SFB 680
Molecular Basis of
Evolutionary Innovations

Statistical mechanics of fitness landscapes

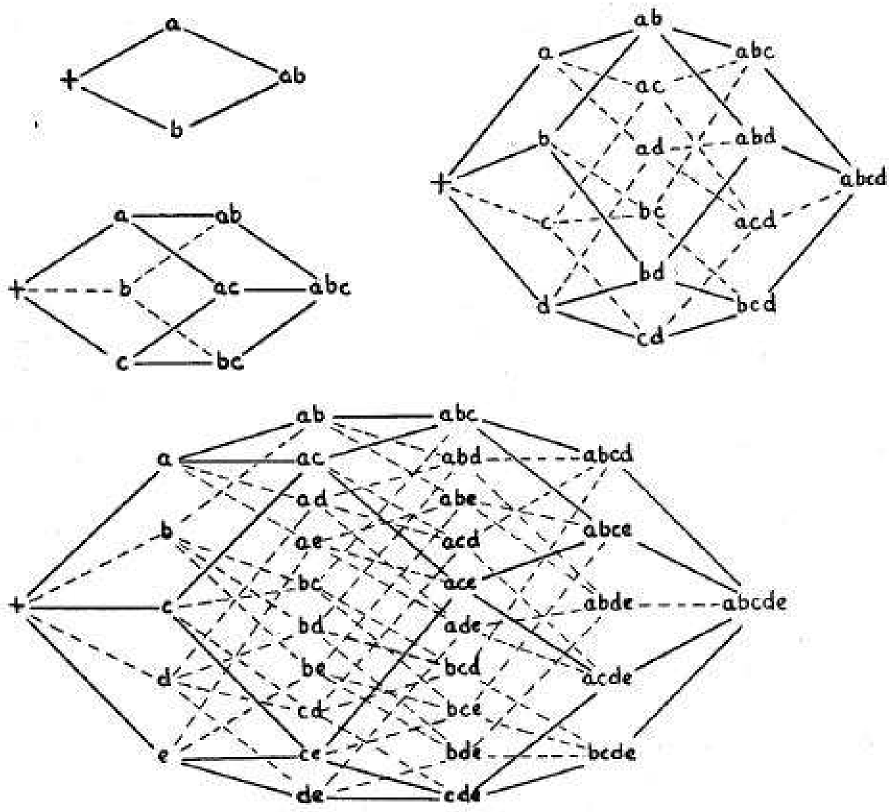
Joachim Krug
Institute for Theoretical Physics, University of Cologne

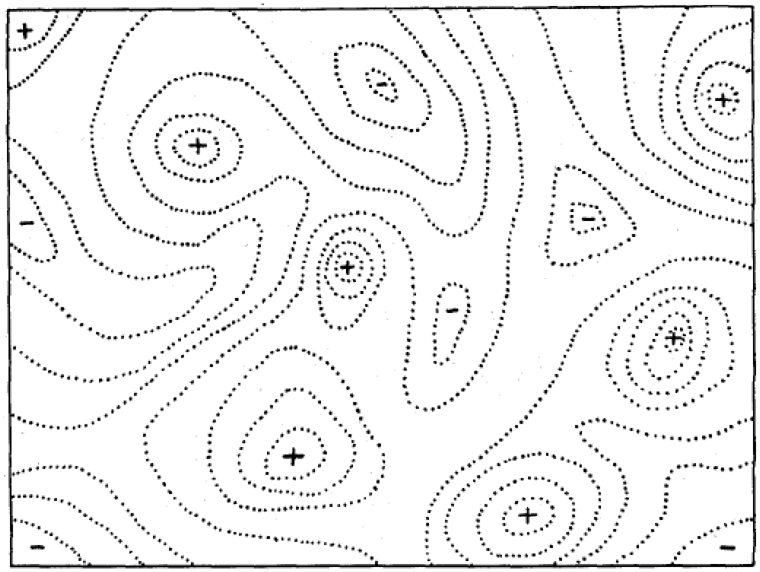
& Jasper Franke, Johannes Neidhart, Stefan Nowak, Benjamin Schmiegelt, Ivan Szendro

Advances in Nonequilibrium Statistical Mechanics
Galileo Galilei Institute, Arcetri, June 6, 2014

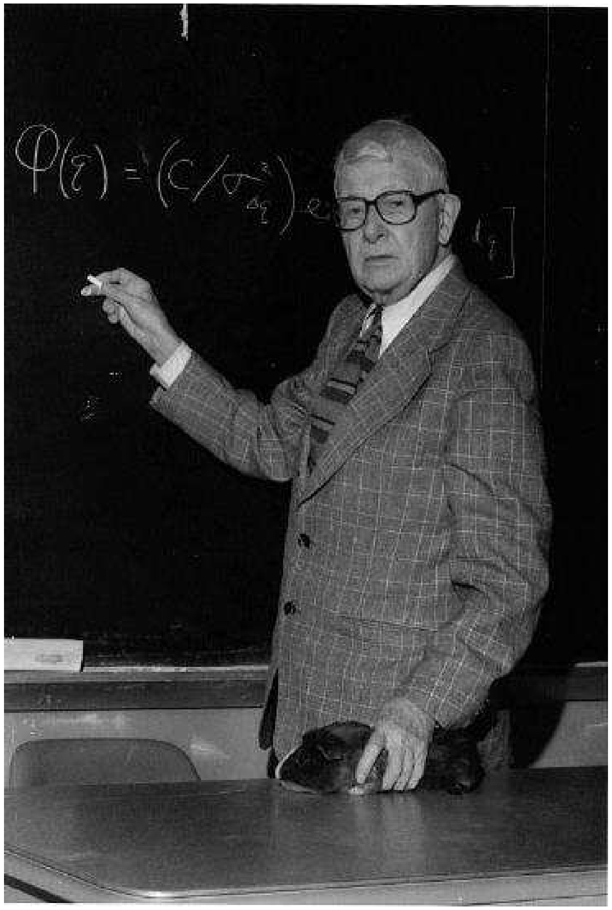
Fitness landscapes

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



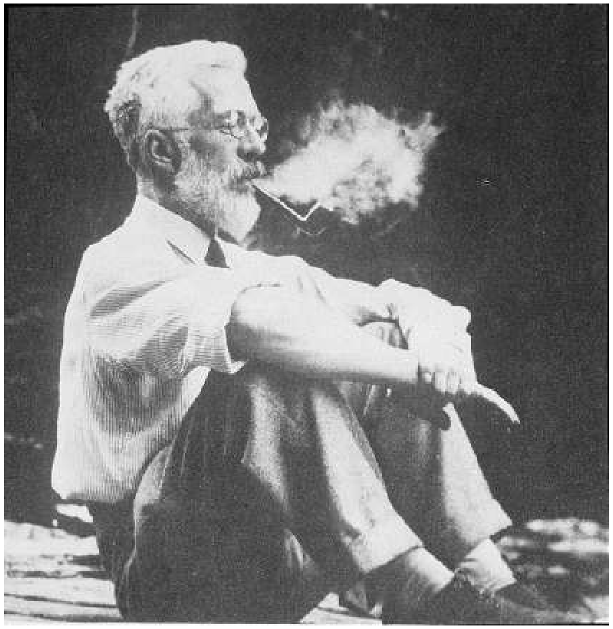


“The two dimensions of figure 2 are a very inadequate representation of such a field.”



Sewall Wright

“In a rugged field of this character, 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 in such a field.”



Ronald A. Fisher

“In one dimension, a curve gives a series of alternate maxima and minima, but in two dimensions two inequalities must be satisfied for a true maximum, and I suppose that only about one fourth of the stationary points will satisfy both. Roughly I would guess that with n factors only 2^{-n} of the stationary points would be stable for all types of displacement, and any new mutation will have a half chance of destroying the stability. This suggests that true stability in the case of many interacting genes may be of rare occurrence, though its consequence when it does occur is especially interesting and important."

Fisher to Wright, 31.5.1931

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 $L \sim 10^2$
- 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
- Genotypic distance: Two sequences are nearest neighbors if they differ in a single letter (mutation)

Mathematical setting

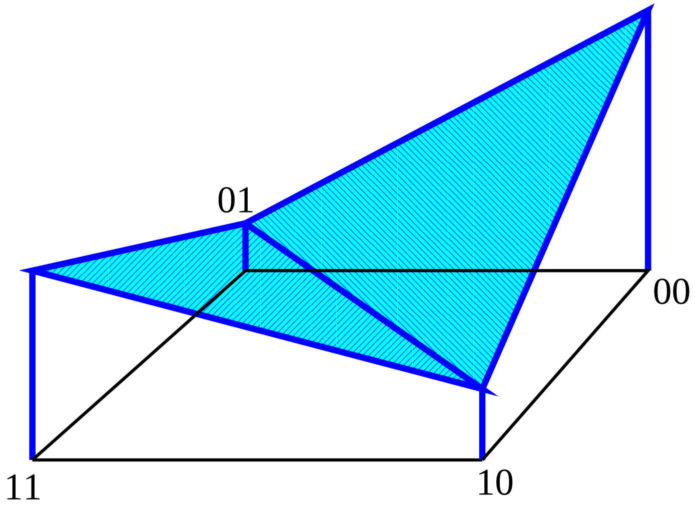
- Genotypes are binary sequences $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_L)$ with $\sigma_i \in \{0,1\}$ or $\sigma_i \in \{-1,1\}$ (presence/absence of mutation).

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

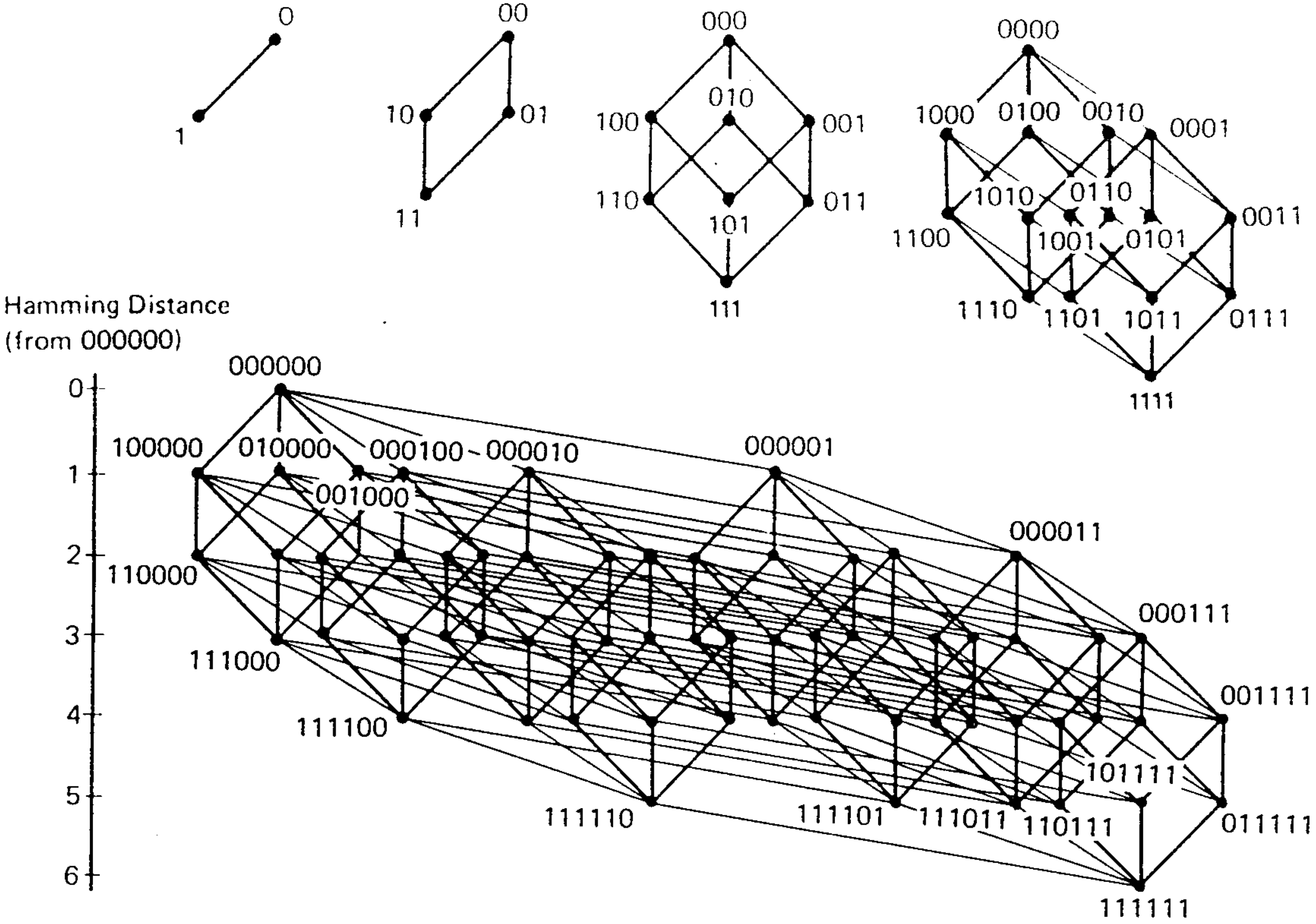
- Epistasis implies interactions between the effects of different mutations

- 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 $L = 2$:



Binary sequence spaces are hypercubes



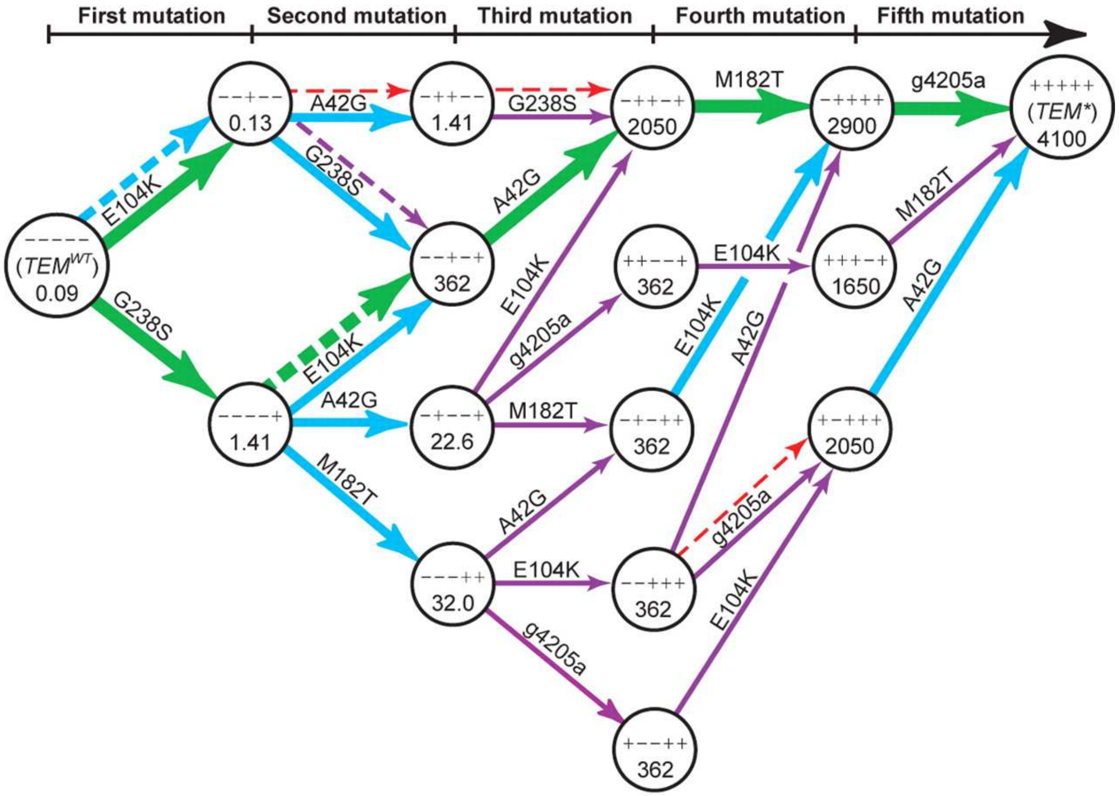
A survey of empirical fitness landscapes

I.G. Szendro, M.F. Schenk, J. Franke, JK, J.A.G.M. de Visser
J. Stat. Mech. P01005 (2013), special issue on Evolutionary Dynamics

J.A.G.M. de Visser, JK
Nature Reviews Genetics (in press)

Pathways to antibiotic resistance

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

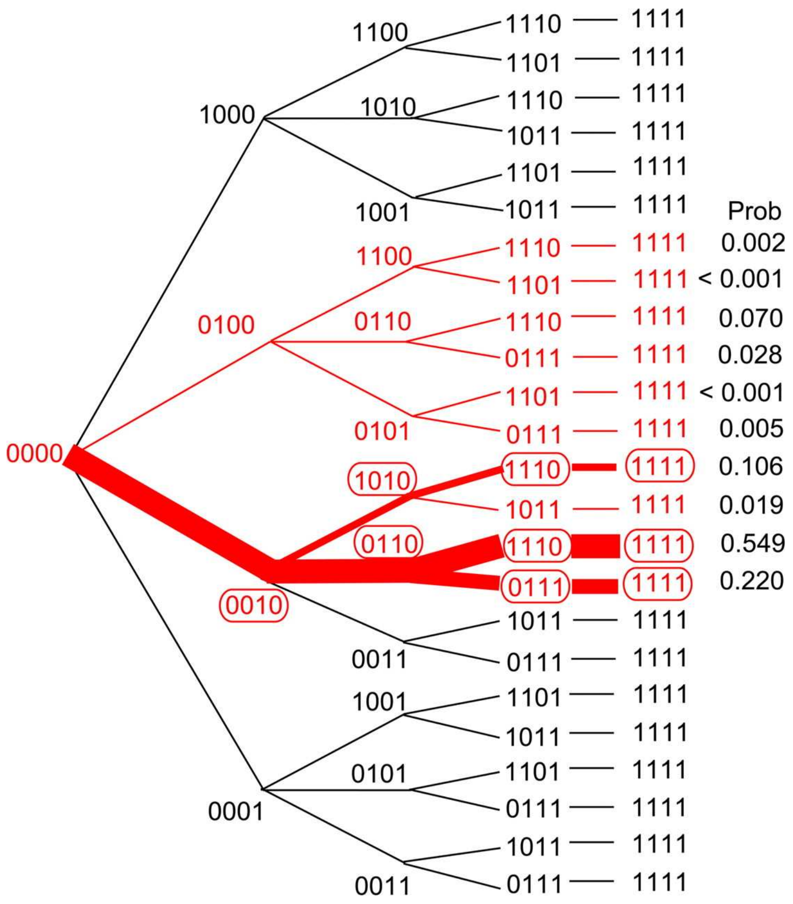


- 5 mutations in the β-lactamase enzyme confer resistance to cefotaxime

- 5! = 120 different mutational pathways, out of which 18 are monotonically increasing in resistance; figure shows 10 “most important” paths

Pyrimethamine resistance in the malaria parasite

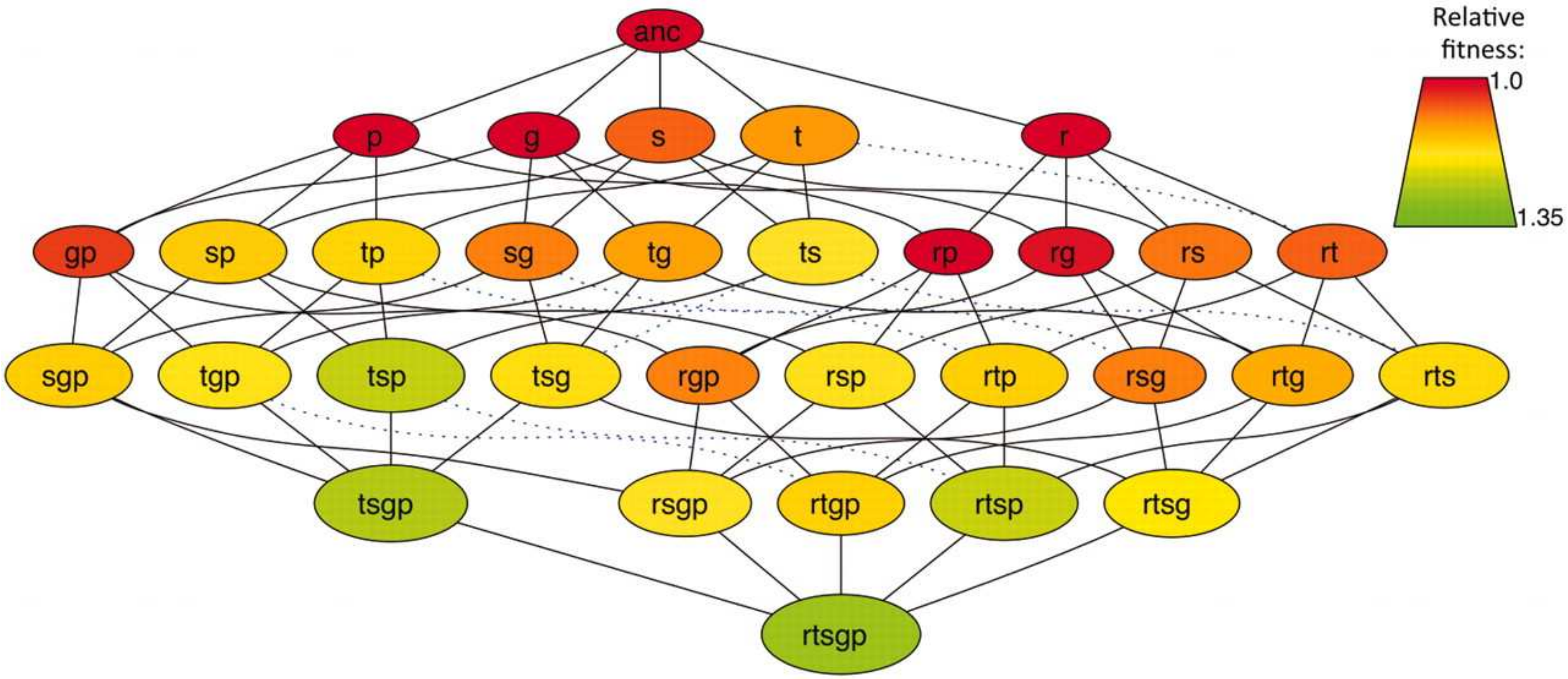
E.R. Lozovsky et al., Proc. Natl. Acad. Sci. USA **106**, 12025 (2009)



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

Five mutations from a long-term evolution experiment with *E. coli*

A.I. Khan et al., Science **332** (2011) 1193

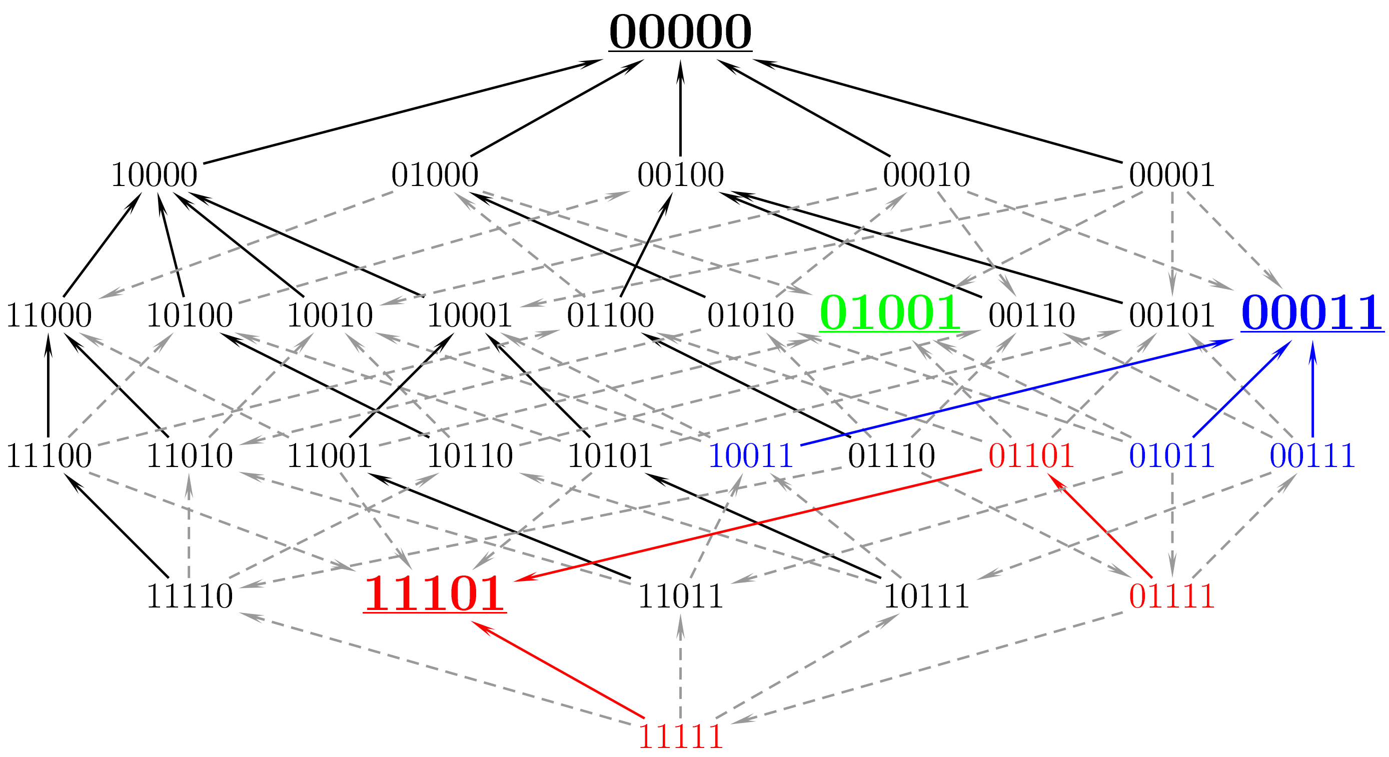


- single fitness peak, 86 out of $5! = 120$ pathways are monotonic

$\Rightarrow$ landscape is rather smooth

The *Aspergillus niger* fitness landscape

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



- Combinations of 8 individually deleterious marker mutations (one out of $\binom{8}{5} = 56$ five-dimensional subsets shown)

- Arrows point to increasing fitness, 3 local fitness optima highlighted

Measures of landscape ruggedness

Local fitness optima

Haldane 1931, Wright 1932

- A genotype σ is a local optimum if $f(\sigma) > f(\sigma')$ for all one-mutant neighbors σ'
- In the absence of sign epistasis there is a single global optimum
- Reciprocal sign epistasis is a necessary but not sufficient condition for the existence of multiple fitness peaks Poelwijk et al. 2011, Crona et al. 2013

Selectively accessible paths

Weinreich et al. 2005

- A path of single mutations connecting two genotypes $\sigma \rightarrow \sigma'$ with $f(\sigma) < f(\sigma')$ 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

Probabilistic models of fitness landscapes

House-of-cards/random energy model

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

  Kingman 1978, Kauffman & Levin 1987

- What is the expected number of fitness maxima?

- A genotype has L neighbors and is a local maxima if its fitness is the largest among $L+1$ i.i.d. random variables, which is true with probability $\frac{1}{L+1}$

$$\Rightarrow \quad \mathbb{E}(n_{\max}) = \frac{2^L}{L+1}$$

- Density of maxima decays algebraically rather than exponentially with L

- Variance of the number of maxima

  Macken & Perelson 1989

$$\mathrm{Var}(n_{\max}) = \frac{2^L(L-1)}{2(L+1)^2} \rightarrow \frac{1}{2}\mathbb{E}(n_{\max}) \quad \text{for} \quad L \rightarrow \infty$$

Accessible pathways in the house-of-cards model

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

- What is the expected number of shortest, fitness-monotonic 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 $f_0, \ldots., f_{d-1}$.

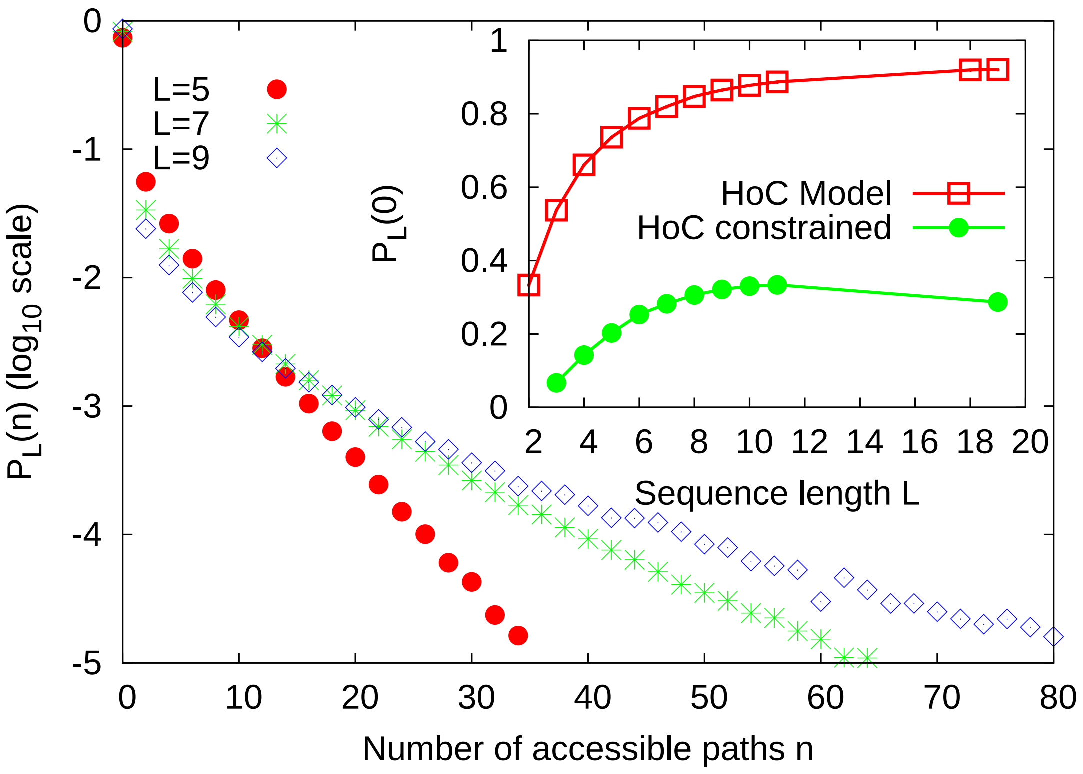
- A path is accessible iff $f_0 < f_1 \ldots. < f_{d-1}$

- 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 holds in particular for the $L!$ paths from the antipodal point of the global optimum.

Distribution of number of accessible paths from antipodal genotype



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

- Characterize the distribution $P_L(n)$ by $\mathbb{E}(n_{\text{acc}})$ and the probability $P_L(0)$ that no path is accessible $\Rightarrow$ define accessibility as $\overline{P}_L \equiv 1 - P_L(0)$

“Accessibility percolation” as a function of initial fitness

- When fitnesses are drawn from the uniform distribution and the fitness of the initial genotype is f_0, then Hegarty & Martinsson, arXiv:1210.4798

$$\lim_{L\to\infty} \overline{P}_L = \begin{cases} 0 & \text{for} \quad f_0 > \frac{\ln L}{L} \\ 1 & \text{for} \quad f_0 < \frac{\ln L}{L}, \end{cases}$$

- This implies in particular that $\lim_{L\to\infty} \overline{P}_L = 0$ for the HoC model with unconstrained initial fitness

- If arbitrary paths with backsteps are allowed, the accessibility threshold becomes independent of L and is conjectured to be $1 - \frac{1}{2}\sinh^{-1}(2) \approx 0.27818\ldots$ Berestycki, Brunet, Shi, arXiv:1401.6894

- On a regular tree of height h and branching number b the accessibility threshold for $h, b \longrightarrow \infty$ occurs at $h/b = e$

Nowak & Krug, EPL 2013; Roberts & Zhao, ECP 2013

Landscapes with tunable ruggedness

Kauffman's NK-model

Kauffman & Weinberger 1989

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

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

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

- $K=0$: Non-interacting $\quad$ $K=L-1$: House-of-cards

Rough Mount Fuji model

Aita et al. 2000; Neidhart et al., arXiv:1402.3065

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

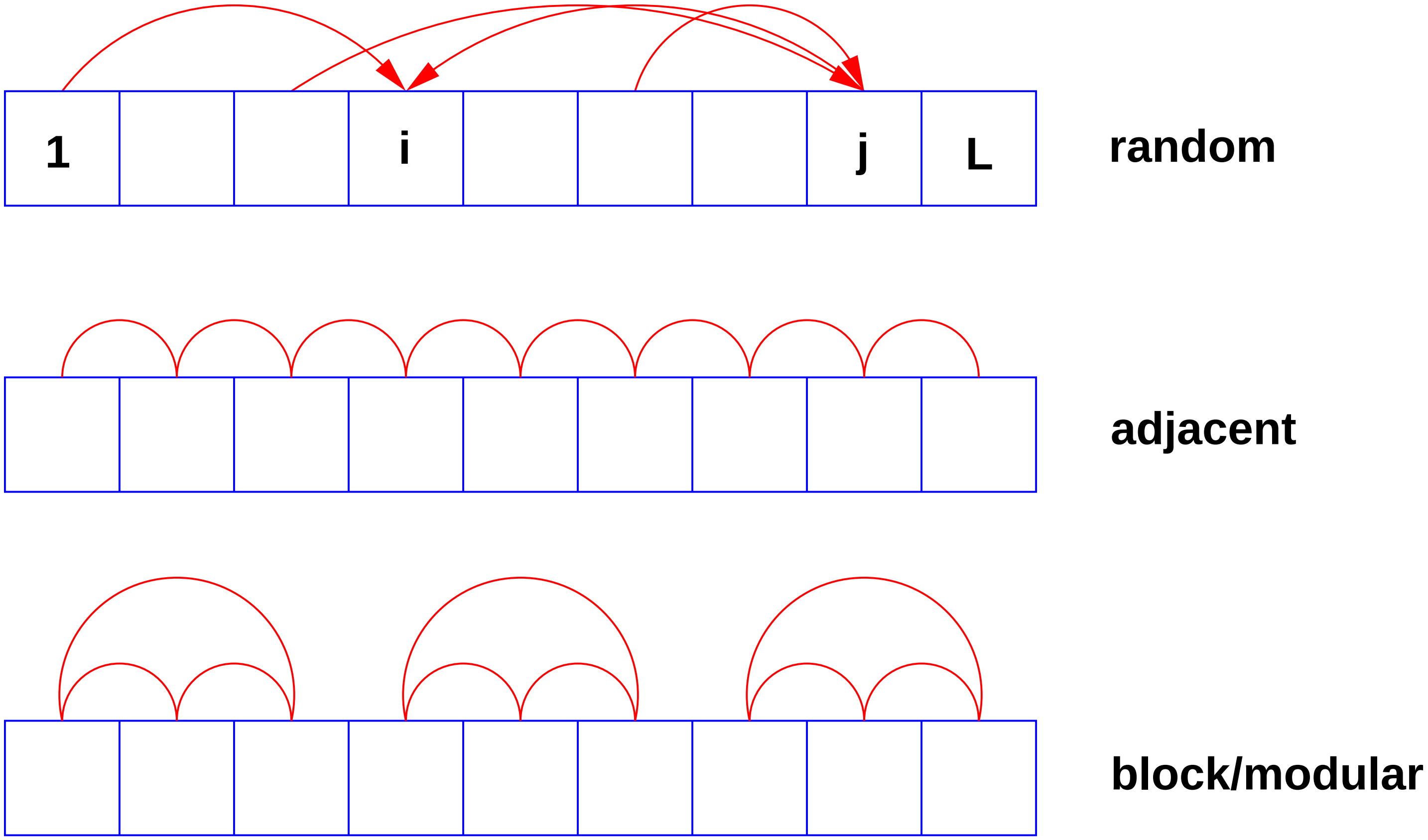
$$f(\sigma) = -cd(\sigma, \sigma^*) + \eta(\sigma) \quad c > 0$$

η: i.i.d. random variables $\quad$ $d(\sigma, \sigma')$: Hamming distance

- Equivalent to a random energy model in a magnetic field

“Genetic architecture” in Kauffman’s NK-model

- Different schemes for choosing the interaction partners:



1 i j L

random

adjacent

block/modular

- Which properties of the fitness landscape are sensitive to this choice?

“Genetic architecture” in Kauffman’s NK-model

- Fitness correlation function is manifestly independent of the neighborhood scheme P.R.A. Campos, C. Adami, C.O. Wilke (2002)

- This implies independence also for the Fourier spectrum of the landscape, which is given by $\tilde{F}_p = 2^{-(K+1)}\binom{K+1}{p}$ J. Neidhart, I.G. Szendro, JK 2013

- In the block model, the mean number of local maxima is given exactly by

$$\mathbb{E}(n_{\text{max}}^{\text{block}}) = \left(\frac{2^{K+1}}{(K+1)+1}\right)^B = \frac{2^L}{(K+2)^{L/(K+1)}}$$

Perelson & Macken 1995

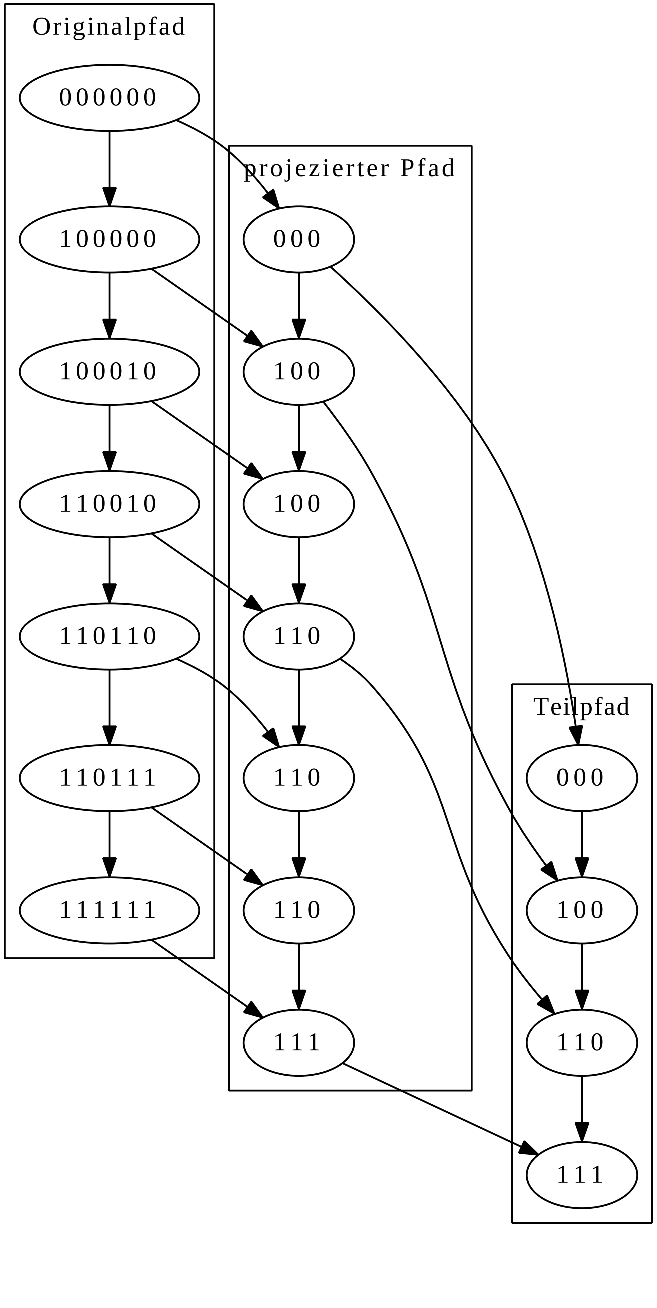
where $B = \frac{L}{K+1}$ is the number of blocks of size $K+1$ each

- Mean number of accessible paths in the block model:

$$\mathbb{E}(n_{\text{acc}}^{\text{block}}) = \frac{L!}{[(K+1)!]^{L/(K+1)}}$$

B. Schmiegelt, JK 2014

Path decomposition for the block model



Evolutionary accessibility in the block model

B. Schmiegelt, JK, J. Stat. Phys. **154**, 334 (2014)

- A given pathway spanning the whole landscape is accessible iff all subpaths within the $B = L/(K+1)$ blocks are accessible

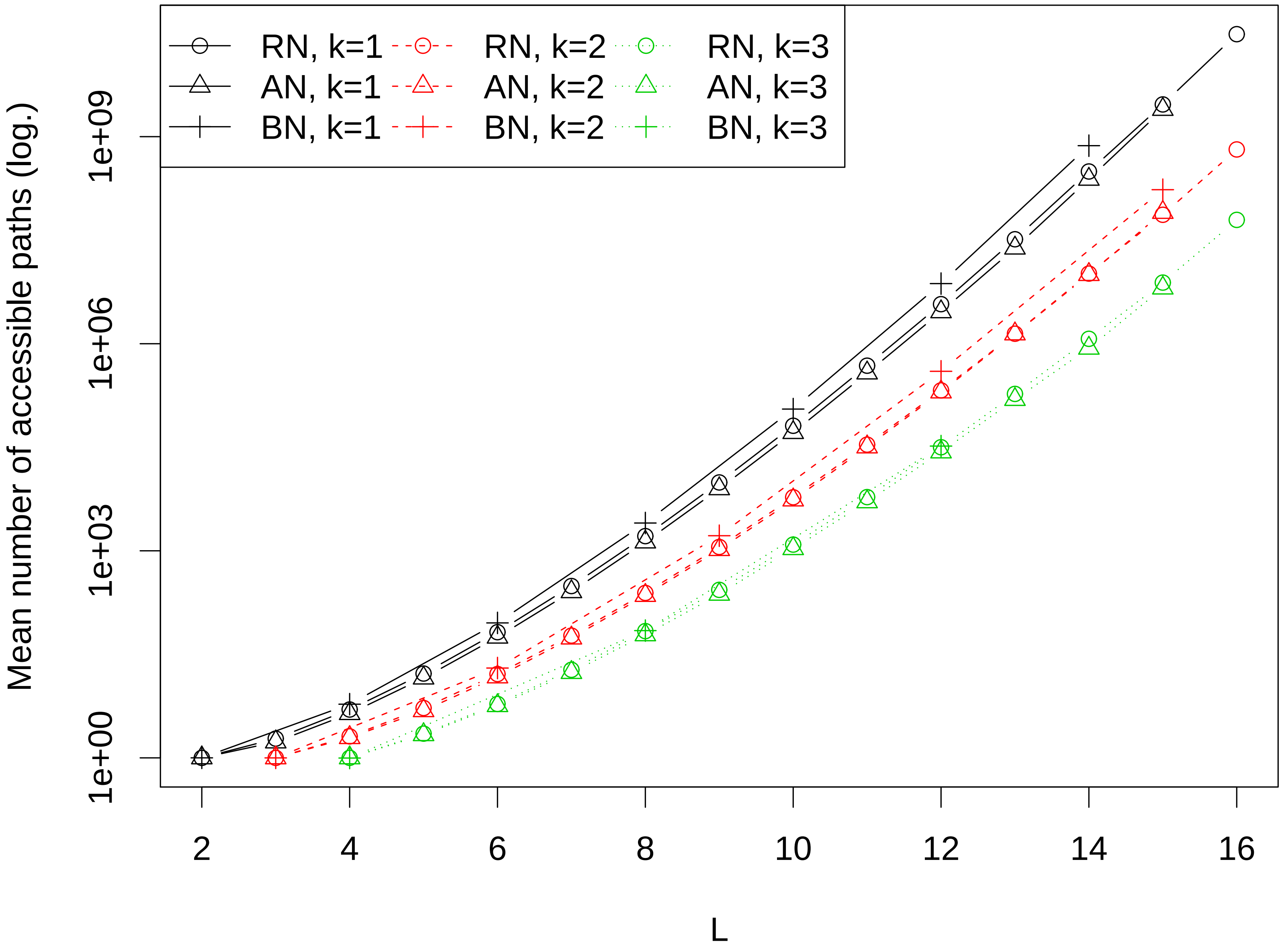
- Each combination of accessible subpaths can be combined into $\frac{L!}{[(K+1)!]^B}$ global paths

$$\Rightarrow \quad n_{\text{acc}}^{\text{block}} = \frac{L!}{[(K+1)!]^B} \prod_{i=1}^{B} n_{\text{acc}}^{(i)}$$

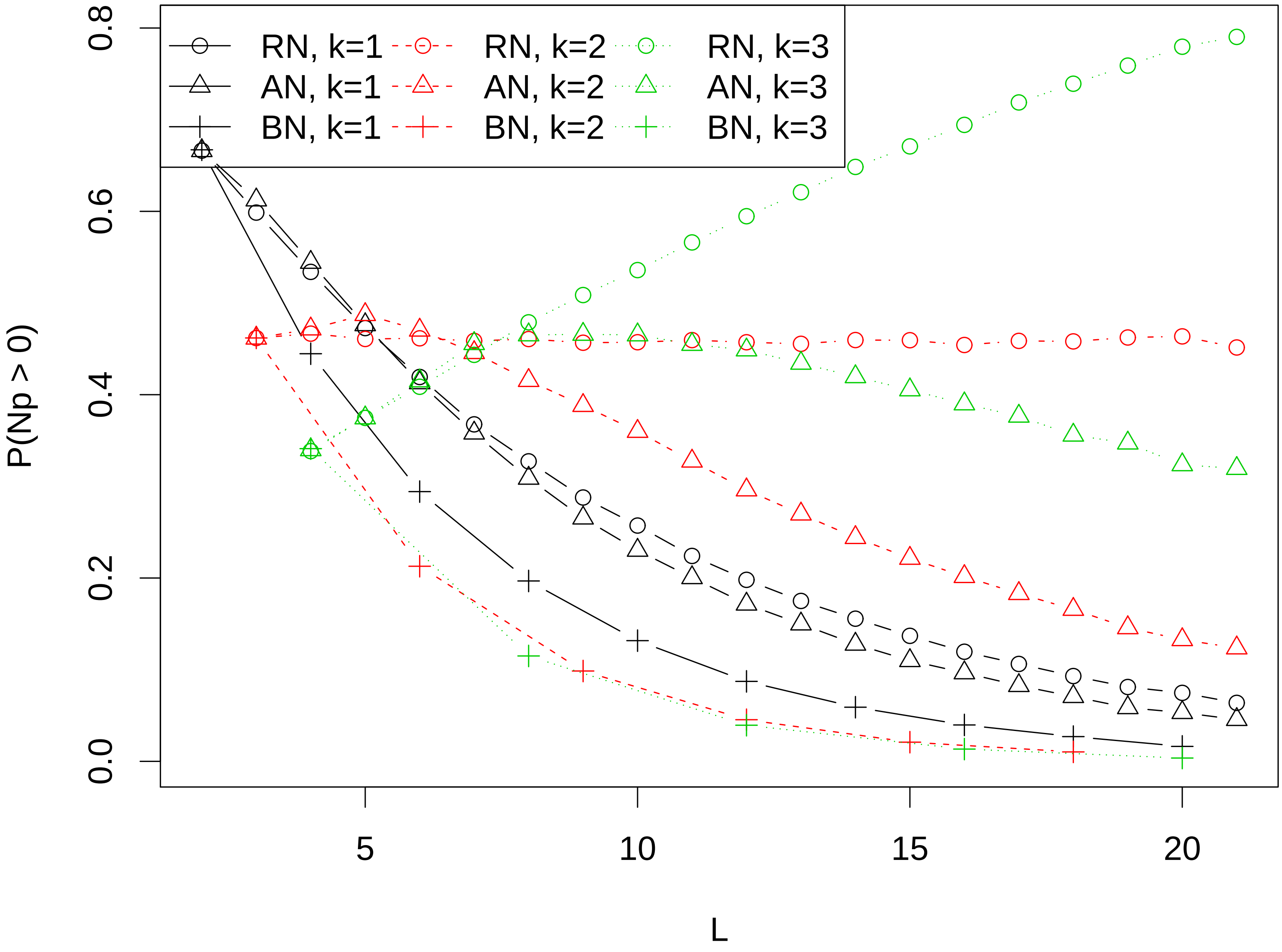
- Since the blocks are HoC-landscapes of size $K+1$, the expected number of accessible paths is $\mathbb{E}(n_{\text{acc}}^{\text{block}}) = \frac{L!}{[(K+1)!]^B}$ and the accessibility is $\overline{P}_L^{\text{block}} = [\overline{P}_{K+1}^{\text{HoC}}]^{\frac{L}{K+1}}$ which approaches zero exponentially fast in L for any K

- This implies that most landscape have no path to the maximum (low accessibility) but those that do have many (low predictability)

Mean number of paths is insensitive to genetic architecture



...but accessibility appears to be very sensitive



Distribution of the number of accessible paths in the block model

- Path number distribution in terms of HoC model:

$$P_L(n) = \begin{cases} \sum\limits_{D_B(z)} \prod\limits_{i=1}^{B} P_L^{\mathrm{HoC}(K+1)}(n_i) & \text{if} \quad z = \dfrac{[(K+1)!]^B}{L!} \cdot N \in \mathbb{N}_0 \\ 0 & \text{else,} \end{cases}$$

where $D_B(z) = \{(n_1,\ldots,n_B) \in \mathbb{N}_0^B \mid \prod_{i=1}^B n_i = z\}$

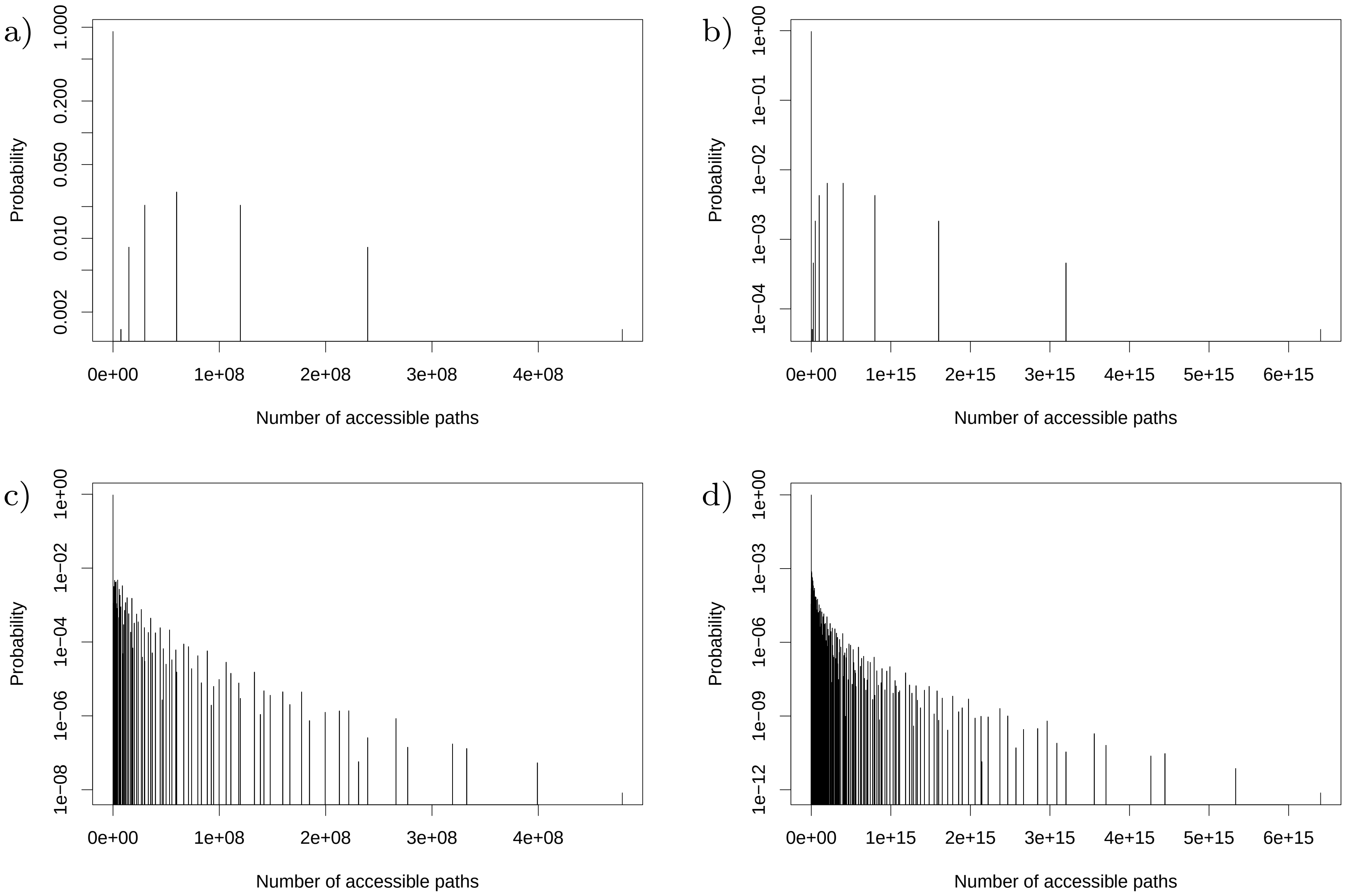
- HoC distribution is exactly known for sequence lengths 2 and 3

- In particular for $K = 1$ the HoC paths numbers are 0, 1 or 2 and

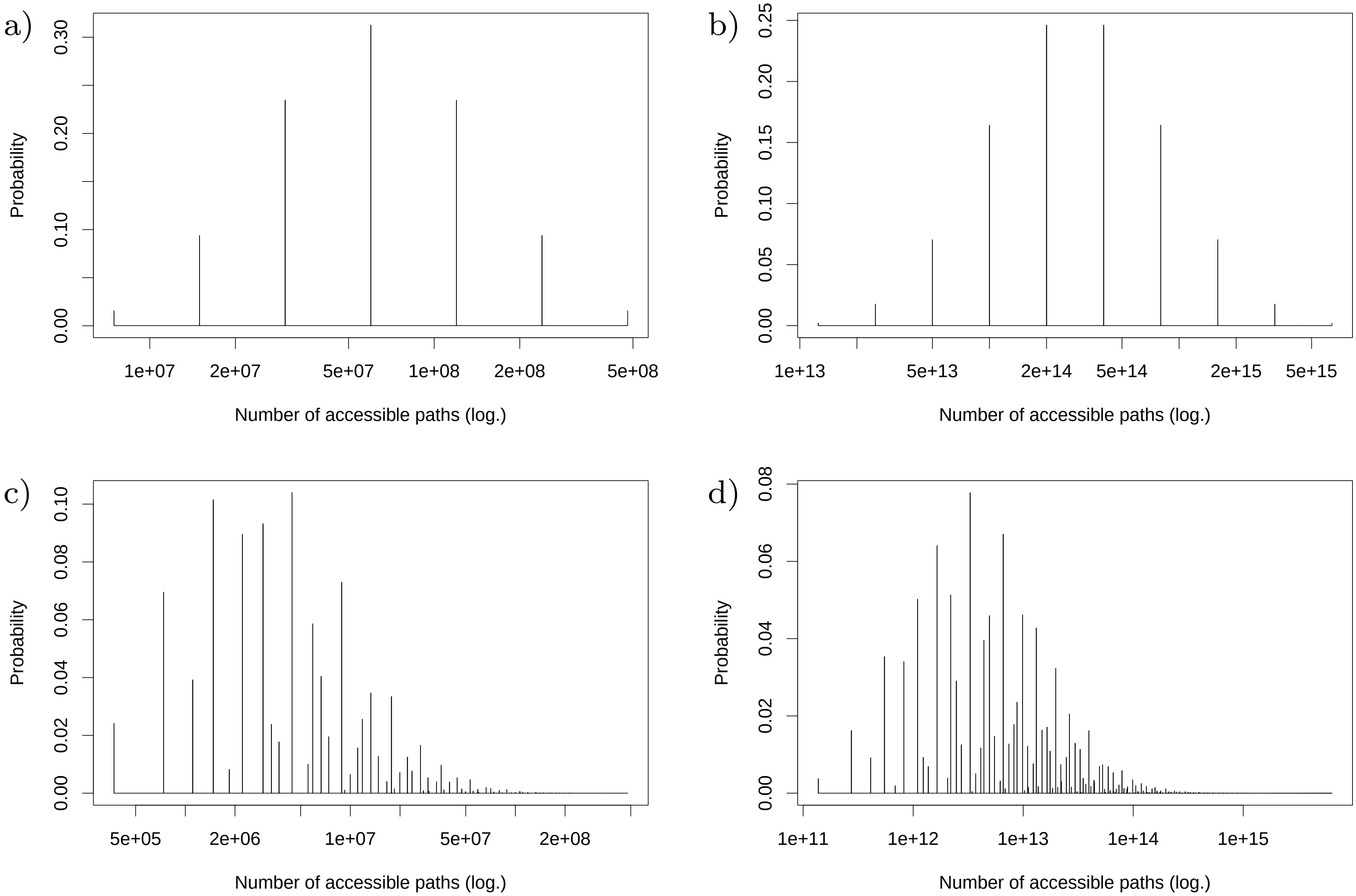
$$P_L(n \neq 0) = 3^{-B} \binom{B}{k} \delta_{n,n_k}, \quad k = 0, 1, \ldots, B = \frac{L}{K+1}$$

with $n_k = L!2^{k-B}$, and $P_L(0) = 1 - \left(\frac{2}{3}\right)^{L/2}$.

Exact path number distributions for $L = 12, 18$ and $K = 1, 2$



Exact path number distributions for $L = 12, 18$ and $K = 1, 2$



Asymptotics of the number of maxima

Number of maxima in the NK-model

- Rigorous work on the NK-model with adjacent neighborhoods shows that for fixed K

  $$\mathscr{M} \equiv \mathbb{E}(n_{\max}) \sim \lambda_K^L \quad \text{for} \quad L \longrightarrow \infty$$

  with constants $\lambda_K \in (1,2)$ Evans & Steinsaltz 2002, Durrett & Limic 2003

- The exact result for the block model is of this form with $\lambda_K = (K+2)^{-\frac{1}{K+1}}$

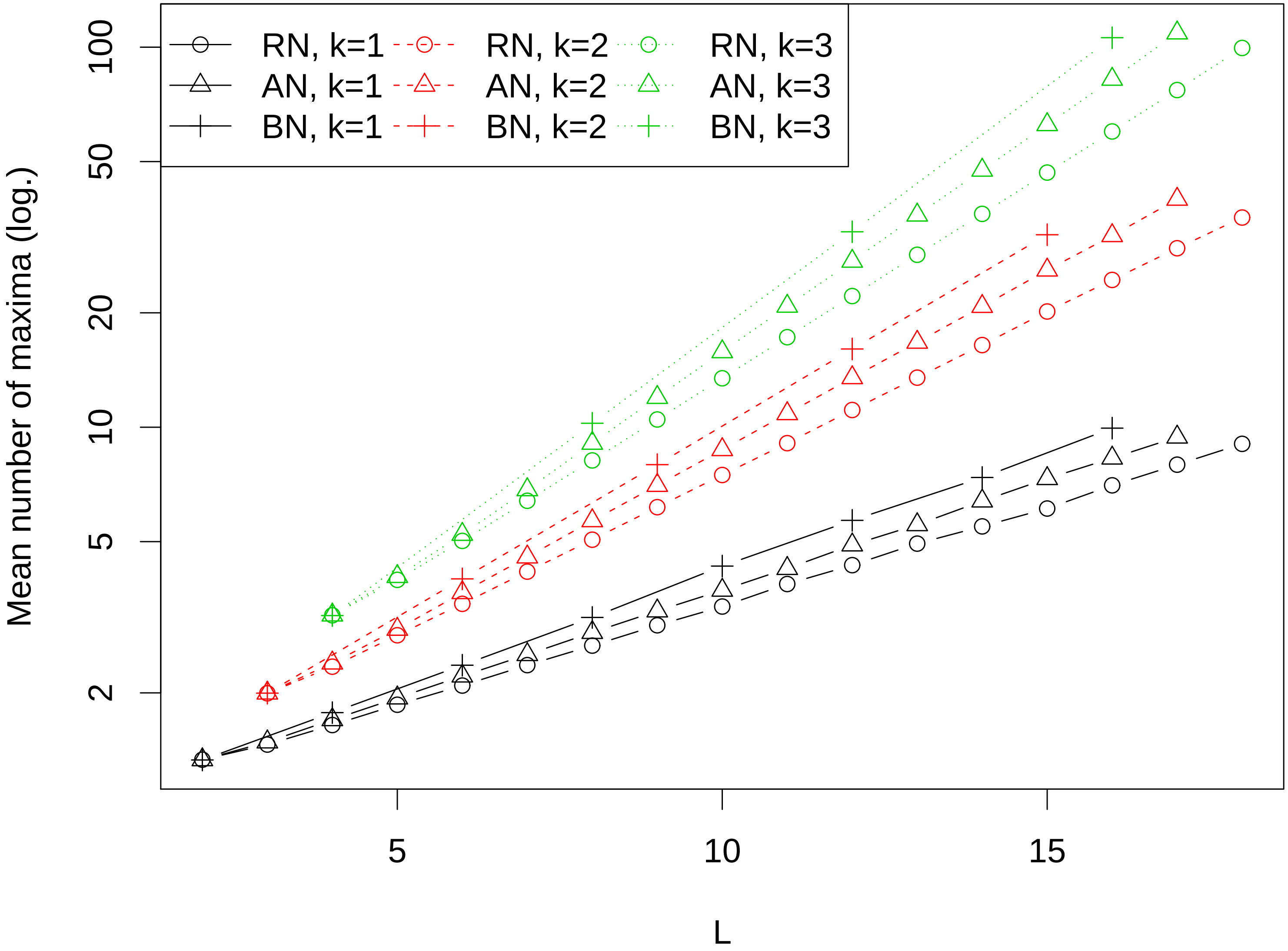
- Known explicit values for λ_K are remarkably close but not identical to the block model result, e.g. for $K = 1$:

  $$0.55463... \leq \lambda_1 \leq 0.5769536... < 3^{-1/2} = 0.57735...$$

- When the limits $L \longrightarrow \infty$ and $K \longrightarrow \infty$ are taken simultaneously with $\alpha = L/K$ fixed, rigorous analysis shows that $\mathscr{M} \sim \frac{2^L}{L^\alpha}$, which is also true for the block model. Limic & Pemantle 2004

Mean number of maxima for different genetic architectures

B. Schmiegelt, JK, J. Stat. Phys. **154**, 334 (2014)



Number of maxima in the Rough Mount Fuji model

J. Neidhart, I.G. Szendro, JK, in preparation

- A genotype at distance d from the reference sequence σ^* has d neighbors in the 'uphill' direction and $L-d$ neighbors in the 'downhill' direction

- The fitness distribution of uphill/downhill neighbors is shifted by $\pm c$ with respect to the fitness distribution of the focal genotype

- Denoting by $P(x) = \mathbb{P}(\eta < x)$ the probability distribution of the random fitness component and by $p(x) = \frac{dP}{dx}$ the corresponding density, the probability that a genotype at distance d is a local maximum is therefore

$$p_{\max}(d) = \int dx \; p(x) P(x-c)^d P(x+c)^{L-d}$$

and the expected total number of maxima is

$$\mathscr{M} = \sum_{d=0}^{L} \binom{L}{d} p_{\max}(d) = \int dx \; p(x) [P(x-c) + P(x+c)]^L$$

Classification in terms of tail behavior of $P(x)$

- For distributions with tail heavier than exponential (power law or stretched exponential) $\mathscr{M} \longrightarrow \frac{2^L}{L}$ for $L \longrightarrow \infty$, which implies that the fitness gradient (c) is asymptotically irrelevant
- For distributions with an exponential tail $\mathscr{M} \longrightarrow \frac{2^L}{\cosh(c)L}$ for large L
- For distributions with tails lighter than exponential such as $1 - P(x) \sim \exp[-x^\beta]$ with $\beta > 1$ the number of maxima behaves to leading order as

$$\mathscr{M} \sim \frac{2^L}{L} \exp[-\beta c (\ln L)^{1-\frac{1}{\beta}}]$$

- For distributions with bounded support on $[0,1]$ and boundary singularity $1 - P(x) \sim (1-x)^\nu$ the asymptotic behavior is of the form

$$\mathscr{M} \sim \frac{(2-c^\nu)^L}{L^\nu}$$

for $c < 1$ and $\mathscr{M} = 1$ for $c > 1$.

Summary

- The fitness landscape over the space of genotypes is a key concept in evolutionary biology that has only recently become accessible to empirical exploration

- Mathematical analysis of probabilistic models can help to extrapolate from the low dimensionality of existing empirical data sets to genome-wide scales

- Notion of pathway accessibility defines a new class of percolation-type problems on hypercubes and other graphs

- Comparison between fitness landscape models with tunable ruggedness shows that similar asymptotics for the number of maxima can arise through different mathematical mechanisms