OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

The Time Scale of Evolutionary Innovation

Krishnendu Chatterjee[1]*, Andreas Pavlogiannis[1], Ben Adlam[2], Martin A. Nowak[2]

1 IST Austria, Klosterneuburg, Austria, **2** Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, Massachusetts, United States of America

Abstract

A fundamental question in biology is the following: what is the time scale that is needed for evolutionary innovations? There are many results that characterize single steps in terms of the fixation time of new mutants arising in populations of certain size and structure. But here we ask a different question, which is concerned with the much longer time scale of evolutionary trajectories: how long does it take for a population exploring a fitness landscape to find target sequences that encode new biological functions? Our key variable is the length, L, of the genetic sequence that undergoes adaptation. In computer science there is a crucial distinction between problems that require algorithms which take polynomial or exponential time. The latter are considered to be intractable. Here we develop a theoretical approach that allows us to estimate the time of evolution as function of L. We show that adaptation on many fitness landscapes takes time that is exponential in L, even if there are broad selection gradients and many targets uniformly distributed in sequence space. These negative results lead us to search for specific mechanisms that allow evolution to work on polynomial time scales. We study a regeneration process and show that it enables evolution to work in polynomial time.

Citation: Chatterjee K, Pavlogiannis A, Adlam B, Nowak MA (2014) The Time Scale of Evolutionary Innovation. PLoS Comput Biol 10(9): e1003818. doi:10.1371/journal.pcbi.1003818

Editor: Niko Beerenwinkel, ETH Zurich, Switzerland

Received December 13, 2013; **Accepted** July 21, 2014; **Published** September 11, 2014

Copyright: © 2014 Chatterjee et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Austrian Science Fund (FWF) Grant No P23499-N23, FWF NFN Grant No S11407-N23 (RiSE), ERC Start grant (279307: Graph Games), and Microsoft Faculty Fellows award. Support from the John Templeton foundation is gratefully acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: Krishnendu.Chatterjee@ist.ac.at

Introduction

Our planet came into existence 4.6 billion years ago. There is clear chemical evidence for life on earth 3.5 billion years ago [1,2]. The evolutionary process generated procaria, eucaria and complex multi-cellular organisms. Throughout the history of life, evolution had to discover sequences of biological polymers that perform specific, complicated functions. The average length of bacterial genes is about 1000 nucleotides, that of human genes about 3000 nucleotides. The longest known bacterial gene contains more than 10^5 nucleotides, the longest human gene more than 10^6. A basic question is what is the time scale required by evolution to discover the sequences that perform desired functions. While many results exist for the fixation time of individual mutants [3–15], here we ask how the time scale of evolution depends on the length L of the sequence that needs to be adapted. We consider the crucial distinction of polynomial versus exponential time [16–18]. A time scale that grows exponentially in L is infeasible for long sequences.

Evolutionary dynamics operates in sequence space, which can be imagined as a discrete multi-dimensional lattice that arises when all sequences of a given length are arranged such that nearest neighbors differ by one point mutation [19]. For constant selection, each point in sequence space is associated with a non-negative fitness value (reproductive rate). The resulting fitness landscape is a high dimensional mountain range. Populations explore fitness landscapes searching for elevated regions, ridges, and peaks [20–27].

A question that has been extensively studied is how long does it take for existing biological functions to improve under natural selection. This problem leads to the study of adaptive walks on fitness landscapes [15,20,21,28,29]. In this paper we ask a different question: how long does it take for evolution to discover a new function? More specifically, our aim is to estimate the expected discovery time of new biological functions: how long does it take for a population of reproducing organisms to discover a biological function that is not present at the beginning of the search. We will discuss two approximations for rugged fitness landscapes. We also discuss the significance of clustered peaks.

We consider an alphabet of size four, as is the case for DNA and RNA, and a nucleotide sequence of length L. We consider a population of size N, which reproduces asexually. The mutation rate, u, is small: individual mutations are introduced and evaluated by natural selection and random drift one at a time. The probability that the evolutionary process moves from a sequence i to a sequence j, which is at Hamming distance one from i, is given by $P_{i,j} = [Nu/(3L)]\rho_{i,j}$, where $\rho_{i,j}$ is the fixation probability of sequence j in a population consisting of sequence i. In the special case of a flat fitness landscape, we have $\rho_{i,j} = 1/N$, and $P_{i,j} = [u/(3L)]$. Thus we have an evolutionary random walk, where each step is a jump to a neighboring sequence of Hamming distance one.

Results

Consider a high-dimensional sequence space. A particular biological function can be instantiated by some of the sequences. Each sequence i has a fitness value f_i, which measures the ability of the sequence i to encode the desired function. Biological fitness landscapes are typically expected to have many peaks [29–31].

EFTA00611781

Author Summary

Evolutionary adaptation can be described as a biased, stochastic walk of a population of sequences in a high dimensional sequence space. The population explores a fitness landscape. The mutation-selection process biases the population towards regions of higher fitness. In this paper we estimate the time scale that is needed for evolutionary innovation. Our key parameter is the length of the genetic sequence that needs to be adapted. We show that a variety of evolutionary processes take exponential time in sequence length. We propose a specific process, which we call 'regeneration processes', and show that it allows evolution to work on polynomial time scales. In this view, evolution can solve a problem efficiently if it has solved a similar problem already.

They can be highly rugged due to epistatic effects of mutations [32–34]. They can also contain large regions or networks of neutrality [20,21]. Empirical studies of short RNA sequences have revealed that the underlying fitness landscape has low peak density [35]: around 15 peaks in 4^{24} sequences.

For the purpose of estimating the expected discovery time we can approximate the fitness landscape with a binary step function over the sequence space. We discuss two different approximations (Figure 1). For the first approximation, we consider the scenario where fitness values below some threshold, $f_{\min}$, have negligible contribution; those sequences do not instantiate the desired function (either not at all or only below the minimum level that could be detected by natural selection). We approximate the rugged fitness landscape as follows: if $f_i < f_{\min}$ then $f_i = 0$; if $f_i \geq f_{\min}$ then $f_i = 1$. The set of sequences with $f_i \geq f_{\min}$ constitutes the target set, and the remaining fitness landscape is neutral.

The second approximation works as follows. Consider the evolutionary process exploring a rugged fitness landscape where the goal is to attain a fitness level f^*. Local maxima below f^* slow down the evolutionary process to attain f^*, because the evolutionary walk might get stuck in those local maxima. In order to derive lower bounds for the expected discovery time, the rugged fitness landscape can be approximated as follows. Let $\hat{f}$ be the fitness value of the highest local maximum below f^*. Then for every sequence in a mountain range with a local maximum below f^* we assign the fitness value $\hat{f}$. The mountain ranges with local maxima above f^* are the target sequences. Note that the target set includes sequences that start at the upslope of mountain ranges with peaks above f^*. Thus, again we obtain a fitness landscape with clustered targets and neutral region, where the neutral region consists of all sequences whose fitness values have been assigned to $\hat{f}$. The two approximations are illustrated in Figure 1. For $f^* = f_{\min}$ the second approximation generates larger target areas than the first approximation and is therefore more lenient.

Our key results for estimating the discovery time can now be formulated for binary fitness landscapes, but they apply to any type of rugged landscape using one of the two approximations. We note that our methods can also be applied for certain non-binary fitness landscapes, and an example of a fitness landscape with a large gradient arising from multiplicative fitness effects is discussed in Sections 6 and 7 of Text S1.

We now present our main results in the following order. We first estimate the discovery time of a single search aiming to find a single broad peak. Then we study multiple simultaneous searches for a single broad peak. Finally, we consider multiple broad peaks that are uniformly randomly distributed in sequence space.

We first study a broad peak of target sequences described as follows: consider a specific sequence; any sequence within a certain Hamming distance of that sequence belongs to the target set. Specifically, we consider that the evolutionary process has succeeded, if the population discovers a sequence that differs from the specific sequence in no more than a fraction c of positions. We refer to the specific sequence as the target center and c as the width (or radius) of the peak. For example, if $L = 100$ and $c = 0.1$, then the target center is surrounded by a cloud of approximately 10^{18} sequences. For a single broad peak with width c, the target set contains at least $2^{cL}/(3L)$ sequences, which is an exponential function of L. The fitness landscape outside the broad peak is flat. We refer this binary fitness landscape as a broad peak landscape. The population needs to discover any one of the target sequences in the broad peak, starting from some sequence that is not in the broad peak. We establish the following result.

Theorem 1. *Consider a single search exploring a broad peak landscape with width c and mutation rate u. The following assertions hold*:

- *if $c < 3/4$, then there exists $L_0 \in \mathbb{N}$ such that for all sequence spaces of sequence length $L \geq L_0$, the expected discovery time is at least* $\exp[(3-4c)\frac{L}{16}\log\frac{6}{4c+3}]$;
- *if $c \geq 3/4$, then for all sequence spaces of sequence length L, the expected discovery time is at most $O(L^3/u)$.*

Our result can be interpreted as follows (see Theorem S2 and Corollary S2 in Text S1): (i) If $c < 3/4$, then the expected discovery time is exponential in L; and (ii) if $c \geq 3/4$, then the expected discovery time is polynomial in L. Thus, we have derived a *strong dichotomy* result which shows a sharp transition from polynomial to exponential time depending on whether a specific condition on c does or does not hold.

For the four letter alphabet most random sequences have Hamming distance $3L/4$ from the target center. If the population is further away than this Hamming distance, then random drift will bring it closer. If the population is closer than this Hamming distance, then random drift will push it further away. This argument constitutes the intuitive reason that $c = 3/4$ is the critical threshold. If the peak has a width of less than $c = 3/4$, then we prove that the expected discovery time by random drift is exponential in the sequence length L (see Figure 2). This result holds for any population size, N, as long as $4^L >> N$, which is certainly the case for realistic values of L and N. In the Text S1 we also present a more general result, where along with a single broad peak, instead of a flat landscape outside the peak we consider a multiplicative fitness landscape and establish a sharp dichotomy result that generalizes Theorem 1 (see Corollary S2 in Text S1).

Remark 1. *We highlight two important aspects of our results.*

1. *First, when we establish exponential lower bounds for the expected discovery time, then these lower bounds hold even if the starting sequence is only a few steps away from the target set.*
2. *Second, we present strong dichotomy results, and derive mathematically the most precise and strongest form of the boundary condition.*

Let us now give a numerical example to demonstrate that exponential time is intractable. Bacterial life on earth has been around for at least 3.5 billion years, which correspond to 3×10^{13} hours. Assuming fast bacterial cell division of 20–30 minutes on average we have at most 10^{14} generations. The expected discovery

EFTA00611782

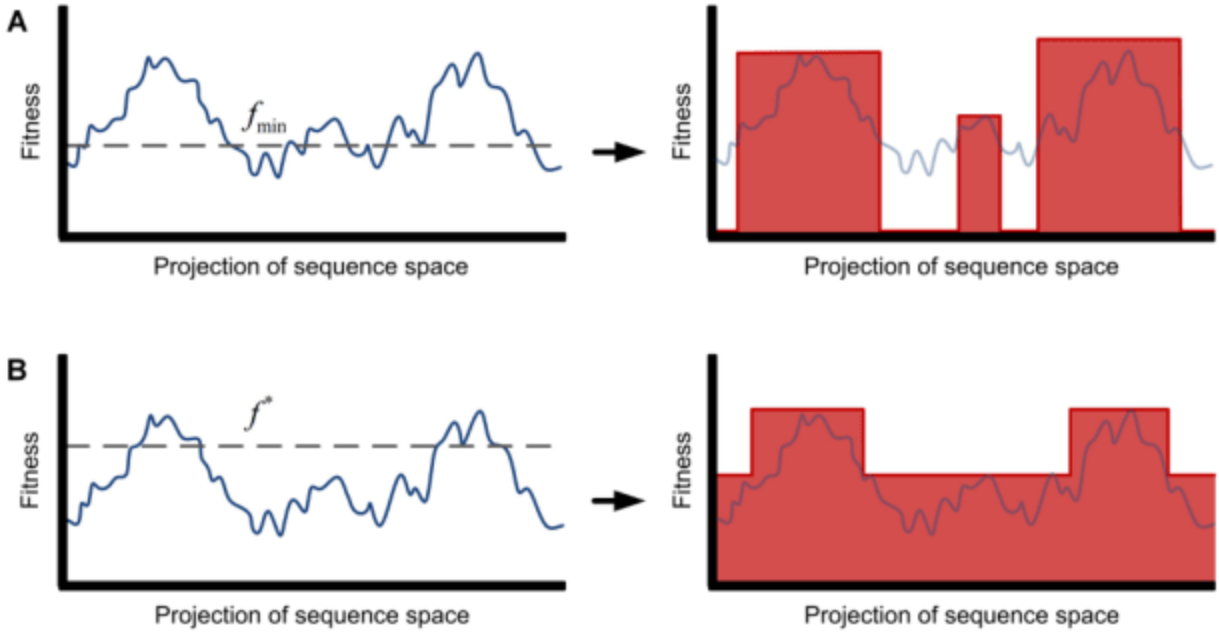

Figure 1. Approximations of a highly rugged fitness landscape by broad peaks and neutral regions. The figures depict examples of highly rugged fitness landscapes where the sequence space has been projected in one dimension. (A) Sequences with fitness below some level f_{min} are functionally very different to the desired function, and selection cannot act upon them. All other sequences are considered as targets. The fitness landscape is approximated by a step function: if $f_i < f_{min}$, then $f_i = 0$, otherwise $f_i = 1$. (B) Local maxima below the desired fitness threshold f^* are known to slow down the evolutionary random walk towards sequences that attain fitness at least f^*. We approximate the fitness landscape by broad peaks and neutral regions by increasing the fitness of every sequence that belongs in a mountain range with fitness below f^* to the maximal local maxima $\tilde{f}$ below f^*. Note that the target set starts from the upslope of a mountain range whose peak exceeds f^*.
doi:10.1371/journal.pcbi.1003818.g001

time for a sequence of length $L = 1000$ with a very large broad peak of $c = 1/2$ is approximately 10^{65} generations; see Table 1.

If individual evolutionary processes cannot find targets in polynomial time, then perhaps the success of evolution is based on the fact that many populations are searching independently and in parallel for a particular adaptation. We prove that multiple, independent parallel searches are not the solution of the problem, if the starting sequence is far away from the target center. Formally we show the following result.

Theorem 2. *In all cases where the lower bound on the expected discovery time is exponential, for all polynomials $p_1(\cdot)$, $p_2(\cdot)$ and $p_3(\cdot)$, for any starting sequence with Hamming distance at least $3L/4$ from the target center, the probability for any one out of $p_3(L)$ independent multiple searches to reach the target set within $p_1(L)$ steps is at most $1/p_2(L)$.*

If an evolutionary process takes exponential time, then polynomially many independent searches do not find the target in polynomial time with reasonable probability (for details see

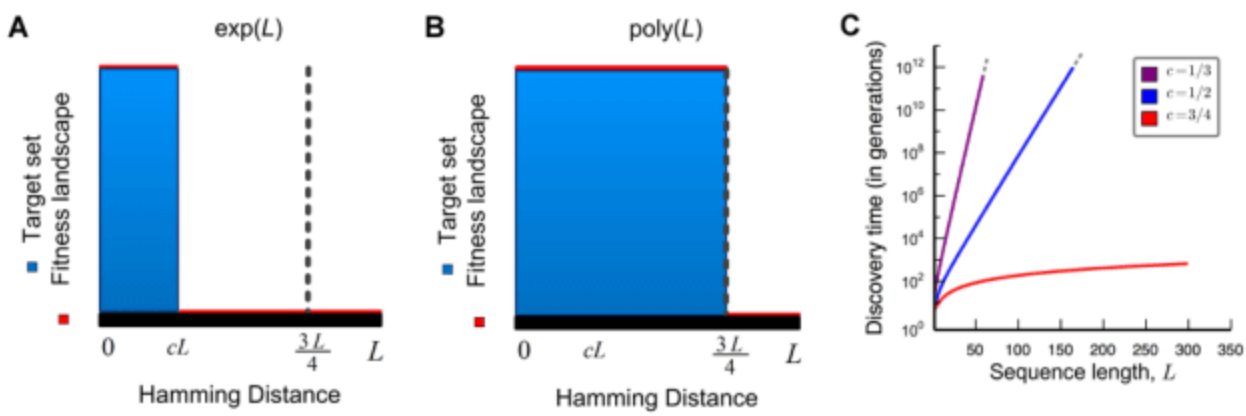

Figure 2. Broad peak with different fitness landscapes. For the broad peak there is a specific sequence, and all sequences that are within Hamming distance cL are part of the target set. The fitness landscape is flat outside the broad peak. (A) If the width of the broad peak is $c < 3/4$, then the expected discovery time is exponential in sequence length, L. (B) If the width of the broad peak is $c \geq 3/4$, then the expected discovery time is polynomial in sequence length, L. (C) Numerical calculations for broad peak fitness landscapes. We observe exponential expected discovery time for $c = 1/3$ and $c = 1/2$, whereas polynomial expected discovery time for $c = 3/4$.
doi:10.1371/journal.pcbi.1003818.g002

EFTA00611783

Table 1. Numerical data for discovery time in flat fitness landscapes.

$r=1$	$c=\frac{1}{3}$	$c=\frac{1}{2}$	$c=\frac{3}{4}$
$L=10^2$	$1.02\cdot10^{18}$	$7.36\cdot10^{7}$	183
$L=10^3$	$5.89\cdot10^{170}$	$1.28\cdot10^{65}$	2666

Numerical data for the discovery time of broad peaks with width $c=1/3, 1/2$, and $3/4$ embedded in flat fitness landscapes. First the discovery time is computed for small values of L as shown in Figure 2(C). Then the exponential growth is extrapolated to $L=100$ and $L=1000$, respectively. We show the discovery times for $c=1/2$, and $1/3$. For $c=3/4$ the values are polynomial in L.
doi:10.1371/journal.pcbi.1003818.t001

Theorem S5 in the Text S1). We also show an informal and approximate calculation of the success probability for M independent searches, as follows: if the expected discovery time is exponential (say, d), then the probability that all M independent searches fail upto b steps is at least $\exp(-(Mb)/d)$ (i.e., the success probability within b steps of any of the searches is at most $1-\exp(-(Mb)/d)$), when the starting sequence is far away from the target center. In such a case, one could quickly exhaust the physical resources of an entire planet. The estimated number of bacterial cells [36] on earth is about 10^{30}. To give a specific example let us assume that there are 10^{24} independent searches, each with population size $N=10^6$. The probability that at least one of those independent searches succeeds within 10^{14} generations for sequence length $L=1000$ and broad peak of $c=1/2$ is less than 10^{-26}.

In our basic model, individual mutants are evaluated one at a time. The situation of many mutant lineages evolving in parallel is similar to the multiple searches described above. As we show that whenever a single search takes exponential time, multiple independent searches do not lead to polynomial time solutions, our results imply intractability for this case as well.

We now explore the case of multiple broad peaks that are uniformly and randomly distributed. Consider that there are m target centers. Around each target center there is a selection gradient extending up to a distance cL. Formally we can consider any fitness function f that assigns zero fitness to a sequence whose Hamming distance exceeds cL from all the target centers, which in particular is subsumed by considering the multiple broad peaks where around each center we consider a broad peak of target set with peak width c. We establish the following result:

Theorem 3. *Consider a single search under the multiple broad peak fitness landscape of $m<<4^L$ target centers chosen uniformly at random, with peak width at most c for each center and $c<3/4$. Then with high probability, the expected discovery time of the target set is at least $(1/m)\exp[2L(3/4-c)^2]$.*

Whether or not the function $(1/m)\exp[2L(3/4-c)^2]$ is exponential in L depends on how m changes with L. But even if we assume exponentially many broad peak centers, m, with peak width cL where $c<3/4$, we need not obtain polynomial time (Figure 3 and Theorem S6 in Text S1).

It is known that recombination may accelerate evolution on certain fitness landscapes [28,37–39], and recombination may also slow down evolution on other fitness landscapes [40]. Recombination, however, reduces the discovery time only by at most a linear factor in sequence length [28,37,38,41,42]. A linear or even polynomial factor improvement over an exponential function does not convert the exponential function into a polynomial one. Hence, recombination can make a significant difference only if the underlying evolutionary process without recombination already operates in polynomial time.

What are then adaptive problems that can be solved by evolution in polynomial time? We propose a "regeneration process". The basic idea is that evolution can solve a new problem efficiently, if it is has solved a similar problem already. Suppose gene duplication or genome rearrangement can give rise to starting sequences that are at most k point mutations away from the target set, where k is a number that is independent of L. It is important that starting sequences can be regenerated again and again. We prove that L^{k+1} many searches are sufficient in order to find the target in polynomial time with high probability (see Figure 4 and Section 10 in Text S1). The upper bound, L^{k+1}, holds even for neutral drift (without selection). Note that in this case, the expected discovery time for any single search is still exponential. Therefore, most of the L^{k+1} searches do not succeed in polynomial time; however, with high probability one of the searches succeeds in polynomial time. There are two key aspects to the "regeneration process": (a) the starting sequence is only a small number of steps away from the target; and (b) the starting sequence can be generated repeatedly. This process enables evolution to overcome the exponential barrier. The upper bound, L^{k+1}, may possibly be further reduced, if selection and/or recombination are included.

Discussion

The regeneration process formalizes the role of several existing ideas. First, it ties in with the proposal that gene duplications and genome rearrangements are major events leading to the emergence of new genes [43]. Second, evolution can be seen as a tinkerer playing around with small modifications of existing sequences rather than creating entirely new ones [44]. Third, the process is related to Gillespie's suggestion [29] that the starting sequence for an evolutionary search must have high fitness. In our theory, proximity in fitness value is replaced by proximity in sequence space. However, our results show that proximity alone is insufficient to break the exponential barrier, and only when combined with the process of regeneration it yields polynomial discovery time with high probability. Our process can also explain the emergence of orphan genes arising from non-coding regions [45]. Section 12 of the Text S1 discusses the connection of our approach to existing results.

There is one other scenario that must be mentioned. It is possible that certain biological functions are hyper-abundant in sequence space [21] and that a process generating a large number of random sequences will find the function with high probability. For example, Bartel & Szostak [46] isolated a new ribozyme from a pool of about 10^{15} random sequences of length $L=220$. While such a process is conceivable for small effective sequence length, it cannot represent a general solution for large L.

Our theory has clear empirical implications. The regeneration process can be tested in systems of in vitro evolution [47]. A

EFTA00611784

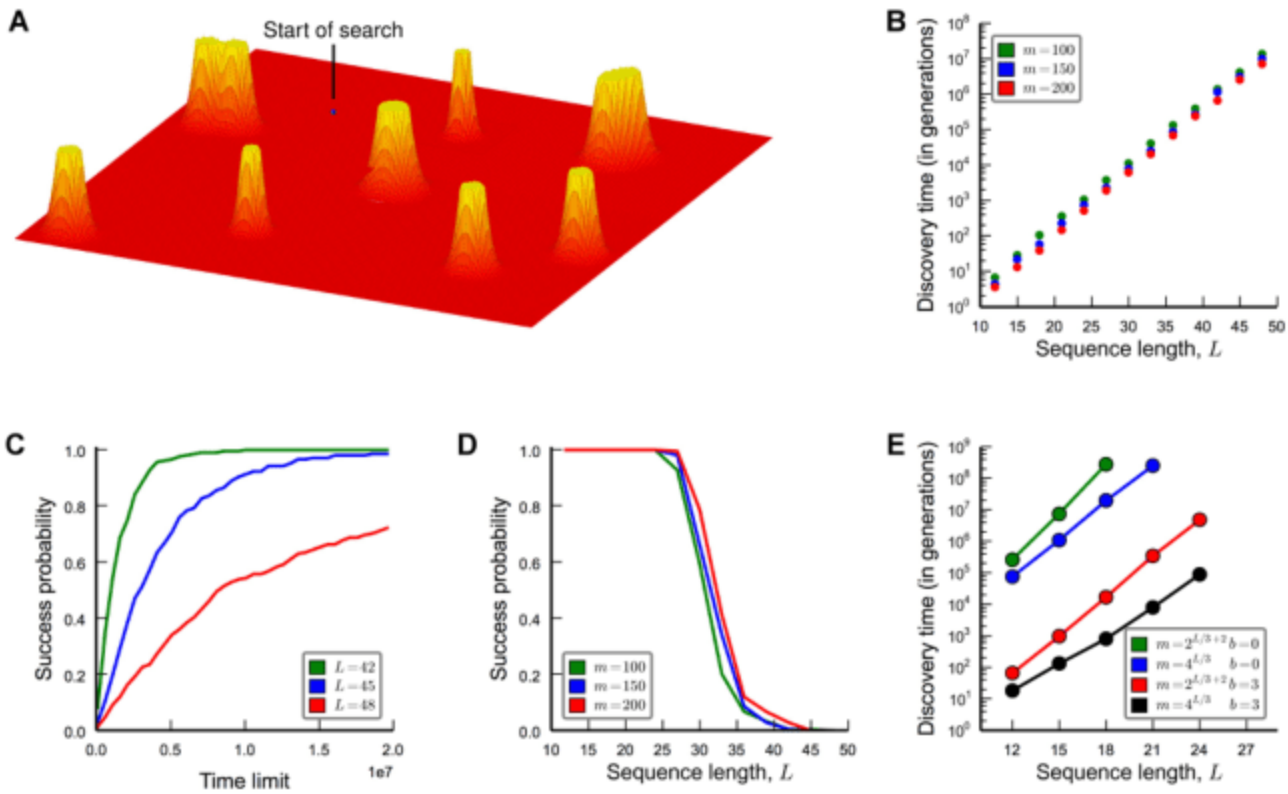

Figure 3. The search for randomly, uniformly distributed targets in sequence space. (A) The target set consists of m random sequences; each one of them is surrounded by a broad peak of width up to cL. The figure shows a pictorial illustration where the L-dimensional sequence space is projected onto two dimensions. From a randomly chosen starting sequence outside the target set, the expected discovery time is at least $(1/m)\exp[2L(3/4-c)^2]$, which can be exponential in L. (B) Computer simulations showing the average discovery time of $m=100$, 150, and 200 targets, with $c=1/3$. We observe exponential dependency on L. The discovery time is averaged over 200 runs. (C) Success probability estimated as the fraction of the 200 searches that succeed in finding one of the target sequences within 10^4 generations. The success probability drops exponentially with L. (D) Success probability as a function of time for $L=42$, 45, and 48. (E) Discovery time for a large number of randomly generated target sequences. Either $m=2^{L/3+2}$ or $m=4^{L/3}$ sequences were generated. For $b=0$ and $b=3$ the target set consists of balls of Hamming distance 0 and 3 (respectively) around each sequence. The figure shows the average discovery time of 100 runs. As expected we observe that the discovery time grows exponentially with sequence length, L.
doi:10.1371/journal.pcbi.1003818.g003

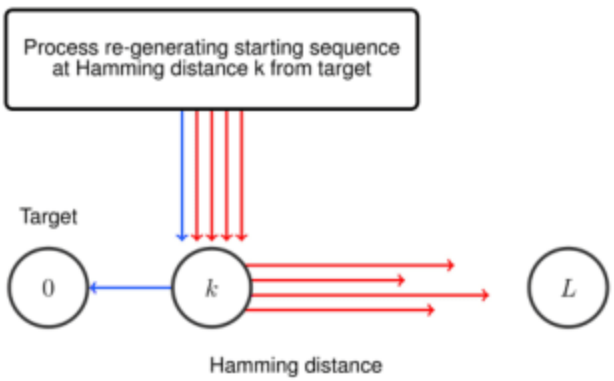

Figure 4. Regeneration process. Gene duplication (or possibly some other process) generates a steady stream of starting sequences that are a constant number k of mutations away from the target. Many searches drift away from the target, but some will succeed in polynomially many steps. We prove that L^{k+1} searches ensure that with high probability some search succeed in polynomially many steps.
doi:10.1371/journal.pcbi.1003818.g004

starting sequence can be generated by introducing k point mutations in a known protein encoding sequence of length L. If these point mutations destroy the function of the protein, then the expected discovery time of any one attempt to find the original sequence should be exponential in L. But only polynomially many searches in L are required to find the target with high probability in polynomially many steps. The same setup can be used to explore whether the biological function can be found elsewhere in sequence space: the evolutionary trajectory beginning with the starting sequence could discover new solutions. Our theory also highlights how important it is to explore the distribution of biological functions in sequence space both for RNA [20,21,35,46] and in the protein universe [48].

In summary, we have developed a theory that allows us to estimate time scales of evolutionary trajectories. We have shown that various natural processes of evolution take exponential time as function of the sequence length, L. In some cases we have established strong dichotomy results for precise boundary conditions. We have proposed a mechanism that allows evolution in polynomial time scales. Some interesting directions of future work are as follows: (1) Consider various forms of rugged fitness landscapes and study more refined approximations as compared to

EFTA00611785

the ones we consider; and then estimate the expected discovery time for the refined approximations. (2) While in this paper we characterize the difference between exponential and polynomial for the expected discovery time, more refined analysis (such as efficiency for polynomial time, like cubic vs quadratic time) for specific fitness landscapes using mechanisms like recombination is another interesting problem.

Materials and Methods

Our results are based on a mathematical analysis of the underlying stochastic processes. For Markov chains on the one-dimensional grid, we describe recurrence relations for the expected hitting time and present lower and upper bounds on the expected hitting time using combinatorial analysis (see Text S1 for details). We now present the basic intuitive arguments of the main results.

Markov chain on the one-dimensional grid

For a single broad peak, due to symmetry we can interpret the evolutionary random walk as a Markov chain on the one-dimensional grid. A sequence of type i is i steps away from the target, where i is the Hamming distance between this sequence and the target. The probability that a type i sequence mutates to a type $i-1$ sequence is given by $ui/(3L)$. The stochastic process of the evolutionary random walk is a Markov chain on the one-dimensional grid $0,1,\ldots,L$.

The basic recurrence relation

Consider a Markov chain on the one-dimensional grid, and let $H(j,i)$ denote the expected hitting time from i to j. The general recurrence relation for the expected hitting time is as follows:

$$H(j,i) = 1 + P_{i,i+1}H(j,i+1) + P_{i,i-1}H(j,i-1) + P_{i,i}H(j,i); \qquad (1)$$

for $j<i<L$, with boundary condition $H(j,j)=0$. The interpretation is as follows. Given the current state i, if $i \neq j$, at least one transition will be made to a neighboring state i', with probability $P_{i,i'}$, from which the hitting time is $H(j,i')$.

Intuition behind Theorem 1

Theorem 1 is derived by obtaining precise bounds for the recurrence relation of the hitting time (Equation 1). Consider that $P_{k,k-1}>0$ for all $j<k\leq i$ (i.e., progress towards state j is always possible), as otherwise j is never reached from i. We show (see Lemma 2 in the Text S1) that we can write $H(j,i)$ as a sum, $H(j,i)=\sum_{n=L-i}^{L-j-1} b_n$, where b_n is the sequence defined as:

$$\begin{aligned} &(i)\ b_0 = \frac{1}{P_{L,L-1}}; \\ &(ii)\ b_n = \frac{1+P_{L-n,L-n+1}b_{n-1}}{P_{L-n,L-n-1}} \quad \text{for } n>0. \end{aligned} \qquad (2)$$

The basic intuition obtained from Equation 2 is as follows: (i) If $\frac{P_{L-n,L-n+1}}{P_{L-n,L-n-1}} \geq \lambda$, for some constant $\lambda>1$, then the sequence b_n grows at least as fast as a geometric series with factor λ. (ii) On the other hand, if $\frac{P_{L-n,L-n+1}}{P_{L-n,L-n-1}} \leq 1$ and $P_{L-n,L-n-1} \geq \alpha$ for some constant $\alpha>0$, then the sequence b_n grows at most as fast as an arithmetic series with difference $1/\alpha$. From the above case analysis the result for Theorem 1 is obtained as follows: If $c<\frac{3}{4}$, then for all $cL<n<\frac{3+4c}{8}L$, we have $\frac{P_{L-n,L-n+1}}{P_{L-n,L-n-1}} \geq \lambda$ for some $\lambda>1$, and hence the sequence b_n grows geometrically for a linear length in L. Then, $H(cL,i) \geq \lambda^{\frac{3-4c}{8}L}$ for all states $i>cL$ (i.e., for all sequences outside of the target set). This corresponds to case 1 of Theorem 1. On the other hand, if $c \geq \frac{3}{4}$, then it is $\frac{P_{L-n,L-n+1}}{P_{L-n,L-n-1}} \leq 1$, and case 2 of Theorem 1 is derived (for details see Corollary 2 in Text S1).

Intuition behind Theorem 2

The basic intuition for the result is as follows: consider a single search for which the expected hitting time is exponential. Then for the single search the probability to succeed in polynomially many steps is negligible (as otherwise the expectation would not have been exponential). In case of independent searches, the independence ensures that the probability that all searches fail is the product of the probabilities that every single search fails. Using the above arguments we establish Theorem 2 (for details see Section 8 in Text S1).

Intuition behind Theorem 3

For this result, it is first convenient to view the evolutionary walk taking place in the sequence space of all sequences of length L, under no selection. Each sequence has $3L$ neighbors, and considering that a point mutation happens, the transition probability to each of them is $\frac{1}{3L}$. The underlying Markov chain due to symmetry has fast mixing time, i.e., the number of steps to converge to the stationary distribution (the mixing time) is $O(L\log L)$. Again by symmetry the stationary distribution is the uniform distribution. If $c<\frac{3}{4}$, then from Theorem 1 we obtain that the expected time to reach a single broad peak is exponential. By union bound, if $m<<4^L$, the probability to reach any of the m broad peaks within $O(L\log L)$ steps is negligible. Since after the first $O(L\log L)$ steps the Markov chain converges to the stationary distribution, then each step of the process can be interpreted as selection of sequences uniformly at random among all sequences. Using Hoeffding's inequality, we show that with high probability, in expectation $\frac{\exp\left(2\cdot(3/4-c)^2\cdot L\right)}{m}$ such steps are required before a sequence is found that belongs to the target set. Thus we obtain the result of Theorem 3 (for details see Section 9 in Text S1).

Remark about techniques

An important aspect of our work is that we establish our results using elementary techniques for analysis of Markov chains. The use of more advanced mathematical machinery, such as martingales [49] or drift analysis [50,51], can possibly be used to derive more refined results. While in this work our goal is to distinguish between exponential and polynomial time, whether the techniques from [49–51] can lead to a more refined characterization within polynomial time is an interesting direction for future work.

Supporting Information

Text S1 Detailed proofs for "The Time Scale of Evolutionary Innovation."
(PDF)

EFTA00611786

Acknowledgments

We thank Nick Barton and Daniel Weissman for helpful discussions and pointing us to relevant literature.

Author Contributions

Conceived and designed the experiments: KC AP BA MAN. Analyzed the data: KC AP BA MAN. Wrote the paper: KC AP BA MAN.

References

1. Allwood AC, Grotzinger JP, Knoll AH, Burch IW, Anderson MS, et al. (2009) Controls on development and diversity of early archean stromatolites. Proc Natl Acad Sci USA 106: 9548–9555.
2. Schopf JW (August 2006) The first billion years: When did life emerge? Elements 2: 229–233.
3. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–626.
4. Ewens WJ (1967) The probability of survival of a new mutant in a uctuating environment. Heredity 22: 438–443.
5. Barton NH (1995) Linkage and the limits to natural selection. Genetics 140: 821–41.
6. Campos PR (2004) Fixation of beneficial mutations in the presence of epistatic interactions. Bull Math Biol 66: 473–486.
7. Antal T, Scheuring I (2006) Fixation of strategies for an evolutionary game in finite populations. Bull Math Biol 68: 1923–1944.
8. Whitlock MC (2003) Fixation probability and time in subdivided populations. Genetics 164: 767–779.
9. Altrock PM, Traulsen A (2009) Fixation times in evolutionary games under weak selection. New J Phys 11. doi:10.1088/1367-2630/11/1/013012
10. Kimura M, Ohta T (1969) Average number of generations until fixation of a mutant gene in a finite population. Genetics 61: 763–771.
11. Johnson T, Gerrish P (2002) The fixation probability of a beneficial allele in a population dividing by binary fission. Genetica 115: 283–287.
12. Orr HA (2000) The rate of adaptation in asexuals. Genetics 155: 961–968.
13. Wilke CO (2004) The speed of adaptation in large asexual populations. Genetics 167: 2045–2053.
14. Desai MM, Fisher DS, Murray AW (2007) The speed of evolution and maintenance of variation in asexual populations. Curr Biol 17: 385–394.
15. Ohta T (1972) Population size and rate of evolution. J Mol Evol 1: 305–314.
16. Papadimitriou C (1994) Computational complexity. Addison-Wesley.
17. Cormen T, Leiserson C, Rivest R, Stein C (2009) Introduction to Algorithms. MIT Press.
18. Valiant LG (2009) Evolvability. J ACM 56: 3:1–3:21.
19. Maynard Smith J (1970) Natural selection and the concept of a protein space. Nature 225: 563–564.
20. Fontana W, Schuster P (1987) A computer model of evolutionary optimization. Biophys Chem 26: 123–147.
21. Fontana W, Schuster P (1998) Continuity in evolution: On the nature of transitions. Science 280: 1451–1455.
22. Eigen M, Mccaskill J, Schuster P (1988) Molecular quasi-species. J Phys Chem 92: 6881–6891.
23. Eigen M, Schuster P (1978) The hypercycle. Naturwissenschaften 65: 7–41.
24. Park SC, Simon D, Krug J (2010) The speed of evolution in large asexual populations. J Stat Phys 138: 381–410.
25. Derrida B, Peliti L (1991) Evolution in a at fitness landscape. Bull Math Biol 53: 355–382.
26. Stadler PF (2002) Fitness landscapes. Appl Math & Comput 117: 187–207.
27. Worden RP (1995) A speed limit for evolution. J Theor Biol 176: 137–152.
28. Crow JF, Kimura M (1965) Evolution in sexual and asexual populations. Am Nat 99: pp. 439–450.
29. Gillespie JH (1984) Molecular evolution over the mutational landscape. Evolution 38: 1116–1129.
30. Kauffman S, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. Journal of Theoretical Biology 128: 11–45.
31. Orr HA (2000) A minimum on the mean number of steps taken in adaptive walks. Journal of Theoretical Biology 220: 241–247.
32. Weinreich DM, Watson RA, Chao L (2005) Perspective:sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59: 1165–1174.
33. Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ (2007) Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445: 383–386.
34. Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, et al. (2011) Second-order selection for evolvability in a large escherichia coli population. Science 331: 1433–1436.
35. Jimenez JI, Xulvi-Brunet R, Campbell GW, Turk-MacLeod R, Chen IA (2013) Comprehensive experimental fitness landscape and evolutionary network for small rna. Proc Natl Acad Sci USA. 110(37):14984–9.
36. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. Proc Natl Acad Sci USA 95: 6578–6583.
37. Smith JM (1974) Recombination and the rate of evolution. Genetics 78: 299–304.
38. Crow JF, Kimura M (1970) An introduction to population genetics theory. Burgess Publishing Company.
39. Park SC, Krug J (2013) Rate of adaptation in sexuals and asexuals: A solvable model of the fishermuller effect. Genetics 195: 941–955.
40. de Visser JAGM, Park S, Krug J (2009) Exploring the effect of sex on empirical fitness landscapes. The American Naturalist 174: S15–S30.
41. Neher RA, Shraiman BI, Fisher DS (2010) Rate of adaptation in large sexual populations. Genetics 184: 467–481.
42. Weissman DB, Hallatschek O (2014) The rate of adaptation in large sexual populations with linear chromosomes. Genetics 196: 1167–1183.
43. Ohno S (1970) Evolution by gene duplication. Springer-Verlag.
44. Jacob F (1977) Evolution and tinkering. Science 196: 1161–1166.
45. Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. Nat Rev Genet 12: 692–702.
46. Bartel D, Szostak J (1993) Isolation of new ribozymes from a large pool of random sequences. Science 261: 1411–1418.
47. Leconte AM, Dickinson BC, Yang DD, Chen IA, Allen B, et al. (2013) A population-based experimental model for protein evolution: Effects of mutation rate and selection stringency on evolutionary outcomes. Biochemistry 52: 1490–1499.
48. Povolotskaya IS, Kondrashov FA (2010) Sequence space and the ongoing expansion of the protein universe. Nature 465: 922–926.
49. Williams D (1991) Probability with Martingales. Cambridge mathematical textbooks. Cambridge University Press.
50. Hajek B (1982) Hitting-time and occupation-time bounds implied by drift analysis with applications. Advances in Applied Probability 14: pp. 502–525.
51. Lehre PK, Witt C (2013) General drift analysis with tail bounds. CoRR abs/1307.2559.

EFTA00611787