Математические структуры и моделирование 2020. № 1(53). С. 56–76

УДК 519.16
DOI 10.24147/2222-8772.2020.1.56-76

ПРИМЕНЕНИЕ ОЦЕНОК ИЗ ТЕОРИИ ЭВОЛЮЦИОННЫХ ВЫЧИСЛЕНИЙ К ПРОЦЕДУРАМ НАПРАВЛЕННОЙ ЭВОЛЮЦИИ

А.В. Еремеев[1,2]
д.ф.-м.н., доцент, e-mail: eremeev@ofim.oscsbras.ru
А.В. Спиров[2,3]
к.б.н., e-mail: sspirov@yandex.ru

[1]Институт математики им. С.Л. Соболева СО РАН, Омск, Россия
[2]Институт научной информации по общественным наукам РАН, Москва, Россия
[3]Институт эволюционной физиологии и биохимии им. И.М. Сеченова, Санкт–Петербург, Россия

Аннотация. Область эволюционных вычислений возникла в компьютерных науках в результате переноса идей из эволюционной биологии и развивалась независимо в течение нескольких десятилетий. Цель настоящей работы состоит в демонстрации возможности переноса некоторых результатов из теории эволюционных вычислений обратно в биологию. Показано, что эволюционные алгоритмы без элитных особей могут рассматриваться как модели эволюционного поиска последовательности синтетического энхансера «с нуля». Эта задача состоит в отыскании одного из небольшого числа предположительно неизвестных последовательностей энхансера, начиная работу с исходного случайного набора последовательностей ДНК. В генной инженерии для таких целей используются методы направленной эволюции, и в частности, метод SELEX. В настоящей работе применяются верхние оценки математического ожидания времени первого достижения целевой области пространства решений, известные для эволюционных алгоритмов, для оценки среднего числа итераций процедуры SELEX до получения искомой серии последовательностей. Кроме того, с использованием теории эволюционных вычислений предлагается верхняя оценка математического ожидания доли последовательностей ДНК с достаточно высокой приспособленностью после заданного числа итераций процедуры SELEX. Оба подхода оцениваются в вычислительном эксперименте с использованием целевой функции Royal Road, с некоторым упрощением определяющей количество сайтов специфического связывания транскрипционного фактора.

Ключевые слова: время первого достижения, процедура SELEX, функция Royal Road, энхансер.

Введение

Область эволюционных вычислений возникла в компьютерных науках в результате переноса идей из биологии и развивалась независимо в течение нескольких десятилетий, будучи обогащаема методами из теории вероятностей, теории сложности и оптимизации. Цель настоящей работы состоит в демонстрации того, как некоторые результаты из теории эволюционных вычислений могут быть перенесены обратно в эволюционную биологию. Поскольку некоторые биологические термины могут оказаться незнакомы читателю, статья снабжена приложением с терминологическими пояснениями.

Процедуры направленной эволюции типа SELEX (систематическая эволюция лигандов путём экспоненциального обогащения) известны как ценный инструмент в поиске ДНК- и РНК-последовательностей с высокой способностью связывания с предварительно определённой молекулой-мишенью, например — молекулой белка. Однако процедура SELEX затратна по времени и стоимости. В связи с этим параллельно с практическими процедурами *in vitro* для отбора и оценки последовательностей ДНК и РНК были разработаны подходы *in silico*, позволяющие сократить время и стоимость получения искомых последовательностей за счёт компьютерного имитационного моделирования процедуры SELEX (см., например, [10, 15]).

Биотехнологические подходы, подобные SELEX, могут рассматриваться как экспериментальные реализации эволюционных алгоритмов (ЭА) [23]. В этих процедурах циклически оценивают, мутируют и применяют отбор к популяциям молекул нуклеиновых кислот с целью получения требуемой последовательности нуклеотидов (например последовательности промотора или энхансера). Энергию свободного связывания транскрипционного фактора с последовательностью нуклеотидов можно рассматривать как молекулярную реализацию функций приспособленности известного семейства Royal Road [20], взятую с обратным знаком, т. к. наиболее адаптированные последовательности соответствуют минимумам энергии свободного связывания. Как и в случае функций Royal Road (см. например функцию $R3$ из [17]), искомая последовательность (в ДНК-алфавите) должна включать несколько коротких подпоследовательностей нуклеотидов (называемых сайтами), точно совпадающих с консенсусной последовательностью, или близких к ним, как показано на рис. 1. Значение высоты букв, представляющих последовательность нуклеотидов сайта связывания, будет пояснено ниже в разделе 1.1. Последовательности спейсеров между сайтами являются произвольными, но длина их может иметь значение. Каждый сайт служит мишенью для специфического связывания некоторого транскрипционного фактора (рис. 1), и такое связывание лежит в основе функционирования регуляторного элемента (транскрипционный фактор контролирует активность гена через его регуляторный элемент). Как и в случае с функцией приспособленности Royal Road в ЭА, поиск последовательности нуклеотидов в процедуре SELEX происходит «с нуля». Нахождение каждого нового сайта повышает уровень приспособленности последовательности дискретно. Порядок, в котором формируются сайты связывания, является произвольным.

В современной литературе необходимость развивать моделирование и теорию экспериментальных подходов типа SELEX широко признана, как и признано то, что численные и аналитические подходы из области эволюционных вычислений должны найти эффективное применение и развитие в биоинформатике. Однако одной из серьёзных трудностей при этом является тот факт, что SELEX может проводиться над популяциями молекул астрономически высокой численности (10^8–10^{15}), тогда как в ЭА типично используются популяции численностью в сотни или тысячи особей. На персональной ЭВМ технически возможно реализовать ЭА и с популяцией из миллионов особей, но популяции порядка 10^9 находятся на грани возможностей современных (супер)компьютеров [16]. Именно это в первую очередь обосновывает необходимость дальнейшего развития теории ЭА для нужд биоинженерии, поскольку мы не можем промоделировать такие эксперименты SELEX численно [16, 18].

Настоящая работа нацелена на *прогноз эффективности* процедур SELEX, если имеются предварительные оценки длины последовательности нуклеотидов отдельного энхансера и её статистических свойств. Аналитические методы исследования, используемые в данной работе, позволяют получать оценки для процедур SELEX с популяциями такого размера, для которых численное моделирование не представляется возможным.

Отметим некоторые различия между оптимизацией функций Royal Road и проблемой поиска последовательности нуклеотидов с помощью SELEX-процедуры. Одно из решающих отличий состоит в том, что каждый участок в функции Royal Road («строительный блок» в терминах ЭА) имеет единственную подходящую для него последовательность, в то время как в биологии каждый сайт представляет собой семейство последовательностей, близких в метрике Хэмминга между собой и дающих близкие к минимуму значения энергии свободного связывания. Также важно, что в SELEX на стадии отбора имеется промежуточная «популяция» молекул, которые связываются с белками-мишенями (с вероятностью отбора, монотонно зависящей от их способности к связыванию), а затем выбираются для построения «молекул-потомков», т. е. амплификации в полимеразной цепной реакции. Не всякая процедура отбора, известная в ЭА, имеет такую промежуточную популяцию. Например, турнирная и пропорциональная селекция в классическом генетическом алгоритме не имеют такой популяции, т. к. в этих процедурах родительский генотип выбирается независимо для каждого нового потомка. Наконец, современные процедуры SELEX могут включать в себя до 50 итераций, в то время как число итераций ЭА, как правило, измеряется тысячами.

Практический случай, аналогичный нашему исследованию, можно найти в результатах генетической селекции на основе in vivo SELEX с последовательностями вируса репы *Turnip Crinkle Virus* (TCV) [25], где 28 пар нуклеотидных оснований (п.н.) вирусного регуляторного элемента *motif1-hairpin* были рандомизированы и затем подвергнуты селекции на растениях. Большинство из «победителей» в этом эксперименте содержали до трёх коротких последовательностей (5–7 п.н.), многие из которых обнаруживаются в известных в природе промоторных элементах этого вируса.

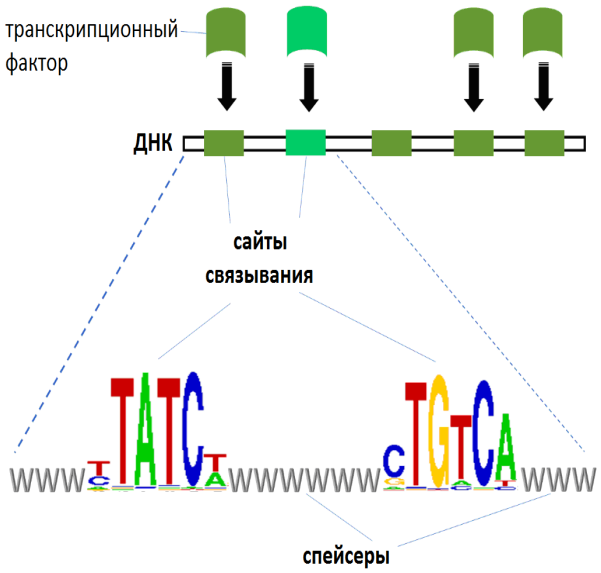

Рис. 1. Пример генно-регуляторного элемента. Энхансер обычно представляет собой кластер сайтов специфического связывания с ДНК-связывающими факторами — активаторами. Каждый сайт представляет собой короткую последовательность нуклеотидов. Сайты обычно разделены спейсерными последовательностями. Сайты связывания представлены прямоугольниками и логотипами последовательностей. W — произвольный нуклеотид.

С одной стороны, верхняя граница [9] на математическое ожидание времени первого достижения целевого участка генотипического пространства для ЭА позволяет оценить сверху среднее время получения достаточно эффективной серии последовательностей (например сайтов связывания транскрипционных факторов) в процедуре SELEX. С другой стороны, теоретический подход из [12] даёт верхнюю оценку средней доли последовательностей ДНК с достаточно высокой приспособленностью на заданной итерации процедуры SELEX. Теоретические границы, найденные с помощью этих подходов, оцениваются в вычислительном эксперименте с использованием функции приспособленности Royal Road, которая рассматривается как упрощённая оценка энергии свободного связывания с транскрипционным фактором FIS в процедуре SELEX, взятая с обратным знаком.

1. Регуляторный элемент гена

Известно, что ген состоит из кодирующей и регуляторной частей. Регуляторная часть просто организованных генов, как правило, включает промотор и энхансер, как показано на рис. 2.

Энхансер — короткий участок ДНК, который может быть специфически связан с белками (факторами транскрипции), чтобы увеличить вероятность того, что начнётся процесс транскрипции определённого гена (обычно факторы, специфически связывающиеся с энхансером, являются активаторами). Как

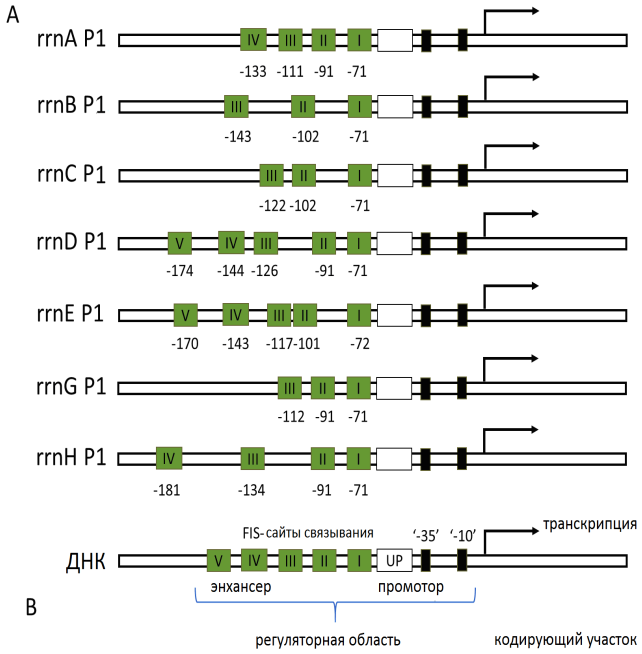

Рис. 2. Оперон рибосомной РНК (rrn) бактерии кишечной палочки *E.coli* как пример прокариотических генов с энхансерами. Каждый из регуляторных регионов состоит из основного промотора и области активации UAR. UAR включает в себя элемент UP и кластер сайтов связывания (3–5 сайтов) для ДНК-связывающего фактора FIS. Расстояния между соседними FIS-сайтами мало отличаются или кратны 20–21 п.н. (A) — схематическое изображение семи регуляторных областей оперона rrn. (B) — схема гена rrn с кластером из пяти сайтов связывания FIS. Отрицательные числа обозначают позиции (в п.н.) от начала транскрипции для различных регуляторных сайтов основного промотора и UAR. Основной промотор состоит из сайтов «-10» и «-35». Стрелка обозначает начало транскрипции и указывает её направление [13].

правило, энхансер представляет собой кластер сайтов для распознавания и связывания факторами транскрипции и другими ДНК-связывающими факторами (как показано на рис. 2). Каждый сайт связывания представляет собой относительно короткую последовательность пар нуклеотидных оснований, сходную или идентичную так называемой консенсусной последовательности для данного ДНК-связывающего фактора. Чем ближе последовательность сайта к консенсусу, тем выше вероятность того, что фактор найдёт и свяжется с ним, и тем ниже будет энергия свободного связывания, и тем выше вероятность того, что ген начнёт транскрибироваться. Здесь мы будем использовать хорошо известный и всесторонне изученный пример прокариотического энхансера для семейства генов рибосомальной РНК *E. coli* (рибосомальный оперон), как показано на рис. 2. Это один из наиболее изученных и относительно просто организованных прокариотических генов с энхансером. Данный регуляторный элемент включает в себя кластер ДНК-связывающих сайтов для транскрипционного фактора FIS. Другой известный пример промотора с кластером сайтов связывания детально

исследован в дрожжах *Yarrowia lipolytica* [7].

1.1. Сайт связывания транскрипционного фактора FIS

Не только последовательности нуклеотидов в сайтах связывания, но также и порядок, и расстояния между соседними сайтами могут иметь решающее значение для правильного функционирования энхансера. Некоторые авторы называют это грамматикой генно-регуляторных элементов [14]. В кластере FIS-сайтов крайне важно, чтобы начальные позиции соседних сайтов разделялись расстояниями, равными или кратными 20–21 п.н. (см. рис. 2). Предполагается, что эта особенность связана с величиной шага двойной спирали ДНК (20–21 п.н. соответствует двум поворотам спирали ДНК). Для нашего рассмотрения мы будем использовать **упрощающее предположение 1:** *расстояние между начальными позициями сайтов связывания кратно константе (большей, чем длина сайта)*. В частности, для сайтов FIS эта константа будет составлять 20 п.н.

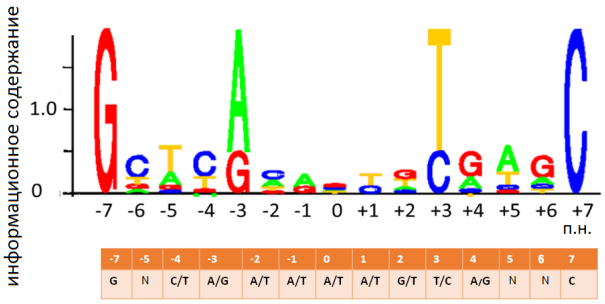

Рис. 3. Логотип консенсусной последовательности для фактора FIS. Положение подсчитывается в п.н. от середины последовательности (сайт является палиндромным) [21].

Рассмотрим более детально сайт FIS. Как уже упоминалось выше, сайты связывания для данного фактора имеют не тождественную последовательность нуклеотидов, а целое семейство похожих последовательностей, которые несколько различаются между собой энергией свободного связывания с фактором. Для представления частот встречаемости нуклеотидных последовательностей с требуемыми свойствами традиционно используется так называемый логотип последовательности. Он состоит из стопки букв в каждой позиции последовательности. Частота встречаемости нуклеотида в каждой позиции представлена относительной высотой соответствующей буквы. Общая высота стопки букв представляет информационное содержание позиции (измеряется в битах), равное исходной энтропии этой позиции минус апостериорное значение энтропии (в смысле Шеннона). Логотип последовательности для сайтов связывания FIS показан на рис. 3. Из этого рисунка видно, что может быть сделано **упрощающее предположение 2**: *каждая позиция сайта связывания FIS имеет один или два подходящих нуклеотида*. В дальнейшем мы будем

рассматривать сайт связывания как активный, если последовательность ДНК имеет подходящие буквы во всех своих позициях.

2. Неэлитарный эволюционный алгоритм с (μ, λ)–селекцией как модель SELEX

2.1. Схема алгоритма

Рассмотрим задачу максимизации:

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

где ϕ — целевая функция (называемая *функцией приспособленности* в литературе по ЭА), $\mathcal{A}$ — алфавит для кодирования решений, например, $\{0, 1\}$ в случае компьютерных систем или $\{A, C, G, T\}$, если (1) рассматривается как модель адаптации в молекулярной генетике.

В области эволюционных алгоритмов задачи вида (1) решаются эвристически посредством имитационного моделирования популяции особей (особями называют последовательности из множества $\mathcal{A}^n$), которые подвергаются случайным мутациям, селекции, а иногда и кроссинговеру (см., например [5]). При этом каждую отдельную позицию в кодированном представлении решений в виде последовательности из $\mathcal{A}^n$ иногда называют «геном». Однако во избежание двусмысленности термин «ген» в настоящей работе будет пониматься только в исходном биологическом смысле, а в контексте эволюционного алгоритма вместо «генов» будем говорить о позициях строки из $\mathcal{A}^n$. Ожидается, что эволюционный процесс будет направлять поиск к оптимальному решению (или локальному оптимуму). Иногда сходимость к оптимуму может быть гарантирована теоретически при времени счёта, стремящемся к бесконечности [19], или могут быть доказаны верхние оценки на математическое ожидание количества пробных решений, просматриваемых до первого достижения оптимума [5, 9].

Популяцию из λ особей на итерации t обозначим через

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

где x^{kt} — особь с номером k в X^t, $k = 1, \ldots, \lambda$.

При мутации некоторое подмножество позиций в строке x изменяется случайным образом. Для любой заданной особи x результат оператора мутации может рассматриваться как случайная величина $\mathrm{Mut}(x) \in \mathcal{X}$ с распределением вероятностей, зависящим от x. Наиболее часто используемый вариант этого оператора, *побитовая мутация*, случайным образом меняет каждую позицию строки x с заданной вероятностью мутации p_m.

В настоящей статье рассматривается только побитовая мутация в предположении, что новое значение для любой мутированной позиции x_i выбирается случайным образом из $\mathcal{A}\backslash\{x_i\}$.

В операторе (μ, λ)–*селекции* родители выбираются равномерно случайным образом среди μ наиболее приспособленных особей в популяции. Общая схема работы ЭА, рассматриваемого в данной статье, имеет следующий вид.

1. Построить начальную популяцию X^0 случайным образом.
2. Для всех t от 0 до $t_{\max} - 1$ выполнять:
   2.1. Для всех k от 1 до λ выполнять:
      2.1.1. Выбрать родительскую особь x из X^t с помощью (μ, λ)-селекции.
      2.1.2. Добавить $x_k^{(t+1)} := \mathrm{Mut}(x)$ в популяцию X^{t+1}.
3. Результат — лучшее из найденных решений $\tilde{x}^t$, т. е. наиболее приспособленная особь из $X^0, \ldots, X^t$.

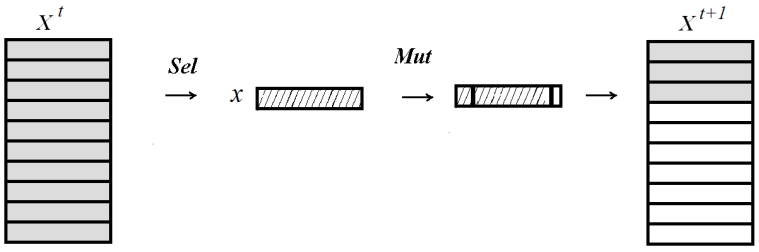

Рис. 4. Итерация t эволюционного алгоритма ЭА.

Одна итерация этого алгоритма показана на рис. 4. В теоретических исследованиях ЭА обычно рассматриваются без критерия остановки. Поэтому также будем считать, что $t_{\max} = \infty$. Описанный ЭА может рассматриваться как упрощённая версия генетического алгоритма (см., например [19]), в котором не используется оператор кроссинговера.

2.2. Моделирование SELEX для регулятроной области гена

2.2.1. Краткое описание SELEX

Процедура SELEX *in vitro* работает следующим образом (см., например [10]). Изначально химически синтезированная «библиотека» разнообразных молекул ДНК заданной длины помещается в раствор вместе с молекулами-мишенями. После достижения равновесного состояния несвязанные молекулы удаляются и молекулярные комплексы мишень/ДНК расщепляются. Освобождённые таким образом последовательности ДНК умножаются ПЦР. При этом возможны некоторые случайные модификации последовательностей, которые могут рассматриваться как мутации, после чего выполняется следующая итерация SELEX. Как правило, этот процесс проходит несколько итераций, а в некоторых случаях — и несколько десятков итераций. Аналогичная процедура может быть применена к молекулам РНК. Процедура SELEX также может быть реализована *in vivo* (в живой природе), или *in silico* (в компьютерном эксперименте), или в их комбинации [15]. В отличие от SELEX *in vitro* процедура SELEX *in vivo* ищет последовательности, которые повышают некоторый функционал качества (понимается как приспособленность) в живых организмах, см.,

например [25]. На завершающей стадии полученные на последней итерации SELEX последовательности нуклеотидов секвенируются и производится анализ кинетики их связывания с молекулой-мишенью.

Стадия мутагенеза в SELEX *in vitro* типично осуществляется экспериментальными процедурами полимеразной цепной реакции (ПЦР). При этом умножение популяции молекул осуществляется полимеразной реакцией и сочетается с внесением точечных замен (ошибок копирования). Точечные мутации неизбежны в процедурах ПЦР, более того, их уровень может быть специально увеличен. ПЦР могут также вызывать и более сложные мутации, но их частота невысока.

Методы направленной эволюции, такие как SELEX *in silico*, направлены на поиск требуемых последовательностей нуклеотидов в вычислительном эксперименте и по своей схеме аналогичны ЭА [23]. В настоящей работе, в отличие от этих методов, ЭА рассматривается как модель процедуры SELEX, а цель анализа состоит в прогнозировании эффективности SELEX *in vitro* или *in vivo*.

Промежуточная «популяция» молекул, которые в SELEX на стадии отбора оказываются связаны с белками-мишенями, может рассматриваться как праобраз подпопуляции из μ особей в операторе (μ, λ)-селекции в ЭА. Совокупность молекул, полученных в результате амплификации в полимеразной цепной реакции, рассматривается как очередная популяция ЭА численности λ.

2.2.2. Семейство псевдо-булевых функций Royal Road

Функции Royal Road были впервые введены для изучения гипотезы об эффективной рекомбинации строительных блоков в операторах кроссинговера генетических алгоритмов [17]. Исходное определение функции Royal Road в [17] было дано для двоичного алфавита $\mathcal{A} = \{0,1\}^n$ в предположении, что задано некоторое множество *схем* S, где под схемой понимается некоторая n-элементная строка символов из алфавита $\mathcal{A} \cup \{\text{“*”}\}$. Строка $x \in \mathcal{A}$ называется *экземпляром* схемы $s \in S$ тогда и только тогда, когда $x_i = s_i$ для всех позиций, где $s_i \neq$ “*”. Предположим, что задан набор положительных весов $c_s,\ s \in S$. В [17] функция Royal Road определяется как

$$\phi(x) := \sum_{s \in S} c_s \cdot [x \text{ является экземпляром } s].$$

Здесь и далее $[\cdot]$ обозначает скобку Иверсона:

$$[P] := \begin{cases} 1, & \text{если } P \text{ истинно;} \\ 0 & \text{– иначе} \end{cases}$$

для любого утверждения P, которое может быть либо истинно, либо ложно.

Одна из наиболее часто используемых версий функций Royal Road (см., например [9]) определена для алфавита $\mathcal{A} = \{0,1\}^n$ в предположении, что задано n/r невзвешенных неперекрывающихся между собой схем с r фиксированными

позициями в каждой (позиции одной схемы при этом называются *блоком*):

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

В настоящей статье определение Royal Road функции обобщается на недвоичные алфавиты и позиции схемы с двумя подходящими буквами с целью моделирования кластера из нескольких сайтов связывания. Без ограничения общности будем считать, что во всех позициях с одним подходящим значением требуется выбор последней буквы $a_{|\mathcal{A}|}$ алфавита $\mathcal{A}$, и такие требования предъявляются к первым r_1 позициям каждого блока. Далее все позиции с двумя подходящими буквами допускают две последние буквы $a_{|\mathcal{A}|-1}, a_{|\mathcal{A}|}$ алфавита, и они занимают оставшиеся r_2 позиций каждого блока, $r = r_1 + r_2$. Мы будем обозначать эту обобщённую функцию Royal Road через $\textsc{RoyalRoad}_{r_1,r_2}(x)$, предполагая, что

$$\textsc{RoyalRoad}_{r_1,r_2}(x) := \sum_{i=0}^{n/r-1} \prod_{j=1}^{r_1} [x_{ir+j} = a_{|\mathcal{A}|}] \prod_{j=r_1+1}^{r} [x_{ir+j} \in \{a_{|\mathcal{A}|}, a_{|\mathcal{A}|-1}\}].$$

Как следует из [7, 13], по мере увеличения числа сайтов связывания фактора транскрипции увеличивается экспрессия соответствующего гена *in vivo*. Эта закономерность была продемонстрирована на примере процедуры SELEX *in vivo* по направленной эволюции энхансера репликациии motif1-hairpin у вируса репы TCV [25]. В наиболее приспособленных решениях содержались серии из трёх последовательностей, сходных с известными «мотивами» в энхансерах вируса TCV в живой природе. Увеличение интенсивности отбора последовательностей по мере увеличения числа сайтов связывания можно ожидать и в процедурах SELEX *in vitro*, т. к. каждый новый сайт связывания увеличивает вероятность связывания участка ДНК с молекулой-мишенью. В связи с этим далее принимается следующее **упрощающее предположение 3**: *критерий отбора молекул, используемый в SELEX, является возрастающей функцией числа действующих сайтов связывания в последовательности нуклеотидов* x. С учётом упрощающих предположений 1–3 приспособленность промотора FIS может быть оценена обобщённой функцией Royal Road с числом блоков от 4 до 6 (каждый блок соответствует отдельному сайту связывания фактора FIS), где $r_1 = 2, r_2 = 6$. Пространство поиска состоит из строк длиной $n = 32, 40$ или 48, содержащих символы из 4-буквенного алфавита $\mathcal{A} = \{A, C, G, T\}$.

В области SELEX *in vitro* близкий к предложенной модели подход представлен в работе [24]. Здесь число блоков равно 2, но при этом, в отличие от функций Royal Road, имеются разные консенсусы у блоков.

3. Теоретический анализ неэлитарного эволюционного алгоритма с (μ, λ)-селекцией

3.1. Верхние границы для доли генотипов с высокой приспособленностью в популяции эволюционного алгоритма

Предположим, что $\phi_0 := \min\{\phi(x) : x \in \mathcal{X}\}$ и задано d линий уровня функции приспособленности, так что $\phi_0 < \phi_1 < \phi_2 \ldots < \phi_d$. Определим $d+1$ подмножеств $\mathcal{X}$

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

Очевидно, что $H_0 = \mathcal{X}$. Для удобства определим $H_{d+1} := \emptyset$. Кроме того, обозначим подмножества линий уровня $A_i := H_i \backslash H_{i+1},\ i = 0, \ldots, d$, которые дают разбиение всего пространства $\mathcal{X}$.

Теперь предположим, что для всех $i = 0, ..., d$ и $j = 1, ..., d$ известны априорные верхние оценки β_{ij} на вероятности мутационного перехода из подмножества A_i в H_j на шаге 2.1.2 алгоритма из раздела 2.1.

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

В дальнейшем $\mathbf{B}$ обозначает матрицу с элементами β_{ij}, $i = 0, ..., d$, $j = 1, ..., d$. Популяция на итерации t может быть представлена *вектором популяции*

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

где $z_i^{(t)} \in [0, 1]$ — доля генотипов из H_i в X^t. Вектор популяции $\mathbf{z}^{(t)}$ является случайным вектором, где $z_i^{(t)} \geqslant z_{i+1}^{(t)}$ для $i = 1, ..., d-1$, так как $H_{i+1} \subseteq H_i$.

Пусть $\Pr\{x^{(t)} \in H_j\}$ — вероятность того, что особь, которая добавляется после селекции и мутации в X^t, имеет генотип в H_j, $j = 0, ..., d$, $t > 0$. Согласно схеме ЭА $\Pr\{x^{(t)} \in H_j\} = \Pr\{x_1^{(t)} \in H_j\} = ... = \Pr\{x_\lambda^{(t)} \in H_j\}$. Следующее утверждение легко доказать (см., например, предложение 1 в [12]).

Утверждение 1. *Для всех* $t > 0$, $i = 1, ..., d$ *имеет место равенство* $\mathbf{E}[z_i^{(t)}] = \Pr\{x^{(t)} \in H_i\}$.

Пусть $P_{\mathrm{ch}}(z_i)$ обозначает вероятность выбора родителя особи из H_i. По определению (μ, λ)–селекции,

$$P_{\mathrm{ch}}(z_i) = \begin{cases} z_i \lambda/\mu, & \text{если } \ z_i \leqslant \mu/\lambda, \\ 1, & \text{иначе.} \end{cases}$$

В результате рассуждений, аналогичных приведённым в разделе 3.1 из [12] (см., например [3]), можно сделать вывод о том, что

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

откуда следуют верхние оценки

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

для ожидаемой доли генотипов с приспособленностью не ниже уровня ϕ_i, где $i = 1, \ldots, d$.

Пусть $((d+1)\times d)$-матрица $\mathbf{B}$ называется *монотонной* тогда и только тогда, когда $\beta_{i-1,j} \leqslant \beta_{ij}$ для всех i, j от 1 до d. Монотонность матрицы $\mathbf{B} = (\beta_{ij})$ означает, что чем больше уровень приспособленности A_i родительского решения, тем больше его нижняя оценка вероятности перехода в любое заданное подмножество H_j, $j = 1, \ldots, d$. Другими словами, это означает, что $\beta_{i,j} - \beta_{i-1,j} \geqslant 0$.

Следующее утверждение доказывается аналогично утверждению 4 из [12] (см. подробнее в [3]).

Утверждение 2. *Если* $\mathbf{B}$ *монотонна, то для всех* $j = 1, \ldots, d$ *выполнено*

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

Итеративным применением неравенства (3) компоненты математического ожидания вектора популяции $\mathbf{E}[\mathbf{z}^{(t)}]$ могут быть ограничены до любого t, начиная с исходного вектора $\mathbf{E}[\mathbf{z}^{(0)}]$, описывающего популяцию X^0. В связи с тем, что на данный момент замкнутой математической формулы для верхней оценки вектора $\mathbf{E}[\mathbf{z}^{(t)}]$ не известно, далее его оценки получаются описанным алгоритмом.

В случае функции $\textsc{RoyalRoad}_{r_1,r_2}$ мы будем использовать термин *1-блок* для любого блока, в котором назначение символов соответствует его схеме (т. е. во всех r_1 позициях, в которых требуется однозначное понимание консенсуса, присвоены требуемые значения, а во всех r_2 позициях, которые допускают два варианта, присвоено одно из двух допустимых значений). В противном случае мы будем называть этот блок *0-блоком*. Вероятности перехода между 0- и 1-состояниями блока под действием мутации описываются следующим образом:

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

$$\Pr(1 \to 1) = \left(1 - p_{\mathrm{m}} + \frac{1}{3}p_{\mathrm{m}}\right)^{r_2}(1 - p_{\mathrm{m}})^{r_1};\quad \Pr(1 \to 0) = 1 - \Pr(1 \to 1).$$

Естественно предположить, что d равно количеству блоков $n/(r_1 + r_2)$, а подмножества $H_0, ..., H_d$ соответствуют линиям уровня $\phi_0 = 0, \phi_1 = 1, ..., \phi_d = d$. Матрица верхних границ $\mathbf{B}$ вычисляется по формуле (20) из [12]. Эта матрица $\mathbf{B}$ удовлетворяет свойству монотонности при условии, что $\frac{2}{3}p_{\mathrm{m}} \leqslant \Pr(1 \to 1)$, например, в случае $r_1 = 2$, $r_2 = 6$ это верно для всех $p_{\mathrm{m}} < 1/4$. Программная реализация описанного алгоритма вычисления верхних оценок для $\mathbf{E}[z_j^{(t)}]$, $j = d-2, d-1, d$, приводится в [4].

3.2. Теоретическая верхняя оценка на среднее время первого получения требуемой последовательности

Предположим, что известна нижняя оценка p_0 для вероятности не снизить уровень приспособленности любого генотипа при мутации, т. е. $\Pr\{\mathrm{Mut}(x) \in H_j \mid x \in A_j\} \geqslant p_0$ для всех $j = 1, \ldots, d-1$. Кроме того, предположим, что для каждого уровня $j = 0, \ldots, d-1$ известны нижние оценки s_j вероятностей улучшающих мутаций, то есть $\Pr\{\mathrm{Mut}(x) \in H_{j+1} \mid x \in A_j\} \geqslant s_j$ для всех $j = 0, \ldots, d-1$. Обозначим $s_* := \min_{j=0,\ldots,d-1} s_j$.

На основе теоремы об уровнях функции приспособленности в [9] получено

Следствие 1. *Если неэлитарный ЭА с (μ, λ)-селекцией применяется при:*

- *достаточно малом отношении μ/λ таком, что $\mu/\lambda \leqslant p_0/(1+\delta)$ при некотором $\delta \in (0, 1]$,*
- *достаточно большом размере популяции таком, что $\lambda \geqslant \left(\frac{4\lambda}{\delta^2\mu}\right) \ln\left(\frac{128(d+1)\lambda}{s_*\delta^2\mu}\right)$,*

то номер итерации t, при котором впервые будет получена особь из H_d, оценивается сверху величиной

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

Легко проверить, что если функция $\textsc{RoyalRoad}_{r_1,r_2}$ используется как функция приспособленности в ЭА, тогда можно предположить:

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

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

Здесь нижняя оценка s_* базируется на «худшем» из возможных сценариев, когда для улучшения приспособленности требуется с помощью мутации получить верные значения во всех позициях одного блока, при этом не изменив содержимое других блоков, если ни одна позиция в рассматриваемом блоке до мутации не соответствовала консенсусу. Чем больше позиций содержится в блоке и чем больше отношение r_1/r_2, тем более заниженной будет оценка (4).

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

Нижняя оценка (5) вероятности улучшения приспособленности в случае, когда j блоков соответствуют консенсусу, получена из того, что имеется $d-j$ способов улучшить ровно один из оставшихся блоков, не изменяя остальных. Эта оценка мало отклоняется от точного значения вероятности улучшения приспособленности, т. к. шансы увеличить два блока сразу очень малы.

Упрощения, использованные при получении (4) и (5), в отличие от упрощающих предположений 1–3, относятся не к построению математической модели, а к вычислению оценок времени достижения искомой последовательности в рамках выбранной модели.

4. Применение оценок из теории ЭА к процедуре SELEX

4.1. Доля оптимальных последовательностей в вычислительном эксперименте и верхняя оценка для неё

Ниже представлены результаты вычислительного эксперимента в сравнении с теоретическими оценками, полученными в разделе 3.1. С этой целью рассмотрим применение ЭА к функции приспособленности $\textsc{RoyalRoad}_{2,6}$, моделирующей SELEX для случая 5 сайтов FIS. Средняя доля особей из подмножеств H_d, H_{d-1} и H_{d-2} представлена на рис. 5. Здесь $\lambda = 10^6$, $\mu = 10^4$, $p_{\mathrm{m}} = 0{,}1$.

Статистика в вычислительном эксперименте набиралась по 1000 прогонам алгоритма, в каждом из которых только одна особь $x_1^{(t)}$ для каждого t проверялась на попадание в целевое подмножество H_d, H_{d-1} или H_{d-2}. Заметим, что $\mathbf{E}[z_i^{(t)}] = \Pr\{x_1^{(t)} \in H_i\}$ по утверждению 1 и, например, для H_d при заданном t мы имеем серию из 1000 испытаний Бернулли, где вероятность успеха $\Pr\{x_1^{(t)} \in H_d\}$ оценивается по результатам описанного эксперимента. Доверительные интервалы уровня 95 % для вероятности успеха в схеме Бернулли были построены с использованием нормального приближения, как описано в [1], гл. 34. На рис. 5 результаты эксперимента показаны пунктирными линиями, а сплошные линии соответствуют верхним оценкам, полученным итеративным применением формулы (3).

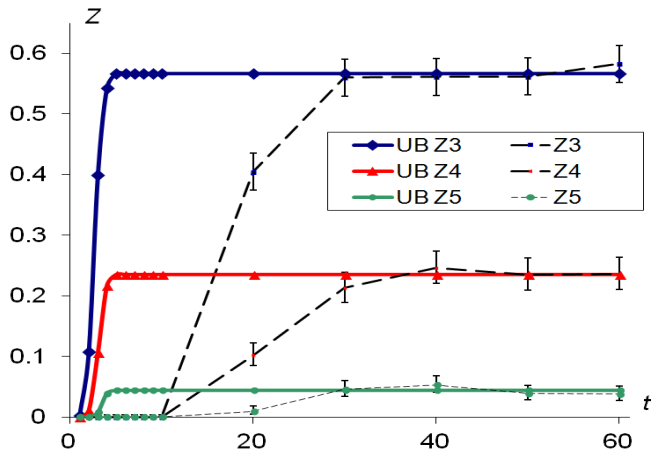

Рис. 5. Сравнение теоретических верхних оценок (UBz_3, UBz_4 и UBz_5) и экспериментального оценивания для средней пропорции оптимальных и субоптимальных особей z_5, z_4 и z_3 в популяции ЭА с $\lambda = 10^6$, $\mu = 10^5$, $p_{\mathrm{m}} = 0,1$, моделирование SELEX для 5 FIS-сайтов (функция приспособленности $\textsc{RoyalRoad}_{2,6}$). Доверительные интервалы рассчитываются с уровнем 95 %.

Как видно из рис. 5, после 30 итераций верхняя оценка из утверждения 2 становится близкой к истинному значению пропорции оптимальных и субоптимальных генотипов (с точностью до ширины доверительного интервала). Общее время счёта в этом вычислительном эксперименте с ЭА на Xeon QuadCore 2.5 ГГц (при его двухпоточном использовании) составило более месяца, в то время как оценка по формуле (3) вычисляется за долю секунды.

4.2. Верхняя оценка математического ожидания времени оптимизации в сравнении с вычислительным экспериментом

Чтобы оценить верхнюю границу ожидаемого времени работы ЭА, полученную в следствии 1, нами проведён вычислительный эксперимент с 1000 независимыми прогонами ЭА с вероятностью мутации $p_{\mathrm{m}} = 0{,}1$. Применение теоретической оценки среднего времени первого получения оптимума к случаю FIS-энхансера с 4 сайтами связывания, смоделированного функцией Royal Road с $r_1 = 2, r_2 = 6$, $n = 32$, в предположении, что в ЭА $\lambda = 10^5$, $\mu = 10^3$, даёт $UB = 3{,}9 \cdot 10^9$, тогда как в эксперименте среднее число итераций до получения оптимума оказалось равным $30{,}75$.

Также было рассмотрено применение теоретической оценки среднего времени первого получения оптимума к случаю с 5 сайтами связывания, при использовании модельной функции Royal Road с $r_1 = 4, r_2 = 2$, $n = 30$ (что соответствует примеру на рис. 1). При настраиваемых параметрах $\lambda = 10^5$, $p_{\mathrm{m}} = 0{,}1$ и $\mu = 500$ оценка даёт $UB = 4{,}85 \cdot 10^7$, тогда как в эксперименте потребовалось $30{,}46$ итераций в среднем.

Общее время счёта в этих экспериментах составило несколько минут, однако при моделировании практически значимых случаев с численностью популяции $\lambda = 10^8$ и более аналогичный вычислительный эксперимент может оказаться проблематичным.

4.3. Обсуждение результатов

Можно предположить, что причина большой переоценки времени поиска, вытекающей из следствия 1, заключается в чрезмерно пессимистическом предположении об изменениях в блоках функции Royal Road при мутации, используемых для вычисления значений s_* и s_j в (4) и (5). Эти вероятностные оценки рассчитываются для наихудшего случая, если предположить, что если сайт не удовлетворяет требованиям связывания, то все позиции в этом сайте отличаются от требуемых значений. Эта гипотеза также согласуется с результатами нашего дополнительного эксперимента, в котором было установлено $r_1 = 0$, $r_2 = 1$ и $n = 40$. При $\lambda = 100$, $p_{\mathrm{m}} = 0{,}025$ и $\mu = 10$ были получены значительно более близкие результаты: $UB = 392{,}6$ и в среднем $45{,}715$ итераций в эксперименте. Таким образом, для улучшения оценки из следствия 1 требуется дополнительный анализ изменчивости в блоках функции Royal Road под действием мутации.

В завершение рассмотрим, как сделанные упрощения могли повлиять на адекватность предложенных оценок.

Упрощающее предположение 1 о равенстве расстояний между начальными позициями всех сайтов связывания исключает некоторые (по-видимому маловероятные) альтернативные варианты возникновения сайтов связывания с другими сдвигами, что, скорее всего, занижает оценки численности наиболее приспособленных последовательностей и увеличивает оценки среднего времени поиска.

Для изучения эффекта от упрощающего предположения 2 (о том, что каждая позиция сайта связывания FIS имеет один или два подходящих нуклеотида) требуется дополнительное исследование. При этом следовало бы учитывать тот факт, что энергия свободного связывания транскрипционного фактора с отдельным сайтом связывания хорошо аппроксимируется аддитивной функцией от последовательности нуклеотидов (см., например [6]), а вероятность связывания в отдельном сайте описывается сигмоидальной функцией от энергии свободного связывания [11].

По упрощающему предположению 3 критерий отбора молекул является возрастающей функцией от числа действующих сайтов связывания. Если данное предположение неприменимо, то возможны как завышенные, так и заниженные оценки при использовании предложенного подхода.

Использование (μ, λ)-селекции в ЭА в качестве модели SELEX также представляет собой упрощение. Детальному исследованию распределениий вероятностей отбора в процедуре SELEX посвящено значительное число публикаций – см., например [11, 16, 22]. Однако теоретический или экспериментальный анализ ЭА с распределениями вероятностей селекции из этих работ представляется чрезвычайно трудоёмким. С целью уточнения модели имеет смысл рассмотреть «релаксированную» версию (μ, λ)-селекции, где решение о включении индивида в промежуточную популяцию численности μ будет не детерминированным, а рандомизированным соответствующим образом. При прочих равных условиях такая «релаксированная» версия селекции будет давать меньшую вероятность выбора наилучших особей в популяции, чем стандартная (μ, λ)-селекция, а значит, можно ожидать и большего времени поиска требуемой последовательности.

5. Заключение

Два подхода из теории эволюционных алгоритмов (оценки доли достаточно приспособленных особей в популяции ЕА и оценки среднего времени первого достижения оптимума) применены для моделирования экспериментальных методов современной биоинженерии. С этой целью теоретические оценки, полученные с помощью обоих подходов, применены к неэлитарному эволюционному алгоритму с (μ, λ)–селекцией и функцией приспособленности Royal Road. Показано, что этот ЭА может рассматриваться как модель процедуры направленной эволюции SELEX для получения генно-регуляторного элемента со многими сайтами связывания. Теоретические оценки сопоставлены с результатами

вычислительных экспериментов.

Проведённый анализ показывает, что теоретические оценки, рассмотренные в разделе 3.1, в принципе могут быть использованы для прогноза числа достаточно эффективных энхансеров после заданного числа итераций процедуры SELEX. Верхние границы на среднее число итераций SELEX до получения требуемой последовательности (раздел 3.2) представляются чрезмерно пессимистичными. Необходимы дальнейшие исследования для улучшения теоретических оценок, чтобы они могли быть применены к процедурам SELEX, состоящим всего из нескольких раундов. Также необходимы дальнейшие исследования для сравнения теоретических прогнозов и вычислительных экспериментов с результатами практических экспериментов.

Благодарности

Работа выполнена при поддержке Российского научного фонда, грант № 17-18-01536.

Приложение

Настоящее приложение содержит пояснения некоторых биологических терминов, использованных в статье.

Амплификация (лат. amplificatio – усиление, увеличение) – процесс образования дополнительных копий молекул ДНК.

Полимеразная цепная реакция (ПЦР) – экспериментальный метод молекулярной биологии, позволяющий добиться значительного увеличения малых концентраций определённых фрагментов нуклеиновой кислоты (ДНК) в биологическом материале (пробе). ПЦР позволяет проводить случайный мутагенез: ошибки в последовательность ДНК вносятся полимеразой в условиях, понижающих её специфичность.

Транскрипционный фактор FIS (Factor for inversion stimulation): белок – транскрипционный фактор, способный специфически связываться со своими сайтами связывания в энхансере и, как следствие, усиливать транскрипцию соответствующего гена (гена-мишени для FIS). «Специфическое связывание» транскрипционного фактора заключается в том, что такой белок способен находить, узнавать и специфически нековалентно связываться с определёнными последовательностями оснований молекулы ДНК. Такие последовательности называются сайтами связывания (это относительно короткие последовательности).

Энхансер (англ. enhancer – усилитель) – небольшой участок ДНК, который после связывания с ним факторов транскрипции стимулирует транскрипцию гена.

Элемент UP (upstream promoter element) – последовательность, расположенная «выше» канонического элемента «-35» у некоторых бактериальных промотеров, усиливающая действие такого промотера.

Литература

1. Крамер Г. Математические методы статистики. М.: Мир, 1975.
2. Нуклеиновые кислоты. От А до Я / Ред. С. Мюллер. М : Бином. Лаборатория знаний, 2013.
3. Рычкова М.А. О математическом ожидании численности особей с высокой приспособленностью в популяции эволюционного алгоритма. Омск : ОмГУ, 2018. URL: `http://iitam.omsk.net.ru/~eremeev/rychkova.pdf` (дата обращения: 24.09.2019).
4. Программа для вычисления верхней оценки вектора популяции. URL: `http://msm.omsu.ru/jrns/jrn53/upper_bound_on_z.zip` (дата обращения: 04.02.2020).
5. Auger A., Doerr B. Theory of Randomized Search Heuristics: Foundations and Recent Developments. River Edge, NJ, USA : World Scientific Publishing Co., Inc., 2011.
6. Benos P.V, Bulyk M.L., Stormo G.D. Additivity in protein–DNA interactions: how good an approximation is it? // Nucleic Acids Res. 2002. V. 30, Issue 20. P. 4442–4451.
7. Blazeck J., Liu L., Redden H., Alper H., Tuning gene expression in *Yarrowia lipolytica* by a hybrid promoter approach // Appl. Environ. Microbiol. 2011. V. 77, No. 22. P. 7905–7914.
8. Borisovsky P., Eremeev A. Comparing evolutionary algorithms to the (1+1)-EA // Theoretical Computer Science. 2008. V. 403, No. 1. P. 33–41.
9. Corus D., Dang D.-C., Eremeev A.V., Lehre P.K. Level-based analysis of genetic algorithms and other search processes // IEEE Transactions on Evolutionary Computation. 2018. V. 22, Issue 5. P. 707–719.
10. Darmostuk M., Rimpelova S., Gbelcova H., Ruml T. Current approaches in SELEX: An update to aptamer selection technology // Biotechnology Advances. 2015. V. 33. P. 1141–1161.
11. Djordjevic M., Sengupta A.M. Quantitative modeling and data analysis of SELEX experiments // Physical Biology. 2005. V. 3, No. 1. P. 13–28.
12. Eremeev A.V. On proportions of fit individuals in population of genetic algorithm with tournament selection // Evolutionary Computation. 2018. V. 26, No. 2. P. 269–297.
13. Hirvonen C.A., Ross W., Wozniak C.E., Marasco E., Anthony J.R., Aiyar S.E., Newburn V.H., Gourse R.L. Contributions of UP elements and the transcription factor FIS to expression from the seven rrn P1 promoters in *Escherichia coli* // J. Bacteriol. 2001. V. 183, No. 21. P. 6305–6314.
14. Gertz J., Siggia E.D., Cohen B.A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters // Nature. 2009. V. 457. P. 215—218.
15. Kinghorn A.B., Fraser L.A., Lang S., Shiu S., Tanner J.A. Aptamer bioinformatics // International journal of molecular sciences. 2017. V. 18, No. 12. P. 2516.
16. Lee Y.-G., McKay B., Kim K.-I., Kim D.-K., Hoai N.-X. Investigating vesicular selection: A selection operator in *in vitro* evolution // Applied Soft Computing. 2011. V. 11, Issue 8. P. 5528–5550.
17. Mitchell M., Forrest S., Holland J.H. The royal road for genetic algorithms: fitness landscapes and GA performance // Proceedings of the 1st European Conf. on Artificial Life. Cambridge, MA: MIT Press, 1992. P. 245–254.

18. Oh I.S., Lee Y., McKay R. Simulating chemical evolution // Proceedings of 2011 IEEE Congress of Evolutionary Computation. New Orleans, LA, 2011. P. 2717–2724.
19. Rudolph G. Finite Markov chain results in evolutionary computation: A tour d'horizon // Fundamental Informaticae. 1998. V. 35, No. 1–4. P. 67–89.
20. Spirov A., Holloway D. New approaches to designing genes by evolution in the computer // Real-World Applications of Genetic Algorithms / O. Roeva. London: InTech, 2012. P. 235–260.
21. Shao Y., Feldman-Cohen L.S., Osuna R. Functional characterization of the *Escherichia coli* FIS-DNA binding sequence // J. Mol. Biol. 2008. V. 376, No. 3. P. 771–85.
22. Spill F., Weinstein Z.B., Shemirani A.I., Ho N., Desai D., Zaman M.H. Controlling uncertainty in aptamer selection // Proceedings of the National Academy of Sciences. 2016. V. 113, No. 43. P. 12076–12081.
23. Voigt C.A., Martinez C., Wang Z.G., Mayo S.L., Arnold F.H. Protein building blocks preserved by recombination // Nat Struct Biol. 2002. V. 9. P. 553-558.
24. Wu L., Curran J.F. An allosteric synthetic DNA // Nucleic Acids Research. 1999. V. 27, No. 6. P. 1512-–1516.
25. Zhang G., Simon A.E. A multifunctional Turnip Crinkle Virus replication enhancer revealed by *in vivo* functional SELEX // J. Mol. Biol. 2003. V. 326. P. 35—48.

ESTIMATES FROM EVOLUTIONARY ALGORITHMS THEORY APPLIED TO DIRECTED EVOLUTION

A.V. Eremeev[1,2]
Dr. Sci. (Phys.-Math.), Associate Professor, e-mail: eremeev@ofim.oscsbras.ru
A.V. Spirov[2,3]
Ph. D. (Biology) , e-mail: sspirov@yandex.ru

[1]Sobolev Institute of Mathematics SB RAS, Omsk
[2]The Institute of Scientific Information, for Social Sciences RAS, Moscow
[3]I.M. Sechenov Institute of Evolutionary Physiology and Biochemistry RAS, St. Petersburg

Abstract. The field of evolutionary computation 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 the theory of evolutionary computation may be transferred back into biology. It has been noted that the non-elitist evolutionary algorithms optimizing Royal Road fitness functions may be considered as models of evolutionary search for the synthetic enhancer sequences "from scratch". 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 in order to upper-bound the expected time to finding a sufficiently fit series of motifs in a SELEX procedure. On the other hand, using the theory of evolutionary computation, we propose an upper bound on the expected proportion of the DNA sequences with sufficiently high fitness at a given round 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.

Keywords: runtime analysis, SELEX procedure, Royal Road function, enhancer.

References

1. Kramer G. Matematicheskie metody statistiki. Moscow, Mir Publ., 1975. (in Russian)
2. Nucleic Acids from A to Z: A Concise Encyclopedia / Ed. by S. Müller. Wiley-VCH Verlag GmbH & Co. KGaA, 2008.
3. Rychkova M.A. O matematicheskom ozhidanii chislennosti osobei s vysokoi prisposoblennost'yu v populyatsii evolyutsionnogo algoritma. Omsk, OmGU Publ., 2018. URL: `http://iitam.omsk.net.ru/~eremeev/rychkova.pdf` (24.09.2019). (in Russian)
4. Supplementary materials – program for computing the upper bound on population vector, 2020. URL: `http://iitam.omsk.net.ru/~eremeev/upper_bound_on_z.zip` (24.09.2019).
5. Auger A., Doerr B. Theory of Randomized Search Heuristics: Foundations and Recent Developments. River Edge, NJ, USA, World Scientific Publishing Co., Inc., 2011.
6. Benos P.V, Bulyk M.L., and Stormo G.D. Additivity in protein–DNA interactions: how good an approximation is it? Nucleic Acids Res., 2002, vol. 30, issue 20, pp. 4442–4451.
7. Blazeck J., Liu L., Redden H., and Alper H. Tuning gene expression in *Yarrowia lipolytica* by a hybrid promoter approac. Appl. Environ. Microbiol., 2011, vol. 77, no. 22, pp. 7905–7914.
8. Borisovsky P. and Eremeev A. Comparing evolutionary algorithms to the (1+1)-EA. Theoretical Computer Science, 2008, vol. 403, no. 1, pp. 33–41.
9. Corus D., Dang D.C., Eremeev A.V., and Lehre P.K. Level-based analysis of genetic algorithms and other search processes. IEEE Transactions on Evolutionary Computation, 2018, vol. 22, issue 5, pp. 707–719.
10. Darmostuk M., Rimpelova S., Gbelcova H., and Ruml T. Current approaches in SELEX: An update to aptamer selection technology. Biotechnology Advances, 2015, vol. 33, pp. 1141–1161.
11. Djordjevic M., Sengupta A.M. Quantitative modeling and data analysis of SELEX experiments. Physical Biology, 2005, vol. 3, no. 1, pp. 13–28.
12. Eremeev A.V. On proportions of fit individuals in population of genetic algorithm with tournament selection. Evolutionary Computation, 2018, vol. 26, no. 2, pp. 269–297.
13. Hirvonen C.A., Ross W., Wozniak C.E., Marasco E., Anthony J.R., Aiyar S.E., Newburn V.H., and Gourse R.L. Contributions of UP elements and the transcription factor FIS to expression from the seven rrn P1 promoters in *Escherichia coli*. J. Bacteriol, 2001, vol. 183, no. 21, pp. 6305–6314.
14. Gertz J., Siggia E.D., and Cohen B.A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature, 2009, vol. 457, pp. 215–218.
15. Kinghorn A.B., Fraser L.A., Lang S., Shiu S., and Tanner J.A. Aptamer bioinformatics. International journal of molecular sciences, 2017, vol. 18, no. 12, pp. 2516.

16. Lee Y.-G., McKay B., Kim K.-I., Kim D.-K., and Hoai N.-X. Investigating vesicular selection: A selection operator in *in vitro* evolution. Applied Soft Computing, 2011, vol. 11, issue 8, pp. 5528–5550.
17. Mitchell M., Forrest S., and Holland J.H. The royal road for genetic algorithms: fitness landscapes and GA performance. Proceedings of the 1st European Conf. on Artificial Life, Cambridge, MA, MIT Press, 1992, pp. 245–254.
18. Oh I.S., Lee Y., and McKay R. Simulating chemical evolution. Proceedings of 2011 IEEE Congress of Evolutionary Computation, New Orleans, LA, 2011, pp. 2717–2724.
19. Rudolph G. Finite Markov chain results in evolutionary computation: A tour d'horizon. Fundamental Informaticae, 1998, vol. 35, no. 1–4, pp. 67–89.
20. Spirov A. and Holloway D. New approaches to designing genes by evolution in the computer. Real-World Applications of Genetic Algorithms, O. Roeva, London, InTech, 2012, pp. 235–260.
21. Shao Y., Feldman-Cohen L.S., and Osuna R. Functional characterization of the *Escherichia coli* FIS-DNA binding sequence. J. Mol. Biol., 2008, vol. 376, no. 3, pp. 771–85.
22. Spill F., Weinstein Z.B., Shemirani A.I., Ho N., Desai D., and Zaman M.H. Controlling uncertainty in aptamer selection. Proceedings of the National Academy of Sciences, 2016, vol. 113, no. 43, pp. 12076–12081.
23. Voigt C.A., Martinez C., Wang Z.G., Mayo S.L., and Arnold F.H. Protein building blocks preserved by recombination. Nat Struct Biol., 2002, vol. 9, pp. 553–558.
24. Wu L. and Curran J.F. An allosteric synthetic DNA. Nucleic Acids Research, 99, vol. 27, no. 6, pp. 1512–1516.
25. Zhang G. and Simon A.E. A multifunctional Turnip Crinkle Virus replication enhancer revealed by *in vivo* functional SELEX. J. Mol. Biol., 2003, vol. 326, pp. 35–48.

Дата поступления в редакцию: 28.09.2019