Математическая биология и биоинформатика. 2013. Т. 8. № 2. С. 480–501.
URL: http://www.matbio.org/2013/Chaley_8_480.pdf

========== БИОИНФОРМАТИКА ==========

УДК: 577.322

Исследование феномена скрытой периодичности в геномах эукариотических организмов

Чалей М.Б.[1*], Кутыркин В.А.[2], Теплухина Е.И.[1], Тюльбашева Г.Э.[1], Назипова Н.Н.[1]

[1] Институт математических проблем биологии, Российская академия наук, Пущино, Московская область, 142290, Россия
[2] Московский Государственный Технический Университет им. Н.А. Баумана, Москва, 107005, Россия

Аннотация. Представлен анализ данных первого выпуска базы HeteroGenome, содержащей выявленные районы скрытой периодичности в геномах ряда эукариотических организмов. Тандемные повторы с различной сохранностью копий паттерна, включая сильно размытые повторы, были идентифицированы в геномах *S. cerevisiae*, *A. thaliana*, *C. elegans* и *D. melanogaster*. Данные были получены с помощью оригинального спектрально-статистического подхода к поиску достоверных районов скрытой периодичности в последовательностях ДНК. Введение двухуровневой структуры представления данных (на первом, неизбыточном, уровне районы скрытой периодичности рассматриваются в целом, на втором уровне – консервативные фрагменты их периодической структуры) позволило оценить долю покрытия (~10% от длины генома) анализируемых геномов районами скрытой периодичности. Оценка выведена на основе данных первого уровня. Анализ количественного и качественного состава (по уровню дивергенции) районов скрытой периодичности по всем хромосомам рассматриваемых организмов выявил характеристические типы периодичности в геноме каждого организма. Получены гистограммы плотности распределения районов скрытой периодичности для каждой хромосомы рассматриваемых геномов. Выявлен репертуар длин периодов в геномах. База данных HeteroGenome предоставляет дополнительные возможности анализа содержащихся в ней данных и доступна по адресу: http://www.jcbi.ru/lp_baze/.

Ключевые слова: *скрытая периодичность, тандемные повторы, анализ генома*.

ВВЕДЕНИЕ

Тандемные повторы (массивы последовательно повторяющихся копий некоторого исходного фрагмента последовательности ДНК, или паттерна) как объекты периодичной структуры генома давно находятся в фокусе внимания исследователей. С одной стороны, это внимание обусловлено стремлением понять молекулярные механизмы возникновения и эволюции повторов, их функциональное значение в геноме; с другой стороны, – возможностью разрабатывать на их основе маркеры для исследований популяционной и эволюционной генетики. Повреждение копий паттерна при замене исходных нуклеотидов, а также при вставках и делециях, как единичных, так и нескольких нуклеотидов, ведёт к образованию размытых тандемных повторов. Размытые тандемные повторы, повреждения копий паттерна в которых ограничены только заменами нуклеотидов (нукл.), принято называть нечеткими тандемными повторами. Размытые тандемные повторы, включая нечёткие повторы, являются участками скрытой периодичности в геноме.

[*]maramaria@yandex.ru

Наиболее исследованной группой тандемных повторов в геноме являются микросателлиты (паттерн не более 10 нукл.) и минисателлиты (паттерн не более 100 нукл.), благодаря их использованию в качестве генетических маркеров в медицинской криминалистике, для установления родства или позиционного клонирования, и также в популяционной и эволюционной генетике [1].

Считается, что причина появления микросателлитов в геноме обусловлена «проскальзыванием» ДНК полимеразы на периодичном шаблоне из-за теплового шума в процессе репликации хромосом [2]. В отличие от микросателлитов, молекулярный механизм появления и распространения более крупных минисателлитов связывают главным образом с процессами рекомбинационного характера, такими как неравный кроссинговер или конверсия генов [3]. Тандемные повторы могут возникать и в результате последовательной дупликации генов при гомологичной рекомбинации хромосом на стадии мейоза [4].

Многочисленные тандемные повторы располагаются в центромерных и теломерных районах хромосом [5]. Тандемные повторы были найдены в хрупких (fragile) сайтах хромосом [6, 7]. В некоторых случаях экспансия триплетных микросателлитов в хрупких сайтах вызывает умственную отсталость у человека [8]. Неврологические нарушения у человека также могут быть вызваны «динамическим» мутациями (сокращением или увеличением копий в тандемном повторе), как в кодирующих, так и в некодирующих районах, причем мутациями не только триплетных микросателлитов [9-11]. Тандемные повторы, расположенные в некодирующих районах, могут оказывать влияние на уровень экспрессии генов, процессы транскрипции и трансляции [12].

Ранее был разработан ряд программ поиска совершенных или почти совершенных тандемных повторов: TRF [13], ACMES [14], MREPATT [15], STRING [16], mreps [17], ATRHunter [18] и др. Различные алгоритмы лежат в основе этих программ, соответственно, и результаты их не всегда совпадают, находясь в зависимости от длины периода, количества копий и степени размытия или дивергенции повтора.

В последние годы разрабатываются программы, направленные на выявление всё более размытых тандемных повторов с тем, чтобы иметь возможность их изучения и в эволюционном аспекте: TandemSWAN [19], IMEX [20], TRStalker [21], программа поиска на основе модели эволюционирующих тандемных повторов [22]. Если совершенные (или почти совершенные) тандемные повторы изменчивы в числе копий и динамичны за счёт механизма проскальзывания при репликации ДНК, то сильно размытые повторы являются гораздо более стабильными элементами структуры генома, функциональная роль которых ещё мало изучена.

Как правило, на основе результатов программ поиска размытых тандемных повторов создаются разнообразные информационные ресурсы, например, база данных TRedD [23] тандемных повторов в геноме человека, найденных с помощью алгоритма [22]. Наиболее известна база данных TRDB [24], содержащая тандемные повторы, найденные методом Tandem Repeats Finder (TRF) [13] в полностью секвенированных геномах прокариот и эукариот, в том числе и в геноме человека. База данных TRbase [25] связывает тандемные повторы в геноме человека, выявленные методом TRF [13], с локализацией генов на хромосомах, особенно выделяя гены, повреждения которых привели к генетическим заболеваниям.

Следует отметить, что предлагаемые эвристические алгоритмы выявления сильно размытых тандемных повторов не решают проблему достоверности получаемых результатов. Чтобы гарантировать обнаружение именно размытого тандемного повтора, обычно вводится дополнительная фильтрация результатов. Например, в программе, основанной на модели эволюционирующих тандемных повторов [22], уровень дивергенции копий паттерна в повторе ограничивается максимальным значением ~30%. Тем не менее, это значение превышает уровень дивергенции, равный

20%, задаваемый в вероятностной модели метода TRF [13]. Кроме того, ранее отмечалась избыточность результатов метода TRF (имеются случаи, когда несколько паттернов тандемного повтора предлагаются для одного участка ДНК) [19] и нестабильность его результатов (при смещении на 1–3 нукл. на одном участке ДНК предлагаются различные оценки паттерна) [26].

В настоящей работе оригинальный спектрально-статистический подход [26-28] был применён для поиска достоверных районов скрытой периодичности в геномах *Saccharomyces cerevisiae, Arabidopsis thaliana, Caenorhabditis elegans* и *Drosophila melanogaster*. Как было показано ранее [26-28], такой подход позволяет избегать неоднозначности при определении структуры скрытой периодичности в размытых тандемных повторах, и оптимизирует оценку размера паттерна периодичности. Теоретически, с помощью спектрально-статистического подхода [28] можно выделять чрезвычайно размытые тандемные повторы, для которых средняя степень дивергенции копий паттерна составляет ~50%. Общее описание этого подхода приводится в следующем разделе. Подход использует χ^2 – статистику проверки однородности последовательности ДНК на уровне значимости, характерном для размытых тандемных повторов, последовательности которых очевидно неоднородные. Однако значимая неоднородность является необходимым, но всё-таки не достаточным условием определения периодичности указанных выше типов. Поскольку результаты поиска скрытой периодичности, полученные в автоматическом режиме, несомненно, являются неоднородными последовательностями, но некоторый дополнительный анализ требуется для подтверждения их периодической структуры, в дальнейшем, для краткости, об этих результатах часто говорится как о неоднородности в последовательностях генома. Поэтому название базы HeteroGenome (неоднородный геном) отражает эту особенность содержащихся в ней данных.

При создании первого выпуска базы HeteroGenome не преследовалась цель накопления данных о скрытой периодичности в геномах как можно большего числа организмов, доступных на сегодняшний день для анализа. Прежде всего, нам хотелось, чтобы данные, собранные в базе, послужили осмыслению феномена скрытой периодичности в геномах живых организмов, исследованию его масштаба, характера и, возможно, целенаправленности. Поэтому в настоящей работе был сделан акцент на количественный и качественный анализ данных о районах скрытой периодичности в геномах четырех вышеперечисленных организмов.

Данная статья представляет первый выпуск базы данных HeteroGenome, в которой содержатся районы скрытой периодичности, выявленные с помощью спектрально-статистического подхода [28] в геномах различных организмов. Благодаря двухуровневому представлению данных, где на первом, неизбыточном, уровне содержатся непересекающиеся последовательности выявленных районов скрытой периодичности, оказалось возможным оценить, какой процент эти участки составляют от длины рассматриваемого генома. Кроме того, с помощью специального параметра, отражающего средний уровень сохранности копий паттерна в повторе, описываемого в следующем разделе, можно проанализировать качество периодической структуры в каждом районе. Такой анализ позволил выявить характерные типы периодичности в каждом из рассматриваемых в настоящей работе геномов.

Таким образом, база данных HeteroGenome может быть полезна как для исследований в областях функциональной и эволюционной геномики при поиске тандемных повторов, характерных для анализируемого генома, так и для изучения феномена скрытой периодичности в последовательностях ДНК. База имеет удобный пользовательский интерфейс, различные опции дополнительного анализа данных, и также позволяет пользователю загрузить результаты запроса на свой компьютер. База

данных HeteroGenome находится в свободном доступе по адресу http://www.jcbi.ru/lp_baze/.

МАТЕРИАЛЫ И МЕТОДЫ

При создании первого выпуска базы данных HeteroGenome поиск районов скрытой периодичности был выполнен в четырёх геномах хорошо изученных модельных организмов [29]: *S. cerevisiae*, *A. thaliana*, *C. elegans* и *D. melanogaster*. Геномы этих организмов были секвенированы одними из первых, они являются достаточно точными и аннотированными последовательностями ДНК. Кроме того, они представляют геномы эукариот от одноклеточных (пекарские дрожжи) до многоклеточных организмов растений (крестоцветное растение резуховидка Таля) и животных (нематода), что способствует общему рассмотрению феномена скрытой периодичности в геноме.

Таблица 1. Ссылки на источники данных по геномам модельных организмов, использованных в работе.

Организм	Хромосомы	Идентификаторы GenBank	
S. cerevisiae	I	NC_001133.7	GI:144228165
	II	NC_001134.7	GI:50593115
	III	NC_001135.4	GI:85666111
	IV	NC_001136.8	GI:93117368
	V	NC_001137.2	GI:7276232
	VI	NC_001138.4	GI:42742172
	VII	NC_001139.8	GI:162949218
	VIII	NC_001140.5	GI:82795252
	IX	NC_001141.1	GI:6322016
	X	NC_001142.7	GI:116006492
	XI	NC_001143.7	GI:83722562
	XII	NC_001144.4	GI:85666119
	XIII	NC_001145.2	GI:44829554
	XIV	NC_001146.6	GI:117937805
	XV	NC_001147.5	GI:84626310
	XVI	NC_001148.3	GI:50593503
	MT	NC_001224.1	GI:6226515
A. thaliana	I	NC_003070.9	GI:240254421
	II	NC_003071.7	GI:240254678
	III	NC_003074.8	GI:240255695
	IV	NC_003075.7	GI:240256243
	V	NC_003076.8	GI:240256493
C. elegans	I	NC_003279.4	GI:86561680
	II	NC_003280.4	GI:86562519
	III	NC_003281.5	GI:86563600
	IV	NC_003282.3	GI:72185816
	V	NC_003283.5	GI:86564547
	X	NC_003284.5	GI:86565306
D. melanogaster	2L	NT_033779.4	GI:116010444
	2R	NT_033778.3	GI:116010442
	3L	NT_037436.3	GI:116010443
	3R	NT_033777.2	GI:56411841
	4	NC_004353.3	GI:116010290
	X	NC_004354.3	GI:116010291

Исходные последовательности ДНК полных геномов были получены с сайта ftp://ftp.ncbi.nih.gov/genomes/. Данные о хромосомах анализируемых организмов представлены в Таблице 1.

В настоящей работе для поиска размытых тандемных повторов используется оригинальный спектрально-статистический подход [26-28]. Выявление скрытой периодичности с помощью этого подхода, по существу, направлено на выделение значимой неоднородности на тест-периодах анализируемой нуклеотидной последовательности. Для каждого тест-периода L анализируемая последовательность разбивается на подстроки длины L (последняя подстрока может иметь меньшую длину). Если n – длина анализируемой последовательности, то $R_L = n/L$ – её тест-экспонент для тест-периода L. Такое разбиение на подстроки позволяет вычислить частоту π_j^i встречаемости i-той буквы алфавита нуклеотидной последовательности в j-той позиции тест-периода. Матрица $\pi = (\pi_j^i)_L^K$ называется выборочной L-профильной матрицей для анализируемой последовательности, где K – размер алфавита нуклеотидной последовательности. Частота p^i встречаемости i-той буквы алфавита в анализируемой последовательности определяется по матрице $\pi = (\pi_j^i)_L^K$:

$$p^i = \frac{1}{L}\sum_{j=1}^{L}\pi_j^i, \quad i = 1,...,K. \tag{1}$$

Для проверки однородности последовательности на тест-периоде L используется нормализованная χ^2-статистика Пирсона [28].

$$\nu_{NP}(L,n) = R_L \sum_{j=1}^{L}\sum_{i=1}^{K}(\pi_j^i - p^i)^2 \big/ p^i(1-p^i). \tag{2}$$

Поскольку поиск тандемных повторов происходит в нуклеотидных базах данных большого объема, для проверки однородности последовательностей ДНК был выбран уровень значимости (ошибка I-го рода) $\alpha = 10^{-6}$. Для фиксированного значения L тест-периода этому уровню соответствует критическое значение $\chi^2_{crit}(\alpha, N)$ с $N = (K-1)(L-1)$ степенями свободы. Следовательно, если значение статистики $\nu_{NP}(L,n)$ анализируемой строки длины n на тест-периоде L удовлетворяет условию

$$\nu_{NP}(L,n) \big/ \chi^2_{crit}(\alpha,(K-1)(L-1)) \le 1, \tag{3}$$

то на тест-периоде L принимается гипотеза об однородности строки, в противном случае строка признаётся неоднородной. Поэтому в качестве спектральной характеристики анализируемой нуклеотидной последовательности используется функция H вида

$$H(L) = \nu_{NP}(L,n) \big/ \chi^2_{crit}(\alpha,(K-1)(L-1)), \tag{4}$$

где $L = 1,...,L_{\max}$ $(L_{\max} \sim n/5K)$.

График функции H (H-спектр), называемый *спектром проявления неоднородности* в анализируемой последовательности, наглядно демонстрирует проявление значимых неоднородностей строки на тех тест-периодах, где $H(L) > 1$. Такие тест-периоды образуют *спектр структуры неоднородности* последовательности, который затем анализируется с помощью дополнительного спектрально-статистического параметра.

На каждом тест-периоде L анализируемой последовательности по выборочной L-профильной матрице $\pi = (\pi_j^i)_L^K$ вычисляется значение параметра

$$pl(L) = \frac{1}{L}\sum_{j=1}^{L} \max\{\pi_j^i : i \in 1,...,K\}, \qquad (5)$$

называемого *уровнем сохранности буквы* на тест-периоде L.

Таким образом, спектр структуры неоднородности анализируется с помощью спектра уровня сохранности буквы (pl-спектра) в рассматриваемой последовательности. Тест-период из спектра структуры неоднородности, на который указывает первый максимум в спектре уровня сохранности буквы, рассматривается как оценка размера паттерна периодичности в размытом тандемном повторе (см. рис. 1). Такое максимальное значение pl-спектра можно интерпретировать как средний индекс сохранности копий предполагаемого паттерна периодичности. Рисунок 1 является своего рода показательным примером того, как совместное использование именно двух параметров (H-спектра и pl-спектра) позволяет однозначно оценить размер паттерна периодичности. В анализируемой на рис. 1 последовательности спектр H выделяет структуру неоднородности на тест-периодах, кратных семи. Максимум в pl-спектре уровня сохранности буквы выделяет среди них тест-период в 21 нукл., который принимается в качестве оценки размера паттерна периодичности.

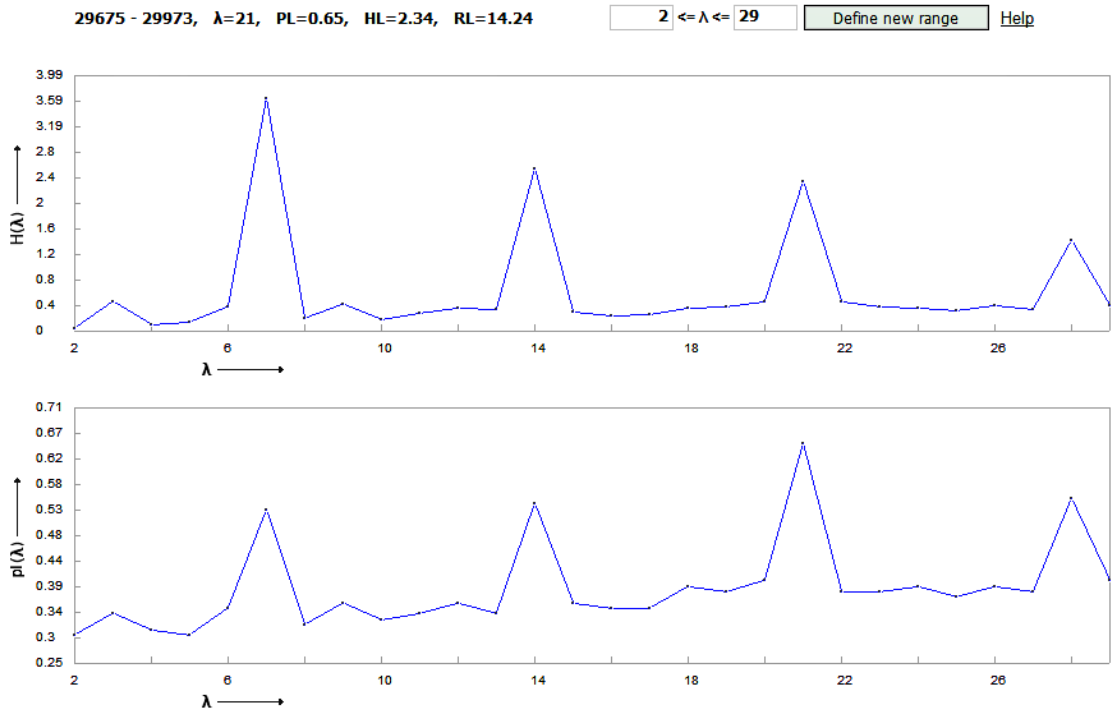

Рисунок 1. Пример спектральных характеристик в базе данных HeteroGenome для последовательности ДНК из генома *C. elegans*, хромосомы V (29675–29973 нукл.). Вверху: спектр проявления неоднородности (H-спектр, см. (4)). Внизу: спектр уровня сохранности буквы (pl-спектр, см. (5)). Наибольший пик pl-спектра соответствует длине паттерна скрытой периодичности в 21 нукл.

Проблема достоверности результатов выявления размытых тандемных повторов в условиях ограниченного статистического материала (достаточно малого числа копий паттерна в повторе) решается в рамках спектрально-статистического подхода на основе стохастической модели проявления неоднородности в текстовых строках, предложенной в [28]. Эта модель позволяет использовать дополнительные

статистические испытания при проверке гипотезы о наличии неоднородности в последовательности ДНК.

Поскольку с алгоритмической точки зрения определение скрытой периодичности является очень трудной задачей, если период априори неизвестен, то для выявления периодичности в последовательностях ДНК был выбран метод выявления районов высоко значимой (на уровне $\alpha = 10^{-6}$) неоднородности с помощью серии перекрывающихся окон, каждое из которых с переменным шагом сканирует анализируемую последовательность ДНК. Длина исходного окна серии составляет 30 нукл., при переходе к следующему окну его длина удваивается. Таким образом, общая стратегия программного комплекса, реализующего спектрально-статистический подход [28] для поиска размытых тандемных повторов, напоминает shotgun-стратегию секвенирования геномов [30]. В рамках такой стратегии сначала секвенируются относительно короткие и перекрывающиеся фрагменты, а затем производится их компьютерная сборка в более протяженные участки. Аналогично, полученные исходные данные о районах значимой неоднородности в геномах рассматриваемых модельных организмов прошли специальные процедуры дополнительной обработки и оптимизации границ выявленных районов неоднородности.

РЕЗУЛЬТАТЫ

Применение спектрально-статистического подхода [26-28] позволило создать комплекс программ, достоверно выявляющих районы скрытой периодичности, включая размытые и нечёткие тандемные повторы. Теоретический предельный уровень дивергенции копий паттерна в тандемных повторах, выявляемых этим комплексом, составляет 50%. Созданный комплекс программ одинаково хорошо выявляет как микросателлиты с длиной паттерна от 2 нукл., так и минисателлиты с паттерном ~ 10 – 100 нукл. наряду с мегасателлитами, имеющими паттерн длиной от 100 до 2000 нукл. Минимальное количество копий в выявляемых размытых тандемных повторах – две. Результаты поиска районов скрытой периодичности в полных геномах модельных организмов *S. cerevisiae, A. thaliana, C. elegans* и *D. melanogaster*, пройдя специальные процедуры по оптимизации их границ, вошли в базу данных HeteroGenome (http://www.jcbi.ru/lp_baze). На странице Database Statistics (см. рис. 2.) представлены результаты анализа данных, содержащихся в базе.

HeteroGenome является реляционной базой данных под управлением СУБД MySQL. Для удобства пользователя организована поисковая система по всем полям базы (см. рис. 2) с возможностью сортировки данных любого из полей (Region Location, Region Length, Period Length, Exponent, Preservation Level). Подробное описание всех полей и их возможных значений выводится в отдельных окнах, открывающихся при установке курсора на название поля. На сайте базы данных имеется руководство пользователя с примерами, демонстрирующими работу с базой. В HeteroGenome предусмотрена возможность загрузки результатов поиска в виде текстового файла на компьютер пользователя. В соответствии с двухуровневой структурой логической записи (см. рис. 3), описанной ниже, интерфейс базы HeteroGenome предлагает выбор уровня запроса информации: выборочный первый уровень (nonredundant) или общий второй уровень (simple), когда выводится список всех последовательностей, параметры которых удовлетворяют запросу.

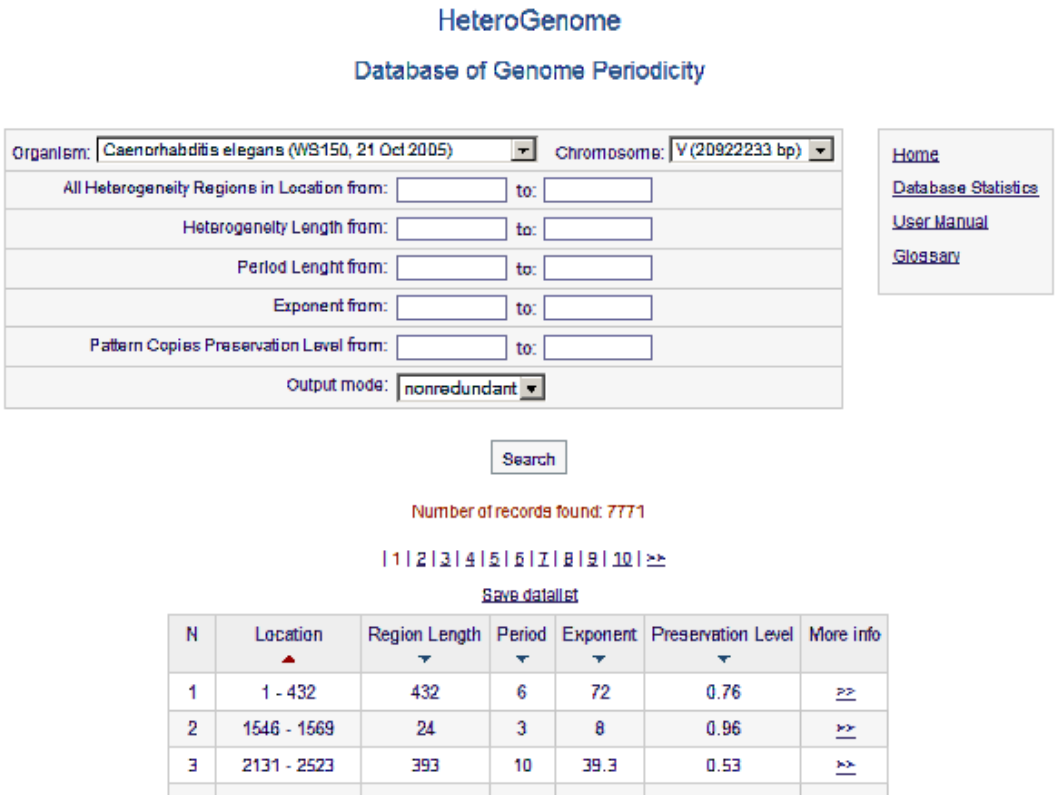

Рисунок 2. Форма запроса пользователя и вывод данных в базе HeteroGenome. Показан вывод всех данных о районах скрытой периодичности на хромосоме V из генома *C. elegans,* соответствующих первому, неизбыточному, уровню записей базы.

Для каждой последовательности в базе HeteroGenome в отдельных окнах (см. рис. 3) организован просмотр *H*- и *pl*-спектров (см. формулы (4), (5) и рис. 1), на основании которых выводится оценка размера паттерна скрытой периодичности. Кроме того, возможен просмотр последовательности в виде профиля, т.е. столбца её сегментов с длиной, равной оценке размера паттерна (см. рис. 4).

Руководствуясь информацией, полученной из визуального анализа спектров, пользователь имеет возможность изменять длину сегментов последовательности, чтобы уточнить оценку размера паттерна в размытом тандемном повторе. Используя встроенный графический интерфейс Sequence Viewer (http://www.ncbi.nlm.nih.gov/projects/sviewer/) (см. рис. 3), пользователь может получить информацию о локализации и функциональном контексте рассматриваемого района ДНК на хромосоме.

Для неизбыточного представления данных в базе HeteroGenome одна логическая запись соответствует группе последовательностей ДНК на хромосоме, обладающих статистически значимой неоднородностью (скрытой периодичностью), пересекающихся и (или) имеющих одинаковую или кратную длину периода. Два уровня представления данных выделяются в группе. На первом уровне рассматривается последовательность ДНК наибольшей длины, называемая представителем группы. Остальные последовательности группы относятся ко второму уровню. Как правило, они соответствуют хорошо детерминированным локальным структурам периодичности в ДНК представителя группы. Пример такой двухуровневой организации записи в HeteroGenome показан на рисунке 3. Значения параметров представителя группы (последовательности 1-го уровня) указаны в верхней таблице. Под заголовком INTRINSIC HETEROGENEITIES в отдельной таблице приведены значения параметров для последовательностей 2-го уровня. Общую структуру группы отражает масштабированная графическая схема, расположенная ниже.

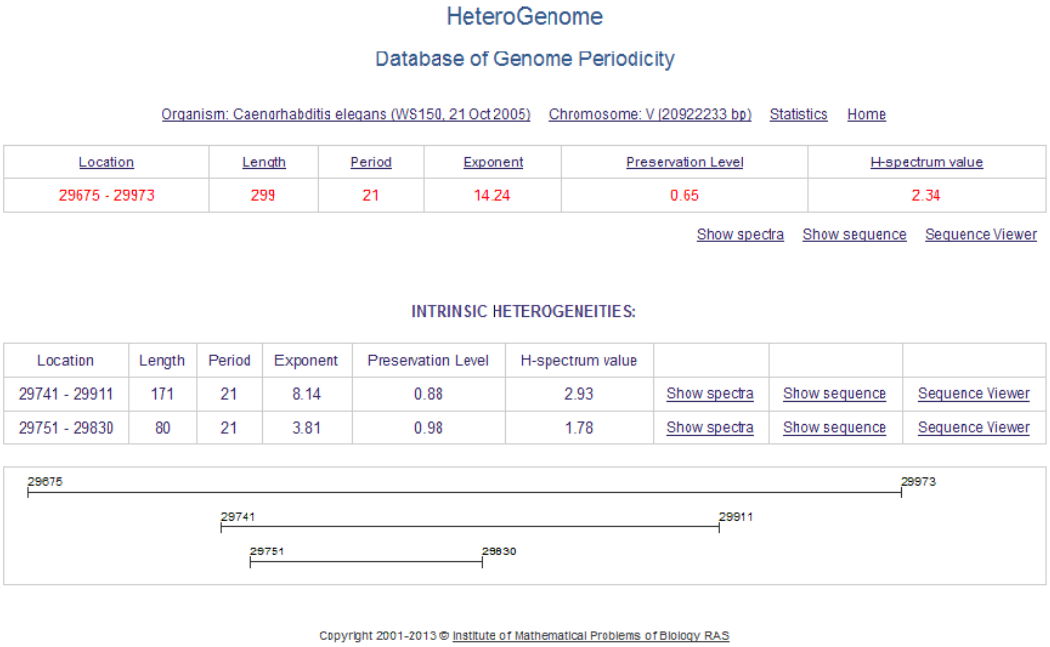

Рисунок 3. Двухуровневая структура записи в базе данных HeteroGenome. Последовательности ДНК из генома *C. elegans* с выявленной скрытой периодичностью в 21 нукл. на хромосоме V (29675–29973 нукл.) составляют группу, в которой выделяются два уровня. На первом уровне находится последовательность наибольшей длины, представляющая группу (её параметры указаны красным шрифтом). Второй уровень группы (INTRINSIC HETEROGENEITIES) образуют фрагменты детерминированной внутренней структуры в последовательности ДНК представителя группы. Внизу показана графическая схема всей группы. Последовательности первого уровня всех групп в базе HeteroGenome составляют неизбыточные данные о скрытой периодичности в геноме.

Кроме групп, содержащих последовательности ДНК обоих уровней (представителя группы и элементов его внутренней неоднородности), в базе имеются группы, содержащие только последовательность ДНК представителя группы. Отдельные группы практически не пересекаются между собой. Поэтому последовательности, представляющие группы, образуют неизбыточное покрытие хромосомы районами значимой неоднородности (скрытой периодичности).

Поиск информации о скрытой периодичности в базе HeteroGenome (с заданием длины периода или длины района периодичности, уровня сохранности копий паттерна и др.), как будет показано далее, можно проводить как на неизбыточном первом уровне, так и среди всех последовательностей обоих уровней.

Длина периода и координаты любого района скрытой периодичности (неоднородности) в базе HeteroGenome могут быть определены более точно в результате визуального анализа спектральных параметров (см. рис. 1), разбиения последовательности ДНК на сегменты предполагаемой длины периода (см. рис. 4) и задания длины участков, фланкирующих исходную последовательность.

В некоторых случаях дополнительное визуальное рассмотрение последовательностей группы способствует более корректному прочтению данных, анализ и распределение которых по группам были выполнены компьютерными программами. Проанализировав состав группы, пользователь может уточнить её размеры или разделить её на независимые подгруппы.

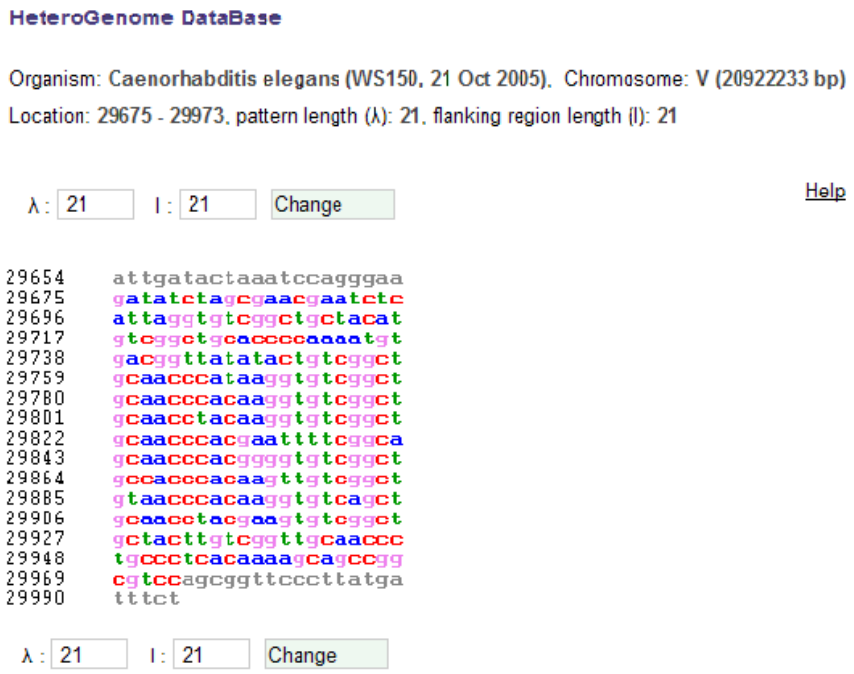

Рисунок 4. Пример профиля анализируемой последовательности в базе данных HeteroGenome. Разбиение на сегменты последовательности ДНК из генома *C. elegans*, хромосомы V (29675–29973 нукл.) выполнено согласно выявленной длине паттерна скрытой периодичности λ = 21 нукл. (см. рис. 1). Район скрытой периодичности представлен цветным шрифтом, а фланкирующие участки – серым. По умолчанию длина фланков равна длине паттерна (l = λ).

Анализ скрытой периодичности в HeteroGenome

Сравнение данных для геномов *S. cerevisiae*, *A. thaliana*, *C. elegans* и *D. melanogaster* из базы HeteroGenome с соответствующими данными из базы TRDB [24] показало, что HeteroGenome содержит практически все тандемные повторы, представленные в TRDB, и вместе с тем существенно дополняет их данными о сильно размытых тандемных повторах.

При исследовании эволюции и функционального значения районов скрытой периодичности в геноме немаловажным количественным показателем является доля длины полного генома, занимаемая такими районами. Неизбыточные данные о районах достоверной неоднородности (скрытой периодичности) в базе HeteroGenome позволяют достаточно точно оценить, какой процент в анализируемых геномах модельных организмов составляют тандемные повторы (включая сильно размытые повторы). Таблица 2 представляет полученные оценки.

Таблица 2. Доля районов неоднородности (скрытой периодичности) в геномах анализируемых модельных организмов

	Длина генома, нукл.	Общая длина выявленных районов неоднородности, нукл.	Районы неоднородности/Длина генома, %
S. cerevisiae	12070900	419909	3.5
A. thaliana	119146348	4247672	3.6
C. elegans	100269917	6692629	6.7
D. melanogaster	120381546	5108483	4.2

Как будет показано далее, большую часть районов периодичности в геномах проанализированных организмов составляют микро- и минисателлиты (длина периода меньше 100 нукл.). Известно, что в геноме человека на их долю приходится 3% [30]. Вместе с другими тандемными повторами (с длиной периода порядка 1000 нукл.) их доля составляет около 10% [19]. Принимая во внимание также и данные таблицы 2, можно предположить, что периодичность в геномах эукариот варьирует в пределах 10%. Возможно, такой процент обусловлен балансом между молекулярными механизмами возникновения тандемных повторов и дивергенции их последовательностей, стабилизирующей длину повторов.

Влияние скрытой периодичности на длину хромосом

Районы периодичности являются нестабильными участками в геноме, способными как увеличиваться, так и уменьшаться в длине за счёт механизмов проскальзывания ДНК репликазы, рекомбинации и дупликации [2-4]. Со временем мутации (точечные замены, вставки/делеции нуклеотидов) нарушают детерминированную структуру ДНК районов периодичности, стабилизируя их длину. Благодаря тому, что используемый в работе метод выявления размытых тандемных повторов позволяет без избыточности оценить их долю в геноме, можно исследовать влияние эволюции районов периодичности на хромосомы. Рассмотрим долю районов скрытой периодичности (достоверной неоднородности, представленной тандемными повторами) в зависимости от длины хромосом анализируемых модельных организмов (см. рис. 5). Для каждого модельного организма наблюдается характерный разброс значений процента покрытия районами периодичности отдельных хромосом.

Наибольшее различие (4.95%) между максимальным (8.91%, хромосома I) и минимальным (3.96%, хромосома X) значениями наблюдается для *C. elegance*. Приблизительно на одну треть меньше диапазон значений процента покрытия хромосом в геноме *D. melanogaster* (3.11%) – от 3.30% (хромосома IV) до 6.41% (хромосома X). Для обоих организмов диапазон значений обусловлен особым положением X хромосомы, причем в случае *C. elegance* хромосома имеет минимальный процент покрытия, а в случае *D. melanogaster* хромосома X имеет максимальный процент покрытия среди всех хромосом генома.

Диапазон значений процента покрытия хромосом районами периодичности (достоверной неоднородности) в геноме *D. melanogaster* сравним с таким диапазоном в геноме *S. cerevisiae*. Последний диапазон составляет 3.57% между максимальным (6.27%, хромосома I) и минимальным (2.7%, хромосома XVI) значениями. Отметим, что хромосома I, лидирующая по проценту периодичности в геноме *S. cerevisiae*, является самой короткой хромосомой дрожжей.

Если в геномах *S. cerevisiae*, *C. elegans* и *D. melanogaster* диапазон значений процента периодичности на хромосомах сравним с его средним значением для отдельных геномов, то в геноме *A. thaliana* такой диапазон не превышает 0.75%. Как можно видеть (см. рис. 5), с ростом длины хромосомы арабидопсиса доля районов периодичности остаётся практически постоянной. Вообще говоря, для всех проанализированных геномов модельных организмов с ростом длины хромосом доля районов периодичности имеет тенденцию оставаться неизменной или сокращаться. Можно полагать, что, благодаря своей нестабильности, способности к удлинению из-за ошибок репликации, тандемные повторы оказали хотя и небольшое (~10%), но всё-таки заметное влияние на рост длины хромосом исследуемых модельных организмов.

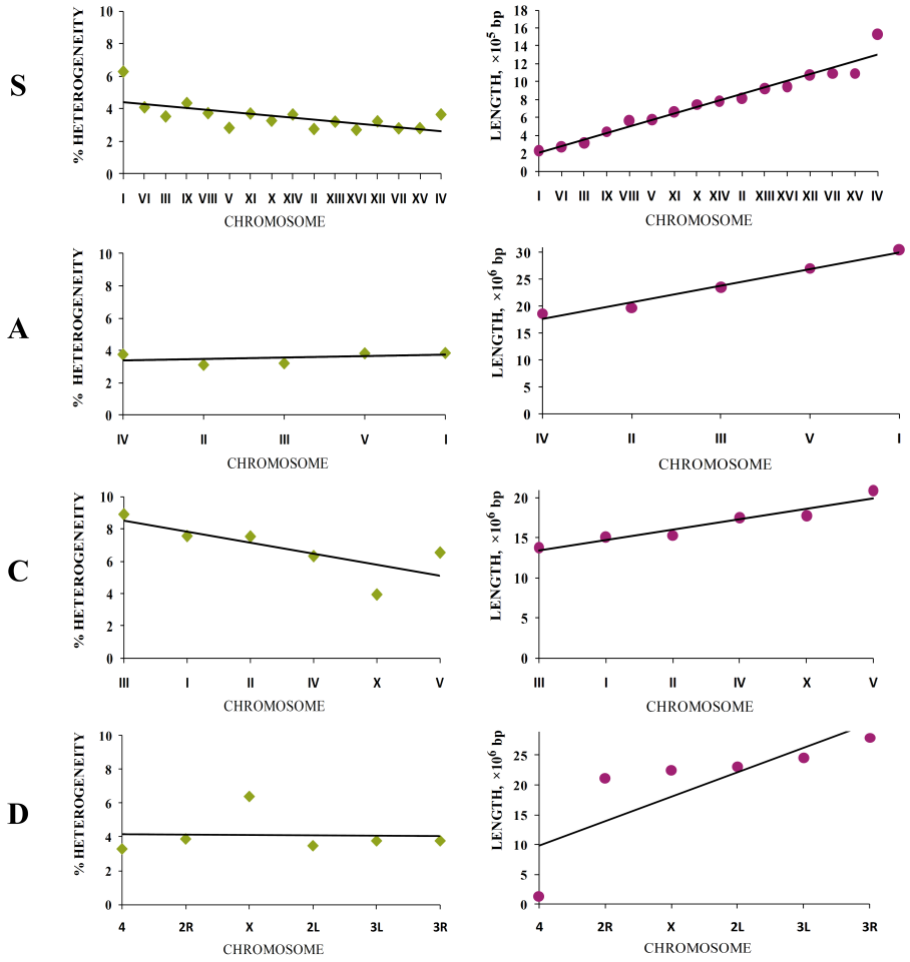

Рисунок 5. Процент покрытия хромосом модельных организмов *S. cerevisiae* (S), *A. thaliana* (A), *C. elegans* (C) и *D. melanogaster* (D) районами скрытой периодичности. Для каждого организма хромосомы упорядочены по возрастанию длины, что показывают соответствующие графики справа. Сплошными прямыми линиями на графиках показаны линии тренда. Процентная доля районов скрытой периодичности на хромосоме была определена в соответствии с неизбыточным уровнем записей данных в базе HeteroGenome.

Анализ сохранности периодической структуры в районах неоднородности

Согласно данным базы HeteroGenome, серия рисунков 6–9 представляет гистограммы распределения выявленных районов скрытой периодичности в соответствии с уровнем сохранности (параметром *pl*, см. формулу (5)) их периодической структуры. Отдельно для групп микро- (длина периода $2 \le L \le 10$), мини- ($10 < L \le 100$) и мегасателлитов ($100 < L \le 2000$) показан процент от длины каждой хромосомы, занимаемый сильно размытыми ($0.4 \le pl \le 0.7$), размытыми ($0.7 < pl \le 0.8$), слабо размытыми ($0.8 < pl \le 0.9$) и совершенными ($0.9 < pl \le 1.0$) тандемными повторами таких типов. Порядок хромосом на рисунках 6–9 определяется видимым сходством их гистограмм.

Как видно из рис. 6, районы скрытой периодичности в геноме *S. cerevisiae* в основном представлены сильно размытыми последовательностями микросателлитов, доля которых в геноме составляет ~2%. Сильно размытые минисателлиты составляют менее 1% генома. За исключением хромосом I и IX, в геноме дрожжей *S. cerevisiae* практически отсутствуют мегасателлитные повторы с длиной периода более 100 нукл.

Некоторые хромосомы в геноме *S. cerevisiae* имеют похожие гистограммы во всех трёх рассматриваемых группах сателлитов, например хромосомы VIII и XII, хромосомы XIII, VI, XIV и также хромосомы XI, XV, VII.

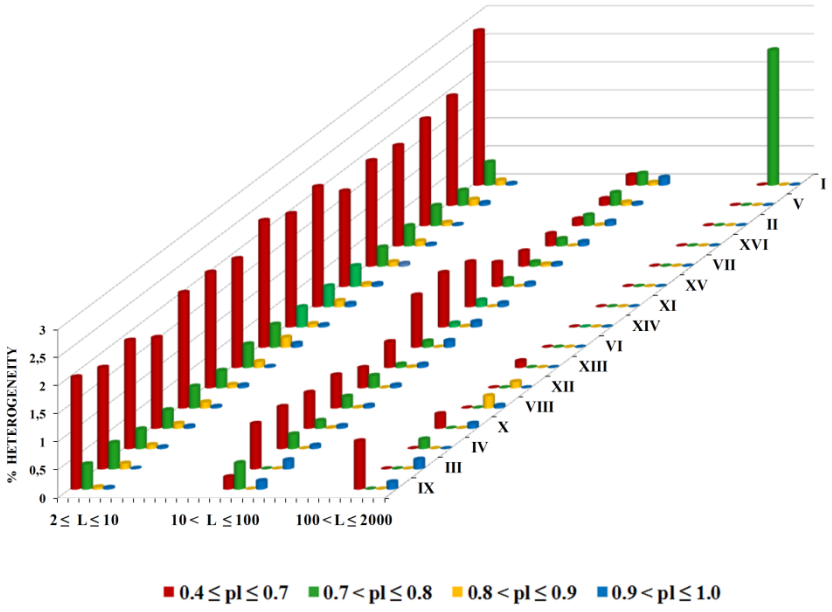

Рисунок 6. Гистограммы структурного состава районов скрытой периодичности в геноме *S. cerevisiae* (хромосомы I – XVI).

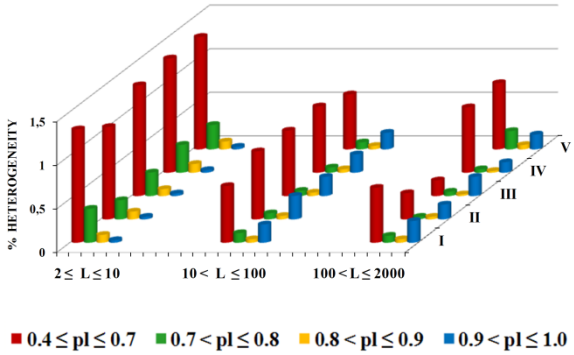

Рисунок 7. Гистограммы структурного состава районов скрытой периодичности в геноме *A. thaliana* (хромосомы I – V).

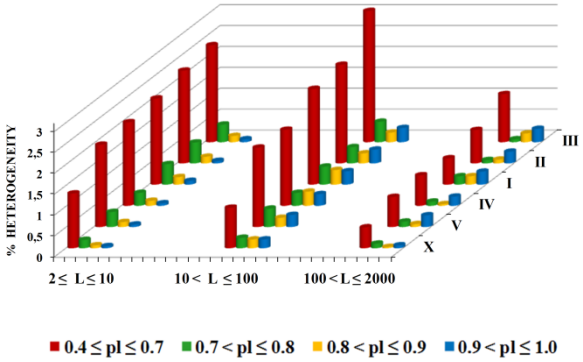

Рисунок 8. Гистограммы структурного состава районов скрытой периодичности в геноме *C. elegans* (хромосомы I – V, X).

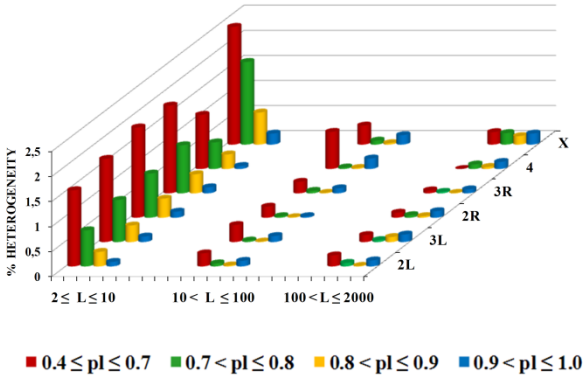

Рисунок 9. Гистограммы структурного состава районов скрытой периодичности в геноме *D. melanogaster* (хромосомы 2L, 3L, 2R, 3R, 4, X).

Для растения резуховидки *A. thaliana* (рис.7) и круглого червя *C. elegans* (рис. 8), длина геномов которых на порядок больше длины генома дрожжей *S. cerevisiae*, представленные гистограммы выявляют сходные тенденции качественного распределения доли тандемных повторов в геноме. Во-первых, сильно размытые минисателлиты составляют в обоих геномах значительную долю ~1–1.5%, сравнимую с долей микросателлитов. Следовательно, у *A. thaliana* и *C. elegans* и минисателлиты, и микросателлиты вносят сопоставимый вклад в структурную и функциональную организацию генома. Во-вторых, доля мегасателлитных повторов, практически отсутствующих в геноме дрожжей (и как будет показано далее, также отсутствующих в геноме дрозофилы), в геномах резуховидки и круглого червя достаточно заметна и составляет ~1%.

Если не принимать во внимание некоторые особенности для хромосом II и III, гистограммы генома *A. thaliana* можно считать похожими друг на друга (см. рис. 7). Также и гистограммы хромосом *C. elegans* (см. рис. 8), за исключением гистограммы для X хромосомы, можно рассматривать как практически идентичные.

Гистограммы, построенные на основе анализа структуры районов периодичности, выявленных на хромосомах плодовой мушки *D. melanogaster*, показаны на рис. 9. Можно видеть, что среди тандемных повторов в геноме *D. melanogaster* доминируют сильно размытые (~1.5% – 2%) и размытые микросателлиты (~0.5% – 1%). Следует также отметить сходство гистограмм для хромосом 2L, 3L, 2R и 3R на рис 9.

Сходство гистограмм, отражающих структурный (качественный) состав районов периодичности для отдельных хромосом модельных организмов, рассмотренных в работе, конечно, не может служить доказательством общего эволюционного происхождения таких хромосом, однако позволяет предположить существование сходных механизмов эволюционного давления и дивергенции, под воздействием которых они формировались. Кроме того, из анализа рисунков 6–9 следует, что в каждом геноме можно выделить один или несколько типов характерных доминирующих периодичностей, например, таких как сильно размытые микросателлиты в геноме *S. cerevisiae*. Для *A. thaliana* и *C. elegans* можно заметить, что их геномы имеют сходный процентный состав характерных типов периодичностей. Можно предположить, что значительный процент (~1.5%) мини- и мегасателлитов является следствием активных рекомбинационных процессов [3] в геномах арабидопсиса и нематоды. Доминирование микросателлитов в геноме дрожжей, возможно, связано с большим числом репликаций ДНК генома при их размножении, а значит и с большей частотой ошибки «проскальзывания» [2], ведущей к удлинению

районов микросателлитов. Установление функциональной роли доминирующих периодичностей в геноме является возможным предметом будущих исследований.

Выявление скрытой периодичности в функциональных районах генома

Используя графический интерфейс Sequence Viewer (http://www.ncbi.nlm.nih.gov/projects/sviewer), для любого района неоднородности в базе HeteroGenome можно получить сведения о его пересечении с аннотированными последовательностями исследуемого генома.

Для представителей групп районов скрытой периодичности в геномах модельных организмов, выявленных с помощью спектрально-статистического подхода, таблицы 3 и 4 представляют общую картину их распределения по аннотированным (функциональным) и неаннотированным (unassigned) в GenBank последовательностям ДНК.

Таблица 3. Распределение групп HeteroGenome по функциональным районам геномов, отмеченным в аннотациях геномов, представленных в GenBank

GenBank характеристики	*S. cerevisiae*[*]	*A. thaliana*	*C. elegans*	*D. melanogaster*
gene	*3276*	*21598*	*25551*	*48657*
mRNA	–	15337	12667	23879
CDS	3269	13370	12046	19463
intron	6	3105	14145	28112
exon	–	–	–	21
STS	–	141	–	150
rep_origin	15	–	–	–
repeat_region	95	–	–	1951
unassigned	738	12935	13773	22851
Полное число групп HeteroGenome	4094	34566	39329	72772

[*]В аннотации генома *S. cerevisiae* отсутствуют характеристики mRNA

Чтобы оценить распределение групп базы HeteroGenome по аннотированным районам генома, использовали достаточно нестрогий критерий. Если область пересечения группы и функционального района составляла не менее 50% от длины меньшей последовательности, в таком случае группу приписывали к рассматриваемому району. Кроме того, при оценке распределения групп по аннотированным районам данные по альтернативному сплайсингу не принимались во внимание, т.е. для одного гена рассматривались только одна мРНК и одна кодирующая последовательность (CDS – Coding DNA Sequence).

Как видно из таблицы 3, согласно выбранному критерию для геномов *S. cerevisiae*, *A. thaliana*, *C. elegans* and *D. melanogaster*, соответственно, 80%, 62%, 65% and 67% групп базы HeteroGenome располагаются в генах. В соответствии с этим списком организмов, 18%, 37.4%, 35% и 31.4% групп базы располагаются в неаннотированных районах (unassigned) их геномов. В целом, распределение по всем остальным функциональным районам, кроме генов, имеет случайный и незначительный характер. Хотя следует отметить, что 2.6% групп, располагаются в районах различных повторов генома *D. melanogaster*.

При расположении групп в генах есть вероятность приписать одну и ту же группу и к интрону, и к экзону, если она пересекает их границу. Поэтому был проведён дополнительный анализ по распределению количества нуклеотидов в группах базы HeteroGenome между интронами и экзонами. Таблица 4 показывает результаты этого анализа в сравнении с полным числом нуклеотидов в рассматриваемых функциональных районах всего генома.

Таблица 4. Общее количество нуклеотидов из групп HeteroGenome, находящихся в составе генов, экзонов и интронов (в скобках приводится доля таких нуклеотидов от общей длины генов, экзонов и интронов в геноме)

Организм	число нуклеотидов из групп HeteroGenome / число соответствующих функциональных нуклеотидов в геноме		
	гены	экзоны*	интроны*
S. cerevisiae	354459 / 8829668, (4%)	352781 / 8737430, (4%)	448 / 64756, (0,7%)
A. thaliana	2037728 / 70751773, (3%)	1601059 / 40167753, (4%)	296614 / 19355644, (1,5%)
C. elegans	3816463 / 60377579, (6,3%)	1479515 / 27267246, (5,4%)	402554 / 32373603, (10%)
D. melanogaster	3786123 / 79344223, (4,8%)	3410523 / 28271547, (12%)	4419548 / 49121309, (9%)

*Данные по экзонам и интронам были получены согласно индексам сборки мРНК (join attribute line). В случае генома *S. cerevisiae* использовались индексы сборки CDS.

Сравнение таблиц 4 и 2 показывает, что доля нуклеотидов, приходящихся на гены, во всех группах базы практически совпадает с долей покрытия геномов районами значимой неоднородности (скрытой периодичности). Распределение нуклеотидов между экзонами и интронами зависит от организма. Так, например, процент общей длины районов скрытой периодичности в экзонах *D. melanogaster* (12%) выше, чем в интронах (9%), несмотря на то, что суммарная длина интронов почти в 2 раза больше суммарной длины экзонов. И наоборот, при сравнимой общей длине экзонов и интронов в геноме *C. elegans* доля районов скрытой периодичности в интронах в 2 раза больше, чем в экзонах (10% и 5,4%, соответственно). Следовательно, прямая зависимость длины районов скрытой периодичности в геномах организмов от длины генома или его функциональных районов не наблюдается.

Распределение плотности районов скрытой периодичности на хромосомах

Исследование распределения плотности районов скрытой периодичности на хромосомах проводилось для всех анализируемых в данной работе организмов. С этой целью каждая хромосома разбивалась на последовательные фрагменты равной длины, соответствующей 0,5% от длины всей хромосомы. Длина фрагмента является шагом разбиения. Для каждого фрагмента определялась суммарная длина (в нуклеотидах) районов скрытой периодичности, выявленных в его границах. Такая длина, нормированная на длину всей хромосомы, и умноженная на 100%, рассматривалась как часть общей доли скрытой периодичности хромосомы, заключённой в рассматриваемом фрагменте. Суммирование по всем фрагментам разбиения даёт оценку общей процентной доли длины районов скрытой периодичности на хромосоме.

При исследовании распределения плотности этих районов учитывались только последовательности – представители групп, дающие неизбыточную оценку длины покрытия хромосом скрытой периодичностью.

Гистограммы на рисунке 10 демонстрируют плотность распределения районов скрытой периодичности на всех хромосомах генома резуховидки Таля (*A. thaliana*). Соответствующие шаги разбиения для хромосом I - V: 152138, 98491, 117299, 92925, 134877 нукл. Результаты по всем хромосомам рассматриваемых организмов представлены на странице Database Statistics базы HeteroGenome.

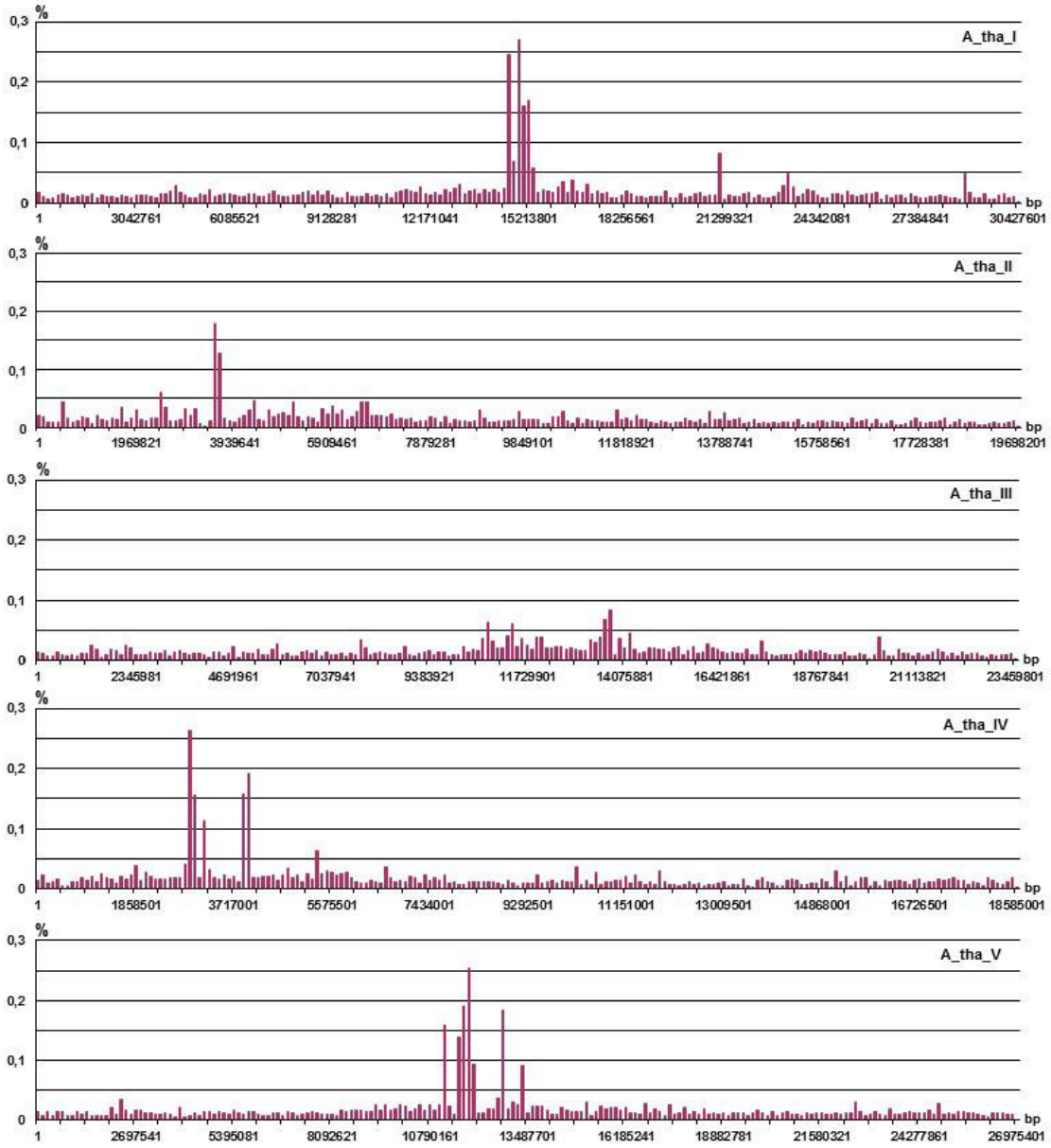

Рис. 10. Распределение плотности районов скрытой периодичности на хромосомах *A. thaliana*. Высота столбца гистограммы показывает процентную долю локальных районов скрытой периодичности, выявленных в пределах соответствующего шага гистограммы, от полной длины хромосомы.

Кроме того, для каждой хромосомы были получены отдельные распределения по трём классам периодичности, т.е. микро- (длина периода $2 \le L \le 10$), мини- ($10 < L \le 100$) и мега- ($L > 100$) сателлитам. На рисунке 11 показан пример таких распределений для хромосомы I из генома *A. thaliana*. Как видно из примеров гистограмм на рис. 10 и рис. 11, плотность распределения районов скрытой периодичности на хромосомах однозначно характеризует каждую хромосому из геномов рассматриваемых организмов. Такое распределение можно рассматривать как

своеобразный ДНК-фингерпринт или индивидуальный штрих-код каждой хромосомы в геномах различных организмов.

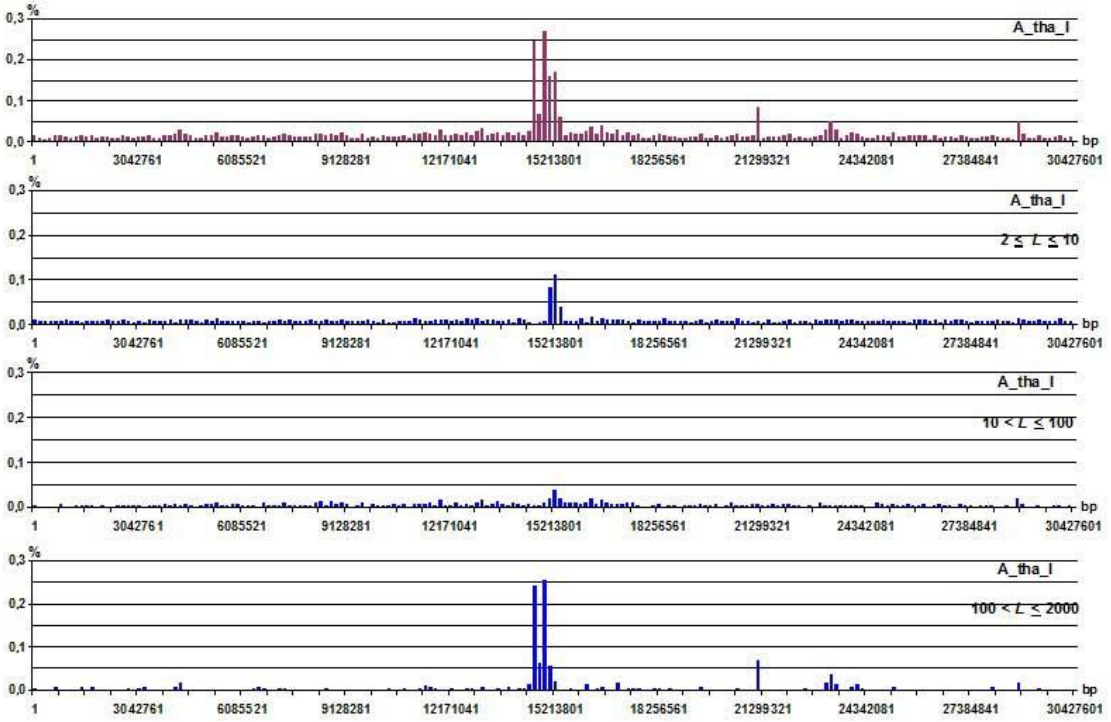

Рис. 11. Пример разложения общего распределения плотности районов скрытой периодичности, выявленных на хромосоме I из генома *A. thaliana* (верхняя гистограмма), по трём классам периодичности (микро-, мини-, и мегасателлиты).

Репертуар длин периодов в геноме

Все полученные оценки для длин паттернов скрытой периодичности были проанализированы по частоте встречаемости на каждой хромосоме для каждого организма. Рисунки 12 и 13 демонстрируют качественную картину частоты встречаемости различных значений длин паттернов периодичности. По горизонтальной оси отложены значения длин паттернов от 2 до 100 нукл., по вертикальной оси – натуральный логарифм от числа участков скрытой периодичности с конкретной длиной периода. Графики на рис. 12, соответствующие геномам *S. cerevisiae* и *D. melanogaster*, показывают, что “репертуар” длин периодов (т.е. ряд значений длин паттернов скрытой периодичности) по всем хромосомам одного генома практически совпадает. Однако, несмотря на специфику репертуара каждого отдельного генома, в некоторых случаях можно выделить общие черты. Так, например, для *S. cerevisiae* и *D. melanogaster* практически все характерные длины периодов кратны трем. Это связано с тем, что длина области пересечения районов скрытой периодичности с генами, кодирующими белки, в геномах *S. cerevisiae* и *D. melanogaster* составляет соответственно 87% и 66% от общей длины выявленных районов. Пики на графиках показывают характерные длины паттернов, среди которых общими для обоих организмов являются длины 3, 6, 12, 15, 18, 21, 24, 27, 33, 36, 45, 57.

Графики на рис. 13, соответствующие геномам *C. elegans* и *A. thaliana*, демонстрируют, что эти организмы имеют более богатые репертуары характерных длин периодов, где помимо значений, кратных трем, встречаются значения, отклоняющиеся в пределах единицы от значений, кратных трем. Это может быть обусловлено двумя факторами. Во-первых, меньшая доля длины выявленного покрытия этих геномов районами скрытой периодичности лежит в генах, кодирующих

белки (55% и 38% для *C. elegans* и *A. thaliana*, соответственно). Во-вторых, можно предположить, что в этих геномах мутационные процессы протекают более активно. Как видно из сравнения рисунков 6, 9 (*S. cerevisiae* и *D. melanogaster*) с рисунками 7, 8 (*C. elegans* и *A. thaliana*), в геномах первых двух организмов доминируют микросателлиты, которые формируются, главным образом, благодаря проскальзыванию ДНК-полимеразы при репликации; а в геномах *C. elegans* и *A. thaliana* помимо микросателлитов выявляется значительная доля мини- и мегасателлитов, распространению которых способствуют различные рекомбинационные процессы и геномные дупликации.

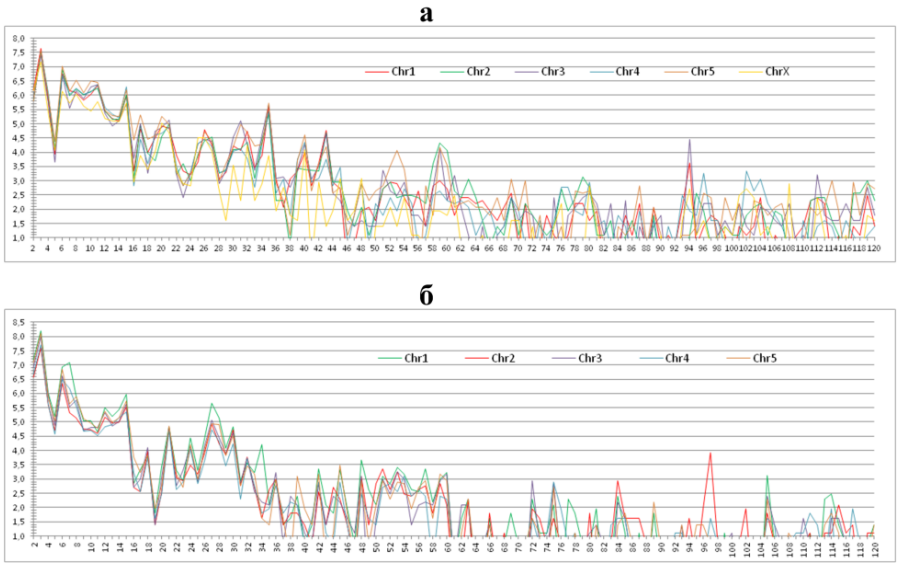

Рис. 12. Графики, представляющие репертуар значений длин паттернов скрытой периодичности в геномах *C. elegans* (а) и *A. thaliana* (б). Разными цветами изображены графики, соответствующие хромосомам.

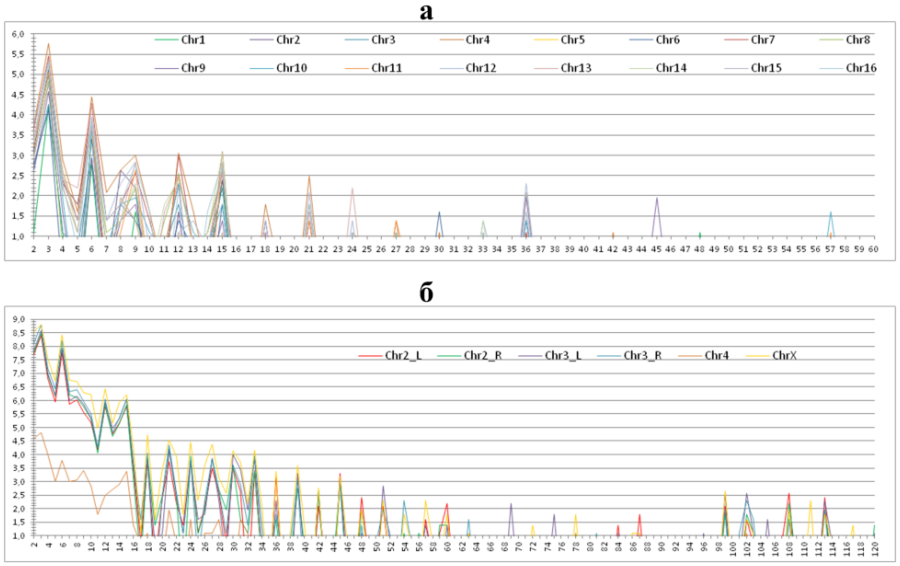

Рис. 13. Графики, представляющие репертуар значений длин паттернов скрытой периодичности в геномах *S. cerevisiae* (а) и *D. melanogaster* (б). Разными цветами изображены графики, соответствующие хромосомам.

ЗАКЛЮЧЕНИЕ

В результате применения спектрально-статистического подхода к выявлению районов неоднородности (скрытой периодичности) в геномах модельных организмов *S. cerevisiae, A. thaliana, C. elegans* и *D. melanogaster* были получены достоверные данные о тандемных повторах, в том числе и сильно размытых повторах, которые вошли в первый выпуск базы данных HeteroGenome (http://www.jcbi.ru/lp_baze/).

Специально разработанная двухуровневая структура логических записей в базе данных позволила представить имеющиеся данные в виде непересекающихся между собой районов скрытой периодичности на хромосомах (неизбыточное представление данных) и вместе с тем указать наиболее консервативные участки периодической структуры таких районов.

Благодаря понятному интерфейсу и возможности дополнительного анализа данных база HeteroGenome может быть полезна для молекулярно-генетических исследований модельных организмов и дальнейшего изучения феномена скрытой периодичности в последовательностях ДНК.

В результате анализа полученных данных для геномов *S. cerevisiae, A. thaliana, C. elegance* и *D. melanogaster* можно сделать вывод, что районы скрытой периодичности составляют ~10% в геномах различных организмов. Во всех перечисленных выше модельных организмах доминируют сильно размытые микросателлиты (с диной периода не более 10 нукл.), составляющие ~2% от длины генома. В работе было показано, что для каждого генома выявляется характерный качественный и количественный состав районов скрытой периодичности. Например, в геномах дрожжей *S. cerevisiae* (за исключением хромосом I и IX) и дрозофилы *D. melanogaster* практически отсутствуют мегасателлитные повторы (с длиной периода более 100 нукл.). Напротив, доля мегасателлитных повторов в геномах резуховидки *A. thaliana* и круглого червя *C. elegance* достаточно заметна и составляет ~1%.

Анализ распределения данных базы HeteroGenome между функциональными (аннотированными в GenBank) и нефункциональными (unassigned) последовательностями ДНК генома показал, что больше половины районов скрытой периодичности в проанализированных геномах выявлено в генах, кодирующих белки.

Полученные распределения плотности районов скрытой периодичности на хромосомах демонстрируют их хромосомоспецифичность. По-видимому, картина распределения районов периодичности является уникальной характеристикой, или идентификатором, каждой хромосомы и отражает, в конечном счете, картину пространственного позиционирования транскрипционно активной ДНК.

Были проанализированы репертуары длин паттернов скрытой периодичности в геномах пекарских дрожжей, резуховидки Таля, круглого червя и плодовой мухи. Показано, что геном каждого организма имеет специфический репертуар, однако внутри генома репертуар всех хромосом практически совпадает. Некоторые организмы имеют небогатые репертуары с четко определяемыми значениями характерных длин периодов, как, например, для дрожжей (*S. cerevisiae*) и мухи (*D. melanogaster*). Для двух других организмов наблюдается большое разнообразие значений длин периодов. Предположительно, оно является следствием активно идущих процессов рекомбинации и дупликации в геноме.

Работа была выполнена при поддержке гранта № 12-07-00530 Российского фонда фундаментальных исследований (РФФИ).

СПИСОК ЛИТЕРАТУРЫ

1. Richard G.F., Kerrest A., Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.* 2008. V. 72. P. 686–727.
2. Kelkar Y.D., Strubczewski N., Hile S.E., Chiaromonte F., Eckert K.A., Makova K.D. What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biol. Evol.* 2010. V. 2. P. 620–635.
3. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 2004. V. 5. P. 435–445.
4. Welch J.W., Maloney D.H., Fogel S. Unequal crossing-over and gene conversion at the amplified CUP1 locus of yeast. *Mol. Gen. Genet.* 1990. V. 222. P. 304–310.
5. Tyler-Smith,C. and Willard,H.F. Mammalian chromosome structure. *Curr. Opin. Genet. Dev.* 1993. V. 3. P. 390–397.
6. Hewett D.R., Handt O., Hobson L., Mangelsdorf M., Eyre H.J., Baker E., Sutherland G.R., Schuffenhauer S., Mao J.I., Richards R.I. FRA10B structure reveals common elements in repeat expansion and chromosomal fragile site genesis. *Mol. Cell.* 1998. V. 1**.** P. 773–781.
7. Yu S., Mangelsdorf M., Hewett D., Hobson L., Baker E., Eyre H.J., Lapsys N., Le Paslier D., Doggett N.A., Sutherland G.R., Richards R.I. Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. *Cell*. 1997. V. 88. P. 367–374.
8. Fu Y.H., Kuhl D.P., Pizzuti A., Pieretti M., Sutcliffe J.S., Richards S., Verkerk A.J., Holden J.J., Fenwick R.G. Jr, Warren S.T., et al. Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell*. 1991. V. 67. P. 1047–1058.
9. Liquori C.L., Ricker K., Moseley M.L., Jacobsen J.F., Kress W., Naylor S.L., Day J.W., Ranum L.P. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science*. 2001. V. 293. P. 864–867.
10. Matsuura T., Fang P., Pearson C.E., Jayakar P., Ashizawa T., Roa B.B., Nelson D.L. Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: repeat purity as a disease modifier? *Am. J. Hum. Genet.* 2006. V. 78. P. 125–129.
11. Lalioti M.D., Scott H.S., Buresi C., Rossier C., Bottani A., Morris M.A., Malafosse A., Antonarakis S.E. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*. 1997. V. 386. P. 847–851.
12. Martin P., Makepeace K., Hill S.A., Hood D.W., Moxon E.R. Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. USA*. 2005. V. 102. P. 3800–3804.
13. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*. 1999. V. 27. P. 573–580.
14. Reneker J., Shyu C.R., Zeng P., Polacco J.C., Gassmann W. ACMES: fast multiple-genome searches for short repeat sequences with concurrent cross-species information retrieval. *Nucleic Acids Res*. 2004. V. 32. P. W649–W653.
15. Roset R., Subirana J.A., Messeguer X. MREPATT: detection and analysis of exact consecutive repeats in genomic sequences. *Bioinformatics*. 2003. V. 19. P. 2475–2476.
16. Parisi V., Fonzo V.D., Aluffi-Pentini F. STRING: finding tandem repeats in DNA sequences. *Bioinformatics*. 2003. V. 19. P. 1733–1738.
17. Kolpakov R., Kucherov G. Mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* 2003. V. 31. P. 3672–3678.

18. Wexler Y., Yakhini Z., Kashi Y., Geiger D. Finding approximate tandem repeats in genomic sequences. *J. Comput. Biol.* 2005. V. 12. P. 928–942.
19. Boeva V., Regnier M., Papatsenko D., Makeev V. Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics*. 2006. V. 22. P. 676–684.
20. Mudunuri S.B., Nagarajaram H.A. IMEx: imperfect microsatellite extractor. *Bioinformatics*. 2007. V. 23. P. 1181–1187.
21. Pellegrini M., Renda M.E., Vecchio A. TRStalker: an efficient heuristic for finding fuzzy tandem repeats. *Bioinformatics*. 2010. V. 26. P. i358–i366.
22. Sokol D., Benson G., Tojeira J. Tandem repeats over the edit distance. *Bioinformatics*. 2007. V. 23. P. e30–e35.
23. Sokol D., Atagun F. TRedD – A database for tandem repeats over the edit distance. *Database*. 2010. Article ID baq003.
24. Gelfand Y., Rodriguez A., Benson G. TRDB – the Tandem Repeats Database. *Nucleic Acids Res.* 2007. V. 35. P. 80–87.
25. Boby T., Patch A., Aves S. TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics*. 2005. V. 21. P. 860–921.
26. 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 Recogn. Image Anal.* 2009. V. 19. P. 358–367.
27. Чалей М.Б., Назипова Н.Н., Кутыркин В.А. Совместное использование различных критериев проверки однородности для выявления скрытой периодичности в биологических последовательностях. *Мат. биол. и биоинформ*. 2007. Т. 2. №1. С. 20-35. URL: http://www.matbio.org/downloads/Chaley2007(2_20).pdf (дата обращения: 28.07.2013).
28. Chaley M., Kutyrkin V. Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences. *Math. Biosci.* 2008. V. 211. P. 186–204.
29. Fields S., Johnston M. Cell biology. Whither model organism research? *Science*. 2005. V. 307. P. 1885–1886.
30. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001. V. 409. P. 860–921.

Материал поступил в редакцию 23.07.2013, опубликован 24.09.2013.