Monotonicity of Fitness Landscapes and Mutation Rate Control[☆]

Roman V. Belavkin[a,*], Alastair Channon[b], Elizabeth Aston[b], John Aston[c], Rok Krašovec[d], Christopher G. Knight[d]

[a]*School of Engineering and Information Sciences, Middlesex University, London NW4 4BT, UK*
[b]*Research Institute for the Environment, Physical Sciences and Applied Mathematics, Keele University, ST5 5BG, UK*
[c]*Department of Statistics, University of Warwick, CV4 7AL, UK*
[d]*Faculty of Life Sciences, University of Manchester, M13 9PT, UK*

Abstract

The typical view in evolutionary biology is that mutation rates are minimised. Contrary to that view, studies in combinatorial optimisation and search have shown a clear advantage of using variable mutation rates as a control parameter to optimise the performance of evolutionary algorithms. Ronald Fisher's work is the basis of much biological theory in this area. He used Euclidean geometry of continuous, infinite phenotypic spaces to study the relation between mutation size and expected fitness of the offspring. Here we develop a general theory of optimal mutation rate control that is based on the alternative geometry of discrete and finite spaces of DNA sequences. We define the monotonic properties of fitness landscapes, which allows us to relate fitness to the topology of genotypes and mutation size. First, we consider the case of a perfectly monotonic fitness landscape, in which the optimal mutation rate control functions can be derived exactly or approximately depending on additional constraints of the problem. Then we consider the general case of

[☆]This work was supported by EPSRC grant EP/H031936/1

[*]Corresponding author

Email addresses: r.belavkin@mdx.ac.uk (Roman V. Belavkin), a.d.channon@epsam.keele.ac.uk (Alastair Channon), e.j.aston@epsam.keele.ac.uk (Elizabeth Aston), j.a.d.aston@warwick.ac.uk (John Aston), rok.krasovec@manchester.ac.uk (Rok Krašovec), chris.knight@manchester.ac.uk (Christopher G. Knight)

non-monotonic landscapes. We use the ideas of local and weak monotonicity to show that optimal mutation rate control functions exist in any such landscape and that they resemble control functions in a monotonic landscape at least in some neighbourhood of a fitness maximum. Generally, optimal mutation rates increase when fitness decreases, and the increase of mutation rate is more rapid in landscapes that are less monotonic (more rugged). We demonstrate these relationships by obtaining and analysing approximately optimal mutation rate control functions in 115 complete landscapes of binding scores between DNA sequences and transcription factors. We discuss the relevance of these findings to living organisms, including the phenomenon of stress-induced mutagenesis.

Keywords: Adaptation, Fitness landscape, Mutation rate, Population Genetics, Phenotypic Plasticity

---

Contents

1 Introduction **3**

2 A Generalisation of Fisher's Geometric Model of Adaptation **5**
- 2.1 Representation and assumptions 6
- 2.2 Fisher's representation in Euclidean space 8
- 2.3 Probability of adaptation and representation in a Hamming space 9
- 2.4 Random mutation 12
- 2.5 Optimal control of mutation rates 14

3 Evolutionary Optimisation of Mutation Rate Control Functions **17**
- 3.1 Inner-GA 17
- 3.2 Meta-GA 18
- 3.3 Evolved control functions 19

4 Locally and Weakly Monotonic Fitness Landscapes **19**
- 4.1 Memoryless communication between fitness and distance . . . 20
- 4.2 Monotonicity of fitness landscapes 22
- 4.3 Weak monotonicity 25

5 Evolving Fitness-Based Mutation Rate Control Functions 27
5.1 Evolved control functions 27
5.2 Landscapes for transcription factors 28
5.3 Monotonicity and controllability 29

6 Discussion 32

Appendix A Memoryless Communication 44

Appendix B Monotonicity 47

Appendix C Point Mutation as Optimal Solution of Variational Problem 47

Appendix D Supplementary Figures 50

1. Introduction

Mutation is one of the most important biological processes that influence evolutionary dynamics. During replication mutation leads to a loss of information between the offspring and its parent, but it also allows the offspring to acquire new features. These features are likely to be deleterious, but have the potential to be beneficial for adaptation. Thus mutation can be seen as a process of innovation, which is particularly important as the number of all living organisms is tiny relative to the number of all possible organisms. A question that naturally arises with regards to mutation is whether there is an optimal balance between the amount of information lost and potential fitness gained.

The seminal mathematical work to investigate biological mutation is by Ronald Fisher [1], who considered mutation as a random motion in Euclidean space, the points of which are vectors representing collections of phenotypic traits of organisms. Using the geometry of Euclidean space, Fisher showed that probability of adaptation decreases exponentially as a function of mutation size (defined using the ratio of mutation radius and distance to the optimum), and concluded therefore that adaptation is more likely to occur by small mutations. Several studies, however, suggested that large mutations can be quite frequent in nature, thereby prompting re-examination of the theory [2]. Thus, Kimura [3] extended the theory to take into account differences in probabilities of fixation for mutations of small and large size.

Subsequently Orr [4] considered the effect of mutation across several replications. Interestingly, while he had a critical role in developing mathematical theory around discrete alleles, Fisher in his geometric model uses Euclidean space, which is uncountably infinite and unbounded. That this is an important issue became apparent only after the realisation that biological evolution occurs in a countable or even finite space of discrete molecular sequences [5]. However, subsequent geometric models based on Fisher's, while they have explicitly modelled discrete mutational steps (e.g. [6]), continue to assume that they occur within the same infinite Euclidean space. This issue may contribute to the fact that the predictions of such models have at best only been partially verified in actual biological systems [7, 8, 9, 10]. One of the contributions of the current work is that we consider mutation using the geometry of other spaces, and in particular the geometry of a Hamming space, which is finite and leads to a radically different view about the role of large mutations.

Mutation size as considered by Fisher is closely related to mutation frequency measured in biology in terms of the number of mutations per replication per DNA base. Mutation rates in biology vary over several orders of magnitude [11]. Nonetheless, mutation rate for any particular species is typically believed to be minimised, within bounds set by physiology [12], or more likely population genetics [13]. Despite this, mutation rates are known to vary within and among populations of a single species [14] and recently, population-genetic models have been developed proposing that variable mutation rates may be in fact adaptive in biology [15].

Independent of such biological concerns, researchers in evolutionary computation and operations research have a longer history of considering variable mutation rates in genetic algorithms (GAs) (e.g. see [16, 17, 18, 19, 20] for reviews). In particular, Ackley suggested in [21] that mutation probability is analogous to temperature in simulated annealing, which decreases with time through optimisation. A gradual reduction of mutation rate was also proposed by Fogarty [22]. In a pioneering work, Yanagiya [23] used Markov chain analysis of GAs to show that a sequence of optimal mutation rates maximising the probability of obtaining global solution exists in any problem. A significant contribution to the field was made by Thomas Bäck [24], who studied the probability of adaptation in the space of binary sequences and suggested that mutation rate should depend on fitness values rather than time. More recently, numerical methods have been used to optimise a mutation operator [20] that was based on the Markov chain model of GA by Nix

and Vose [25]. The complexity of this model, however, restricted the application of this method to small spaces and populations. It is these insights regarding mutation rate variation from evolutionary computation and operations research which we develop here towards the particular issues presented by biological systems.

We develop theory in the following directions:

1. Generalise Fisher's geometric model of adaptation for metric spaces, and in particular for discrete spaces of sequences, such as the Hamming spaces with arbitrary alphabets.
2. Define problems of optimal mutation rate control within such spaces, and study how different problem formulations (e.g. time horizon, objective function) affect the solutions.
3. Extend the theory to more biologically realistic (i.e. rugged) fitness landscapes.

Some relevant results have already been reported. For example, results for general Hamming spaces were first reported in [26, 27]. We develop these results towards biology in Section 2. Various optimisation problems were considered in [28, 29], deriving theoretical optimal mutation rate control functions. We address how such control functions may also be obtained numerically in Section 3. In Section 4, we develop theory to consider a fitness landscape as a memoryless communication channel between fitness values and distance from an optimal sequence. We introduce the ideas of local and weak monotonicity of a landscape. This allows us to formulate hypotheses about monotonicity and mutation rate control in biological fitness landscapes. We test these hypotheses by numerically obtaining optimal mutation rate control functions for 115 published complete landscapes of transcription factor binding [30]. Our results presented in Section 5 show that all the optimal mutation rate control functions in these biological landscapes do indeed converge to non-trivial forms consistent with the theory developed here. We also observe differences among optimal mutation rate control functions, variation that relates to variation in the landscapes' monotonic properties. We conclude in Section 6 by discussing how mutation rate control as considered here may be manifested in living organisms.

2. A Generalisation of Fisher's Geometric Model of Adaptation

In this section, we consider an abstract problem, in which organisms are viewed as points in some metric space and adaptation as a motion in this

space towards some target point (an optimal organism). In such formulation, maximisation of biological fitness corresponds to a minimisation of distance to the target, and geometry of the metric space allows us to solve the optimisation problem precisely. These abstract results will be used in the following sections to develop the theory further bringing it closer to biology.

2.1. Representation and assumptions

Let Ω be a set of all possible organisms. Environment defines a preference relation $\lesssim$ on Ω (a total pre-order), so that $a \lesssim b$ means b is better adapted to or has a higher replication rate in a particular environment than a. Throughout this paper we shall consider only the case of countable or even finite Ω, although the theory can be easily extended with certain care to the uncountable case. It is well-known from game theory (e.g. [31]) that in the countable case the preference relation always has a utility representation: there exists a real function $f : \Omega \to \mathbb{R}$ such that $a \lesssim b$ if and only if $f(a) \leq f(b)$. In the biological context, the utility function is called *fitness*, and it is usually defined to have non-negative values (i.e. if $f(\omega)$ is the replication rate of ω). Having positive fitness values is not essential, because the preference relation does not change under a strictly increasing transformation of f, such as adding a constant $\varepsilon \in \mathbb{R}$ to f or multiplying it by a positive number (i.e. representation $f(\omega)$ is equivalent to $\lambda f(\omega) + \varepsilon$ for any $\lambda > 0$ and $\varepsilon \in \mathbb{R}$). Thus, our interpretation of fitness simply as a numerical representation of a preference relation on organisms is distinct from population genetic definitions of fitness (e.g. see [32]). We shall assume also that there exists a top (optimal) element $\top \in \Omega$ such that $f(\top) = \sup f(\omega)$, which is the most adapted and quickly replicating specie in the current environment. Note that a finite set Ω always contains at least one top (optimal) element $\top$ as well as at least one bottom element $\bot$.

Generally, one can consider also the set Θ of all environments (including other organisms), because different environments $\theta \in \Theta$ impose different preference relations $\lesssim_\theta$ on Ω, which have to be represented by different fitness functions $f_\theta(\omega) := f(\omega, \theta)$. In this paper, however, we shall assume that a particular environment has been fixed, and therefore consider only one preference relation and one fitness function.

During the replication, organism a can mutate into b with probability $P(b \mid a)$, and the products $P(b \mid a) \cdot f(a)$ define the *selection-mutation* matrix — the infinitesimal generator of the replicator-mutator dynamics (generally non-linear Markov evolution). Mutation can have different effects on

fitness of the offspring. Mutation $a \mapsto b$ can be deleterious, if $f(a) > f(b)$, neutral, if $f(a) = f(b)$, or beneficial, if $f(a) < f(b)$. We shall analyse how the probability of beneficial mutation can be related to the 'geometry' of mutation.

Fitness is defined by the interaction of an organism with its environment, and therefore it is a property of a *phenotype*. Thus the set Ω, which is the domain of the fitness function, can be thought of as not just the set of all organisms, but the set of all possible phenotypes. Reproduction of organisms, however, involves passing of information about the phenotypes in the form of codes, which can be elements of some other set. Consider a representation of phenotypes $\omega \in \Omega$ by points of a topological vector space $\mathcal{H}$ (e.g. a space of traits, a space of DNA sequences and so on). In information theory, a mapping $\kappa : \Omega \to \mathcal{H}$ is called a *code*, and we shall assume here that it is uniquely decodable: $\kappa(a) = \kappa(b)$ implies $a = b$. That is, $\omega \mapsto \kappa(\omega)$ is an injection of Ω into a possibly larger space $\mathcal{H}$. In biological terms, each genotype has either one or no phenotype, and each phenotype has precisely one genotype. In addition, we shall assume that the image of κ is closed under the operation of addition in $\mathcal{H}$, which implies that for all a, $b \in \Omega$, there exists $c \in \Omega$ such that $\kappa(a) + \kappa(c) = \kappa(b)$. Thus, mutation $a \mapsto b$ in Ω can be represented in $\mathcal{H}$ by addition of codes $\kappa(a)$ and $\kappa(c) = \kappa(b) - \kappa(a)$, as shown on the following diagram:

$$\begin{array}{ccc} \Omega \ni a & \xrightarrow{\text{Mutation}} & b \in \Omega \\ {\scriptstyle\kappa}\downarrow & & \uparrow{\scriptstyle\kappa^{-1}} \\ \mathcal{H} \ni \kappa(a) & \xrightarrow{+\kappa(c)} & \kappa(b) \in \mathcal{H} \end{array}$$

We shall assume that the topology in $\mathcal{H}$ is defined by a metric $d : \mathcal{H} \times \mathcal{H} \to [0, \infty)$ (i.e. $\mathcal{H}$ is a metric vector space). Under a uniquely decodable mapping κ, the metric on $\mathcal{H}$ induces an equivalent metric on Ω representing 'dissimilarity' of two phenotypes. Thus, abusing notation, we shall identify phenotypes ω with their codes $\kappa(\omega)$ and write $d(a, b)$ and $b = a + c$ instead of $d(\kappa(a), \kappa(b))$ and $\kappa(b) = \kappa(a) + \kappa(c)$. A sphere and a ball of radius $r \in [0, \infty)$ around every point $a \in \Omega$ is defined as usual:

$$S(a, r) := \{b \in \Omega : d(a, b) = r\}, \quad B(a, r) := \bigcup_{n \in [0, r]} S(a, n)$$

If a mutates into b, then we call $r = d(a, b)$ a *mutation radius*.

More generally, a representation may be non-uniquely decodable or even stochastic, in which case Ω is not a metric space, but this will not be considered here. Thus we consider a simplified picture of uniquely decodable genotypes. The motivation for distinguishing genotype and phenotype however will become apparent in Section 4 when we define the monotonic properties of general fitness landscapes. In particular, the radius r is the dissimilarity of the codes (e.g. genotypes) $\kappa(a)$ and $\kappa(b)$, and it depends on the choice of a representation space $\mathcal{H}$, its metric and the encoding-decoding schemes κ and κ^{-1}, all of which may influence landscape monotonicity.

2.2. Fisher's representation in Euclidean space

In this section, we identify fitness $f(\omega)$ with the negative distance $-d(\top, \omega)$ from the top element, but later we shall generalise the relation between fitness and the topology of a representation space. Thus, adaptation (beneficial mutation) corresponds to a transition from a sphere of radius $n = d(\top, a)$ into a sphere of a smaller radius $m = d(\top, b)$, which is depicted in Figure 1.

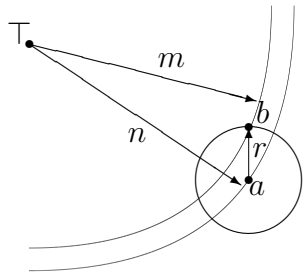


Figure 1: Mutation of a into b by radius $r = d(a, b)$. The distances $n = d(\top, a)$ and $m = d(\top, b)$ from an optimal element $\top$ define the fitnesses of a and b. The intersection of spheres $S(a, r)$ and $S(\top, m)$ define the probability $P(m \mid n, r)$.

This geometric view of mutation and adaptation is based on Ronald Fisher's idea [1], which was, perhaps, the earliest mathematical work on the role of mutation in adaptation. Fisher represented phenotypes by points of Euclidean space $\mathcal{H} \equiv \mathbb{R}^l$ of $l \in \mathbb{N}$ traits, and therefore equipped Ω with Euclidean metric $d_E(a, b) = \|a - b\|_2$ (here $\|\cdot\|_2$ denotes the standard ℓ_2-norm in $\mathbb{R}^l$). The top element $\top$ was identified with the origin in $\mathbb{R}^l$, and fitness $f(\omega)$ with the negative distance $-d_E(\top, \omega)$. Then Fisher used geometry of the Euclidean space to show that probability of beneficial mutation decreases

exponentially as mutation radius increases, and therefore mutations of small radii are more likely to be beneficial. Despite subsequent development of the theory [2], the use of Euclidean space for representation was not revised.

Euclidean space is infinite, and the interior of any ball has always smaller volume than its exterior. Therefore, assuming mutation in random directions, an organism on the surface of a ball around an optimum is always more likely to mutate into the exterior than the interior of this ball. This obvious and simple property is key for the conclusion that adaptation is more likely to occur by small mutations. Recently, we showed that the geometry of a finite space, such as the Hamming space of sequences, implies a different relation between the radius of mutation and adaptation [26, 27]. In particular, mutation radius maximising the probability of adaptation varies as a function of the distance to the optimum.

2.3. Probability of adaptation and representation in a Hamming space

One of the most common examples of a finite metric space is a Hamming space of sequences. Let us denote by $\mathcal{H}_\alpha^l := \{1, \ldots, \alpha\}^l$ the set of all sequences of letters from a finite alphabet $\{1, \ldots, \alpha\}$ and length l. The alphabet can be equipped with operations of addition and multiplication such that it becomes a finite field $GF(\alpha)$ (a Galois field) and $\mathcal{H}_\alpha^l$ becomes a linear algebra over $GF(\alpha)$. A linear algebra is also a vector space, and $\mathcal{H}_\alpha^l$ is an example of a finite vector space (there are α^l points in $\mathcal{H}_\alpha^l$). The space $\mathcal{H}_\alpha^l$ can be equipped with the Hamming metric $d_H(a, b) := |\{i : a_i \neq b_i\}|$ counting the number of different letters. The Hamming metric can also be defined as $d_H(a, b) := \|a - b\|_H$, where $\| \cdot \|_H : \mathcal{H}_\alpha^l \to \{0, 1, \ldots, l\}$ is the Hamming weight counting the number of letters in a sequence not equal to the additive unit of the field $GF(\alpha)$ (zero of the field). Thus, addition of sequences results in a substitution of some letters, which corresponds to a simple mutation, and the Hamming distance counts the number of substitutions.

Consider mutation of sequence $a \in S(\top, n)$ into sequence $b \in S(\top, m)$ by radius $r = d_H(a, b)$, as shown on Figure 1. Assuming equal probabilities for all sequences in the sphere $S(a, r)$, the probability that the offspring sequence is in the sphere $S(\top, m)$ is given by the number of elements in the intersection of spheres $S(\top, m)$ and $S(a, r)$:

$$P(m \mid n, r) = \frac{|S(\top, m) \cap S(a, r)|_{d(\top, a)=n}}{|S(a, r)|} \tag{1}$$

where $|\cdot|$ denotes cardinality of a set (the number of its elements). The cardinality of the intersection $S(\top, m) \cap S(a, r)$ with condition $d(\top, a) = n$ is computed as follows

$$|S(\top, m) \cap S(a, r)|_{d(\top,a)=n} = \sum_{\substack{r_0+r_-+r_+=\min\{r,m\} \\ r_+-r_-=n-\max\{r,m\}}} (\alpha-2)^{r_0} \binom{n-r_+}{r_0} (\alpha-1)^{r_-} \binom{l-n}{r_-} \binom{n}{r_+} \quad (2)$$

The summation runs over indexes r_0, r_- and r_+ satisfying conditions $r_0 + r_- + r_+ = \min\{r, m\}$ and $r_+ - r_- = n - \max\{r, m\}$. These conditions follow from the triangle inequalities for r, m and n, such as

$$|n - m| \leq r \leq n + m$$

When $r \leq m$ then r_0, r_- and r_+ count respectively the numbers of neutral, deleterious and beneficial substitutions in $r \in [0, l]$. They also satisfy the following constraints $r_- \in [0, \lfloor (r+m-n)/2 \rfloor]$ and $r_+ \in [0, \lfloor (n-|r-m|)/2 \rfloor]$, where $\lfloor \cdot \rfloor$ denotes the floor operation.

The cardinality of sphere $S(a, r) \subset \mathcal{H}_\alpha^l$ is

$$|S(a, r)| = (\alpha - 1)^r \binom{l}{r} \quad (3)$$

Equations (1)-(3) allow us to compute the probability of adaptation in the Hamming space $\mathcal{H}_\alpha^l$, which is the probability that the offspring is in the interior of ball $B(\top, n)$:

$$P(m < n \mid n, r) = \sum_{m=0}^{n-1} P(m \mid n, r) \quad (4)$$

Figure 2 shows the probability of adaptation in Hamming space H_4^{100} (i.e. alphabet of size $\alpha = 4$ and length $l = 100$) as a function of mutation radius r for different values of $n = d(\top, a)$. One can see that when $n < 75$ (more generally when $n < l(1-1/\alpha)$), the probabilities of adaptation decrease with r, similar to Fisher's conclusion for the Euclidean space. However, for $n = 75$ there is no such decrease, and when $n > 75$ (i.e. for $n > l(1 - 1/\alpha)$), the probability of adaptation actually increases with r. This is due to the fact that, unlike Euclidean space, Hamming space is finite, and the interior of

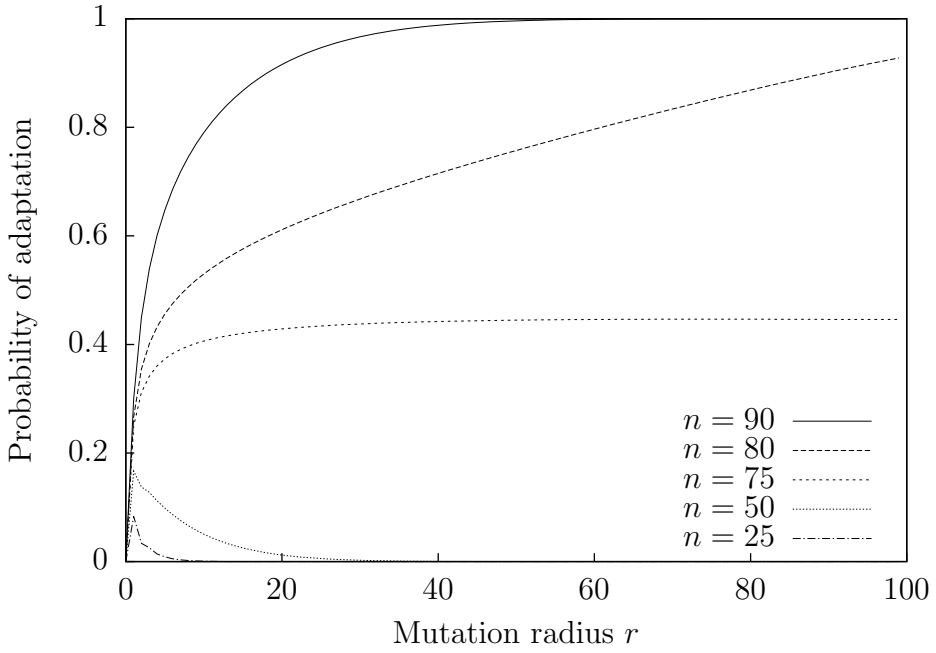


Figure 2: Probability of adaptation $P(m < n \mid n, r)$ in the Hamming space $\mathcal{H}_4^{100}$ as a function of mutation radius r. Different curves show $P(m < n \mid n, r)$ for different distances $n = d_H(\top, a)$ of the parent sequence from the optimum $\top$.

ball $B(\top, n)$ can be larger than its exterior. The geometry of a Hamming space has a number of interesting properties [33]. For example, every point ω has $(\alpha - 1)^l$ diametric opposite points $\neg\omega$, such that $d_H(\omega, \neg\omega) = l$, and the complement of a ball $B(\omega, r)$ in $\mathcal{H}_\alpha^l$ is the union of $(\alpha - 1)^l$ balls $B(\neg\omega, l - r - 1)$.

Remark (Representation space). Using arbitrary alphabets is important not only because DNA molecules are sequences with $\alpha = 4$ bases, but also because it allows us to consider different representations, where the letters of the alphabet may correspond not to DNA base-pairs, but to higher-level structures such as triplets of DNA bases (encoding amino acids) or genes. Changing the representation by considering subsequences of a sequence as letters from an alphabet of a larger size α corresponds to decreasing the length l of the sequence. The Hamming metric, measuring the distance between sequences, takes values in $\{0, \ldots, l\}$, and changing the alphabet and length changes the geometry of the representation space $\mathcal{H}_\alpha^l$.

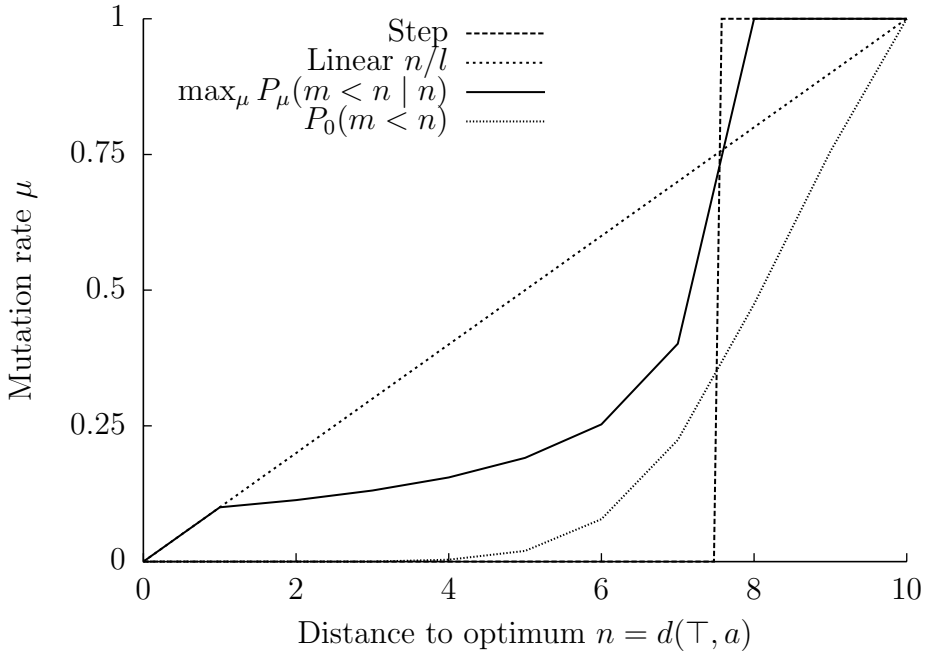


Figure 3: Optimal mutation rate control functions derived mathematically to minimise expected distance to optimum in Hamming space $\mathcal{H}_4^{10}$. Different control functions are optimal solutions to different optimisation problems described in Section 2.5.

Remark (Variable lengths). Hamming metric compares sequences of the same lengths, and it counts the least number of substitutions between a pair of sequences, which is the main mutation mechanism that we consider here. Variable length sequences can be compared using, for example, the Levenshtain metric, which counts the least number of substitutions, insertions and deletions. The space of all variable length finite alphabet sequences is countably infinite, and it can be considered as a vector space over an extended Galois field [34]. Hamming spaces are finite subspaces of this space, and one can consider the set of nested Hamming spaces, where increasing complexity corresponds to an increasing sequence length [27]. We note also that every such finite subspace has a top element $\top$, but the whole space of variable length sequences may fail to have one.

2.4. Random mutation

By mutation of parent sequence a into b we understand a random process, so that the mutation radius is a random variable. The simplest form of

mutation, called *point mutation*, is the random process of independently substituting each letter in a to any of the other $\alpha - 1$ letters with probability μ. At its simplest, with one parameter, there is an equal probability $\mu/(\alpha-1)$ of mutating to each other base. Such mutation corresponds also to additive noise: $b = a + c$, where c is a sequence obtained by point mutation of the origin in $\mathcal{H}_\alpha^l$ (the sequence with all letters equal to zero — the additive unit of the field $GF(\alpha)$). The parameter μ is called *mutation rate*. For point mutation, the probability of mutating by radius $r \in [0, l]$ is given by the binomial distribution:

$$P_\mu(r \mid n) = \binom{l}{r} \mu(n)^r (1 - \mu(n))^{l-r} \tag{5}$$

The expected value and variance of the mutation radius are respectively $\mathbb{E}_\mu\{r\} = l\mu$ and $\sigma^2(r) = l\mu(1-\mu)$. Note in the equation above that we assume that mutation rate μ may depend on the distance $n = d_H(\top, a)$ from the top sequence.

Optimisation of the mutation rate requires knowledge of the probability $P_\mu(m \mid n)$ that the offspring sequence b is in the sphere $S(\top, m)$ that can be expressed as follows:

$$P_\mu(m \mid n) = \sum_{r=0}^{l} P(m \mid n, r)\, P_\mu(r \mid n) \tag{6}$$

Equations (1)–(3) and (5) can be substituted into equation (6) to obtain the precise expression for transition probability $P_\mu(m \mid n)$ in $\mathcal{H}_\alpha^l$.

Remark (Optimal mutation operator). Mutation in biology is much more complex than described above, and its precise mathematical modelling involves many parameters. One parameter point mutation, however, is optimal in a certain sense: it is a solution of one specific variational problem of minimisation of expected distance between points a and b in a Hamming space subject to a constraint on mutual information between a and b. We define and solve this problem in Appendix C. The optimal solutions are conditional probabilities having exponential form $P_\beta(b \mid a) \propto \exp[-\beta\, d_H(a, b)]$, where parameter $\beta > 0$, called the inverse temperature, is related to the constraint on mutual information. Because Hamming metric $d_H(a, b) = \|a - b\|_H$ is computed as the sum $\sum_{i=1}^{l} \delta_{a_i}(b_i)$ of elementary distances $\delta_{a_i}(b_i)$ between letters a_i and b_i in ith position in the sequence, and the values of $\delta_{a_i}(b_i)$ do not depend on the position i, the exponential conditional probability factorises into

the product $P_\beta(b \mid a) \propto \prod_{i=1}^{l} e^{-\beta\,\delta_{a_i}(b_i)}$ corresponding to independent substitution of letters a_i into b_i with equal probabilities $\mu/(\alpha-1)$, where μ is related to the inverse temperature β. Changing the representation space $\mathcal{H}$ and its metric will result in a different optimal mutation operation. For example, if $\mathcal{H}$ is the space of variable length sequences with Levenshtain metric, then optimal mutation $P_\beta(b \mid a)$ will involve independent substitutions, insertions and deletions. If elementary distances $\delta_{a_i}(b_i)$ are different between different pairs of letters, then there will be different parameters for different pairs. If elementary distances depend on the position i in a sequence or the metric $d(a, b)$ is not the sum of elementary distances, then the optimal mutation is a more complex process with non-independent substitutions, insertions or deletions, the phenomenon known in biology as epistasis.

2.5. Optimal control of mutation rates

The fact that we have shown above that the probability of adaptation depends on mutation rate introduces the possibility of organisms maximising the expected fitness of their offspring by controlling mutation rate. The exact form of the optimal mutation rate control functions depends on a number of factors, such as the time horizon. Here we cover the principal elements required, developed in [28].

Let $P_t(a)$ be the distribution of parent sequences in $\mathcal{H}_\alpha^l$ at time t, and let $P_t(n) = \sum_{a:d(\top,a)=n} P_t(a)$ be the distribution of their distances $n = d_H(\top, a)$ from the optimum. Transition probabilities $P(m \mid n)$ define linear transformation of $P_t(n)$ into distribution $P_{t+1}(m)$ of distances $m = d_H(\top, b)$ of their offspring from the optimum:

$$P_{t+1}(m) = \sum_{n=0}^{l} P(m \mid n) P_t(n)$$

If this linear transformation $T(\cdot) := \sum_{n=0}^{l} P(m \mid n)(\cdot)$ does not change with time and assuming that distance to the optimum has Markov property (i.e. distance at $t+1$ depends only on distance at t, but not at $t-1$, $t-2$, etc), then the distribution $P_{t+s}(m)$ after s generations is defined by $T^s(\cdot)$, the sth power of $T(\cdot)$.

According to equation (6) transition probabilities $P_\mu(m \mid n)$ from sphere $S(\top, n)$ to $S(\top, m)$ depend on the mutation rate parameter μ for each distance n from top sequence $\top$, and we call the collection of pairs (n, μ) the

mutation rate *control function* $\mu(n)$. The expressions for the transition probabilities $P_\mu(m \mid n)$ between spheres around optimal element $\top \in \mathcal{H}_\alpha^l$ can be used to optimise this function. This optimisation, however, can be done with respect to different criteria leading to different optimal functions. For example, after one replication, the conditional expected distance to the optimum $\mathbb{E}\{m \mid n\} = \sum_{m=0}^{l} m P_\mu(m \mid n)$ is minimised if the mutation rate μ depends on n according to the following *step function*:

$$\mu(n) = \begin{cases} 0 & \text{if } n < l(1 - 1/\alpha) \\ \frac{1}{2} & \text{if } n = l(1 - 1/\alpha) \\ 1 & \text{otherwise} \end{cases} \tag{7}$$

This function is shown on Figure 3 for Hamming space $\mathcal{H}_4^{10}$. The sudden change of the optimal mutation rate from $\mu = 0$ to $\mu = 1$ at $n = l(1 - 1/\alpha)$ corresponds to the sudden change of the effect of the mutation radius on the probability of adaptation shown on Figure 2. If parent sequences are uniformly distributed $P_t(a) = \alpha^{-l}$ in $\mathcal{H}_\alpha^l$, then mutation of sequences with this control function achieves the greatest decrease $\mathbb{E}\{n\} - \mathbb{E}\{m\} = \sum_{n=0}^{l} n P_t(n) - \sum_{m=0}^{l} m P_{t+1}(m)$ of the expected distance to the optimum. Note, however, that sequences with $n = d_H(\top, a) < l(1-1/\alpha)$ do not mutate. Therefore, if after several generations all sequences are closer to $\top$ than $l(1-1/\alpha)$, then their offspring cannot get closer to $\top$. In the space of binary sequences ($\alpha = 2$) this occurs after only one replication. For this reason, the control of mutation by the step function is not optimal for adaptation in more than one generation.

Deriving a mutation rate control function minimising the expected distance to the optimum after several generations is not a trivial task. However, for a sufficiently large number of generations this problem is equivalent to minimising the expected time at which individuals achieve maximum fitness. The expected convergence times can be computed using techniques for absorbing Markov chains, and numerical methods show that the optimal mutation rate control changes in this case from a step to a more smooth, sigmoid-like function [28].

A simpler but closely related problem is maximisation of probability $P_\mu(b = \top \mid a)$ of mutating directly to the optimum, or maximisation of the probability $P_\mu(m = 0 \mid n)$, which has the following expression:

$$P_\mu(m = 0 \mid n) = (\alpha - 1)^{-n} \mu^n (1 - \mu)^{l-n} \tag{8}$$

Conditions $dP_\mu/d\mu = 0$ and $d^2P_\mu/d\mu^2 \leq 0$ defining the mutation rate maximising this probability lead to the equation $n - l\mu = 0$ and the following linear mutation rate control function shown on Figure 3 for $\mathcal{H}_4^{10}$:

$$\mu(n) = \frac{n}{l} \tag{9}$$

This variation of optimal control functions illustrates the importance of the number of generations (time horizon) for which the expected fitness is maximised, as pointed out previously by Orr [4].

Another approach to mutation rate control is to maximise the probability of adaptation:

$$P_\mu(m < n \mid n) = \sum_{m=0}^{n-1} P_\mu(m \mid n)$$

Bäck obtained the mutation rate function $\mu(n)$ maximising this probability (which he called the probability of *success*) in the space $\mathcal{H}_2^l$ of binary sequences [24]. The expressions from the previous section allow us to obtain similar functions for general Hamming spaces $\mathcal{H}_\alpha^l$. Figure 3 shows this function for $\mathcal{H}_4^{10}$. We note that the comparison $m < n$ used in the probability of adaptation and its maximisation effectively changes fitness from being absolute (i.e. depending only on an individual) to relative (e.g. depending also on the parent of an individual). Indeed, maximisation of $P(m < n \mid n)$ is equivalent to maximisation of the expected value $\mathbb{E}\{f_2(m,n) \mid n\}$ of a two-valued relative fitness function $f_2(m,n) = 1$ if $m < n$; $f_2(m,n) = 0$ otherwise.

Another approach that we pursued elsewhere is based on information theory [27, 29]. In brief, the optimisation of expected fitness is performed subject to constraints on information divergence of distribution $P_{t+1}(m)$ from distribution $P_t(n)$. The resulting optimal mutation rates $\mu(n)$ correspond to cumulative probabilities $P_0(m < n) = \sum_{m=0}^{n-1} P_0(m)$, where $P_0(m)$ is the distribution of $m = d(\top, a)$ assuming uniform distribution $P(\omega) = \alpha^{-l}$ of sequences in $\mathcal{H}_\alpha^l$. Figure 3 shows this function for $\mathcal{H}_4^{10}$. We point out that this control not only achieves a very fast decrease of the expected distance $\mathbb{E}\{m\}$ to the optimum, but the resulting populations also have the smallest variance $\sigma^2(m)$ of the distances.

There are other optimisation criteria, such as maximisation of cumulative expected fitness, that may lead to different optimal control functions. Thus, Figure 3 and this discussion illustrates the fact that there is no single optimal mutation rate control function, but a variety of functions each of which solves

a specific optimisation problem. However, it is also evident from Figure 3 that all these control functions have a common property of monotonically increasing mutation rate with increasing distance of parent sequence from the optimum. Where an evolutionary system optimises a particular criterion, such as one of those considered in this section, on a monotonic landscape, the optimal mutation rate control function will be the corresponding derived function. In Section 3 we shall consider an approximation technique applicable to a more general class of problems including cases where derivation is impractical. In Section 4 we relax the assumption of a monotonic landscape.

3. Evolutionary Optimisation of Mutation Rate Control Functions

Analytical approaches cannot always be applied to derive optimal mutation rate control functions due to high problem complexity. Another approach is to use numerical optimisation or evolutionary techniques to obtain approximately optimal solutions. In this section, we introduce such an evolutionary technique that uses two genetic algorithms. The first, which we refer to as the Inner-GA, evolves sequences with the mutation rate controlled by some function that maps fitness to mutation rate. The second, which we refer to as the Meta-GA, evolves a population of such mutation rate control functions for better performance of the Inner-GA. In this section, we describe the details of these algorithms and report results of experiments. The Inner-GA can use any fitness function. First, we shall apply the techique to the case when fitness of an individual is its negative distance from a selected point in a Hamming space. Later we shall apply the technique to more general non-monotonic fitness landscapes.

3.1. Inner-GA

The Inner-GA is a simple generational genetic algorithm that uses no selection and no recombination. Each genotype in the Inner-GA is a sequence $\omega \in \mathcal{H}_\alpha^l$, and we used populations of 100 individuals. The initial population had equal numbers of individuals at each fitness value, and they were evolved by the Inner-GA for 500 generations using simple point mutation, according to a mutation rate control function specified by the Meta-GA. The fitness can be defined by an arbitrary real function $y = f(\omega)$, and the average fitness $\bar{y}(t)$ of the population is calculated at each generation, in order that expected fitness $\mathbb{E}\{y\}(t)$ may be maximised by the Meta-GA.

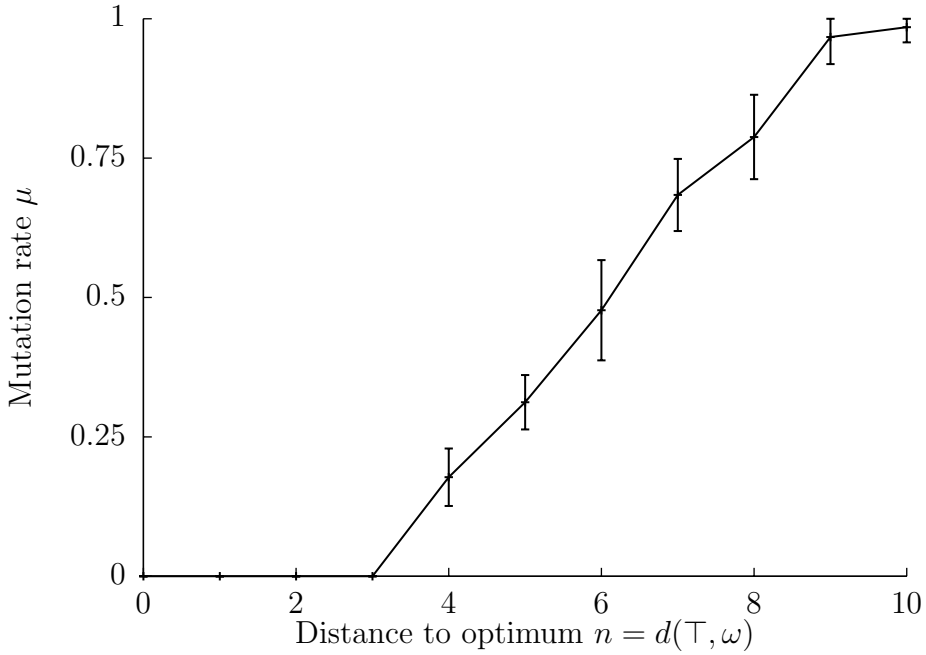


Figure 4: Means and standard deviations of mutation rates for the fittest control functions evolved in each of 20 runs of the Meta-GA using Inner-GAs with individuals in $\mathcal{H}_4^{10}$ evolved to minimise expected distance to the optimum after 500 generations.

3.2. Meta-GA

The Meta-GA is a simple generational genetic algorithm that uses tournament selection (a good choice when little is known or assumed about the structure of the landscape). Each genotype in the Meta-GA is a mutation rate function $\mu(y)$ of fitness values y. The domain of $\mu(y)$ is an ordered partition of the range $\{y : f(\omega) = y,\ \omega \in \mathcal{H}_\alpha^l\}$ of the Inner-GA fitness function. Thus, individuals in the Meta-GA are sequences of real values $\mu \in [0, 1]$ representing probabilities of mutation at different fitnesses, as used in the Inner-GA.

At each generation of the Meta-GA, multiple copies of the Inner-GA were evolved for 500 generations, with the mutation rate in each copy controlled by a different function $\mu(y)$ taken from the Meta-GA population. We used populations of 100 individual functions, which were initialised to $\mu(y) = 0$. All runs within the same Meta-GA generation were seeded with the same

initial population of the Inner-GA. The Meta-GA evolved functions $\mu(y)$ for $5 \cdot 10^5$ generations to maximise the average fitness $\bar{y}(t) \approx \mathbb{E}\{y\}(t)$ in the final generation of the Inner-GA.

The Meta-GA used the following selection, recombination and mutation:

- Randomly select three individuals from the population and replace the least fit of these with a mutated crossover of the other two; repeat with the remaining individuals until all individuals from the population have been selected or fewer than three remain.

- Crossover recombines the start of the numerical sequence representing one mutation rate function with the end of another using a single cut point chosen randomly, excluding the possibility of being at either end so that there are no clones.

- Mutation adds a uniform-random number $\Delta\mu \in [-.1, .1]$ to one randomly selected value μ (mutation rate) on the individual mutation rate function but then bounds that value to be within $[0, 1]$.

The Meta-GA returns the fittest mutation rate function $\mu(y)$.

3.3. Evolved control functions

The kind of mutation rate control function the Meta-GA evolves depends greatly on properties of the fitness landscape used in the Inner-GA. In Section 2.5 we showed theoretically that for $f(\omega)$ corresponding to negative distance to optimum $-d_H(\top, \omega)$, the optimal mutation rate increases with $n = d_H(\top, \omega)$. Therefore, the population of mutation rate functions in the Meta-GA should evolve the same characteristics in such a landscape. Figures 4 shows the average and standard deviations of the fittest control functions evolved in 20 runs of the Meta-GA using Inner-GAs with individuals in $\mathcal{H}_4^{10}$ (i.e. $\alpha = 4$, $l = 10$) and fitness defined by $f(\omega) = -d_H(\top, \omega)$. As predicted, the mutation rate increases with $n = d_H(\top, \omega)$. We shall now consider more complex landscapes.

4. Locally and Weakly Monotonic Fitness Landscapes

The logic behind the variation and optimal control of mutation rates described in the previous section was based on the assumption that fitness $f(\omega)$ is isomorphic to negative Hamming distance $-d_H(\top, \omega)$ from the top

sequence, which allowed us to derive optimal control functions using the geometry of the space of sequences. As detailed below, this assumption implies global monotonicity of the fitness landscape, and it is highly unlikely in real biological landscapes, which can be rugged [35]. In this section, we define the concept of a local and weak monotonicity relative to a chosen metric and show that all landscapes are weakly monotonic at least in some small but non-trivial neighbourhood of the top sequence. This relation between fitness and distance allows one to implement a control of mutation rate using feedback from fitness values. We then consider how monotonicity of different landscapes may influence these fitness-based optimal control functions.

4.1. Memoryless communication between fitness and distance

If fitness $y = f(\omega)$ is not isomorphic to the negative distance $n = d_H(\top, \omega)$ from the optimum, then fitness values of the sequences do not provide full information about their distances. Thus, in order to employ the optimal control $\mu(n)$ of mutation rate based on the distance from the top sequence, one has to estimate the distance from fitness values. The estimation of unobserved random variable $n = d_H(\top, \omega_t)$ at time t from a sequence $y_t, y_{t-1}, \ldots, y_0$ of observed random variables is known as the *filtering* problem [36]. Note that generally the observed process $\{y_t\}_{t\geq 0}$ is not Markov (i.e. $P(y_{t+1} \mid y_t, \ldots, y_0) \neq P(y_{t+1} \mid y_t)$), even if the unobserved process $\{n\}_{t\geq 0}$ and the joint process $\{(n, y_t)\}_{t\geq 0}$ are. For this reason, the optimal control of mutation rate should be a function $\mu(y_t, \ldots, y_0)$ of the entire history of observations. It seems unlikely, however, that such a control has biological relevance, as its implementation would require information about fitness values in all previous generations. Instead, we shall consider a control based only on the current fitness value y_t. Our analysis will focus on monotonic properties of the fitness landscape that will allow us to relate transition probability $P_\mu(y_{t+1} \mid y_t)$ between fitness values of the parent and offspring with probability $P_\mu(m \mid n)$ of transitions between spheres of different radii around the optimum. We shall demonstrate that the 'similarity' between these transition probabilities increases as sequences evolve closer to the optimum, and for this reason the optimal control function $\mu(y_t)$ based on the current fitness values should closely resemble the distance-based optimal control function $\mu(n)$ in some neighbourhood of the optimum.

By a *fitness landscape*, we understand it to mean a graph of a function $f \circ \kappa^{-1} : \mathcal{H}^l_\alpha \to \mathbb{R}$ which associates representations $\kappa(\omega) \in \mathcal{H}$ (codes) of individuals with their fitness values $y = f(\omega)$. The landscape defines a joint

distribution $P(y, n)$ of the fitness values $y = f(\omega)$ and distances $n = d_H(\top, \omega)$ from the nearest global optimum. This joint distribution defines conditional probabilities $P(n \mid y)$ and $P(y \mid n)$. Let us consider mutation of sequence a into sequence b, and let us denote by $n = d_H(\top, a)$ and $m = d_H(\top, b)$ their distances from the nearest optimum and by $y_t = f(a)$ and $y_{t+1} = f(b)$ their fitness values. Thus, given sequence b, its fitness and distance values y_{t+1} and m are independent of the parent sequence a. We shall assume further that given distance m, the fitness value y_{t+1} is also independent of distance n: $P(y_{t+1} \mid m, n) = P(y_{t+1} \mid m)$. One can show that this is equivalent to conditional independence of y_{t+1} and y_t given distances m and n: $P(y_{t+1}, y_t \mid m, n) = P(y_{t+1} \mid m)\, P(y_t \mid n)$. The transition probability $P_\mu(y_{t+1} \mid y_t)$ in this case can be expressed as a composition of transition probabilities $P(n \mid y_t)$, $P_\mu(m \mid n)$ and $P(y_{t+1} \mid m)$ in the following way:

$$P_\mu(y_{t+1} \mid y_t) = \sum_{m=0}^{l} \sum_{n=0}^{l} P(y_{t+1} \mid m) P_\mu(m \mid n) P(n \mid y_t)$$

Thus, we assume that the fitness landscape acts as a memoryless communication channel between distances of individuals to the nearest global optimum and their fitness values. The amount of information communicated through this channel defines how 'similar' the conditional probabilities $P_\mu(y_{t+1} \mid y_t)$ and $P_\mu(m \mid n)$ are and how effective a mutation control function $\mu(y)$ is.

If fitness values $y = f(\omega)$ of sequences and their distances $n = d_H(\top, \omega)$ from the nearest global optimum are statistically independent, then $P(n \mid y_t) = P(n)$, $P(y_{t+1} \mid m) = P(y_{t+1})$ and therefore $P_\mu(y_{t+1} \mid y_t) = P(y_{t+1})$. This means that fitness y_{t+1} of the offspring is independent of fitness y_t of its parent, and therefore a control of mutation rate will have *no* effect on fitness of the offspring. On the other hand, if there is a one-to-one correspondence between the fitness values $y = f(\omega)$ and distances $n = d_H(\top, \omega)$ (i.e. there is a bijection $g : \mathbb{R} \to \mathbb{R}$ such that $f(\omega) = g \circ d_H(\top, \omega)$ and $g^{-1} \circ f(\omega) = d_H(\top, \omega)$), then $P_\mu(y_{t+1} \mid y_t) = P_\mu(m = g^{-1}(y_{t+1}) \mid n = g^{-1}(y_t))$, and the optimal mutation rate control function is $\mu \circ g^{-1}(y)$, where $\mu(n)$ is an optimal control function obtained using $P_\mu(m \mid n)$. In particular, the identity $f(\omega) = -d_H(\top, \omega)$ used in previous section is established by $g(\cdot) = -1 \times (\cdot)$. In Appendix A the relationship between transition probabilities $P(y_{t+1} \mid y_t)$ and $P(m \mid n)$ is explained in more detail.

4.2. Monotonicity of fitness landscapes

Let us consider landscapes in which fitness and distance to nearest global optimum are not isomorphic but there is a deterministic mapping between them. Moreover, we shall consider monotonic properties of these mappings, which allow us to clarify notions of 'smooth' or 'rugged' fitness landscapes, used in biological literature. Note that these monotonic properties are relative to (i.e. depend on) the choice of a representation space, its metric d and encoding-decoding scheme. Below we introduce the definitions of various monotonic properties of landscapes which later allow us to analyse rugged biological landscapes and address optimal control of mutation rate in such landscapes.

Definition 1 (Local monotonicity)**.** Let (Ω, d) be a metric space, and let $f : \Omega \to \mathbb{R}$ be a real function. Then, if all a and b inside some ball $B(\omega, \delta)$ satisfy the properties below, we say that:

- d is *locally monotonic* relative to f at ω if:

$$-d(\omega, a) \leq -d(\omega, b) \quad \Longleftarrow \quad f(a) \leq f(b)$$

- f is *locally monotonic* relative to metric d at ω if:

$$-d(\omega, a) \leq -d(\omega, b) \quad \Longrightarrow \quad f(a) \leq f(b)$$

- f and d are *locally isomorphic* at ω if:

$$-d(\omega, a) \leq -d(\omega, b) \quad \Longleftrightarrow \quad f(a) \leq f(b)$$

- We say that d or f are *globally monotonic* (*isomorphic*) at $\top$ relative to each other if the relevant property holds over $B(\top, \delta) \equiv \Omega$.

The three monotonic relations between fitness and distance defined above are illustrated on Figure 5. These cases represent idealised situations, but they help in understanding the properties of real and biologically relevant landscapes. Let us first consider global monotonicity, that is when the monotonic properties hold for the entire Ω.

The monotonic relationships between distance $d(\top, \omega)$ and fitness $f(\omega)$ can be represented by real monotonic functions $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$

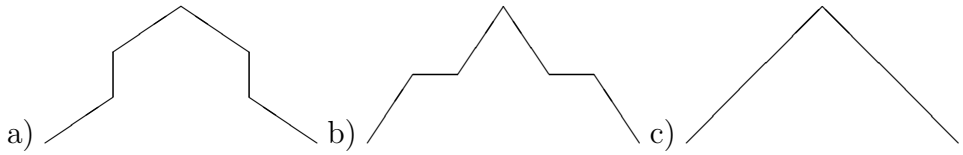


Figure 5: Schematic representation of monotonic properties described in Definition 1. Abscissae represent sequence space, ordinates represent fitness. a) Distance to optimum is monotonic relative to fitness (fitness landscape can have 'cliffs'); b) fitness is monotonic relative to distance to optimum (landscape can have 'plateaus'); c) fitness and distance to optimum are isomorphic (neither cliffs nor plateaus are allowed).

such that $h \circ f(\omega) = d(\top, \omega)$ and $g \circ d(\top, \omega) = f(\omega)$. These mappings are shown in the commutative diagrams in Figure 6. It is clear from the diagrams that mappings h and g are adjoint to encoding κ and decoding κ^{-1} schemes. Thus, for these diagrams to commute, these mappings as well as the representation space with its topology must satisfy certain properties. This represents the fact that monotonicity of fitness and distance (i.e. monotonicity of h and g) is relative to the choice of a representation space, its metric d and encoding-decoding scheme.

$$\begin{array}{ccc} (\Omega, \lesssim) & \xleftarrow{f^{-1}} & (\mathbb{R}, \leq) \\ \downarrow{\scriptstyle \kappa} & & \downarrow{\scriptstyle h} \\ (\mathcal{H}_\alpha^l, \lesssim) & \xrightarrow{-d(\top,\cdot)} & (\mathbb{R}, \leq) \end{array} \qquad \begin{array}{ccc} (\Omega, \lesssim) & \xrightarrow{f} & (\mathbb{R}, \leq) \\ \uparrow{\scriptstyle \kappa^{-1}} & & \uparrow{\scriptstyle g} \\ (\mathcal{H}_\alpha^l, \lesssim) & \xleftarrow{-d^{-1}(\top,\cdot)} & (\mathbb{R}, \leq) \end{array}$$

Figure 6: Commutative diagrams connecting the preference relationship among phenotypes (upper left), fitness values (upper right), distances from the top sequence (lower right) and the representation space (lower left). The arrows give the functions relating these sets including the fitness/inverse-fitness function (top) and encoding-decoding function (left). The *landscape* is the mapping from the representation space (lower left) to the fitness values (upper right). The diagrams show that the relationships h and g between fitness and distance to the optimum impose certain properties on the metric d, the representation space and the encoding-decoding scheme. Note that $d(\top, a)$ is used as short-hand for $d(\kappa(\top), \kappa(a))$; see Section 2.1.

If the metric d is monotonic relative to fitness f, then the distance to optimum is overdetermined, because there are generally more fitness values $f(\omega)$ than spheres $S(\top, n)$ around the optimum (see Figure 5a). This follows directly from the fact that in this case sequences with the same fitness must

have the same distances to the optimum, but not necessarily vice versa (see Proposition 2 in Appendix B). Transition probabilities $P(y_{t+1} \mid y_t)$ between fitness values are easily determined by transition probabilities $P(m \mid n)$ between spheres around $\top$ and monotonic function $h \circ f(\omega) = d(\top, \omega)$ (see Proposition 1 in Appendix A):

$$P(y_{t+1} \mid y_t) = \frac{1}{|h^{-1} \circ h(y_{t+1})|} P(m = h(y_{t+1}) \mid n = h(y_t))$$

where $h^{-1}(y) := \{x : h(x) = y\}$ is the pre-image of y, and cardinality $|h^{-1} \circ h(y)| \geq 1$ represents degeneracy of the mapping h (i.e. the number of fitness values at the same distance from $\top$ as y). Thus, generally $P(y_{t+1} \mid y_t) \leq P(m = h(y_{t+1}) \mid n = h(y_t))$, when distance to optimum is monotonic. In addition, it is easy to show that in the case of a globally monotonic metric there can be only one optimal element. Indeed, applying the definition to $\top_1$ and $\top_2$ we have:

$$f(\top_1) = f(\top_2) \implies d(\top_2, \top_1) = d(\top_2, \top_2) = 0 \iff \top_1 = \top_2$$

The case of distance being overdetermined has little practical interest for our theory. In addition, this property does not allow for fitness plateaus as can be seen from Figure 5a. Such plateaus may be important in biology [37]. It is therefore particularly interesting to look at the case where f is monotonic to d, which allows for plateaus. In this case distance to optimum is underdetermined, because there can be fewer fitness values $f(\omega)$ than spheres $S(\top, n)$ around the optimum (see Figure 5b). It follows directly from the fact that in this case sequences with the same distance from the optimum must have the same fitness values, but not necessarily vice versa (see Proposition 2 in Appendix B). Transition probabilities $P(y_{t+1} \mid y_t)$ between fitness values can be computed from transition probabilities $P(m \mid n)$ between spheres around $\top$ and monotonic function $g \circ d(\top, \omega) = f(\omega)$ (see Proposition 1 in Appendix A):

$$P(y_{t+1} \mid y_t) = \frac{1}{|g^{-1}(y_t)|} \sum_{m \in g^{-1}(y_{t+1})} \sum_{n \in g^{-1}(y_t)} P(m \mid n)$$

One can see that the relation between two transition probabilities is more complicated than in the previous case, and captures a model of 'noisy' communication between fitness and distance simply in the mapping g. The

amount of noise in this case depends on the average degeneracy of the mapping g, represented by the average number of distance values $|g^{-1}(y)|$ corresponding to each fitness value $y = f(\omega)$. The extreme case is a constant fitness function, which has only one value so that all sequences are optimal. A non-trivial example of a highly degenerate landscape is a Boolean landscape, where fitness can have only two values, a situation close to many in biology where a single, non-lethal aspect of the environment is critical for determining fitness (e.g. a nutrient that either can or cannot be utilised, an absent vitamin that is or is not required or, resistance or not to a pathogen or stressor). We now combine the results obtained in Section 2 with those in this section to derive transition probabilities between fitness values on this Boolean landscape where fitness is not isomorphic to distance as in the landscapes used in Section 2 and how this leads on to optimal mutation rate control even in this degenerate case.

Example 1 (Boolean landscapes)**.** Boolean fitness landscape is defined by $f(\omega) = 1$ if $\omega = \top$; $f(\omega) = 0$ otherwise. There can be multiple optima $\top \in \Omega$ with $f(\top) = 1$, and the domain is partitioned into two disjoint subsets $f^{-1}(1) = \{\omega : f(\omega) = 1\}$ and $f^{-1}(0) = \{\omega : f(\omega) = 0\}$. Because there are only two fitness values, there are only four transition probabilities $P(y_{t+1} \mid y_t)$ between them, the most important of which for optimisation purposes is probability $P(y_{t+1} = 1 \mid y_t = 0)$. This probability is related to probability $P(\omega_{t+1} \mid \omega_t)$ of transitions between any two points ω_t, $\omega_{t+1} \in \Omega$ in the following way:

$$P(y_{t+1} = 1 \mid y_t = 0) = \frac{1}{|f^{-1}(0)|} \sum_{\omega_{t+1} \in f^{-1}(1)} \sum_{\omega_t \in f^{-1}(0)} P(\omega_{t+1} \mid \omega_t)$$

One can see that the size of subsets $f^{-1}(1)$ and $f^{-1}(0)$ relative to each other plays an important role, and this characteristic can be used to study different types of Boolean landscapes. When ω are represented by sequences in a Hamming space $\mathcal{H}_\alpha^l$, the probability $P(\omega_{t+1} \mid \omega_t)$ with $d_H(\omega_{t+1}, \omega_t) = n$ is given by equation (8): $P_\mu(\omega_{t+1} \mid \omega_t) = (\alpha - 1)^{-n} \mu^n (1 - \mu)^{l-n}$. This expression can be used to maximise the transition probability above by optimising the mutation rate $\mu(0)$.

4.3. Weak monotonicity

Generally, fitness landscapes may have different local monotonic properties, described above, and the relationship between fitness and distance to

an optimum may not be described by any function, but rather it is non-deterministic, described by conditional probabilities $P(n \mid y_t)$ and $P(y_{t+1} \mid m)$. In this case, we can still define monotonicity in a *weak* sense (i.e. on average) using conditional expected fitness values within spheres of a given radius from point ω:

$$\mathbb{E}\{f \mid n\} = \frac{1}{|S(\omega, n)|} \sum_{a:d(\omega,a)=n} f(a)$$

Definition 2 (Weak local monotonicity)**.** Let (Ω, d) be a metric space, and let $f : \Omega \to \mathbb{R}$ be a real function. Then we call f *weakly locally monotonic* at ω relative to metric d if there exists a ball $B(\omega, \delta)$ such that for all a, b within this ball, the following condition holds:

$$-d(\omega, a) = -n \leq -d(\omega, b) = -m \quad \Longrightarrow \quad \mathbb{E}\{f \mid n\} \leq \mathbb{E}\{f \mid m\}$$

It is not difficult to show that all fitness landscapes are weakly and locally monotonic at $\top$. To see this, assume the opposite, that $\mathbb{E}\{f \mid n\} > \mathbb{E}\{f \mid m\}$ holds for all neighbourhoods of $\top$. Then clearly $\sup f(\omega)$ cannot be attained at $\top$ (i.e. $\top$ is not the optimum). Thus, there must be some neighbourhood $B(\top, \delta)$, containing elements other than $\top$, where weak monotonicity holds. Our analysis in Section 5 suggests that biological landscapes may exhibit weak monotonicity in large neighbourhoods of the optimum.

As discussed previously, if f is locally monotonic relative to d, then spheres $S(\top, \delta)$ cannot contain elements with different values $y = f(\omega)$. This is not true in the case of weak monotonicity. The variation of fitness within the spheres $S(\top, n)$ can be measured by the conditional variance of fitness:

$$\sigma^2(f \mid n) = \frac{1}{|S(\top, n)|} \sum_{\omega:d(\top,\omega)=n} |f(\omega) - \mathbb{E}\{f \mid n\}|^2$$

Clearly, stronger monotonicity implies smaller variance $\sigma^2(f \mid n)$. It is not difficult to see that an increase of expected fitness $\mathbb{E}\{f \mid n\} \to f(\top)$ coincides with a decrease of the variance $\sigma^2(f \mid n) \to 0$. Because of these weak locally monotonic properties of general fitness landscapes, the probabilities of transitions $P(y_{t+1} \mid y_t)$ between fitness values that are close to the optimum $y_t,\ y_{t+1} > f(\top) - \varepsilon$ will be similar to transition probabilities $P(m \mid n)$ between spheres with $n,\ m = d(\top, \omega) < \delta$. Therefore, we formulate the following hypotheses:

Hypothesis 1 Optimal mutation rate increases with a decrease in fitness in some neighbourhood of an optimum for realistic fitness landscapes (e.g. biological landscapes) where fitness is not isomorphic to distance, similar to the monotonic increase in optimal mutation rate derived for the isomorphic case.

Hypothesis 2 Real and biological landscapes exhibit weak monotonicity in large neighbourhoods of an optimum.

Hypothesis 3 The larger the neighbourhood of weak monotonicity, the more mutation rate control may contribute to evolution towards high fitness.

5. Evolving Fitness-Based Mutation Rate Control Functions

To test the relevance of our predictions about the optimal mutation rate control functions more widely in biologically realistic sequence-fitness landscapes, we used the described earlier Meta-GA technique (see Section 3) to evolve approximately optimal functions for 115 published complete landscapes of transcription factor binding [30]. Transcription factors have evolved over very long periods to bind to specific DNA sequences. The landscapes show experimentally measured strengths of interaction (DNA-TF binding score) between the double-stranded DNA sequences of length $l = 8$ of base pairs each and a particular transcription factor. Because these landscapes represent results of direct interaction between the DNA sequences and the transcription factors, the DNA sequences can be thought of as both 'phenotypes' and their codes, which allows us to identify the space Ω of phenotypes with the representation space, which in this case is the Hamming space $\mathcal{H}_4^8$ ($\alpha = 4$, $l = 8$). The DNA-TF binding score, however, which plays the role of fitness, is clearly not identical to the negative Hamming distance of a sequence from the top sequence (a sequence with the maximum DNA-TF binding score). In this section, we show that the mutation rate control functions obtained for these landscapes using evolutionary technique conform well to our theoretical predictions about the optimal mutation rate control.

5.1. Evolved control functions

We used the Meta-GA evolutionary optimisation technique, described in Section 3, to obtain for each landscape the best possible mutation rate control function that maximises the average DNA-TF binding score in the population

(expected fitness) after 500 replications. The Meta-GA algorithm converged within a small margin of statistical error to a specific mutation rate control function in each landscape. To get sufficiently significant results as well as an estimate of the convergence, 16 replicate runs were performed in each of the 115 transcription factor landscapes.

Figure 7 shows the average values and standard deviations of the evolved mutation rates for three transcription factors: Srf, Glis2 and Zfp740. Evolved functions for all landscapes are shown on Figure D.10 in supplementary material. One can see that the evolved functions for each transcription factor landscape is monotonic in the direction predicted: close to zero mutation at the maximum fitness, rising to high levels further from the maximum fitness value. Once the mutation rate has peaked near the maximum value $\mu = 1$, the mutation rates tend to decrease and become chaotic. As will be shown in the next section, this occurs at lower fitness values at which the landscape is no longer monotonic (i.e. further from the peak of fitness). Small standard deviations indicate good convergence to a particular control function. Observe that there is poor convergence at low fitness areas of the landscape that are poorly explored by the genetic algorithm.

5.2. Landscapes for transcription factors

The variation in the evolved mutation rate control function is clearly related to a variation in the properties of the landscapes. Our theoretical analysis suggests that the main property affecting the mutation rate control is monotonicity of the landscape relative to a metric measuring the mutation radius. In particular, the radius of point-mutation is measured by the Hamming metric, and we shall look into the local and weak monotonic properties of the transcription factors landscapes relative to the Hamming metric.

Figure 8 shows average DNA-TF binding scores within spheres $S(\top, n)$ around the optimal sequence as a function of Hamming distance $n = d_H(\top, \omega)$ from the optimum. Data is shown for three transcription factors: Srf, Glis2 and Zfp740. Lines connect average values at discrete distances for visualisation purpose. Errorbars show standard deviations of the DNA-TF binding scores within the spheres. Distributions of fitness with respect to Hamming distance $d_H(\top, \omega)$ for all 115 transcription factors are shown on Figure D.11 (supplementary material).

One can see from Figure 8 that the landscape for the Srf factor has monotonic properties: The average values increase steadily for sequences that are closer to the optimum, and the deviations from the mean within the

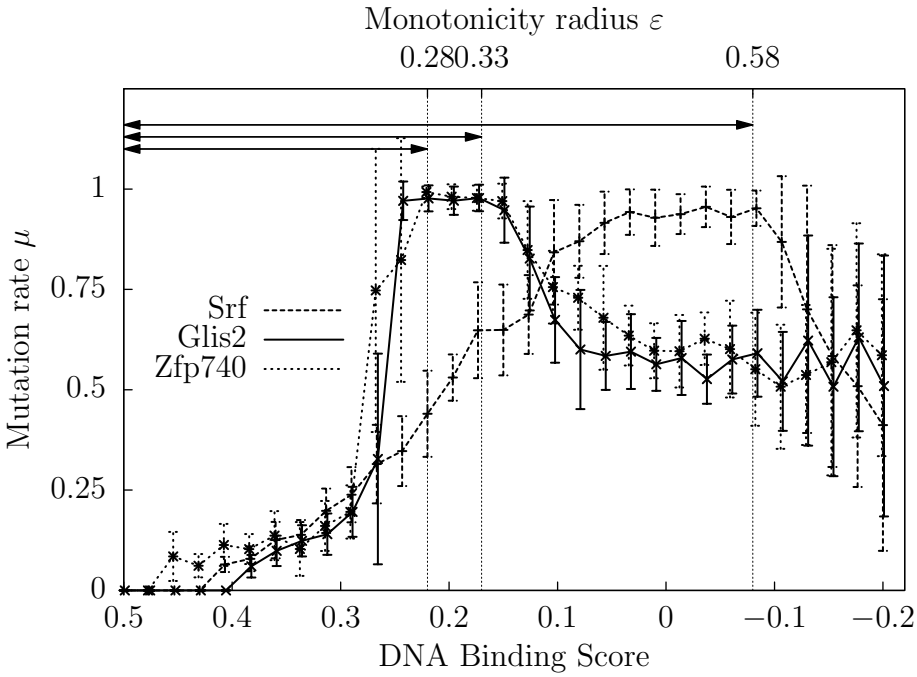


Figure 7: Examples of GA-evolved optimal mutation rate control functions. Data are shown for the transcription factors Srf, Glis2 and Zfp740. Each curve represents the average of 16 independently evolved optimal mutation rate functions on a particular transcription factor DNA-binding landscape [30]. Errorbars represent standard deviations from the mean. Similar curves for all 115 landscapes are shown in supplementary Fig. D.10. The arrows indicate the *edge of monotonicity* ε, that defines an interval of fitness values below the maximum, where mutation rate monotonically increases.

spheres are relatively small. This is in contrast to the other two landscapes. We note also that the average values for Glis2 decrease quite sharply around the optimum, while the landscape for Zfp740 has a relatively flat plateau area around the optimum, which means that there are many sequences with high DNA-TF binding score. This difference may explain different gradients of optimal mutation rates near the maximum fitness shown on Figure 7.

5.3. Monotonicity and controllability

Our results have confirmed that the evolved optimal mutation rates rise from zero to very high levels as fitness decreases from the maximum value $f(\top)$ to some value $f(\top) - \varepsilon$ (see supplementary Fig. D.10). We refer to the corresponding value $\varepsilon > 0$ as the *monotonicity radius*, as it defines the

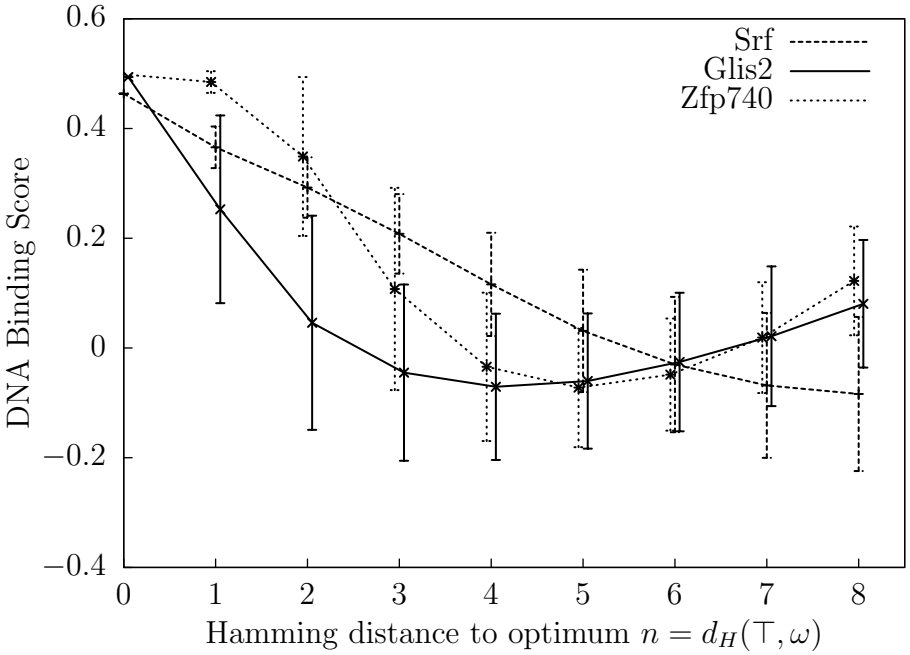


Figure 8: Examples of fitness landscapes based on the binding score between DNA sequences and transcription factors (TF) from [30]. Data are shown for the transcription factors: Srf, Glis2 and Zfp740. Lines connect mean values of the binding score shown as functions of the Hamming distance from the top sequence (a sequence with the highest DNA-TF binding score). Errorbars represent standard deviations. Similar curves for all 115 landscapes are shown in supplementary Fig. D.11.

neighbourhood of $\top$ in terms of fitness values in which the evolved mutation rate control function has monotonic properties. We find substantial variation in monotonicity radius among transcription factors (see Fig. 7 and supplementary Fig. D.10)

We hypothesised that the variation in the optimal mutation rate control functions relates to variation in the monotonicity of the transcription factor landscapes. Various measures have been proposed for the roughness of biological landscapes [35]. Here we focus on the Kendall's τ correlation, which is directly concerned with monotonicity; specifically, τ measures the proportion of mutations that, in moving closer to the optimum in sequence space, also increase in fitness. As shown in Figure 9, we find that τ of the landscape does indeed have a relationship with the monotonicity radius ε of the evolved mutation rate control functions (Spearman's $\rho = 0.77$, $P \approx 10^{-16}$, $N = 115$).

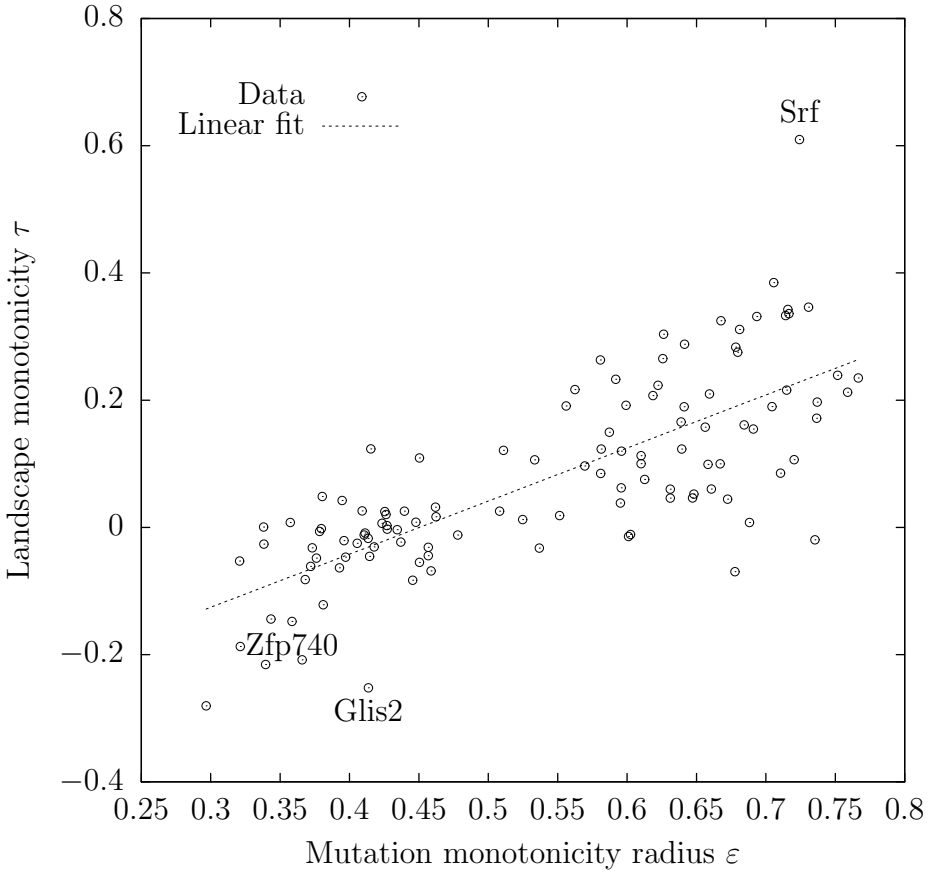


Figure 9: Linear relation between monotonicity of the landscapes measured by the Kendall's τ correlation (ordinates) and the edge of monotonicity ε (abscissae) of the corresponding evolved mutation rate control functions. Three labels show data for three transcription factors shown in Figs. 7 and 8.

Finally, we hypothesise that these related features of the transcription factor landscape and mutation rate function themselves relate to the biological evolution of this transcription factor system. To test this we looked at the evolutionary age of transcription factor families [38]. We find the suggestion of a relationship between the monotonicity of a landscape (τ) and the age of the transcription factor family, implying that the more recently a transcription factor family evolved, the more monotonic is its landscape (Spearman's

$\rho = 0.23$, $P = 0.061$, $N = 115$). We find a more substantial relationship between this evolutionary age and the monotonicity radius ε (Spearman's $\rho = 0.36$, $P = 0.0032$, $N = 115$).

6. Discussion

In this paper we have developed and tested theory relating to the control of mutation rate in biological sequence landscapes. To do so, we had to move the theory closer to the biology in three ways. Firstly (in Section 2), we generalised Fisher's geometric model of adaptation, from its Euclidean space (continuous and infinite) to discrete, finite Hamming spaces of sequences. Doing so demonstrated that, in contrast to the behaviour in Euclidean space, where the probability of beneficial mutation behaves similarly at different distances from the optimum [39], the probability of beneficial mutation, for a given mutation size, varies markedly depending on the distance from the optimum (Figure 2). Secondly, we analytically derived functions for optimal control of the mutation rate minimising the expected Hamming distance to a particular point (optimal sequence). We demonstrated also a variation of these control functions dependent on specific formulations of the optimisation problem. Nonetheless we observed consistency: all optimal functions increase monotonically (Figure 3). Thirdly, we developed theory concerning locally and weakly monotonic landscapes, demonstrating that all possible landscapes, including biologically rugged landscapes, can be included in these categories and thus that, at some level, our theoretical findings regarding mutation rate control may be applied to realistic biological landscapes. The most striking differences from existing theory in Euclidean spaces occur when sequences are short and far from a peak. We therefore used transcription factors binding to DNA sequences [30] as a test case, which involves both short sequences (eight base-pairs) and highly evolved binding specificity (i.e. we expect that many sequences will bind much more weakly than the best). We tested hypotheses arising from the theory, relating to the nature of optimal mutation rate functions (Hypothesis 1 and Figure 7), the monotonic properties of landscapes (Hypothesis 2 and Figure 8) and the relationship between the two (Hypothesis 3 and Figure 9). In each case we find evidence to support the hypothesis, implying that our theory is relevant to these biological landscapes.

We have considered the possibility of varying the general mutation rate for a single genotype, that is *mutation rate plasticity*, and identified forms that such plasticity may be expected to take as a function of fitness in biological

fitness landscapes. This raises a number of important questions about how this theory might relate to living organisms. The primary question is whether such control of mutation rate plasticity actually occurs in nature. Variation in mutation rate is well known, and organisms with a genetically encoded raised mutation rate, termed mutators, are found at appreciable frequencies in various real populations, apparently via their association (that is 'hitch-hiking') with beneficial mutations [40, 41]. Mutation rate plasticity is a more subtle effect than simply being a mutator. However, as with the evolution of mutators, for mutation rate control to have evolved at all might be expected to require so-called 'second-order selection', that is selection not directly on a trait's effect on an individual's fitness, but indirectly via the genetic effects it produces [42]. While rare, there are clear examples of second-order selection occurring in biology [43], and in our more abstracted system of genetic algorithms we do see mutation rate plasticity rapidly evolving to particular forms (Figure 7). This implies that mutation rate control of the sort we have considered may reasonably be hypothesised in biology.

Most existing discussion of mutation rate plasticity in nature concerns the observed phenomenon of stress-induced mutagenesis [44]. It has long been postulated, and most recently argued from a population genetic model [15], that such plasticity might indeed be adaptive. Such adaptationist hypotheses for stress-induced mutation have been subject to protracted debate [45], but here are present two difficulties. First, it is necessary to exclude alternative, non-adaptive hypotheses. For instance, it seems likely that a raised mutation rate could be a physiologically unavoidable direct effect of stress. This has long been speculated, for instance Muller remarked that the kinetics of temperature's effect on mutation rate resembles that of an ordinary chemical reaction [46]. Second, there needs to be a connection between the imposed or measured variable, stress, and the variable considered by the theory, (inverse) fitness. The first difficulty is ameliorated by the development of explicit theory around non-adaptive hypotheses of mutation rate variation [13]. However, this population genetic theory is currently defined as an alternative to physiological hypotheses of mutation rate variation, whereas real organisms experience both physiological and population-genetic constraints. Integrating the two would help understand what might be expected in terms of stress-induced mutagenesis without recourse to adaptive hypotheses. Regarding the second issue, the connection between stress and fitness, 'stress', as typically defined, can be difficult to separate from 'normal' physiological processes [47]. This means that stress is not a simple inverse

of fitness. Indeed, stress may actually be associated with increased fitness (e.g. in the phenomenon of hormesis [48]). Therefore, while it is possible that stress-induced mutagenesis is an example of mutation rate control as discussed in this paper, further work is required to clarify how exactly the theory relates to that example and perhaps to look for new examples of mutation rate control.

Given the current uncertainties about the existence of mutation rate control in nature, it is important to ask whether, nonetheless, mechanisms exist whereby the processes discussed in this paper *could* be exercised. The very existence of mutator phenotypes demonstrates that, physiologically, increasing mutation rates from the low values typical of biology is possible. If it is possible via genetic change in mutators, it seems highly likely also to be possible in a controlled way via plastic changes. Indeed, several different mechanisms for modulating mutation rates have been proposed, notably by regulating particular DNA repair mechanisms [49, 50, 51] or up regulating mutagenic repair [52, 53, 54].

While regulation of mutation rate is mechanistically feasible, a more challenging issue for the relevance of the theory presented here is whether feedback mechanisms exist for an individual organism to assess its own fitness against which to set its mutation rate. Stress is one indicator that may be assessed by an individual and is known to induce regulatory responses (e.g. the SOS response in bacteria [55]), but as discussed above, stress may not be a clear indicator of fitness. We propose three possible alternative feedback mechanisms, assessing either absolute or relative fitness. Absolute fitness is the scale used in the theory developed in this paper and concerns the number of offspring left in subsequent generations. For some organisms it may be possible to assess absolute fitness by assessing their own reproductive period relative to an internal or external clock. It is notable that one of the best characterised examples of stress-induced mutation [14] actually relates to mutagenesis in ageing bacterial colonies (MAC) and ageing may be an appropriate biological clock for this mechanism, one that is known to be associated with mutation rates in human males [56]. Secondly, for organisms with limited dispersal rates, the number of live organisms of the same genotype in the near vicinity may be a proxy for absolute fitness. Thirdly, while the fitness scale we have worked with is absolute, we have demonstrated elsewhere [26, 27] that appropriate mutation rate functions may be approximated by the cumulative distribution function of the population fitnesses through evolution (also shown in Figure 3). That is, information about an

organism's fitness relative to others in its population could in principle act as feedback, allowing an individual to set its mutation rate in a good approximation to what would be optimal if absolute fitness were known. These latter two mechanisms raise the intriguing possibility that population-level or social effects could be important in determining individual mutation rates. Testing which, if any, of these processes actually occurs in biology will give important insights into evolutionary mechanisms.

We have focused on fitness-associated control of mutation rate. However, mutation is only one evolutionary process where fitness-associated control may be beneficial. Recombination and dispersal are also evolutionary processes that may be under the control of the individual and therefore open to similar effects. Fitness-associated recombination has been demonstrated to be advantageous theoretically [57, 58] and identified in biology [59, 60]. Similarly, the idea that dispersal associated with low fitness might be advantageous has a basis in simulation of spatially differentiated populations [61, 62]. This association might perhaps be framed more generally in terms of 'fitness associated dispersal'. Thus the framework for control of mutation rate in response to fitness that we have developed here may in future be applicable to both recombination and dispersal.

To conclude, our development of theory and testing its predictions *in silico* not only clarifies ideas around the monotonicity of fitness landscapes and mutation rate control, it leads directly to questions testable in living organisms. At the same time there is the potential for greater insight through further development of the theory. Three directions seem particularly likely to be fruitful.

First, while it is striking how effective mutation rate control is for adaptive evolution without invoking selection in our *in silico* experiments, it will be important to consider the role of a selection strategies. Such strategies may implicitly modify fitness functions. For instance, one of the analytically derived functions shown in Figure 3 is the mutation rate function for a DNA space ($\mathcal{H}_4^{10}$) which maximises the probability of adaptation (as derived by Bäck for binary sequences [24]). As outlined in the corresponding section, maximising the probability of adaptation is equivalent to maximising expected fitness of the offspring relative to its parent. This effect may be implicit in a selection strategy that removes the offspring of reduced fitness that will inevitably be produced by maximising offspring expected fitness. Given the importance of selection in biology, we therefore anticipate that such functions may be closer to putative mutation rate control functions in

living organisms. This requires further work.

A second area for development is in variable adaptive landscapes. The importance of time-varying adaptive landscapes in biological evolution is becoming increasingly appreciated [63, 64] and variable mutation rates have a particular role here [65]. It is worth noticing however that our derivation of optimal mutation rate functions is *not* dependent on a fixed landscape, as it depends only on the fitness values. Nonetheless, as we demonstrate for the transcription factor landscapes, variation in landscapes' monotonic properties relates to the shape of mutation rate functions in predictable ways (Figure 9). This deserves further exploration both theoretically and empirically: measuring variation in the monotonic properties of real biological landscapes will be informative about optimal mutation rate functions and *vice versa*.

Finally, there is potential to develop theory around the role of encoding-decoding schemes. Landscape monotonicity, as explored here, is not absolute; it depends on the encoding-decoding scheme (see Figure 6). That is, if the encoding changes, it may be possible to convert a non-monotonic landscape into a monotonic one. Biology uses a variety of such encoding schemes which may themselves evolve. For the transcription factor landscapes used here, the encoding-decoding scheme is defined by the biochemical interactions between the transcription factor (a protein molecule) and DNA. Thus, evolution of transcription factors constitutes evolution of the encoding-decoding scheme, and indeed we do find a relationship between that evolution (age of families) and the monotonic properties of the associated landscapes. A more familiar example of a biological encoding-decoding scheme is the genetic code where there is much existing work on its evolution (e.g. [66]). Determining how evolution of such codes affects the monotonic properties of biological landscapes as explored here may therefore provide novel insights into large-scale evolutionary patterns. Ultimately, theory such as this that identifies analytically or empirically optimal mutation rate control functions may help make predictions about evolutionary responses to future environmental change [67] or inferences about the environment(s) within which particular organisms evolved. In the meantime mutation rate control as developed here will assist directed evolution within biological and other complex landscapes, for instance in the evolution of DNA-protein binding [68].

References

[1] R. A. Fisher, The Genetical Theory of Natural Selection, Oxford University Press, Oxford, 1930.

[2] H. A. Orr, The genetic theory of adaptation: a brief history, Nat Rev Genet 6 (2) (2005) 119–27.

[3] M. Kimura, Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods, Proceedings of the National Academy of Sciences 77 (1) (1980) 522–526.

[4] H. A. Orr, The population genetics of adaptation: the distribution of factors fixed during adaptive evolution, Evolution 52 (4) (1998) 935–949.

[5] J. M. Smith, Natural selection and concept of a protein space, Nature 225 (5232) (1970) 563–564.

[6] H. A. Orr, The population genetics of adaptation: the adaptation of DNA sequences, Evolution 56 (7) (2002) 1317–30.

[7] M. J. McDonald, T. F. Cooper, H. J. Beaumont, P. B. Rainey, The distribution of fitness effects of new beneficial mutations in pseudomonas fluorescens, Biol Lett 7 (1) (2011) 98–100.

[8] T. Bataillon, T. Zhang, R. Kassen, Cost of adaptation and fitness effects of beneficial mutations in pseudomonas fluorescens, Genetics 189 (3) (2011) 939–49.

[9] R. Kassen, T. Bataillon, Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria, Nat Genet 38 (4) (2006) 484–8.

[10] D. R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, H. A. Wichman, Beneficial fitness effects are not exponential for two viruses, Journal of Molecular Evolution 67 (4) (2008) 368–76.

[11] J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, Rates of spontaneous mutation, Genetics 148 (4) (1998) 1667–86.

[12] J. W. Drake, A constant rate of spontaneous mutation in DNA-based microbes, Proceedings of the National Academy of Sciences of the United States of America 88 (16) (1991) 7160–4.

[13] M. Lynch, Evolution of the mutation rate, Trends in genetics : TIG 26 (8) (2010) 345–52.

[14] I. Bjedov, O. Tenaillon, B. Gerard, V. Souza, E. Denamur, M. Radman, F. Taddei, I. Matic, Stress-induced mutagenesis in bacteria, Science 300 (5624) (2003) 1404–9.

[15] Y. Ram, L. Hadany, The evolution of stress-induced hypermutation in asexual populations, Evolution Int J Org Evolution 66 (7) (2012) 2315–28.

[16] A. E. Eiben, R. Hinterding, Z. Michalewicz, Parameter control in evolutionary algorithms, IEEE Transactions on Evolutionary Computation 3 (2) (1999) 124–141.

[17] G. Ochoa, Setting the mutation rate: Scope and limitations of the $1/l$ heuristics, in: Proceedings of Genetic and Evolutionary Computation Conference (GECCO-2002), Morgan Kaufmann, San Francisco, CA, 2002, pp. 315–322.

[18] I. D. Falco, A. D. Cioppa, E. Tarantino, Mutation-based genetic algorithm: performance evaluation, Applied Soft Computing 1 (4) (2002) 285–299.

[19] J. Cervantes, C. R. Stephens, "optimal" mutation rates for genetic search, in: M. Cattolico (Ed.), Proceedings of Genetic and Evolutionary Computation Conference (GECCO-2006), ACM, Seattle, Washington, USA, 2006, pp. 1313–1320.

[20] F. Vafaee, G. Turán, P. C. Nelson, Optimizing genetic operator rates using a Markov chain model of genetic algorithms, ACM, 2010, pp. 721–728.

[21] D. H. Ackley, An empirical study of bit vector function optimization, in: L. Davis (Ed.), Genetic Algorithms and Simulated Annealing, Pitman, 1987, Ch. 13, pp. 170–204.

[22] T. C. Fogarty, Varying the probability of mutation in the genetic algorithm, in: J. D. Schaffer (Ed.), Proceedings of the 3rd International Conference on Genetic Algorithms, Morgan Kaufmann, 1989, pp. 104–109.

[23] M. Yanagiya, A simple mutation-dependent genetic algorithm, in: S. Forrest (Ed.), Proceedings of the 5th International Conference on Genetic Algorithms, Morgan Kaufmann, 1993, p. 659.

[24] T. Bäck, Optimal mutation rates in genetic search, in: S. Forrest (Ed.), Proceedings of the 5th International Conference on Genetic Algorithms, Morgan Kaufmann, 1993, pp. 2–8.

[25] A. E. Nix, M. D. Vose, Modeling genetic algorithms with Markov chains, Annals of Mathematics and Artificial Intelligence 5 (1) (1992) 77–88.

[26] R. V. Belavkin, A. Channon, E. Aston, J. Aston, C. G. Knight, Theory and practice of optimal mutation rate control in Hamming spaces of DNA sequences, in: T. Lenaerts, M. Giacobini, H. Bersini, P. Bourgine, M. Dorigo, R. Doursat (Eds.), Advances in Artificial Life, ECAL 2011: Proceedings of the 11th European Conference on the Synthesis and Simulation of Living Systems, MIT Press, 2011, pp. 85–92.

[27] R. V. Belavkin, Mutation and optimal search of sequences in nested Hamming spaces, in: IEEE Information Theory Workshop, IEEE, 2011.

[28] R. V. Belavkin, Dynamics of information and optimal control of mutation in evolutionary systems, in: A. Sorokin, R. Murphey, M. T. Thai, P. M. Pardalos (Eds.), Dynamics of Information Systems: Mathematical Foundations, Vol. 20 of Springer Proceedings in Mathematics and Statistics, Springer, 2012, pp. 3–21.

[29] R. V. Belavkin, Minimum of information distance criterion for optimal control of mutation rate in evolutionary systems, in: L. Accardi, W. Freudenberg, M. Ohya (Eds.), Quantum Bio-Informatics V, QP-PQ: Quantum Probability and White Noise Analysis, World Scientific, Forthcoming.

[30] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen,

H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, M. L. Bulyk, Diversity and complexity in DNA recognition by transcription factors, Science 324 (5935) (2009) 1720–3.

[31] J. von Neumann, O. Morgenstern, Theory of games and economic behavior, 1st Edition, Princeton University Press, Princeton, NJ, 1944.

[32] H. A. Orr, Fitness and its role in evolutionary genetics, Nat Rev Genet 10 (8) (2009) 531–539.

[33] R. Ahlswede, G. Katona, Contributions to the geometry of Hamming spaces, Discrete Mathematics 17 (1) (1977) 1–22.

[34] R. Sánchez, R. Grau, Vector space of the extended base-triplets over the Galois field of five DNA bases alphabet, International Journal of Biological and Life Sciences 3 (2) (2007) 89–96.

[35] A. E. Lobkovsky, Y. I. Wolf, E. V. Koonin, Predictability of evolutionary trajectories in fitness landscapes, PLoS Computational Biology 7 (12) (2011) e1002302.

[36] R. L. Stratonovich, On the theory of optimal non-linear filtration of random functions, Theory of Probability and its Applications 4 (1959) 223–225, English translation.

[37] A. Wagner, Neutralism and selectionism: a network-based reconciliation, Nature Reviews Genetics 9 (12) (2008) 965–74.

[38] M. T. Weirauch, T. R. Hughes, A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution, in: T. R. Hughes (Ed.), A Handbook of Transcription Factors, Vol. 52 of Subcellular biochemistry, Springer, 2011, pp. 25–73.

[39] H. A. Orr, The distribution of fitness effects among beneficial mutations, Genetics 163 (4) (2003) 1519–26.

[40] F. Taddei, M. Radman, J. Maynard-Smith, B. Toupance, P. H. Gouyon, B. Godelle, Role of mutator alleles in adaptive evolution, Nature 387 (6634) (1997) 700–2.

[41] P. D. Sniegowski, P. J. Gerrish, R. E. Lenski, Evolution of high mutation rates in experimental populations of E. coli, Nature 387 (6634) (1997) 703–5.

[42] O. Tenaillon, F. Taddei, M. Radmian, I. Matic, Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation, Research in Microbiology 152 (1) (2001) 11–16.

[43] R. J. Woods, J. E. Barrick, T. F. Cooper, U. Shrestha, M. R. Kauth, R. E. Lenski, Second-order selection for evolvability in a large Escherichia coli population, Science 331 (6023) (2011) 1433–6.

[44] R. S. Galhardo, P. J. Hastings, S. M. Rosenberg, Mutation as a stress response and the regulation of evolvability, Critical Reviews in Biochemistry and Molecular Biology 42 (5) (2007) 399–435.

[45] O. Tenaillon, E. Denamur, I. Matic, Evolutionary significance of stress-induced mutagenesis in bacteria, Trends in microbiology 12 (6) (2004) 264–70.

[46] H. Muller, The measurement of gene mutation rate in Drosophila, its high variability, and its dependence upon temperature, Genetics 13 (4) (1928) 279.

[47] J. M. Koolhaas, A. Bartolomucci, B. Buwalda, S. F. de Boer, G. Flugge, S. M. Korte, P. Meerlo, R. Murison, B. Olivier, P. Palanza, G. Richter-Levin, A. Sgoifo, T. Steimer, O. Stiedl, G. van Dijk, M. Wohr, E. Fuchs, Stress revisited: a critical evaluation of the stress concept, Neuroscience and biobehavioral reviews 35 (5) (2011) 1291–301.

[48] D. Costantini, N. B. Metcalfe, P. Monaghan, Ecological processes in a hormetic framework, Ecology Letters 13 (11) (2010) 1435–47.

[49] G. Feng, H. C. Tsui, M. E. Winkler, Depletion of the cellular amounts of the MutS and MutH methyl-directed mismatch repair proteins in stationary-phase Escherichia coli K-12 cells, Journal of Bacteriology 178 (8) (1996) 2388–96.

[50] J. W. Drake, Avoiding dangerous missense: thermophiles display especially low mutation rates, PLoS Genetics 5 (6) (2009) e1000520.

[51] A. Deem, A. Keszthelyi, T. Blackgrove, A. Vayl, B. Coffey, R. Mathur, A. Chabes, A. Malkova, Break-induced replication is highly inaccurate, PLoS Biology 9 (2) (2011) e1000594.

[52] R. G. Ponder, N. C. Fonville, S. M. Rosenberg, A switch from high-fidelity to error-prone DNA double-strand break repair underlies stress-induced mutation, Molecular Cell 19 (6) (2005) 791–804.

[53] A. Slack, P. C. Thornton, D. B. Magner, S. M. Rosenberg, P. J. Hastings, On the mechanism of gene amplification induced under stress in escherichia coli, PLoS Genetics 2 (4) (2006) e48.

[54] S. van der Veen, T. Abee, Generation of variants in listeria monocytogenes continuous-flow biofilms is dependent on radical-induced DNA damage and RecA-mediated repair, PLoS ONE 6 (12) (2011) e28590.

[55] J. Courcelle, A. Khodursky, B. Peter, P. O. Brown, P. C. Hanawalt, Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli, Genetics 158 (1) (2001) 41–64.

[56] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. W. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, K. Stefansson, Rate of de novo mutations and the importance of father's age to disease risk, Nature 488 (7412) (2012) 471–475.

[57] L. Hadany, T. Beker, On the evolutionary advantage of fitness-associated recombination, Genetics 165 (4) (2003) 2167–79.

[58] A. F. Agrawal, L. Hadany, S. P. Otto, The evolution of plastic recombination, Genetics 171 (2) (2005) 803–12.

[59] A. F. Agrawal, A. D. Wang, Increased transmission of mutations by low-condition females: evidence for condition-dependent DNA repair, PLoS Biology 6 (2) (2008) e30.

[60] W. H. Zhong, N. K. Priest, Stress-induced recombination and the mechanism of evolvability, Behavioral Ecology and Sociobiology 65 (3) (2011) 493–502.

[61] C. A. Aktipis, Know when to walk away: contingent movement and the evolution of cooperation, Journal of Theoretical Biology 231 (2) (2004) 249–60.

[62] C. A. Aktipis, Is cooperation viable in mobile organisms? simple walk away rule favors the evolution of cooperation in groups, Evolution and human behavior : official journal of the Human Behavior and Evolution Society 32 (4) (2011) 263–276.

[63] V. Mustonen, M. Lassig, From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation, Trends Genet 25 (3) (2009) 111–9.

[64] S. Collins, Many possible worlds: Expanding the ecological scenarios in experimental evolution, Evolutionary Biology 38 (1) (2011) 3–14.

[65] M. Stich, S. C. Manrubia, E. Lazaro, Variable mutation rates as an adaptive strategy in replicator populations, PLoS ONE 5 (6) (2010) e11186.

[66] S. J. Freeland, R. D. Knight, L. F. Landweber, L. D. Hurst, Early fixation of an optimal genetic code, Molecular Biology and Evolution 17 (4) (2000) 511–8.

[67] L. M. Chevin, R. Lande, G. M. Mace, Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory, PLoS Biology 8 (4) (2010) e1000357.

[68] C. G. Knight, M. Platt, W. Rowe, D. C. Wedge, F. Khan, P. J. Day, A. McShea, J. Knowles, D. B. Kell, Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape, Nucleic Acids Res 37 (1) (2009) e6.

[69] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423 and 623–656.

[70] R. L. Stratonovich, On value of information, Izvestiya of USSR Academy of Sciences, Technical Cybernetics 5 (1965) 3–12, in Russian.

[71] R. V. Belavkin, Optimal measures and Markov transition kernels, Journal of Global Optimization (2012) 1–30.

Appendix A. Memoryless Communication

Let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be measurable sets. We shall now consider an $X \times Y$-valued stochastic process $\{(x_t, y_t)\}_{t\geq 0}$ and the 'similarity' between the marginal processes $\{x_t\}_{t\geq 0}$ and $\{y_t\}_{t\geq 0}$ under special assumptions on the communication between X and Y. Recall that a Markov *transition kernel* from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$ is a conditional probability measure $P(Y_i \mid x)$ on $(Y, \mathcal{Y})$, which is $\mathcal{X}$-measurable for each $Y_i \in \mathcal{Y}$. We shall often use measure-theoretic notation $dP(y \mid x)$ for transition kernel $P(Y_i \mid x)$, especially when it appears under the integral.

Proposition 1. *Let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be measurable sets, and let $\{(x_t, y_t)\}_{t\geq 0}$ be a $X \times Y$-valued stochastic process such that elements of the marginal process $\{y_t\}_{t\geq 0}$ are conditionally independent given corresponding elements of $\{x_t\}_{t\geq 0}$:*

$$dP(y_t, \ldots, y_0 \mid x_t, \ldots, x_0) = dP(y_t \mid x_t) \times \cdots \times dP(y_0 \mid x_0)$$

Then transition kernel $dP(y_{t+1} \mid y_t)$ can be expressed as a composition of transition kernels $dP(x_t \mid y_t)$, $dP(x_{t+1} \mid x_t)$ and $dP(y_{t+1} \mid x_{t+1})$ as follows:

$$dP(y_{t+1} \mid y_t) = \int\limits_{x_{t+1}\in X} \int\limits_{x_t \in X} dP(y_{t+1} \mid x_{t+1})\, dP(x_{t+1} \mid x_t)\, dP(x_t \mid y_t)$$

This transition kernel has the following properties:

1. *If X and Y are statistically independent, then $y_{t+1} \in Y$ is independent of $y_t \in Y$: $dP(y_{t+1} \mid y_t) = dP(y_{t+1})$*
2. *If $dP(x \mid y)$ corresponds to a function $x = h(y)$ and y are uniformly distributed in the preimage $h^{-1}(x)$, then*

$$dP(y_{t+1} \mid y_t) = \frac{1}{|h^{-1} \circ h(y_{t+1})|}\, dP(x_{t+1} = h(y_{t+1}) \mid x_t = h(y_t))$$

3. *If $dP(y \mid x)$ corresponds to a function $y = g(x)$ and x are uniformly distributed in the preimage $g^{-1}(y)$, then*

$$dP(y_{t+1} \mid y_t) = \frac{1}{|g^{-1}(y_t)|} \int\limits_{x_{t+1}\in g^{-1}(y_{t+1})} \int\limits_{x_t \in g^{-1}(y_t)} dP(x_{t+1} \mid x_t)$$

4. *If $dP(y \mid x)$ corresponds to a bijection $y = h(x)$, then*

$$dP(y_{t+1} \mid y_t) = dP(x_{t+1} = h(y_{t+1}) \mid x_t = h(y_t))$$

Proof. Transition kernel $dP(x_{t+1} \mid x_t)$ can generally be expressed as follows:

$$\begin{aligned} dP(y_{t+1} \mid y_t) &= \int\limits_{x_{t+1} \in X} \int\limits_{x_t \in X} dP(y_{t+1}, x_{t+1}, x_t \mid y_t) \\ &= \int\limits_{x_{t+1} \in X} \int\limits_{x_t \in X} dP(y_{t+1} \mid x_{t+1}, x_t, y_t)\, dP(x_{t+1} \mid x_t, y_t)\, dP(x_t \mid y_t) \end{aligned}$$

Using the Bayes formula and conditional independence $dP(y_{t+1}, y_t \mid x_{t+1}, x_t) = dP(y_{t+1} \mid x_{t+1})\, dP(y_t \mid x_t)$ one can show that $dP(y_{t+1} \mid x_{t+1}, x_t, y_t) = dP(y_{t+1} \mid x_{t+1})$ and $dP(x_{t+1} \mid x_t, y_t) = dP(x_{t+1} \mid x_t)$. Indeed

$$\begin{aligned} dP(y_{t+1} \mid x_{t+1}, x_t, y_t) &= \frac{dP(y_{t+1}, y_t \mid x_{t+1}, x_t)}{\int\limits_{y_{t+1} \in Y} dP(y_{t+1}, y_t \mid x_{t+1}, x_t)} \\ &= \frac{dP(y_{t+1} \mid x_{t+1})\, dP(y_t \mid x_t)}{\int\limits_{y_{t+1} \in Y} dP(y_{t+1} \mid x_{t+1})\, dP(y_t \mid x_t)} = dP(y_{t+1} \mid x_{t+1}) \end{aligned}$$

$$\begin{aligned} dP(x_{t+1} \mid x_t, y_t) &= \int\limits_{y_{t+1} \in Y} dP(y_{t+1}, x_{t+1} \mid x_t, y_t) \\ &= \int\limits_{y_{t+1} \in Y} \frac{dP(y_{t+1}, y_t \mid x_{t+1}, x_t)\, dP(x_{t+1} \mid x_t)}{dP(y_t \mid x_t)} \\ &= \int\limits_{y_{t+1} \in Y} \frac{dP(y_{t+1} \mid x_{t+1})\, dP(y_t \mid x_t)\, dP(x_{t+1} \mid x_t)}{dP(y_t \mid x_t)} \\ &= dP(x_{t+1} \mid x_t) \end{aligned}$$

Thus, $dP(y_{t+1} \mid y_t)$ can be expressed using the composition of transition kernels $dP(y_{t+1} \mid x_{t+1})\, dP(x_{t+1} \mid x_t)\, dP(x_t \mid y_t)$. We now consider four important cases.

1. If X and Y are independent, then $dP(y_{t+1} \mid x_{t+1}) = dP(y_{t+1})$ and $dP(x_t \mid y_t) = dP(x_t)$, and therefore

$$dP(y_{t+1} \mid y_t) = dP(y_{t+1}) \int\limits_{x_{t+1} \in X} \int\limits_{x_t \in X} dP(x_{t+1} \mid x_t)\, dP(x_t) = dP(y_{t+1})$$

2. If $x = h(y)$ and y are uniformly distributed in the preimage $h^{-1}(x)$, then

$$dP(x_t \mid y_t) = \delta_{h(y_t)}(x_t)\,, \qquad dP(y_{t+1} \mid x_{t+1}) = \frac{1}{|h^{-1} \circ h(y_{t+1})|}$$

   which gives the resulting expression.
3. If $y = g(x)$ and x are uniformly distributed in the preimage $g^{-1}(y)$, then

$$dP(x_t \mid y_t) = \frac{1}{|g^{-1}(y_t)|}\,, \qquad dP(y_{t+1} \mid x_{t+1}) = \delta_{g(x_{t+1})}(y_{t+1})$$

   The resulting expression is obtained by integrating $dP(x_{t+1} \mid x_t)$ for each $x_{t+1} \in g^{-1}(y_{t+1})$ and $x_t \in g^{-1}(y_t)$.
4. Follows trivially from the fact that $|h^{-1} \circ h(y)| = 1$ for a bijection.

□

Remark. It is not required in Proposition 1 for any of the stochastic processes $\{(x_t, y_t)\}_{t \geq 0}$, $\{x_t\}_{t \geq 0}$ or $\{y_t\}_{t \geq 0}$ to be Markov. It is well-known, however, that if $\{x_t\}_{t \geq 0}$ is Markov (i.e. $dP(x_{t+1} \mid x_t, \ldots, x_0) = dP(x_{t+1} \mid x_t)$) and y_t are conditionally independent given the corresponding x_t, then the combined process $\{(x_t, y_t)\}_{t \geq 0}$ is Markov as well, because in this case $dP(x_{t+1}, y_{t+1} \mid x_t, y_t, \ldots, x_0, y_0) = dP(y_{t+1} \mid x_{t+1})\, dP(x_{t+1} \mid x_t)$. The unobserved process $\{x_t\}_{t \geq 0}$ is often referred to as a hidden Markov model, and x_t is estimated from observed values $y_0, \ldots, y_t$ of the related process $\{y_t\}_{t \geq 0}$ (this is called the *filtering* problem [36]). Note that the observed process $\{y_t\}_{t \geq 0}$ is usually non-Markov (i.e. $dP(y_{t+1} \mid y_t, \ldots, y_0) \neq dP(y_{t+1} \mid y_t)$). In the context of Section 4, the unobserved variable $x \in X$ is distance to optimum $d(\top, \omega)$, and observed variable $y \in Y$ is fitness.

Appendix B. Monotonicity

Proposition 2. *Let (Ω, d) be a metric space, and let $f : \Omega \to \mathbb{R}$ be a function with $f(\top) = \sup f(\omega)$ for some $\top \in \Omega$. If the metric d is monotonic at $\top$ relative to f, then all ω with the same values $f(\omega)$ have the same distance $d(\top, \omega)$ from the optimum. Conversely, if f is monotonic at $\top$ relative to d, then all ω with the same distance $d(\top, \omega)$ from the optimum have the same values $f(\omega)$.*

Proof. Indeed, using the definition of monotonic d:

$$\begin{aligned} f(a) = f(b) &\iff f(a) \le f(b) \ \wedge \ f(a) \ge f(b) \\ &\implies -d(\top, a) \le -d(\top, b) \ \wedge \ -d(\top, a) \ge -d(\top, b) \\ &\iff d(\top, a) = d(\top, b) \end{aligned}$$

Using the definition of monotonic f:

$$\begin{aligned} d(\top, a) = d(\top, b) &\iff -d(\top, a) \le -d(\top, b) \ \wedge \ -d(\top, a) \ge -d(\top, b) \\ &\implies f(a) \le f(b) \ \wedge \ f(a) \ge f(b) \\ &\iff f(a) = f(b) \end{aligned}$$

□

Appendix C. Point Mutation as Optimal Solution of Variational Problem

Let (Ω, d) be a metric space, $dQ(a \in \Omega)$ be a probability measure of the 'parent' points, and let $dP(b \in \Omega)$ be a probability measure of their 'offspring' points obtained by a stochastic transformation defined by the transition kernel $dP(b \mid a)$. The product $dP(b \mid a)\, dQ(a)$ defines a joint probability measure of parents and their offspring. The expected distance between parents and offspring is

$$\mathbb{E}\{d(a, b)\} = \int\limits_{\Omega \times \Omega} d(a, b)\, dP(b \mid a)\, dQ(a)$$

The mutual information between parents and offspring is defined as

$$I\{a, b\} = \int\limits_{\Omega \times \Omega} \left[\ln \frac{dP(b \mid a)}{dP(b)} \right] dP(b \mid a)\, dQ(a)$$

We remind that $I\{a,b\} \geq 0$ with zero if and only if a and b are statistically independent. The supremum of $I\{a,b\}$ corresponds to the case when b is obtained from a deterministically using some injective function on Ω (i.e. a one-to-one mapping). For example, if b is identical to a (i.e. $dP(b \mid a)$ corresponds to the identity mapping on Ω), then $I\{a,b\} = \sup I\{a,b\} = |\Omega|$ and $d(a,b) = 0$. Consider the following variational problem

$$\text{minimise} \quad \mathbb{E}\{d(a,b)\} \quad \text{subject to} \quad I\{a,b\} \leq \lambda \tag{C.1}$$

where optimisation is over all joint probability measures $dP(b \mid a)\, dQ(a)$ or over all transition probabilities $dP(b \mid a)$, if $dQ(a)$ is fixed. Because of the constraint on mutual information, the transition probabilities $dP(b \mid a)$ cannot correspond to any injective function on Ω, and therefore generally b cannot be identical to a so that $\mathbb{E}\{d(a,b)\} > 0$. Note that problem (C.1) has the following 'inverse' problem:

$$\text{minimise} \quad I\{a,b\} \quad \text{subject to} \quad \mathbb{E}\{d(a,b)\} \leq \upsilon \tag{C.2}$$

The constraint on the expected distance implies that a and b are not independent so that $I\{a,b\} > 0$. It is well-known in information theory (e.g. see [69, 70] or [71] for generalisations) that solutions to these variational problems are members of an exponential family

$$dP_\beta(b \mid a) = e^{-\beta\, d(a,b) - \Psi(\beta,a)}\, dP(b)\,, \qquad e^{\Psi(\beta,a)} = \int\limits_B e^{-\beta\, d(a,b)}\, dP(b)$$

where parameter β (called the *inverse temperature*) is defined from one of the conditions:

$$I\{a,b\} = \lambda\,, \qquad \mathbb{E}\{d(a,b)\} = \upsilon$$

Moreover, if the metric space Ω is also a group $(\Omega, +)$ with invariant measure ν, and the metric is translation invariant $d(a,b) = d(a+c, b+c)$, then these exponential transition kernels have the following simplified form

$$dP_\beta(b \mid a) = e^{-\beta\, d(a,b) - \Psi_0(\beta)}\, d\nu(b)\,, \qquad e^{\Psi_0(\beta)} = \int\limits_B e^{-\beta\, d(a,b)}\, d\nu(b)$$

In particular, this is the case when Ω is a normed vector space, and the metric is defined using the difference of two vectors: $d(a,b) = \|a - b\|$. For example,

the Hamming space $\mathcal{H}_\alpha^l := \{1, \ldots, \alpha\}^l$ is a finite vector space over a finite field $GF(\alpha)$ with the Hamming metric defined as $d_H(a, b) = \|a - b\|_H$, where $\|\cdot\|_H$ is the Hamming weight. The invariant measure on a Hamming space is the counting measure $\nu(b) = 1$. Thus, for a Hamming space the optimal transition kernel solving problems (C.1) and (C.2) is

$$P_\beta(b \mid a) = e^{-\beta \|a-b\|_H - \Psi_0(\beta)}, \qquad e^{\Psi_0(\beta)} = \sum_{b \in \mathcal{H}_\alpha^l} e^{-\beta \|a-b\|_H}$$

We now show that the above exponential transition kernel implements point mutation.

Indeed, because $e^{-\beta \|a-b\|_H} = e^{-\beta r}$ for all sequences in the sphere $S(a, r) := \{b : \|a - b\|_H = r\}$ around point a and radius r, the summation of $e^{-\beta \|a-b\|_H}$ over all sequences $b \in \mathcal{H}_\alpha^l$ can be replaced by the summation of $|S(a, r)|e^{-\beta r}$ over the spheres of all radii $r \in \{0, \ldots, l\}$. The number of sequences in a sphere of the Hamming space $\mathcal{H}_\alpha^l$ is $|S(a, r)| = (\alpha - 1)^r \binom{l}{r}$, and therefore

$$e^{\Psi_0(\beta)} = \sum_{b \in \mathcal{H}_\alpha^l} e^{-\beta \|a-b\|_H} = \sum_{r=0}^{l} (\alpha - 1)^r \binom{l}{r} e^{-\beta r} = [1 + (\alpha - 1)e^{-\beta}]^l$$

Thus, $P_\beta(b \mid a)$ has the following simple expression:

$$P_\beta(b \mid a) = \frac{e^{-\beta \|a-b\|_H}}{[1 + (\alpha - 1)e^{-\beta}]^l}$$

Given a sequence that is $n = \|\top - a\|_H$ letters away from $\top$, the probability of mutation by radius $r = \|a - b\|_H$ is:

$$P_\beta(r \mid n) = |S(a, r)|P_\beta(b \mid a) = (\alpha - 1)^r \binom{l}{r} \frac{e^{-\beta r}}{[1 + (\alpha - 1)e^{-\beta}]^l}$$

The inverse temperature parameter β is determined either from condition $I\{a, b\} = \lambda$ or $\mathbb{E}\{\|a - b\|_H\} = \upsilon$. In particular, it is convenient to use the latter condition in conjunction with the following expression for the expected mutation radius

$$\mathbb{E}\{r\} = \frac{d}{d\beta}\Psi_0(\beta) = \frac{l}{1 + e^\beta/(\alpha - 1)}$$

Inverting the equation $\mathbb{E}\{r\}(\beta) = \upsilon$ gives the result

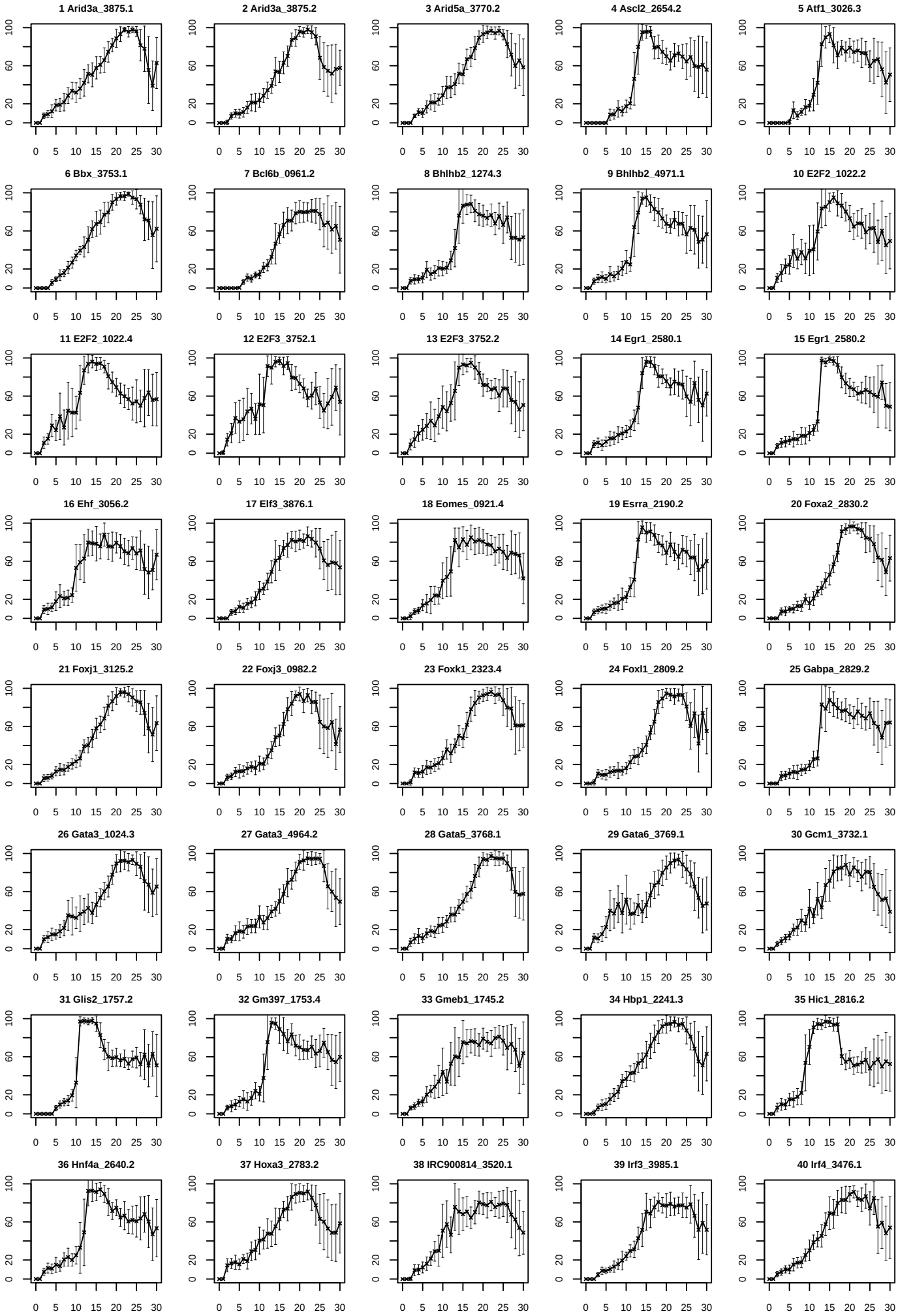
$$\beta = \ln\left(\frac{l-\upsilon}{\upsilon}\right) + \ln(\alpha - 1)$$

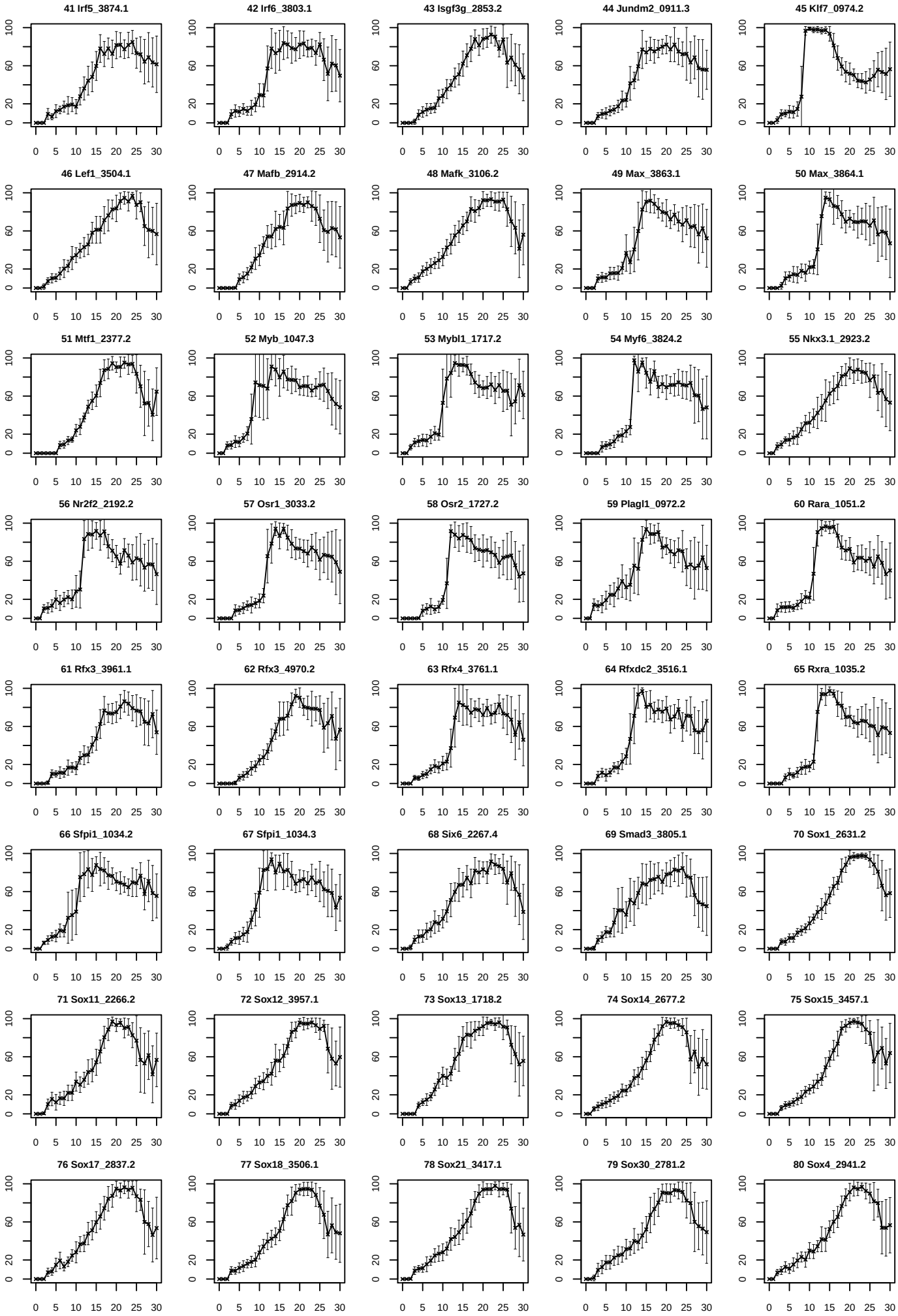
Changing parametrisation from β to υ, the probability $P_\beta(r \mid n)$ can be written as binomial distribution with probability of success $\mu = \upsilon/l$:

$$P_\upsilon(r \mid n) = \binom{l}{r}\left(\frac{\upsilon}{l-\upsilon}\right)^r\left(1+\frac{\upsilon}{l-\upsilon}\right)^{-l} = \binom{l}{r}\left(\frac{\upsilon}{l}\right)^r\left(1-\frac{\upsilon}{l}\right)^{l-r}$$

Therefore, exponential transition kernel that solves optimisation problems (C.1) and (C.2) in the Hamming space corresponds to independent substitution of each letter in a sequence to any other of the $\alpha - 1$ letters with probability $\mu/(\alpha - 1)$, and this process is known as point mutation.

Appendix D. Supplementary Figures


1 Arid3a_3875.1
2 Arid3a_3875.2
3 Arid5a_3770.2
4 Ascl2_2654.2
5 Atf1_3026.3
6 Bbx_3753.1
7 Bcl6b_0961.2
8 Bhlhb2_1274.3
9 Bhlhb2_4971.1
10 E2F2_1022.2
11 E2F2_1022.4
12 E2F3_3752.1
13 E2F3_3752.2
14 Egr1_2580.1
15 Egr1_2580.2
16 Ehf_3056.2
17 Elf3_3876.1
18 Eomes_0921.4
19 Esrra_2190.2
20 Foxa2_2830.2
21 Foxj1_3125.2
22 Foxj3_0982.2
23 Foxk1_2323.4
24 Foxl1_2809.2
25 Gabpa_2829.2
26 Gata3_1024.3
27 Gata3_4964.2
28 Gata5_3768.1
29 Gata6_3769.1
30 Gcm1_3732.1
31 Glis2_1757.2
32 Gm397_1753.4
33 Gmeb1_1745.2
34 Hbp1_2241.3
35 Hic1_2816.2
36 Hnf4a_2640.2
37 Hoxa3_2783.2
38 IRC900814_3520.1
39 Irf3_3985.1
40 Irf4_3476.1
0 20 60 100
0 5 10 15 20 25 30


41 Irf5_3874.1
42 Irf6_3803.1
43 Isgf3g_2853.2
44 Jundm2_0911.3
45 Klf7_0974.2
46 Lef1_3504.1
47 Mafb_2914.2
48 Mafk_3106.2
49 Max_3863.1
50 Max_3864.1
51 Mtf1_2377.2
52 Myb_1047.3
53 Mybl1_1717.2
54 Myf6_3824.2
55 Nkx3.1_2923.2
56 Nr2f2_2192.2
57 Osr1_3033.2
58 Osr2_1727.2
59 Plagl1_0972.2
60 Rara_1051.2
61 Rfx3_3961.1
62 Rfx3_4970.2
63 Rfx4_3761.1
64 Rfxdc2_3516.1
65 Rxra_1035.2
66 Sfpi1_1034.2
67 Sfpi1_1034.3
68 Six6_2267.4
69 Smad3_3805.1
70 Sox1_2631.2
71 Sox11_2266.2
72 Sox12_3957.1
73 Sox13_1718.2
74 Sox14_2677.2
75 Sox15_3457.1
76 Sox17_2837.2
77 Sox18_3506.1
78 Sox21_3417.1
79 Sox30_2781.2
80 Sox4_2941.2
0 20 60 100
0 5 10 15 20 25 30

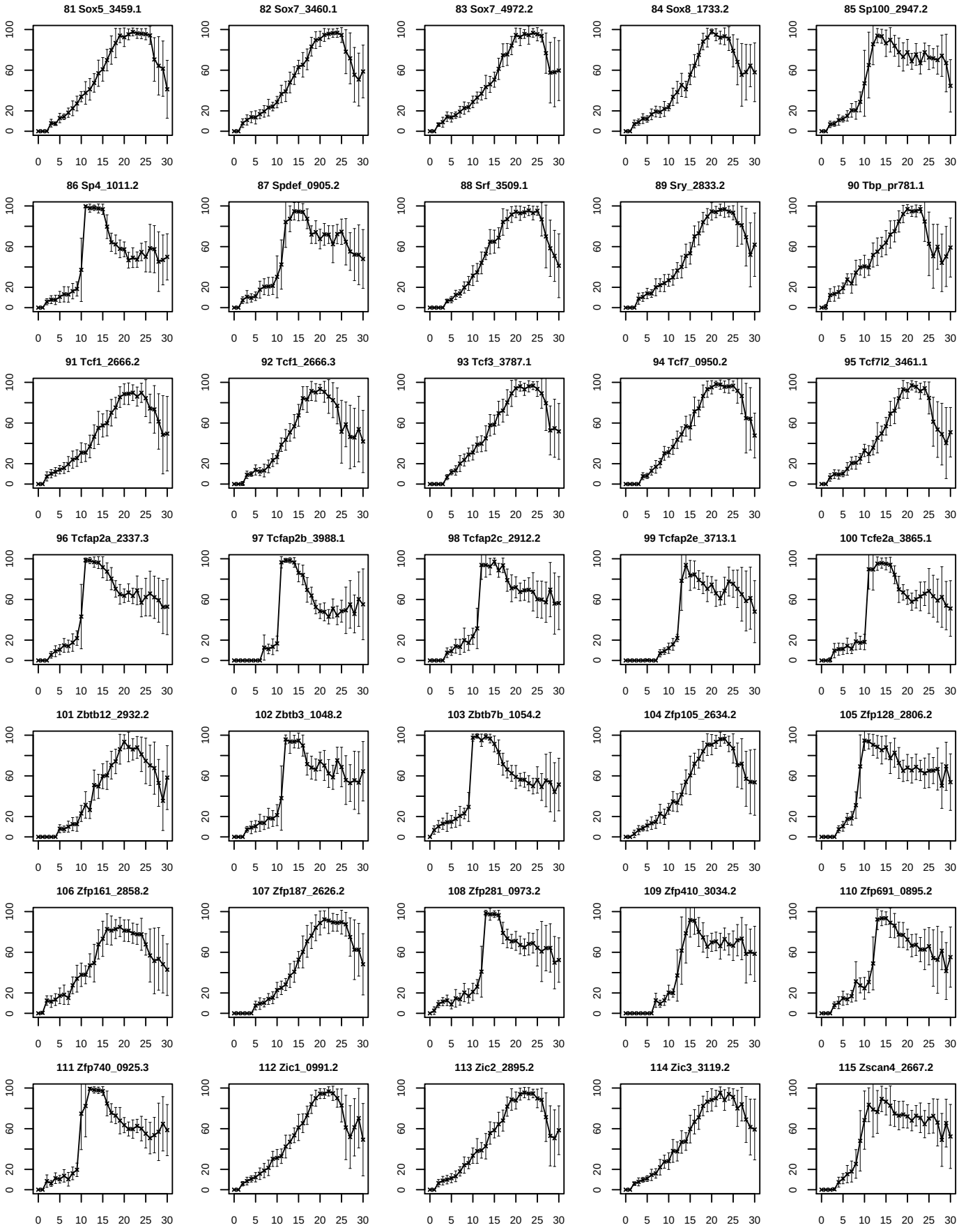
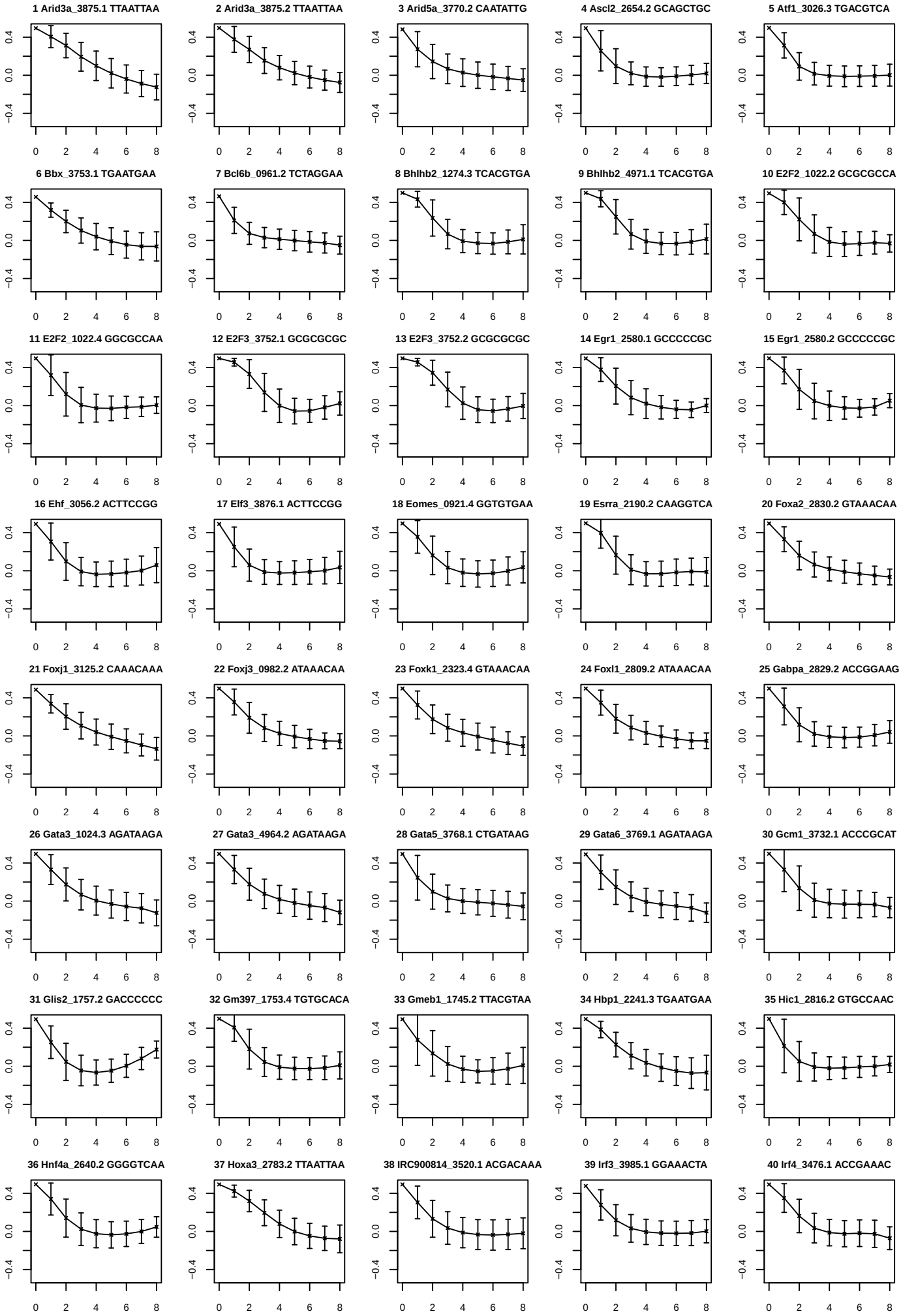
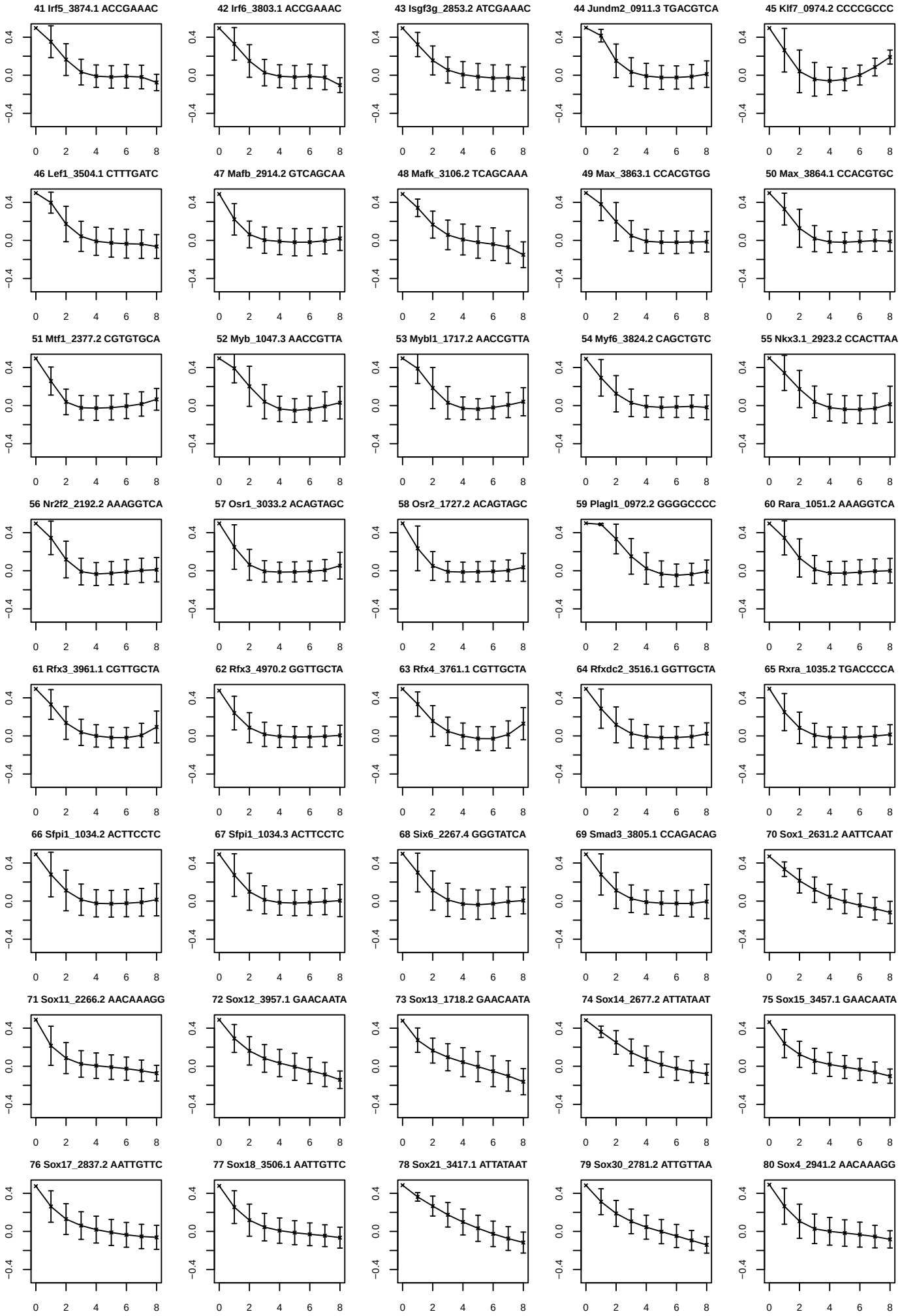

81 Sox5_3459.1
82 Sox7_3460.1
83 Sox7_4972.2
84 Sox8_1733.2
85 Sp100_2947.2
86 Sp4_1011.2
87 Spdef_0905.2
88 Srf_3509.1
89 Sry_2833.2
90 Tbp_pr781.1
91 Tcf1_2666.2
92 Tcf1_2666.3
93 Tcf3_3787.1
94 Tcf7_0950.2
95 Tcf7l2_3461.1
96 Tcfap2a_2337.3
97 Tcfap2b_3988.1
98 Tcfap2c_2912.2
99 Tcfap2e_3713.1
100 Tcfe2a_3865.1
101 Zbtb12_2932.2
102 Zbtb3_1048.2
103 Zbtb7b_1054.2
104 Zfp105_2634.2
105 Zfp128_2806.2
106 Zfp161_2858.2
107 Zfp187_2626.2
108 Zfp281_0973.2
109 Zfp410_3034.2
110 Zfp691_0895.2
111 Zfp740_0925.3
112 Zic1_0991.2
113 Zic2_2895.2
114 Zic3_3119.2
115 Zscan4_2667.2
0 20 60 100
0 5 10 15 20 25 30

Figure D.10: Optimal mutation rate control functions evolved by a Meta-GA on the transcription factor DNA-binding landscapes from [30]. Ordinates show mutation rates, and abscissae show the binding scores. Each panel corresponds to a different transcription factor. Lines connect the average mutation rates obtained in 16 independent trials on a particular landscape. Errorbars represent standard deviations from the mean. The GAs do not spend much time at low binding scores meaning that the results become more random.


1 Arid3a_3875.1 TTAATTAA
2 Arid3a_3875.2 TTAATTAA
3 Arid5a_3770.2 CAATATTG
4 Ascl2_2654.2 GCAGCTGC
5 Atf1_3026.3 TGACGTCA
6 Bbx_3753.1 TGAATGAA
7 Bcl6b_0961.2 TCTAGGAA
8 Bhlhb2_1274.3 TCACGTGA
9 Bhlhb2_4971.1 TCACGTGA
10 E2F2_1022.2 GCGCGCCA
11 E2F2_1022.4 GGCGCCAA
12 E2F3_3752.1 GCGCGCGC
13 E2F3_3752.2 GCGCGCGC
14 Egr1_2580.1 GCCCCCGC
15 Egr1_2580.2 GCCCCCGC
16 Ehf_3056.2 ACTTCCGG
17 Elf3_3876.1 ACTTCCGG
18 Eomes_0921.4 GGTGTGAA
19 Esrra_2190.2 CAAGGTCA
20 Foxa2_2830.2 GTAAACAA
21 Foxj1_3125.2 CAAACAAA
22 Foxj3_0982.2 ATAAACAA
23 Foxk1_2323.4 GTAAACAA
24 Foxl1_2809.2 ATAAACAA
25 Gabpa_2829.2 ACCGGAAG
26 Gata3_1024.3 AGATAAGA
27 Gata3_4964.2 AGATAAGA
28 Gata5_3768.1 CTGATAAG
29 Gata6_3769.1 AGATAAGA
30 Gcm1_3732.1 ACCCGCAT
31 Glis2_1757.2 GACCCCCC
32 Gm397_1753.4 TGTGCACA
33 Gmeb1_1745.2 TTACGTAA
34 Hbp1_2241.3 TGAATGAA
35 Hic1_2816.2 GTGCCAAC
36 Hnf4a_2640.2 GGGGTCAA
37 Hoxa3_2783.2 TTAATTAA
38 IRC900814_3520.1 ACGACAAA
39 Irf3_3985.1 GGAAACTA
40 Irf4_3476.1 ACCGAAAC


41 Irf5_3874.1 ACCGAAAC
42 Irf6_3803.1 ACCGAAAC
43 Isgf3g_2853.2 ATCGAAAC
44 Jundm2_0911.3 TGACGTCA
45 Klf7_0974.2 CCCCGCCC
46 Lef1_3504.1 CTTTGATC
47 Mafb_2914.2 GTCAGCAA
48 Mafk_3106.2 TCAGCAAA
49 Max_3863.1 CCACGTGG
50 Max_3864.1 CCACGTGC
51 Mtf1_2377.2 CGTGTGCA
52 Myb_1047.3 AACCGTTA
53 Mybl1_1717.2 AACCGTTA
54 Myf6_3824.2 CAGCTGTC
55 Nkx3.1_2923.2 CCACTTAA
56 Nr2f2_2192.2 AAAGGTCA
57 Osr1_3033.2 ACAGTAGC
58 Osr2_1727.2 ACAGTAGC
59 Plagl1_0972.2 GGGGCCCC
60 Rara_1051.2 AAAGGTCA
61 Rfx3_3961.1 CGTTGCTA
62 Rfx3_4970.2 GGTTGCTA
63 Rfx4_3761.1 CGTTGCTA
64 Rfxdc2_3516.1 GGTTGCTA
65 Rxra_1035.2 TGACCCCA
66 Sfpi1_1034.2 ACTTCCTC
67 Sfpi1_1034.3 ACTTCCTC
68 Six6_2267.4 GGGTATCA
69 Smad3_3805.1 CCAGACAG
70 Sox1_2631.2 AATTCAAT
71 Sox11_2266.2 AACAAAGG
72 Sox12_3957.1 GAACAATA
73 Sox13_1718.2 GAACAATA
74 Sox14_2677.2 ATTATAAT
75 Sox15_3457.1 GAACAATA
76 Sox17_2837.2 AATTGTTC
77 Sox18_3506.1 AATTGTTC
78 Sox21_3417.1 ATTATAAT
79 Sox30_2781.2 ATTGTTAA
80 Sox4_2941.2 AACAAAGG

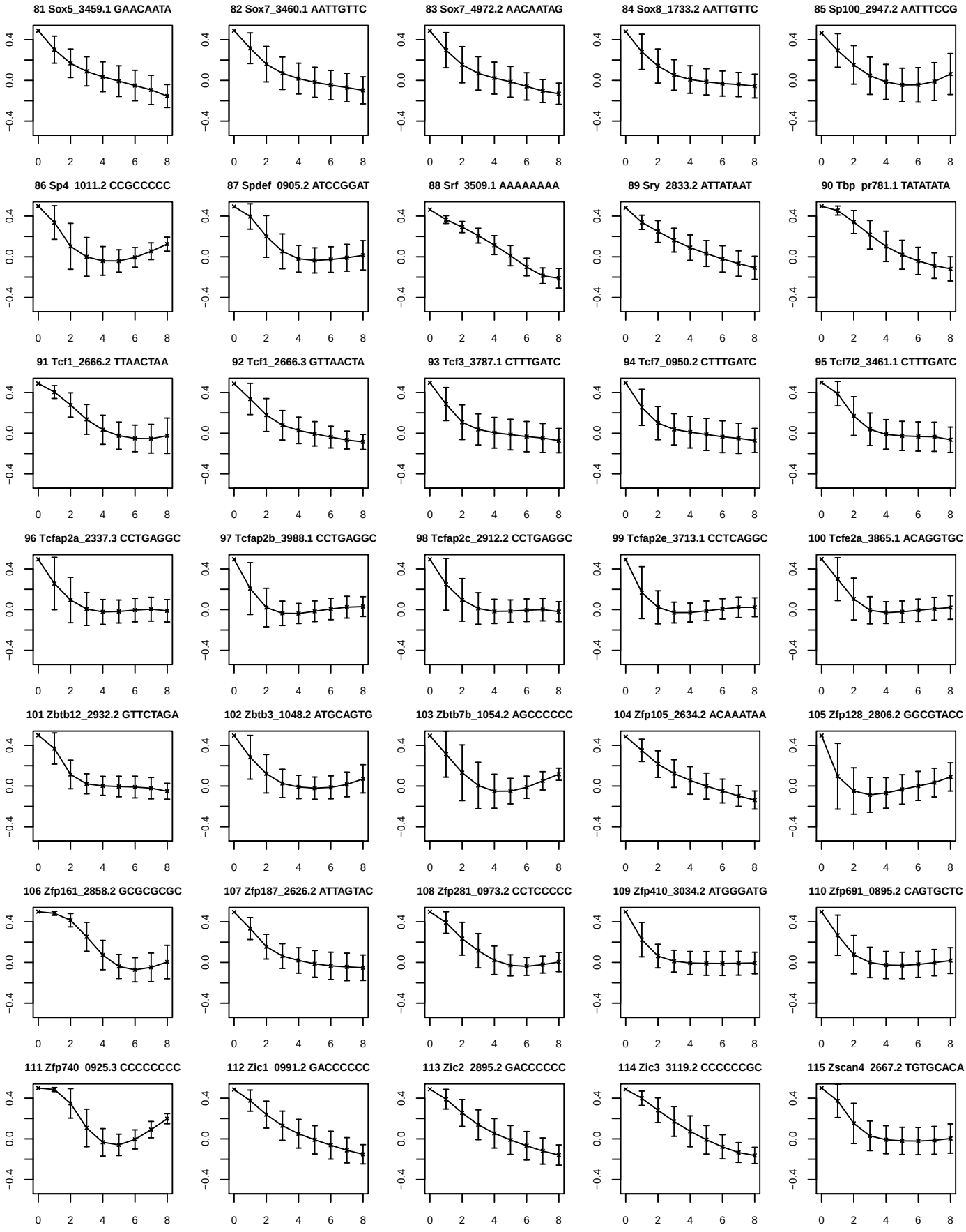

81 Sox5_3459.1 GAACAATA
82 Sox7_3460.1 AATTGTTC
83 Sox7_4972.2 AACAATAG
84 Sox8_1733.2 AATTGTTC
85 Sp100_2947.2 AATTTCCG
86 Sp4_1011.2 CCGCCCCC
87 Spdef_0905.2 ATCCGGAT
88 Srf_3509.1 AAAAAAAA
89 Sry_2833.2 ATTATAAT
90 Tbp_pr781.1 TATATATA
91 Tcf1_2666.2 TTAACTAA
92 Tcf1_2666.3 GTTAACTA
93 Tcf3_3787.1 CTTTGATC
94 Tcf7_0950.2 CTTTGATC
95 Tcf7l2_3461.1 CTTTGATC
96 Tcfap2a_2337.3 CCTGAGGC
97 Tcfap2b_3988.1 CCTGAGGC
98 Tcfap2c_2912.2 CCTGAGGC
99 Tcfap2e_3713.1 CCTCAGGC
100 Tcfe2a_3865.1 ACAGGTGC
101 Zbtb12_2932.2 GTTCTAGA
102 Zbtb3_1048.2 ATGCAGTG
103 Zbtb7b_1054.2 AGCCCCCC
104 Zfp105_2634.2 ACAAATAA
105 Zfp128_2806.2 GGCGTACC
106 Zfp161_2858.2 GCGCGCGC
107 Zfp187_2626.2 ATTAGTAC
108 Zfp281_0973.2 CCTCCCCC
109 Zfp410_3034.2 ATGGGATG
110 Zfp691_0895.2 CAGTGCTC
111 Zfp740_0925.3 CCCCCCCC
112 Zic1_0991.2 GACCCCCC
113 Zic2_2895.2 GACCCCCC
114 Zic3_3119.2 CCCCCCGC
115 Zscan4_2667.2 TGTGCACA
0.4
0.0
−0.4
0
2
4
6
8

Figure D.11: Landscapes of binding score between 8 base-pair DNA sequences and transcription factors (TF) from [30]. Ordinates show binding scores, and abscissae show Hamming distances from the top sequence (a sequence with the highest DNA-TF binding score). Each panel corresponds to a different transcription factor. Lines connect mean values of the binding score for each value of the Hamming distance from the top sequence. Errorbars represent standard deviations. Note that this dataset does not distinguish between sequences on opposite strands of the DNA. Therefore, a sequence and its reverse complement are shown only once and the Hamming distance shown is either to the top sequence or its reverse complement, whichever is the closer.