SFB 680
Molecular Basis of
Evolutionary Innovations

Sign epistasis and evolutionary accessibility

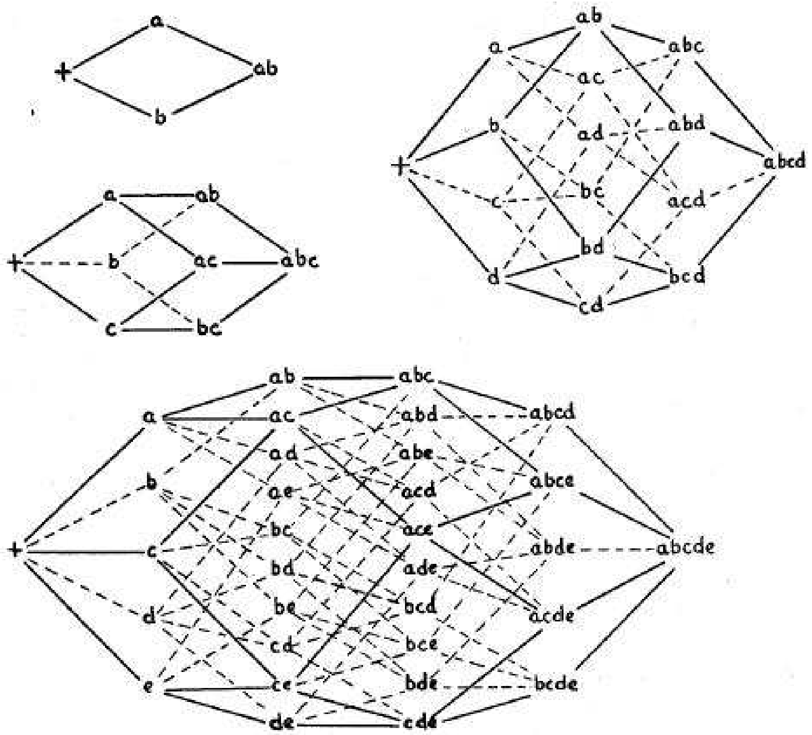
Joachim Krug
Institute of Theoretical Physics, University of Cologne, Germany

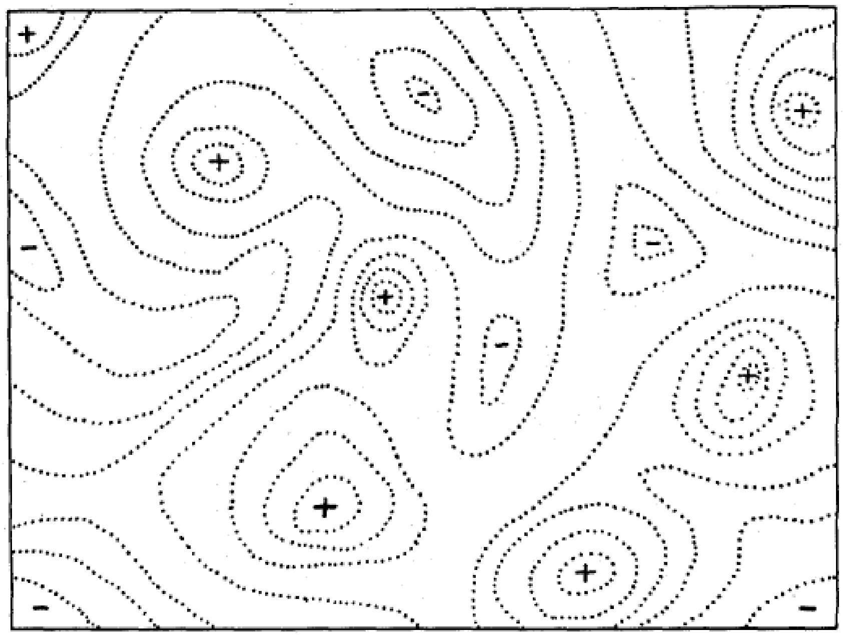
with Jasper Franke, Alexander Klözer and Arjan de Visser (Wageningen)

KITP Santa Barbara, 3/9/2011

Fitness landscapes

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





”...selection will easily carry the species to the nearest peak, but there will be innumerable other peaks that will be higher but which are separated by ‘valleys’. 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...”

Epistasis and sign epistasis

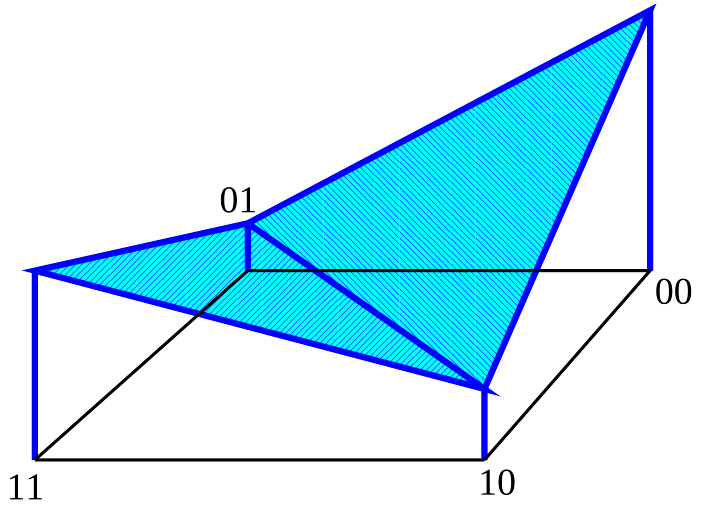
- **General setting**: L haploid loci $i = 1, \ldots, L$ at which a mutation can be present ($\sigma_i = 1$) or absent ($\sigma_i = 0$)

- A fitness landscape is a function $w(\sigma)$ on the set of 2^L genotypes

- Epistasis implies interactions between the effects of different mutations

- Sign epistasis: Mutation at a given locus is beneficial or deleterious depending on the state of other loci Weinreich, Watson & Chao (2005)

- Reciprocal sign epistasis for $L = 2$:



Two manifestations of sign epistasis

Local fitness optima

Haldane 1931, Wright 1932

- Reciprocal sign epistasis is a necessary but not sufficient condition for the existence of multiple fitness peaks
  Poelwijk et al. 2011

- Local optima are probably common, but their existence cannot be empirically proven and their evolutionary importance remains controversial
  Whitlock et al. 1995; Gavrilets 2004

Accessibility of mutational pathways

Weinreich et al. 2005

- A path of single mutations connecting two genotypes $\sigma \longrightarrow \sigma'$ with $w(\sigma) < w(\sigma')$ is selectively accessible if fitness increases monotonically along the path

- In the absence of sign epistasis all paths to the global optimum are accessible, and vice versa

A caveat

- Accessibility of pathways as defined here makes no statement about the probability that a pathway will actually be realized under a given evolutionary scenario

- In the SSWM regime of strong selection ($Ns \gg 1$) and weak mutation ($N\mu \ll 1$) adaptation can proceed only along accessible paths, and the weight of a path is given by the product of fixation probabilities

  Orr 2002; Weinreich et al. 2006

- Here we focus on existence of accessible paths, a property that depends only on the ordering of fitness values

The punchline

- Statistics of accessible mutational pathways as a measure of fitness landscape ruggedness

- Quantities of interest:

  (i) probability to find at least one accessible path $\Rightarrow$ accessibility
  (ii) mean number of accessible paths $\Rightarrow$ predictability

  to the global fitness maximum of the landscape

- Address genome-wide accessibility (number of loci $L \longrightarrow \infty$)

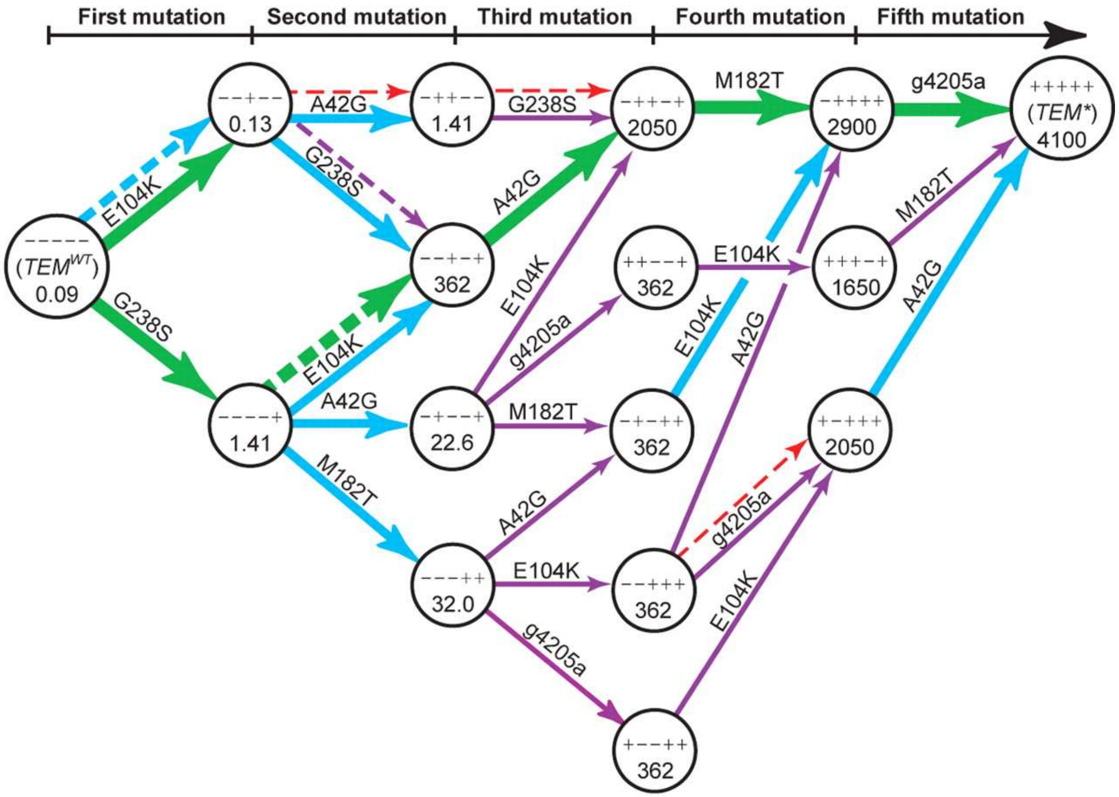
- Across a wide range of models, accessibility is high, in the sense that the probability of finding at least one path tends to unity, and predictability is low, in the sense that many alternative pathways exist

- Subgraph analysis of an empirical multilocus fitness landscape confirms these features and allows to estimate epistasis parameters

Empirical fitness landscapes

Example 1: The TEM1 β-lactamase resistance landscape

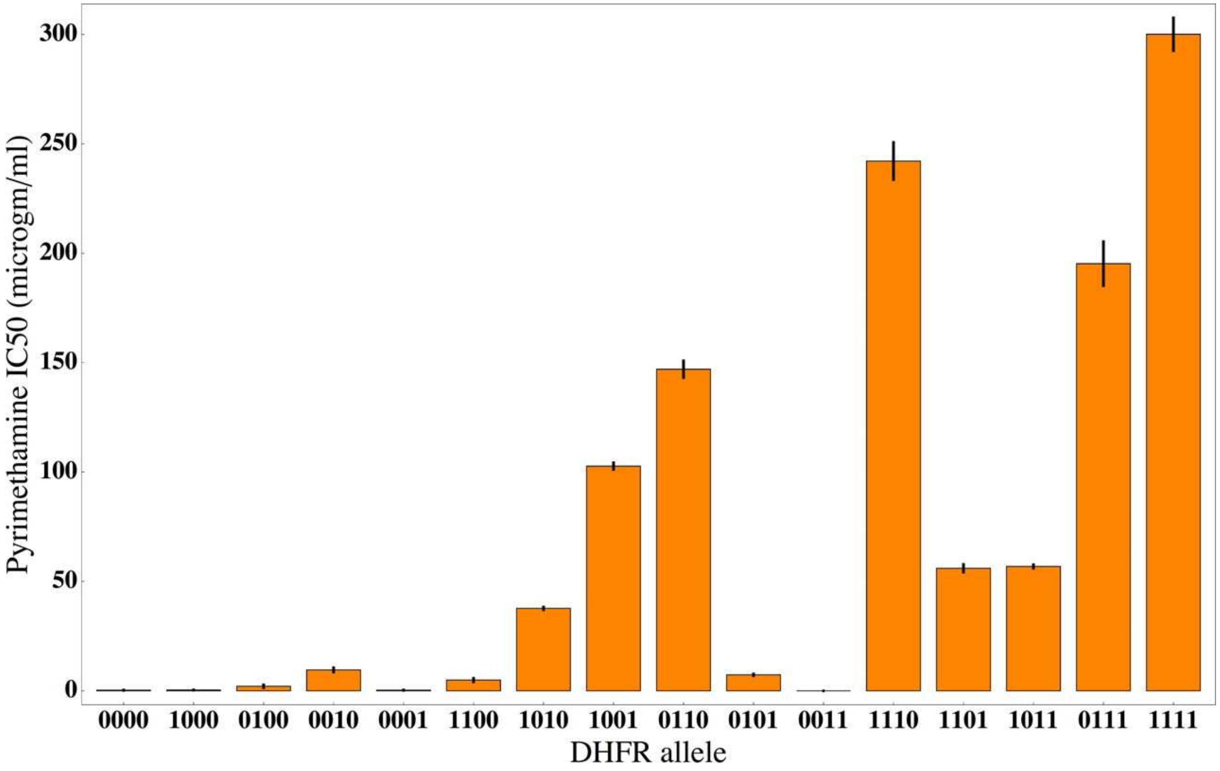
D.M. Weinreich, N.F. Delaney, M.A. De Pristo, D.L. Hartl, Science **312**, 111 (2006)



- 5 mutations in the β-lactamase enzyme confer resistance to cefotaxime
- 18 out of L!=120 paths from the wildtype to the fivefold mutant are accessible (10 most important paths shown); single ’fitness’ peak

Example 2: Pyrimethamine resistance in the malaria parasite

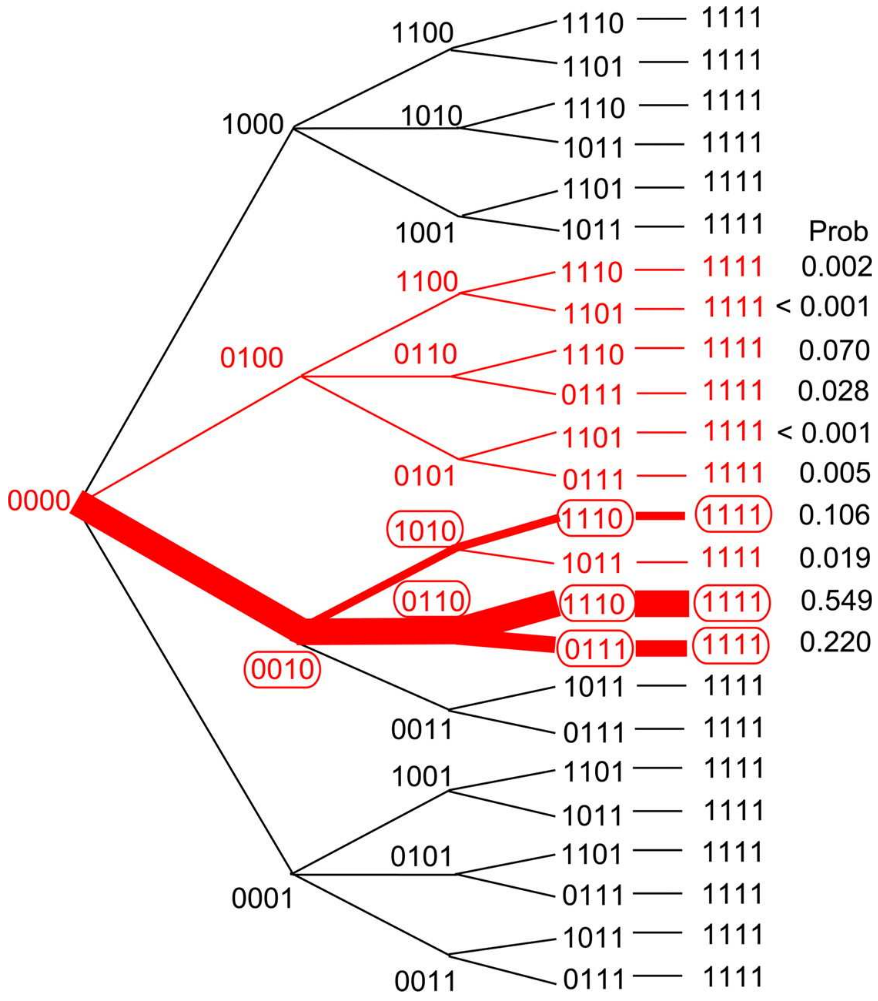
E.R. Lozovsky et al., Proc. Natl. Acad. Sci. USA **106**, 12025 (2009)



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

Example 2: Pyrimethamine resistance in the malaria parasite

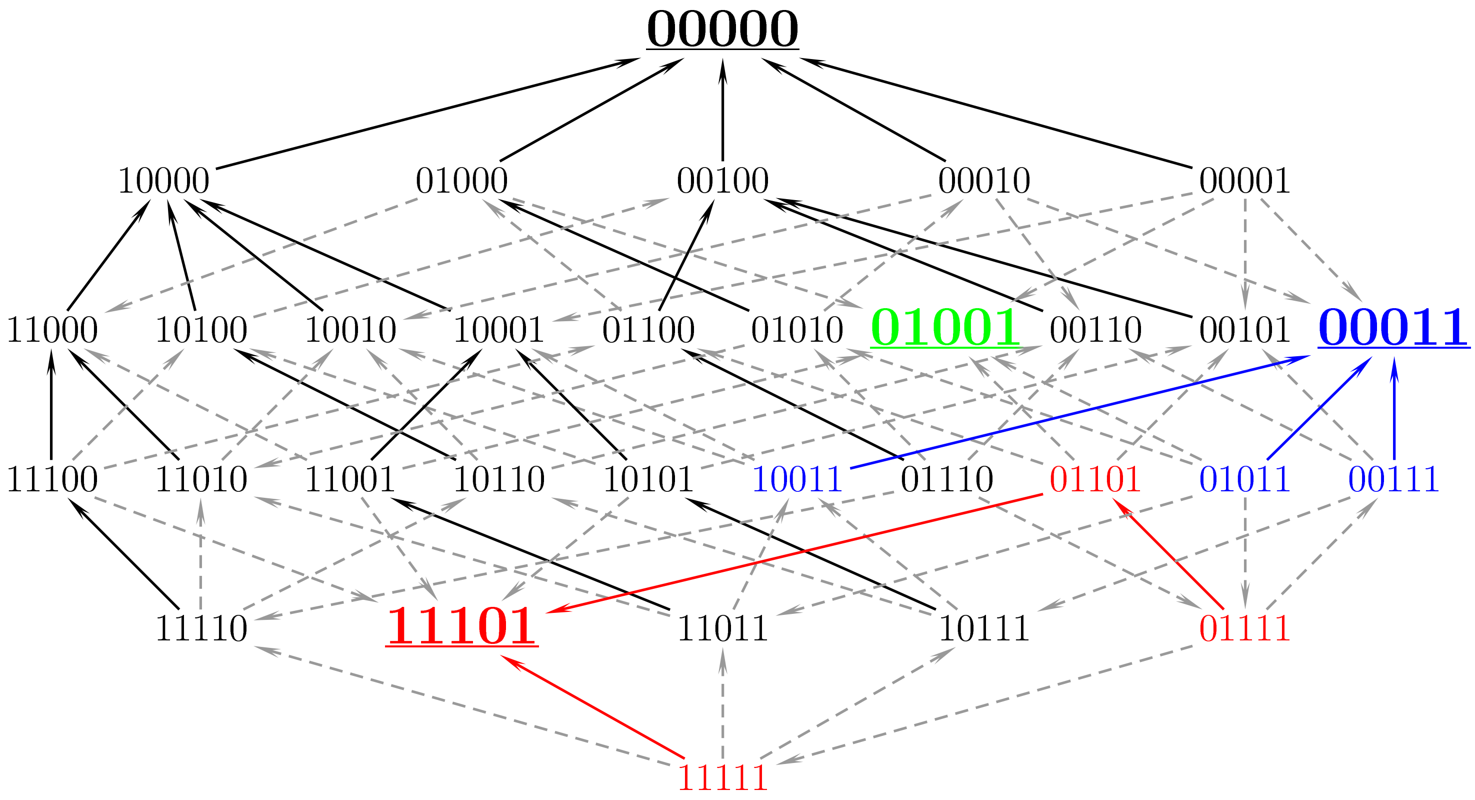
E.R. Lozovsky et al., Proc. Natl. Acad. Sci. USA **106**, 12025 (2009)



- Dominant pathways consistent with occurrence 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)



- All combinations of 5 individually deleterious marker mutations
- 3 local fitness optima, 25 out of 120 paths are accessible

The *Aspergillus niger* data set

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

- 8 marker mutations residing on different chromosomes
  (1 spore color mutation, 5 auxotrophies, 2 resistances)

- 186 out of $2^8 = 256$ possible combinations were isolated among ~ 2500 segregants

- Fitness (= growth rate) was measured for two replicates per strain

- Fitness relative to wild type falls in the range $w_{\mathrm{min}} = 0.274 \leq w \leq 1$

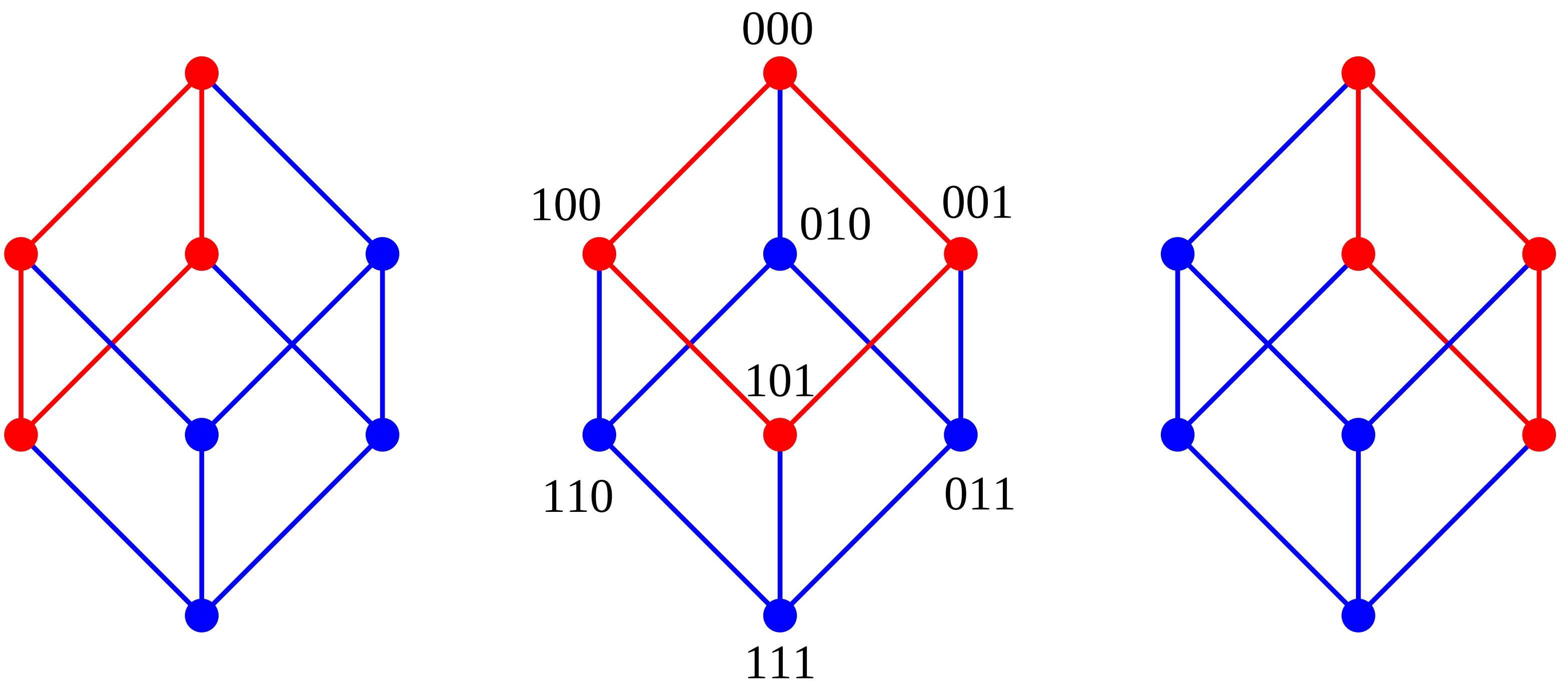
- Likelihood of missing more than one strain with fitness $> w_{\mathrm{min}}$ is < 5 %
  $\Rightarrow$ assign zero fitness to missing strains ("lethals")

- Lethals are mostly associated with lysine deficiency (62 out of 70)

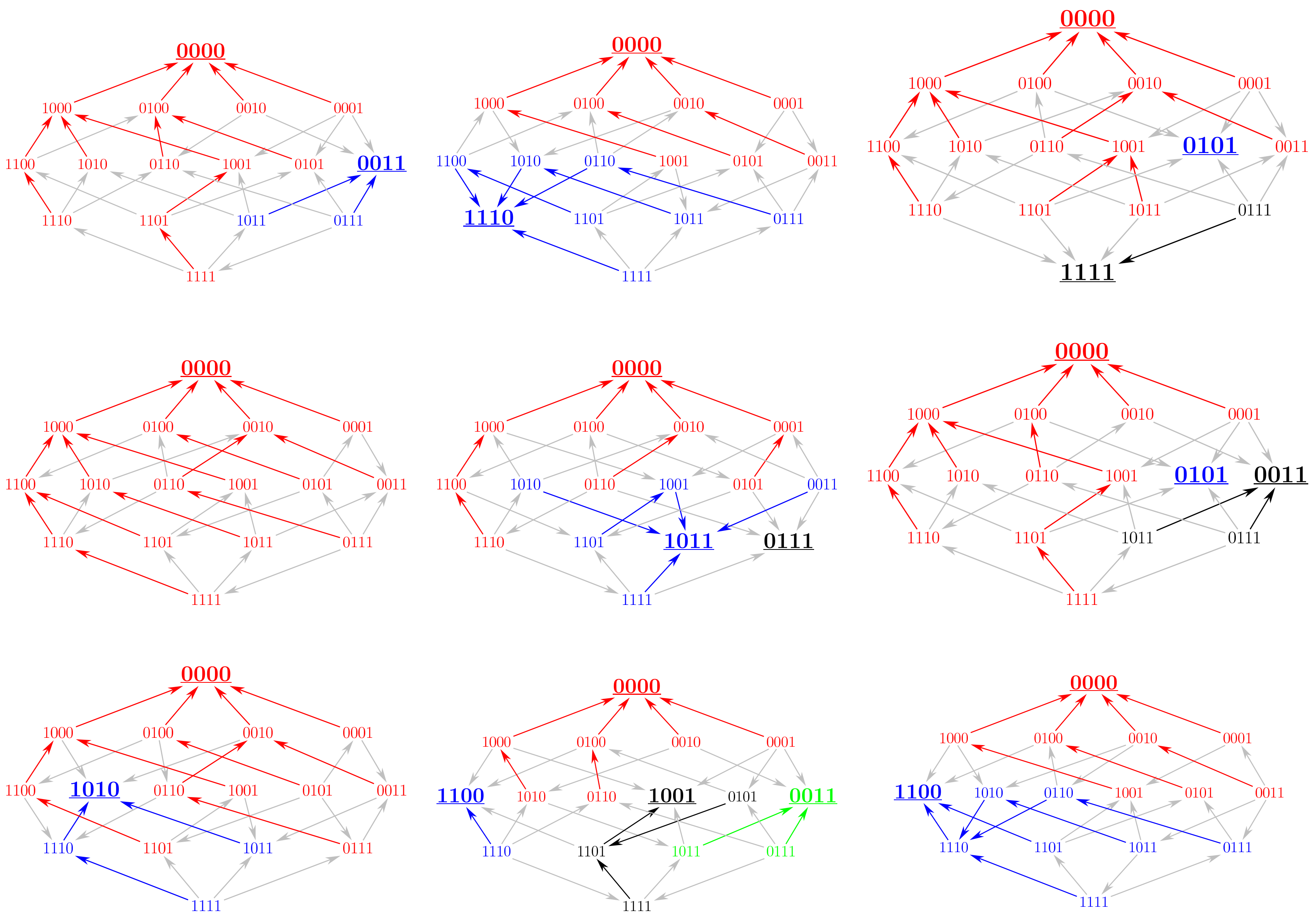
Subgraph analysis

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

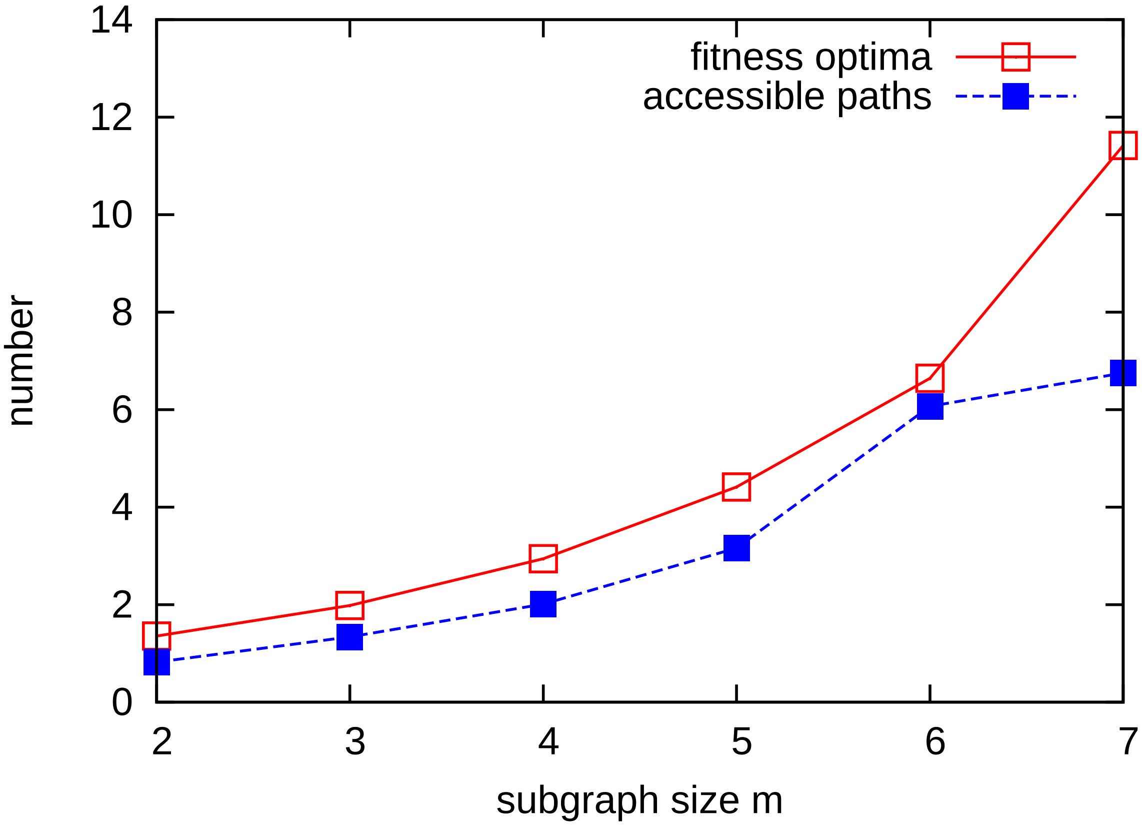
- Example: $L = 3, m = 2$



A selection of m=4 subgraphs



Subgraph properties as a function of subgraph size



- What do these numbers mean?
- Can they be reproduced by fitness landscape models?

Random models of fitness landscapes

Null model: House-of-cards

- In the house-of-cards model fitness is assigned randomly to genotypes

  Kingman 1978, Kauffman & Levin 1987

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

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

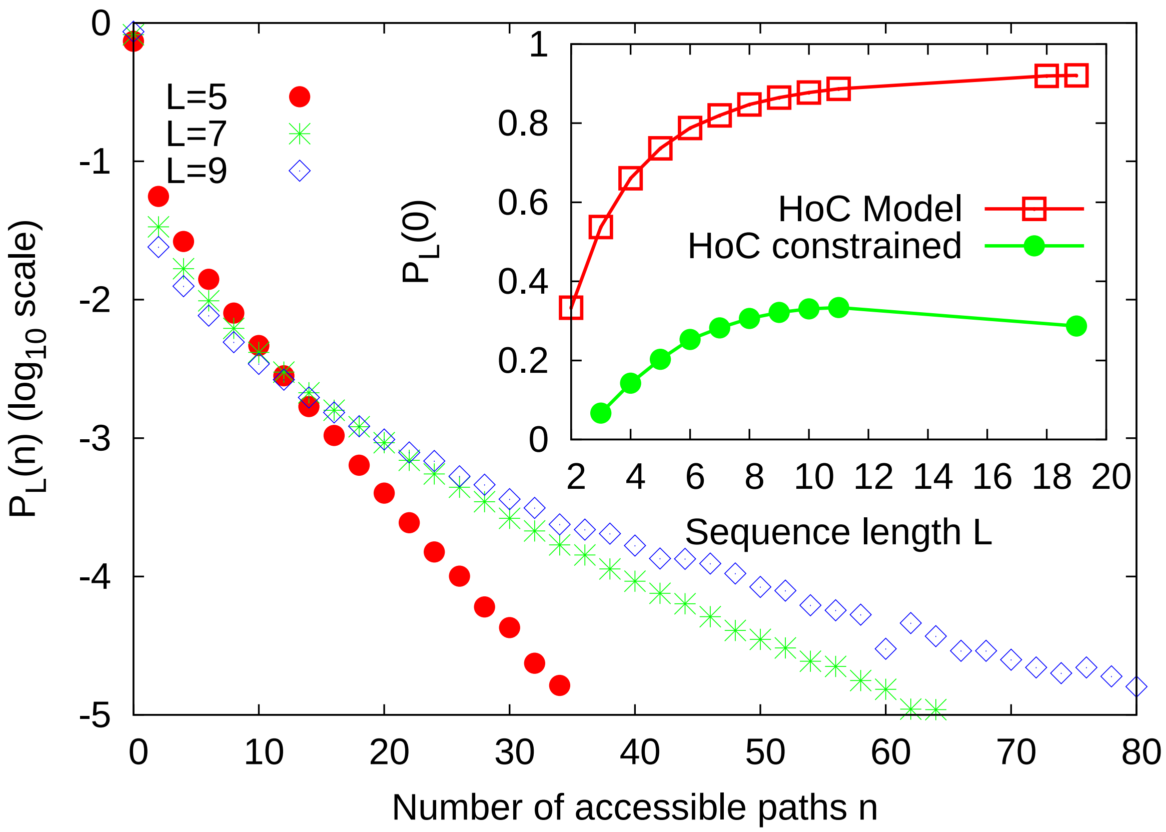
- A path is accessible iff $w_0 < w_1 \ldots. < w_{d-1}$

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

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

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

Distribution of number of accessible paths from reversal genotype



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

- Characterize distribution by $\langle n_{\rm acc} \rangle$ and the probability $P_L(0)$ that no path is accessible; for HoC model $P_L(0) \to 1$ for large L

Landscapes with tunable ruggedness

Kauffman's LK-model

Kauffman & Weinberger 1989

- Each locus interacts randomly with $K \leq L-1$ other loci:

$$\ln w(\sigma) = \sum_{i=1}^{L} f_i(\sigma_i | \sigma_{i_1}, \ldots, \sigma_{i_K})$$

f_i: Uncorrelated RV's assigned to each of the 2^{K+1} possible arguments

- $K = 0$: Non-epistatic $K = L-1$: House-of-cards

Rough Mt. Fuji landscapes

Aita et al. 2000

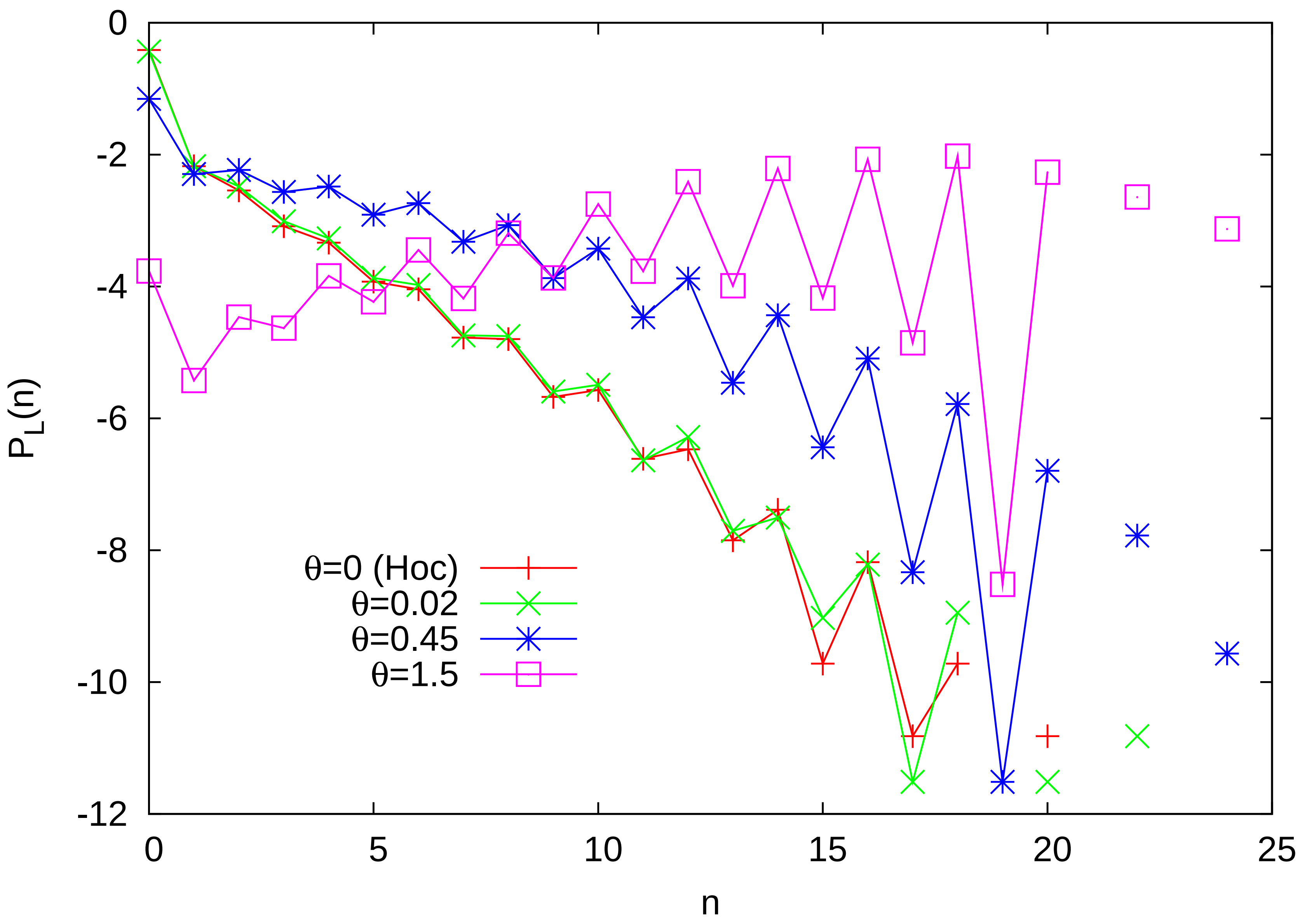
- Non-epistatic ("Mt. Fuji") landscape perturbed by a random component:

$$\ln w(\sigma) = -\theta d(\sigma, \sigma^{(0)}) + \eta(\sigma)$$

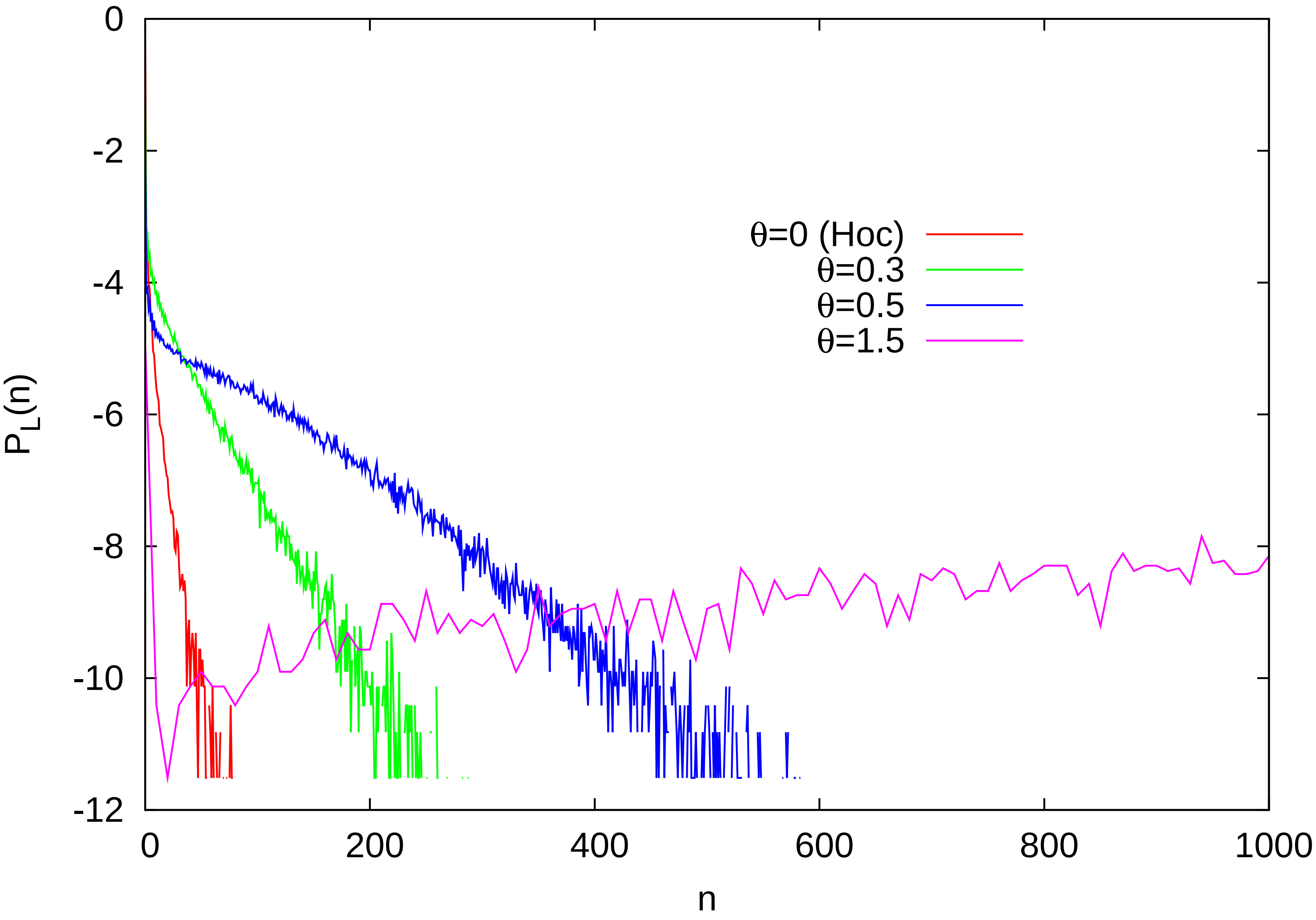
η: (Gaussian) RV's with unit variance $d(\sigma, \sigma')$: Hamming distance

- $\theta = 0$: House-of-cards $\theta \to \infty$: Non-epistatic

Distribution of accessible paths in the Mt.Fuji model ($L = 4$)



Distribution of accessible paths in the Mt.Fuji model ($L = 7$)



Mean number of paths in the rough Mt. Fuji model

J. Franke, G. Wergen, JK, JSTAT P10013 (2010)

- Probability p_{acc} for a path to be accessible is equal to the probability for the L RV's $w_k = \eta_k + ck$ to be ordered in the sense of $w_0 < w_1 < \ldots < w_{L-1}$.

- When the η_k are drawn from the Gumbel distribution $\mathrm{Prob}[\eta < x] = \exp[-e^{-x}]$ this probability is given by

$$p_{\mathrm{acc}} = \frac{(1-e^{-\theta})^L}{\prod_{k=1}^{L}(1-e^{-\theta k})} \approx \sqrt{\frac{\theta}{2\pi}}\, e^{\pi^2/6\theta}\, (1-e^{-\theta})^L \quad \text{for } L \to \infty$$

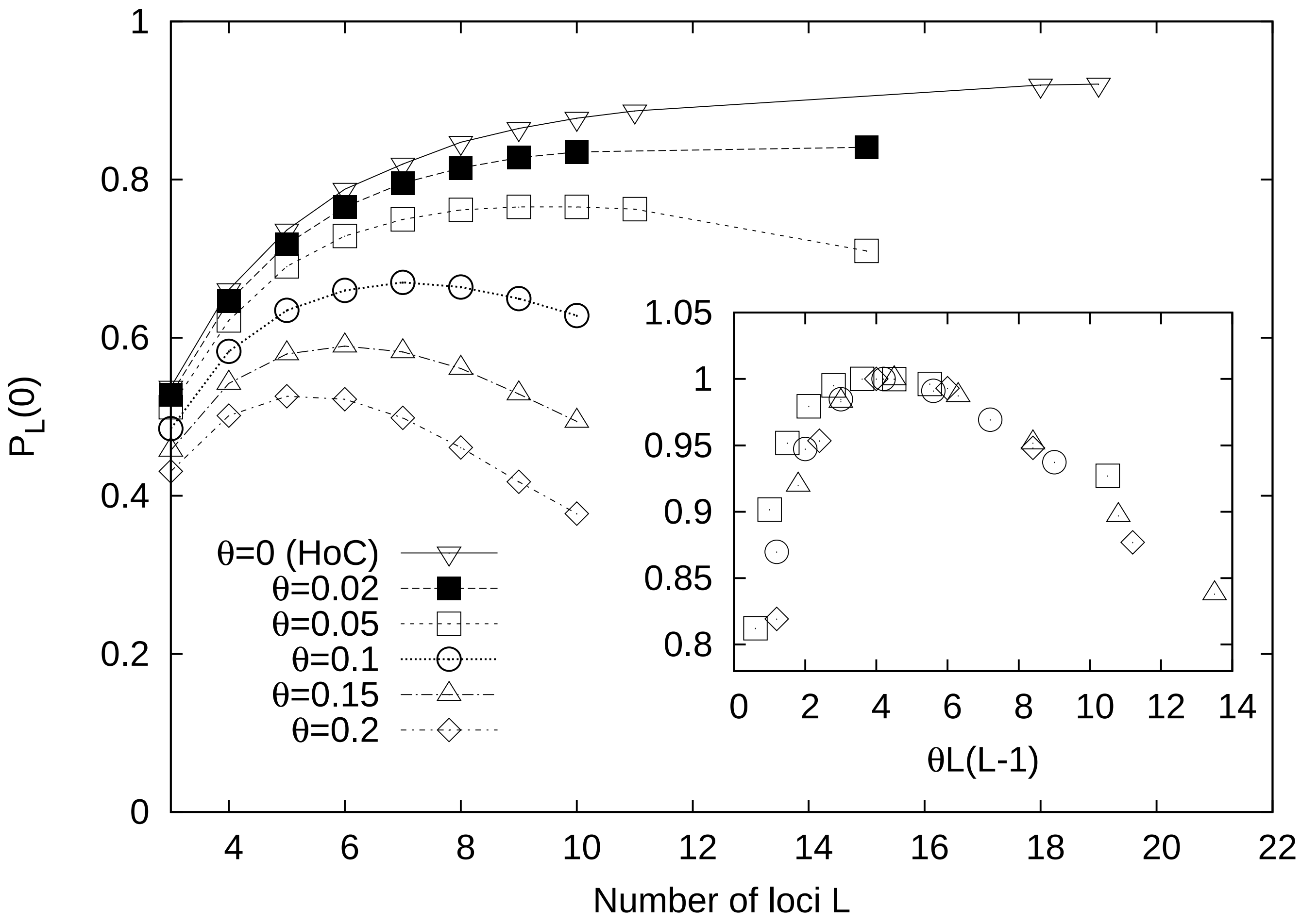
- For general distributions an expansion for small θ yields

$$p_{\mathrm{acc}} \approx \frac{1}{L!} + \frac{\theta}{(L-2)!} \int d\eta\, p(\eta)^2$$

- Since the total number of paths is $L!$, this implies that for any $\theta > 0$

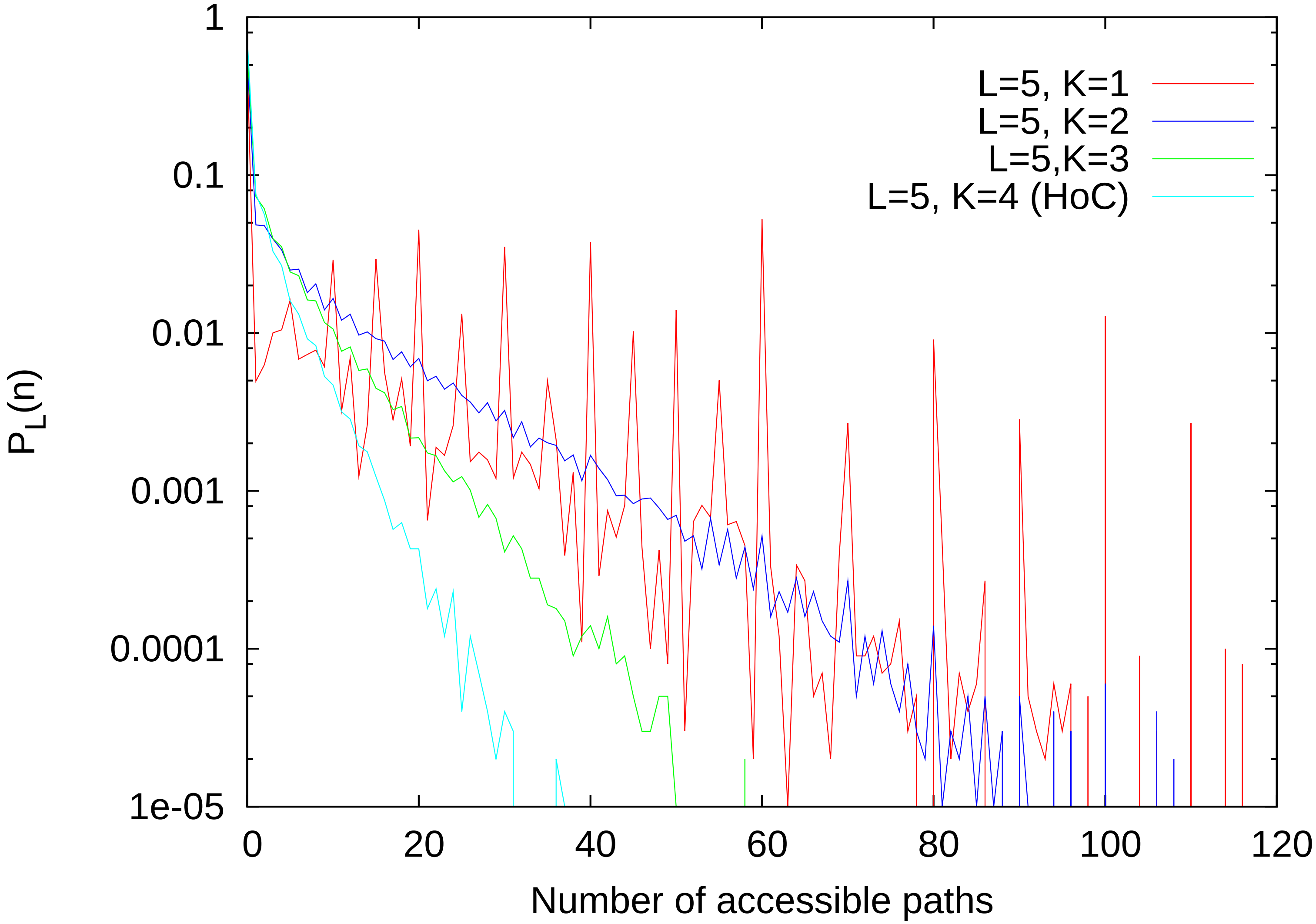
$$\langle n_{\mathrm{acc}} \rangle = L!\, p_{\mathrm{acc}} \to \infty \quad \text{for } L \to \infty$$

Probability of no accessible path in the Mt.Fuji model



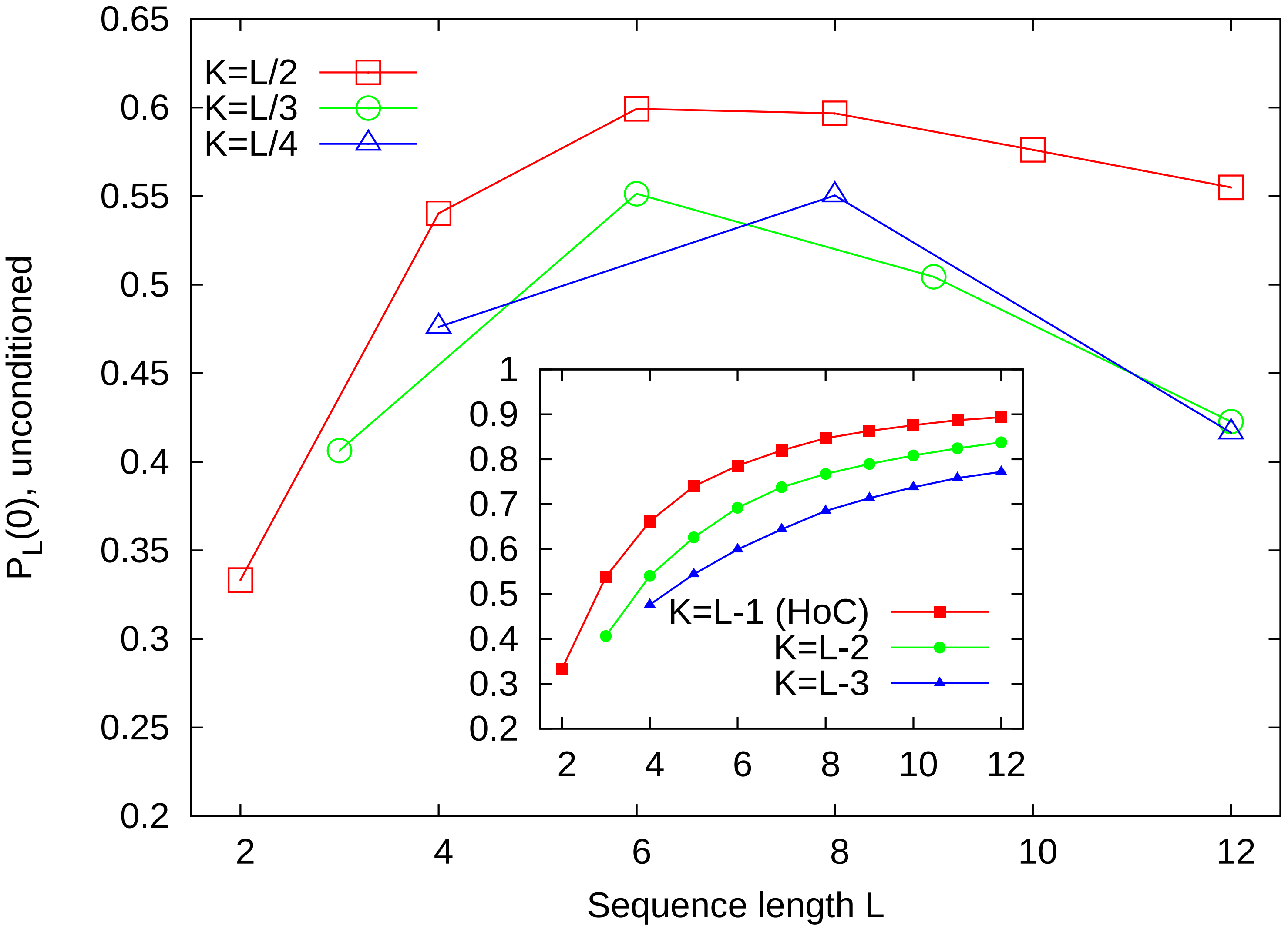
- $P_L(0)$ is generically a non-monotonic function of L
- Beyond the scale $L \sim 1/\sqrt{\theta}$ accessibility increases with increasing L

Distribution of accessible paths in the Kauffman model



- $n_{\mathrm{acc}} = 0$ is the most likely outcome for any $K \geq 1$

Probability of no accessible path in the Kauffman model



- $P_L(0)$ is non-monotonic for K/L fixed but increasing for $L-K$ fixed

Application to the *A. niger* landscape

Effect of lethal mutations

m	# CSG	$\langle n \rangle_{\mathrm{leth}}$	$\langle n \rangle$	$P_m(0)$
2	20 (19.5)	1.61 (1.72)	0.82	0.36
3	29 (28.1)	4.05 (4.22)	1.34	0.39
4	19 (19.5)	12.53 (13.19)	2.01	0.50
5	4 (4.9)	55.32 (48.81)	3.16	0.63
6	0 (0.2)	246.0 (201.16)	6.07	0.68

- CSG: Complete subgraphs not containing any lethal genotypes

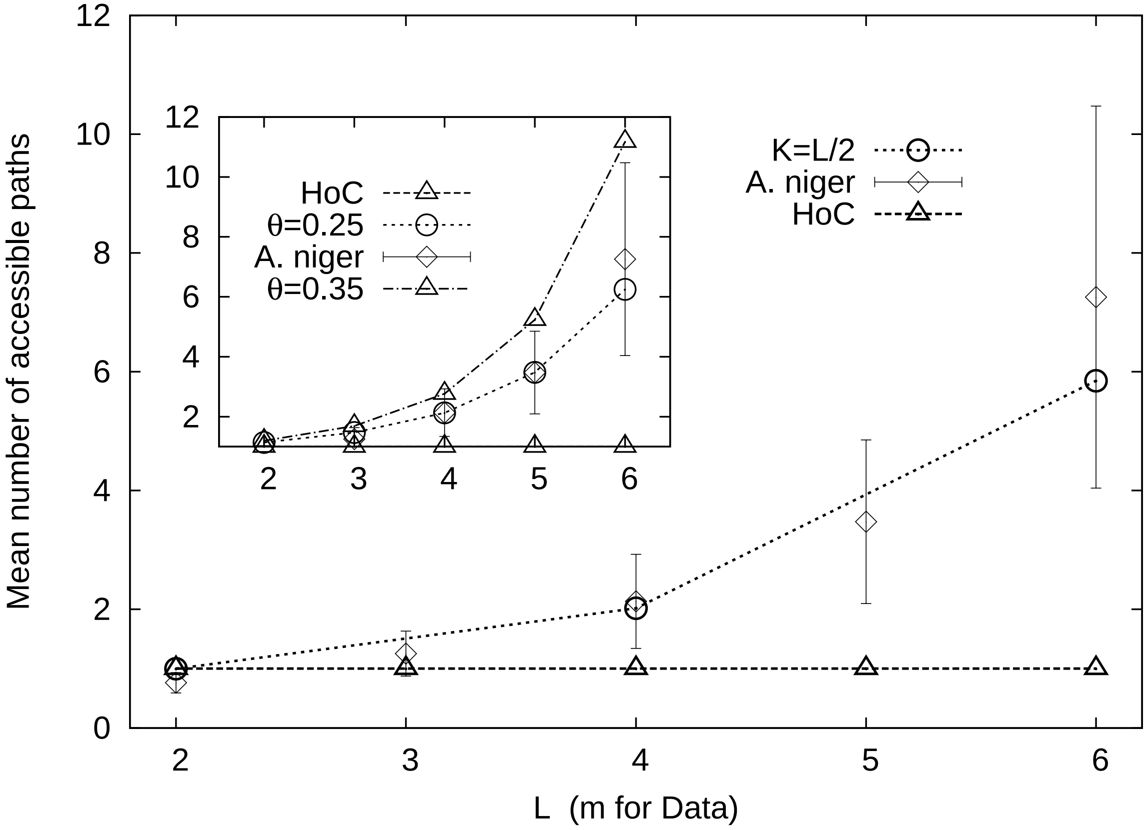
- $\langle n \rangle_{\mathrm{leth}}$: Number of remaining accessible paths if only blocking by lethals is taken into account

- Numbers in brackets show predictions of a simple multiplicative model of lethality

- $\langle n \rangle_{\mathrm{leth}} \gg \langle n \rangle \Rightarrow$ accessibility is limited mainly by epistasis among viable genotypes

- Comparison to models without lethals is therefore meaningful

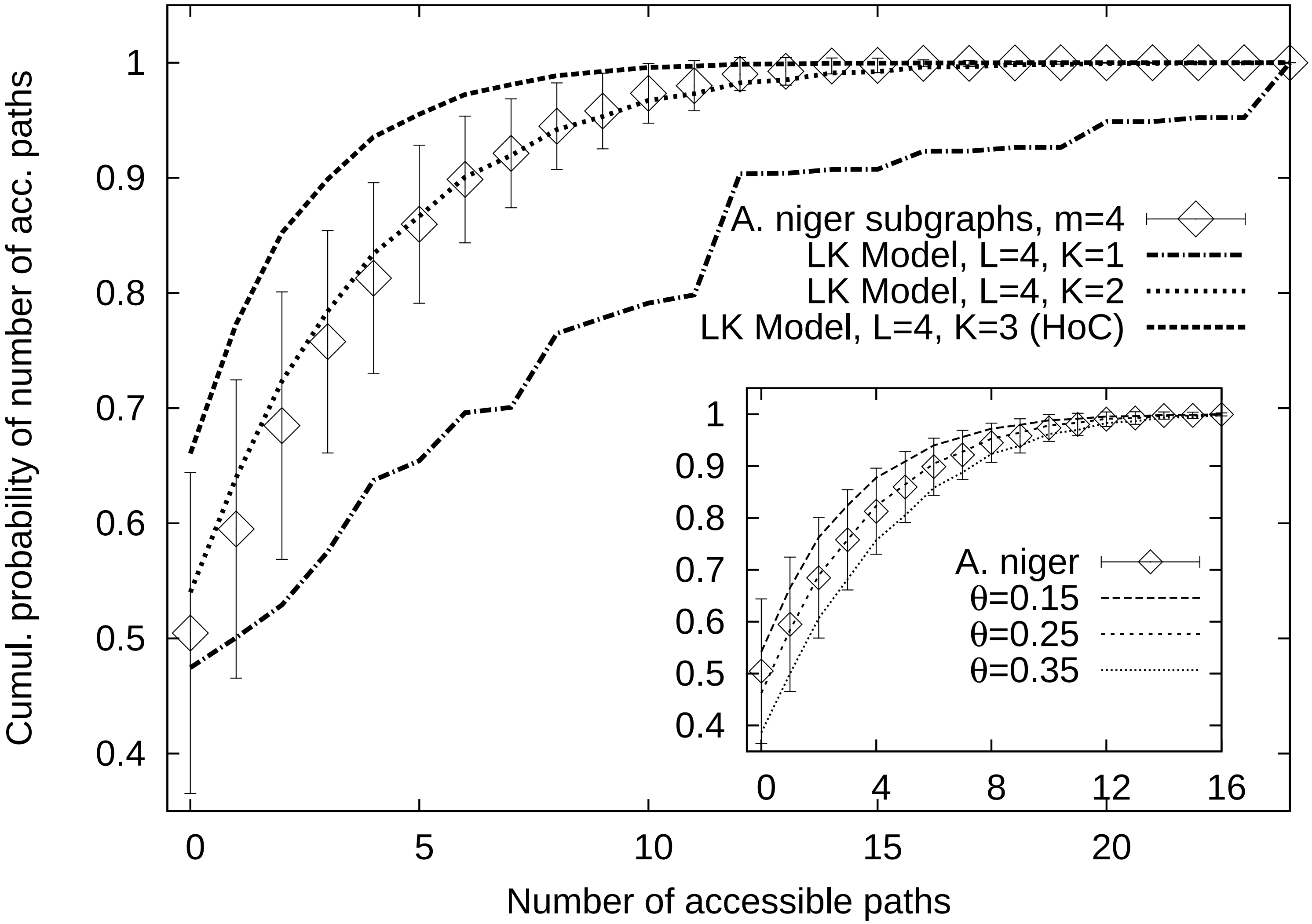
Mean number of accessible paths from subgraph analysis



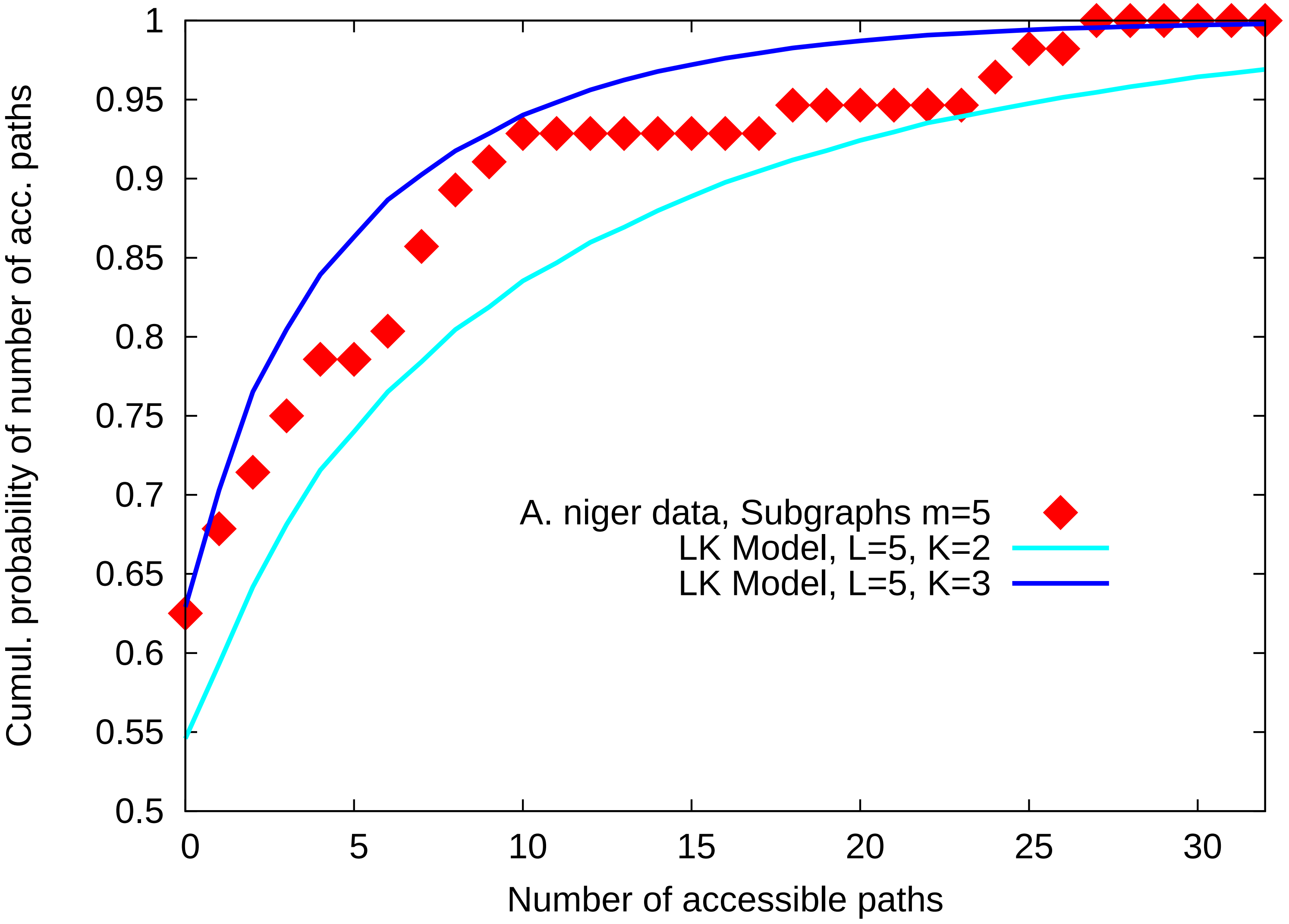
- Error bars from resampling analysis

- 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 ($m = 4$)



Cumulative distribution of the number of paths ($m = 5$)



Summary

- Accessibility of mutational pathways as a measure of fitness landscape ruggedness and predictability of evolution
- Across a wide range of models, accessibility is high (in the sense of $P_L(0) \longrightarrow 0$) and predictability is low (in the sense of $\langle n_{\mathrm{acc}} \rangle \longrightarrow \infty$) for $L \longrightarrow \infty$
- Subgraph analysis of an empirical multilocus fitness landscape confirms these features and allows to estimate epistasis parameters
- Mechanism may be related to percolation on the hypercube:
  Exponential suppression of long paths is overwhelmed by the factorial proliferation in the number of paths