Accessibility percolation with backsteps

Julien Berestycki, Éric Brunet, Zhan Shi*

January 28, 2014

Abstract

Consider a graph in which each site is endowed with a value called *fitness*. A path in the graph is said to be "open" or "accessible" if the fitness values along that path is strictly increasing. We say that there is accessibility percolation between two sites when such a path between them exists. Motivated by the so called House-of-Cards model from evolutionary biology, we consider this question on the L-hypercube $\{0,1\}^L$ where the fitness values are independent random variables. We show that, in the large L limit, the probability that an accessible path exists from an arbitrary starting point to the (random) fittest site is no more than $x^*_{1/2} = 1 - \frac{1}{2}\sinh^{-1}(2) = 0.27818\ldots$ and we conjecture that this probability does converge to $x^*_{1/2}$. More precisely, there is a phase transition on the value of the fitness x of the starting site: assuming that the fitnesses are uniform in $[0,1]$, we show that, in the large L limit, there is almost surely no path to the fittest site if $x > x^*_{1/2}$ and we conjecture that there are almost surely many paths if $x < x^*_{1/2}$. If one conditions on the fittest site to be on the opposite corner of the starting site rather than being randomly chosen, the picture remains the same but with the critical point being now $x^*_1 = 1 - \sinh^{-1}(1) = 0.11863\ldots$. Along the way, we obtain a large L estimation for the number of self-avoiding paths joining two opposite corners of the L-hypercube.

2000 Mathematics Subject Classification: Primary 60J80; Secondary 60G18
Keywords: Evolutionary biology, percolation, trees, branching processes

1 Introduction

1.1 Definition of the model

We consider the following mathematical model inspired by evolutionary biology:

1. The genome of an organism is made of L sites which can each be in two states (or alleles): 0 (the wild state) or 1 (the mutant state). There are therefore 2^L possible genomes, which are coded as an L-bit binary word or as a corner of the L-hypercube [6].

2. During reproduction (supposed asexual and without recombination), we assume that the only mutations that can occur consist in changing the state at one single site, either from the wild to the mutant state or from the mutant to the wild state. With our representation, a mutation is flipping one single bit in the L-bit word or traveling along one edge of the L-hypercube [5, 6].

*emails: Julien.Berestycki@upmc.fr, Zhan.Shi@upmc.fr, EricBrunet@lps.ens.fr.
J.B. and Z.S.: Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7599, LPMA, F-75005, Paris France
É.B.: Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 8550, LPS-ENS, F-75005, Paris France

3. We assume that we are in a regime with a low mutation rate, high selection and a population which is not too large. In this regime, when a mutation occurs, the new genome either fixates (*i.e.* it invades the whole population and becomes the new resident type) if its fitness value is better than the value of the resident population, or is eliminated if it is lower. This happens (in the regime we assume) fast enough that a new mutation has no time to appear before the population is homogeneous again.

In this model, the evolutionary history of the population as a whole can be described as a path along the edges of the L-hypercube, with the constraint that the fitness value must increase at each step. We call such paths "open" or "selectively accessible" [3, 12, 13]. We emphasize that we allow paths of arbitrary length, where bits can flip from 1 to 0 as well as from 0 to 1. The question we wish to address is the following: assuming that the population is initially in the state $(0, 0, \ldots, 0)$, is there an evolutionary path allowing it to evolve to the fittest site available?

To answer this question we need a model for the fitness values of each site. As a first approach, we consider the House-of-Cards [7] model (which is equivalent [1] to the NK model [6] with $K = N - 1$) where the fitness values of the 2^L sites are independent random numbers. For the purpose of discussing the existence of open paths, the actual fitness values of each site are not relevant; the only useful information are how the fitness values are ordered. This means that the answer to our question does not depend on the chosen distribution of the fitness values, and that we can safely choose the most convenient distribution:

4. The fitness value of the fittest site is 1 and that the fitness values of the other $2^L - 1$ sites are independent random numbers chosen uniformly between 0 and 1.

As we explained, this is equivalent to the House-of-Cards model if, furthermore, the fittest site is chosen uniformly at random amongst the 2^L sites of the hypercube. In this paper, we first consider the case where the fittest site is deterministically chosen to be $(1, 1, \ldots, 1)$, then the case where it is a fixed arbitrary site σ_{fittest} and, finally, the case where the fittest site is random.

1.2 Notations

- Sites are coded as L bit binary word. The initial state of the population is $(0, 0, \ldots, 0)$.
- H (as in "Hamming distance") is the number of bits set to 1 in the fittest site σ_{fittest}.
- The fitness of the starting site $(0, 0, \ldots, 0)$ is noted x.
- Θ is the number of open (selectively accessible) paths from $(0, 0, \ldots, 0)$ to the fittest site. To compare with previous results, we also write as $\tilde{\Theta}$ the number of open paths of minimal length to the fittest site.
- The probability of an event is written $\mathbb{P}$ and its expectation $\mathbb{E}$. We often need to condition on the value x of the fitness in the starting site; when we do we write the conditional probability and expectation as $\mathbb{P}^x$ and $\mathbb{E}^x$. The values of H and L are implicit in the notation.

1.3 Results of previous works

Similar models have been studied by several groups in the past few years, either directly on the hypercube as above [2, 4, 9, 11] or in the geometrically simpler setting of a tree [2, 10, 8].

Except for [9], all the previous studies focused only on the number $\tilde{\Theta}$ of open paths going from the starting position $(0, 0, \ldots, 0)$ to the opposite corner $(1, 1, \ldots, 1)$ with minimal length,

meaning that a mutation can only flip a bit from 0 to 1 and not the other way around. In this setting, one only needs to consider $H = L$ as direct paths to a fittest site at Hamming distance H cannot leave anyway the H-hypercube .

There are $L!$ minimal length paths (open or not) connecting the starting site $(0, 0, \ldots, 0)$ to the opposite corner $(1, 1, \ldots, 1)$, each of these minimal length paths go through L random fitnesses between 0 and 1 (including the starting site, but excluding the end site which is assumed to have fitness 1) and the probability that a given minimal length path is open is the probability that these L random numbers are in order, which is $1/L!$. Therefore

$$\mathbb{E}(\tilde{\Theta}) = 1. \tag{1}$$

This expectation is however misleading, as the typical number of open minimal length paths is not 1. Indeed, if one conditions on the fitness x of the starting site, the probability that a given minimal length path is open is $(1-x)^{L-1}/(L-1)!$ because the path meets $L-1$ random values (excluding both the starting and end sites) and these values must be all between x and 1 and in ascending order. Therefore

$$\mathbb{E}^x(\tilde{\Theta}) = L(1-x)^{L-1}. \tag{2}$$

This conditional expectation is a decreasing function of x which is equal to 1 for $x = x_c(L)$ with $x_c(L) := 1 - \exp[-\frac{\ln L}{L-1}] \sim \frac{\ln L}{L}$ for large L. This implies that

$$\mathbb{P}(\tilde{\Theta} \geq 1) \leq x_c(L) + \int_{x_c(L)}^{1} \mathbb{E}^x(\tilde{\Theta}) = x_c(L) + (1 - x_c(L))^L \sim \frac{\ln L}{L} \qquad \text{as } L \to \infty. \tag{3}$$

By a clever second moment argument, Hegarty and Martinsson [4] proved that the above bound is tight:

$$\mathbb{P}(\tilde{\Theta} \geq 1) \sim \frac{\ln L}{L} \qquad \text{as } L \to \infty. \tag{4}$$

More precisely, they show that if $a(L)$ is a positive diverging function of L, then

$$\mathbb{P}^{\frac{\ln(L)+a(L)}{L}}(\tilde{\Theta} \geq 1) \to 0, \qquad \mathbb{P}^{\frac{\ln(L)-a(L)}{L}}(\tilde{\Theta} \geq 1) \to 1. \tag{5}$$

In [2], we showed that there were of order of L open minimal length paths when the starting position x is of order $1/L$ and we gave the limiting law of $\tilde{\Theta}/L$.

Informally, (5) means that if the starting fitness x is larger than $(\ln L)/L$, then there are no open minimal length path, and if x is smaller than $(\ln L)/L$, then there are some open minimal length paths. Even more informally, the expectation (2) tells the truth: when the expectation goes to zero, there are no path (which is obvious); when the expectation diverges, there are some paths (which is not automatic).

1.4 Our results

In this paper, we consider paths which are no longer of minimal length: a mutation can change a 1 into a 0 as well as a 0 into a 1. We compute bounds for the expected number of open paths connecting $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$ given the starting fitness x which lead to

Theorem 1 *When $H = L$ (that is, when the fittest site is $(1, 1, \ldots, 1)$),*

$$\left[\mathbb{E}^x(\Theta)\right]^{1/L} \to \sinh(1-x) \qquad \text{as } L \to \infty.$$

In particular there is a critical value x_1^* for the fitness of the starting position,

$$x_1^* = 1 - \sinh^{-1}(1) = 1 - \ln(\sqrt{2}+1) = 0.11863\ldots,$$

such that

- *For $x > x_1^*$, $\mathbb{E}^x(\Theta)$ goes to zero exponentially fast as $L \to \infty$ and, therefore $\mathbb{P}^x(\Theta \geq 1) \to 0$.*
- *For $x < x_1^*$, $\mathbb{E}^x(\Theta)$ diverges exponentially fast as $L \to \infty$.*

As a consequence,

$$\limsup_{L\to\infty} \mathbb{P}(\Theta \geq 1) \leq x_1^*. \tag{6}$$

We conjecture that the expectation "tells the truth" and that:

Conjecture 1 *when $H = L$, for $x < x_1^*$,*

$$\mathbb{P}^x(\Theta \geq 1) \to 1 \qquad \text{as } L \to \infty.$$

and, as a consequence,

$$\lim_{L\to\infty} \mathbb{P}(\Theta \geq 1) = x_1^*.$$

As an illustration, Figure 1 shows the result of numerical simulations measuring the probability $\mathbb{P}^x(\Theta \geq 1)$ that there are some open paths on the L-hypercube as a function of x for different values of L. Our theorem is that, for large L, the probability goes to zero on the right of the black line and our conjecture is that it goes to 1 on the left. One might guess such a scenario from this picture alone, with, however, a critical value around 0.15 rather than the actual $x_1^* \approx 0.12$. Our work proves however that the critical value cannot be larger than x_1^*.

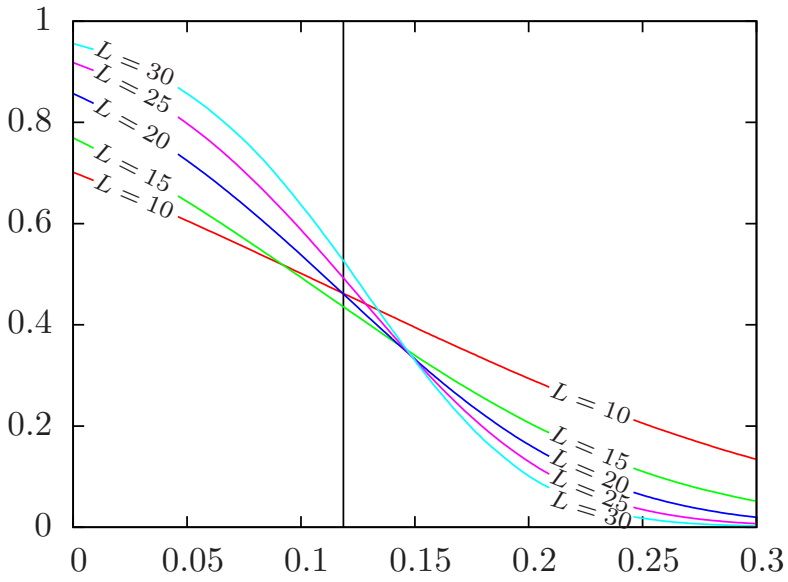

Figure 1: Probability to have an open path from $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$ in the L-hypercube as a function of the starting fitness x. The critical x_1^* is represented by the black vertical line. These curves were obtained by Monte-Carlo simulation on 10^6 samples per size (only 10^5 for the largest size).

When the fittest site is not $(1, 1, \ldots, 1)$, we have the following more general result:

Theorem 2 *Let $\alpha \in [0,1]$ and consider a function $L \mapsto H(L)$ such that $H(L)/L \to \alpha$ as $L \to \infty$. Then, when $\sigma_{fittest}$ is chosen such that its Hamming distance is $H = H(L)$,*

$$\left[\mathbb{E}^x(\Theta)\right]^{1/L} \to \sinh(1-x)^\alpha \cosh(1-x)^{1-\alpha} \qquad \text{as } L \to \infty.$$

In particular, for each α, there is a critical value x^*_α for the fitness of the starting position, which is the unique solution of

$$\sinh(1-x^*_\alpha)^\alpha \cosh(1-x^*_\alpha)^{1-\alpha} = 1,$$

such that

- *For $x > x^*_\alpha$, $\mathbb{E}^x(\Theta)$ goes to zero exponentially fast as $L \to \infty$ and, therefore $\mathbb{P}^x(\Theta \geq 1) \to 0$.*
- *For $x < x^*_\alpha$, $\mathbb{E}^x(\Theta)$ diverges exponentially fast as $L \to \infty$.*

As a consequence,

$$\limsup_{L\to\infty} \mathbb{P}(\Theta \geq 1) \leq x^*_\alpha. \tag{7}$$

We conjecture again that the expectation "tells the truth" and that:

Conjecture 2 *when $H = H(L)$ with $H(L)/L \to \alpha$, for $x < x^*_\alpha$,*

$$\mathbb{P}^x(\Theta \geq 1) \to 1 \qquad \text{as } L \to \infty.$$

and, as a consequence,

$$\lim_{L\to\infty} \mathbb{P}(\Theta \geq 1) = x^*_\alpha.$$

Figure 2 gives the critical value x^*_α as a function of α. Noteworthy points are $x^*_1 = 1 - \sinh^{-1}(1) = 0.118626\ldots$ as already noted and $x^*_{1/2} = 1 - \frac{1}{2}\sinh^{-1}(2) = 1 - \frac{1}{2}\ln(2+\sqrt{5}) = 0.278182\ldots$.

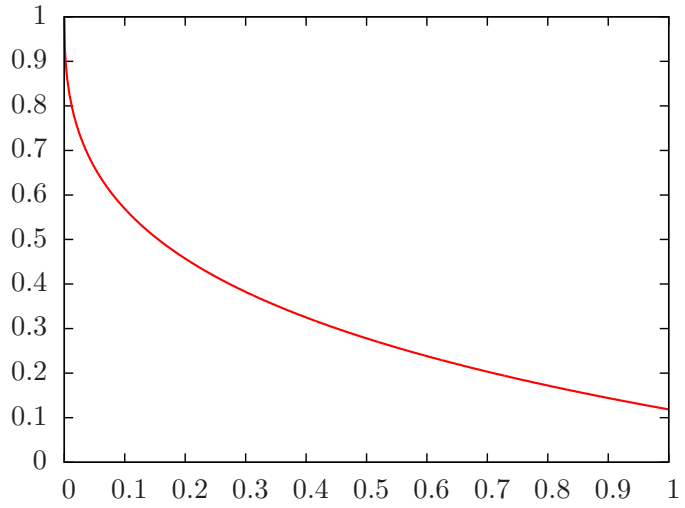

Figure 2: The critical point x^*_α as a function of α.

Finally, when the fittest site is chosen uniformly at random (this is the model which is truly equivalent to the House-of-Cards model) it is clear that for large L the value of H/L converges to $1/2$. This leads to the following result:

Theorem 3 *When the fittest site is chosen uniformly at random, one has*

- *For* $x > x^*_{1/2}$, $\mathbb{P}^x(\Theta \geq 1) \to 0$ *as* $L \to \infty$.

As a consequence,

$$\lim_{L\to\infty} \mathbb{P}(\Theta \geq 1) \leq x^*_{1/2}. \tag{8}$$

Furthermore, if one assumes Conjecture 2,

- *For* $x < x^*_{1/2}$, $\mathbb{P}^x(\Theta \geq 1) \to 1$ *as* $L \to \infty$*; therefore* $\mathbb{P}(\Theta \geq 1) \to x^*_{1/2}$.

As a by-product of this work, we also found that the number of self-avoiding paths (just plain paths, without any notion of fitness, openness or accessibility) joining $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$ in the L-hypercube grows as a double exponential, see Theorem 4 at the end of Section 2.1.

Theorem 1 if proved in Section 2, Theorem 4 in Section 3, Theorem 2 in Section 4 and Theorem 3 in Section 5.

2 Proof when the fittest site is $(1, 1, \ldots, 1)$

We consider here the case $H = L$, *i.e.* when the fittest site, the one with a fitness equal to 1, is $(1, 1, \ldots, 1)$. The generalization to an arbitrary fittest site is described in Section 4.

The minimum length of a path from $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$ is L, as each one of the L bits has to be switched from 0 to 1. There exists however longer paths which have backsteps, *i.e.* steps where a bit is flipped from 1 to 0. The length of a path with p backsteps is clearly $L + 2p$ as each backstep must be compensated by an extra forward step.

We only need to consider paths that do not self-intersect, as it is obvious that a path going twice to the same site cannot see its fitness increase strictly. We define

$$\begin{aligned} a_L &= \text{the number of self-avoiding paths connecting } (0, 0, \ldots, 0) \text{ to } (1, 1, \ldots, 1), \\ a_{L,p} &= \text{the number of self-avoiding paths connecting } (0, 0, \ldots, 0) \text{ to } (1, 1, \ldots, 1) \\ &\qquad \text{with a length } L + 2p \text{ (that is, with } p \text{ backsteps).} \end{aligned} \tag{9}$$

As an illustration, Figure 3 shows all self-avoiding paths on the 3-hypercube which begin by "right, up": there are three of them with respective lengths 3 ($p = 0$), 5 ($p = 1$) and 7 ($p = 2$). But there are 6 choices for the first two steps so, by symmetry, $a_{3,0} = 6$, $a_{3,1} = 6$, $a_{3,2} = 6$ and, of course, $a_3 = 18$.

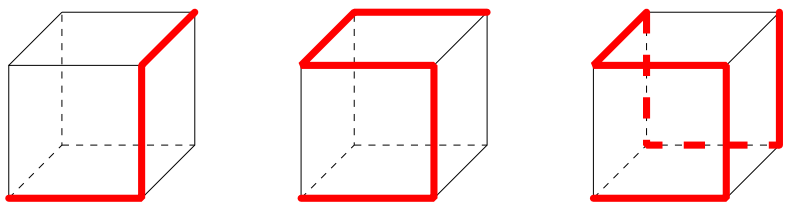

Figure 3: The three self-avoiding paths on the cube connecting $(0, 0, \ldots, 0)$ =(bottom, left, front) to the opposite corner $(1, 1, \ldots, 1)$ =(top, right, back) which begin by (right, up).

When the starting site has a fixed fitness x, the probability that a self-avoiding path of length $L + 2p$ is open is simply (as in the introduction) $(1 - x)^{L+2p-1}/(L + 2p - 1)!$: the $L + 2p - 1$

interior sites must be between x and 1 and they must be in order. Thus, the expected number of open paths when the starting value is x is

$$\mathbb{E}^x(\Theta) = \sum_{p\geq 0} a_{L,p} \frac{(1-x)^{L+2p-1}}{(L+2p-1)!}. \tag{10}$$

The problem therefore reduces to finding good estimates on $a_{L,p}$.

2.1 Some remarks about $a_{L,p}$

There seems to be little literature on the number a_L of self-avoiding paths on a L-hypercube joining $(0,0,\ldots,0)$ to $(1,1,\ldots,1)$. The sequence is referenced in the Online Encyclopedia of Integer Sequences [14] and values are given up to $L=5$:

$$\begin{gathered} a_1 = 1, \quad a_2 = 2, \quad a_3 = 3! \times 3 = 18, \\ a_4 = 4! \times 268 = 6\,432, \quad a_5 = 5! \times 155\,429\,607 = 18\,651\,552\,840. \end{gathered} \tag{11}$$

Furthermore, the value for a_5 is mentioned in an evolutionary biology paper [9]. These numbers were obtained by brute force computer enumeration; due to the combinatorial explosion, a_6 is out of reach by this method.

For a given L, the length of a self-avoiding path cannot exceed the number $2^L - 1$ of sites to explore; hence one must have $L + 2p \leq 2^L - 1$. Just to give a flavour of the structure of paths, here are all the numbers $a_{5,p}$:

$$\begin{gathered} a_{5,0} = 5!, \quad a_{5,1} = 5! \times 10, \quad a_{5,2} = 5! \times 107, \quad a_{5,3} = 5! \times 1\,097, \quad a_{5,4} = 5! \times 9\,754, \\ a_{5,5} = 5! \times 72\,305, \quad a_{5,6} = 5! \times 448\,536, \quad a_{5,7} = 5! \times 2\,243\,671, \\ a_{5,8} = 5! \times 8\,631\,118, \quad a_{5,9} = 5! \times 24\,044\,702, \quad a_{5,10} = 5! \times 44\,617\,008, \\ a_{5,11} = 5! \times 48\,280\,086, \quad a_{5,12} = 5! \times 24\,000\,420, \quad a_{5,13} = 5! \times 3\,080\,792. \end{gathered} \tag{12}$$

It is clear that $a_{L,p}$ must be a multiple of $L!$ as from a given path more paths can be built by simply applying a permutation of the L directions of the edges. For $p=0$, one has obviously $a_{L,0} = L!$. For $p=1$, the paths have length $L+2$ with one backstep at some position k with $3 \leq k \leq L$. (The first step cannot be a backstep as all the bits are still 0. The second step cannot be a backstep as it would bring the path back to the starting position. The steps $L+1$ and $L+2$ cannot be the backstep as the system would have already reached the end point at step L.) For a given backstep position k, there are $L!/(L-k+1)!$ possible paths up to step $k-1$, and $k-2$ possible choice for the backstep (because there are $k-1$ bits set to 1 but one cannot choose the bit that was just set), and $L-k+1$ choices for the step following the backstep (because there are $L-k+2$ bits set to 0 but one cannot choose the bit that was just unset) and $(L-k+1)!$ choices for all the subsequent steps. The number of paths with a given backstep position k is then $L! \times (k-2)(L-k+1)$; summing over all k gives

$$a_{L,1} = L! \times \frac{L(L-1)(L-2)}{6}. \tag{13}$$

A similar (but much more strenuous) derivation leads for $p=2$ to

$$a_{L,2} = L! \times \frac{(L-1)(L-2)(5L^4+3L^3+34L^2-264L+180)}{360}. \tag{14}$$

It is easy to convince oneself that for fixed p, as $L \to \infty$,

$$a_{L,p} \sim L! \frac{L^{3p}}{6^p p!}. \tag{15}$$

Indeed, one needs to choose p backsteps at positions $3 \leq k_1 < k_2 < \cdots < k_p \leq L + 2p - 2$ in a sequence of $L+2p \approx L$ steps (we are dropping all the non-dominant terms). At the j-th backstep there are of order k_j choices to choose the bit we set to 0 (actually: $k_j - 2j$ choices if the previous step was not a backstep, but $k_j \propto L$ and $j \leq p$ and we are dropping all the non-dominant terms). The step after backstep j has to leading order $L - k_j$ bits 0 which can be switched to 1, and all the other steps combine to build $L!$. With the k_j given, one therefore gets a number of paths of order $L! \times k_1(L - k_1) \times k_2(L - k_2) \times \cdots \times k_p(L - k_p)$. Summing over the ordered k_j's leads to (15). The expression for fixed k_j's is far from being correct if there are backsteps too close to each other but all of this only contribute to the next order term in the expression.

It would be extremely interesting to understand better how a_L grows with L. In this work, we give upper and lower bounds for $a_{L,p}$ to study the problem defined in the introduction which, as a by-product, also lead to the following theorem proved in Section 3.

Theorem 4 *Recall that a_L is the number of self-avoiding paths on the L-hypercube from $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$. Then*

$$\lim_{L \to \infty} \frac{\ln \ln a_L}{L} = \ln 2.$$

More precisely, there exists two positive constants c and c' such that, for L large enough,

$$c \leq \frac{\ln a_L}{2^L} \leq c' \ln L.$$

2.2 Coding of a path

We use the following representation for a path on the hypercube: it is a string of numbers between 1 and L where each number indicates the position of the bit being flipped by the corresponding step. The first time a particular number is met, the bit is flipped from 0 to 1; the next time from 1 to 0, etc. To take an example, the paths of Figure 3 would be respectively coded "123", "12131" and "1213212" assuming that 1 is the left/right direction, 2 is the up/down direction and 3 is the front/back direction.

Clearly, for paths going from $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$, each number must appear an odd number of times in the string. A path visits twice the same site if there exists a non-empty substring[1] of the path where each number appears an even number of times (including zero, of course), as all the bits encoding the position are clearly the same before and after the substring. Examples of minimal forbidden substrings include "11", "1212", "12313424", etc. A self-avoiding path is then, of course, such that there is no such substring.

2.3 Upper bound

In this section we prove the following:

Lemma 1 *When $H = L$,*

$$\mathbb{E}^x(\Theta) \leq L \sinh(1 - x)^L \operatorname{cotanh}(1 - x).$$

[1] We recall that a substring is a subsequence of consecutive terms.

This implies that $\limsup_{L\to\infty}\left[\mathbb{E}^x(\Theta)\right]^{1/L} \leq \sinh(1-x)$, which is the first half of the proof of Theorem 1.

Let

$$M_{L,p} = \text{the set of paths connecting } (0,0,\ldots,0) \text{ to } (1,1,\ldots,1) \text{ with a length } L+2p \text{ (that is, with } p \text{ backsteps) where intersections are allowed.} \tag{16}$$

Clearly,

$$a_{L,p} \leq M_{L,p}, \tag{17}$$

(by an abuse of notation, the cardinal of $M_{L,p}$ is also noted $M_{L,p}$) and it turns out that this very simple upper bound is sufficient for our purpose.

With our representation, $M_{L,p}$ is the set of strings with length $L+2p$ made of numbers between 1 and L where each number appears an odd number of times. We build $M_{L,p}$ by recurrence: for any p, there is one path in $M_{1,p}$: it is the path "111..." where one walks back and forth between the two sites of the 1-hypercube.

To construct a path in $M_{L+1,p}$ (with length $L+1+2p$), we

- choose how many times the number $L+1$ appears. This number is odd, let it be $2q+1$ with $0 \leq q \leq p$,
- choose the positions of the $2q+1$ numbers $L+1$ amongst the $L+1+2p$ possible positions in the string,
- fill in the remaining $L+2p-2q$ positions with the string coding an arbitrary path chosen in $M_{L,p-q}$.

In equations, this construction gives

$$M_{1,p} = 1, \qquad M_{L+1,p} = \sum_{q=0}^{p} \binom{L+1+2p}{2q+1} M_{L,p-q}. \tag{18}$$

Writing the binomial with factorials one gets

$$\frac{M_{L+1,p}}{(L+1+2p)!} = \sum_{q=0}^{p} \frac{M_{L,p-q}}{(L+2(p-q))!} \times \frac{1}{(2q+1)!}. \tag{19}$$

Let $G_L(X)$ be the generating function defined by

$$G_L(X) := \sum_{p\geq 0} \frac{M_{L,p}}{(L+2p)!} X^{L+2p}. \tag{20}$$

Notice that from (10) and (17), one has

$$\mathbb{E}^x(\Theta) \leq G_L'(1-x). \tag{21}$$

The recurrence on $M_{L,p}$ translates into a recurrence on $G_L(X)$:

$$\begin{aligned} G_{L+1}(X) &= \sum_{p\geq 0}\sum_{q=0}^{p} \frac{M_{L,p-q}}{(L+2(p-q))!} X^{L+2(p-q)} \times \frac{X^{2q+1}}{(2q+1)!}, \\ &= \sum_{q\geq 0}\sum_{p\geq q} \frac{M_{L,p-q}}{(L+2(p-q))!} X^{L+2(p-q)} \times \frac{X^{2q+1}}{(2q+1)!}, \\ &= \sum_{q\geq 0}\sum_{p\geq 0} \frac{M_{L,p}}{(L+2p)!} X^{L+2p} \times \frac{X^{2q+1}}{(2q+1)!}, \\ &= G_L(X)\sinh(X). \end{aligned} \tag{22}$$

But $G_1(X) = \sinh(X)$, hence

$$G_L(X) = \sinh(X)^L \tag{23}$$

which, with (21), concludes the proof of Lemma 1.

2.4 Lower bound

In this section we proove the second half of Theorem 1:

Lemma 2 *When $H = L$,*

$$\limsup_{L\to\infty} \left[\mathbb{E}^x(\Theta)\right]^{1/L} \geq \sinh(1-x).$$

For this lower bound, we construct a subset m_L of all the self-avoiding paths on the L-hypercube joining $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$. The construction is recursive:

- For $L = 1$, there is only one self-avoiding path joining the two corners of the 1-hypercube. In our coding, this path is represented by the string "1".
- A path (coded by a string) is in m_{L+1} if (a) the number $L + 1$ appears an odd number of times in the string, (b) the number $L + 1$ never appears at two consecutive positions, and (c) the string where all the number $L + 1$ are removed codes a path in m_L.

It is clear by recurrence that paths in m_L are valid self-avoiding paths on the L-hypercube. To illustrate, $m_1 = \{$"1"$\}$, $m_2 = \{$"12", "21"$\}$, $m_3 = \{$"123", "132", "312", "213", "231", "321", "31323", "32132"$\}$. Notice how we lost symmetry in the paths: "31323" is in m_3 but not "13121".

We now define

$$m_{L,p} = \text{the set of paths of length } L + 2p \text{ in } m_L. \tag{24}$$

Clearly

$$m_{L,p} \leq a_{L,p}, \tag{25}$$

(once again, by an abuse of notation we write the cardinal of $m_{L,p}$ also as $m_{L,p}$).

One has $m_{1,0} = \{$"1"$\}$ and $m_{1,p} = \emptyset$ for $p > 0$. The sets $m_{L,p}$ are then built recursively: to construct a path in $m_{L+1,p}$, we

- choose how many times the number $L + 1$ appears. This number is odd, let it be $2q + 1$ with $0 \leq q \leq p$,
- choose the positions of the $2q + 1$ numbers $L + 1$ amongst the $L + 1 + 2p$ possible positions in the string in such a way that there are no two consecutive numbers $L + 1$,
- fill in the remaining $L + 2(p - q)$ positions with the string coding an arbitrary path chosen in $m_{L,p-q}$.

Let us recall that the number of ways of choosing P items out of a sequence of N such that two consecutive items in the sequence cannot be both chosen is $\binom{N-P+1}{P}$. Indeed, each configuration can be bijectively obtained by first choosing P items out of a sequence of $N - P + 1$ and then expanding the sequence to size N by inserting one unchosen item "$\cdot$" before each chosen item "$\bullet$" except the first one; for instance, with $P = 4$ and $N = 11$: $(\bullet\cdot\bullet\bullet\cdot\cdot\bullet\cdot \to \bullet\cdot\cdot\bullet\cdot\bullet\cdot\cdot\cdot\bullet\cdot)$.

With this result, the construction of $m_{L,p}$ leads to

$$m_{1,p} = \mathbb{1}_{p=0}, \qquad m_{L+1,p} = \sum_{q=0}^{p} \binom{L+1+2p-2q}{2q+1} m_{L,p-q}. \tag{26}$$

In order to write a generating function similar to the $G_L(X)$ defined in the previous section, we need to replace the binomial in (26) by the same binomial as in (18). Let us write, for $0 \le q \le p$,

$$\begin{aligned}
\binom{L+1+2p-2q}{2q+1} &= \frac{1}{(2q+1)!} \times \quad (L+1+2p-2q)! \quad \times \frac{1}{(L+2p-4q)!} \\
&= \frac{1}{(2q+1)!} \times \frac{(L+1+2p)!}{\prod_{k=0}^{2q-1}(L+1+2p-k)} \times \frac{\prod_{k=0}^{2q-1}(L+2p-2q-k)}{(L+2p-2q)!} \\
&= \binom{L+1+2p}{2q+1} \times \prod_{k=0}^{2q-1} \frac{L+2p-2q-k}{L+1+2p-k} \qquad (27) \\
&= \binom{L+1+2p}{2q+1} \times \prod_{k=0}^{2q-1} \Big[1 - \frac{2q+1}{L+1+2p-k}\Big] \\
&\ge \binom{L+1+2p}{2q+1} \times \Big[1 - \frac{2q+1}{L+2}\Big]^{2q} \mathbb{1}_{2q<L+1} \qquad (\text{using } p \ge q).
\end{aligned}$$

Then, defining $\tilde{m}_{L,p}$ by the recurrence

$$\tilde{m}_{1,p} = m_{1,p} = \mathbb{1}_{p=0}, \qquad \tilde{m}_{L+1,p} = \sum_{q=0}^{p} \tilde{m}_{L,p-q} \binom{L+1+2p}{2q+1} \times \Big[1 - \frac{2q+1}{L+2}\Big]^{2q} \mathbb{1}_{2q<L+1} \quad (28)$$

it is clear that

$$\tilde{m}_{L,p} \le m_{L,p} \le a_{L,p}. \qquad (29)$$

As for the upper bound, let $g_L(X)$ be the generating funcion of the $\tilde{m}_{L,p}$ defined by the finite sum

$$g_L(X) := \sum_{p \ge 0} \frac{\tilde{m}_{L,p}}{(L+2p)!} X^{L+2p}. \qquad (30)$$

Notice that from (10) and (29), one has

$$g'_L(1-x) \le \mathbb{E}^x(\Theta). \qquad (31)$$

The recurrence on $\tilde{m}_{L,p}$ translates into a recurrence on $g_L(X)$. One gets easily, by the same argument as in (22)

$$g_{L+1}(X) = g_L(X) \times \sum_{q \ge 0} \frac{X^{2q+1}}{(2q+1)!} \Big[1 - \frac{2q+1}{L+2}\Big]^{2q} \mathbb{1}_{2q<L+1}\,. \qquad (32)$$

Defining

$$\sinh_l(X) := \sum_{q \ge 0} \frac{X^{2q+1}}{(2q+1)!} \Big[1 - \frac{2q+1}{l+1}\Big]^{2q} \mathbb{1}_{2q<l}, \qquad (33)$$

then (32) reads $g_{L+1}(X) = g_L(X) \sinh_{L+1}(X)$. Furthermore, $g_1(X) = X = \sinh_1(X)$ so that

$$g_L(X) = \prod_{l=1}^{L} \sinh_l(X). \qquad (34)$$

The derivative of g_L can then be written

$$g'_L(X) = \Big(\sum_{l=1}^{L} \frac{\sinh'_l(X)}{\sinh_l(X)} \Big) \times g_L(X) \tag{35}$$

It is clear that $X \leq \sinh_l(X) \leq \sinh(X)$ and that $1 \leq \sinh'_l(X) \leq \cosh(X)$. The sum in (35) is therefore bounded between $L/\sinh(X)$ and $L\cosh(X)/X$, and the sum to the power $1/L$ converges to 1 as $L \to \infty$.

Furthermore, by dominated convergence, $\sinh_l(X) \to \sinh(X)$ (and $\sinh'_l(X) \to \cosh(X)$) as $l \to \infty$. This is sufficient to imply that $g_L(X)^{1/L}$ converges by Cesaro to $\sinh(X)$. Finally, $g'_L(X)^{1/L} \to \sinh(X)$ and, with (31),

$$\sinh(1-x) \leq \liminf_{L\to\infty} \big[\mathbb{E}^x(\Theta) \big]^{1/L}, \tag{36}$$

which is the second half of Theorem 1.

3 Proof of Theorem 4

In the previous section, we used the bounds $m_{L,p} \leq a_{L,p} \leq M_{L,p}$ to obtain an estimate on $\mathbb{E}^x(\Theta)$ with (10). We now use the same bounds to obtain an estimate on $a_L = \sum_p a_{L,p}$ and prove Theorem 4.

lower bound We define another generating function of the $m_{L,p}$. Let

$$\phi_L(X) = \sum_{p \geq 0} m_{L,p} X^{L+2p}. \tag{37}$$

Then one finds easily from (26) that

$$\phi_1(X) = X, \qquad \phi_{L+1}(X) = \frac{1+X}{2}\phi_L\big(X+X^2\big) - \frac{1-X}{2}\phi_L\big(X-X^2\big). \tag{38}$$

From its definition $\phi_L(X)$ is a polynomial in X (recall that $a_{L,p}$ and, therefore, $m_{L,p}$ is zero if p is too large). This polynomial is an odd function of X if L is odd and an even function if L is even. Let d_L be the degree of this polynomial. By considering the highest degree term in (38) one gets easily

$$d_1 = 1, \qquad d_{L+1} = \begin{cases} 2d_L & \text{if } L \text{ is odd,} \\ 2d_L + 1 & \text{if } L \text{ is even.} \end{cases} \tag{39}$$

This can be solved into

$$d_L = \begin{cases} \frac{2^{L+1}-1}{3} & \text{if } L \text{ is odd,} \\ \frac{2^{L+1}-2}{3} & \text{if } L \text{ is even.} \end{cases} \tag{40}$$

By definition, $m_L = \sum_p m_{L,p} = \phi_L(1)$. From (38) one has, furthermore,

$$m_L = \phi_L(1) = \phi_{L-1}(2). \tag{41}$$

As the polynomials ϕ_L have non-negative integer coefficients, one clearly have

$$a_L \geq m_L \geq 2^{d_{L-1}} \tag{42}$$

which is enough for the lower bound of Theorem 4.

Upper bound There are infinitely many paths in M_L, but one knows that paths in a_L have a maximum length of 2^L (one could be more precise: at most $2^L - 1$ if L is odd and at most $2^L - 2$ if L is even) so that

$$a_L \leq \sum_p M_{L,p} \, \mathbb{1}_{L+2p \leq 2^L} \, . \tag{43}$$

Let us write for an analytical function f

$$T_n f = \text{The Taylor polynomial of } f \text{ of degree } n. \tag{44}$$

Our upper bound is then

$$a_L \leq \int_0^\infty e^{-X} T_{2^L} G_L(X) \, \mathrm{d}X, \tag{45}$$

as seen from the definition (20) of G_L.

For any absolutely increasing function f (all derivatives are non-negative), any order n and any cutoff point C, one has

$$T_n f(X) \leq \begin{cases} f(X) & \text{if } X \leq C, \\ f(C) \left(\frac{X}{C}\right)^n & \text{if } X \geq C. \end{cases} \tag{46}$$

The first line is trivial as we removed some non-negative terms. The second line is also trivial because $X^k \leq C^k (X/C)^n$ for $X \geq C$ and $k \leq n$, so the inequality holds for all the terms in the polynomial $T_n f$. Then

$$\int_0^\infty e^{-X} T_n f(X) \, \mathrm{d}X \leq f(C) C + f(C) \frac{n!}{C^n}. \tag{47}$$

We apply this to $f(X) = G_L(X) = \sinh(X)^L \leq e^{LX}/2^L$ and to $C = n/L$:

$$\int_0^\infty e^{-X} T_n G_L(X) \, \mathrm{d}X \leq \frac{n e^n}{L 2^L} + \frac{n! e^n}{n^n} \frac{L^n}{2^L}. \tag{48}$$

Remember that by Stirling $e^n n!/n^n \sim \sqrt{2\pi n}$. The second term on the right hand-side is much larger than the first (L^n vs e^n). Replace n by 2^L, and it is easy to check that the bound (45) gives the second half of Theorem 4.

4 Proof for an arbitrary H

We assume now that the fittest site, the one with a fitness equal to 1, is no longer $(1, 1, \ldots, 1)$ but rather an arbitrary given site σ_{fittest}. By symmetry, the accessibility of σ_{fittest} depends only on the number of bits set to 1 in σ_{fittest}; let H (as in "Hamming distance") be this number of bits. To simplify the discussion, we assume that the bits 1 to H in σ_{fittest} are set to 1 and that the bits $H + 1$ to L are set to 0.

We emphasize that we consider paths on the L-hypercube and not on the H-hypercube: valid paths may leave the H-hypercube and then do backsteps to go back to σ_{fittest}. In previous studies where only shortest paths where considered (no backstep), considering the case $\sigma_{\text{fittest}} \neq (1, 1, \ldots, 1)$ was meaningless as it was simply equivalent to changing the dimension of the hypercube.

The minimum length of a path from $(0,0,\dots,0)$ to σ_{fittest} is H. A path with p backsteps has a length $H+2p$. We define

$$a_{L,H,p} = \text{the number of self-avoiding paths connecting } (0,0,\dots,0) \text{ to } \sigma_{\text{fittest}} \\ \text{with a length } H+2p \text{ (that is, with } p \text{ backsteps).} \tag{49}$$

Then, with the same argument as before,

$$\mathbb{E}^x(\Theta) = \sum_{p\geq 0} a_{L,H,p} \frac{(1-x)^{H+2p-1}}{(H+2p-1)!}. \tag{50}$$

With the coding introduced in Section 2.2, a self-avoiding path from $(0,0,\dots,0)$ to σ_{fittest} is a string of numbers between 1 and L such that (a) the numbers between 1 and H appear an odd number of times, (b) the numbers between $H+1$ and L appear an even number of times (including zero), (c) in any non-empty substring, there must be at least one number which appears an odd number of times.

Using the same strategy as in the previous section, we bound $a_{L,H,p}$:

$$m_{L,H,p} \leq a_{L,H,p} \leq M_{L,H,p}, \tag{51}$$

where

- $M_{L,H,p}$ is the number (or the set) of paths on the L-hypercube of length $H+2p$ from $(0,0,\dots,0)$ to σ_{fittest} where intersections are authorized. Naturally, $M_{H,H,p} = M_{H,p}$ (as defined in Section 2.3). For $L \geq H$, we obtain recursively $M_{L+1,H,p}$ as the number of strings of length $H+2p$ such that $L+1$ appears an even number of times $2q$ *at arbitrary positions* and such that if one removes all the occurrences of $L+1$, the resulting string is in $M_{L,H,p-q}$.
- $m_{L,H,p}$ is defined recursively: for $H=L$ one has $m_{H,H,p} = m_{H,p}$ (as defined in Section 2.4). For $L \geq H$, we obtain recursively $m_{L+1,H,p}$ as the number of strings of length $H+2p$ such that $L+1$ appears an even number of times $2q$ *but never at two consecutive positions* and such that if one removes all the occurrences of $L+1$, the resulting string is in $m_{L,H,p-q}$.

These definitions translate directly into the following equations for $L \geq H$:

$$M_{L+1,H,p} = \sum_{q=0}^{p} \binom{H+2p}{2q} M_{L,H,p-q}, \qquad m_{L+1,H,p} = \sum_{q=0}^{p} \binom{H+2p-2q+1}{2q} m_{L,H,p-q}, \tag{52}$$

to be compared with (18) and (26). In each case, $2q$ is the number of times $L+1$ appears in the string of length $H+2p$. The two binomials correspond respectively to the number of ways of choosing $2q$ elements in $H+2p$, and the number of ways of choosing $2q$ in $H+2p$ such that two consecutive elements may not be chosen.

For the upper bound we define as before the generating function $G_{L,H}(X)$:

$$G_{L,H}(X) := \sum_{p\geq 0} \frac{M_{L,H,p}}{(H+2p)!} X^{H+2p}. \tag{53}$$

Using the same technique as in (22), one gets, for $L \geq H$,

$$G_{L+1,H}(X) = G_{L,H}(X)\cosh(X). \tag{54}$$

Furthermore, since $G_{H,H}(X) = G_H(X) = \sinh(X)^H$, one has

$$G_{L,H}(X) = \sinh(X)^H \cosh(X)^{L-H}. \tag{55}$$

Now, for the lower bound, we make the same transformation as before and we obtain

$$\binom{H+2p-2q+1}{2q} \geq \binom{H+2p}{2q}\Big[1-\frac{2q}{H+1}\Big]^{2q-1}\mathbb{1}_{2q<H+1}\,. \tag{56}$$

Obviously by recurrence, $m_{L,H,p} \geq \tilde{m}_{L,H,p}$ where we define

$$\tilde{m}_{H,H,p} = \tilde{m}_{H,p}, \qquad \tilde{m}_{L+1,H,p} = \sum_{q=0}^{p}\binom{H+2p}{2q}\tilde{m}_{L,H,p-q}\Big[1-\frac{2q}{H+1}\Big]^{2q-1}\mathbb{1}_{2q<H+1}\,. \tag{57}$$

Introducing the generating function $g_{L,H}(X)$

$$g_{L,H}(X) := \sum_{p\geq 0}\frac{\tilde{m}_{L,H,p}}{(H+2p)!}X^{H+2p}, \tag{58}$$

one gets from (57)

$$g_{H,H}(X) = g_H(X) = \prod_{l=1}^{H}\sinh_l(X), \qquad g_{L+1,H}(X) = g_{L,H}(X)\cosh_H(X), \tag{59}$$

where $\sinh_l(X)$ was defined in (33) and where

$$\cosh_H(X) = \sum_{q\geq 0}\frac{X^{2q}}{(2q)!}\Big[1-\frac{2q}{H+1}\Big]^{2q-1}\mathbb{1}_{2q<H+1}\,. \tag{60}$$

This gives

$$g_{L,H}(X) = \Big(\prod_{l=1}^{H}\sinh_l(X)\Big)\cosh_H(X)^{L-H}. \tag{61}$$

Finally, collecting the bits, one has the bounds

$$g'_{L,H}(1-x) \leq \mathbb{E}^x(\Theta) \leq G'_{L,H}(1-x). \tag{62}$$

If one chooses a function $L \mapsto H(L)$ such that $H(L)/L \to \alpha$ as L goes to infinity, then it is very easy to check that $\big[g'_{L,H(L)}(1-x)\big]^{1/L}$ and $\big[G'_{L,H(L)}(1-x)\big]^{1/L}$ both converges as $L\to\infty$ to the same quantity $\sinh(1-x)^\alpha\cosh(1-x)^{1-\alpha}$, which proves Theorem 2.

5 Proof when the fittest point is random

In this short section we prove Theorem 3. We consider the situation in which the fittest site σ_{fittest} is chosen uniformly at random in the hypercube (this is the case which is equivalent to the House-of-Cards model). In this case, H is obviously a binomial with parameters L and $1/2$, and therefore, $\alpha := H/L \to \frac{1}{2}$ in probability. Thus, when $x > x^*_{1/2}$, for $\epsilon > 0$ small enough so that $x > x^*_{1/2-\epsilon}$ we have that

$$\begin{aligned}\mathbb{P}^x(\Theta \geq 1) &\leq \mathbb{P}^x(\Theta \geq 1, \alpha \in \frac{1}{2}\pm\epsilon) + \mathbb{P}^x(\alpha \notin \frac{1}{2}\pm\epsilon),\\ &\to 0 \qquad \text{as } L\to\infty.\end{aligned} \tag{63}$$

To see this, just observe that the second term on the right-hand side tends to 0 independently of the value of x while the first term also goes to 0 as for any $\alpha > 1/2 - \epsilon$ we always have $x > x^*_\alpha$.

Moreover, if we assume that Conjecture 2 holds, *i.e.* that for all α fixed, $\mathbb{P}^x(\Theta \geq 1) \to 1$ when $x < x^*_\alpha$, the second part of Theorem 3 follows by the same argument.

However, be wary that the expected number of paths is lying. Just looking at the upper bound (but the lower bound should be the same) one can write with (55) and (62):

$$\mathbb{E}^x(\Theta) \leq \sum_{H=0}^{L} \frac{1}{2^L} \binom{L}{H} G'_{L,H}(1-x) = L\Big[\frac{e^{1-x}}{2}\Big]^L, \tag{64}$$

which diverges exponentially if and only if $x \leq 1 - \ln 2 = 0.30685\ldots$. This seems to give a critical point which is larger than $x^*_{1/2}$, but what happens is that with an exponentially small probability, α is much smaller than $1/2$ which generates exponentially many paths thereby contributing to the expectation.

References

[1] Altenberg L. (1997). NK Fitness Landscapes, in *Handbook of Evolutionary Computation*, pp B2.7:5–B2.7:10, T. Bäck and D. Fogel and Z. Michalewicz editors, IOP Publishing Ltd and Oxford University Press

[2] Berestycki J., Brunet É., Shi Z. (2013+). The number of accessible paths in the hypercube `ArXiv 1304.0246`

[3] Franke, J., Klözer, A., de Visser, J.A.G.M. and Krug, J. (2011). Evolutionary accessibility of mutational pathways, *PLoS Comput. Biol.* **7**, e1002134, 9pp

[4] Hegarty, P. and Martinsson, A. (2012+). On the existence of accessible paths in various models of fitness landscapes, `ArXiv 1210.4798` [math.PR]

[5] Gillespie J.H. (1983). A simple stochastic gene substitution model. *Theor. Pop. Biol0.* **23** 202

[6] Kauffman, S. and Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J. Theoret. Biol.* **128**, 11–45

[7] Kingman, J.F.C. (1978). A simple model for the balance between selection and mutation. *J. Appl. Probab.* **15**, 1–12

[8] Nowak, S. and Krug, J. (2013). Accessibility percolation on n-trees. *Europhys. Lett.* **101**, 66004

[9] DePristo, M.A., Hartl, D.L. and Weinreich, D.M. (2007). Mutational Reversions During Adaptive Protein Evolution *Mol Biol Evol.* **24**, 1608–1610

[10] Matthew I.R. and Zhao L.Z. (2013+). Increasing paths in trees `ArXiv 1305.0814` [math.PR]

[11] Schmiegelt B., Krug J. (2013+). Evolutionary accessibility of modular fitness landscapes `ArXiv 1306.1938` [q-bio]

[12] Weinreich, D.M., Watson, R.A. and Chao, L. (2005). Perspective: Sign epistasis and genetic constraints on evolutionary trajectories. *Evolution* **59**, 1165–1174

[13] Weinreich, D.M., Delaney, N.F., DePristo, M.A. and Hartl, D.M. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114

[14] *On-line Encyclopedia of Integer Sequences*, **A059783**, `http://oeis.org/A059783`