Evaluation of Runtime Bounds for SELEX Procedure with High Selection Pressure

Anton Eremeev
Sobolev Institute of Mathematics SB RAS
Omsk, Russia Institute of Scientific Information for Social Sciences RAS
Moscow, Russia

Alexander Spirov
I.M. Sechenov Institute of Evolutionary Physiology and Biochemistry RAS
St. Petersburg, Russia Institute of Scientific Information for Social Sciences RAS
Moscow, Russia

ABSTRACT

The aim of this paper is to apply theoretical bounds known for the Evolutionary Algorithms (EAs) to the genetic engineering technique of Systematic Evolution of Ligands by EXponential enrichment (SELEX). We discuss how the EAs optimizing Royal Road or Royal Staircase fitness functions may be considered as models of evolutionary search "from scratch". We consider the design of synthetic enhancers and promoters in SELEX. This problem asks for a tight cluster of supposedly unknown motifs from the initial random set of DNA sequences using SELEX. 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 number of rounds of SELEX until a series of binding sites for protein factors is found. The theoretical bounds are compared to the results of computational experiments modelling bacterial promoters and enhancers. Our results suggest that for some cases with large population size, theoretical bounds give favorable prediction, while computational experiments require prohibitive CPU resource.

CCS CONCEPTS

• **Computing methodologies** → **Artificial life**; *Discrete-event simulation*; • **Applied computing** → **Systems biology**.

KEYWORDS

Runtime Analysis, SELEX procedure, Royal Road, Royal Staircase

ACM Reference Format:
Anton Eremeev and Alexander Spirov. 2019. Evaluation of Runtime Bounds for SELEX Procedure with High Selection Pressure. In *Proceedings of the Genetic and Evolutionary Computation Conference 2019 (GECCO '19 Companion).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3319619.3321906

INTRODUCTION

SELEX (Systematic Evolution of Ligands by EXponential enrichment) procedures are known as a valuable tool for finding DNA or RNA sequences with high affinity for a pre-specified target molecules. SELEX may be considered as an experimental implementation of Evolutionary Algorithm (EA). Experimenters iteratively apply selection, reproduction and mutation to populations of nucleic acid molecules to breed a desired DNA or RNA sequence.

In this paper, we aim at the *prediction of efficiency* of SELEX for gene-regulatory elements (promoters and enhancers), if the parameters of the optimal sequence can be predicted to some degree, e.g. on the basis of existing precedents. Efficiency of a typical gene-regulatory element with some degree of simplification can be described by a Royal Road [5] or a Royal Staircase [7] fitness function, where the four-letter alphabet of nucleotides is used instead of the binary alphabet. The desired sequence in DNA-alphabet must include several short subsequences of nucleotides (binding sites) with exact match in some fixed positions and several acceptable options in others. The contents of spacers between the binding sites are arbitrary, but the length of the spacers is often important. Each binding site serves as a target for a specific protein (DNA-binding factor), and if such a binding occurs, it can influence the activity of the gene, adjacent to the regulatory element. We assume that the desired subsequences of binding sites are sought "from scratch". If the order of the sites finding is arbitrary and all sites have identical binding constraints, we can take a four-letter version of the Royal Road function as a fitness model. If the order of the sites finding is pre-determined, we can use a four-letter version of the Royal Staircase fitness function. This approach to modelling SELEX for regulatory elements by means of EAs was proposed in [3, 6], where further biological details may be found. In the present paper, we apply another theoretical technique [2] for the EA analysis, which is more appropriate in the case of high selection pressure. The upper bound on expected first hitting time of optimal solutions (the EA runtime) from [2] allows to upper-bound the expected time to finding a sufficiently efficient series of motifs (e.g. binding sites) in a SELEX. We evaluate this approach in computational experiment on the examples of bacterial ribosomal RNA (rRNA) operon promoter rrnB P1 of E. coli and its enhancer [4]. It is expected that development of such methods will be helpful for prediction of efficiency of *in vitro* evolution, combined with rational design strategies.

1 NON-ELITIST EA AS A MODEL OF SELEX

SELEX procedure for DNA sequences *in vitro* works as follows (see e.g. [1]). Initially a chemically synthesized DNA library is incubated with target molecules. Unbound molecules are removed and the target-DNA complex is split. Bound DNA sequences (fittest individuals in terms of the EA) are amplified by the PCR reaction with possible mutations and the next round of SELEX is performed.

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3321906

This is repeated for several rounds. SELEX may be applied to RNAs and may also be implemented *in vivo* or *in silico*.

Here we consider a non-elitist EA with (μ, λ)-selection (see e.g. [2]) as a model of SELEX for gene-regulatory elements with many binding sites. Let $\mathcal{A}$ be an alphabet for solutions encoding, e.g. $\{A, C, G, T\}$ as in genetics. The space of genotypes is $\mathcal{A}^n$. A population of λ individuals $x^{1t}, \ldots, x^{\lambda t}$ from $\mathcal{A}^{n\lambda}$ on the EA iteration t is denoted by X^t. We assume that the mutation operator randomly changes each component of x with a given mutation probability p_{m}, assuming that a new value for any mutated component x_i is chosen at random from $\mathcal{A}\backslash\{x_i\}$. In the (μ, λ)-*selection* operator, parents are sampled uniformly at random among the μ fittest individuals in X^t.

An upper bound on the runtime of this EA in case of sufficiently large λ and selection pressure may be found by Corollary 4.1 [2].

Generalized Royal Road fitness for modelling enhancer. The Royal Road functions were defined in [5] 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 $\mathcal{A} \cup \{"*"\}$. A string $x \in \mathcal{A}$ is *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. One of the frequently used versions of Royal Road functions is defined for $\mathcal{A} = \{0, 1\}$, assuming n/r non-overlapping schema with r fixed positions per scheme (these positions are called *a block*).

In [3], the Royal Road functions are extended to non-binary alphabets and schema positions with two appropriate letters are allowed. W.l.o.g. 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, $r = r_1 + r_2$. The generalized Royal Road function $\mathrm{RR}_{r_1,r_2}(x)$ is defined as

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

A simplifying assumption can be made that the selection criterion in the SELEX procedure is an increasing function of the number of active binding sites in a string x (see [3]). Then the generalized Royal Road function may be used to model an enhancer. In the example of FIS factors of rRNA rrnB P1 enhancer, one can put $r_1 = 2, r_2 = 6$, where each of the four blocks corresponds to a separate binding site of enhancer (see the details in [3]).

Generalized Royal Staircase fitness for modelling promoter. The class of Royal Staircase fitness functions, defined for the bitstirngs [7] is a generalization of the well-known LeadingOnes function. Each function in this class is defined by two parameters, the number of blocks and the number r of bits per block. Starting from the first position of a given string s, the number $I(s)$ of consecutive 1s in a string is counted and the fitness is assumed to be $\lfloor I(s)/r \rfloor$, i.e. the number of consecutive fully-set blocks starting from the left.

The generalized Royal Staircase functions are based on the same extension of the alphabet size as in the case of $\mathrm{RR}_{r_1,r_2}(x)$, and besides that we allow to choose the values r_1, r_2 specifically for each block. So a generalized Royal Staircase function with k blocks is denoted by $\mathrm{RS}^{r_{21},\ldots,r_{2k}}_{r_{11},\ldots,r_{1k}}(x)$, where r_{1i} is the number of positions with one appropriate value in block i, and r_{2i} is the number of positions with two appropriate letters in block i, $i = 1, \ldots, k$. The effects of the binding sites are enabled sequentially, ordered by their distance to the gene. Therefore it is expected that a promoter evolves in SELEX by sequential finding and adding of building blocks, with each addition raising the transcriptional efficiency (fitness) of the promoter sequence. In the example of rRNA rrnB P1 promoter in E. coli, one can assume $k = 6$ (the FIS binding site with two domains, as well as the UP site), defining $r_{11} = r_{21} = r_{12} = r_{22} = 3, r_{13} = r_{14} = 6, r_{23} = 2, r_{24} = 5, r_{15} = r_{16} = 1, r_{24} = r_{25} = 4$.

2 THEORY VS SIMULATION

The runtime bound [2] in the case of EA for the FIS enhancer ($\mathrm{RR}_{2,6}(x)$ function) with $p_{\mathrm{m}} = 0.03, \lambda = 6 \cdot 10^{15}$, and $\mu = 10$, predicts at most 13.6 generations on average. This population size is practically implementable in the *in vitro* SELEX but prohibitive for the EA simulation aimed at forecasting the average number of SELEX iterations till the required sequence will be found. A computational experiment with $p_{\mathrm{m}} = 0.03, \mu = 17$ and much smaller population size $\lambda = 10^5$ gave 53 generations on average (SEM=5).

For the rrnP1 promoter ($\mathrm{RS}(x)$ function), $p_{\mathrm{m}} = 0.2, \lambda = 5 \cdot 10^{14}$, and $\mu = 12$, the upper bound [2] gives 13.6 rounds on average. A computational experiment with $p_{\mathrm{m}} = 0.2, \mu = 12$ and much smaller population size $\lambda = 10^5$ gives 125.7 rounds on average (SEM=3.5).

Comparing these results we conclude that in the case of large (but practically appropriate in SELEX) population sizes with very high selection pressure, the proposed theoretical bound may give favorable prediction, while computational experiments require prohibitive computational resource. In the EA simulations described above, the average number of iterations till finding an optimum turns out to be larger than a usual number of rounds of SELEX in practice. Further research is needed to tighten runtime bounds because in many other cases the EA simulations demonstrate the runtime by several orders of magnitude smaller than the bound.

On one hand, in the case of high selection pressure, the new upper bound on the runtime is tighter than the one discussed in [3]. On the other hand, the upper bounds [3] on the *proportion* of optimal solutions in case of $\mathrm{RR}_{r_1,r_2}(x)$ seem to be tighter and more robust.

ACKNOWLEDGMENTS

Supported by the Russian Science Foundation grant 17-18-01536.

REFERENCES

[1] M. Darmostuk, S. Rimpelova, H. Gbelcova, and T. Ruml. 2015. Current approaches in SELEX: An update to aptamer selection technology. *Biotech. Advances* 33, 6, Part 2 (2015), 1141 – 1161.
[2] A. Eremeev. 2017. Hitting times of local and global optima in genetic algorithms with very high selection pressure. *Yugoslav J. of Oper. Res.* 27, 3 (2017), 323–339.
[3] A. Eremeev and A. Spirov. 2018. Estimates from Evolutionary Algorithms Theory Applied to Gene Design. In *11th Int. Multiconf. BGRS\SB.* 33–38.
[4] S. T. Estrem, W. Ross, T. Gaal, Z.W. Chen, W. Niu, R.H. Ebright, and R.L. Gourse. 1999. Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes & Development* 13, 16 (1999), 2134–2147.
[5] M. Mitchell, S. Forrest, and J.H. Holland. 1992. The royal road for genetic algorithms: fitness landscapes and GA performance. In *1st European Conf. on Artificial Life.* MIT Press, Cambridge, MA, 245–254.
[6] A. Spirov and E. Myasnikova. 2018. Evolutionary Computations and Modular Organization of the Gene Regulatory Regions. In *11th Int. Multiconf. BGRS\SB.* 94–99.
[7] E. van Nimwegen and J.P. Crutchfield. 2000. Optimizing Epochal Evolutionary Search Population-Size Independent Theory. *Computer Methods in Applied Mech. and Engineering* 186, 2-4 (2000), 171–194.