***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 8. Пущино: ИМПБ РАН, 2020. Статья № e28. doi: 10.17537/icmbb20.36*

Обнаружение в геномах высших организмов сильно размытых мегасателлитных повторов

Назипова Н.Н.*, Тюльбашева Г.Э., Теплухина Е.И.

Институт математических проблем биологии РАН – филиал ИПМ РАН им. М.В. Келдыша

*nnn@impb.ru

Создано программное обеспечение для обнаружения нового вида периодичности в геномах – размытых мегасателлитов с переменной длиной паттерна. Это тандемные (идущие непрерывно друг за другом) повторы с характерной длиной паттерна от 1000 нуклеотидов. В сочетании с имеющейся в нашем распоряжении вычислительной технологией, которая хорошо обнаруживает размытые микро-, мини- и макросателлиты, это программное обеспечение позволит получать новые данные о некодирующей части генома, где локализуется большое количество периодичностей разного вида. Это приблизит нас к пониманию их роли в геноме.

Ключевые слова: *геном, гены, межгенные промежутки, периодичность ДНК, дивергенция копий паттерна.*

Detection in the Genomes of Higher Organisms of Highly Fuzzy Megasatellite Repeats

Nazipova N., Tyulbasheva G., Teplukhina E.

Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics RAS

We created a software for detecting a new type of periodicity in genomes - approximate megasatellites with variable pattern length. These are tandem (continuous one after another) repeats with a pattern length of 1000 nucleotides and longer. The computational technology, which we developed earlier, is good at detecting fuzzy micro-, mini- and macrosatellites. The created software supply us with new tool that allows obtaining new data on the non-coding part of the genome, where a large number of periodicities of different types localize. This will bring us closer to understanding their role in the genome.

Key words: genome, genes, intergenic regions, DNA periodicity, pattern copy divergence.

1. Введение

Изучение длинных периодичностей в геноме важно для того, чтобы найти подтверждения гипотезам о роли периодической ДНК в геноме. Считается, что всего повторяющейся ДНК в геноме человека – около 50 %, в этой фракции находятся все виды периодичности: повторяющиеся участки в кодирующей и в не кодирующей белки частях генома. Последняя занимает более 90 % генома человека. До сих пор считается, что некодирующая часть генома несет многочисленные эгоистичные элементы и «прочую ДНК, не подверженную отбору, и в меру нашего понимания, являющуюся «мусором» [1].

Нами ранее был разработан метод обнаружения размытых тандемных повторов в геномах с любой длиной паттерна и любой степенью сохранности паттерна [2–4]. Реализация этого подхода функционирует в виде сервера (http://www.mathcell.ru/model5.php?l=ru) и решает проблему нахождения периодичностей с постоянной длиной копий паттерна. В основу технологии извлечения информации о периодичности в геномах положен разработанный коллективом спектрально-статистический подход [2], применение которого позволило создать комплекс программ, достоверно выявляющих статистически значимые размытые тандемные повторы.

Этот подход основан на применении двух ключевых характеристик наличия периодичности. Если задана символьная последовательность длины n над алфавитом длины K, то можно исследовать ее на предмет наличия периодичности любой длины L (значение тест-периода) с помощью характеристики сохранности букв в позициях тест-периодов:

$$pl(L) = \frac{1}{L}\sum_{j=1}^{L} \max\left\{\pi_j^i : i \in 1,...K\right\}, \qquad (1)$$

где π_j^i – число встречаемости i-ой буквы алфавита последовательности в j-ых позициях тест-периода.

***Proceedings of the International Conference "Mathematical Biology and Bioinformatics".** Ed. V.D. Lakhno. Vol. 8. Pushchino: IMPB RAS, 2020. Paper No. e28. doi: 10.17537/icmbb20.36*

Характеристика $pl(L)$ является средним значением максимальных частот букв в позициях тест-периода.

Чтобы убедиться в неоднородности анализируемой последовательности, вычисляется спектрально-статистическая характеристика с использованием нормализованного χ^2_{crit} – критерия Пирсона с $(K\text{-}1)(L\text{-}1)$ степенями свободы:

$$H(L) = \frac{\nu_{NP}(L,n)}{\chi^2_{\text{crit}}\left(\alpha,(K-1)(L-1)\right)}, \qquad (2)$$

где $\nu_{NP}(L,n) = R_L \sum_{j=1}^{L}\sum_{i=1}^{K}(\pi_j^i - p^i)^2 \Big/ p^i\left(1-p^i\right)$,

$\alpha = 10^{-6}$ – уровень значимости,

p^i – частота встречаемости i-ой буквы алфавита в анализируемой последовательности,

$R_L = n/L$ – количество копий тест-периода (может быть нецелое число).

Этот подход реализован в виде многоэтапной вычислительной технологии, на первом этапе которой проводится многократное сканирование геномных последовательностей для выявления участков скрытой периодичности. Каждый последующий этап технологии имеет целью подробный анализ возможности удаления вхождений, а также пересечений и объединений участков с различными паттернами периодичности с целью последовательной борьбы с избыточностью результатов, неизбежной при реализации автоматизированной полномасштабной сканирующей технологии. Теоретический предельный уровень дивергенции копий паттерна в тандемных повторах, выявляемых нашим подходом, составляет 50 %. Этот метод одинаково хорошо выявляет как микро- и минисателлиты с длиной паттерна от 2 нукл. и более, так и периодичности с длиной паттерна свыше 100 нукл. (макросателлиты).

В процессе анализа геномов мыши и крысы с помощью спектрального анализа [5–7] нами был обнаружен новый вид скрытой периодичности – мегасателлитные тандемные повторы. Они характерны тем, что имеют длину паттерна порядка более 2 тысяч позиций, а также большую вариабельность в длинах повторов паттерна, которая обусловлена наличием треков переменной длины, представленных простыми повторами, например, ATAT…, это т.н. SSR (simple sequence repeats), при достаточно консервативной базовой последовательности (core sequence) паттерна периодичности. Для выявления такого вида периодичностей в геномах планируется использовать подходы, разработанные для выравнивания коротких прочтений в процессе сборки генома.

Для сборки консенсусной последовательности (генома) из набора коротких многократно перекрывающихся между собой фрагментов прочтений разработаны алгоритмы выравнивания, использующие преобразование Барроуза – Уиллера (Burrows-Wheeler Transform, BWT) [8] при обратном обходе префиксных деревьев [9], и позволяющие эффективно выравнивать участки ДНК с тем, чтобы определить степень их сходства. При этом допускаются несовпадения, вставки и выпадения фрагментов. Этот подход быстрее в 1000 раз широко используемого алгоритма Нидлмана-Вунша (Needleman-Wunsch) [10]. Но мы используем другой способ выравнивания: находим затравки (seeds) и пытаемся расширить участки совпадения насколько это возможно в обе стороны от затравки.

2. Описание метода

В основе метода поиска мегасателлитных последовательностей лежит применение суффиксных массивов. В принципе, этот метод хорошо подходит и для поиска всех разнесенных неточных повторов переменной длины, имеющих один и тот же паттерн.

Сначала строится суффиксный массив всей геномной последовательности. Суффиксы располагаются в лексикографическом порядке [11]. В дальнейшем работа для экономии памяти будет идти не с суффиксами, а с их начальными позициями в геноме.

Также создаётся l-граммный словарь (лексикографически упорядоченный в порядке убывания массив всех слов заданной длины l с количеством вхождений каждой l-граммы в исследуемую последовательность). У нас $l = 14$. Затем он переупорядочивается по убыванию числа вхождений, при этом исключаются l-граммы, которые встречаются менее трёх раз.

На следующем шаге для повышения скорости счета поэтапно сокращается массив l-грамм. Для каждого слова просматриваются все лексикографически меньшие слова и удаляются те l-граммы, которые совпадают с рассматриваемым словом с точностью до двух несовпадений. Получаем массив представителей групп похожих l-грамм.

Для каждой l-граммы с помощью алгоритма обратного поиска в BWT [12, 9] находим все координаты вхождений с двумя разрешенными несовпадениями. Получаем P вхождений. Так как нас интересуют тандемные повторы, мы берем в рассмотрение только $p <= P$ координат, лежащих в разумных границах близости друг от друга, чтобы образовать тандемный повтор, расстояние между которыми не превышает заданную максимальную длину паттерна тандемного повтора. Этот параметр в настоящей реализации алгоритма равен 4000 нуклеотидов.

Имея p координат вхождений l-граммы, делаем $m = C_p^2$ попарных продолжений выравнивания таким образом, чтобы сохранить плотность несовпадений (они должны быть разбросаны

равномерно, а не скапливаться на концах выравниваемых участков). Поэтому расширение затравки осуществляется довольно строгой процедурой, которая основана, по сути, на алгоритме Нидлмана – Вунша с простой системой весов (цена совпадения – 1, несовпадения, вставки/выпадения символа – –1). Эта трудоемкая процедура дает нам, во-первых, левую и правую границы района периодичности, а также максимальное значение длины паттерна периодичности L_{max}. И таких районов, находящихся на любом расстоянии друг от друга, можно получить одновременно несколько.

Полученный район возможной периодичности исследуется с целью определения истинной длины паттерна с использованием модифицированного нами метода глобального оберточного программирования (Global Wraparound Dynamic Programming) [13]. При выборе характерной длины паттерна предпочтение отдаётся значению $L = 100, \ldots, L_{max}$, при котором величина $H(L)$ (2) максимальна, а величина сохранности паттерна $pl(L)$ (1) не уменьшилась.

3. Результаты

Результаты показаны на паре примеров, полученных при анализе I хромосомы *C. elegans* (идентификатор GenBank NC_003279.4). Длины периодичностей, представленные в примере, не являются характерными для мегасателлитов. Мы их намеренно взяли небольшими, чтобы была возможность наглядно представить результат.

Пример 1. Для затравки caaaaaaatttccc в результате работы программы нашелся участок с координатами 14258256–14258787 и характерной длиной паттерна 103. На рисунке 1 он выписан с разбивкой на блоки длины 103, красным выделены вхождения самой *l*-граммы и ее гомологов. В нижней строке приведена консенсусная последовательность.

```
14258256 tttcggaaaatttgaattcccgccgaaaagctttctcagaaaatttgaatttcccgccaaaaagttttatcggataaatttgaatttcccgcctaaaaatttt
14258359 atcggaaaatttgagtttcccgccaaaaaattctcacagaaaatttgaatttcccgccaaaattcgttttctcagaaaatttgaattttccgccaaaaagtat
14258462 tctcagaaagtttgaatttcccgccaataagttttatcggaaaatttgagtttcccgccataaaaatttctctcagaaaatttaaatttccagccaaaaagtt
14258565 atctcagaaaatttaaatttcccgccaaaaaatttctcagaaaatttaaatttcccgccaaactttttctcagaaaatttgaatttcccgccaaaaagtat
14258668 tctcggaatatttgaatttccagccaaaacctttttcgcggaaaattagaattttccgccaaaaaattttcccagaaaatttgaatttcccgccaaaattggt
14258771 tgggttcgccacaattg
Consens  TTTCGGAAAATTTGAATTTCC*CCCAAAAACTTTTTCAGAAAAATTTAA*TTTCCCCCCAAAAA**TTTTC*CAGAAAATTTGAATTTCCCGCCAAAAA*T*T
```

Рис. 1. Участок из примера 1, разбитый на блоки одинаковой длины, которая равна длине паттерна размытой периодичности в 103 нукл. Показатели качества периодичности $pl(L) = 0.721$, $H(L) = 1.936$.

На рисунке 2 приведен результат выравнивания этого участка с сильно размытыми копиями повтора. Из него видно, что копии имеют разную длину. Показатели качества такого представления периодичности изменились: $pl(L) = 0.821$, $H(L) = 3.535$.

```
14258256 tttcggaaaatttgaatt-c-cc-gccgaaaagctttctcagaaaatttgaatttcccgccaaaaagttttat-cggataaatttgaatttcccgcc-taaaaatttt
14258359 at-cggaaaatttgagtttc-cc-gccaaaaaattctcacagaaaatttgaatttcccgccaaaa----------ttc--g----------ttt
14258428 tctcagaaaatttgaatt-ttcc-gccaaaaagtattctcagaaagtttgaatttcccgccaataagttttat-cgga-aaatttgagtttcccgccataaaaatttc
14258532 tctcagaaaatttaaatt-t-ccagccaaaaagttatctcagaaaatttaaatttcccgccaaaaaaatttct-caga-aaatttaaatttcccgcc--aaacttttt
14258634 tctcagaaaatttgaatttc-cc-gccaaaaagtattctcggaatatttgaatttccagccaaaaccttttcgcgga-aaattagaattttccgcc-aaaaaatttt
14258738 cc-cagaaaatttgaatttc-cc-gcc-aaaa---tt----gg----ttgggtt---cgccacaa----ttg

Consens  TCTCAGAAAATTTGAATTTC-CC-GCCAAAAAGTTTTCTCAGAAAATTTGAATTTCCCGCCAAAAAGTTTTAT-CGGA-AAATTTGAATTTCCCGCC-TAAAAATTTT
```

Рис. 2. Результат работы модифицированного алгоритма обертывающего выравнивания для участка, приведенного на рис. 1.

Пример 2. Для *l*-граммы atttgccaaaaaaa в результате работы программы нашелся участок с координатами 11369740–11370339 и характерной длиной паттерна 70. На рисунке 3 он выписан с разбивкой на блоки длины 70, красным выделены 6 вхождений гомолога затравки attttccaaaaaaa. В нижней строке приведена консенсусная последовательность. Показатели качества периодичности $pl(L) = 0.743$, $H(L) = 2.665$.

```
11369740    ttttttcaagctaaacagaacaaaattttccaaaaaaaaaatttctttttaggctcaaaaatttttttagg
11369810    cccaaatttttttaggcttaaaaaaatttttcgaaaacaaaaatttttttttttaggctcaaaaatttttta
11369880    ggcccaaatttttttatgcttaaaaaaattttccaaaaaaaaaaaatttttttttaggctcaaaaatttttta
11369950    ggcccaaatttttcaggcttaaaaaaattttccaaaaaaaaaaaaatttctttttaggctcaaaaatttttt
11370020    taggcccaaatttttaaggcttaaaaaaattttccaaaaaaaaaaatgtctttttaggctcaaaaatttttt
11370090    tagacccaaatttttttaggcttaaaaaaatttttttttttttaataaatttttttaggctcaaaaatttttta
11370160    ggcccaaatttttcaggcttaaaaaaattttacaaaaaaaaaaatttttttttttaggctcaaaaatttttta
11370230    ggcccaaatttttttaggcttaaaaaaattttacaaaaaaaaaaatttttttttttaggctcaaaaatttttta
11370300    ggcccaaatttttttaggcttaaaaaaatttttccaaaaaaa

Consensus   GGCCCAAATTTTTTAGGCTTAAAAAAATTTT*CAAAAAAAAAA*TTTTTTTTAGGCTCAAAAA*TTTTTA
```

Рис. 3. Разбитый на блоки одинаковой длины, которая равна длине паттерна размытой периодичности, участок из примера 2.

На рисунке 4 приведен результат выравнивания этого участка с почти идеальными копиями повтора. Из него видно, что копии почти не отличаются по длине, но в начале участка имеется

дивергировавший фрагмент конца предыдущей копии паттерна. Показатели качества такого представления периодичности $pl(L) = 0.934$, $H(L) = 7.299$.

```
11369740   ttttttcaagctaaacagaacaaaatttttccaaaaa-aaaaa--tttctttttt-aggctcaaaaatttttttaggcccaaatttttt
11369822   ------------aggcttaaaaaaaatttttcgaaaacaaaaa--tttttttttttaggctcaaaaatttttttaggcccaaatttttt
11369894   ------------atgcttaaaaaaaatttttccaaaaa-aaaaa--attttttttt-aggctcaaaaatttttttaggcccaaatttttc
11369964   ------------aggcttaaaaaaaatttttccaaaaa-aaaaaaatttctttttt-aggctcaaaaatttttttaggcccaaatttttta
11370036   ------------aggcttaaaaaaaatttttccaaaaa-aaaaa--tgtctttttt-aggctcaaaaatttttttagacccaaatttttt
11370106   ------------aggcttaaaaaaaattttttttttttt-aataa--att-ttttt-aggctcaaaaatttttt-aggcccaaatttttc
11370174   ------------aggcttaaaaaaaattttacaaaaa-aaaaa--ttttttttt-aggctcaaaaatttttttaggcccaaatttttt
11370244   ------------aggcttaaaaaaaattttacaaaaa-aaaaa--ttttttttt-aggctcaaaaatttttttaggcccaaatttttt
11370314   ------------aggcttaaaaaaaatttttccaaaa-aa

Consensus  ------------AGGCTTAAAAAAATTTTCCAAAAA-AAAAA--TTTTTTTTT-AGGCTCAAAAATTTTTTAGGCCCAAATTTTTT
```

Рис. 4. Результат работы модифицированного алгоритма обертывающего выравнивания для участка, приведенного на рис. 3.

Приведенные примеры демонстрируют тот факт, что в одной и той же геномной последовательности есть периодичности с различной степенью дивергенции копий паттерна. А есть такие, в которых паттерн испорчен слабо. Возможно последние находятся под давлением отбора или появились в геноме недавно. Это относится к примеру 2. Другие периодичности, примером тому первый приведенный нами пример, несут в себе явные следы длинной мутационной истории. Изучение механизмов разрушения копий паттерна, связи истории развития копий паттерна с длиной паттерна периодичности, а также установление начального паттерна в сильно размытых тандемных повторах является сложной задачей.

4. Обсуждение

Мы создали программное обеспечение для обнаружения нового вида периодичности в геномах – размытых мегасателлитов с переменной длиной паттерна. Это тандемные (идущие непрерывно друг за другом) повторы с длиной паттерна от 1000 нуклеотидов. В сочетании с имеющейся в нашем распоряжении вычислительной технологией, которая хорошо обнаруживает размытые микро-, мини- и макросателлиты, это программное обеспечение позволит получать новые данные о некодирующей части генома, где может локализоваться большое количество периодичностей любого вида. Это приблизит нас к пониманию их места и истинной роли в геноме.

Роль эта может быть очень значительной. Возможно, что мегасателлиты – это резерв генома для оперативного формирования механизмов защиты организма от угроз путем внутригеномного переноса генетического материала. Они могут служить для горизонтального переноса генетического материала между организмами. Есть свидетельства, что периодические участки обеспечивают специфические формы укладки хроматина, что является еще одним механизмом регуляции экспрессии генов, когда одни гены экспонируются на поверхность хроматина, а другие, наоборот, прячутся внутрь. Известно, что места локализации мегасателлитов в геномах одного вида консервативны, непостоянно лишь число копий, этот вид изменчивости индивидуумов одного вида называется Variable Copy Number Polymorphism (VCN Polymorphism). Индивидуумы мало отличаются по нуклеотидному составу, они отличаются структурными вариациями геномов. Возможно также, что и в процессе развития одного индивида число копий в мегасателлитах меняется, что может быть еще одним механизмом регуляции генов развития. Тандемные повторы, которые находятся в межгенных промежутках – это временно молчащие блоки для накопления мутаций, эволюционные резервуары. В них могут созревать новые для эволюции генома элементы.

5. Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00996.

6. Список литературы

1. Кунин Е. *Логика случая. О природе и происхождении биологической эволюции*. М.: ЗАО Издательство Центрполиграф, 2017. 527 с.
2. Чалей М.Б., Кутыркин В.А., Теплухина Е.И., Тюльбашева Г.Э., Назипова Н.Н. *Математическая биология и биоинформатика*. 2013. V. 8. № 2. P. 480–501. doi: 10.17537/2013.8.480.
3. Chaley M., Kutyrkin V., Tulbasheva G., Teplukhina E., Nazipova N. *Database: the journal of biological databases and curation.* 2014. V. 2014. P. 1-18. doi: 10.1093/database/bau040.
4. Chaley M.B., Nazipova N.N., Kutyrkin V.A. *Pattern Recogn. Image Anal.* 2009. V. 19. P. 358–367. doi: 10.1134/S1054661809020217.
5. Tetuev R.K., Nazipova N.N. *Repbase Reports*. 2010. V. 10. № 5. P. 776–776.
6. Tetuev R.K., Dedus F.F., Nazipova N.N. *Repbase Reports*. 2010. V. 10. № 8. P. 1185–1185.
7. Pyatkov M.I., Filippov V.V., Pankratov A.N. *Repbase Reports*. 2012. V. 12. № 3. P. 256–256.
8. Burrows M., Wheeler D.J. *A Blocksorting Lossless Data Compression Algorithm*. Palo Alto, Calif: 1994. (SRC Research Reports, Vol. 124).

9. Li H., Durbin R. *Bioinformatics*. 2009. V. 25. P. 1754–1760. doi: 10.1093/bioinformatics/btp324.
10. Needleman S.B., Wunsch C.D. *J. Mol. Biol*. 1970. V. 48. № 3. P. 443–453. doi: 10.1016/0022-2836(70)90057-4.
11. Hon W., Lam T., Sadakane K., Sung W., Yiu S. *Algorithmica*. 2007. V. 48. P. 23–36. doi: 10.1007/s00453-006-1228-8.
12. Ferragina P., Manzini G. *Proc. IEEE Symposium on Foundations of Computer Science*. 2000. P. 390–398.
13. Benson G. *J. Comput. Biol*. 1997. V. 4. № 3. P. 351–367.