PROBABILISTIC STRUCTURES IN EVOLUTION

DFG SPP 1590

COLLABORATIVE RESEARCH CENTER | SFB 680

Molecular Basis of Evolutionary Innovations

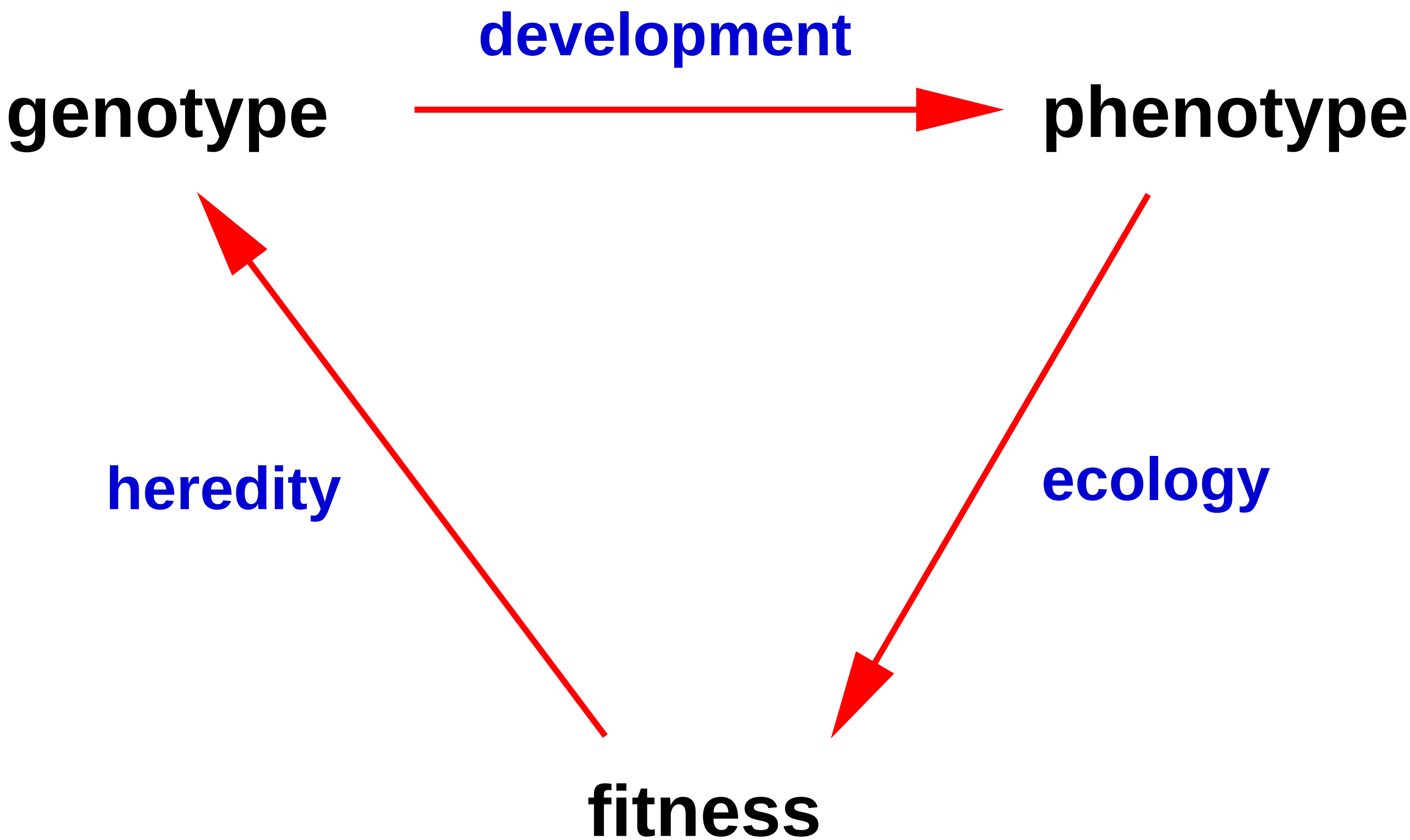
Genotypes, phenotypes and Fisher's geometric model

Joachim Krug
Institute for Theoretical Physics
University of Cologne

117th Statistical Mechanics Conference, Rutgers University, May 7, 2017

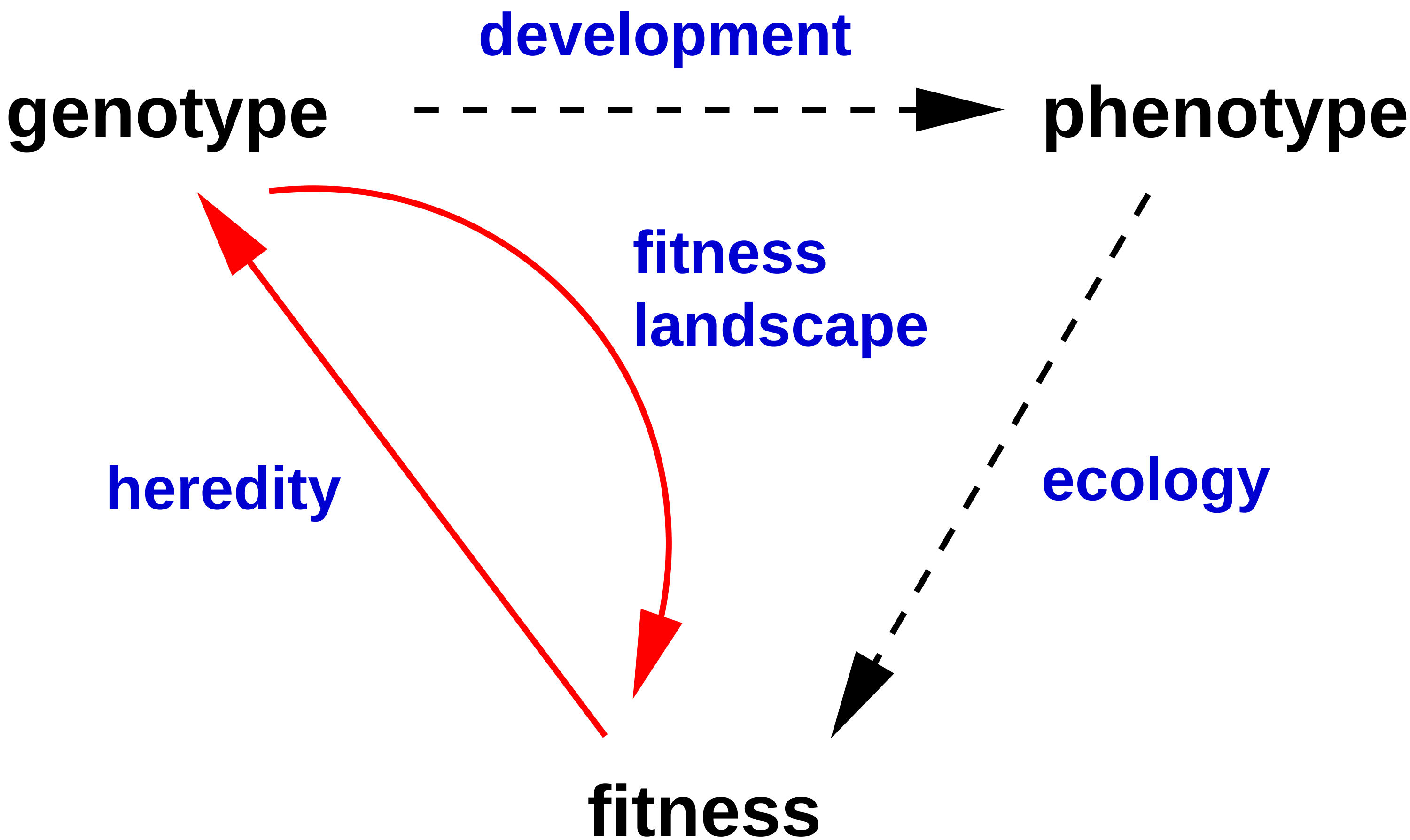
Biology in a nutshell

Courtesy Amitabh Joshi



Biology in a nutshell

Courtesy Amitabh Joshi



- Fitness landscape concept introduced by S. Wright (1932)

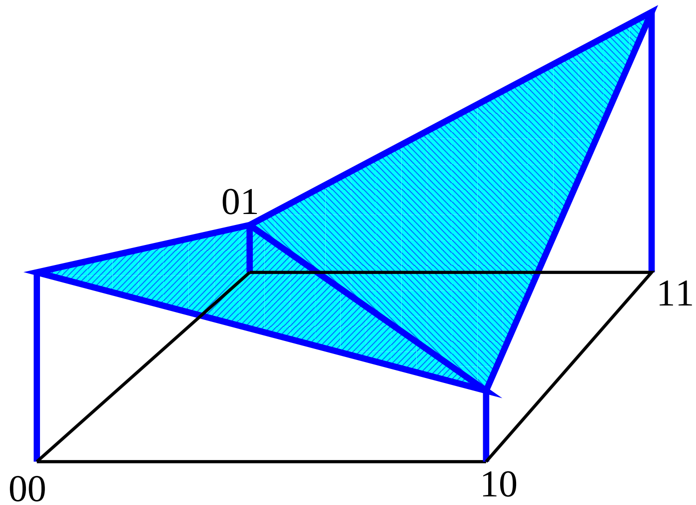
Fitness landscapes

J.A.G.M. de Visser, JK, Nature Reviews Genetics **15**, 480-490 (2014)

- General setting: L binary genetic loci τ_i at which a mutation can be present $(\tau_i = 1)$ or absent $(\tau_i = 0)$.

- A fitness landscape is a function on the set of 2^L genotypes

- A fitness landscape is complex/rugged if it has multiple fitness maxima:

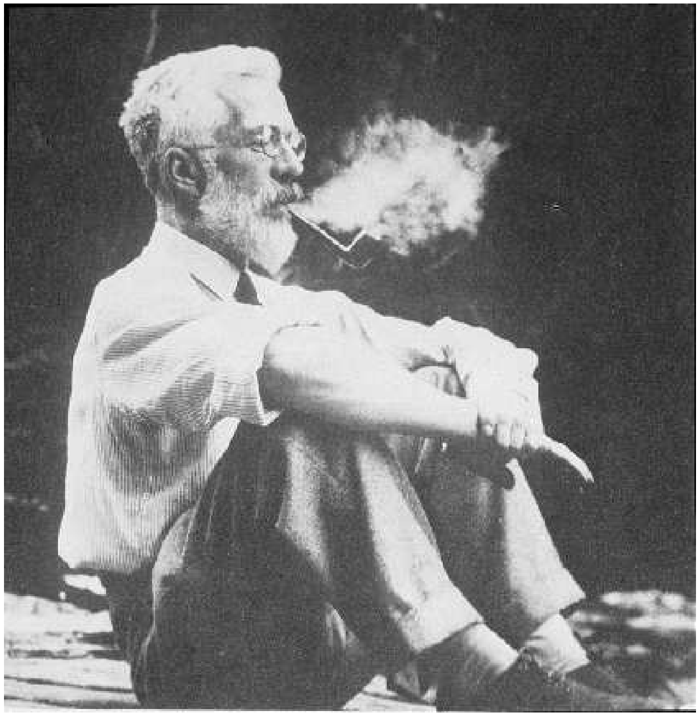


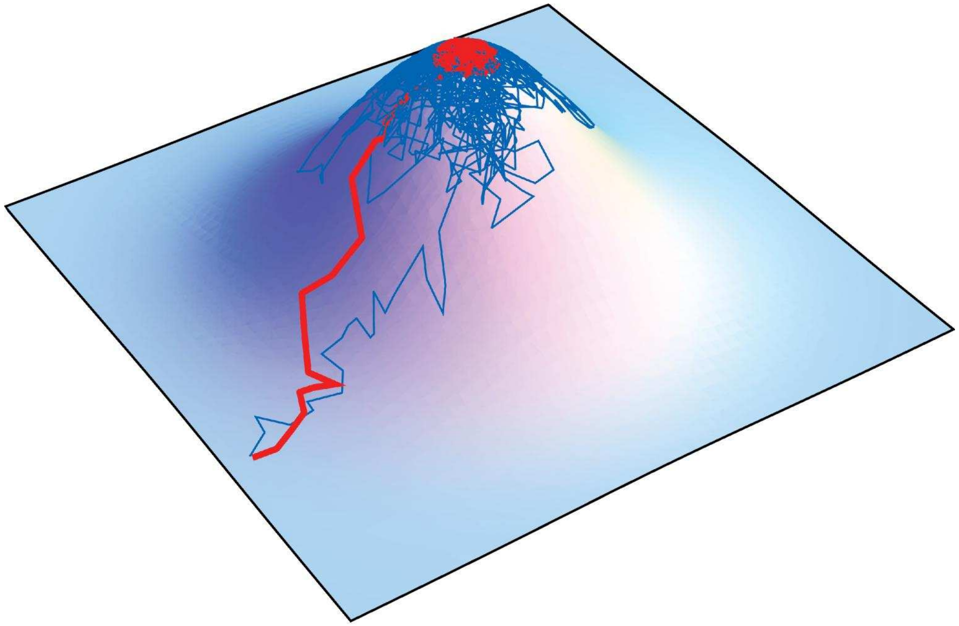
- Question for this talk: How do rugged fitness landscapes arise from a nonlinear phenotype-fitness map?

Fisher’s geometric model

"The statistical requirements of the situation, in which one thing is made to conform to another in a large number of different respects, may be illustrated geometrically..."

R.A. Fisher, The Genetical Theory of Natural Selection (1930)





O. Tenaillon, Annu. Rev. Ecol. Evol. Sys. (2014)

From simple phenotypes to complex genotypes

- Organism is characterized by n real-valued phenotypic traits x_i which form a vector $\vec{x} = (x_1, x_2, \ldots, x_n)$ in a n-dimensional Euclidean space

- Fitness is a (nonlinear) function $F(\vec{x})$ of the phenotype with a unique optimum at the origin $x_1 = x_2 = \ldots = x_n = 0$

- Universal pleiotropy: Mutations are isotropic random displacements in phenotypic space (univariate Gaussian)

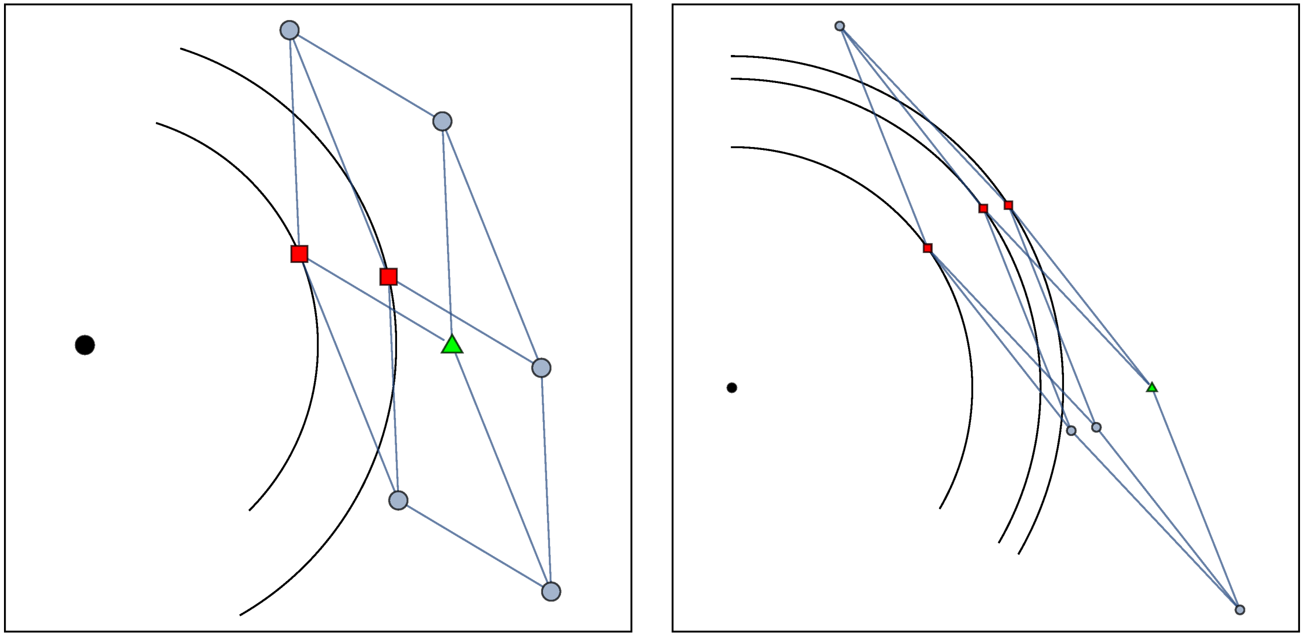
- Additivity of phenotypes: Given two phenotypic mutations $\vec{m}_1$, $\vec{m}_2$, the phenotypic effect of the double mutant is $\vec{m}_{12} = \vec{m}_1 + \vec{m}_2$ Martin et al. 2007

- Then the phenotypic landscape $F(\vec{x})$ induces a genotypic landscape

$$f(\tau_1, \ldots, \tau_L) = F\left(\vec{Q} + \sum_{i=1}^{L} \tau_i \vec{m}_i\right)$$

where $\vec{Q}$ represents the wildtype and the $\vec{m}_i$ are a fixed set of mutations

Geometry of the genotype-phenotype map



- The mapping

$$\tau \longrightarrow \vec{z}(\tau) = \vec{Q} + \sum_{i=1}^{L} \tau_i \vec{m}_i$$

projects L-dimensional hypercube onto n-dimensional phenotype space

- Figure shows the wild type phenotype (green triangle) and genotypic fitness maxima (red squares) for $L = 3, n = 2$

FGM as a spin glass model

- For a parabolic phenotypic fitness function $F(\vec{x}) = -|\vec{x}|^2$ the genotypic fitness landscape becomes

$$f(\tau) = -|\vec{Q}|^2 - 2\sum_{i=1}^{L}(\vec{Q}\cdot\vec{m}_i)\tau_i - \sum_{i,j=1}^{L}(\vec{m}_i\cdot\vec{m}_j)\tau_i\tau_j$$

which corresponds to an antiferromagnetic Hopfield model with n continuous patterns and random fields of strength $\sim |\vec{Q}|$

- The linear part dominates for large $|\vec{Q}| \Rightarrow$ fitness landscape is less rugged when wildtype phenotype is far from the origin

- The model displays a zero temperature phase transition at

$$q = \frac{|\vec{Q}|}{L} = q_0 = \frac{1}{\sqrt{2\pi}} \approx 0.39894$$

where the extensive part of the ground state entropy vanishes

S. Hwang, D. Dean, JK (unpublished)

Genotypic complexity of FGM

S. Hwang. S.-C. Park, JK, Genetics (Early Online)

Number of genotypic maxima

- A common global quantifier of genotypic complexity is the expected number of genotypic fitness maxima $\langle \mathcal{N} \rangle$

- Experience with random field models shows that in many cases

$$\langle \mathcal{N} \rangle \sim \exp[\Sigma^* L] \quad \text{for} \quad L \to \infty$$

  which defines the genotypic complexity $\Sigma^* \geq 0$

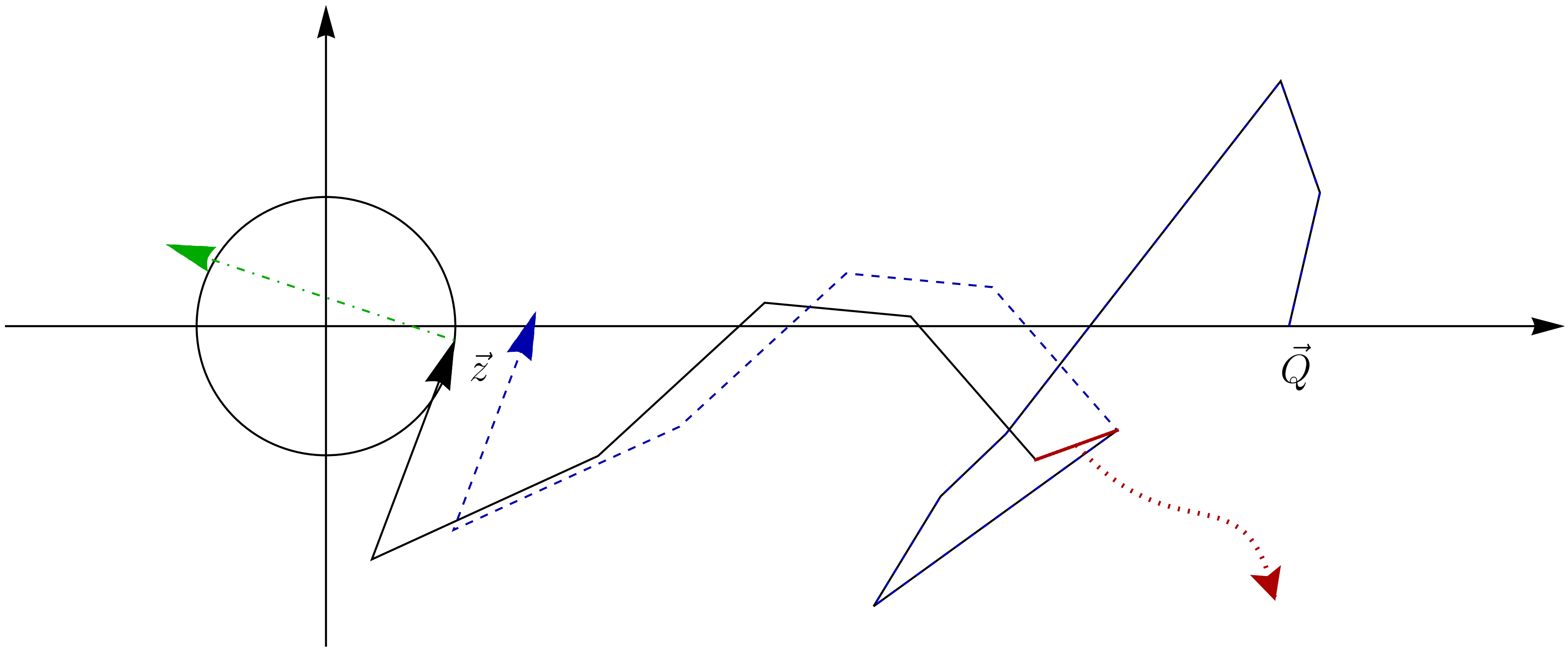
- Within FGM, a genotype $\tau = (\tau_1, \tau_2, \ldots, \tau_L)$ with phenotype

$$\vec{z} = \vec{Q} + \sum_{i=1}^{L} \tau_i \vec{m}_i$$

  is a fitness maximum iff $|\vec{z}| < |\vec{z} + (1 - 2\tau_j)\vec{m}_j|$ for all $j = 1, \ldots, L$

- This is true with unit probability if the corresponding phenotype is optimal, i.e. if $\vec{z} = 0 \quad \Rightarrow \quad$ genotypic maxima arise from near-optimal phenotypes

Number of genotypic maxima: Geometry



- Composition of mutation vectors defines a random walk (“polymer”) in phenotype space with endpoint $\vec{z}$

- To generate genotypic maxima, the polymer needs to be “stretched” towards the origin

Number of genotypic maxima: Asymptotics

- Expected number of maxima for large L is given by $\langle \mathcal{N} \rangle \sim L^{-(1+n/2)} \exp[\Sigma^* L]$ where Σ^* is the solution of the variational problem

$$\Sigma^* = \max_{\phi \in [0,1]} \left\{ -\phi \log \phi - (1-\phi) \log(1-\phi) - \frac{q^2}{2\phi} \right\}$$

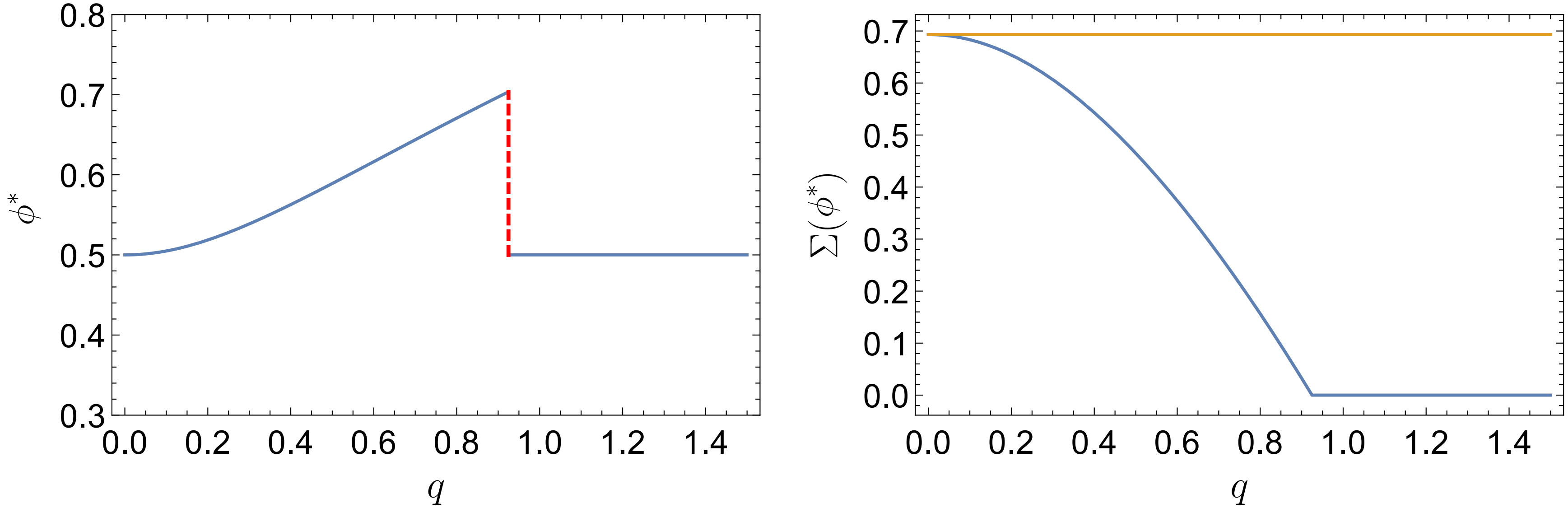
  with

  - ϕ: fraction of mutations that are present (= have $\tau_i = 1$)
  - $q = |\vec{Q}|/L$: scaled distance of the wild type phenotype to the optimum

- Variational problem encodes a tradeoff between the abundance of genotypes (“entropy”) and their likelihood to reach the phenotypic optimum (“energy”)

- The number of maxima decreases with increasing phenotypic dimension, but to leading (exponential) order it is independent of n

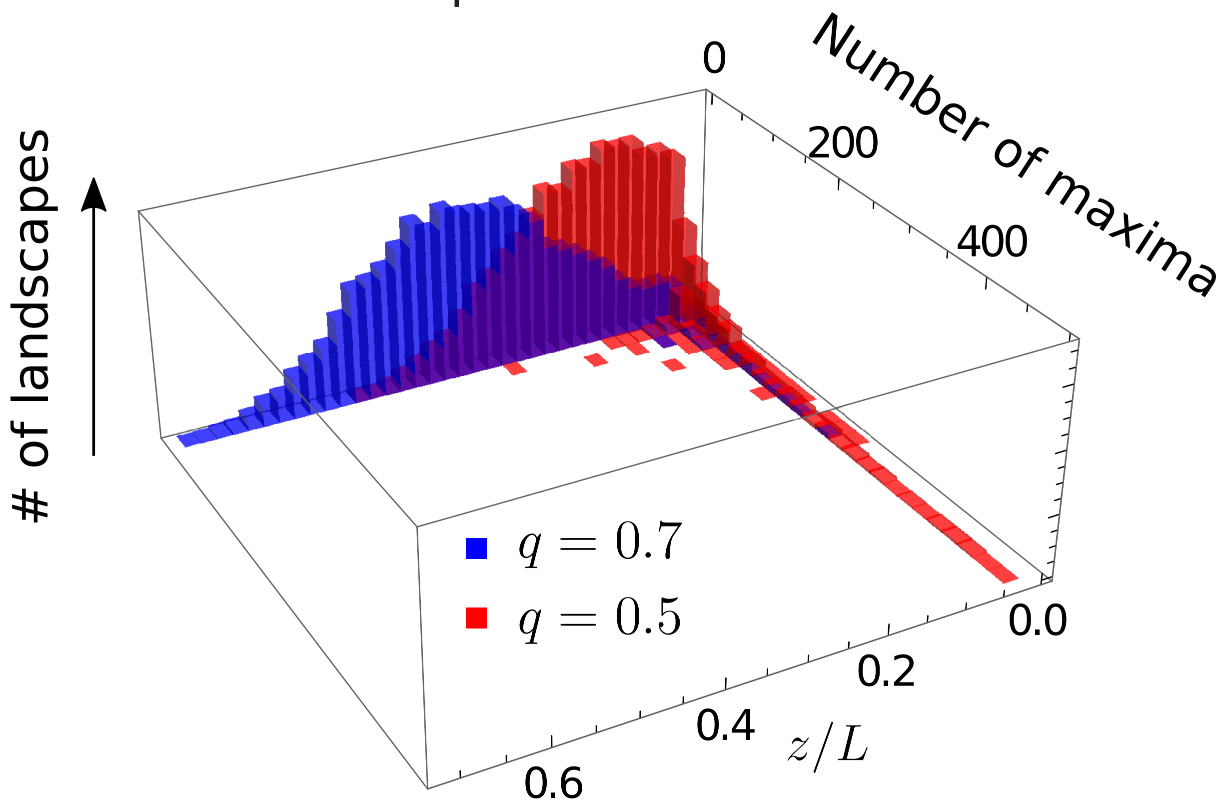
Number of genotypic maxima: Phase transition



- $\Sigma^*(q=0) = \ln 2 \Rightarrow \langle \mathcal{N} \rangle \sim \frac{2^L}{L^{1+n/2}}$, to be compared to an uncorrelated random fitness landscape ("random energy model") with $\langle \mathcal{N} \rangle \sim \frac{2^L}{L}$
- Σ^* vanishes at a first order phase transition at $q = q_c \approx 0.924809 > q_0$
- For $q > q_c$ the number of maxima reaches a finite limit for $L \to \infty$ which however grows exponentially with n

Coexistence and rare events

- In the coexistence region $q_0 < q < q_c$, $\langle \mathcal{N} \rangle$ is dominated by rare realizations with exponentially many maxima, whereas typical realizations have a moderate number of peaks



- These rare realizations are those for which the phenotypic displacements approach close to the optimum $z = 0$

Interactions between beneficial mutations in Aspergillus nidulans

S. Schoustra, S. Hwang, JK and J.A.G.M. de Visser, Proc. Roy. Soc. B (2016)

Experimental system

- 244 beneficial mutants of *A. nidulans* collected from the boundary of growing colonies in complex (rich) or minimal (poor) medium

- Generated 55 pairwise combinations between mutations of similar effect using sexual crosses

- Goal: Quantify the dependence of pairwise epistatic interaction

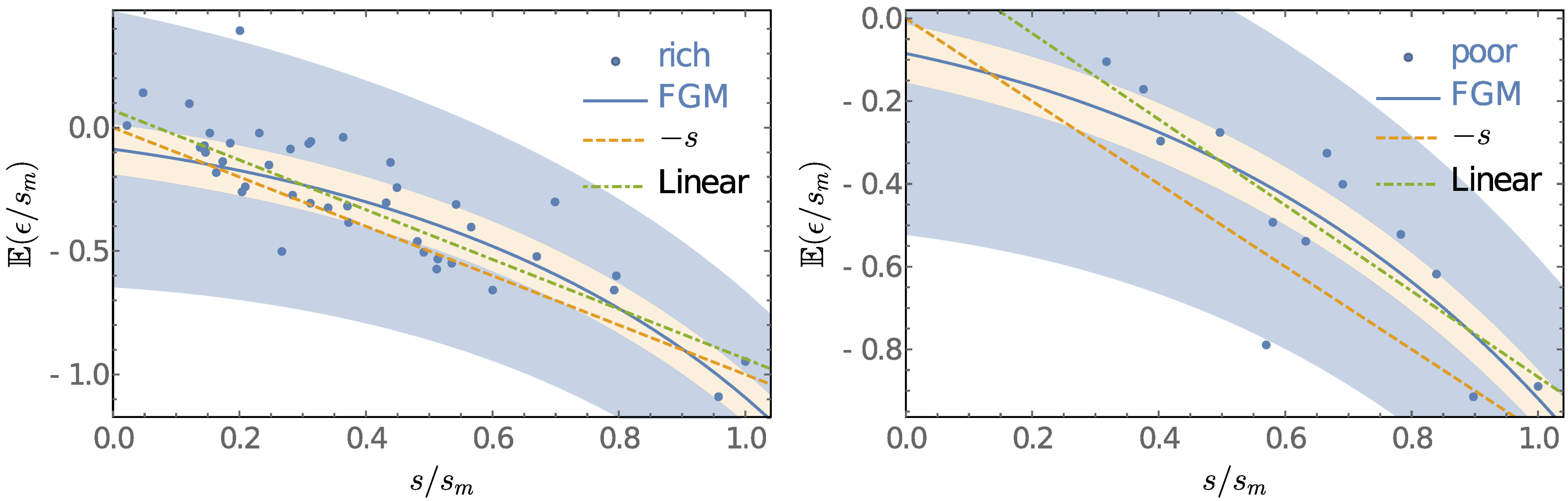
$$\varepsilon_{ab} = \Delta f_{ab} - (\Delta f_a + \Delta f_b)$$

  on the strength $s = \Delta f_a = \Delta f_b$ of single mutations

- Data show clearly that $\varepsilon_{ab} < 0$ and is negatively correlated with s ("diminishing returns epistasis")

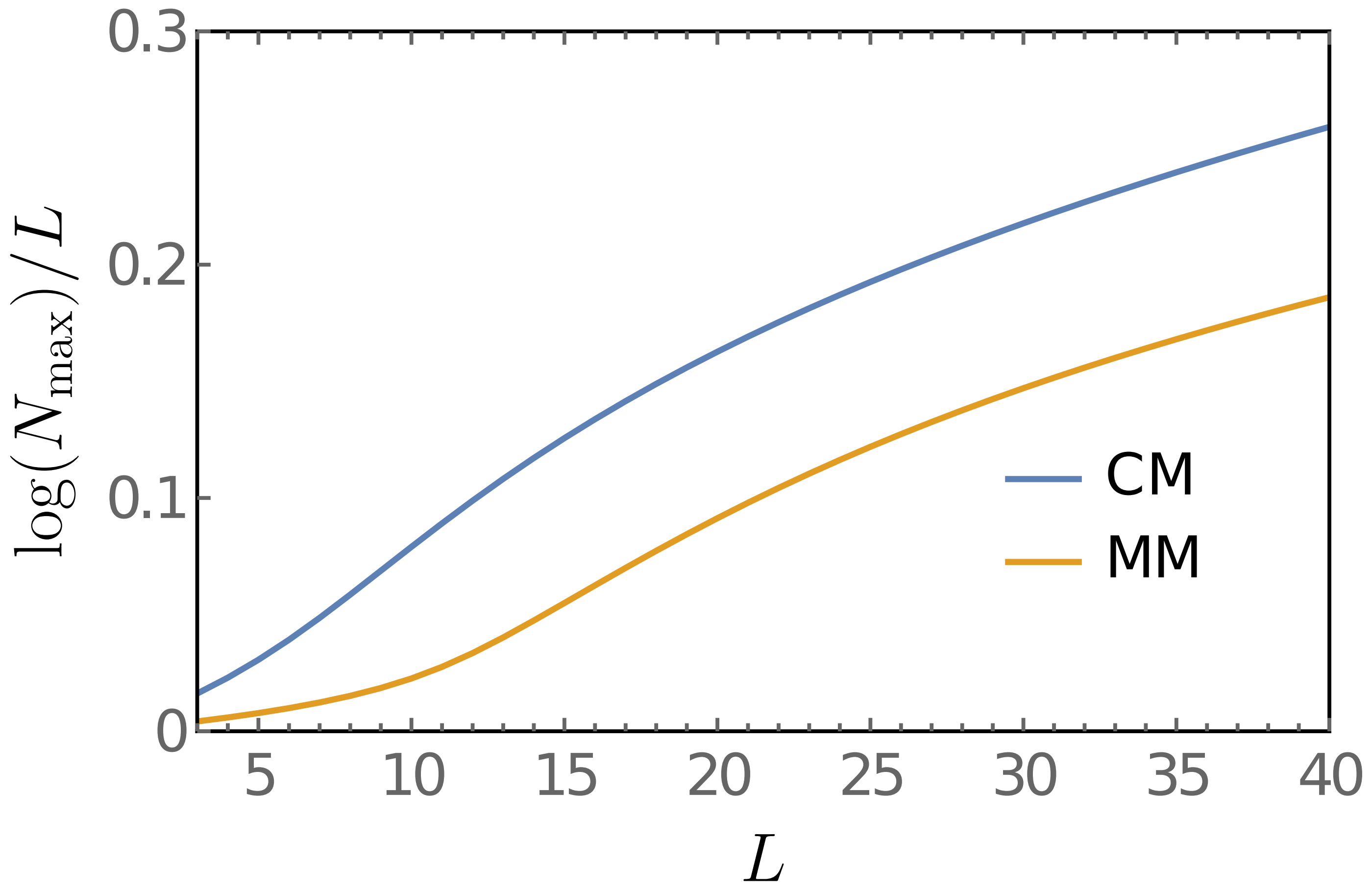
- FGM predicts the distribution of ε_{ab} conditioned on s, the first two moments of which can be computed analytically

Fit of FGM to data



- ε and s normalized to largest observed mutational effect s_m
- Measurement error (inner pink region) is insufficient to explain the observed variability $\Rightarrow$ importance of intrinsic stochasticity of FGM
- FGM parameters: $Q = 6.89,\ n = 19.3,\ s_0/s_m = 1.41$ (rich)
  $Q = 9.81,\ n = 34.8,\ s_0/s_m = 1.62$ (poor)
- How to interpret the differences in n?

Genotypic complexity of the A. nidulans landscapes



- Rich medium landscape (CM) is more rugged, despite having lower phenotypic dimension

Conclusions

- Fisher’s geometric model is a good example of a “proof-of-concept” model in biology
  Servedio et al., PLOS Biol. 2014

- It demonstrates how genotypic complexity can be explained in terms of additive phenotypes combined with a simple nonlinear phenotype-fitness map

- The model also provides a framework for condensing experimental data into a few phenomenological parameters, but their interpretation is not straightforward

- From the viewpoint of statistical physics, questions related to the phase structure and the role of rare events remain to be understood