EVOLUTIONARY TRAJECTORIES IN RUGGED FITNESS LANDSCAPES

Kavita Jain, Weizmann Institute

(Work done in collaboration with J. Krug, Köln University)

Outline

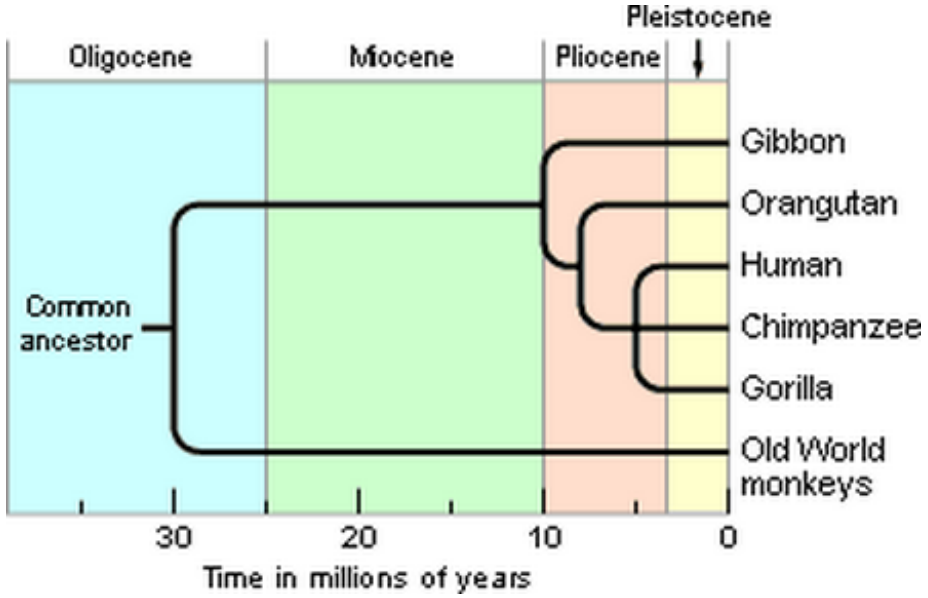
- Set the stage: Evolution, Fitness, Phenotype, Genotype
- Define the model in genotype space
- Phase transition in the steady state
- Dynamics of the model in the ordered phase

Macroevolution

Studies of DNA and proteins of various species has shown that over a period of billions of years,

$$\text{Simple molecules} \xrightarrow{10^9 \text{yrs}} \text{Complex multicellular organisms}$$

During this process, extinction and speciation has occurred.



(but apparently the word has not reached Kansas)

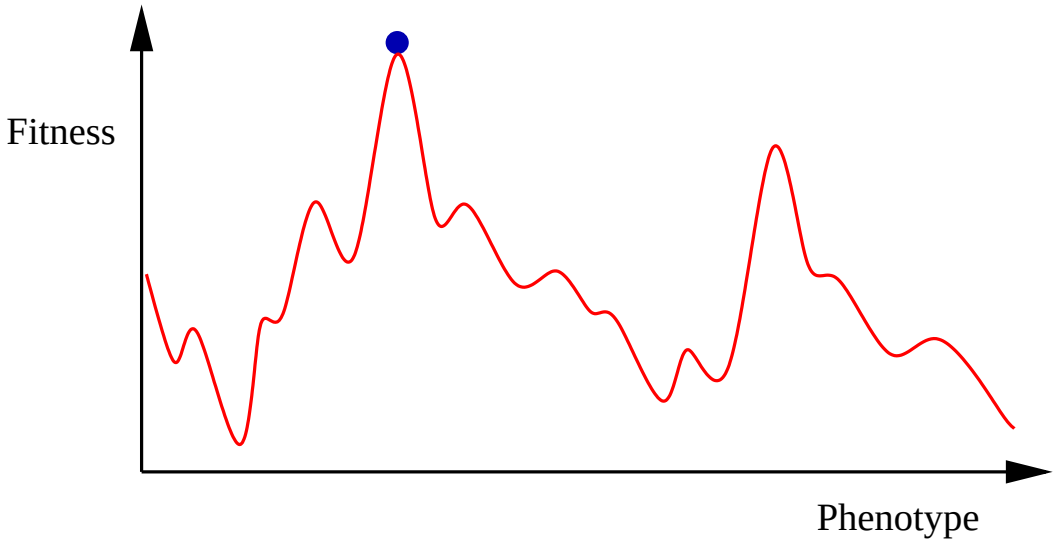
Microevolution

- Consider a species well “adapted” to an environment. If the external conditions are altered, it will evolve.

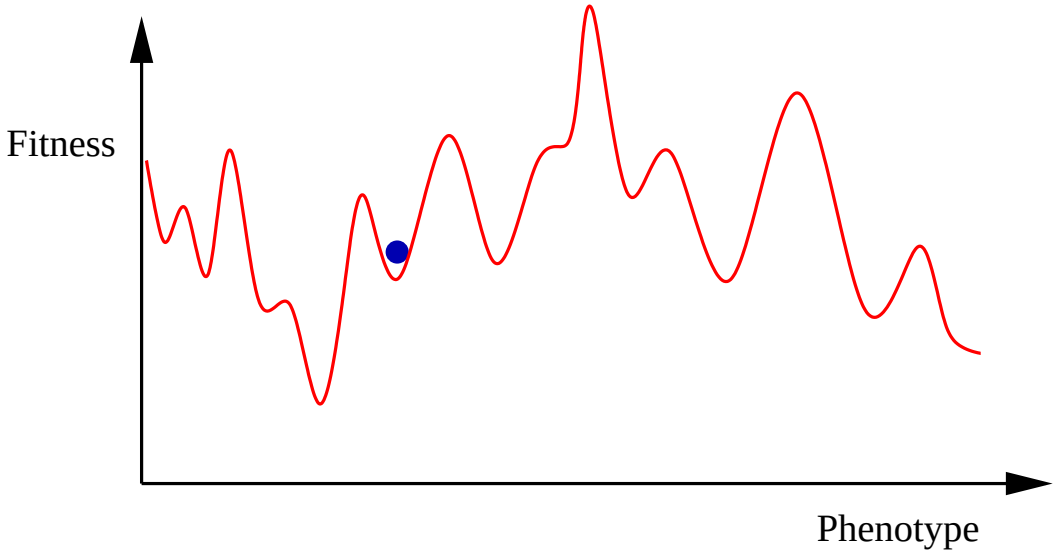
- A measure of adaptation is the fitness or reproductive success. A well adapted population has high fitness.

- But the selection acts on the phenotypic traits such as cell size, ability to infect etc. For e.g., a virus with high infectivity has a better chance of leaving progeny.

A population well-adapted to a given environment resides at the optimum of the fitness landscape.



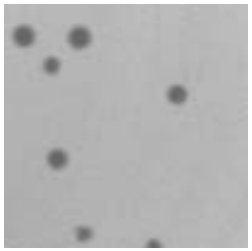
Due to a change in the environment, it finds itself in a fitness valley.



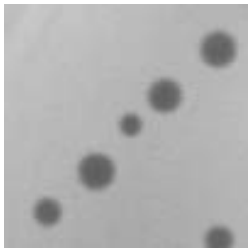
Measuring Fitness

In experiments, the replication rate of a species is a measure of the fitness.

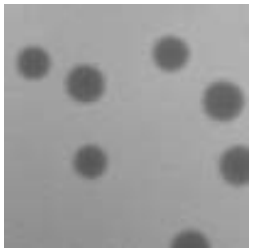
For e.g., size of the viral plaques on bacterial lawn (Burch and Chao, 1999).



$t = 0$

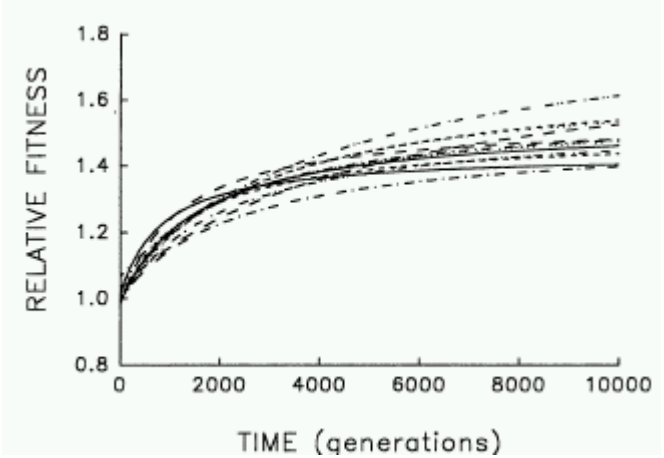


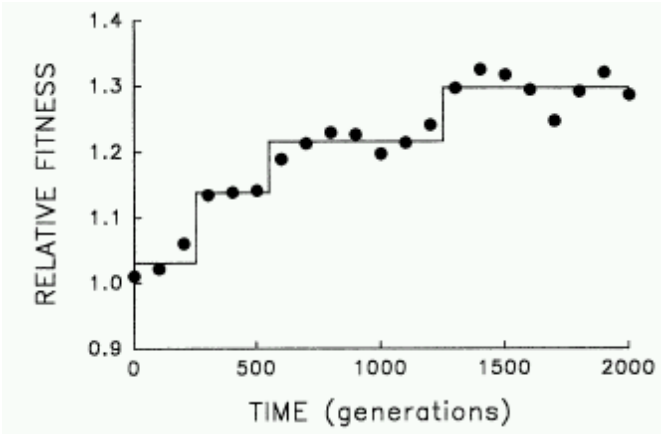
$t = 50$



$t = 100$

The fitness of a starved bacterial colony increases with time (Lenski and Travisano, 1994).





The information is coded in the genome ...

- Genes are passed from one generation to the next.
- Phenotype is a function of the genotype.

However, phenotype-genotype mapping is not known except for few cases. For e.g., with a RNA sequence (Genotype), a planar structure (Phenotype) that minimises energy (Fitness) can be found (Fontana *et al.*, 1993).

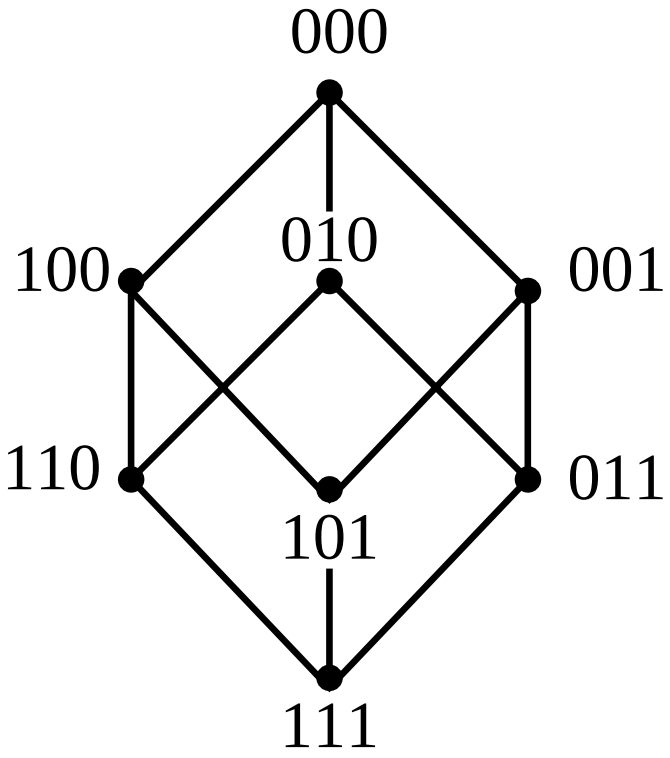
A microscopic theory of evolution works with RNA/DNA sequence $\sigma \equiv \{\sigma_1, ..., \sigma_N\}$, $\sigma_i =$ A, U/T, C, G.

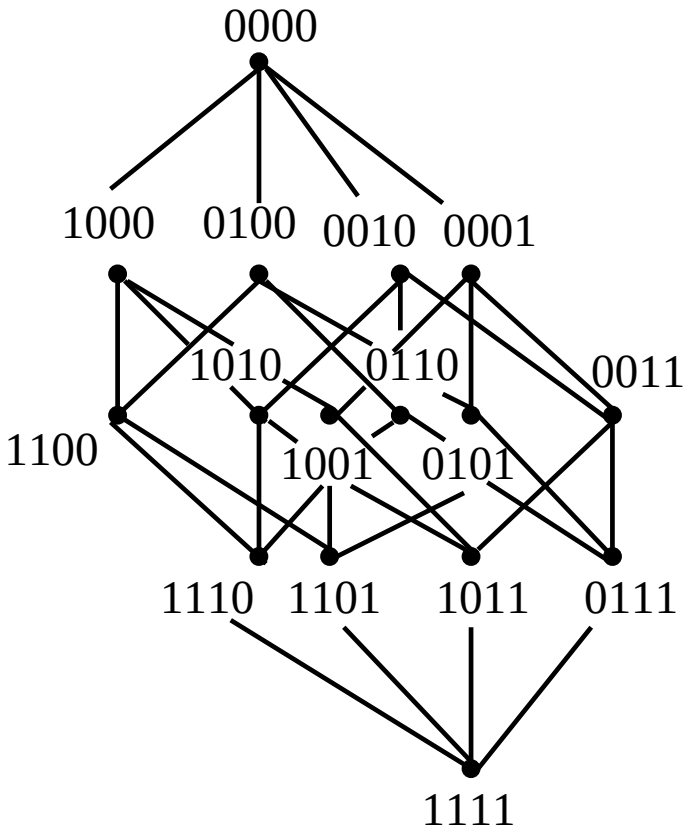
Genome length N of various organisms (Drake *et al.*, 1998) :

RNA virus	*E. Coli*	*C. Elegans*	Mouse	Human
10^3-10^4	4.6 $\times 10^6$	8.0 $\times 10^7$	2.7 $\times 10^9$	3.2 $\times 10^9$

Sequence Space

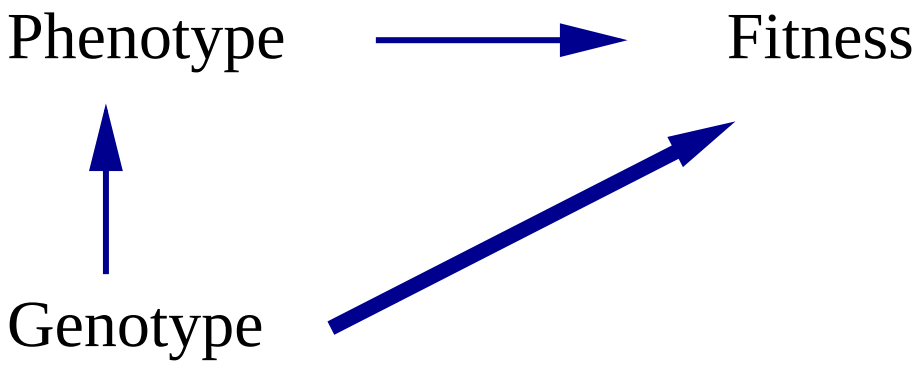
The 2^N binary sequences can be arranged on the Hamming space with Hamming distance $d(\sigma, \sigma') = \sum_{i=1}^{N}(1 - \delta_{\sigma_i, \sigma_i'})$.





Fitness Landscape in Sequence Space

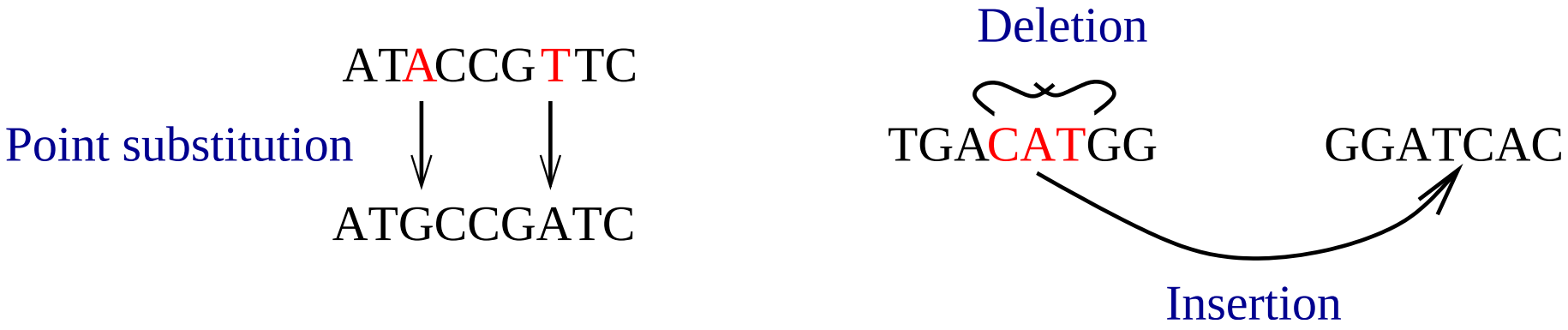
In principle:



In practice:

- With sequence σ, assign an i.i.d. variable W_σ chosen from $p(W)$.
- The resulting landscape is rugged which is consistent with experiments.
- Due to reproduction: Population(σ,t+1)=W_σ Population(σ,t)

But Reproduction is Error Prone ...



Mutation rates for various organisms (Drake *et al.*, 1998) :

RNA virus	*E. Coli*	*C. Elegans*	Mouse	Human
10^{-3}-10^{-4}	5.4×10^{-10}	2.3×10^{-10}	1.8×10^{-10}	5.0×10^{-11}

We will consider point mutations occurring with probability

$$p_{\sigma \leftrightarrow \sigma'} = \mu^{d(\sigma,\sigma')}(1-\mu)^{N-d(\sigma,\sigma')}$$

Eigen's Model

Selection localises population, while mutation delocalises it.

One may anticipate a phase transition !!

That this indeed is the case is captured by a class of deterministic models :

$$X_\sigma(t+1) = \frac{\sum_{\sigma'} p_{\sigma \leftarrow \sigma'} W_{\sigma'} X_{\sigma'}(t)}{\sum_{\sigma'} W_{\sigma'} X_{\sigma'}(t)}$$

where $X_\sigma(t)$ is the average fraction of type σ at time t (Eigen, 1971).

The Model Applies ...

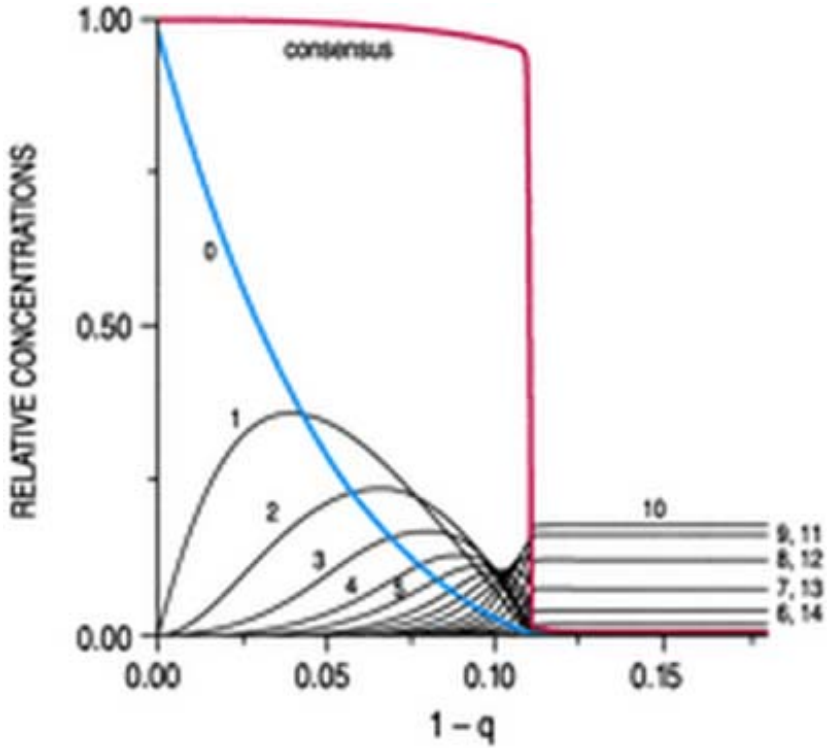
- When the population reproduces asexually as is the case for microbes.
- When the population M scales with the volume of the sequence space :

$$X_\sigma(0) = \delta_{\sigma,\sigma^{(0)}}, X_\sigma(1) \sim \mu^{d(\sigma,\sigma^{(0)})}$$

The fraction $X_\sigma(1)$ is detectable if $\mu^N \geq 1/M$ (infinite population limit).

Phase Transition

Consider single peak landscape: $W(\sigma) = W_0\delta_{\sigma,0} + (1 - \delta_{\sigma,0})$



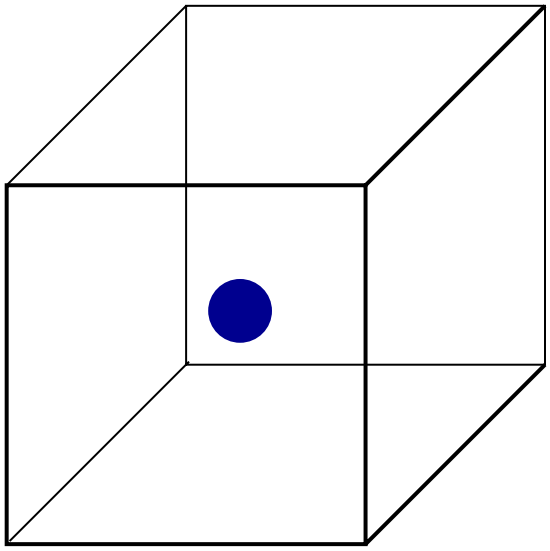
(Tarazona, 1992)

For $\mu \to 0, N \to \infty$, μN fixed,

$$X_0 = \frac{W_0 e^{-\mu N} - 1}{W_0 - 1}$$

Phase transition at $\mu_c = \ln W_0 / N$

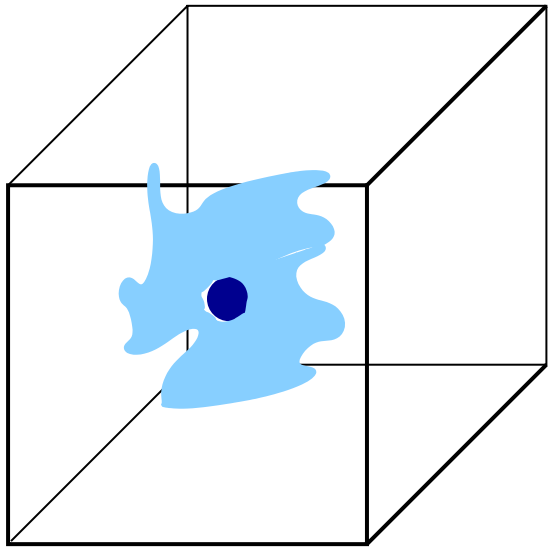
Schematically ...



$\mu = 0$

"Survival of the fittest"

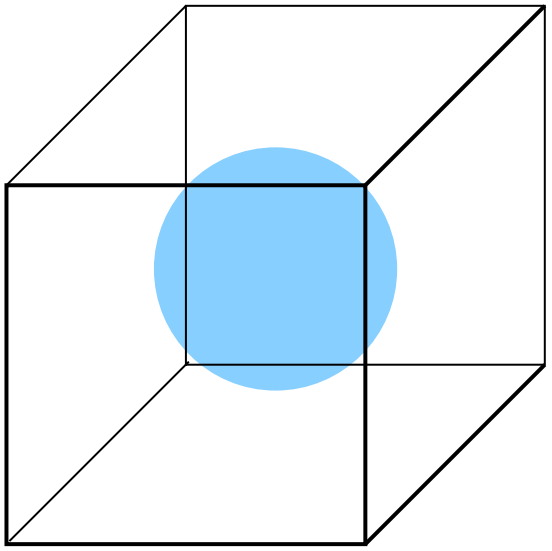
All the population at the master sequence



$0 < \mu < \mu_c$

Quasispecies

Closely related mutants centred about the master sequence



$\mu \geq \mu_c$

Delocalised population

Mutants all over the sequence space

Quasispecies: some remarks

- Non-Darwinian concept.

- The extremely heterogeneous makeup of the quasispecies has been seen in experiments. In Qβ phage, only 14$\%$ of the population was found to be wild-type (Domingo *et al.*, 1978).

- Viral diseases like common cold and AIDS are hard to tackle due to this reason, and new antiviral strategies have been proposed.

- Unlike in the condensate phase in driven-diffusive systems, the density is not spread all over the Hamming space (for e.g., upto 4 mutants in Qβ phage of genome size $\sim 10^3$).

So Far:

- Defined a mutation-selection model.
- Steady state has a localising-delocalising phase transition.

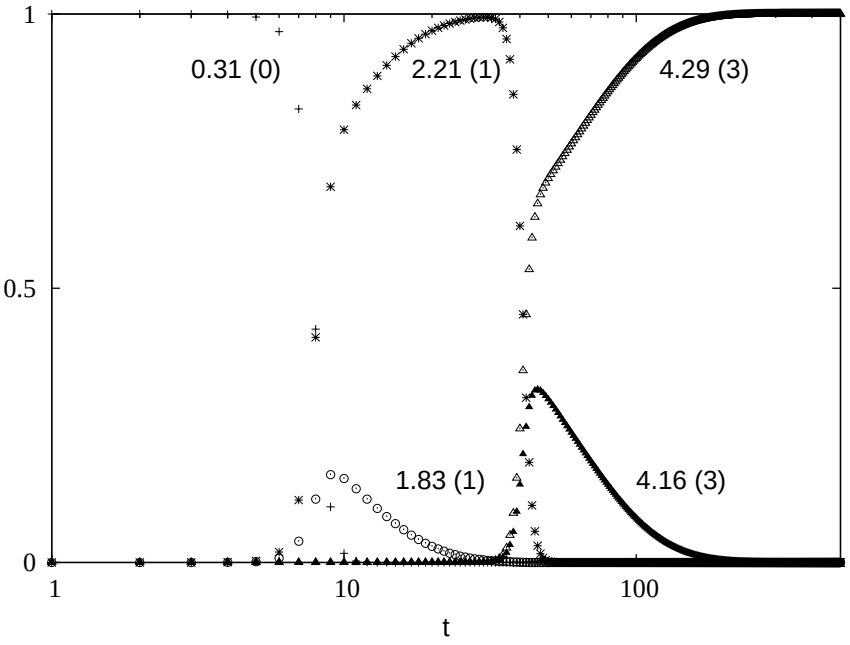
Now:

- What are the dynamics of the evolutionary process?

Numerical Iteration

- Start with a randomly chosen $\sigma^{(0)}$.
- Since the population is infinite, all mutants are present at $t = 1$!!
- Mutants better than the parent grow faster; parent population drops.
- Process repeats until the fittest is found.
- Mutants with large Hamming distance appear later and vice versa. Some mutant classes do not appear at all.

Which is the most populated sequence at t?



$$N = 6, \mu = 10^{-6},$$
$$p(W) = e^{-W}$$

Linear Equation

$$X_\sigma(t+1) = \frac{\sum_{\sigma'} p_{\sigma\leftarrow\sigma'} W_{\sigma'} X_{\sigma'}(t)}{\sum_{\sigma'} W_{\sigma'} X_{\sigma'}(t)}$$

Define the unnormalised variable Z through

$$X_\sigma(t) = \frac{Z_\sigma(t)}{\sum_{\sigma'} Z'_\sigma(t)}$$

which obeys

$$Z_\sigma(t+1) = \sum_{\sigma'} p_{\sigma\leftarrow\sigma'} W_{\sigma'} Z_{\sigma'}(t)$$

Thus, a linear equation is obtained which can be diagonalised with the initial condition $Z_\sigma(0) = \delta_{\sigma,\sigma^{(0)}}$.

Random Slope Model (Krug and Karl, 2003)

We first note that

$$Z_\sigma(1) = e^{-|\ln \mu| d(\sigma,\sigma^{(0)})}\ W_{\sigma^{(0)}}$$

- all mutants get available immediately
- concentration depends only on the distance from $\sigma^{(0)}$
- higher the distance, smaller is the population

For mutants with sufficiently high fitness, now turn off the mutations :

$$Z_\sigma(t) = Z_\sigma(1)\ e^{(t-1)\ln W_\sigma} \text{ for } t > 1$$

Random Slope Model (Contd.)

Taking logarithms on both sides :

$$E_\sigma(t) = -d(\sigma, \sigma^{(0)}) + t\ F_\sigma$$

We have $\binom{N}{d}$ lines with random slopes at intercept $-d$. For purposes of σ^*, only the best amongst these matter.

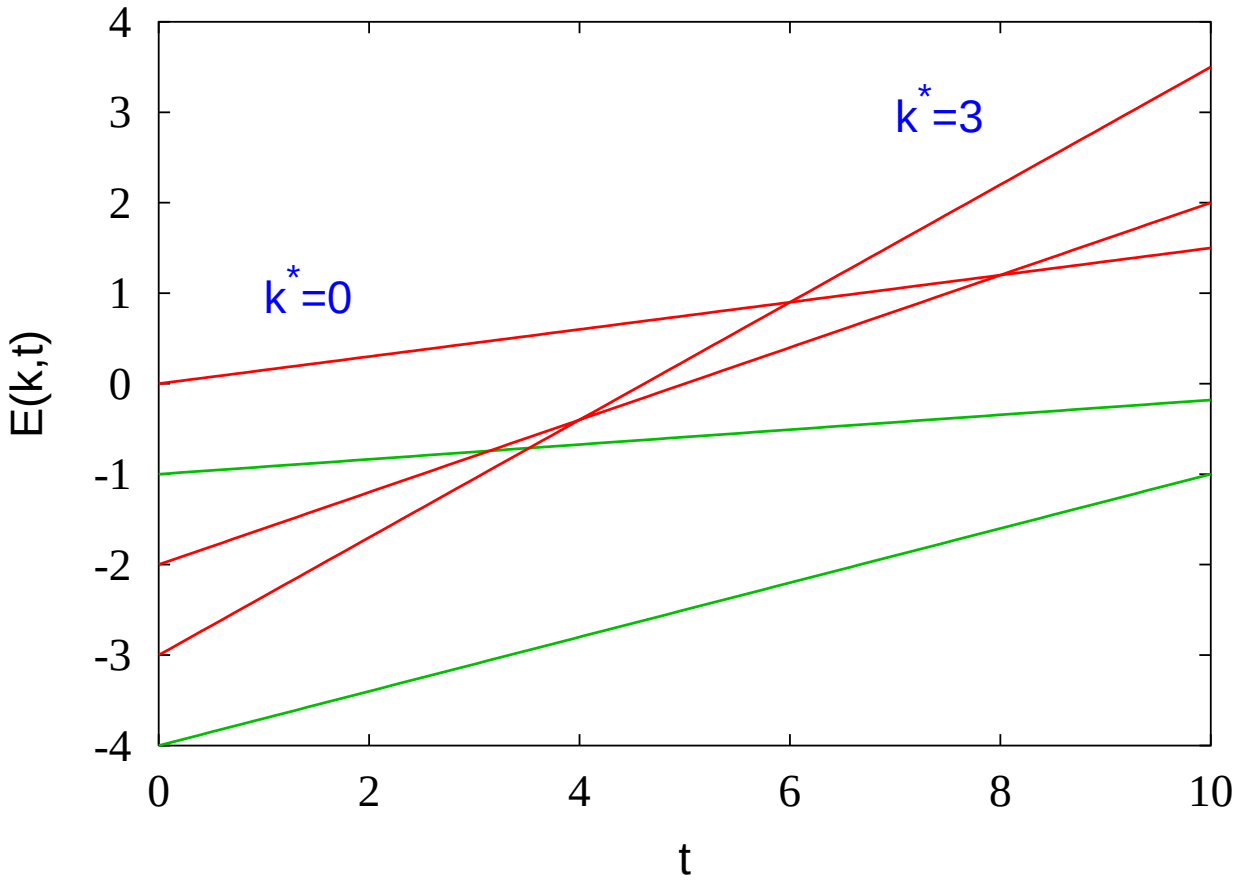
$$E_k(t) = -k + t\ F_k \quad , \quad k = 0, ..., N$$

where F_k is non-identically distributed variable chosen from

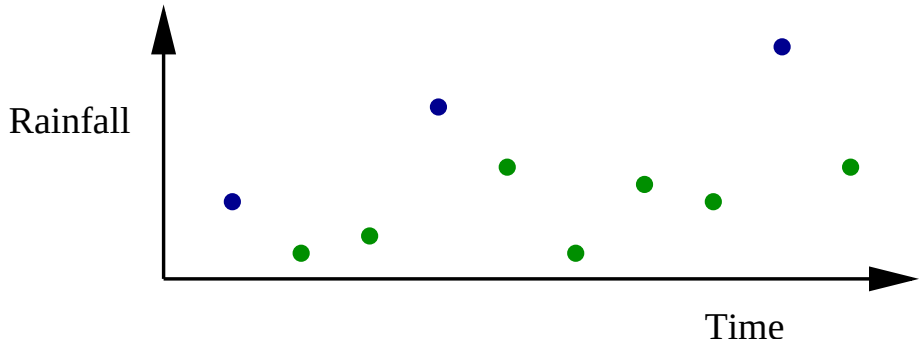
$$P_k(F) = \binom{N}{k}\ p(F)\ q(F)^{\binom{N}{k}-1}$$

Evolutionary Race

The population can be classified as : Spectator, Contender, Winner



- Spectator has a slope lower than that of the current winner k^*.
- Contender is a record since it has a slope higher than that of all the lines above it.



- Winner is a record that minimises the overtaking time also.

Thus, the prescription for σ^* is :

Find the fittest at constant Hamming distance from the initial sequence.

Winner is the one that overcomes the initial disadvantage in minimum time.

(It works.)

Traffic on a Single Lane Highway

(Ben-Naim *et al.*, 1994)

- Each car has an initial speed v_0.
- It moves ballistically with v_0 until it overtakes the preceding car.
- The overtaking car assumes the speed of the car leading the cluster.

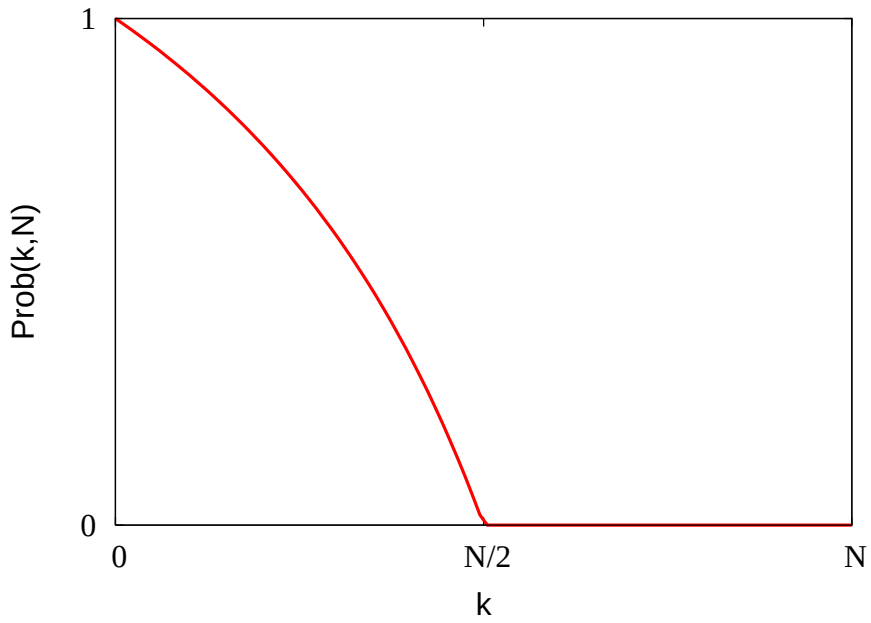
- The overtaking line behaves like the leading car.

Contenders vs. Winners (Jain and Krug, 2005)

Winners are a subset of contenders. How many of each type are there?

$$\text{Prob}(k^{th} \text{ slope is a record}) \approx \frac{N-2k}{N-k}$$

This distribution is universal and vanishes at $k = N/2$ since global maximum is typically located at $N/2$.

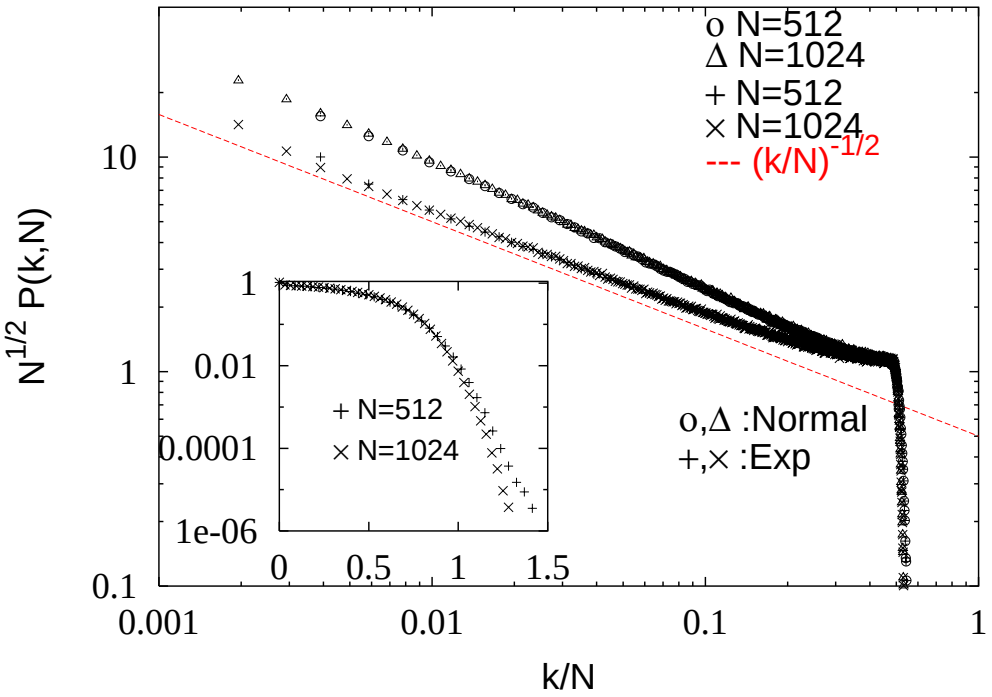


Average number of records $\approx (1 - \ln 2)N$

For $p(F)$ decaying as or faster than the exponential,

$$\text{Prob}(k^{th} \text{ slope is a winner}) \approx N^{-1/2} f(k/N)$$

so that the average number of winners scales as $\sqrt{N}$.



Thus, most of the 2^N sequences do not participate in the evolutionary race.

Approach to the Fittest

- Intersection time $T(k,k') = (k-k')/(F_k - F_{k'})$
- An estimate of the typical time T to find the fittest :

  T(last winner, last-but-one) $\sim \sqrt{N}/\epsilon$ since

—-most of the sequences are located within $\sqrt{N}$ distance of $N/2$

—-$\epsilon \sim O(1)$ using extremal statistics for Gumbel class

- Universal tails of the evolution time distribution :

$$P(T) \sim P\left(\frac{\sqrt{N}}{\epsilon}\right) \sim \frac{\sqrt{N}}{T^2}\mathrm{prob}(\epsilon = 0)$$

Fat tails imply that the expected time diverges.

What Next?

Study of the dynamics of stochastic, finite population model.