UNIVERSITY
OF COLOGNE

COLLABORATIVE RESEARCH CENTER 1310

Predictability in Evolution

Complexity and accessibility of random landscapes

Joachim Krug
Institute for Biological Physics
University of Cologne

Les Houches School on Large Deviations, July 2024

Outline

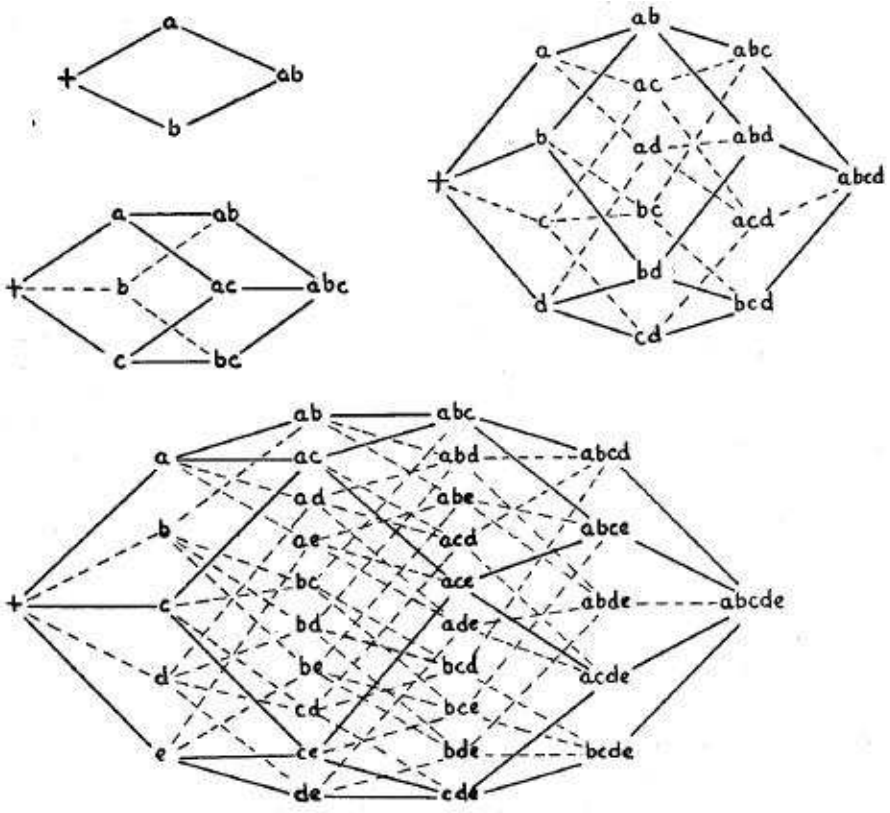
1. Fitness landscapes
2. Counting peaks: House-of-Cards/Random Energy model
3. Counting accessible paths
4. Structured fitness landscapes

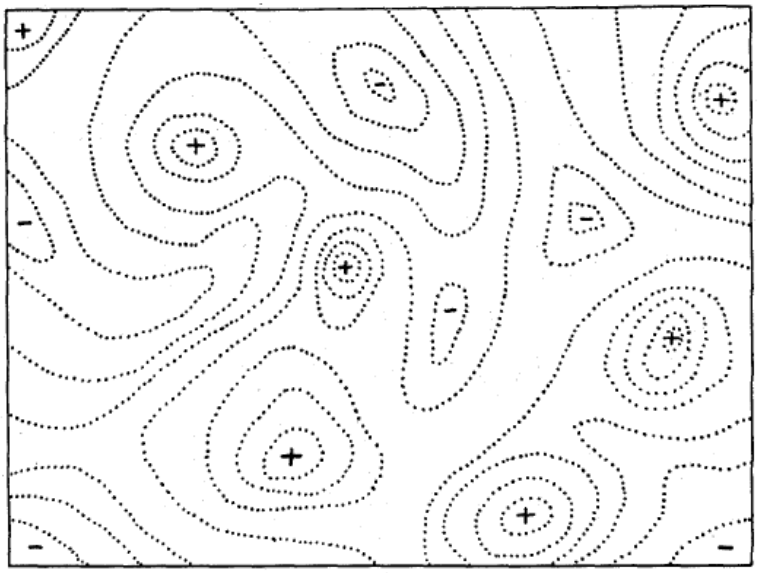
Literature

- JK, D. Oros, JSTAT 034003 (2024)
- J.A.G.M. de Visser, JK, Nat. Rev. Genet. 15:480 (2014)
- C. Bank, Ann. Rev. Ecology, Evolution & Systematics 53:457 (2022)

Fitness landscapes

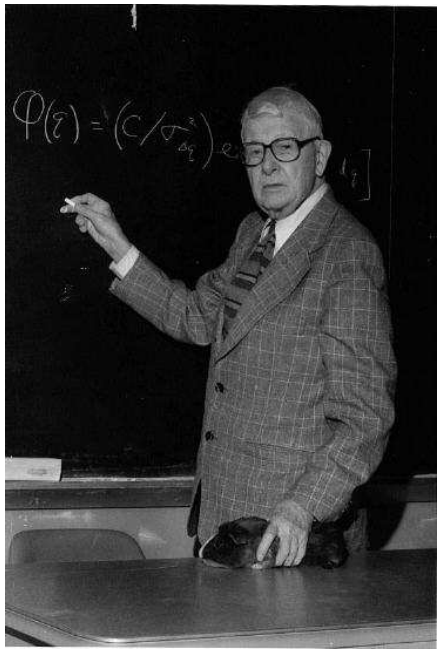
S. Wright, Proc. 6th Int. Congress of Genetics (1932)





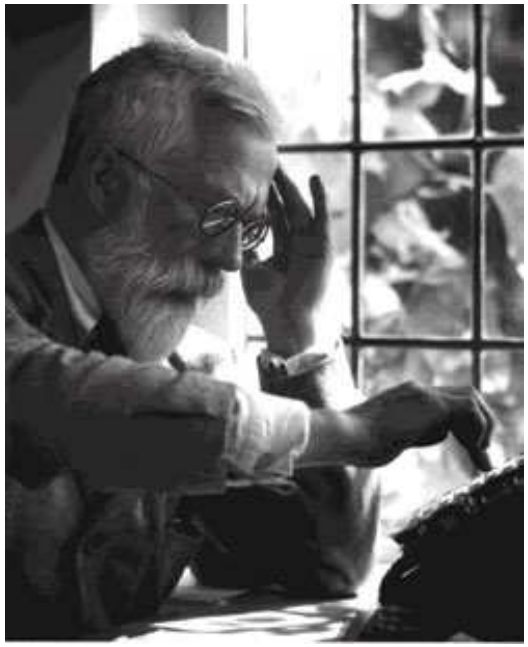
"The two dimensions of figure 2 are a very inadequate representation of such a field."

Ruggedness and accessibility



S. 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

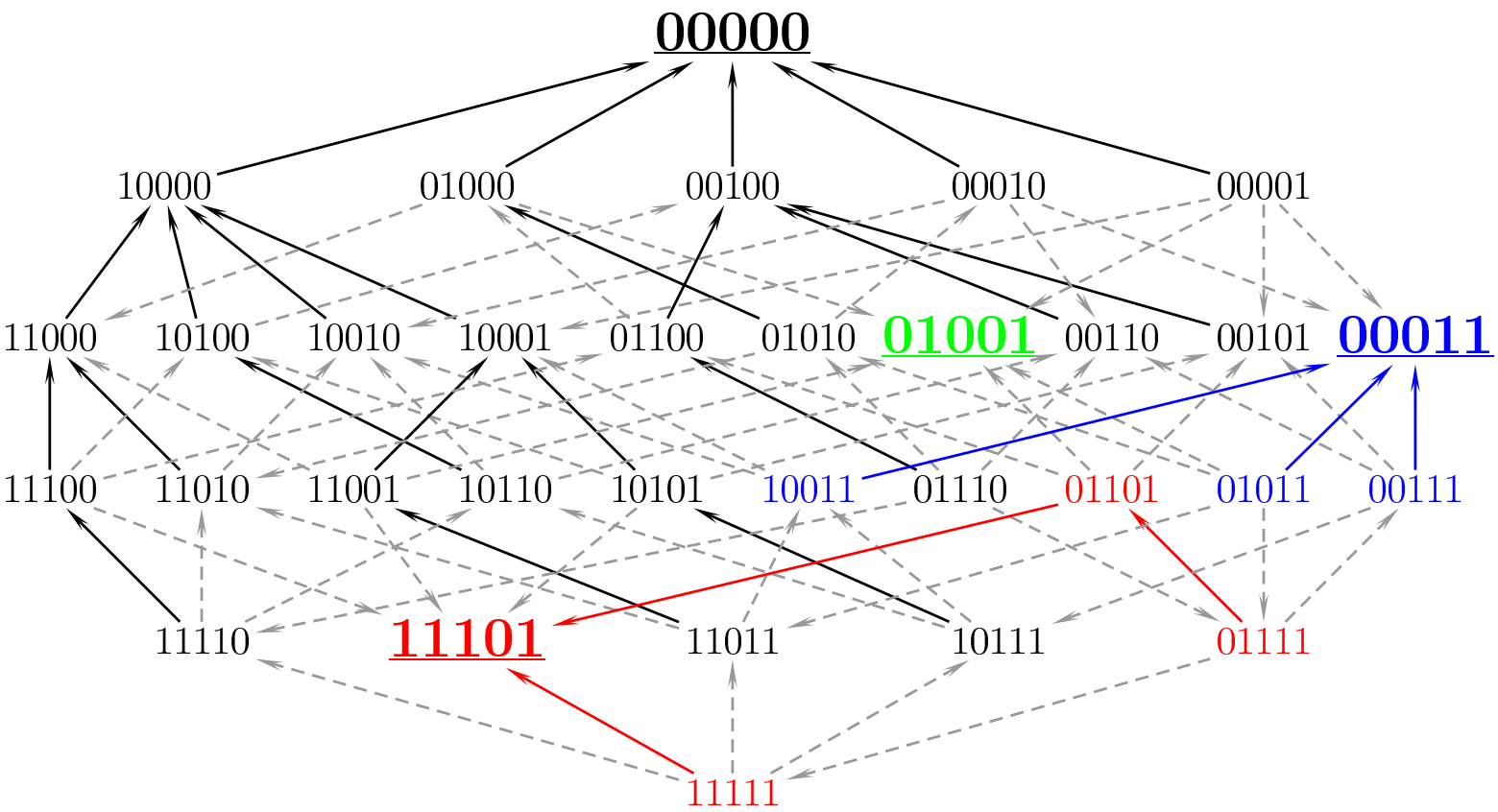
In one dimension, a curve gives a series of alternate maxima and minima, but in two dimensions two inequalities must be satisfied for a true maximum, and I suppose that only about one fourth of the stationary points will satisfy both. Roughly I would guess that with n factors only 2^{-n} of the stationary 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

Empirical fitness landscapes

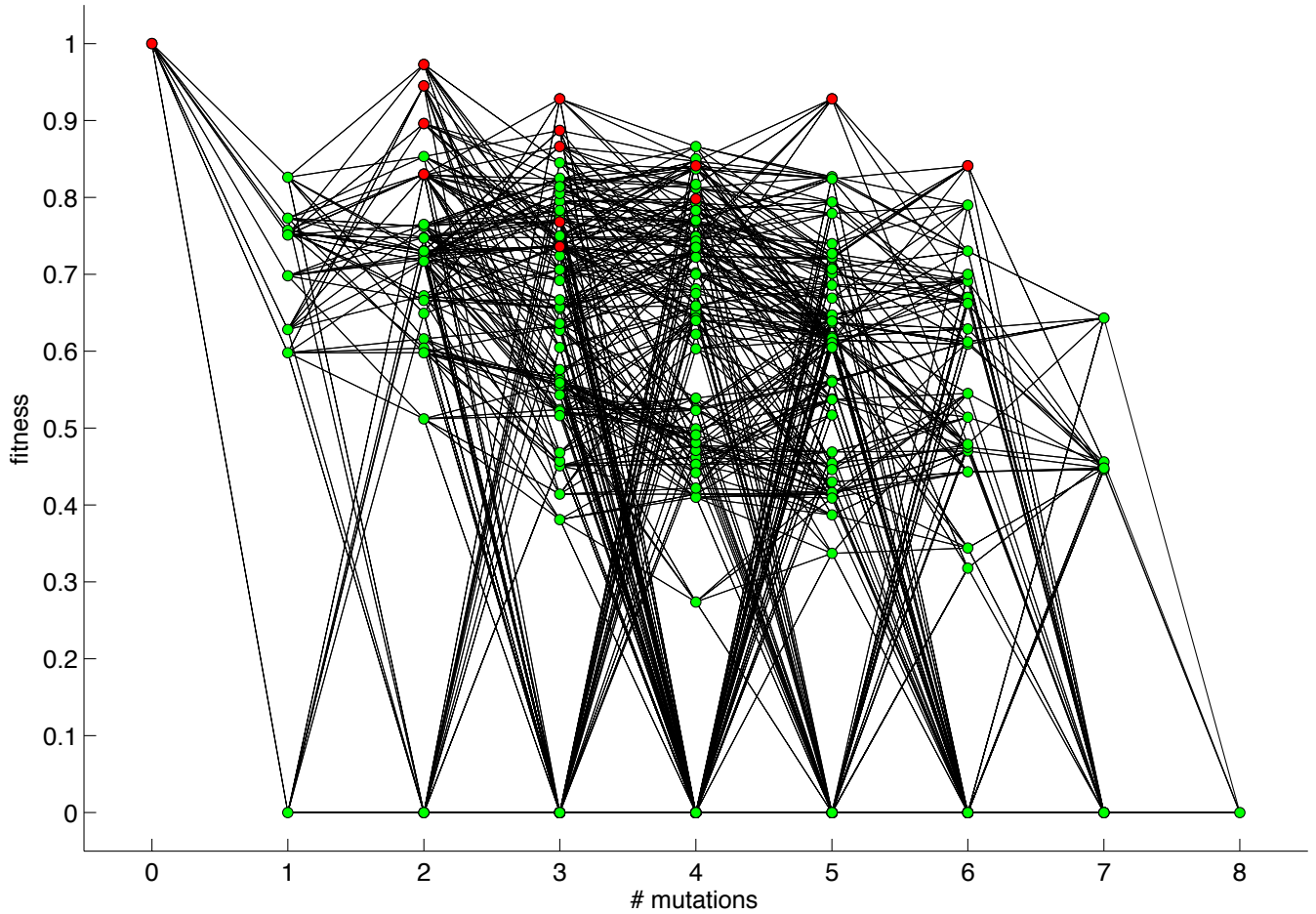
The *Aspergillus niger* fitness landscape

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



- All combinations of 5 mutations residing on different chromosomes
- Fitness graph representation, 3 peaks marked in color

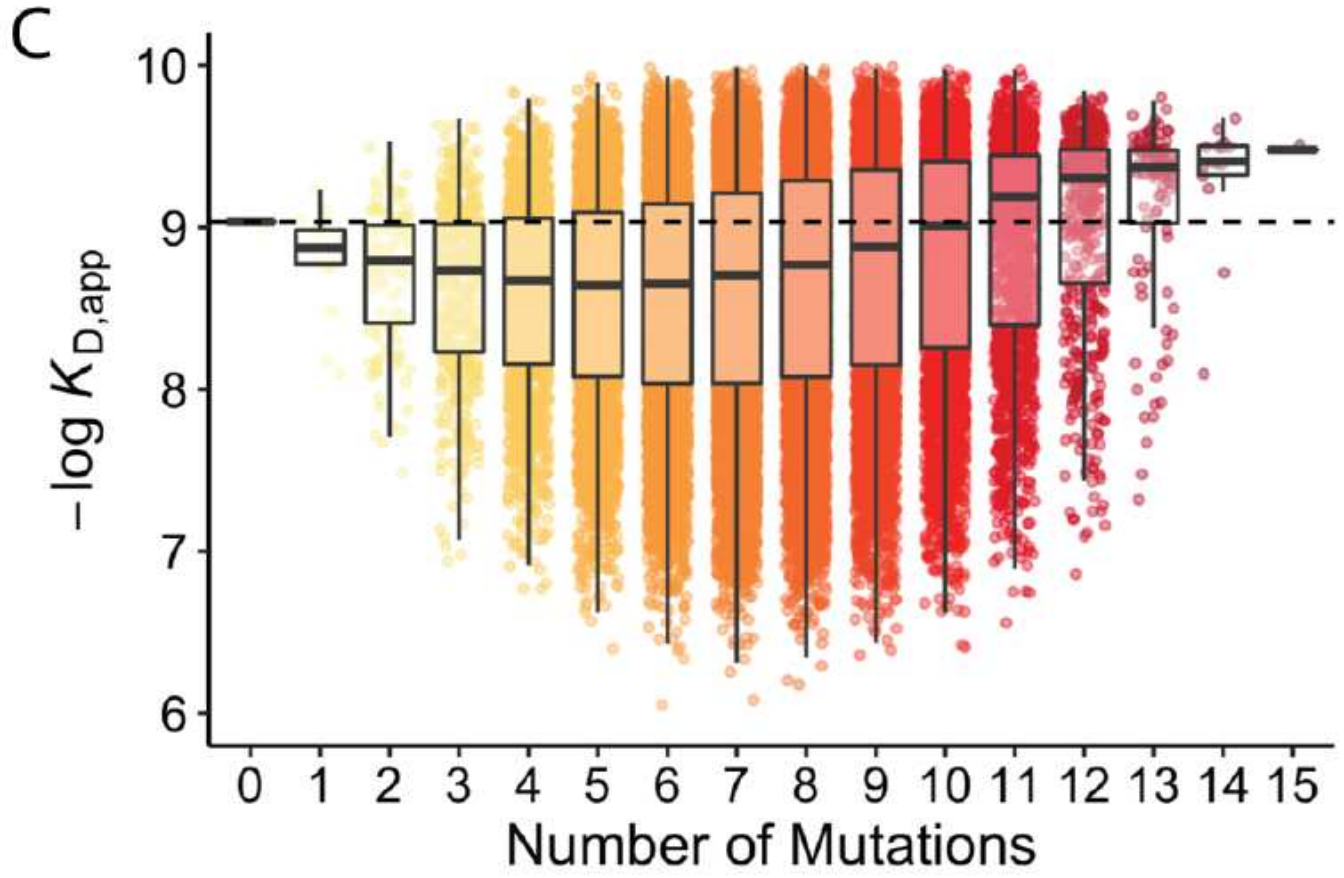
The *Aspergillus niger* fitness landscape



- All combinations of 8 mutations residing on different chromosomes
- Peaks marked in red; 86 out of 256 combinations are lethal

Affinity landscape of the SARS-CoV2 spike protein

Moulana et al., Nat. Comm. 2022

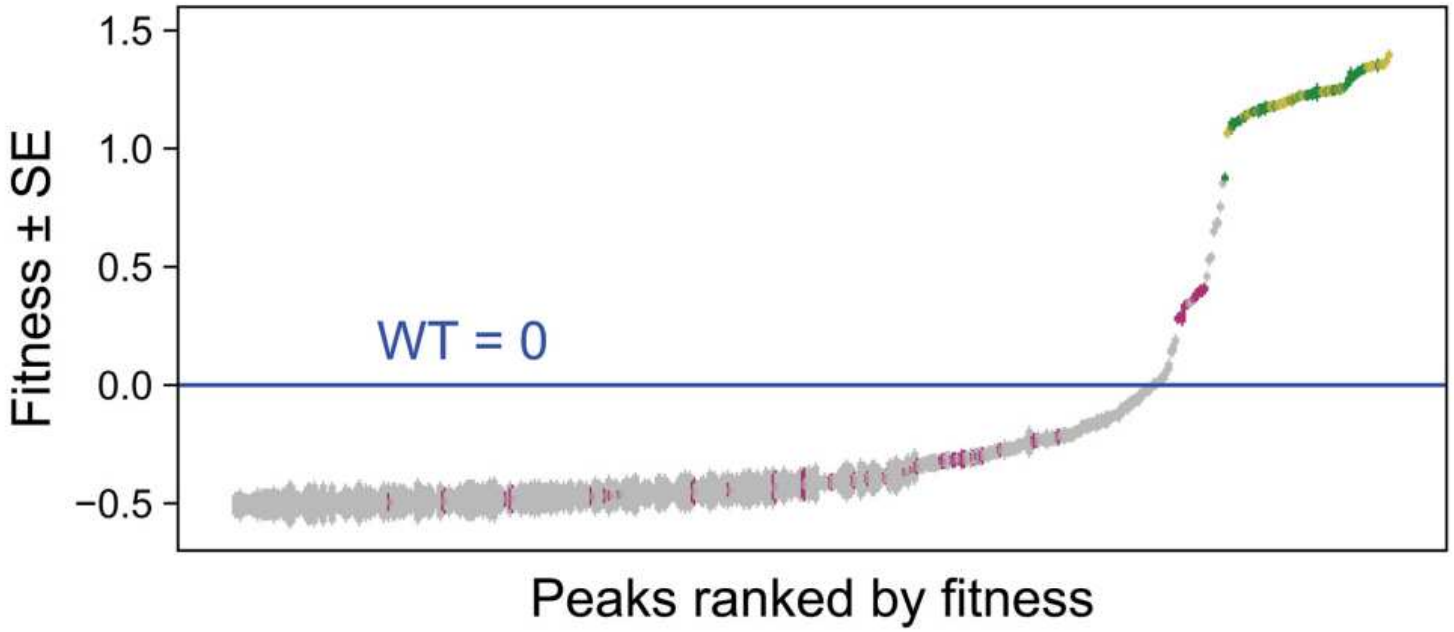


- All $2^{15} = 32768$ combinations of $L = 15$ mutations separating the ancestral Wuhan strain from Omicron BA.1

“A rugged yet easily navigable fitness landscape”

Papkou. . . Wagner, Science 2023

- $4^9 = 262,144$ combinations of nucleotides at 9 positions of the *fol A* gene in *E. coli* coding for dihydrofolate reductase (DHFR)
- Fitness measurements in trimethoprime yield 18,018 functional sequences

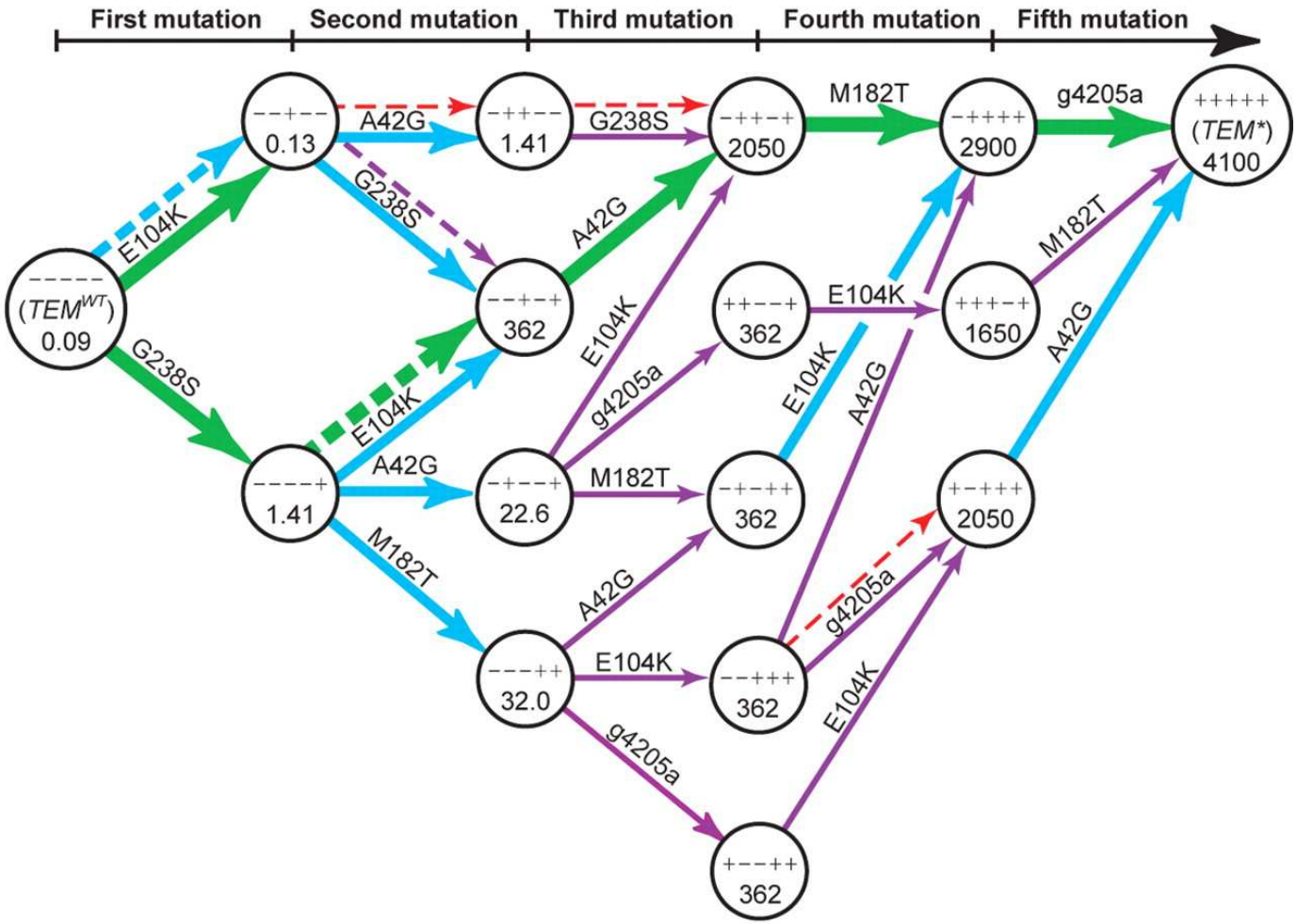


- 514 fitness peaks, 73 have high fitness

Accessibility of HoC landscapes

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

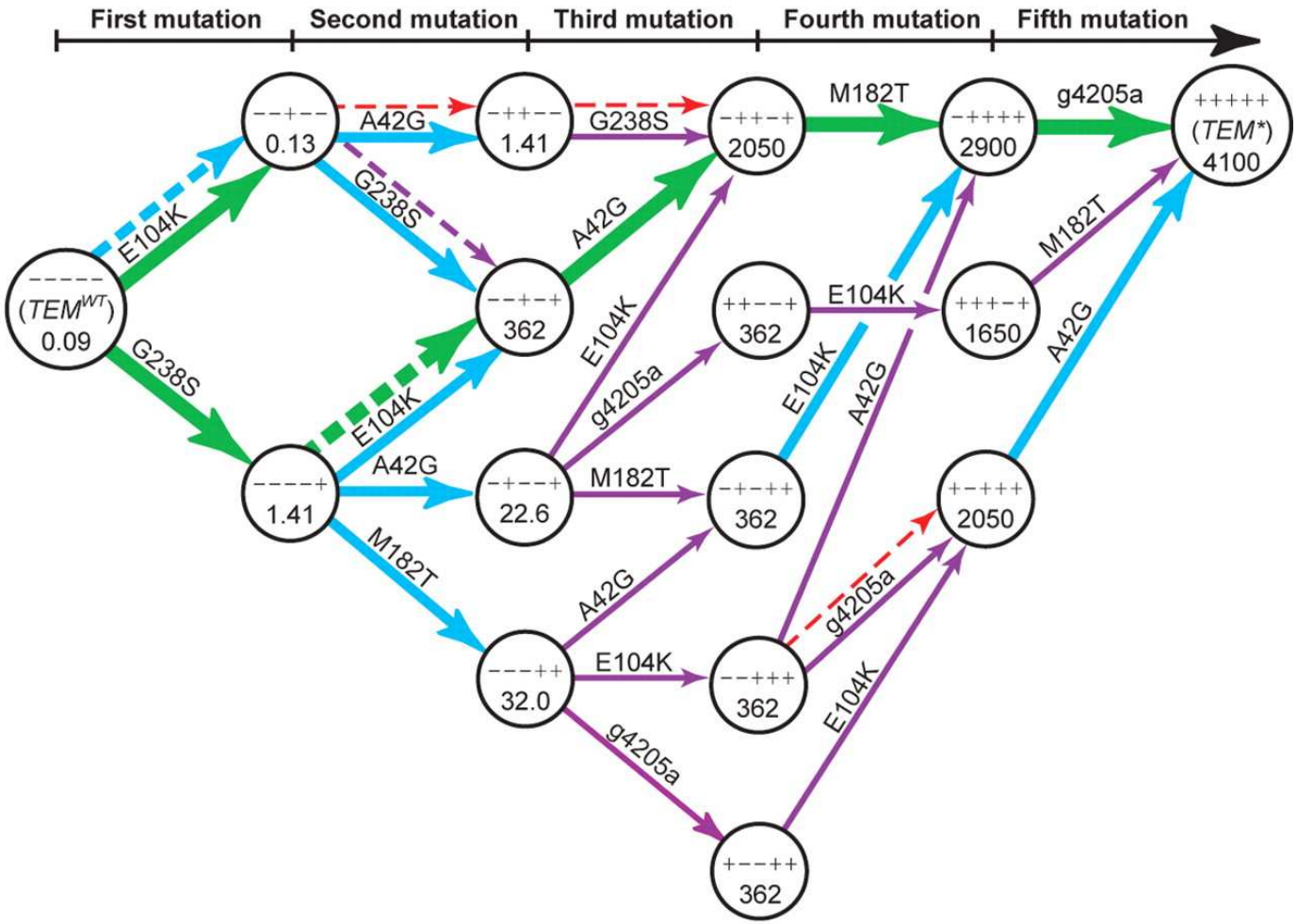
D.M. Weinreich et al., Science **312**, 111 (2006)



- 5 mutations in an enzyme increase antibiotic resistance by $\sim 4.5 \times 10^4$

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

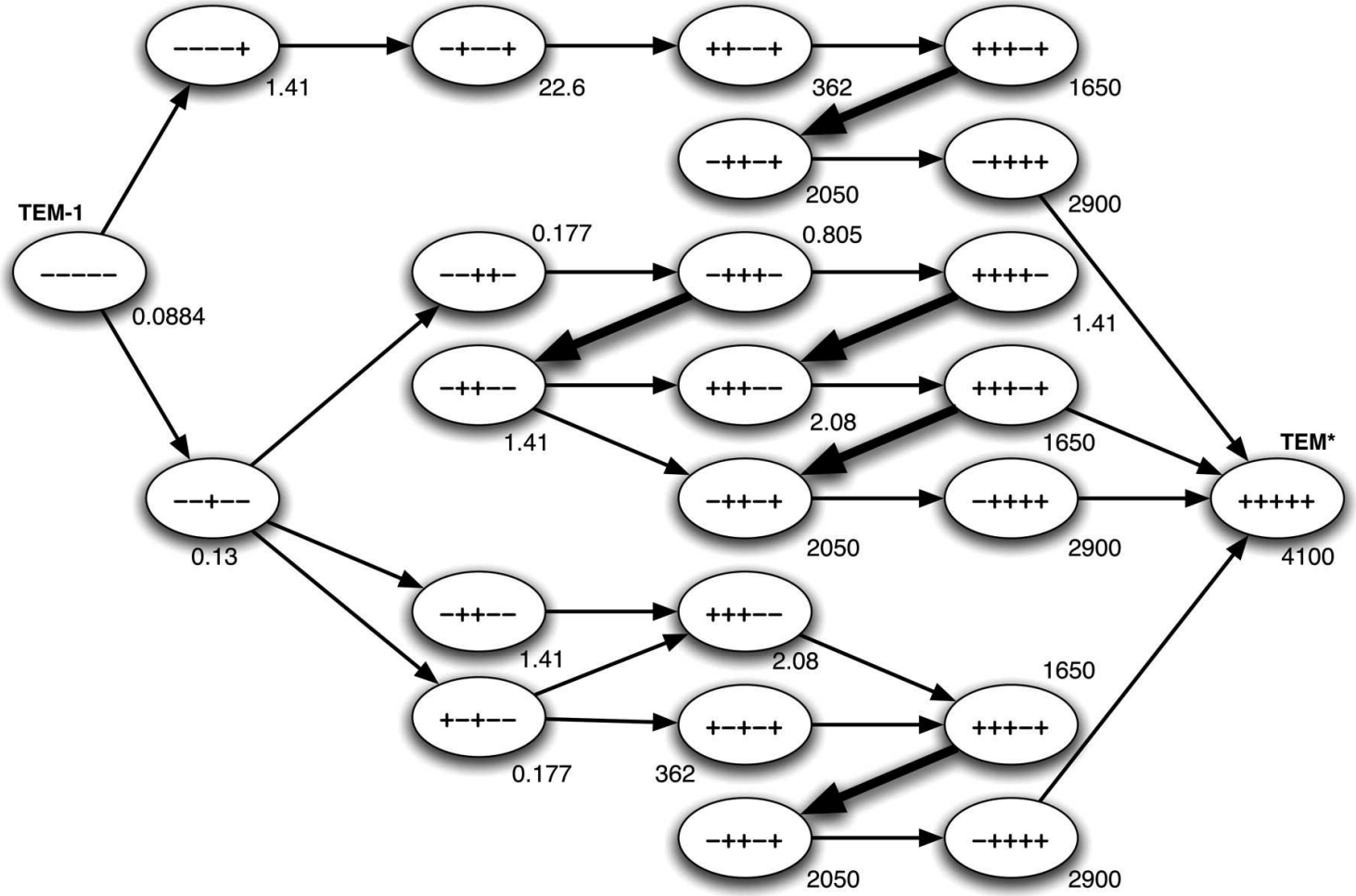
D.M. Weinreich et al., Science **312**, 111 (2006)



- 18 out of $5! = 120$ direct mutational pathways are accessible...

Including indirect paths

De Pristo et al. 2007



- ...and 27 out of 18651552840 indirect pathways

Accessibility percolation

S. Nowak, JK, EPL 2013

- Take fitness values to be i.i.d. $U[0,1]$ random variables
- A path of length ℓ between genotypes α, ω with $g(\omega) - g(\alpha) = \beta \in [0,1]$ is accessible if all $\ell - 1$ intermediate fitness values are in $(g(\alpha), g(\omega))$ and increasingly ordered, which occurs with probability

$$P_{\beta,\ell} = \frac{\beta^{\ell-1}}{(\ell-1)!}$$

- The number of accessible paths is a non-negative integer-valued random variable $X_{\alpha,\omega}$
- Is there a sharp accessibility threshold β_c in $\mathbb{P}[X_{\alpha,\omega} \geq 1]$ when $L \to \infty$ and

$$\delta \equiv \lim_{L\to\infty} \frac{d(\alpha,\omega)}{L} > 0 \; ?$$

Direct paths on the binary hypercube

P. Hegarty, A. Martinsson, Ann. Appl. Probab. 2014

- The total number of direct paths of length ℓ is $\ell!$, thus the expected number of accessible paths is

$$\mathbb{E}(X_{\alpha,\omega}) = \ell! P_{\beta,\ell} = \ell \beta^{\ell-1}$$

  which vanishes asymptotically for large ℓ when $\beta < 1$

- By Markov's inequality

$$\mathbb{P}[X_{\alpha,\omega} \geq 1] = \sum_{k=1}^{\infty} \mathbb{P}[X_{\alpha,\omega} = k] \leq \sum_{k=1}^{\infty} k\mathbb{P}[X_{\alpha,\omega} = k] = \mathbb{E}[X_{\alpha,\omega}]$$

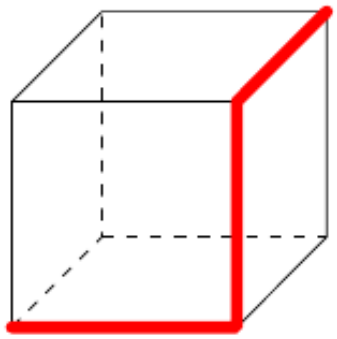
  it then follows that $\lim_{\ell\to\infty} \mathbb{P}[X_{\alpha,\omega} \geq 1] = 0$

- Analysis of the second moment $\mathbb{E}(X_{\alpha,\omega}^2)$ shows that, conversely, $\lim_{\ell\to\infty} \mathbb{P}[X_{\alpha,\omega} \geq 1] = 1$ for $\beta = \beta_\ell$ with $1 - \beta_\ell < \frac{\ln \ell}{\ell}$

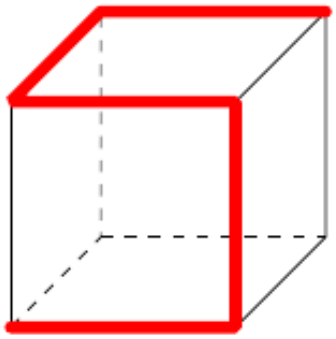
Indirect paths on the binary hypercube

Berestycki et al. 2014; Martinsson 2015; Li 2018

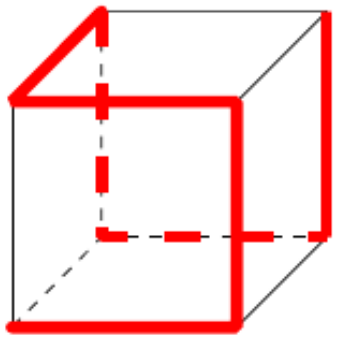
- Paths on the 3-cube with p backsteps and length $\ell = 3 + 2p$



$p = 0, \ell = 3$



$p = 2, \ell = 5$



$p = 4, \ell = 7$

- The accessibility threshold $\beta_c(\delta) < 1$ is the solution of

$$\lim_{L\to\infty} [\mathbb{E}(X_{\alpha,\omega})]^{1/L} = \sinh(\beta)^{\delta} \cosh(\beta)^{1-\delta} = 1$$

- The expectation $\mathbb{E}(X_{\alpha,\omega})$ "tells the truth"

HoC model with $a > 2$

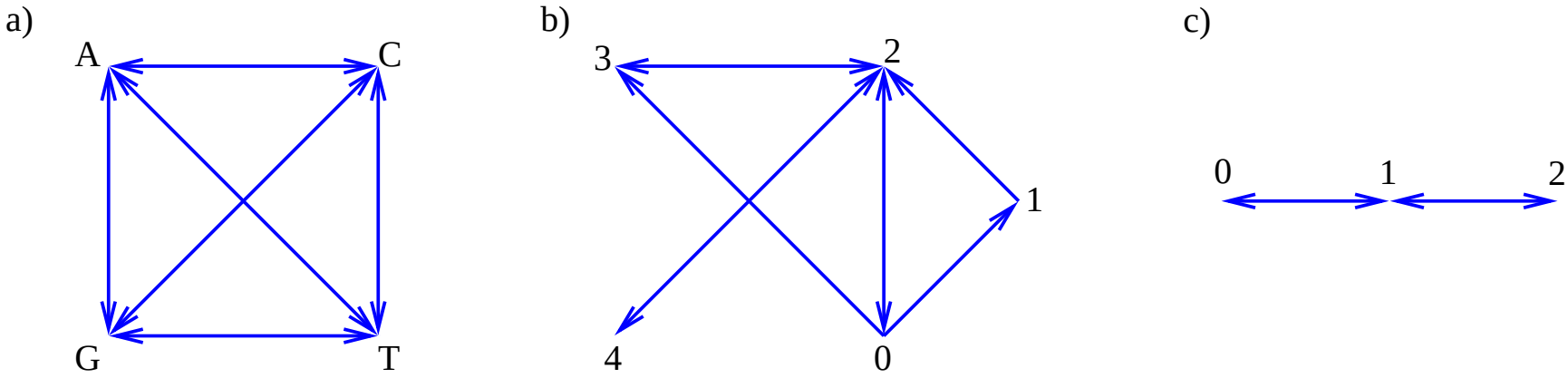
B. Schmiegelt, JK, J. Math. Biol. 2023

- Generalize the binary hypercube to Hamming graphs $\mathbb{H}_a^L$ with $a > 2$
- Biologically relevant cases are $a = 4$ (DNA, RNA) and $a = 20$ (proteins)
- Allowed mutational transitions between alleles are encoded by the $a \times a$ adjacency matrix $\mathbf{A}$ of the mutation graph
- Consider a sequence of initial and endpoints $\alpha^{(L)}, \omega^{(L)}$ such that the fraction of sites at which $\alpha_i^{(L)} = k$ and $\omega_i^{(L)} = l$ is given by p_{kl} for $L \to \infty$
- Theorem: The accessibility threshold β_c is given by the solution β^* of

$$\lim_{L\to\infty} [\mathbb{E}(X_{\alpha,\omega})]^{1/L} = \prod_{k,l=0}^{a-1} [(e^{\beta \mathbf{A}})_{kl}]^{p_{kl}} = 1$$

for most (but not all) mutation graphs. In general, β^* is a lower bound on β_c, and there are no accessible paths if $\beta^* > 1$

Examples of mutation graphs



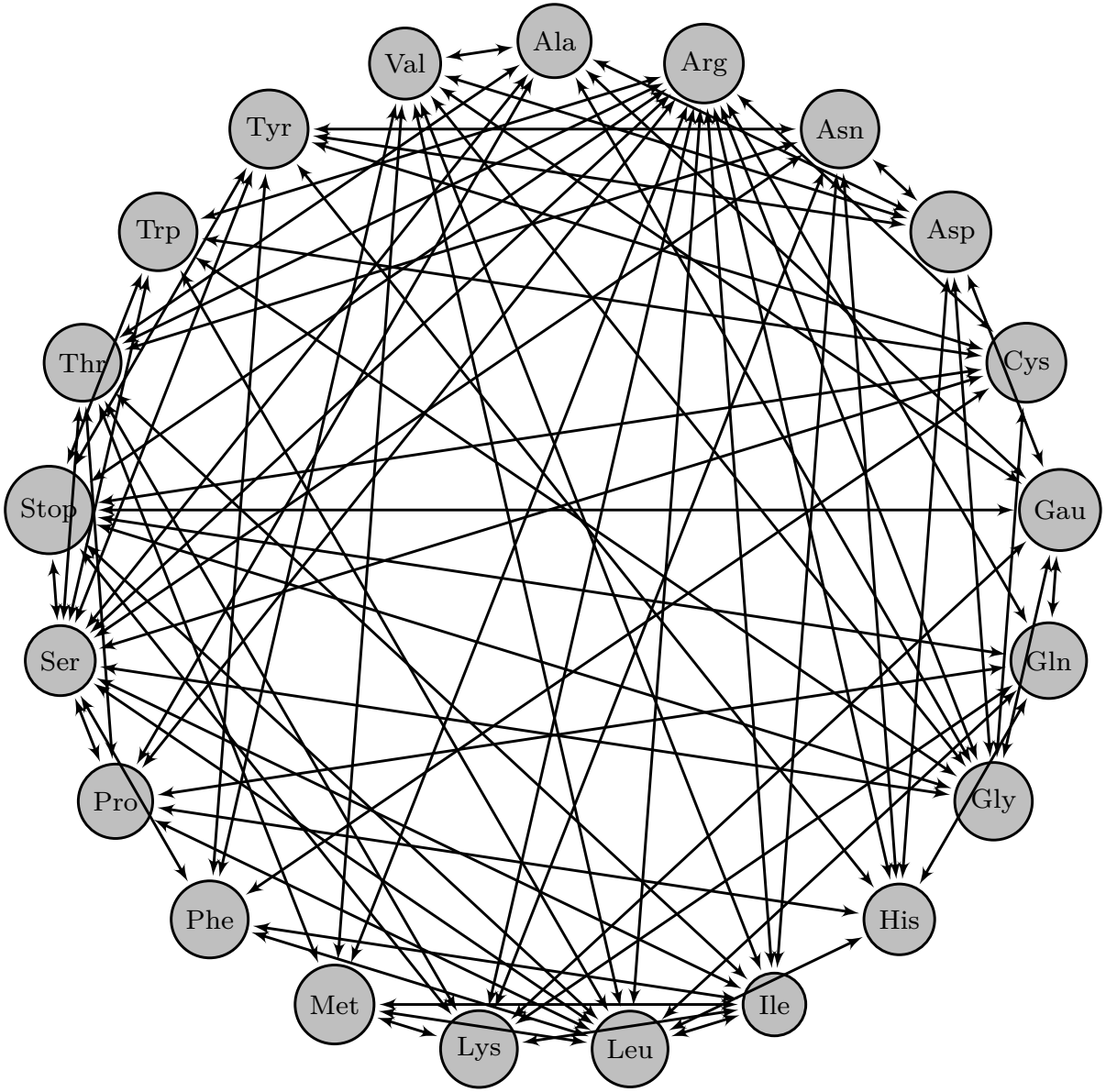
a) Nucleotide mutation graph $(a = 4)$:

$$\beta_c(\delta = 1) = \ln\left(\frac{1}{\sqrt{2}} + \sqrt{\sqrt{2} - \frac{1}{2}}\right) \approx 0.509$$

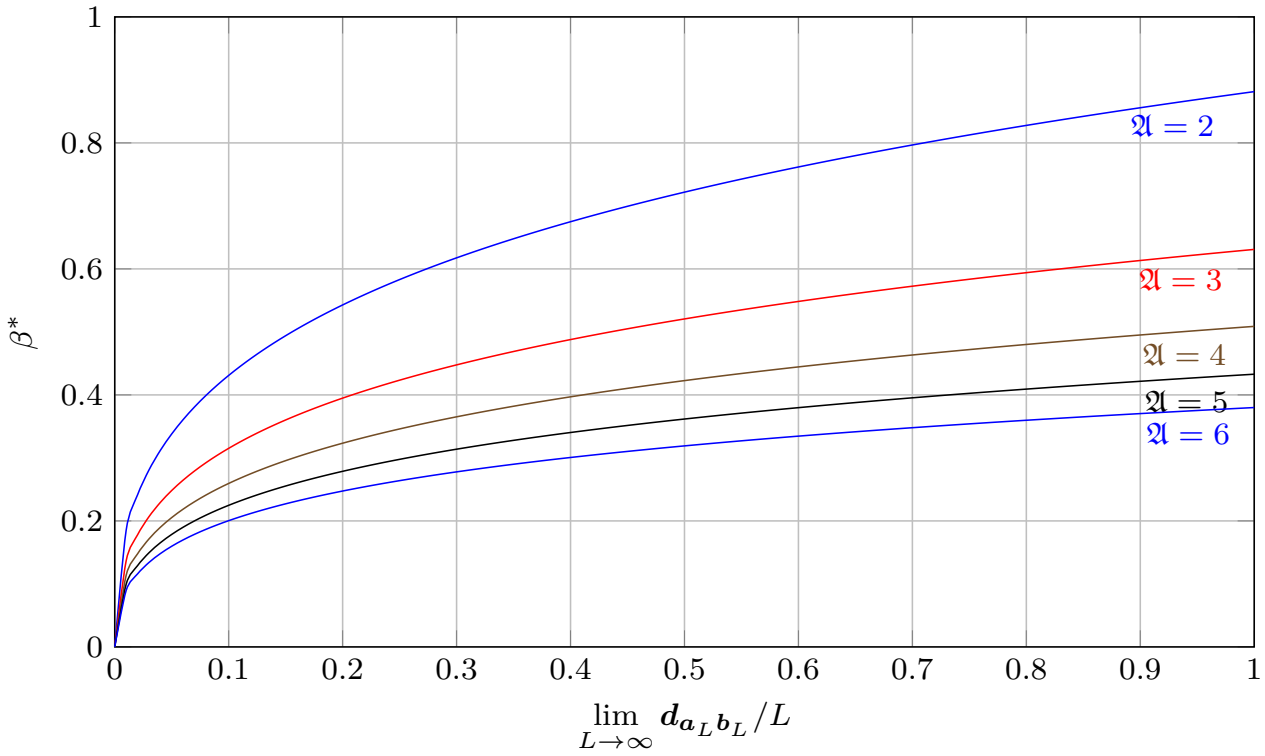
b) Smallest known mutation graph for which $\beta_c > \beta^*$ and $\beta^* < 1$

c) Path graph with $a = 3$: $\beta^*(\delta = 1) = \sqrt{2}^{-1} \ln(3 + 2\sqrt{2}) \approx 1.25 > 1$

The amino acid mutation graph ($a = 21$)



Accessibility threshold for the complete graph



- Accessibility threshold at full distance ($\delta = 1$) is

$$\beta_c(a) = \frac{\ln(a)}{a} + \frac{1+\ln(a)}{a^2} + \mathscr{O}\left(\frac{\ln(a)}{a^3}\right) \quad \text{for large } a$$

and the path length ℓ_c at the threshold is $\frac{\ell_c}{L} \approx \ln a + \frac{1+\ln a}{a}$

Structured fitness landscapes

Models of structured fitness landscapes

- Kauffman's NK model — Kauffman & Weinberger 1989; Hwang et al. 2018

$$g(\sigma) = \sum_i g_i(\sigma_{b_{i,1}}, \sigma_{b_{i,2}}, ..., \sigma_{b_{i,k}}) \text{ with } \{b_{i,1}, ..., b_{i,k}\} \subseteq \{1, 2, ..., L\}, 1 \leq k \leq L$$

g_i: HoC fitness landscape on the k-dimensional hypercube

- Rough Mt. Fuji model — Aita et al. 2000; Neidhart et al. 2014

$$g(\sigma) = -cd(\sigma, \sigma^*) + \xi_\sigma$$

with $c > 0$, i.i.d. RV's ξ_σ and a reference genotype σ^*

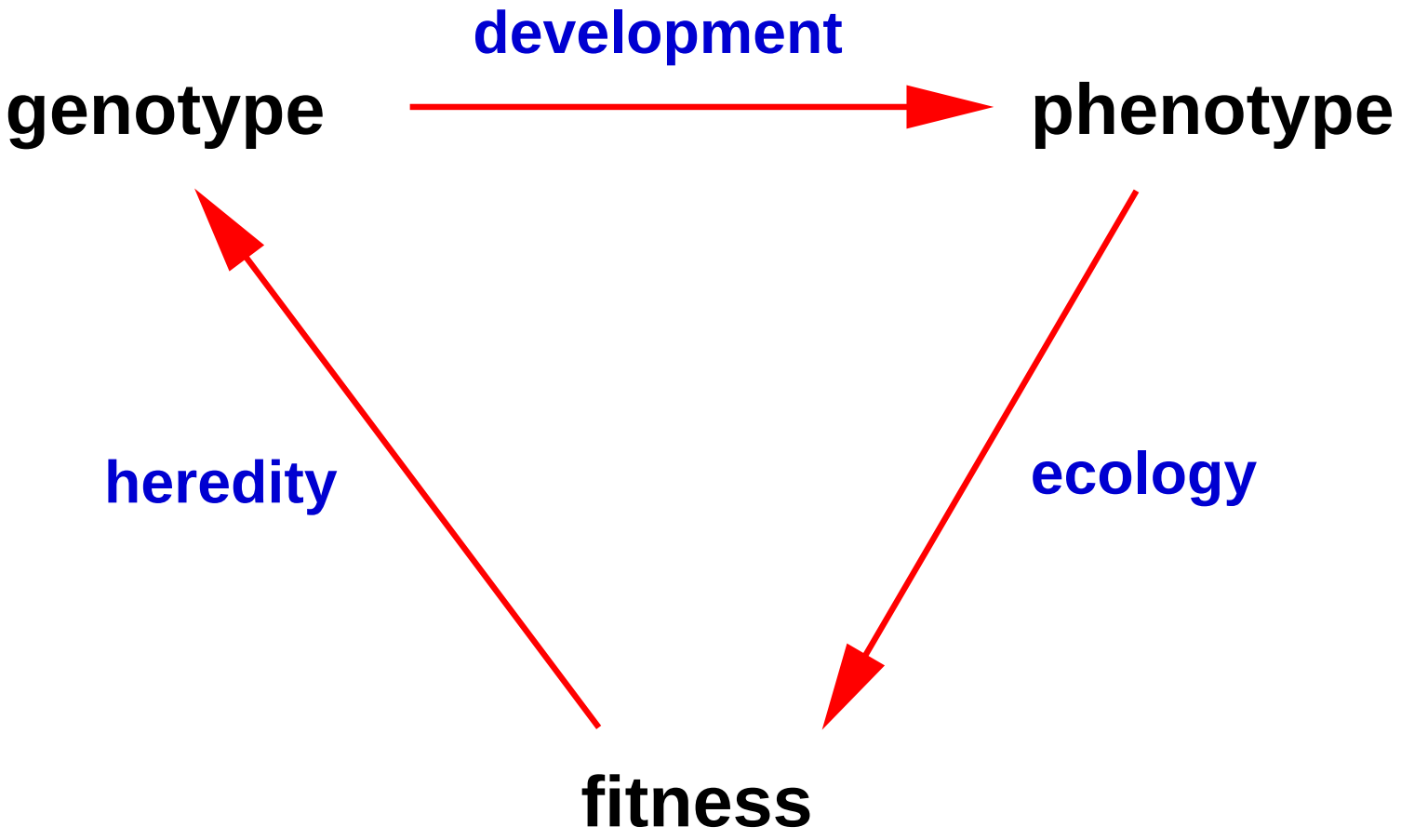
- Landscapes with an intermediate phenotype — Fisher 1930; Hwang et al. 2017

$$g(\sigma) = \Phi\left(\sum_{i=1}^{L} a_i \sigma_i\right)$$

with random coefficients a_i and a nonlinear phenotype-fitness map Φ

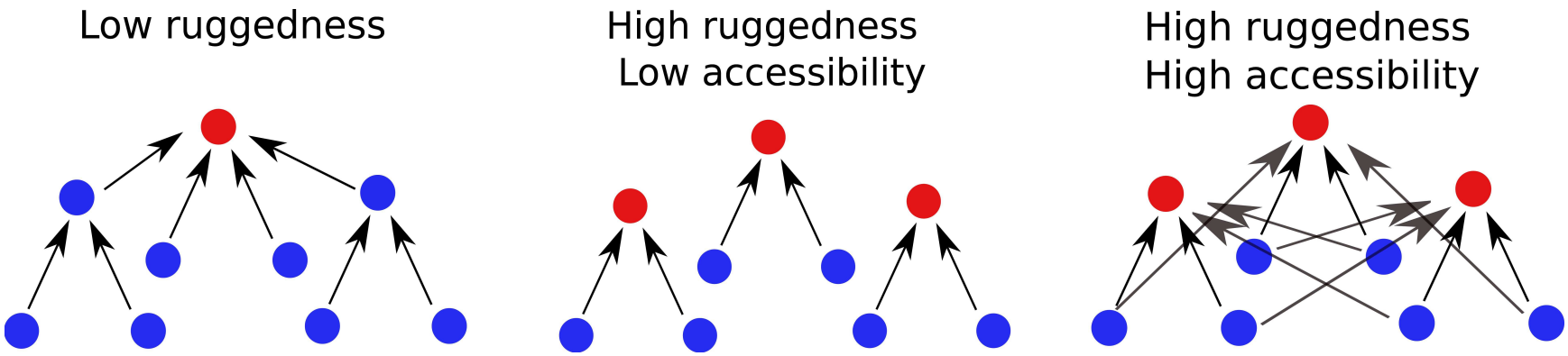
Genotype-phenotype-fitness maps

Courtesy Amitabh Joshi, JNCASR



Main messages of this part

- There is a large class of fitness landscapes that combine high ruggedness with high accessibility



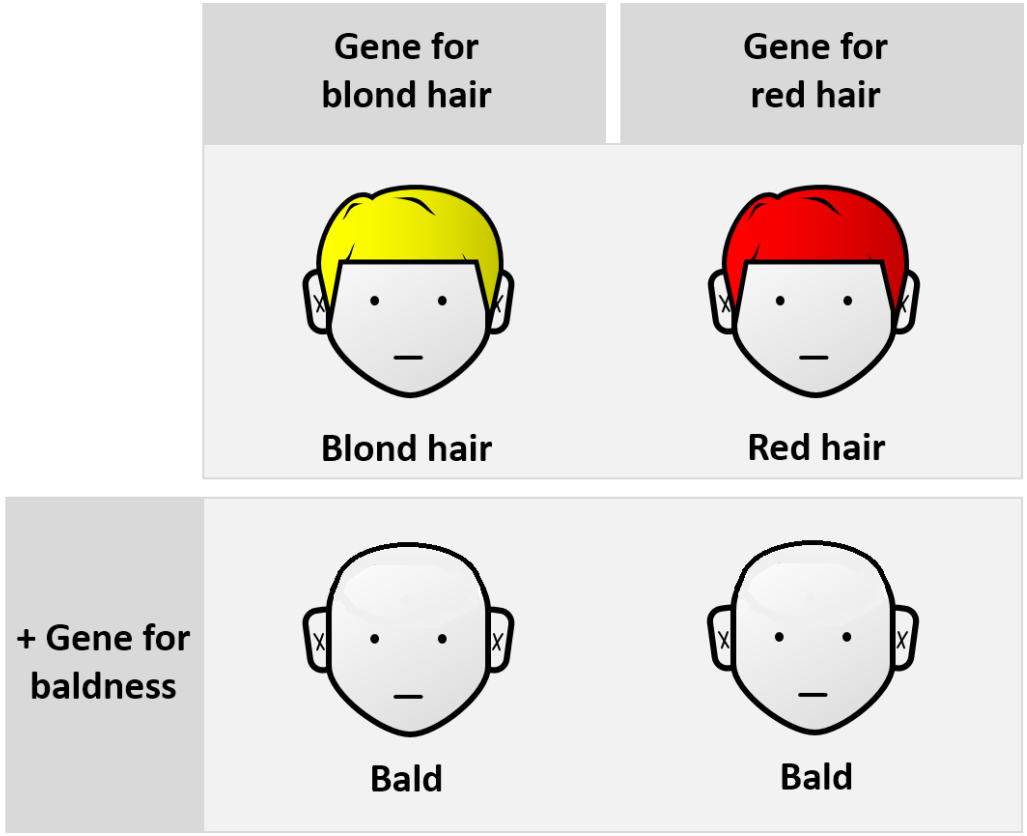
S.G. Das, S. Direito, B. Waclaw, R. Allen, JK, eLife 9:e55155 (2020)

- Such landscapes display universal negative epistasis (UNE), a property of set functions known in discrete mathematics as submodularity

- UNE and high accessibility arise naturally when a nonlinear phenotype-fitness map acts on one or several linear phenotypes

Epistasis: Historical definition

- William Bateson (1909): Epistasis implies that one mutation masks ("stands above") the phenotypic effect of another



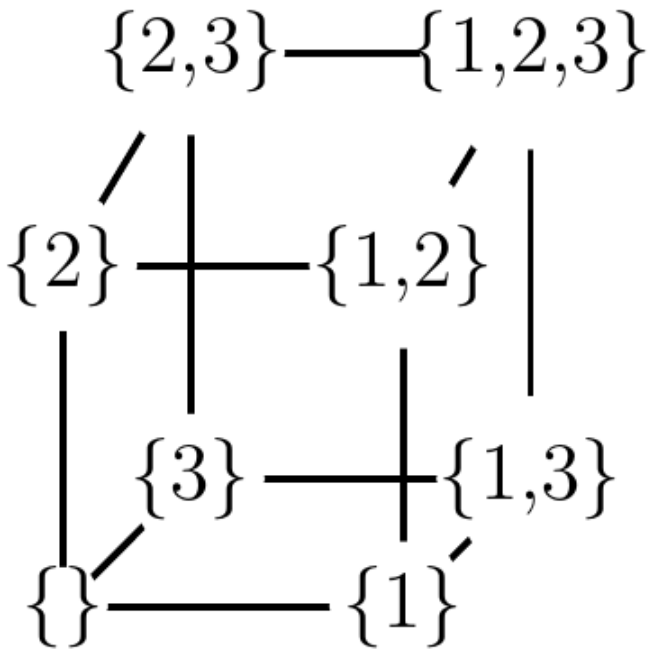
Credit: Wikipedia/Thomas Shafee

Genotypes as sets

Das et al, eLife 2020

- The hypercube $\{0,1\}^L$ is isomorphic to the power set $\mathscr{P}(\{1,\ldots,L\})$:

$$\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_L) \in \{0,1\}^L \;\rightarrow\; \{i : \sigma_i = 1\} \in \mathscr{P}(\{1,\ldots,L\})$$



credit: D. Oros

- In the following $\mathscr{L} = \{1,\ldots,L\}$ denotes the locus set

Universal epistasis

K. Crona, JK, M. Srivastava, J. Math. Biol. 2023

- A fitness landscape displays universal negative epistasis, if for any two genotypes σ, σ' with $\sigma' \subset \sigma \subset \mathscr{L}$, and any subset $\tau \subseteq \mathscr{L} \setminus \sigma$

$$g(\sigma \cup \tau) - g(\sigma) \leq g(\sigma' \cup \tau) - g(\sigma') \qquad \text{(UNE)}$$

  i.e. the fitness effect of adding the mutations in τ is smaller in the background σ than in the background σ', if σ' is a subset of σ

- Defining $\sigma = A$ and $\sigma' \cup \tau = B$, it follows that

$$\sigma \cup \tau = A \cup B, \qquad \sigma' = A \cap B$$

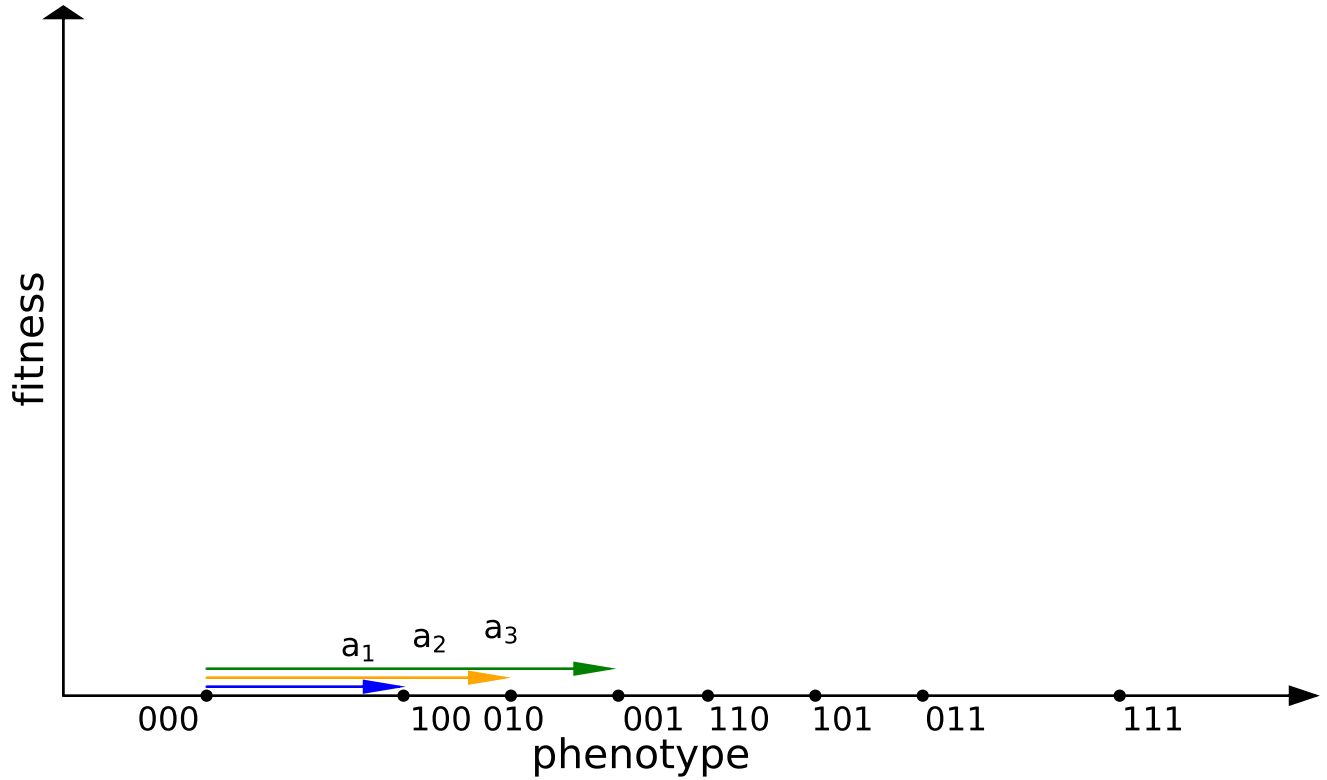
  and the condition can be rewritten as

$$g(A \cup B) + g(A \cap B) \leq g(A) + g(B) \quad \forall A, B \in \mathscr{P}(\mathscr{L})$$

  which is known as submodularity for set functions Edmonds 1970

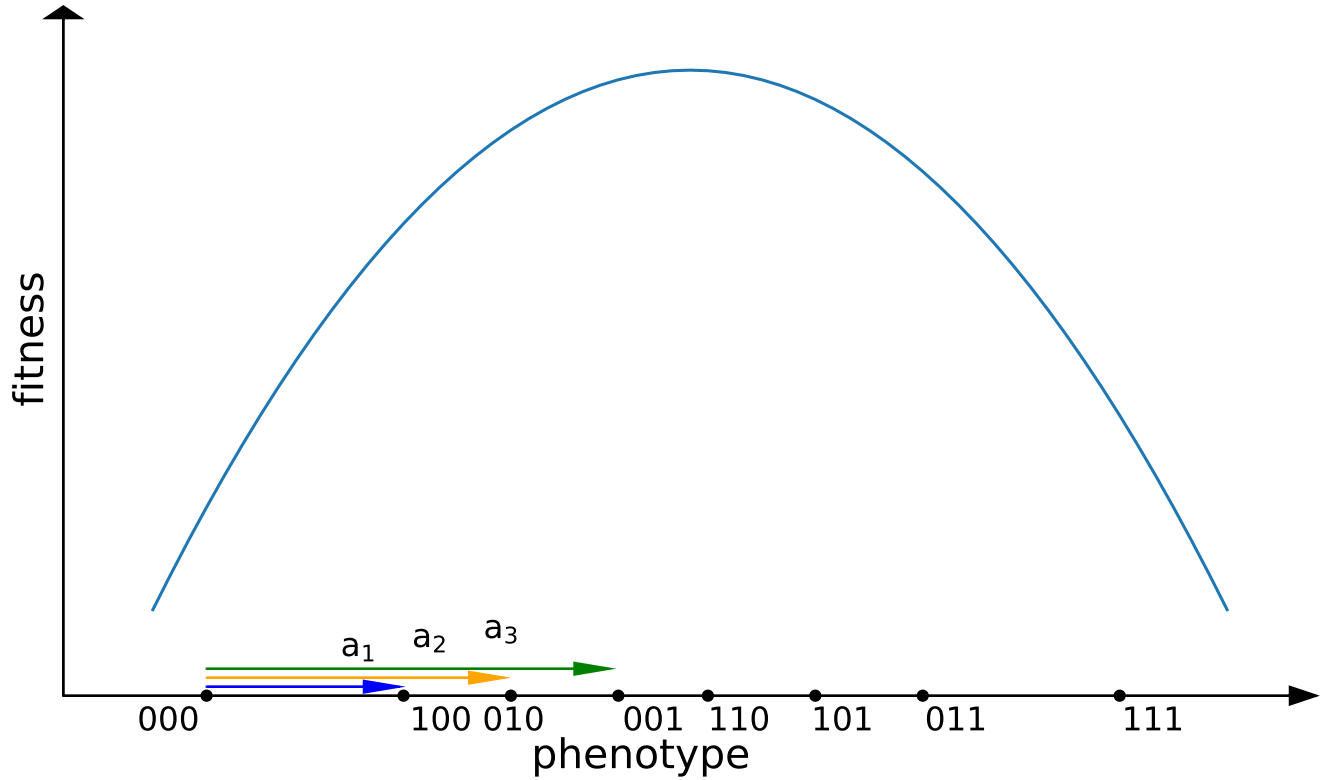
- Universal positive epistasis/supermodularity are defined in the same way

Constructing submodular landscapes



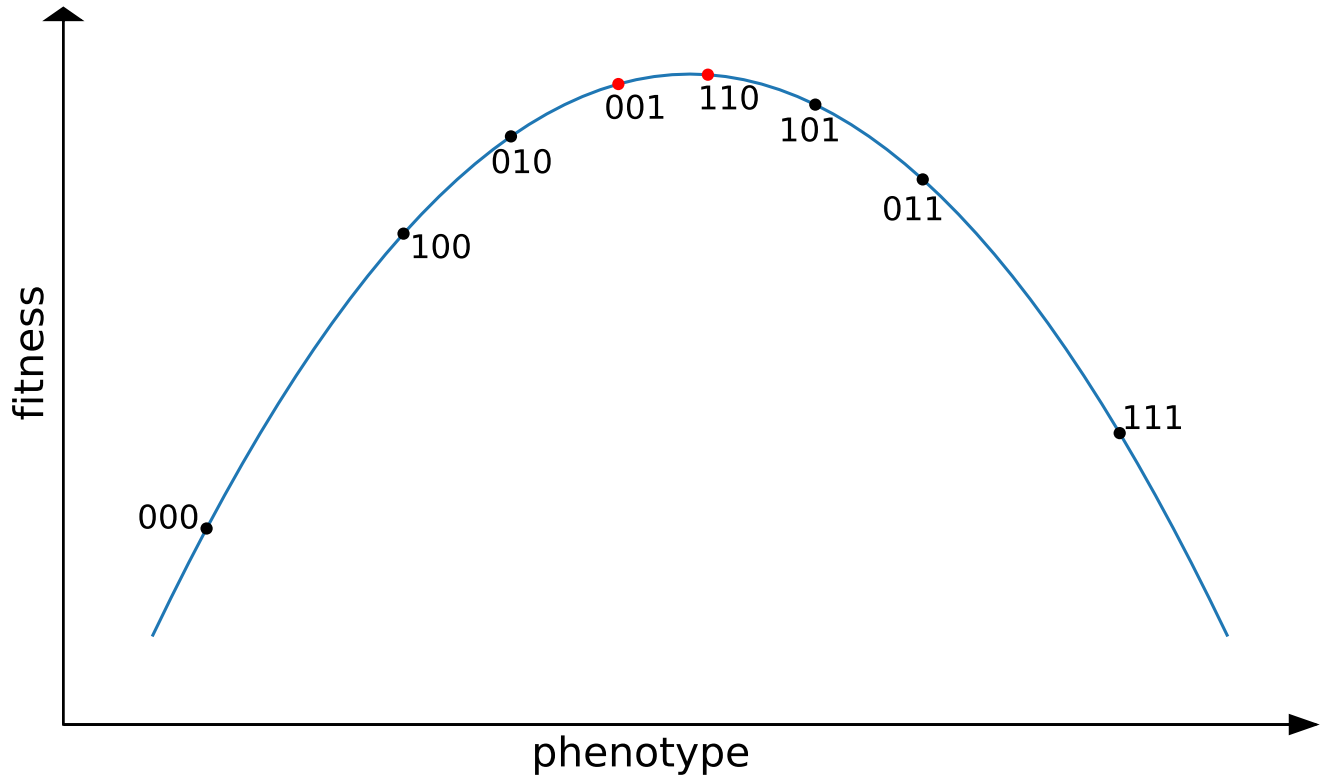
- phenotype: $z(\sigma) = a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3$

Constructing submodular landscapes



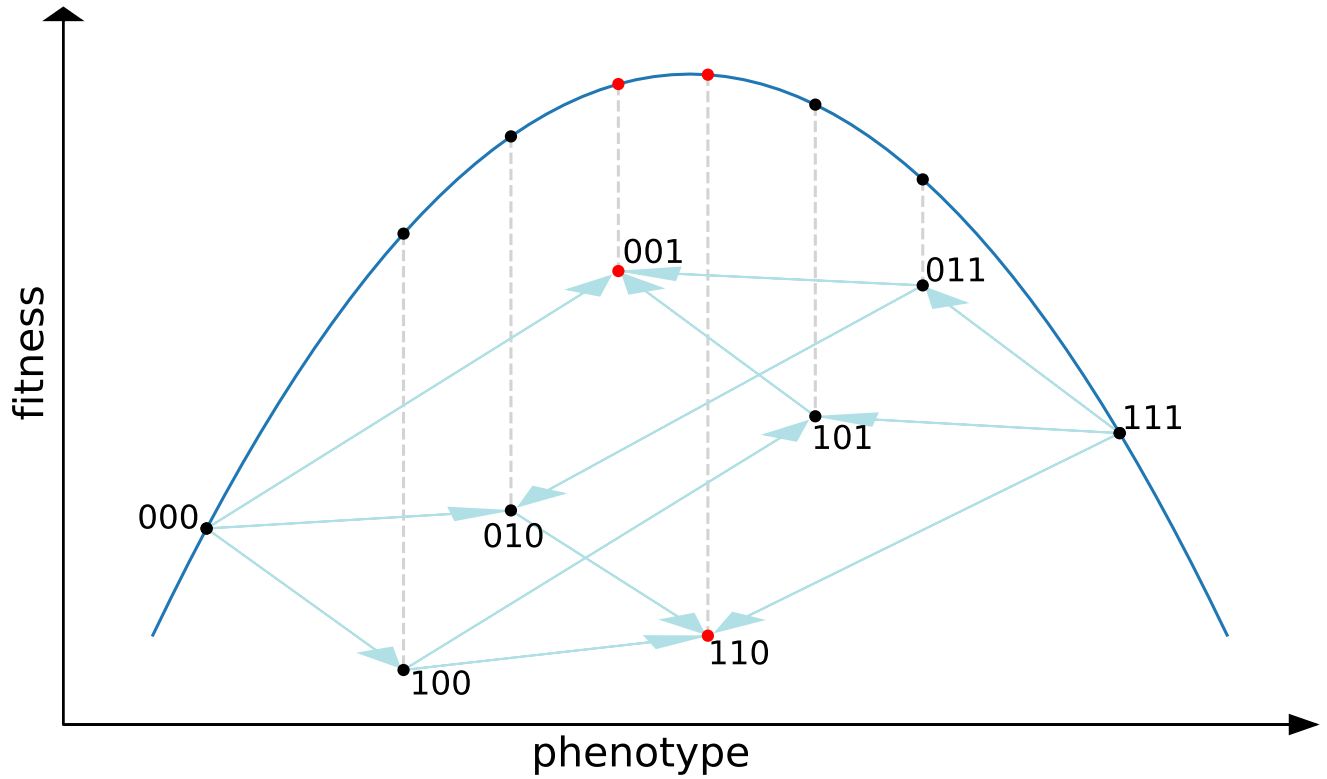
- phenotype: $z(\sigma) = a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3$
- fitness $\Phi(z) = -(z - z_0)^2$

Constructing submodular landscapes



- phenotype: $z(\sigma) = a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3$
- fitness $\Phi(z) = -(z - z_0)^2$
- genotype-phenotype-fitness map: $g(\sigma) = \Phi[z(\sigma)]$

Constructing submodular landscapes



- phenotype: $z(\sigma) = a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3$
- fitness $\Phi(z) = -(z - z_0)^2$
- genotype-phenotype-fitness map: $g(\sigma) = \Phi[z(\sigma)]$

Constructing submodular landscapes

- Fisher's geometric model (FGM) generates rugged fitness landscapes by composing a linear genotype-phenotype map with a non-monotonic, single-peaked phenotype-fitness map Φ:

$$\sigma \quad \rightarrow \quad z(\sigma) = \sum_{i=1}^{L} a_i \sigma_i \quad \rightarrow \quad g(\sigma) = \Phi\left(\sum_{i=1}^{L} a_i \sigma_i \right)$$

- FGM satisfies (UNE) if Φ is concave and the a_i are positive:

$$g(\sigma \cup \tau) - g(\sigma) = \Phi[z(\sigma) + z(\tau)] - \Phi[z(\sigma)] <$$

$$< \Phi[z(\sigma') + z(\tau)] - \Phi[z(\sigma')] = g(\sigma' \cup \tau) - g(\sigma')$$

because

$$z(\sigma') = \sum_{i \in \sigma'} a_i < \sum_{i \in \sigma} a_i = z(\sigma) \text{ if } \sigma' \subset \sigma$$

- The positivity condition on the a_i can be relaxed

Mapping to a Hopfield model

Park et al., J. Phys. A 2020

- The fitness function $\Phi(z) = -z^2$ corresponds to a Hamiltonian

$$H = -\Phi = \left(\sum_{i=1}^{L} a_i \sigma_i \right)^2 = \sum_{i,j} a_i a_j \sigma_i \sigma_j$$

- Transforming to spin variable $\eta_i = 1 - 2\sigma_i \in \{-1, 1\}$ gives

$$H = \sum_{ij} J_{ij} \eta_i \eta_J + \sum_i h_i \eta_i$$

with

$$J_{ij} = \frac{1}{4} a_i a_j, \quad h_i = -\frac{1}{2} \left(\sum_j a_j \right) a_i$$

- This is an antiferromagnetic Hopfield model with a real-valued pattern

Nokura 1998

The accessibility property

The accessibility property

- Recall that a genotype σ is accessible from genotype τ if there is a fitness-increasing (accessible) direct or indirect path $\tau \to \sigma$

- Definition: A fitness landscape has the subset-superset accessibility property (AP) if any peak is accessible from all its sub- and supersets along all direct paths Das et al., eLife 2020

- In particular, all peaks are always accessible from the 0-string $\sigma = \emptyset$ and the 1-string $\sigma = \mathscr{L}$

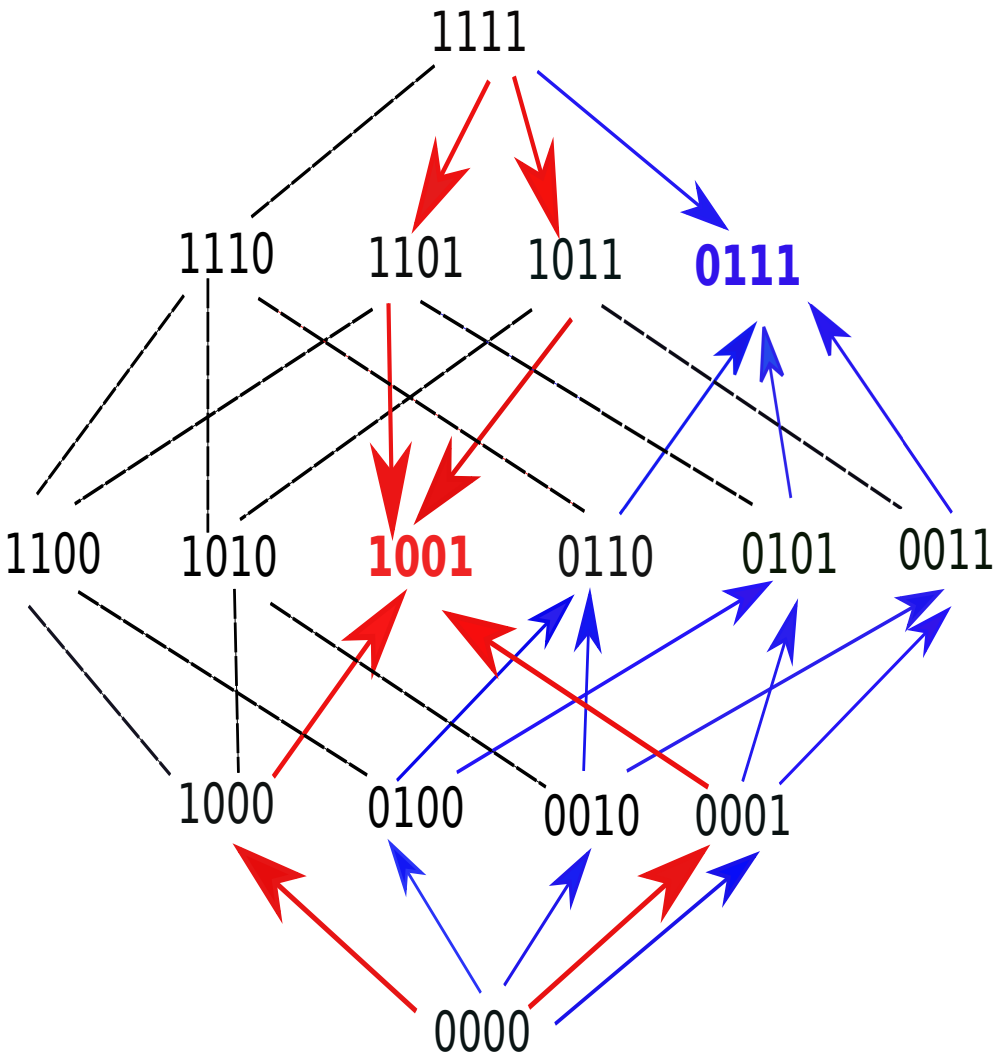
- The accessibility property implies a lower bound

$$S_\sigma \geq 2^{|\sigma|} + 2^{L-|\sigma|} - 1$$

on the size S_σ of the basin of attraction of a peak genotype σ

- The AP depends only on the rank ordering of fitness values and is therefore invariant under arbitrary monotonic transformations of fitness

Illustration of the accessibility property for $L = 4$



red: sub/supersets of 1001

blue: sub/supersets of 0111

Submodular fitness landscapes are highly accessible

Cherenin 1962; Goldengorin 2009; Krug & Oros 2024

- For any peak genotype σ

$$g(\sigma \cup \{i\}) - g(\sigma) < 0 \quad \text{and} \quad g(\sigma) - g(\sigma \setminus \{j\}) > 0$$

  for all $j \in \sigma$, $i \in \mathscr{L} \setminus \sigma$

- Consider a subset genotype $\sigma' \subset \sigma$ and a mutation $k \in \sigma \setminus \sigma' \subset \sigma$

- Then by (UNE)

$$g(\sigma' \cup \{k\}) - g(\sigma') \geq g(\sigma) - g(\sigma \setminus \{k\}) > 0$$

  which implies that the mutation k is beneficial on the background σ', and hence the corresponding step is accessible

- Accessibility from superset genotypes is proved in the same way

Summary

- Progress in the experimental exploration of biological fitness landscapes motivates the study of random landscape models
- The goal is the characterization of landscape topography through quantitative measures such as complexity, accessibility and epistasis
- The field offers opportunities for statistical physics approaches, and connects to glass physics and combinatorial optimization

Thanks to:

- Arjan de Visser (Wageningen), Su-Chan Park (Seoul), Kristina Crona (Washington DC)
- Sungmin Hwang (Paris), Suman Das (Bern), Muhittin Mungan (Cologne)
- Alexander Klözer, Jasper Franke, Johannes Neidhart, Stefan Nowak, Benjamin Schmiegelt, Alexander Klug, Daniel Oros, Muna Turki