

Королев С.А., Горбунов К.Ю., Зверков О.А., Селиверстов А.В., Любецкий В.А.

Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, г. Москва, Россия

ВЫРОЖДЕННЫЕ ИНВЕРТИРОВАННЫЕ ПОВТОРЫ В ГЕНОМАХ МИКОБАКТЕРИЙ

АННОТАЦИЯ

В статье обсуждается частота инвертированных повторов на ДНК в зависимости от их длины, числа несовпадений и положения относительно кодирующих областей.

КЛЮЧЕВЫЕ СЛОВА

Биоинформатика; микобактерии; ДНК; инвертированный повтор.

Korolev S.A., Gorbunov K.Yu., Zverkov O.A., Seliverstov A.V., Lyubetsky V.A.

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia

DEGENERATE INVERTED REPEATS IN THE GENOMES OF MYCOBACTERIUM

ABSTRACT

The article discusses frequency of inverted repeats in DNA, depending on their length, number of mismatches, and position relative to the coding regions.

KEYWORDS

Bioinformatics; Mycobacterium; DNA; inverted repeat.

Введение

Основная часть статьи направлена на изучение некодирующих участков геномов микобактерий. Среди микобактерий много возбудителей опасных социально значимых инфекций, в том числе, туберкулёза (*Mycobacterium bovis*, *M. tuberculosis*, и другие) и проказы (*M. leprae*). Поэтому их исследование может иметь значение для медицины и эпидемиологии. В частности, важно оценивать риск возникновения новых штаммов возбудителей, устойчивых к действию антибиотиков. В то же время существуют и свободно живущие виды, например, *M. smegmatis*.

С одной стороны, некодирующие участки генома играют важную роль в регуляции экспрессии генов. В частности, их исследование позволяет понять механизмы ответа на стресс, включая различные виды терапевтического воздействия на возбудителей опасных инфекций. С другой стороны, эти участки несут информацию о хромосомных перестройках, которые служат важным фактором эволюции, в том числе – механизмом возникновения устойчивости к антибиотикам и изменению состава мембранных белков, распознаваемых иммунной системой хозяина.

Мы отождествляем каждую цепь ДНК с последовательностью букв в алфавите {A, C, G, T}. Нуклеотид А комплементарен нуклеотиду Т, нуклеотид С комплементарен нуклеотиду G. Два слова равной длины n комплементарны, если для всех позиций, начиная с первой, k -й нуклеотид первого слова комплементарен $(n-k+1)$ -му нуклеотиду второго слова. Инвертированным повтором называется участок ДНК чётной длины, начало которого комплементарно концу. Вырожденным инвертированным повтором называется участок произвольной длины, близкий в метрике Левенштейна к инвертированному повтору [1]. Расстояние Левенштейна (также редакционное расстояние или дистанция редактирования) между двумя строками — это минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую. Реализации алгоритма для вычисления расстояния Левенштейна на разных языках программирования доступны в [2]. Это расстояние между двумя строками длины m и n вычисляется методом динамического программирования [3], выполняющим $O(mn)$ операций с линейной памятью $O(\min(n, m))$.

Напомним, что ДНК состоит из двух комплементарных цепей. Поэтому в зависимости от цепи, каждый ген транскрибируется в определённом направлении, а некодирующие области различаются в зависимости от взаимного направления транскрипции фланкирующих генов.

Участок РНК, соответствующий вырожденному инвертированному повтору на ДНК, может образовать шпильку, то есть вторичную структуру, в которой комплементарные нуклеотиды из начала и из конца соединяются между собой. Середина образует петлю, в которой цепь РНК изгибается в пространстве. Нуклеотиды петли не обязательно комплементарны. Петля не может быть короче трёх и обычно содержит не меньше четырёх нуклеотидов. Большие петли уменьшают стабильность шпильки. Нуклеотиды, не входящие в состав петли, образуют плечи шпильки. Некомплементарность нуклеотидов плеч также уменьшает стабильность шпильки, поскольку некомплементарные нуклеотиды образуют выпячивания. Мы классифицируем шпильки по трём параметрам: длине плеча, длине петли и расстоянию между одним плечом и участком, комплементарным другому плечу в метрике Левенштейна.

Шпильки играют важную роль в регуляции экспрессии генов, поскольку во многих случаях служат терминаторами транскрипции, то есть прерывают процесс создания РНК по ДНК. Примеры рассмотрены в работах [4-9]. Шпильки участвуют в регуляции экспрессии генов, часто образуя сложные структуры; они могут служить для предотвращения конфликтов, возникающих в ходе транскрипции генов на комплементарных цепях ДНК [10-11]. Шпилька на 3'-конце РНК служит для стабилизации транскрипта, предотвращая его разрушение ферментами РНКазы.

Также вырожденные инвертированные повторы могут служить сайтами кооперативного связывания транскрипционных факторов с ДНК. В этом случае две копии фактора связывают два участка ДНК, расположенные на комплементарных цепях ДНК. С другой стороны, инвертированные повторы возникают в результате хромосомных перестроек, в частности, на краях вставок мобильных элементов. Сравнительный анализ хромосомных структур даёт важную информацию об эволюции генома [12].

Материалы

Геномные данные получены из базы данных GenBank. Мы рассмотрели полные геномы следующих сорока видов микобактерий:

NC_000962 *Mycobacterium tuberculosis H37Rv*
NC_002677 *Mycobacterium leprae TN*
NC_002755 *Mycobacterium tuberculosis CDC1551*
NC_002944 *Mycobacterium avium subsp. paratuberculosis K-10*
NC_002945 *Mycobacterium bovis AF2122/97*
NC_008146 *Mycobacterium sp. MCS*
NC_008595 *Mycobacterium avium 104*
NC_008596 *Mycobacterium smegmatis str. MC2 155*
NC_008611 *Mycobacterium ulcerans Agy99*
NC_008705 *Mycobacterium sp. KMS*
NC_008726 *Mycobacterium vanbaalenii PYR-1*
NC_008769 *Mycobacterium bovis BCG str. Pasteur 1173P2*
NC_009077 *Mycobacterium sp. JLS*
NC_009338 *Mycobacterium gilvum PYR-GCK*
NC_009525 *Mycobacterium tuberculosis H37Ra*
NC_009565 *Mycobacterium tuberculosis F11*
NC_010397 *Mycobacterium abscessus ATCC 19977*
NC_010612 *Mycobacterium marinum M*
NC_011896 *Mycobacterium leprae Br4923*
NC_012207 *Mycobacterium bovis BCG str. Tokyo 172*
NC_012943 *Mycobacterium tuberculosis KZN 1435*
NC_014814 *Mycobacterium gilvum Spyr1*
NC_015564 *Amycolicococcus subflavus DQS3-9A1*
NC_015576 *Mycobacterium sp. JDM601*
NC_015758 *Mycobacterium africanum GM041182*
NC_015848 *Mycobacterium canettii CIPT 140010059*
NC_016604 *Mycobacterium rhodesiae NBB3*
NC_016768 *Mycobacterium tuberculosis KZN 4207*
NC_016804 *Mycobacterium bovis BCG str. Mexico*
NC_016946 *Mycobacterium intracellulare ATCC 13950*
NC_016947 *Mycobacterium intracellulare MOTT-02*
NC_016948 *Mycobacterium intracellulare MOTT-64*
NC_017522 *Mycobacterium tuberculosis CCDC5180*

NC_017523 *Mycobacterium tuberculosis* CCDC5079
NC_017524 *Mycobacterium tuberculosis* CTRI-2
NC_017904 *Mycobacterium* sp. MOTT36Y
NC_018027 *Mycobacterium chubuense* NBB4
NC_018078 *Mycobacterium tuberculosis* KZN 605
NC_018143 *Mycobacterium tuberculosis* H37Rv
NC_018289 *Mycobacterium smegmatis* str. MC2 155

Методы

Независимо рассматривались кодирующие области (гены) и некодирующие (межгенные) области трёх типов:

- между сходящимися генами;
- между расходящимися генами;
- между последовательно расположенными генами.

Учитывались только кодирующие области, размеченные в аннотациях геномов. Для этого исследования написана программа, реализующая оригинальный алгоритм поиска вырожденных инвертированных повторов и привязки их к областям генома. Программа написана на языке Python и работает следующим образом:

- Из основного кода вызывается функция `find_hairpins_in_file`, она получает на вход gbk-файл, с помощью вспомогательных функций получает из него нуклеотидную последовательность, находит на ней координаты генов. Затем для каждой межгенной области (добавляя 20 н. с каждой стороны) вызывается функция `find_cross_hairpins`, после чего результаты делятся по типам межгенных областей и возвращаются в основную программу, где для всех файлов уже вычисляются средние величины и т.д.
- Функция `find_cross_hairpins` получает на вход последовательность нуклеотидов, минимальную длину плеча шпильки (по умолчанию это 7 н.), максимальную величину петли (по умолчанию это 14 н.). В цикле по величине петли ищутся шпильки с минимальной длиной плеча (минимальная длина повышается, если петля достигает ее значения) и удовлетворяющие заданному расстоянию Левенштейна между плечами. Каждая найденная шпилька, передается в функцию `find_all_possible_hairpins` (с условием, чтобы в итоге максимальная длина плеча была ограничена 35 н.). Параметры шпилек и их расстояния до генов записываются в массив. Вызывается функция `check_and_delete_hairpins`. Этот результат возвращается в `find_hairpins_in_file`.
- Функция `find_all_possible_hairpins` получает на вход последовательность, у которой посередине найдена шпилька, и параметры этой шпильки. Функция добавляет по одному нуклеотиду с каждой стороны и проверяет, можно ли ожидать шпильку большей длины. Возвращает все возможные варианты более длинных шпилек.
- Функция `check_and_delete_hairpins` получает на вход массив из шпилек с их параметрами и расстояниями до генов. Сортирует шпильки по расстоянию до левого гена. Вычисляет отношение перекрытия соседних шпилек к длине (оба плеча и петля) более короткой из них, если перекрытие превышает порог (по умолчанию - 70%), то удаляет короткую шпильку. При одинаковой длине удаляет ту, в которой больше расстояние Левенштейна между плечами. Возвращает оставшиеся шпильки.

Время счёта на процессоре с двумя ядрами составило примерно 10 минут на один геном.

Результаты и обсуждение

Всего рассмотрено 187759 межгенных областей, из них 123737 последовательных, 31994 сходящихся и 32028 расходящихся (рис. 1).

Большое число вырожденных инвертированных повторов, соответствующих шпилькам с длиной петли четыре нуклеотида, позволяет предполагать, что значительная доля этих повторов действительно соответствует шпилькам на РНК. Две зависимости расстояния между шпилькой и ближайшим геном от параметров шпильки существенно различаются между собой для двух типов межгенных областей. Это говорит о различной роли шпилек в зависимости от типа области. Они либо играют регуляторную роль в экспрессии генов, либо служат для стабилизации транскриптов, располагаясь на 3'-конце РНК. Среднее расстояние от шпильки до ближайшего гена также зависит от типа области и составляет около 100 п.н. для областей между сходящимися фланкирующими генами и около 70 п.н. для областей между расходящимися фланкирующими генами.

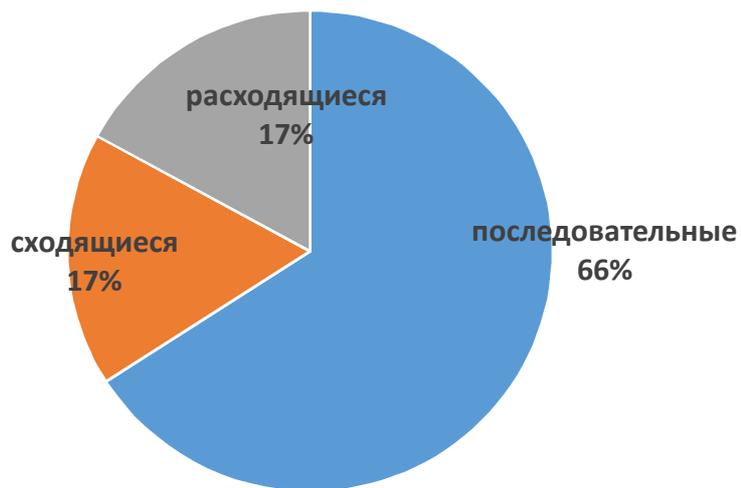


Рис.1. Соотношение чисел межгенных областей трёх типов в геномах рассмотренных микобактерий

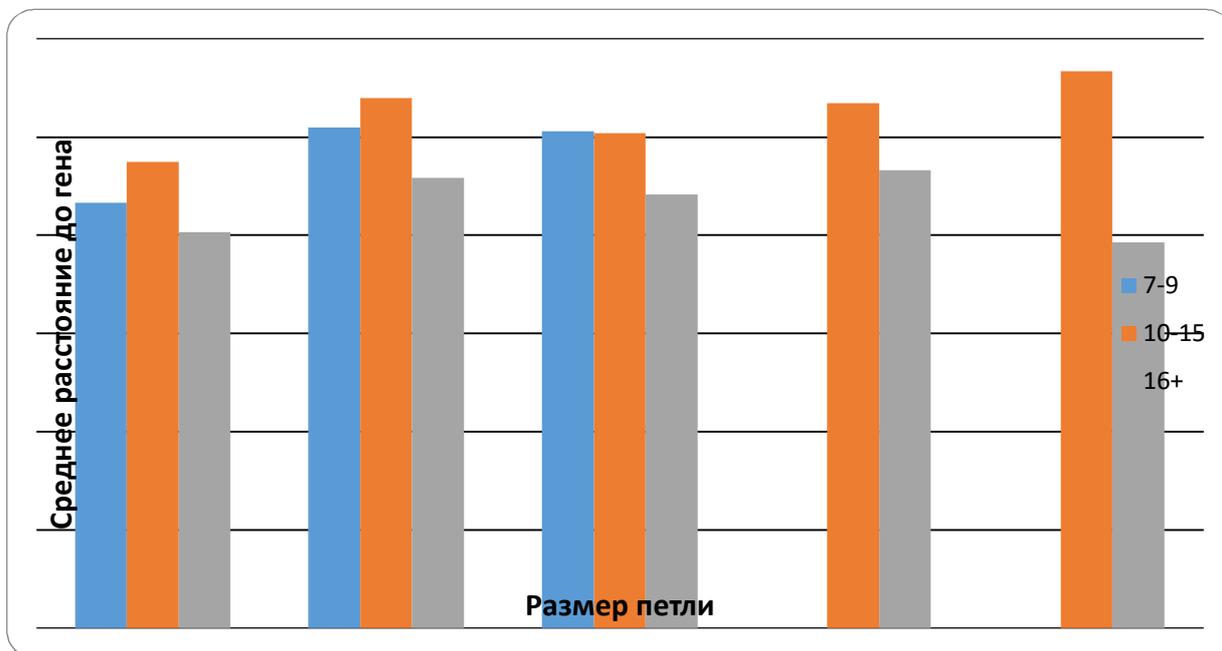


Рис.2. Зависимость среднего расстояния между шпилькой и ближайшим к ней геном от размера петли для разных размеров плеча шпильки в области между сходящимися генами

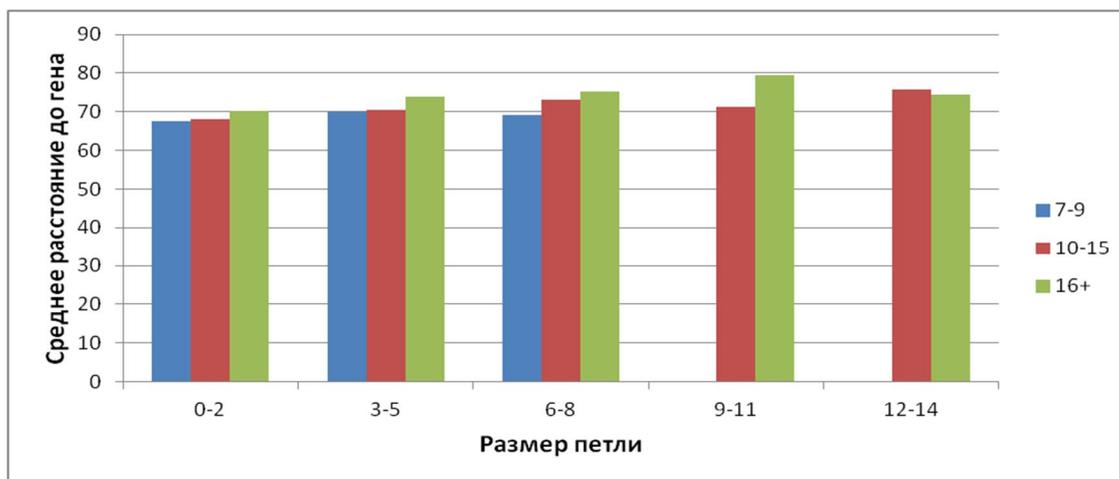


Рис.3. Зависимость среднего расстояния между шпилькой и ближайшим к ней геном от размера петли для разных размеров плеча шпильки в области между расходящимися генами

Для последовательных генов заметен рост расстояния между шпильками всех размеров и началом гена при увеличении петли. Также с увеличением петли растёт расстояние между короткими шпильками и концом гена. При величине петли в 6-8 нуклеотидов, среднее расстояние между короткими шпильками и началом гена, становится больше, чем между началом гена и длинными или средними шпильками. В областях между расходящимися генами шпильки находятся ближе к генам, чем в областях между сходящимися генами.

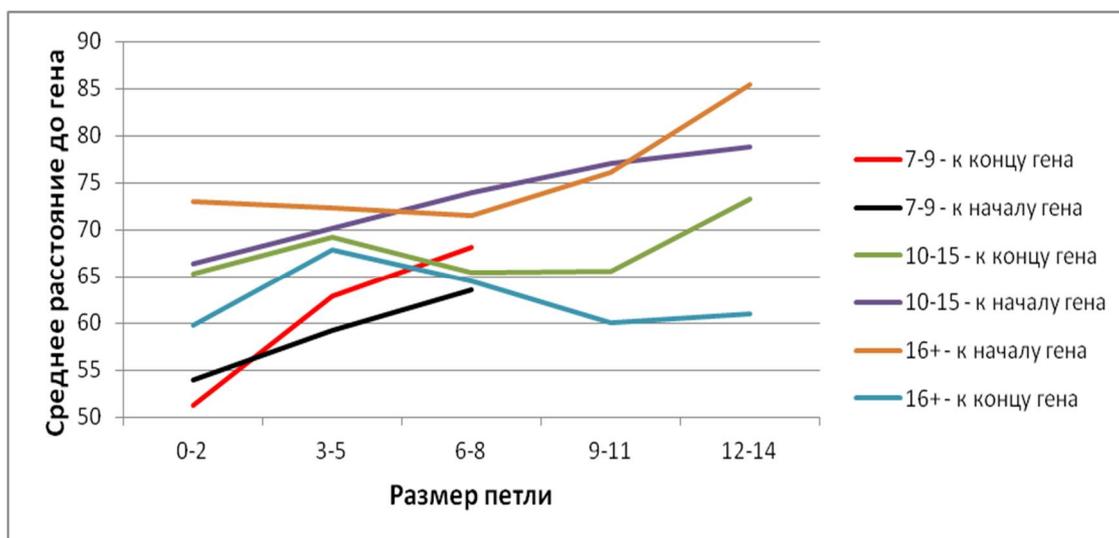


Рис.4. Зависимость среднего расстояния между шпилькой и ближайшим к ней геном от размера петли для разных размеров плеча шпильки и разных вариантов примыкания шпильки к гену в области между последовательными генами

В областях между последовательными генами короткие шпильки в среднем находятся ближе к началам генов, а средние и длинные ближе к концам. Средние длины некодирующих областей между последовательно расположенными и сходящимися генами приблизительно равны для разных таксономических групп (табл.1). А некодирующие области между расходящимися генами в среднем значительно длиннее у всех групп. В то время как среднее расстояние до ближайшего гена более короткое именно для расходящихся генов. То есть этот эффект нельзя объяснить простым увеличением длин некодирующих областей.

Табл.1. Среднее расстояние между генами в зависимости от типа расположения генов для разных таксономических групп

	Последовательные гены	Расходящиеся гены	Сходящиеся гены
Микобактерии	77,1	174,6	71,4
Актинобактерии	100,4	211,0	104,7
Цианобактерии	135,9	251,8	124,4
Фирмикуты	112,9	259,0	146,4

Выводы

Определены интервалы типичных значений параметров шпилек и расстояний от них до ближайших генов у микобактерий. Полученные результаты могут служить основой для дальнейшего предсказания регуляции экспрессии генов. Также полученные результаты могут быть использованы для предсказания частоты хромосомных перестроек, в результате которых возникают инвертированные повторы. Это позволяет уточнить ранее рассмотренную модель эволюции генома.

Работа выполнена за счёт гранта Российского научного фонда (проект 14-50-00150).

Литература

1. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии Наук СССР. — 1965. — Т. 163, № 4. — С. 845–848.
2. Levenshtein distance. URL: https://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Levenshtein_distance
3. Wagner R.A., Fischer M.J. The string-to-string correction problem // J. ACM. — 1974. — V. 21, no. 1. — P. 168–173.
4. Лопатовская К.В., Селиверстов А.В., Любецкий В.А. Аттеноаторная регуляция оперонов биосинтеза аминокислот и аминоацил-тРНК у бактерий: сравнительный геномный анализ // Молекулярная биология. — 2010. — Т. 44, № 1. — С. 140–151.
5. Любецкая Е.В., Селиверстов А.В., Любецкий В.А. У актинобактерий число длинных шпилек в межгенных трейлерных областях велико по сравнению с другими областями генома // Молекулярная биология. — 2007. — Т. 41, № 4. — С. 739–742.
6. Lyubetsky V.A., Pirogov S.A., Rubanov L.I., Seliverstov A.V. Modeling classic attenuation regulation of gene expression in bacteria // Journal of Bioinformatics and Computational Biology. — 2007. — V. 5, no. 1. — P. 155–180.
7. Селиверстов А.В., Любецкий В.А. Механизм регуляции транспорта марганца у Brucella с участием длинной спирали РНК // Биофизика. — 2009. — Т. 54, № 2. — С. 222–225.
8. Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria // BMC Microbiology. — 2005. — V. 5, no. 54, 14 pages. DOI: 10.1186/1471-2180-5-54.
9. Grundy F.J., Henkin T.M. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. // Mol Microbiol. — 1998. — V. 30, no. 4. — P. 737–749.
10. Lyubetsky V.A., Zverkov O.A., Pirogov S.A., Rubanov L.I., Seliverstov A.V. Modeling RNA polymerase interaction in mitochondria of chordates // Biology Direct. — 2012. — V. 7, no. 26. DOI: 10.1186/1745-6150-7-26.
11. Lyubetsky V.A., Zverkov O.A., Rubanov L.I., Seliverstov A.V. Modeling RNA polymerase competition: the effect of σ -subunit knockout and heat shock on gene transcription level // Biology Direct. — 2011. — V. 6, no. 3. DOI: 10.1186/1745-6150-6-3.
12. Lyubetsky V.A., Gershgorin R.A., Seliverstov A.V., Gorbunov K.Yu. Algorithms for reconstruction of chromosomal structures // BMC Bioinformatics. — 2016. — V. 17, no. 40, 23 pages. DOI: 10.1186/s12859-016-0878-z.

References

1. Levenshtein V.I. Binary codes capable of correcting deletions, insertions, and reversals // Soviet Physics Doklady. — 1966. — V. 10, no. 8. — P. 707–710.
2. Levenshtein distance. URL: https://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Levenshtein_distance
3. Wagner R.A., Fischer M.J. The string-to-string correction problem // J. ACM. — 1974. — V. 21, no. 1. — P. 168–173.
4. Lopatovskaya K.V., Seliverstov A.V., Lyubetsky V.A. Attenuation regulation of the amino acid and aminoacyl-tRNA biosynthesis operons in bacteria: a comparative genomic analysis // Molecular Biology. — 2010. — V. 44, no. 1. — P. 128–139. DOI: 10.1134/S0026893310010164.
5. Lyubetskaya E.V., Seliverstov A.V., Lyubetsky V.A. The number of long hairpins in intergenic trailer regions of actinobacteria is far greater than in other genomic regions // Molecular Biology. — 2007. — V. 41, no. 4. — P. 670–673. DOI: 10.1134/S002689330704022X.
6. Lyubetsky V.A., Pirogov S.A., Rubanov L.I., Seliverstov A.V. Modeling classic attenuation regulation of gene expression in bacteria // Journal of Bioinformatics and Computational Biology. — 2007. — V. 5, no. 1. — P. 155–180.
7. Seliverstov A.V., Lyubetsky V.A. Mechanism of manganese transport regulation in Brucella involving a long RNA helix // Biophysics. — 2009. — V. 54, no. 2. — P. 152–155. DOI: 10.1134/S0006350909020055.
8. Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria // BMC Microbiology. — 2005. — V. 5, no. 54, 14 pages. DOI: 10.1186/1471-2180-5-54.
9. Grundy F.J., Henkin T.M. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. // Mol Microbiol. — 1998. — V. 30, no. 4. — P. 737–749.
10. Lyubetsky V.A., Zverkov O.A., Pirogov S.A., Rubanov L.I., Seliverstov A.V. Modeling RNA polymerase interaction in mitochondria of chordates // Biology Direct. — 2012. — V. 7, no. 26. DOI: 10.1186/1745-6150-7-26.
11. Lyubetsky V.A., Zverkov O.A., Rubanov L.I., Seliverstov A.V. Modeling RNA polymerase competition: the effect of σ -subunit knockout and heat shock on gene transcription level // Biology Direct. — 2011. — V. 6, no. 3. DOI: 10.1186/1745-6150-6-3.
12. Lyubetsky V.A., Gershgorin R.A., Seliverstov A.V., Gorbunov K.Yu. Algorithms for reconstruction of chromosomal structures // BMC Bioinformatics. — 2016. — V. 17, no. 40, 23 pages. DOI: 10.1186/s12859-016-0878-z.

Поступила 21.10.2016

Об авторах:

Королев Семен Александрович, лаборатория № 6 Института проблем передачи информации им. А.А. Харкевича Российской академии наук, korolev@iitp.ru;

Горбунов Константин Юрьевич, лаборатория № 6 Института проблем передачи информации им. А.А. Харкевича Российской академии наук, кандидат физико-математических наук, gorbunov@iitp.ru;

Зверков Олег Анатольевич, лаборатория № 6 Института проблем передачи информации им. А.А. Харкевича Российской академии наук, кандидат физико-математических наук, zverkov@iitp.ru;

Селиверстов Александр Владиславович, лаборатория № 6 Института проблем передачи информации им. А.А. Харкевича Российской академии наук, кандидат физико-математических наук, slvstv@iitp.ru;

Любецкий Василий Александрович, заведующий лабораторией № 6 Института проблем передачи информации им. А.А. Харкевича Российской академии наук, доктор физико-математических наук, lyubetsk@iitp.ru.