Математическая биология и биоинформатика. 2013. Т. 8. № 2. С. 502–512.
URL: http://www.matbio.org/2013/Chaley_8_502.pdf

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

УДК: 577.2:519.23

Распознавание скрытой периодичности в последовательностях ДНК

Чалей М.Б.[*1], Кутыркин В.А.[2]**

[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. На основе разработанного ранее спектрально-статистического подхода предложены прямые методы распознавания нового типа периодичности в последовательности ДНК – скрытой профильной периодичности, обобщающей известное понятие размытого тандемного повтора. Проведено сравнение предлагаемых методов с другими подходами, опирающимися только на косвенные признаки наличия скрытой периодичности в последовательности ДНК.

Ключевые слова: *размытый тандемный повтор, скрытая профильная периодичность, статистический анализ текстовой структуры последовательности ДНК.*

1. ВВЕДЕНИЕ

Проблеме выявления скрытой периодичности в ДНК посвящено достаточно много работ. Многие из этих работ ориентированы на выделение необходимых признаков, которые могут быть связаны с наличием скрытой периодичности в анализируемых последовательностях ДНК. К таким признакам можно отнести высокие значения амплитуд в Фурье-анализе, выявляемый значительный уровень отклонения от однородности или от других усреднённых показателей и т. п. Исследованию таких количественных характеристик посвящены многие работы, например, [1–11] и др. При этом во многих случаях возникает противоречивая картина результатов. В то время как Фурье-анализ выделяет в качестве оценки скрытого периода одно значение, альтернативный метод предлагает другое значение, многократно отличающееся от первого. В частности, в работе [2] такие оценки скрытого периода различаются в 10 и более раз. Отдать предпочтение какой-либо из этих оценок не представляется возможным, так как при попытке выявления скрытой периодичности используются только необходимые признаки. Без достоверного распознавания эталона они не являются достаточными. В таком случае следует называть эти признаки косвенными.

Для распознавания скрытой периодичности необходимо привести эталон периодической строки, к которому достаточно близка анализируемая последовательность ДНК. Например, в базе TRDB [12] накапливаются последовательности со скрытой периодичностью, основанной на понятии размытого тандемного повтора [13]. В этом случае в качестве эталона предлагается совершенный текстовой тандемный повтор с паттерном периодичности, однозначно определяемым найденным консенсус-паттерном. Близость анализируемого размытого тандемного

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

повтора к эталону оценивается с помощью небольшого числа (~20%) фиксируемых повреждений (вставок, делеций и замен букв алфавита ДНК) по отношению к эталону. Тем самым, для пополнения базы TRDB используется метод распознавания размытых тандемных повторов [13], в котором найденный консенсус-паттерн обеспечивает достаточный признак наличия скрытой периодичности в последовательностях ДНК. Аналогичный подход используется в работе [14], где для районов скрытой периодичности, найденных с помощью необходимых косвенных признаков, предлагается подтверждение их периодичности на основе визуального анализа профиля последовательности ДНК, представленного столбцом непрерывных подстрок с длиной предполагаемого паттерна [9–11]. В методах, упомянутых выше [1–11], используются только косвенные признаки. Далее будет продемонстрирована недостаточность одних только косвенных признаков для распознавания скрытой периодичности.

Ранее в работе [11] был предложен новый тип скрытой периодичности, названный скрытой профильной периодичностью, или профильностью. Было показано, что такое новое понятие скрытой периодичности обобщает понятие размытого тандемного повтора. В качестве эталона служит совершенная периодическая случайная строка, индуцируемая случайным паттерном, который определяет статистическую структуру реализаций эталона. Такой паттерн подробно описывается в следующем разделе. Следовательно, для распознавания скрытой профильности необходимо создать статистические критерии, которые будут фиксировать близость анализируемой последовательности ДНК к эталонной случайной строке.

В настоящей работе описываются методы распознавания скрытой профильности в последовательностях ДНК. Для оценки периода скрытой профильности используется количественный критерий, выступающий в качестве необходимого признака наличия скрытой периодичности в последовательности ДНК. На основе этой оценки и статистической структуры анализируемой последовательности предлагается случайный паттерн периодичности, индуцирующий эталонную случайную строку. Далее предлагаются достаточные признаки наличия скрытой профильности в анализируемой последовательности ДНК. Эти достаточные признаки основаны, в частности, на статистических критериях, оценивающих близость анализируемой последовательности к эталонной случайной строке. Следовательно, такие методы позволяют рассматривать анализируемую последовательность как реализацию совершенно случайной периодической строки (эталона).

2. МОДЕЛЬ ЭТАЛОНА ДЛЯ РАСПОЗНАВАНИЯ СКРЫТОЙ ПРОФИЛЬНОСТИ В ПОСЛЕДОВАТЕЛЬНОСТИ ДНК

В основе модели скрытой профильности рассматривается мульти-полиномиальная схема испытаний. Эта схема индуцируется упорядоченным набором из L последовательных полиномиальных схем с четырьмя исходами (4 буквы алфавита ДНК). Следовательно, в каждом испытании реализуется случайная буква, которая описывается вероятностным распределением букв алфавита ДНК. В первом испытании этой мульти-полиномиальной схемы реализуется первая случайная буква (полиномиальная схема), во втором – вторая, …, в L-том – L-ая случайная буква (полиномиальная схема). Затем такая последовательность из набора L испытаний повторяется. На заключительном этапе испытаний может повториться менее L реализаций таких случайных букв. Для всей мульти-полиномиальной схемы описанный выше упорядоченный набор из L полиномиальных схем образует её профиль, или паттерн, составленный из независимых случайных букв и называемый далее паттерном профильной периодичности. Эту мульти-полиномиальную схему можно рассматривать как случайную строку из перечисленных независимых случайных букв. Следовательно, эталоном для скрытой профильности является совершенный тандемный повтор,

паттерн которого представлен случайной строкой из независимых случайных букв. Такая случайная строка далее будет называться профильной строкой.

2.1. Определение эталона профильной периодичности

Эталоном скрытой профильной периодичности является профильная строка. Дадим её строгое определение.

Пусть $Chr(\mathbf{p})$ – случайная буква со столбцом частот $\mathbf{p} = (p^1,..., p^K)^T$, которая является случайной величиной, принимающей с вероятностью p^i значение i-той буквы алфавита $A = \langle a_1,..., a_K \rangle$. Специальная случайная строка $Str = Str_n(\boldsymbol{\pi}) = Chr(\mathbf{p}_1)\ldots Chr(\mathbf{p}_n)$ из n независимых случайных букв индуцируется матрицей $\boldsymbol{\pi} = (\mathbf{p}_1,..., \mathbf{p}_n) = (\pi_j^i)_n^K$, называемой n-профильной матрицей. Любое целое число L из диапазона $1,..., L_{\max}$, где $L_{\max} \sim n/5K$, называется тест-периодом строки Str. Букву $a_i \in A$ можно отождествить со случайной буквой, все компоненты столбца частот которой нулевые, за исключением i-той единичной компоненты. Поэтому любую текстовую строку в алфавите A можно отождествлять с соответствующей специальной случайной строкой той же длины.

Пусть L – тест-период, $0 \le M < L$ и $Str = Str_n(\boldsymbol{\pi}) = Str_L(\boldsymbol{\pi}_1)\ldots Str_L(\boldsymbol{\pi}_m) Str_M(\boldsymbol{\pi}_{m+1})$ – разложение строки Str на подстроки длины L. Тогда, если $M = 0$, матрица $\Pi_{Str}(L) = m^{-1}\sum_{i=1}^{m} \boldsymbol{\pi}_i$ называется L-профильной матрицей строки Str. Если $M \neq 0$, то в матрицу $\Pi_{Str}(L)$ вносятся соответствующие поправки. Таким образом, для строки Str введён профильно-матричный спектр Π_{Str}, определённый на каждом тест-периоде. Если $\boldsymbol{\pi}_1 = \ldots = \boldsymbol{\pi}_m = \boldsymbol{\pi}_0$ и $\boldsymbol{\pi}_0 = (\boldsymbol{\pi}_{m+1}, \boldsymbol{\pi}_{01})$, то строка Str называется L-профильной строкой со случайным паттерном периодичности $Str_L(\boldsymbol{\pi}_0)$. В этом случае для строки Str используется обозначение $Tdm_L(\boldsymbol{\pi}_0, n)$ и матрица $\boldsymbol{\pi}_0$ называется её главной профильной матрицей, поскольку она индуцирует весь профильно-матричный спектр этой строки.

3. МЕТОДЫ РАСПОЗНАВАНИЯ СКРЫТОЙ ПРОФИЛЬНОЙ ПЕРИОДИЧНОСТИ В ТЕКСТОВЫХ СТРОКАХ

Для распознавания скрытой профильности в анализируемой последовательности ДНК необходимо найти такую профильную строку (эталон), для которой анализируемую последовательность ДНК можно рассматривать в качестве её реализации. Поэтому скрытая профильная периодичность в последовательности ДНК может быть выявлена только с помощью статистических критериев. При анализе последовательности ДНК в этих критериях используются различные статистики [9–11], имеющие вид функциональных зависимостей (спектров) от тестируемых периодов последовательности. Тестируемый период (тест-период) – это натуральное число, не превышающее половины длины анализируемой последовательности ДНК.

3.1. Оценки паттерна эталона скрытой профильной периодичности на основе критериев согласия

Пусть str – текстовая строка длины n в алфавите $A = \langle a_1, ..., a_K \rangle$, анализируемая на наличие скрытой профильности. Для того, чтобы проверить, является ли тест-период λ периодом скрытой профильной периодичности, используется статистика

$$\psi(\Pi_{str}(\lambda), \Pi_{Tdm_\Lambda}(\lambda), n) = \frac{n}{\lambda} \sum_{j=1}^{\lambda} \sum_{i=1}^{K} (\pi^{*i}_j - \pi^i_j)^2 \big/ \pi^i_j \sim \chi^2_{(K-1)\lambda}, \quad (1)$$

где $\Pi_{str}(\lambda) = (\pi^{*i}_j)^K_\lambda$ и $\Pi_{Tdm_\Lambda}(\lambda) = (\pi^i_j)^K_\lambda$ – λ-профильные матрицы текстовой строки str и Λ-профильной строки $Tdm_\Lambda = Tdm_\Lambda(\Pi_{str}(\Lambda), n)$, соответственно. Кроме того, в формуле (1) указано, что вводимая статистика имеет χ^2-распределение с $(K-1)\lambda$ степенями свободы при $n/\lambda > 5K$, если строку str можно рассматривать как реализацию строки Tdm_Λ. В этом случае используются критерии согласия Пирсона [15, с. 483].

Статистика (1) позволяет создать спектр D_1 отклонения строки str от однородности (1-профильности), который на тест-периоде λ принимает значение:

$$D_1(\lambda) = \psi(\Pi_{str}(\lambda), \Pi_{Tdm_1}(\lambda), n) / \chi^2_{crit}((K-1)\lambda, \alpha), \quad (2)$$

где $Tdm_1 = Tdm_1(\Pi_{str}(1), n)$, и $\chi^2_{crit}(N, \alpha)$ – критическое значение χ^2_N-распределения на уровне значимости $\alpha = 0.05$.

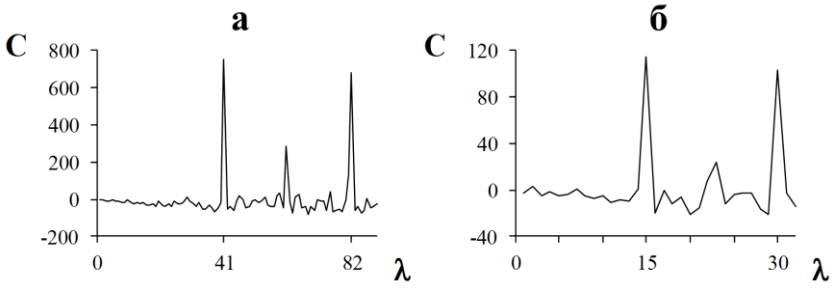

Рис. 1. Характеристические спектры размытых тандемных повторов из базы данных TRDB.
(**а**) Тандемный повтор на хромосоме I из генома человека (индексы: 2956–4750, размер паттерна 41 нукл., % несовпадений между копиями – 13, % вставок и делеций между копиями – 5).
(**б**) Тандемный повтор на хромосоме XI из генома мыши *M. musculus* (индексы: 6611441–6612079, размер паттерна 15 нукл., % несовпадений между копиями – 15, % вставок и делеций между копиями – 5).

Если в формуле (2) $D_1 < 1$, то строка str признаётся однородной, т. е. её период профильной периодичности равен 1. В противном случае, строка признаётся неоднородной. Неоднородную строку можно анализировать на наличие в ней скрытой L-профильной периодичности, когда скрытый период $L > 1$. В работах [16–19] показано, что для получения оценки такого скрытого периода может использоваться характеристический спектр C, принимающий на тест-периоде λ значение:

$$C(\lambda) = \psi(\Pi_{str}(\lambda), \Pi_{Tdm_\Lambda}(\lambda), n) - M(\chi^2_{(K-1)\lambda}), \quad (3)$$

где $M(\chi^2_N)$ – математическое ожидание χ^2-распределения с N степенями свободы. В работах [16–19] для оценки скрытого периода было предложено следующее эффективное правило. Первый тест-период L с ярко выраженным максимальным

значением спектра C служит оценкой скрытого периода в строке str. В частности, эффективность этого правила была проверена на размытых тандемных повторах из базы TRDB [12] (см. рис. 1).

На рисунке 1 приведены характеристические спектры размытых тандемных повторов с периодом 41 нуклеотид (нукл.) (рис. 1,**а**) и с периодом 15 нукл. (рис. 1,**б**). Эти спектры, согласно сформулированному правилу, позволяют получить правильные оценки скрытых периодов для рассматриваемых тандемных повторов.

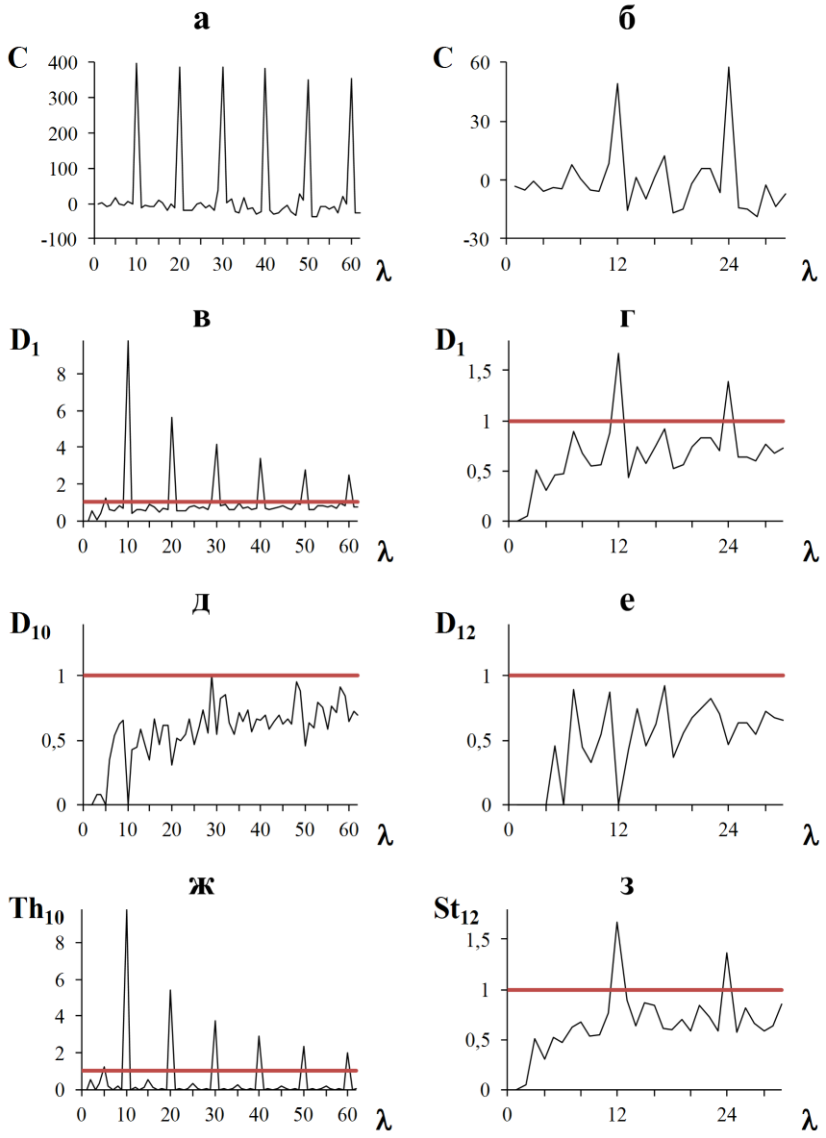

Рис. 2. Спектры в левой половине рисунка соответствуют последовательности ДНК из генома *C. elegans*, хромосомы III (индексы: 307381–308580). Спектры, представленные в правой половине рисунка, построены для последовательности ДНК из генома *A. thaliana*, хромосомы IV (индексы: 4904045 – 4904399). (**а**, **б**) Характеристические спектры. (**в**, **г**) Спектры D_1 отклонения от однородности (1-профильности). (**д**, **е**) Спектры D_L отклонения от L-профильности, где $L=10$, $L=12$, соответственно. (**ж**) Теоретическая реконструкция Th_L ($L=10$) спектра D_1 на рис. 2,**в**. (**з**) Статистическая реконструкция St_L ($L=12$) спектра D_1 на рис. 2,**г**.

В работах [16–19] была подтверждена эффективность указанного выше правила для оценки периода скрытой профильной периодичности в последовательностях ДНК, не являющихся размытыми тандемными повторами. На рисунках 2,**а** и 2,**б** приведены примеры характеристических спектров для нуклеотидных последовательностей, не

являющихся размытыми тандемными повторами. Далее будет показано, что в этих последовательностях выявляется скрытая периодичность с периодами $L=10$ нукл. (рис. 2,**а**) и $L=12$ нукл. (рис. 2,**б**).

Пусть L – оценка скрытого периода, полученная указанным выше способом. Тогда с помощью статистики (1) вводится спектр D_L отклонения строки *str* от L-профильности, принимающий на тест-периоде λ значение:

$$D_L(\lambda) = \psi(\Pi_{str}(\lambda), \Pi_{Tdm_L}(\lambda), n) / \chi^2_{crit}((K-1)\lambda, \alpha), \quad (4)$$

где $Tdm_L = Tdm_L(\Pi_{str}(L), n)$ и $\chi^2_{crit}(N, \alpha)$ – критическое значение χ^2_N-распределения на уровне значимости α ($\alpha = 0.05$). Если $D_L < 1$, то принимается гипотеза о том, что в строке *str* наблюдается скрытая L-профильность, т. е. её период профильной периодичности равен L. В противном случае, принимается гипотеза об отсутствии скрытой L-профильности в строке *str*.

На рисунках 2,**в** и 2,**г** показаны спектры отклонения от однородности для двух последовательностей ДНК, не являющихся, как было отмечено выше, размытыми тандемными повторами. Согласно выработанному правилу (см. формулу (2) и текст к ней) эти последовательности признаются неоднородными. Анализ характеристических спектров этих последовательностей ДНК позволяет выбрать в качестве оценок периодов скрытой профильной периодичности значения $L=10$ (рис. 2,**а**) и $L=12$ (рис. 2,**б**). Спектры отклонения от L-профильности, приведённые на рисунках 2,**д** и 2,**е**, согласно предложенному критерию согласия (см. формулу (4) и текст к ней), подтверждают правильность предложенных оценок периодов скрытой профильности.

3.2. Методы реконструкции спектра отклонения от однородности для подтверждения оценки паттерна эталона скрытой профильной периодичности

Пусть для строки *str* принята гипотеза о наличии в ней скрытой L-профильности, следовательно, строку *str* можно рассматривать как реализацию L-профильной строки $Tdm_L = Tdm_L(\Pi_{str}(L), n)$, так как профильно-матричный спектр строки *str*, согласно критерию согласия Пирсона (см. формулу (4) и текст к ней), статистически неотличим от профильно-матричного спектра строки Tdm_L. В этом случае паттерн строки Tdm_L служит оценкой паттерна эталона для скрытой L-профильности в строке *str*.

Поскольку для анализируемой строки *str* при создании эталона $Tdm_L = Tdm_L(\Pi_{str}(L), n)$ L-профильной периодичности используется критерий согласия, то для оценки полученного паттерна этой скрытой профильной периодичности можно получить дополнительное подтверждение. В работах [16–19] в качестве такого подтверждения был предложен метод сравнения спектров D_1 (см. (2)) и Th_L отклонения от однородности строк *str* и Tdm_L, соответственно. Спектр D_1 выбран как наиболее информативный по отношению к статистической структуре строки *str*. По аналогии с формулой (2) в качестве спектра Th_L отклонения от однородности строки Tdm_L на тест-периоде λ принимается формула:

$$Th_L(\lambda) = \psi(\Pi_{Tdm_L}(\lambda), \Pi_{Tdm_1}(\lambda), n) / \chi^2_{crit}((K-1)\lambda, \alpha). \quad (5)$$

Фактически, спектр Th_L является теоретической реконструкцией спектра D_1 строки *str*, и, в случае корректности оценки паттерна эталона скрытой периодичности, спектр Th_L имеет наглядное сходство со спектром D_1. На рисунке 2,**ж** приведён спектр теоретической реконструкции спектра отклонения от однородности для последовательности ДНК из генома *Caenorhabditis elegans*. Наблюдаемое наглядное

сходство этой реконструкции с исходным спектром отклонения от однородности (рис. 2,**в**) подтверждает выявленную скрытую периодичность.

Для подтверждения выявленной скрытой L-профильной периодичности можно использовать и статистическую реконструкцию спектра D_1 анализируемой строки str. В этом случае с помощью выборочной L-профильной матрицы $\Pi_{str}(L)$, используя датчик случайных чисел, создаётся строка str_* – статистический аналог строки str. Для строки str_* вычисляется спектр отклонения от однородности, обозначаемый St_L. При достаточной длине строки str, в случае корректности оценки паттерна эталона скрытой периодичности, спектр St_L имеет наглядное сходство со спектром D_1 (см. рис. 2,**з** и рис. 2,**г**).

4. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В настоящей работе предложены методы распознавания скрытой профильной периодичности в последовательностях ДНК. Понятие скрытой профильной периодичности обобщает понятие размытого тандемного повтора для последовательности ДНК. В отличие от размытого тандемного повтора в качестве паттерна скрытой L-профильной периодичности предлагается не текстовой консенсус-паттерн, а случайная строка $Str_L(\boldsymbol{\pi}_0)$, состоящая из независимых случайных букв. Профильная матрица $\boldsymbol{\pi}_0$ этой случайной строки однозначно определяет такой паттерн и называется профильной матрицей паттерна $Str_L(\boldsymbol{\pi}_0)$. Например, для рассмотренных в предыдущем разделе последовательностей ДНК *C. elegans* и *A. thaliana* (см. рис. 2), не являющихся размытыми тандемными повторами, были найдены оценки паттернов скрытой профильной периодичности, матрицы которых показаны на рисунке 3.

а

	1	2	3	4	5	6	7	8	9	10
A	0.58	0.38	0.49	0.53	0.32	0.07	0.09	0.12	0.27	0.39
T	0.08	0.12	0.21	0.33	0.57	0.75	0.77	0.62	0.30	0.10
G	0.28	0.38	0.15	0.09	0.08	0.16	0.12	0.22	0.38	0.45
C	0.05	0.11	0.15	0.06	0.03	0.02	0.02	0.04	0.06	0.06

б

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.17	0.13	0.27	0.23	0.37	0.53	0.53	0.59	0.34	0.24	0.28	0.14
T	0.47	0.47	0.47	0.40	0.20	0.13	0.30	0.28	0.48	0.66	0.59	0.69
G	0.23	0.33	0.20	0.23	0.37	0.13	0.07	0.07	0.03	0.03	0.07	0.03
C	0.13	0.07	0.07	0.13	0.07	0.20	0.10	0.07	0.14	0.07	0.07	0.14

Рис. 3. Матрицы – оценки паттернов скрытой профильной периодичности, выявленной в (**а**) последовательности ДНК *C. elegans*, хромосомы III (индексы: 307381 – 308580) и (**б**) последовательности ДНК *A. thaliana*, хромосомы IV (индексы: 4904045 – 4904399).

В методах, использующих Фурье-анализ, оценку периода скрытой периодичности получают в соответствии с частотой, на которой достигается наибольшее значение амплитуды в Фурье спектре анализируемой последовательности. Рисунок 4 иллюстрирует сравнение результатов выявления скрытой периодичности в последовательности ДНК с помощью Фурье-анализа и предлагаемого в настоящей работе спектрально-статистического подхода. Спектры Фурье были построены с

помощью FFT программ [6, 7], доступных на сайте [20]. Согласно Фурье-спектру на рис. 4,**а**, наибольшие пики достигаются на частотах 0.43, 0.11 и 0.33, соответствующих периодам 2.3, 3 и 9 нукл. Спектры D_1, D_2, D_3 и D_9 отклонения от соответствующей профильности, показанные на рис. 4,**в**, **д**, **ж**, **е**, позволяют говорить, что рассматриваемая последовательность является неоднородной (спектр D_1) и в ней нет профильной периодичности с периодами 2, 3 и 9 нукл. (спектры D_2, D_3 и D_9). Однако характеристический спектр последовательности (рис. 4,**б**) даёт оценку периода скрытой периодичности 84 нукл. Спектр D_{84} на рис. 4,**з** позволяет сделать вывод о наличии в последовательности скрытого периода в 84 нукл., что подтверждается теоретической реконструкцией Th_{84} (см. рис. 4,**г**) спектра D_1 (рис. 4,**в**). Рассмотренный пример показывает, что благодаря возможности верификации оценки периода, предлагаемый спектрально-статистический подход обладает преимуществом по сравнению с Фурье-анализом последовательности ДНК.

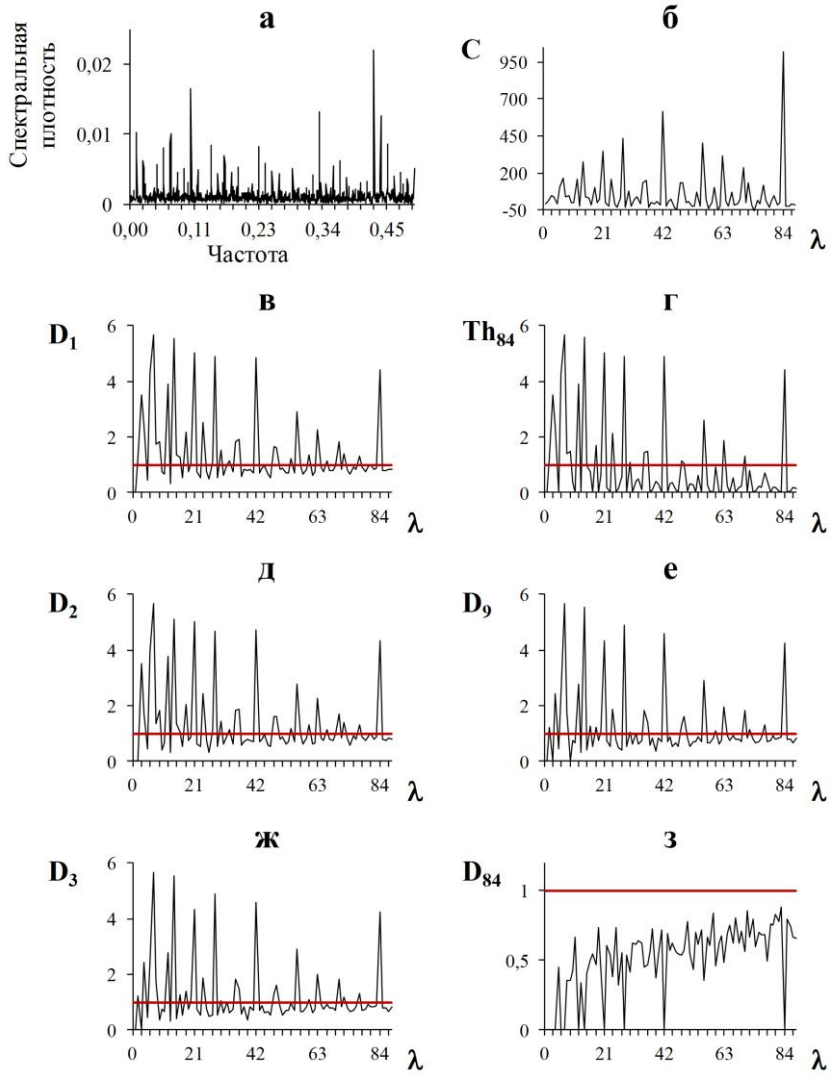

Рис. 4. Различные спектры для анализа нуклеотидной последовательности на скрытую периодичность. Рассматривается CDS (кодирующая последовательность ДНК) белка 285A цинковых пальцев человека из базы KEGG (hsa:26974, 1773 нукл.).

Аналогичный вывод можно сделать и при сравнении предлагаемого подхода с другим статистическим методом [2–4], основанном на Z-score статистике. Тест-период,

на котором достигается максимальное значение такой статистики, предлагался в качестве оценки периода скрытой периодичности без её необходимого подтверждения. В работе [4] на основе спектров Z-score статистики была получена оценка скрытого периода в 2 нукл. для четырёх последовательностей ДНК. Для двух из них из геномов бактерий *Burkholderia cepacia* и *Campylobacter jejuni* на рисунках 5,**а** и 5,**б** приведены характеристические спектры, и на рисунках 5,**в**, 5,**г** – спектры D_1 отклонения от однородности. Согласно спектру D_1 на рис. 5,**г,** одна из этих последовательностей является однородной, т. е. 1-профильной, и, следовательно, обладает скрытой мононуклеотидной периодичностью. Другая последовательность неоднородна (см. рис. 5,**в**), и, согласно её характеристическому спектру (см. рис. 5,**а**), оценка её скрытого периода равна 3 нукл.

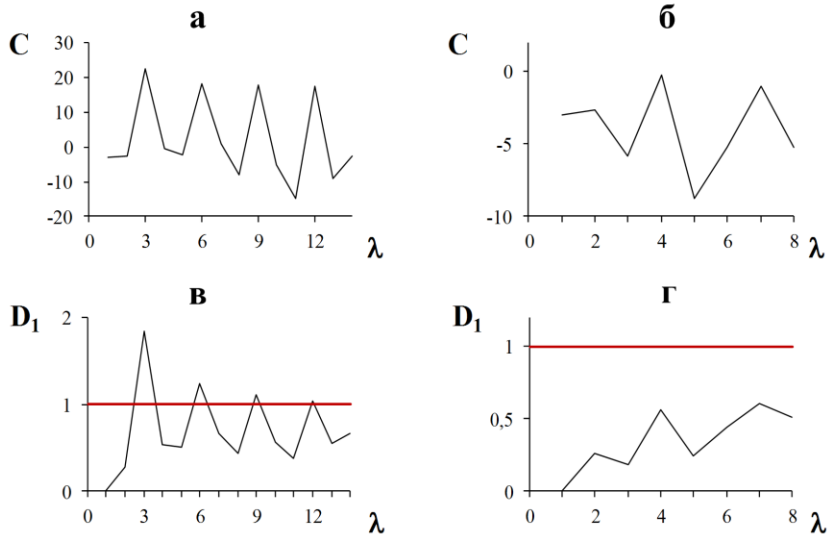

Рис. 5. Характеристические спектры (**а**, **б**) и спектры D_1 отклонения от однородности (**в**, **г**) для последовательностей ДНК из геномов бактерий: (**а**, **в**) *B. cepacia* (GenBank AF453480, GI:18182536, индексы: 4166 – 4368), (**б**, **г**) *C. jejuni*, (GenBank AL111168, GI:3047139, индексы: 176413 – 176535. В работе [4] для этой последовательности указан не существующий на сегодняшний день локус CJ11168X1).

Для рассматриваемой последовательности из генома *B. cepacia* на рис. 6,**а** и рис. 6,**б** приведены спектры D_3 и D_2 отклонения от 3- и 2-профильности, соответственно. Из этих рисунков следует, что в последовательности наблюдается скрытая 3-, но не 2-профильность. Этот вывод подтверждается теоретической реконструкцией Th_3 спектра D_1, что показано на рис. 6,**в**. Сравнение теоретической реконструкции Th_2 спектра D_1, приведённое на рис. 6,**г,** дополнительно опровергает предположение о наличии в последовательности скрытого периода в 2 нукл. Кроме того, рассматриваемая последовательность является кодирующей, и для неё трудно ожидать периода скрытой периодичности в 2 нукл.

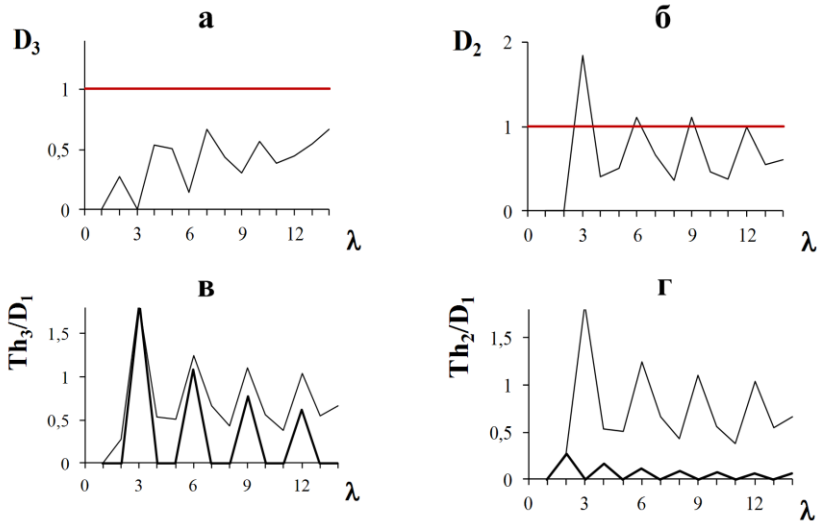

Рис. 6. Для последовательности ДНК из генома бактерии *B. cepacia* (GenBank AF453480, GI:18182536, индексы: 4166 – 4368) показаны: (**а**) спектр D_3 отклонения от 3-профильности и (**б**) спектр D_2 отклонения от 2-профильности; (**в**, **г**) теоретические реконструкции Th_3 и Th_2 (показаны жирной линией) спектра D_1 отклонения от однородности (тонкая линия) в предположении наличия в последовательности 3- и 2-профильности, соответственно.

Таким образом, результаты оценки скрытого периода, полученные в работе [4] без их верификации, могут оказаться некорректными. Аналогичный вывод справедлив и для двух других из упоминаемых выше четырёх последовательностей ДНК.

5. ВЫВОДЫ

В работе предложены методы распознавания нового типа периодичности – скрытой профильной периодичности, которая обобщает известное понятие размытого тандемного повтора. При распознавании скрытой профильной периодичности в качестве эталона предлагается специальная периодическая случайная строка, однозначно определяемая её паттерном. Этот паттерн состоит из независимых случайных букв, каждая из которых определяет вероятностное распределение букв текстового алфавита (алфавита ДНК). В этом случае с помощью разработанных статистических критериев (с заданным 5%-м уровнем значимости) показывается, что анализируемая текстовая строка (последовательность ДНК) может рассматриваться как реализация случайной эталонной строки. Сравнение предложенного спектрально-статистического подхода по распознаванию скрытой периодичности показало его преимущества по отношению к другим подходам, основанным на косвенных признаках наличия скрытой периодичности в последовательности ДНК.

СПИСОК ЛИТЕРАТУРЫ

1. Korotkov E.V., Korotkova M.A., Tulko J.S. Latent sequence periodicity of some oncogenes and DNA-binding protein genes. *CABIOS*. 1997. V. 13. P. 37–44.
2. Korotkov E.V., Korotkova M.A., Kudryashov N.A. Information decomposition method for analysis of symbolical sequences. *Physical Letters A*. 2003. V. 312. P. 198–210.
3. Shelenkov A., Korotkov A., Korotkov E. MMsat-a database of potential micro-and minisatellites. *Gene*. 2008. V. 409. P. 53–60.
4. Shelenkov A., Skryabin K., Korotkov E. Search and classification of potential minisatellite sequences from bacterial genomes. *DNA Res*. 2006. V. 13. P. 89–102.

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. Issac B., Singh H., Kaur H., Raghava G.P.S. Locating probable genes using Fourier transform approach. *Bioinformatics*. 2002. V. 18. P. 196–197.
7. Sharma D., Issac B., Raghava G.P.S., Ramaswamy R. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics*. 2004. V. 20. P. 1404–1412.
8. Лобзин В.В., Чечеткин В.Р. Порядок и корреляции в геномных последовательностях ДНК. Спектральный подход. *УФН*. 2000. Т. 170. С. 57–81.
9. Чалей М.Б., Назипова Н.Н., Кутыркин В.А. Совместное использование различных критериев проверки однородности для выявления скрытой периодичности в биологических последовательностях. *Мат. биология и биоинформатика*. 2007. Т. 2. №1. С. 20–35. URL: http://www.matbio.org/downloads/Chaley2007(2_20).pdf (дата обращения: 03.09.2013).
10. 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.
11. 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.
12. Gelfand Y., Rodriguez A., Benson G. TRDB – the Tandem Repeats Database. *Nucleic Acids Res*. 2007. V. 35. P. 80–87. URL: http://tandem.bu.edu/cgi-bin/trdb/trdb.exe (дата обращения 31.07.2013).
13. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res*. 1999.V. 27. P. 573-580.
14. Чалей М.Б., Кутыркин В.А., Тюльбашева Г.Э., Теплухина Е.И., Назипова Н.Н. Исследование феномена скрытой периодичности в геномах эукариотических организмов. *Математическая биология и биоинформатика*. 2013. Т. 8, №2. С. 480–501. URL: http://www.matbio.org/2013/Chaley_8_480.pdf (дата обращения: 03.09.2013).
15. Крамер Г. *Математические методы статистики*. М.: Мир, 1975. 648 С. (Перевод с англ. под ред. Колмогорова А.Н. Cramer H. *Mathematical methods of statistics*. Stockholm, 1946).
16. Chaley M.B., Kutyrkin V.A. Structure of proteins and latent periodicity in their genes. *Moscow Univ. Biol. Sci. Bull*. 2010. V. 65. P. 133–135.
17. Chaley M., Kutyrkin V. Profile-Statistical Periodicity of DNA Coding Regions. *DNA Res.* 2011. V. 18. P. 353–362.
18. Кутыркин В.А., Чалей М.Б. Распознавание различных уровней в организации кодирования генетической информации. *Вестник МГТУ им. Н.Э.Баумана. Серия Естественные науки*. 2011. Спец. выпуск №2 «Математическое моделирование». С. 200–215.
19. Кутыркин В.А., Чалей М.Б. Структурные различия кодирующих и некодирующих районов последовательностей ДНК генома человека. *Вестник МГТУ им. Н.Э.Баумана*. *Сер. Естественные науки*. 2012. Спец.выпуск № 3 «Математическое моделирование». С. 146-157.
20. FTG: Fast Fourier Transform based GENE Prediction Server. URL: http://www.imtech.res.in/raghava/ftg/ (дата обращения 31.07.2013).

Материал поступил в редакцию 16.08.2013, опубликован 24.09.2013.