Annals of Mathematics and Artificial Intelligence
https://doi.org/10.1007/s10472-025-09987-5

RESEARCH

Analysis and optimization of probabilities of beneficial mutation and crossover recombination in a Hamming space

Roman V. Belavkin[1]

Accepted: 20 May 2025

Abstract
Inspired by Fisher's geometric approach to study beneficial mutations, we analyse probabilities of beneficial mutation and crossover recombination of strings in a general Hamming space with arbitrary finite alphabet. Mutations and recombinations that reduce the distance to an optimum are considered as beneficial. Geometric and combinatorial analysis is used to derive closed-form expressions for transition probabilities between spheres around an optimum giving a complete description of Markov evolution of distances from an optimum over multiple generations. This paves the way for optimization of parameters of mutation and recombination operators. Here we derive optimality conditions for mutation and recombination radii maximizing the probabilities of mutation and crossover into the optimum. The analysis highlights important differences between these evolutionary operators. While mutation can potentially reach any part of the search space, the probability of beneficial mutation decreases with distance to an optimum, and the optimal mutation radius or rate should also decrease resulting in a slow-down of evolution near the optimum. Crossover recombination, on the other hand, acts in a subspace of the search space defined by the current population of strings. However, probabilities of beneficial and deleterious crossover are balanced, and their characteristics, such as variance, are translation invariant in a Hamming space, suggesting that recombination may complement mutation and boost the rate of evolution near the optimum.

Keywords Mutation · Crossover · Recombination · Evolutionary algorithm · Optimal parameter control

Mathematics Subject Classification (2010) 05B25 · 68W20 · 68T05 · 60C05 · 68R05 · 68R15

Notation

$\mathbb{N}$	the set of natural numbers $\{1, 2, 3, \ldots\}$
$\mathbb{R}$	the field of real numbers
$l \in \mathbb{N}$	'length' of tuples or strings
$\alpha \in \mathbb{N}$	size of a finite alphabet

✉ Roman V. Belavkin
r.belavkin@mdx.ac.uk

[1] Faculty of Science and Technology, Middlesex University, The Burroughs, NW4 4BT London, United Kingdom

Published online: 23 June 2025

$\{1,\ldots,\alpha\}^l$	the set of all α^l strings of length l over alphabet of size α
$\mathbb{R}^l$	l-dimensional real vector space
d_E	Euclidean metric on $\mathbb{R}^l$
d_H	or d Hamming metric on $\{1,\ldots,\alpha\}^l$
$\mathcal{H}_\alpha^l$	Hamming space — set $\{1,\ldots,\alpha\}^l$ with the Hamming metric
x, y, z	points in $\mathcal{H}_\alpha^l$, which are l-tuples or strings $x = (x_1,\ldots,x_l)$
i, j	positions in strings, such as $x = (x_1,\ldots,x_i,x_j,\ldots,x_l)$
$\top$	top or greatest string in $\mathcal{H}_\alpha^l$ with respect to some preference relation $\lesssim$ on $\mathcal{H}_\alpha^l$
k, m, n	values of Hamming distance from $\top$, such as $d(\top, x) = n$
r	*mutation radius* $d(x, y) = r$ in the context of mutation or *recombination radius* in the context of recombination (the number of letters substituted)
h	Hamming distance $d(x, y) = h$ between two parent strings in crossover recombination and referred to as *recombination capacity*
$S(x, r)$	the sphere of radius r around x $\{y : d(x, y) = r\}$
$B(x, r)$	the closed ball of radius r around x $\{y : d(x, y) \le r\}$
μ	mutation rate
ν	recombination rate
$P\{\cdot\}$	probability measure
$P(n)$	probability mass function equal to $P\{d(\top, x) = n\}$
$\mathbb{E}_P\{n\}$	the expected value of random variable with respect to measure P
$\sigma_P^2\{n\}$	the variance of random variable with respect to measure P

1 Introduction

Natural evolution can be viewed as a search for an optimal genotype $\top$ (top) in the space $\{1,\ldots,\alpha\}^l$ of all genetic codes of finite alphabet $\{1,\ldots,\alpha\}$ of size $\alpha \in \mathbb{N}$ and length $l \in \mathbb{N}$. Optimality can be defined by some fitness function $f : \{1,\ldots,\alpha\}^l \to \mathbb{R}$ maximized at $\top$. Some mathematicians, however, simplified the analysis by replacing fitness $f(x)$ of a genotype with its distance $d(\top, x)$ from $\top$. For example, Roland Fisher [1] used Euclidean space $\mathbb{R}^l$ of l traits to represent species by vectors of l traits and Euclidean distance $d_E(\top, x)$ from an optimum to represent (negative) fitness of x. This simplification allowed him to analyse the probability of beneficial mutations, which in this geometric model meant that mutation of x resulted in an offspring y closer to the optimum: $d_E(\top, y) \le d_E(\top, x)$. Fisher's famous result was that beneficial mutations are always more rare than deleterious, and that the only way to equalize their chances is to minimize the mutation radius $d_E(x, y) = r$. This result follows from the geometry of Euclidean space, where every closed ball $B(\top, n) = \{x \in \mathbb{R}^l : d_E(\top, x) \le n\}$ around $\top$ is compact (and has finite volume), while its complement is always unbounded. Thus, a random mutation of x with $d_E(\top, x) = n$ in all directions by radius $d(x, y) = r$ should more likely end outside the ball $B(\top, n)$ and further from the optimum resulting in a deleterious mutation.

The discovery of DNA and RNA molecules lead to the realization that mutations occur on the level of genetic codes, which are better represented as strings $x = (x_1,\ldots,x_l)$ of length $l \in \mathbb{N}$ over some finite alphabet $\{1,\ldots,\alpha\} \ni x_i$. Thus, Fisher's theory of beneficial mutations had to be reconsidered for spaces of strings with alphabets of arbitrary size $\alpha \in \mathbb{N}$ and variable lengths $l \in \mathbb{N}$ [2–5]. Furthermore, this geometric approach (i.e. replacing fitness with distance) had limited appeal for practical applications, because distances to an optimum are usually not known. However, the values $f(x)$ of a fitness function can often provide some information about the distance $d(\top, x)$, and the correlation between fitness and distance has

been discussed in the literature, for example, as a measure of problem difficulty [6, 7]. Various notions of monotonicity of fitness landscapes have been defined and proven to hold in a broad class of landscapes if they are continuous at least at the optimum $\top$ (see Theorem 1 in [8]). While it is always possible to construct counter examples, fitness landscapes in real-world applications or biology often exhibit some forms of monotonicity around optimum, as was demonstrated in [8] for 115 complete landscapes of transcription factor bindings [9].

The generalization of Fisher's geometric model of beneficial mutations to spaces of strings with alphabets of arbitrary size $\alpha \in \mathbb{N}$ was used to derive several optimal mutation rate control functions [2–5]. They showed that optimal mutation rates should have a decreasing relation to fitness in monotonic fitness landscapes [8]. These theoretical predictions lead to the discovery of mutation rates plasticity first in *e. coli* [10] and then in other microbes and potentially across all domains of life [11]. The role of *quorum sensing* in this phenomenon and the relation of population density (as a fitness proxy) and stress to mutation rate [12] suggest a broad scope for applications in many areas including antimicrobial resistance (AMR). Another potential area of applications of this geometric approach is operational research, where many nature-inspired algorithms [13] are used to solve complex combinatorial optimization problems.

Evolutionary algorithms, such as genetic algorithms (GA), encode candidate solutions by finite length strings $x = (x_1, \ldots, x_l)$ with letters from a finite alphabet $x_i \in \{1, \ldots, \alpha\}$, and operators of selection, mutation and recombination are applied iteratively to search the space $\{1, \ldots, \alpha\}^l$ [14]. Mutation is a random substitution of some letters in the parent string by any of the $\alpha - 1$ letters from the alphabet. Recombination, on the other hand, is a substitution of some letters in one parent string by the letters from another string (e.g. in the corresponding positions for crossover recombination). Thus, mutation searches across the entire space $\{1, \ldots, \alpha\}^l$, while recombination can only search in a subspace defined by the current population. However, recombination of different strings makes some directions of the search more likely (i.e. a kind of pseudo-gradient).

Many heuristics have been identified to improve the search efficiency by finding optimal settings or optimal controls of certain parameters, such as the mutation rate. In particular, one popular heuristic is to set the mutation rate to $\mu = 1/l$, where l is the string length [15]. Other works showed the advantage of using a variable mutation rate that may depend on time or fitness of individuals [14, 16–21]. Many of these works considered only binary codes ($\alpha = 2$), because their combinatorics is more tractable. More recent studies in the theory of evolutionary algorithms have considered arbitrary finite alphabets and self-adjusting mutation rates [22–24].

Different heuristic recombination operators have also been employed, such as one-point crossover or a uniform crossover operators, and its important role in maintaining diversity has been recognized [25, 26]. While there are many other types of recombination operators considered in the literature, including mixtures of codes from more than two parents [27], this paper will only consider crossover between two parent strings. Even in this basic case, however, combinatorial analysis of crossover is more challenging than that for mutation, because it involves more points and more parameters. Many studies have considered recombination only for binary codes [18, 28–32].

The analysis of evolutionary operators for codes with alphabets of size $\alpha > 2$ should have a broader scope of applications not only in the context of DNA or RNA molecules with $\alpha = 4$, but also for larger alphabets, such as $\alpha = 22$ for the number of canonical amino acids. In addition, some recombination operators substitute entire substrings of length r (e.g. $r = \lfloor l/2 \rceil$ in one-point crossover). Therefore, recombination can be considered acting on the space of strings $\{1, \ldots, \alpha^r\}^{l/r}$ (i.e. alphabet of size α^r).

This work develops a geometric approach to evolution of strings in Hamming spaces extending it from mutation to crossover recombination. In the next section, we start by reviewing some of the basic properties of a Hamming space and formulate the problem of finding the probability of mutation onto a sphere of a given radius. Its closed-form solution is given in Theorem 1, which has been previously presented in [2–5, 8]. These results about mutation are included here not only for completeness, but also because they are used in the analysis of recombination in Section 3, and in particular Lemma 2 for intersection of spheres. We also derive new formulae for the conditional expected value and variance of Hamming distance after mutation. In Section 3, we formulate analogous problem for probability of crossover recombination onto a sphere in a Hamming space, and then derive closed-form solution in Theorem 2. As with mutation, we also derive new formulae for the expected value and variance of distance after recombination. We conclude each section by analysing the effects of parameters on probabilities of beneficial mutation and recombination and deriving optimality conditions for mutation and crossover recombination into optimum. We discuss how our results open new possibilities for a long-term analysis and optimization of mutation and recombination operators.

2 Mutation

2.1 Mutation in a Hamming space

Consider the space $\{1, \dots, \alpha\}^l$ of strings (or codes) of length $l \in \mathbb{N}$ and finite alphabet of size $\alpha \in \mathbb{N}$, and let us equip it with the Hamming metric counting the number of different letters:

$$d(x, y) = |\{i \in \{1, \dots, l\} : x_i \neq y_i\}| = \sum_{i=1}^{l} (1 - \delta_{x_i y_i}), \quad \delta_{x_i y_i} = \begin{cases} 1 \text{ if} x_i = y_i \\ 0 \text{ otherwise} \end{cases}. \tag{1}$$

This metric space is referred to as Hamming space and denoted by $\mathcal{H}_\alpha^l$. The definition of Hamming metric (1) as the sum of elementary distances $1 - \delta_{x_i y_j}$ leads to the following useful result.

Lemma 1 *(Mean and variance of Hamming distance) Let $\mathcal{H}_\alpha^l$ be a Hamming space, and let $P : 2^{\mathcal{H}_\alpha^l \times \mathcal{H}_\alpha^l} \to [0, 1]$ be a joint probability distribution. Then the expected value and variance of the Hamming distance $d(x, y)$ between pairs of strings $x, y \in \mathcal{H}_\alpha^l$ are*

$$\mathbb{E}_P\{d(x, y)\} = l\langle P_i \rangle, \tag{2}$$

$$\sigma_P^2\{d(x, y)\} = l\langle P_i \rangle + l(l-1)\langle P_{ij} \rangle - (l\langle P_i \rangle)^2, \tag{3}$$

where

$$\langle P_i \rangle := \frac{1}{l} \sum_{i=1}^{l} \mathbb{E}_P\{1 - \delta_{x_i y_i}\} = \mathbb{P}\{x_i \neq y_i\},$$

$$\langle P_{ij} \rangle := \frac{1}{l(l-1)} \sum_{i=1}^{l} \sum_{\substack{j=1 \\ j \neq i}}^{l} \mathbb{E}_P\{(1 - \delta_{x_i y_i})(1 - \delta_{x_j y_j})\} = \mathbb{P}\{x_i \neq y_i \wedge x_j \neq y_j \mid i \neq j\}.$$

are respectively the average probability of non-identical letters at positions $i \in \{1, \ldots, l\}$ *and the average joint probability of non-identical letters at two different positions* i *and* $j \neq i$ *under the distribution* P.

See Appendix A.1 for the proof. The average probabilities $\langle P_i \rangle := \mathbb{P}\{x_i \neq y_i\}$ and $\langle P_{ij} \rangle := \mathbb{P}\{x_i \neq y_i \wedge x_j \neq y_j \mid i \neq j\}$ can often be estimated from additional information about distances of strings x and y. For example, assume a joint distribution $P(x, y)$ such that $d(x, y) = n$ for all pairs of strings such that $P(x, y) > 0$ (i.e. exactly n letters $x_i \neq y_i$). Then $\langle P_i \rangle = n/l$ and $\langle P_{ij} \rangle = (n/l)[(n-1)/(l-1)]$. Substituting these probabilities into (2) and (3) gives the desired result $\mathbb{E}_P\{d(x, y)\} = n$ and $\sigma_P^2\{d(x, y)\} = 0$. More interesting and useful formulae will be obtained in Proposition 1 for mutation and Proposition 5 for crossover.

The geometry of Hamming space is different from that of the Euclidean space $\mathbb{R}^l$ employed by Fisher [33]. In particular, $\mathcal{H}_\alpha^l$ is finite, has finite diameter l, and every point $x \in \mathcal{H}_\alpha^l$ has $(\alpha - 1)^l$ diametric opposite points $\neg x$ (i.e. such that $d(x, \neg x) = l$). The number of elements in a sphere $S(\top, n) := \{x \in \mathcal{H}_\alpha^l : d(\top, x) = n\}$ of radius n around $\top$ is

$$|S(\top, n)| = (\alpha - 1)^n \binom{l}{n}.$$

The number of elements in a closed ball $B(\top, n) := \{x \in \mathcal{H}_\alpha^l : d(\top, x) \leq n\}$ is the sum $\sum_{k=0}^{n} |S(\top, k)|$ for all the spheres it contains. The complement $H_\alpha^l \setminus B(\top, n)$ is the union of all balls $B(\neg\top, l - n)$ around $(\alpha - 1)^l$ diametric opposite points $\neg\top$ (see [33] for details and many other properties of Hamming space). The number of elements in the complement $H_\alpha^l \setminus B(\top, n)$ is the sum $\sum_{k=l-n}^{l} |S(\top, k)|$, and it may contain fewer elements than the ball itself, unlike in the Euclidean space.

The 'equator' of a Hamming space is the radius equal to $\lfloor l(1 - 1/\alpha) \rceil$ (here $\lfloor \cdot \rceil$ denotes the nearest integer), which corresponds to the median of the binomial distribution $P(n; l, p)$ with parameter $p = 1 - 1/\alpha$. Indeed, under a uniform distribution $P_0(x) = \alpha^{-l}$ of strings $x \in \mathcal{H}_\alpha^l$, the probability $P_0(n)$ of distances $d(\top, x) = n$ from $\top$ (or from any other point) is

$$P_0(x) = \frac{1}{\alpha^l} \quad \Longrightarrow \quad P_0(n) := P_0\{x \in S(\top, n)\} = \frac{|S(x, n)|}{\alpha^l} = \frac{(\alpha - 1)^n}{\alpha^l} \binom{l}{n},$$

which can be written as the binomial distribution $P_0(n) = \binom{l}{n}(1 - 1/\alpha)^n (1/\alpha)^{l-n}$. Its expected value and variance are

$$\mathbb{E}_{P_0}\{n\} = l(1 - 1/\alpha), \qquad \sigma_{P_0}^2(n) = l(1 - 1/\alpha)(1/\alpha),$$

which can also be obtained using formulae (2) and (3) with $\langle P_i \rangle = 1 - 1/\alpha$ and $\langle P_{ij} \rangle = (1 - 1/\alpha)^2$. The median is the nearest integer of the above expected value. For alphabets of size $\alpha > 2$ the distribution of distances is skewed towards the end of the range $[0, l]$.

Asexual reproduction of species corresponds to a transformation $x \mapsto y$ of their genetic codes due to a random substitution of $r \in [0, l]$ letters — a process which we shall generally refer to as *mutation*. The resulting distance $d(x, y) = r$ from the parent string in this context is referred to as the *mutation radius* shown on Fig. 1. If distance $d(\top, \cdot)$ from the optimum $\top$ is taken as a model of (negative) fitness, then *beneficial* mutation is a transition from sphere $S(\top, n) \ni x$ onto sphere $S(\top, m) \ni y$ of a smaller radius $m < n$, as shown on Fig. 1. Mutation is *neutral* if $m = n$, and *deleterious* if $m > n$.

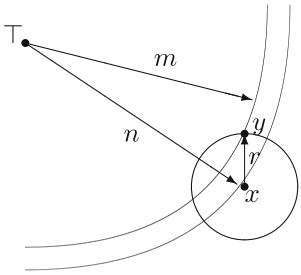


Fig. 1 Mutation of string $x \in S(\top, n)$ into $y \in S(\top, m)$ by substitution of $r = d(x, y)$ letters. The number of strings in the intersection $S(\top, m) \cap S(x, r)$ defines the geometric probability $P(m \mid n, r)$ (9)

Example 1 (Mutation) Let $\top = (\mathrm{AAAAA}) \in \mathcal{H}_3^5$ and consider string $x = (\mathrm{BBBAA})$ mutating into $y = (\mathrm{BACBA})$ by a substitution of the second, third and fourth letters:

$$\begin{array}{ccc} & \top = (\mathrm{AAAAA}) & \\ {\scriptstyle n=3}\ \cdots & & \cdots\ {\scriptstyle m=3} \\ x = (\mathrm{BBBAA}) & \xrightarrow{\quad r=r_+ + r_0 + r_- = 3 \quad} & (\mathrm{B}\underset{r_+}{\mathbf{A}}\underset{r_0}{\mathbf{C}}\underset{r_-}{\mathbf{B}}\mathrm{A}) = y \end{array}$$

Thus, the mutation radius is $d(x, y) = 3$, and the mutation is neutral, because $d(\top, x) = d(\top, y) = 3$. Notice that there was $r_+ = 1$ beneficial, $r_0 = 1$ neutral and $r_- = 1$ deleterious substitution. A substitution of three letters in $x = (\mathrm{BBBAA})$ may also result in string $z = (\mathrm{BAACA})$, which is closer to the optimum, $d(\top, z) = 2$ (i.e. beneficial mutation).

Henceforth we shall denote by r_+, r_- and r_0 the numbers of beneficial, deleterious and neutral substitutions respectively. These numbers add up to the mutation radius $r = d(x, y)$, while the difference $r_+ - r_-$ is equal to the difference $n - m$ of distances:

$$r_+ + r_- + r_0 = r\,, \tag{4}$$

$$r_+ - r_- = n - m\,. \tag{5}$$

If string $x \in S(\top, n)$ mutates into $y \in S(\top, m)$, then the range of the mutation radius is defined by the triangle inequalities:

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

At the extreme values $r = |n-m|$ or $r = n+m$ of the mutation radius, there are no neutral substitutions. Indeed, for the maximum value $r = n+m$ there are exactly $r_+ = n = d(\top, x)$ beneficial and $r_- = m = d(\top, y)$ deleterious substitutions, so that $r_0 = r - r_+ - r_- = 0$. For the minimum value $r = |n - m|$ there are $r_+ = \max\{0, n - m\}$ beneficial and $r_- = \max\{0, m - n\}$ deleterious substitutions. Then (4) and $r = |n - m| = \max\{n - m, m - n\}$ imply $r_0 = 0$. In both extreme cases we also have $r_- = r - r_+$ and $r_+ = \frac{1}{2}(r + n - m)$ (if the latter is integer). Clearly, neutral substitutions are impossible for binary strings ($\alpha = 2$).

2.2 Evolutionary dynamics under mutation

A random mutation $x \mapsto y$ corresponds to some transition probability $P(y \mid x)$, where x is the 'parent', and y is its 'offspring', and it induces a transformation of distribution $P_t(x)$ of

the parent codes into the distribution $P_{t+1}(y)$ of the offspring codes:

$$P_{t+1}(y) = \sum_{x \in \mathcal{H}_\alpha^l} P(y \mid x)\, P_t(x)\,.$$

The corresponding distance distributions $P_t(n) := P_t\{x \in S(\top, n)\}$ and $P_{t+1}(m) := P_{t+1}\{y \in S(\top, m)\}$ are transformed as well:

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

Here, $P(m \mid n)$ is the transition probability between spheres around $\top$ due to mutation:

$$P(m \mid n) := P\{y \in S(\top, m) \mid x \in S(\top, n)\}\,.$$

If $P(m \mid n)$ is time invariant, then the linear operator

$$M(\cdot) = \sum_{n=0}^{l} P(m \mid n)\,(\cdot) \tag{6}$$

acting on distributions $P_t(n)$ of distances $d(\top, x) = n \in [0, l]$ generates the entire evolution $\{P_t\}_{t \geq 0}$ due to mutation as $P_{t+s} = M^s P_t$. This can be used in simulations to analyse the effects of mutation and adaptation over several generations.

The transition probability $P(m \mid n)$ can be factorized in the following way:

$$P(m \mid n) = \sum_{r=0}^{l} P(m \mid n, r)\, \underbrace{P(r \mid n)}_{\text{Mutation}}, \tag{7}$$

where $P(r \mid n) := P\{y \in S(x, r) \mid x \in S(\top, n)\}$ is the probability of mutation radius $r \in [0, l]$ conditioned on distance $n = d(\top, x)$. This probability can be determined from the mutation operator.

Example 2 (Point mutation) In a simple point mutation, each letter x_i is substituted independently with probability $\mu \in [0, 1]$ called the *mutation rate* (i.e. μ is fixed for all $i \in \{1, \ldots, l\}$). Each letter x_i can be substituted to any of the $\alpha - 1$ letters y_i with uniform probability $1/(\alpha - 1)$. In this case, the probability that $r \in [0, l]$ letters are substituted has binomial distribution:

$$P_\mu(r \mid n) = \binom{l}{r} \mu^r(n)[1 - \mu(n)]^{l-r}\,. \tag{8}$$

Note that the mutation rate μ may depend on the distance $d(\top, x) = n \in [0, l]$ of the parent string from the optimum. The mutation operator (6) in this case is parameterized by the mutation rate control function $\mu(n)$:

$$M_{\mu(n)}(\cdot) = \sum_{n=0}^{l} \left[\sum_{r=0}^{l} P(m \mid n, r) P_\mu(r \mid n) \right] (\cdot)\,.$$

2.3 Geometric probability of mutation onto a sphere

The probability $P(m \mid n, r)$ in factorization (7) is independent of the mutation operator, and it represents a purely geometric problem depicted on Fig. 1:

$$P(m \mid n, r) := P\{y \in S(\top, m) \mid x \in S(\top, n), d(x, y) = r\}\,. \tag{9}$$

Fisher considered this geometric probability in Euclidean space $\mathbb{R}^l$ [1]. For a Hamming space $\mathcal{H}_\alpha^l$, this problem was considered in [2–5, 8], and here we review its solution, because it will be useful for the analysis of recombination in Section 3. One can see from Fig. 1 that probability (9) depends on the number of strings in the intersection of spheres $S(\top, m)$ and $S(x, r)$.

Lemma 2 *(Intersection of spheres [2–5, 8]) The number of elements in the intersection $S(\top, m) \cap S(x, r)$ of spheres around points $\top$, $x \in \mathcal{H}_\alpha^l$ with $d(\top, x) = n$ is*

$$\Big|S(\top, m) \cap S(x, r)\Big|_{d(\top,x)=n} = \sum_{r_+=0}^{r} (\alpha-2)^{r_0} \binom{n-r_+}{r_0} (\alpha-1)^{r_-} \binom{l-n}{r_-} \binom{n}{r_+}, \quad (10)$$

where $r_+ \in [0, r]$, and indices $r_- \geq 0$, $r_0 \geq 0$ satisfy the equations:

$$r_- = r_+ - (n-m)\,, \qquad r_0 = r - 2r_+ + (n-m)\,.$$

Distances $n = d(\top, x)$, $m = d(\top, y)$ and $r = d(x, y)$ must satisfy the triangle inequalities:

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

Otherwise, the number is zero.

See Appendix A.2 for the proof.

Remark 1 The summation in (10) is shown across all $r_+ \in [0, r]$, but it is important to check also that $r_- = r_+ - (n-m) \geq 0$ and $r_0 = r - 2r_+ + (n-m) \geq 0$. The triangle inequalities $|n-m| \leq r \leq n+m$ imply the following bounds $\max\{0, n-m\} \leq r_+ \leq \frac{1}{2}(r+n-m) \leq r$, which can be used for a more efficient implementation.

Formula (10) for the binary case $\alpha = 2$ was previously analysed in [19] (see also [14, 28]). The solution for arbitrary alphabets was first given in [2] (see also [3–5]). We now have all information required to find probability (9).

Theorem 1 *(Geometric probability of mutation onto a sphere [2–5]) The probability $P(m \mid n, r)$ that a substitution of $r \in [0, l]$ letters in string $x \in S(\top, n) \subset \mathcal{H}_\alpha^l$ results in string $y \in S(\top, m)$ is*

$$P(m \mid n, r) = \frac{\sum_{r_+=0}^{r} (\alpha-2)^{r_0} \binom{n-r_+}{r_0} (\alpha-1)^{r_-} \binom{l-n}{r_-} \binom{n}{r_+}}{(\alpha-1)^r \binom{l}{r}} \quad (11)$$

with $r_+ \in [0, r]$ and the numbers $r_- \geq 0$, $r_0 \geq 0$ defined by the equations

$$r_- = r_+ - (n-m)\,, \qquad r_0 = r - 2r_+ + (n-m)\,.$$

The probability is zero if the triangle inequalities $|n-r| \leq m \leq n+r$ are not satisfied.

Proof The probability is given by the proportion of strings in sphere $S(x, r)$ that are also in the sphere $S(\top, m)$:

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

The number in the intersection is given by (10), and the number of elements in $S(x, r)$ is $(\alpha-1)^r \binom{l}{r}$. □

Remark 2 The proof above makes an implicit assumption about a uniform distribution within subsets (spheres), which is common if no other information about the distribution is given (i.e. the principle of insufficient reason).

Example 3 (Binary case $\alpha = 2$) For binary strings, formula (11) reduces to:

$$P(m \mid n, r) = \frac{\binom{l-n}{r-r_+}\binom{n}{r_+}}{\binom{l}{r}}, \tag{12}$$

where $r_+ = \frac{1}{2}(r + n - m)$ must be non-negative integer (otherwise, the probability is zero). Note that the right-hand-side of (12) is the hypergeometric distribution $\mathbb{P}\{X = r_+\}$ of $r_+ \in [0, r]$ if it is considered as a random variable. The above formula is also valid for $\alpha > 2$ when the mutation radius $d(x, y) = r$ is minimized ($r = |n - m|$) or maximized ($r = n + m$), because there are no neutral substitutions in these cases.

Conditional probability (11) was implemented in a digital computer using Common Lisp programming language, and Fig. 2 illustrates its dependency on parameters n and r in Hamming space $\mathcal{H}_4^{40}$ ($\alpha = 4$, $l = 40$). Three charts correspond to three values of distance $n \in \{5, 20, 35\}$ of the parent string. Abscissae show mutation radii $r = d(x, y)$, while ordinates show the resulting distances $m = d(\top, y)$ of offsprings after mutation. The grayscale represents different values of probability $P(m \mid n, r)$ with white corresponding to $P = 0$ and black to $P = 1$. One can see that the mutation radius has different effects on the probability depending on whether the parent's distance $n = d(\top, x)$ is less or greater than the 'equator' $l(1 - 1/\alpha)$: for $n < l(1 - 1/\alpha)$ higher mutation radius makes larger distances $m = d(\top, y)$ more likely, but the effect reverses for $n > l(1 - 1/\alpha)$. This corresponds to the fact that in Hamming space closed ball with radius $n > l(1 - 1/\alpha)$ is larger than its complement. One may also notice from Fig. 2 that the expected distance $\mathbb{E}_P\{m \mid n, r\}$ may have a simple relation to the mutation radius $r = d(x, y)$. This relation is given below.

Proposition 1 *The expected value and variance of the conditional probability distribution* (11) *for Hamming distance* $m = d(\top, y)$ *of string* $y \in S(x, r)$ *obtained by a substitution of* $d(x, y) = r$ *letters in string* $x \in S(\top, n) \subset \mathcal{H}_\alpha^l$ *are*

$$\mathbb{E}_P\{m \mid n, r\} = n + \left(1 - \frac{n}{l(1 - 1/\alpha)}\right) r, \tag{13}$$

$$\sigma_P^2\{m \mid n, r\} = n \left[\frac{\alpha - 2}{(\alpha - 1)^2} + \frac{(l - n)(l - r)}{(1 - 1/\alpha)^2 l(l - 1)}\right] \frac{r}{l}. \tag{14}$$

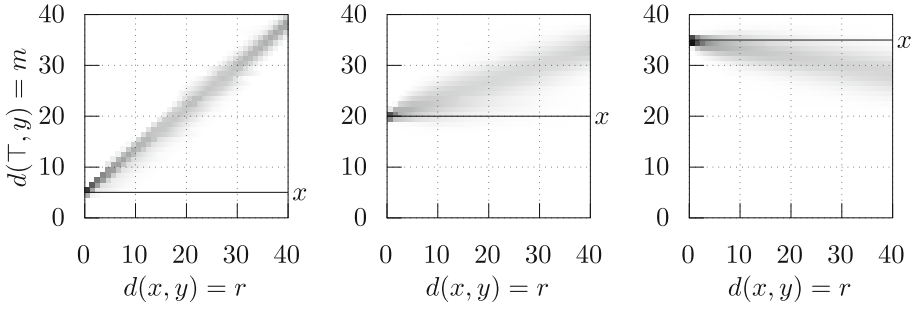


Fig. 2 Dependency of the probability $P(m \mid n, r)$ (11) on the mutation radius r (abscissae) in space $\mathcal{H}_4^{40}$. Ordinates show the resulting distance $d(\top, y) = m$ after mutation. Three charts correspond to three distances $d(\top, x) = n \in \{5, 20, 35\}$ of the parent string. Grayscale represents probability $P \in [0, 1]$

The proof uses formulae (2) and (3) from Lemma 1, where the average probabilities $\langle P_i \rangle$ and $\langle P_{ij} \rangle$ are defined as functions of parameters $n = d(\top, x)$ and $r = d(x, y)$. See Appendix A.3 for details.

Formula (13) confirms the linear dependency of posterior expectation of distance $m = d(\top, y)$ on the mutation radius as can be seen on Fig. 2. The slope of this dependency is $1 - n/l(1 - 1/\alpha)$, which is positive if $d(\top, x) = n < l(1 - 1/\alpha)$ (i.e. if x is closer to $\top$ than the 'equator') and negative if $n > l(1 - 1/\alpha)$. One can see also from (13) that at distance $n = l(1 - 1/\alpha)$ the expected value of $m = d(\top, y)$ becomes independent of the mutation radius $r = d(x, y)$ and is equal to $l(1 - 1/\alpha)$. Also, the value $r = l(1 - 1/\alpha)$ of the mutation radius makes the expectation of distance equal to $l(1 - 1/\alpha)$ and independent of the parent's distance $n = d(\top, x)$, which corresponds to the uniform distribution $P_0(x) = \alpha^{-l}$ of strings.

Differentiation of (14) over r and setting the derivative to zero

$$\frac{\partial}{\partial r}\sigma^2\{m \mid n, \hat{r}\} = \frac{n}{l}\left[\frac{(\alpha - 2)}{(\alpha - 1)^2} + \frac{(l - 2\hat{r})(l - n)}{l(1 - 1/\alpha)^2(l - 1)}\right] = 0$$

gives the saddle point, which is the maximum, because the second derivative is negative (observe that the radius appears only in $(l - 2r)$ with the minus sign). The mutation radius maximizing the variance of offspring's distance $d(\top, y) = m$ is

$$\hat{r}(n) = \frac{l}{2}\left[1 + \frac{(\alpha - 2)(l - 1)}{\alpha^2(l - n)}\right].$$

One can see that generally, unless $\alpha = 2$, the maximizing mutation radius depends on the distance $d(\top, x) = n$ of the parent string, and it is not equal to the Hamming space equator $l(1 - 1/\alpha) = \frac{l}{2}[1 + (\alpha - 2)/\alpha]$.

2.4 Maximization of probability of beneficial mutation

The closed-form expression (11) combined with probability $P(r \mid n)$ of the mutation radius gives complete solution to transition probability (7) between spheres around the optimum under mutation. This makes it possible to study Markov evolution of distance distributions and solve related optimization problems. Here we consider two simple problems that have exact solutions.

Proposition 2 *(Minimization of the expected distance after mutation) The expected value* (13) *of Hamming distance $d(\top, y) = m$ after mutation of $x \in S(\top, n) \subset \mathcal{H}^l_\alpha$ into $y \in S(\top, m)$ is minimized if the mutation radius $d(x, y) = r \in [0, l]$ has the values $r = 0$ for $n < l(1-1/\alpha)$, $r = l$ for $n > l(1 - 1/\alpha)$, and any value for $n = l(1 - 1/\alpha)$.*

Proof The minimization $\mathbb{E}_P\{m \mid n, r\} < n$ over the mutation radius $r \in [0, l]$ follows from (13). □

It follows that for the simple point mutation operator (Example 2) the mutation rate minimizing the expected distance after one mutation is

$$\hat{\mu}(n) = \begin{cases} 0 & \text{if } n < l(1 - 1/\alpha) \\ 1 - 1/\alpha & \text{if } n = l(1 - 1/\alpha) \\ 1 & \text{if } n > l(1 - 1/\alpha) \end{cases}.$$

Although the offspring's expected distance $\mathbb{E}\{m \mid n, r\}$ is independent of the mutation radius at distance $n = l(1 - 1/\alpha)$ (so that μ can have any value), we use the value $1 - 1/\alpha$,

because several known mutation rate functions $\mu(n)$ are monotonic and pass through this value (e.g. the linear function $\mu(n) = n/l$, derived below, passes through $1 - 1/\alpha$ at $n = l(1 - 1/\alpha)$).

Note 'step' mutation rate function $\hat{\mu}(n)$ above has the following problem. Notice that for binary strings ($\alpha = 2$) at distances $d(\top, x) = n > l/2$ the mutation rate $\mu = 1$ changes the distances to $d(\top, y) = m = l - n < l/2$. Thus, all bitstrings will be at distance $d(\top, y) \leq l/2$ after just one generation. A similar effect will occur for strings with alphabets $\alpha > 2$, but it may take several generations because of the possibility of neutral substitutions. Therefore, after multiple generations the distribution of distances $P_{t+s}(m) = M^s_{\mu(n)} P_t(n)$, $s > 1$, will not change due to mutation with rate $\mu = 0$. It is clear that the step mutation rate function is not optimal for evolution over multiple generations. Computer simulations show that a sigmoid type mutation rate functions achieve optimality for multiple generations [4], but analytic derivation of such results is not straightforward. Another approach is to maximize the probability of mutation directly into the optimum: $x \mapsto y = \top$.

Proposition 3 *(Mutation into the optimum) The probability $P(m = 0 \mid n, r)$ that string $x \in S(\top, n) \subset \mathcal{H}^l_\alpha$ mutates into string $y = \top$ is zero unless $d(x, y) = r(n) = n$, in which case the probability is*

$$P(m = 0 \mid n, r = n) = \frac{1}{|S(x, n)|} = \frac{1}{(\alpha - 1)^n \binom{l}{n}} .$$

Proof This obvious result can be obtained formally by substituting the values $r_+ = n$, $r_- = r_0 = 0$ into (11). □

Substitution of the above probability, which is reciprocal of the number of elements in $S(x, n)$, into (7) and using binomial distribution (8) for the mutation radius gives the following formula for the probability of transition into optimum under simple point mutation (Example 2):

$$P_\mu(m = 0 \mid n) = (\alpha - 1)^{-n} \mu^n(n)[1 - \mu(n)]^{l-n} .$$

The optimal mutation rate $\hat{\mu}$ maximizing this probability is

$$\frac{\partial}{\partial \mu} P_\mu = (\alpha - 1)^{-n} \hat{\mu}^n [1 - \hat{\mu}]^{l-n} \underbrace{\left(\frac{n}{\hat{\mu}} - \frac{l - n}{1 - \hat{\mu}} \right)}_{=0} = 0 \qquad \Longrightarrow \qquad \hat{\mu}(n) = \frac{n}{l} , \quad (15)$$

$$\frac{\partial^2}{\partial \mu^2} P_\mu = (\alpha - 1)^{-n} \hat{\mu}^n [1 - \hat{\mu}]^{l-n} \left[\underbrace{\left(\frac{n}{\hat{\mu}} - \frac{l - n}{1 - \hat{\mu}} \right)^2}_{=0} - \frac{l - 2n/\hat{\mu} + n/\hat{\mu}^2}{(1 - \hat{\mu})^2} \right]$$

$$= (\alpha - 1)^{-n} \hat{\mu}^n [1 - \hat{\mu}]^{l-n} \left[\frac{l(1 - l/n)}{(1 - n/l)^2} \right] \leq 0 .$$

Indeed, one can see that for $\hat{\mu} = n/l$ the first derivative is zero, and the second derivative is negative, because $1 - l/n \leq 0$. Observe also that as the diameter of Hamming space increases $l \to \infty$ the optimal mutation rate $\hat{\mu}(n) = n/l$ converges to zero for each $n \in \mathbb{N}$.

These examples show that minimization of the mutation radius or mutation rate may not always be optimal in Hamming space. The heuristic value $\mu = 1/l$ [15] is optimal only at distance $d(\top, x) = 1$ with respect to the criterion of Proposition 3. Although the exact shapes of the mutation rate functions optimal subject to additional constraints may differ

(e.g. see examples in [3–5, 8] for constraints on the number of generations or information constraints), these functions have the common property of monotonically increasing optimal mutation rates $\mu(n)$ with distance $d(\top, \cdot) = n$ from the optimum.

3 Recombination

3.1 Crossover recombination in a Hamming space

Recombination is a substitution of some letters in string $x \in \mathcal{H}_\alpha^l$ by letters from another string $y \in \mathcal{H}_\alpha^l$. In this paper, we shall only consider *crossover* recombination when letters x_i are substituted by letters y_i with the same index $i \in \{1, \ldots, l\}$. This corresponds to Hamming metric (1) accounting for differences of letters only at the same indices.

Because in recombination we have to consider two parent strings x, y and their distances from $\top$, we have a triangle $(x, y, \top)$ and three distances:

$$n = d(\top, x)$$
$$k = d(\top, y)$$
$$h = d(x, y)$$

as shown on Fig. 3. The number $r \in [0, l]$ of letters that are exchanged during crossover between x and y will be referred to as the *recombination radius*, and it can be larger than the distance $h = d(x, y)$ between two strings, because crossover may recombine identical letters. After crossover of string x with y (the parents), the resulting new string z (the offspring) is a mixture of $l - r$ letters from x and r letters from y, and it is created 'between' its parents in the sense of Hamming distance: $d(z, y) \leq d(x, y)$ and $d(x, z) \leq d(x, y)$. It is convenient to denote the result of recombination as $z = (1 - \lambda)x \oplus \lambda y$, where $\lambda = r/l$, by analogy with convex combination in a real space, although this is only a notational convenience (hence the use of symbol $\oplus$ instead of $+$).

Recombination of x with y into z using r letters has a dual recombination z', which can be formed as a substitution of the remaining $l - r$ letters from y into x. Thus, the dual recombination is $z' = \lambda x \oplus (1-\lambda)y$ in our 'convex combination' notation. Contrary to a real space, a mixture $z = (1 - \lambda)x \oplus \lambda y$ in a Hamming space is not unique, because it depends on positions at which r letters are substituted. The totality of all possible recombinations of r letters between two strings has been called a *recombination potential* [29]:

$$I(x, y, r) := \{z = (1 - \lambda)x \oplus \lambda y : \lambda = r/l\}.$$

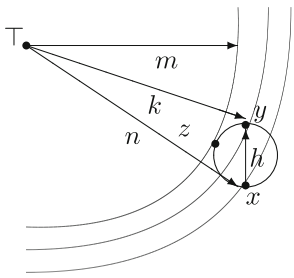


Fig. 3 Recombination of string $x \in S(\top, n)$ with $y \in S(\top, k)$ into string $z \in S(\top, m)$ by crossover of $r \in [0, l]$ letters. The number of strings in the intersection $S(\top, m) \cap I(x, y, r)$ defines the geometric probability $P(m \mid n, k, h, r)$ (22)

The number of strings in $I(x, y, r)$ is

$$|I(x, y, r)| = \binom{l}{r}.$$

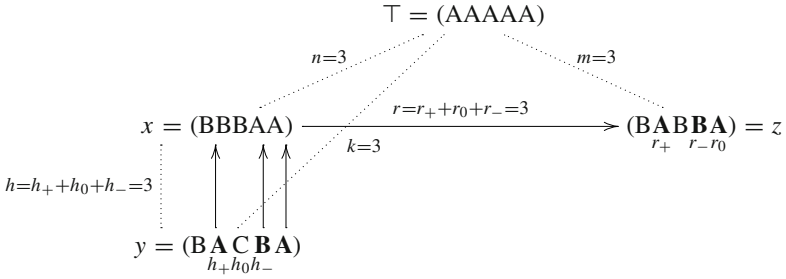
Note that some strings in $I(x, y, r)$ may appear more than once, because different recombinations may result in the same offspring. This makes $I(x, y, r)$ a multiset. Also, because substitution of r letters from y into x is the same as substitution of the remaining $l - r$ letters from x into y, we have the following equality:

$$I(x, y, r) = I(y, x, l - r).$$

Exchanging different letters $x_i \neq y_i$ at the same indices cannot make them equal, so that the Hamming distance between two parent strings x, y and between their recombinations $z = (1-\lambda)x \oplus \lambda y$ and $z' = (1-\lambda)y \oplus \lambda x$ remain the same: $d(x, y) = d(z, z')$. This implies that recombination potential $I(x, y, r)$ has a round shape: its elements belong to a sphere of diameter $h = d(x, y)$ as shown on Fig. 3.

If distance $d(\top, \cdot)$ from the optimum $\top$ is taken as a model of (negative) fitness, then recombination is called *beneficial* for parent $x \in S(\top, n)$ if it corresponds to a transition onto sphere $S(\top, m) \ni z$ of a smaller radius $m < n$, as shown on Fig. 3. Recombination is *neutral* if $m = n$, and *deleterious* if $m > n$. Note that a recombination can be beneficial for $x \in S(\top, n)$, but not necessarily for $y \in S(\top, k)$.

Example 4 (Crossover recombination) Let $\top = (\mathrm{AAAAA}) \in \mathcal{H}_3^5$ and consider strings $x = (\mathrm{BBBAA})$ and $y = (\mathrm{BACBA})$ recombining into string $z = (\mathrm{BABBA})$, which can be obtained by a substitution of the second, fourth and the last letters in x by the corresponding letters from y:



The dual offspring is $z' = (\mathrm{BBCAA})$. Thus, the recombination radius is $r = 3$, and the recombination is neutral, because $d(\top, x) = d(\top, z) = 3$. Note that although the recombination radius was $r = 3$, only two of these substitutions occurred out of $d(x, y) = h = 3$ different letters; the third substitution was made for identical letters. A substitution of the second, third and the last letters in $x = (\mathrm{BBBAA})$ by the corresponding letters from $y = (\mathrm{BACBA})$ results in string $v = (\mathrm{BACAA})$, which is closer to the optimum, $d(\top, v) = 2$.

As with mutation, we denote by r_+, r_- and r_0 the numbers of beneficial, deleterious and neutral substitutions respectively. They satisfy the same equations as (4) and (5):

$$r_+ + r_- + r_0 = r, \tag{16}$$

$$r_+ - r_- = n - m. \tag{17}$$

Here, $n = d(\top, x)$ and $m = d(\top, z)$. The numbers r_+ and r_- count beneficial or deleterious substitutions for parent x, but not necessarily for parent y. These substitutions can

only occur for $h = d(x, y)$ different letters. Therefore, $r_+ \leq h_+$ and $r_- \leq h_-$, where h_+ and h_- are the maximum possible numbers of beneficial and deleterious substitutions for x and $h_+ + h_- \leq h = d(x, y)$. Denoting by h_0 the maximum number of possible neutral substitutions out of h different letters, the equations for these maximal numbers are

$$h_+ + h_- + h_0 = h\,, \tag{18}$$

$$h_+ - h_- = n - k\,. \tag{19}$$

These equations are identical to (4) and (5) if parent y is considered as a mutated version of parent x with $h = d(x, y)$ considered as the mutation radius. As mentioned earlier, substitutions during crossover may occur also among the $l - d(x, y) = l - h$ identical letters (as in Example 4). Such substitutions are neutral, and therefore it is possible that $r_0 > h_0$. In fact, the range of recombination radius is $r \in [0, l]$, and it can be larger than $h = d(x, y)$. Changes between the parent and offspring strings are defined only by the distance $h = d(x, y)$, and we shall refer to it as *recombination capacity*.

The range of recombination capacity is defined by the triangle inequalities:

$$|n - k| \leq h \leq n + k$$

As with mutation radius, there are no potential neutral substitutions ($h_0 = 0$) at the extreme values $h = |n - k|$ or $h = n + k$.

Proposition 4 *(Dual recombination) If crossover recombination of $x \in S(\top, n)$ with $y \in S(\top, k)$ by exchanging $r \in [0, l]$ letters results in string $z \in S(\top, m)$, then the dual recombination $z' \in S(\top, m')$ is at distance $d(\top, z') = m'$:*

$$m' = n + k - m\,.$$

See Appendix A.4 for the proof. Intuitively, if x receives $n - m$ letters $y_i = \top_i$ from y, then the Hamming distance reduces from $d(\top, x) = n$ to $d(\top, z) = n - (n - m) = m$. At the same time, string y receives $n - m$ letters $x_i \neq \top_i$ from x (the dual recombination), which means that $d(\top, y) = k$ changes to $d(\top, z') = k + n - m$.

3.2 Evolutionary dynamics under recombination

Crossover recombination can be viewed as a transition from the pair $(x, y) \in \mathcal{H}_\alpha^l \times \mathcal{H}_\alpha^l$ of two parent strings to the pair $(z, z') \in \mathcal{H}_\alpha^l \times \mathcal{H}_\alpha^l$ of recombination z and its dual z'. The corresponding transition probability $P(z, z' \mid x, y)$ induces a transformation of the distribution $P_t(x, y)$ of the parent pairs into the distribution $P_{t+1}(z, z')$ of the offspring pairs:

$$P_{t+1}(z, z') = \sum_{(x,y) \in \mathcal{H}_\alpha^l \times \mathcal{H}_\alpha^l} P(z, z' \mid x, y)\, P_t(x, y)\,.$$

Joint distributions $P_t(x, y)$ represent pairing probabilities of the parent strings for recombination, and they may depend on various properties such as fitness of individuals as well as their similarity. Thus, generally $P_t(x, y) = P_t(y \mid x) P_t(x)$ and $P_t(y \mid x) \neq P_t(y)$.

The joint pairing distributions $P_t(x, y)$ induce the corresponding joint distributions of distances $n = d(\top, x)$ and $k = d(\top, y)$ of the paired strings:

$$P_t(n, k) := P_t\{x \in S(\top, n), y \in S(\top, k)\}\,.$$

These joint distributions can be formed as products $P_t(k \mid n)P_t(n)$, where $P_t(n) := P_t\{x \in S(\top, n)\}$. If strings (x, y) are paired independently, then $P_t(k, n) = P_t(k)P_t(n)$. However, generally $P_t(k \mid n) \neq P_t(k)$.

Example 5 (Matching) If string $x \in S(\top, n)$ is paired with string $y \in S(\top, k)$ at equal distances $d(\top, x) = n = k = d(\top, y)$, then

$$P(k \mid n) = \delta_n(k)\,, \qquad P_t(n, k) = \begin{cases} P_t(n) & \text{if } k = n \\ 0 & \text{otherwise} \end{cases}.$$

This joint distribution may occur as an equilibrium solution to a minimax problem [34], when both parents minimize distances $d(\top, \cdot)$ of the strings they are recombined with (i.e. maximizing fitness of their partners).

Apart from the distances from $\top$, each pair of strings $x, y \in \mathcal{H}_\alpha^l$ is characterized also by their distance $h = d(x, y)$ (recombination capacity). The joint distribution $P_t(n, k, h) = P_t(h \mid n, k)P_t(n, k)$ is also defined by the pairing distribution $P_t(x, y)$.

Example 6 (Random pairing from a sphere) Consider string $x \in S(\top, n)$ paired with string $y \in S(\top, k)$ (i.e. distances $d(\top, x) = n$ and $d(\top, y) = k$ are fixed). If string $y \in S(\top, k)$ is chosen uniformly at random from $S(\top, k)$, then conditional probability $P(h \mid n, k)$ is defined by the intersection of spheres $S(x, h)$ and $S(\top, k)$ (see Fig. 3):

$$\begin{aligned} P(h \mid n, k) &= \frac{|S(x, h) \cap S(\top, k)|_{d(\top,x)=n}}{|S(\top, k)|} \\ &= \frac{\sum_{h_+=0}^{h} (\alpha-2)^{h_0} \binom{n-h_+}{h_0} (\alpha-1)^{h_-} \binom{l-n}{h_-} \binom{n}{h_+}}{(\alpha-1)^k \binom{l}{k}}\,. \end{aligned}$$

The latter formula is obtained using (10) for intersection of spheres with distance $d(x, y) = h = h_+ + h_- + h_0$ treated as the mutation radius $r = r_+ + r_- + r_0$ and substituting k for m. The probability is zero if the triangle inequalities $|n - k| \leq h \leq n + k$ are not satisfied.

Example 7 (Pairing at specific distance) One can try to pair strings choosing a specific value of the distance $d(x, y) = h$ between the parent strings (assuming that the current population has individuals satisfying this equality constraint). The range of $h \in [0, l]$ is defined by the triangle inequalities: $|n - k| \leq h \leq n + k$ (and $n + k \leq l$). For example, choosing the maximum value $h = \min\{n + k, l\}$ corresponds to the probability

$$P(h \mid n, k) = \delta_{\min\{n+k,l\}}(h)\,.$$

Minimization of h corresponds to $\delta_{|n-k|}(h)$, and the average $h = \frac{1}{2}(|n - k| + n + k) = \max\{n, k\}$ to $\delta_{\max\{n,k\}}(h)$.

Recombination of strings (x, y) into (z, z'), where z' denotes the dual recombination, results in a transformation of distances $n = d(\top, x)$ and $k = d(\top, y)$ into distances $m = d(\top, z)$ and $m' = d(\top, z')$. The joint distributions of distance pairs $P_t(n, k) := P_t\{x \in S(\top, n), y \in S(\top, k)\}$ and $P_{t+1}(m, m') := P_{t+1}\{z \in S(\top, m), z' \in S(\top, m')\}$ are transformed as follows:

$$P_{t+1}(m, m') = \sum_{n=0}^{l} \sum_{k=0}^{l} P(m, m' \mid n, k)\, P_t(n, k)\,,$$

where $P(m, m' \mid n, k)$ is the transition probability between the pairs of spheres of radii (n, k) and (m, m'):

$$P(m, m' \mid n, k) := P\{z \in S(\top, m), z' \in S(\top, m') \mid x \in S(\top, n), y \in S(\top, k)\}.$$

The analysis is similar to mutation, but the transformations are now applied to joint distributions of distance pairs. One can also obtain the transformation of distance distribution $P_t(n) := P_t\{x \in S(\top, n)\}$ into $P_{t+1}(m) := P_{t+1}\{z \in S(\top, m)\}$:

$$P_{t+1}(m) = \sum_{m'=0}^{l} \sum_{n=0}^{l} \sum_{k=0}^{l} P(m, m' \mid n, k)\, P(k \mid n) P_t(n).$$

Note that $m' = n + k - m$ for crossover recombination (Proposition 4), so that

$$P(m, m' \mid n, k) = \begin{cases} P(m \mid n, k) & \text{if } m' = n + k - m \\ 0 & \text{otherwise} \end{cases}.$$

Thus, the summation over $m' \in [0, l]$ above is not necessary, and it is sufficient to derive the expressions using only probability $P(m \mid n, k)$.

If the transition kernels $P(m \mid n, k)$ and $P(k \mid n)$ are time invariant, then the linear operator

$$R(\cdot) = \sum_{n=0}^{l} \left[\sum_{k=0}^{l} P(m \mid n, k)\, P(k \mid n) \right] (\cdot) \tag{20}$$

acting on distributions $P_t(n)$ of distances $d(\top, x) = n \in [0, l]$ generates the entire evolution $\{P_t\}_{t \geq 0}$ due to recombination as $P_{t+s} = R^s P_t$. This can be used in simulations to analyse the effects of recombination and pairing strategies on evolution.

In Section 2 on mutation, we expanded transition probability $P(m \mid n)$ over all values of the mutation radius $r \in [0, l]$ (7). Similarly, here we expand the transition probability $P(m \mid n, k)$ over all values of the recombination radius $r \in [0, l]$ and recombination capacity $h = d(x, y)$:

$$P(m \mid n, k) = \sum_{r=0}^{l} \sum_{h=0}^{l} P(m \mid n, k, h, r) \underbrace{P(r \mid n, k, h)}_{\text{Recombination}} \underbrace{P(h \mid n, k)}_{\text{Pairing}}. \tag{21}$$

The probability $P(h \mid n, k)$ has been discussed in Examples 6 and 7. The probability $P(r \mid n, k, h)$ of recombination radius $r \in [0, l]$ can be determined from the analysis of the recombination operator.

Example 8 (Uniform crossover) In this form of recombination, letters x_i and y_i at each position $i \in \{1, \ldots, l\}$ in the parent strings are swapped with probability $v \in [0, 1]$, called the *recombination rate*, independently of letters x_j and y_j at other positions. In this case, $P(r \mid n, k, h)$ is the binomial distribution:

$$P_v(r \mid n, k, h) = \binom{l}{r} v^r(n, k, h)[1 - v(n, k, h)]^{l-r}.$$

The rate v may be different for different values of $n, k, h \in [0, l]$, so that the recombination operator depends on the recombination rate control function $v(n, k, h)$.

Example 9 (One point crossover) In this form of recombination a single index $i \in \{1, \ldots, l\}$ is selected in the parent strings x and y, and all letters x_j, y_j with $j \geq i$ are swapped. Thus, if

$i = \lfloor l/2 \rceil$, then approximately half of the letters are swapped, and the recombination radius is $r = \lfloor l/2 \rceil$ (here $\lfloor \cdot \rceil$ denotes the nearest integer). In this case, $P(r \mid n, k, h)$ is the Dirac distribution:

$$P(r \mid n, k, h) = \delta_{\lfloor l/2 \rceil}(r) .$$

Observe that the recombination radius in this case is equal to $r = l - i$, where i is the index of one point crossover. Potentially, one can define one-point crossover with variable index $i = l - r$, where the value $r = r(n, k, h)$ of the recombination radius may depend on distances n, k and h.

Using factorization (21), the recombination operator (20) acting on distributions of distances from $\top$ takes the form

$$R(\cdot) = \sum_{n=0}^{l} \left[\sum_{k=0}^{l} \sum_{r=0}^{l} \sum_{h=0}^{l} P(m \mid n, k, h, r)\, P(r \mid n, k, h)\, P(h \mid n, k) P(k \mid n) \right] (\cdot) ,$$

where probabilities $P(k, h \mid n) = P(h \mid n, k) P(k \mid n)$ are defined by the pairing strategy (Examples 5, 6, 7) and $P(r \mid n, k, h)$ by the crossover process (Examples 8, 9). The unknown conditional probability $P(m \mid n, k, h, r)$ will be determined in the next section.

3.3 Geometric probability of recombination onto a sphere

Similar to geometric probability $P(m \mid n, r)$ defined in (9) for mutation, the probability $P(m \mid n, k, h, r)$ in factorization (21) represents a purely geometric problem depicted on Fig. 3:

$$P(m \mid n, k, h, r) := P\{z \in S(\top, m) \mid x \in S(\top, n),\, y \in S(\top, k) \cap S(x, h),\, r \in [0, l]\} \quad (22)$$

Solution to this geometric problem requires counting the number of elements in certain subsets. The sought offspring strings are in the intersection of recombination potential $I(x, y, r)$ and sphere $S(\top, m)$. The difference $n - m$ of the radii of spheres $S(\top, n)$ and $S(\top, m)$ defines the difference $r_+ - r_-$ of beneficial and deleterious substitutions into string x from y (17). The upper bounds $r_+ \le h_+$ and $r_- \le h_-$ can be defined from the conditions on the parent strings: $x \in S(\top, n)$ and $y \in S(\top, k) \cap S(x, h)$ (i.e. by the distances n, k and h for the triangle $(\top, x, y)$). Indeed, $h_+ \le h = d(x, y)$ and $h_- = h_+ - (n - k)$ by (19), where $n = d(\top, x)$ and $k = d(\top, y)$. It is convenient to count strings in the intersection $S(\top, m) \cap I(x, y, r)$ by grouping them based on the values $h_+ \in [0, h]$ of possible beneficial substitutions. We shall denote by $[h_+]$ the class of all strings y that have $r_+ \le h_+$ beneficial substitutions into x.

Lemma 3 *(Intersection of sphere and recombination potential) Let $x \in S(\top, n) \subset \mathcal{H}_\alpha^l$, and let $[h_+]$ be the class of all strings $y \in S(\top, k) \cap S(x, h)$ such that the number of beneficial substitutions from y into x (reducing the distance $d(\top, x) = n$) be at most $h_+ \le h$. Then the number of elements in the intersection of $S(\top, m)$ with recombination potential $I(x, y, r)$ for $y \in [h_+]$ is*

$$\left| S(\top, m) \cap I(x, y, r) \right|_{y \in [h_+]} = \sum_{r_+=0}^{h_+} \binom{l - h_+ - h_-}{r - r_+ - r_-} \binom{h_-}{r_-} \binom{h_+}{r_+} , \quad (23)$$

where $h_- = h_+ - (n - k) \ge 0$, $r_- = r_+ - (n - m) \ge 0$ and $r_+ \in [0, h_+]$, $r - r_+ - r_- \ge 0$.

Proof Given the maximum numbers h_+ and $h_- = h_+ - (n-k)$ of beneficial and deleterious substitutions (observe that h_- depends on $n = d(\top, x)$ and $k = d(\top, y)$), the remaining $l - h_+ - h_-$ letters can only make neutral substitutions. Thus, for specific values of $r_+ \le h_+$, $r_- \le h_-$, the total number of combinations is

$$\binom{l-h_+-h_-}{r-r_+-r_-}\binom{h_-}{r_-}\binom{h_+}{r_+},$$

where $r_- = r_+ - (n-m)$ from (17), and $r - r_+ - r_- \ge 0$ is the total number of neutral recombinations from (16). Feasible values of $r_+ \in [0, h_+]$ are determined from non-negativity of r_- and $r - r_+ - r_-$. The total number is obtained by summing over all feasible values of $r_+ \in [0, h_+]$. □

The maximum number of beneficial substitutions is bounded above $h_+ \le h = d(x, y)$, and adding the numbers in (23) for all $h_+ \in [0, h]$ would account for all strings in the intersection $S(\top, m) \cap I(x, y, r)$. However, the classes $[h_+]$ are not distributed uniformly as they have different sizes. Indeed, the totality of all strings y to be recombined with $x \in S(\top, n)$ is the intersection of spheres $S(\top, k)$ and $S(x, h)$, as shown on Fig. 3. The number of strings in this intersection is given by formula (10) in Lemma 2, where instead of mutation radius $r = r_0 + r_- + r_0$ we have to use recombination capacity $h = h_+ + h_- + h_0$ and substituting k for m. In fact, the number of strings y in the intersection $S(\top, k) \cap S(x, h)$ with specific values of the maximum numbers h_+, h_- and h_0 is

$$(\alpha-2)^{h_0}\binom{n-h_+}{h_0}(\alpha-1)^{h_-}\binom{l-n}{h_-}\binom{n}{h_+}.$$

This can be used to derive probability (22).

Theorem 2 *(Geometric probability of recombination onto a sphere) The probability $P(m \mid n, k, h, r)$ that crossover recombination of $r \in [0, l]$ letters in string $x \in S(\top, n) \subset \mathcal{H}_\alpha^l$ with string $y \in S(\top, k) \cap S(x, h)$ results in string $z \in S(\top, m)$ is*

$$P(m \mid n, k, h, r) = \frac{\sum_{h_+=0}^{h}(\alpha-2)^{h_0}\binom{n-h_+}{h_0}(\alpha-1)^{h_-}\binom{l-n}{h_-}\binom{n}{h_+}\sum_{r_+=0}^{h_+}\binom{l-h_+-h_-}{r-r_+-r_-}\binom{h_-}{r_-}\binom{h_+}{r_+}}{\binom{l}{r}\sum_{h_+=0}^{h}(\alpha-2)^{h_0}\binom{n-h_+}{h_0}(\alpha-1)^{h_-}\binom{l-n}{h_-}\binom{n}{h_+}} \tag{24}$$

with $r_+ \in [0, h_+]$, $h_+ \in [0, h]$ and the numbers $h_- \ge 0$, $h_0 \ge 0$, $r_- \ge 0$, $r - r_+ - r_- \ge 0$ defined by the equations

$$h_- = h_+ - (n-k), \qquad h_0 = h - 2h_+ + (n-k), \qquad r_- = r_+ - (n-m).$$

The probability is zero if the triangle inequalities $|n-k| \le h \le n+k$ or $|n-m| \le \min\{r, h\} \le n+m$ are not satisfied.

Proof The probability $P(m \mid n, k, h, r)$ can be expressed as the following sum of products of conditional probabilities $P(m \mid n, k, h, r, h_+)$ and $P(h_+ \mid n, h, k)$ for all values of $h_+ \in [0, h]$:

$$P(m \mid n, k, h, r) = \sum_{h_+=0}^{h} P(m \mid n, k, h, r, h_+) P(h_+ \mid n, k, h). \tag{25}$$

Observe that the recombination radius r does not occur in probability $P(h_+ \mid n, h, k)$. This is because the maximum number of possible beneficial recombinations $h_+ \in [0, h]$ is defined solely by the parent strings, and it is independent of r.

Here, $P(m \mid n, k, h, r, h_+)$ is the probability that an offspring z in the recombination potential $I(x, y, r)$ is at distance $m = d(\top, z)$ from the optimum subject to the condition that the parent string y has at most h_+ potential beneficial recombinations (fixing specific three points $(\top, x, y)$ in a Hamming space also fixes their distances $n = d(\top, x), k = d(\top, y)$ and $h = d(x, y)$). This probability is the ratio of strings in the intersection of sphere $S(\top, m)$ with potential $I(x, y, r)$ with the condition $y \in [h_+]$ over all offspring strings in $I(x, y, r)$:

$$P(m \mid n, k, h, r, h_+) = \frac{|S(\top, m) \cap I(x, y, r)|_{y \in [h_+]}}{|I(x, y, r)|} .$$

The number $|S(\top, m) \cap I(x, y, r)|_{y \in [h_+]}$ is given by (23), and the number of elements in the potential $I(x, y, r)$ is $\binom{l}{r}$, so that

$$P(m \mid n, k, h, r, h_+) = \frac{\sum_{r_+=0}^{h_+} \binom{l-h_+-h_-}{r-r_+-r_-}\binom{h_-}{r_-}\binom{h_+}{r_+}}{\binom{l}{r}} , \tag{26}$$

where $r_+ - r_- = n - m$ with the constraints $r_- \geq 0, r - r_+ - r_- \geq 0$ and $h_- = h_+ - (n-k) \geq 0$.

Probability $P(h_+ \mid n, k, h)$ is the ratio of parent strings y with at most h_+ beneficial recombinations in the intersection $S(\top, k) \cap S(x, h)$ out of all parent strings in this intersection:

$$P(h_+ \mid n, k, h) = \frac{|S(\top, k) \cap S(x, h)|_{d(\top,x)=n,\, d(\top,y)=k,\, y \in [h_+]}}{|S(\top, k) \cap S(x, h)|_{d(\top,x)=n}} .$$

Using (10) for intersection of the spheres and making the described earlier substitutions this probability is

$$P(h_+ \mid n, k, h) = \frac{(\alpha-2)^{h_0}\binom{n-h_+}{h_0}(\alpha-1)^{h_-}\binom{l-n}{h_-}\binom{n}{h_+}}{\sum_{h_+=0}^{h} (\alpha-2)^{h_0}\binom{n-h_+}{h_0}(\alpha-1)^{h_-}\binom{l-n}{h_-}\binom{n}{h_+}} , \tag{27}$$

where $h_+ - h_- = n - k$ and $h_+ + h_- + h_0 = h$ with the constraints $h_- \geq 0$, $h_0 \geq 0$. The final formula (24) is obtained by the substitution of (26) and (27) into (25). □

Remark 3 The summations in (24) are shown across all $r_+ \in [0, h_+]$ and $h_+ \in [0, h]$, but it is important to check also that all other indices are non-negative: $r_- \geq 0$, $r - r_+ - r_- \geq 0$, $h_- \geq 0$ and $h_0 \geq 0$. The triangle inequalities $|n - k| \leq h \leq n + k$ imply the following bounds $\max\{0, n - k\} \leq h_+ \leq \frac{1}{2}(h + n - k) \leq h$ (and similar for r_+), which can be used for a more efficient implementation.

Example 10 (Binary case $\alpha = 2$) In the binary case there are no neutral substitutions among $h = d(x, y)$ different letters, and therefore $h_0 = 0$, $h = h_+ + h_-$ and $h_+ - h_- = n - k$ define one possible value $h_+ = \frac{1}{2}(h + n - k)$. Probability (27) becomes

$$P(h_+ \mid n, k, h) = \delta_{\frac{1}{2}(h+n-k)}(h_+) .$$

Substituting $h_+ = \frac{1}{2}(h+n-k)$ and using the conditions $r_+ \in [0, h_+], r_- = r_+ - (n-m) \geq 0, r_0 = r - 2r_+ + (n-m) \geq 0$ formula (24) reduces to

$$P(m \mid n,k,h,r) = \frac{\sum_{r_+=0}^{h_+} \binom{l-h}{r-2r_++(n-m)} \binom{h-h_+}{r_+-(n-m)} \binom{h_+}{r_+}}{\binom{l}{r}}.$$

This formula is also valid for $\alpha > 2$ when the distance $d(x,y) = h$ is minimized ($h = |n-k|$) or maximized ($h = n+k$), because there are $h_0 = 0$ possible neutral substitutions among $h = d(x,y)$ different letters in these cases.

Conditional probability (24) was implemented in a digital computer using Common Lisp programming language, and Figs. 4–6 illustrate its dependency on parameters n, k, h and r in Hamming space $\mathcal{H}_4^{40}$ ($\alpha = 4$, $l = 40$). Ordinates on all charts show the resulting distance $m = d(\top, z)$ of offspring after crossover. The grayscale represents different values of probability $P(m \mid n,k,h,r)$ with white corresponding to $P = 0$ and black to $P = 1$.

Figure 4 shows the effect of distance $k = d(\top, y)$ of the second parent (abscissae) relative to the distance $n = d(\top, x)$ of the first parent shown on three charts for $n \in \{10, 20, 30\}$. On all charts recombination radius was $r = \lfloor l/2 \rceil = 20$ and capacity $d(x,y) = h = \max\{n,k\}$. One can see that the resulting distribution appears to have linear dependency on distance $d(\top, y) = k$, and (as expected) crossover with $d(\top, y) < d(\top, x)$ increases the chance of beneficial recombination.

Figure 5 shows the effect of recombination capacity $h = d(x,y)$ (abscissae). Three charts correspond to distances $d(\top, y) = k \in \{10, 20, 30\}$ and $d(\top, x) = n = 20$. Recombination radius was $r = \lfloor l/2 \rceil = 20$ on all charts. One can see that recombination capacity increases the variance of the resulting distribution of $d(\top, z) = m$ with the maximum variance achieved for $h = n + k$.

Figure 6 shows the effect of recombination radius r (abscissae). Three charts correspond to three distances $d(\top, y) = k \in \{10, 20, 30\}$ of the second parent and distance $d(\top, x) = n = 20$. Recombination capacity was $d(x,y) = h = \max\{n,k\}$ on all charts. One can see that small recombination radius concentrates the probability at distance of the first parent, while increasing the radius concentrates the probability at distance of the second parent (as expected). One can see also the variance of the offspring's distance $d(\top, z) = m$ appears to be maximized at $r = \lfloor l/2 \rceil = 20$.

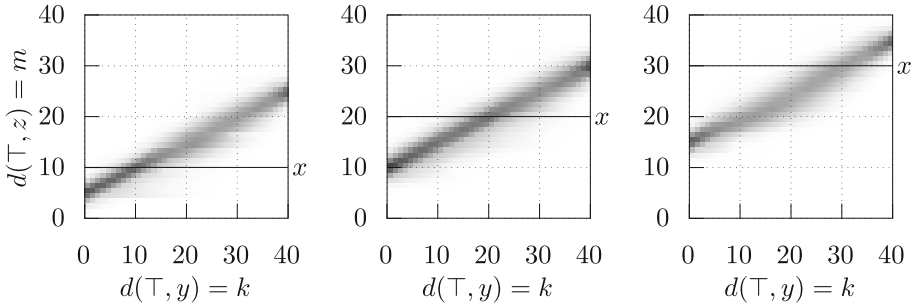


Fig. 4 Dependency of the probability $P(m \mid n,k,h,r)$ (24) on the distance $d(\top, y) = k$ of the second string (abscissae) in space $\mathcal{H}_4^{40}$. Ordinates show the resulting distance $d(\top, z) = m$ after crossover. Three charts correspond to three distances $d(\top, x) = n \in \{10, 20, 30\}$ of the first parent string. Grayscale represents probability $P \in [0, 1]$. Parameters $d(x,y) = h = \max\{n,k\}$ and $r = 20$

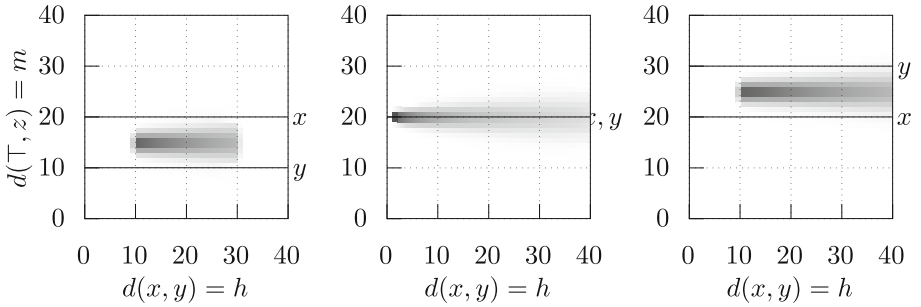


Fig. 5 Dependency of the probability $P(m \mid n, k, h, r)$ (24) on the recombination capacity $d(x, y) = h$ (abscissae) in space $\mathcal{H}_4^{40}$. Ordinates show the resulting distance $d(\top, z) = m$ after crossover. Three charts correspond to three distances $d(\top, y) = k \in \{10, 20, 30\}$. Parameters $d(\top, x) = n = 20$ and $r = 20$. Grayscale represents probability $P \in [0, 1]$

The above observations about the expected value and variance of distance $d(\top, z) = m$ after crossover are confirmed by the corresponding formulae below.

Proposition 5 *The expected value and variance of conditional probability distribution* (24) *for Hamming distance $m = d(\top, z)$ of string z obtained by a crossover recombination of $r \in [0, l]$ letters in string $x \in S(\top, n)$ from string $y \in S(\top, k) \cap S(x, h)$ in a Hamming space $\mathcal{H}_\alpha^l$ are*

$$\mathbb{E}_P\{m \mid n, k, h, r\} = n + \frac{(k - n)}{l} r \,, \tag{28}$$

$$\sigma_P^2\{m \mid n, k, h, r\} = \left[h - \langle h_0 \rangle - \frac{(n - k)^2}{l}\right] \frac{r(l - r)}{l(l - 1)} \,. \tag{29}$$

Here $\langle h_0 \rangle := \mathbb{E}\{h_0 \mid n, k, h\}$ is the expected maximum number of neutral substitutions among $h = d(x, y)$ different letters, which is computed using formula (27) *for conditional distribution $P(h_+ \mid n, k, h)$ and using the relation $h_0 = h - 2h_+ + n - k$.*

The proof uses formulae (2) and (3) from Lemma 1, where probabilities $\langle P_i \rangle$ and $\langle P_{ij} \rangle$ are defined as functions of parameters $d(\top, x) = n$, $d(\top, y) = k$, $d(x, y) = h$ and recombination radius $r \in [0, l]$. See Appendix A.5 for details.

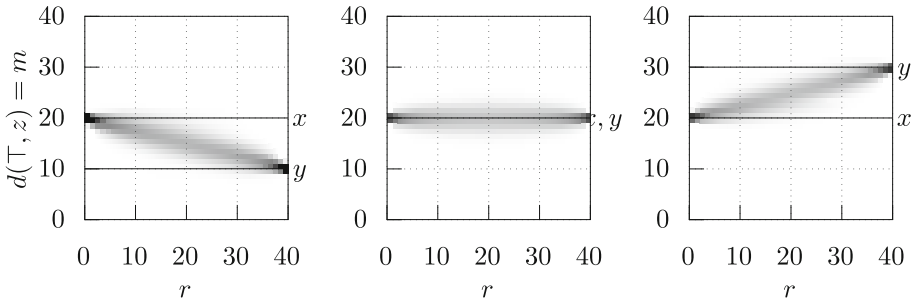


Fig. 6 Dependency of the probability $P(m \mid n, k, h, r)$ (24) on the recombination radius $r \in [0, l]$ (abscissae) in space $\mathcal{H}_4^{40}$. Ordinates show the resulting distance $d(\top, z) = m$ after crossover. Three charts are shown for three values of distance $d(\top, y) = k \in \{10, 20, 30\}$. Parameters $d(\top, x) = n = 20$ and $d(x, y) = h = \max\{n, k\}$. Grayscale represents probability $P \in [0, 1]$.

Formula (28) confirms 'linear' dependency that can be seen on Fig. 6. The slope of this dependency is defined by the difference $d(\top, y) - d(\top, x) = k - n$ of distances of the two parent strings. Similarly, the variance depends only on the squared difference $(n-k)^2$ in (29). Thus, the slope and the variance of distance distribution after crossover are invariant under translation $n \mapsto n + m$ and $k \mapsto k + m$ of distances from the optimum (since $n + m - k - m = n - k$).

Formula (29) shows also that the variance is maximized at equal distances $d(\top, x) = d(\top, y)$ (i.e. $(n-k)^2 = 0$) and when recombination capacity $d(x, y) = h$ is maximized: $h = n + k$ (in which case $\langle h_0 \rangle = 0$). These effects of distances $d(\top, x) = n, d(\top, y) = k$ and capacity $d(x, y) = h$ on the variance are shown on Fig. 7.

The dependency of variance (29) on the recombination radius $r \in [0, l]$ is particularly interesting, as this parameter is independent of the others. Maximization gives the following result:

$$\frac{\partial}{\partial r}\sigma^2\{m \mid n, k, h, r\} = \left[h - \langle h_0 \rangle - \frac{(n-k)^2}{l}\right]\frac{l - 2\hat{r}}{l(l-1)} = 0 \qquad \Longrightarrow \qquad \hat{r} = \frac{l}{2}\,,$$

(the second derivative is negative). Therefore, the variance of distances after crossover recombination is maximized when exactly half of the letters in the strings are recombined. This corresponds to the one-point crossover at index $i = \lfloor l/2 \rceil$. For uniform crossover with rate $\nu = 1/2$ recombination radius is random with the mean value $l/2$.

Finally, let us show and discuss the following symmetry property of beneficial and deleterious crossover recombinations.

Proposition 6 *Geometric probability* (24) *has the following symmetry:*

$$P(m \mid n, k, h, r) = P(n + k - m \mid n, k, h, l - r)\,.$$

See Appendix A.6 for the proof. Notice that for strings at equal distances $d(\top, x) = d(\top, y) = n$ the probability that recombinations is beneficial $m = n - (n - m) < n$ is equal to the probability that the dual recombination is deleterious $m' = n + n - m > n$. Thus, chances of beneficial and deleterious recombinations are in a certain sense equal, and this property is uniform across the entire Hamming space $\mathcal{H}_\alpha^l$. This is different from mutation, because beneficial mutations are less frequent than deleterious mutations for all strings with $d(\top, x) < l(1 - 1/\alpha)$.

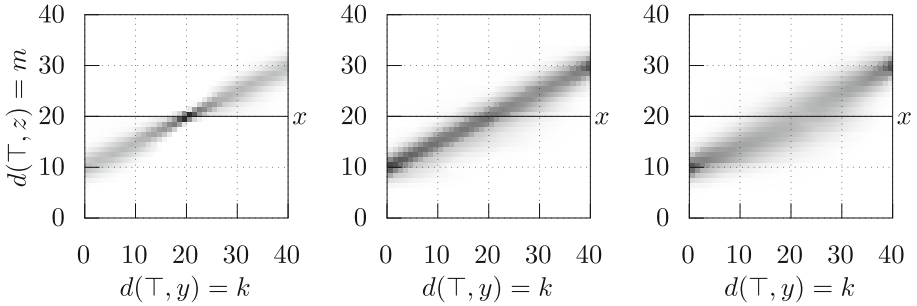


Fig. 7 Dependency of the probability $P(m \mid n, k, h, r)$ (24) in space $\mathcal{H}_4^{40}$ on the distance $d(\top, y) = k$ of the second string (abscissae) and using three strategies for recombination capacity $d(x, y) = h$: minimization $h = |n - k|$ (left), mean $h = \max\{n, k\}$ (centre), maximization $h = n + k \leq l$ (right). Ordinates show the resulting distance $d(\top, z) = m$ after crossover. Recombination radius $r = 20$. Grayscale represents probability $P \in [0, 1]$

3.4 Maximization of probability of beneficial recombination

The closed-form expression (24) combined with probabilities of the recombination radius $P(r \mid n, k, h)$ and recombination capacity $P(h \mid n, k)$ gives complete solution to transition probability (21) between pairs of spheres around optimum after crossover. This makes it possible to maximize the probability of beneficial crossover recombination. As with mutation, however, this optimization problem can be defined in many different ways (e.g. subject to a constraint on the number of generations). Below we consider two simplified problems that have exact solutions.

Proposition 7 *(Minimization of the expected distance after crossover) The expected value* (28) *of Hamming distance $d(\top, z) = m$ after crossover of $x \in S(\top, n)$ with $y \in S(\top, k) \cap S(x, h)$ into $z \in S(\top, m)$ is minimized if the recombination radius $r \in [0, l]$ has the values $r = 0$ for $n < k$, $r = l$ for $n > k$, and any value for $n = k$.*

Proof The minimization $\mathbb{E}_P\{m \mid n, k, h, r\} < n$ over the recombination radius $r \in [0, l]$ follows from (28). □

This simple result implies the following recombination rate function for the uniform crossover operator (Example 8):

$$\hat{\nu}(n, k) = \begin{cases} 0 & \text{if } n < k \\ 1/2 & \text{if } n = k \\ 1 & \text{if } n > k \end{cases}.$$

The application of such a strategy for recombination can be limited if the population has no individuals with equal distances $d(\top, x) = d(\top, y)$. Another approach is to maximize the probability of crossover recombination directly into optimum. As for mutation, this problem has exact solution.

Proposition 8 *(Recombination into the optimum) The probability $P(m = 0 \mid n, k, h, r)$ that crossover recombination of string $x \in S(\top, n) \subset \mathcal{H}_\alpha^l$ with $y \in S(\top, k) \cap S(x, h)$ results in string $z = \top$ is*

$$P(m = 0 \mid n, k, h, r) = \frac{|\{\top \in I(x, y, r)\}|}{|I(x, y, r)|} = \begin{cases} \frac{\binom{l-h}{r-n}}{\binom{l}{r}} & \text{if } h = n + k \le l, n \le r \le l - k \\ 0 & \text{otherwise} \end{cases},$$

where $|\{\top \in I(x, y, r)\}|$ is the number of copies of element $\top$ in the recombination potential $I(x, y, r)$ (recall that it is a multiset); $h = d(x, y)$ is recombination capacity, and $r \in [0, l]$ is recombination radius. The optimal recombination radius maximizing the above probability is

$$\hat{r} = \left\lfloor l \left(\frac{n}{n + k} \right) \right\rceil, \tag{30}$$

where $n = d(\top, x)$ and $k = d(\top, y)$ (here $\lfloor \cdot \rceil$ denotes the nearest integer).

Proof Crossover recombination into $z = \top$, $d(\top, z) = m = 0$, implies that the recombination radius is $r \ge n = d(\top, x)$ letters, of which there should be exactly $r_+ = n$ beneficial substitutions, and which is also their maximum number $h_+ = n$. For deleterious substitutions $r_- = 0$ and $h_- = h_+ - (n - k) = k = d(\top, y)$ by (19). There can be no neutral substitutions among $h = d(x, y)$ different letters, which implies $h_0 = 0$ and $h = h_+ + h_- = n + k$. Thus, recombination capacity $h = d(x, y)$ must be maximized (this also maximizes the variance

of distances after crossover (29)). There can be $r_0 = r - r_+ - r_- = r - n \geq 0$ neutral substitutions among $l - d(x, y) = l - h$ identical letters, which also must be identical to the corresponding letters in $\top$. Because $r_0 = r - n \leq l - h = l - n - k$, we also obtain the upper bound $r \leq l - k$. The formula for the probability $P(m = 0 \mid n, k, h, r)$ is obtained by substituting the values $r_+ = h_+ = n$, $r_- = 0$, $r_0 = r - n$, $h_- = k$, $h = h_+ + h_-$ into formula (24) and considering the constraints $h = d(x, y) = n + k \leq l$ and $n \leq r \leq l - k$. This probability is the proportion of all optimal elements $\top$ in the recombination potential $I(x, y, r)$, which is a multiset (i.e. it may contain multiple copies of $\top$).

The optimal recombination radius $r \in [n, l-k]$ maximizing the probability can be found by setting its derivative over r to zero. Using the following formula for the derivative of the binomial coefficient [35]:

$$\frac{\partial}{\partial r}\binom{l}{r} = \binom{l}{r}[H_{l-r} - H_r]\,,$$

where H_r is the rth harmonic number, and employing simple approximation $H_r \approx \ln r$ gives the following necessary optimality condition:

$$\begin{aligned}\frac{\partial}{\partial r}\frac{\binom{l-n-k}{r-n}}{\binom{l}{r}} &= \frac{\binom{l-n-k}{\hat{r}-n}}{\binom{l}{\hat{r}}}\left[H_{l-k-\hat{r}} - H_{\hat{r}-n} - H_{l-\hat{r}} + H_{\hat{r}}\right] \\ &\approx \frac{\binom{l-n-k}{\hat{r}-n}}{\binom{l}{\hat{r}}}\left[\ln\frac{(l-k-\hat{r})\hat{r}}{(\hat{r}-n)(l-\hat{r})}\right] = 0\,.\end{aligned}$$

The root $\hat{r}$ of the above equation is obtained by setting the logarithm to zero resulting in the following equations:

$$(l-k-\hat{r})\hat{r} = (\hat{r}-n)(l-\hat{r}) \quad\Longrightarrow\quad -k\hat{r} = -n(l-\hat{r}) \quad\Longrightarrow\quad \hat{r} = l\left(\frac{n}{n+k}\right).$$

One can check also that the second derivative is negative for $\hat{r}$:

$$\frac{\partial^2}{\partial r^2}\frac{\binom{l-n-k}{r-n}}{\binom{l}{r}} \approx \frac{\binom{l-n-k}{\hat{r}-n}}{\binom{l}{\hat{r}}}\left[\left(\ln\frac{(l-k-\hat{r})\hat{r}}{(\hat{r}-n)(l-\hat{r})}\right)^2 + \frac{\partial}{\partial r}\ln\frac{(l-k-\hat{r})\hat{r}}{(\hat{r}-n)(l-\hat{r})}\right].$$

The square of the logarithm on the left is zero at $\hat{r}$. The derivative of the logarithm on the right is

$$\begin{aligned}\frac{\partial}{\partial r}\ln\frac{(l-k-\hat{r})\hat{r}}{(\hat{r}-n)(l-\hat{r})} &= -\frac{1}{l-k-\hat{r}} + \frac{1}{l-\hat{r}} - \frac{1}{\hat{r}-n} + \frac{1}{\hat{r}}\Big|_{\hat{r}=l\left(\frac{n}{n+k}\right)} \\ &= (n+k)\left[\frac{1}{n} + \frac{1}{k}\right]\left[\frac{1}{l} - \frac{1}{l-(n+k)}\right] \leq 0\,,\end{aligned}$$

because $l \geq l - (n+k)$. Formula (30) is the nearest integer of $l\left(\frac{n}{n+k}\right)$. □

The procedure for maximizing the probability $P(m = 0 \mid n, k, h, r)$ of crossover recombination into the optimum $\top$ can now be outlined:

1. Let $x \in S(\top, n)$ be the first parent string (i.e. the distance $d(\top, x) = n$ is fixed).
2. Choose the set $\{y \in S(\top, k)\}$ of second parents as close as possible to the optimum $\top$, because decreasing the distance $d(\top, y) = k$ increases the numerator $\binom{l-h}{r-n}$ in the probability $P(m = 0 \mid n, k, h, r)$.

3. Choose the second parent $y \in S(\top, k)$ with the maximum recombination capacity $d(x, y) = h = n + k \le l$ (otherwise, the probability $P(m = 0 \mid n, k, h, r)$ is zero).
4. Recombine precisely $\hat{r} = \left\lfloor l \left(\frac{n}{n+k}\right) \right\rceil$ letters to maximize $P(m = 0 \mid n, k, h, r)$.

One can see from formula (30) that the optimal recombination radius increases with distance $d(\top, x) = n$ of the first string and decreases with distance $d(\top, y) = k$ of the second string. Observe also that at equal distances $n = k$ the optimal value is $\hat{r} = l/2$, which also maximizes the distance variance (29).

In the end of this section, let us compare two recombination operators — the uniform (Example 8) and the one-point crossover (Example 9). In the case of uniform crossover, the recombination radius is a binomial random variable, and taking into account the conditions $h = n + k$ and $r \in [n, l - k]$ we obtain

$$P_\nu(m = 0 \mid n, k) = \sum_{r=n}^{l-k} \binom{l-n-k}{r-n} \nu^r(n,k)[1 - \nu(n,k)]^{l-r} .$$

Optimization of the recombination rate is complicated due to the range of possible recombination radii $r \in [n, l - k]$. For simplicity, let us assume that r takes only one value $r \in [n, l - k]$. In this case, the maximum is found by differentiation:

$$\frac{\partial}{\partial \nu} P_\nu = \binom{l-n-k}{r-n} \hat{\nu}^r [1-\hat{\nu}]^{l-r} \underbrace{\left(\frac{r}{\hat{\nu}} - \frac{l-r}{1-\hat{\nu}}\right)}_{=0} = 0 \qquad \Longrightarrow \qquad \hat{\nu} = \frac{r}{l} ,$$

$$\frac{\partial^2}{\partial \nu^2} P_\nu = \binom{l-n-k}{r-n} \hat{\nu}^r [1-\hat{\nu}]^{l-r} \left[\underbrace{\left(\frac{r}{\hat{\nu}} - \frac{l-r}{1-\hat{\nu}}\right)^2}_{=0} - \frac{l - 2r/\hat{\nu} + r/\hat{\nu}^2}{(1-\hat{\nu})^2} \right]$$

$$= \binom{l-n-k}{r-n} \hat{\nu}^r [1-\hat{\nu}]^{l-r} \left[\frac{l(1-l/r)}{(1-r/l)^2} \right] \le 0 ,$$

because $1 - l/r \le 0$. Substituting the optimal recombination radius (30) results in the following recombination rate function:

$$\hat{\nu}(n, k) = \frac{n}{n+k} .$$

Thus, uniform crossover with the above recombination rate makes the mean value $\mathbb{E}\{r\} = l\nu$ of the recombination radius equal to the optimal value (30). However, its variance $\sigma^2(r) = l\nu(1-\nu)$ is generally not zero, meaning that the exact optimal value of recombination radius is not guaranteed.

In the case of one-point crossover, the distribution of recombination radius is the Dirac $\delta_{l-i}(r)$ with zero variance, where $i \in [1, \ldots, l]$ is the index of one-point crossover. Therefore, if the index of one-point crossover is set to $i = l - \hat{r} = \lfloor lk/(n+k) \rceil$, then it is feasible to guarantee the optimal value given by (30). This could be the advantage of one-point crossover over the uniform crossover. Interestingly, a process similar to one-point crossover is used in nature to exchange genetic material between pairs of homologous non-sister chromatids.

4 Discussion

We have analysed geometry and combinatorics of mutation and crossover operators in Hamming spaces. The new formula for geometric probability (24) of crossover recombination of two strings onto a sphere around an optimum now complements a similar formula (11) for mutation that was derived previously in [2–5]. Combined with the information about specific mutation and crossover operators (e.g. Examples 2, 8, 9) one can compute stochastic matrices M and R to represent Markov operators, defined by (6) and (20), transforming distance distributions $p(t) := P_t\{x \in S(\top, n)\}$ under mutation and recombination in a Hamming space $\mathcal{H}_\alpha^l$. Their product together with the diagonal $(l+1) \times (l+1)$ matrix S representing selection gives complete Markov evolution of distance distributions:

$$p(t) = (MRS)^t\, p(0)\,, \qquad p(0) = P_0\,.$$

This opens up the possibility for computer simulations and numerical optimization of long-term evolutionary dynamics under various control functions and strategies, such as the variable mutation rates $\mu(n)$ (e.g. as in (15)), recombination radii $r(n,k)$ (e.g. as in (30)) and pairing strategies (e.g. Examples 5, 6, 7). In some cases, analytic solutions are also possible, such as the optimality conditions given in Propositions 2 and 3 for mutation (previously presented in [3–5]) and in Propositions 7 and 8 for crossover recombination. It is important to note, however, that such solutions that are optimal for these specific criteria may not be optimal for other criteria (e.g. see the discussion in [36] or various optimality criteria and constraints in [4]). Simulations using the above Markov process can be used for optimization of evolutionary dynamics over multiple generations and considering other characteristics, such as the rate of convergence or the running time. The latter can be estimated as time to absorption for stochastic matrix MRS by considering the optimum $\top \in \mathcal{H}_\alpha^l$ as an absorbing state. Such a programme has already been realized in [4], but only for the mutation operator M. This work extends the range of tools suitable for a more complete study. Future work may consider potential applications to the run-time analysis of evolutionary algorithms and optimization of mutation and recombination operators in more complex fitness landscapes (i.e. when fitness is not the negative distance to optimum). Such analysis can be facilitated by the formulae derived here and considering monotonic relations between fitness and distance that can often be postulated [8].

Another interesting direction to explore is the interaction between mutation and crossover operators. Our analysis suggests that mutation and crossover recombination may have different and in some sense complementary properties. Mutation has the advantage that its range is the entire space $\{1, \dots, \alpha\}^l$ of strings. However, it lacks direction, and when strings are closer to an optimum the majority of mutations are deleterious. Maximization of the probability of beneficial mutation requires that mutation rates decrease as strings evolve closer to a fitness peak, which has the inevitable effect of slowing down the evolution. Recombination, on the other hand, acts in a subspace defined by the current population. However, unlike mutation, recombination can have a direction towards higher fitness, and it equalizes the chances of beneficial and deleterious recombinations. It has properties, such as variance of distance distribution, that are translation invariant. These observations suggest that mutation can be more important for diversity and adaptation of the population that is far away from the fitness peak. Once the population has evolved closer to the fitness peak and the mutation rate reduced, recombination may become more important to maintain the rate of adaptation. These hypotheses can be tested using simulations.

It is important to emphasize that the theory and probability formulae presented here are exact and not approximate or asymptotic. At the same time the model concerns a Markov process with only $l+1$ states (i.e. the range $\{0,\ldots,l\}$ of Hamming distance on $\mathcal{H}_\alpha^l$) and square $(l+1)\times(l+1)$ matrices. Even though l can be large in some cases, this model is more computationally tractable than other approaches, such as [37, 38] that used Markov processes with states corresponding to all possible variable size populations of strings. Furthermore, many properties can be derived from simulations with small l. In addition, the formulae for the first two moments (i.e. (13), (28) for the means and (14), (29) for the variances) can be used to infer some approximate properties. All formulae are valid for strings with arbitrary alphabet size $\alpha \in \mathbb{N}$ broadening their scope of applications to different areas including biological systems.

Although the analysis presented here is based on information about Hamming distances between strings, the conclusions can be translated into more practical or biologically relevant notions of fitness and similarity. Previously we showed that fitness is related to distance from an optimum at least in some neighbourhoods of a local optimum in a broad class of fitness landscapes [8]. Theoretical predictions about optimal mutation rates were tested in computational experiments with transcription factor binding landscapes as well as experiments *in vivo* with various microbes [10–12]. They discovered that mutation rates are strongly anti-correlated with population density, which microbes can sense and that is related to biological fitness (i.e. the replication rate). Thus, organisms may use a control strategy of mutation parameters similar to that predicted by the theory in order to increase their adaptability. This trait appears to exist across all domains of life [11]. It is reasonable to assume that similar strategies may exist for controlling parameters of recombination operators. Testing these hypotheses experimentally is an exciting prospect.

Appendix A Proofs

A.1 Proof of Lemma 1 for the mean and variance of Hamming distance

Proof Using the definition of Hamming distance (1) as a sum of elementary distances $1-\delta_{x_i y_i} \in \{0,1\}$ we have

$$\begin{aligned}
\mathbb{E}_P\{d(x,y)\} &= \mathbb{E}_P\left\{\sum_{i=1}^{l}(1-\delta_{x_i y_i})\right\} = \sum_{i=1}^{l}\mathbb{E}_P\{1-\delta_{x_i y_i}\},\\
\sigma_P^2\{d(x,y)\} &= \mathbb{E}_P\left\{\left(\sum_{i=1}^{l}(1-\delta_{x_i y_i})\right)^2\right\} - \left(\mathbb{E}_P\left\{\sum_{i=1}^{l}(1-\delta_{x_i y_i})\right\}\right)^2\\
&= \sum_{i=1}^{l}\mathbb{E}_P\{(1-\delta_{x_i y_i})^2\} + \sum_{i=1}^{l}\sum_{\substack{j=1\\ j\neq i}}^{l}\mathbb{E}_P\{(1-\delta_{x_i y_i})(1-\delta_{x_j y_j})\} -\\
&\quad - \left(\sum_{i=1}^{l}\mathbb{E}_P\{1-\delta_{x_i y_i}\}\right)^2,
\end{aligned}$$

where we used additivity of the expected value and the square of the sum formula: $\left(\sum_i a_i\right)^2 = \sum_i a_i^2 + 2\sum_{i<j} a_i a_j = \sum_i a_i^2 + \sum_i \sum_{j\neq i} a_i a_j$. Noticing that $(1-\delta_{x_i y_i})^2 = 1-\delta_{x_i y_i}$ and

denoting the average values of the expectations under summations by $\langle P_i \rangle$ and $\langle P_{ij} \rangle$ (see their definitions in Lemma 1) we obtain the following formulae:

$$\mathbb{E}_P\{d(x,y)\} = l\langle P_i \rangle\,,$$
$$\sigma_P^2\{d(x,y)\} = l\langle P_i \rangle + l(l-1)\langle P_{ij} \rangle - (l\langle P_i \rangle)^2\,.$$

□

The symbol $\langle P_i \rangle$ introduced above is the average (over all positions $i \in \{1,\dots,l\}$) of the probability that $x_i \neq y_i$. Indeed, the differences $1-\delta_{x_i y_i}$ are non-zero (and equal to one) only when $x_i \neq y_i$, and therefore the expectations $\mathbb{E}_P\{1-\delta_{x_i y_i}\}$ are the probabilities that letters $x_i \neq y_i$ at positions $i \in \{1,\dots,l\}$ (and these probabilities can be different for different i).

Similarly, $\langle P_{ij} \rangle$ is the average of joint probability that $x_i \neq y_i$ and $x_j \neq y_j$ at two different positions $i \neq j$ (there are $l(l-1)$ off-diagonal elements). Indeed, the products $(1-\delta_{x_i y_i})(1-\delta_{x_j y_j})$ are non-zero only when both $x_i \neq y_i$ and $x_j \neq y_j$, and the expectations $\mathbb{E}_P\{(1-\delta_{x_i y_i})(1-\delta_{x_j y_j})\}$ are the corresponding joint probabilities.

A.2 Proof of Lemma 2 for intersection of spheres in $\mathcal{H}_\alpha^l$

Proof A substitution of letter x_i at position $i \in \{1,\dots,l\}$ may result in one of three possibilities:

- Letter $x_i \neq \top_i$ is substituted to $y_i = \top_i$ resulting in a beneficial substitution. Such substitutions can only occur in $n = d(\top, x)$ letters $x_i \neq \top_i$. Denoting by r_+ the total number of beneficial substitutions gives $\binom{n}{r_+}$ combinations.
- Letter $x_i = \top_i$ is substituted to any of $\alpha - 1$ letters $y_i \neq \top_i$ resulting in a deleterious substitution. Such substitutions can only occur in $l - d(\top, x) = l - n$ letters $x_i = \top_i$. Denoting by r_- the total number of deleterious substitutions gives $(\alpha-1)^{r_-}\binom{l-n}{r_-}$ possibilities.
- Letter $x_i \neq \top_i$ is substituted to one of the remaining $\alpha - 2$ letters $y_i \neq \top_i$ resulting in a neutral substitution. Such substitutions can only occur in $n = d(\top, x)$ letters $x_i \neq \top_i$ minus the number r_+ of beneficial substitutions. Denoting by r_0 the total number of neutral substitutions gives $(\alpha-2)^{r_0}\binom{n-r_+}{r_0}$ possibilities.

For specific values of r_+, r_- and r_0, the total number of combinations is

$$(\alpha-2)^{r_0}\binom{n-r_+}{r_0}(\alpha-1)^{r_-}\binom{l-n}{r_-}\binom{n}{r_+}\,.$$

By (4) and (5), the values of r_- and r_0 are related to $r_+ \in [0, r]$:

$$r_- = r_+ - (n-m)\,,$$
$$r_0 = r - (r_+ + r_-) = r - 2r_+ + (n-m)\,.$$

Formula (10) is obtained by summing over all feasible values of r_+ subject to constraints $r_- \geq 0$ and $r_0 \geq 0$. □

A.3 Proof of Proposition 1

Proof Here we use formulae (2) and (3) with the following average probabilities:

$$\langle P_i \rangle := \mathbb{P}\{y_i \neq \top_i \mid x \in S(\top, n),\, y \in S(x, r)\}\,,$$
$$\langle P_{ij} \rangle := \mathbb{P}\{y_i \neq \top_i \wedge y_j \neq \top_j \mid i \neq j, x \in S(\top, n),\, y \in S(x, r)\}\,,$$

where the conditions are defined by the fact that string $y \in S(x, r) \subset \mathcal{H}_\alpha^l$ is obtained from $x \in S(\top, n)$ by a substitution of $d(x, y) = r$ letters. Indices i and $j \in \{1, \ldots, l\}$ are positions of letters in the strings $x = (x_1, \ldots, x_l)$, $y = (y_1, \ldots, y_l)$ and $\top = (\top_1, \ldots, \top_l)$. For each position $i \in \{1, \ldots, l\}$ there are three mutually exclusive possibilities for event $y_i \neq \top_i$:

- Letter $x_i \neq \top_i$ is not substituted, so that $y_i = x_i \neq \top_i$. The number of letters $x_i \neq \top_i$ is $n = d(\top, x)$, and there are $l - r$ letters that remain not substituted in y, which means that the corresponding probability is $(n/l)(1 - r/l)$.
- Letter $x_i \neq \top_i$ is substituted by $y_i \neq \top_i$, $y_i \neq x_i$. There are $\alpha - 1$ letters in the alphabet not equal to $\top_i$, and there are $\alpha - 2$ remaining possibilities for $y_i \neq \top_i$. Given that there are $n = d(\top, x)$ letters $x_i \neq \top_i$ and r letters are substituted, the corresponding probability is $(n/l)[(\alpha - 2)/(\alpha - 1)](r/l)$.
- Letter $x_i = \top_i$ is substituted by any other letter $y_i \neq \top_i$. The number of letters $x_i = \top_i$ is $l - n = l - d(\top, x)$, and there are r letters that are substituted, which corresponds to the probability $(1 - n/l)(r/l)$.

Adding the above probabilities for disjoint events gives the desired average probability:

$$\langle P_i \rangle = \frac{n}{l}\left(1 - \frac{r}{l}\right) + \frac{n}{l}\left(\frac{\alpha - 2}{\alpha - 1}\right)\frac{r}{l} + \left(1 - \frac{n}{l}\right)\frac{r}{l}\,.$$

Because $\langle P_i \rangle$ is the average probability across all positions $i \in \{1, \ldots, l\}$, its value is the same for all i, and the expected value $\mathbb{E}_P\{m \mid n, r\}$ can be computed as $l\langle P_i \rangle$. Its expression can be simplified as follows:

$$l\langle P_i \rangle = n\left(1 - \frac{1}{\alpha - 1}\frac{r}{l}\right) + (l - n)\frac{r}{l}\,.$$

The expression above can also be transformed into formula (13).

The average joint probability $\langle P_{ij} \rangle := \mathbb{P}\{y_i \neq \top_i \wedge y_j \neq \top_j \mid i \neq j, n, r\}$ is factorized into the product $\mathbb{P}\{y_i \neq \top_i \mid n, r\}\mathbb{P}\{y_j \neq \top_j \mid y_i \neq \top_i, i \neq j, n, r\}$, where the first probability for the first event $z_i \neq \top_i$ is the average probability $\langle P_i \rangle$ derived above. The second is the average conditional probability of the second event $z_j \neq \top_j$ for $j \neq i$ (and conditioned on the first event $z_i \neq \top_i$). This conditional probability can be determined from the following considerations.

For each of the three possibilities for $y_i \neq \top_i$ in the string of length l, there are three possibilities for $y_j \neq \top_j$ in the remaining string of length $l - 1$. Thus, there are $3 \times 3 = 9$ joint events, the probabilities of which are defined similarly, but using numbers $n - 1$ or n and $r - 1$ or r depending on the type and position of the first event $y_i \neq \top_i$. Thus, the product

$l(l-1)\langle P_{ij}\rangle$ is as follows:

$$\begin{aligned} l(l-1)\langle P_{ij}\rangle = n\left(1-\frac{r}{l}\right)\left[(n-1)\left(1-\frac{1}{\alpha-1}\frac{r}{l-1}\right)+(l-n)\frac{r}{l-1}\right] \\ +n\left(\frac{\alpha-2}{\alpha-1}\right)\frac{r}{l}\left[(n-1)\left(1-\frac{1}{\alpha-1}\frac{r-1}{l-1}\right)+(l-n)\frac{r-1}{l-1}\right] \\ +(l-n)\frac{r}{l}\left[n\left(1-\frac{1}{\alpha-1}\frac{r-1}{l-1}\right)+(l-1-n)\frac{r-1}{l-1}\right]. \end{aligned}$$

Formula (14) for the variance is obtained by substituting the expressions for $l\langle P_i\rangle$ and $l(l-1)\langle P_{ij}\rangle$ into equation $l\langle P_i\rangle + l(l-1)\langle P_{ij}\rangle - (l\langle P_i\rangle)^2$. □

A.4 Proof of Proposition 4

Proof Let us denote by $r' = l - r$ the recombination radius of the dual recombination z', and let r'_+, r'_- and r'_0 be respectively the numbers of beneficial, deleterious and neutral substitutions into x. Because the dual recombination is a substitution of the remaining $l - r$ letters, we have

$$r_+ + r'_+ = h_+\,, \qquad r_- + r'_- = h_-\,.$$

Using the above equations and (17), (19) we have

$$\begin{aligned} r'_+ - r'_- &= h_+ - h_- - (r_+ - r_-) \\ &= (n-k) - (n-m)\,. \end{aligned}$$

On the other hand, by analogy with (17), we have

$$r'_+ - r'_- = n - m'\,. \tag{31}$$

Therefore, $n - k - (n - m) = m - k = n - m'$. □

A.5 Proof of Proposition 5

Proof Here we use formulae (2) and (3) with the following average probabilities

$$\begin{aligned} \langle P_i\rangle &:= \mathbb{P}\{z_i \neq \top_i \mid x \in S(\top,n),\, y \in S(\top,k),\, d(x,y) = h, r\}\,, \\ \langle P_{ij}\rangle &:= \mathbb{P}\{z_i \neq \top_i \wedge z_j \neq \top_j \mid i \neq j, x \in S(\top,n),\, y \in S(\top,k),\, d(x,y) = h, r\}\,, \end{aligned}$$

where the conditions are defined by the fact that string $z \in S(\top,m) \subset \mathcal{H}^l_\alpha$ is obtained from $x \in S(\top,n)$ by a substitution of $r \in [0,l]$ letters from string $y \in S(\top,k) \cap S(x,h)$. Indices i and $j \in \{1,\ldots,l\}$ are positions of letters in the strings. For each position $i \in \{1,\ldots,l\}$ there are two mutually exclusive possibilities for $z_i \neq \top_i$:

- Letter $x_i \neq \top_i$ is not substituted by y_i, so that in the offspring $z_i = x_i \neq \top_i$. The number of letters $x_i \neq \top_i$ is $n = d(\top, x)$, and there are $l - r$ letters that remain not substituted in z, which means that the corresponding probability is $(n/l)(1 - r/l)$.
- Letter $x_i = \top_i$ is substituted by letter $y_i \neq \top_i$, so that in the offspring $z_i = y_i \neq \top_i$. The number of letters $y_i \neq \top_i$ is $k = (\top, y)$, and there are r letters that are substituted, which corresponds to probability $(k/l)(r/l)$.

Adding the above probabilities for disjoint events gives the desired average probability:

$$\langle P_i \rangle = \frac{n}{l}\left(1 - \frac{r}{l}\right) + \frac{k}{l}\frac{r}{l}\,.$$

Because $\langle P_i \rangle$ is the average probability across all positions $i \in \{1, \ldots, l\}$, its value is the same for all i, and the expected value $\mathbb{E}_P\{m \mid n, k, h, r\}$ can be computed as $l\langle P_i \rangle$:

$$l\langle P_i \rangle = n\left(1 - \frac{r}{l}\right) + k\frac{r}{l}\,.$$

The above expression gives formula (28).

The average joint probability $\langle P_{ij} \rangle := \mathbb{P}\{z_i \neq \top_i \wedge z_j \neq \top_j \mid i \neq j, n, k, h, r\}$ is factorized into the product $\mathbb{P}\{z_i \neq \top_i \mid n, k, h, r\}\mathbb{P}\{z_j \neq \top_j \mid z_i \neq \top_i, i \neq j, n, k, h, r\}$, where the first probability for the first event $z_i \neq \top_i$ is the average probability $\langle P_i \rangle$ derived above. The second is the average conditional probability of the second event $z_j \neq \top_j$ for $j \neq i$ (and conditioned on the first event $z_i \neq \top_i$). This conditional probability can be determined from the following considerations.

Let us decompose each of the two cases of the first event $z_i \neq \top_i$ into three subcases resulting in $2 \times 3 = 6$ mutually exclusive subcases of event $z_i \neq \top_i$. These subcases were not considered for the probability $\langle P_i \rangle$, because it concerns only one index i. When two indices i and $j \neq i$ are considered for the joint probability $\langle P_{ij} \rangle$, these subcases are important, because they influence the numbers that are required for the probability of the second event $z_j \neq \top_j$.

First, let us consider when any of n letters $x_i \neq \top_i$ are not substituted by y_i. There are three subcases for such non-substitutions. They can be among

1. h_+ letters $y_i = \top_i$ (i.e. some of h_+ possible beneficial substitutions do not occur).
2. h_0 letters $y_i \neq \top_i$, $y_i \neq x_i$ (i.e. some of h_0 neutral substitutions do not occur).
3. $n - h_0 - h_+$ identical letters $y_i = x_i \neq \top_i$.

If a non-substitution occurs at position i, then the number n reduces to $n - 1$, but the recombination radius r remains the same. The number k of letters $y_i \neq \top_i$ remains the same in the first subcase (because $x_i \neq \top_i$ was not substituted by one of $l - k$ letters $y_i = \top_i$), but it reduces to $k - 1$ in the last two subcases. The length l of the remaining string is $l - 1$.

Second, let us consider when letters x_i are substituted by any of k letters $y_i \neq \top_i$. Again, there are three subcases for such substitutions. They can be among

1. h_- letters $y_i \neq \top_i$ (i.e. some of h_- possible deleterious substitutions occur).
2. h_0 letters $y_i \neq \top_i$, $y_i \neq x_i$ (i.e. some of h_0 neutral substitutions occur).
3. $k - h_0 - h_-$ identical letters $y_i = x_i \neq \top_i$.

Note that $k - h_0 - h_- = n - h_0 - h_+$, which follows from $h_+ - h_- = n - k$. If a substitution by letter $y_i \neq \top_i$ occurs at position i, then the number k reduces to $k - 1$, and the recombination radius r reduces to $r - 1$. The number n of letters $x_i \neq \top_i$ remains the same in the first subcase (because one of $l - n$ letters $x_i = \top_i$ was substituted by $y_i \neq \top_i$), but it reduces to $n - 1$ in the last two subcases. The length l of the remaining string is $l - 1$.

For each of the six subcases, there are two possibilities for the second event $z_j \neq \top_j$ in the remaining string of length $l - 1$, so that there are $6 \times 2 = 12$ joint events. Note that the values h_+, h_0 and h_- are generally random and related by (18) and (19). Thus, initially we derive the formula for the product $l(l-1)\langle P_{ij} \rangle$ assuming that specific values of h_+, h_0 and

h_- have been fixed:

$$\begin{aligned} l(l-1)\langle P_{ij}\rangle = \left(1-\frac{r}{l}\right)\Bigg\{ & h_+\left[(n-1)\left(1-\frac{r}{l-1}\right)+k\frac{r}{l-1}\right] \\ & +h_0\left[(n-1)\left(1-\frac{r}{l-1}\right)+(k-1)\frac{r}{l-1}\right] \\ & +(n-h_0-h_+)\left[(n-1)\left(1-\frac{r}{l-1}\right)+(k-1)\frac{r}{l-1}\right]\Bigg\} \\ & +\frac{r}{l}\Bigg\{h_-\left[n\left(1-\frac{r-1}{l-1}\right)+(k-1)\frac{r-1}{l-1}\right] \\ & +h_0\left[(n-1)\left(1-\frac{r-1}{l-1}\right)+(k-1)\frac{r-1}{l-1}\right] \\ & +(k-h_0-h_-)\left[(n-1)\left(1-\frac{r-1}{l-1}\right)+(k-1)\frac{r-1}{l-1}\right]\Bigg\} \end{aligned}$$

The formula above contains six lines corresponding to six subcases of the first event $z_i \neq \top_i$: three non-substitutions of $x_i \neq \top_i$ and three substitutions of $x_i = \top_i$. Expressions in square brackets on each line correspond to two cases of the second event $z_j \neq \top_j$, $j \neq i$: non-substitutions of $x_j \neq \top_j$ and substitutions of $x_j = \top_j$. Factoring and noticing that $h_+ + h_0 + n - h_0 - h_+ = n$ and $h_- + h_0 + k - h_0 - h_- = k$ we obtain:

$$\begin{aligned} l(l-1)\langle P_{ij}\rangle = & \left(1-\frac{r}{l}\right)\left\{n\left[(n-1)\left(1-\frac{r}{l-1}\right)+(k-1)\frac{r}{l-1}\right]+h_+\frac{r}{l-1}\right\} \\ & +\frac{r}{l}\left\{k\left[(n-1)\left(1-\frac{r-1}{l-1}\right)+(k-1)\frac{r-1}{l-1}\right]+h_-\left(1-\frac{r-1}{l-1}\right)\right\} \end{aligned}$$

The right-hand-side of the above equation can be written as

$$\begin{aligned} & \left(1-\frac{r}{l}\right)\left\{n\left[(n-1)\left(1-\frac{r-1}{l-1}\right)+(k-1)\frac{r-1}{l-1}\right]+n\left(\frac{k-1}{l-1}-\frac{n-1}{l-1}\right)+h_+\frac{r}{l-1}\right\} \\ & \quad +\frac{r}{l}\left\{k\left[(n-1)\left(1-\frac{r-1}{l-1}\right)+(k-1)\frac{r-1}{l-1}\right]+h_-\left(1-\frac{r-1}{l-1}\right)\right\}, \end{aligned}$$

which allows us to factor the members as follows:

$$\begin{aligned} & \left[\left(1-\frac{r}{l}\right)n+\frac{r}{l}k\right]\left[\left(1-\frac{r-1}{l-1}\right)(n-1)+\frac{r-1}{l-1}(k-1)\right]+ \\ & \qquad +\left(1-\frac{r}{l}\right)\left(\frac{n}{l-1}(k-n)+h_+\frac{r}{l-1}\right)+\frac{r}{l}h_-\left(1-\frac{r-1}{l-1}\right). \end{aligned}$$

Substituting $h_- = h_+ + (k-n)$ the expression for $l(l-1)\langle P_{ij}\rangle$ becomes

$$\begin{aligned} & \left[\left(1-\frac{r}{l}\right)n+\frac{r}{l}k\right]\left[\left(1-\frac{r-1}{l-1}\right)(n-1)+\frac{r-1}{l-1}(k-1)\right]+ \\ & \qquad +\left(1-\frac{r}{l}\right)\frac{(n+r)(k-n)+2rh_+}{l-1}. \end{aligned}$$

It can now be combined with the expression for $l\langle P_i\rangle$ in equation $l\langle P_i\rangle + l(l-1)\langle P_{ij}\rangle - (l\langle P_i\rangle)^2$ to derive the variance formula with fixed h_+:

$$\begin{aligned}\sigma_P^2\{m \mid n,k,h,r,h_+\} &= \frac{l2h_+ - k^2 + lk + 2nk - ln - n^2}{l^2(l-1)} r(l-r) \\ &= \frac{l[2h_+ - (n-k)] - k^2 + 2nk - n^2}{l^2(l-1)} r(l-r) \\ &= \left[2h_+ - (n-k) - \frac{(n-k)^2}{l}\right] \frac{r(l-r)}{l(l-1)} \\ &= \left[h - h_0 - \frac{(n-k)^2}{l}\right] \frac{r(l-r)}{l(l-1)}\,.\end{aligned}$$

Formula (29) for the variance is obtained by averaging over all possible values of h_+ or $h_0 = h - 2h_+ + n - k$, which means that they are replaced by their expected values $\langle h_+\rangle$ or $\langle h_0\rangle := \mathbb{E}\{h_0 \mid n,k,h\}$ with respect to $P(h_+ \mid n,k,h)$ (27). □

A.6 Proof of Proposition 6

Proof Looking at factorization (25) of probability $P(m \mid n,k,h,r)$, one can see that probability $P(h_+ \mid n,k,h)$ does not influence the result, because it does not include variables m and r. Thus, we only need to consider probability $P(m \mid n,k,h,r,h_+)$ given by (26). Using symmetry of binomial coefficients $\binom{l}{r} = \binom{l}{l-r}$, one can see that

$$\binom{l-h_+-h_-}{r-r_+-r_-}\binom{h_-}{r_-}\binom{h_+}{r_+} = \binom{l-h_+-h_-}{r'-r'_+-r'_-}\binom{h_-}{r'_-}\binom{h_+}{r'_+},$$

where $r' = l - r$, $r'_+ = h_+ - r_+$ and $r'_- = h_- - r_-$. Note that $r' = l - r$ is recombination radius of the dual recombination $z' = (1-\lambda)y \oplus \lambda x$, and $r' = r'_+ + r'_- + r'_0$. Therefore, $P(m \mid n,k,h,r,h_+) = P(m' \mid n,k,h,l-r,h_+)$, where $m' = n + k - m$ by Proposition 4. □

Acknowledgements The author deeply acknowledges Alastair Channon, John Aston, Christopher Knight, Rok Krašovec, Elizabeth Aston and Danna Gifford for their contributions to previous joint work and collaborations on evolutionary dynamics of mutation and discussions of the new results on crossover recombination. Anton Eremeev is acknowledged for discussing early drafts of this work in 2016. Professors Satoru Miyazaki is deeply acknowledged for inviting and hosting the author at the Bioinformatics seminars in Tokyo University of Science between 2012–2024, where results of this work were presented and discussed on multiple occasions. The author is grateful to Professor Keiko Sato for additional discussions of mathematical results. The author deeply acknowledges very useful comments of three anonymous reviewers.

Author Contributions Roman Belavkin developed and derived the mathematical results and wrote the main manuscript text.

Funding No datasets were generated or analysed during the current study.

Declarations

Competing Interests The authors declare no competing interests

References

1. Fisher, R.A.: The genetical theory of natural selection. Oxford University Press, Oxford (1930)
2. Belavkin, R.V., Channon, A., Aston, E., Aston, J., Knight, C.G.: Theory and practice of optimal mutation rate control in Hamming spaces of DNA sequences. In: Lenaerts, T., Giacobini, M., Bersini, H., Bourgine, P., Dorigo, M., Doursat, R. (eds.) Advances in artificial life, ECAL 2011: Proceedings of the 11th European conference on the synthesis and simulation of living systems, pp 85–92. MIT Press, Cambridge, MA, USA (2011)
3. Belavkin, R.V.: Mutation and optimal search of sequences in nested Hamming spaces. In: 2011 IEEE information theory workshop, pp 90–94 (2011)
4. Belavkin, R.V.: Dynamics of information and optimal control of mutation in evolutionary systems. In: Sorokin, A., Murphey, R., Thai, M.T., Pardalos, P.M. (eds.) Dynamics of information systems: Mathematical foundations. Springer Proceedings in Mathematics and Statistics, vol. 20, pp. 3–21. Springer, Switzerland (2012)
5. Belavkin, R.V.: Minimum of information distance criterion for optimal control of mutation rate in evolutionary systems. In: Accardi, L., Freudenberg, W., Ohya, M. (eds.) Quantum bio-informatics v. QP-PQ: Quantum probability and white noise analysis, vol. 30, pp 95–115. World Scientific, Singapore (2013)
6. Jones, T., Forrest, S.: Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In: Eshelman, L. (ed.) Proceedings of the 6th international conference on genetic algorithms, San Francisco, CA, pp 184–192 (1995)
7. Poli, R., Galvan-Lopez, E.: The effects of constant and bit-wise neutrality on problem hardness, fitness distance correlation and phenotypic mutation rates. IEEE Trans. Evol. Comput. **16**(2), 279–300 (2012)
8. Belavkin, R.V., Channon, A., Aston, E., Aston, J., Krašovec, R., Knight, C.G.: Monotonicity of fitness landscapes and mutation rate control. J. Math. Biol. **73**(6), 1491–1524 (2016)
9. Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.F., Coburn, D., Newburger, D.E., Morris, Q., Hughes, T.R., Bulyk, M.L.: Diversity and complexity in DNA recognition by transcription factors. Sci. **324**(5935), 1720–3 (2009)
10. Krašovec, R., Belavkin, R.V., Aston, J.A.D., Channon, A., Aston, E., Rash, B.M., Kadirvel, M., Forbes, S., Knight, C.G.: Mutation rate plasticity in rifampicin resistance depends on escherichia coli cell-cell interactions. Nat. Commun. **5**(3742) (2014)
11. Krašovec, R., Richards, H., Gifford, D.R., Hatcher, C., Faulkner, K.J., Belavkin, R.V., Channon, A., Aston, E., McBain, A.J., Knight, C.G.: Spontaneous mutation rate is a plastic trait associated with population density across domains of life. PLoS Biol. **15**(8) (2017)
12. Krašovec, R., Richards, H., Gifford, D.R., Belavkin, R.V., Channon, A., Aston, E., McBain, A.J., Knight, C.G.: Opposing effects of final population density and stress on escherichia coli mutation rate. ISME J. **12**, 2981–2987 (2018)
13. Yang, X.-S.: Nature-inspired metaheuristic algorithms. Luniver Press, Frome, UK (2010)
14. Bäck, T.: Optimal mutation rates in genetic search. In: Forrest, S (ed.) Proceedings of the 5th international conference on genetic algorithms, pp 2–8. Morgan Kaufmann, San Francisco, CA (1993)
15. Ochoa, G.: Setting the mutation rate: Scope and limitations of the 1// heuristics. In: Proceedings of genetic and evolutionary computation conference (GECCO-2002), pp 315–322. Morgan Kaufmann, San Francisco, CA (2002)
16. Fogarty, T.C.: Varying the probability of mutation in the genetic algorithm. In: Schaffer, J.D (ed.) Proceedings of the 3rd international conference on genetic algorithms, pp 104–109. Morgan Kaufmann, San Francisco, CA (1989)
17. Yanagiya, M.: A simple mutation-dependent genetic algorithm. In: Forrest, S (ed.) Proceedings of the 5th international conference on genetic algorithms, p. 659. Morgan Kaufmann, San Francisco, CA (1993)
18. Srinivas, M., Patnaik, L.M.: Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Trans. Syst. Man Cybern. **24**(4), 656–667 (1994)

19. Braga, A.D.P., Aleksander, I.: Determining overlap of classes in the n-dimensional Boolean space. In: Neural networks, 1994. IEEE world congress on computational intelligence., 1994 ieee international conference on, vol. 7, pp 8–13 (1994)
20. Eiben, A.E., Hinterding, R., Michalewicz, Z.: Parameter control in evolutionary algorithms. IEEE Trans. Evol. Comput. **3**(2), 124–141 (1999)
21. Falco, I.D., Cioppa, A.D., Tarantino, E.: Mutation-based genetic algorithm: performance evaluation. Appl. Soft Comput. **1**(4), 285–299 (2002)
22. Doerr, B., Johannsen, D., Schmidt, M.: Runtime analysis of the $(1 + 1)$ evolutionary algorithm on strings over finite alphabets. In: FOGA'11: Proceedings of the 11th workshop proceedings on foundations of genetic algorithms, pp 119–126 (2011)
23. Doerr, B., Pohl, S.: Run-time analysis of the $(1+1)$ evolutionary algorithm optimizing linear functions over a finite alphabet. In: GECCO'12: Proceedings of the 14th annual conference on genetic and evolutionary computation, pp 1317–1324 (2012)
24. Doerr, B., Doerr, C., Kötzing, T.: Static and self-adjusting mutation strengths for multi-valued decision variables. Algorithmica **80**, 1732–1768 (2018)
25. Kötzing, T., Sudholt, D., Theile, M.: How crossover helps in pseudo-boolean optimization. In: GECCO '11: Proceedings of the 13th annual conference on genetic and evolutionary computationJuly 2011, pp 989–996 (2011)
26. Dang, D.-C., Friedrich, T., Kötzing, T., Krejca, M.S., Lehre, P.K., Oliveto, P.S., Sudholt, D., Sutton, A.M.: Escaping local optima using crossover with emergent diversity. IEEE Trans. Evol. Comput. **22**(3), 484–497 (2018)
27. Moraglio, A., Kim, Y.-H., Yoon, Y., Moon, B.-R.: Geometric crossovers for multiway graph partitioning. Evol. Comput. **15**(4), 445–474 (2007)
28. Eremeev, A.V.: Modeling and analysis of genetic algorithm with tournament selection. In: Fonlupt, C., Hao, J.K., Lutton, E., Schoenauer, M., Ronald, E (eds.) Artificial evolution: 4th European conference, AE'99, Dunkerque, France, November 3-5, 1999. Selected Papers. Lecture notes in computer science, vol. 1829, pp 84–95. Springer, Berlin, Heidelberg (2000)
29. Eremeev, A.V.: On complexity of optimal recombination for binary representations of solutions. Evol. Comput. **16**(1), 127–147 (2008)
30. Eremeev, A.V.: On Complexity of the Optimal Recombination for the Travelling Salesman Problem. In: Merz, P., Hao, J.K (eds.) Evolutionary computation in combinatorial optimization: 11th european conference, EvoCOP 2011, Torino, Italy, April 27-29, 2011. Proceedings, pp 215–225. Springer, Berlin, Heidelberg (2011)
31. Eremeev, A.V., Kovalenko, J.V.: Optimal recombination in genetic algorithms for combinatorial optimization problems: Part I. Yugoslav J. Oper. Res. **24**(1), 1–20 (2014)
32. Eremeev, A.V., Kovalenko, J.V.: Optimal recombination in genetic algorithms for combinatorial optimization problems: Part II. Yugoslav J. Oper. Res. **24**(2), 165–186 (2014)
33. Ahlswede, R., Katona, G.O.H.: Contributions to the geometry of Hamming spaces. Discret. Math. **17**(1), 1–22 (1977)
34. Neumann, J., Morgenstern, O.: Theory of games and economic behavior. Princeton University Press, Princeton, NJ (1944)
35. Binomial.: Wolfram Research, Inc. http://functions.wolfram.com/06.03.20.0003.01 (2001)
36. Buskulic, N., Doerr, C.: Maximizing drift is not optimal for solving onemax. In: nez, M.L.I (ed.) GECCO'19: Proceedings of the genetic and evolutionary computation conference companion, pp 425–426 (2019)
37. Nix, A.E., Vose, M.D.: Modeling genetic algorithms with Markov chains. Ann. Math. Artif. Intell. **5**(1), 77–88 (1992)
38. Vafaee, F., Turán, G., Nelson, P.C.: Optimizing genetic operator rates using a Markov chain model of genetic algorithms. In: Pelikan, M., Branke, J (eds.) Proceedings of genetic evolutionary computation conference (GECCO-2010), pp 721–728. Association for Computing Machinery, New York, NY, USA (2010)