Вавиловский журнал генетики и селекции. 2015;19(6):661-667
Doi 10.18699/VJ15.087

Использование графических ускорителей для выявления функциональных сигналов в регуляторных районах генов прокариот

О.В. Вишневский[1, 2], А.В. Бочарников[2], А.А. Романенко[2]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
2 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия

Различные методы выявления значимых контекстных сигналов широко используются для поиска сайтов связывания транскрипционных факторов и выявления структурно-функ-циональной организации регуляторных районов генов. Такие методы не требуют ни предварительного выравнивания выборки анализируемых последовательностей, ни экспери-ментальной информации о точном расположении сайтов связывания транскрипционных факторов. Широкое распро-странение получили методы поиска контекстных сигналов, основанные на выявлении вырожденных олигонуклеотидных мотивов, записанных в 15-буквенном коде номенклатуры IUPAC (International Union of Pure and Applied Chemistry). Существенной сложностью использования вырожденных мотивов является их огромное разнообразие, что заставляет исследователей применять различные эвристические подходы, не гарантирующие нахождение наиболее значимого сигнала. Появление высокопроизводительных вычислитель-ных систем, основанных на использовании графических ускорителей, сделало возможным применение точных полнопереборных методов для выявления значимых мотивов. Нами разработана новая система выявления значимых вырожденных олигонуклеотидных мотивов заданной длины в регуляторных районах генов, основанная на использовании широко распространенных графических ускорителей и обеспечивающая поиск сигнала с наибольшей значимостью. Показана высокая эффективность использования графических ускорителей (GPU) в сравнении с расчетами на центральном процессоре (CPU). С использованием предложенного подхода проанализированы регуляторные районы генов B. subtilis, E. coli, H. pylori, M. gallisepticum, M. genitalium и M. pneumoniae. Для каждого вида прокариот были выявлены наборы вырожденных мотивов и проведена их классификация на основе сходства с сайтами связывания транскрипционных факторов E. coli.

Ключевые слова: вырожденный олигонуклеотидный мотив; регуляция транскрипции; регуляция трансляции; CUDA; графические ускорители.

HOW TO CITE THIS ARTICLE?
Vishnevsky o.V., Bocharnikov A.V., r omanenko A.A. The use of graphics accelerators to detect functional signals in the regulatory regions of prokaryotic genes. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):661-667. Doi 10.18699/V J15.087

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?
Вишневский О.В., Бочарников А.В., Романенко А.А. Использование графических ускорителей для выявления функциональных сигналов в регуляторных районах генов прокариот. Вавиловский журнал генетики и селекции. 2015;19(6):661-667. Doi 10.18699/VJ15.087

The use of graphics accelerators to detect functional signals in the regulatory regions of prokaryotic genes

O.V. Vishnevsky[1, 2], A.V. Bocharnikov[2], A.A. Romanenko[2]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Novosibirsk State University, Novosibirsk, Russia

Various methods for identification of significant contextual signals are widely used to search for transcription factor binding sites and to identify the structural and functional organization of regulatory regions. These methods do not require any pre-alignment of the sample sequences analyzed or experimental information about the exact location of transcription factor binding sites. Methods of searching for contextual signals, based on the identi-fication of degenerate oligonucleotide motives recorded in the 15-letter IUPAC code have become widespread. An essential problem with degenerate motifs is their great diversity, which makes the researchers apply heuristics which do not guarantee that the most significant signal will be found. The development of high-performance computing systems based on the use of graphics cards has made it possible to use the exact exhaustive methods to identify significant motifs. We have developed a new system for identifying significant degenerate oligonucleo- tide motifs of a given length in the regulatory regions based on the use of widespread graphics cards that provides a search for the signal with the greatest signifi-cance. High efficiency of the GPU compared with CPU was demonstrated. Using the proposed approach, we ana-lyzed the regulatory regions of B. subtilis, E. coli, H. pylori, M. gallisepticum, M. genitalium and M. pneumoniae genes. Sets of degenerate motifs have been identified for each species of prokaryotes. They were classified on the basis of similarity with the transcription factor binding sites of E. coli.

Key words: degenerated oligonucleotide motif; transcription regulation; translation regulation; CUDA; GPU.

Received 22.09.2015
Accepted for publication 16.10.2015

e-mail: oleg@bionet.nsc.ru

Выявление функциональных сигналов в регуляторных районах генов является важной задачей современной биоинформатики и необходимо как для понимания базовых механизмов регуляции транскрипции и трансляции, так и для определения специфических особенностей функционирования регуляторных районов.

Существующие методы, как правило, основываются на использовании ранее полученной экспериментальной информации о локализации сайтов связывания транскрипционных факторов (ССТФ), собранной в специализированных базах данных (Kolchanov et al., 2002; Matys et al., 2006; Portales-Casamar et al., 2010), или на сравнении анализируемых последовательностей и выявлении в них относительно схожих участков (Vishnevsky, Kolchanov, 2005).

Гигантский рост баз данных в связи с появлением методов высокопроизводительного секвенирования ДНК (Elnitski et al., 2003) и огромное разнообразие регуляторных сигналов требуют разработки новых высокопроизводительных компьютерных методов для их выявления и анализа.

Наибольшее распространение получили методы выявления значимых контекстных сигналов, основанные на анализе частот *l*-плетов (*l*-letter substrings) (Pesole et al., 2000); деревьев суффиксов (suffix trees) (Marsan et al., 2000); поиске максимальной клики в графе, построенном на основе дистанции редактирования (edit distance) *l*-плетов (Pevzner, Sze, 2000); локальном множественном выравнивании, основанном на «жадном» (greedy) алгоритме (Hertz, Stormo, 1999); методе EM (Expectation–Maximization) (Grundy et al., 1996) и стохастическом отборе (stochastic sampling) (Lawrence et al., 1993). Результатом работы таких методов, как правило, являются позиционные весовые матрицы или олигонуклеотидные мотивы, записанные в 4- (A, T, G, C) или 15-буквенном коде IUPAC (A, T, G, C, R = G/A, Y = T/C, M = A/C, K = G/T, W = A/T, S = G/C, B = T/G/C, V = A/G/C, H = A/T/C, D = A/T/G, N = A/T/G/C).

Применение олигонуклеотидных мотивов является одним из наиболее ранних и широко распространенных подходов, но их использование затруднено гигантским разнообразием возможных вариантов. Так, мотив длиной 8, записанный в 15-буквенном коде IUPAC, имеет $15^8 \sim 2{,}5 \times 10^9$ различных вариантов записи в 4-буквенном коде. Это приводит к необходимости использования различных эвристических подходов (Mrázek et al., 2002; Vishnevsky, Kolchanov, 2005). Однако такие подходы не гарантируют достижения глобального минимума и нахождения наиболее представленного и значимого мотива. Для применения точного полнопереборного метода необходимо использование высокопроизводительных вычислительных систем, основанных на массивно-параллельных алгоритмах.

На данный момент наиболее популярными системами для параллельных вычислений в биоинформатике являются:

- FPGA (Field-Programmable Gate Array) – программируемая логическая интегральная схема (ПЛИС) (Baker, Prasanna, 2006; Yooseph et al., 2007).
- Cell/BE (Cell Broadband Engine) – процессор, используемый в Sony PlayStation 3 (Fomin, Alemasov, 2009).
- GPU (Graphics Processing Unit) – графический ускоритель (Manavski, Valle, 2008; Sukhwani, Herbordt, 2009).
- CPU (Central Processing Unit) – универсальный центральный процессор и компьютерные кластеры на его основе (Vishnevsky, Kolchanov, 2005).

Хотя FPGA и кластеры CPU обладают очень высокими вычислительными мощностями, они до сих пор являются очень дорогим и не для всех доступным решением. Cell/BE более доступны, но не обладают необходимой вычислительной мощью. По соотношению цена/качество наиболее приемлемым и распространенным решением являются вычислительные платформы, основанные на графических ускорителях. Поэтому для разработки метода выявления вырожденных олигонуклеотидных мотивов мы использовали графические ускорители и технологию CUDA.

Графические ускорители стали одной из наиболее распространенных компонент современных компьютеров. Изначально разработанные для обработки компьютерной графики, они стали одним из наиболее мощных инструментов компьютерного анализа, благодаря своей относительно небольшой стоимости и высокой производительности. Это было достигнуто за счет фундаментальных изменений в архитектуре чипа. Она была оптимизирована для одновременного параллельного расчета огромного числа относительно простых операций. За счет уменьшения размера кэш-памяти на чипе, упрощения арифметико-логических устройств (Arithmetically-Logic Unit, ALU), относительного уменьшения количества ALU, работающих с двойной точностью, было значительно увеличено общее количество ALU на чипе. В результате чип GPU может содержать сотни вычислительных ядер, одновременно обрабатывающих поток данных.

Одной из проблем, возникающих при работе с большим количеством вычислительных потоков, является синхронизация. Поэтому все ALU на GPU, именуемые скалярными процессорами (Scalar Processor, SP), собраны в группы, называемые потоковыми мультипроцессорами (Streaming Multiprocessor, SM). Все программные нити внутри одного SM имеют доступ к общей разделяемой (shared) памяти. Доступ к ней осуществляется в сотни раз быстрее, чем к глобальной памяти GPU.

Фактически графические ускорители являются процессорами, построенными на основе технологии SIMD (Single Instruction Multiple Data). Это означает, к сожалению, что не все алгоритмы могут быть эффективно реализованы с использованием GPU.

Несмотря на высокую производительность графических ускорителей, долгое время использование их для решения задач обработки неграфических данных было затруднено. Применение платформ OpenGL и DirectX API требовало переформулирования математических алгоритмов в терминах обработки графики. Появление технологии CUDA (Compute Unified Device Architecture), предложенной NVIDIA, позволило существенно упростить разработку программ для графических ускорителей. CUDA является расширением языка C, позволяющим создавать многопотоковые приложения для устройств, поддерживающих эту архитектуру (NVIDIA CUDA programming guide 3.2. http://developer.download.nvidia.com/).

Table 1. Binary representation of the 15-letter IUPAc code for letters in a motif)

l etters of the IUPAc code		A	T	G	c	r	Y	M	K	w	S	B	H	V	D	n
Transcription of letters in the IUPAc code		**a**	**t**	**G**	**C**	**G/a**	**t /C**	**a /C**	**G/t**	**a /t**	**C/G**	**!a**	**!G**	**!t**	**!C**	**n**
n ucleotide	**a**	1	0	0	0	1	0	1	0	1	0	0	1	1	1	1
	t	0	1	0	0	0	1	0	1	1	0	1	1	0	1	1
	G	0	0	1	0	1	0	0	1	0	1	1	0	1	1	1
	C	0	0	0	1	0	1	1	0	0	1	1	1	1	0	1
c ode		1	2	4	8	5	10	9	6	3	12	14	11	13	7	15

Поскольку программа на CUDA запускается на GPU, процессоре с SIMD архитектурой, она должна содержать ядро, состоящее из вычислительных операций, одновременно рассчитываемых на GPU в виде множественных нитей (threads). Каждая нить имеет уникальный идентификатор, который может быть использован для получения соответствующих данных из памяти для обработки. Нити объединены в блоки (blocks), в свою очередь объединенные в сеть (grid). Это сделано для облегчения программной реорганизации нитей на структуру обрабатываемых данных. Например, нити могут быть организованы в виде как одномерной, так и двух- и трехмерной решетки. Нити, объединенные в блок, могут взаимодействовать друг с другом, синхронизироваться и использовать одну разделяемую память. Нити в блоке выполняются на мультипроцессоре SM одновременно группами, называемыми *warp*.

На основе технологии CUDA нами разработан новый метод выявления значимых вырожденных олигонуклеотидных мотивов, записанных в 15-буквенном коде IUPAC, позволяющий использовать в расчетах широко распространенные графические ускорители. С использованием предложенного подхода были проанализированы регуляторные районы генов *B. subtilis*, *E. coli*, *H. pylori*, *M. gallisepticum*, *M. genitalium* и *M. pneumoniae*. Для каждого вида прокариот были выявлены наборы вырожденных мотивов и проведена их классификация на основе сходства с сайтами связывания транскрипционных факторов *E. coli*. Мотивы, полученные таким образом, могут являться целью дальнейшего экспериментального анализа.

Материалы и методы

Нами предложен алгоритм расчета представленности мотива M длины l в выборке D, состоящей из N_{seq} последовательностей длины L_{seq}, основанный на оценке соответствия мотива M каждой из $N_{seq} \cdot (L_{seq} - l + 1)$ позиций выборки D. Для этого каждый символ мотива в 15-буквенном коде IUPAC записывается в виде целого числа от 1 до 15 (табл. 1), а каждый нуклеотид выборки анализируемых последовательностей D записывается в виде целого числа от 0 до 3 (табл. 2).

В этом случае соответствие между мотивом M длины l и районом $[i; i+l]$ анализируемой последовательности, записанной 4-буквенном коде, может быть оценено с помощью операции побитового сдвига вправо. При этом если буквы в позиции мотива M и анализируемой нуклеотидной последовательности соответствуют друг другу, то побитовый сдвиг вправо бинарного представления символа мотива M (табл. 1) на число, соответствующее бинарному представлению нуклеотида (табл. 2), выдаст 1, в противном случае – 0. Таким образом, если все символы мотива и сравниваемого участка последовательности соответствуют друг другу, произведение результатов побитового сдвига для всех позиций будет равным 1. Подобный подход позволяет существенно ускорить оценку соответствия мотива и нуклеотидной последовательности.

Table 2. Binary representation of 15-letter IUPAc code for letters in a nucleotide sequence

n ucleotide	**a**	**t**	**G**	**C**
c ode	0	1	2	3

Для оценки представленности всех 15^l возможных мотивов все рассматриваемые мотивы разбиваются на группы, равные количеству потоков в потоковом блоке, а каждый потоковый блок обрабатывает свою нуклеотидную последовательность. При этом все нуклеотидные последовательности анализируемой выборки D размещались в текстурной памяти, что позволило существенно ускорить доступ к этим последовательностям. Последовательность, с которой работает блок, копировалась в разделяемую память, поскольку доступ к ней существенно быстрее, чем к глобальной памяти.

Загрузка последовательностей из текстурной памяти производится всеми потоками блока. Размер разделяемой памяти на мультипроцессоре ограничивает длину последовательностей в ~14 тыс. нуклеотидов, что достаточно для решения большинства задач по анализу регуляторных районов генов. Сократить количество итераций обращения к текстурной памяти можно за счет использования упакованных типов данных. В нашем случае вместо char (один 8-битный символ) использовался uchar4 (четыре 8-битных символа), то есть, например, для загрузки одной последовательности длины $L = 2\,000$ нуклеотидов 512 потоками нам потребуется четыре итерации обращений к текстурной памяти для char и только одна – для uchar4.

Затем каждый поток в блоке проверяет встречаемость одного мотива в одной последовательности нуклеотидов и запоминает результат в глобальной памяти. В случае использования запакованных типов данных (uchar4) каждый поток может обрабатывать одновременно четыре последо-

Table 3. Properties of samples from regulatory regions of prokaryotic genes

Species	N_{seq}	N_{mot}	N_{TFBS}
B. subtilis	4 109	388	27
E. coli	4 173	334	42
H. pylori	1 565	454	35
M. gallisepticum	725	486	24
M. genitalium	212	452	28
M. pneumoniae	687	423	35

вательности. После этого запускается другое ядро на GPU, которое вычисляет встречаемость обработанной порции мотивов во всех последовательностях нуклеотидов. Пока на GPU идет обработка мотивов, на CPU готовится следующая их порция, и процесс повторяется.

После того как процесс расчета представленности для всего множества мотивов проведен, производятся оценка значимости полученных мотивов согласно биномиальному критерию (Vishnevsky, Kolchanov, 2005) и расчет их представленности в выборке случайных последовательностей. Случайная выборка генерировалась с частотами нуклеотидов, соответствующими частотам нуклеотидов в анализируемой выборке. Мотивы, не удовлетворяющие граничным критериям, удалялись из рассмотрения, а среди оставшихся мотивов выбирался наиболее значимый. Позиции этого мотива маскировались в выборке анализируемых последовательностей, и процесс оценки значимости оставшихся мотивов производился заново. Затем среди найденных мотивов выявлялся следующий по значимости мотив, производилась маскировка позиций его расположения в выборке последовательностей, и цикл поиска значимых мотивов повторялся до тех пор, пока в анализе оставались мотивы, удовлетворяющие граничным критериям.

Предложенный алгоритм был реализован в виде компьютерной программы на языке CUDA. Программа может работать в операционных системах Windows и Linux и позволяет оценивать представленность в заданной выборке нуклеотидных последовательностей всех вырожденных олигонуклеотидных мотивов длиной 8, записанных в 15-буквенном коде IUPAC. Программа обладает интерфейсом, в котором пользователь может задать границы окна в анализируемой выборке, граничный уровень значимости и представленности в выборке случайных последовательностей. Можно указать такие параметры случайной выборки, как количество последовательностей в ней и необходимость использования частот нуклеотидов, характерных для анализируемой выборки последовательностей. Поиск может проводиться как в прямой, так и комплементарной цепях ДНК. На вход программы подается выборка нуклеотидных последовательностей, записанных в формате FASTA. На выходе – набор полученных вырожденных олигонуклеотидных мотивов, удовлетворяющих заданным критериям.

В качестве примера использования предложенного метода нами проведен поиск вырожденных олигонуклеотидных мотивов длиной 8 в регуляторных районах генов шести видов прокариот. Для этого из базы данных GenBank (Benson et al., 2013) были экстрагированы выборки [–100; +25] районов относительно старта трансляции для *B. subtilis*, *E. coli*, *H. pylori*, *M. gallisepticum*, *M. genitalium* и *M. pneumoniae*. Поиск проводился в трех окнах шириной 50 нуклеотидов: [–100; –50], [–75; –25] и [–25; +25] относительно старта трансляции. Хорошо видно (табл. 3), что построенные выборки значительно различались по количеству содержащихся в них последовательностей N_{seq} – от 212 для *M. genitalium* до 4 173 для *E. coli*.

В каждой из построенных выборок проводился поиск значимых вырожденных олигонуклеотидных мотивов. В качестве достоверных рассматривались мотивы, чья представленность в выборке регуляторных районов превышала 10 % (для отбора относительно слабо вырожденных мотивов), а вероятность наблюдения по случайным причинам не превышала 10^{-8}. Затем была проведена классификация выявленных мотивов с использованием базы данных сайтов связывания транскрипционных факторов *E. coli* (Osada et al., 2004).

результаты и обсуждение

Оценка производительности работы программы на разных вычислительных устройствах

Для того чтобы получить оценки производительности работы программы, не зависящие ни от длины и количества анализируемых последовательностей, ни от длины мотивов, будем измерять производительность в количестве сравнений позиций мотивов с позициями выборки последовательностей за единицу времени. Для набора мотивов M и выборки последовательностей D производительность G вычисляется следующим образом:

$$G = \frac{|M| \cdot |D|}{t \cdot 10^9} = \frac{l \cdot N_{mot} \cdot N_{seq} \cdot (L_{seq} - l + 1)}{t \cdot 10^9},$$

где $|M|$ – суммарное количество всех букв в наборе вырожденных олигонуклеотидных мотивов, $|D|$ – суммарное количество всех позиций (в выборке последовательностей), в которых возможно сравнение с мотивами, t – время работы в секундах, N_{seq} – количество последовательностей в выборке, L_{seq} – длина последовательностей, l – длина мотивов.

С помощью предложенной меры мы провели оценку эффективности использования разработанного метода на различных графических ускорителях и CPU. В качестве CPU использовался четырехъядерный процессор i7-950

с частотой 3 ГГц. Распараллеливание расчетов между различными ядрами CPU проводилось с использованием библиотеки POSIX Threads. На рис. 1 приведены сравнительные оценки производительности *G* на различных вычислительных платформах на выборке из 10 последовательностей в зависимости от их длины.

Из рис. 1 хорошо видно, что наиболее производительным решением является сдвоенная Tesla C1060, которая суммарно обладает 480 процессорами. Отметим, что все графические ускорители имеют худшую производительность при малой длине последовательности. Это можно объяснить тем, что внутренний цикл становится достаточно коротким и запуск вычислительного ядра происходит чаще. Tesla C2050 обладает лучшей производительностью при длине последовательности 32 символа. При длине последовательностей от 50 нуклеотидов производительность всех GPU стабилизируется. Показано, что метод хорошо шкалируется между различными вычислительными устройствами и его производительность растет с ростом выборки. В среднем отставание CPU i7-950 от сдвоенной GPU C1060 составляет 18 раз. Кроме того, при наличии 4 ядер и частоты 3ГГц i7-950 в 1,5 раза быстрее, чем бюджетная видеокарта geForce210 с 16 процессорами и частотой 1,4 ГГц (данные не приведены). Производительность одного графического ускорителя Tesla C2050 примерно в 70 раз выше, чем производительность одного ядра CPU i7-950.

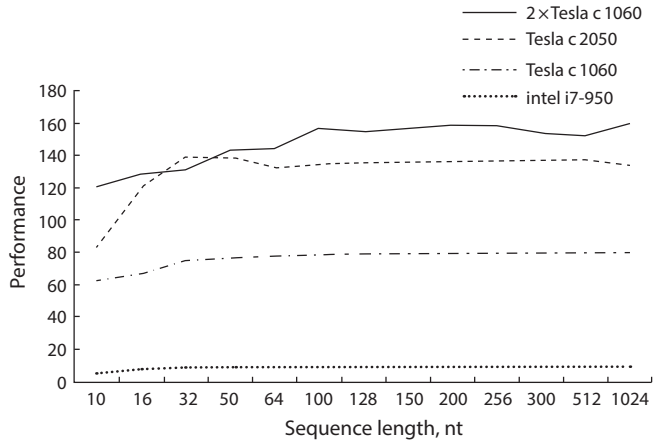

Fig. 1. Performance evaluation *G* of the program identifying degenerate oligonucleotide motifs vs. sequence length. The performance was evaluated on various devices.

Выявление вырожденных олигонуклеотидных мотивов в регуляторных районах генов прокариот

Для каждой из построенных выборок регуляторных районов были получены сотни вырожденных мотивов N_{mot} (табл. 3). Их количество варьировало от 334 для *E. coli* до 486 для *M. gallisepticum*. Отметим, что количество выявляемых мотивов N_{mot} и размер анализируемой выборки N_{seq} не демонстрируют явной корреляции между собой. В табл. 4 приведен пример мотивов, найденных в регуляторных районах генов *M. genitalium*. Например, мотив **AAAYWAAA** представлен в 11 % ($F = 0{,}11$) анализируемых последовательностей, в 3 % ($Q = 0{,}03$) случайных последовательностей, а вероятность его наблюдения по случайным причинам в анализируемой выборке – 10^{-989} ($-P = 984$). Для оценки величины Q использовались выборки, состоящие из 700 случайных последовательностей, сгенерированных с частотами нуклеотидов, характерными для частот нуклеотидов выборки регуляторных районов генов соответствующего вида прокариот. Отметим, что следующий за ним по значимости мотив **AAAAYWAR** является практически полным подобием **AAAYWAAA** со сдвигом и добавлением к нему нуклеотида **A** на 3′-конце, т. е. функционально значимым для регуляторных районов *M. genitalium*, видимо, является контекстный сигнал длины большей, чем восемь нуклеотидов. Такие сигналы можно выявлять в наборе полученных мотивов с использованием различных методов оценки подобия и рассматривать отдельно.

Затем мы провели классификацию полученных мотивов с использованием базы данных сайтов связывания транскрипционных факторов *E. coli* (Osada et al., 2004) с уровнем значимости $p < 10^{-4}$. Как и ожидалось, наибольшее количество ССТФ, демонстрирующих значимое сходство с выявленными мотивами, наблюдалось для *E. coli* ($N_{TFBS} = 42$). Можно предположить, что часть ССТФ-специфичных мотивов, полученных для других видов прокариот, соответствуют транскрипционным факторам – ортологам транскрипционных факторов, найденных в *E. coli* (табл. 5). Оставшаяся часть выявленных мотивов может соответствовать как видоспецифичным ССТФ, отсутствующим в базе данных ССТФ *E. coli*, так и некоторым структурным физико-химическим особенностям регуляторных районов генов прокариот, таким, например, как короткие поли-A / поли-T тракты, приводящие к формированию участков повышенной «плавкости» (easily melting sites) и специфическому изгибу ДНК.

Нам показалось интересным оценить относительное сходство и взаиморасположение вырожденных мотивов в регуляторных районах генов эволюционно близких и эволюционно удаленных видов прокариот. Ранее (Vishnevsky et al., 2011) нами был предложен метод усредненной оценки межвидового олигонуклеотидного сходства регуляторных районов генов H_{Oli}. Этот метод учитывает как степень вырожденности мотивов, так и характер их расположения в регуляторных районах. На рис. 2 показаны оценки такого сходства H_{Oli}, полученные для разных видов, рассчитанные на основе вырожденных мотивов, полученных для *M. genitalium*. Для нижней оценки величины олигонуклеотидного сходства использовалась выборка, состоящая из 700 случайных последовательностей, сгенерированных с частотами нуклеотидов, характерными для частот нуклеотидов выборки регуляторных районов генов *M. genitalium*. Хорошо видно, что максимальные значения олигонуклеотидного сходства с регуляторными районами генов *M. genitalium* наблюдаются для наиболее близких к нему видов, таких как *M. gallisepticum* ($H_{Oli} = 0{,}53$) и *M. pneumoniae* ($H_{Oli} = 0{,}5$), в то время как эволюционно удаленные от него виды, такие как *B. subtilis* и *E. coli*, имеют существенно меньшие значения H_{Oli} – 0,22 и 0,08 соответственно. Кроме того, с использованием метода парного выравнивания мы оценили усредненную меж-

Table 4. An example of motifs found in [–100; –50] regions of *M. genitalium* with reference to the translation start

Motif	*F*	*Q*	*–P*
AAAYWAAA	0.11	0.03	989
AAAAYWAR	0.14	0.05	804
AAAMAASM	0.12	0.02	603
SAAAAAMW	0.12	0.04	558
AWYTNWTT	0.21	0.10	392
WAAYTRWT	0.14	0.07	374
NWAGMAAA	0.14	0.07	348
TRCWANWK	0.12	0.08	34
RKMMTTWK	0.12	0.08	34
YAASYTWN	0.12	0.08	34
MWYTWTKS	0.12	0.05	34
WKWWGSWK	0.18	0.08	30
SMASMANT	0.11	0.05	30
AASNSYTW	0.11	0.05	26

Designations: *F*, motif occurrences in the sample analyzed; *Q*, motif occurrences in the sample of random sequences generated with the nucleotide frequencies specific o the analyzed sample; *P*, decimal logarithm of the probability of observing the motif by chance.

Table 5. An example of motifs found in the [–100; –50] regions of *H. pylori* relative to the translation start that demonstrates a significa t similarity to *E.coli* TFBSs

Motif	TFBS	*–P*
SNSAWRAA	metr	6
KRRGAWWR	lrp	5
NWMTTWAA	nagc	4
WWKSSMTT	metJ	4
AWAAYNSM	argr	4

P, decimal logarithm of observation probability of the motif in E. coli TFBSs by chance.

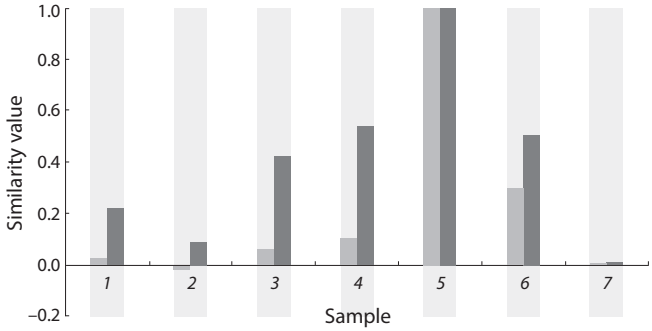

Fig. 2. Values of the average oligonucleotide similarity H_{Oli} (dark box) and the averaged interspecies homology H_{align} (light box) in the regulatory regions of *M. genitalium* genes with the regulatory regions of genes of other prokaryotes.

1, *B. subtilis*; *2*, *E. coli*; *3*, *H. pylori*; *4*, *M. gallisepticum*; *5*, *M. genitalium*; *6*, *M. pneumoniae*; *7*, random sample.

видовую гомологию H_{align} регуляторных районов генов различных видов прокариот согласно Vishnevsky et al., (2011). На рис. 2 показаны оценки усредненной гомологии регуляторных районов генов 5 видов прокариот с регуляторными районами генов *M. genitalium*. Видно, что усредненная гомология с *M. genitalium* резко снижена даже для эволюционно близких к нему видов, а для эволюционно далеких видов она находится на случайном уровне.

Таким образом, на основе анализа величин H_{Oli} и H_{align} можно сделать вывод, что общий контекст регуляторных районов генов может практически полностью изменяться в ходе эволюции и разделения видов, в то время как регуляторный код, основанный на присутствии в этих районах специфических контекстных сигналов, остается в значительной степени консервативным. Можно предположить, что это обусловлено как относительной консервативностью сайтов связывания транскрипционных факторов, так и сходством физико-химических структурных особенностей регуляторных районов генов.

На основе технологии CUDA нами разработан программный комплекс для итерационного выявления в регуляторных районах генов наборов значимых вырожденных

олигонуклеотидных мотивов фиксированной длины, записанных в 15-буквенном коде IUPAC. Предложенный метод может использовать для вычислений широко распространенные бюджетные графические ускорители. Было показано, что применение GPU позволяет в десятки раз увеличить производительность системы в сравнении с распространенными CPU процессорами. На базе предложенного подхода с использованием языков Perl, C++ и CUDA нами разрабатывается Интернет-доступный сайт, который даст исследователям новый инструмент для изучения регуляторных районов генов.

Проведенный анализ регуляторных районов генов *B. subtilis*, *E. coli*, *H. pylori*, *M. gallisepticum*, *M. genitalium* и *M. pneumoniae* выявил сотни значимых олигонуклеотидных мотивов. Оценка сходства полученных мотивов с сайтами связывания транскрипционных факторов *E. coli* показала, что только часть мотивов демонстрируют достоверное сходство с этими ССТФ. Мы полагаем, что оставшиеся неклассифицированные мотивы могут соответствовать сайтам связывания видоспецифичных транскрипционных факторов или разнообразным структурным особенностям регуляторных районов, необходимым для нормального протекания процессов транскрипции или трансляции. Полученные мотивы могут являться мишенями для дальнейшего экспериментального анализа.

Выявленные нами мотивы были использованы для оценки усредненного межвидового сходства регуляторных районов генов прокариот различной эволюционной удаленности друг от друга. Оказалось, что, хотя регуляторные районы генов разных видов прокариот в ходе эволюционного расхождения и накопления мутаций различаются крайне сильно, они продолжают сохранять в относительно консервативном виде специфические контекстные сигналы, обеспечивающие регуляцию базовых молекулярно-генетических процессов.

Acknowledgments

Computations were done at the Bioinformatics Shared Access Center of the Institute of the Cytology and Genetics; Siberian Supercomputer Center of SB RAS; and the Supercomputer Center of the Novosibirsk State University. This work was supported by the Institute of Cytology and Genetics, Budgeted Project VI.61.1.2.

Conflict of interest

The authors declare no conflict of interest.

References

Baker Z.K., Prasanna V.K. An architecture for efficient hardware data mining using reconfigurable computing systems. 14th Annual IEEE Symp. on Field-Programmable Custom Computing Machines, 2006.

Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. GenBank. Nucl. Acids Res. 2013;41(Database issue):D36-42.

Elnitski L., Hardison R.C., Yang S., Kolbe D., Eswara P., O'Connor M.J., Schwartz S., Miller W. Chiaromonte F. Distinguishing regulatory DNA from neutral sites. Genome Res. 2003;13(1):64-72.

Fomin E.S., Alemasov N.A. Implementation of a non-bonded interaction calculation algorithm for the cell architecture. Lect. Notes Comput. Sci. 2009;5698:399-405.

Grundy W.N., Bailey T.L., Elkan C.P. ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. CABIOS. 1996;12:303-310.

Hertz G.Z, Stormo G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 1999;15:563-577.

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 2002;30:312-317.

Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., Wootton J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993;262:208-214.

Manavski S.A., Valle G. CUDA compatible GPU cards as efficient hardware accelerators for Smith–Waterman sequence alignment. BMC Bioinformatics. 2008;26;9 Suppl 2:S10.

Marsan L., Sagot M.F. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. J. Comput. Biol. 2000;7:345-362.

Matys V., Kel-Margoulis O.V., Fricke E., Liebich I., Land S., Barre-Dirrie A., Reuter I., Chekmenev D., Krull M., Hornischer K., Voss N., Stegmaier P., Lewicki-Potapov B., Saxel H., Kel A.E., Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucl. Acids Res. 2006;34:D108-10.

Mrázek J., Gaynon L.H., Karlin S. Frequent oligonucleotide motifs in genomes of three streptococci. Nucl. Acids Res. 2002;19:4216-4221.

NVIDIA CUDA programming guide 3.2. [http://developer.download.nvidia.com/compute/cuda/3_2/toolkit/docs/CUDA_C_Programming_Guide.pdf]

Osada R., Zaslavsky E., Singh. M. Comparative analysis of methods for representing and searching for transcription factor binding sites. Bioinformatics 2004;20(18):3516-3525.

Pesole G., Liuni S., Dsouza M. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. Bioinformatics. 2000;16:439-450.

Pevzner P.A., Sze S.H. Combinatorial approaches to finding subtle signals in DNA sequences. Proc. of the 8th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB). 2000.

Portales-Casamar E., Thongjuea S., Kwon A.T., Arenillas D., Zhao X., Valen E., Yusuf D., Lenhard B., Wasserman W.W., Sandelin A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucl. Acids Res. 2010;38:D105-10.

Sukhwani B., Herbordt M.C. GPU acceleration of a production molecular docking code. Proc. of 2nd Workshop on General Purpose Processing on Graphics Processing Units. 2009.

Vishnevsky O.V., Gunbin K.V., Bocharnikov A.V., Berezikov E.V. Analysis of the conservative motifs in promoters of miRNA genes, expressed in different tissues of mammalians. Evolutionary Biology Concepts, Molecular and Morphological Evolution. 2011.

Vishnevsky O.V., Kolchanov N.A. ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. Nucl. Acids Res. 2005;33(Web Server issue):417-22.

Yooseph S., Sutton G., Rusch D.B., Halpern A.L., Williamson S.J., Remington K., Eisen J.A., Heidelberg K.B., Manning G., Li W., Jaroszewski L., Cieplak P., Miller C.S., Li H., Mashiyama S.T., Joachimiak M.P., van Belle C., Chandonia J.M., Soergel D.A., Zhai Y., Natarajan K., Lee S., Raphael B.J., Bafna V., Friedman R., Brenner S.E., Godzik A., Eisenberg D., Dixon J.E., Taylor S.S., Strausberg R.L., Frazier M., Venter J.C. The sorcerer II global ocean sampling expedition: expanding the universe of protein families. PLoS Biol. 2007:5(3):e16.