Математическая биология и биоинформатика. 2011. Т. 6. № 2. С. 147–160.
URL: http:// www.matbio.org/2011/Porozov2011(6_147).pdf

========== **БИОИНФОРМАТИКА** ==========

УДК: 577.213+519.246.87

Применение вейвлет-преобразования к задаче поиска сайтов начала репликации в ДНК человека

Щербаков А.А.[*1], Порозов Ю.Б.[†2]

[1] *Саратовский государственный технический университет, 410054, Саратов, ул. Политехническая, 77*
[2] *Санкт-Петербургский государственный университет информационных технологий, механики и оптики, 197101 Санкт-Петербург, Кронверкский пр., 49*

Аннотация. Поиск сайтов начала репликации (ORI) в ДНК человека – актуальная для современной биологии проблема. Предложенный нами подход для предсказания положения ORI в человеческом геноме базируется на поиске скрытых зависимостей в последовательностях нуклеотидов при помощи метода сравнения вейвлет-спектров. Подбор параметров вейвлет-преобразования был осуществлен на основе данных о положении некоторых известных ORI в геноме человека. Применение вейвлет-преобразования с найденными параметрами для предсказания ORI в последовательностях с неизвестным расположением сайтов начала репликации и в случайных последовательностях нуклеотидов показало хорошие результаты. Результаты анализа вейвлет-спектров последовательностей нуклеотидов могут быть применены как самостоятельно в качестве индикаторов положения ORI, так и как один из факторов в различных классификаторах, таких, как байесовский классификатор, метод опорных векторов и других. В работе указаны преимущества и недостатки метода и приведены возможные пути повышения его эффективности.

Ключевые слова: *вейвлет-спектр, ДНК, спектральный анализ, сайт начала репликации.*

ВВЕДЕНИЕ

Репликация ДНК является одним из важнейших клеточных процессов, ответственных за точную передачу генетической информации в последовательных поколениях клеток. Этот процесс начинается с определенного места в последовательности ДНК, называемого сайтом начала репликации (origin of replication, ORI). Репликация ДНК может быть как однонаправленной, так и двунаправленной, что определяется характером перемещения одной или двух расходящихся вилок репликации вдоль молекулы ДНК. Инициирование ORI у эукариотов связано с фазами клеточного цикла и может сильно зависеть от расстояния и времени активации соседних ORI, от транскрипционной активности, а также от локальной структуры хроматина [1]. Следует отметить, что последовательность нуклеотидов, распознаваемая белками, инициирующими процесс репликации, существенно различается как в различных эукариотических организмах, так и в пределах одной молекулы ДНК. В одноклеточных эукариотах *Saccharomyces cerevisiae* ORI представлены

[*] andrei_050724@mail.ru
[†] porozov@ifc.cnr.it

консервативными последовательностями (Autonomously replicating sequence, ARS) длиной около 150–400 пар нуклеотидов [2, 3] и 50 пар нуклеотидов у *Kluyveromyces lactis* [4]. В многоклеточных организмах сайты начала репликации в геноме могут располагаться двояко. Первый вариант расположения – сайт начала репликации располагается на протяжении нескольких тысяч нуклеотидов. Во втором случае несколько сайтов репликации находятся в “зонах инициации” протяженностью от 10 до 50 тысяч нуклеотидов [5]. Известно, что в процессе репликации у многоклеточных эукариот очень важное значение играет не столько определенная последовательность нуклеотидов, сколько эпигенетическая регуляция этого процесса. Их клетки могут не только изменять длину S-фазы клеточного цикла во время развития, но и менять порядок, в котором их геном реплицируется в течении клеточной дифференциации, и менять положение активных ORI в зависимости от положения экспрессированных генов. Сайты инициации репликации могут находиться вблизи A-T богатых областей, CpG островов или сайтов связывания факторов транскрипции [6]. Было показано, что структура и реорганизация хроматина, так же как эпигенетическая регуляция, играют очень важную роль в специфичном распознавании комплексом белков сайта начала репликации [7, 8]. Кроме того, имеются данные о зависимости между изменением структуры хроматина во время S-фазы клеточного цикла, репликацией ДНК и наследованием эпигенетической информации у высших эукариот [9].

Методы анализа сигналов в последние годы начинают находить применение и в анализе биологических последовательностей – ДНК и белков [10]. В частности, показано, что спектральные методы могут быть полезны для локализации экзонов в геноме [11, 12], повышения точности предсказания положения генов [13, 14], поиска повторов и корреляций в последовательностях нуклеотидов [15-17], определения типа мембранных белков [18], предсказания их структурных особенностей [19] и идентификации семейств [20].

В настоящей работе рассмотрены различные способы использования вейвлетов и кодирования при поиске ORI в последовательности ДНК человека, приведены результаты применения вейвлет-преобразования для предсказания положения сайтов начала репликации в ДНК человека, и проведен анализ результатов такого поиска. Показано, что частотный анализ нуклеотидной последовательности может быть полезен при обнаружении биологически важных функциональных участков ДНК, которые, однако, не имеют специфичности набора и порядка следования нуклеотидов.

МАТЕРИАЛЫ И МЕТОДЫ

Последовательность нуклеотидов в ДНК или РНК – это последовательность символов из четырехбуквенного алфавита. В то же время спектральные методы анализа сигналов (и/или временных рядов), такие как преобразование Фурье и вейвлет-преобразование, оперируют численными, а не символьными, значениями. Поэтому одной из задач было нахождение кодировки – правил замен символов алфавита ДНК (A, T, G и C) на числа.

Вторая сложность заключалась в том, что на сегодняшний день нет убедительных данных о характере спектров в местах локализации сайтов начала репликации в ДНК человека. В связи с этим требовалось ручная инспекция результатов вейвлет-анализа известных ORI с целью выявления характерных для этих областей спектральных паттернов при одновременном подборе кодировки.

Для нахождения кодировки, которая в дальнейшем могла бы служить индикатором положения сайтов начала репликации и определения характера изменений спектра в областях ORI, были использованы обучающие последовательности из работы Карнани с соавт. [21]. Для увеличения вероятности нахождения сайтов начала репликации в определенных позициях из работы были отобраны данные DNA-microarrays с результатами гибридизации ORC (origin recognition complex), то есть данные прямого

теста на сайты связывания комплексов распознавания ORI. Исходные данные получены при помощи геномного браузера [22, 23] и сборки генома версии NCBI35/hg17 (май 2004).

Исходная последовательность ДНК представлена строкой символов 'A', 'T', 'G' и 'C', соответствующих нуклеотидам ДНК. Обозначим эту входную строку как S:

$$S = \{s_1, s_2, \ldots, s_N\}, \quad s_i \in \{'A','T','C','G'\}, \tag{1}$$

где N – длина последовательности.

Поскольку входная строка состоит из символов, то она непригодна для исследования методами для цифровой обработки сигнала. Обозначим оператором L отображение исходной строки S длиной N в последовательность чисел. Тогда мы получим дискретную функцию F:

$$L : S \to F . \tag{2}$$

Самым простым методом преобразования символьной строки в цифровой сигнал является замена символа строки на некоторое заданное для данного символа число. Мы применили этот подход в работе.

Для получения вейвлет-спектра полученной функции F необходимо было выбрать базовый вейвлет. Наиболее распространенными вейвлетами являются вейвлет Хаара, Добеши, Мейера, Гаусса [24, 25]. Поскольку функция F является дискретной, то для преобразования было целесообразно использовать вейвлет Хаара (рис.1) [26, 27]. Он не является непрерывно дифференцируемой функцией и не подходит для непрерывного вейвлет-преобразования, но при анализе дискретных функций может дать хорошие результаты. Вейвлет Хаара вычисляется достаточно просто, что позволяет повысить скорость преобразования. Хорошие результаты были также получены при использовании в качестве вейвлета скейлинг-функции вейвлета Хаара. Скейлинг-функция вейвлета Хаара задается выражением:

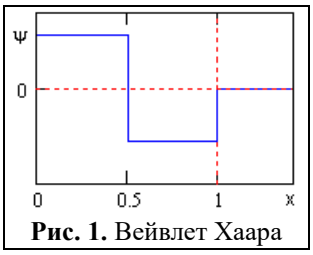

Рис. 1. Вейвлет Хаара

$$\varphi(t) = \begin{cases} 1, & 0 < t < 1, \\ 0, & t < 0, t > 1. \end{cases} \tag{3}$$

Для покрытия всей области значений функции F базовый вейвлет растягивался и перемещался вдоль оси абсцисс. Здесь и далее под абсциссой понимается порядковый номер (индекс) нуклеотида в анализируемой последовательности. Таким образом, вейвлет-спектр являлся функцией двух переменных: сдвига и масштаба. Сдвиг – перемещение базового вейвлета вдоль оси абсцисс. Масштаб – это коэффициент растяжения вейвлета. Например, если вейвлет Хаара задан на интервале [0; 1], то при масштабе m он будет задан на интервале [0; m]. Большие значения масштаба соответствуют глобальному, более общему представлению сигнала, а низкие значения масштаба позволяют различить детали на спектре. В терминах частоты низкие частоты соответствуют глобальной информации о сигнале (которая может быть распределена на всей его протяженности), а высокие частоты – детальной информации и особенностям, которые имеют малую протяженность, т. е. масштаб вейвлета, как единица шкалы частотно-временного представления сигналов, обратен частоте [24]. Таким образом, меняя масштаб преобразования, мы можем находить функционально важные участки разной длины.

Значение вейвлета с масштабом m и сдвигом k в позиции нуклеотида t вычислялось как функция ψ:

$$\psi_{mk}(t) = |m|^{-1/2} \psi[(t-k)/m] \tag{4}$$

где ψ – базовый вейвлет, ψ_{mk} – значение вейвлета масштаба m со сдвигом k.

Множитель перед ψ нужен для сохранения общей площади окна вейвлета. При такой записи на больших масштабах значения не будут уменьшаться.

Для получения всего спектра необходимо было выбрать границы изменения масштаба преобразования p. Поскольку анализируемая функция F дискретна, то минимальным масштабом, соответствующим самой высокой частоте спектра является 1. Минимальный шаг изменения масштаба также принимался равным 1. Поскольку функция F дискретна, то нами использовалось дискретное вейвлет-преобразование.

Обозначим коэффициенты искомого двумерного спектра как C:

$$C_{mk} = \sum_{t=-\infty}^{+\infty} F(t) \cdot \psi_{mk}(t) \tag{5}$$

где F – исходная функция, t – позиция нуклеотида, ψ_{mk} – выбранный вейвлет.

Подставляя в формулу (5) выражение (4), и учитывая, что функция $F(t)$ обращается в ноль за пределами интервала $[1;N]$, получим коэффициенты спектра:

$$C_{mk} = |m|^{-1/2} \sum_{t=1}^{t=N} F(t) \cdot \psi[(t-k)/m]\,. \tag{6}$$

Таким образом, параметры преобразования представляют собой набор значений:

$$P = \{L, \Psi, s1, s2\}, \tag{7}$$

где $s1$ и $s2$ – соответственно нижняя и верхняя граница интервала изменения масштаба вейвлет-преобразования.

Набор параметров (7) полностью характеризует параметры преобразования, обеспечивающие получение спектра участка ДНК.

Для проверки работы метода и поиска характерных особенностей спектров ДНК в областях ORI нами была разработана программа DNAAnalyser на языке C# (рис. 2), которая реализует описанный выше алгоритм преобразования. Ввод нуклеотидных последовательностей может быть произведен из файла формата fasta или непосредственно с сервера геномного браузера UCSC [22, 23]. В случае загрузки с сервера пользователь должен указать хромосому и индекс первого и последнего загружаемого нуклеотида. Кроме того, загруженная таким образом последовательность нуклеотидов доступна для редактирования в специальном окне программы. Это позволяет добавить к загруженному участку случайную или любую реальную последовательность.

Программа позволяет задавать следующие параметры преобразования: базовый вейвлет (выбирается из списка – вейвлетов Хаара, Морлета, МНАТ-вейвлета (Давыдов, 2007) и скейлинг функции вейвлета Хаара), числовые значения для кодирования нуклеотидов, границы масштаба преобразования, вид масштаба (линейный и логарифмический, при логарифмическом масштабе каждое последующее значение в 2 раза больше предыдущего, что позволяет с меньшими затратами провести исследование при большом диапазоне изменения масштаба). Программа позволяет сохранять проект в файл формата XML с расширением .dnapr. В файл сохраняются исходная последовательность нуклеотидов, все параметры преобразования, а также кодированная последовательность и вейвлет-спектр последовательности, если они были рассчитаны. Полученный файл может быть открыт DNAAnalyser для последующего анализа и повторения расчетов. Исследованные последовательности из работы [21] приведены в приложении 1 (таблица 1, см. дополн. материалы к статье). Представление результатов преобразования в удобной для анализа графической форме позволило визуально выявлять особенности спектров.

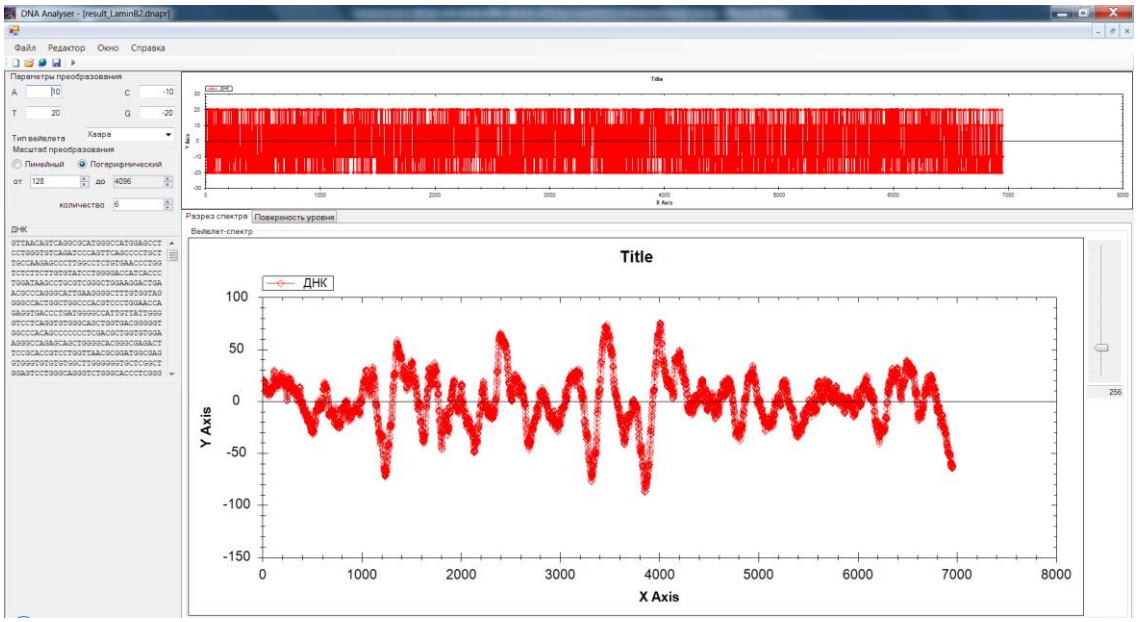

Рис. 2. Общий вид окна программы DNAAnalyser.

Полученный в результате вейвлет-преобразования спектр представляется в окне программы как поверхность уровня (рис. 4 и далее в тексте), цвета которой определяются значениями коэффициентов C_{mk} из формулы (6). Цвета подбираются автоматически исходя из максимального и минимального значения коэффициента C_{mk}, при этом максимальному значению соответствует красный цвет, минимальному – синий. Используемая цветовая шкала является стандартной при отображении высот и глубин на картах и позволяет визуально оценить спектр. По оси абсцисс откладывается сдвиг (переменная k), равный индексу нуклеотида, который располагался в середине вейвлет-функции. По оси ординат откладывается масштаб преобразования (переменная m). Шкала масштабов линейна, даже если при расчетах использовался логарифмический вид масштаба. При этом промежуточные значения получаются путем аппроксимации.

Программа позволяет просмотреть разрез данной поверхности по любому масштабу, использованному при вычислении. График разреза представлен на рис.3.

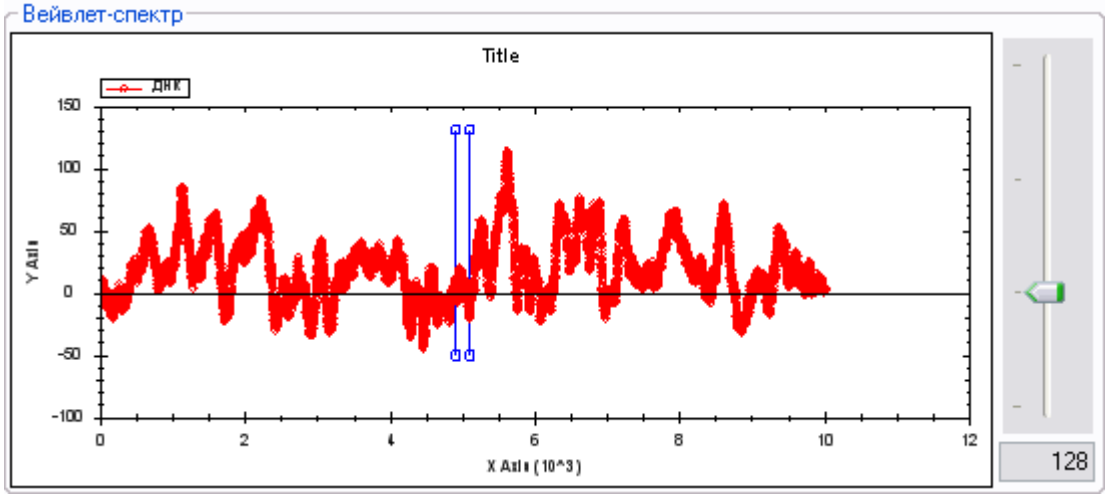

Рис. 3. Представление разреза спектра по масштабу 128

По оси абсцисс, как в предыдущем случае, откладывается позиция нуклеотида в последовательности. По оси ординат откладываются значения коэффициентов C_{Mk}, где

M – значения масштаба, по которому строится разрез. Значение *M* задается ползунком справа от графика.

Сложность автоматизированного поиска сайтов начала репликации в ДНК человека заключается в ограниченном количестве информации для подбора таких параметров преобразования, при которых спектры в местах ORI имели бы характерные отличия от таковых для участков цепи нуклеотидов без сайтов начала репликации. В работе использовались следующие варианты для составления кодировок:

1. Пуриновые основания (аденин и гуанин) имеют равные значения, а пиримидиновые основания (тимин и цитозин) имеют кратные им равные значения, или противоположные по знаку значения.
2. Комплементарные нуклеотиды имеют равные значения, при этом разные пары могут иметь либо кратные значения, либо противоположные по знаку значения.
3. Значения в комплементарной паре кратны друг другу и противоположны по знаку значениям в другой комплементарной паре.

При этом сами значения для нуклеотидов перебирались из диапазона [1…100]. В рамках предлагаемого в работе метода перебор большого количества параметров преобразования сопровождался большими вычислительными затратами. Поэтому в качестве основной задачи мы ставили проверку возможности применения аппарата вейвлет-преобразования к анализу последовательностей нуклеотидов ДНК и выявления областей расположения ORI на основе предложенного метода.

В качестве исходных данных использовалась нуклеотидная последовательность гена Lamin B2 длиной 6953 bp (Genbank ID M94363.1), в позиции 3933 которой находится сайт начала репликации [28], и также 15 участков ДНК человека длиной 10 kbp, содержащих в разных позициях последовательности Orc-HCO1 – Orc-HCO15 (приложение 1, таблица 1, см. доп. материалы к статье) из работы [21].

Для решения поставленной задачи предварительно выбирались параметры преобразования $P = \{L, \Psi, s1, s2\}$. Таким образом, эксперимент включал в себя следующие этапы:

1. Проведение серии вейвлет-преобразований последовательности Lamin B2 и последовательностей Orc-HCO1 – Orc-HCO15 (приложение 1, таблица 1) с различными параметрами *L*, Ш, s1, s2;
2. Определение оптимальных наборов параметров, при которых наибольшее количество исследуемых спектров имеют выраженные особенности в области расположения ORI;
3. Проведение поиска известного участка ORI, помещенного в случайные последовательности с равномерным дискретным распределением нуклеотидов;
4. Проведение поиска известного участка ORI в областях ДНК человека, значительно превышающих искомый сайт по длине;
5. Проведение анализа полученных спектров. Оценка возможности применения предложенного метода для предсказания расположения сайтов начала репликации в ДНК человека. Планирование направлений дальнейших исследований для повышения эффективности работы метода.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

1. Определение параметров преобразования

Для определения параметров вейвлет-преобразования, при которых области ORI приобретали бы характерные особенности, мы использовали разработанную нами программу DNAAnalyser, реализующую описанный выше алгоритм преобразования.

На первом этапе в качестве тестовой использовалась последовательность HUMLAMBBB Human Lamin B2 (Genbank ID M94363.1). После этого каждый

полученный спектр инспектировался на предмет наличия уникальных особенностей в области ORI (нуклеотид 3933).

В процессе анализа спектров последовательности Lamin B2, полученных при различных кодировках L, было найдено правило 7 и кодировка 8:

$$A = -C, \qquad T = -G, \tag{7}$$

$$L(S) = \begin{cases} 10, & если(s_i = 'A'); \\ 20, & если(s_i = 'T'); \\ -10, & если(s_i = 'C'); \\ -20, & если(s_i = 'G'); \end{cases} \tag{8}$$

В качестве базовых вейвлетов использовались вейвлет Хаара и скейлинг-функция вейвлета Хаара. Этим достигалось достаточно наглядное выявление особенностей спектра и экономия вычислительных ресурсов – вейвлет Хаара и его скейлинг-функция являются простыми ступенчатыми функциями.

Одним из важнейших параметров при поиске отличительных особенностей спектра в местах расположения ORI является масштаб преобразования. При большем масштабе получается более гладкая функция, скрывающая спектральные отличия коротких участков. При меньшем масштабе получается резко изменяющаяся функция, и это, в свою очередь, затрудняет анализ. Поэтому было выбрано кратномасштабное преобразование с диапазоном изменения масштаба от 2^7 до 2^{13}. Кратномасштабное преобразование предполагает, что значения масштаба равны степени двойки, т. е. каждый последующий больше предыдущего в 2 раза. Применение кратномасштабного преобразования позволило проверить большой диапазон масштабов при сравнительно небольших вычислительных затратах. Выбор границ обуславливался тем, что спектр функции в окне шириной менее 128 очень часто не покрывал область сайта начала репликации в ДНК, которая по разным данным может колебаться от 10 до 150–200 нуклеотидов, а окно более 16384 уже больше исходной последовательности.

На рис. 4. представлена поверхность, полученная в результате вейвлет-преобразования последовательности Lamin B2 с выбранными параметрами. Области больших значений коэффициентов представлены красным цветом, области малых значений – синим цветом. На рисунке 4 видно, что в области, близкой к позиции 3933 происходит резкое увеличение значений коэффициентов – они переходят через 0 и достигают максимальных значений в приведенном спектре. Особенно это заметно на масштабах 2048 и 4096.

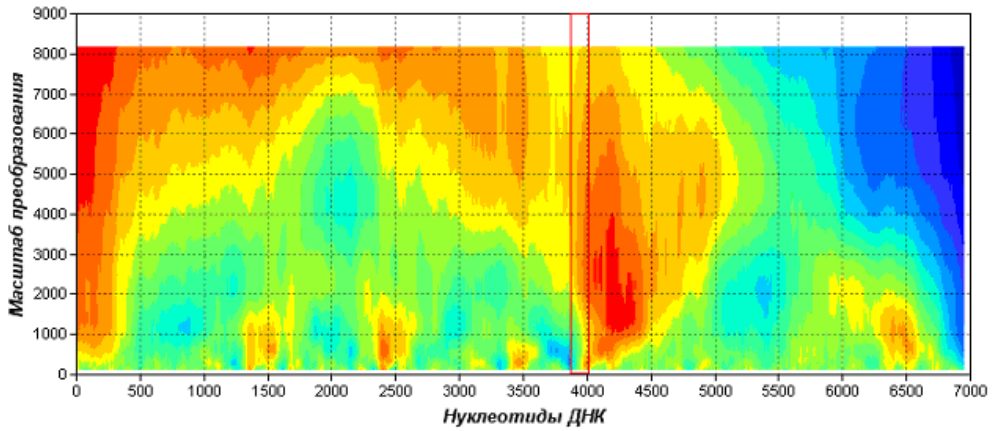

Рис. 4. Результаты кратномасштабного вейвлет-анализа последовательности |M94363.1| HUMLAMBBB Human lamin B2 (LAMB2) длиной 6952 bp [28]. Интервал масштаба от 128 до 8192. Вблизи позиции сайта начала репликации (3933) и на масштабах преобразования 2048 и 4096 значения коэффициентов спектра резко возрастают.

2. Исследование случайных последовательностей

Для проверки гипотезы о наличии уникальных спектральных особенностей в областях ORI, участки с сайтами начала репликации размером 100–200 нуклеотидов были помещены в сгенерированную случайную последовательность с равномерным законом распределения нуклеотидов. Длина образованной таким образом тестовой последовательности вместе с участком ORI составляла 10 kb.

Исследование показало, что на больших масштабах и при выбранных параметрах преобразования спектральная картина и значения коэффициентов спектра не позволяли выявить область сайта начала репликации (рис. 5). Это объяснялось, в частности, случайным характером распределения нуклеотидов и масштабом преобразования, значительно превышающим по длине область ORI.

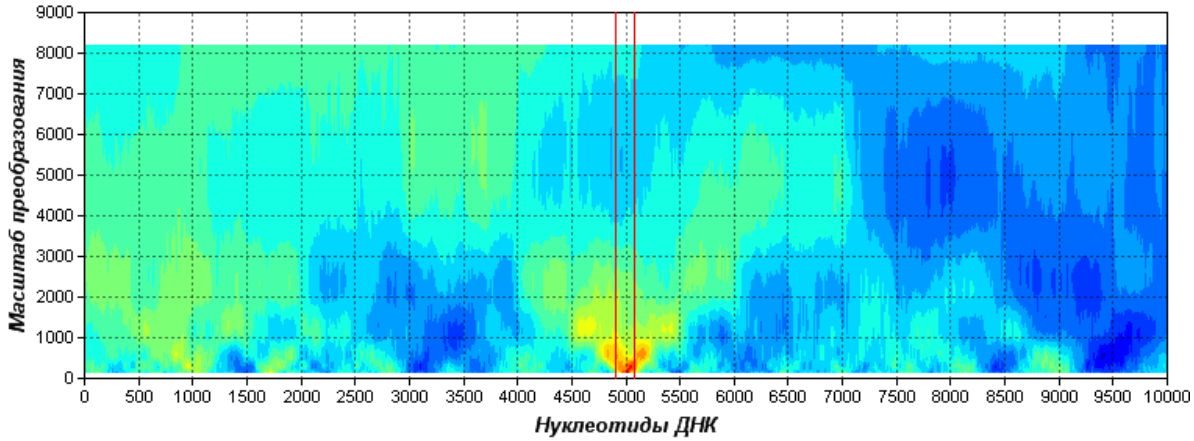

Рис. 5. Результаты кратномасштабного вейвлет-анализа случайной последовательности с помещенным в нее сайтом начала репликации Orc-HCO5 [21]. Длина последовательности 10 kb Сайт инициации репликации – 4911–5090. Приведены коэффициенты скейлинг-функции Хаара при различных масштабах преобразования. Интервал изменения масштаба от 128 до 8192. На малых масштабах преобразования виден всплеск коэффициентов в области ORI.

В то же время, на малых масштабах преобразования на рис. 6 отчетливо виден пик спектра в области помещенного в случайную последовательность ORI Orc-HCO5 (границы области отмечены двумя линиями). Таким образом, в дальнейших исследованиях было принято целесообразным применение масштаба вейвлет-спектра от 64 до 512.

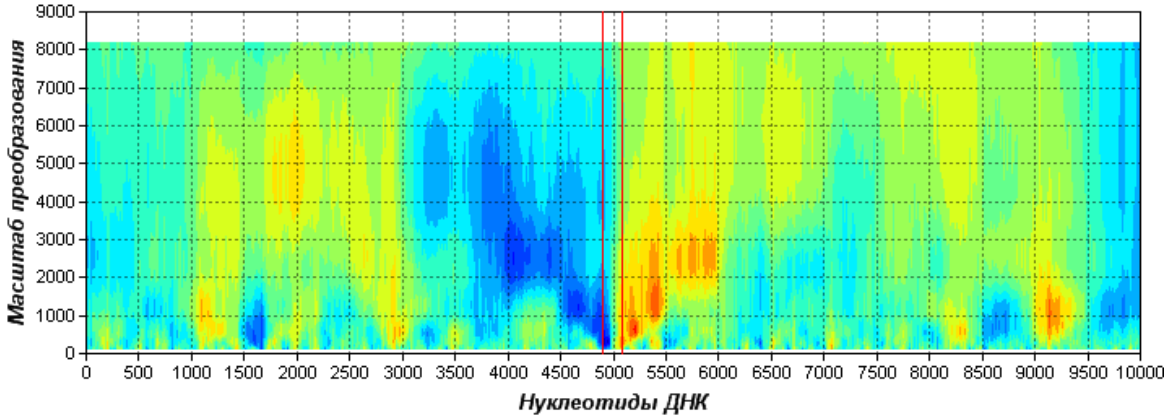

Рис. 6. Результаты кратномасштабного вейвлет-анализа последовательности Orc-HCO5 [21] длиной 10 kb вейвлетом Хаара. Интервал изменения масштаба от 128 до 8192. Вертикальными линиями ограничена область ORI.

Результаты преобразования для случайной последовательности с равномерным распределением с помещенным в нее Orc-HCO9 и Orc-HCO7 [21] представлены соответственно на рисунках 7 и 8.

Наличие резких пиков спектра было обнаружено для последовательностей Orc-HCO4, Orc-HCO5, Orc-HCO7, Orc-HCO8, Orc-HCO9, Orc-HCO10, Orc-HCO11, Orc-HCO14, Orc-HCO15 из статьи [21], см. приложение 2 в доп. материалах к статье.

Следует заметить, что характеристики спектров, полученных при использовании как вейвлета Хаара, так и скейлинг-функции Хаара, а именно их особенности в областях ORI, были одинаково хорошо применимы для определения положения последовательности сайта начала репликации. На рис. 6 представлен спектр той же последовательности, что и на рис.5. В качестве вейвлета использовался вейвлет Хаара. Отличие спектров, представленных на рис. 5 и рис. 6 состоит в том, что на спектре на рис. 5 виден пик, а на рис. 6 – нулевое значение при переходе от минимума до максимума спектра.

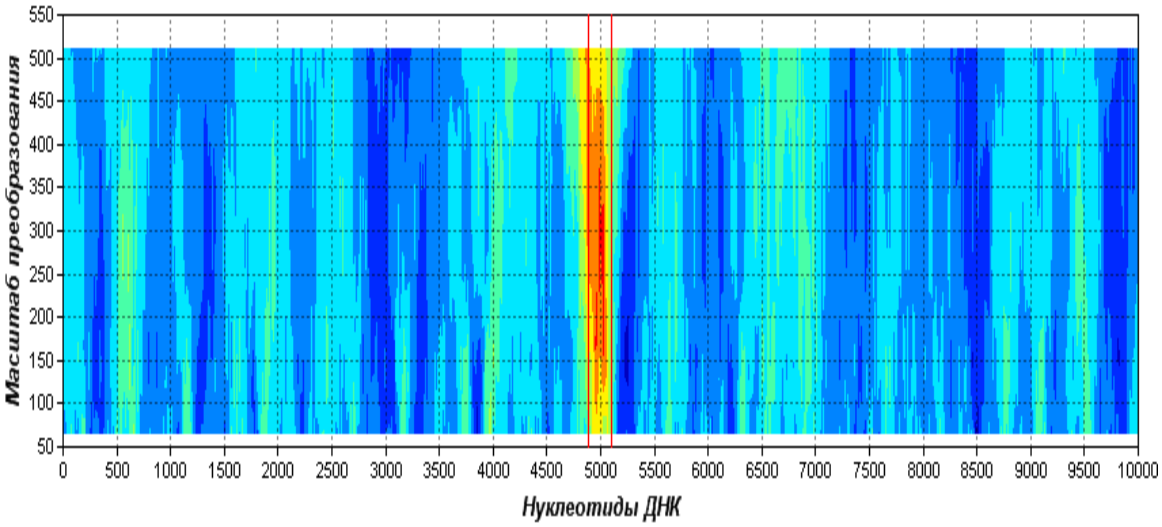

Рис. 7. Результаты кратномасштабного вейвлет-анализа последовательности Orc-HCO9 длиной 10 kb скейлинг-функцией Хаара. Интервал изменения масштаба от 64 до 512. Границы Orc-HCO9 отмечены двумя линиями.

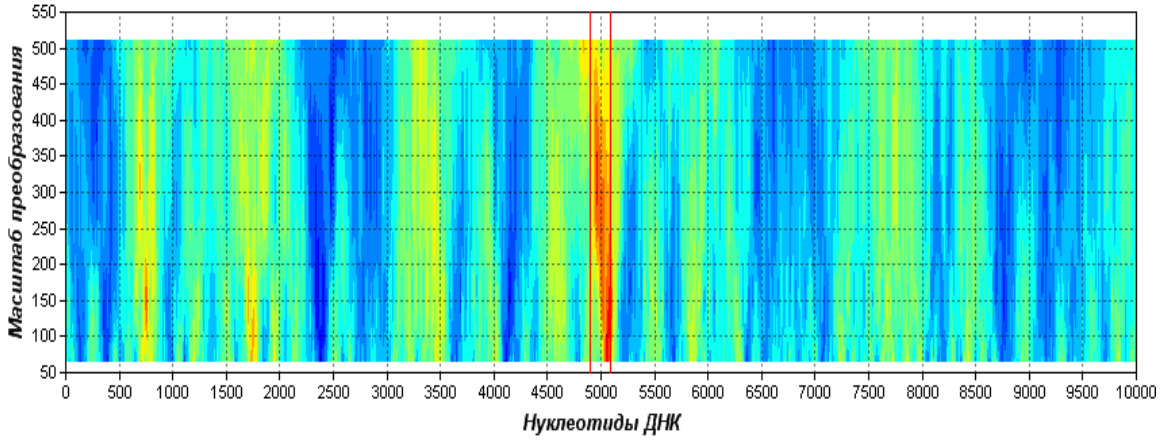

Рис. 8. Результаты кратномасштабного вейвлет-анализа последовательности Orc-HCO7 длиной 10 kb скейлинг-функцией Хаара. Интервал изменения масштаба от 64 до 512. Границы Orc-HCO7 отмечены двумя линиями.

3. Исследование реальных последовательностей

Сложность спектрального анализа последовательностей нуклеотидов состоит в наличии функциональных взаимосвязей между участками исследуемой последовательности, повторов различной длины, периодических отклонений от нормального распределения по длине последовательности, повторяющихся регуляторных участков. Спектры реальных последовательностей гораздо сложнее, чем спектр случайной нуклеотидной цепи с введенным в нее участком ORI. Поэтому выявление на спектре реальной ДНК нужных пиков и перепадов – задача гораздо более сложная, чем поиск функционально важной последовательности, взятой из эволюционно сформировавшейся ДНК и помещенной внутрь сгенерированной последовательности нуклеотидов с равномерным дискретным распределением.

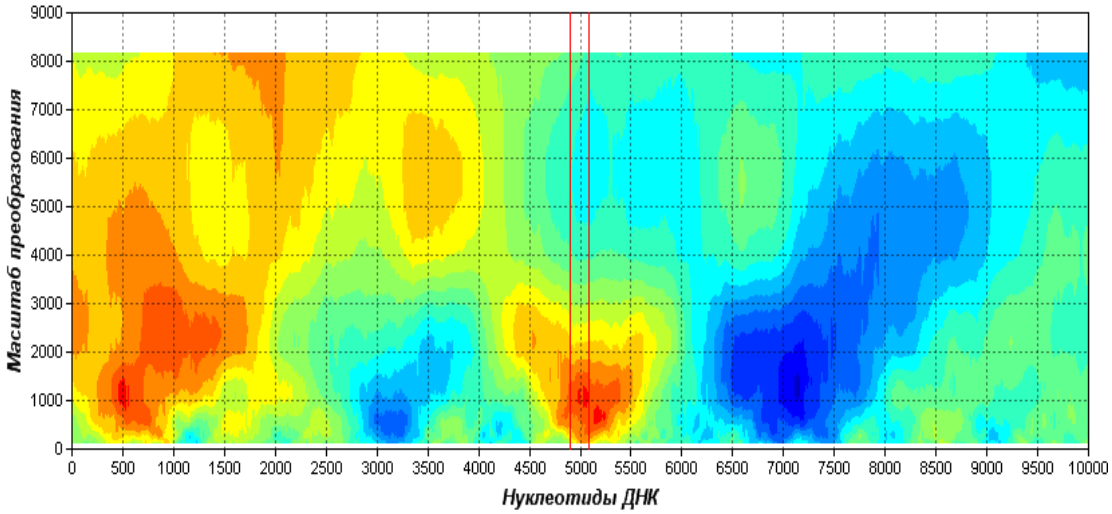

Рис. 9. Результаты кратномасштабного вейвлет-анализа последовательности Orc-HCO8 длиной 10 kb скейлинг-функцией Хаара. Масштаб от 128 до 8192.

Результаты исследований последовательностей из работы [21], содержащих ORI (приложение 1, табл. 1), приведены ниже. Было обнаружено, что применение найденных параметров преобразования позволяет локализовать в последовательностях ДНК человека следующие области: Orc-HCO5, Orc-HCO8, Orc-HCO13, Orc-HCO14, Orc-HCO15 [21]. На спектрах данных последовательностей в области расположения сайтов начала репликации присутствуют пики (рис. 9, рис. 10, приложение 3 в доп. материалах к статье).

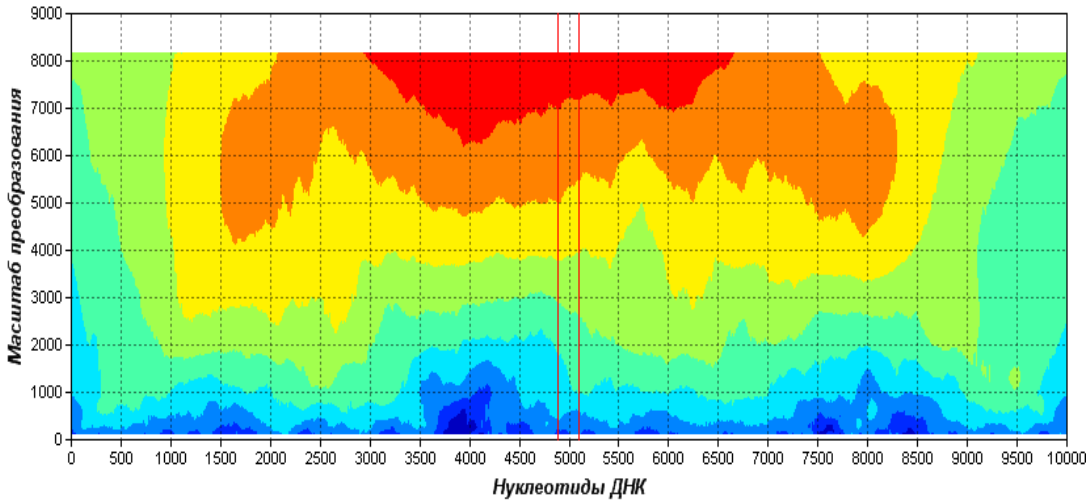

Рис. 10. Результаты кратномасштабного вейвлет-анализа последовательности Orc-HCO14 длиной 10 kb скейлинг-функцией Хаара. Масштаб от 128 до 8192.

Очевидно, что спектр последовательности Orc-HCO8 (рис. 9) похож на спектр последовательности Lamin B2 (рис. 4).

В 5 из 15 проанализированных реальных последовательностей нуклеотидов при использовании выбранных параметров преобразования нам удалось выявить отличительные особенности спектров и, следовательно, локализовать сайты начала репликации. Это последовательности Orc-HCO5, Orc-HCO8, Orc-HCO13, Orc-HCO14, Orc-HCO15 (приложение 3, рис. 26-30).

Очевидной причиной значительного различия эффективности работы метода на случайных и реальных последовательностях является сложный характер скрытых периодичностей, имеющихся в эволюционно сформировавшихся последовательностях, которые часто не позволяют обнаружить спектральные особенности, характерные для ORI-участка.

Следует отметить, что параметры преобразования последовательности в спектр подбирались эмпирически. Кроме того, в работе не применялись специальные алгоритмы распознавания образов для локализации ORI. Использование этих техник является одним из путей повышения эффективности работы метода на реальных последовательностях.

ЗАКЛЮЧЕНИЕ

Поиск и предсказание участков ДНК, не обладающих выраженной sequence-специфичностью, но в то же время являющихся исключительно важными для функционирования клетки, передачи генетической информации и для самых различных регуляторных процессов, протекающих на уровне ДНК, является одной из ключевых задач современной биологии и биоинформатики. Разработка и применение методов, позволяющих с большой точностью находить такие участки, несмотря на существенные различия в нуклеотидной последовательности, имеет очень большое значение для молекулярной биологии и генной медицины.

Одними из таких областей в геноме являются сайты начала репликации. Вопросам синхронизации репликации с фазами клеточного цикла, изучению ее роли в развитии, связи с HOX-белками [29], регуляции процесса репликации и ее нарушениям, в частности, при канцерогенезе [30, 31] уделяется огромное внимание.

Методы Фурье и вейвлет-преобразования, давно и с успехом применяемые в цифровой обработке сигнала, все еще недостаточно широко используются при анализе нуклеотидных и аминокислотных последовательностей, что следует из анализа литературных источников [32].

Нами разработан подход и соответствующее ПО, позволяющее проводить подбор параметров вейвлет-преобразования и исследования последовательностей ДНК преобразованием с различными параметрами и вейвлет-функциями. В работе проведен поиск параметров и анализ результатов вейвлет-преобразования, примененного к участкам ДНК человека, содержащим сайты начала репликации. Эмпирический анализ вейвлет-спектров тестового набора последовательностей позволил подобрать параметры преобразования, при использовании которых зоны ORI обладают уникальными спектральными характеристиками. Это, в свою очередь, сделало возможным локализацию указанных небольших областей на участке ДНК. Применение этого метода для предсказания ORI, помещенного в случайную последовательность нуклеотидов, позволил находить сайты начала репликации в 60% случаев. При этом характерно, что все спектры либо имели узкий ярко выраженный пик в области расположения ORI, либо не имели вообще ярко выраженного пика.

Эффективность работы метода на реальных последовательностях составила 30%. При этом, в отличие от анализа случайных последовательностей, в большинстве случаев пики расплывались, не давая определить положение ORI. Это может быть

следствием сложного характера скрытых периодичностей в последовательностях, состоящих из областей ORI, различных регуляторных последовательностей, повторов, областей экзонов и интронов, сайтов специфического связывания.

Следует отметить, что специальные методики анализа спектров, распознавания образов и методы машинного обучения в работе при анализе спектров и поверхностей сигналов не применялись.

Для повышения чувствительности метода могут оказаться полезными исследования и эксперименты с реальными последовательностями ДНК, направленные на поиск и проверку других параметров преобразования и с использованием большего набора исходных данных. Результаты вейвлет-преобразования во многом зависят от базового вейвлета. Поэтому особое внимание при проведении дополнительных исследований необходимо уделить подбору вейвлета, который обеспечит лучшее качество предсказания определенных участков в последовательности нуклеотидов. Поскольку вейвлетом может являться любая функция, удовлетворяющая определенным условиям [24, 25], то помимо использования большого количества широко известных вейвлетов возможно создание своего особенного для решения конкретной задачи. Можно предположить, что, несмотря на функциональную похожесть последовательностей, где инициируется репликация ДНК, их спектральные характерисики могут сильно варьировать. Поэтому вполне вероятна ситуация, при которой наилучшую чувствительность покажет метод с несколькими проходами-преобразованиями с разными базовыми вейвлетами и, соответственно, их параметрами. Поиск и применение различных схем кодирования последовательности (например, использовать вместо простой замены кодировку, предлагаемую в статье Е.Б. Броди [33]) может улучшить прогностический потенциал метода. Для различных функциональных участков ДНК возможно нахождение различных уникальных параметров преобразований. Комбинация вейвлет-преобразования с методами машинного обучения и искусственного интеллекта и, возможно, распознавания образов на спектрах может увеличить предиктивный потенциал спектральных методов для нахождения функциональных участков ДНК и, в частности, сайтов начала репликации в ДНК человека.

СПИСОК ЛИТЕРАТУРЫ

1. Bogan J.A., Natale D.A., and Depamphilis M.L. Initiation of eukaryotic DNA replication: conservative or liberal? *J. Cell. Physiol.* 2000. V. 184. № 2. P. 139–150.
2. Raghuraman M.K., Winzeler E.A., Collingwood D., Hunt S., Wodicka L, Conway A., Lockhart D.J., Davis R.W., Brewer B.J., Fangman W.L. Replication Dynamics of the Yeast Genome. *Science.* 2001. V. 294. № 5540. P. 115–121.
3. Bell S.P., Dutta A. DNA replication in eukariotic cells. *Annual Review of Biochemistry.* 2002. V. 71. № 1. P. 333–374.
4. Liachko I., Bhaskar A., Lee C., Chung S.C., Tye B.K., Keich U. A Comprehensive Genome-Wide Map of Autonomously Replicating Sequences in a Naive Genome. *PLoS Genet.* 2010. V. 6. № 5. P. e1000946.
5. Gilbert D.M. Making sense of eukaryotic DNA replication origins. *Science.* 2001. V. 294. № 5540. P. 96–100.
6. Bogan J.A., Natale D.A., and Depamphilis M.L. Initiation of eukaryotic DNA replication: conservative or liberal? *Journal of Cellular Physiology.* 2000. V. 184. № 2. P. 139–150.
7. Demeret C., Vassetzky Y., and Mechali M. Chromatin remodelling and DNA replication: from nucleosomes to loop domains. *Oncogene.* 2001. V. 20. № 24. P. 3086–3093.
8. Mechali M. DNA replication origins: from sequence specificity to epigenetics. *Nature reviews. Genetics.* 2001. V. 2. № 8. P. 640–645.

9. McNairn A.J. and Gilbert D.M. Epigenomic replication: linking epigenetics to DNA replication. *BioEssays : news and reviews in molecular, cellular and developmental biology*. 2003. V. 25. № 7. P. 647–656.
10. Bajic V.B., Bajic I.V., Hide W. A new method of spectral analysis of DNA/RNA and protein sequences. Plenary lecture. In: *Proc. First International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'98)*. Novosibirsk, Russia. 1998. V. 1. P. 120–123.
11. Oueslati A.E., Lachiri Z., Ellouze N. Spectral analysis of DNA sequence: The exon's location method. In: *Proceedings of the 2007 15th International Conference on Digital Signal Processing*. Eds. S. Sanei et al. New York: IEEE, 2007. P. 115–118.
12. Haimovich A.D., Byrne B., Ramaswamy R., Welsh W.J. Wavelet analysis of DNA walks. *Journal of Computational Biology*. 2006. V. 13. № 7. P. 1289–1298.
13. Chang C.Q., Fung P.C.W., Hung Y.S. Improved gene prediction by resampling-based spectral analysis of DNA sequence. In: *2008 International Special Topic Conference on Information Technology and Applications in Biomedicine*. 2008. P. 221–224.
14. Mena-Chalco J.P., Carrer H., Zana Y., Cesar R.M. Identification of protein coding regions using the modified Gabor-wavelet transform. *IEEE-ACM Transactions on Computational Biology and Bioinformatics*. 2008. V. 5. № 2. P. 198–207.
15. Bucur A., van Leeuwen J., Dimitrova N., Mittal C. Frequency Sorting Method for Spectral Analysis of DNA Sequences. In: *2008 IEEE International Conference on Bioinformatics and Biomedicine*. Los Alamitos: IEEE Computer Soc. 2008. V. 1. P. 43–50.
16. Berger J.A., Mitra S.K., Astola J. Power spectrum analysis for DNA sequences. In: *Seventh International Symposium on Signal Processing and Its Applications*. New York: IEEE. 2003. V. 2. P. 29–32.
17. Лобзин В.В., Чечеткин В.Р. Порядок и корреляции в геномных последовательностях ДНК. Спектральный подход. *Успехи физических наук.* 2000. Т. 170. № 1. С. 57–81.
18. Qiu J.D., Sun X.Y., Huang J.H., Liang R.P. Prediction of the Types of Membrane Proteins Based on Discrete Wavelet Transform and Support Vector Machines. *Protein Journal*. 2010. V. 29. № 2. P. 114–119.
19. Zhang S.L. and Wang T.M. Feature analysis of protein structure by using discrete Fourier transform and continuous wavelet transform. *Journal of Mathematical Chemistry*. 2009. V. 46. № 2. P. 562–568.
20. Турутина В.П., Ласкин А.А., Кудряшов Н.А., Скрябин К.Г., Коротков Е.В. Идентификация скрытой периодичности в аминокислотных последовательностях белковых семейств. *Биохимия*. 2006. Т. 71. № 1. С. 26–41.
21. Karnani N., Taylor C.M., Malhotra A., Dutta A. Genomic Study of Replication Initiation in Human Chromosomes Reveals the Influence of Transcription Regulation and Chromatin Structure on Origin Selection. *Mol. Biol. of Cell*. 2010. V. 21. № 3. P. 393–404.
22. Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. The Human Genome Browser at UCSC. *Genome Research*. 2002. V. 12. № 6. P. 996–1006.
23. Rosenbloom K.R., Dreszer T.R., Pheasant M., Barber G.P., Meyer L.R., Pohl A., Raney B.J., Wang T., Hinrichs A.S., Zweig A.S., Fujita P.A., Learned K., Rhead B., Smith K.E., Kuhn R.M., Karolchik D., Haussler D., Kent W.J. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Research*. 2010. V. 38. № suppl 1. P. D620–D625.
24. Давыдов А.В. *Вейвлеты в вейвлетный анализ сигналов. Курс лекций*. 2007 URL: http://prodav.narod.ru/wavelet/ (дата обращения: 21.05.2011).
25. Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения. *Успехи физических наук*. 1996. Т. 166. № 11. С. 1145–1170.

26. Haar A. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen.* 1910. V. 69. № 3. P. 331–371.
27. Chui C.K. *An introduction to wavelets*. San-Diego, Ca, USA: Academic Press Professional, Inc. 1992. 266 p.
28. Abdurashidova G., Deganuto M., Klima R., Riva S., Biamonti G., Giacca M., Falaschi A. Start sites of bidirectional DNA synthesis at the human lamin B2 origin. *Science.* 2000. V. 287. № 5460. P. 2023–2026.
29. Marchetti L., Comelli L., D'Innocenzo B., Puzzi L., Luin S., Arosio D., Calvello M., Mendoza-Maldonado R., Peverali F., Trovato F., Riva S., Biamonti G., Abdurashidova G., Beltram F., Falaschi A. Homeotic proteins participate in the function of human-DNA replication origins. *Nucleic Acids Research.* 2010. V. 38. № 22. P. 8105–19.
30. Falaschi A., Abdurashidova G., and Biamonti G. DNA replication, development and cancer: a homeotic connection? *Critical Reviews in Biochemistry and Molecular Biology.* 2010. V. 45. № 1. P. 14–22.
31. Del Bene F., Wittbrodt J. Cell cycle control by homeobox genes in development and disease. *Seminars in Cell & Developmental Biology.* 2005. V. 16. № 3. P. 449–460.
32. Liò P. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics.* 2003. V. 19. № 1. P. 2–9.
33. Brodie Of Brodie E.B., Nicolay S., Touchon M., Audit B., d'Aubenton-Carafa Y., Thermes C., Arneodo A. From DNA sequence analysis to modeling replication in the human genome. *Physical Review Letters.* 2005. V. 94. № 24. P. 248103.

Материал поступил в редакцию 14.06.2011, опубликован 21.07.2011.