Математическая биология и биоинформатика. 2012. Т. 7. №2. С. 476–492.
URL: http://www.matbio.org/2012/Pankratov_7_476.pdf

==========БИОИНФОРМАТИКА==========

УДК:519

Поиск протяженных повторов в геномах на основе спектрально-аналитического метода

©2012 Панкратов А.Н.*[1,2], Пятков М.И.[1], Тетуев Р.К.[1], Назипова Н.Н.[1], Дедус Ф.Ф.[2,1]

[1]*Институт математических проблем биологии РАН, 142290, Пущино, ул.Институтская, д.4*
[2]*Факультет ВМК МГУ имени М.В.Ломоносова, 119991 ГСП-1, Москва, Ленинские горы*

Аннотация. Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном интегральном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Ключевые слова: *сравнение геномов, аппроксимация, матрица сходства, распознавание образов, мегасателлиты, разнесенные повторы*

ВВЕДЕНИЕ

Поиск повторяющихся последовательностей является одной из задач, которые решает биоинформатика. Множество различных типов повторов, отличающихся друг от друга механизмами распространения, пространственной ориентацией относительно друг друга, содержащие как кластеры генов и отдельные регуляторные элементы, так и некодирующие участки, составляют 50% генома [1].

Различают два вида повторяющейся ДНК: разнесенные и тандемные повторы. Разнесенные, или диспергированные, повторы обязаны своим происхождением действию мобильных генетических элементов (транспозонов). Наличие таких повторов может иметь различные последствия для генома. Перемещение участка генома на новое место может инактивировать соседние гены или иначе влять на их активность. Разбросанные по геному копии транспозонов, независимо друг от друга подвергаются мутированию, создавая "горячие точки"для реализации генетического разнообразия и изменчивости генома.

Тандемные повторы являются результатом дублирования фрагментов ДНК, когда копия фрагмента следует сразу за образцом. В зависимости от размера образца тандемные повторы подразделяются на четыре класса – мегасателлиты, сателлитная

*pan@impb.psn.ru

ДНК, минисателлиты и микросателлиты. Мегасателлиты имеют длину паттерна периодичности свыше 1000 нуклеотидных пар (н.п.). Длина сателлитной ДНК может составлять от 100 тысяч до более 1 миллиона нуклеотидов с длиной повторяющегося мотива от 100 до 1000 н.п. Минисателлиты – тандемно повторяющиеся фрагменты ДНК длиной от 7 до 100 нуклеотидов. Большая часть сателлитов и минисателлитов локализована в гетерохроматиновых областях хромосом. Микросателлиты (или простые короткие тандемные повторы)– это повторяющиеся фрагменты ДНК длиной до 6 н.п.

Появление полных геномов разных видов организмов открывает широкие возможности для их сравнительных исследований. Интригующей задачей является выяснение причин морфологических различий видов млекопитающих, а также фенотипических различий особей в популяции при практически одинаковом наборе генов. В настоящее время укрепилось мнение, о том, что механизмы изменчивости у млекопитающих связаны с цис-регуляторными районами - фрагментами ДНК, которые находятся вне кодирующих областей и включают и выключают гены. В качестве таких регуляторов экспрессии генов часто выступают тандемные повторы.

Совсем недавно повторы считались не функциональной частью генома, а "мусорной"("эгоистичной") ДНК. Обнаружение в геномах носителей внутривидового разнообразия организмов, таких, как VNTRs и CNVs, добавили знаний о природе изменчивости. VNTRs (Variable Number of Tandem Repeats) - массивы тандемных повторов с консервативным паттерном длиной от 5 до 100 н.п. и строго определенным местоположением в геноме, количество копий в которых индивидуально для каждой особи в популяции или для разных хромосом у одной особи. CNVs (Copy Number Variations) - вид генетической изменчивости, связанной с различным числом копий мегасателлитных фрагментов генома, содержащих в себе и гены, и регуляторные элементы, и спейсеры. Считается, что роль тандемных повторов состоит в том, чтобы обеспечивать в данный момент времени и места такую укладку эухроматина, чтобы участки транскрибируемой ДНК сформировали специфический паттерн экспрессии [2].

В 2004 году [3] впервые были представлены подтверждения гипотезы о прямой связи между длиной паттернов периодичности и морфологической вариабельностью видов. Тандемные повторы распространены в кодирующих последовательностях генов позвоночных, особенно в генах, обеспечивающих развитие. Ортологичные районы периодичности часто оказываются законсервированными даже у удаленных таксонов. Применением знаний о повторяющихся фрагментах геномов может быть использование их в задачах систематики биологических организмов. Таксоноспецифичность определенных повторяющихся участков генома позволяет уточнить существующую классификацию организмов, используя геномные данные. Например, до сих пор идет дискуссия о необходимости выделения грызунов вида *Cavia porcellus* в парвотряд *Caviamorpha* [11, 12]. Интересно отметить, что впервые аргументы обе стороны получают при исследовании геномов, до появления возможности исследования на молекулярном уровне спор о выделении отряда зайцеобразных [13] был разрешён при исследовании морфологических признаков. При наличии отсеквенированных геномов задачу классификации организмов можно решать на уровне похромосомного сравнения последовательностей ДНК, используя цифровые методы обработки генетических текстов.

На данный момент разработано множество алгоритмов и программ [32, 33, 34, 35] для вычислительной оценки подобия фрагментов ДНК и её производных (белков, РНК). Однако, за последние 20 лет наметилось отставание возможностей вычислительного анализа генетических данных от стремительно развивающейся экспериментальной

базы в современной биологии: большинство алгоритмов обработки генетических последовательностей основаны на нескольких базовых принципах обработки текстовой информации, таких как расстояние Хэмминга или редакционное расстояние Левенштейна [36]. Временная сложность алгоритмов существенно нелинейна, главным замедляющим фактором при сравнении схожих генетических данных являются точечные мутации (типа замен, изъятий и вставок букв), "исправление" которых увеличивает время анализа. При больших размерах сравниваемых фрагментов разумно предполагать большее количество мутаций, и, как следствие, наблюдать резкое снижение эффективности алгоритмов на больших участках (длиной от 10000 н.п.). Такое положение вещей существует, если следовать по общепринятому пути дискретного анализа. В настоящей работе предлагается метод, в котором существенно используются идеи из области непрерывного анализа данных.

Предложение использовать спектральный подход к задаче поиска повторов в геномах было впервые сделано в работах [18, 19], в которых были выдвинуты идеи использовать непрерывные методы для анализа цифровых последовательностей. В качестве функционального аналога геномной последовательности было предложено использовать кривые GC-содержания, которые характеризуют силу связывания двойной спирали ДНК. Следующим этапом развития метода стало построение матрицы сходства и вывод решающего правила при распознавания неточных повторов, основанного на среднеквадратичном отклонении проекций кривых на первые базисные функции [20]. Для поиска протяженных тандемных повторов был предложен алгоритм [21], основанный на оценке периодичности кривой GC-содержания. В ходе вычислительных экспериментов были найдены протяженные повторы [25, 26], которые поставили задачу автоматизации поиска повторяющихся последовательностей такого типа.

Попытки поиска повторов в нуклеотидных последовательностях с использованием непрерывных методов предпринимались и в других работах [14, 15, 16]. Однако, поскольку масштаб измерений остался тем же, то есть шаг дискретизации был равен одному нуклеотиду, данные подходы не оказались эффективнее классических методов обработки строк.

В данной работе представлено дальнейшее развитие вычислительного метода для поиска протяженных ($\geq$ 1000 н.п.) повторов в нуклеотидных последовательностях: повышена устойчивость распознавания повторов за счёт введения дополнительной кривой GA-содержания, которая делает информационное описание последовательности более полным; выведено усовершенствованное решающее правило, инвариантное к выбору масштаба; определены условия аппроксимации, при которых повторы отображаются на матрице гомологии в виде характерных паттернов, что позволило найти и опубликовать неизвестный ранее повтор [27]; построен алгоритм поиска инвертированных повторов в пространстве коэффициентов разложения, а также представлены первые результаты по полногеномному сравнению геномов *Mus musculus* и *Rattus norvegicus*.

1. ОПИСАНИЕ СПЕКТРАЛЬНОГО МЕТОДА ПОИСКА ПОВТОРОВ

1. Основные этапы метода

Преимущество непрерывных методов проявляется тогда, когда мы сравниваем не одиночные нуклеотиды, а целые блоки нуклеотидов, где каждый блок можно представить в виде некоторой дискретной функции. Далее сравнение функций можно производить спектральными методами, применяя разложения по ортогональным

базисам и сравнивая коэффициенты разложения.

В общем виде алгоритм поиска повторов состоит из пяти этапов:

1. Представление нуклеотидной последовательности в виде набора функций-аналогов.
2. Преобразование функций-аналогов в спектральное представление.
3. Сравнение спектров разложения.
4. Отображение и анализ матрицы спектральной гомологии.

Далее каждый этап описан более подробно.

2. Расчет функции-аналога последовательности ДНК

Для того чтобы воспользоваться аппроксимативными возможностями полиномиальных ортогональных базисов, необходимо преобразовать генетическую последовательность $S = s_1 s_2 s_3 \dots s_i \dots s_N$ из алфавита $s_i \in A = \{a_1, \dots, a_n\}$, где N – количество нуклеотидов в последовательности S, в числовую функцию. Разобьем алфавит A на два подмножества A_1 и A_2, так что $A_1 \cup A_2 = A$:

$$g^{A_1}(s_i) = \begin{cases} 1, & \text{если } s_i \in A_1 \\ 0, & \text{если } s_i \notin A_1 \end{cases} \tag{1}$$

Пусть $A = \{\text{A,T,G,C}\}$ алфавит нуклеотидных последовательностей, в качестве A_1 мы можем взять его подмножество $\{\text{G,C}\}$. Если окно длиной W_1 двигать с шагом d_1, символьная последовательность перекодируется в числовую функцию-аналог следующим образом:

$$f_j^{GC} = \sum_{i=jd_1}^{jd_1+W_1-1} g^{GC}(s_i),\ j = 1, \dots, \left[\frac{N - W_1 + 1}{d_1}\right]. \tag{2}$$

f^{GC} - дискретная функция, равная количеству букв алфавита $A_1 = \{\text{G,C}\}$ в окне W_1 последовательности S, является широко известной функцией GC-содержания. Параметр d_1 вводится для того, чтобы прореживать последовательность значений функции-аналога. Это необходимо для ускорения времени счета при больших размерах обрабатываемых последовательностей. Для обеспечения однозначного декодирования нуклеотидной последовательности из функции-аналога необходимо, чтобы $d_1 = 1$, а количество линейно-независимых функций-аналогов должно быть две.

Покажем процедуру декодирования бинарной последовательности, заданной в алфавите $\{0,1\}$ по функции-аналогу f_i, полученной при $d_1 = 1$, если известен начальный фрагмент последовательности длиной $W_1 - 1$. Считаются известными символы s_i при $i = 1, \dots, W_1 - 1$, тогда каждый последующий символ последовательности определяется следующим образом:

$$\begin{aligned} s_{W_1} &= f_1 - \sum_{j=1}^{W_1-1} s_i; \\ s_i &= f_{i-W_1+1} - f_{i-W_1} + s_{i-W_1}, \quad i = W_1 + 1, \cdots, N. \end{aligned} \tag{3}$$

Для последовательности в произвольном алфавите рассмотрим минимальный двоичный код, в качестве длины кодового слова для которого будет двоичный

вектор длины $\lceil \log_2 n \rceil$, где n - длина алфавита. Тогда построение функции-аналога и восстановление можно проводить отдельно для каждого бита этого кода по процедуре, описанной выше. Таким образом, мы получаем $\lceil \log_2 n \rceil$ линейно независимых функций-аналогов для произвольной символьной последовательности, по которым можно восстановить всю последовательность, зная начальный сегмент последовательности длиной $W_1 - 1$. В случае четырехбуквенного алфавита, соответствующего нуклеотидным последовательностям, для обеспечения однозначного декодирования последовательности потребуются $\lceil \log_2 4 \rceil = 2$ функции-аналога. Рассмотрим двоичное кодирование символов нуклеотидной последовательности:

$$\begin{array}{ccc} A & 1 & 0 \\ T & 0 & 0 \\ G & 1 & 1 \\ C & 0 & 1 \end{array}$$

Тогда вычисление функции-аналога для младшего бита соответствует функции GC-содержания f^{GC}, а функция-аналог для старшего бита - функции GA-содержания f^{GA}. Другие типы кодирования последовательности будут приводить к построению других линейно-независимых функций-аналогов.

Заметим, что имеет место соотношение:

$$f^{CT} = W_1 - f^{GA} \tag{4}$$

которое можно использовать для оценивания комплементарных повторов.

3. Получение спектров разложения

На данном этапе функцию f^{GC} нужно разделить на фрагменты для преобразования в коэффициенты разложения по ортогональному базису. Для этого фиксируется окно W_2, которое двигается по функции f^{GC} с шагом d_2, и на каждом шаге фрагмент, попавший в окно, преобразуется в коэффициенты разложения. Вектора коэффициентов разложения сохраняются для дальнейшей оценки близости между ними.

На данном этапе возникает вопрос о выборе системы ортогональных полиномов, по которым раскладываются фрагменты функции f^{GC}. Рассмотрим базисы полиномов Лежандра и Фурье непрерывного аргумента и Чебышева дискретного аргумента [22]. Каждый из этих базисов был протестирован в данной задаче, и все базисы оказались пригодными для оценивания среднеквадратичного отклонения фрагментов функций-аналогов. При этом каждый из этих базисов имеет свои достоинства и недостатки в рамках алгоритма решения данной задачи.

Базис Лежандра относится к базисам непрерывного аргумента, при вычислении коэффициентов разложения по которому требуется интерполяция аппроксимируемой функции с постоянной сетки на неравномерную сетку узлов квадратурной формулы Гаусса, что приводит к дополнительным вычислениям.

Базис Чебышева дискретного аргумента является дискретным аналогом базиса Лежандра и состоит из функций, ортогональных на равномерной сетке с единичным весом, поэтому процесс интерполяции можно исключить при вычислении коэффициентов разложения. Однако известно, что вычисление полиномов Чебышева дискретного аргумента по рекуррентным формулам обладает неустойчивостью [23], что связано с отказом от интерполяции на сетку Гаусса. Кроме того, рекуррентное соотношение для полиномов Чебышева дискретного аргумента является более сложным по количеству вычислительных операций. В данной задаче неустойчивость

рекуррентного алгоритма не оказывает существенного влияния на результат, поскольку окно аппроксимации значительно превосходит количество вычисляемых коэффициентов разложения.

Базис тригонометрических полиномов Фурье состоит из функций, ортогональных на постоянной сетке, и имеет самое простое для вычисления рекуррентное соотношение. Особенностью базиса Фурье является то, что он предназначен для аппроксимации периодических сигналов, а фрагменты функций-аналогов последовательности не являются периодическими. Но поскольку в данной задаче используется оценка интегрального среднеквадратичного отклонения функций, то эффект Гиббса, который возникает при аппроксимации разрывных сигналов, имеет локальный характер (в данном случае проявляется на концах интервала) и практически не влияет на оценку интеграла отклонения функций.

В конечном итоге наиболее рациональное решение - использовать для представления функции в спектральном виде разложение по базису тригонометрических функций:

$$B = \left\{ \frac{1}{\sqrt{2}}, \sin kx, \cos kx, \ldots \right\}, k = 1, 2, \ldots \tag{5}$$

где k – номер гармоники. Система функций $\phi_i(x) \in B, i = 0, 1, \ldots,$ удовлетворяет условию ортогональности в рамках скалярного произведения, определённого следующим образом:

$$(\phi_i, \phi_j) = \int_{-\pi}^{\pi} \phi_i(x)\phi_j(x)dx = \begin{cases} \pi, & \text{если } i = j \\ 0, & \text{если } i \neq j \end{cases} \tag{6}$$

Скалярное произведение в дискретной форме на равномерной сетке из W_2 узлов $x_k = \frac{\pi}{W_2}(2k - 1 - W_2)$ при $k = 1, \ldots, W_2$ [24]:

$$(\phi_i, \phi_j) = \frac{2\pi}{W_2} \sum_{k=1}^{W_2} \phi_i(x_k)\phi_j(x_k) = \begin{cases} \pi, & \text{если } i = j \\ 0, & \text{если } i \neq j \end{cases} \tag{7}$$

Коэффициенты разложения вычисляются по формулам:

$$C_i = \frac{(f, \phi_i)}{(\phi_i, \phi_i)} \tag{8}$$

где f – некоторый фрагмент функции-аналога длиной W_2.

4. Сравнение спектров разложения

Для оценки близости двух фрагментов f и g используется метрика, согласованная с нормой и скалярным произведением:

$$\rho(f, g) = \| f - g \| = \sqrt{(f - g, f - g)}. \tag{9}$$

При этом скалярное произведение может быть вычислено в пространстве коэффициентов разложения следующим образом:

$$(f - g, f - g) = \sum_{k=0}^{L-1} (C_k - D_k)^2 (\phi_k, \phi_k), \tag{10}$$

где C_k и D_k - коэффициенты разложения f и g соответственно. Оценить скалярное произведение можно так:

$$(f-g, f-g) = \int_{-\pi}^{\pi} (f-g)^2 dt \le 2\pi W_1^2, \tag{11}$$

поскольку для любой функции-аналога выполняется неравенство $f \le W_1$ по определению.

Таким образом, решающее правило основано на проверке следующего неравенства:

$$\theta(f,g) \le \varepsilon, \tag{12}$$

где $\varepsilon \in [0,1]$ пороговое значение решающего правила, L - количество коэффициентов разложения и

$$\theta(f,g) = \frac{1}{2W_1^2} \sum_{k=0}^{L-1} (C_k - D_k)^2 \tag{13}$$

Среднеквадратичное отклонение является монотонно возрастающим по числу коэффициентов разложения, что позволяет прервать вычисление суммы квадратов, если пороговое значение ε превышено.

5. Формулы для инвертированных повторов

Для поиска участков с прямыми повторами векторы коэффициентов разложения сравниваются выбранной метрикой по формуле (12), а для того, чтобы из коэффициентов, отвечающих за прямое сходство, получить коэффициенты, которые будут соответствовать инвертированным повторам, требуются провести два типа преобразований над кривой содержания. Смысл первого преобразования сделать комплементарным один из сравниваемых фрагментов кривой к другому, а второе "разворачивает" этот фрагмент в противоположном направлении. Оба преобразования происходят в пространстве коэффициентов разложения.

Рассмотрим два взаимно-инвертированных фрагмента функции GA-содержания (Рис. 1).

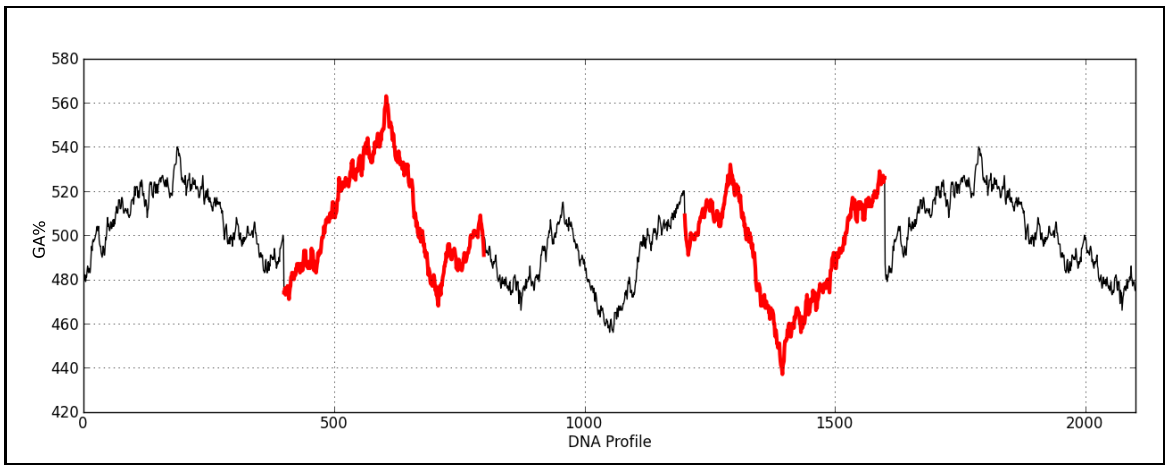


Рис. 1: Кривые GA-содержания. Красным обозначены участки, взаимно-инвертированные относительно друг друга.

Пользуясь соотношениями (4) и (7), мы можем выразить коэффициенты C_k^{CT} через коэффициенты C_k^{GA} следующим образом:

$$C_k^{CT} = \frac{(W_1 - f^{GA}, \phi_k)}{(\phi_k, \phi_k)} = \frac{(W_1, \phi_k) - (f^{GA}, \phi_k)}{(\phi_k, \phi_k)} = \frac{(W_1, \phi_k)}{(\phi_k, \phi_k)} - C_k^{GA} \tag{14}$$

так как для всех ортогональных базисов, которые мы используем (Фурье, Чебышева дискретного аргумента, Лежандра), справедливо:

$$\frac{(W_1, \phi_k)}{(\phi_k, \phi_k)} = 0 \tag{15}$$

то для всех коэффициентов, кроме нулевого, будет выполняться равенство:

$$C_k^{CT} = -C_k^{GA}. \tag{16}$$

Нулевой коэффициент в зависимости от базиса будет принимать значение, согласно выражению (14). Представим W_1 в виде $W_1 = \sqrt{2}W_1\phi_0$, тогда для базиса Фурье нулевой коэффициент будет выражаться соотношением:

$$C_0^{CT} = \frac{(\sqrt{2}W_1\phi_0, \phi_0)}{(\phi_0, \phi_0)} - C_0^{GA} = \sqrt{2}W_1 - C_0^{GA} \tag{17}$$

Второе преобразование, связанное с “разворотом” кривой в противоположном направлении, что необходимо делать при решении задачи поиска инвертированных повторов, можно сделать, воспользовавшись тем, что базис (5) состоит из чётных и нечётных функций. Смена знака у нечётных коэффициентов, приведет к тому, что кривая “развернется” в противоположном направлении. Применяя данное свойство к выражению (16), мы получаем итоговое выражение:

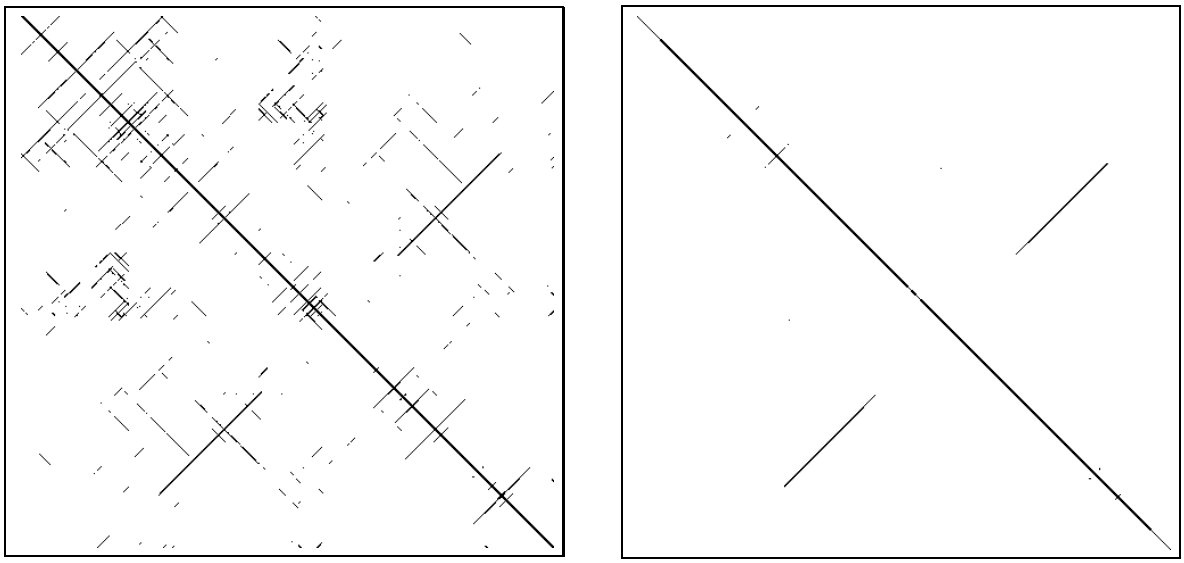
$$\begin{cases} C_0^{CT} = \sqrt{2}W_1 - C_0^{GA} \\ C_{2k}^{CT} = -C_{2k}^{GA} \\ C_{2k+1}^{CT} = C_{2k+1}^{GA} \end{cases} \tag{18}$$

6. Матрица спектральной гомологии

Для отображения результатов сравнения векторов коэффициентов разложения по аналогии с точечной матрицей гомологии [17] создаётся матрица спектральной гомологии.

Точка на пересечении строки и столбца ставится в случае близости спектров коэффициентов полученных для функций-аналогов фрагментов. Протяженные участки сходства, как и в случае с точечной матрицей, отображаются параллельными (в случае прямых повторов) или перпендикулярными (в случае инвертированных повторов) отрезками линий, парадллельными главной диагонали. Автоматический анализ матрицы спектральной гомологии позволяет выделить наиболее существенные участки сравнения. Для более точного определения координат повторов требуется этап верификации. Для того чтобы повысить качество распознавания, одновременно используются две кривые f^{GC} и f^{GA}, при этом ширина окна W_2 не меняется.

На Рис. 2 изображена фильтрация матрицы гомологии за счёт добавления дополнительной кривой, построенной по нуклеотидам G и A. Оценка близости между спектрами коэффициентов разложения происходит по двум функциям-аналогам



а) б)

Рис. 2: Зашумленный повтор на матрице спектральной гомологии. а) Оценивание только по кривой GC-содержания. б) Оценивание по кривым GC- и GA-содержания.

раздельно, после чего повтор отображается на матрице, если обе пары функций-аналогов, для каждого фрагмента, близки.

$$(\theta(f^{GC}, g^{GC}) \leq \varepsilon) \wedge (\theta(f^{GA}, g^{GA}) \leq \varepsilon) \tag{19}$$

где f и g - функции-аналоги сравниваемых последовательностей. Из рисунка видно, что совмещение двух матриц спектрального сходства позволяет существенно уменьшить уровень шума на изображении.

2. ПРИЛОЖЕНИЯ СПЕКТРАЛЬНОГО АЛГОРИТМА ПОИСКА ПОВТОРОВ

1. Поиск протяженных тандемных повторов

Поиск мегасателлитных тандемных повторов с помощью спектрального алгоритма является задачей распознавания с обучением по образцу. В качестве образца были выбраны структуры ранее найденных повторов IMPB_01 [25] и IMPB_02 [26]. Анализ матрицы сходства данных участков геномов позволил найти наиболее оптимальные параметры, при которых шаблонные структуры IMPB_01 и IMPB_02 выделяются на спектральной матрице сходства с минимальным шумом (Рис. 3) в виде квадрата.

Дальнейшая задача сводилась к сканированию геномов с определенными параметрами и поиску подобных шаблонов. В результате анализа *Mus musculus* и *Rattus norvegicus*, было выявлено некоторое количество протяженных повторов, тем не менее под искомый шаблон также попадали ранее известные многократно повторяющиеся сателлиты длиной порядка 300 п.н.

Анализ 17-й хромосомы кролика выявил регион, в котором находятся протяженные тандемные повторы. Длина мотива приблизительно равняется 2623 н.п. Данный повтор был опубликован в RepBase [30] под именем MSU1 [27]. Изображение кривой GA-содержания повтора MSU1 (Рис.4) отражает несовершенность периодичности. Копии повтора отличаются друг от друга минимально на 4.2%, максимально на 22.7%, в консенсусную последовательность входит 47% строго консервативных позиций.

Для изображения тандемного повтора в виде сплошного квадрата шаг окна аппроксимации должен приблизительно соответствовать длине мотива. Также было

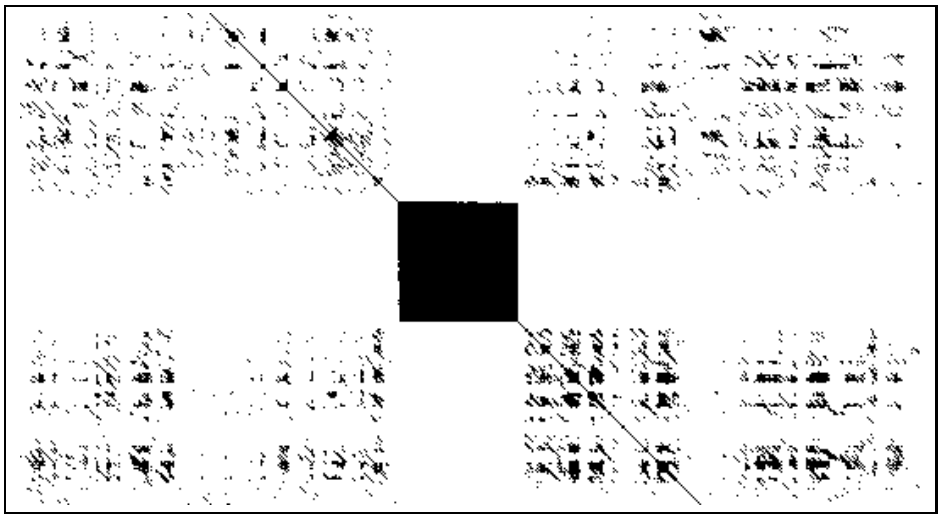


Рис. 3: Тандемный повтор IMPB_01 на спектральной матрице сходства четко виден в виде квадрата.

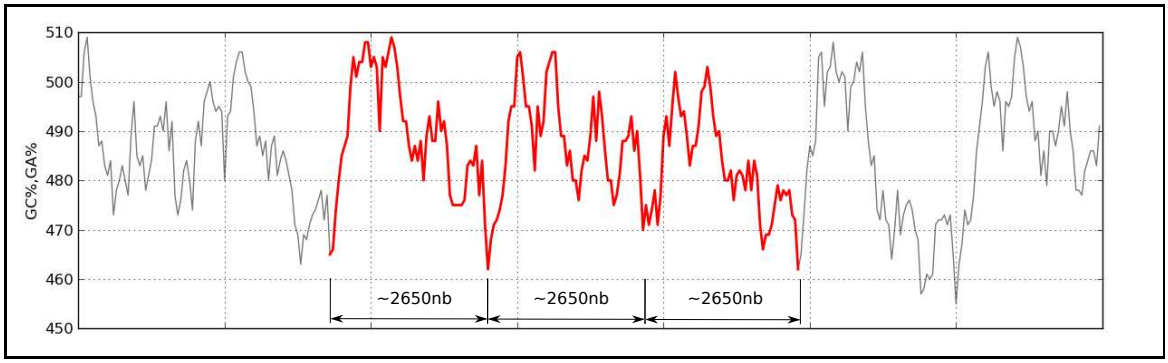


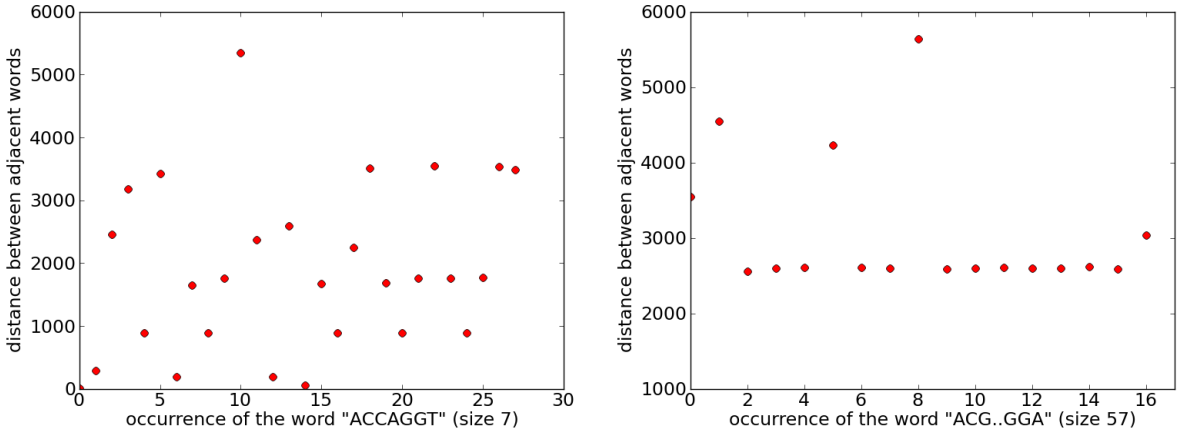
Рис. 4: Фрагмент наложения функций GC- и GA-содержания для тандемного повтора MSU1

обнаружено чисто эмпирически соотношение между длиной окна GC-содержания W_1 и окна аппроксимации W_2 как $1:4$. В частности, для поиска повтора MSU1 были использованы следующие параметры: $W_1 = 2500, W_2 = 10000, d_2 = 2500, L = 15, \varepsilon = 0.0011$.

После того, как квадрат локализован, т.е. определены координаты всего повторяющегося участка, его необходимо нарезать на отдельные фрагменты. В этой задаче исходно имеются только предварительные данные о среднем размере паттерна повтора (параметр d_2) и кратности повтора, вычисляемой путем деления общей длины повтора на величину среднего размера паттерна. Вообще говоря, дивергировавшие элементы повтора могут не только отличаться друг от друга текстуально, но и существенно отличаться друг от друга по длине. Для такой нарезки используется подход, подобный используемым в программах BLAST [31] и TRF [32], в основе которого лежит поиск якорных (консервативных) последовательностей. Задача состоит в поиске наибольшей общей подпоследовательности, входящей в большинство элементов повтора. Тогда по координатам вхождений этого якоря можно нарезать участок повтора на отдельные элементы. Отличительной особенностью используемого метода является то, что предварительные данные, полученные спектральным методом

создают ограничения, которые позволяют отфильтровать большую часть якорных последовательностей, не вводя дополнительных эвристик.

Суть подхода заключается в поиске наиболее точного фрагмента максимальной длины, удовлетворяющего условию кратности. В качестве подготовительного этапа процедуры создаётся полный словарь вхождений четырехнуклеотидных слов, после чего из него удаляются те слова, число встречаемости которых меньше ожидаемого значения кратности тандемного повтора. Итерационная процедура состоит в наращивании длины l оставшихся в словаре слов, путём создания из каждого слова четырех новых длины $l+1$ и подсчета количеств вхождений для каждого из слов с удалением тех, которые недостаточное число раз представлены на анализируемом участке. Первые несколько итераций увеличивают размер словаря, но очевидно, что в конце процедуры остаются несколько наиболее протяженных якорных последовательностей, опираясь на координаты которых можно решить задачу выделения повторяющихся элементов. После того, как якорные последовательности подобрана, текст нарезается на отдельные строки и выравнивается с помощью программы ClustalW2 [37].



а) б)

Рис. 5: Встречаемость и разброс по тексту слов с различной длиной а) якорная последовательность размером 7 нуклеотидов в повторе MSU1. б) якорная последовательность размером 57 нуклеотидов в повторе MSU1.

Для выявления более полной картины повторяемости можно использовать короткие якорные последовательности, это позволит выделить внутренние повторяющиеся фрагменты, которые входят в состав крупных. Тем не менее, выбор таких последовательностей не всегда очевиден. С уменьшением размера последовательности возрастает вероятность её случайного возникновение в тексте (см. Рис.5,а). Очевидно, что в случае сильно дивергировавших элементов повтора полностью автоматизировать процедуру точной нарезки повтора на элементы не удается. Использование протяженных якорных последовательностей в большинстве случаев позволяет выделить искомый повторяющийся фрагмент. Построение графика расстояний между соседними вхождениями якорей помогает выявить наилучший из всех возможных якорей. Таковым будет тот олигонуклеотид, расстояния между смежными вхождениями которого будут наиболее компактно разбросаны вокруг среднего значения длины паттерна периодичности. На Рис.5, показаны графики, построенные для олигонуклеотидов длины 7 и 57, по горизонтальной оси отложены координаты олигонуклеотида в общем повторе, а по вертикальной оси отложено расстояние,

на котором находится следующее точное вхождение этого же олигонуклеотида. Большинство точек сосредоточено в районе реального размера повторяющегося фрагмента в MSU1(2600 н.п.), однако график для той якорной последовательности, которая длиннее, более наглядно это показывает. Отдельные выбросы могут быть связаны как со отдельными спейсерами, разбивающими тандемный повтор, так и с тем, что в якорной последовательности имеются точечные мутации. Во втором случае можно повторить процедуру выделения периода для отдельно взятого фрагмента.

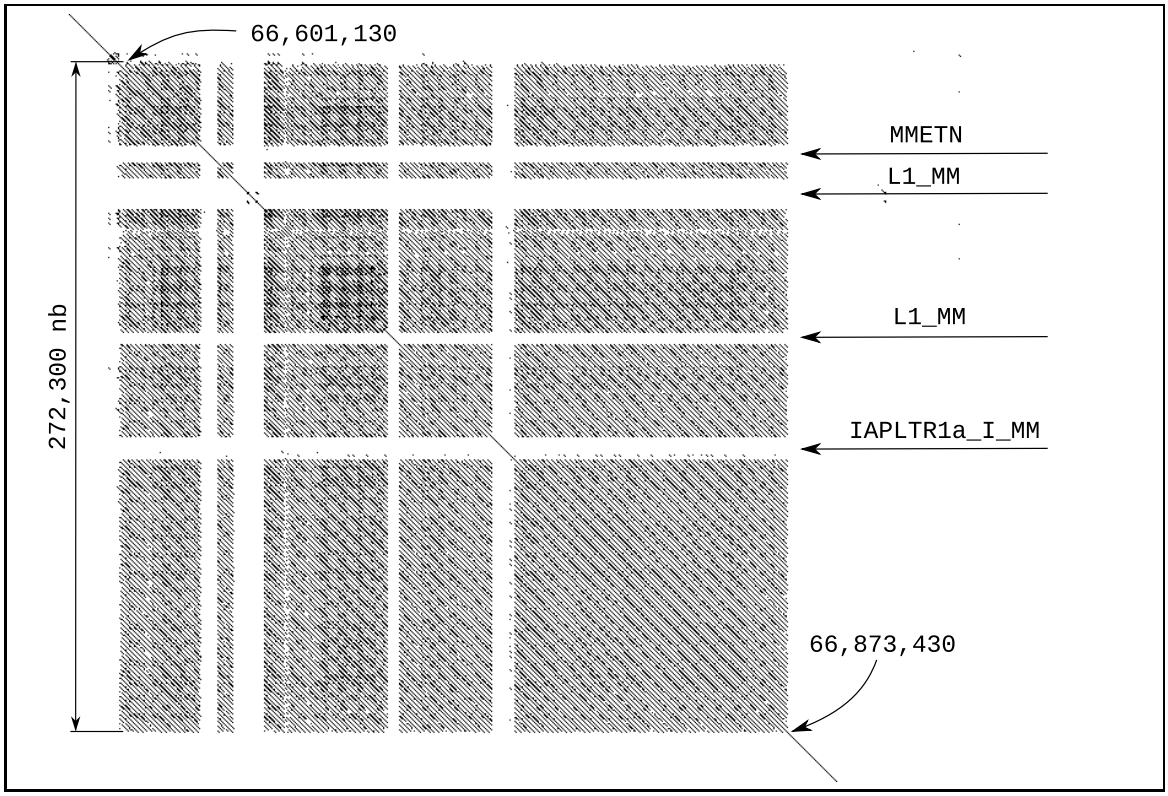


Рис. 6: Кластер генов SNORD115, обнаруженный на 7-й хромосоме мыши. Кластер представлен тандемными повторами, каждая копия содержит в себе нуклеотидную последовательность гена малой ядрышковой РНК SNORD115 длиной 89 н.п., расположенную в последней четверти паттерна периодичности, при средней длине паттерна, равной 1873. Кластер разбит фрагментами транспозонов, наиболее крупные из них обозначены стрелками. Обозначения транспозонов взяты из базы данных Repbase.

При анализе 7 хромосомы *Mus musculus* найден кластер, состоящий из тандемных повторов длиной порядка 1873 н.п. и количеством копий более 130. Каждая копия содержит в себе ген SNORD115 протяженностью 89 н.п. [29] (Рис. 6).

2. Полногеномное сравнение

Логичным продолжением поиска мегасателлитных тандемных повторов стало полнохромосомное сравнение организмов *Mus musculus* и *Rattus norvegicus*.

На Рис.7 изображена генерализованная таблица подобия, полученная после полногеномного сравнения ДНК крысы и мыши. На этом рисунке отображены повторы длиной не менее 200 тысяч нуклеотидов. Такие повторы хорошо согласуются с известными районами синтении [38]. Интересно, что на таком масштабе анализа проявляются сложности, связанные с более крупными мутациями — делециями, вставками и инверсиями - разворотом фрагментов ДНК на 180 градусов.

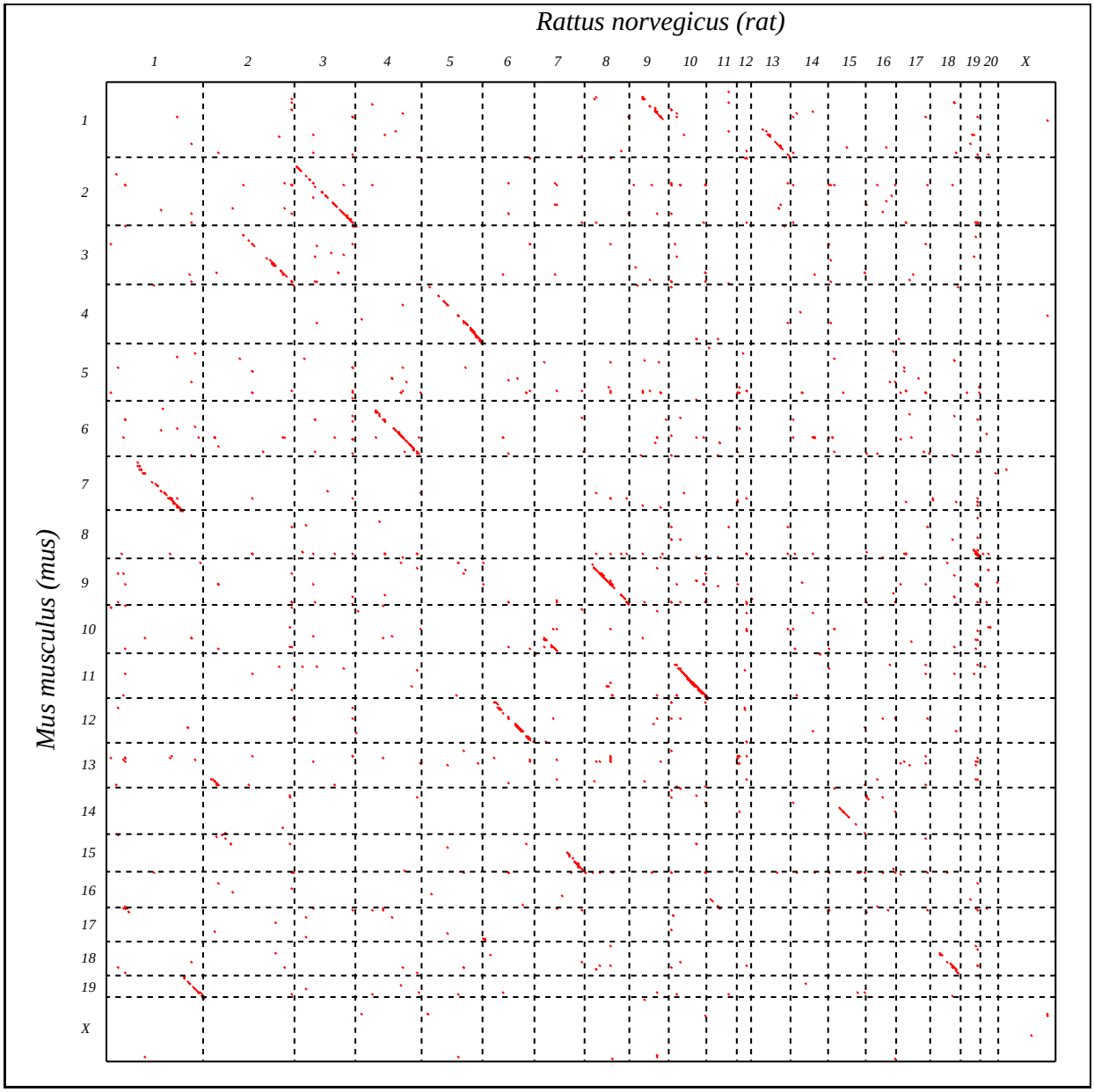
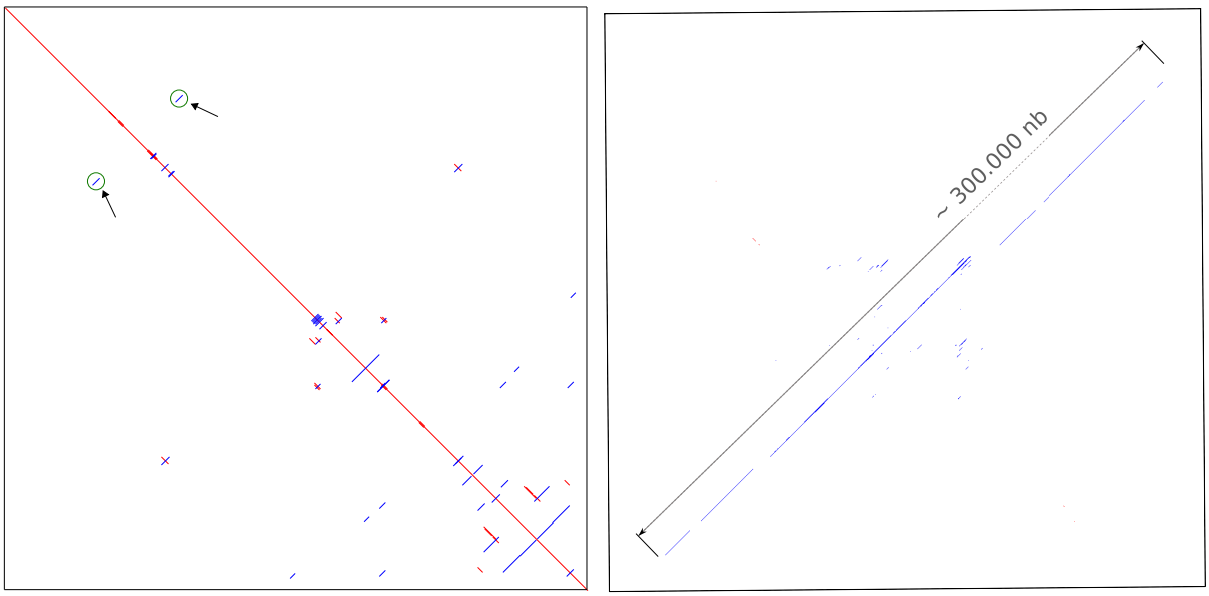


Рис. 7: Матрица межгеномного сходства. Красными точками показаны все регионы, содержащие протяженные повторы. Хромосома Y *Mus musculus* не показана из-за ее малой протяженности относительно выбранного масштаба.

3. Сравнение с ДНК-гибридизацией

Параметры метода могут изменятся в широком диапазоне, что позволяет исследовать последовательности на разных масштабах. Масштабирование позволяет построить предварительную карту повторов, а затем рассматривать наиболее интересные фрагменты. К примеру, на Рис.8,а) изображена общая карта крупных повторов человеческой Y хромосомы. Одним из интересных мест данной хромосомы является протяженный инвертированный повтор длиной порядка 300 тыс. н.п. (отмечен стрелками), более подробно его структуру можно рассмотреть на Рис.8,б), из которого видно, что повтор сильно разбит, между схожими участками имеются области, где сходство минимально или отсутствует.



а) б)

Рис. 8: а) Приближённая карта человеческой Y хромосомы. Стрелками обозначен протяженный инвертированный повтор, ограничивающий область размером 3.5 млн.н.п. б) Обозначенный повтор при меньшем масштабе. Структура показывает, что повтор сильно разбит, между схожими участками имеются протяженные области, где сходство минимально или отсутствует.

Повтор ограничивает область размером в 3.5 млн.н.п. и, вероятно, является причиной инверсии этой области [28].

ВЫВОДЫ

Обобщённый спектрально-аналитический метод обладает следующими свойствами, потенциал которых был реализован при построении представленного в этой работе метода анализа нуклеотидных последовательностей:

1. интегральное оценивание, которое позволяет не сосредотачиваться на локальных особенностях последовательности;

2. масштабирование, т.е. выбор масштаба, при котором анализируется последовательность, за счёт варьирования размера и шага скользящих окон получения функций-аналогов и спектров разложения;

3. операции над функциями-аналогами в пространстве коэффициентов разложения, которые позволяют находить инвертированные и комплементарные повторы, не изменяя саму последовательность.

Отличием от классических алгоритмов поиска повторов является то, что непосредственная работа с символьной последовательностью ведется только на этапе предобработки, а основную часть алгоритма составляют спектральные методы, основанные на численных методах с плавающей точкой. Это позволяет наиболее полно использовать возможности процессоров при реализации алгоритмов, поскольку задействуются как целочисленные операции, так и операции с плавающей точкой.

Для поиска протяженных тандемных повторов изначально был предложен алгоритм [21], основанный на оценке периодичности функции-аналога последовательности.

В данной работе для автоматизации поиска тандемных повторов применяется алгоритм, основанный на построении и анализе матрицы гомологии, что приводит к большему количеству вычислений. Тем не менее, оба алгоритма являются по сложности линейными в зависимости от длины обрабатываемой последовательности, поскольку анализ матрицы гомологии проводится только вдоль диагонали.

Работа выполнена при поддержке грантов РФФИ №11-07-00716, №12-07-00530, №10-01-00609, а также компаний Intel и "Т-платформы".

СПИСОК ЛИТЕРАТУРЫ

1. Collins F.S., Morgan M., Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science*. 2003. V. 300. P. 286–290.

2. Подгорная О.И., Остромышенский Д.И., Кузнецова И.С., Матвеев И.В., Комиссаров А.С. Парадоксы организации центромера и гетерохроматина. *Цитология*. 2009. Т. 51. № 3. С. 204—211.

3. Fondon J.W. III, Garner H.R. Molecular origins of rapid and continuous morphological evolution. *Proc. Nat. Acad. Sci*. 2004. V. 101. № 52. P. 18058—18062.

4. Lakich D., Kazazian H.H. Jr, Antonarakis S.E., Gitschier J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat. Genet.*. 1993. V. 5. P. 236–241.

5. Emery A.E.H. Emery-Dreifuss syndrome. *J. Med. Genet.*. 1989. V. 26. P. 637–641.

6. Small K., Iber J., Warren S.T. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat. Genet.*. 1997. V. 16. P. 96–99.

7. 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.

8. Sutherland G.R., Richards I.R. Simple tandem DNA repeats and human genetic disease. *Proc. Natl. Acad. Sci. USA*. 1995. V. 92. P. 3636–3641.

9. Mitas M. Trinucleotide repeats associated with human disease. *Nucleic Acids Res*. 1997. V. 25. P. 2245–2253.

10. Toth G., Gaspari Z., Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res*. 2000. V. 10. P. 967–981.

11. Graur D., Hide W.A., Li W.H. Is the guinea-pig a rodent? *Nature*. 1991. V. 351. P. 649–652.

12. Saitou N., Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. 1987. V. 4. P. 406–425.

13. Gidley J.W.. The lagomorphs an independent order. *Science*. 1912. V. 36. P. 285–286.

14. Волков В.В., Леонтьев А.Ю. Исследование симметрии генетических текстов методом Фурье-анализа. *Биополимеры и клетка*. 1990. Т. 6. № 6. С. 68–72.

15. Benson D. Fourier method for biosequence analysis. *Nucl. Acid Res*. 1991. V. 18. P. 6305–6310.

16. Лобзин В.В., Чечеткин В.Р. Порядок и корелляция в геномных последовательностях ДНК. Спектральный подход. *УФН*. 2000. Т. 170. № 1. С. 57–81.

17. Gibbs A.J., McIntyre G.A. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* 1970. V. 16. P. 1–11.

18. 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. *Доклады Академии Наук*. 2006. Т. 411. № 5. С. 599–602.

19. Tetuev R.K., Dedus F.F., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Pankratov A.N. Recognition of the structural-functional organization of genetic sequences. *Moscow University Computational Mathematics and Cybernetics*. 2007. V. 31. № 2. P. 49–53.

20. 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., 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.

21. Тетуев Р.К., Назипова Н.Н., Панкратов А.Н., Дедус Ф.Ф. Поиск мегасателлитных тандемных повторов в геномах эукариот по оценке осцилляций кривых GC-содержания. *Математическая биология и биоинформатика*. 2010. Т. 5. № 1. С. 30–42. URL: http://www.matbio.org/downloads/Tetuev2010(5_30).pdf (дата обращения: 20.04.2012).

22. Никифоров А.Ф., Суслов С.К., Уваров В.Б. *Классические ортогональные полиномы дискретной переменной*. М.: Наука, 1985.

23. Никифоров А.Ф., Скачков М.В. Методы вычисления q-полиномов. *Матем. моделирование*. 2001. Т. 13. № 8. С. 85–94.

24. Хэмминг Р.В., *Численные методы для научных работников и инженеров*. М.: Наука, 1972. 400 с. (Перевод с англ. Hamming R.W. *Numerical methods for scientists and engineers*. MC GRAW-HILL BOOK COMPANY, 1962).

25. Tetuev R.K., Nazipova N.N. Consensus of repeated region of mouse chromosome 6 containing 60 tandem copies of a complex pattern. *Repbase Reports*. 2010. V. 10. №5. P. 776.

26. Tetuev R.K., Nazipova N.N., Dedus F.F. Consensus of repeated region of rat chromosome 4 similar to mouse chromosome 6 repeated region, enclosed in the intergenic region between genes Hrh1 and Atg7. *Repbase Reports*. 2010. V. 10. № 8. P. 1185.

27. Pyatkov M.I., Filippov V.V., Pankratov A.N. Consensus of repeated region of rabbit chromosome 17 containing over 15 huge approximate tandem repeats. *Repbase Reports*. 2012. V. 12. № 3.

28. Tilford C., Kuroda-Kawaguchi T., Skaletsky H., Rozen S., Brown L., Rosenberg M., McPherson J., Wylie K., Sekhon M., Kucaba A., Waterston R., Page D. A physical map of the human Y chromosome. *Nature*. 2001. V. 409. P. 943–945.

29. Cavaillé J., Buiting K., Kiefmann M., Lalande M., Brannan C.I., Horsthemke B., Bachellerie J.P., Brosius J., Hüttenhofer A. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *PNAS*. 2000. V. 97. № 26. P. 14311–14316.

30. 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.

31. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol.* 1990. V. 215. № 3. P. 403–410.

32. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999. V. 27. P. 573–578.

33. Kolpakov R., Bana G., Kucherov G. Mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acid Research*. 2003. № 31. P. 3672–3678.

34. Ogurtsov A.Y., Roytberg M.A, Shabalina S.A., Kondrashov A.S. OWEN: aligning long collinear regions of genomes. *Bioinformatics*. 2002. V. 18. P. 1703–1704.

35. Landau G.M., Schmidt J.P. and Sokol D. An Algorithm for Approximate Tandem Repeats. *Journal of Computational Biology*. 2001. V. 8. P. 1–18.

36. Levenshtein V.I. Binary codes capable of correcting, deletions, insertions and reversals. *Soviet Phys. Dokl*. 1966. № 10. P. 707–710.

37. Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. ClustalW and ClustalX version 2.0. *Bioinformatics*. 2007. V. 23. P. 2947–2948.

38. Loots G.G., Ovcharenko I. ECRbase: Database of Evolutionary Conserved Regions, Promoters, and Transcription Factor Binding Sites in Vertebrate Genomes. *Bioinformatics*. 2007. V. 23. P. 122–124.

Материал поступил в редакцию 15.07.2012, опубликован 08.08.2012.