Estimates from Evolutionary Algorithms Theory Applied to Gene Design

Anton Eremeev[1,2], Alexander Spirov[2,3]

December 24, 2018

[1] Omsk Department of Sobolev Institute of Mathematics SB RAS, Omsk, Russia.

[2] The Institute of Scientific Information for Social Sciences RAS, Moscow, Russia.

[3] I.M. Sechenov Institute of Evolutionary Physiology and Biochemistry RAS, St. Petersburg, Russia.
E-mails: eremeev@ofim.oscsbras.ru, sspirov@yandex.ru

Abstract

The field of evolutionary algorithms (EAs) emerged in the area of computer science due to transfer of ideas from biology and developed independently for several decades, enriched with techniques from probability theory, complexity theory and optimization methods. Our aim is to consider how some recent results form EA theory may be transferred back into biology. It has been noted that the EAs optimizing Royal Road fitness functions may be considered as models of evolutionary search for the gene promoter sequences Yfrom scratch¥. Here we consider the design of synthetic promoters from the EAs methodology viewpoint. This problem asks for a tight cluster of supposedly unknown motifs from the initial random (or partially random) set of DNA sequences using SELEX approaches. We apply the upper bounds on the expected hitting time of a target area of genotypic space, the EA runtime, in order to upper-bound the expected time to finding a sufficiently fit series of motifs (e.g. binding sites for transcription factors) in a SELEX procedure. On the other hand, using the EA theory we propose an upper bound on expected proportion of the DNA sequences with sufficiently high fitness at a given iteration of a SELEX procedure. Both approaches are evaluated in computational experiment, using a Royal Road fitness function as a model of the SELEX procedure for regulatory FIS factor binding site. Our results suggest that some theoretically provable bounds for EA performance may be used, at least in principle, for a-priory estimation of efficiency of SELEX-based approaches.

Keywords. Runtime analysis, SELEX procedure, Royal Road function

1 Introduction

The field of evolutionary algorithms (EAs) emerged in the area of computer science as a transfer of ideas from biology and developed independently for several decades, enriched with techniques from probability theory, complexity theory and optimization methods. Our aim is to consider how some recent results in theory of EAs may be transferred back into biology.

SELEX (Systematic Evolution of Ligands by EXponential enrichment) procedures are known as a valuable tool in identifying DNA and RNA sequences with high affinity for a pre-specified target proteins/molecules. However, SELEX procedure is time-consuming and costly. Therefore, in parallel with the practical *in vitro* procedures for selection and evaluation of DNA and RNA sequences, *in silico* approaches have been developed to predict the high-affinity sequences, allowing to reduce time and cost (see e.g. [4]). Most of the *in silico* studies focus on selecting the configurations with the highest probability to be found *in vitro*. In contrast to this, our analysis aims at the prediction of *efficiency* of the SELEX procedures, if the contents of individual enhancer are already known or expected to some degree.

Here we proceed from the consideration that the biotechnological approaches like SELEX, could be treated as the experimental implementations of EA. Experimenters cyclically evaluate, mutate and apply selection to the populations of nucleic acid molecules to breed the desired sequence (particularly, the promoter sequence). We noticed that typical organization of the gene-regulatory element can be treated as the molecular implementation of the well-known Royal Road functions in EA [10]. As it is the case of the Royal Road functions (namely the function $R3$ from [8]), the desired sequence (in DNA-alphabet) must include several short stretches of nucleotides (called as sites) with exact consensus sequence (or closely related to it), as shown in Fig. 1. The sequences of the spacers between the sites are arbitrary, but the size of the spacers is often important. Each site serves as the target for specific binding of the DNA-binding factor (Fig. 1) and this specific binding is laid in the basis of the regulatory element functioning (regulation of the gene activity by its regulatory factors). As it is the case of the Royal Road fitness functions in EA, the desired sequence is sought by experimenters from scratch. The finding of each new site raises the whole sequence fitting level by discrete step. The order of the sites finding is arbitrary. There are some dissimilarities between optimization of the Royal Road functions and the problem to breed the gene-regulatory element via SELEX. One of the crucial differences is that each site in a Royal Road function (a building block in the EA terms) has the only sequence, while in biology each site is the family of closely related sequences with (slightly or moderately) different affinity (fitness level). Here we do not study the difference in details.

On the one hand, the general upper bound on expected hitting time of a target area of genotypic space by EA (the EA runtime) from [3] allows to upper-bound the expected time to finding a sufficiently efficient series of motifs (e.g. binding sites for transcription factors) in a SELEX procedure. On the other hand, the theoretical approach [5] yields upper bounds on expected proportion of the DNA sequences with sufficiently high fitness

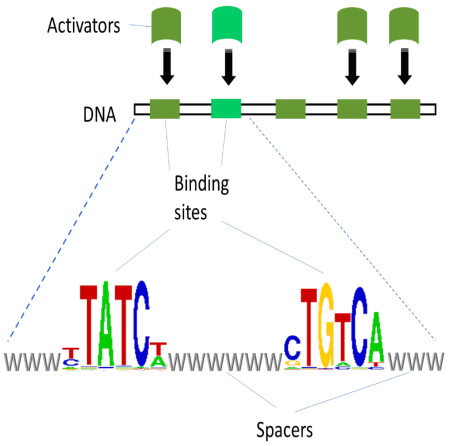

Figure 1: An example of gene-regulatory element ЦI enhancer and its consideration from the EA viewpoint. Enhancer typically is a cluster of specific sites for binding by the DNA-binding factors ЦI activators. Each site is a set of short sequences varying by its exact affinity levels for their factor. The sites are usually separated by spacer sequences. The binding sites are presented by rectangles and by sequences logos. W is arbitrary nucleotide

at a given iteration of SELEX procedure. Both approaches are evaluated in computational experiment, using a Royal Road fitness function as a model of SELEX procedure applied to regulatory element for FIS factors.

A practical case, resembling to our study, may be found in the results of *in vivo* genetic selection (function-based in vivo SELEX) with Turnip Crinkle Virus sequences [12], where 28 bases of the viral regulatory element motif1-hairpin were randomized and then subjected to selection in plants. Most of the winners in this experiment contained up to three short motifs (5-7 bp), many of which are found in other promoter elements of the virus.

2 Gene Regulatory Regions

It is known that a gene consists of coding and regulatory parts. Regulatory part of simply organized genes typically include promoter and enhancer, as illustrated by Fig. 2. We will focus on the enhancer.

An enhancer is a short region of DNA that can be specifically bound by proteins (transcription factors) to increase the likelihood that transcription of a particular gene will occur. Usually the DNA-binding factors, having affinity to an enhancer, are activators. Typically an enhancer is a cluster of sites for recognizing and binding by transcription factors and other DNA-binding factors (as illustrated by Fig. 2). Each DNA-binding site is a relatively short sequence of base pairs (bp) similar with or identical to the, so called, consensus sequence for a given DNA-binding factor. The closer the site sequence

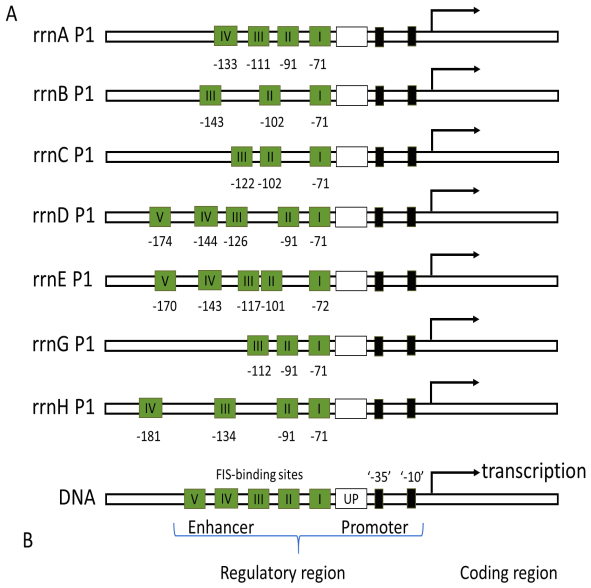

Figure 2: The E.coli ribosomal RNA (rrn) operon as the example of the prokaryotic genes with enhancers. Each of the regulatory regions consists of the core promoter and Upstream Activation Region (UAR). The UAR includes UP element and the cluster of binding sites (3-5 sites) for the DNA-binding factor FIS. The distances between the neighbor FIS-sites are equal or multiple to 20-21 base pairs (bp). (A) Schematic representation of seven rrn operon regulatory regions. (B) The scheme of idealized rrn gene with the cluster of five equally-distanced FIS binding sites. P1 - first promoter. Negative numbers mark positions (in bp) from the transcription initiation site for the diverse regulatory sites of the core promoter and UAR. Core promoter consists of the Ç-10Ÿ and Ç-35Ÿ sites. Broken arrow indicates the transcription initiation site and the transcription direction [6].

to the consensus, the higher is the probability that the factor find and bind to it and the higher is the binding strength. The more effective the factor binding to its DNA-binding site, the higher the probability for the gene to start transcription. Here we will use the well known and exhaustively studied example of the prokaryotic enhancer (also known as the Upstream Activation Region) for the family of the E.coli ribosomal RNA genes (ribosomal operon), as shown in Fig. 2. This is one of the best studied and relatively simply organized prokaryotic genes with enhancer. This regulatory element includes a cluster of DNA-binding sites for the regulatory factor FIS (Fig. 2).

2.1 Binding Sites for the FIS Factor

Not only the DNA-binding site sequences, but also the order and the distances between the neighbor sites can be crucial for the enhancerŸs proper functioning. Some authors call it as the grammar for the gene-regulatory elements [7]. In the FIS-site cluster it is

crucial that the beginning points for the neighbor sites tend to be separated by distances equal or multiple to 20-21 bp (see 2). This feature is supposed to be related with the DNA double helix turn (20-21 bp corresponds to two turns). For our consideration we will use the *Simplifying Assumption 1: the distance between the beginnings of the sites is multiple to the constant (large than the site length).* For the FIS sites the constant would be 20 bp.

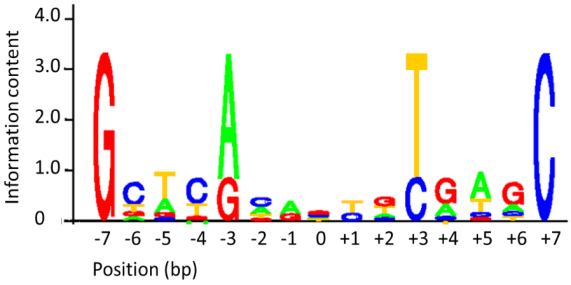

Figure 3: Consensus logo for the factor FIS. Position is counted in bp from the middle of the consensus sequence: the site is palindromic. Position Y-4Y can keep any base except A, position Y+4Y can keep any base except T [11].

Let us now point the attention to the FIS-site details crucial for our following considerations. As we mentioned, a binding site for a given factor is not a unique sequence but a family of similar sequences with different affinities for the factor. To represent the set of sequences and their (estimated) affinity levels, the sequence logo representation was introduced. It consists of a stack of letters at each position. The relative preference for different letters at each position is represented by their relative heights. The total height of the stack of the letters represents the information content of the position (measured in bits), equal to the original entropy of this position minus its a posteriori enthropy value (in the sense of Shannon). The sequence logo for the FIS binding sites is shown in Fig. 3. It can be seen from this figure that we can make *Simplifying Assumption 2*: *Each position of the FIS binding site has either one or two appropriate letters.* In what follows, we will consider a binding site as active if the DNA sequence has appropriate letters in all of its positions.

3 The Non-Elitist Evolutionary Algorithm with (μ, λ)-Selection as a Model of SELEX

3.1 Outline of the Algorithm

Consider a maximization problem:

$$\max\{\phi(x) : x \in \mathcal{A}^n\}, \tag{1}$$

where ϕ is the objective function (called *fitness function* in the EA literature), $\mathcal{A}$ is an alphabet for solutions encoding, e.g. $\{0, 1\}$ in the case of computer systems or $\{A, C, G, T\}$ if (1) is considered as a model of adaptation in genetics.

In the field of evolutionary algorithms, the problems of the form (1) are solved heuristically by modelling a population of individuals that undergo random mutation, selection, and sometimes crossover (see e.g. [1]). The evolutionary process that is associated with such transformations of the population is expected to guide the search towards an optimal (or locally optimal) solution. Sometimes the convergence to optimum may be guaranteed as time tends to infinity [9] or upper bounds may be proven for the expected number of tentative solutions evaluated until an optimum was found [1, 3].

Let $\mathcal{A}^n$ denote the space of genotypes and let λ be the population size. A population of λ individuals (represented by their genotypes) on the EA iteration t is denoted by

$$X^t = (x^{1t}, \ldots, x^{\lambda t}) \in \mathcal{A}^{n\lambda},$$

where x^{kt} is an individual number k in X^t, $k = 1, \ldots, \lambda$. During the mutation, a subset of genes in the genotype string x is randomly altered. Given a genotype x, the output of a mutation operator may be viewed as a random variable $\mathrm{Mut}(x) \in \mathcal{X}$ with the probability distribution depending on x. The most frequently used type of this operator, the *bitwise mutation* randomly changes each gene of x with a given mutation probability p_m. In this paper, we will consider only the bitwise mutation, assuming that a new value for any mutated gene x_i is chosen at random from $\mathcal{A}\backslash\{x_i\}$. In (μ, λ)-*selection* operator, parents are sampled uniformly at random among the fittest μ individuals in the population. The overall scheme of the EA considered in this paper is as follows.

1. Generate the initial population X^0.
2. For $t := 0$ to $t_\mathrm{max} - 1$ do
   2.1. For $k := 1$ to λ do
      2.1.1. Choose a parent genotype x from X^t
         by (μ, λ)-selection.
      2.1.2. Add $x_k^{(t+1)} = \mathrm{Mut}(x)$ to the population X^{t+1}.
3. Output the best incumbent $\tilde{x}^t$, i.e.
   the fittest genotype in $X^0, \ldots, X^t$.

In theoretical studies, the EAs are usually treated without a stopping criterion. Therefore we will also assume that $t_\mathrm{max} = \infty$. The described EA may be considered as a simplified version of a genetic algorithm (see e.g. [9]), which does not use a crossover operator in our case.

3.2 Modeling SELEX for Gene Regulatory Region

3.2.1 SELEX described in brief

SELEX procedure for DNA sequences *in vitro* works as follows [4]. Initially a chemically synthesized DNA library is incubated with target molecules. Unbound molecules are re-

moved and the target/DNA complex is split. Released DNA sequences (μ best individuals in terms of the EA) are amplified by the PCR reaction with possible modifications by mutations (the next population of λ individuals is built) and the next round of selection is performed. Typically, this process is repeated for several or more than 20 rounds. Analogous procedure may be applied to RNAs. In some cases the SELEX procedure may be implemented *in vivo* or *in silico*, or in combination. Unlike the *in vitro* SELEX, the function-based *in vivo* SELEX searches for functional sequences that contribute to the fitness of the molecule [12].

EAs generally and genetic algorithms in particular can be used as the approach to simulate SELEX procedures, as well as, some other related experimental techniques in modern bioengineering. Here we consider the EA as a model of SELEX for gene-regulatory elements with many sites.

3.2.2 Royal Road functions as a model of FIS promoter

Royal Road functions were introduced and used to study the significance of the *building block hypothesis* for crossover operators [8]. Here we use the Royal Road functions to model the design of the gene regulatory elements from scratch in the experimental approaches of in vitro evolution, such as SELEX.

The original definition of a Royal Road function [8] is formulated for the binary alphabet $\mathcal{A} = \{0,1\}^n$, assuming that a set S of *schemata* is given. Each scheme $s \in S$ is an n-element string of symbols from the alphabet $\mathcal{A} \cup \{\text{"*"}\}$. A string $x \in \mathcal{A}$ is called *an instance* of scheme s iff $x_i = s_i$ for all positions where $s_i \neq$ "*". Suppose that a set of positive weights $c_s,\ s \in S$ is given. In [8], the Royal Road function is defined as $\phi(x) := \sum_{s \in S} c_s [x \text{ is an instance of } s]$. Here and below, $[\cdot]$ denotes the Iverson bracket:

$$[P] := \begin{cases} 1, & \text{if } P \text{ is true;} \\ 0 & \text{otherwise.} \end{cases}$$

for any statement P that can be true or false.

One of the frequently used versions of Royal Road functions in computer science (see e.g [3]) is defined for $\mathcal{A} = \{0,1\}$, assuming n/r unweighted non-overlapping schema with r fixed positions per scheme (these positions are called *a block*):

$$\textsc{RoyalRoad}_r(x) := \sum_{i=0}^{n/r-1} \prod_{j=1}^{r} x_{ir+j}.$$

In this paper, we generalize the definition of Royal Road functions, including the non-binary alphabets and schema positions with two appropriate letters, in order to model the FIS binding sites. Without loss of generality, we will assume that all positions with a single appropriate value require the last letter $a_{|\mathcal{A}|}$ of the alphabet $\mathcal{A}$ and they occupy the first r_1 positions of each block, all positions with two appropriate letters admit the last two letters $a_{|\mathcal{A}|-1}, a_{|\mathcal{A}|}$ of the alphabet and they occupy the remaining r_2 positions

of each block, $r = r_1 + r_2$. We will denote this generalized Royal Road function as $\textsc{RoyalRoad}_{r_1,r_2}(x)$, assuming it to equal

$$\sum_{i=0}^{n/r-1} \prod_{j=1}^{r_1} [x_{ir+j} = a_{|\mathcal{A}|}] \prod_{j=r_1+1}^{r} [x_{ir+j} \in \{a_{|\mathcal{A}|}, a_{|\mathcal{A}|-1}\}].$$

Considering the example of [12], where the SELEX procedure yielded up to three active binding sites, we make *Simplifying Assumption 3*: *The selection criterion employed in the SELEX procedure is an increasing function of the number of active binding sites in a string x.* In view of the simplifying assumptions 1–3, the binding strength of FIS promoter may be modeled by the generalised Royal Road function with 4 to 6 blocks (each block corresponds to a separate binding site of FIS factor) where $r_1 = 2, r_2 = 6$. The search space consists of strings of length $n = 32, 40$ or 48 with symbols from the 4-letter alphabet $\mathcal{A} = \{A, C, G, T\}$.

4 Theoretical Analysis of the Non-Elitist Evolutionary Algorithm with (μ, λ)-Selection

4.1 Upper Bounds on Proportion of Fit Genotypes in Population of Evolutionary Algorithm

Assume that $\phi_0 := \min\{\phi(x) : x \in \mathcal{X}\}$ and there are d level lines of the fitness function fixed so that $\phi_0 < \phi_1 < \phi_2 \ldots < \phi_d$. Let us define $d+1$ subsets of $\mathcal{X}$

$$H_i := \{x : \phi(x) \geq \phi_i\}, \quad i = 0, \ldots, d.$$

Obviously, $H_0 = \mathcal{X}$. For the sake of convenience, we define $H_{d+1} := \emptyset$. Also, we denote the level sets $A_i := H_i \backslash H_{i+1},\ i = 0, \ldots, d$ which give a partition of $\mathcal{X}$. Now suppose that for all $i = 0, ..., d$ and $j = 1, ..., d$, the a priori upper bounds β_{ij} on mutation transition probabilities from subset A_i to H_j are known in step 2.1.2 of the EA:

$$\Pr\{\mathrm{Mut}(x) \in H_j \mid x \in A_i\} \leq \beta_{ij}.$$

Fig. 4 illustrates the transitions considered here.

In what follows, $\mathbf{B}$ denotes the matrix with elements β_{ij}, $i = 0, ..., d$, $j = 1, ..., d$. Let the population on iteration t be represented by the *population vector*

$$\mathbf{z}^{(t)} = (z_1^{(t)}, z_2^{(t)}, \ldots, z_d^{(t)})$$

where $z_i^{(t)} \in [0, 1]$ is the proportion of genotypes from H_i in X^t. The population vector $\mathbf{z}^{(t)}$ is a random vector, where $z_i^{(t)} \geq z_{i+1}^{(t)}$ for $i = 1, ..., d-1$ since $H_{i+1} \subseteq H_i$.

Let $\Pr\{x^{(t)} \in H_j\}$ be the probability that an individual, which is added after selection and mutation into X^t, has a genotype in H_j, $j = 0, ..., d$, $t > 0$. According to the scheme of the EA, $\Pr\{x^{(t)} \in H_j\} = \Pr\{x_1^{(t)} \in H_j\} = ... = \Pr\{x_\lambda^{(t)} \in H_j\}$. The following proposition is easy to prove (see e.g. Proposition 1 in [5]).

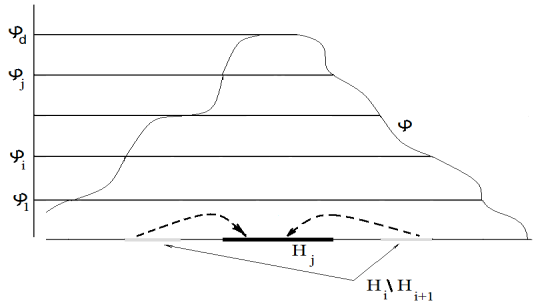

Figure 4: Transitions from $H_i \backslash H_{i+1}$ to H_j during mutation.

Proposition 1 *For all* $t > 0$, $i = 1, ..., d$ *holds* $\mathbf{E}[z_i^{(t)}] = \Pr\{x^{(t)} \in H_i\}$.

Let $P_{\mathrm{ch}}(z_i)$ denote the probability to choose a parent individual from H_i. By the definition of (μ, λ)-selection,

$$P_{\mathrm{ch}}(z_i) = \begin{cases} z_i \lambda/\mu, & \text{if } \ z_i \leq \mu/\lambda, \\ 1, & \text{otherwise.} \end{cases}$$

By a reasoning similar to that in Subsection 3.1 of [5] we conclude that

$$\Pr\{x^{(t+1)} \in H_j | \mathbf{z}^{(t)} = \mathbf{z}\} \leq \sum_{i=0}^{d} \beta_{ij} (P_{\mathrm{ch}}(z_i^{(t)}) - P_{\mathrm{ch}}(z_{i+1}^{(t)})),$$

implying the upper bounds

$$\mathbf{E}[z_j^{(t+1)}] \leq \beta_{dj} - \sum_{i=1}^{d} (\beta_{i,j} - \beta_{i-1,j}) \ \mathbf{E}[1 - P_{\mathrm{ch}}(z_i^{(t)})] \tag{2}$$

for the expected proportion of genotypes with fitness above each of the given levels $\phi_1, \ldots, \phi_d$.

A $((d+1) \times d)$-matrix $\mathbf{B}$ is called *monotone* iff $\beta_{i-1,j} \leq \beta_{ij}$ for all i, j from 1 to d. Monotonicity of a matrix $\mathbf{B} = (\beta_{ij})$ means that the greater fitness level A_i a parent solution has, the greater is its bound on transition probability to any subset H_j, $j = 1, \ldots, d$. In other words, it means that $\beta_{i,j} - \beta_{i-1,j} \geq 0$.

The following proposition is proved analogously to Proposition 4 in [5].

Proposition 2 *If* $\mathbf{B}$ *is monotone then for all* $j = 1, \ldots, d$

$$\mathbf{E}[z_j^{(t+1)}] \leq \beta_{dj} - \sum_{i=1}^{d} (\beta_{ij} - \beta_{i-1,j}) \left(1 - P_{\mathrm{ch}}(\mathbf{E}[z_i^{(t)}]\right). \tag{3}$$

By an iterative application of inequality (3), the components of the expected population vectors $\mathbf{E}[\mathbf{z}^{(t)}]$ may be bounded up to an arbitrary t, starting from the initial

vector $\mathbf{E}[\mathbf{z}^{(0)}]$, describing the population X^0. Note that the obtained estimate is independent of the population size and valid for arbitrary λ.

In the case of $\textsc{RoyalRoad}_{r_1,r_2}$ function, we will use the term *1-block* for any block complying with its scheme (i.e. all r_1 positions that require the concensus value, are given the concensus vlaue and in all r_2 positions that admit two options, one of the two admissible values is assigned). Otherwise we will call the block *a 0-block*. The transition probabilities between the 0- and 1-states of the block during mutation are as follows:

$\Pr(0 \to 1) \le \frac{2}{3}p_{\mathrm{m}}$; $\Pr(0 \to 0) = 1 - \Pr(0 \to 1)$;

$\Pr(1 \to 1) = \left(1 - p_{\mathrm{m}} + \frac{1}{3}p_{\mathrm{m}}\right)^{r_2} (1 - p_{\mathrm{m}})^{r_1}$;

$\Pr(1 \to 0) = 1 - \Pr(1 \to 1)$.

It is natural to assume that d equals to the number of blocks $n/(r_1 + r_2)$ and the subsets $H_0, ..., H_d$ correspond to the level lines $\phi_0 = 0$,$\phi_1 = 1$,..., $\phi_d = d$. The matrix of upper bounds $\mathbf{B}$ may be found by the explicit formula (20) from [5], denoting $\tilde{r} := \frac{2}{3}p_{\mathrm{m}}$. This matrix $\mathbf{B}$ satisfies the monotonicity property, provided that $\tilde{r} \le \Pr(1 \to 1)$, e.g. in the case of $r_1 = 2$, $r_2 = 6$ this holds for all $p_{\mathrm{m}} < 1/4$.

4.2 Theoretical Upper Bound on the EA Runtime

Let T denote the random variable, equal to the number of tentative solutions evaluated until some element of the current population is sampled from H_d for the first time. In the case when H_d is the set of optimal solutions, T is usually called the *runtime* of an evolutionary algorithm. The topic of constructing lower and upper bounds on the runtime of different evolutionary algorithms is much more popular, compared to bounding the abundance of sufficiently fit individuals as it was presented in Subsection 4.1.

Suppose that a lower bound p_0 is known for the probability not to reduce the fitness-level of any genotype under mutation, i.e. $\Pr\{\mathrm{Mut}(x) \in H_j \mid x \in A_j\} \ge p_0$ for all $j = 1, \ldots, d-1$.

Besides that, let us suppose that for each level $j = 0, \ldots, d-1$ some lower bounds s_j are known for the improving probabilities, i.e. $\Pr\{\mathrm{Mut}(x) \in H_{j+1} \mid x \in A_j\} \ge s_j$ for all $j = 0, \ldots, d-1$. Denote $s_* := \min_{j=0,\ldots,d-1} s_j$.

The Level Theorem for bounding the runtime of non-elitist EAs [3] implies the following

Corollary 1 *If the EA with (μ, λ)-selection is used with*

- *sufficiently small ratio μ/λ, such that $\mu/\lambda \le p_0/(1+\delta)$ for some $\delta \in (0, 1]$,*
- *sufficiently large population size, such that $\lambda \ge \left(\frac{4\lambda}{\delta^2\mu}\right) \ln\left(\frac{128(d+1)\lambda}{s_*\delta^2\mu}\right)$*

then the expected EA runtime satisfies the inequality

$$\mathbf{E}[T] < UB := \left(\frac{8}{\delta^2}\right) \sum_{j=0}^{d-1} \left(\lambda \ln\left(\frac{6\delta\lambda}{4 + \mu s_j \delta}\right) + \frac{\lambda}{\mu s_j}\right).$$

It is easy to verify that if $\textsc{RoyalRoad}_{r_1,r_2}$ function is used as the fitness function in the EA, then we can assume:

$$p_0 := (1-p_{\rm m})^{(d-1)r_1}\left(1-\frac{2p_{\rm m}}{3}\right)^{(d-1)r_2},$$

$$s_* := \left(\frac{p_{\rm m}}{3}\right)^{r_1}\left(\frac{2p_{\rm m}}{3}\right)^{r_2}\left((1-p_{\rm m})^{r_1}\left(1-\frac{2p_{\rm m}}{3}\right)^{r_2}\right)^{(d-1)} \quad (4)$$

$$s_j := (d-j)s_*,\ j=0,\ldots,d-1. \quad (5)$$

5 Application of Bounds from EA Theory to SELEX-Type Procedure

5.1 Ratio of Optimal Individuals in Computational Experiment and the Upper Bound for It

Below we present the experimental results in comparison with the theoretical estimates obtained in Subsection 4.1. To this end we consider an application of the EA to the $\textsc{RoyalRoad}_{2,6}$ fitness function, modelling SELEX for 5 FIS sites. The average proportion of sufficiently fit genotypes for three different fitness levels is presented in Fig. 5. Here $\lambda = 10000$, $\mu = 100$, $p_{\rm m} = 0.1$. (Note: it can be shown analogously to Theorem 3 [5] that with increasing population size λ, upper bound (3) becomes tighter.)

The statistics is accumulated over 1000 runs of the algorithm and one individual $x_1^{(t)}$ for each t is checked for hitting the target subset H_d, H_{d-1} or H_{d-2}. Note that $\mathbf{E}[z_i^{(t)}] = \Pr\{x_1^{(t)} \in H_i\}$ by Proposition 1 and e.g. for H_d at a given t we have a series of 1000 Bernoully trials with a success probability $\Pr\{x_1^{(t)} \in H_d\}$, estimated from the experimental data, together with the 95%-confidence interval.

The experimental results are shown in dashed lines. The solid lines correspond to the upper bounds obtained by the iterative application of (3).

As it can be seen from Fig. 5, after approximately 50 iterations the upper bound from Proposition 2 becomes relatively close to the true value of proportion of near-optimal genotypes.

5.2 Expected Runtime Bound Compared to Computational Experiment

In order to evaluate the upper bound on the EA expected runtime, suggested by Corollary 1, we have carried out a computational experiment with 1000 independent runs of the EA, given the mutation probability $p_{\rm m} = 0.1$.

Application of the runtime bound to the case of FIS-enhancer, modelled by the Royal Road function with $r_1 = 2, r_2 = 6$, $n = 32$, i.e. 4 blocks, assuming $\lambda = 5000$, $\mu = 500$, in the EA gives $UB = 3.9 \cdot 10^{13}$, while the experimental average runtime is $\hat{T}_{exp} = 172190$.

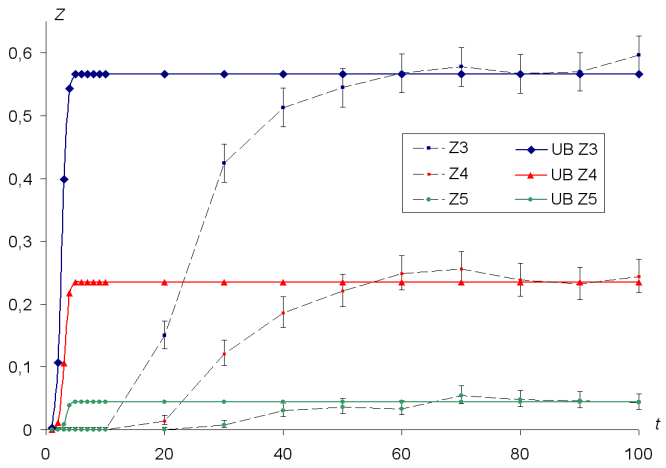

Figure 5: Comparison of upper bounds (UBz_3, UBz_4 and UBz_5) and experimental estimates for the average proportion of optimal and sub-optimal individuals z_5, z_4 and z_3 in the population of EA with $\lambda = 10000$, $\mu = 100$, $p_{\rm m} = 0.1$, modelling SELEX for 5 FIS sites (i.e. the fitness function is $\textsc{RoyalRoad}_{2,6}$). Confidence intervals are computed with level 95%.

Application of the runtime bound to the case of $r_1 = 4, r_2 = 2$ with 5 blocks and $n = 30$ (corresponds to the example from Fig. 1), assuming $\lambda = 2000, p_{\rm m} = 0.1$ and $\mu = 100$ gives $UB = 4.852 \cdot 10^{11}$, while the experiment gives $\hat{T}_{exp} = 98270$.

Most likely, the reason for the large over-estimation in the runtime upper bound implied by Corollary 1 is in the over-pessimistic assumption of the block updates under mutation, used for computing the values s_* and s_j in (4) and (5). These probability bounds are computed for the worst-case scanrio, assuming that if a binding site does not meet the binding requirements, then all positions in this site are different from the consensus. This also agrees with the outcome of our additional experiment where we set $r_1 = 0, r_2 = 1$ and $n = 5$, obtaining much tighter results: $UB = 2293.1$ and $\hat{T}_{exp} = 1202.4$ (in fact the upper bound of Corollary 1 becomes assymptotically tight in this special case of Royal Road function, known as the OneMax function [3]). Thus, further problem-specific EA analysis is required in order to keep track of variability of the components in blocks of the Royal Road function.

6 Conclusions

Two approaches from the theory of evolutionary algorithms (bounding the proportions of fit individuals in the EA population and bounding the EA runtime) are applied to model the experimental techniques in modern bioengineering. To this end, the theoretical bounds obtained by both of the approaches are applied to a non-elitist mutation-based evolutionary algorithm with (μ, λ)-selection optimizing a generalized Royal Road fitness function. We argue that this EA may be considered as a model of SELEX procedure of

in vitro evolution "from scratch" for FIS promoter with many binding sites. Theoretical predictions are compared to the results of computational experiments.

Our analysis indicates that the considered theoretical bounds (see Subsection 4.1), in principle, may be used for prediction of abundance of promoter sequences with sufficiently high affinity to a target protein after a given number of iterations of a SELEX procedure. The upper bounds on the average number of SELEX rounds until the required sequence is found (see Subsection 4.2) appear to be over-pessimistic. Further research is needed in order to improve the theoretical bounds so that they may be applied to the SELEX procedures consisting of just several rounds. More detailed models of the binding sites may be developed in order to incorporate more details into the structure of the EA fitness function. Further research is needed further the theoretical prediction and the estimates found in computational experiments are compared against the results of practical experiments.

Acknowledgement

The research was supported by the Russian Science Foundation (grant 17-18-01536).

References

[1] A. Auger, B. Benjamin, "Theory of Randomized Search Heuristics: Foundations and Recent Developments". World Scientific, 2011.

[2] P. Borisovsky and A. Eremeev, "Comparing evolutionary algorithms to the (1+1)-EA". Theoretical Computer Science, vol. 403(1), pp. 33Ц-41, 2008.

[3] D. Corus , D.-C. Dang, A.V. Eremeev and P.K. Lehre. (2017, September) Level-based analysis of genetic algorithms and other search processes. IEEE Transactions on Evolutionary Computation. [Online]. Available: https://doi.org/10.1109/TEVC.2017.2753538

[4] M. Darmostuk, S. Rimpelova, H. Gbelcova, T. Ruml, "Current approaches in SELEX: An update to aptamer selection technology", *Biotechnology Advances,* vol. 33, pp. 1141Ц-1161, 2015.

[5] A.V. Eremeev, "On proportions of fit individuals in population of genetic algorithm with tournament selection", *Evolutionary Computation*, vol. 26(2), pp. 269–297, 2018.

[6] C.A. Hirvonen, W. Ross, C.E. Wozniak, E. Marasco, J.R. Anthony, S.E. Aiyar, V.H. Newburn, R.L. Gourse, "Contributions of UP elements and the transcription factor FIS to expression from the seven rrn P1 promoters in Escherichia coli" *J Bacteriol.,* vol. 183(21), pp. 6305-6314, 2001.

[7] J. Gertz, E.D. Siggia, B.A. Cohen, "Analysis of combinatorial cis-regulation in synthetic and genomic promoters", *Nature*, vol. 457, pp. 215Ц-218, 2009.

[8] M. Mitchell, S. Forrest, J.H. Holland, "The royal road for genetic algorithms: fitness landscapes and GA performance" In: Proc. of the 1st European Conf. on Artificial Life. MIT Press. Cambridge, MA, pp. 245-254, 1992.

[9] G. Rudolph, "Finite markov chain results in evolutionary computation: A tour d'horizon", *Fundamental Informaticae,* vol. 35(1-4), pp. 67–89, 1998.

[10] A. Spirov and D. Holloway, "New approaches to designing genes by evolution in the computer". In: Real-World Applications of Genetic Algorithms (ed. by O. Roeva). InTech, pp. 235–260, 2012.

[11] Y. Shao, L.S. Feldman-Cohen, R. Osuna, "Functional characterization of the Escherichia coli Fis-DNA binding sequence". *J. Mol. Biol.*, vol. 376(3), pp. 771–85, 2008.

[12] G. Zhang and A.E. Simon, "A multifunctional Turnip Crinkle Virus replication enhancer revealed by in vivo functional SELEX". *J. Mol. Biol.* vol. 326, pp. 35-Ц48, 2003.