ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Quasispecies

Niko Beerenwinkel

Outline

- HIV infection
- Sequence space
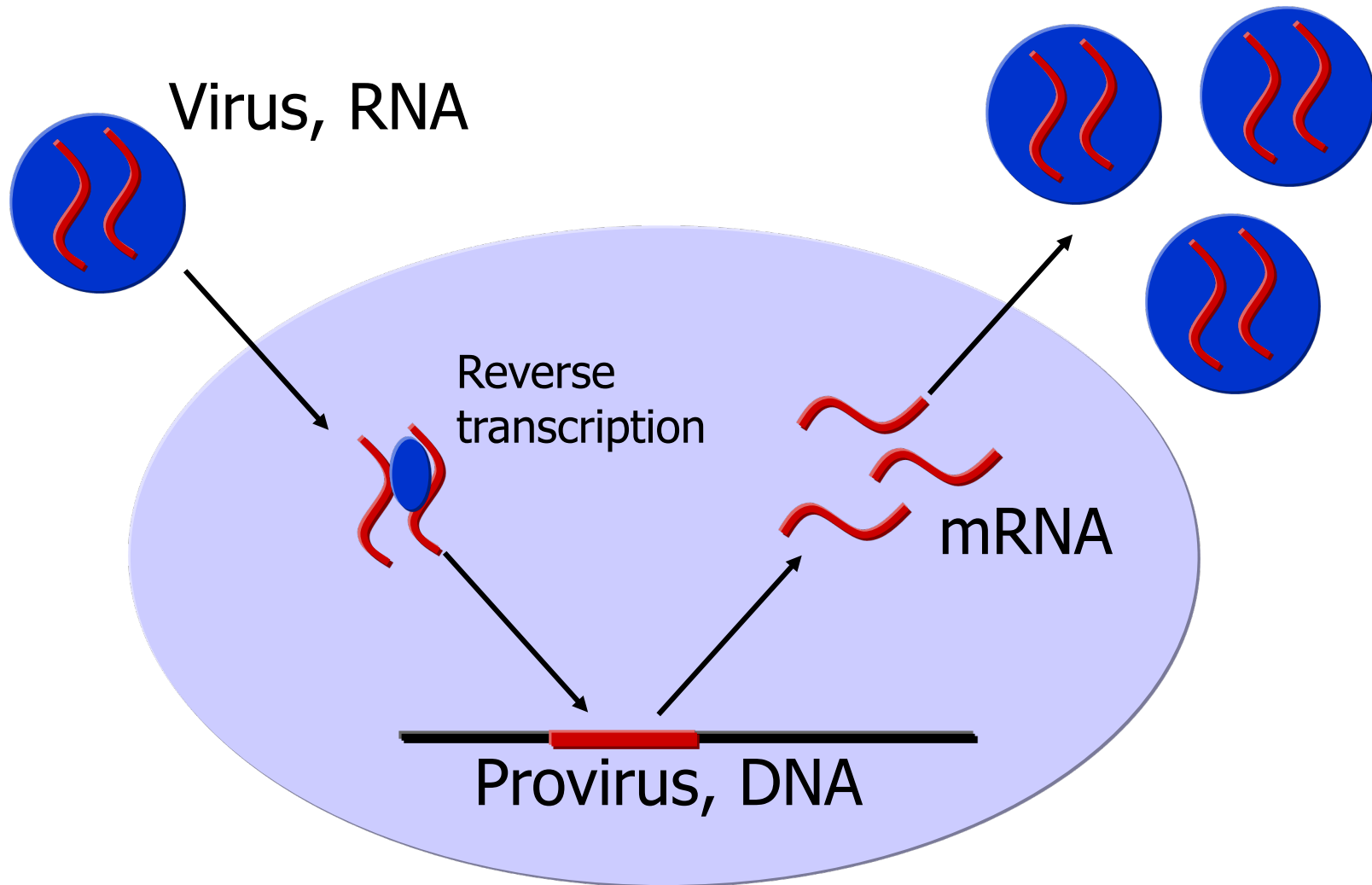- Fitness landscapes
- Quasispecies equation

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

The central dogma of molecular biology

CCT CAG GTC ACT CTT ...

CCU CAG GUC ACU CUU ...

Pro Gln Val Thr Leu ...
 P   Q   V   T   L

The genetic code

HIV infection

Virus, RNA

Reverse transcription

mRNA

Provirus, DNA

Parameters of HIV dynamics

- Short genome, ~10,000 bases
- Very high mutation rate (no proof-reading of reverse transcription), $3 \cdot 10^{-5}$ per base per replication
- Large population sizes
- Short generation time
- Exposed to strong selective forces (immune response, antiretroviral therapy)

→ extreme evolutionary dynamics, rapid adaptation

HIV evolutionary escape from selective pressure
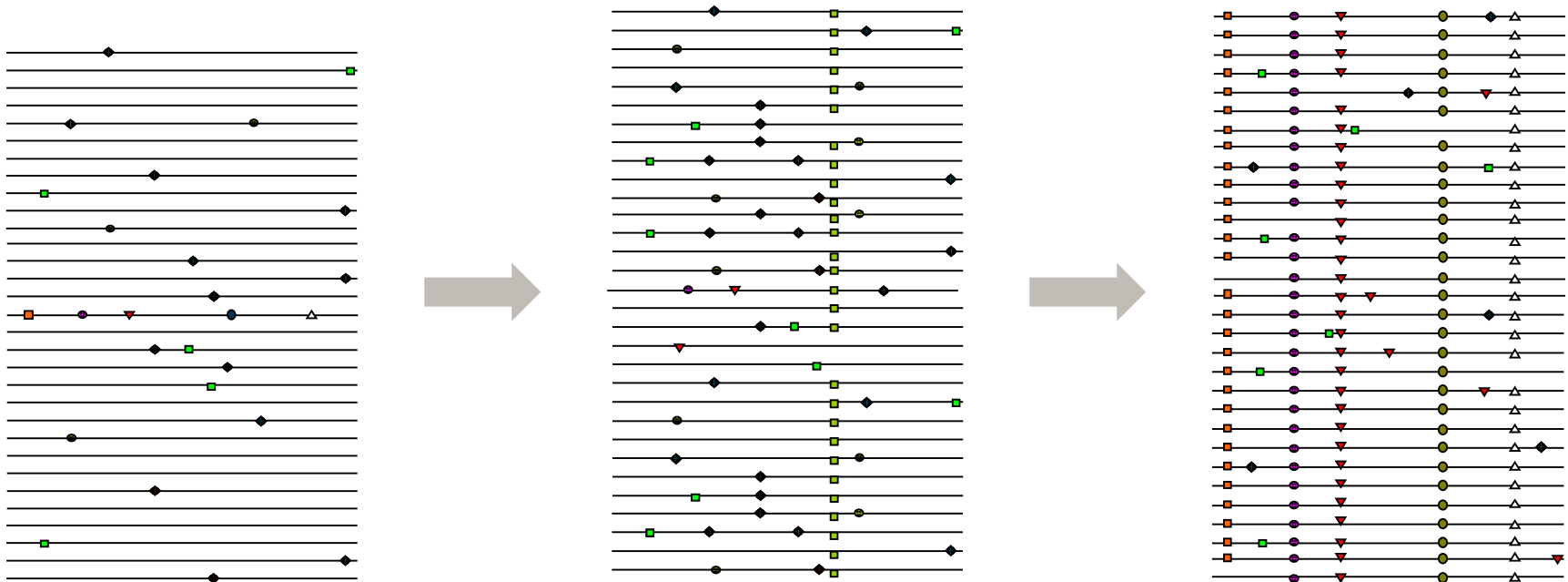
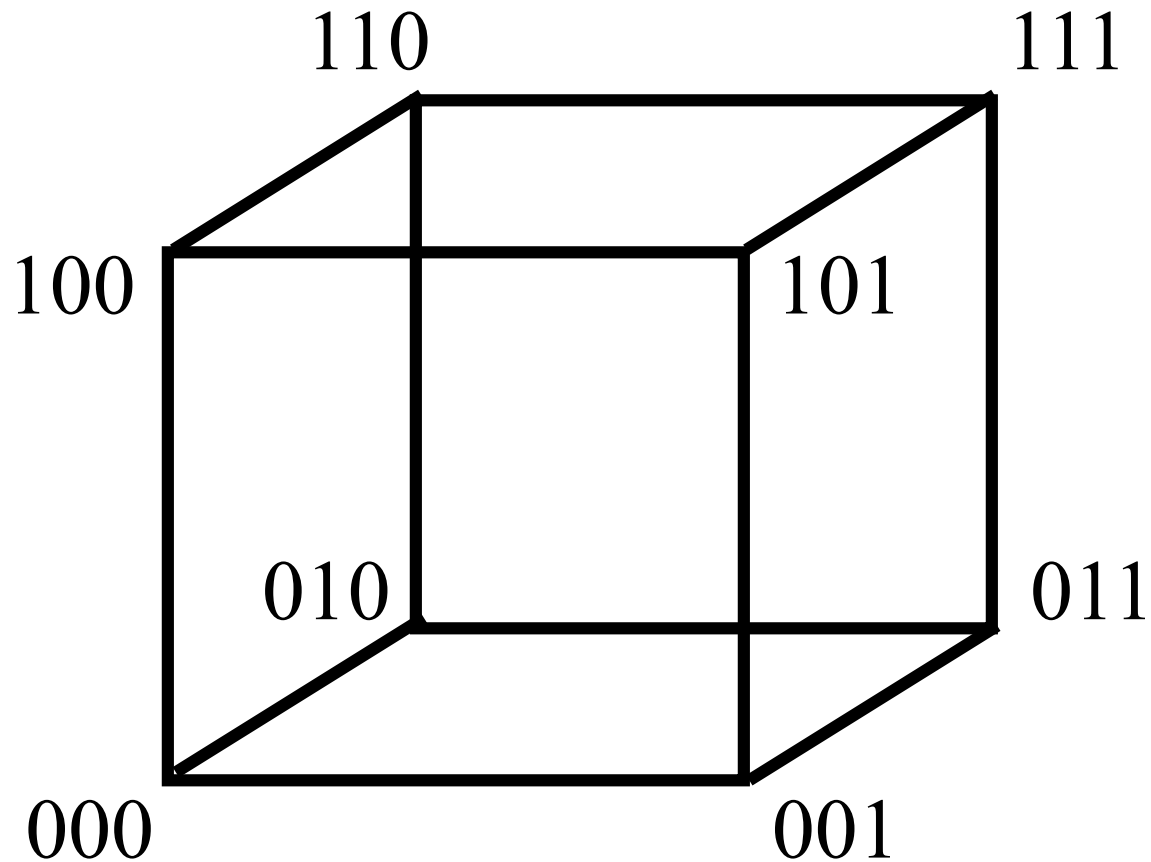Genetic progression (drug resistance development, immune escape)

Sequence space

- Alphabet, A
  - DNA: A, C, G, T
  - Protein: Ala, Arg, Asn, Asp, Cys, Glu, Gln, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val
  - Binary: 0, 1

- A *sequence* (or genome, or genotype if haploid populations are considered) of length L is an element of

$$\mathcal{A}^L = \{(a_1, \ldots, a_L) | a_i \in \mathcal{A}\}$$

- We define sequence space as the union $\mathcal{A}^\star = \cup_{L \geq 0} \mathcal{A}^L$

- Genome sizes L range from 10^4 to 10^{11}, but we often focus on a subset of loci (and characters).

Binary sequences of length *L* = 3

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
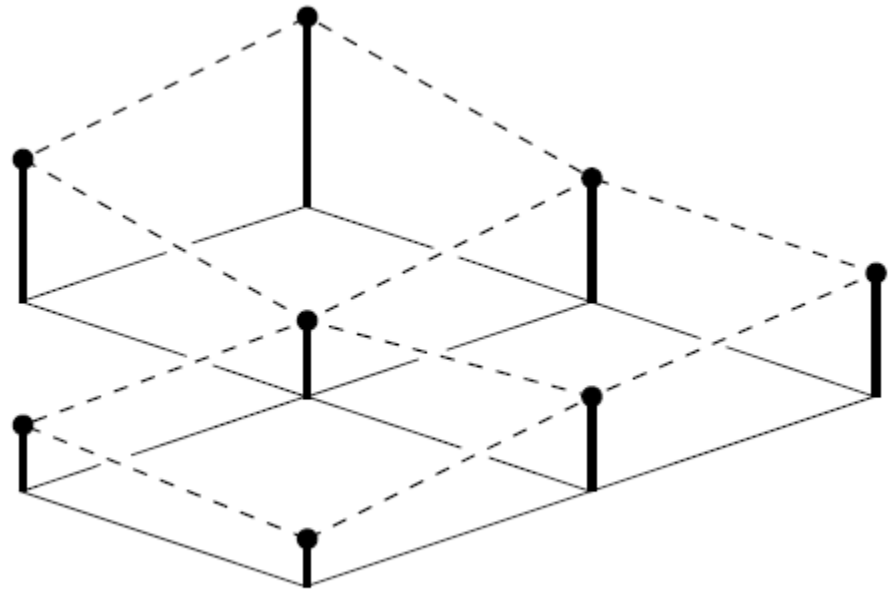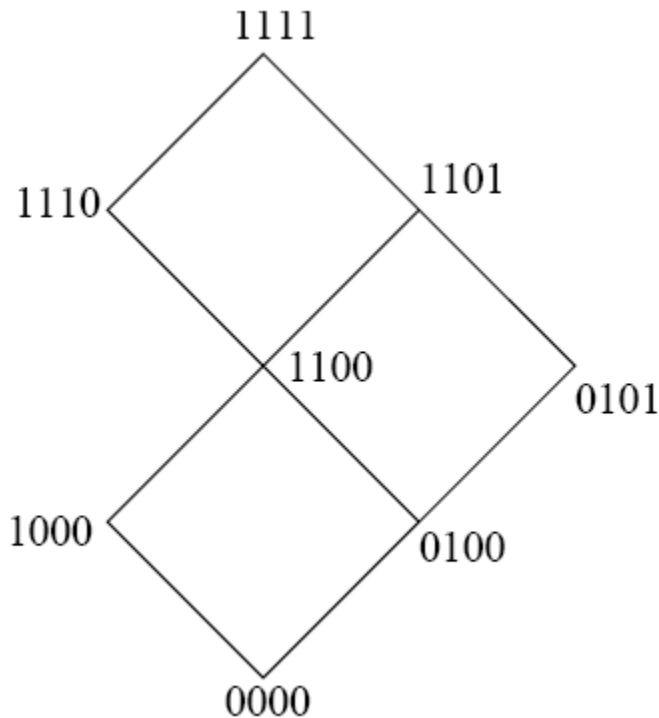Department of Biosystems
Science and Engineering
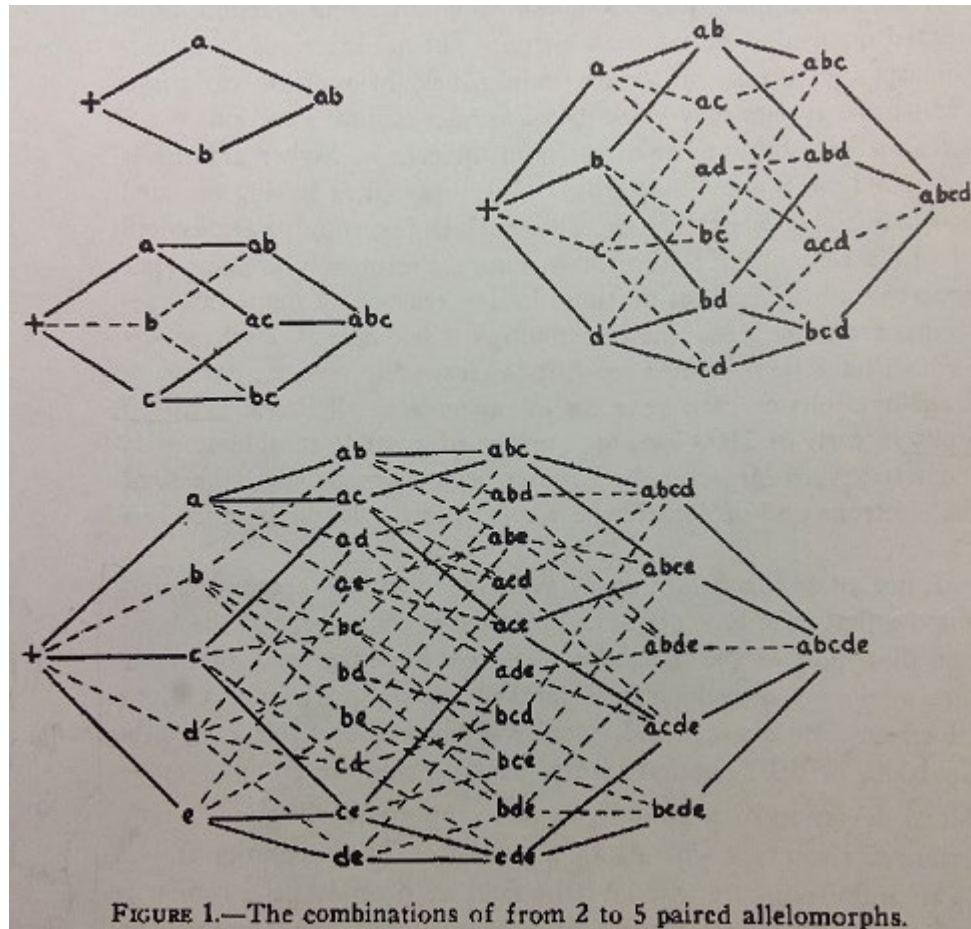
Geometry of sequence space

- Sequence space A^L has dimension L, and consists of $|A|^L$ sequences.

- Distance is measured by the Hamming distance, or Manhattan distance, i.e., the number of mismatches.

- In this metric, all sequences are close to each other: the distance between any pair of sequences is bounded by L.

- For example, if $A = \{0,1\}$ and $L = 1000$, then there are $2^{1000} \approx 10^{301}$ sequences, but the maximum length of a shortest path from any one sequence to another is only 10^3.

- Evolution is a trajectory of a population in sequence space.

- We will often call a subset $\mathcal{G} \subset \mathcal{A}^\star$ the genotype space.
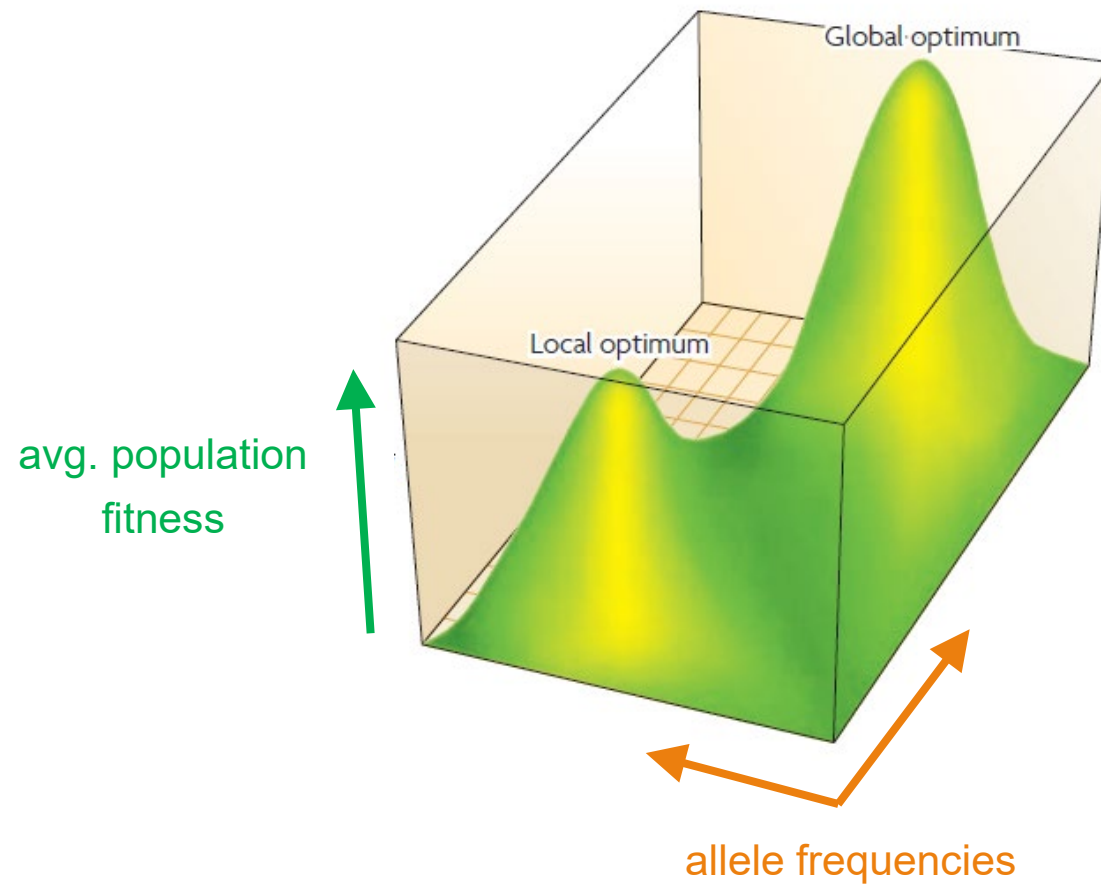
Fitness landscapes

- A (static) fitness landscape is a mapping $f : \mathbf{G} \to \mathbf{IR}$, such that $f(g)$ is the fitness of genotype g.

Fitness landscape (S. Wright, 1931)

FIGURE 1.—The combinations of from 2 to 5 paired allelomorphs.

Fitness landscape

avg. population fitness

allele frequencies

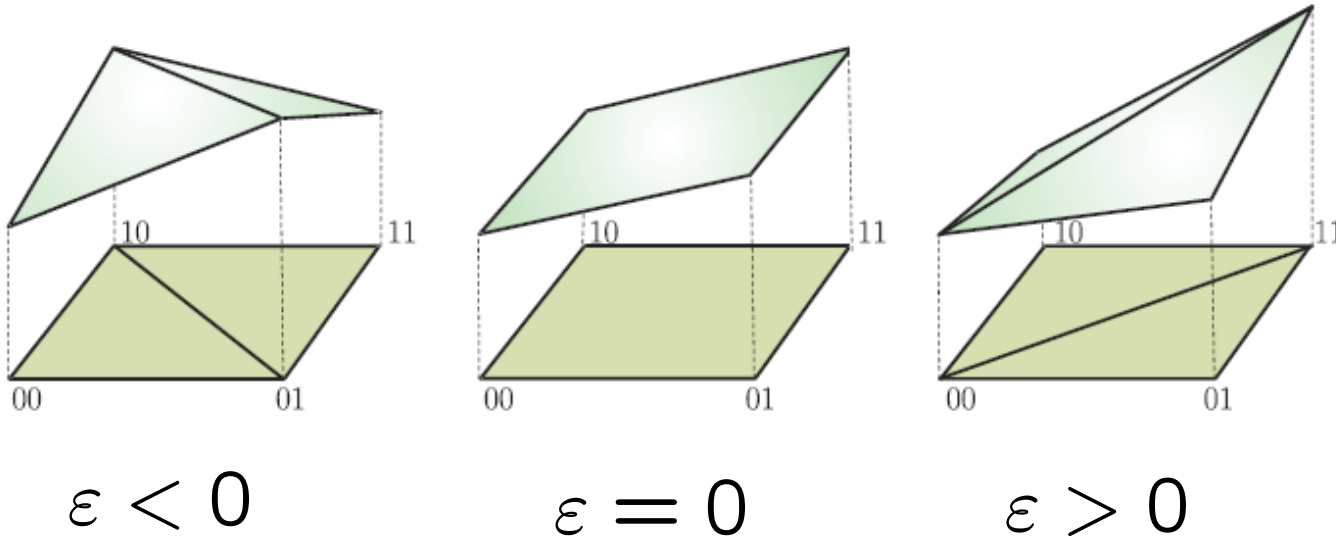# Fitness landscapes encode epistatic interactions

- Two-locus, biallelic system

$$\varepsilon < 0 \qquad \varepsilon = 0 \qquad \varepsilon > 0$$

- Epistasis:

$$\varepsilon = f_{00} + f_{11} - f_{01} - f_{10}$$

Genotype $\rightarrow$ phenotype $\rightarrow$ fitness

- More precisely, $f : \mathbf{G} \rightarrow \mathbf{P} \rightarrow \mathbf{IR}$ because fitness depends on the phenotype of the organism (size, shape, behavior, metabolism, …).

- Phenotype space is more complex (and less well defined). Phenotypes can be discrete or continuous.

- In general, fitness depends on how the individual interacts with the ecological environment and with other individuals in the population.

- In general, fitness is difficult (often impossible) to measure.

- For some viruses (e.g., HIV) and bacteria (e.g., E. coli), there are in vitro assays for measuring fitness.

The quasispecies equation

- We consider (geno-)types i = 0, 1, 2, …, n.

- Let x(t) = (x_0(t), …, x_n(t)) denote the genotype frequencies.

- Let Q = (q_{ij}) = ($q_{i \to j}$) be a mutation matrix, f = (f_0, …, f_n) a fitness landscape, and $\phi = x \cdot f$ the average fitness of the population.

- The quasispecies equation is

$$\dot{x}_i = \sum_{j=0}^{n} x_j f_j q_{ji} - \phi x_i, \quad i = 0, \dots, n$$

Properties of the quasispecies equation

$$\dot{x}_i = \sum_{j=0}^{n} x_j f_j q_{ji} - \phi x_i, \quad i = 0, \ldots, n$$

- The term "species" refers to chemical species (RNA molecule types), rather than biological species.

- If replication is error free, Q = I, then we recover the selection equation ("survival of the fittest").

- If f = (1, ..., 1), we recover the mutation equation.

- If Q is irreducible, there is a single, globally stable equilibrium x^* in the interior of the probability simplex.

- In general, the equilibrium x^* does not maximize the fitness ϕ. Mutation reduces population fitness.

Solving the quasispecies equation (1/2)

- The change of variables

$$X_i(t) = x_i(t)e^{\psi(t)} \quad \text{with } \psi(t) = \int_0^t \phi(s)ds$$

  leads to the linear ODE system

$$\dot{X}_i = \sum_{j=0}^{n} X_j f_j q_{ji}$$

  of absolute abundances X_i, because

$$X = \sum_{j=0}^{n} X_j = e^{\psi} \quad \text{and thus} \quad x_i = X_i/X$$

- The total population size X grows exponentially at rate ϕ,

$$\dot{X} = \dot{\psi}e^{\psi} = \phi X$$

Solving the quasispecies equation (2/2)

- With the mutation-selection matrix $W = (w_{ij}) = (f_j q_{ji})$ the quasispecies equation can be written in vector notation as

$$\dot{x} = xW - \phi x$$

- The equilibrium x^* is the solution of the eigenvalue problem

$$xW = \phi x$$

- The average fitness ϕ is the largest eigenvalue of the matrix W, and x^* is the corresponding (normalized) left eigenvector.
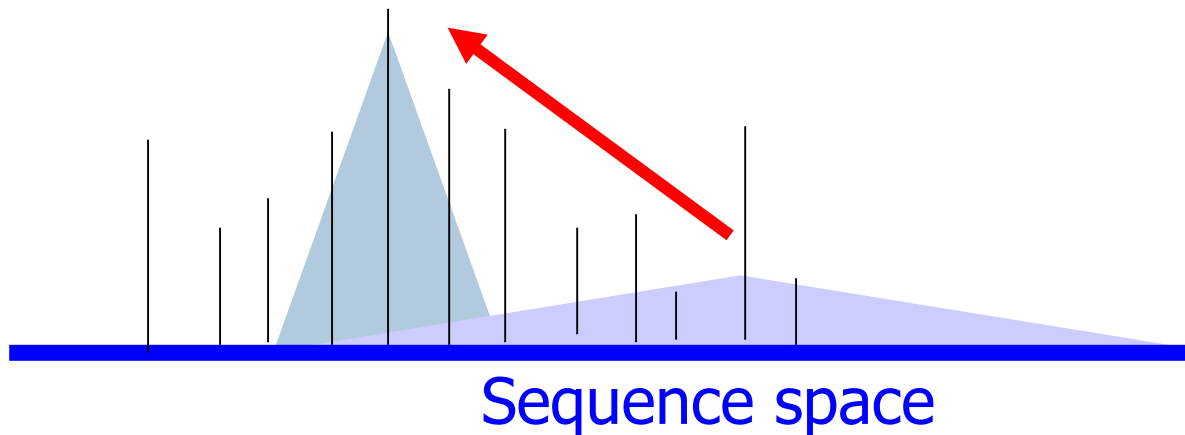
Binary sequences

- Consider binary sequences of fixed length L.

- We assume that only point mutations occur in this population (ignoring insertion, deletions, recombination) and that they are independent and occur at the same rate u.

- $q_{ij} = u^{d(i,j)}(1-u)^{L-d(i,j)}, \quad d(i,j)$ Hamming distance

- For HIV, L = 10^4, u = $3 \cdot 10^{-5}$. Thus:

  - The whole genome is copied error-free with prob. $(1-u)^L = 0.74$.

  - The copy has one (arbitrary) error with prob. $Lu(1-u)^{L-1} = 0.22$.

  - The copy has a particular error with prob. $u(1-u)^{L-1} = 2.2 \cdot 10^{-5}$.

  - With 10^9 new viruses per day, each one-point mutant occurs 22,000 times each day!

0000110011000110
0000110011000110
0000110011100110
0000110011000110
0000110011100110
0000100011000110
0100110011001110
0000110011000110
1000110011000010
0000110111000110
0000110011000110
0001110011000100
0000110011000110
….

Adaptation

- Adaptation of a population is localization in sequence space at a (local) maximum of the fitness landscape.

Sequence space

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Adaptation of a quasispecies

- The equilibrium x^* of the quasispecies equation is a *mutation-selection balance* of the population: selection drives individuals towards a local peak of the fitness landscape, while mutation generates types of lower fitness.

- Localization is no longer possible if mutation removes the fittest types faster than they are selected.

- There should be a necessary condition on the mutation rate and the selection strength for adaptation to occur.

- This condition is called the *error threshold*, and it exists for many, but not all, fitness landscapes.

- We look at the simplest possible, interesting case.

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

A simplified quasispecies equation

We consider binary sequences of length L. Type 0, the sequence 0...0 called *wild type* or *master sequence*, has fitness $f_0 > 1$. All other types have fitness 1.

The wild type is copied without error with probability $q = (1 - u)^L$. We denote by x_0 the frequency of the wild type and by x_1 the sum of the frequencies of all other types. Ignoring back mutations to the wild type, the quasispecies equation is

$$\dot{x}_0 = x_0 f_0 q - \phi x_0$$
$$\dot{x}_1 = x_0 f_0 (1 - q) + x_1 - \phi x_1$$

or equivalently, $\dot{x}_0 = x_0[f_0 q - 1 - x_0(f_0 - 1)]$. The equilibrium is

$$x_0^* = \begin{cases} \frac{f_0 q - 1}{f_0 - 1} & \text{if } f_0 q > 1 \\ 0 & \text{otherwise} \end{cases}$$

A simplified quasispecies equation

wild type

mutants

$f_0 \, q$

x_0

$f_0(1-q)$

x_1
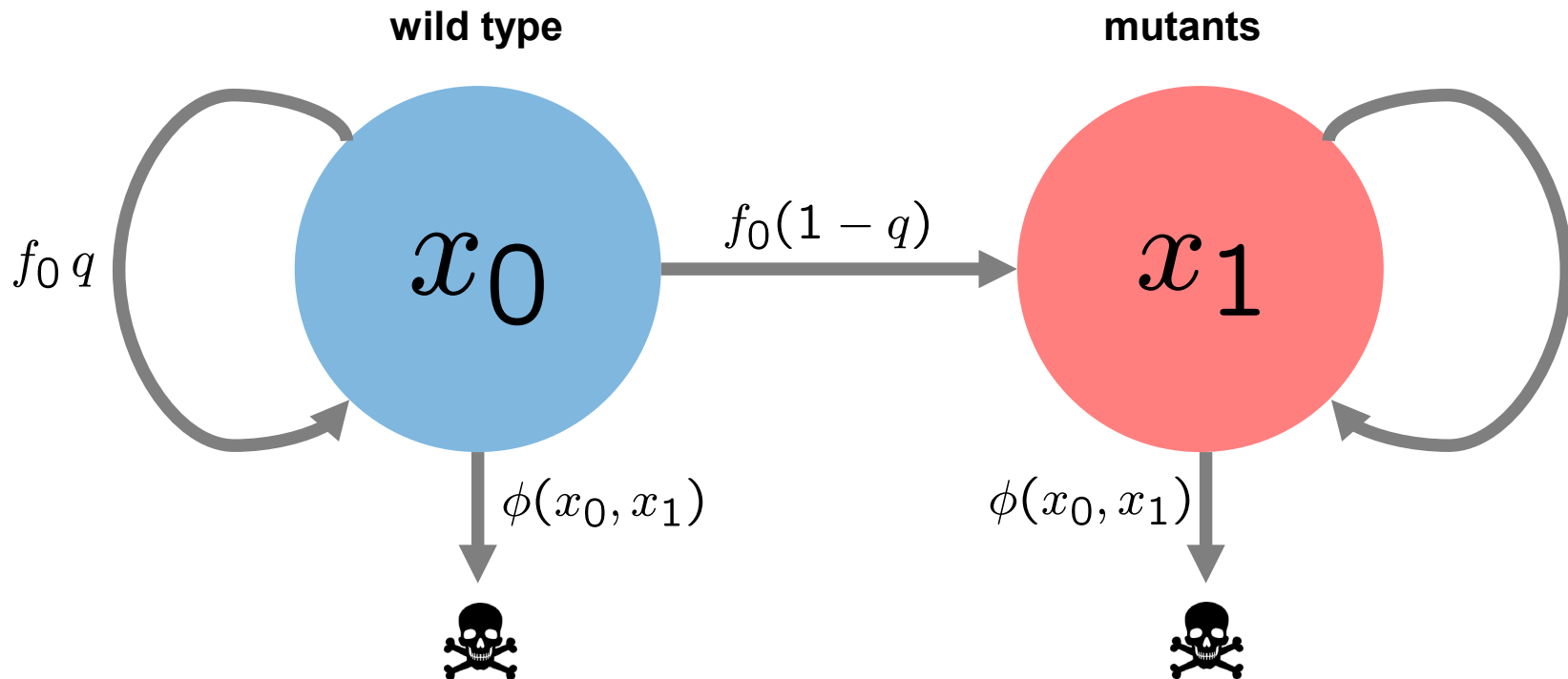
$\phi(x_0, x_1)$

$\phi(x_0, x_1)$

$$\dot{x}_0 = x_0 f_0 q - \phi x_0$$
$$\dot{x}_1 = x_0 f_0 (1-q) + x_1 - \phi x_1$$

Calculating the error threshold

The virus population can populate the wild type $(x_0^* > 0)$ only if $f_0 q > 1$. For small mutation rates, $u \ll 1$, and moderate fitness values, $\log f_0 \approx 1$, this inequality becomes

$$\log(f_0 q) = 1 + L \log(1 - u) \approx 1 - Lu > \log 1 = 0$$

or

$$uL < 1$$

where uL is the genomic mutation rate, i.e., the expected total number of mutations per replication.

The quasispecies model predicts that adaptation can occur only below the critical mutation rate $u_c := 1/L$, the *error threshold*.

Many RNA viruses appear to live close to the error threshold. For HIV, $uL = 3 \cdot 10^{-5} \times 10^4 = 0.3$.

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Per-site mutation rate versus genome size
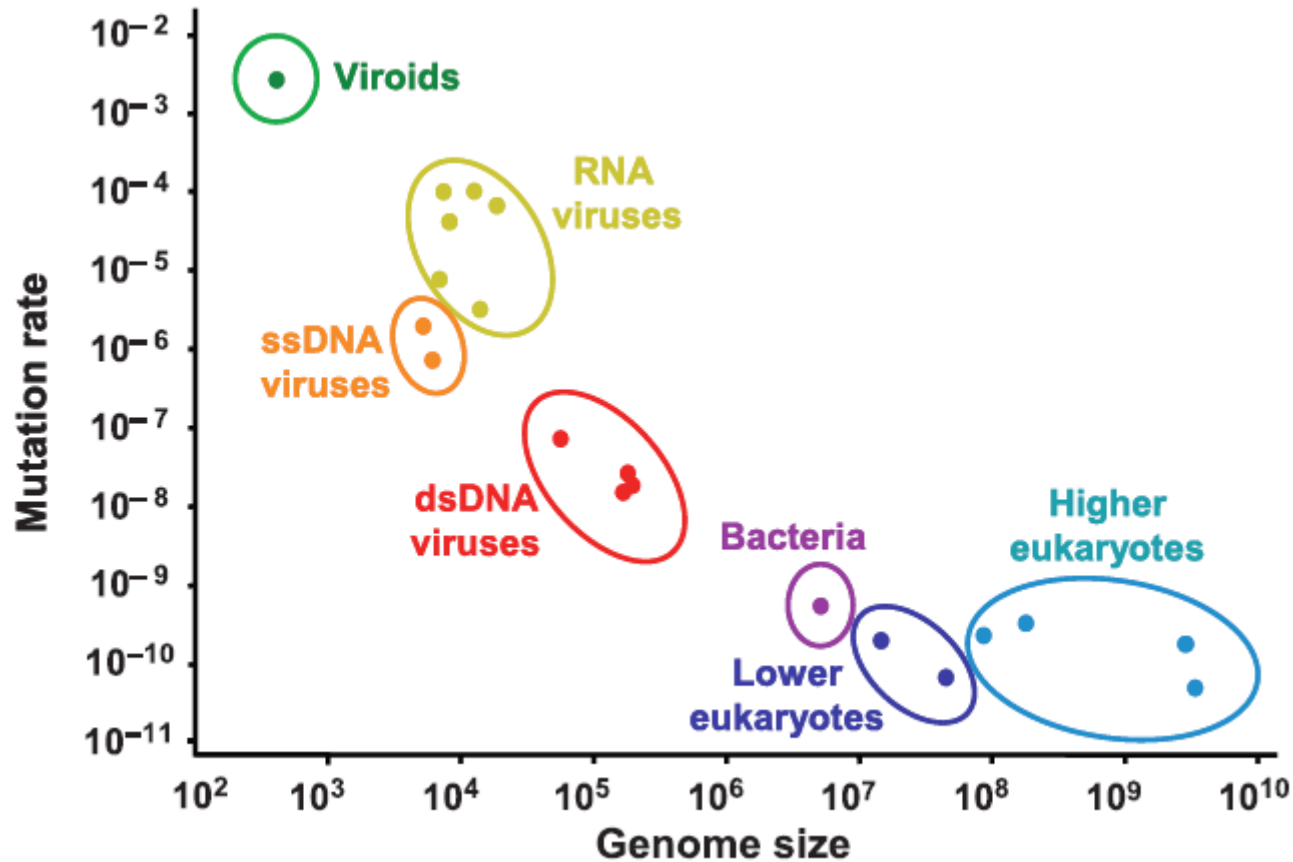
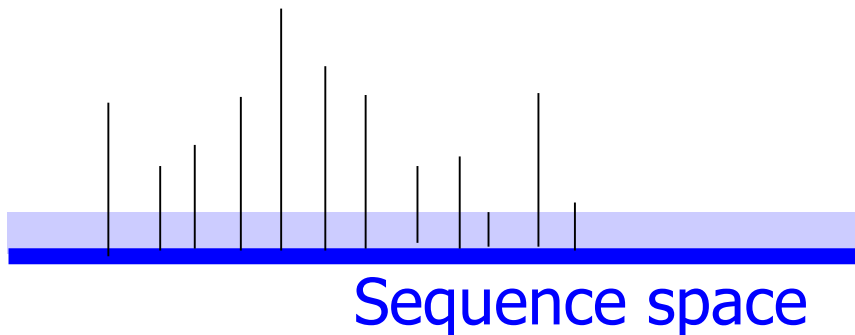
Gago et al (Science 2009)
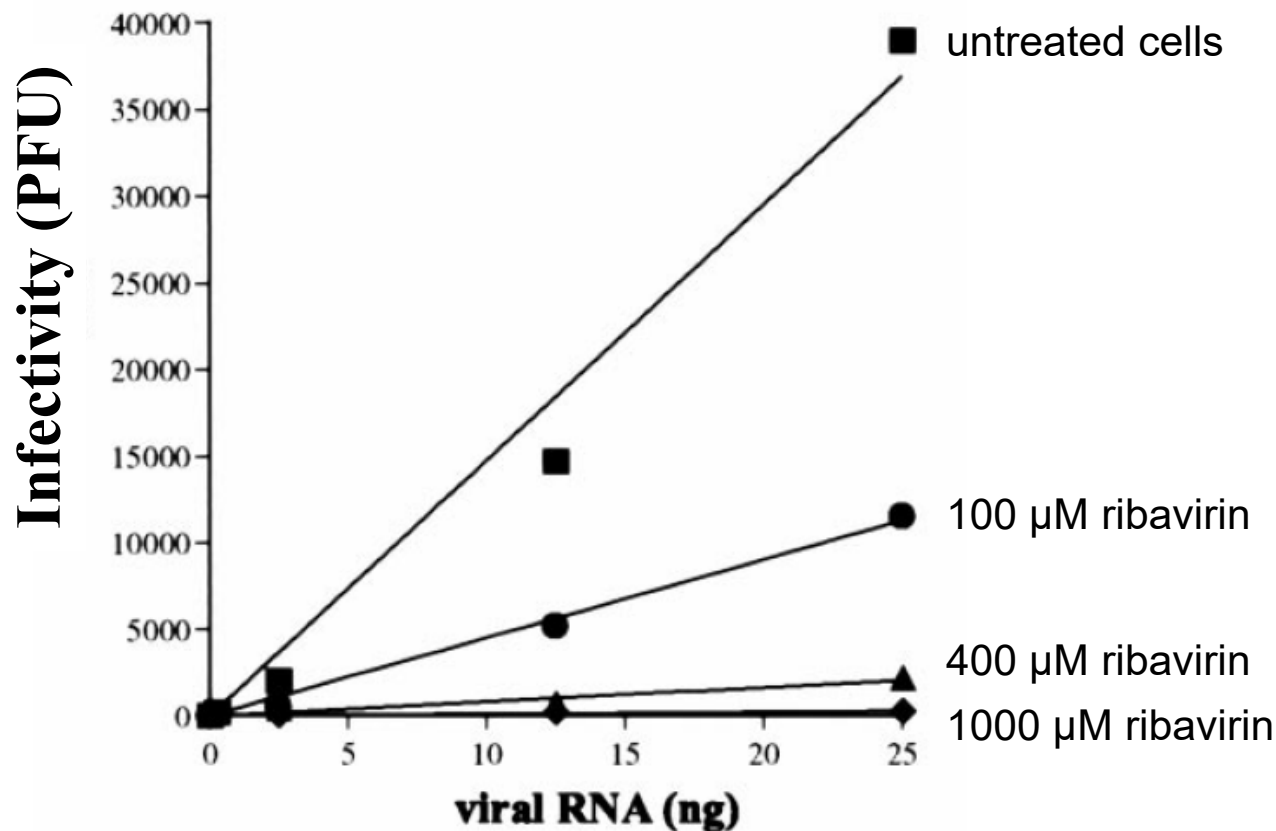
# Error threshold

$$uL < 1$$

Sequence space

$$uL > 1$$

Sequence space
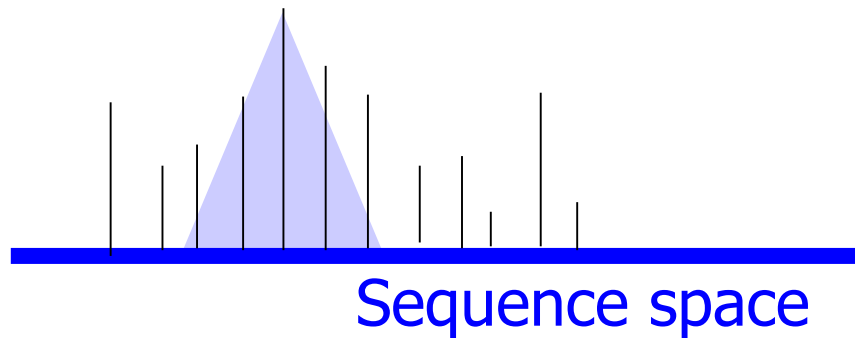
- The case uL > 1, is known as *mutational meltdown*.
- Some antiretroviral drugs seem to work by increasing u.

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Molecular evidence for the error catastrophe

- Poliovirus and ribavirin [Crotty et al., PNAS 2001]
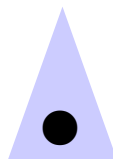
Selection of the quasispecies

$$u < u_1$$

$$u_1 < u < u_2$$

$$u_2 < u$$
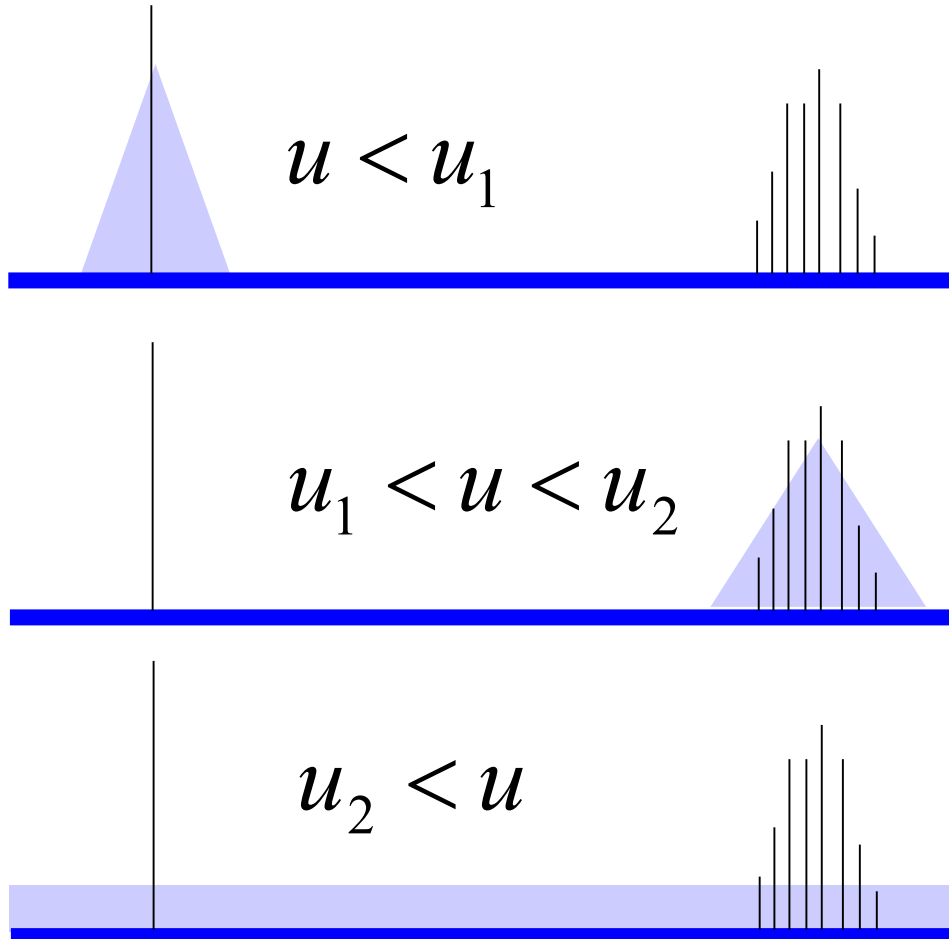
- Selection can favor low and wide peaks over high and narrow peaks.

- For any given mutation rate, the equilibrium distribution (the quasispecies) with maximum average fitness is selected.
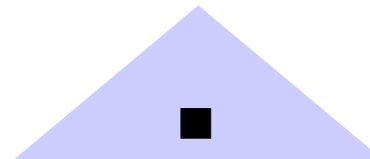
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Genomic diversity in quasispecies is critical for viral tissue tropism and pathogenesis
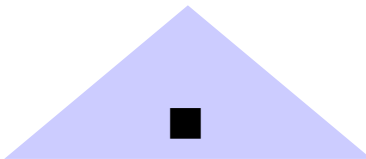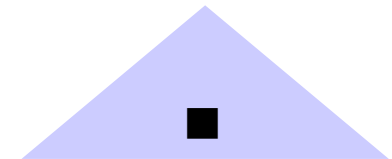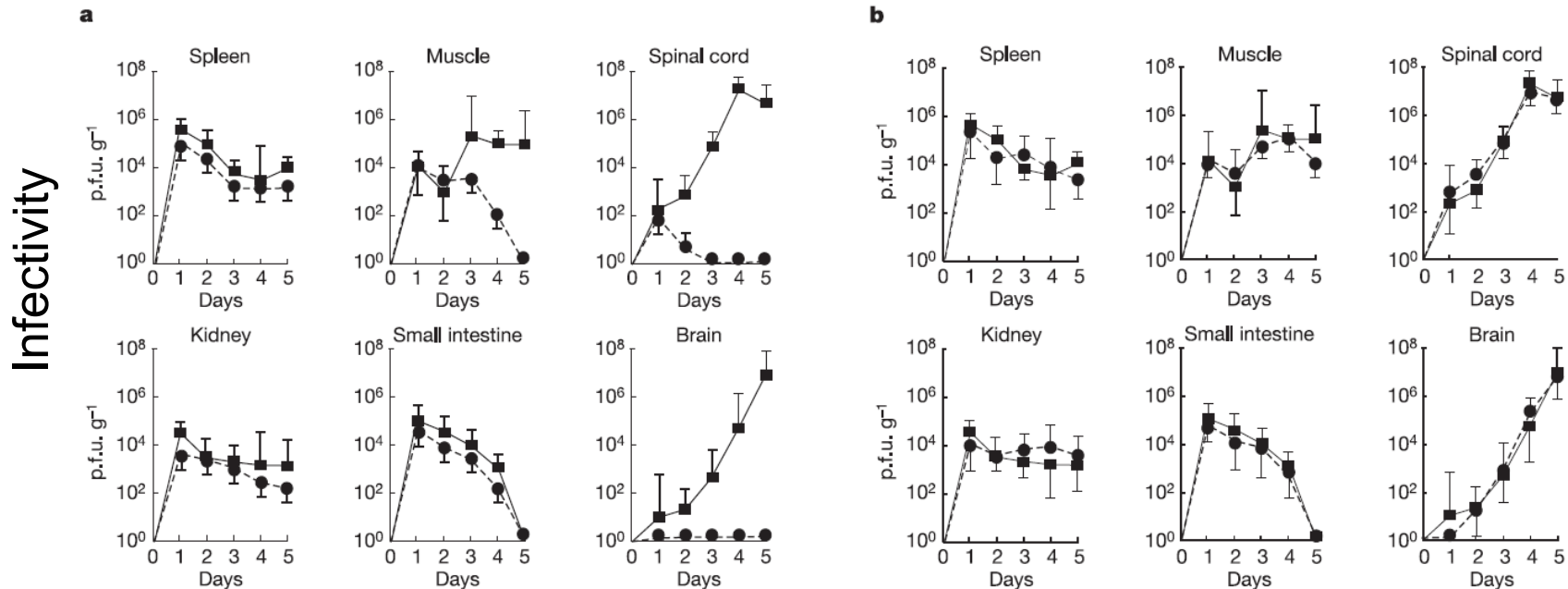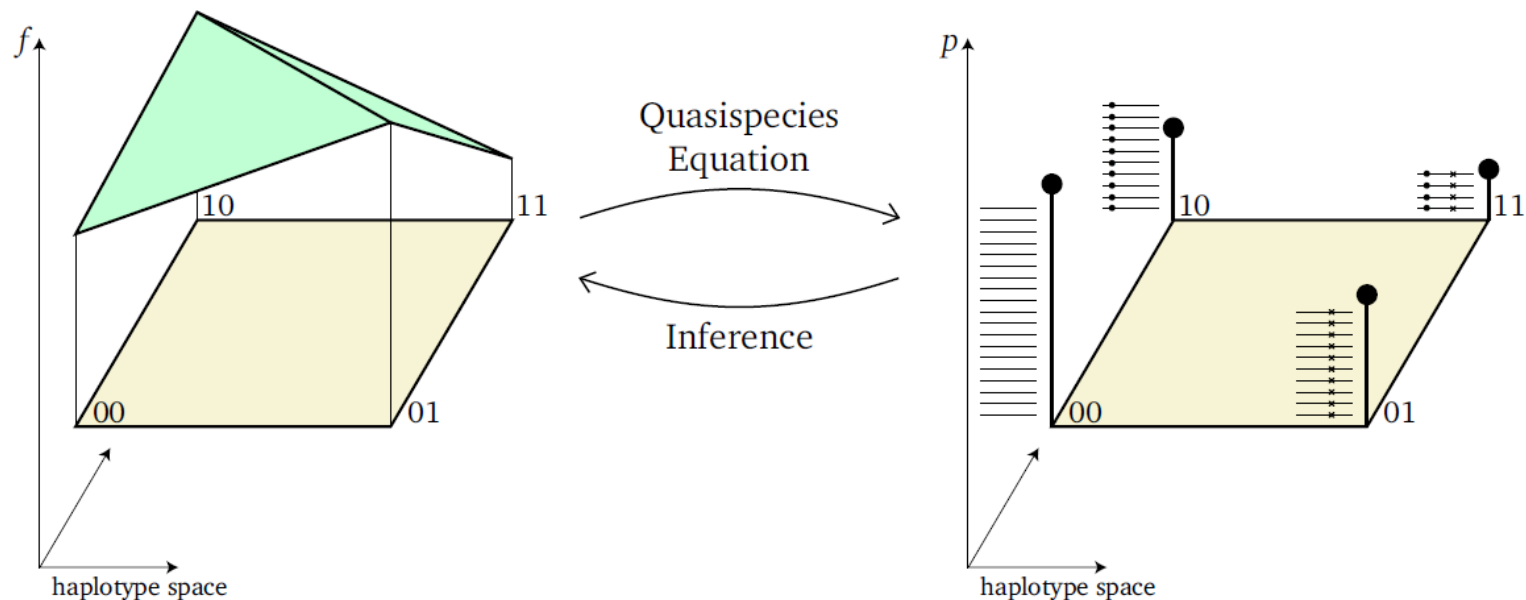

Vignuzzi et al., Nature 2006

Estimating viral fitness from NGS data

- Assuming the quasispecies model and the viral population to be in equilibrium, fitness can be estimated from observed haplotype frequencies:

Fitness landscape   Probability simplex

Summary

- Evolution can be thought of as a time trajectory in the probability simplex.

- The quasispecies equation describes a mutation-selection balance. It predicts an error threshold, a critical mutation rate beyond which genetic information can not be maintained.

- The quasispecies distribution can be calculated by solving an eigenvalue problem.

- Quasispecies theory has been applied to RNA viruses (and to tumor cell populations).

References

- Nowak M. Evolutionary Dynamics.