PROBABILISTIC STRUCTURES
IN EVOLUTION
DFG SPP 1590

COLLABORATIVE RESEARCH CENTER | SFB 680
Molecular Basis of
Evolutionary Innovations

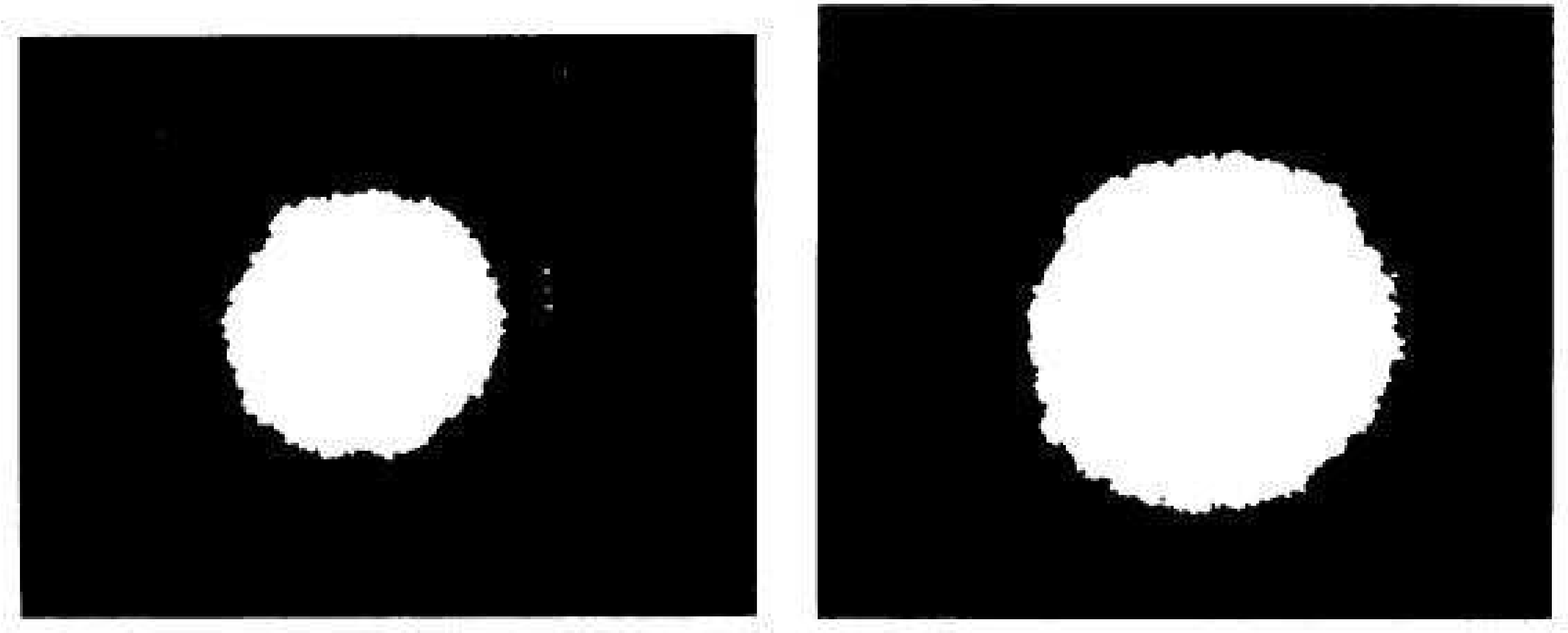
Random paths in evolutionary biology

Joachim Krug
Institute for Theoretical Physics, University of Cologne

- Biological contexts for the KPZ equation
- Paths on the hypercube
- Accessibility percolation

“Nonequilibrium dynamics of stochastic and quantum integrable systems”
KITP Santa Barbara, February 18, 2016

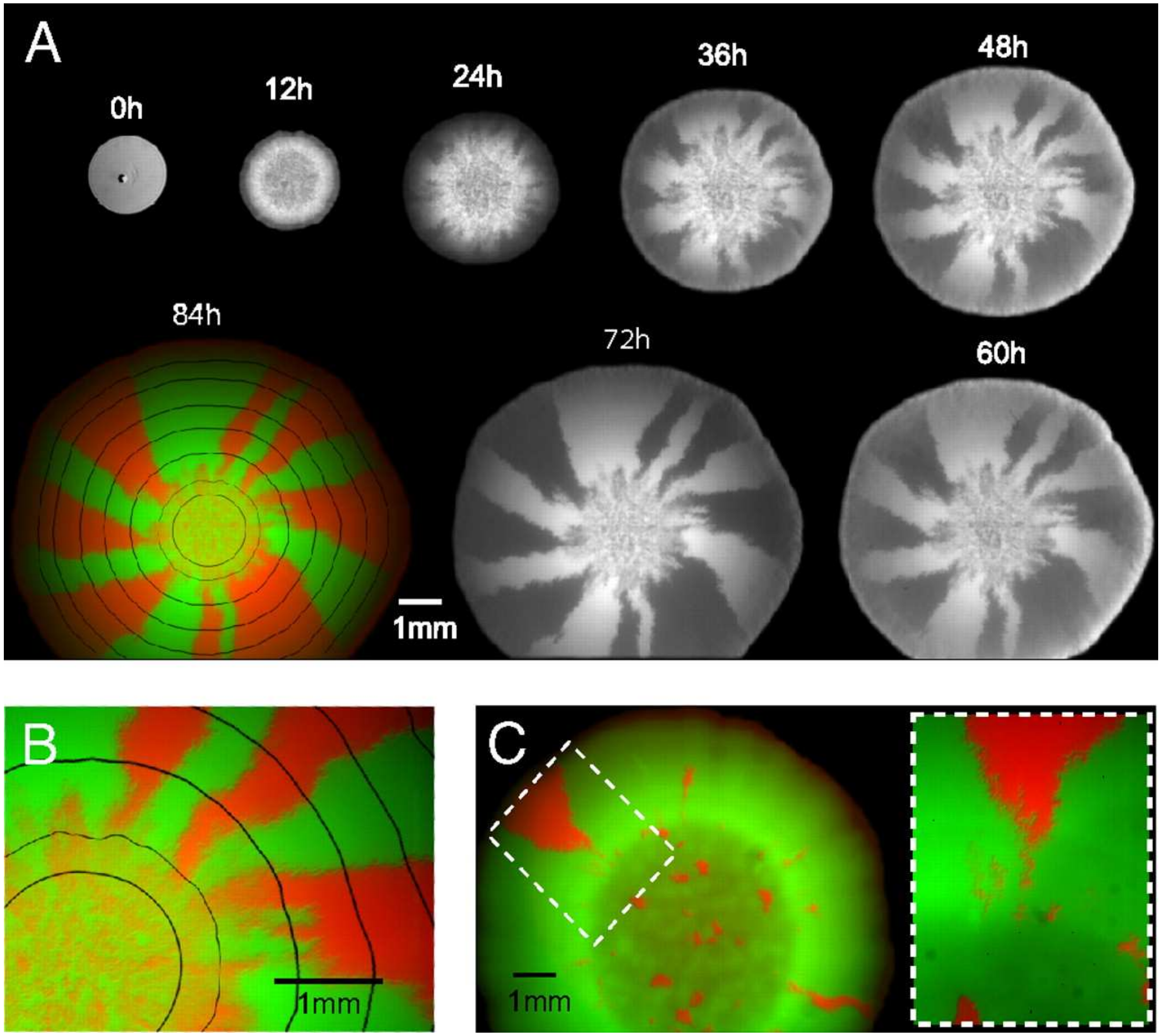
Eden growth



Eden 1961

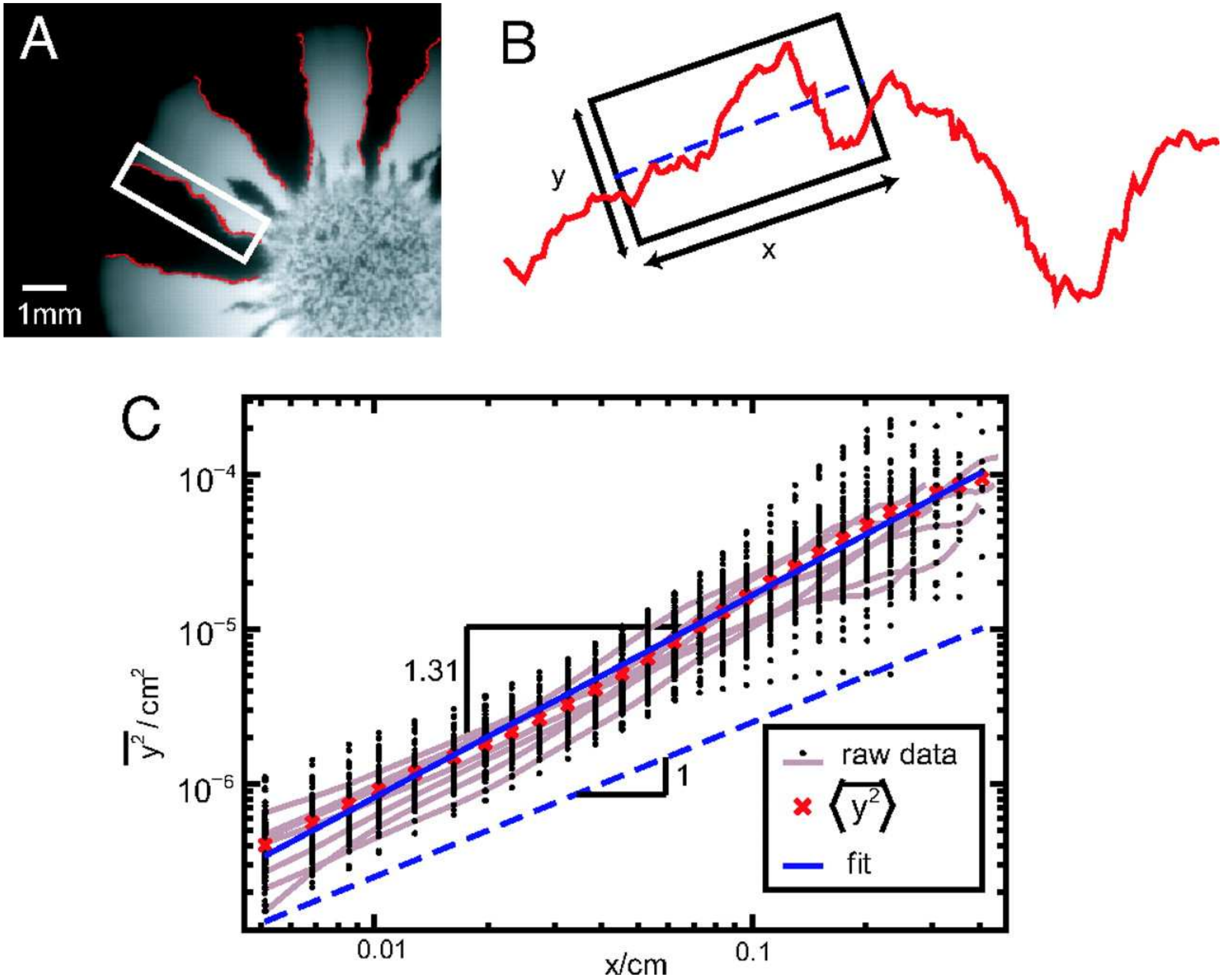
Genetic segregation in growing bacterial colonies

Hallatschek et al., PNAS 2007



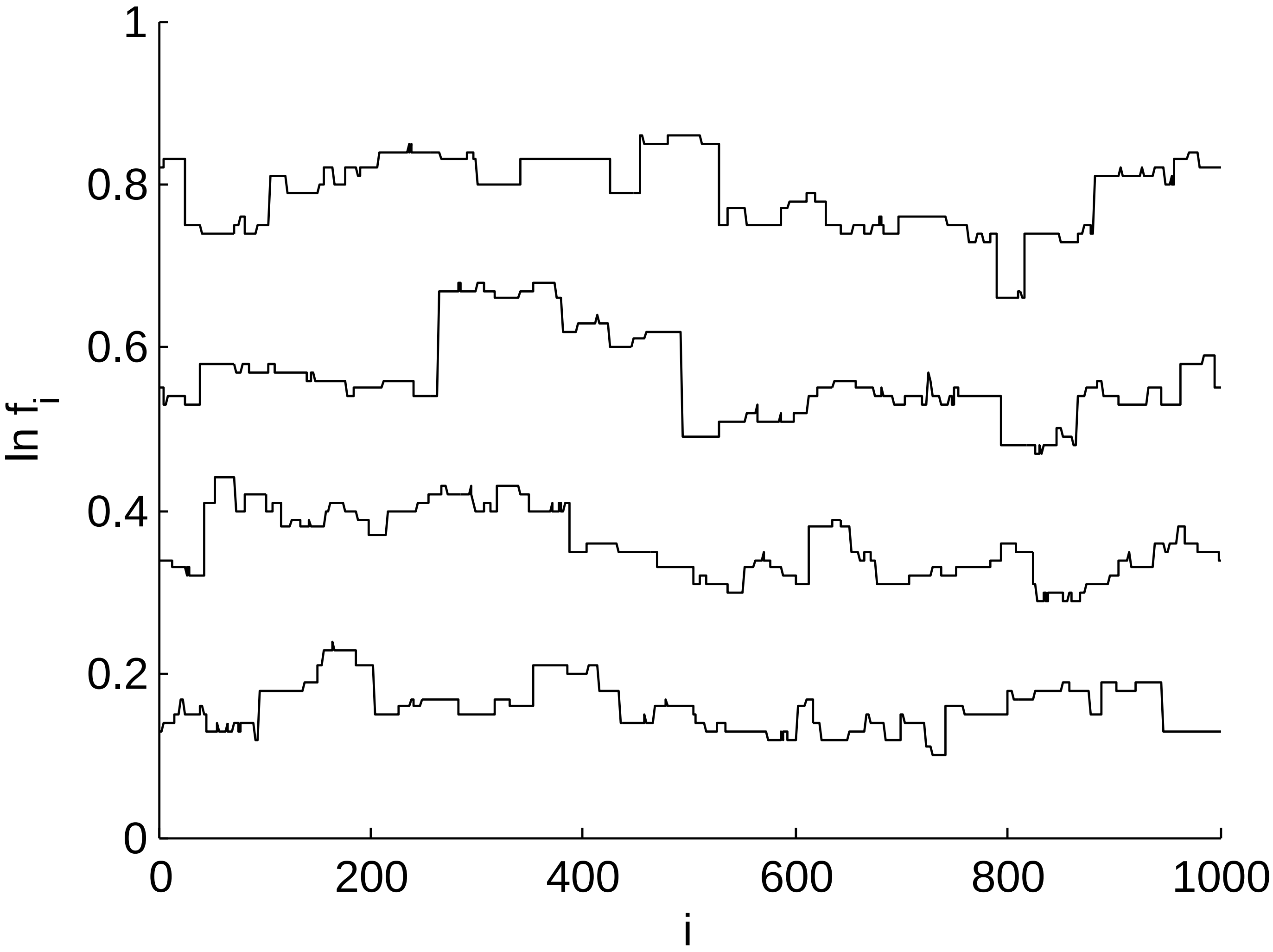
Sector boundaries display superdiffusive KPZ fluctuations

Hallatschek et al., PNAS 2007



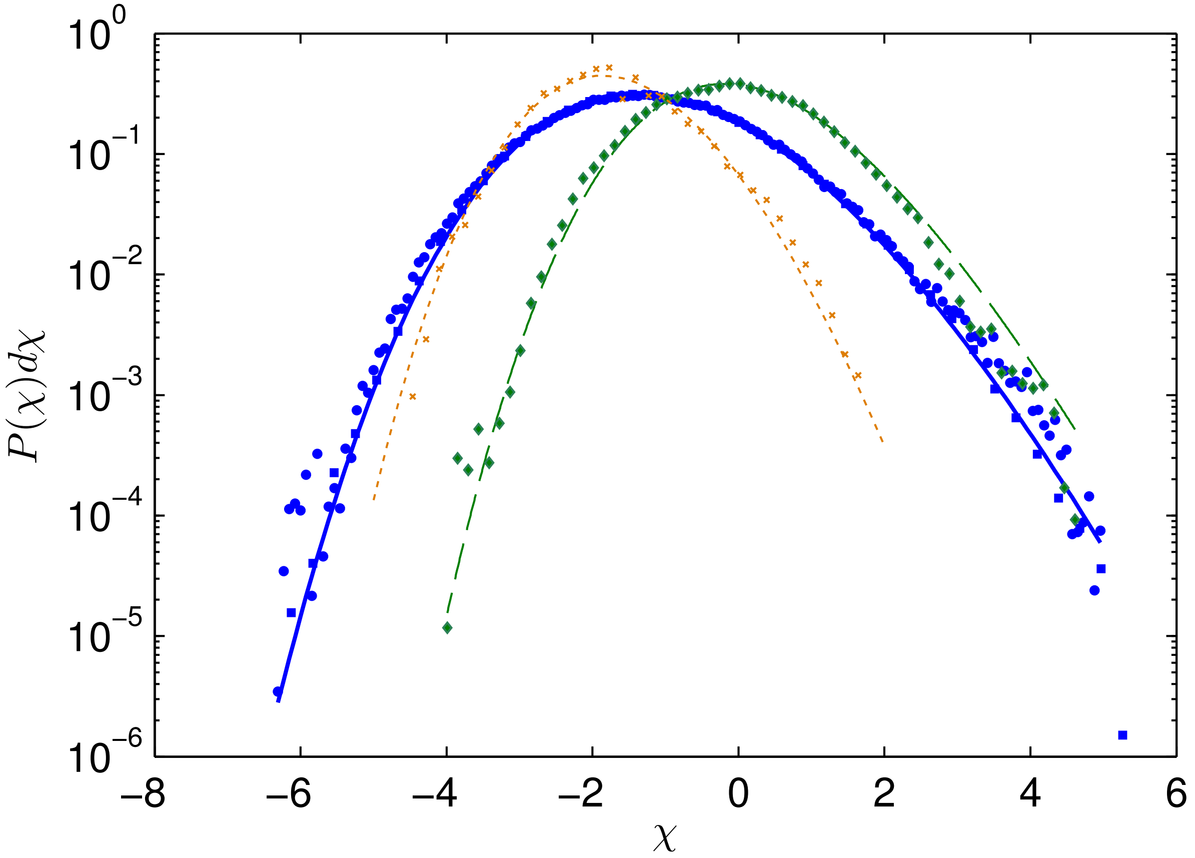
Fitness of a population in a linear habitat

J. Otwinowski, S. Boettcher, PRE 84:011925 (2011)



Universal fitness distributions

J. Otwinowski, JK, Phys. Biol. 11:056003 (2014)



- Three universal distributions for flat, droplet and stationary initial conditions

SHE and PAM

- Under the Cole-Hopf transformation $\psi(x,t) = \exp\left[\frac{\lambda}{2\nu}h(x,t)\right]$ the KPZ equation tranforms into the stochastic heat equation (SHE)

$$\frac{\partial \psi}{\partial t} = \nu\nabla^2\psi + \frac{\lambda}{2\nu}\eta(x,t)\psi$$

- Via the Feynman-Kac formula this establishes the relation to directed polymers in random media (DPRM) and first passage percolation (FPP)

- When the noise is independent of time (“columnar DPRM”) the problem is known as the parabolic Anderson model (PAM) with a natural biological intepretation: Ebeling, Engel, Esser, Feistel JSP 1984

$x \longrightarrow$ phenotype, $\psi(x,t) \longrightarrow$ population density, $\eta(x) \longrightarrow$ fitness

- However in that context the dynamics should properly be defined on the space of genetic sequences rather than on $\mathbb{R}^d$ or $\mathbb{Z}^d$

Sequence spaces

- Genetic information is encoded in DNA-sequences consisting of four different nucleotide bases

..ACTATCCATCTACTACTCCCAGGAATCTCGATCCTACCTAC...

- The sequence space consists of all 4^L sequences of length L

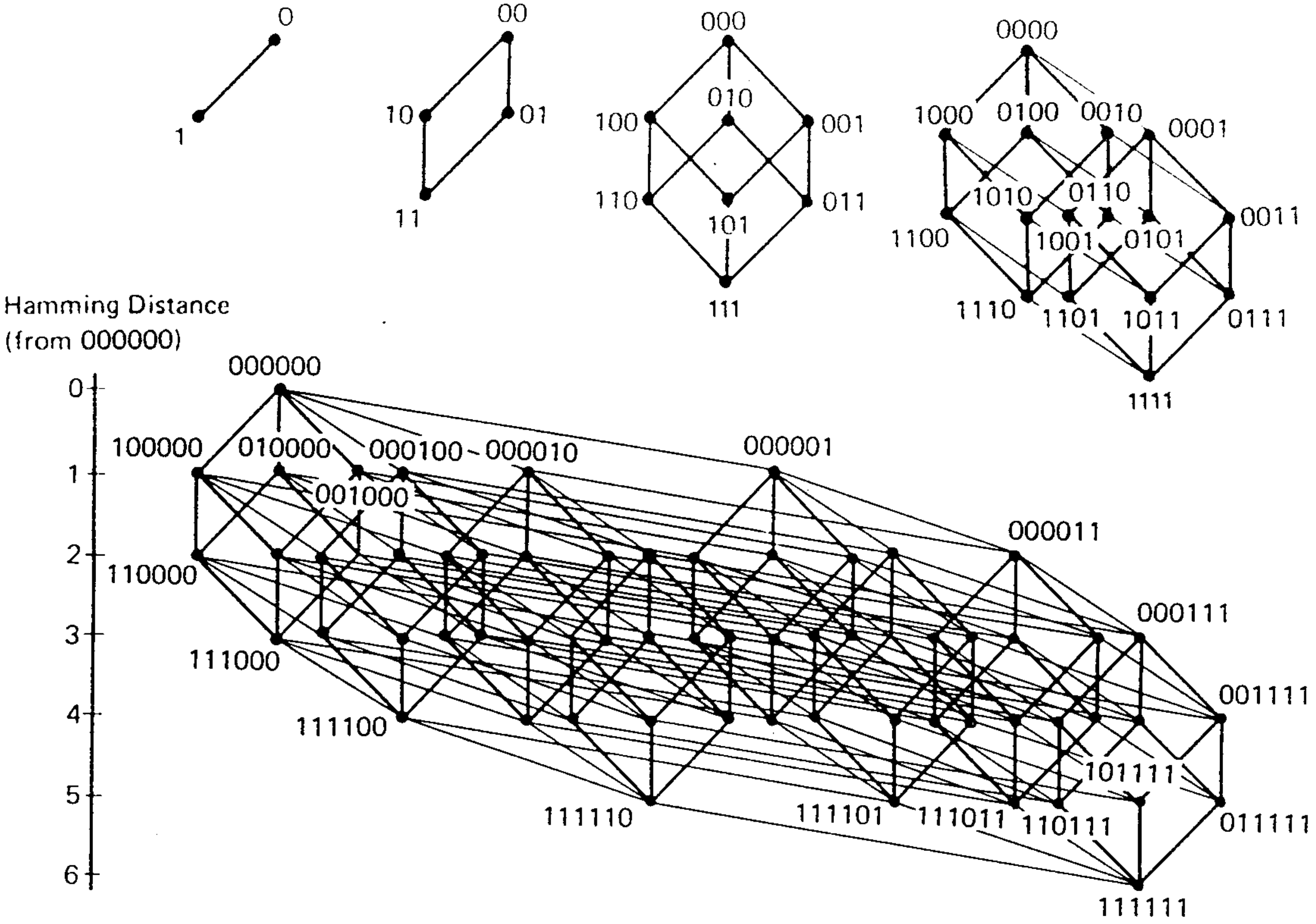
- Typical genome lengths:
  $L \sim 10^3$ (viruses), $L \sim 10^6$ (bacteria), $L \sim 10^9$ (higher organisms)

- Proteins are sequences of 20 amino acids with $L \sim 10^2$

- Coarse-grained representation of classical genetics: L genes that are present as different alleles; often it is sufficient to distinguish between wild type (0) and mutant (1) $\Rightarrow$ binary sequences

- Hamming distance: Two sequences are nearest neighbors if they differ in a single letter (mutation)

Hamming spaces/hypercubes for $L = 1 - 6$

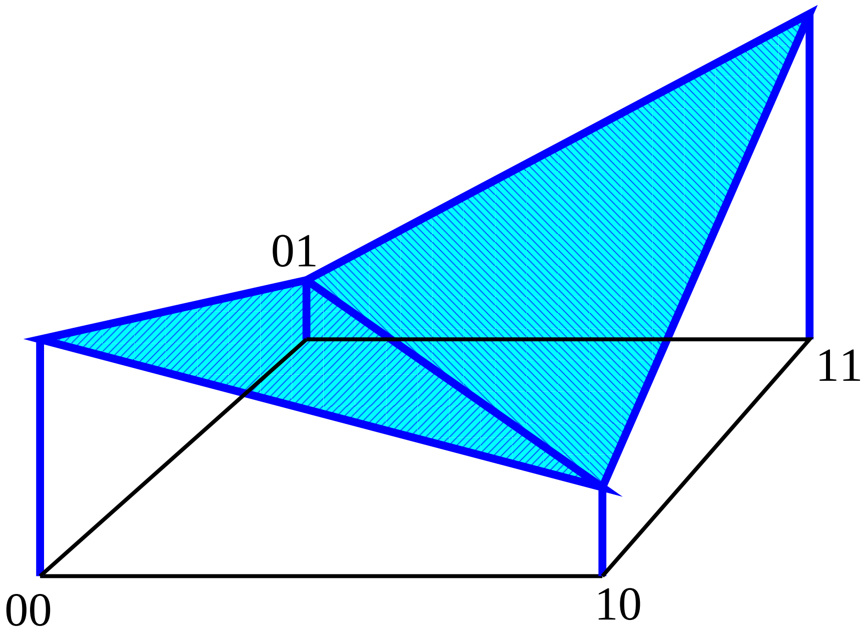


Fitness landscapes

- A fitness landscape assigns a fitness value $f(\sigma)$ to each genotype sequence $\sigma = (\sigma_1\sigma_2..\sigma_L)$ with $\sigma_i \in \{0,1\}$

- Evolution is a hill-climbing process in the fitness landscape

- Example: $L = 2$

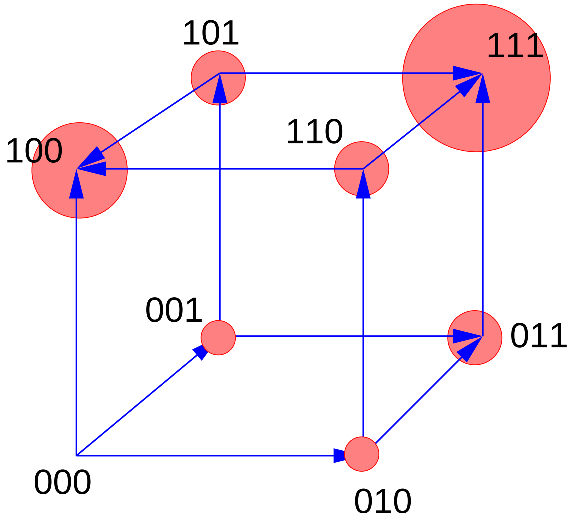


Fitness landscapes

- A fitness landscape assigns a fitness value $f(\sigma)$ to each genotype sequence $\sigma = (\sigma_1\sigma_2..\sigma_L)$ with $\sigma_i \in \{0,1\}$

- Evolution is a hill-climbing process in the fitness landscape

- Example: $L = 3$

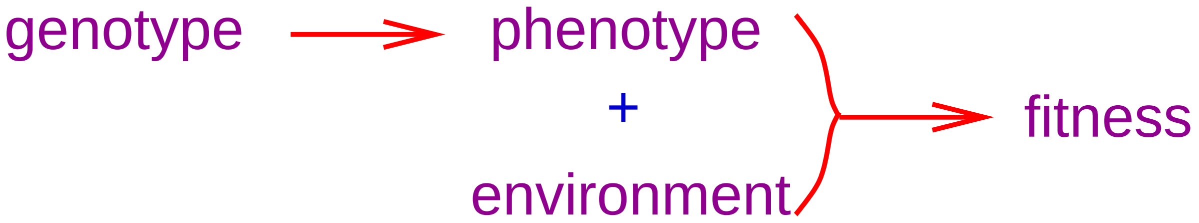


- 3! = 6 directed (0 → 1) and 18 undirected mutational pathways from $\sigma^{(0)} = (000)$ to $\sigma^{(1)} = (111)$

Random fitness landscapes

- The fitness $f(\sigma)$ of genotype σ is the expected number of offspring of an individual carrying σ

- The mapping $\sigma \to f(\sigma)$ is very complicated:



Simple choice: Assign fitnesses at random to genotypes

- Fitnesses as i.i.d. random variables $\Rightarrow$ Kingman's house-of-cards model

  Kingman 1978, Kauffman & Levin 1987

- Equivalent to Derrida's Random Energy Model of spin glasses Derrida 1981

- Correlated landscapes can be generated along similar lines (e.g., the spin-glass-like NK-models)

PAM on the random hypercube: Adaptive flights

K. Jain, JK, JSTAT 2005; K. Jain, PRE 2007

- Under PAM dynamics the population concentrates on sites with exceptionally high fitness

- An adaptive trajectory consists of a sequence of long-ranged “tunneling” events between such sites that terminates at the global maximum

- The number of jumps is $\mathscr{O}(\sqrt{L})$ for Gumbel-class fitness distributions and $\mathscr{O}(1)$ for power-law distributions

- The distribution of the time T_k of the k’th last jump has a universal power law tail
$$\mathrm{Prob}[T_k > t] \sim t^{-k}, \quad k = 1, 2, 3, \ldots$$
which implies that the expected time to the maximum is infinite

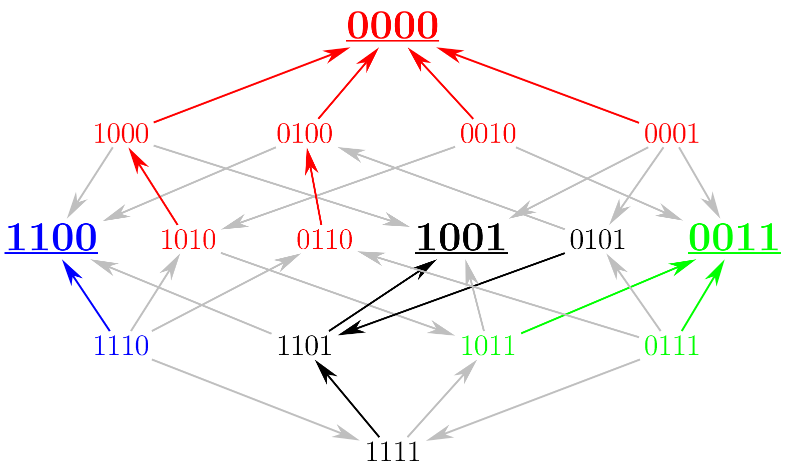
- This scenario is however biologically meaningless because it relies on exponentially small population densities

Evolutionary accessibility

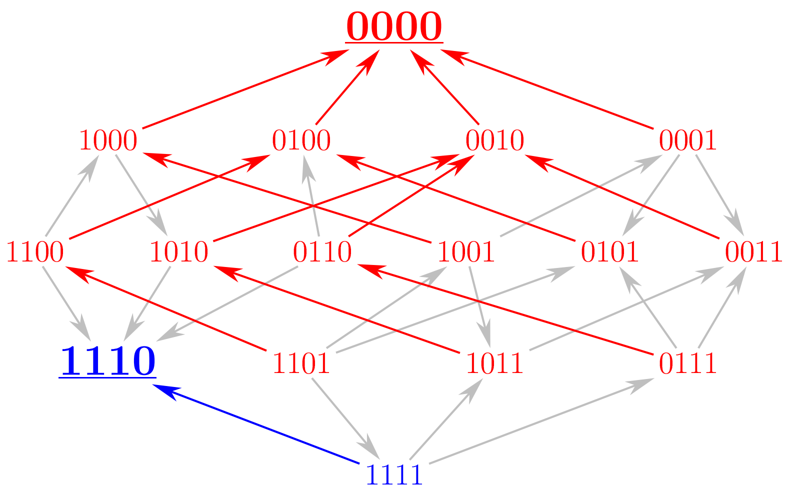
J. Franke et al., PLoS Comp. Biol. **7** (2011) e1002134

- In moderately large populations, adaptive trajectories are constrained to move uphill in single mutational steps ⇒ a pathway connecting two genotypes is accessible if fitness increases monotonically in each step

- Example: Mutational pathways from (1111) to (0000) in two 4-locus subgraphs of an 8-dimensional empirical fitness landscape for the filamentous fungus *Aspergillus niger*



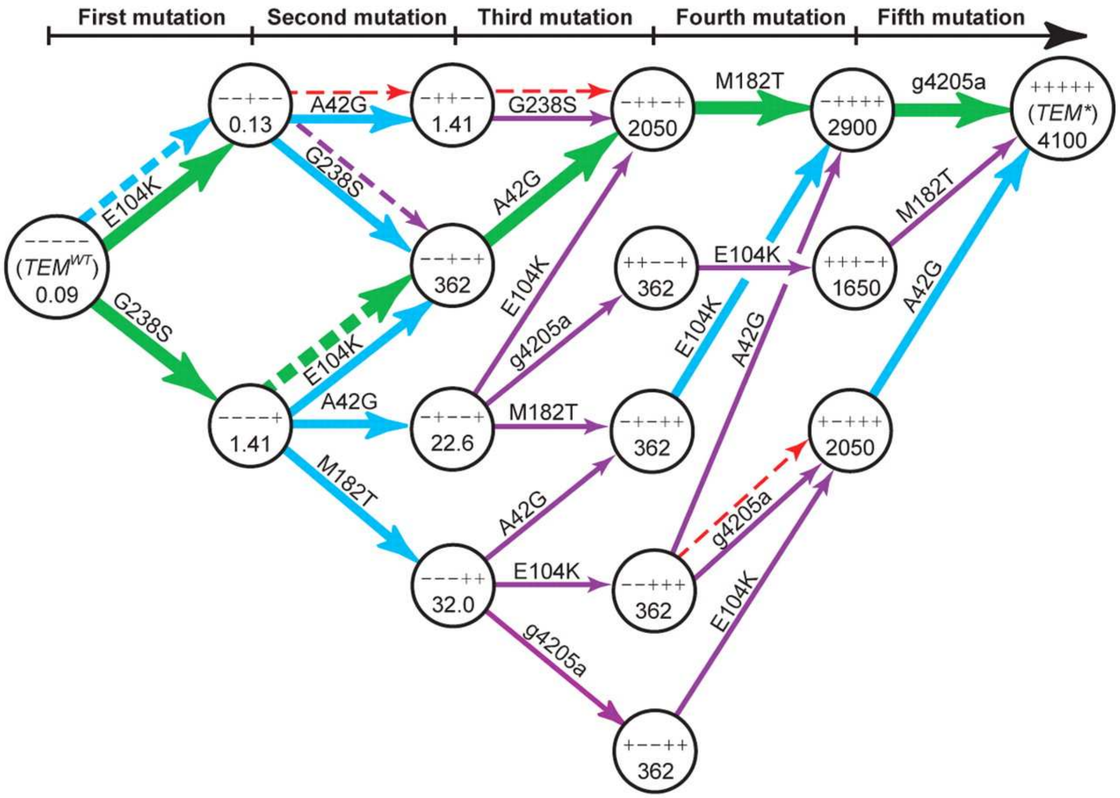
no directed pathway accessible



6 out of 24 pathways accessible

Pathways to antibiotic resistance

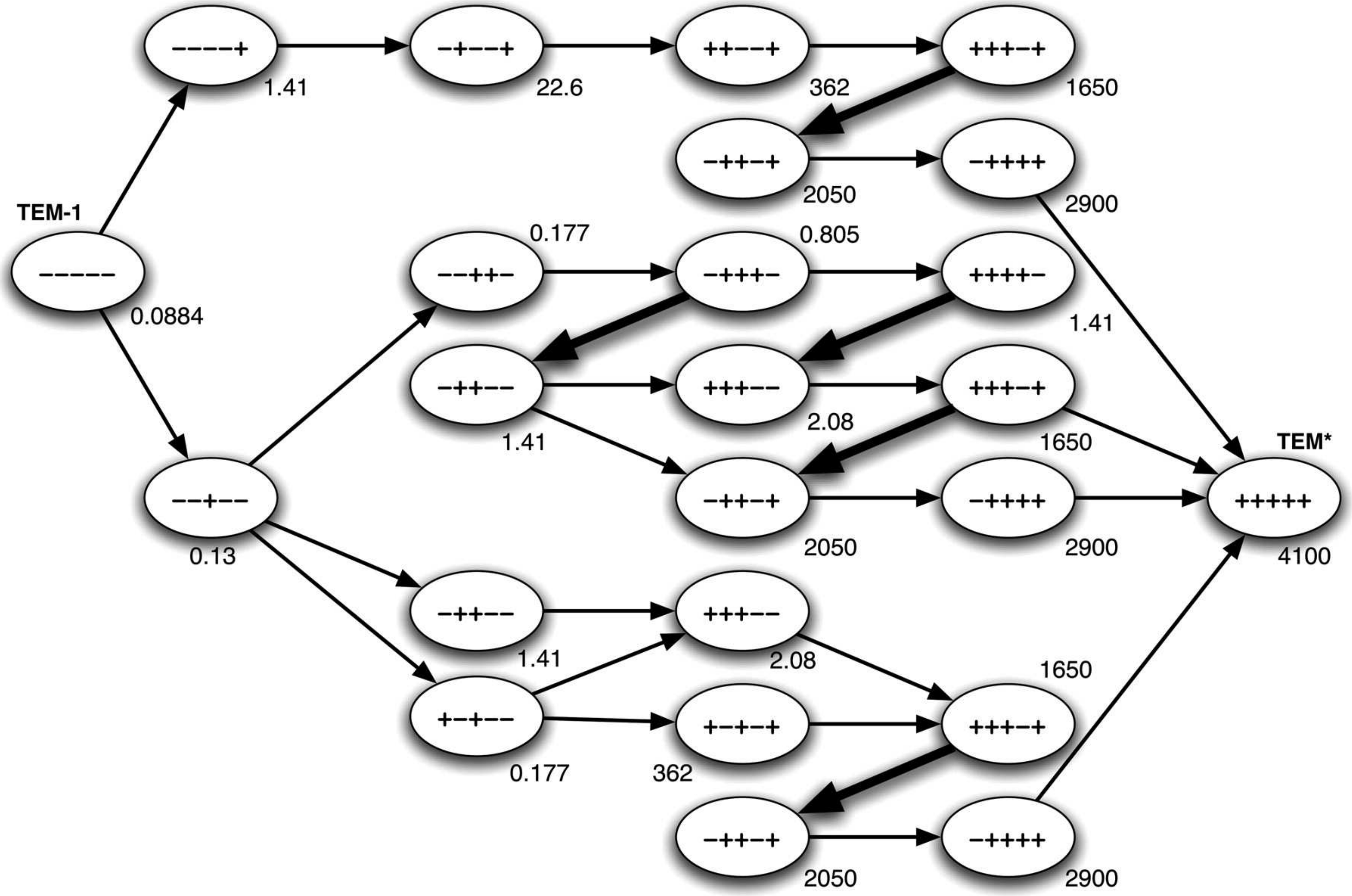
D.M. Weinreich et al., Science **312**, 111 (2006)



- 5 mutations increase resistance to a new drug by $\sim 10^5$
- 18 out of $5! = 120$ directed mutational pathways are accessible, and only few of them have appreciable weight

Pathways to antibiotic resistance

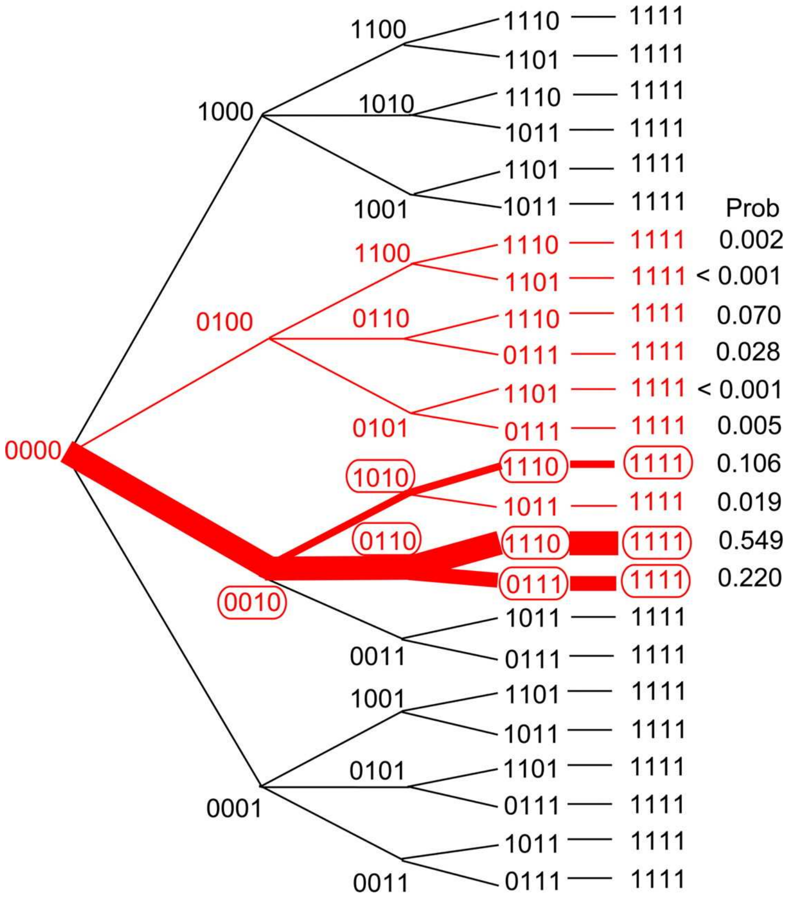
De Pristo et al., Mol. Biol. Evol. 24:1608 (2007)



- 27 out of 18651552840 undirected pathways are accessible

Pathways to drug resistance in malaria

E.R. Lozovsky et al., Proc. Natl. Acad. Sci. USA **106**, 12025 (2009)



- 4! = 24 pathways, 10 (red) are monotonic in resistance
- Dominating pathways are realized in natural populations

Accessibility percolation

S. Nowak, JK, EPL **101**, 66004 (2013)

- Directed or undirected graph G with nodes $x \in G$ and distance $d(\cdot,\cdot)$
- Assign a nondegenerate real random variable $f(x)$ to each node
- A path is a string of nodes $x_0 \longrightarrow x_1 \longrightarrow x_2 \longrightarrow \ldots \longrightarrow x_N$ such that $d(x_i, x_{i+1}) = 1$ for all i
- A path is called accessible if f increases monotonically along the path, i.e. $f(x_0) < f(x_1) < \ldots < f(x_N)$
- Accessibility percolation is concerned with the existence of global paths that connect the global maximum $x_{\max}$ of $f(x)$ to the node at maximal distance $D \equiv \max_{x \in G} d(x_{\max}, x)$
- In the standard setting G is the hypercube and the $f(x)$ are i.i.d. random variables

Directed random hypercube

- Assign maximal fitness $f = 1$ to $\sigma^{(1)} \equiv (1, 1, ..., 1)$ and i.i.d. $U(0,1)$ RV's to all other sites

- What is the expected number of directed accessible pathways from a site at distance d to $\sigma^{(1)}$?

- The total number of paths is $d!$, and a given path consists of d i.i.d. fitness values $f_0, ..., f_{d-1}$; it is accessible iff $f_0 < f_1 .... < f_{d-1}$

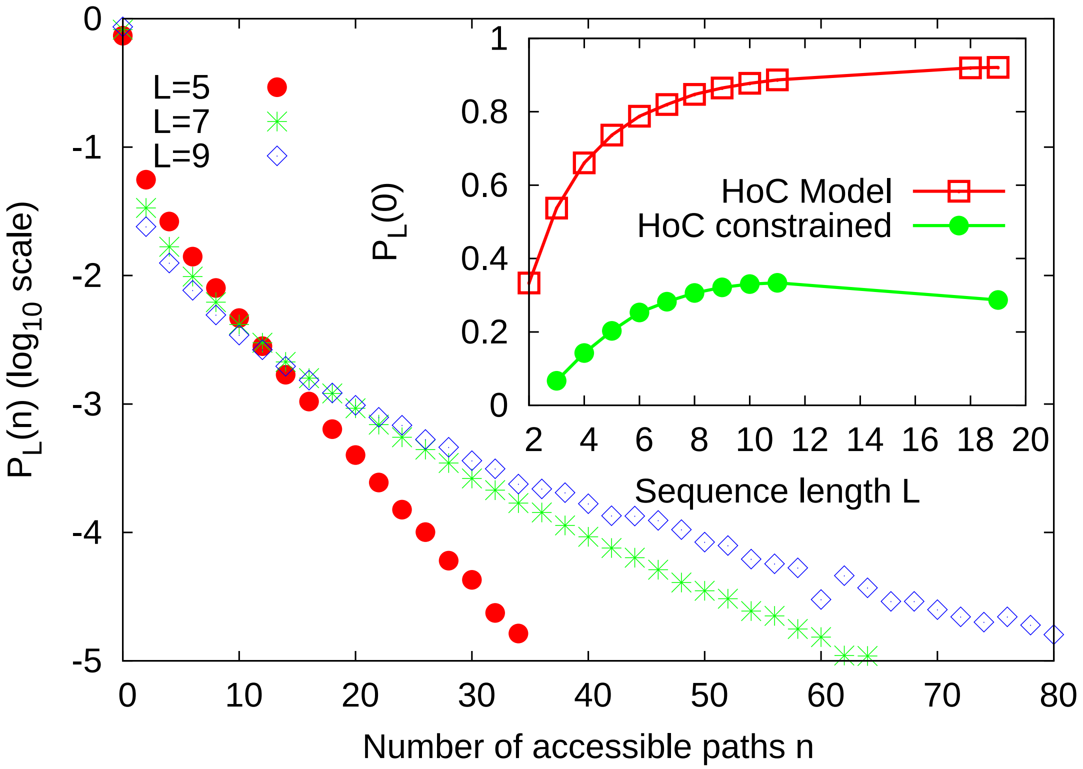
- Since all $d!$ permutations of the d random variables are equally likely, the probability for this event is $1/d!$

$$\Rightarrow \mathbb{E}(n_{\text{acc}}) = \frac{1}{d!} \times d! = 1$$

- This applies in particular for $d = L$

Distribution of the number of accessible paths

J. Franke et al., PLoS Comp. Biol. **7** (2011) e1002134



- "Condensation of probability" at $n_{\text{acc}} = 0$

  $\Rightarrow$ mean is not representative of the typical behavior

- Constraining initial fitness to $f_0 = 0$ massively increases the accessibility

Transition as a function of initial fitness

- Conditioned on initial fitness $f_0 \in [0,1)$ the expected number of accessible paths is

$$\mathbb{E}(n_{\mathrm{acc}}) = \frac{(1-f_0)^{L-1}}{(L-1)!} \times L! = L(1-f_0)^{L-1}$$

  which diverges/vanishes asymptotically for large L when $f_0 < \frac{\ln L}{L}$ / $f_0 > \frac{\ln L}{L}$

- This implies that the existence of accessible paths becomes likely at $f_0 \sim \frac{\ln L}{L}$, in the sense that

  Hegarty & Martinsson, Ann. Appl. Prob. 2014

$$\lim_{L\to\infty} \mathrm{Prob}[n_{\mathrm{acc}} > 0] = \begin{cases} 0 & \text{for} \quad f_0 > \dfrac{\ln L}{L} \\ 1 & \text{for} \quad f_0 < \dfrac{\ln L}{L}. \end{cases}$$

- Proof uses estimate of second moment of n_{acc} and the bounds

$$\mathbb{E}(n_{\mathrm{acc}}) \geq \mathrm{Prob}[n_{\mathrm{acc}} > 0] \geq \frac{\mathbb{E}(n_{\mathrm{acc}})^2}{\mathbb{E}(n_{\mathrm{acc}}^2)}$$

Accessibility percolation on trees

S. Nowak, JK, EPL **101**, 66004 (2013)

- Consider a regular tree with branching number b and height h equipped with i.i.d. RV's on the nodes
- Let $n_{\rm acc}$ denote the number of accessible paths from the root to the leaves
- First and second moments are given by

$$\mathbb{E}(n_{\rm acc}) = \frac{b^h}{h!}, \qquad \mathbb{E}(n_{\rm acc}^2) = \mathbb{E}(n_{\rm acc}) + \frac{b-1}{b} \sum_{k=1}^{h} \binom{2k}{k} \frac{b^{h+k}}{(h+k)!}$$

- Scaling $b, h \to \infty$ at fixed $\alpha = b/h$ it follows that accessibility percolation occurs at some $\alpha_c \in [1/e, 1]$,
- Refined analysis shows that $\alpha_c = 1/e$ which corresponds exactly to the hypercube geometry

Roberts & Zhao, ECP 2013

Effect of downhill steps

É. Brunet, L. Deecke, JK, in preparation

- Two scenarios for allowing downhill steps along the path:
  - unconditional: $\ldots f_{i-2} < f_{i-1} > f_i < f_{i+1} < \ldots.$ for some i
  - conditional: $f_{i-1} > f_i < f_{i+1}$ but $f_{i+1} > f_{i-1}$

- Expected number of accessible paths in the two cases are

$$\mathbb{E}^{\mathrm{uc}}(n_{\mathrm{acc}}) = 2^L - L, \qquad \mathbb{E}^{\mathrm{c}}(n_{\mathrm{acc}}) = 1 + \frac{1}{2}L(L-1)$$

- In the unconditional case accessible paths exist almost surely for any initial fitness when $L \to \infty$, whereas in the conditional case the accessibility threshold is

$$f_0 \sim \frac{(2p+1)\ln L}{L}$$

  when p downhill steps are allowed for

Accessibility percolation on the undirected hypercube

J. Berestycki, É. Brunet, Z. Shi, arXiv:1401.6894

- A general undirected path from $\sigma^{(0)}$ to $\sigma^{(1)}$ consists of $L+2p$ steps where $p \geq 0$ is the number of backsteps (mutational reversions)

- The expected number of accessible paths conditioned on starting fitness f_0 is

$$\mathbb{E}(n_{\text{acc}}) = \sum_{p \geq 0} a_{L,p} \frac{(1-f_0)^{L+2p-1}}{(L+2p-1)!}$$

  where $a_{L,p}$ is the number of paths with p backsteps.

- Analyzing the asymptotics of the $a_{L,p}$ it is shown that

$$\lim_{L \to \infty} [\mathbb{E}(n_{\text{acc}})]^{1/L} = \sinh(1-f_0)$$

  which suggests a finite accessibility threshold $f_0^* = 1 - \sinh^{-1}(1) \approx 0.11863\ldots$

A link to first passage percolation

A. Martinsson, arXiv:1501.02206

- Graph G with i.i.d. U(0,1) random waiting times $\tau(x)$ assigned to nodes x
- The first passage time from a distinguished node $x^{(0)}$ to x is

$$T(x) = \min_{\pi} \left[\sum_{y \in \pi \setminus \{x^{(0)}, x\}} \tau(y) \right]$$

  where π is a path from $x^{(0)}$ to x

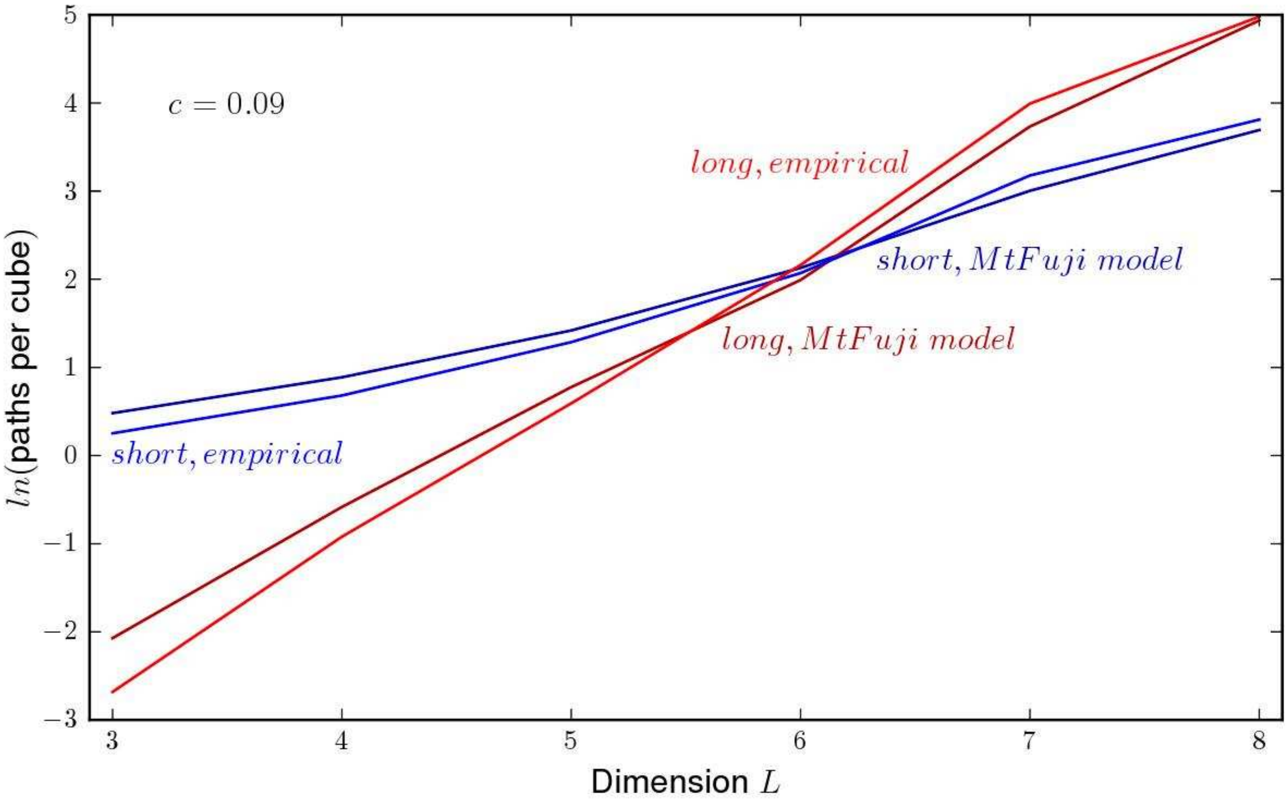
- Then the fitnesses $f(\sigma)$ defined as the fractional part of $f_0 + T(x)$ are i.i.d. $U(0,1)$ RV's, and as a consequence

$$\mathrm{Prob}[n_{\mathrm{acc}}(x^{(0)} \rightarrow x) > 0] = \mathrm{Prob}[T(x) < 1 - f_0].$$

- It follows that the first passage time on the oriented (unoriented) hypercube converges to 1 ($1 - f_0^* = \sinh^{-1}(1) \approx 0.88137...$) for large L.

The role of backsteps in empirical data

M. Josupeit, JK, in preparation



- Comparison of subgraph analysis of an empirical data set with the rough Mt. Fuji model defined by $f(\sigma) = cd(\sigma, \sigma^{(0)}) + \eta_\sigma$ with $U(0,1)$ RV's η_σ
- Accessibility is dominated by direct paths for small L

Summary

- A new type of random path problem motivated by evolutionary biology
- “Critical” role of hypercube geometry
- Provides a tool to interpret empirical fitness landscapes

  J.A.G.M. de Visser, JK, Nat. Rev. Gen. 15:480 (2014)
- Focus so far on the existence of paths rather than on the distribution of path weights

Thanks to:

- Lucas Deecke, Jasper Franke, Mario Josupeit, Stefan Nowak
- Julien Berestycki, Éric Brunet, Peter Hegarty, Anders Martinsson
- Arjan de Visser