Математическая биология и биоинформатика
2015. Т. 10. № 2. С. 387–397. doi: 10.17537/2015.10.387

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

УДК: 577.2:519.23

Структурно-статистические свойства кодирующих районов ДНК

Кутыркин В.А.[*1], Чалей М.Б.[2]**

[1]*Московский государственный технический университет им. Н.Э. Баумана, Москва, Россия*
[2]*Институт математических проблем биологии, Российская академия наук, Пущино, Московская область, Россия*

Аннотация. С помощью спектрально-статистического подхода (2С-подхода) в работе исследуются структурно-статистические свойства кодирующих последовательностей ДНК (CDS) генома человека, к которым относятся свойство 3-регулярности и скрытая профильная периодичность. Особая значимость и объективное существование этих свойств подтверждаются исследованием бинарно перекодированных CDS. Только один вид вырожденной бинарной перекодировки, соответствующий отождествлению комплементарных нуклеотидов, способствует сохранению свойств исходных CDS. Использование невырожденной бинарной перекодировки подтверждает, что скрытая триплетная периодичность в CDS генома человека относится к ранее неизвестному типу профильной периодичности.

Ключевые слова: *скрытая профильная периодичность, спектрально-статистический подход, свойство 3-регулярности CDS, триплетная периодичность.*

1. ВВЕДЕНИЕ

Идея о существовании триплетной периодичности в кодирующих последовательностях ДНК (CDS), генах и экзонах возникла ещё в начале 80-х годов 20-го столетия с началом изучения корреляций в нуклеотидных последовательностях [1–3]. Она получила широкое распространение в связи с задачами поиска районов ДНК, кодирующих белки, и различения кодирующих и некодирующих последовательностей [4–8]. На сегодняшний день использование факта существования скрытой триплетной периодичности для выявления в ДНК геномов последовательностей с функцией кодирования по-прежнему остаётся привлекательным для широкого круга исследователей и создателей компьютерных программ [9–11].

Достаточно долго оставался невыясненным ответ на вопрос о типе скрытой триплетной периодичности в кодирующих последовательностях ДНК. Ранее к достоверным методам распознавания скрытой периодичности относились только методы, распознающие размытые тандемные повторы. Однако, размытые тандемные повторы занимают в геномах организмов достаточно малую часть (~ 10%) [12]. Поэтому вывод о наличии скрытой триплетной периодичности, как правило, обосновывался различными косвенными (оценочными) методами, самым распространенным представителем которых является Фурье-анализ [5, 9]. Однако,

[*]vkutyrkin@yandex.ru
[**]maramaria@yandex.ru

такие косвенные методы не являются методами распознавания скрытой периодичности и их результаты требуют дальнейшего подтверждения.

В работе [13] для последовательностей ДНК был предложен новый тип периодичности, названный профильной периодичностью. В дальнейшем, в работах [14–16] для достоверного распознавания скрытой профильной периодичности был разработан спектрально-статистический подход (2С-подход). Первичный анализ последовательностей генома человека [15, 17] показал, что не менее, чем в 60% CDS генома человека распознаётся скрытая триплетная периодичность. Кроме того, практически во всех CDS (с учетом статистической погрешности) наблюдается свойство 3-регулярности последовательности [17], обусловленное размером кодонов триплетного генетического кода. Также было показано, что практически все интроны генома человека не обладают свойством 3-регулярности. Фактически, наличие 3-регулярности в последовательности ДНК является характеристическим признаком кодирующей последовательности в геноме человека.

В настоящей работе, опираясь на статистики и формулы работы [18], был выполнен более точный количественный анализ свойств CDS генома человека. Показано, что свойства 3-регулярности, скрытой триплетной профильной периодичности и скрытой профильной периодичности с размером периода, кратным трём, являются характерными структурно-статистическими свойствами CDS генома человека. К таким характерным свойствам относится и, проявляемая в CDS генома человека, двухуровневая организация кодирования [14, 15]. Первый уровень определяется свойством 3-регулярности последовательности CDS. В такой последовательности второй уровень организации кодирования проявляется как скрытая профильная периодичность с размером периода, кратным, но не равным трём.

Кроме того, в настоящей работе выполняется анализ сохранности указанных выше структурно-статистических свойств в бинарно перекодированных рассматриваемых CDS при различных вариантах перекодирования. Переход от четырёхбуквенного к бинарному алфавиту позволяет в два раза расширить диапазон тестируемых периодов. Это приводит к повышению надёжности используемых статистических методов. Сохранность спектрально-статистических свойств в бинарно перекодированных CDS будет дополнительным подтверждением их объективного существования в исходных последовательностях.

Сначала рассматривается бинарное перекодирование CDS с потерей информации при обозначении двух различных букв из алфавита ДНК символами 0 и 1. Возможны три варианта такого перекодирования: 1) A, T = 1, G, C = 0; 2) A, G = 1, T, C = 0; 3) A, C = 1, T, G = 0. Как будет показано далее, при первом варианте перекодирования наблюдается наибольшая сохранность структурно-статистических свойств CDS генома человека. Следует отметить, что первый вариант соответствует рассмотрению кодирующего участка как фрагмента двунитевой ДНК, представленного последовательностью пар комплементарных нуклеотидов. На основе этого факта можно предполагать, существенную роль, которую играет распределение двух типов пар комплементарных нуклеотидов (A/T и G/C) в процессе распознавания кодирующих участков в двунитевой ДНК хромосомы.

Далее рассматривается бинарное перекодирование CDS без потери информации, когда буквы алфавита ДНК обозначаются упорядоченными бинарными парами в виде: A = 00, T = 01, G = 10 и C = 11. Показано, что практически во всех, перекодированных таким образом, CDS генома человека выявляются свойство 6-регулярности и профильная периодичность с размером периода, кратным шести. Тем самым, согласно распространённому в литературе мнению о скрытой триплетной периодичности в кодирующих районах генома, в CDS генома человека распознаётся тип скрытой периодичности, названный в работах [14–18] скрытой профильной периодичностью.

Фактически, сохранность распределения структурно-статистических свойств исходных CDS генома человека при различных бинарных перекодировках подтверждает существование характерной профильной периодичности и регулярности в кодирующих районах ДНК.

2. КРАТКОЕ ОПИСАНИЕ СПЕКТРАЛЬНО-СТАТИСТИЧЕСКОГО ПОДХОДА ДЛЯ РАСПОЗНАВАНИЯ РЕГУЛЯРНОСТИ И СКРЫТОЙ ПРОФИЛЬНОЙ ПЕРИОДИЧНОСТИ В ТЕКСТОВЫХ СТРОКАХ

Полное описание спектрально-статистического подхода (2С-подхода) ранее было дано в работах [15, 16, 18]. Этот подход основывается на анализе спектрально-статистических характеристик текстовых строк. Наглядно эти характеристики представлены графиками зависимости соответствующих статистических параметров от тестируемых периодов (тест-периодов) рассматриваемой текстовой строки. Целевое назначение каждого спектра отражено в его названии. Для распознавания скрытой профильной периодичности в текстовой строке последовательно анализируются спектры строки: спектр $\mathbf{D}_1$ отклонения от однородности (1-профильности), характеристический спектр $\mathbf{C}$, спектр $\mathbf{D}_\lambda$ отклонения от тестируемой профильной периодичности (λ-профильности, $\lambda > 1$). При анализе спектров отклонения применяется статистический критерий согласия, использующий соответствующую статистику Пирсона на уровне значимости 5%. В графиках спектров отклонения фиксируется пороговое значение, равное единице. Если пороговое значение превышается более чем для 5% тест-периодов, то принимается гипотеза о значимом отклонении от тестируемой профильности в анализируемой текстовой строке. Если для текстовой строки отвергнута гипотеза об её однородности (скрытой 1-профильности), то далее анализируется характеристический спектр $\mathbf{C}$ этой строки. Первый тест-период строки, на котором проявляется ярко выраженное значение характеристического спектра, предлагается в качестве оценки размера периода L её скрытой профильной периодичности. Если для тест-периода L пороговое значение в спектре отклонения $\mathbf{D}_L$ от L-профильности не превышается для более чем 95% тест-периодов, то принимается гипотеза о том, что в этой текстовой строке наблюдается скрытая L-профильность. Отметим, что, если размер алфавита анализируемой текстовой строки длиной n равен K, то диапазон её тест-периодов имеет вид $(1, 2, ..., \lambda_{\max})$, где $\lambda_{\max} \le n/5K$. На рис. 1 приведён пример распознавания скрытой 24-профильности в CDS ($n = 3006$, $K = 4$) из генома человека. Из рис. 1,а, на котором представлен график спектра отклонения от однородности, следует, что рассматриваемая последовательность является неоднородной. Согласно рис. 1,б, на котором показан график характеристического спектра, оценкой размера скрытого профильного периода является тест-период $L = 24$ нукл., поскольку на этом тест-периоде проявляется первое, ярко выраженное максимальное значение характеристического спектра рассматриваемой последовательности. Корректность этой оценки подтверждает график спектра отклонения от 24-профильности, показанный на рис. 1,в. Свойство 3-регулярности этой последовательности проявляется в том, что пики (максимумы) её характеристического спектра на рис. 1,б наблюдаются на тест-периодах, кратных трём. Однако, из графика отклонения от 3-профильности на рис. 1,г следует, что в анализируемой CDS отсутствует 3-профильность. Таким образом, в рассматриваемой CDS наблюдается двухуровневая организация кодирования. На первом уровне проявляется 3-регулярность и на втором уровне – 24-профильность этой последовательности.

В общем случае, для фиксации R-регулярности в работах [17, 18] с помощью индекса I_R введён критерий R-регулярности для анализируемой текстовой строки. Если для рассматриваемой строки $\mathrm{I}_R \ge 0.7$, то эта строка признаётся R-регулярной.

Исследование CDS и интронов генома человека показало, что практически все (с учетом статистической погрешности) CDS последовательности обладают свойством 3-регулярности, в отличие от интронов, которые этим свойством не обладают [17]. В качестве примера на рис. 1,б приведён график характеристического спектра CDS генома человека, согласно которому эта последовательность обладает свойством 3-регулярности, поскольку индекс регулярности $I_3 = 0.99$.

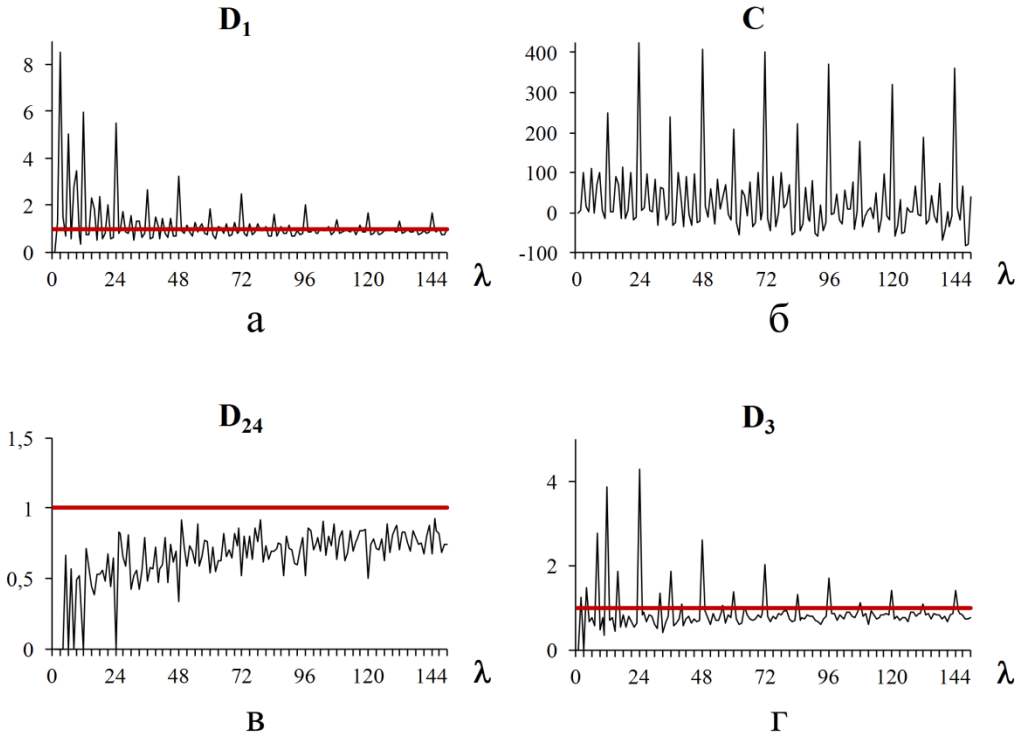

Рис. 1. Распознавание с помощью спектров 2С-подхода скрытой 24-профильности в CDS (KEGG, hsa:389677, 3006 нукл.) белка, содержащего РНК-связывающий мотив.

3. АНАЛИЗ СОХРАННОСТИ СТРУКТУРНО-СТАТИСТИЧЕСКИХ СВОЙСТВ ПРИ РАЗЛИЧНЫХ БИНАРНЫХ ПЕРЕКОДИРОВКАХ CDS ГЕНОМА ЧЕЛОВЕКА С ПОТЕРЕЙ ИНФОРМАЦИИ

Для выборки 17652 CDS генома человека из базы данных KEGG [19], функциональная активность которых получила подтверждение, на рис. 2,а приведена дендрограмма, отражающая распределение структурно-статистических свойств этих последовательностей. На рис. 2,б–г приведены аналогичные дендрограммы для бинарно перекодированных CDS из этой выборки. Дендрограмма на рис. 2,б соответствует бинарному перекодированию, при котором производится замена букв A и T символом 1, букв G и C символом 0. Дендрограмма на рис. 2,в соответствует заменам букв A и G символом 1 и букв T и C символом 0, дендрограмма на рис. 2,г – заменам букв A и C символом 1, букв G и C символом 0. Поскольку при рассматриваемых бинарных перекодировках размер алфавита сокращается в два раза, диапазон тест-периодов каждой перекодированной CDS увеличивается в два раза по сравнению с аналогичным диапазоном исходных CDS.

Из сравнения дендрограммы на рис. 2,а, отражающей распределение структурно-статистических свойств неперекодированных последовательностей из исходной выборки, с аналогичными дендрограммами на рис. 2,б–г для бинарно перекодированных последовательностей исходной выборки, следует, что наибольшая сохранность структурно-статистических свойств проявляется при бинарном перекодировании, когда происходит замена букв A и T символом 1 и букв G и C

символом 0 (см. рис. 2,б). Следовательно, из рассматриваемых здесь трёх вариантов бинарного перекодирования, наименьшая потеря информации происходит при отождествлении комплементарных нуклеотидов, находящихся на одной нити ДНК. Этот факт позволяет предположить, что участки двунитевой ДНК, содержащей кодирующие последовательности, имеют характерное распределение двух типов пар комплементарных нуклеотидов (A/T и G/C).

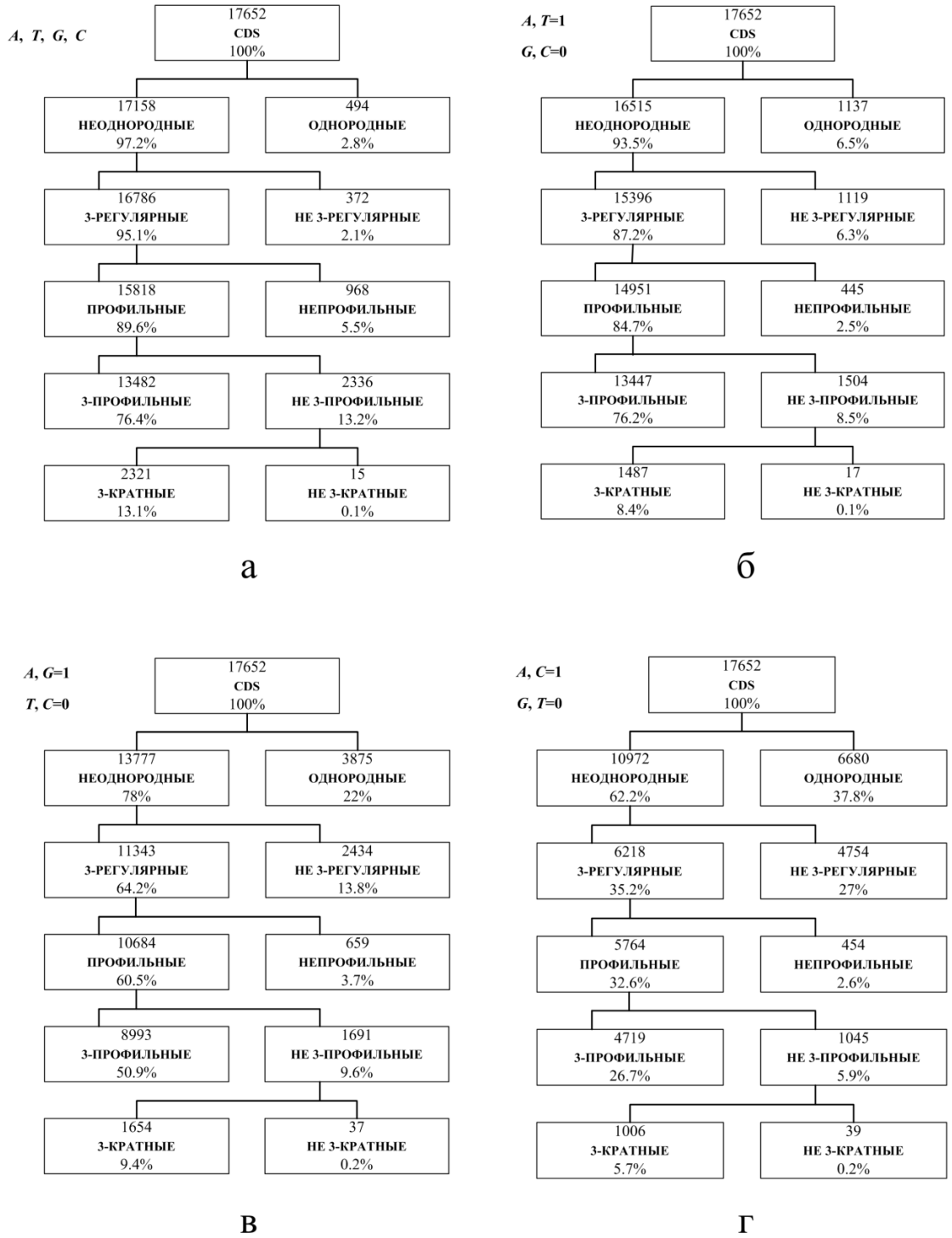

Рис. 2. Дендрограммы, отражающие распределение структурно-статистических свойств CDS генома человека. а) Исходные CDS в алфавите $\{A,T,G,C\}$. Аналогичные дендрограммы для бинарно перекодированных CDS с заменами букв: б) $A,T=1$ и $G,C=0$; в) $A,G=1$ и $T,C=0$; г) $A,C=1$ и $G,T=0$.

Сравнение дендрограмм на рис. 2,а и рис. 2,б демонстрирует качественное сходство распределений структурно-статистических свойств исходных и бинарно перекодированных CDS генома человека. Таким образом, несмотря на потерю информации, выбранный способ бинарного перекодирования CDS генома человека подтверждает наличие в CDS свойств 3-регулярности, триплетной профильности и двухуровневой организации кодирования, когда в последовательности на фоне 3-регулярности распознаётся скрытая профильная периодичность с размером периода, кратным и не равным трём.

4. АНАЛИЗ СОХРАННОСТИ СТРУКТУРНО-СТАТИСТИЧЕСКИХ СВОЙСТВ ПРИ БИНАРНОЙ ПЕРЕКОДИРОВКЕ CDS ГЕНОМА ЧЕЛОВЕКА БЕЗ ПОТЕРИ ИНФОРМАЦИИ

Рассмотрим случай бинарной перекодировки CDS генома человека из указанной выше выборки, выполненной без потери информации, когда для букв исходного алфавита используются замены: $A = 00, T = 01, G = 10, C = 11$. Поскольку при такой перекодировке длина последовательности увеличивается в два раза, а размер алфавита сокращается в два раза, диапазон тест-периодов для бинарно перекодированных CDS увеличивается в четыре раза по сравнению с диапазоном тест-периодов исходных CDS. Дендрограмма распределения структурно-статистических свойств перекодированных CDS приведена на рис. 3.

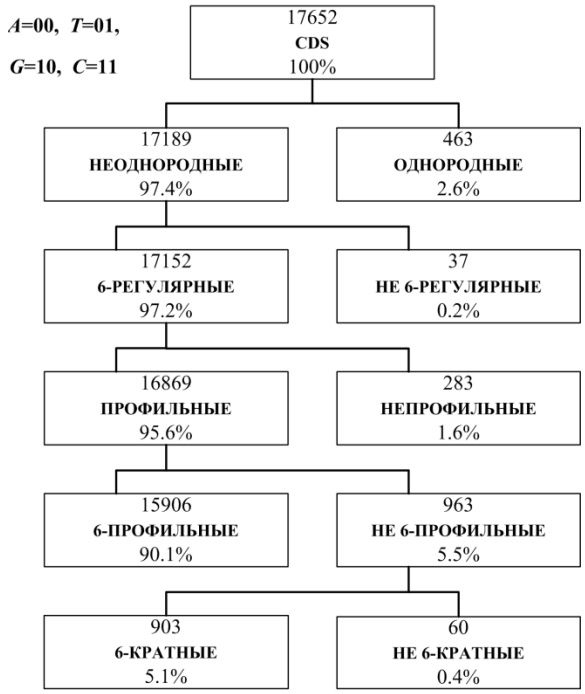

Рис. 3. Дендрограмма распределения структурно-статистических свойств бинарно перекодированных CDS с заменами букв: $A = 00, T = 01, G = 10, C = 11$.

Сравнение дендрограмм распределения структурно-статистических свойств в исходных CDS (рис. 2,а) и в бинарно перекодированных, без потери информации, CDS (рис. 3) показывает их качественное сходство. Как следует из дендрограмм, практически все исходные и перекодированные CDS являются неоднородными и обладают соответствующим свойством регулярности. При заданном способе перекодирования свойство 3-регулярности исходных CDS трансформируется в свойство 6-регулярности бинарно перекодированных последовательностей. Отметим, что среди всех бинарно перекодированных CDS, не являющихся 6-регулярными,

свойством 2-регулярности и 3-регулярности обладают всего лишь девять и пять последовательностей, соответственно. Для 6-регулярных CDS индекс 6-регулярности доминирует над индексами 2- и 3-регулярности. С расширением в четыре раза диапазона тест-периодов в спектрах бинарно перекодированных CDS может быть связано увеличение доли профильных последовательностей (в сравнении с исходными CDS, см. рис.2,а). Возможно, по этой же причине увеличивается доля 6-профильных в бинарно перекодированных CDS по сравнению с долей 3-профильных последовательностей в исходных CDS. Среди бинарно перекодированных профильных CDS, обладающих свойством 6-регулярности, наблюдается значимая доля последовательностей, размер периода которых кратен, но не равен шести нуклеотидам. На рис. 1 приведены спектры для исходной (неперекодированной) CDS (KEGG, hsa:389677). Из этих спектров следует, что анализируемая последовательность обладает свойством 3-регулярности, в ней отсутствует 3-профильность и распознаётся 24-профильность. На рис. 4 для этой бинарно перекодированной последовательности приведены аналогичные спектры. Из анализа спектров на рис. 4 следует, что бинарно перекодированная последовательность неоднородна (см. рис. 4,а), обладает свойством 6-регулярности $(I_6 = 0.92, I_3 = 0.79, I_2 = 0.50)$, в ней отсутствует 6-профильность (см. рис. 4,г) и распознаётся 48-профильность (см. рис. 4,в). Результаты анализа говорят о существовании в последовательности двухуровневой организации кодирования. Первый уровень кодирования проявляется в наличии свойства 6-регулярности при отсутствии 6-профильности. Второй уровень кодирования соответствует 48-профильности.

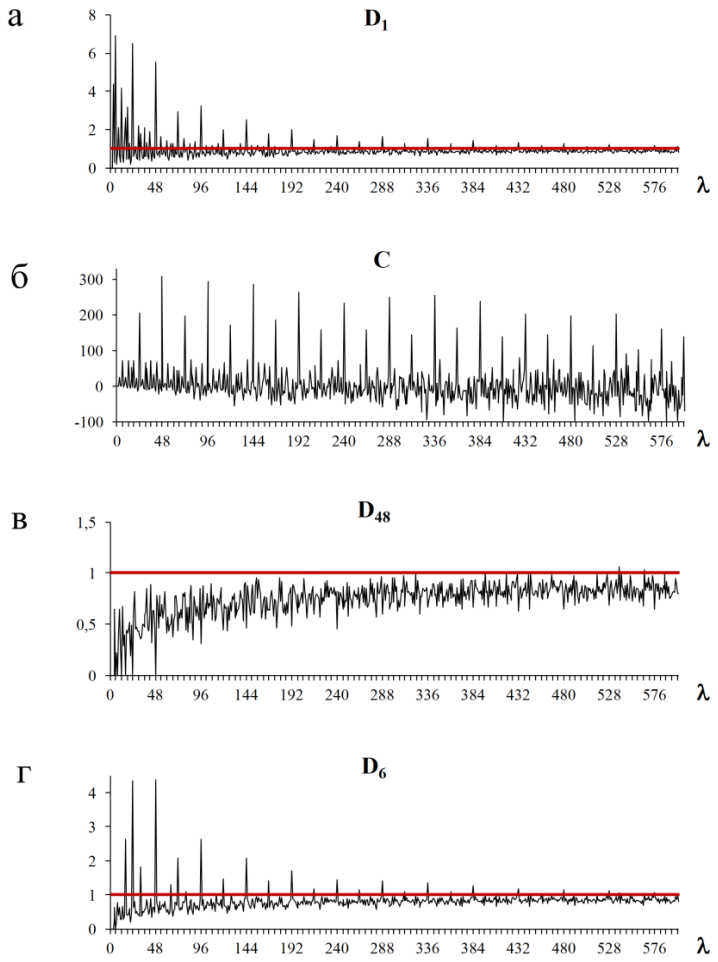

Рис. 4. Распознавание скрытой 48-профильности в бинарно перекодированной CDS из генома человека (KEGG, hsa:389677, 3006 нукл.) с помощью спектров 2С-подхода.

Таким образом, наблюдаемое практически во всех перекодированных CDS генома человека свойство 6-регулярности и профильная периодичность с размером периода, кратным шести, статистически достоверно подтверждают распространённое в литературе мнение о скрытой триплетной периодичности в кодирующих районах генома. Кроме того, рассматриваемое в этом разделе бинарное перекодирование CDS генома человека подтверждает наличие в кодирующих районах генома человека двухуровневой организации кодирования.

5. АНАЛИЗ РАСПРЕДЕЛЕНИЯ ДЛИН ПРОФИЛЬНЫХ CDS ИЗ ГЕНОМА ЧЕЛОВЕКА

В настоящем разделе анализируется распределение длин в двух группах профильных CDS из генома человека. Первую группу составляют 3-профильные последовательности. Во второй группе собраны последовательности с двухуровневой организацией кодирования, которые обладают свойством 3-регулярности и имеют размер периода, кратный и не равный трём. Количественный состав этих групп приведён на дендрограмме рис. 2,а. На рис. 5,а показано распределение по длинам CDS в группе 3-профильных последовательностей. Поскольку 3-профильные последовательности составляют ~ 80% CDS генома человека, распределение длин в этой группе повторяет аналогичное распределение среди всех CDS. На рис. 5,б показано распределение длин CDS в группе с двухуровневой организацией кодирования. Сравнение рис. 5,а и рис. 5,б показывает качественное сходство распределения длин последовательностей в этих группах.

Как было показано в разделе 3, при бинарной перекодировке с потерей информации наибольшая сохранность распределения структурно-статистических свойств в CDS генома человека наблюдалась при замене букв: $A,T = 1$ и $G,C = 0$. Перекодированные таким образом CDS с распознанной профильностью также разбиваются на две группы. Как и ранее, в первой группе собраны 3-профильные последовательности, во второй – последовательности с двухуровневой организацией кодирования. Распределения длин последовательностей в обеих группах показаны на рис. 6.

Из сравнения рис. 5 и рис. 6 можно сделать вывод о качественном сходстве распределений длин в двух аналогичных группах исходных и перекодированных CDS. Рассматриваемая бинарная перекодировка основана на отождествлении пар одного типа (A/T = T/A и G/C = C/G) комплементарных нуклеотидов в двойной нити ДНК. Поэтому такое сходство распределений можно рассматривать в качестве косвенного подтверждения особой роли пар комплементарных нуклеотидов в формировании структурно-статистических свойств кодирующих участков генома.

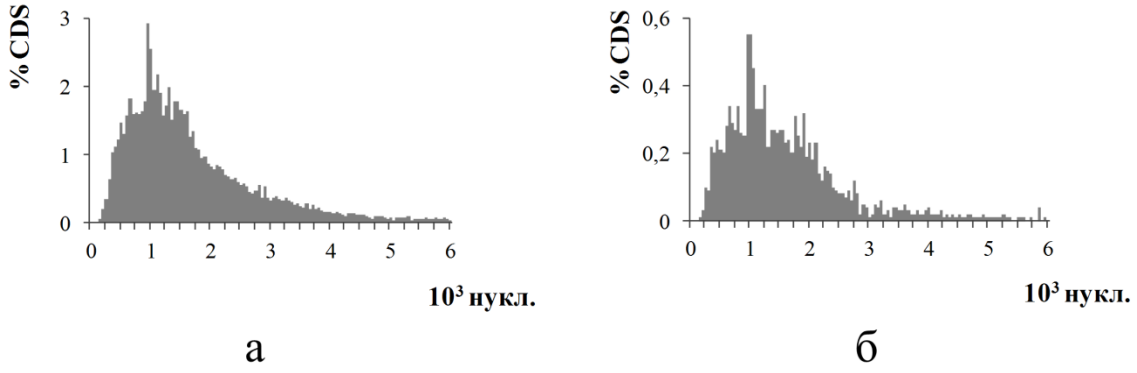

Рис. 5. Для последовательностей CDS, анализируемых в исходном четырехбуквенном алфавите, показано распределение длин в группах из (а) 13482 3-профильных CDS и (б) 2336 CDS с двухуровневой организацией кодирования.

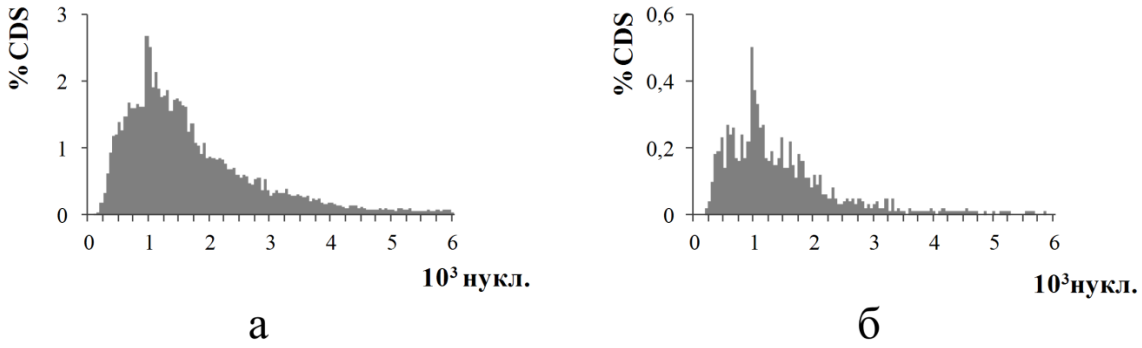

Рис. 6. Для бинарно перекодированных (A, T=1 и G, C=0) последовательностей CDS показано распределение длин в группах из (а) 13447 3-профильных CDS и (б) 1504 CDS с двухуровневой организацией кодирования.

6. ЗАКЛЮЧЕНИЕ

В работе исследовались структурно-статистические свойства кодирующих последовательностей ДНК (CDS) такие, как 3-регулярность, скрытая профильная периодичность (профильность) и скрытая триплетная профильная периодичность (3-профильность). Особо рассматривается обнаруженное явление двухуровневой организации кодирования в CDS. Первый уровень кодирования определяется свойством 3-регулярности последовательности. Второй уровень проявляется на фоне 3-регулярности как скрытая профильная периодичность с размером периода, кратным, но не равным трём. Для подтверждения объективного существования этих свойств в CDS выполнялось бинарное перекодирование их последовательностей.

Было показано, что при бинарном перекодировании CDS с потерей информации при замене A, T = 1, G, C = 0 достигается наибольшая сохранность структурно-статистических свойств CDS генома человека. Такой вариант перекодировки соответствует рассмотрению кодирующего участка как фрагмента двунитевой ДНК, представленного последовательностью пар комплементарных нуклеотидов. На основе этого факта можно предполагать существенную роль, которую играет распределение пар комплементарных нуклеотидов в процессе распознавания кодирующих участков в двунитевой ДНК хромосомы.

Бинарное перекодирование CDS без потери информации при замене A = 00, T = 01, G = 10, C = 11 дало возможность в два раза расширить диапазоны тестируемых периодов исходных CDS. Это позволило выявить профильную периодичность практически во всех (с учётом статистической погрешности) CDS генома человека. Кроме того, в этом случае триплетная профильная периодичность распознаётся в 90% CDS генома человека.

Таким образом, в работе показано, что скрытая триплетная периодичность в кодирующих районах ДНК генома человека относится к типу профильной периодичности. Как отмечалось ранее [15, 16], профильная периодичность расширяет понятие тандемного повтора и размытые тандемные повторы являются частным случаем проявления скрытой профильной периодичности.

Работа была выполнена при частичной поддержке гранта № 15-07-05783 Российского фонда фундаментальных исследований (РФФИ).

СПИСОК ЛИТЕРАТУРЫ

1. Trifonov E.N., Sussman J.L. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. USA*. 1980. V. 77. P. 3816–3820.
2. Shepherd J.C. Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. *J. Mol. Evol*. 1981. V. 17. P. 94–102.
3. Fickett J.W. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*. 1982. V. 10. P. 5303–5318.
4. Fickett J.W., Tung C.S. Assessment of protein coding measures. *Nucleic Acids Res*. 1992. V. 20. P. 6441–5640.
5. Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Biosci*. 1997. V. 13. P.263–270.
6. Grosse I., Herzel H., Buldyrev S.V., Stanley H.E. Species independence of mutual information in coding and noncoding DNA. *Phys. Rev. E*. 2000. V. 61. P. 5624–5629.
7. Yin C., Yau S.S.T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J. Theor. Biol.* 2007. V. 247. P. 687–694. doi: 10.1016/j.jtbi.2007.03.038.
8. Wang L., Stein L.D. Localizing triplet periodicity in DNA and cDNA sequences. *BMC Bioinformatics*. 2010. V. 11. doi: 10.1186/1471-2105-11-550.
9. Marhon S.A., Kremer S.C. Gene prediction based on DNA spectral analysis: a literature review. *J. Comput. Biol*. 2011. V. 18. P. 639–676. doi: 10.1089/cmb.2010.0184.
10. Hua W., Wang J., Zhao J. Discrete Ramanujan transform for distinguishing the protein coding regions from other regions. *J. Mol. Cell Probes*. 2014. V. 28. P. 228–236. doi: 10.1016/j.mcp.2014.04.002.
11. Yin C. Representation of DNA sequences in genetic codon context with applications in exon and intron prediction. *J. Bioinform. Comput. Biol.* 2015. V. 13. № 2. doi: 10.1142/S0219720015500043.
12. Chaley M., Kutyrkin V., Tulbasheva G., Teplukhina E., Nazipova N. HeteroGenome: database of genome periodicity. *Database*. 2014. V. 2014. P. 1–18. doi: 10.1093/database/bau040.
13. 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. doi: 10.1016/j.mbs.2007.10.008.
14. Chaley M.B., Kutyrkin V.A. Structure of proteins and latent periodicity in their genes. *Moscow Univ. Biol. Sci. Bull*. 2010. V. 65. № 4. P. 133–135. doi: 10.3103/S0096392510040012.
15. Chaley M., Kutyrkin V. Profile-Statistical periodicity of DNA coding regions. *DNA Res.* 2011. V. 18. P. 353–362.
16. Чалей М.Б., Кутыркин В.А. Распознавание скрытой периодичности в последовательностях ДНК. *Математическая биология и биоинформатика.* 2013. Т. 8. № 2. С. 502–512. doi: 10.17537/2013.8.502.
17. Кутыркин В.А., Чалей М.Б. Структурные различия кодирующих и некодирующих районов последовательностей ДНК генома человека. *Вестник МГТУ им. Н.Э.Баумана. Серия Естественные науки*. 2012. Спец. выпуск № 3 «Математическое моделирование». С. 146–157.
18. Кутыркин В.А., Чалей М.Б. Спектрально-статистический подход к распознаванию скрытой профильной периодичности в последовательностях ДНК. *Математическая биология и биоинформатика.* 2014. Т. 9. № 1. С. 33–62. doi: 10.17537/2014.9.33.

19. Kanehisa M., Goto S., Sato Y., Furumichi M., Tanabe M. KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res*. 2012. V. 40. № D1. P. D109–D114. doi: 10.1093/nar/gkr988.

Материал поступил в редакцию 03.09.2015, опубликован 02.10.2015.