Математическая биология и биоинформатика. 2010. Т. 5. № 1. С. 30-42.
URL: http://www.matbio.org/downloads/Tetuev2010(5_30).pdf

========== БИОИНФОРМАТИКА ==========

УДК: 612. 017:612.12+616.12:616.45

Поиск мегасателлитных тандемных повторов в геномах эукариот по оценке осцилляций кривых GC-содержания

©2010 Тетуев Р.К.[1*], Назипова Н.Н.[1], Панкратов А.Н.[1, 2], Дедус Ф.Ф.[1, 2]

[1]*Учреждение Российской академии наук Институт математических проблем биологии РАН*
[2]*Московский государственный университет им. М.В. Ломоносова, факультет вычислительной математики и кибернетики*

Аннотация. Разработан эффективный метод для решения задачи распознавания участков протяженных (длиной от 1000 н. п.) размытых тандемных сегментных дупликаций в геномах высших эукариот. Основу метода составляет многократное сканирование генома с использованием скользящего окна с длинами рамки, равными степеням двойки начиная с 256 н. п. Для каждого окна подсчитывается процент GC-содержания, а последовательные значения этой характеристики определяют GC-профиль. Создано программное обеспечение, которое выявляет участки устойчивых осцилляций GC-профиля и определяет характеристики обуславливающих эти осцилляции паттернов периодичности. Преимущества нового подхода, использующего комбинацию численно-аналитических методов, позволили выявить мегасателлитные участки в геноме мыши.

Ключевые слова: *размытые тандемные повторы, тандемные сегментные дупликации, геномы эукариот, осцилляции GC-профиля, мегасателлиты в геноме мыши.*

ВВЕДЕНИЕ

Самым значимым достижением последнего десятилетия является секвенирование генома человека и модельных эукариотических организмов. Появившиеся в распоряжении исследователей огромные хранилища генетической информации позволяют проводить полномасштабные исследования геномов и продвигаться вперед в понимании структурно-функциональной организации геномов.

Повторяющиеся фрагменты самой разной природы (дупликации целых геномов, сегментные дупликации, повторы генов) в геномах различных организмов встречаются достаточно часто. Конкретные механизмы появления повторов могут быть разными (неравный кроссинговер, скользящая гиперрепликация, обратная транскрипция, транспозиция и т. д.), но сам процесс размножения повторов представляет собой типичный автокаталитический процесс. Различают два типа повторяющейся ДНК: разнесенные (dispersed) и тандемные (tandem) повторы.

* ruslan.tetuev@gmail.com

Разнесенные повторы, которые не являются паралогичными копиями генов, считаются результатом действия мобильных генетических элементов, то есть таких элементов ДНК, которые могут вырезаться из одного места генома и встраиваться в другие места генома или другой геном. При этом вырезание идет по одной цепи ДНК, на другой цепи комплементарный фрагмент остается на месте. Образовавшаяся в первой цепи брешь заделывается путем синтеза комплементарной копии участка второй цепи (механизм репарации ДНК). Получается, что на исходном месте все остается, как было, а копия мобильного элемента встраивается в произвольное место генома. Там тоже происходит репарация комплементарной нити ДНК. Таким образом происходит удвоение мобильного участка ДНК. Мобильные фрагменты геномов называются *транспозонами*. Транспозоны осуществляют механизмы горизонтального и вертикального переноса генетической информации.

Наименее понятными являются периодические участки геномов. Молекулы ДНК постоянно испытывают на себе действие различных мутационных процессов. Одним из наименее изученных типов мутаций считается копирование фрагментов ДНК, когда один кусок последовательности (образец) дублируется несколько раз, при этом начало каждой копии образца точно пристраивается в конец предыдущей копии. Результатом такой тандемной дупликации образца является участок последовательности, который называется *тандемным повтором*.

Тандемный повтор характеризуется длиной (длина образца) и кратностью. Каждая копия образца (и сам образец) независимо подвергаются дальнейшему мутированию (заменам, вставкам, удалениям отдельных элементов последовательности, новым тандемным дупликациям внутренних фрагментов и др.). Через определенное время каждая копия повтора дивергирует, точный тандемный повтор становится размытым. По уровню сохранности образца различают *точные*, *несовершенные* (неточные) и *скрытые* (размытые) тандемные повторы. Неточные повторы отличаются от точных наличием замен при одинаковой длине повторяющихся паттернов. Размытые тандемные повторы - это неточные тандемные повторы, искаженные вставками и делециями символов. Т. е. у размытых тандемных повторов паттерны имеют не только замены символов, но и варьирующуюся длину.

Примером точной периодичности могут служить следующие друг за другом повторы олигонуклеотида $(ATCGCT)_n$, где n может принимать значения от нескольких единиц до нескольких десятков: ATCGCT°ATCGCT…ATCGCT. Такая периодичность, особенно если она имеет небольшие длины повторов, легко обнаруживается невооруженным глазом.

Несовершенная периодичность возникает тогда, когда в повторяющихся образцах в различных позициях появляются одиночные замены символов, например: ATCGCT°ATGGCT°ATCGCT°ATCCCT. Несмотря на неточность повторов, несовершенная периодичность все еще видна невооруженным глазом.

В случае же размытой периодичности вычленить повторяющийся мотив трудно. Скрытую периодичность можно идентифицировать лишь на основе анализа частот оснований (или аминокислот в белках) в отдельных позициях периода. Причем вовсе необязательно преобладание какого-то одного типа оснований в таких позициях, но преобладающими могут быть группы типов оснований. Пример скрытой периодичности - $\{(A/G)_3N(C/T)(G/C/T)(T/A)\}_n$, где A - аденин, G - гуанин, C - цитозин, T - тирозин, N - равновероятно любая из четырех букв, $(A/G)_n$ означает, что с одинаковой частотой могут встречаться символы A или G несколько раз (n раз),

Существует классификация тандемных повторов по длине образца. Выделяют микросателлиты (SSR – simple sequence repeats, длина мотива не больше 6 нуклеотидных пар (н. п.)), минисателлиты (от 7 до 100 н. п.), сателлиты (от 100 н. п. до

1000 н. п.) и мегасателлиты. Длина повторяющегося элемента у последних выше 1000 н. п.

Микросателлиты населяют центромерные и теломерные области хромосом, их главная функция – участие в образовании гетерохроматина и сегрегации хромосом. Более 12-ти неврологических генетических болезней человека связано с тандемными тринуклеотидными повторами [1 - 5]. В нормальной популяции такие повторы разнообразны и относительно коротки. У больных людей они имеют кратность от 5 до 2000 копий в зависимости от локуса, вызывающего заболевание.

Иногда с наличием микросателлитов в белках связывают конкретные функции (связывание субстрата, межбелковые взаимодействия), влияние на вид пространственной структуры или ее характерные свойства, например, такие, как гибкость белков или отдельных участков ДНК. Есть данные об усилении экспрессии генов, обусловленной наличием вблизи начала гена определенного вида динуклеотидных микросателлитных последовательностей [6]. Кроме того, с тандемными повторами связывают такие болезни, как рассеянный склероз [7], болезнь Альцгеймера [8], шизофрения [9], и рак [10]. Для всех этих случаев появление повторов в определенной части генома означает патологию.

Основным применением микро- и минисателлитов является ДНК-типирование (определение иммунологической принадлежности клетки, ткани или организма по результатам анализа состава антигенов) в криминалистике [11].

Сателлиты – это, чаще всего, кластеры генов. Известны некоторые виды генов, которые кодируются несколькими копиями, расположенными тандемно. Считается, что такая генная организация обеспечивает одновременную экспрессию нужного количества продуктов этих генов. Так располагаются, например, гены рибосомальных РНК и глобиновые гены [12]. Известно [13], что почти треть предсказанных генов на 19-й хромосоме человека сгруппирована в тандемные массивы.

Мегасателлиты – это огромные (размером от 1 до 200 kb) тандемно повторяющиеся сегменты геномов. Их главная функция состоит в обеспечении генетического полиморфизма, они играют фундаментальную роль в передаче по наследству генетических болезней и эволюции генома [14]. Появляются они в геномах в результате внутрихромосомальных рекомбинационных процессов. Мегасателлиты являются менее распространенным видом, так называемых, сегментных дупликаций (segmental duplications), наличие которых в геномах было обнаружено недавно сначала в геноме человека [15], а затем и в геномах других млекопитающих и растениях.

Повторяющиеся сегменты по содержанию не отличаются от остальной части генома – они содержат внутри себя и высококопийные повторяющиеся последовательности, и гены с экзон-интронной структурой. Большинство сегментных дупликаций, известных на момент появления первой сборки генома человека, были открыты экспериментально [14].

Полные геномы модельных эукариотических организмов один за другим стали появляться, начиная с 2000 года [16], в то же время стали активно развиваться и продолжают развитие методы поиска повторяющихся фрагментов геномов, в частности методы обнаружения периодичностей (тандемных повторов) в полных геномах.

Самыми известными и общепризнанными программными разработками для выявления тандемных повторов в ДНК являются Tandem Repeat Finder (TRF) [17] и RepeatMasker [18]. Именно они используются самой компетентной на сегодняшний день базой данных по структурным элементам геномов UCSC [19] для картирования всех видов периодичностей, исключая мегасателлиты. У каждой их этих разработок своя область применения и свои ограничения на длину паттерна. Наиболее универсальным является спектрально-статистический подход [20, 21], он выявляет

тандемные повторы с паттерном периодичности любой длины от 3 н. п. до нескольких тысяч нуклеотидов.

До сих пор не существует специализированного метода выявления размытых мегасателлитных участков геномов. Созданы методы распознавания сегментных дупликаций (т. е. разнесенных протяженных повторов). Предполагается, что этими методами можно обнаруживать и тандемные сегментные дупликации. Один из наиболее известных методов [14] основан на нарезании генома на большие куски и последующем выравнивании каждого куска со всеми остальными. При этом отмечаются все пары участков длиной более 1 kb и имеющие хорошее (более 90% сходства) выравнивание. С помощью этого метода были выявлены сегментные дупликации, которые в геноме мыши (Build36) занимают 4.94% его длины, а в геноме человека (Build36) - 5.52%. На тандемные дупликации у мыши по данным авторов приходится 35.2% от общего числа повторов, у человека – 21.6%.

Ограничением такого рода методов, основанных на широкомасштабном применении алгоритма BLAST, является то, что они неплохо работают при обнаружении точных и несовершенных разнесенных повторов, но непригодны для поиска размытых мегасателлитных участков в геномах.

Нами разработан достаточно эффективный и перспективный метод для решения задачи распознавания участков протяженных (длиной от 1000 н. п.) размытых тандемных сегментных дупликаций. Преимущества этого метода, использующего комбинацию численно-аналитических подходов, позволили получить результаты, которые не обнаружились другими методами. Наш метод специально разработан для применения к анализу генетических текстов, имеющих большое количество несовпадений и существенные искажения генетического текста путем вставок фрагментов.

ОПИСАНИЕ АЛГОРИТМА

В основу нашего метода [22] положено использование такой широко используемой характеристики последовательности ДНК, как ее GC-содержание. Если представить молекулу ДНК длиной N нуклеотидов в виде последовательности символов $S°=°x_1x_2\ldots x_i\ldots x_{i+W}\ldots x_N$ и обозначить количество букв G и C в подпоследовательности $x_i\ldots x_{i+W-1}$ как $F_W^{(G,C)}(x_i)$, то функция процентного содержания G и C для окна длины W в позиции x_i будет:

$$f_W^{(G,C)}(x_i) = F_W^{(G,C)}(x_i)/W \times 100\%,\ i = 1,\ldots, N\text{-}W\text{+}1.$$

График функции $f_W^{(G,C)}$ на участках, соответствующих местам тандемных повторов, обычно имеет выраженный осциллирующий вид. Особенно ярко это проявляется в тех случаях, когда длина периода этих осцилляций ν принадлежит диапазону $W/4 < \nu < W/2$. Основываясь на этих наблюдениях, можно искать места возможных тандемных повторов, ориентируясь на поиск участков осцилляции функции $f_W^{(G,C)}$.

Известно, что многие осциллирующие кривые приближенно описываются графиком гармонических колебаний:

$$f_W^{(G,C)}(x_i) \approx a_0 + A \cdot \sin\left(2\pi \frac{x_i - \varphi}{\lambda}\right),$$

где параметрами a_0, A, φ, λ, соответственно, являются среднее значение, амплитуда колебаний, фазовый сдвиг и длина периода. На рис. 1 представлен пример приближения функции $f_{4096}^{(G,C)}$, вычисленной для третьей хромосомы мыши на участке

113 040 000 … 113 175 000. Видно, что поведение функции хорошо описывается гармоническим колебанием:

$$f \approx 40\% + 10\% \cdot \sin\left(2\pi \frac{x_i - 113\,060\,000}{33\,000}\right).$$

Интересно, что период повтора и амплитуда колебаний этих фрагментов в действительности довольно сильно варьируются в абсолютных величинах, но это почти никак не отражается на форме функции. Последнее объясняется тем, что даже наличие в каждом участке около 350 - 500 точечных мутаций при ширине окна в 4096 пар нуклеотидов не способно сильно изменить относительные и интегральные оценки.

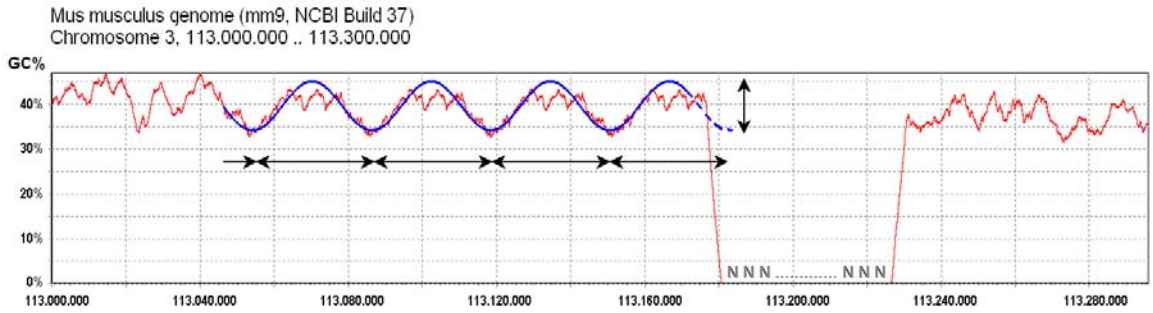


Рис.°1. Пример приближения функции $f_{4096}^{(G,C)}$ функцией $f \approx 40\% + 10\% \cdot \sin\left(2\pi \frac{x_i - 113\,060\,000}{33\,000}\right)$ на участке 3-й хромосомы *Mus musculus*. В районе 113°180°000 - 113°225°000 видна неопределенность (gap) в последовательности генома, возможно, на этом участке хромосомы периодичность имеет продолжение.

Ясно, что при поиске повторов с разными периодами ширина окна W выступает в качестве параметра, который можно выбирать, например, из ряда значений: $256, 512, 1024, \ldots$. Для автоматического сканирования ДНК необходимо ввести количественную оценку степени осцилляции кривых, которая будет вычисляться для каждого фрагмента функции $f_W^{(G,C)}$. Для решения этой задачи используется математический аппарат аналитического приближения функций [23, 24].

Пусть аналитическое приближение исследуемой функции на некотором отрезке (для некоторой подпоследовательности $x_i \ldots x_{j+W-1}$) имеет вид:

$$f_W^{(G,C)}(x_i) \approx a_0 + \tilde{f}(x_i),$$

где $a_0 = \dfrac{\sum_{k=i}^{j} f_W^{(G,C)}(x_k)}{j - i + 1}$ – среднее значение функции $f_W^{(G,C)}$ на заданном интервале. Тогда следует ожидать, что в случае, когда интервал содержит лишь участки осцилляции, функция $\tilde{f}(x_i)$ ведет себя как некоторая зашумленная гармоника:

$$\tilde{f}(x_i) \approx A \sin 2\pi \frac{x_i - \varphi}{\lambda} = A \sin \omega(x_i - \tau),$$

где A, $\omega = \dfrac{2\pi}{\lambda}$, $\tau = \dfrac{2\pi\varphi}{\lambda}$ – параметры гармонического колебания. Зная, что любое аналитически заданное гармоническое колебание

$$f(x) = A \sin \omega(x - \tau)$$

удовлетворяет равенству

$$f''(x) = -\omega^2 f(x),$$

приходим к количественной оценке степени осцилляции функции $f_W^{(G,C)}$, выраженной через коэффициент корреляции функций $\tilde{f}$ и $\tilde{f}'' = -\omega^2 \tilde{f}$. Действительно, в таком случае ожидается выполнение условия:

$$R_{\tilde{f},\tilde{f}''} = \frac{\operatorname{cov}\left(\tilde{f},\tilde{f}''\right)}{\sqrt{D\left[\tilde{f}\right]}\cdot\sqrt{D\left[\tilde{f}''\right]}} \approx -1.$$

Так как последнее условие является необходимым, но не достаточным, алгоритм поиска тандемных повторов следует дополнить *верификацией* полученных результатов. Традиционно этап верификации является наиболее трудоемким и длительным в сравнении с предварительным этапом оценки возможного присутствия повтора. Одним из очевидных преимуществ излагаемого подхода можно считать наличие этапа предварительной оценки, который в большинстве существующих методов отсутствует.

При аппроксимации дискретных данных функциональными рядами наиболее трудной является проблема аналитического описания функций и точного вычисления второй производной. Одним из наиболее популярных аналитических представлений периодических функций является разложение в ряд Фурье:

$$f(x) \approx a_0 + \sum_{n=1}^{N}\left(b_n \sin nx + a_n \cos nx\right).$$

производные функции в таком представлении легко находятся аналитически:

$$f' \approx \sum_{n=1}^{N}\left(-na_n \sin nx + nb_n \cos nx\right),$$
$$f'' \approx -\sum_{n=1}^{N}\left(n^2 b_n \sin nx + n^2 a_n \cos nx\right).$$

Но, к сожалению, такой способ вычисления производных не может иметь практического значения вследствие хорошо известного эффекта Гиббса [25]. В работе [26] описаны дополнительные соображения, вынудившие авторов отказаться от подобного представления в пользу разложения функций в *обобщенный* ряд Фурье по ортогональным полиномам Чебышева I-го рода:

$$f(x) \approx \sum_{n=1}^{N} C_n T_n(x),$$

где

$$T_0(x) = 1,$$
$$T_1(x) = x,$$
$$\ldots$$
$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

Отметим, что при фиксированной длине ряда $n_{const} << W$ такое представление выигрывает, особенно, при работе с несимметричными функциями. На рис. 2 хорошо видна разница между разложениями в ряд Фурье и в ряд Чебышева при длине ряда $n = 20$. Резкий подъем функции на одной (левой) границе привел к появлению характерного «дребезга» и серьезных отклонений в первом описании, в то время, как

разложение по ортогональным полиномам Чебышева аккуратно и точно повторяет все изменения поведения функции.

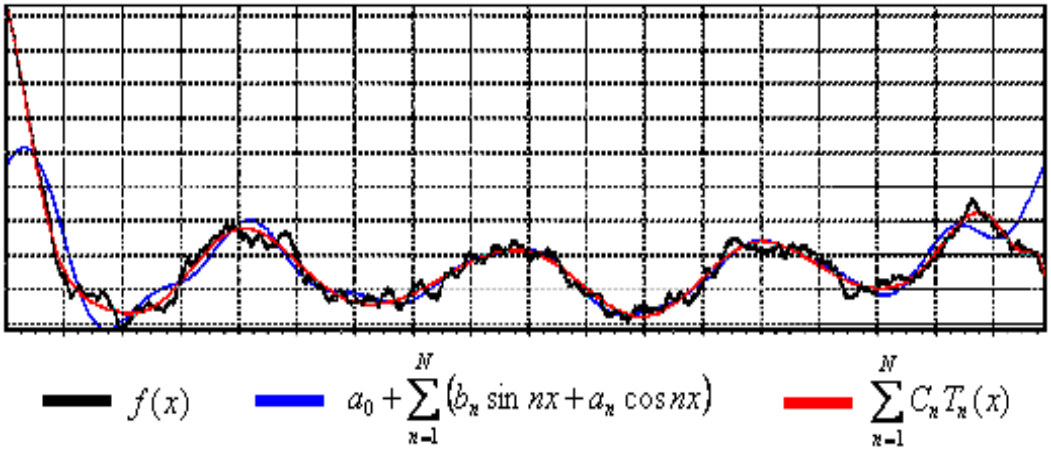


Рис. 2. Сравнение качества аппроксимации периодической функции с помощью ортогональных базисов Фурье и Чебышева I-го рода. Черным цветом изображена исходная функция, синим - аппроксимация с помощью функций Фурье, красным цветом – полиномов Чебышева. На рисунке видны погрешности приближения с помощью рядов Фурье на краях отрезка (эффект Гиббса).

Разложение функций в обобщенный ряд Фурье по полиномам Чебышева I-го рода приводит не только к качественной аппроксимации функций, но и позволяет просто и точно вычислять ее производные [27, 28, 29], делая возможным применение этого подхода на практике. Действительно, степень конечного ряда, полученного после дифференцирования, легко удается понизить:

$$f'(x) = \sum_{n=0}^{N+1} C_n T'_n = 2NC_{N+1}T_N + C_{N+1}T'_{N-1} + \sum_{n=0}^{N} C_n T'_n ,$$

воспользовавшись рекуррентной формулой [30]:

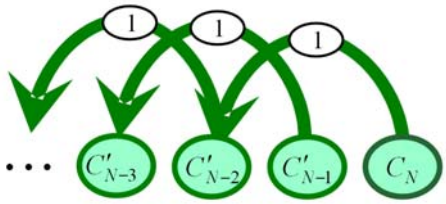
$$T'_{n+1} = T'_{n-1} + 2nT_n .$$

Далее, последовательно применяя этот прием к соответствующим слагаемым, получим итерационную линейную процедуру, используемую для точного определения значений спектра производной $f'(x)$.

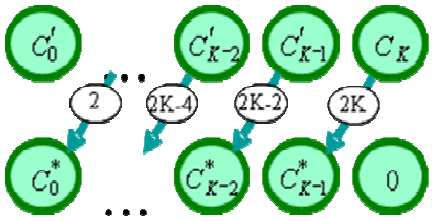
Алгоритм преобразования коэффициентов ряда [28] может быть представлен в виде двух независимых этапов, которые схематически показаны на рис. 3. Эти вычислительные этапы получили название *каскада и диффузии спектра*. На первом шаге (каскад) каждый коэффициент спектра C_0, C_1, C_2,…,C_j, …C_{N-2}, C_{N-1}, C_N, начиная с $j = N - 2$, увеличивается на значение коэффициента с номером $j + 2$. При этом коэффициенты C_{N-1}, C_N остаются неизменными, и ход вычислений производится в направлении от старшего коэффициента спектра к младшему. Второй шаг процедуры (диффузия) сводится к умножению значений спектра на удвоенное значение индекса (т. е. на $2k$, где k – индекс спектрального коэффициента) и к последующему сдвигу всего спектра на одну позицию в сторону нулевого индекса.

Разделение алгоритма преобразования спектральных коэффициентов на два этапа при дифференцировании функций оправдано не только в случае разложения их в ряд по полиномам Чебышева 1-го рода, но и в самом общем случае. Предложенная схема спектрально-аналитического дифференцирования функций (т. е. дифференцирования не самой функции, а ее аппроксимации) хорошо показала себя для всех классических ортогональных полиномов и функций непрерывного аргумента: Якоби, Гегенбауэра, Чебышева (I-го и II-го рода), Лежандра, Лагерра, Сонина-Лагерра (обобщенного

Лагерра), Эрмита [28]. Ранее был известен только частный случай спектрально-аналитического дифференцирования для вычисления производных от функций, аппроксимированных полиномами Гаусса-Чебышева (в отечественной терминологии – Чебышева I-го рода) [30].



А.



Б.

Рис. 3. Схема преобразования спектра функции для вычисления спектра ее производной. Стрелки на схеме обозначают операции взвешенного суммирования: значение-источник прибавляется к значению-приемнику с указанным на стрелке весовым коэффициентом. **А. Спектральный каскад.** К каждому коэффициенту спектра, начиная со второго с конца, прибавляется значение вышестоящего через один коэффициента, при этом ход вычислений производится в направлении от старшего к младшему. **Б. Спектральная диффузия** сводится к умножению каждого коэффициента на удвоенное значение своего индекса и сдвигу всего спектра на одну позицию в сторону младшего коэффициента.

РЕЗУЛЬТАТЫ

Разработанная методика была реализована и тестировалась на геноме лабораторной мыши (Mus musculus, build 37.1, reference assembly C57BL/6J). В качестве доказательства эффективности метода в данной статье приводится лишь несколько примеров его применения.

Один из повторов обнаружен применением нашего подхода на пределе возможностей метода – почти четырехкратный тандемный повтор, найденный в первой хромосоме мыши на участке с координатами, приблизительно равными 172 990°000…173 015 000. Этот повтор имеет длину периода около 7600 нуклеотидов и проявляется при исследовании генома мыши посредством функции $f_{2048}^{(G,C)}$ (рис. 4).

Поскольку наличие в ДНК протяженных многократных тандемных повторов долгое время явного подтверждения не имело, актуальность развития подходов, предназначенных для их обнаружения, вызывала некоторые сомнения. Однако появившиеся в последнее время данные свидетельствуют о том, что подобные повторы в геномах различных организмов отнюдь не редкость. В частности, известно, что большинство сегментных дупликаций в геноме мыши организовано в тандемные кластеры с очень высокой (более 89%) степенью гомологии. В этих гигантских тандемных повторах мало генов, но много ретротранспозонов LINE- и LTR-типа [31].

Нами, наряду с такими периодичностями, обнаружен другой тип тандемного повтора в геноме мыши, который не подпадает под описание авторов вышеупомянутой работы.

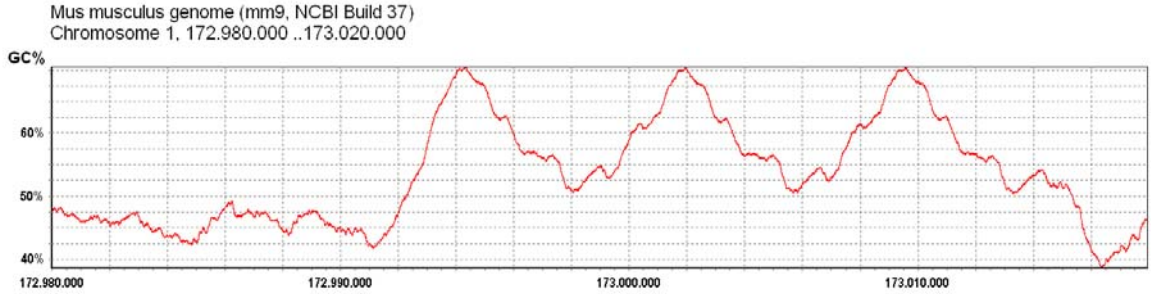


Рис. 4. Вид функции $f_{2048}^{(G,C)}$ на участке 1-й хромосомы *Mus musculus*. Видны четкие осцилляции функции при длине окна W = 2048, которые свидетельствуют о наличии тандемной сегментной дупликации.

С помощью созданного программного обеспечения, реализующего новый подход [35], обнаружен 60-кратный тандемный повтор, располагающийся приблизительно на участке 114 440 000…114 590 000 шестой хромосомы мыши. Данный повтор вызывает протяженную осцилляцию функции $f_{1024}^{(G,C)}$ (рис. 5), имеет период с переменной длиной около 2400 н. п. и общую длину свыше 150000 н. п. Этот участок имеет паттерн периодичности с четкой внутренней организацией, внутри него нет следов ретротранспозонов, но есть протяженные динуклеотидные треки сильно варьирующейся длины. Эти треки разбивают паттерн на 3 консервативные ядра, которые можно представить грузиками, соединенными пружинами с различной степенью натяжения. Наличие простых последовательностей (low complexity треков), очевидно, затрудняет выявление участков такого рода методами, существенно использующими попарное выравнивание сегментов хромосом.

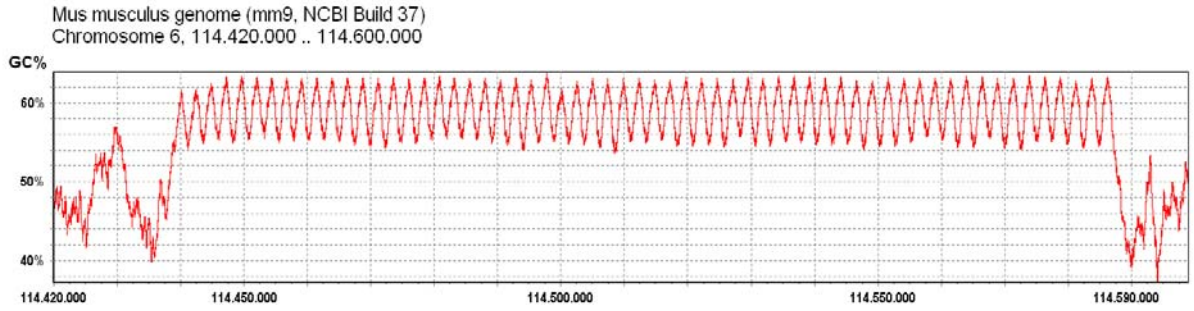


Рис.°5. Вид функции $f_{1024}^{(G,C)}$ на участке 6-й хромосомы *Mus musculus*. Длина окна W = 1024, осцилляции свидетельствуют о наличии тандемной сегментной дупликации кратностью 60 и длиной паттерна периодичности около 2400 н. п.

ОБСУЖДЕНИЕ

С помощью разработанного метода был просканирован геном мыши Mus musculus (сборка NCBI37/mm9, датированная июлем 2007 года, источник: UCSC - University of California, Santa Cruz, Genome Browser database). В результате полногеномного сканирования были найдены многочисленные размытые мегасателлиты, данные о которых не найдены ни в одной из наиболее компетентных баз данных (TRDB [32], UCSC [19]).

Наш метод позволяет находить участки, повторяющиеся фрагменты которых сильно отличаются друг от друга по длине. В качестве примера приведем один из найденных

участков. Он имеет общую протяженность около 150°000 н. п. и длину паттерна около 2400 н. п. При этом отдельные повторяющиеся элементы имеют сильный разброс в длинах (от 2320 до 2450 н. п.). Этот разброс обеспечивается, главным образом, вставками микросателлитных (динуклеотидных тандемных повторов) участков переменной протяженности, которые затрудняют обнаружение какой-либо периодичности более высокого порядка другими методами. В частности, в базе данных UCSC в этом месте генома помечены лишь эти самые динуклеотидные повторы, обнаруженные программой RepeatMasker [18]. Мы обнаружили нашим методом периодическую структуру более высокого порядка, которая целиком укладывается в межгенном промежутке. Периодический характер этого участка 6-й хромосомы мыши наглядно виден на рис. 5. Простые динуклеотидные повторы переменной длины являются ограничителями более консервативных (до 98% сходства) участков. Общая схема структуры обнаруженных паттернов периодичности выглядит следующим образом:

$$(ca)_n \ \mathbf{core1}\ (ct)_m \ \mathbf{core2}\ (ct)_k \ \mathbf{core3},$$

где $12 <= n <= 40$, $30 <= m <= 166$, $k = 12$, а максимальные длины консервативных ядер повтора core1, core2 и core3, соответственно равны 88, 663 и 1566. Паттерны периодичности описанного вида повторяются целиком с небольшими изменениями ровно 60 раз, если не считать обломков по краям этого фрагмента хромосомы. В консервативных ядрах не обнаружены следы транспозонов.

Другая обнаруженная нами область тандемных сегментных дупликаций в геноме мыши локализуется на первой хромосоме, точные координаты ее паттернов:
172°993°815 - 173°001°442 (длина 7628 н. п.),
173°001°443 - 173°009°056 (длина 7613 н. п.),
173°009°057 - 173°016°570 (длина 7513 н. п.).

Попарное сходство этих участков около 97%. Последний фрагмент короче остальных примерно на 1.4%. Кратность повтора - 2.986 при средней длине паттерна повторяемости 7620. Каждая копия имеет сложное строение, паттерн содержит псевдогены тРНК, одну сателлитную последовательность SAT и 8 транспозонов различных типов (эндогенные ретровирусы ERV2 и ERV3, ретротранспозоны SINE и SINE2/tRNA). Всего на 11 повторяющихся элементов, зарегистрированных в базе данных RepBase [33], приходится около 20% протяженности паттерна.

Третий найденный мегасателлит простирается с позиции 113°052°803 до позиции 113°180°645 на 3-й хромосоме мыши. Точные границы повторяющихся паттернов:
113°052°803 - 113°085°385 (длина 32°583 н. п.),
113°085°386 - 113°118°005 (длина 32°620 н. п.),
113°118°006 - 113°150°605 (длина 32°600 н. п.),
113°150°606 - 113°180°645 (длина 30°040 н. п.).

Попарное сходство между паттернами - 99%. Последний сегмент примерно на 9.21% короче остальных. Кратность найденного повтора - 3.91 при средней длине паттерна 32°600 н. п. Он характеризуется тем, что в каждом из повторяющихся паттернов с помощью сервиса CENSOR [34] мы обнаружили, что более 60% паттерна занимают 29 повторяющихся элементов разных типов (эндогенные ретровирусы ERV2 и ERV3, ретротранспозоны L1, SINE1, SINE1/7SL).

Важность поиска и подробной аннотации тандемных повторов большой длины трудно переоценить. Они являются важнейшими носителями генетической изменчивости. Показано, что в некоторых определенных местах геномов находится непостоянное число тандемных повторов с различной длиной повторяющегося фрагмента (STRs – Short Tandem Repeats, VNTRs - Variable Number of Tandem Repeats и CNVs – Copy Number Variations) и непостоянным значением кратности. Это означает,

что число повторений мотива является переменной величиной, оно не постоянно внутри популяции (полиморфизм), не передается по наследству от предка к потомку (генетическая нестабильность), не постоянно даже внутри одного организма (мозаичность). Определение числа копий тандемных повторов иногда играет критическую роль, т. к. некоторые генетические болезни связаны именно с избыточным количеством копий того или иного мотива в определенном месте генома.

Приведенные примеры показывают, что новый метод обнаружения тандемных сегментных дупликаций позволяет эффективно находить участки скрытой периодичности различной степени сохранности. Это могут быть и неточные тандемные дупликации хромосомных участков, и сильно испорченные полинуклеотидными вставками повторяющиеся образцы. Несмотря на безусловную эффективность излагаемого подхода у него все же имеется ряд характерных ограничений. В частности, вероятность обнаружения тандемных повторов тем выше, чем больше их кратность, для эффективного распознавания желательно иметь не менее 3-4-х повторений мотива. В противном случае накладываются дополнительные ограничения на количество возможных мутаций и т. п.

Еще одно тонкое место разработанного метода – выбор длины окна сканирования генома. Длина сканирования генома W накладывает ограничения на длину найденных паттернов периодичности. Если длина паттерна меньше или больше длины окна, то такой тандемный повтор может быть при сканировании генома пропущен. Многократное сканирование геномов с длинами окна, равными последовательным степеням двойки, начиная с минимальной длины 256 н. п., исключает эту возможность.

По протяженным районам периодичности можно, кроме выявления предрасположенности к различного рода генетическим болезням, перечисленным выше, реконструировать эволюционную историю генома, можно изучать дифференциацию между отдельными индивидуумами и географически или по времени изолированными популяциями. Тандемные повторы могут служить генетическими маркерами (участками ДНК с известной локализацией), используя которые можно изучать эпидемии инфекционных болезней. Сравнение обнаруженных у различных организмов паттернов периодичности может позволить открыть специфические для группы организмов мегасателлиты. Не исключена также возможность открытия совсем еще не известной функции, которая возложена природой на протяженные тандемные участки скрытой периодичности ДНК. Работа в этом направлении уже ведется.

Работа частично поддержана грантами РФФИ № 09-07-00455, № 10-07-00112, № 10-01-00609 и № 08-07-00353.

ЛИТЕРАТУРА

1. Kramer P.R. and Pearson C.E. Stability of triplet repeats of myotonic dystrophy and fragile X loci in human mutator mismatch repair cell lines. *Hum. Genet*. 1996. V. 98. P. 151-157.
2. Mitas M. Trinucleotide repeats associated with human disease. *Nucleic Acids Res*. 1997. V. 25. P. 2245-2253.
3. Ranum PW.L. and Day W.J. Dominantly inherited, non-coding microsatellite expansion disorders. *Curr. Opin. Genet. Dev*. 2002. V. 12. P. 266-271.
4. Richards R.I., Holman K., Yu S. and Sutherland G.R. Fragile X syndrome unstable element, P(CCG)N, and other simple tandem repeat sequences are binding-sites for specific nuclear proteins. *Hum. Mol. Genet*. 1993. V. 2. P. 1429-1435.
5. Sutherland G.R. and Richards I.R. Simple tandem DNA repeats and human genetic disease. *Proc. Natl. Acad. Sci. USA*. 1995. V. 92. P. 3636-3641.

6. Hamada H., Seidman M., Howard B. and Gorman C. Enhanced gene-expression by the poly(DT-DG) poly (DC-DA) sequence. *Mol. Cell. Biol*. 1984. V. 4. P. 2622-2630.
7. Guerini F.R. et al. Myelin basis protein gene is associated with ms in DR4- and DR5-positive Italians and Russians. *Neurology*. 2003. V. 61. P. 520–526.
8. Licastro F. et al. Interleukin-6 gene alleles affect the risk of Alzheimer's disease and levels of the cytokine in blood and brain. *Neurobiol. Aging*. 2003. V. 24. P. 921–926.
9. Brzustowicz L.M. et al. Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21–q22. *Science*. 2000. V. 288. P. 678–682.
10. Sidransky D. Nucleic acid-based methods for the detection of cancer. *Science*. 1997. V. 278. P. 1054–1058.
11. Butler J. Forensic *DNA Typing: Biology and Technology Behind STR Markers*. London: Academic Press, 2003.
12. Льюин Б. *Гены*. М.: Мир, 1987. 544 с.
13. Kim J et al Homology-driven assembly of sequence-ready mouse BAC contig map spanning regions related to the 46-Mb gene rich euchromatic segments of human chromosome 19. *Genomics*. 2001. V. 74. P. 129-141.
14. Bailey J.A., Yavor A.M., Massa H.F., Trasl B.J., Eichler E.E. Segmental Duplications: Organization and Impact within the Current Human Genome Project Assembly. *Genome Res*. 2001. V. 11. P. 1005-1017.
15. Eichler E.E. Masquerading repeats: Paralogous pitfalls of the Human Genome. *Genome Res*. 1998. V. 8. P. 758-762.
16. Venter J.C. et al. The sequence of the human genome. *Science*. 2001. V. 291. № 5507. P. 1304-1351.
17. Benson G. Tandem repeat finder: a program to analyse DNA sequences. *Nucleic Acids Res*. 1999. V. 27. P. 573-580. URL: http://tandem.bu.edu/trf/trf.html (дата обращения: 20.04.2010).
18. Smit A.F.A., Hubley R. & Green P. *RepeatMasker*. URL: http://repeatmasker.org/ (дата обращения: 20.04.2010).
19. UCSC Genome Bioinformatics Site. URL: http://genome.ucsc.edu/ (дата обращения: 20.04.2010).
20. Чалей М.Б., Назипова Н.Н., Кутыркин В.А. Совместное использование различных критериев проверки однородности для выявления скрытой периодичности в биологических последовательностях. *Математическая биология и биоинформатика* (электронный журнал). 2007. Т. 2. № 1. С. 20-35. URL: http://www.matbio.org/downloads/Chaley2007(2_20).pdf (дата обращения: 20.04.2010).
21. Chaley M.B., Nazipova N.N., Kutyrkin V.A. Statistical Methods for Detecting Latent Periodicity Patterns in Biological Sequences: The Case of Small-Size Samples. *Pattern Recognition and Image Analysis*. 2009. V. 19. №. 2. P. 358-367.
22. Pankratov A.N., Gorchakov M.A., Dedus F.F., Dolotova N.S., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Novikova D.A., Olshevets M.M., Pyatkov M.I., Rudnev V.R., Tetuev R.K. and Filippov V.V. Spectral Analysis for Identification and Visualization of Repeats in Genetic Sequences. *Pattern Recognition and Image Analysis*. 2009. V. 19. №. 4. P. 687–692.
23. Колмогоров А.Н., Фомин С.В. *Элементы теории функций и функционального анализа*. М.: Наука, 1968.
24. Дедус Ф.Ф., Махортых С.А., Устинин М.Н., Дедус А.Ф. *Обобщенный спектрально-аналитический метод обработки информационных массивов*. М.: Машиностроение, 1999. 356 с.
25. Болтнев А.А., Калиткин Н.Н., Качер О.А. Эффект Гиббса в разностных схемах. *ДАН*. 2006. Т. 411. № 5. С. 594-598.

26. Дедус Ф.Ф., Куликова Л.И., Махортых С.А., Назипова Н.Н., Панкратов А.Н., Тетуев Р.К. Аналитические методы распознавания повторяющихся структур в геномах. *ДАН*. 2006. Т. 411. № 5. С. 599-602.
27. Дедус Ф.Ф., Куликова Л.И., Панкратов А.Н., Тетуев Р.К. *Классические ортогональные базисы в задачах аналитического описания и обработки информационных сигналов.* М: Издат. отд. фак. ВМиК МГУ им. Ломоносова, 2004. 172 с.
28. Тетуев Р.К., Дедус Ф.Ф. *Классические ортогональные полиномы. Применение в задачах обработки данных*. Пущино: препринт ИМПБ РАН, 2007.
29. Новикова Д.А., Поволоцкий А.В. Формулы для преобразования функций в пространстве коэффициентов разложения по базису Чебышева 2-го рода. *Сборник статей молодых ученых факультета ВМиК МГУ*. 2007. № 4. С. 1-8.
30. Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. Numerical Recipes in C: The Art of Scientific Computing. Second Edition. Cambridge University Press, 1997. P. 195-196.
31. She X., Cheng Ze, Zőllner S., Church D.M., Eichler E.E. Mouse Segmental Duplication and Copy-Number Variation. *Nat. Genet*. 2008. V. 40. P. 909-914.
32. Gelfand Ye., Rodriguez A., Benson G. TRDB—The Tandem Repeats Database. *Nucleic Acids Res*. 2007. V. 35. Database issue. P. D80–D87. URL: https://tandem.bu.edu/cgi-bin/trdb/trdb.exe (дата обращения: 20.04.2010).
33. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research.* 2005. V. 110. P. 462-467.
34. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*. 2006. V. 7. P. 474.
35. Тетуев Р.К., Дедус Ф.Ф., Назипова Н.Н., Махортых С.А., Куликова Л.И., Панкратов А.Н., Ольшевец М.М. *Спектральный анализ данных, поиск неточных периодов в сигналах «SpectralRevisor».* Свидетельство Роспатента об официальной регистрации программы для ЭВМ № 2007611639. 2007.

Материал поступил в редакцию 26.04.2010, опубликован 04.05.2010.