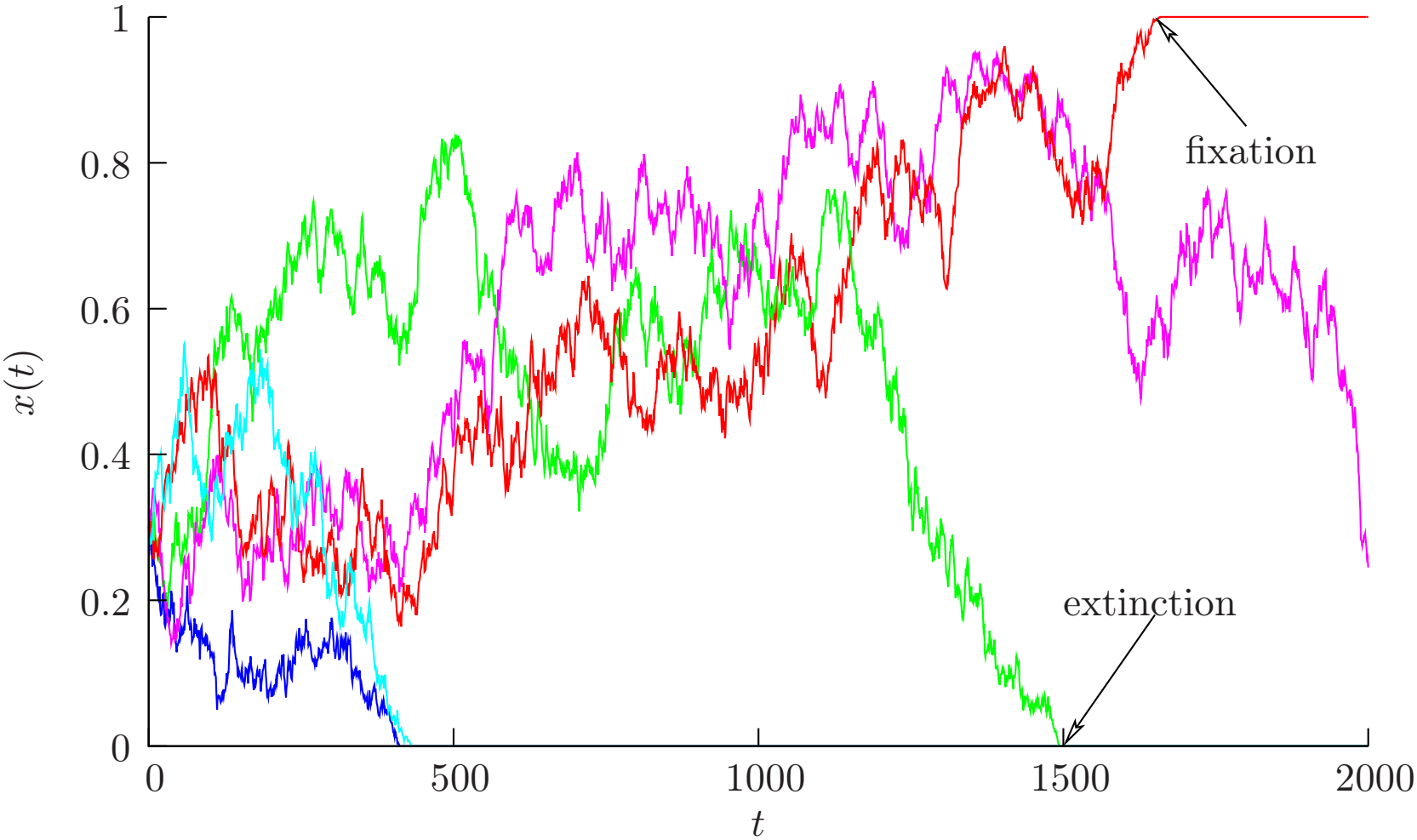
Finite population

The Wright-Fisher model: one realization (neutral)



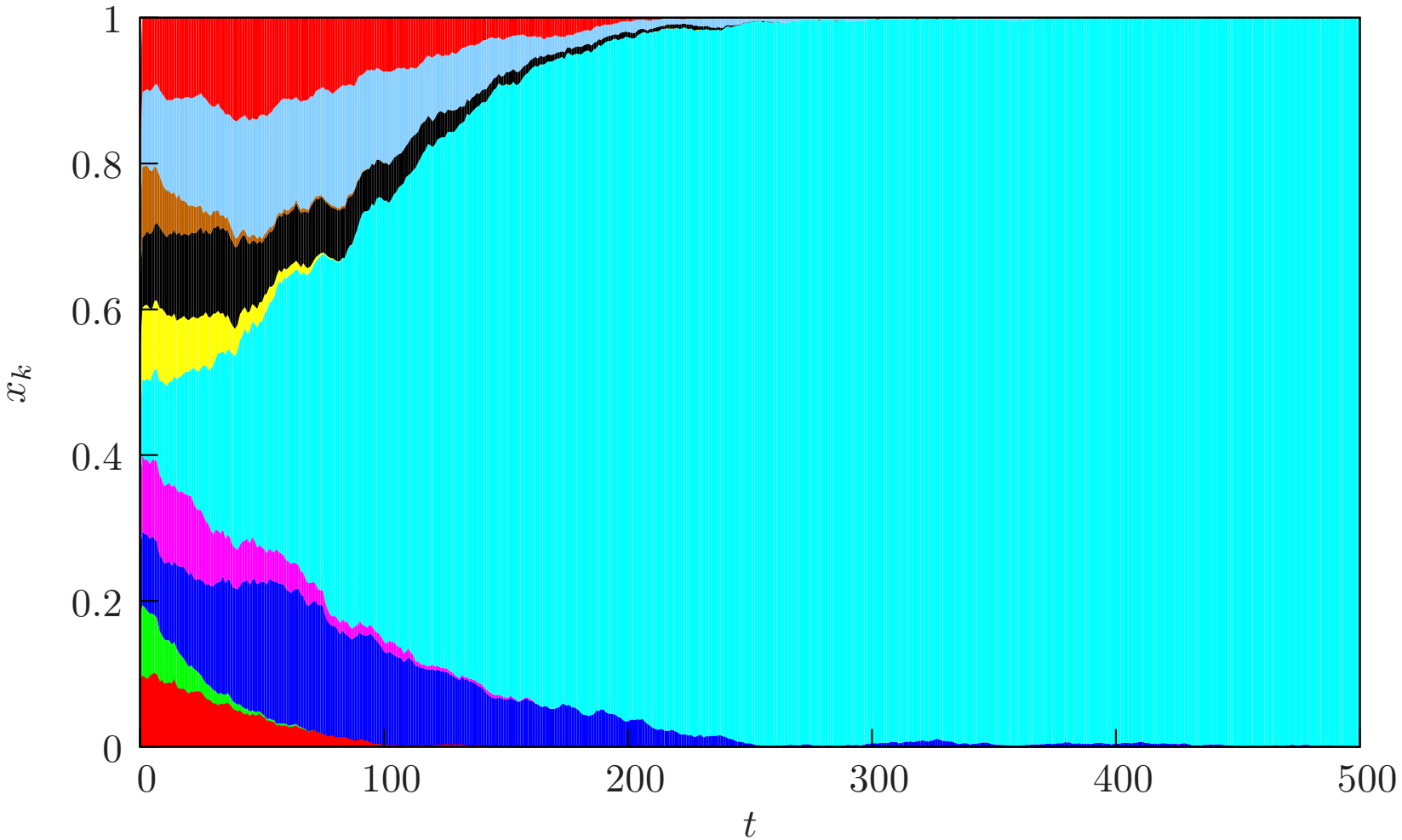
Finite population

The Wright-Fisher model: several realizations (neutral)



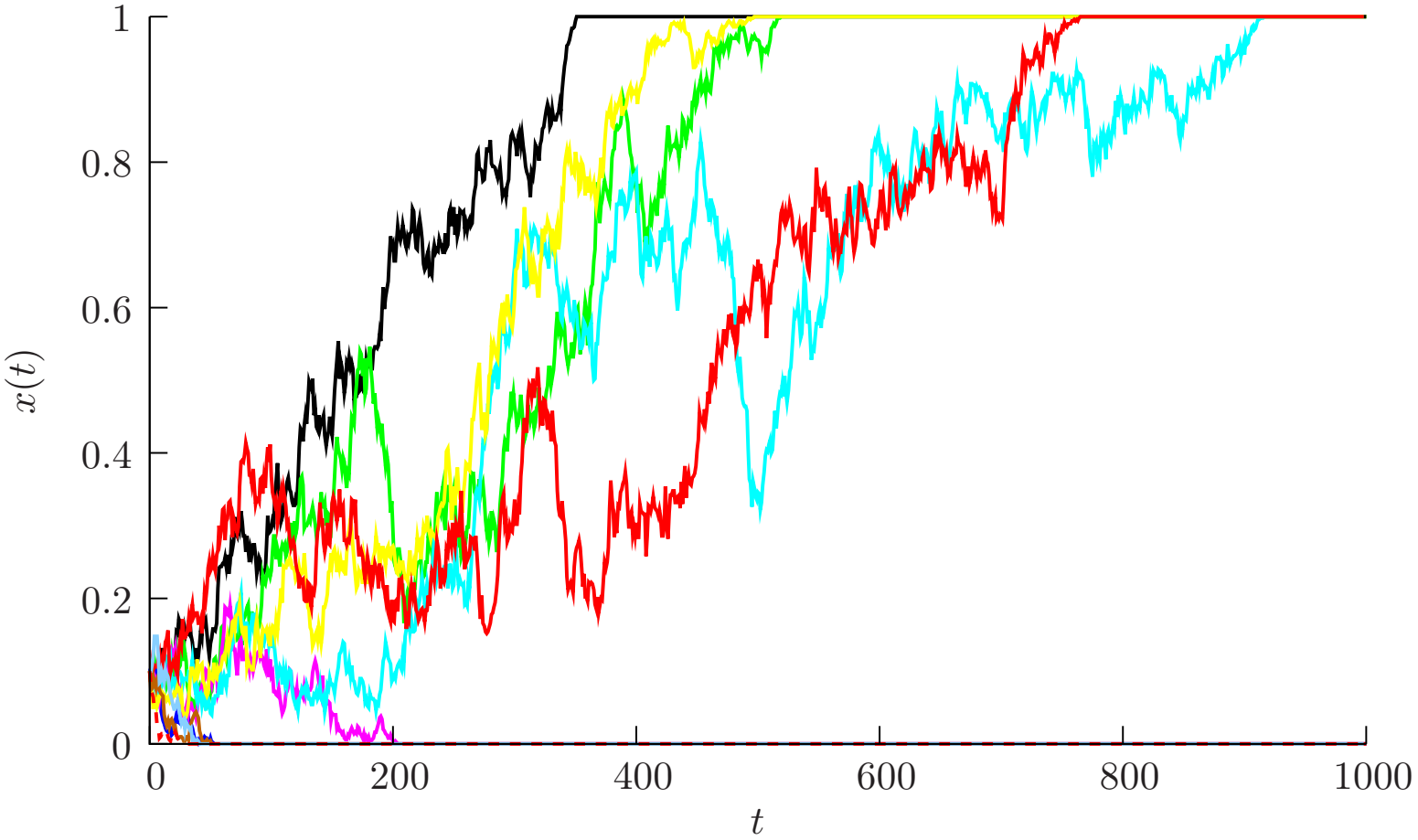
Finite population

The Wright-Fisher model: one realization (selective: $N = 10\,000$, $w_k \in [1.0, 1.1]$, $x_k(0) = 0.1$)



Finite population

The Wright-Fisher model: several realizations (selective: $N = 500$, $s = 0.01$, $x(0) = 0.1$)



Fixation in 5 cases out of 10

Drift

. . . it is often convenient to consider a natural population not so much as an aggregate of living individuals as an aggregate of gene ratios. Such a change of viewpoint is similar to that familiar in the theory of gases. . .

R. A. FISHER, 1953

Drift

We will start our discussion from the simplest situation where the gene frequency fluctuates from generation to generation because of the random sampling of gametes in a finite population. Since Wright's work, the term drift has become quite popular among biologists. However, in the mathematical theory of Brownian motion, the term drift originally connotes directional movement of the particle; therefore in our context the adjective random should be attached to it.

M. KIMURA, 1964 (ABRIDGED)

Drift

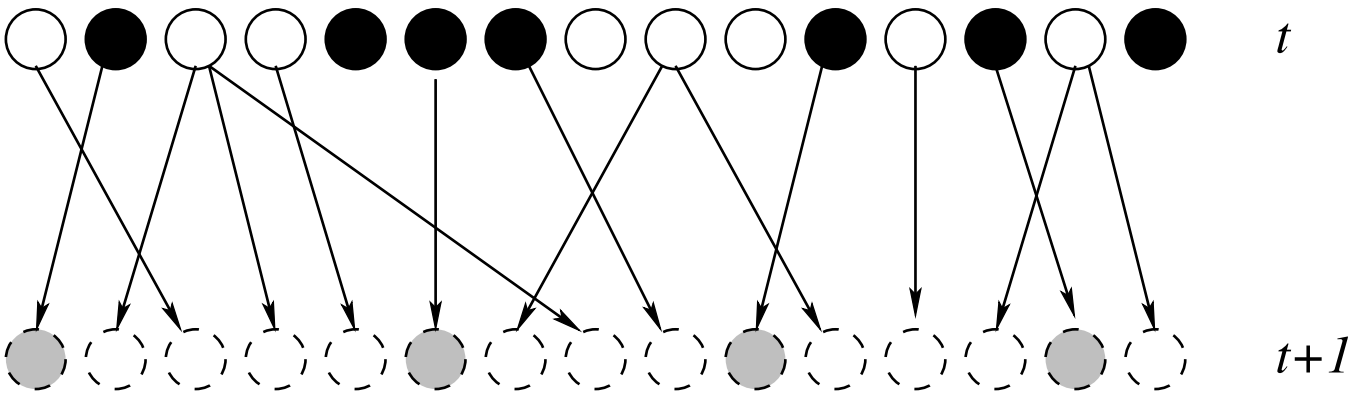
- Finite population implies different outcomes for different experiments in the same conditions (lack of *self-averaging*)
- Necessity to describe an *ensemble* of populations
- Use of the theory of Markov processes
- Simplification by means of *diffusion equations*

Random drift in the neutral case

- Population of N haploid individuals, 2 neutral alleles: A, a
- Frequency of the A allele: $x = n_{\mathrm{A}}/N$
- Wright-Fisher model: At each time step, each individual i of the new generation picks up a parent at random and copies it

Random drift in the neutral case

The Wright-Fisher model



Random drift in the neutral case

- Probability that $n_{\mathrm{A}}(t+1) = n$, given $n_{\mathrm{A}}(t) = Nx(t)$:

$$p_n(t+1) = \binom{N}{n} (x(t))^n (1-x(t))^{N-n}$$

- Assume $N \gg 1$, $\frac{1}{N} \ll x \ll 1 - \frac{1}{N}$, then

$$\mathrm{Prob}\left(x(t+1)=x\right) \propto \exp\left(-\frac{(x-x(t))^2}{2Nx(t)(1-x(t))}\right)$$

- $\Delta x(t) = x(t+1) - x(t)$:

$$\langle \Delta x(t) \rangle = 0 \qquad \left\langle (\Delta x(t))^2 \right\rangle = \frac{x(t)(1-x(t))}{N}$$

The diffusion equation

Fokker-Planck equation:

$$\boxed{\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}\left(\langle \Delta x \rangle_x \, p(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\left\langle \Delta x^2 \right\rangle_x p(x,t)\right)}$$

In our case

$$\frac{\partial p}{\partial t} = \frac{1}{2N}\frac{\partial^2}{\partial x^2}\left(x(1-x)\,p(x,t)\right)$$

The solution in the neutral case

- Set $p(x,t \mid x_0,0) = \sum_n c_n(x_0)\chi_n(x)\,\mathrm{e}^{-\lambda_n t/(2N)}$
- Eigenvalue equation:

$$x(1-x)\chi_n''(x) + (1-2x)\chi_n'(x) + \lambda_n\chi_n(x) = 0$$

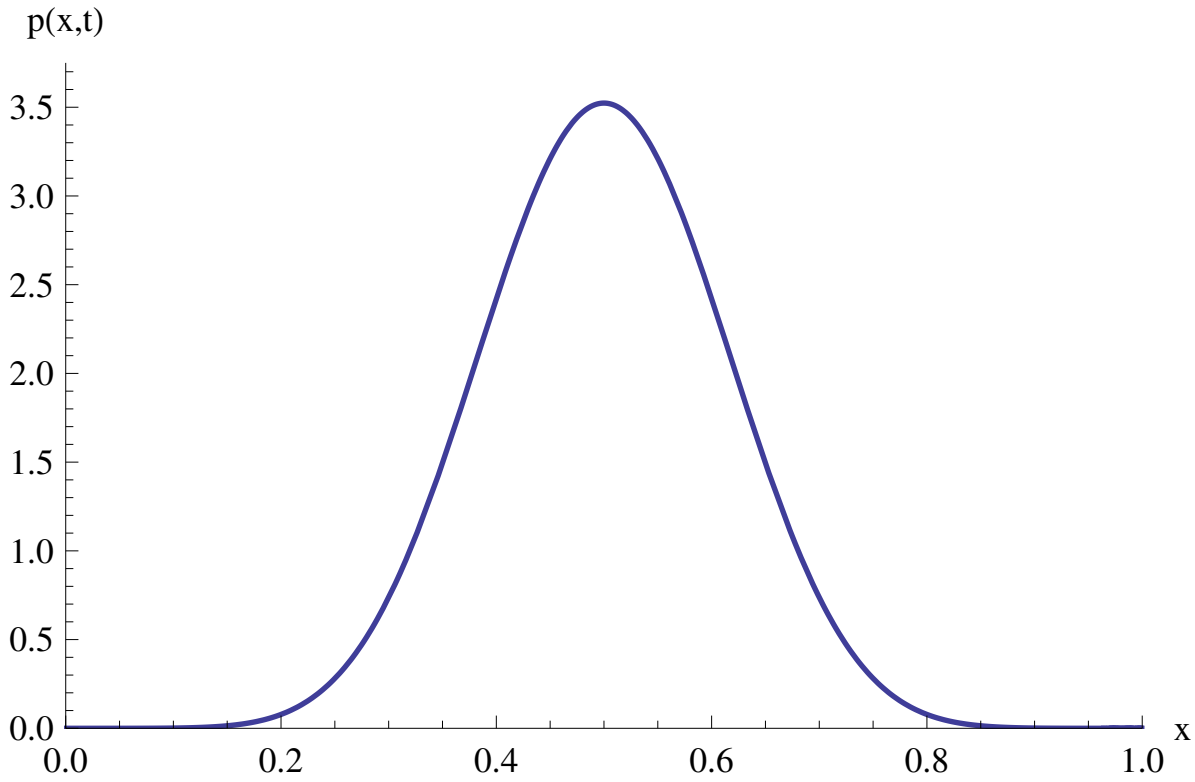
- Boundary conditions: $x = 0, 1$ are singular points; we require $\chi_n(0,1)$ finite $\forall n$
- Initial condition:

$$p(x,0 \mid x_0,0) = \sum_n c_n(x_0)\chi_n(x) = \delta(x-x_0)$$

- Solution in terms of hypergeometric functions:

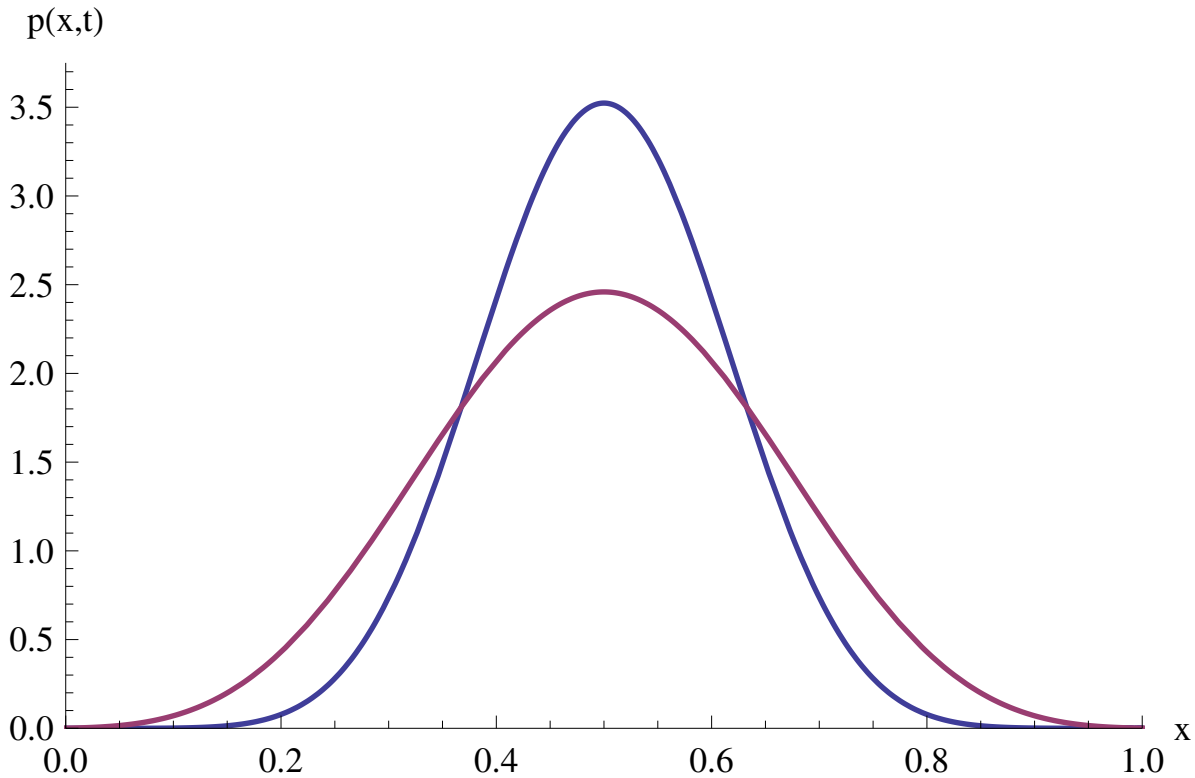
$$\chi_n(x) = F(1-n, n+2, 2, x) \qquad \lambda_n = n(n+1)$$

The solution in the neutral case



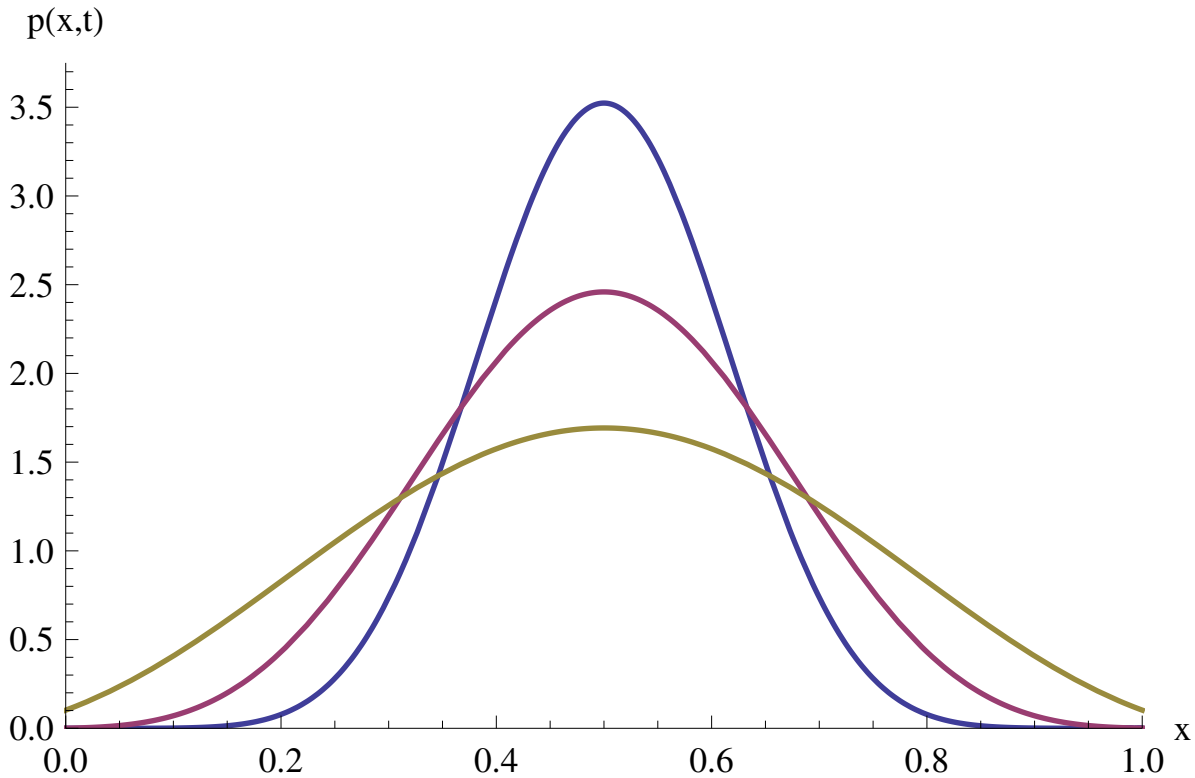
$t = 0.05N$

The solution in the neutral case



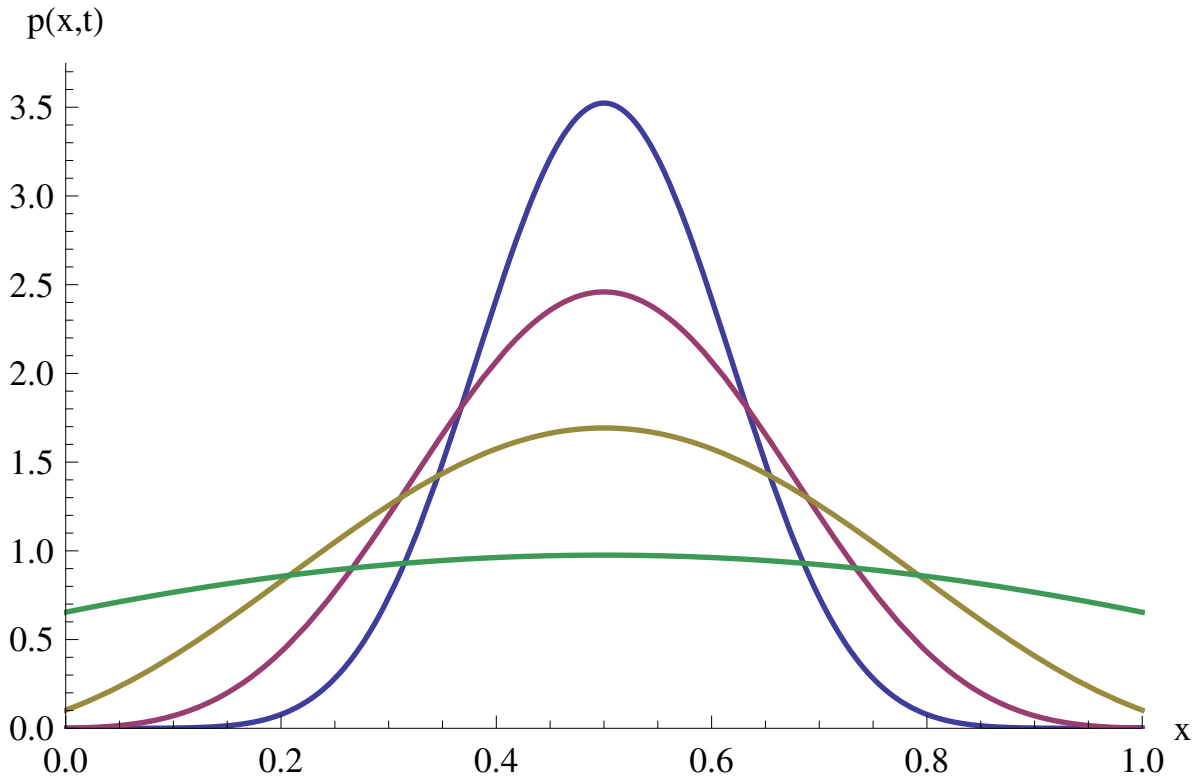
$t = 0.1N$

The solution in the neutral case



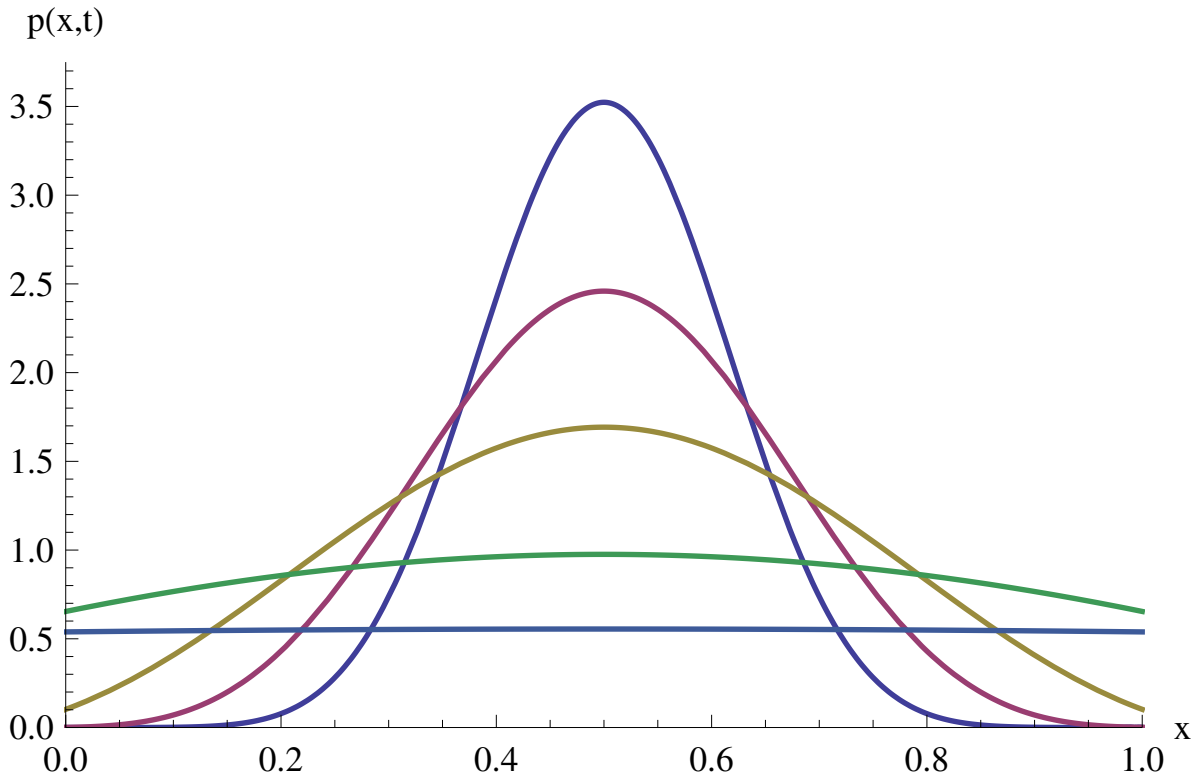
$t = 0.2N$

The solution in the neutral case



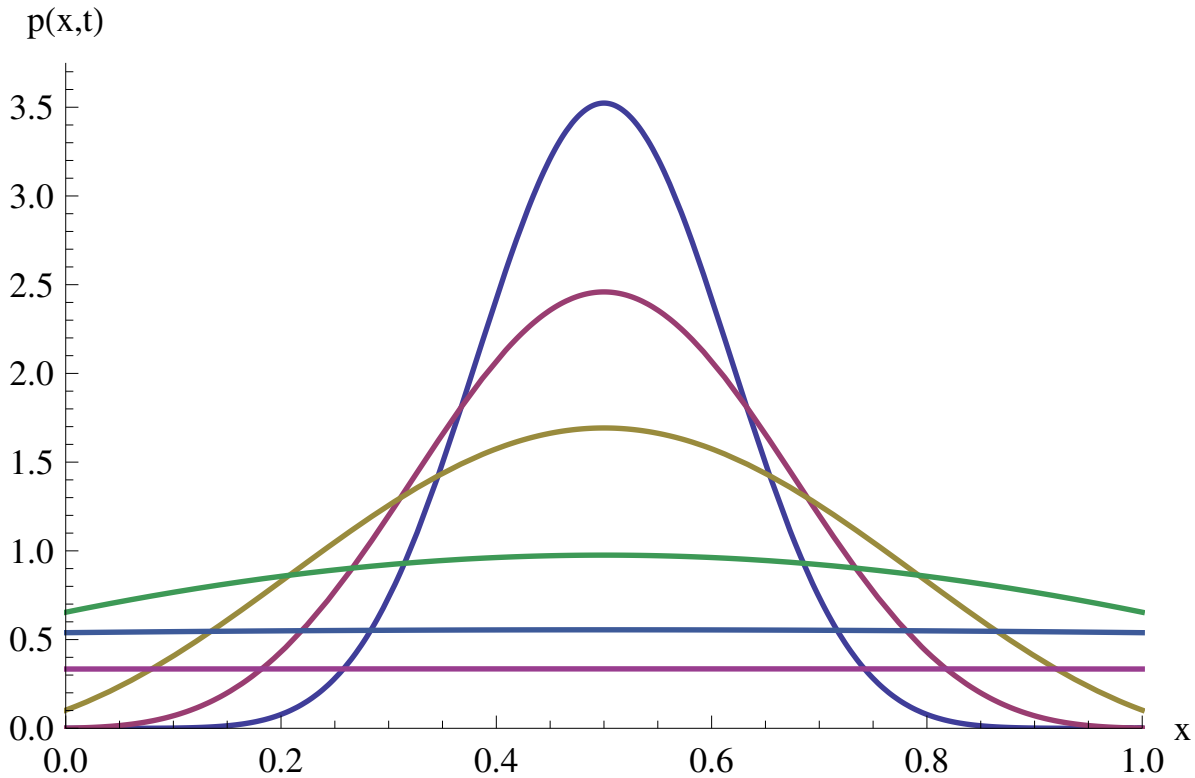
$t = 0.5N$

The solution in the neutral case



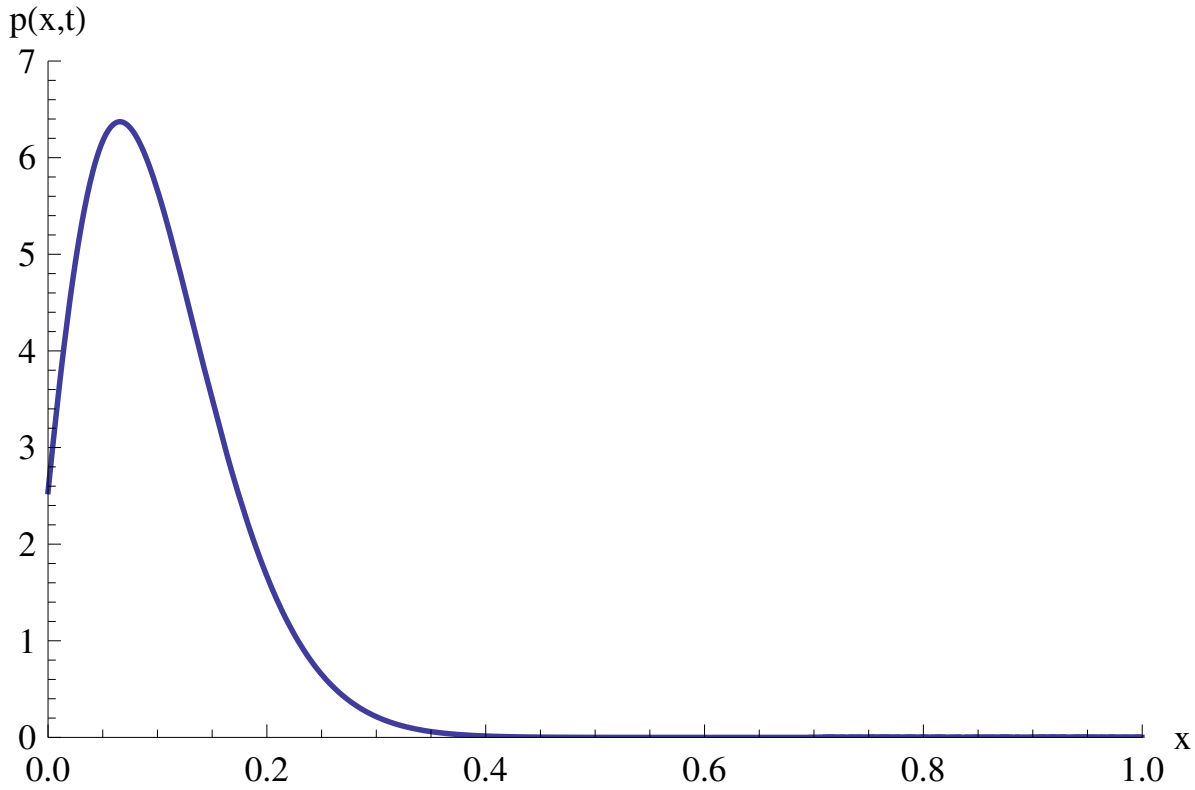
$t = N$

The solution in the neutral case



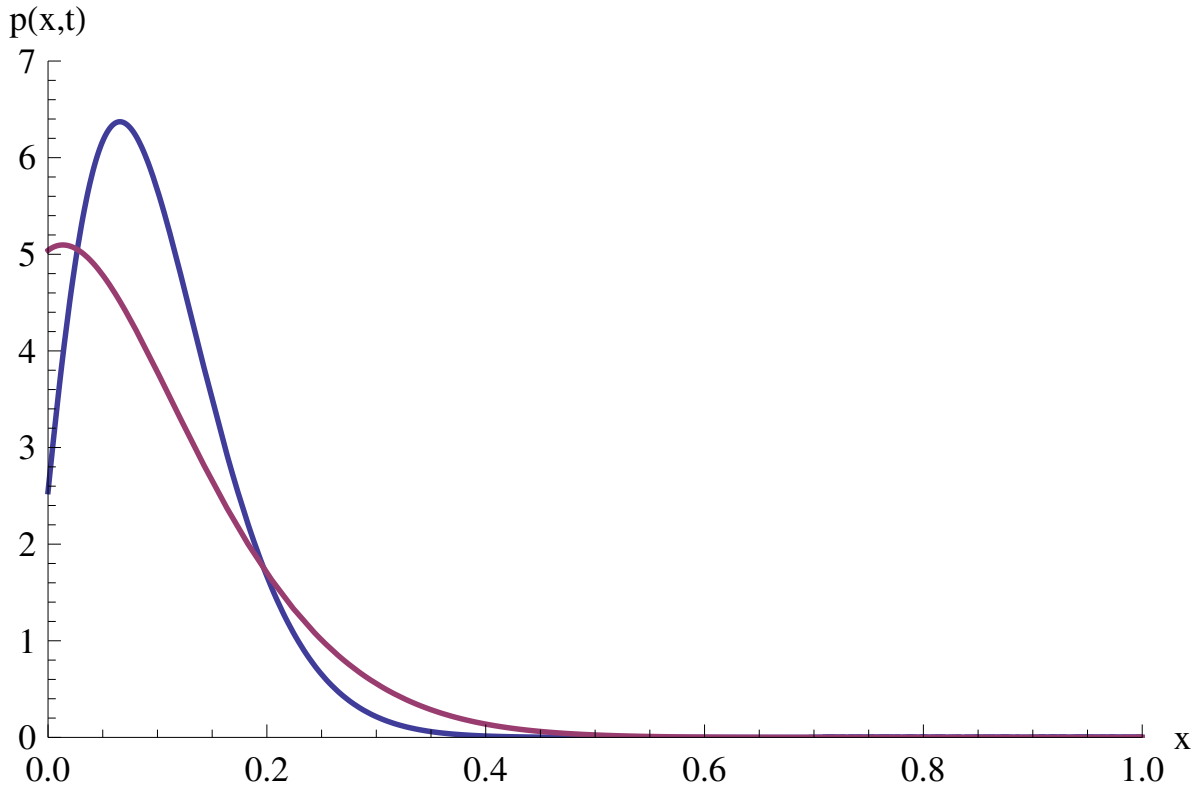
$t = 1.5N$

Initial condition $x(0) = 0.1$



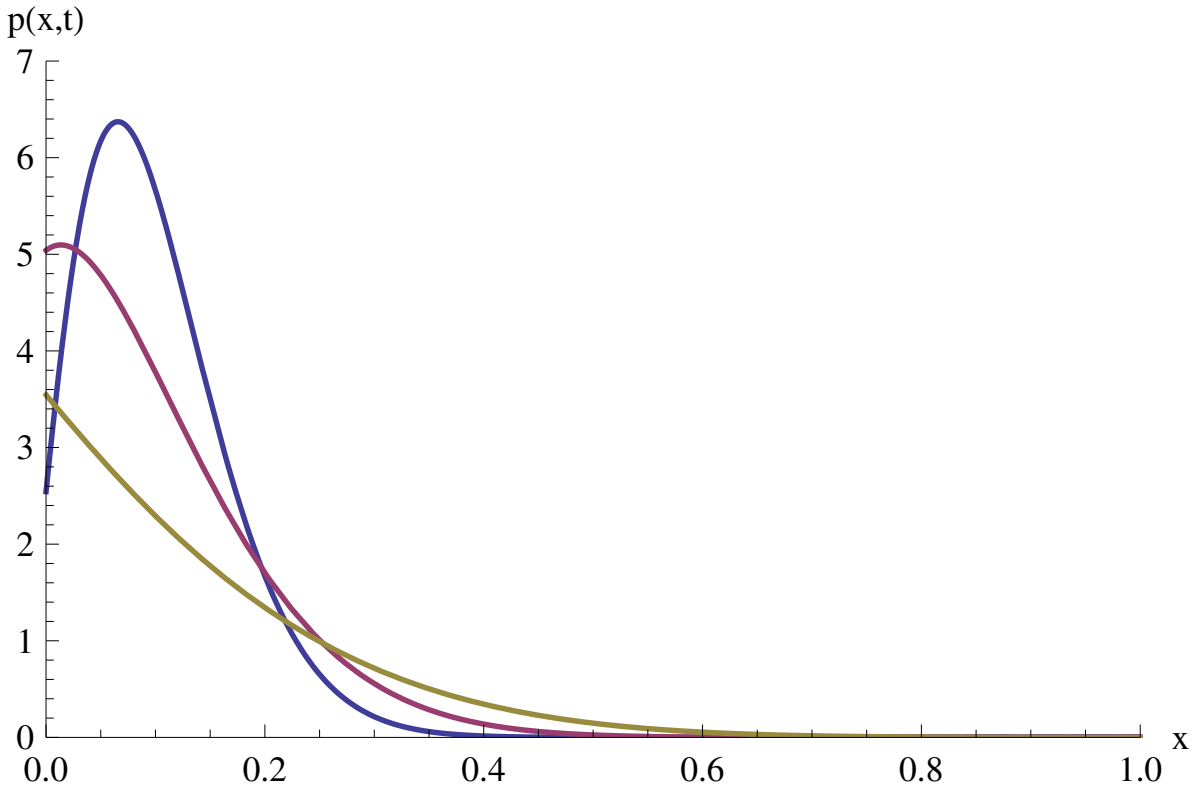
$t = 0.05N$

Initial condition $x(0) = 0.1$



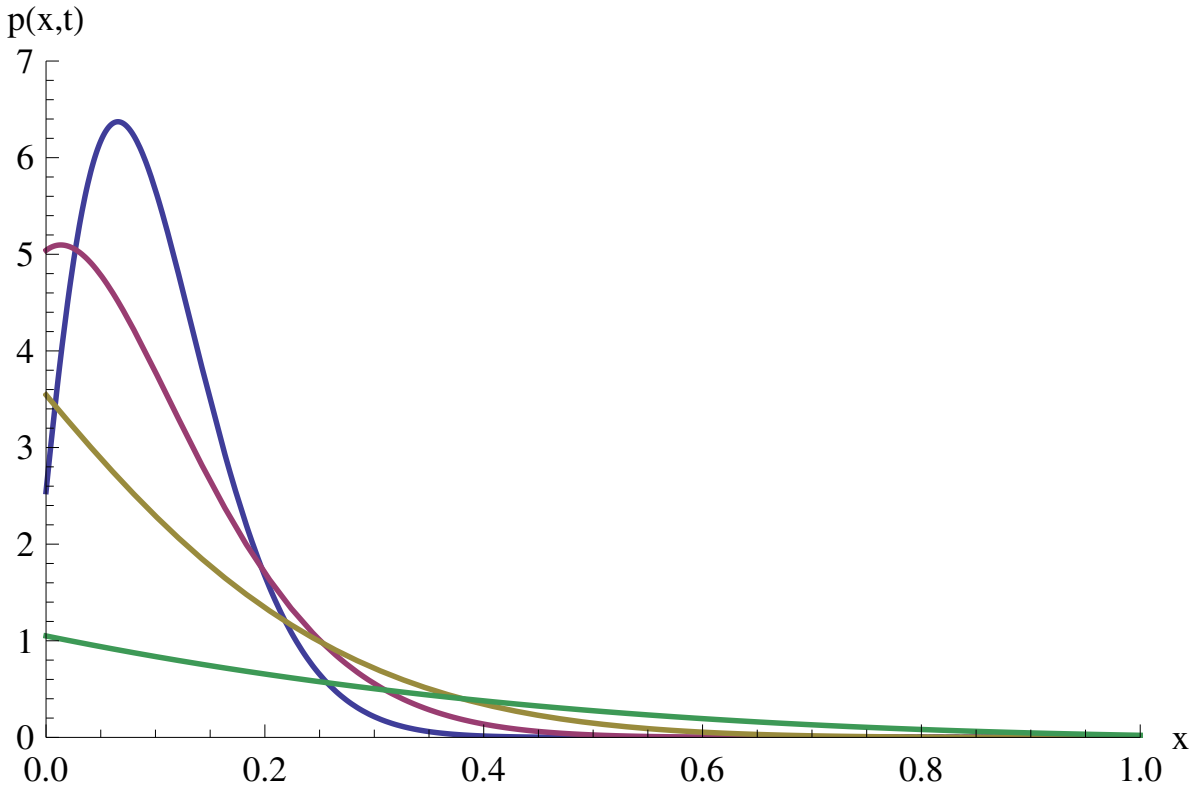
$t = 0.1N$

Initial condition $x(0) = 0.1$



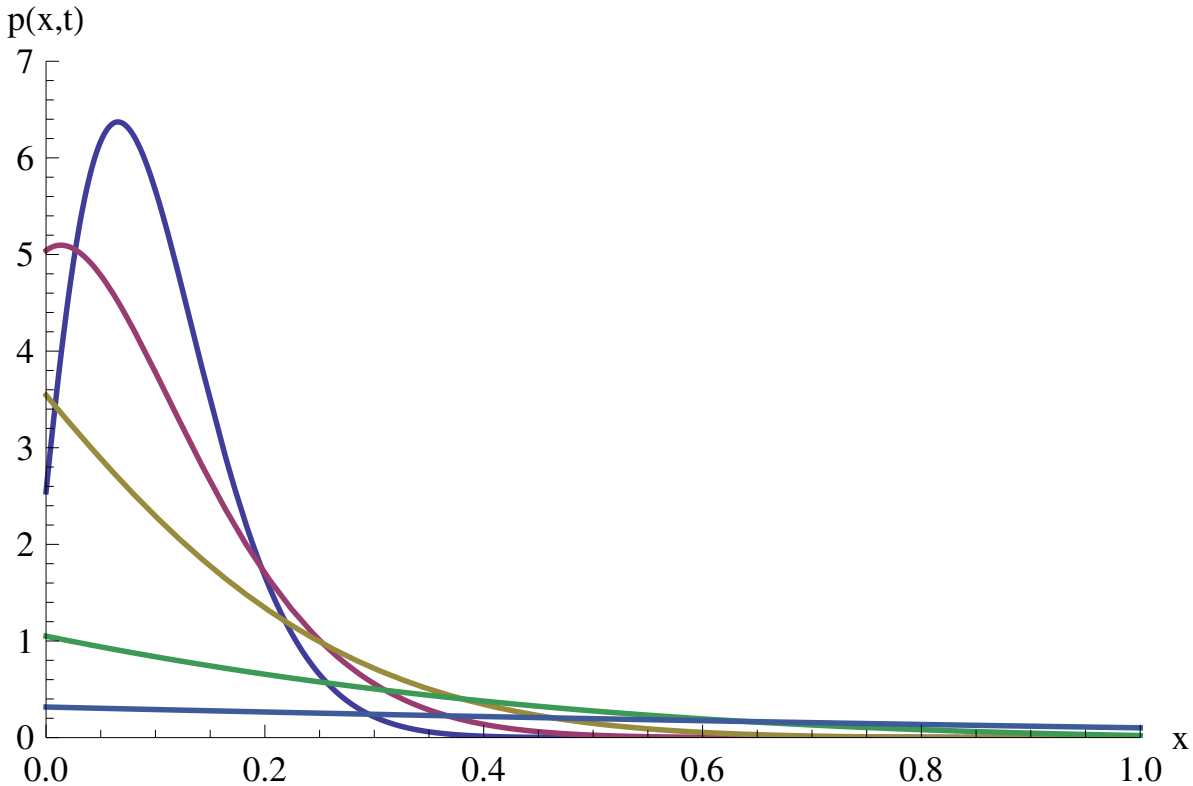
$t = 0.2N$

Initial condition $x(0) = 0.1$



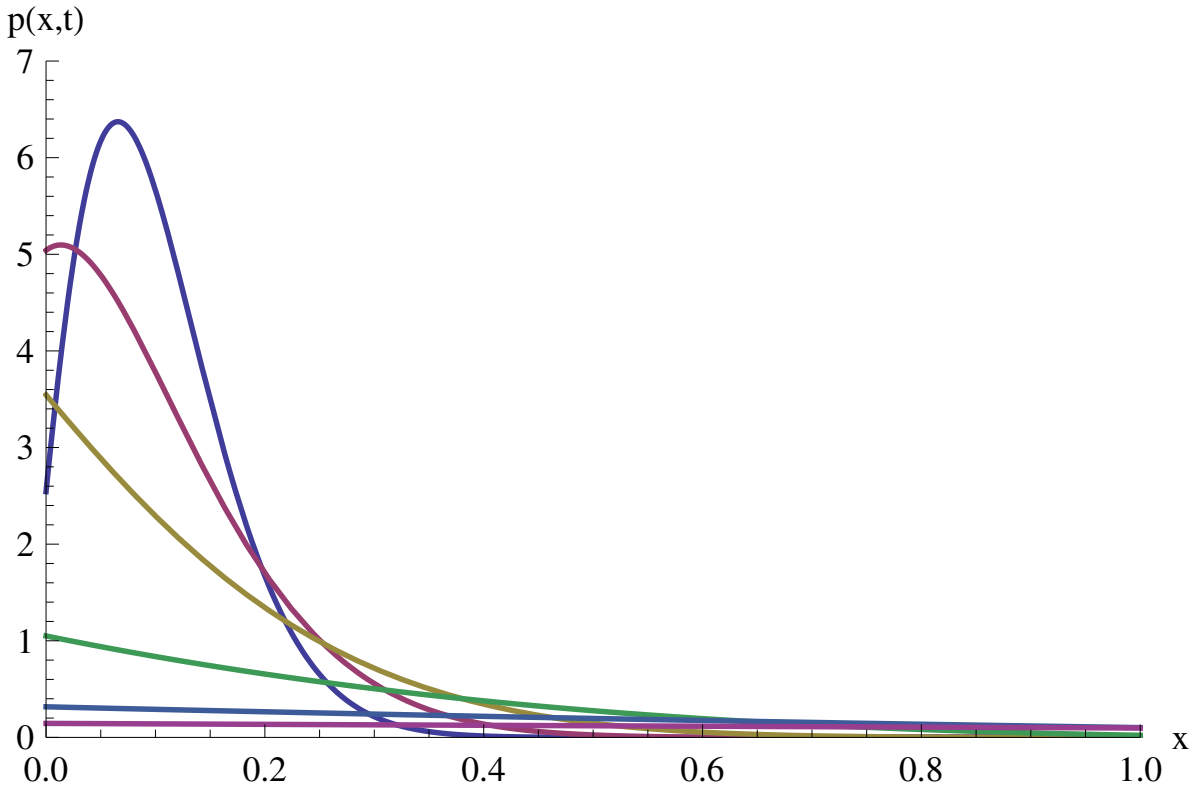
$t = 0.5N$

Initial condition $x(0) = 0.1$



$t = N$

Initial condition $x(0) = 0.1$



$t = 1.5N$

Results

- $p(x,t)$ decays exponentially:
  $p(x,t) \simeq 6x(0)(1-x(0))\mathrm{e}^{-t/N}$ for $t \gg N$
- Probability that A and a coexist at generation t:
  $\Omega(t) = \int_0^1 \mathrm{d}x\, p(x,t)$ decays with the same rate ($p(x,t)$ is flat)
- However, $p(x,t)$ becomes flat later when $x(0) \neq \frac{1}{2}$
- What is the probability of fixation of allele A as a function of $x(0)$?

The backward equation

- $p(x, t \mid x_0, t_0)$: Conditional probability that $x(t) = x$ *given* that $x(t_0) = x_0$
- Consider the effect of a single-generation sampling near t_0: $x(t_0) + 1 = x_0 + \Delta x_0$
- Equation for $p(x, t \mid x_0, t_0)$:

$$\boxed{-\frac{\partial p}{\partial t_0} = \langle \Delta x_0 \rangle_{x_0} \frac{\partial p}{\partial x_0} + \frac{1}{2} \left\langle \Delta x_0^2 \right\rangle_{x_0} \frac{\partial^2 p}{\partial x_0^2}}$$

- In our case

$$-\frac{\partial p}{\partial t_0} = \frac{x_0(1 - x_0)}{2N} \frac{\partial^2 p}{\partial x_0^2}$$

N.B. In the stationary case, p depends only on $t - t_0$, thus $\partial p / \partial t_0 = -\partial p / \partial t$

The fixation probability

- $P(t, x_0, t_0) = p(1, t \mid x_0, t_0)$: probability of being fixed by time t
- "Ultimate" fixation probability: $p^{\text{fix}}(x_0) = \lim_{t\to\infty} P(t, x_0, t_0)$
- From the backward equation we obtain

$$\frac{\mathrm{d}^2 p^{\text{fix}}}{\mathrm{d}x_0^2} = 0 \qquad x \in [0, 1]$$

- Boundary conditions: $p^{\text{fix}}(x_0{=}0) = 0$ and $p^{\text{fix}}(x_0{=}1)$
- Solution:

$$p^{\text{fix}}(x_0) = x_0$$

Wright-Fisher model with selection

- Population of N haploid individuals, two alleles A and a
- Fitnesses: w_{A}, w_{a}
- Probability that an individual with allele A is chosen as a parent:

$$\xi_{\mathrm{A}} = \frac{n_{\mathrm{A}} w_{\mathrm{A}}}{\sum_{j=1}^{N} w_j} = \frac{n_{\mathrm{A}} w_{\mathrm{A}}}{n_{\mathrm{A}} w_{\mathrm{A}} + n_{\mathrm{a}} w_{\mathrm{a}}} = \frac{x\, w_{\mathrm{A}}}{x w_{\mathrm{A}} + (1-x) w_{\mathrm{a}}}$$

- Probability that $n_{\mathrm{A}}(t+1) = n$:

$$p_n(t+1) = \binom{N}{n} \xi_{\mathrm{A}}^{n} \left(1 - \xi_{\mathrm{A}}\right)^{N-n}$$

- Average and variance:

$$\begin{aligned} \langle x_{\mathrm{A}}(t+1) \rangle &= \xi_{\mathrm{A}} \\ \left\langle \left(x_{\mathrm{A}}(t+1) - \langle x_{\mathrm{A}}(t+1) \rangle \right)^2 \right\rangle &= \xi_{\mathrm{A}} \left(1 - \xi_{\mathrm{A}}\right) / N \end{aligned}$$

Selection and drift

If the first human infant with a gene for levitation were struck by lightning in its pram, this would not prove the new genotype to have low fitness, but only that the particular child was unlucky.

JOHN MAYNARD SMITH

Selection and drift

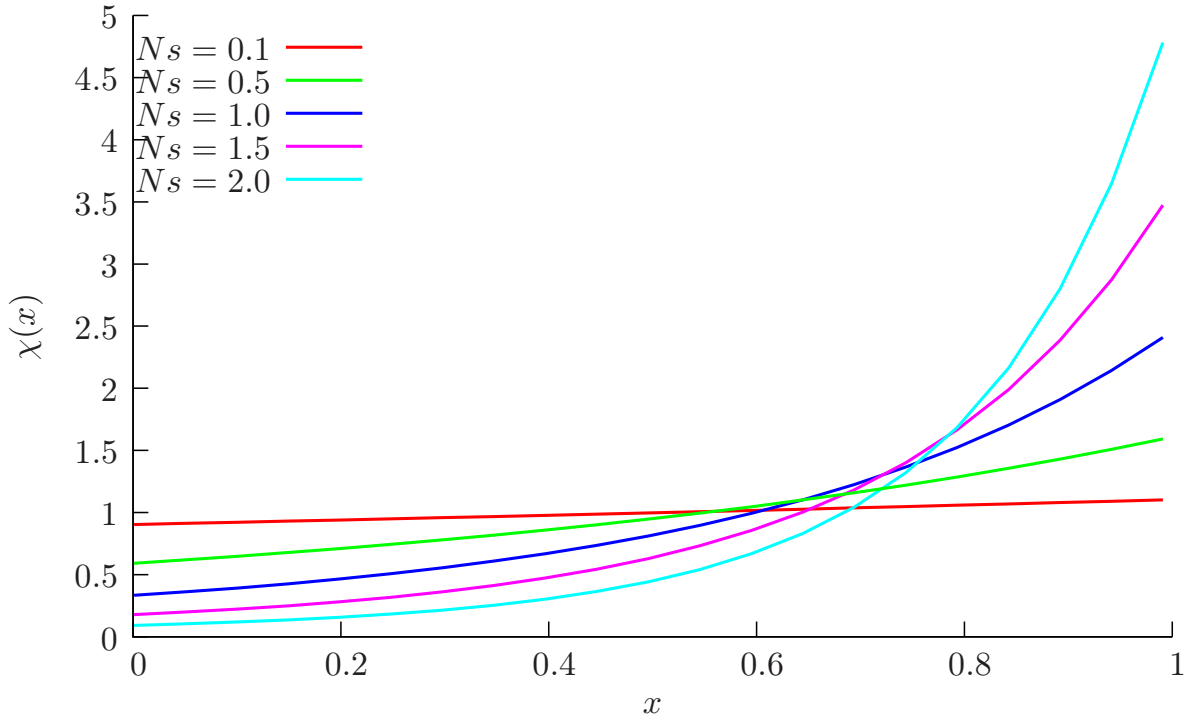
- Set $w_{\mathrm{A}} = 1 + s$, $w_{\mathrm{a}} = 1$, $s \ll 1$
- Then $\xi_{\mathrm{A}} = x w_{\mathrm{A}} / (x w_{\mathrm{A}} + w_{\mathrm{a}}(1 - x)) = (1 + s)x/(1 + sx)$
- Then
  $\langle \Delta x \rangle_x = \langle x(t + 1) \rangle - x = sx(1 - x)/(1 + sx) \simeq sx(1 - x)$
  and $\left\langle \Delta x^2 \right\rangle \simeq (x(1 - x)/N)$
- Diffusion equation for $p(x, t)$:

$$\frac{\partial p}{\partial t} = -s \frac{\partial}{\partial x} (x(1 - x)p) + \frac{1}{2N} \frac{\partial^2}{\partial x^2} (x(1 - x)p)$$

- Solution in terms of spheroidal functions...
- Asymptotically $p(x, t) \propto \chi(x)\, \mathrm{e}^{-\lambda t/N}$

Solution with selection

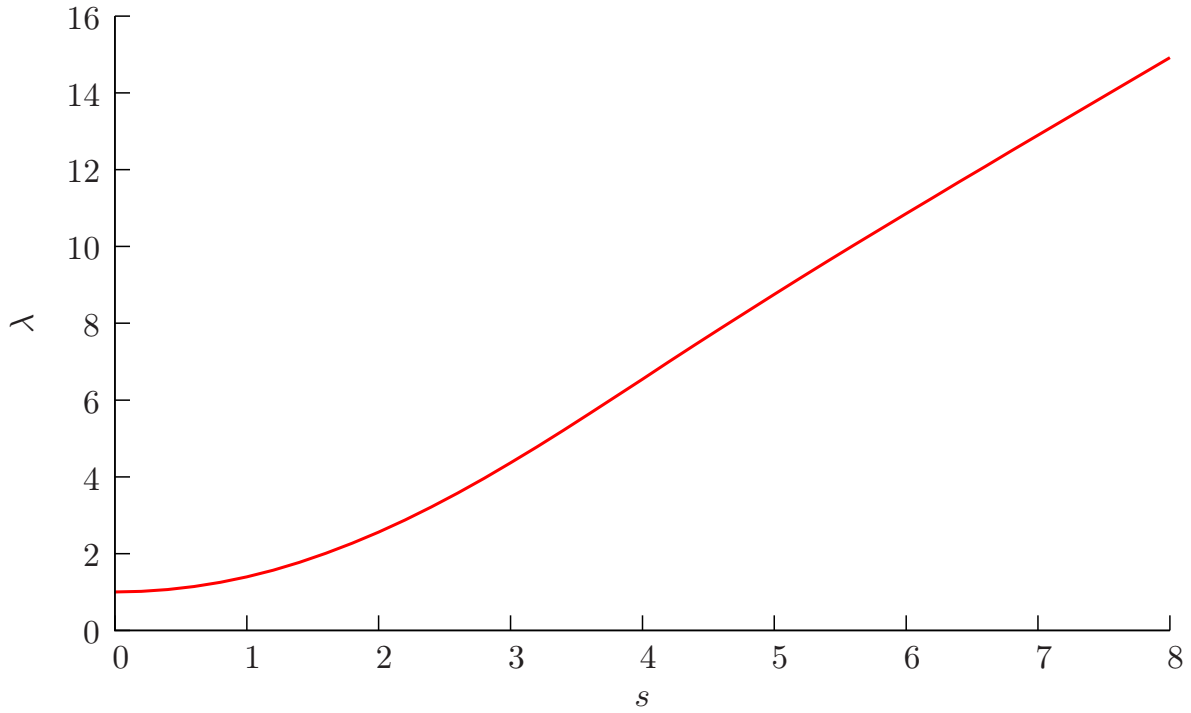
The long-living eigenfunction:



The leading eigenfunction $\chi(x)$ for several values of s

Solution with selection

The decay rate:



Leading eigenvalue λ as a function of Ns; decay rate: λ/N

The fixation probability with selection

- The backward equation:

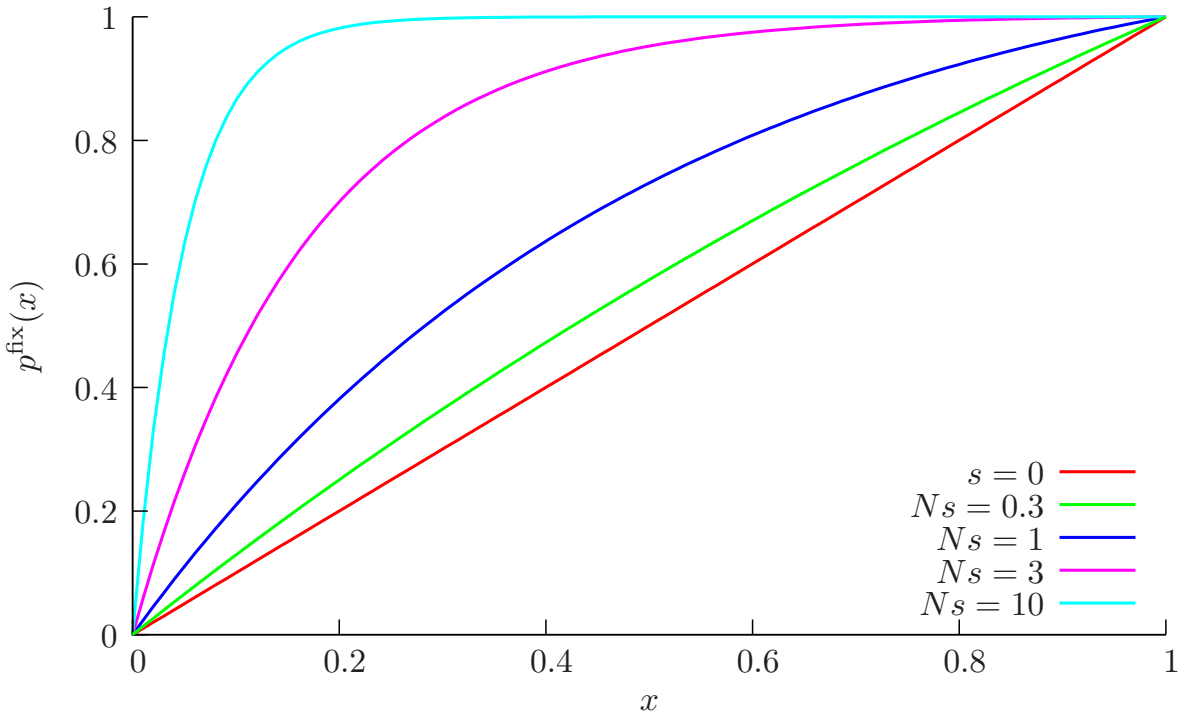
$$\frac{\partial p}{\partial t_0} = s x_0 (1 - x_0) \frac{\partial p}{\partial x_0} + \frac{x_0 (1 - x_0)}{2N} \frac{\partial^2 p}{\partial x_0^2}$$

- Stationary solution:

$$\begin{aligned} \frac{\partial p^{\text{fix}}}{\partial x_0} &= C_1 \mathrm{e}^{-2Nsx_0} \\ p^{\text{fix}}(x_0) &= C_0 - C_1 \mathrm{e}^{-2Nsx_0} \\ &= \frac{1 - \mathrm{e}^{-2Nsx_0}}{1 - \mathrm{e}^{-2Ns}} \end{aligned}$$

- In particular, for $s \to 0$, $p^{\text{fix}} \to x_0$

The fixation probability with selection



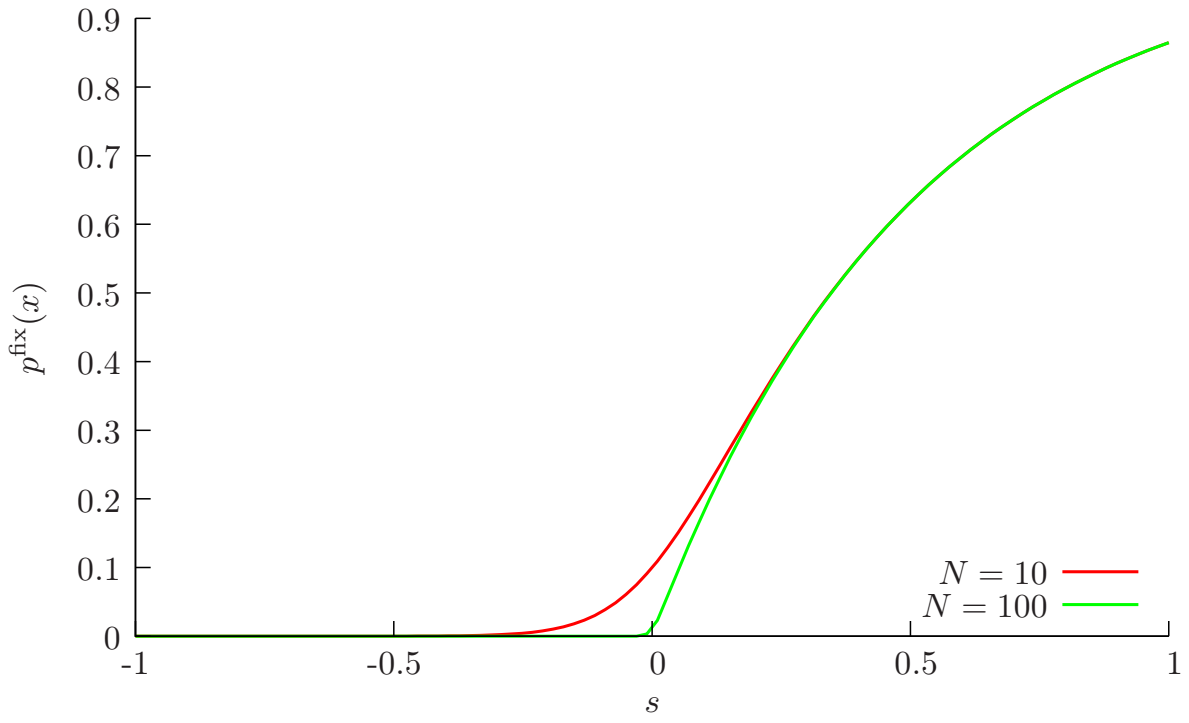
Fixation probability of a single mutant

- For a single mutant $x_0 = \frac{1}{N}$
- Thus

$$\boxed{p^{\text{fix}} = \frac{1 - \mathrm{e}^{-2s}}{1 - \mathrm{e}^{-2Ns}}}$$

- Limits:
  - $s > 0$, $Ns \gg 1$: $p^{\text{fix}} \simeq 1 - \mathrm{e}^{-2s}$ (for $s \ll 1$, $p^{\text{fix}} \simeq 2s$)
  - $s < 0$, $|Ns| \gg 1$, $p^{\text{fix}} \simeq 0$
  - $|Ns| \lesssim 1$, $p^{\text{fix}} \simeq \frac{1}{N}$

Fixation probability of a single mutant



Frequency needed to obtain fixation

- How large must be x to be "almost sure" that a beneficial mutant fixes?
- Solve
$$p^{\mathrm{fix}}(x^*) = 1 - \gamma$$
- For $Ns \gg 1$ we have $p^{\mathrm{fix}}(x) \simeq 1 - \mathrm{e}^{-2Nsx}$, thus
$$x^* = -\frac{\log \gamma}{2Ns} \qquad \text{or} \qquad n^* = -\frac{\log \gamma}{2s}$$
- The fate of the mutant is determined in its initial phase, where it undergoes a branching process—the size of N is irrelevant!

Substitution rate

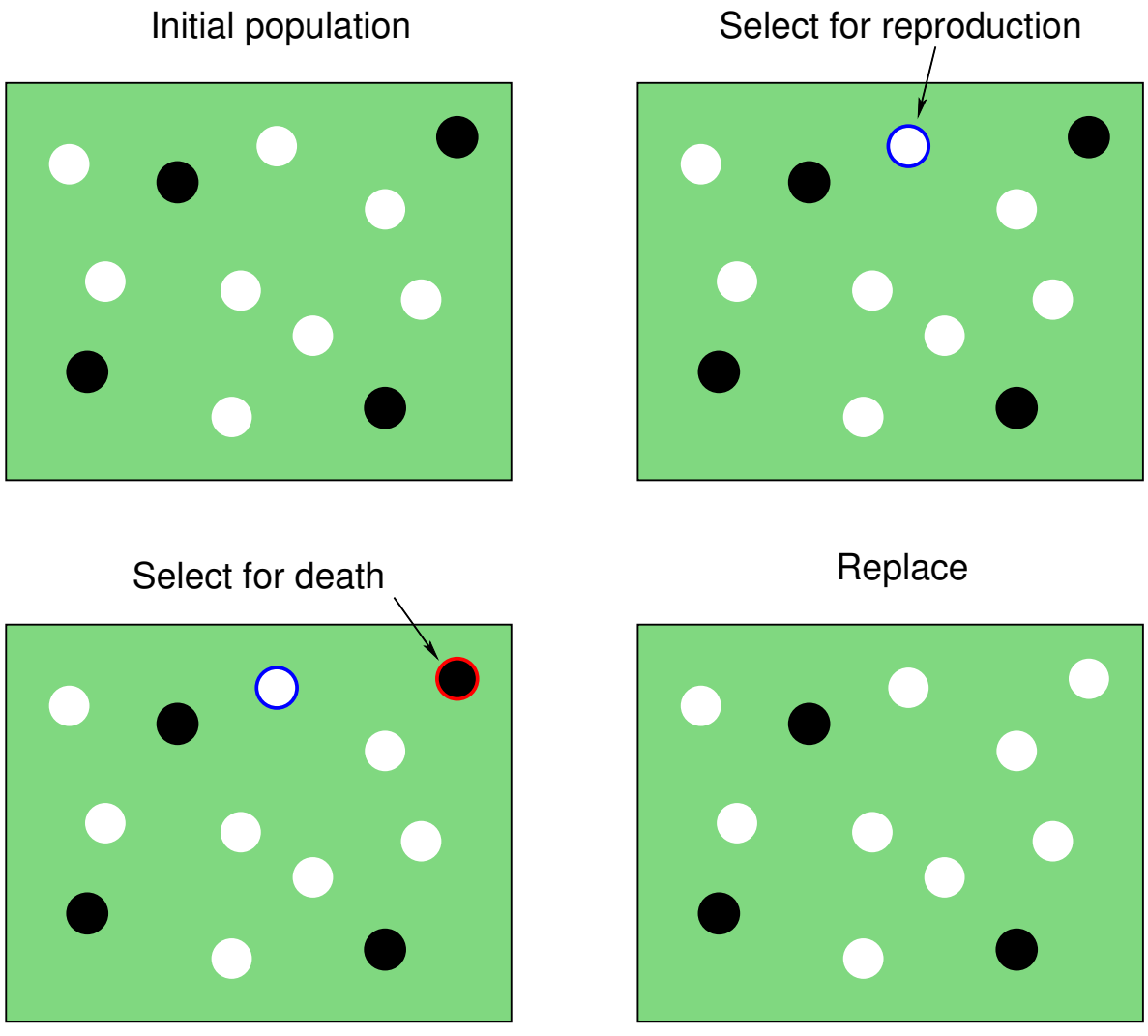
- For a new mutant, $x_0 = \frac{1}{N}$
- For a neutral mutant, $s = 0$, thus $p^{\text{fix}} = x_0 = \frac{1}{N}$
- If u is the mutation probability per genome and generation, the expected number of mutants per generations is uN
- Of these, only a fraction $\frac{1}{N}$ reaches fixation, i.e., produces a *substitution*
- Therefore the rate ν of *neutral* substitutions in a population with mutation rate u is equal to u:

$$\boxed{\text{substitution rate} = \text{mutation rate}}$$

independently of the population size N

The Moran model

Overlapping generations individual-based model:



The Moran model

- Selection: $p_{\text{kill}}(\text{A}) = 1 - s$, $p_{\text{kill}}(\text{a}) = 1$
- $\Delta t = \frac{1}{N}$; $\Delta n_{\text{A}} \in \{-1, 0, +1\}$
- Probabilities:

$$
\begin{aligned}
P_{-1} &= \underbrace{\frac{n_{\text{a}}}{N}}_{\text{Prob}_{\text{repr}}(a)} \underbrace{(1-s)\frac{n_{\text{A}}}{N}}_{\text{Prob}_{kill}(A)} \\
&= (1-s)x(1-x) \\
P_{+1} &= \frac{n_{\text{A}}}{N}\frac{n_{\text{a}}}{N} = x(1-x) \\
P_0 &= 1 - (P_{+1} + P_{-1})
\end{aligned}
$$

The Moran model

- Thus, for $\Delta t = \frac{1}{N}$, $s \ll 1$:

$$\begin{aligned} \langle \Delta n_A \rangle &= P_{+1} - P_{-1} = sx(1-x) \\ \left\langle (\Delta n_A)^2 \right\rangle &= P_{+1} + P_{-1} = (2-s)x(1-x) \simeq 2\,x(1-x) \end{aligned}$$

- The diffusion equation for the Moran model:

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}\left(sx(1-x)p\right) + \underbrace{\frac{1}{N}}_{= 1/2N \text{ for WF}} \frac{\partial^2}{\partial x^2}\left(x(1-x)p\right)$$

- The devil (or God?) is in the details...

Adaptation and drift

MUSTONEN AND LÄSSIG, 2005–2010

Finite population of size N, r alleles, Moran model. Effects of mutation and selection:

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \sum_k \Gamma_{jk} \frac{\partial \Phi}{\partial x_k}; \qquad \Phi = \langle f \rangle_x + \sum_\alpha \mu_\alpha \log x_\alpha$$

- Random drift: $\boldsymbol{x} \longrightarrow \boldsymbol{x} + \boldsymbol{\xi}$

$$\left\langle \xi^j \right\rangle_{\boldsymbol{x}} = 0; \qquad \left\langle \xi^j \xi^k \right\rangle = 2 \frac{\Gamma_{jk}(\boldsymbol{x})}{N}$$

- Fokker-Planck equation for the pdf $P(\boldsymbol{x})$:

$$\begin{aligned} \frac{\partial P}{\partial t} &= \sum_{jk} \frac{\partial}{\partial x_j} \left[-\frac{\partial \Phi}{\partial x_k} \left(\Gamma_{jk} P \right) + \frac{1}{N} \frac{\partial}{\partial x_k} \left(\Gamma_{jk} P \right) \right] \\ &= \sum_{jk} \frac{\partial}{\partial x_j} \Gamma_{jk} \left(-\frac{\partial \tilde{\Phi}}{\partial x_k} P + \frac{1}{N} \frac{\partial P}{\partial x_k} \right) \end{aligned}$$

Adaptation and drift

MUSTONEN AND LÄSSIG, 2005–2010

- $\tilde{\Phi} = \Phi - \frac{1}{N} \log \det \Gamma$; $\det \Gamma = \prod_\alpha x_\alpha$
- Stationary solution:

$$\begin{aligned} P^{\mathrm{eq}}(\boldsymbol{x}) &\propto \mathrm{e}^{N\tilde{\Phi}} = (\det \Gamma)^{-1}\, \mathrm{e}^{N\Phi} = P_0\, \mathrm{e}^{N\langle f \rangle_{\boldsymbol{x}}} \\ P_0(\boldsymbol{x}) &\propto \prod_\alpha x^{-1+N\mu_\alpha} \end{aligned}$$

- Thus, for a *static* fitness function $\boldsymbol{f}$,

$$[N \langle f \rangle_{\boldsymbol{x}}]^{\mathrm{eq}}_{\mathrm{av}} = \int \mathrm{d}\boldsymbol{x}\; P^{\mathrm{eq}}(\boldsymbol{x}) \log \frac{P^{\mathrm{eq}}(\boldsymbol{x})}{P_0(\boldsymbol{x})} = D_{\mathrm{KL}}\left(P^{\mathrm{eq}} \| \; P_0\right)$$

cAMP-response protein binding loci in *E. Coli*

MUSTONEN AND LÄSSIG, 2005

- Factor binding sites are *short* DNA sequences which bind activating factors
- Small mutation rates: $\mu N \ll 1 \Rightarrow$ Population becomes monomorphic ($\boldsymbol{x} = (x_\alpha) \to \delta_{\alpha\beta}$)

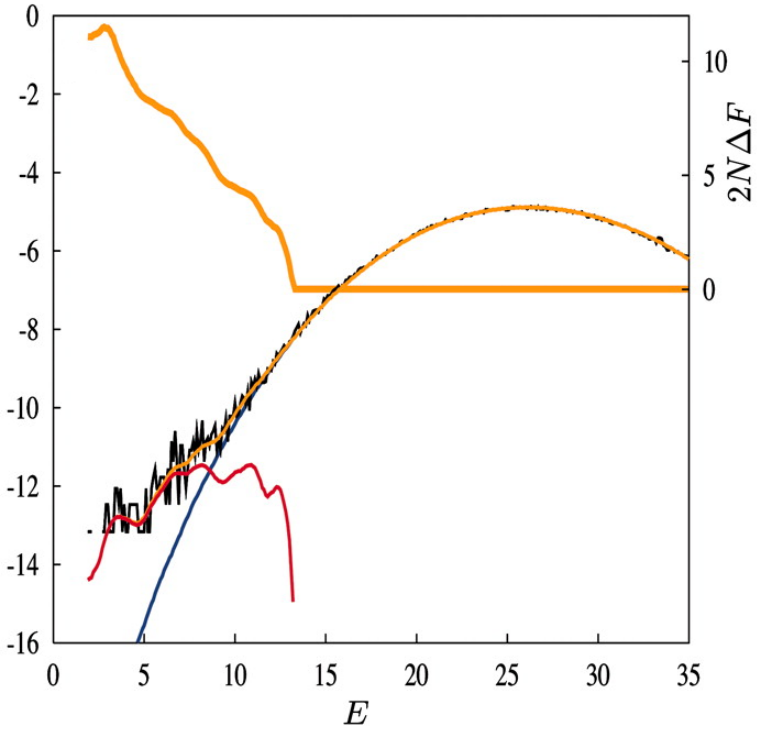
$$p_\beta = \text{Prob}\left(\boldsymbol{x} = \delta_{\alpha_\beta}\right) \propto \mathrm{e}^{Nf_\beta}$$

- It is reasonable to assume that their fitness depends on their binding energy E
- One can expect a *linear model* for $E(\sigma)$, $\sigma = (\sigma_1, \ldots, \sigma_\ell)$, $\sigma_i \in \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$

$$E(\sigma) = \sum_{i=1}^{\ell} \epsilon_i(\sigma_i) \qquad \text{with } \epsilon_i(\sigma) = \epsilon_0 \log \frac{q_i(\sigma)}{p_0(\sigma)}$$

cAMP-response protein binding loci in *E. Coli*

MUSTONEN AND LÄSSIG, 2005



Log histogram $P(E)$ of binding energy E for 520 729 CRP-binding loci in *E. Coli*. Compared with $P(E) = (1-\lambda)P_0(E) + \lambda P_0(E)\mathrm{e}^{2NF(E)}$. The inferred form of $2NF(E)$ is also plotted. (W-F model)