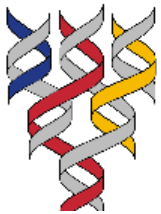


SFB 680
Molecular Basis of
Evolutionary Innovations

The geometry of evolution:
Statistical topography of biological fitness landscapes

Joachim Krug
Institute of Theoretical Physics, University of Cologne, Germany

joint work with Jasper Franke, Ivan Szendro, Arjan de Visser and Martijn Schenk

General Physics Colloquium, Göteborg, March 15, 2012

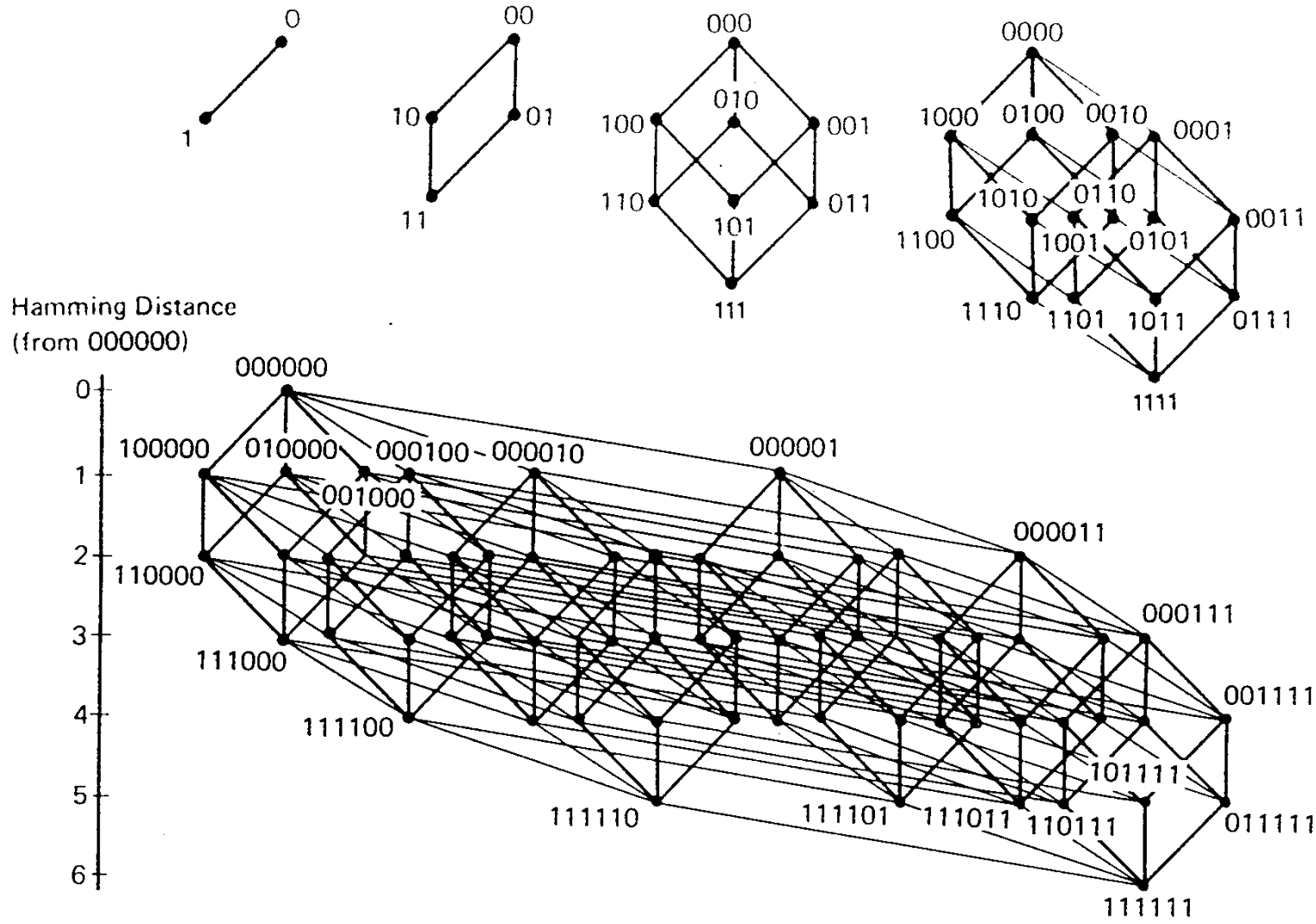
Sequence spaces

- Watson & Crick 1953: Genetic information is encoded in DNA-sequences consisting of **A**denine, **C**ytosine, **G**uanine and **T**hymine

..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 $L \sim 10^2$
- Classical genetics: L genes that are present as different alleles;
  distinguish between wild type (0) and mutant (1) $\Rightarrow$ binary sequences
- Genotypic distance: Two sequences are nearest neighbors if they differ in a single letter (mutation)

Binary sequences spaces are hypercubes

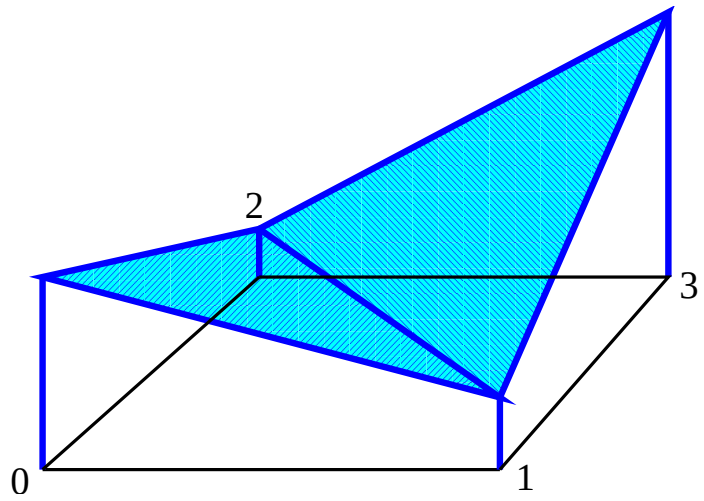


Fitness landscapes

- **Fitness** is a measure of reproductive success of an organism (e.g., number of offspring in the next generation)

- A fitness landscape assigns a fitness value $w(\sigma)$ to each genotype sequence σ

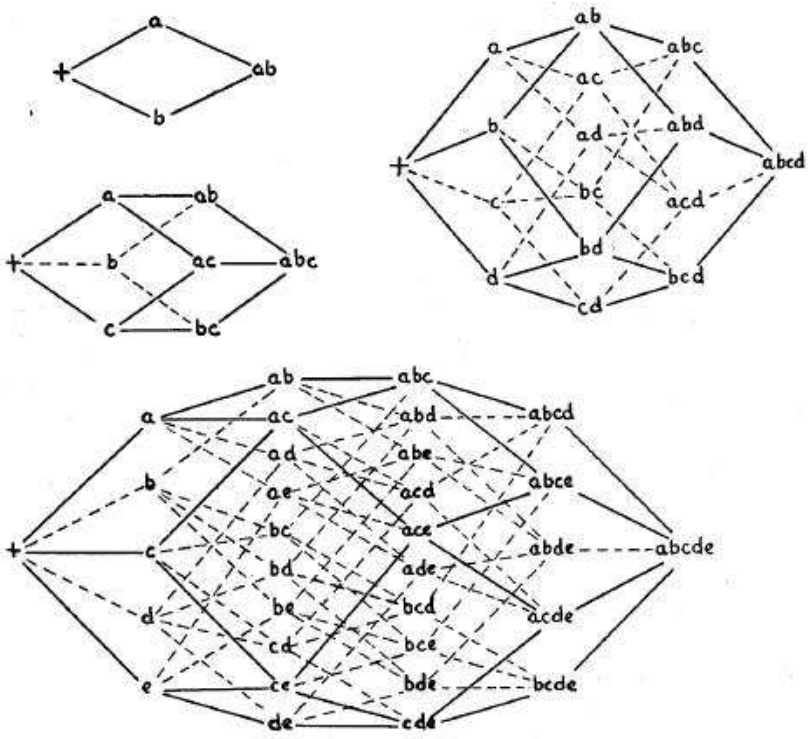
- Example: $L = 2$

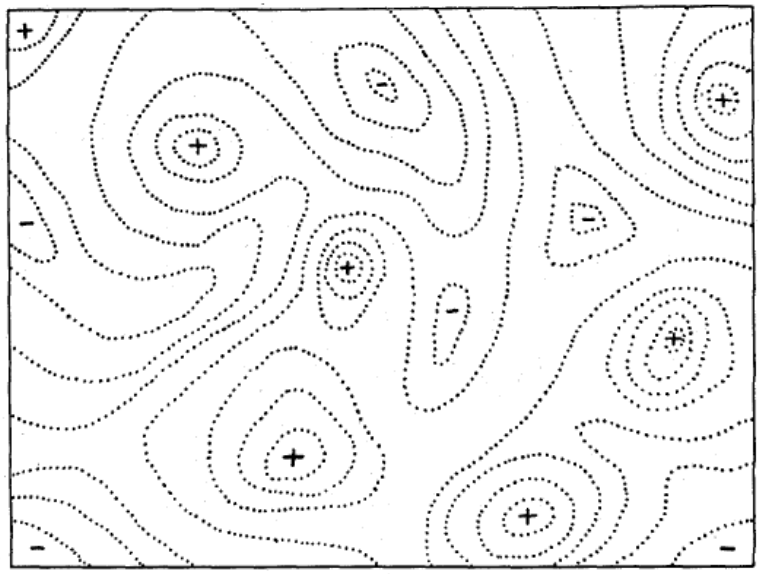


- Evolution is a hill-climbing process in the fitness landscape

Fitness landscapes

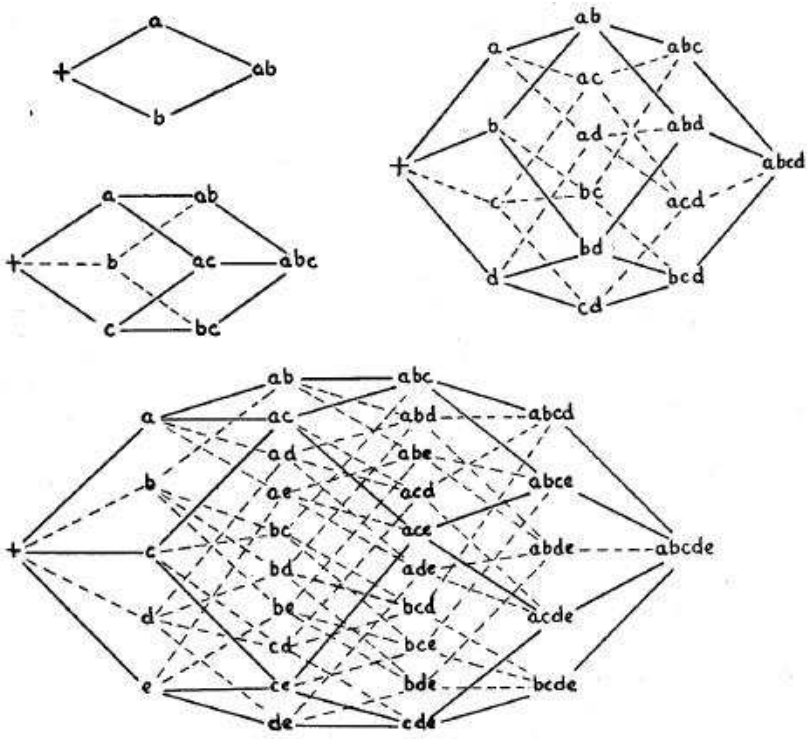
S. Wright, Proc. 6th Int. Congress of Genetics (1932)

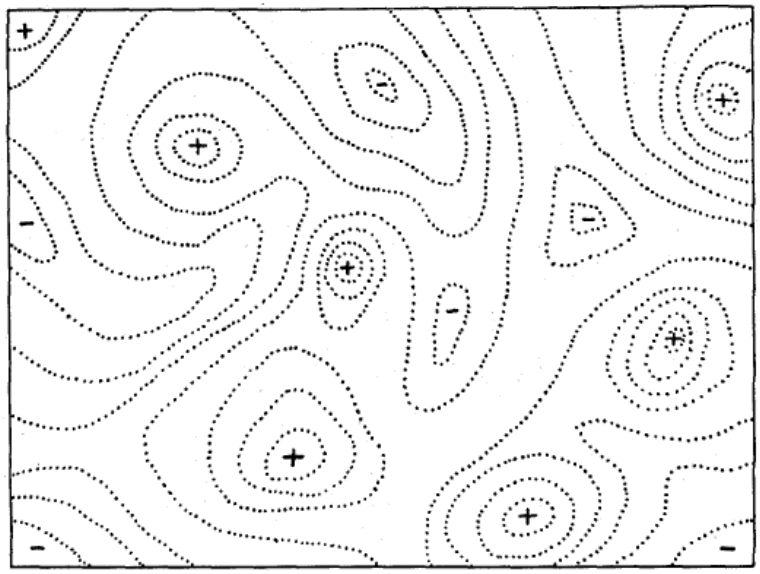




Fitness landscapes

S. Wright, Proc. 6th Int. Congress of Genetics (1932)

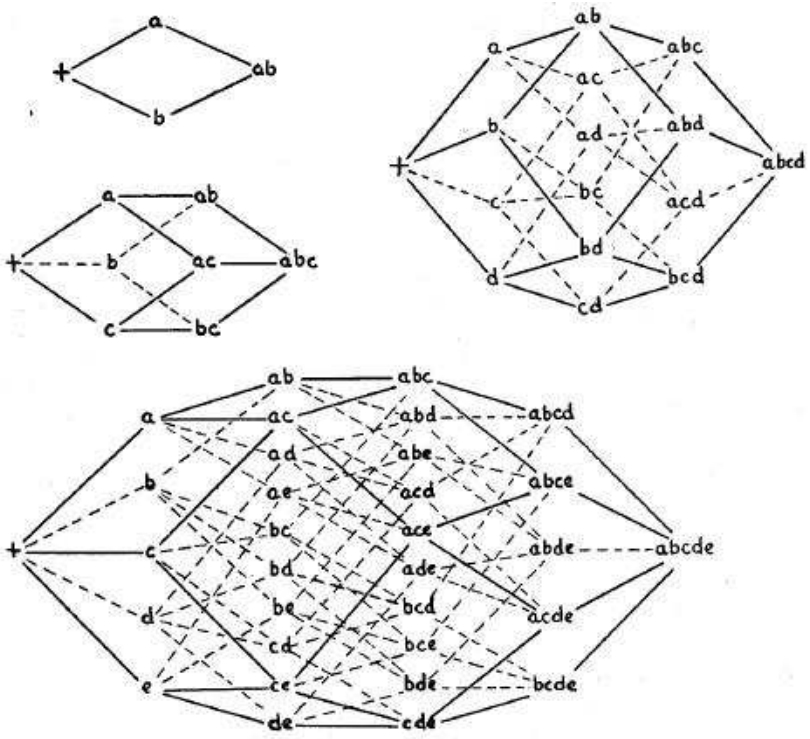


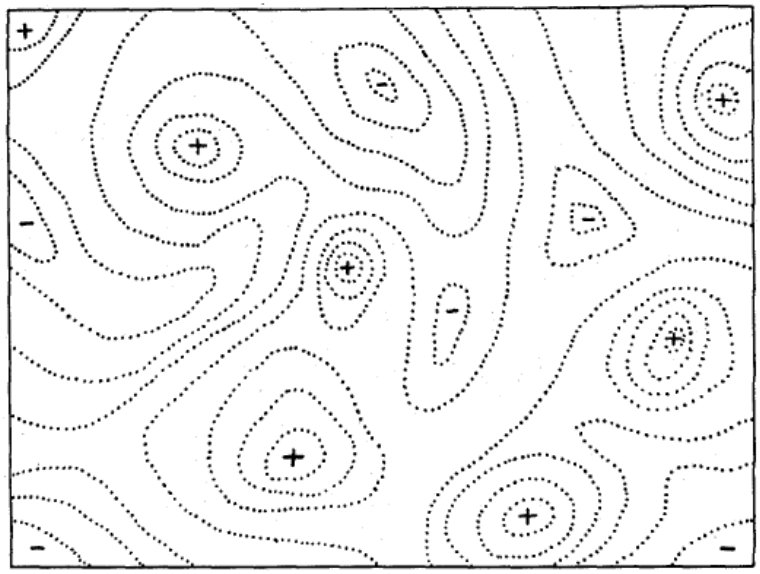


”The problem of evolution as I see it is that of a mechanism by which the species may continually find its way from lower to higher peaks in such a field.”

Fitness landscapes

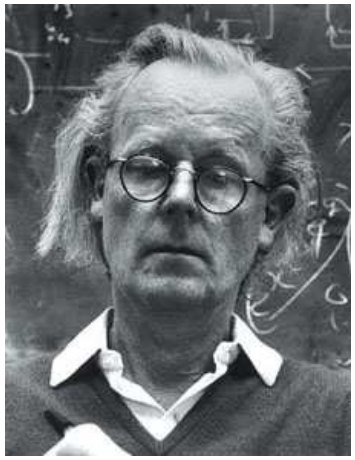
S. Wright, Proc. 6th Int. Congress of Genetics (1932)





"The two dimensions of figure 2 are a very inadequate representation of such a field."

Evolutionary pathways



John Maynard Smith

”The model of protein evolution I want to discuss is best understood by analogy with a popular word game. The object of the game is to pass from one word to another of the same length by changing one letter at the time, with the requirement that all the intermediate words are meaningful in the same language. Thus WORD can be converted into GENE in the minimum number of steps, as follows:

$$\text{WORD} \rightarrow \text{WORE} \rightarrow \text{GORE} \rightarrow \text{GONE} \rightarrow \text{GENE}$$

This is an analogue of evolution, in which the words represent proteins.”

Nature 225:563 (1970)

Two conflicting intuitions:

- Proliferation of fitness peaks severely hampers evolution in high-dimensional genotype spaces; valley crossing is crucial.
  (Wright, Kauffman,...)

- Proliferation of possible pathways implies high evolutionary accessibility in high dimensional spaces; valley crossing is not an issue.
  (Fisher, Gavrilets,...)

Questions:

- How can the structure of high-dimensional fitness landscapes be quantified and modeled?

- How does the fitness landscape topography constrain the evolutionary process?

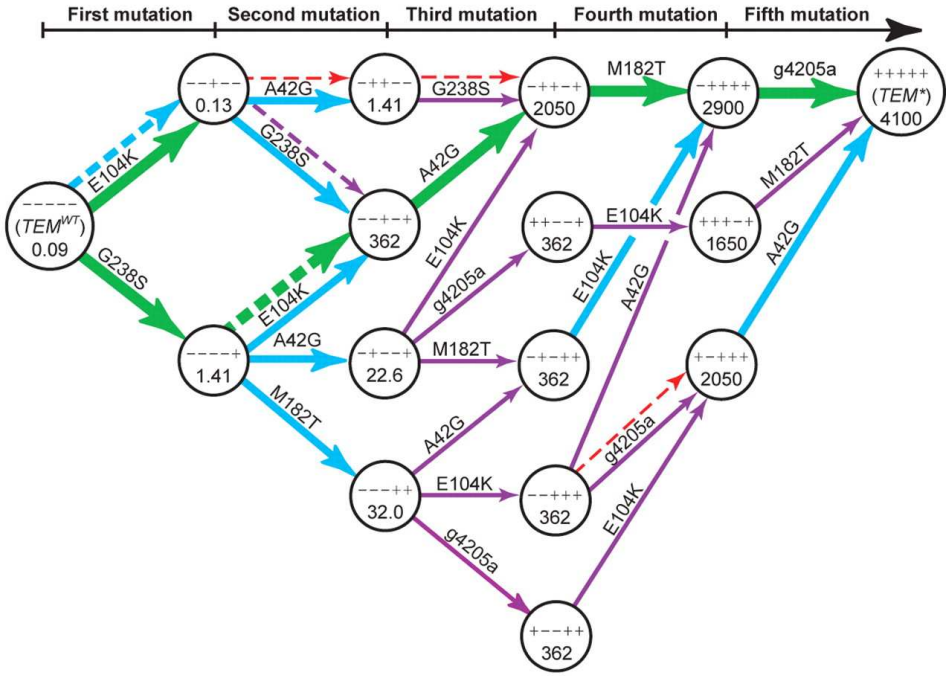
- What do real fitness landscapes look like?

Empirical fitness landscapes

Review: I.G. Szendro, M.F. Schenk, J. Franke, JK, J.A.G.M. de Visser, arXiv:1202.4378

Example 1: The TEM1 β-lactamase resistance landscape

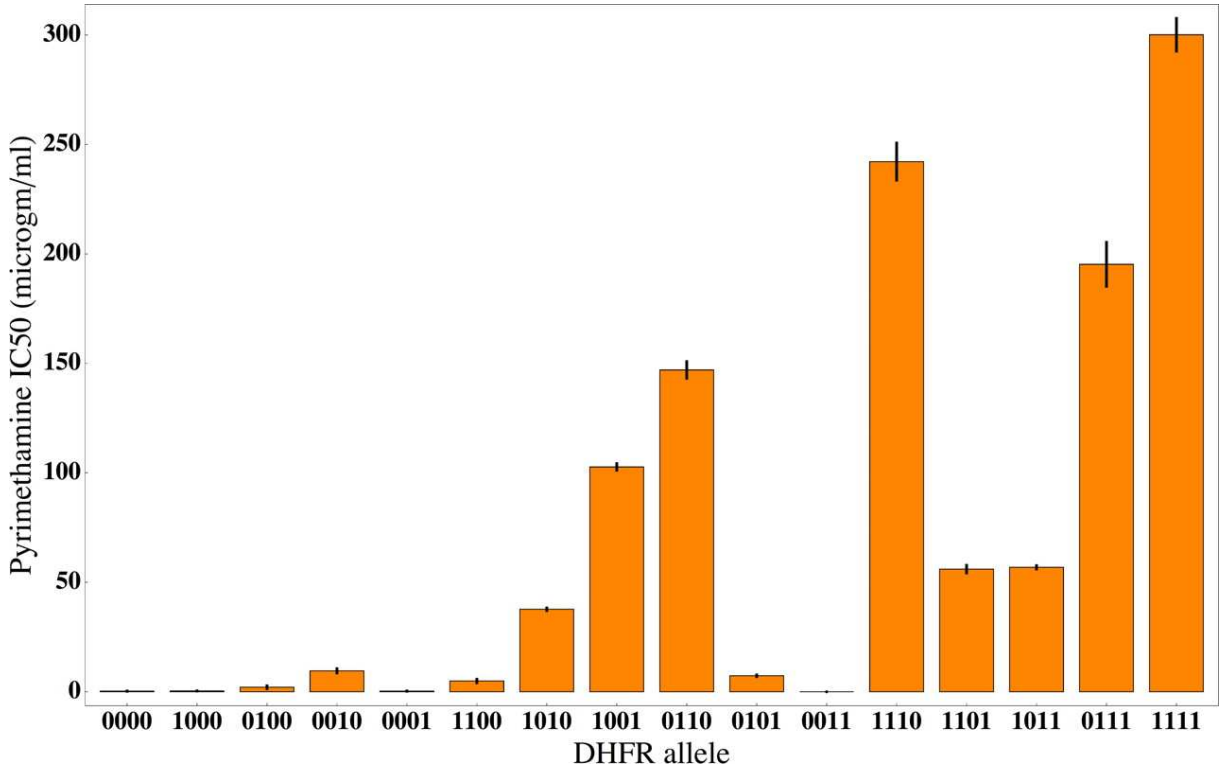
D.M. Weinreich, N.F. Delaney, M.A. De Pristo, D.L. Hartl, Science **312**, 111 (2006)



- Accessible mutational pathways are monotonically increasing in resistance
- 102 out of $5! = 120$ paths from the wildtype to the fivefold mutant are inaccessible

Example 2: Pyrimethamine resistance in the malaria parasite

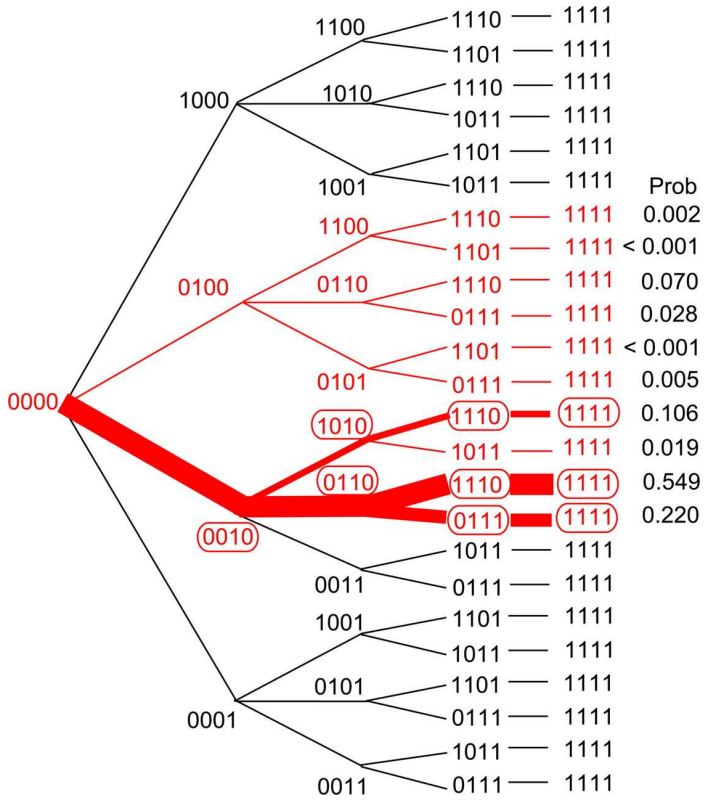
E.R. Lozovsky et al., Proc. Natl. Acad. Sci. USA **106**, 12025 (2009)



- 4 mutations in the dihydrofolate reductase confer resistance to an important malaria drug
- One local maximum at 1001

Example 2: Pyrimethamine resistance in the malaria parasite

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

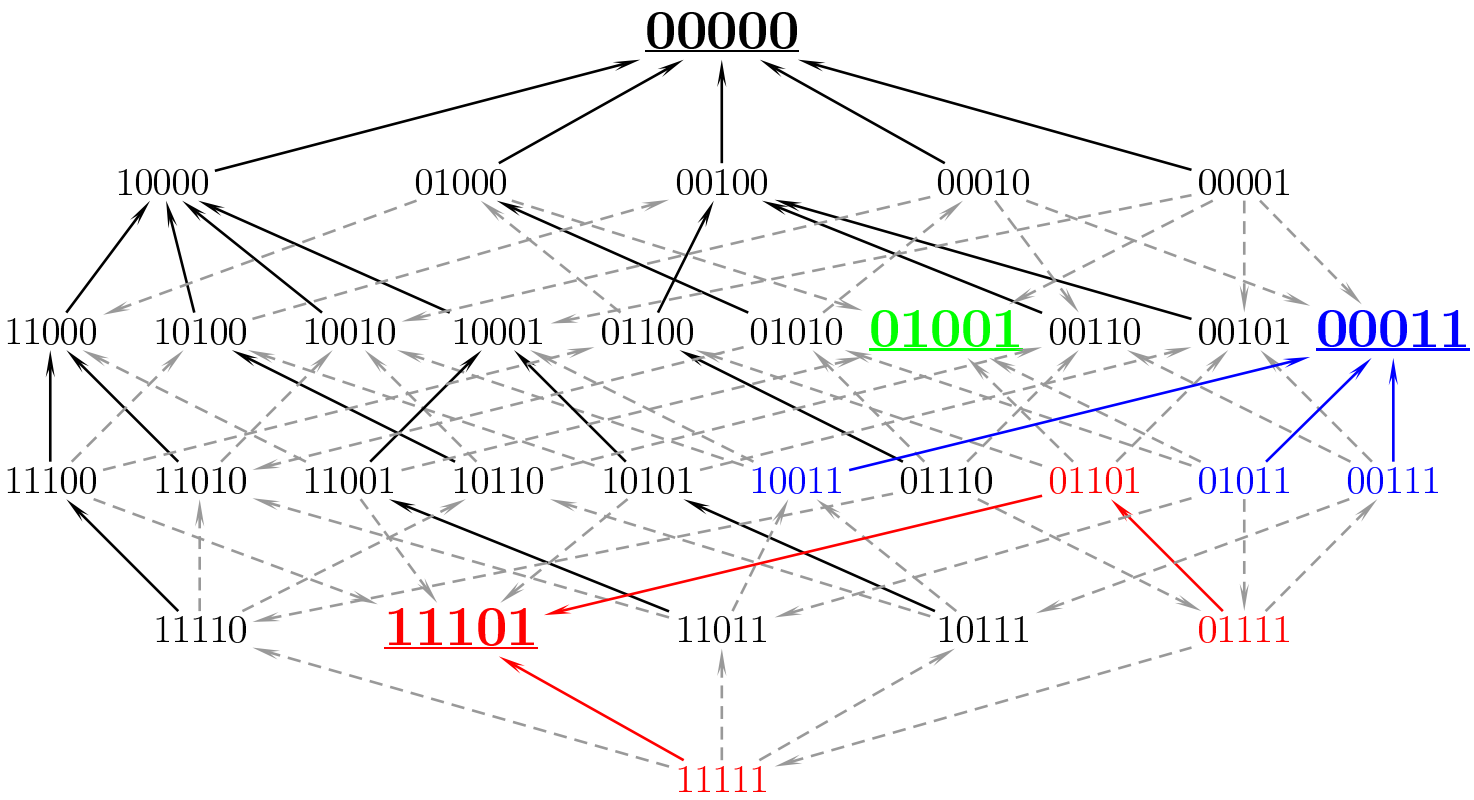


- 14 out of $4! = 24$ paths from the wildtype to the fourfold mutant are inaccessible

- Dominating pathways consistent with polymorphisms in natural populations

Example 3: The *Aspergillus niger* fitness landscape

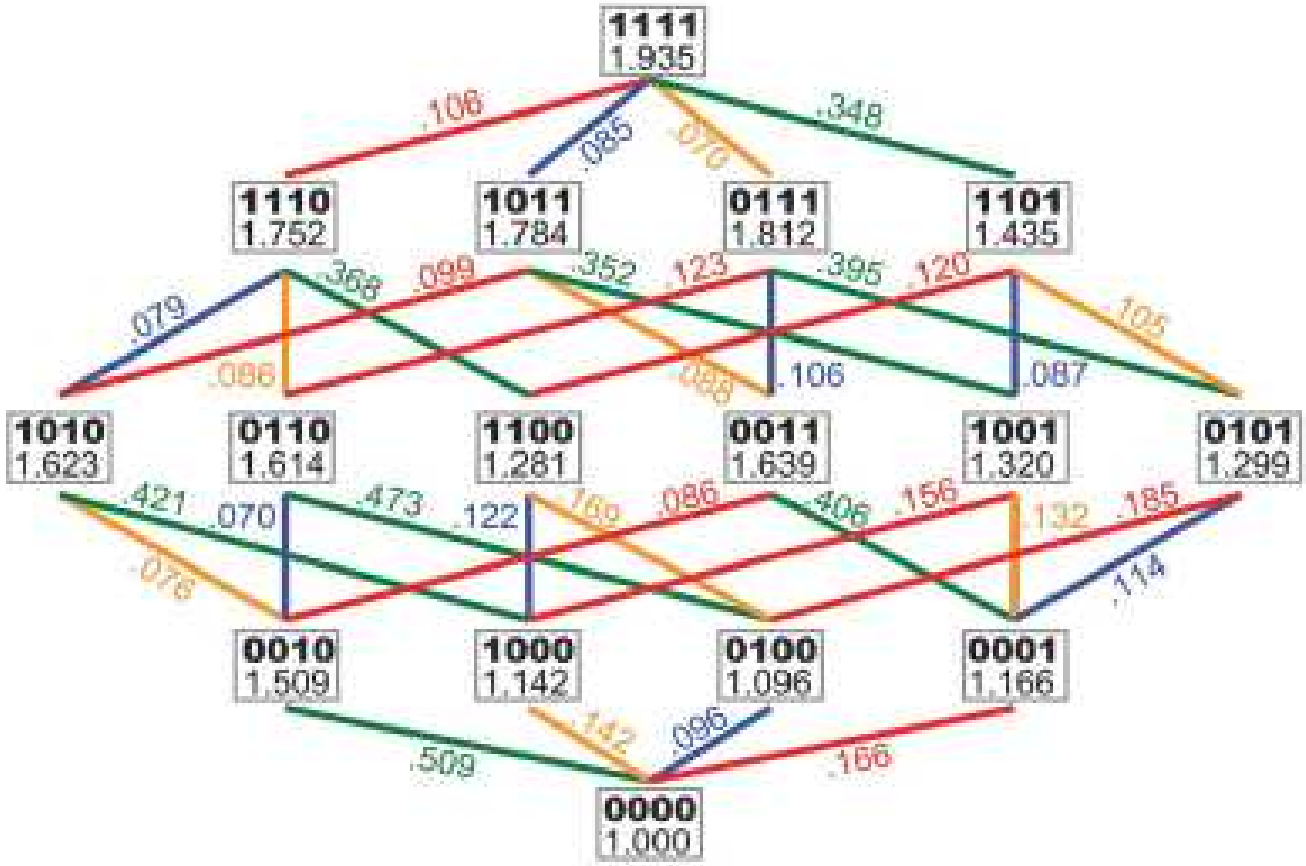
J.A.G.M. de Visser, S.C. Park, JK, American Naturalist **174**, S15 (2009)



- 5 individually deleterious marker mutations in different chromosomes
- 95 out of 120 paths from the fivefold mutant to the wildtype are inaccessible

Example 4: Adaptive mutations in *Methylobacterium extorquens*

Chou et al., Science **332**, 1190 (2011)



- All pathways are accessible

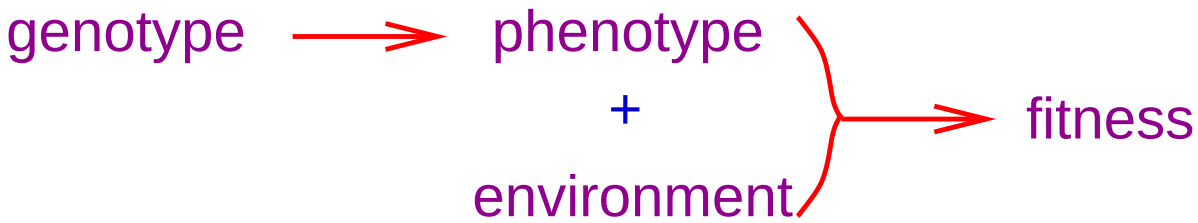
Evolutionary accessibility in model landscapes

J. Franke, A. Klözer, J.A.G.M. de Visser, JK, PLoS Comp. Biol. **7** (2011) e1002134

Random fitness landscapes

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

- The mapping $\sigma \rightarrow w(\sigma)$ 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

- REM/HoC landscape is uncorrelated but maximally rugged (epistatic)

Evolutionary accessibility in the house-of-cards model

- What is the mean number of shortest, selectively accessible paths $n_{\rm acc}$ from an arbitrary genotype at distance d to the global maximum?

- The total number of paths is $d!$, and a given path consists of d independent, identically distributed fitness values $w_0, w_1, \ldots, w_{d-1}$ with $w_d > \max\{w_0, w_1, \ldots, w_{d-1}\}$.

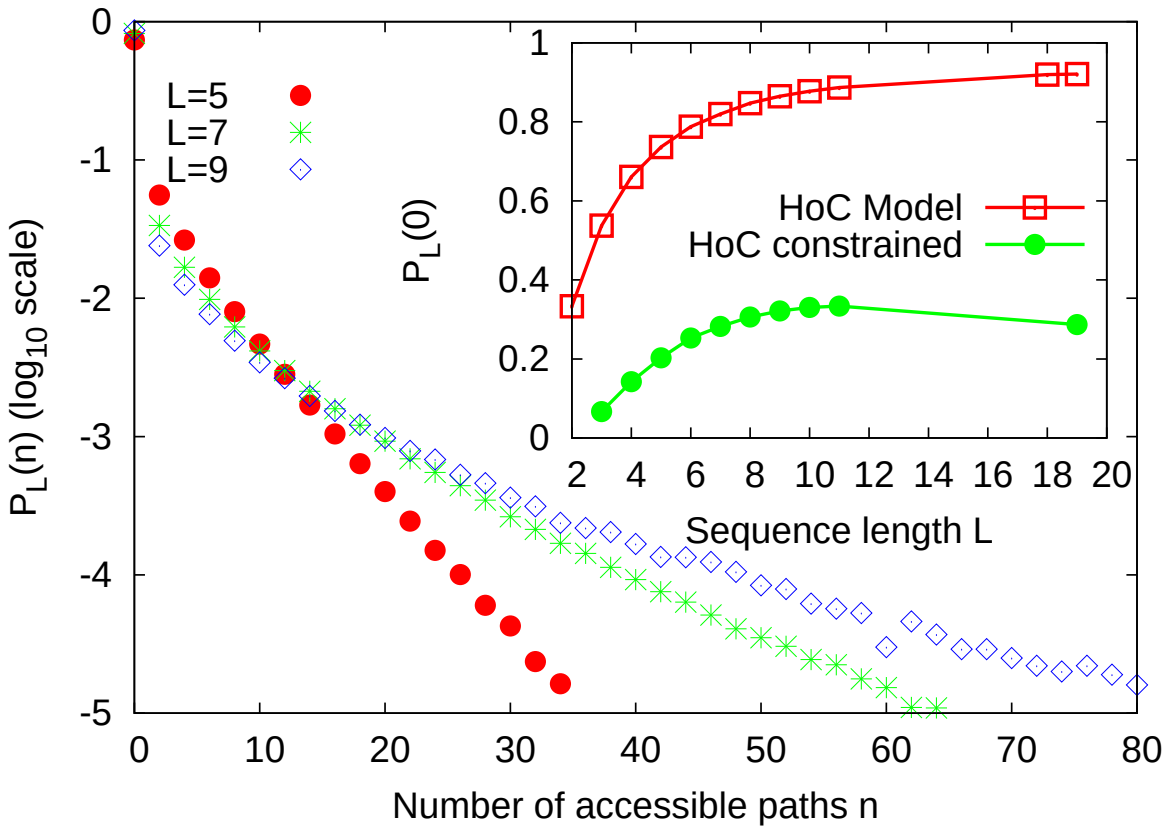
- A path is accessible iff $w_0 < w_1 \ldots < w_{d-1}$

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

$$\Rightarrow \langle n_{\rm acc} \rangle = \frac{1}{d!} \times d! = 1$$

- This holds in particular for the $L!$ paths from the "antipodal" reversal genotype of the global maximum.

Distribution of number of accessible paths from reversal genotype



- "Condensation of probability" at $n_{\mathrm{acc}} = 0$, $P_L(0)$ tends to unity for $L \to \infty$
- Constraining the antipodal sequence to be the global fitness minimum leads to nonmonotonic behavior of $P_L(0)$

Landscapes with tunable ruggedness

Kauffman's LK-model

Kauffman & Weinberger 1989

- Each site interacts randomly with $K \leq L-1$ other sites:

$$f(\sigma) = \sum_{i=1}^{L} f_i(\sigma_i | \sigma_{i_1}, ..., \sigma_{i_K}) \qquad f_i \text{ : i.i.d. RV's}$$

- Diluted spin glass with random fields and interactions up to order $p = K+1$

Rough Mt. Fuji landscapes

Aita et al. 2000

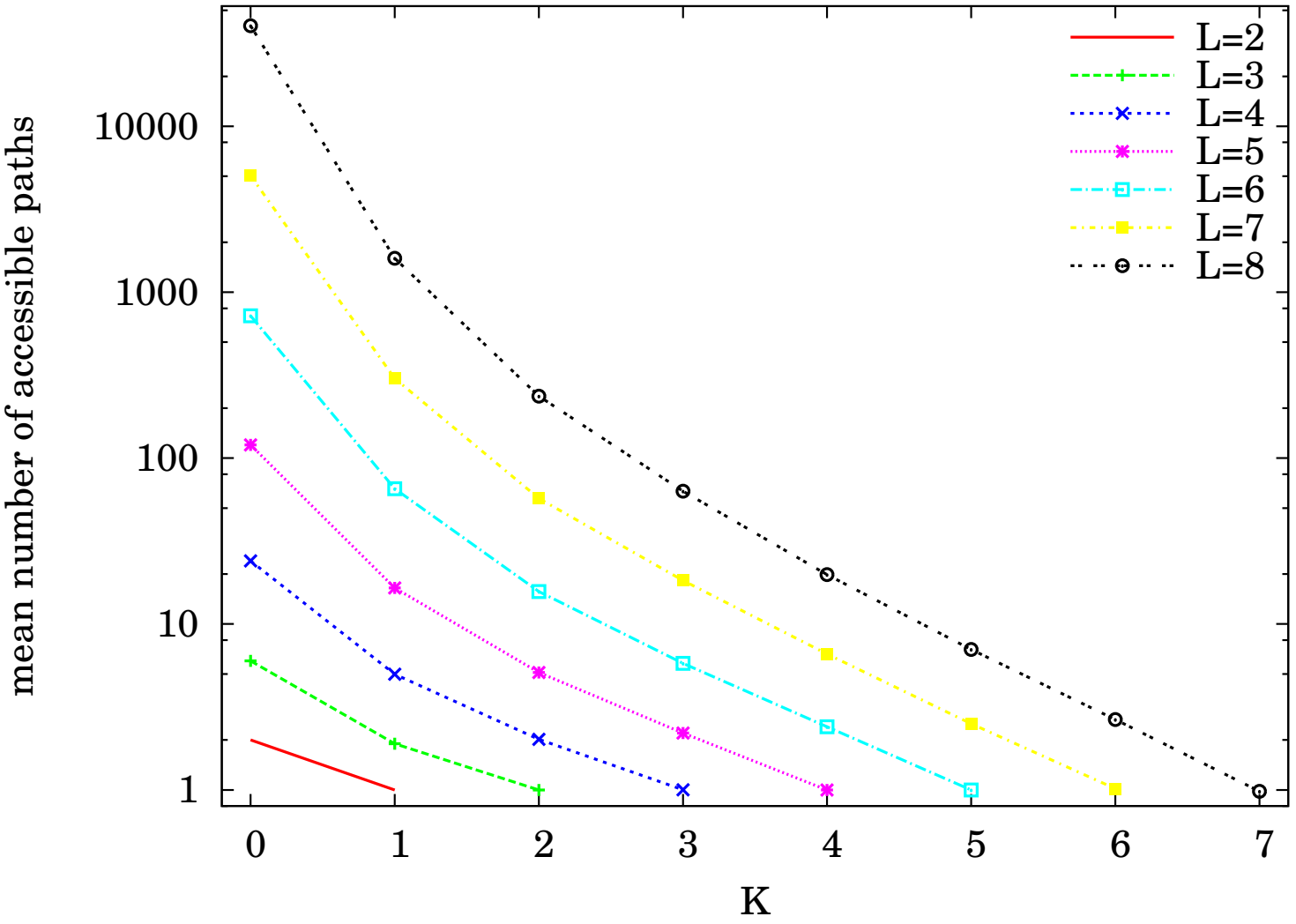
- Average fitness decreases linearly with distance from reference genotype:

$$f(\sigma) = -\theta d(\sigma, \sigma^{(\mathrm{ref})}) + \eta(\sigma)$$

η: (Gaussian) RV's with unit variance $\quad d(\sigma, \sigma')$: Hamming distance

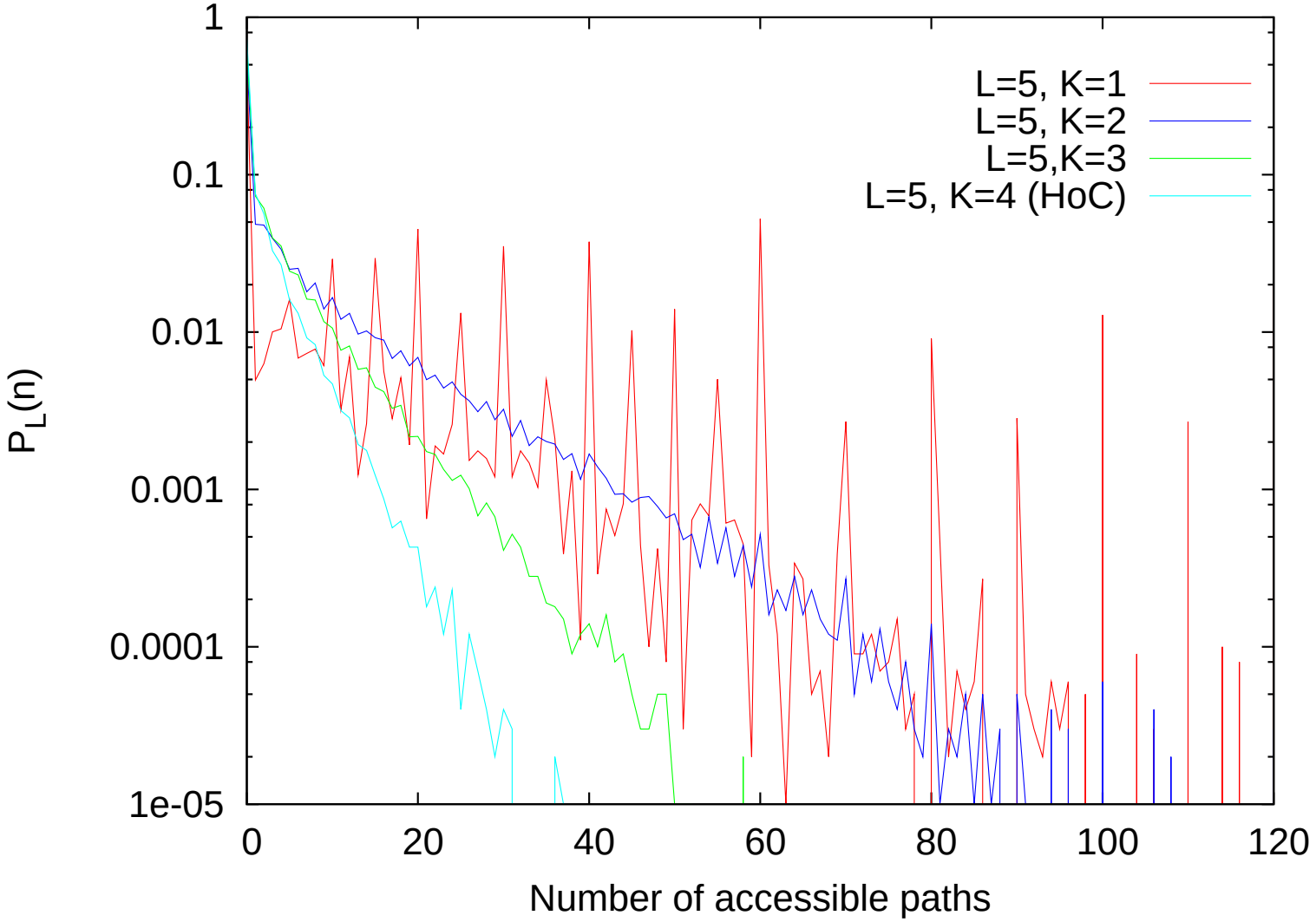
- Equivalent to REM in an external magnetic field

Kauffman model: Mean number of accessible paths



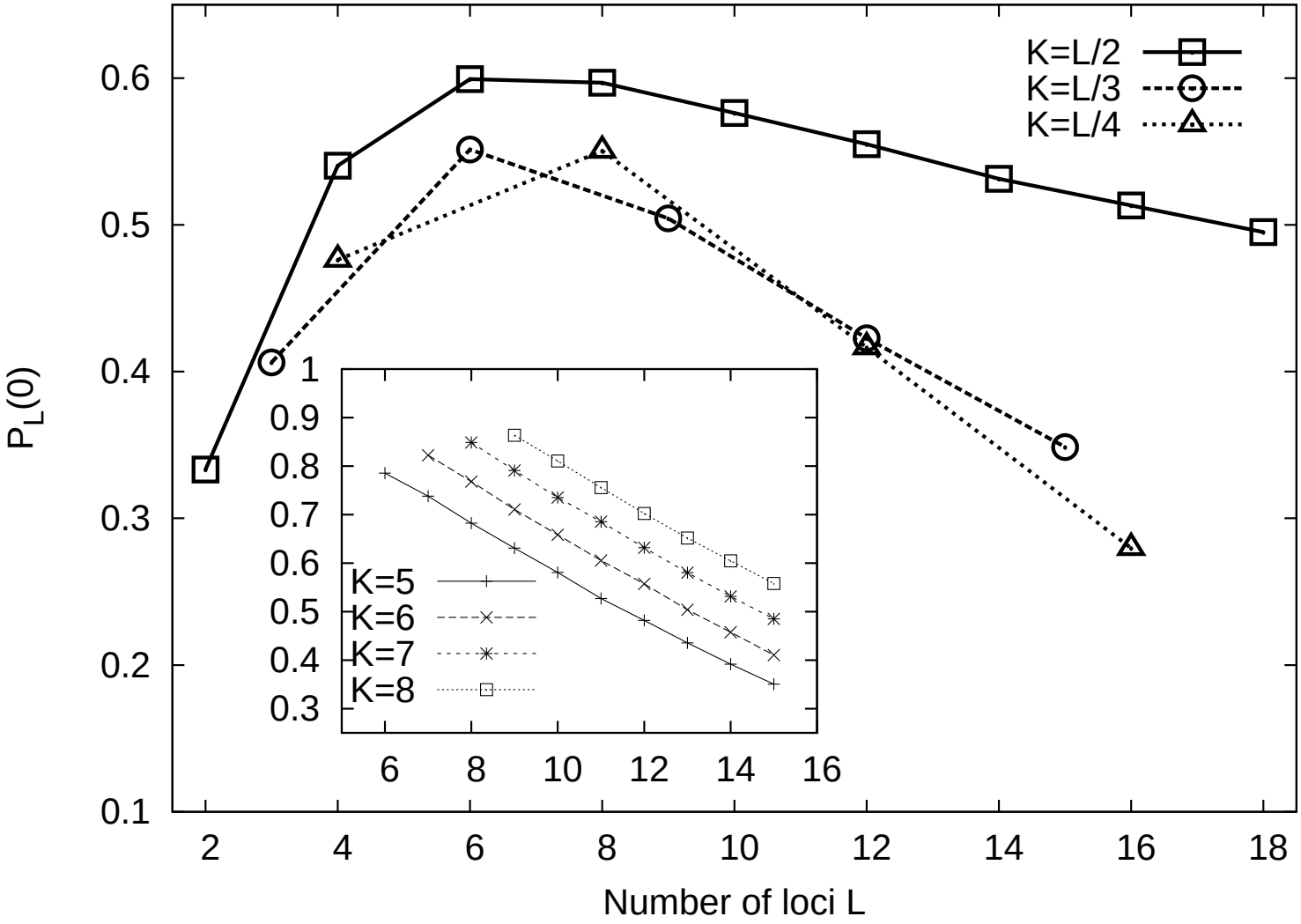
Simple special cases: • $K = 0 : \langle n_{\mathrm{acc}} \rangle = L!$ • $K = L - 1 : \langle n_{\mathrm{acc}} \rangle = 1$

Kauffman model: Distribution of the number of accessible paths



- Roughly exponential decay with combinatorial structure for $K = 1$

Kauffman model: Probability of no accessible path



- Accessibility increases with L both for fixed $K \geq 3$ and fixed K/L

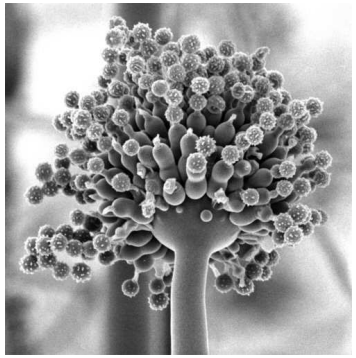
Application to empirical data

J. Franke, A. Klözer, J.A.G.M. de Visser, JK, PLoS Comp. Biol. **7** (2011) e1002134

The *A. niger* data set

J.A.G.M. de Visser et al., Evolution **51**, 1499 (1997)

- $L = 8$ individually deleterious marker mutations residing on different chromosomes of *Aspergillus niger* (black mold)

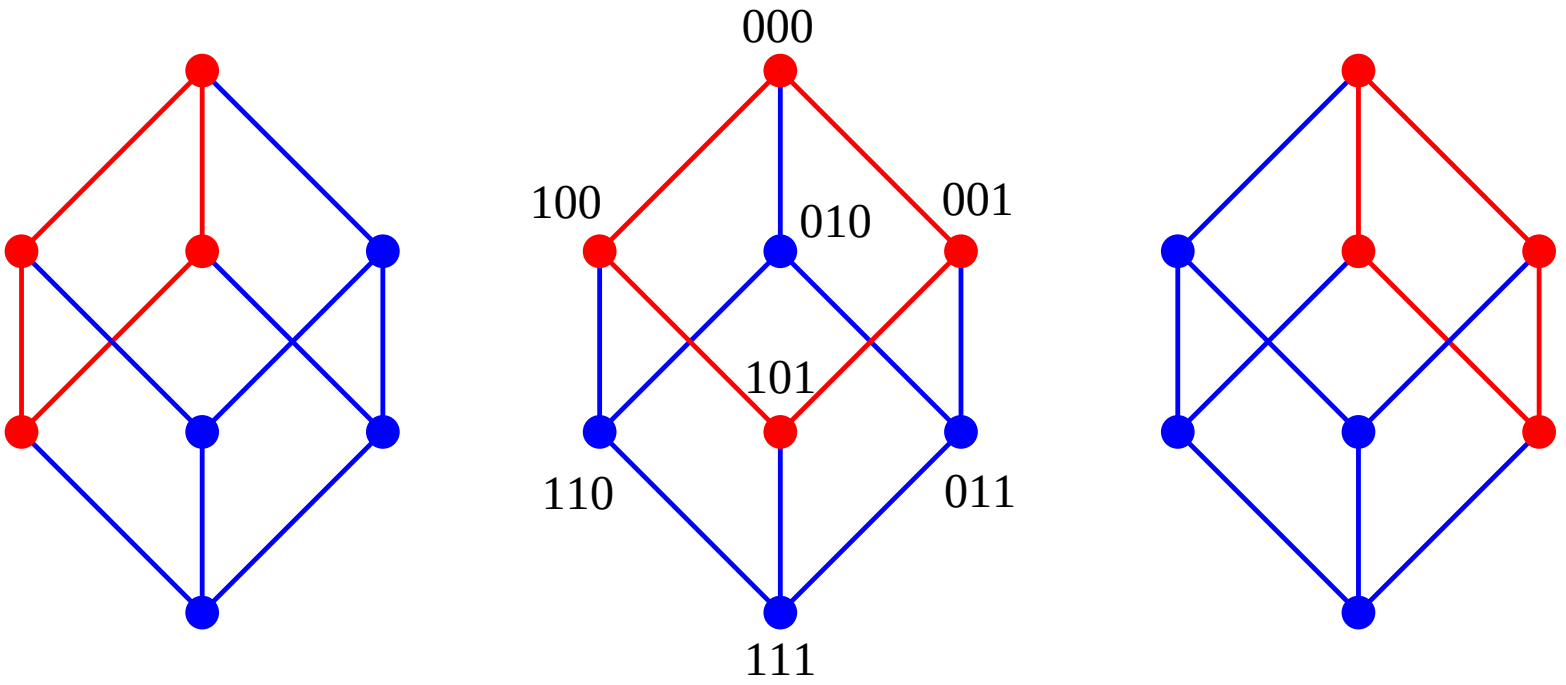


- 186 out of $2^8 = 256$ possible combinations were isolated in $\sim$ 2500 trials
- Fitness (= growth rate) was measured for two replicates per strain
- Fitness relative to wild type falls in the range $w_{\min} = 0.274 \leq w \leq 1$
- Likelihood of missing more than one strain with fitness $> w_{\min}$ is < 5 %
  $\Rightarrow$ assign zero fitness to missing strains ("lethals")

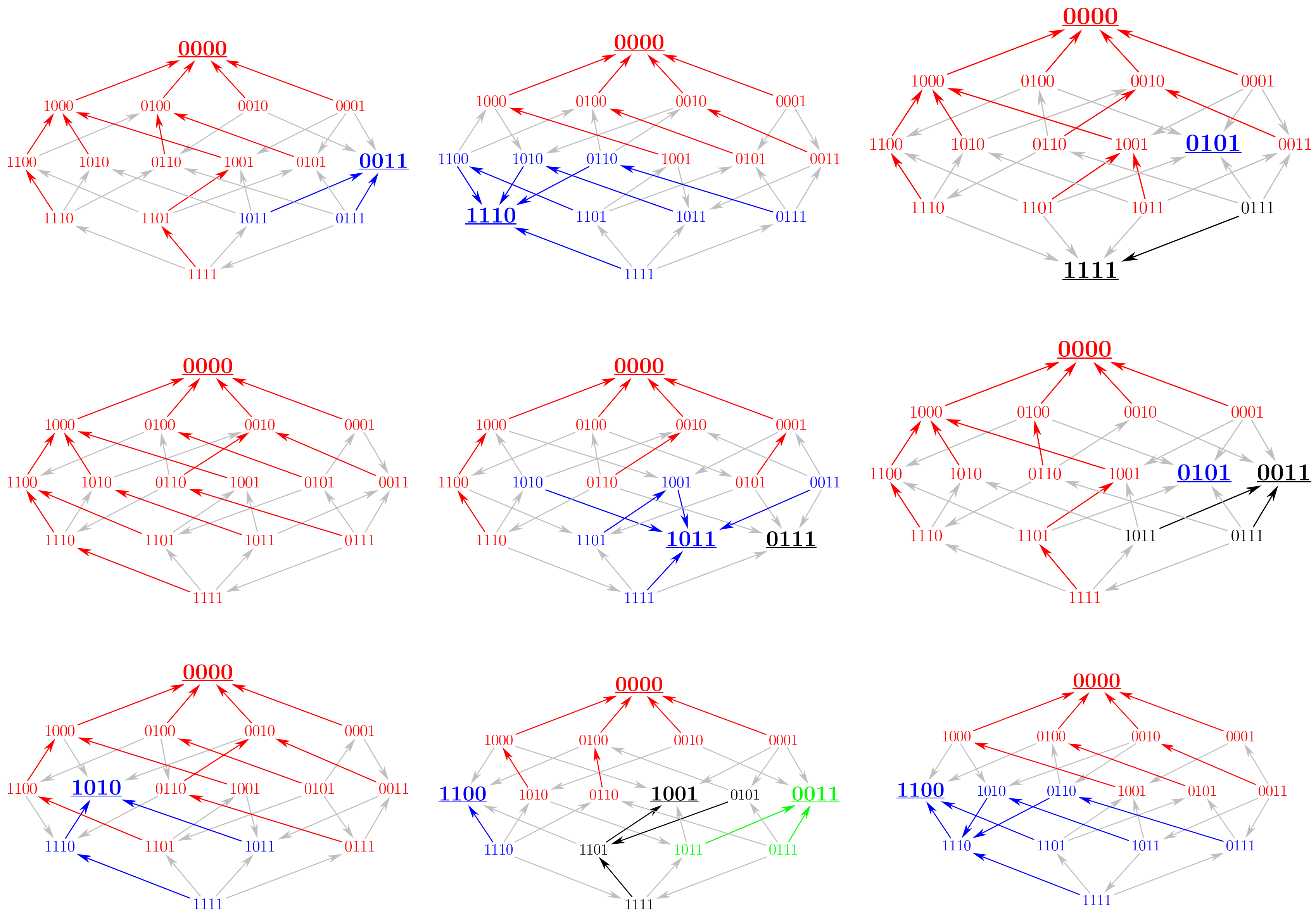
Subgraph analysis

- Probe effect of scale by analyzing ensembles of $\binom{L}{m}$ subgraphs containing subsets of m mutations ($2 \leq m \leq L$)

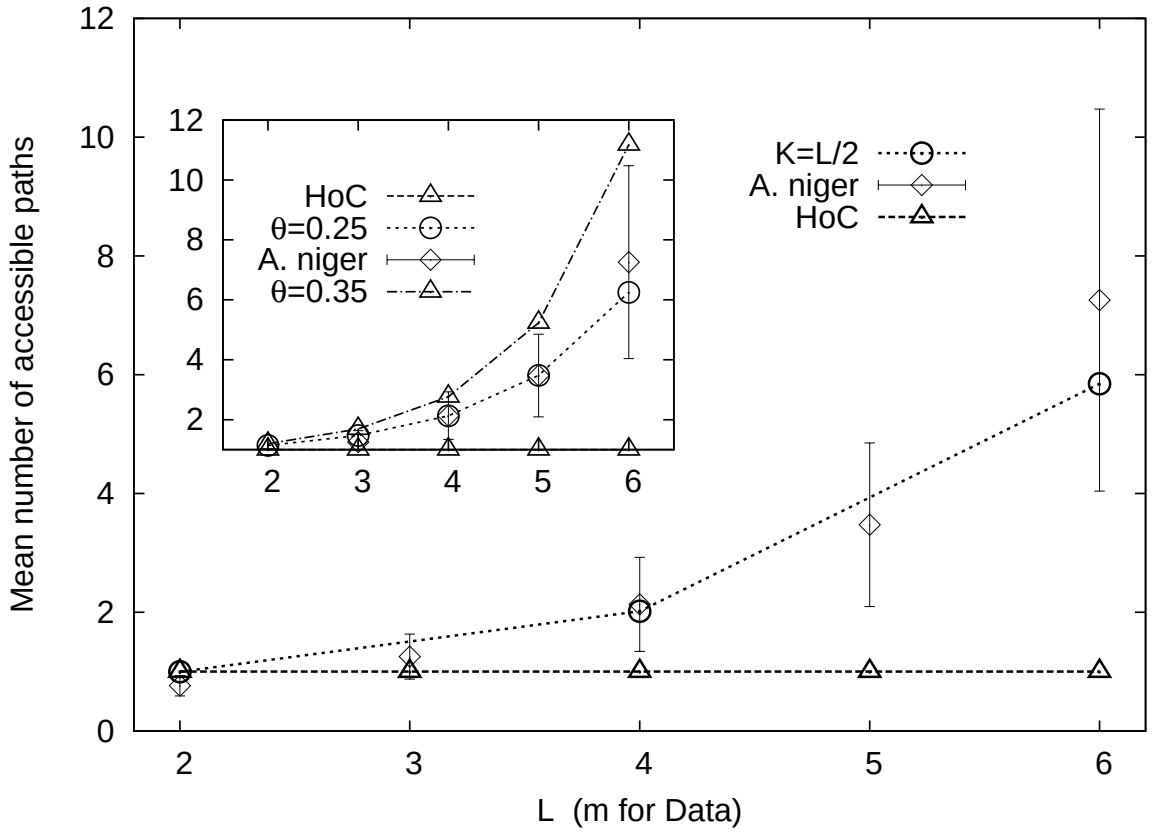
- Example: $L = 3, m = 2$



A selection of m=4 subgraphs of the *A. niger* landscape

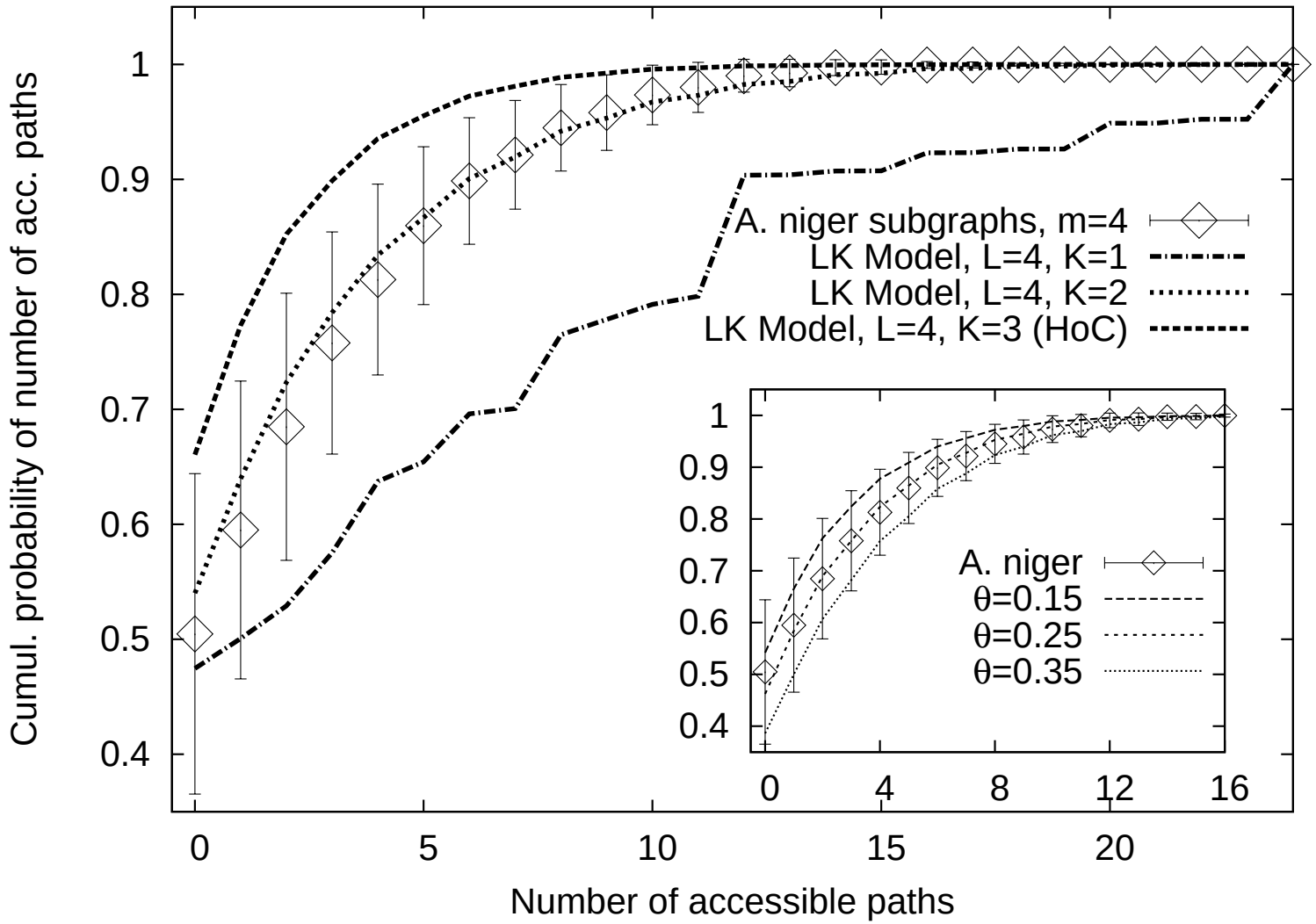


Mean number of accessible paths from subgraph analysis



- Error bars from resampling analysis
- Data are reasonably well described by Kauffman model with $K = L/2$ or rough Mt. Fuji model with $\theta \approx 0.25$

Cumulative distribution of the number of paths ($m = 4$)



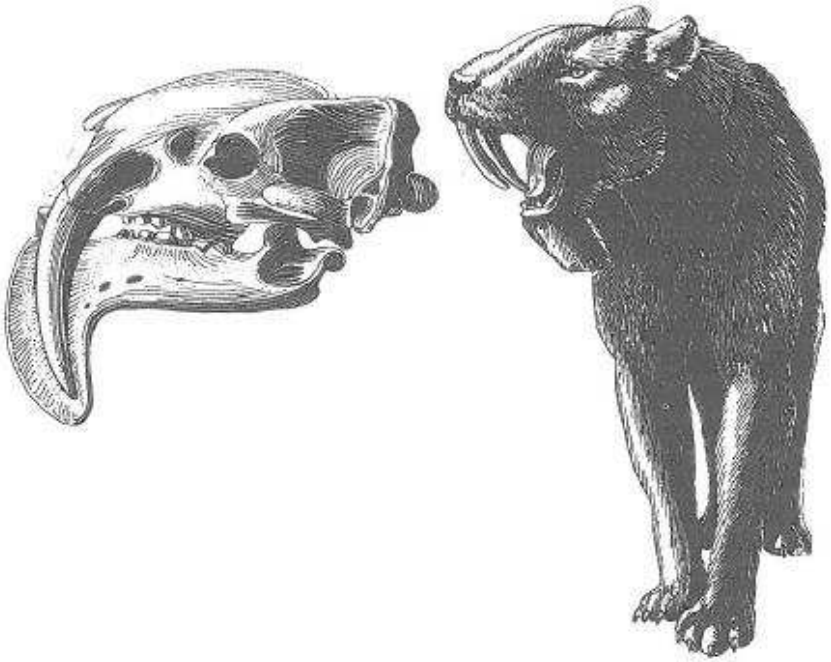
Predictability of evolution on an empirical fitness lanscape

I.G. Szendro, J. Franke, J.A.G.M. de Visser, JK (in preparation)

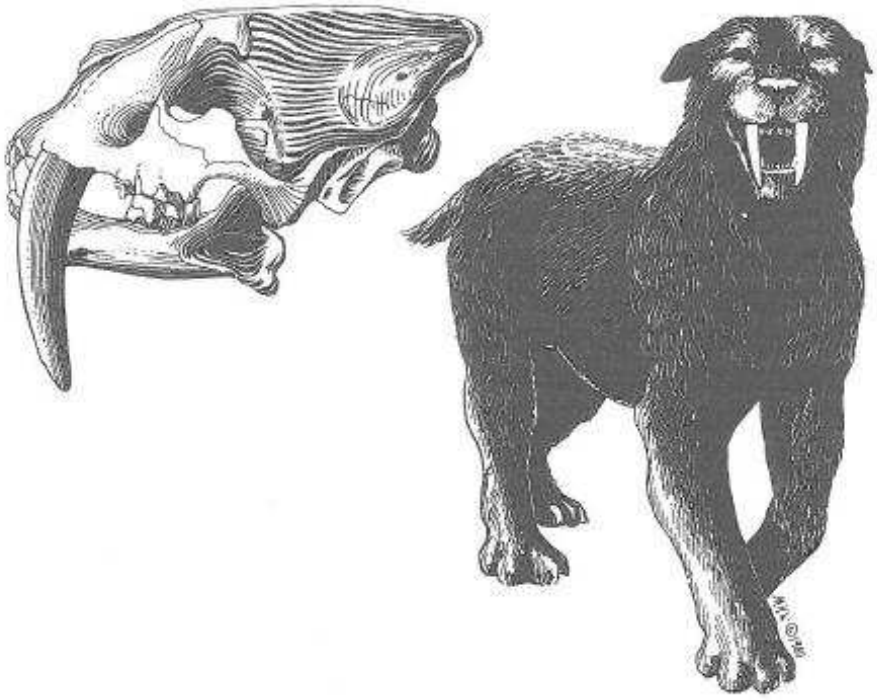
Convergent evolution

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

Simon Conway Morris



marsupial



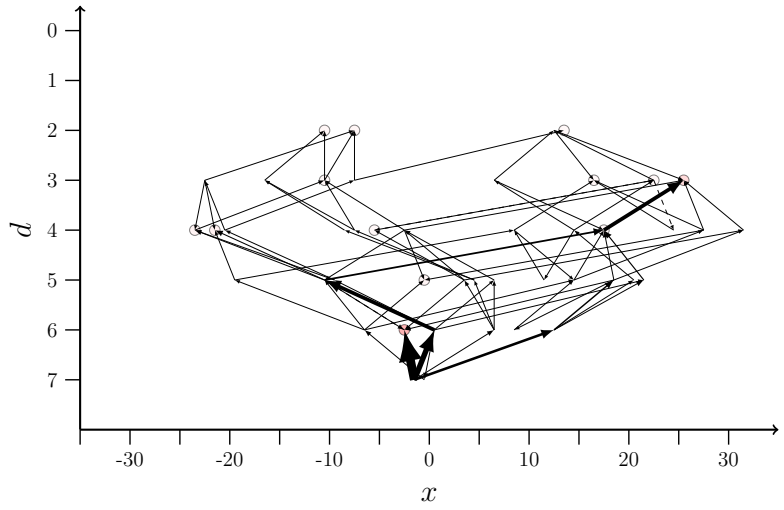
placental

Evolutionary dynamics on an empirical fitness landscape

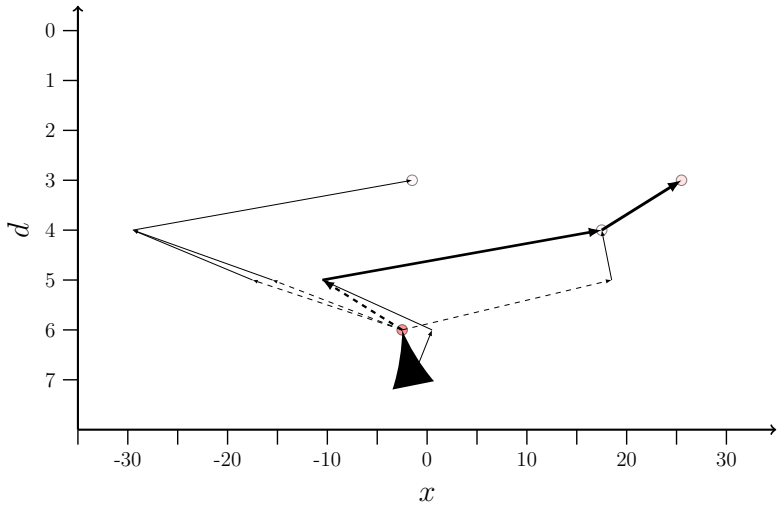
- Fixed number N of individuals reproduce asexually in discrete generations (Wright-Fisher model)
- Mutations occur with probability μ per site and generation
- Evolution starts from a viable genotype at distance d_0 from the wildtype
- Two types of evolutionary trajectories:
  - Lines of descent: Track 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
- Expect: Evolution becomes more predictable with increasing N

K. Jain, JK, Genetics 2007

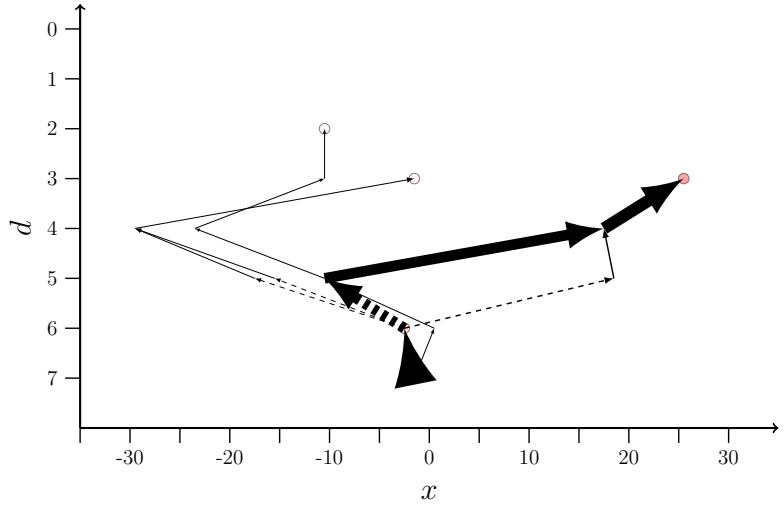
Lines of descent at different population sizes



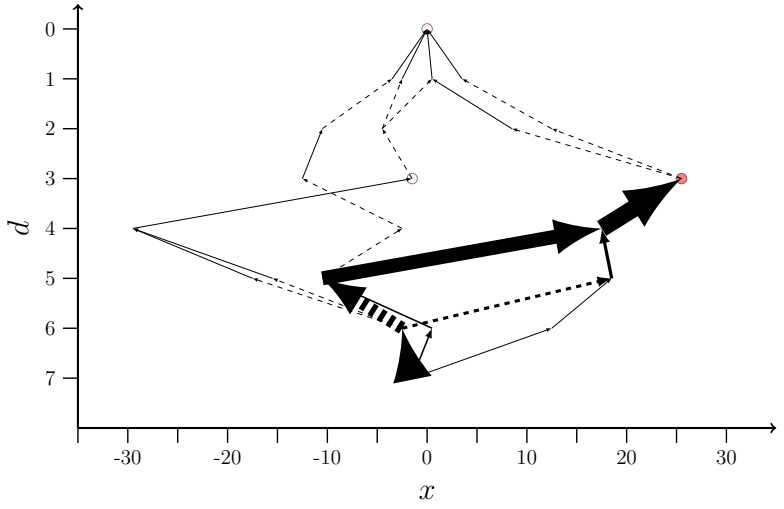
(a) $N = 2^7$



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



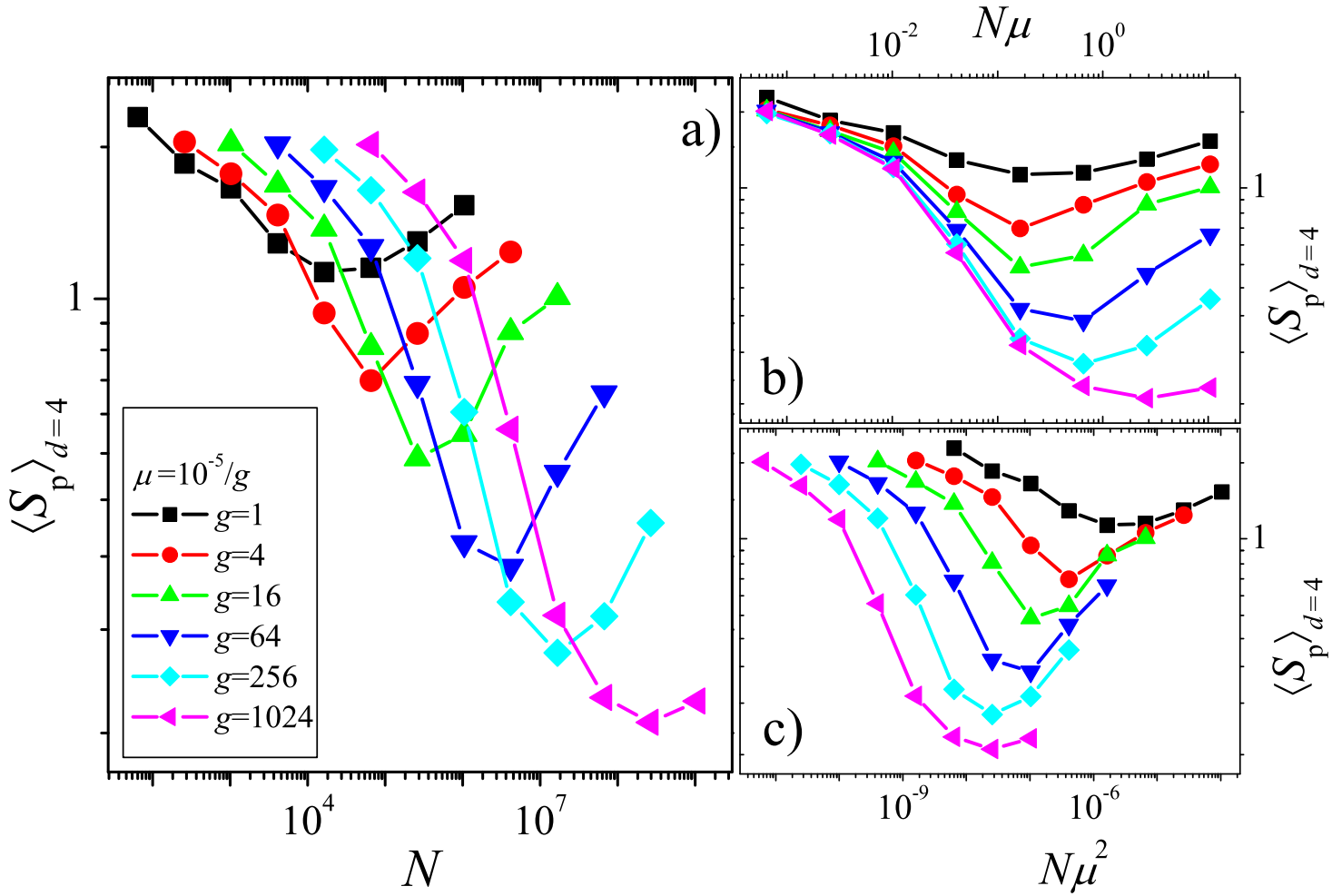
(c) $N = 2^{17}$



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

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

Pathway entropy



$\Rightarrow$ non-monotonic because new pathways become available when $N\mu^2 \sim 1$

Summary

- The concept of evolutionary accessibility motivates new questions about fitness landscape models

- Simulations suggest simple dichotomy in the behavior of accessibility with increasing genotype dimensionality L:
  - For uncorrelated landscapes (REM) $P_L(0) \to 1$ and $\langle n_{\mathrm{acc}} \rangle = 1$
  - For correlated landscapes (LK, Mt. Fuji) $P_L(0) \to 0$ and $\langle n_{\mathrm{acc}} \rangle \to \infty$

- Empirical fitness landscapes are of intermediate ruggedness, with distinct patterns depending on the type of mutations under consideration

  Szendro et al. 2012

- Predictability of evolution on realistic fitness landscapes depends non-monotonically on population size