Fitness landscapes in theory and application

Joachim Krug
Institute of Biological Physics, University of Cologne
E-mail: krug@thp.uni-koeln.de

Problem 1: Geometry of Hamming spaces

The Hamming space H_L^a is a set of all sequences of length L with letters drawn from an alphabet of size a and endowed with the natural Hamming distance as metric. For $a = 2$ the Hamming spaces H_L^2 are hypercubes. Here we consider arbitrary $a \geq 2$.

a.) What is the "diameter" of H_L^a, i.e. the maximal distance between two points? How many direct paths are there connecting two sequences at distance d?

b.) Compute the number $n(d, L, a)$ of sequences that have distance d from a given sequence. Evaluate the resulting expression for large L and d. At which distance $d_{\max}$ do most of the sequences lie, and what is the width of this region?

c.) What is the largest number of local maxima that a fitness landscape defined on H_L^a can have?

Problem 2: Fourier analysis of fitness landscapes

In the lectures, the Fourier basis of the graph Laplacian Δ of the hypercube was introduced. Here the statements presented in the lectures will be proved.

a.) Show that the functions

$$\phi_I(\sigma) = \phi_{i_1,\ldots,i_p}(\sigma) = 2^{-L/2}\sigma_{i_1}\ldots\sigma_{i_p}, \quad \sigma_i = \pm 1,$$

are orthonormal eigenfunctions of Δ with eigenvalues $\lambda_p = -2p$.

b.) Given a fitness landscape $f(\sigma)$, determine the coefficients in the expansion

$$f(\sigma) = \sum_{p=0}^{L} \sum_{0\leq i_1<i_2<\cdots<i_p\leq L} a_{i_1,\ldots,i_p}\phi_{i_1,\ldots,i_p}(\sigma).$$

As an application, compute the amplitude spectrum B_p of the two-locus landscape defined by

$$f(--) = 0, f(+-) = f(-+) = f_1, f(++) = f_2.$$

Under what condition on f_1 and f_2 is this fitness landscape non-epistatic?

Problem 3: The House-of-Cards model

In the House-of-Cards (HoC) model, the fitness values $f(\sigma)$ assigned to genotypes σ are independent, identically distributed (i.i.d.) random variables drawn from a continuous probability density $p(f)$. In this problem, the probability density does not need to be specified, as all properties of interest depend only on the ranked fitness landscape. Throughout we restrict ourselves to the case of binary sequences ($\sigma \in H_L^2$), and focus on the statistics of local fitness maxima. A sequence σ is a local maximum if $f(\sigma) > f(\sigma')$ for all σ' with $d(\sigma, \sigma') = 1$. The number of local fitness maxima $n_{\max}$ is a random variable that takes values between 1 and 2^{L-1}.

a.) Consider first the case $L = 2$ with 4 genotypes. Determine the probability distribution of $n_{\max}$ and compute its mean and variance.
Hint: Use the fact that all 4! orderings of fitness values are equally likely.

b.) The expected number of maxima $\mathbb{E}(n_{\max})$ can be easily obtained for general L from the following argument: The fitness of a local maximum is the largest in the set of $L + 1$ i.i.d. random variables formed by the sequence itself and its L neighbors. Since all variables are equivalent, the probability that any one of them is the largest is $\frac{1}{L+1}$. Hence $\mathbb{E}(n_{\max}) = \frac{2^L}{L+1}$. Here we want to derive the expression

$$\mathrm{Var}(n_{\max}) = \frac{2^L(L-1)}{2(L+1)^2} \tag{1}$$

for the variance of $n_{\max}$. To this end, we write

$$n_{\max} = \sum_{\sigma} \eta(\sigma),$$

where $\eta(\sigma) = 1$ if σ is a local fitness maximum, and $\eta(\sigma) = 0$ else. In this notation, the expected value of $n_{\max}^2$ becomes

$$\mathbb{E}(n_{\max}^2) = \sum_{\sigma,\sigma'} \mathbb{E}[\eta(\sigma)\eta(\sigma')]. \tag{2}$$

Using this representation, show that

$$\mathrm{Var}(n_{\max}) = \mathbb{E}(n_{\max}) + 2^L \binom{L}{2} p_2 - \frac{2^L}{(L+1)^2}\left[1 + L + \binom{L}{2}\right],$$

where p_2 denotes the probability that two sequences at distance $d = 2$ are both local maxima.
Hint: Subdivide the double sum (2) into four parts containing the pairs of sequences at distances $d(\sigma, \sigma') = 0$, $d(\sigma, \sigma') = 1$, $d(\sigma, \sigma') = 2$ and $d(\sigma, \sigma') \geq 3$, respectively.

c.) To complete the proof of (1), show that $p_2 = \frac{1}{L(L+1)}$.

d.) In the *constrained* HoC model, the global fitness maximum is placed at $\sigma^{(1)} = (1, 1, \cdots, 1)$ and the global fitness minimum at the antipodal sequence $\sigma^{(-1)} = (-1, -1, \cdots, -1)$. The fitnesses of the remaining $2^L - 2$ sequences are assigned randomly between $f(\sigma^{(-1)})$ and $f(\sigma^{(1)})$. Show that this modification does not change the expected number of fitness maxima and minima (compared to the standard HoC model).

Problem 4: The complexity catastrophe

We consider the HoC model and assume that the fitness values $f(\sigma)$ are distributed according to a Gaussian with zero mean and standard deviation S.

a.) We begin with a simple application of extremal statistics. A useful way to estimate the expected value $X_{\max}(N)$ of the largest among N independent, identically distributed random variables drawn from a probability density $p(x)$ is through the relation

$$\int_{X_{\max}(N)}^{\infty} p(x) \simeq \frac{1}{N}.$$

Use this relation to show that for the Gaussian, up to corrections of order $\ln(\ln N)$,

$$X_{\max}(N) \simeq \sqrt{2S \ln N}. \tag{3}$$

b.) Use (3) to estimate the expected fitness $f_{\rm loc}(L)$ of a local fitness maximum and the corresponding fitness $f_{\rm glob}(L)$ of the global fitness maximum in the HoC model with Gaussian fitness values. Show that $f_{\rm loc}(L)/f_{\rm glob}(L) \to 0$ for large L. This is Kauffman's complexity catastrophe: Local evolutionary searches get stuck at local fitness maxima, whose fitness falls far below that of the "true" global optimum. Could the catastrophe be avoided by choosing another fitness distribution $p(f)$?

Problem 5: Fitness maxima in the block model

The block model is a special case of Kauffman's NK-model, where the L loci are subdivided into B disjoint, non-interacting sets of $k = K+1$ loci each and fitness values are i.i.d. random variables within each set (for this L has to be an integer multiple of k, $L = Bk$). The fitness of a genotype can thus be written as

$$f(\sigma) = \sum_{b=1}^{B} f_b(\sigma_{(b-1)k+1}, \sigma_{(b-1)k+2}, ..., \sigma_{bk})$$

where each f_b is a HoC landscape on the k-dimensional hypercube.

a.) Let $n_{\max}$ denote the number of fitness maxima in a given realization of the block model landscape and $n_{\max}^{(b)}$ the number of maxima of the b'th block landscape f_b. Argue that

$$n_{\max} = \prod_{b=1}^{B} n_{\max}^{(b)} \tag{4}$$

where the $n_{\max}^{(b)}$ are independent.

b.) Using (4) and the known result for the HoC model, derive an expression for the expected number of maxima for the block model. Show that it behaves as $\mathbb{E}(n_{\max}) \sim (2\lambda_k)^L$ for large L and fixed k, and determine the constant λ_k. Next consider the limit $L, k \to \infty$ at fixed $\gamma = k/L$, and show that $\mathbb{E}(n_{\max})/2^L \sim L^{-1/\gamma}$.

c.) Using the results of Problem 3 and Eq. (4), derive an expression for the variance $\mathrm{Var}(n_{\max})$ for the block model. Determine the *coefficient of variation* defined as

$$C_V(n_{\max}) = \frac{\sqrt{\mathrm{Var}(n_{\max})}}{\mathbb{E}(n_{\max})}$$

which is a dimensionless measure of the width of the distribution of $n_{\max}$. Compare the asymptotic behavior of C_V for large L for (i) the HoC-model corresponding to $B = 1$, $k = L$ and (ii) the block model with fixed k.

Problem 6: Rough Mount Fuji model

In the Rough Mount Fuji (RMF) model the fitness of a genotype is defined by

$$f(\sigma) = -cd(\sigma, \sigma^*) + \xi(\sigma),$$

where $c > 0$ is a parameter, the random fitness components $\xi(\sigma)$ are i.i.d. random variables with probability density $p(\xi)$, and $\sigma^* = (1, 1, 1, ..., 1)$ is a reference sequence assumed here to consist of all 1's. Thus $d(\sigma, \sigma^*)$ is the number of -1's in the sequence σ.

a.) The expected number of maxima for the RMF-model can be shown to be given by the expression

$$\mathbb{E}(n_{\max}) = \int dx\, p(x)[P(x + c) + P(x - c)]^L \tag{5}$$

where $P(f) = \int^f dx\, p(x)$ is the cumulative probability distribution function of the random fitness component, and the integral is taken over the support of p. Show that (5) reduces to the known result for the HoC model for $c = 0$. Also argue that $\mathbb{E}(n_{\max}) \to 1$ for $c \to \pm\infty$. Why is (5) an even function of c?

b.) Evaluate (5) for the case when $p(f)$ is the uniform distribution on the unit interval $[0, 1]$. Show that $\mathbb{E}(n_{\max}) \sim \lambda(c)^L$ for large L, and determine the function $\lambda(c)$. *Hint:* Note that the cases $c < 1/2$ and $c > 1/2$ have to be considered separately. What happens when $c > 1$?

Problem 7: Random adaptive walk

We consider the random adaptive walk on the L-dimensional hypercube where the fitness values are are drawn independently from the uniform distribution on [0,1]. We are interested in the length of the adaptive path until the walk terminates at a local fitness maximum.

a.) Let $p_\ell(f)$ be the probability density that an adaptive walk contains *at least* ℓ steps, and has fitness f after these ℓ steps. Then, argue that this quantity satisfies the following recursion relation for large $L \gg 1$,

$$p_{\ell+1}(f) = \int_0^f \frac{df'}{1 - f'}(1 - f'^{L-1})p_\ell(f'). \tag{6}$$

b.) Introducing the function

$$H_L(f) = \sum_{k=1}^{L} \frac{1}{k} f^k,$$

show that a change of variables to $H = H_{L-1}(f)$ in (6) gives

$$p_{\ell+1}(H) = \int_0^H dH' p_\ell(H').$$

Together with the initial condition $p_{-1}(f) = \delta(f)$, which means that the walk starts from the lowest possible fitness, show that

$$p_\ell(f) = \frac{1}{\ell!} H_{L-1}(f)^\ell.$$

c.) Now, consider the probability Q_ℓ that an adaptive walk contains *exactly* ℓ steps. By definition this satisfies

$$Q_\ell \equiv \int_0^1 df\, (p_\ell(f) - p_{\ell+1}(f)).$$

By integrating (6) over f from 0 to 1, show that $Q_\ell = \int_0^1 df\, f^{L-1} p_\ell(f)$.

d.) Finally, show that the total probability Q_ℓ converges to the Poisson distribution for large L , i.e.,

$$Q_\ell \to \frac{(\ln L)^\ell}{L\ell!}$$

and thus the first moment is given by $\langle \ell \rangle \equiv \sum_{\ell=0}^{\infty} \ell Q_\ell = \ln L$.
Hint: Consider a change of variable $f = e^{-\frac{x}{L}}$ and only keep the leading order in L from the integrand. You will need to consider the series expansion of the logarithm, i.e., $\lim_{L\to\infty} H_L(f) = -\ln(1-f)$ for large L.