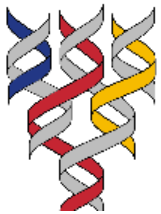
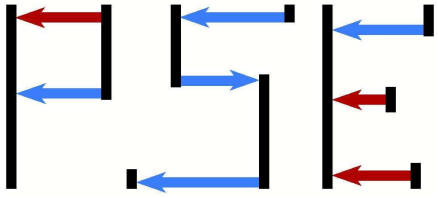


SFB 680
Molecular Basis of
Evolutionary Innovations



Adaptive walks on correlated fitness landscapes

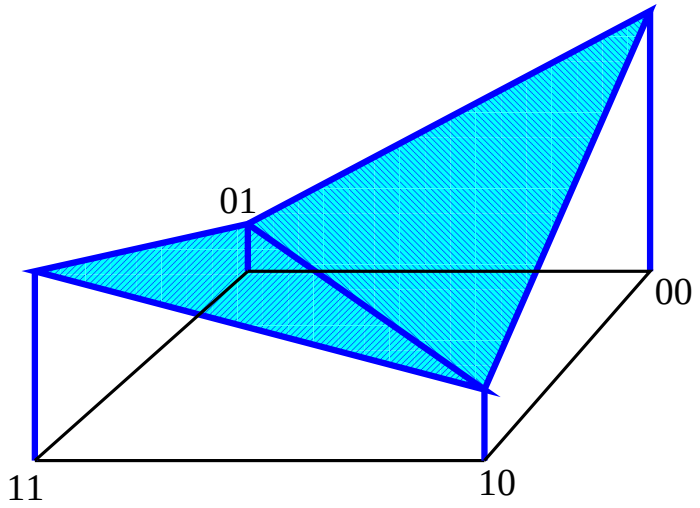
Joachim Krug
Institute for Theoretical Physics, University of Cologne

with J. Neidhart, S. Nowak, I. Szendro (Cologne)
& S.-C. Park (Seoul)

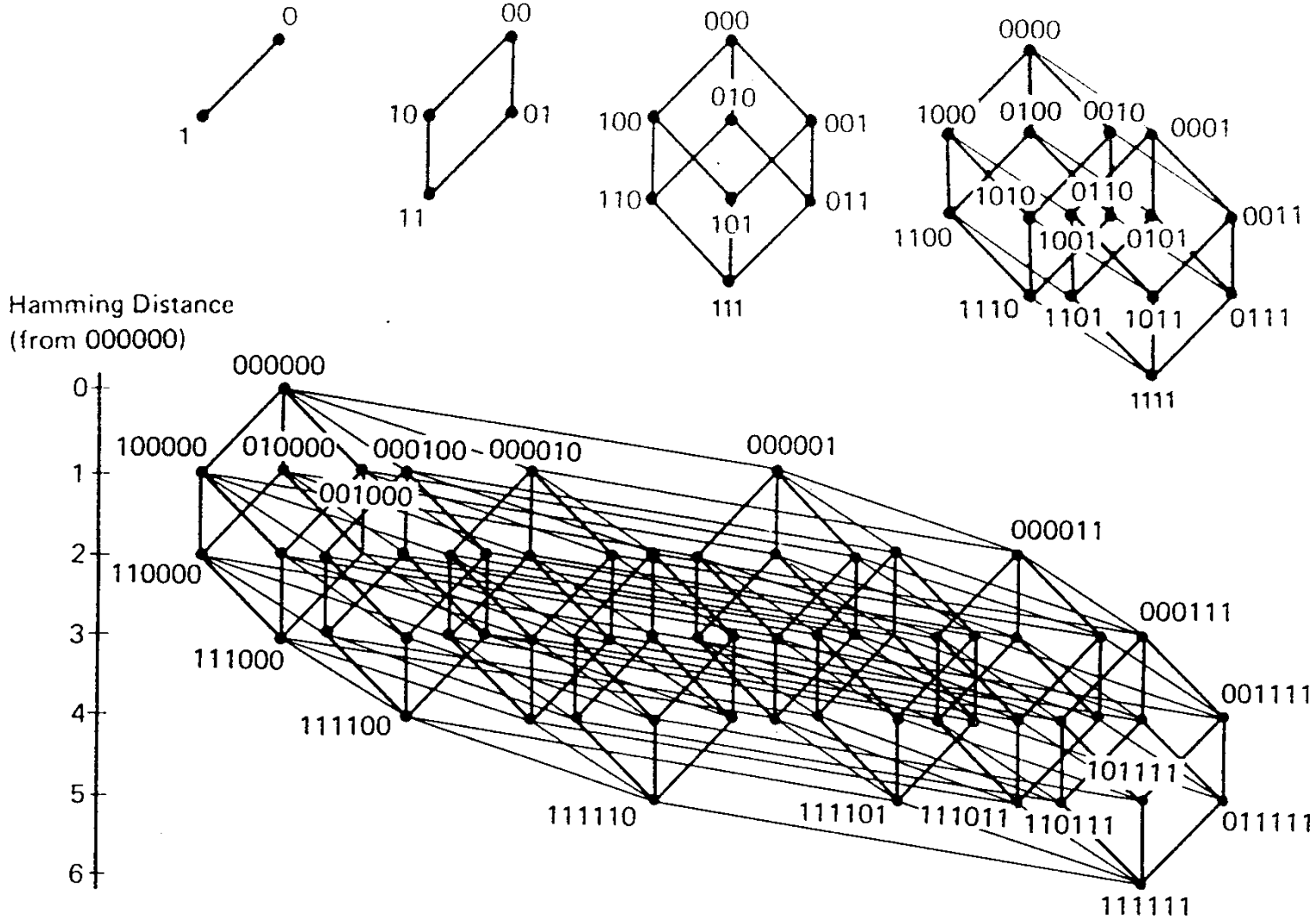
CIRM, Luminy, June 18, 2015

Fitness landscapes

- Genotypes are binary sequences $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_L)$ with $\sigma_i \in \{0, 1\}$ (presence/absence of mutation).
- Together with the Hamming distance $d(\sigma, \sigma') = \sum_{i=1}^{L} 1 - \delta_{\sigma_i, \sigma_i'}$ this defines the Hamming space $\mathbb{H}_2^L$ which is the L-dimensional hypercube
- A fitness landscape is a real-valued function $f(\sigma)$ on $\mathbb{H}_2^L$
- Interactions between the fitness effects of different mutations may induce multiple adaptive peaks:

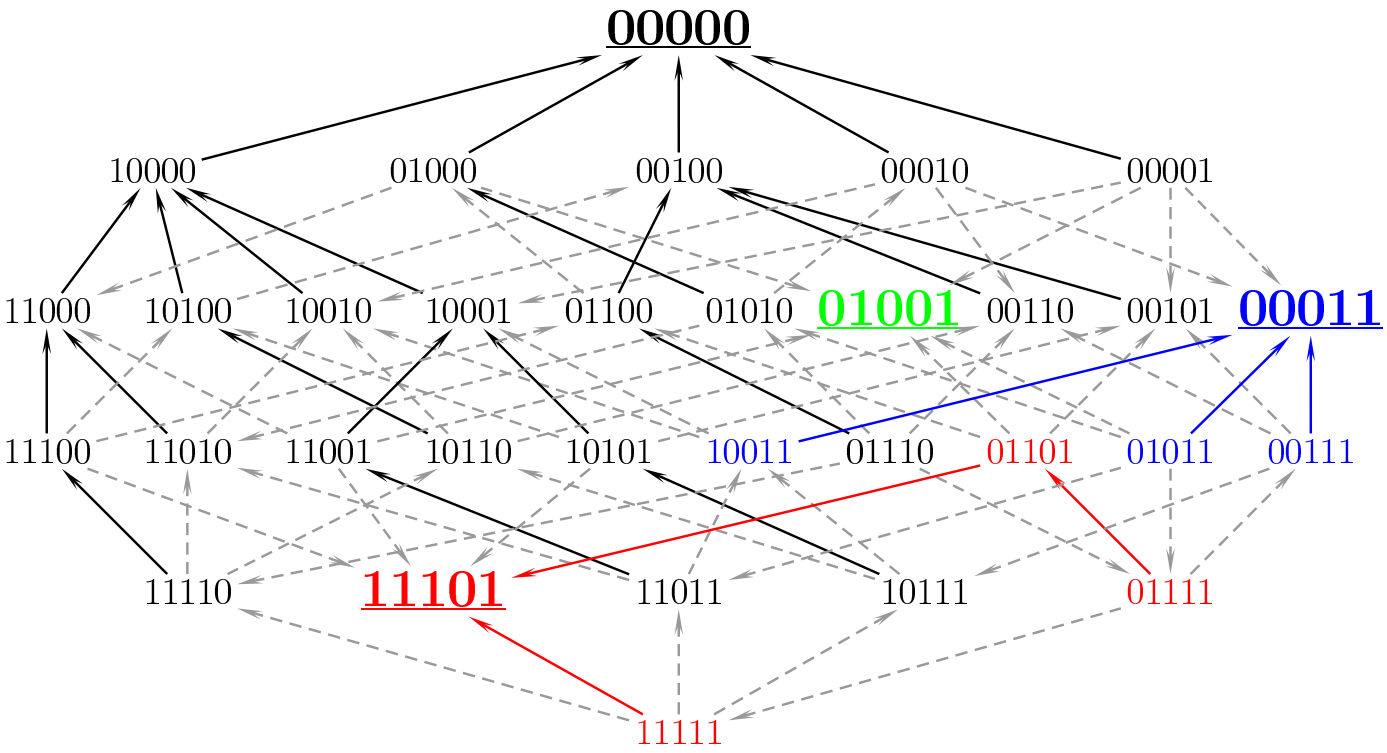


Hypercubes


0
1
00
10
01
11
000
010
100
001
110
011
101
111
0000
1000
0100
0010
0001
1010
0110
0011
1100
1001
0101
1110
1101
1011
0111
1111
Hamming Distance
(from 000000)
000000
0
100000
010000
000100
000010
000001
1
001000
000011
2
110000
000111
3
111000
001111
4
111100
101111
5
111110
111101
111011
110111
011111
6
111111

Example: The *Aspergillus niger* fitness landscape

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



- Combinations of 5 individually deleterious marker mutations
- Arrows point towards higher fitness
- For a survey of other examples see J.A.G.M. de Visser, JK, Nat. Rev. Gen. 2014

Evolutionary accessibility

"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..."

S. Wright, 1932

- Accessibility of fitness landscapes can be quantified by the number of local fitness peaks or the number of fitness-monotonic pathways

  $\Rightarrow$ see talk by Éric Brunet

- However, even if uphill pathways exist it is not clear if populations can find them

- Here we take a dynamic viewpoint and consider populations navigating a rugged fitness landscape through adaptive walks with local rules

- Such walks also serve as a tool for exploring large-scale fitness data sets

  e.g. Kouyos et al., PLoS Genetics 2012

SSWM dynamics

- SSWM = Strong Selection/Weak Mutation Gillespie 1983, Orr 2002
- Weak mutation: Each new mutation goes to fixation or is lost before the next one arrives
- Strong selection: The fixation probability of a mutation of selective advantage s is

$$\pi(s,N) \approx \frac{1-\exp[-2s]}{1-\exp[-2Ns]} \approx 1-\exp[-2s] \approx 2s$$

for $s > 0$ and $\pi = 0$ for $s \leq 0$

- Under these conditions the population performs an uphill adaptive walk in sequence space that terminates at a local fitness maximum
- Formally, an adaptive walk is a Markov chain on $\mathbb{H}_2^L$ with absorption at local maxima

Adaptive walks

- Four flavors of adaptive walks differing in their transition probabilities:

  True Adaptive Walk (TAW)
  Transition rate is proportional to the fitness difference between the resident and mutant genotype Gillespie 1983, Orr 2002

  Random Adaptive Walk (RAW) Macken & Perelson 1989
  All fitter genotypes are chosen with equal probability

  Greedy Adaptive Walk (GAW) Orr 2003
  The most fit genotype is chosen deterministically

  Reluctant Adaptive Walk (RELAW)
  The least fit among the fitter genotypes is chosen deterministically
  Bussolari et al. 2003

- Of interest is the length ℓ (= mean number of steps) and height f^* (= mean achieved fitness) of such walks

Walk length in uncorrelated landscapes

In the uncorrelated House-of-Cards/Mutational Landscape model fitness values are i.i.d. random variables. The following results refer to walks starting at a low fitness genotype:

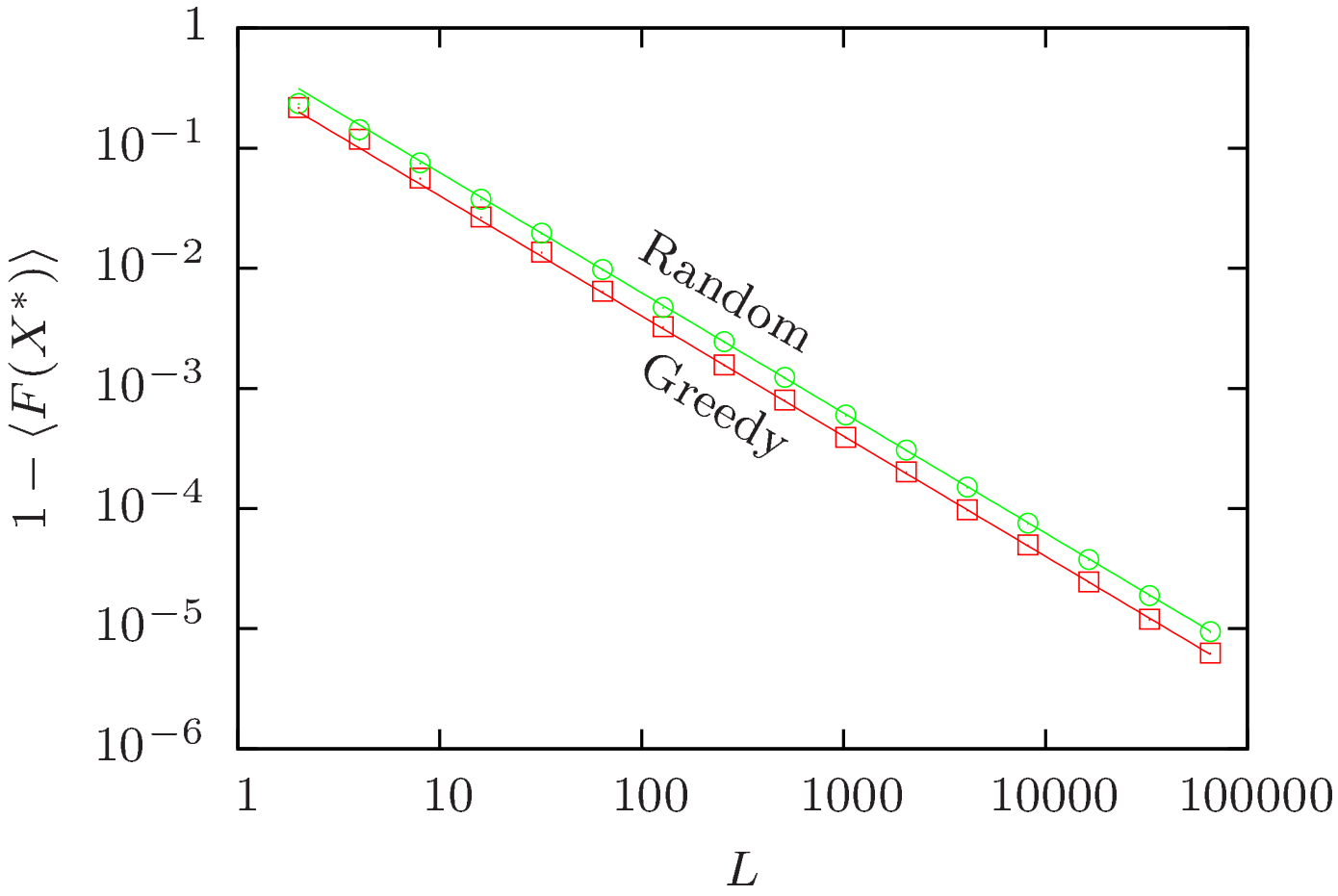
- RAW: $\ell \approx \ln(L) + 1.1$ for large L Flyvbjerg & Lautrup 1992
- GAW: $\ell \to \sum_{k=1}^{\infty} \frac{1}{k!} = e - 1 \approx 1.71828...$ Orr 2003
- RELAW: $\ell \to L + \mathcal{O}(1)$ É. Brunet, JK, 2015
- TAW length asymptotics depends on the extreme value index κ of the fitness distribution according to J. Neidhart & JK 2011, Jain 2011

$$\ell \approx \frac{1-\kappa}{2-\kappa} \ln(L) + c_\kappa \quad \text{for} \quad \kappa < 1.$$

- For relative initial fitness $f_0 \in [0,1]$ let $L \to (1 - f_0)L$

Walk height in uncorrelated landscapes

S. Nowak, JK, JSTAT P06014 (2015)



- For uniform fitness distribution the expected final fitness is of the form $1 - f^* \approx \frac{\beta}{L}$ with $\beta_{\mathrm{RAW}} \approx 0.6243..$, $\beta_{\mathrm{GAW}} \approx 0.4003...$ and $\beta_{\mathrm{RELAW}} = 1$

Models of correlated fitness landscapes

Kauffman's NK-model

Kauffman & Weinberger 1989

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

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

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

- $K = 0$: Non-interacting $\qquad$ $K = L-1$: House-of-Cards

Rough Mount Fuji model

Aita et al. 2000; Neidhart et al. 2014

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

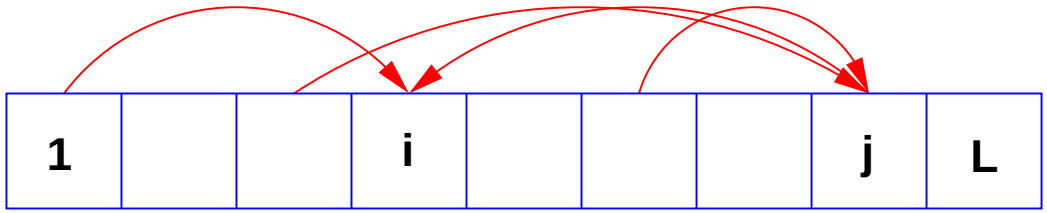
$$f(\sigma) = -cd(\sigma, \sigma^*) + \eta(\sigma) \quad c > 0$$

η: i.i.d. random variables $\qquad$ σ^*: reference sequence

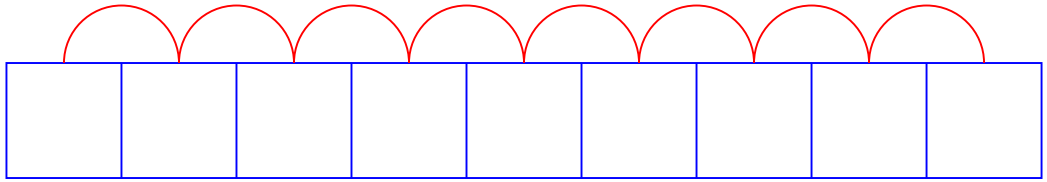
- Landscape roughness is tuned by the ratio $c/\sqrt{\mathrm{Var}(\eta)}$

“Genetic architecture” in Kauffman’s NK-model

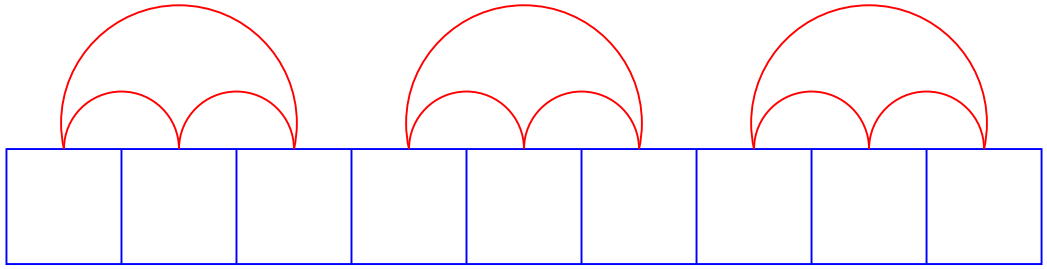
- Different schemes for choosing the interaction partners $(K = 2, L = 9)$:



random



adjacent



block/modular

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

Number of local maxima

Number of maxima in the NK-model

- For the uncorrelated HoC model $\mathscr{M} \equiv \mathbb{E}(n_{\max}) = \frac{2^L}{L+1}$ by symmetry

  Kauffman & Levine 1987

- Rigorous work on the NK-model with adjacent neighborhoods shows that, for fixed K, $\mathscr{M} \sim (2\lambda_K)^L \quad \text{for} \quad L \to \infty$ with constants $\lambda_K \in (\frac{1}{2}, 1)$

  Evans & Steinsaltz 2002, Durrett & Limic 2003

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

  Perelson & Macken 1995

- 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 $L \to \infty$ and $K \to \infty$ are taken simultaneously with $\gamma = L/K$ fixed, rigorous analysis shows that $\mathscr{M} \sim \frac{2^L}{L^\gamma}$, which is also true for the block model.

  Limic & Pemantle 2004

Number of maxima in the Rough Mount Fuji model

J. Neidhart, I.G. Szendro, JK, Genetics **198**, 699 (2014)

- 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

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

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

- For distributions with tail heavier than exponential (power law or stretched exponential) $\mathscr{M} \to \frac{2^L}{L}$ for $L \to \infty$, which implies that the fitness gradient (c) is asymptotically irrelevant

- For distributions with an exponential tail $\mathscr{M} \to \frac{2^L}{\cosh(c)L}$ 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

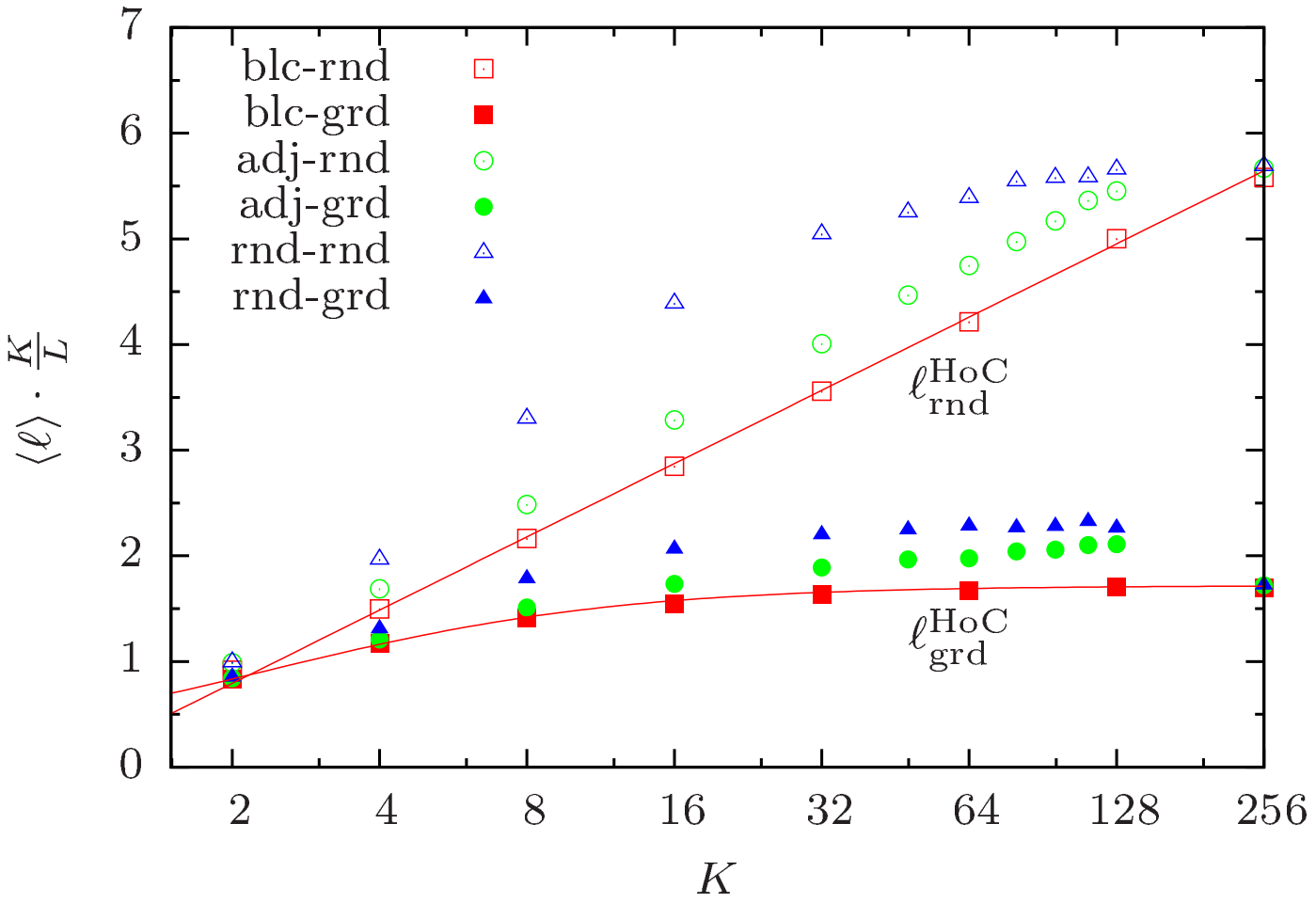
$$\mathscr{M} \sim \frac{(2-c^\nu)^L}{L^\nu}$$

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

Adaptive walks on correlated landscapes

Walk length in the NK landscape

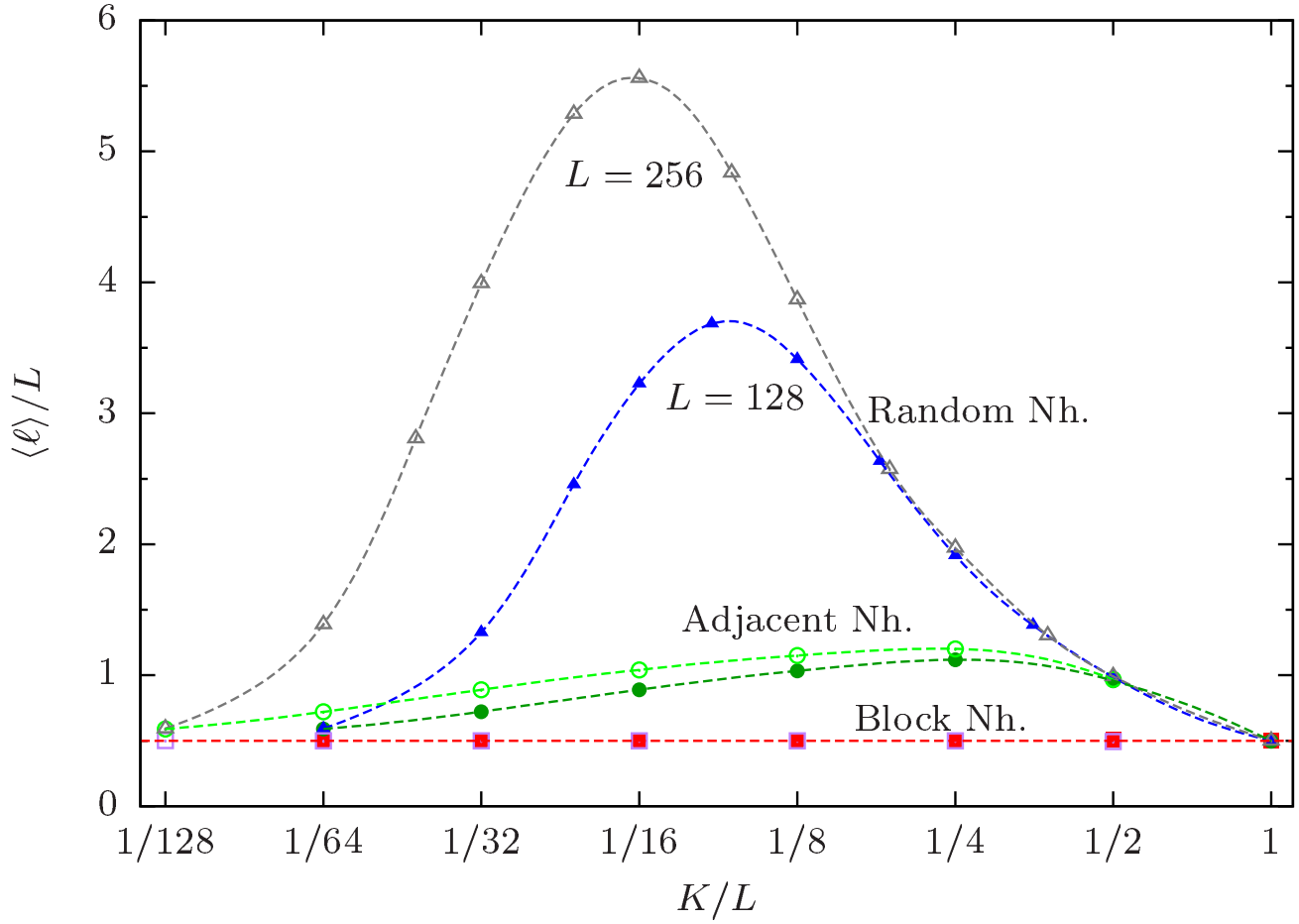
S. Nowak, JK, JSTAT P06014 (2015)



- For the block model $\ell = \frac{L}{K+1}\,\ell_{\mathrm{HoC}}(K+1)$ exactly

Reluctant walks in the NK landscape

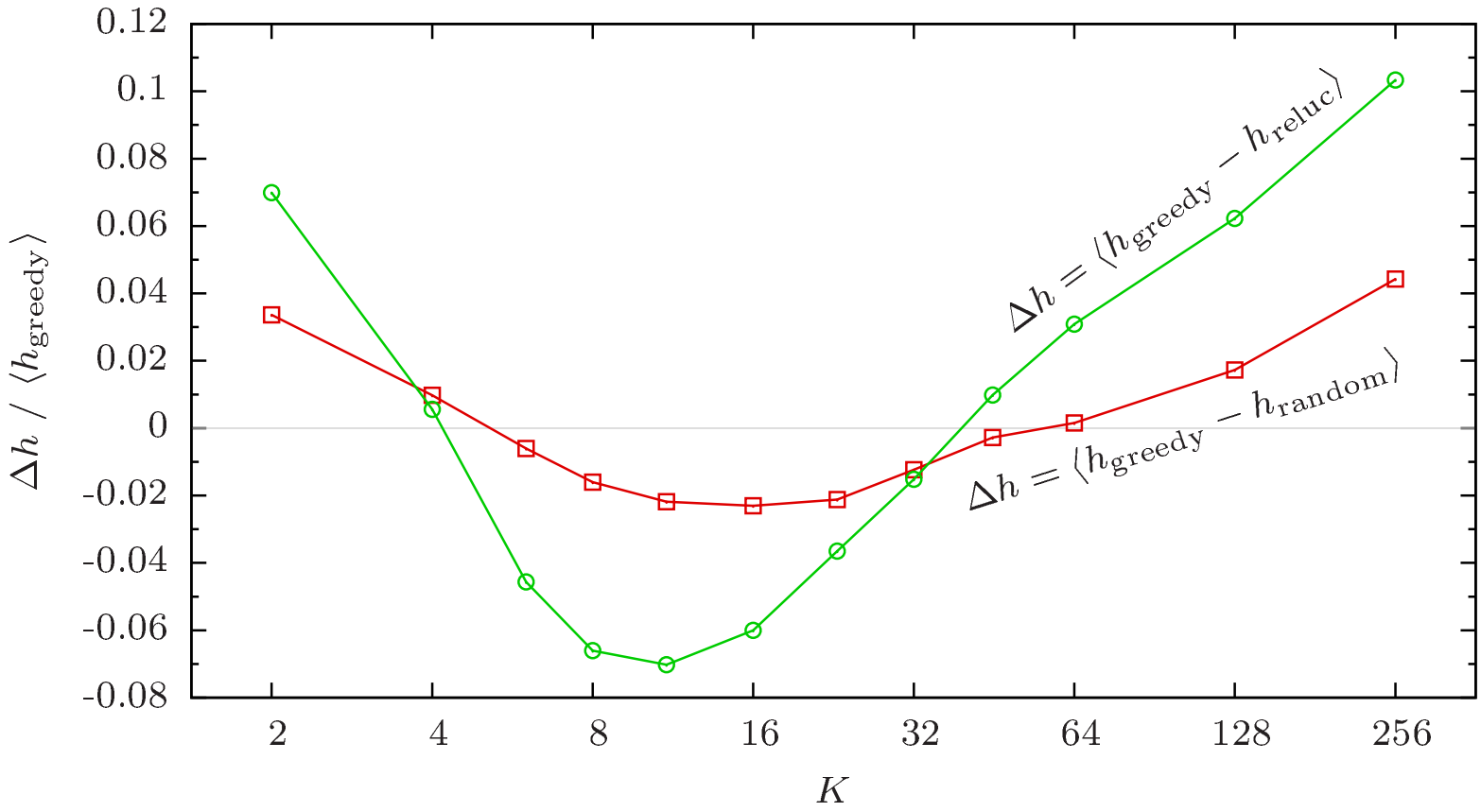
S. Nowak, JK, JSTAT P06014 (2015)



- Reluctant walks are longest for intermediate K....

Reluctant walks in the NK landscape

S. Nowak, JK, JSTAT P06014 (2015)



- ...and may achieve higher fitness than GAW's or RAW's.

Random adaptive walks in the RMF landscape

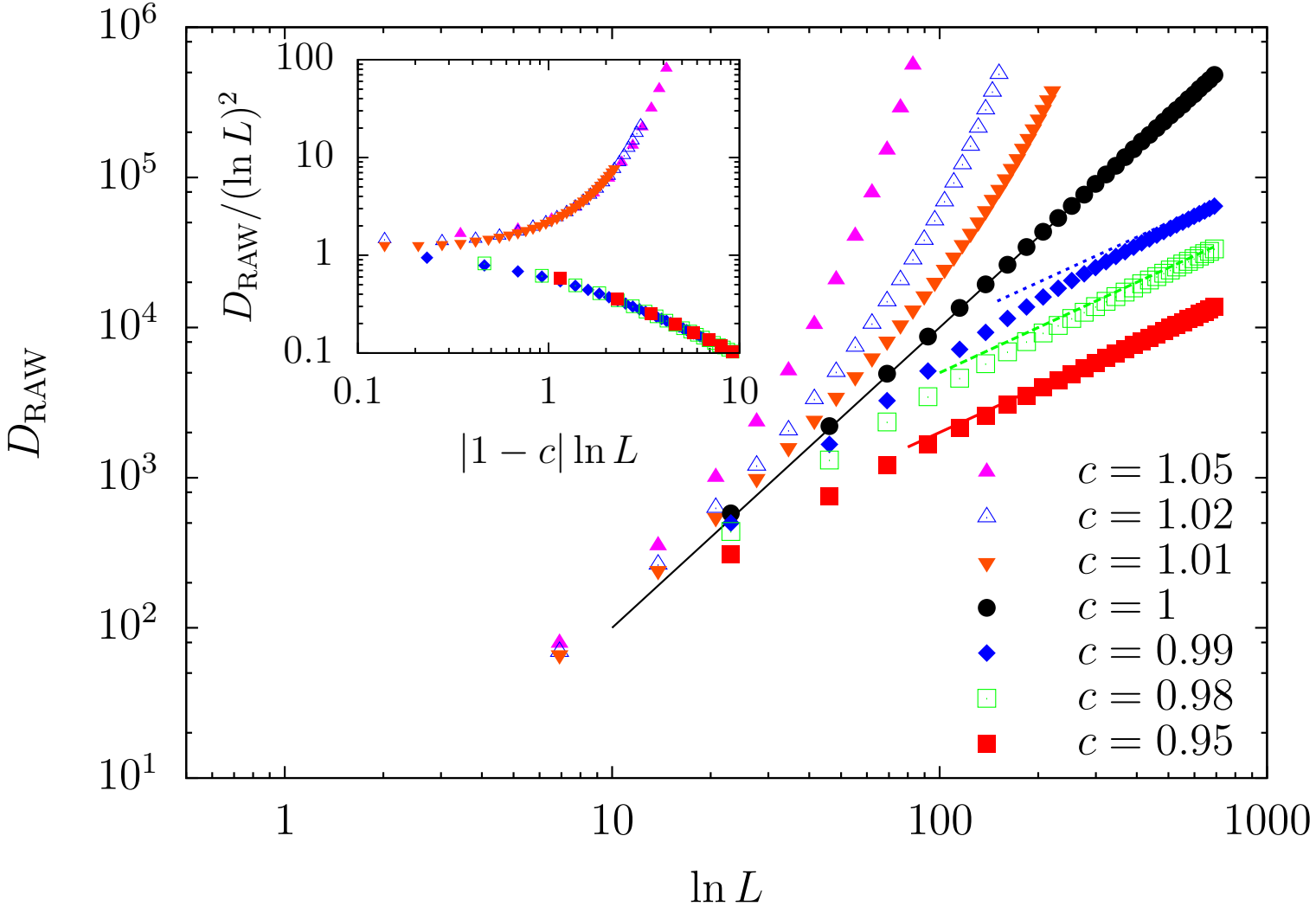
S.-C. Park, I.G. Szendro, J. Neidhart, JK, Physical Review E **91**, 042707 (2015)

- RAW's starting at antipode (maximal distance L from reference sequence)
- Assume RAW takes only 'uphill' steps that decrease $d(\sigma, \sigma^*)$, and draw random fitness component from exponential distribution with mean 1
- Then the mean walk length can be computed analytically and displays a phase transition at $c = 1$:

$$\ell \propto \begin{cases} \ln L/(1-c), & c < 1 \\ (\ln L)^2, & c = 1, \\ O(L), & c > 1. \end{cases}$$

- For tails thinner (fatter) than exponential, $\ell \sim L$ ($\ell \sim \ln L$) for all $c > 0$
- The same scenario applies to the TAW

Random adaptive walks in the RMF landscape



- Numerical verification by simulations

Sketch of derivation

- Let $Q_l(y_l, L)$ denote the probability to take at least l steps and arrive at fitness $-c(L-l)+y_l$. This satisfies the recursion relation

$$Q_{l+1}(y,L) = p(y)\int_{-\infty}^{y+c} dx\, Q_l(y,L)\frac{1-P(x-c)^{L-l}}{1-P(x-c)}$$

  which is explicitly solvable only for $c=0$ Flyvbjerg & Lautrup 1992

- Therefore consider $Q_l(y) \equiv \lim_{L\to\infty} Q_l(y,L)$ and estimate the stopping condition for the walk from $P(z_l-c)^{L-l} \approx 1/e$ where z_l is the mean of $Q_l(y)$

- An explicit solution for $Q_l(y)$ can be found for $p(x)=e^{-x}$, which reads

$$Q_l(y) = -\frac{d}{dy}\left[\sum_{n=0}^{l} y\frac{(y+cn)^{n-1}}{n!}e^{-y-cn}\right]$$

- Other distributions can be treated approximately through a self-consistent equation for z_l

Greedy adaptive walks in the RMF landscape

S.-C. Park, J. Neidhart, JK, in preparation

- For Gumbel-distributed random fitness components, the length of GAW's starting from the antipode of the reference sequence satisfies

$$H_l \equiv \mathbb{P}(\text{length} \geq l) = \prod_{k=1}^{l} \frac{1-e^{-c}}{1-e^{-kc}} = \frac{1}{[l]_{e^{-c}}!}$$

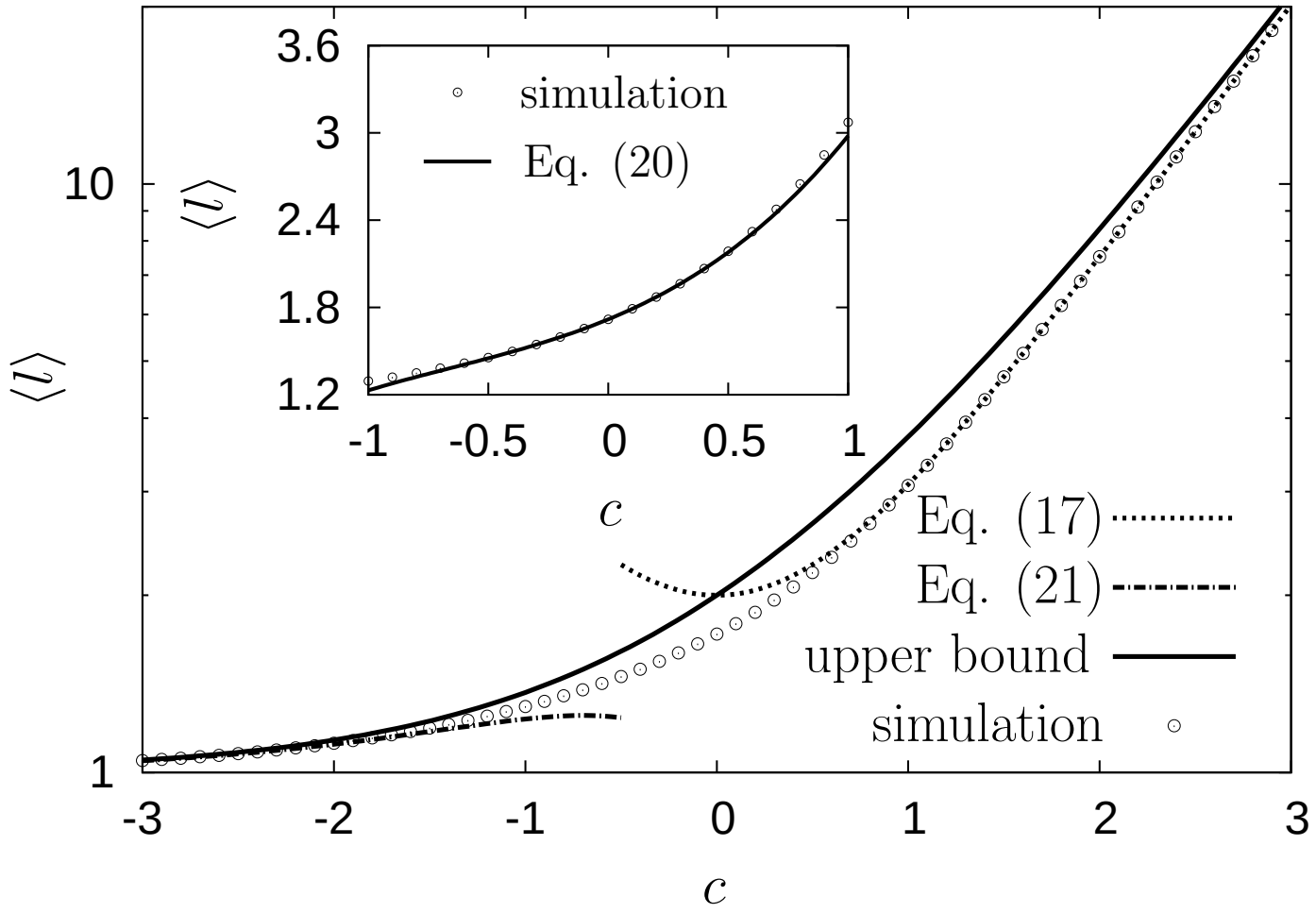
where $[n]_q = \frac{1-q^n}{1-q}$ is the q-number and $L \to \infty$ is implied.

- Correspondingly the mean walk length is given by the q-exponential

$$\ell = \exp_{e^{-c}}(1) - 1 \to e - 1 \quad \text{for} \quad c \to 0$$

- For distributions with non-exponential tails the walk length is either $\ell = \ell(c = 0) = e - 1$ or $\ell = L$ asymptotically.

Mean GAW length (Gumbel case)



- upper bound: $\ell = 1 + e^c$

GAW's with general starting point

- If the walk starts at distance $d = \alpha L$ from the reference sequence with $\alpha < 1$, both uphill and downhill steps have to be taken into account

- Asymptotically for large L the distribution of walk lengths is then given by

$$H_l = \sum_{\{\tau_i = \pm 1\}} \prod_{k=1}^{l} \frac{s_{\tau_k}}{1 + \sum_{m=1}^{k-1} \exp\left[-c \sum_{j=1}^{m} \tau_j\right]}$$

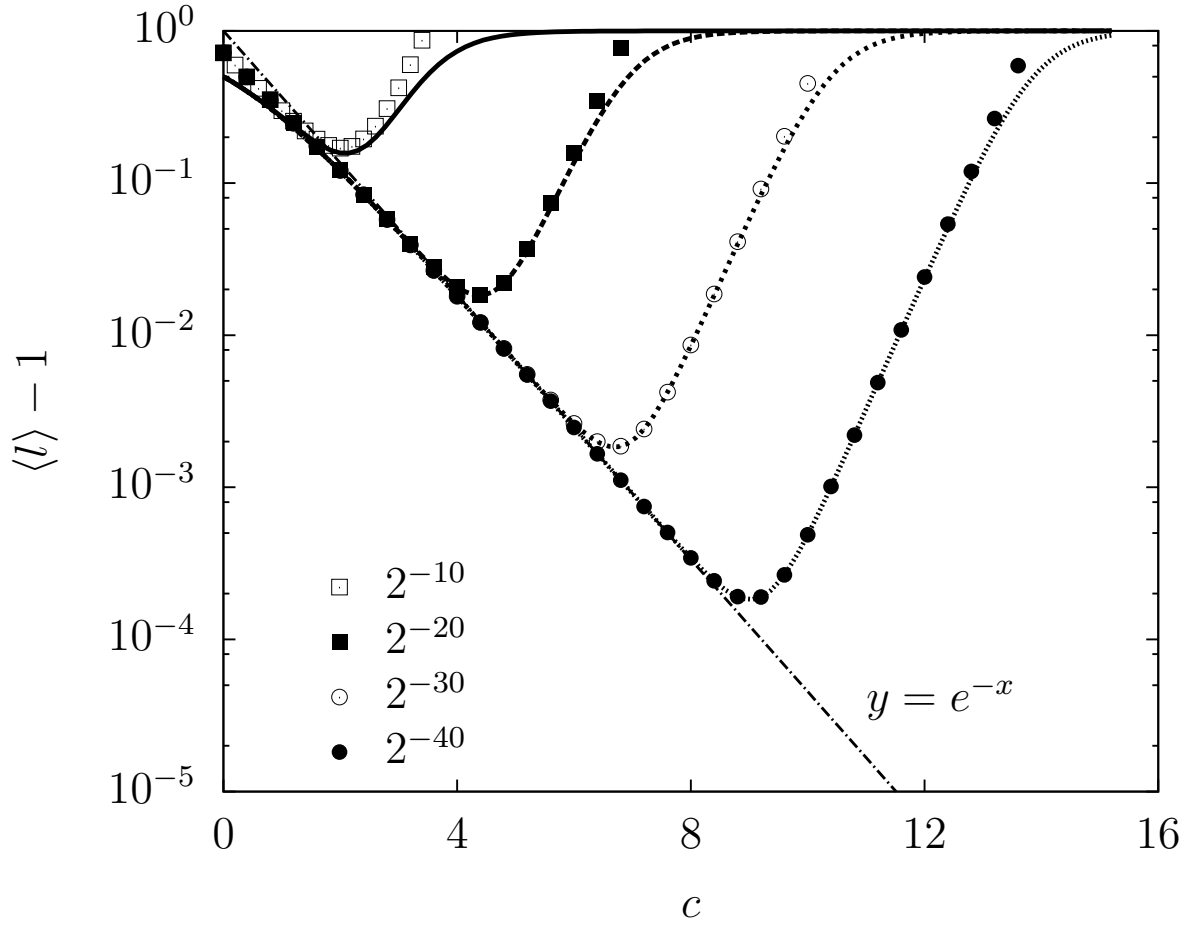
  where $\tau_1, \tau_2, \ldots, \tau_l = \pm 1$ encodes the sequence of uphill and downhill steps and

$$s_1 = \frac{\alpha e^c}{\alpha e^c + (1-\alpha) e^{-c}}, \quad s_{-1} = 1 - s_1$$

- For $\alpha < \frac{1}{2}$ the walk length is non-monotonic in c, i.e. greedy walks on the correlated landscapes can be shorter than on the uncorrelated landscape

- This is related to a similar non-monotonicity in the density of maxima near the reference sequence

Minimum in walk length



- $\alpha = 2^{-10}....2^{-40}$, location of minimum varies as $c_{\min} \sim -\frac{1}{3}\ln(2\alpha)$

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

- Different walk types serve as caricatures of adaptive regimes and have strikingly different properties

- Complementary view of evolutionary accessibility is provided by the analysis of fitness-monotonic pathways (Éric Brunet)