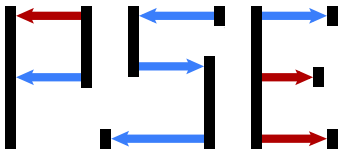


PROBABILISTIC STRUCTURES
IN EVOLUTION

DFG SPP 1590

COLLABORATIVE RESEARCH CENTER | SFB 680
Molecular Basis of
Evolutionary Innovations

How predictable is evolution?

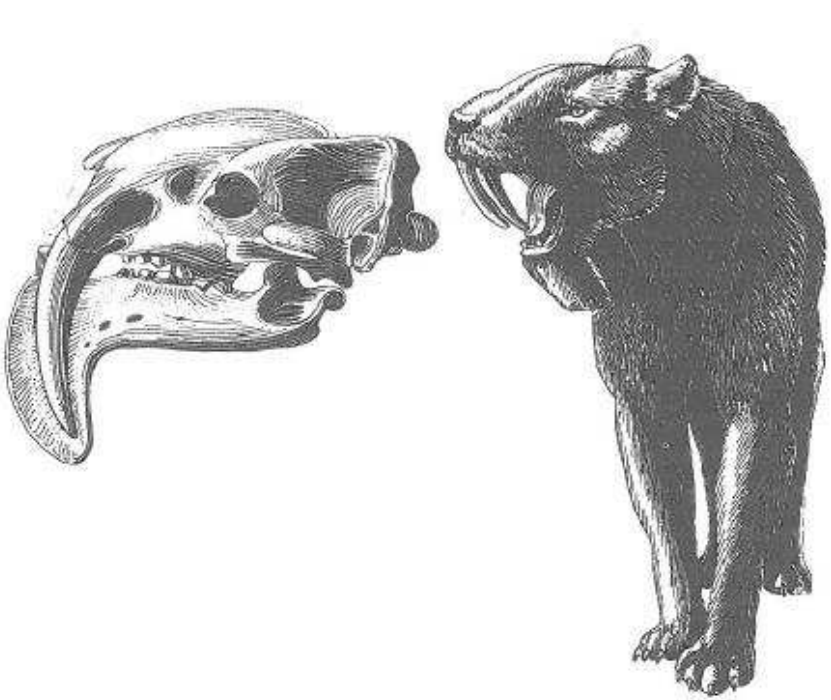
Joachim Krug
Institute for Theoretical Physics, University of Cologne

Higgs Centre Colloquium, Edinburgh, December 11, 2015

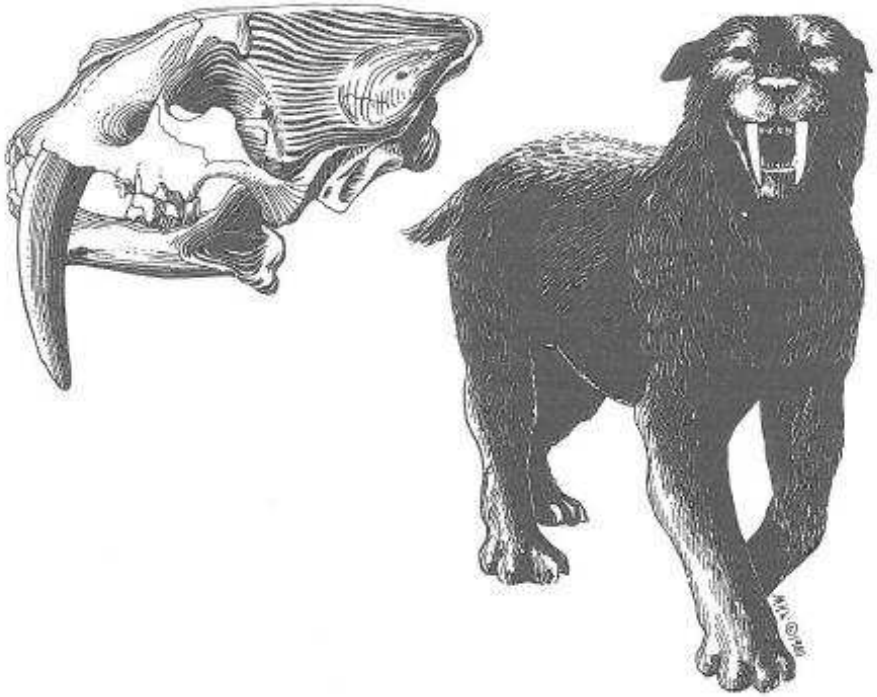
How predictable is evolution?

"The evolutionary routes are many, but the destinations are limited."

Simon Conway Morris, Life's Solution (2003)



marsupial



placental

How predictable is evolution?

- The evolutionary process is an intricate interplay of deterministic selection and stochastic mutational and reproductive events

- If we could replay the ‘tape of life’, would the outcome be similar to the current biosphere or something completely different? S.J. Gould (1989)

How predictable is evolution?

- The evolutionary process is an intricate interplay of deterministic selection and stochastic mutational and reproductive events

- If we could replay the 'tape of life', would the outcome be similar to the current biosphere or something completely different? S.J. Gould (1989)

Notions of predictability

- Strong predictability implies the ability to forecast evolution forward in time (e.g., to predict the dominant strain in the next influenza season or the emergence of antibiotic resistance)

- Weak (*a posteriori*) predictability implies repeatability in replicate realizations of the process

How predictable is evolution?

- The evolutionary process is an intricate interplay of deterministic selection and stochastic mutational and reproductive events

- If we could replay the ‘tape of life’, would the outcome be similar to the current biosphere or something completely different? S.J. Gould (1989)

Experimental evolution

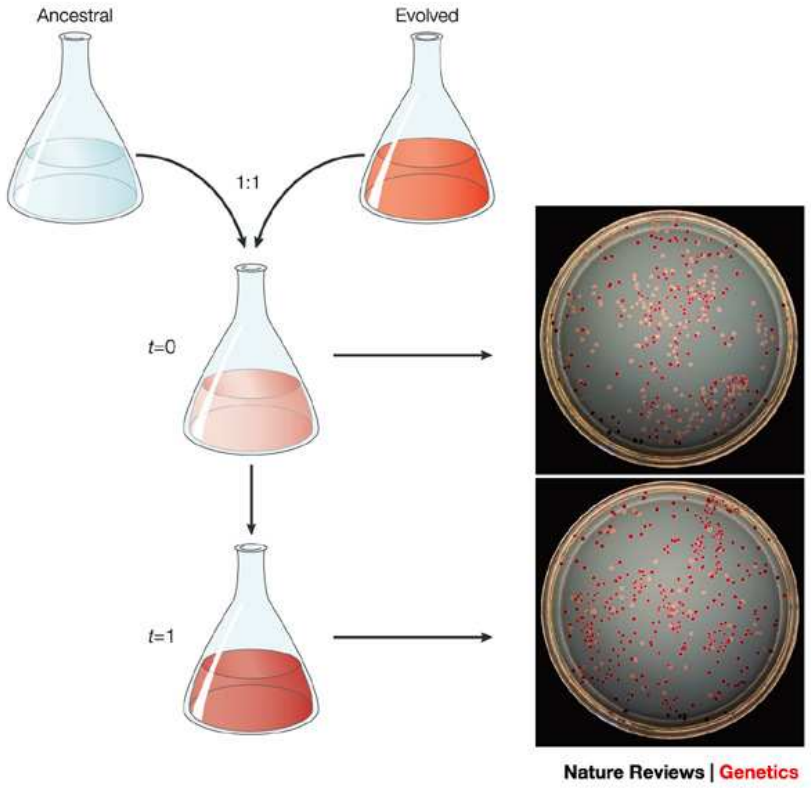
- Evolution experiments with microbes systematically address predictability in the weak sense of repeatability of adaptive trajectories

- A further distinction can be made according to whether the initial point or the endpoint of the process is specified

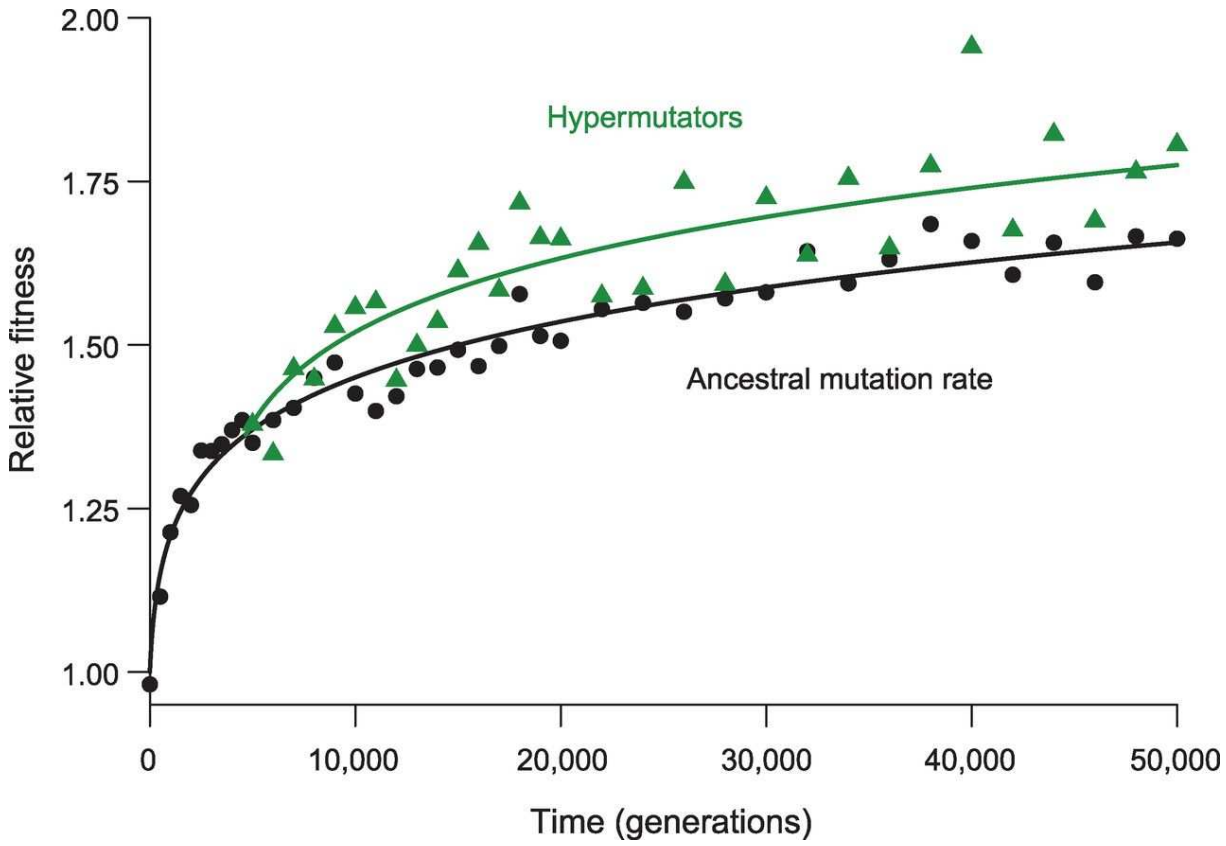
Experimental evolution with microbes

S.F. Elena, R.E. Lenski, Nature Reviews Genetics **4**, 457 (2003)

- Populations of bacteria or viruses propagated in the lab over tens of thousands of generation
- Controlled environment, mutation rate & population size
- Fitness monitored by competition against ancestral population
- Phenotypic characterization & sequencing of mutants
- Repeatability of evolution can be quantified on the phenotypic and genotypic level



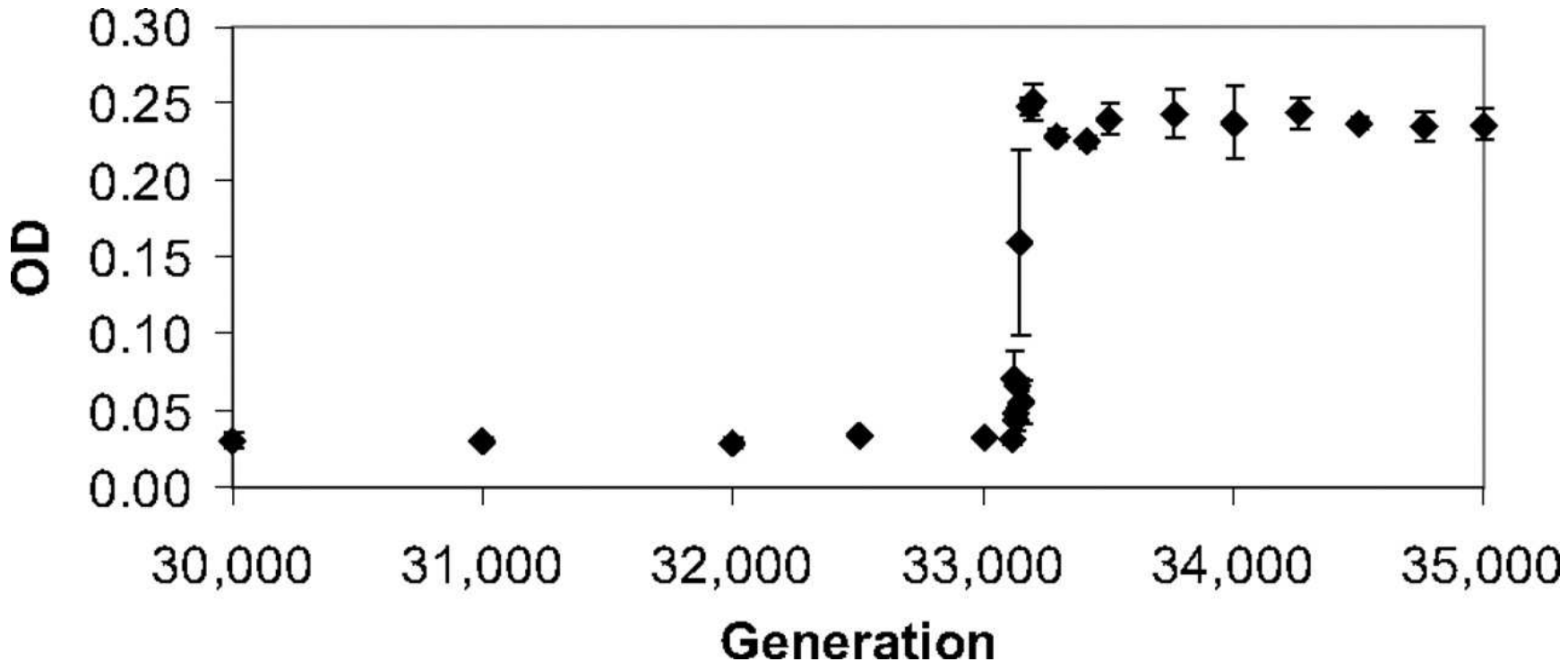
The long-term evolution experiment with *E. coli*



Wiser et al., Science 2013

- Started in 1988 with 12 populations in a glucose-limited medium
- Figure shows mean fitness of 6 normal and 3 hypermutator populations

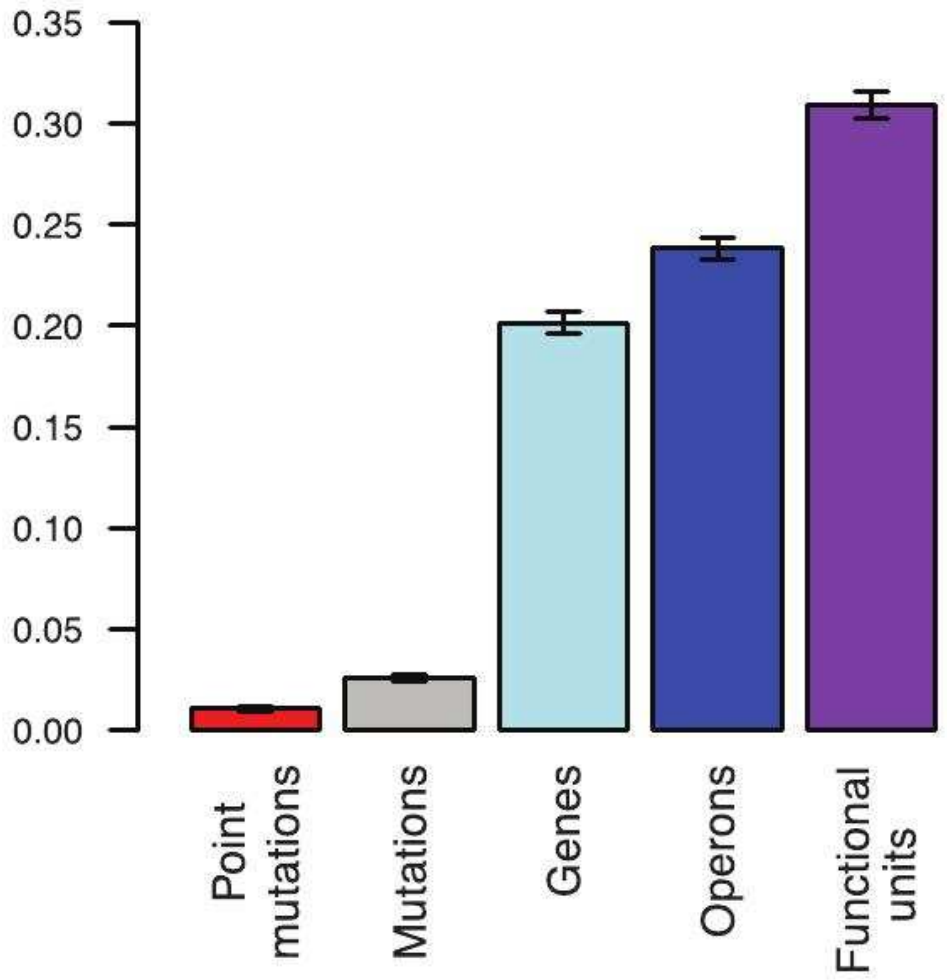
An example of historical contingency



Blount et al., PNAS **105**, 7899 (2008)

- Ability to exploit citrate evolved in one population after 31,500 generations

115 populations of *E. coli* adapting to 42.2^{O} C



Tenaillon et al. Science 335 (2012) 457

- Pairwise genetic convergence increases with organizational level

Goal of this talk

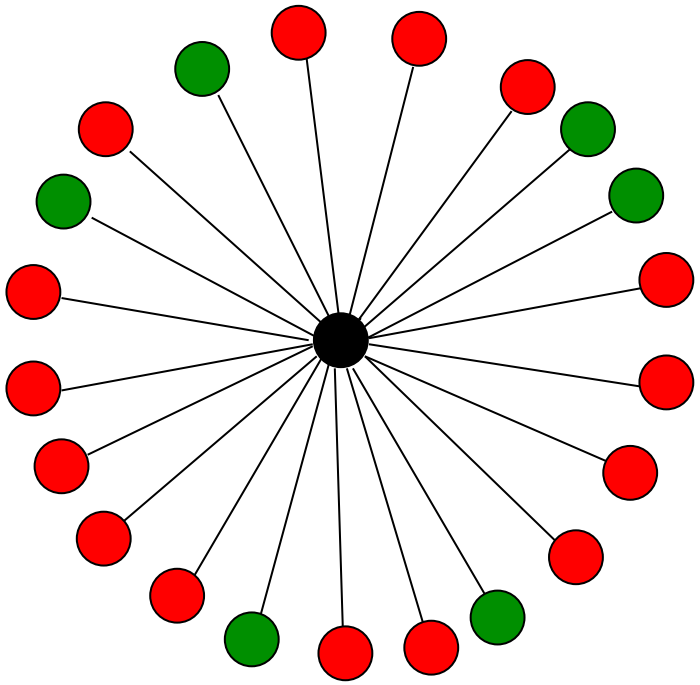
Describe three case studies where the effect of different factors on evolutionary predictablity could be quantified in a simple way

Outline

- A single step of adaptation
- Mutational pathways
- The effect of population size on evolutionary predictability

A single step of adaptation

A single step of adaptation



- The current genotype has access to a set of deleterious and beneficial mutations
- A step of adaptation occurs by fixation of one of the beneficial mutations
- What is the probability that the same mutation is fixed in two replicate populations?

An analogy

- What is the probability that two fair dice show the same number of dots?

An analogy

- What is the probability that two fair dice show the same number of dots?
- What happens to this probability if the dice are loaded?

The probability of parallel evolution

H.A. Orr, Evolution **59**, 216 (2005)

- n beneficial single step mutations are available from the initial genotype
- Each mutant is characterized by its selective advantage $s_i > 0$
- The fixation probability for the i'th mutant is $2s_i$ (Haldane 1927), hence the probability that the i'th mutant is the first to fix is given by

$$\pi_i = \frac{s_i}{\sum_{j=1}^{n} s_j}$$

and the same mutation is fixed in two replicate populations with probability

$$P_2 = \sum_{i=1}^{n} \pi_i^2$$

- This quantity is determined by the distribution of beneficial fitness effects

The extreme value hypothesis

- Gillespie 1983, Orr 2002: Because viable organisms are already very well adapted, fitness distributions of beneficial mutations can be described by extreme value theory (EVT)

- Any distribution falls into one of three EVT classes:
  - Weibull with bounded tails
  - Gumbel with exponential-like unbounded tails (also normal distribution)
  - Fréchet with power-law like heavy tails: $\mathrm{Prob}[s > x] \sim x^{-\alpha}$

- Gumbel class: $P_2 = \frac{2}{n+1}$ which is twice the value $\frac{1}{n}$ expected for fair dice

- P_2 is massively enhanced for heavy-tailed distributions:

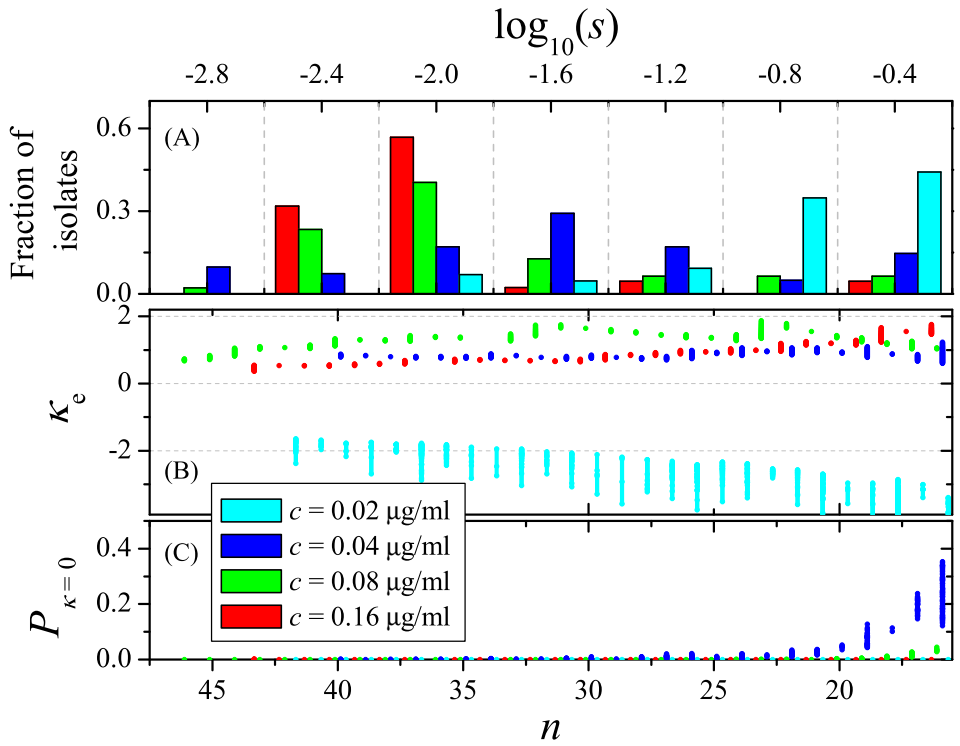
$$P_2 \sim n^{-(\alpha-1)} \gg n^{-1} \text{ for } 1 < \alpha < 2$$

$$P_2 = 1 - \alpha \text{ independent of } n \text{ for } \alpha < 1 \qquad \text{Derrida 1994}$$

Empirical example: The TEM-1 β-lactamase enzyme

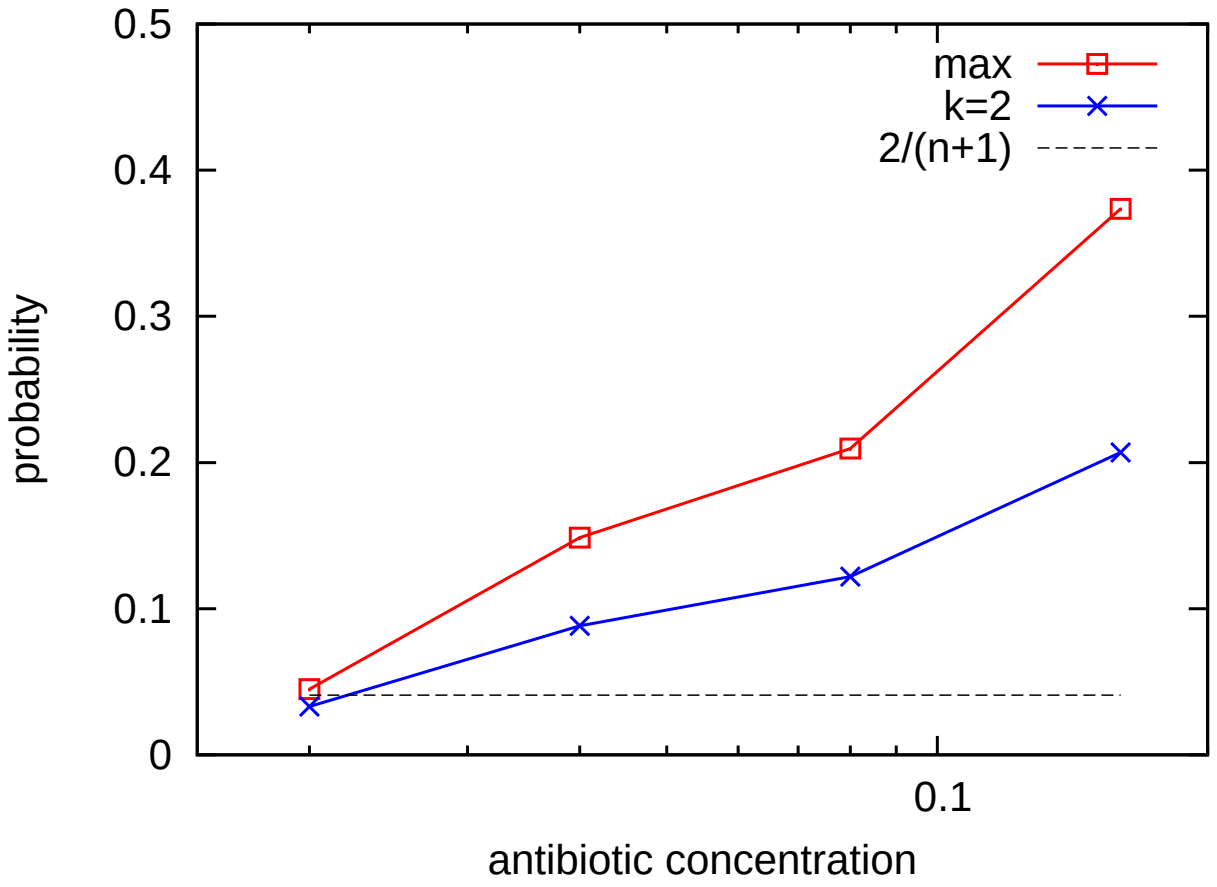
M.F. Schenk, I.G. Szendro, JK, J.A.G.M. de Visser, PLoS Genet. **8** (2012) e1002783

- β-lactamase confers resistance against penicillin to *E. coli*
- 48 out of 2583 point mutations increase resistance against cefotaxime



- Maximum likelihood analysis yields power law distribution with $\alpha \sim 1$

Repeatability measures

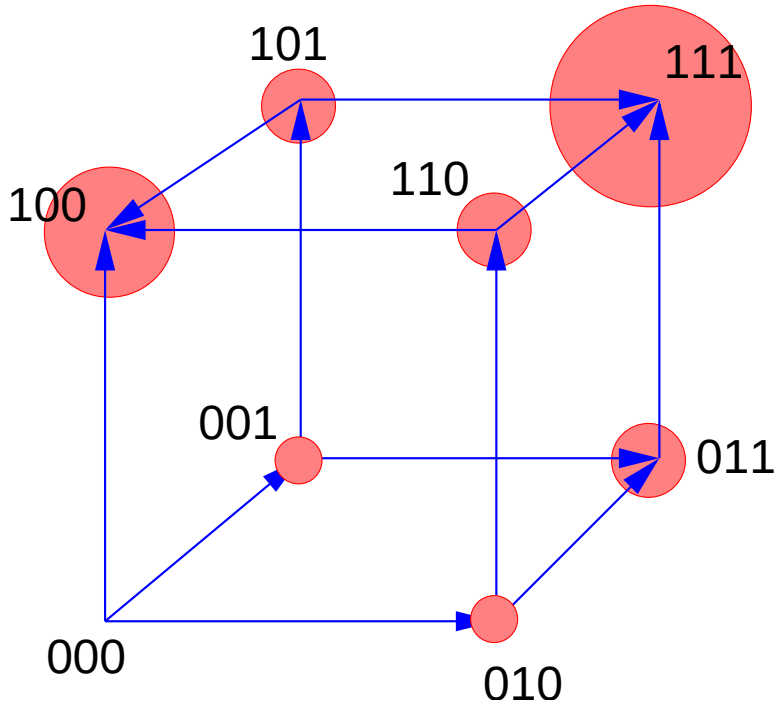


- P_2 and $P_{\max} = \max_i \pi_i > P_2$ increase with antibiotic concentration
- 10 out of 12 replicate lines substitute the largest effect mutation first *in vivo*

Salverda et al., PLoS Genet.**7** (2011) e1001321

Mutational pathways

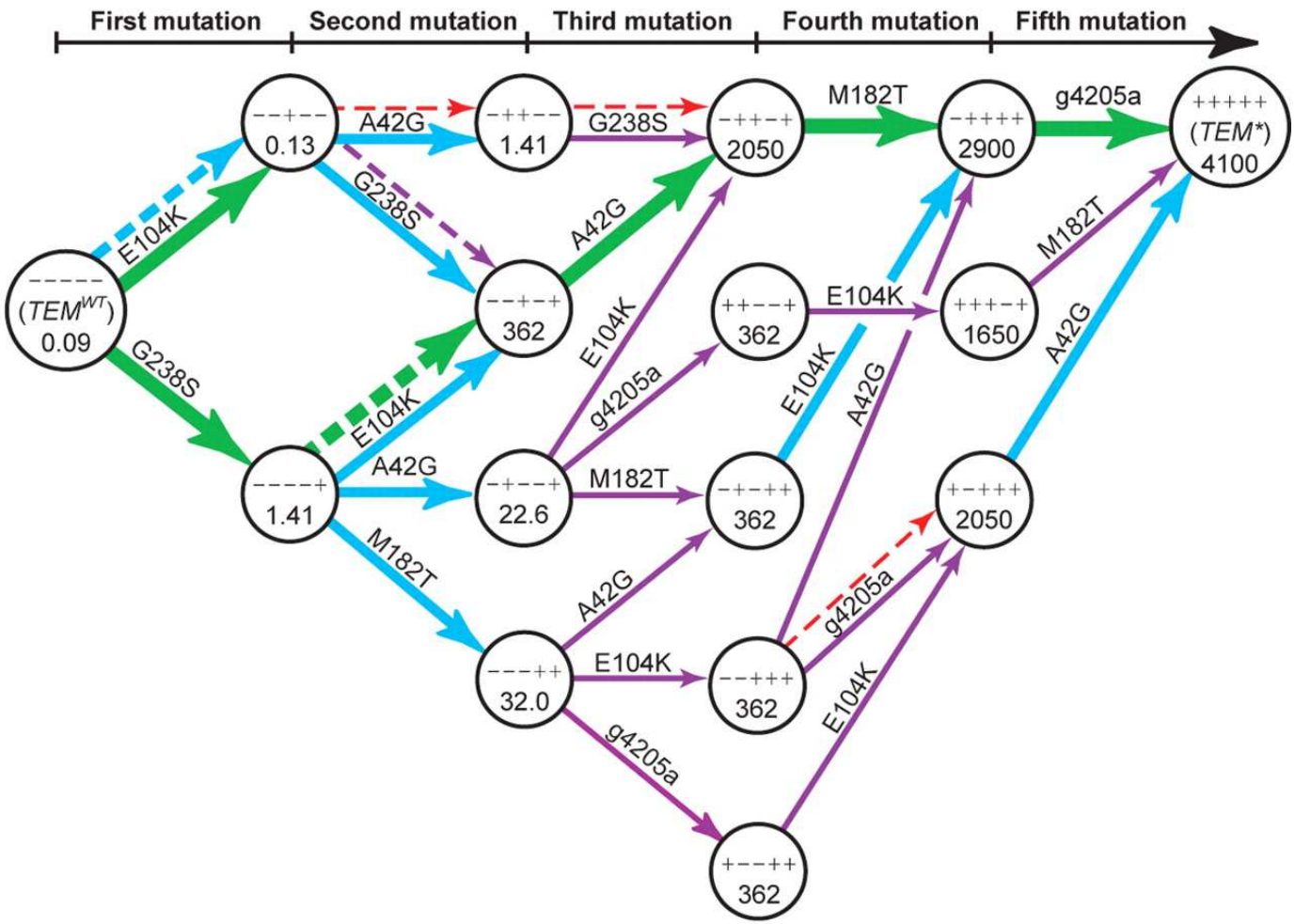
Mutational pathways in fitness landscapes



- $L = 3$ mutational steps from wildtype 000 to adapted type 111
- Mutations can occur in $3 \times 2 \times 1 = 3! = 6$ different orders corresponding to 6 possible pathways
- Only a subset of pathways are “uphill” (= increasing in fitness)

“Darwinian evolution can follow only very few mutational paths to fitter proteins”

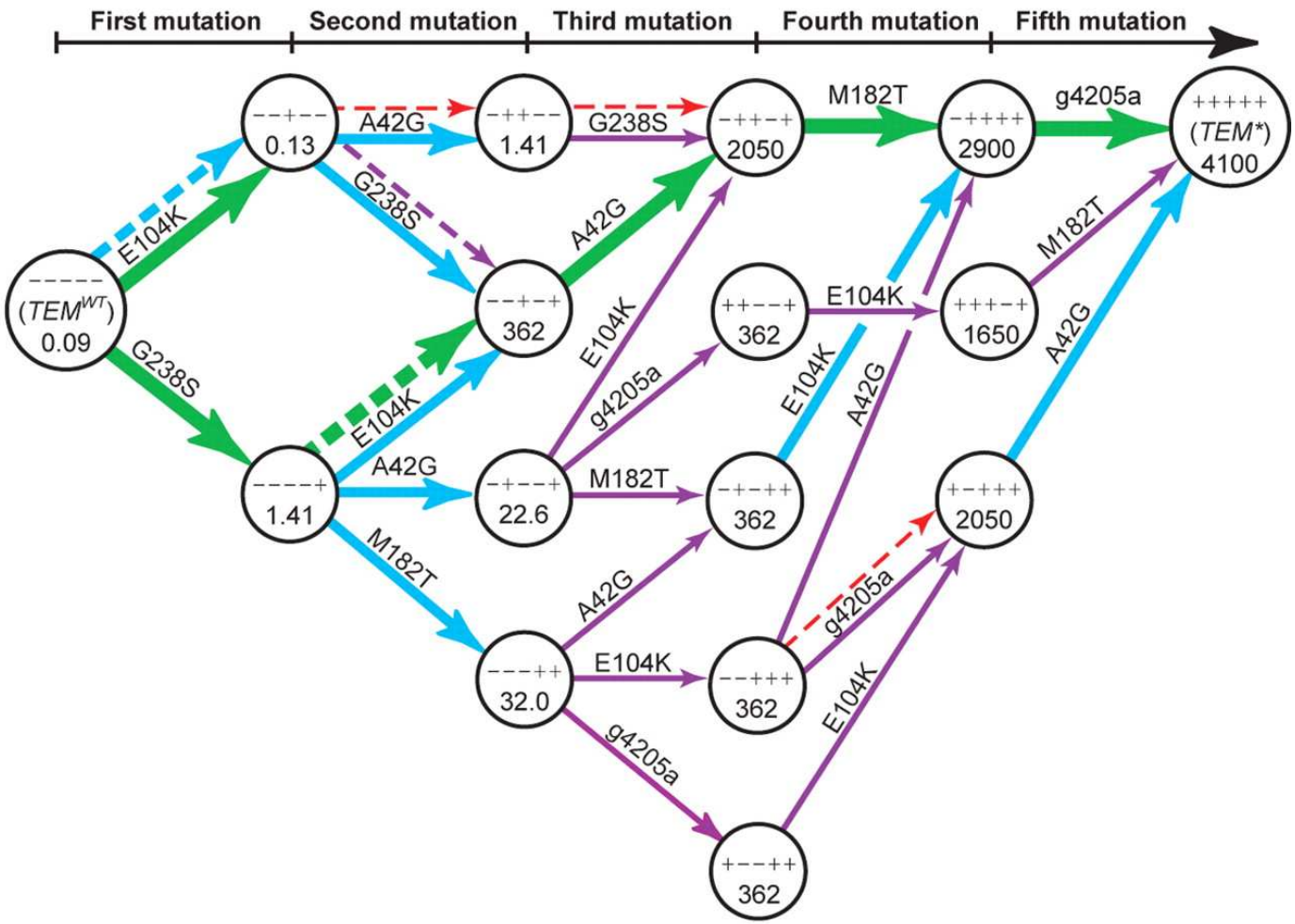
D.M. Weinreich et al., Science **312**, 111 (2006)



- 5 mutations increase resistance of TEM-1 β-lactamase by $\sim 10^5$

“Darwinian evolution can follow only very few mutational paths to fitter proteins”

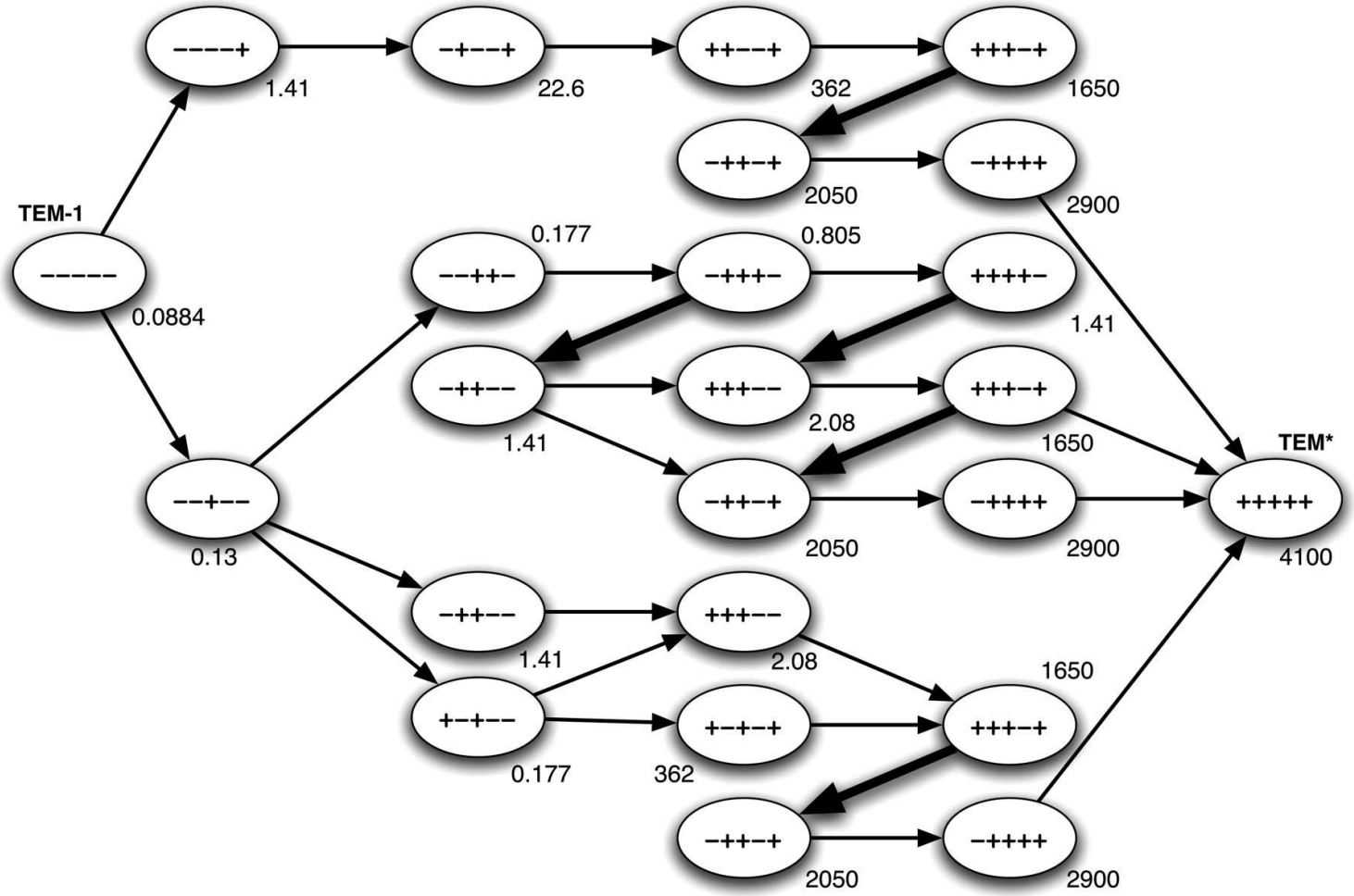
D.M. Weinreich et al., Science **312**, 111 (2006)



- 18 out of $5! = 120$ directed mutational pathways are increasing...

“Darwinian evolution can follow only very few mutational paths to fitter proteins”

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

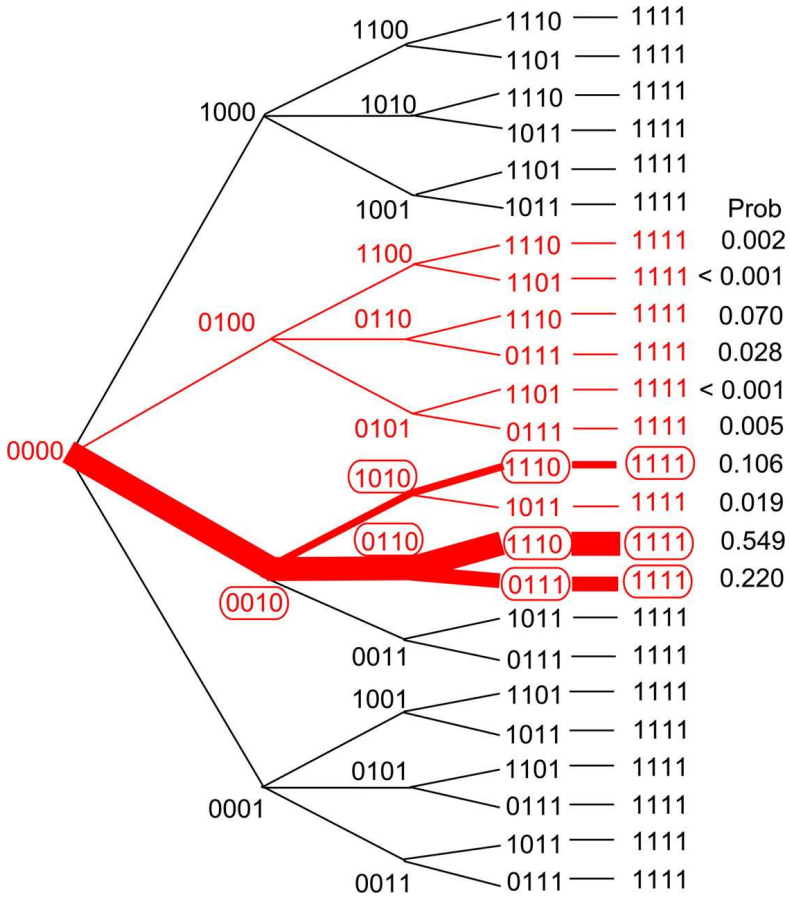


- ...and 27 out of 18651552840 undirected pathways

De Pristo et al. 2007

Pyrimethamine resistance in the malaria parasite

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



- 4! = 24 pathways, 10 (red) are increasing in resistance
- 3 dominant pathways consistent with polymorphisms in natural populations

Accessibility and predictability

- Pathways are accessible if fitness/resistance increases monotonically
- Existence of a small but nonzero fraction of accessible pathways implies high (retrospective) predictability

Accessibility and predictability

- Pathways are accessible if fitness/resistance increases monotonically
- Existence of a small but nonzero fraction of accessible pathways implies high (retrospective) predictability

Questions for theory

- How does accessibility depend on the genetic interactions and on the boundary conditions of the paths?
- How typical is it that a small but nonzero fraction of pathways are accessible?

Null model: House-of-Cards

- In the house-of-cards model fitness is assigned randomly to genotypes, for example, from a uniform distribution Kingman 1978, Kauffman & Levin 1987

- Then the probability that a given path is accessible is $1/L!$ and hence the expected number of accessible paths is Franke et al. 2011

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

Null model: House-of-Cards

- In the house-of-cards model fitness is assigned randomly to genotypes, for example, from a uniform distribution Kingman 1978, Kauffman & Levin 1987

- Then the probability that a given path is accessible is $1/L!$ and hence the expected number of accessible paths is Franke et al. 2011

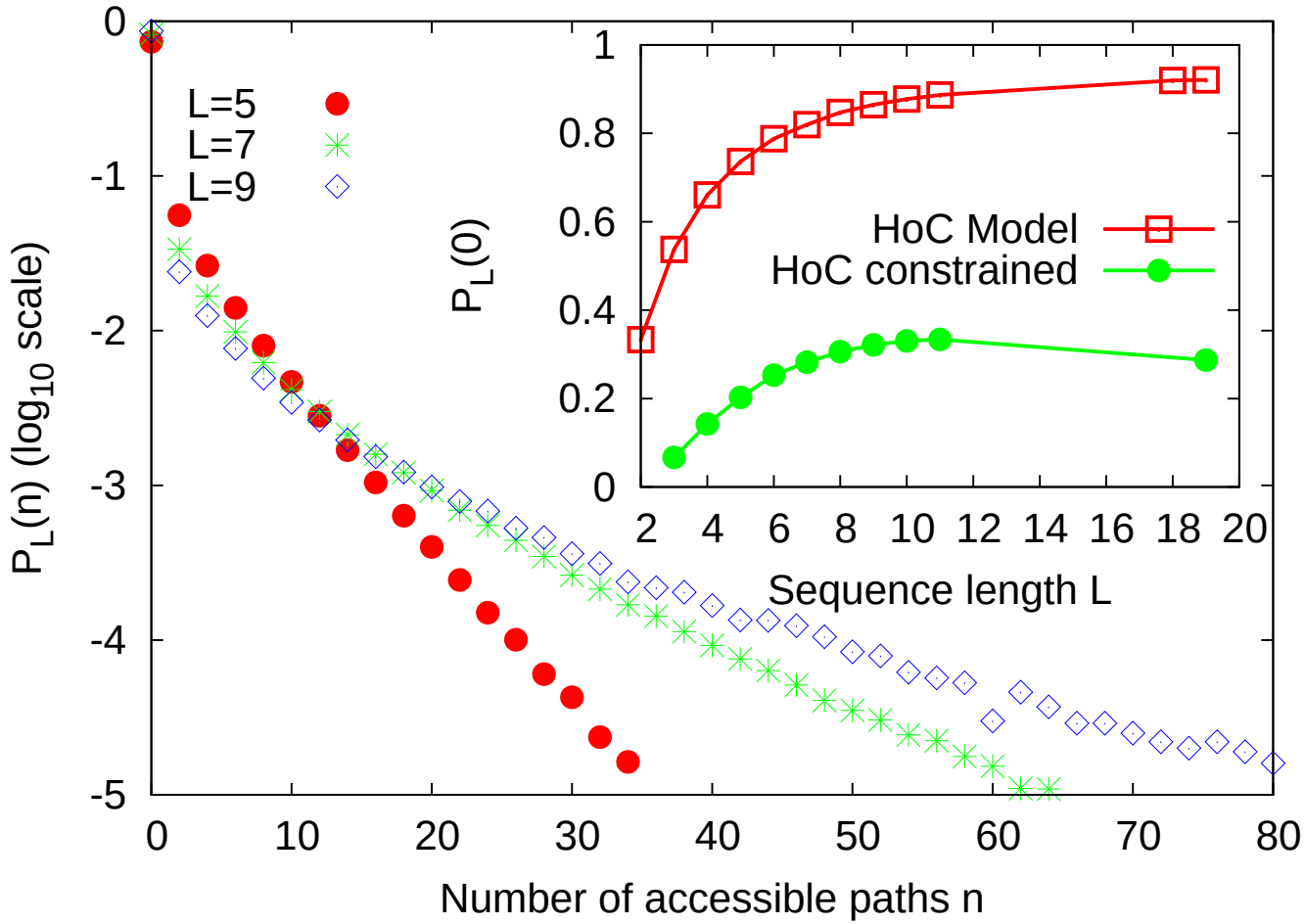
$$\mathbb{E}(n_{\mathrm{acc}}) = L! \times \frac{1}{L!} = 1$$

- This is however misleading, because most landscapes do not possess a single accessible path. The probability for existence of accessible paths decreases asymptotically with the number of mutations as $\log L/L$

- As a consequence, conditioned on accessibility (or low initial fitness) the typical number of paths is of order L

Distribution of the number of accessible paths

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



- Constrained House-of-Cards model has zero fitness at initial genotype

Accessibility percolation

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

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

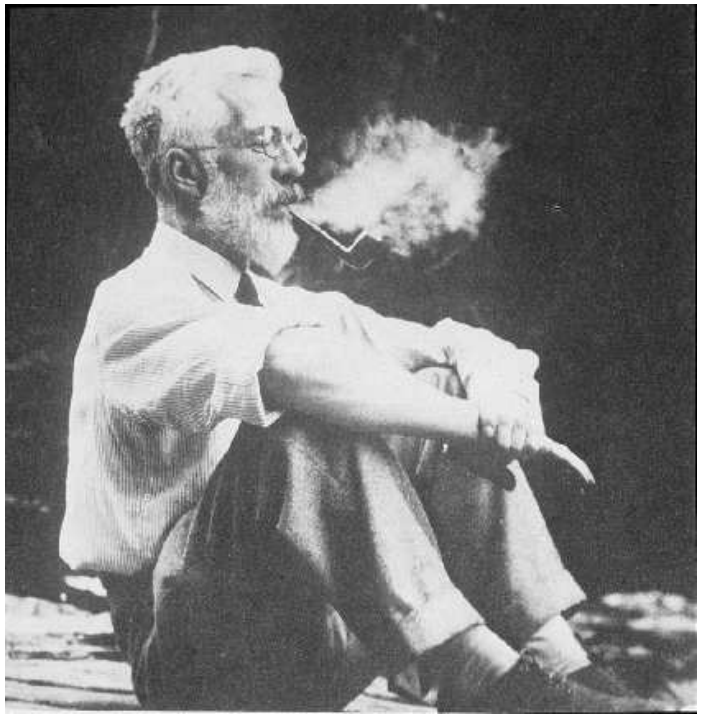
  which diverges/vanishes asymptotically for large L when $f_0 < \frac{\ln L}{L}$ / $f_0 > \frac{\ln L}{L}$

- 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

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

- Introducing correlations into the fitness landscape may or may not increase accessibility B. Schmiegelt, JK, J. Stat. Phys. **154**, 334 (2014)

Predictability and population size

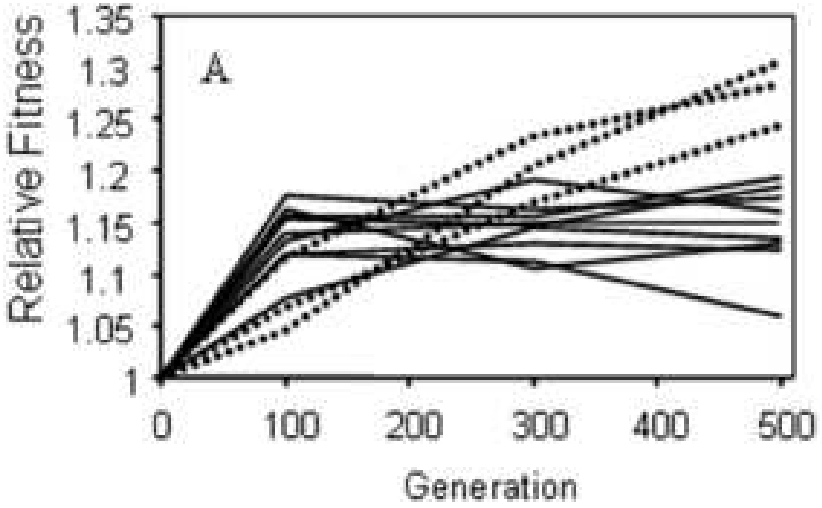


"The regularity of the [rate of adaptation] is in fact guaranteed by the same circumstance which makes a statistical assemblage of particles, such as a bubble of gas obey, without appreciable deviation, the law of gases. A visible bubble will indeed contain several billions of molecules, and this would be a comparatively large number for an organic population, but the principle ensuring regularity is the same." Ronald A. Fisher (1958)

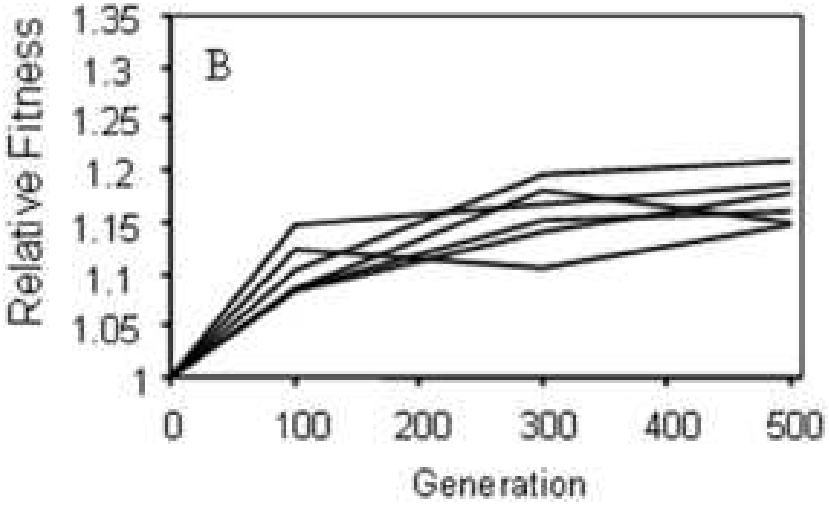
Determinism of fitness trajectories increases with population size

Rozen, Habets, Handel, de Visser, PLoS ONE **3**, e1715 (2008)

$N = 5 \times 10^5$



$N = 2.5 \times 10^7$



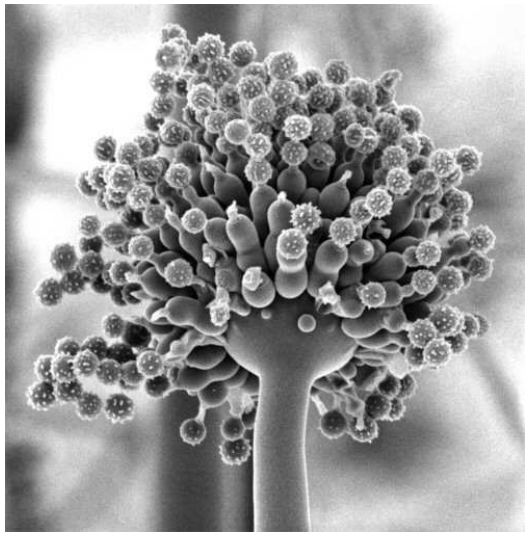
- Small and large populations of *E. coli* evolved in Luria-Bertani nutrient
- Fitness trajectories of large populations are more uniform
- This is to be expected, because the competition of multiple clones in a large population makes adaptation more greedy

K. Jain, S.-C. Park, JK, 2011

The *Aspergillus niger* fitness landscape

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

- $L = 8$ individually deleterious marker mutations residing on different chromosomes of *Aspergillus niger* (black mold) de Visser et al. 1997

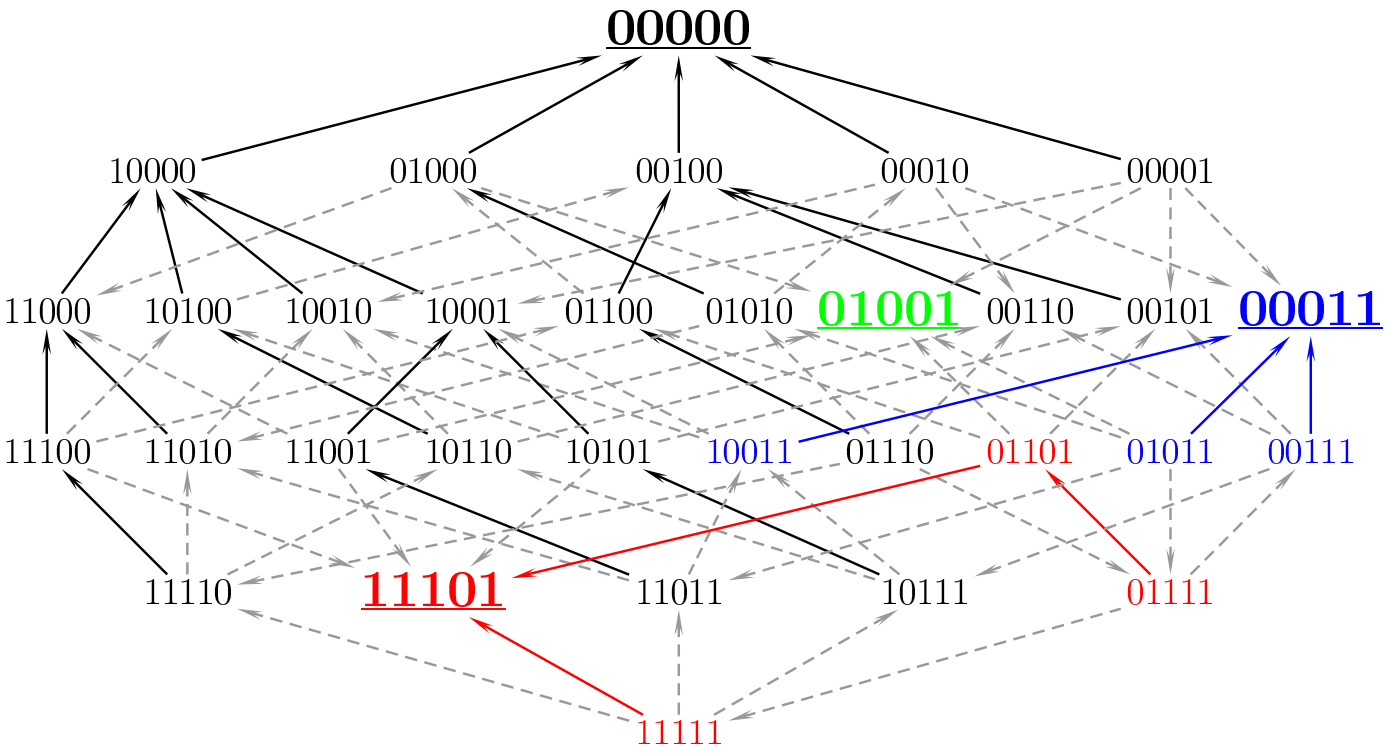


- 186 out of $2^8 = 256$ possible combinations were isolated in $\sim$ 2500 trials

- Fitness (= growth rate) was measured for two replicates per strain, and zero fitness assigned to missing strains

The *Aspergillus niger* fitness landscape

J.A.G.M. de Visser, S.C. Park, JK, American Naturalist **174**, S15 (2009)



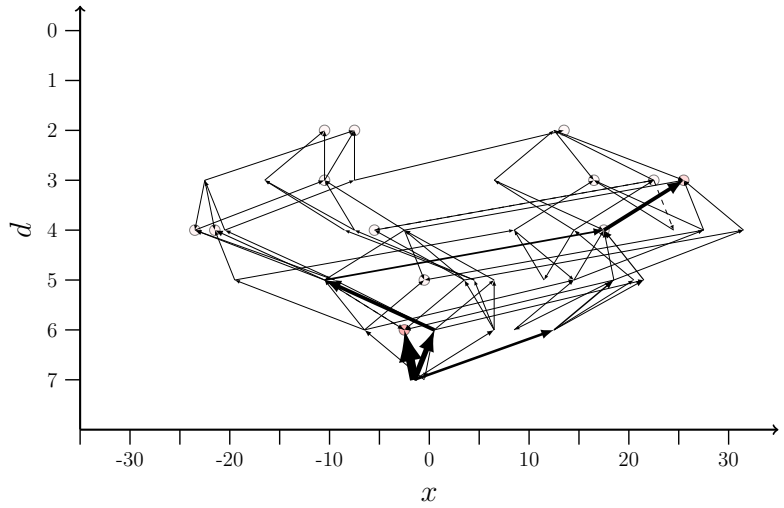
- One out of $\binom{8}{5} = 56$ five-dimensional subsets
- Arrows point to increasing fitness, 3 local fitness optima highlighted

Evolutionary dynamics on the *A. niger* landscape

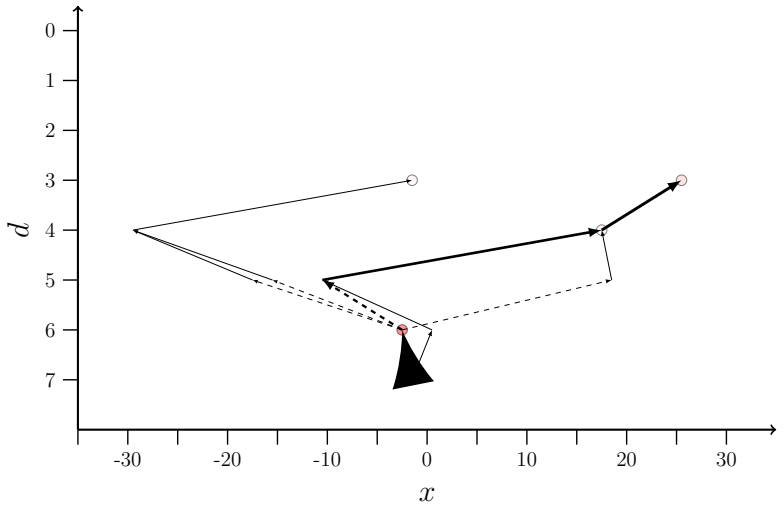
I.G. Szendro, J. Franke, J.A.G.M. de Visser, JK, PNAS **110**:571 (2013)

- Fixed number N of individuals reproduce asexually in discrete generations (Wright-Fisher model)
- Mutations occur with probability μ per site and generation
- Evolution starts from a viable genotype at mutational distance d_0 from the global optimum
- Two types of evolutionary trajectories:
  - Lines of descent: Track first appearance of mutations
  - Paths of the most populated genotype (not necessarily continuous)
- Quantify predictability by the entropy of the distribution of pathways or endpoints, averaged over a large number of evolutionary runs

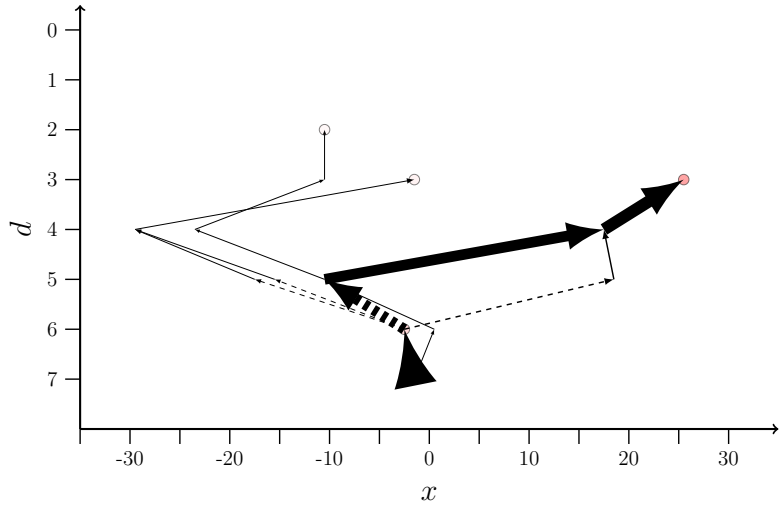
Lines of descent at different population sizes



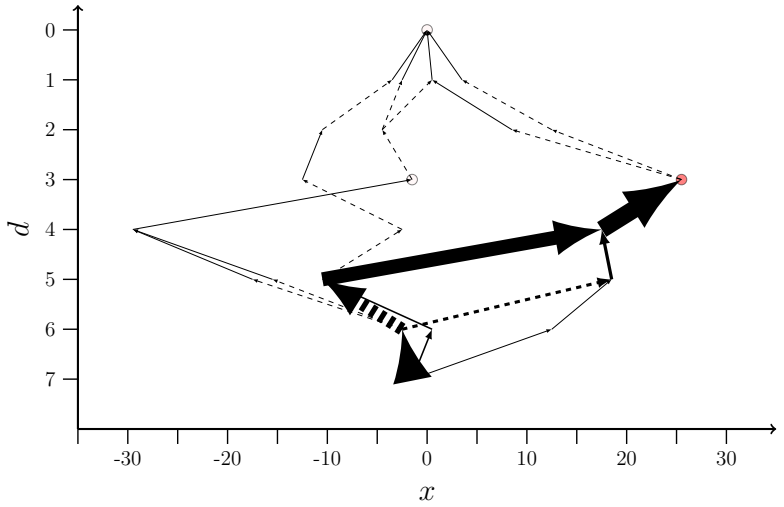
(a) $N = 2^7$



(b) $N = 2^{14}$



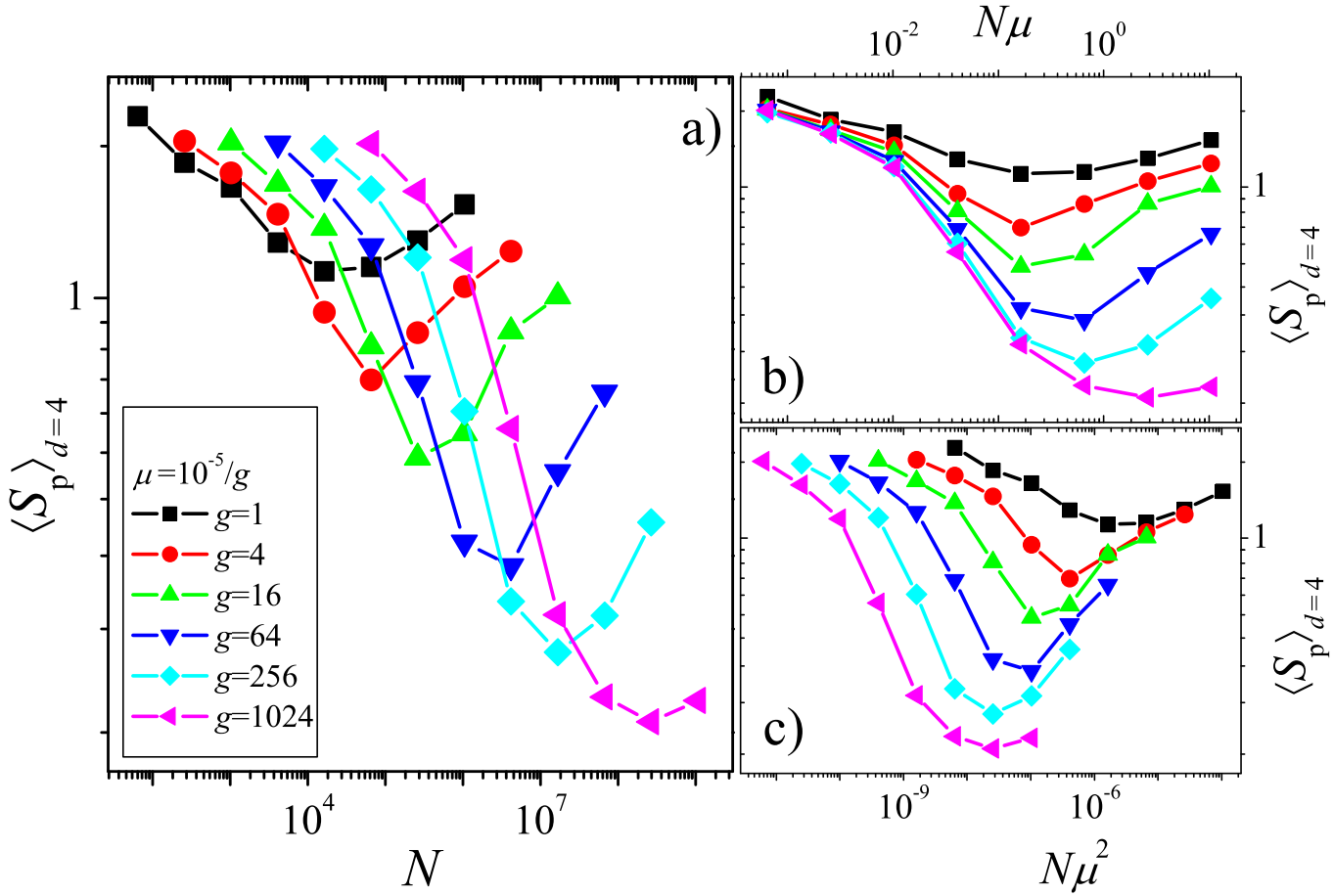
(c) $N = 2^{17}$



(d) $N = 2^{23}$

$d_0 = 7$, $\mu = 10^{-5}$, $N = 128, \ldots, 8 \times 10^6$, 32768 generations

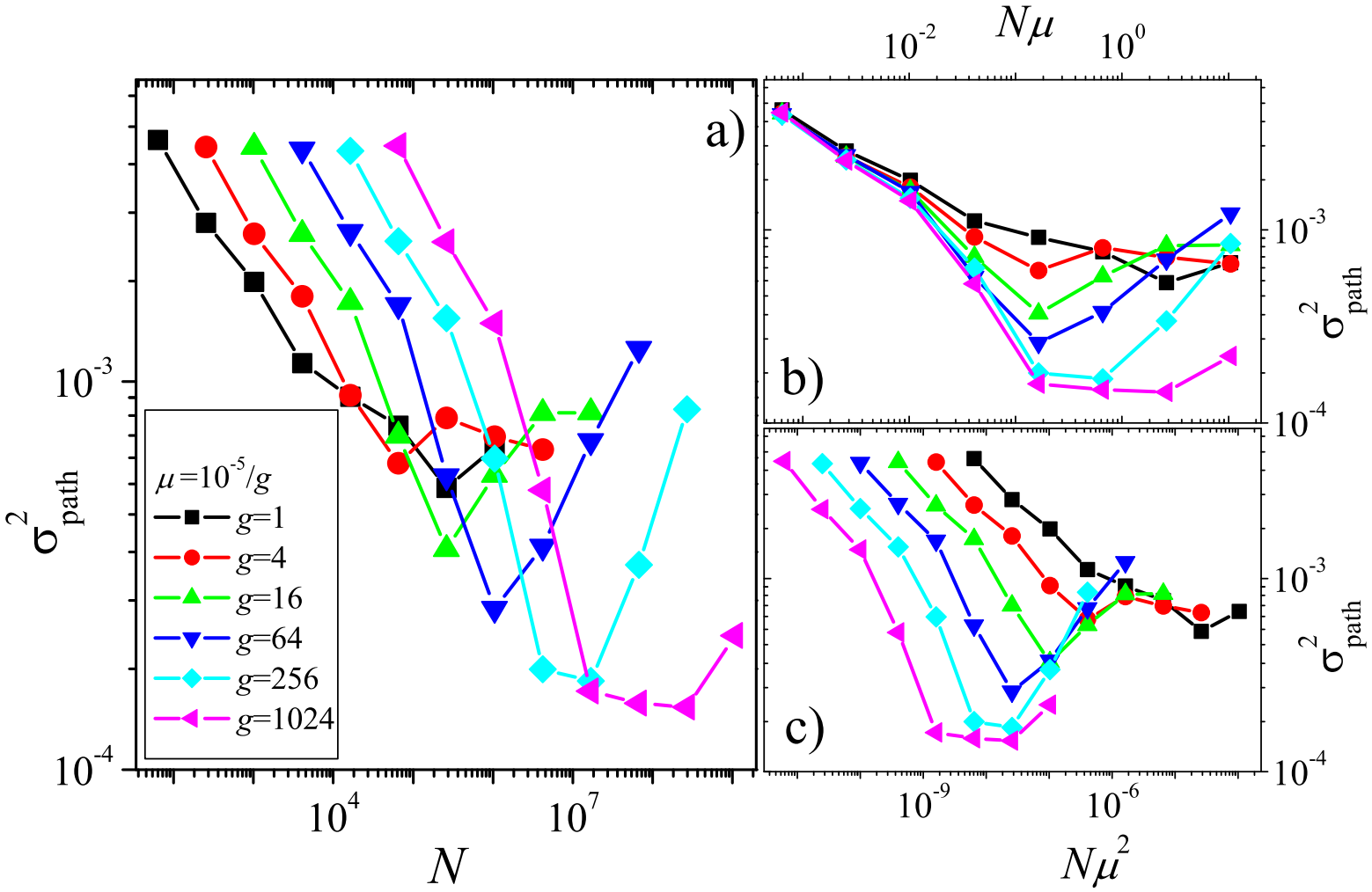
Pathway entropy



- Pathway entropy averaged over 46 starting points at distance $d_0 = 4$
- Entropy varies non-monotonically with population size because double mutants occurring at rate $\sim N\mu^2$ open up new pathways

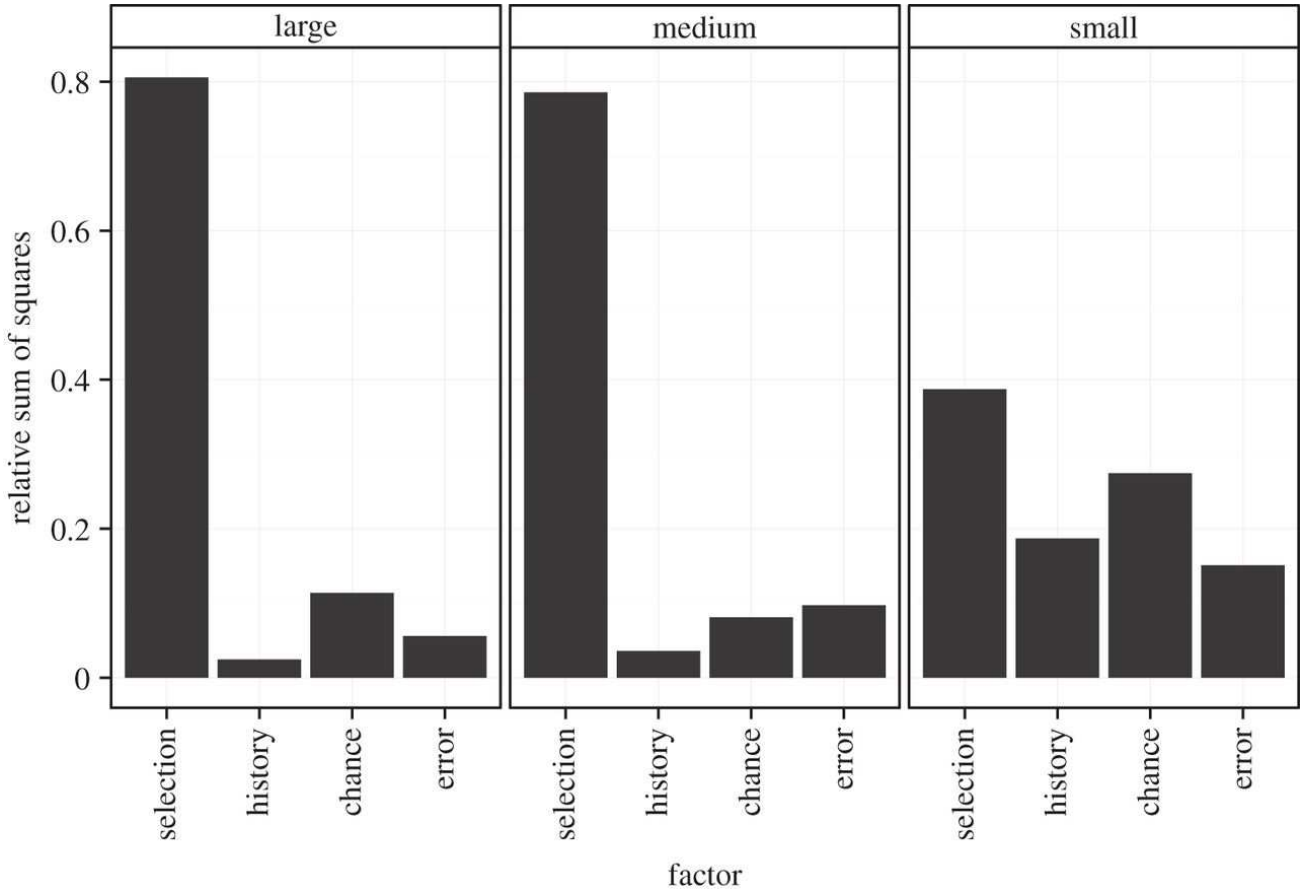
Variability of fitness trajectories

I.G. Szendro (unpublished)



An experimental test

Lachapelle et al., Proc. Roy. Soc B (2015)



- 108 populations of *Chlamydomonas reinhardtii* adapting to salt from 6 different initial genotypes
- Population sizes $N = 5 \times 10^3$, 5×10^4 and 4×10^5,

Summary

Factors contributing to evolutionary predictability:

- Distribution of beneficial fitness effects
- Epistatic interactions determining the accessibility of the fitness landscape
- Mutation supply, as determined by population size and mutation rate

For further discussion see J.A.G.M. de Visser, JK, Nat. Rev. Gen. 15 (2014) 480

Summary

Factors contributing to evolutionary predictability:

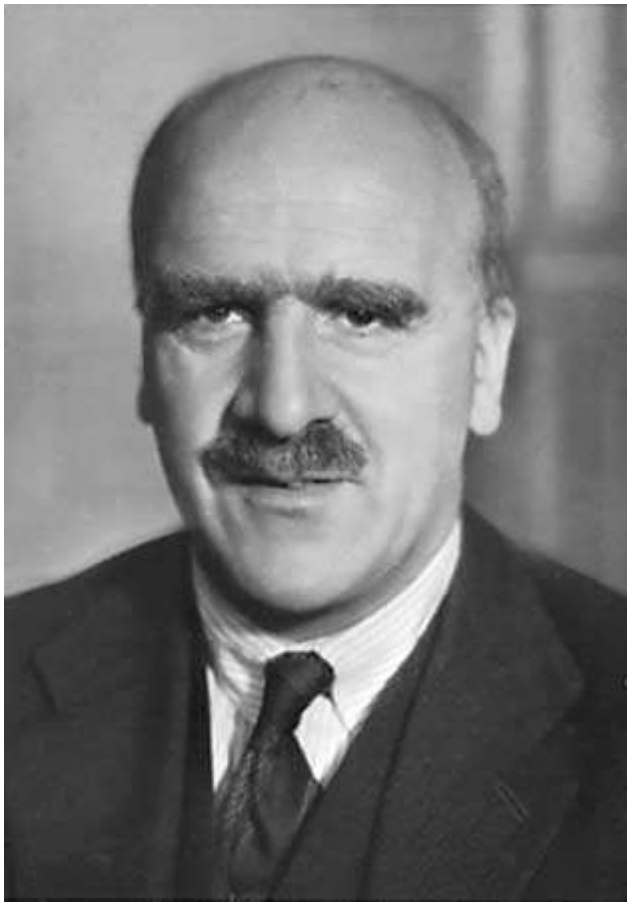
- Distribution of beneficial fitness effects
- Epistatic interactions determining the accessibility of the fitness landscape
- Mutation supply, as determined by population size and mutation rate

For further discussion see J.A.G.M. de Visser, JK, Nat. Rev. Gen. 15 (2014) 480

Thanks to

- Jasper Franke, Johannes Neidhart, Stefan Nowak, Benjamin Schmiegelt, Ivan Szendro (Cologne)
- Arjan de Visser (Wageningen)

Why predictability?



“No scientific theory is worth anything unless it enables us to predict something which is actually going on. Until that is done, theories are a mere game of words, and not such a good game as poetry.”

J.B.S. Haldane (1937)