

УДК: 577.2:519.23

Спектрально-статистический подход к распознаванию скрытой профильной периодичности в последовательностях ДНК

Кутыркин В.А.^{*1}, Чалей М.Б.^{**2}

¹Московский Государственный Технический Университет им. Н.А. Баумана, Москва, 107005, Россия

²Институт математических проблем биологии, Российская академия наук, Пущино, Московская область, 142290, Россия

Ранее, авторы настоящей работы ввели понятие скрытой профильной периодичности в последовательностях ДНК. В журнале «Математическая биология и биоинформатика» за 2013 год авторами была опубликована статья, в которой был изложен метод распознавания такого типа скрытой периодичности. Этот метод основан на применении спектрально-статистического подхода (2С-подхода). В своей статье авторы показали эффективность разработанного ими метода и провели его сравнение с известными методами, в частности, с методом информационного разложения (ИР-методом). В ответ на публикацию этой статьи, в журнал был представлен комментарий, авторами которого являются разработчики ИР-метода. Комментарий посвящён критике метода распознавания скрытой профильной периодичности, основанного на 2С-подходе. Из этого комментария следует, что авторы критики либо невнимательно читали текст нашей статьи, либо не знакомы с некоторыми понятиями использованного в ней математического аппарата. Поэтому мы сочли необходимым более подробно изложить теоретические основы метода и обратить внимание на ключевые моменты его применения, которые не были поняты критиками 2С-подхода. В настоящей работе показано, что, в отличие от 2С-подхода, ИР-метод не является самодостаточным методом распознавания скрытой периодичности в последовательностях ДНК и может выдавать некорректные результаты.

Ключевые слова: *скрытая периодичность, спектрально-статистический подход, скрытая профильная периодичность.*

1. ВВЕДЕНИЕ

В работе авторов [1], посвящённой методам распознавания скрытой периодичности в последовательностях ДНК, было проведено сравнение спектрально-статистического подхода (2С-подхода) с другими методами оценки периода скрытой периодичности. Оценки на основе 2С-подхода отличались от оценок, полученных ранее [2] с помощью метода информационного разложения (ИР-метода) [3]. В ответ на работу [1] появилась публикация [4], в которой 2С-подход, как показано в настоящей работе, подвергается математически некорректной критике, на основе которой, в частности, отвергается значимость нового понятия профильной периодичности. Настоящая работа является ответом авторов 2С-подхода на критическую публикацию [4].

* vkutyркиn@yandex.ru

** maramaria@yandex.ru

Авторы публикации [4] отмечают, что 2С-подход анализировал исходные последовательности ДНК из GenBank, в то время как ИР-метод применялся к предварительно выравненным последовательностям [2]. При этом процесс выравнивания не является частью ИР-метода. В настоящей работе с помощью 2С-подхода сравниваются результаты распознавания скрытой периодичности для четырёх последовательностей ДНК из работы [2, табл. 4] с использованием приведенного авторами [2] выравнивания и без него. Результаты проведённого анализа говорят о том, что ИР-метод не является методом распознавания скрытой периодичности, и полученные с его помощью оценки длины периода, как и оценки, выводимые на основе Фурье-спектров, должны дополнительно подтверждаться другими методами.

Для доказательства эффективности 2С-подхода и некорректности его критики в настоящей работе более подробно излагается его теоретическое обоснование и практическая значимость. Показано, что введённый новый тип периодичности в нуклеотидных последовательностях, названный скрытой профильной периодичностью, является естественным обобщением понятия размытых тандемных повторов.

Анализ работ авторов публикации [4] показал, что ими не было описано ни одной новой математической модели скрытой периодичности. Поэтому рассуждения авторов об общих свойствах скрытой периодичности (особенно о наличии различных периодов скрытой периодичности в одной последовательности ДНК) ничем не обоснованы.

Разбор критики, содержащейся в публикации [4], помимо искажений утверждений из ранее опубликованных работ авторов настоящей статьи [5-7], выявил непонимание авторами [4] стандартных методов математической статистики и основных понятий теории распознавания образов. Показано, что авторы критики неверно трактуют процесс распознавания скрытой периодичности в последовательностях ДНК и выявленную в последовательностях кДНК двухуровневую организацию кодирования [7], ошибочно обвиняя авторов настоящей работы в некорректном заимствовании замечаний из работы [3] на стр.204.

В разделе 4 авторы подробно отвечают на все критические замечания публикации [4], показывая их несостоятельность.

2. РАСПОЗНАВАНИЕ СКРЫТОЙ ПЕРИОДИЧНОСТИ В ПОСЛЕДОВАТЕЛЬНОСТЯХ ДНК

2.1. Общие принципы распознавания скрытой периодичности в текстовых строках

Для того чтобы в текстовых строках распознавать скрытую периодичность, необходимо иметь общую модель (общий эталон) их совершенной периодичности. Такая эталонная модель определяет тип скрытой периодичности в анализируемых текстовых строках (последовательностях нуклеотидов ДНК). При распознавании скрытой периодичности определяется конкретный эталон совершенной периодичности, индуцированный повторением соответствующего паттерна периодичности, и показывается, что к этому эталону близка, в некотором предварительно определённом смысле, анализируемая текстовая строка. В этом случае признаётся, что в анализируемой строке существует скрытая периодичность указанного типа, представленная найденной оценкой паттерна периодичности, который индуцирует конкретный эталон периодичности. Следует отметить, что можно предложить много типов скрытой периодичности при условии, что для них определены математически строгие эталонные модели совершенной периодичности, каждая из которых индуцируется своим типом паттерна периодичности. При распознавании скрытой периодичности заданного типа можно использовать предварительную процедуру оценки размера паттерна периодичности, т. е. периода скрытой периодичности в

анализируемой строке. Для этой цели используются различные косвенные показатели, которые могут указывать на существование скрытой периодичности в анализируемой строке, но достоверно не гарантируют её наличия. Следовательно, найденная с помощью таких косвенных показателей, оценка скрытого периода должна быть дополнительно подтверждена в процессе распознавания паттерна скрытой периодичности рассматриваемого типа. Может оказаться, что такая оценка скрытого периода не подтверждается дальнейшей процедурой распознавания, что демонстрируется в настоящей работе.

Следует отметить, что Фурье-спектры и аналогичные им косвенные показатели наличия скрытой периодичности в последовательности ДНК, как правило, имеют сложную картину со многими максимумами. Однако, на этом основании нельзя делать вывод о существовании в анализируемой последовательности нескольких скрытых периодов.

Исторически первый тип периодичности был определён в виде совершенных тандемных повторов. Реальные последовательности ДНК с таким скрытым типом периодичности были названы размытыми тандемными повторами [8]. В этом случае предполагается, что размытый тандемный повтор был получен из совершенного путём небольшой доли (не более 20%) повреждений нуклеотидов в его позициях в виде замен (mismatches), вставок и делеций (indels). Если среди повреждений нет вставок и делеций, то такой размытый повтор называется нечётким (fuzzy) тандемным повтором. Совершенный тандемный повтор задаётся последовательным повторением своего паттерна, поиск которого может осуществляться методами динамического программирования [8], по результатам которых производится выравнивание анализируемой последовательности. В этом случае паттерн периодичности совершенного тандемного повтора корректно распознаётся в виде так называемого консенсус-паттерна.

Следует отметить, что при использовании методов динамического программирования предварительной оценки скрытого периода (размера паттерна) не производится. Использование такой оценки может повысить эффективность выравнивания. Например, в работе [9] предварительная оценка изменила размер паттерна, найденного методом TRF [8], с 19 до 35 нуклеотидов (нукл.), что позволило получить выравнивание со значительно меньшим количеством повреждений. В работе [9] также предложен критерий, использующий стандартную статистику и позволяющий принять или отвергнуть гипотезу об однородности анализируемой последовательности на испытываемом периоде (тест-периоде). Если на основании этого критерия гипотеза об однородности текстовой строки отвергается, то в диапазоне тест-периодов строки производится анализ спектра другой статистики, названной уровнем сохранности букв паттерна, на каждом из тест-периодов. Исследование базы TRDB [10] показало, что в размытых тандемных повторах максимальное значение спектра этой статистики во многих случаях достигается на скрытом периоде размытого тандемного повтора. Такой способ получения предварительной оценки скрытого периода в последовательностях, несущественно отличающихся от размытых тандемных повторов, был назван спектрально-статистическим подходом (2С-подходом) [9, 11, 12]. На основе 2С-подхода была создана база данных HeteroGenome [13], подвергнутая необоснованной критике авторами публикации [4] в разделе 8.

В дальнейшем 2С-подход получил развитие в работах [5–7] с введением нового типа периодичности, основанного на строгой математической модели профильной периодичности (профильности). Чтобы показать несостоятельность критики 2С-подхода авторы вынуждены ещё раз и более подробно изложить его теоретическое обоснование и основные положения.

2.2. Определение эталона профильной периодичности и соответствующего ему типа скрытой периодичности

В качестве основы 2С-подхода [5, 6, 9] была предложена модель профильной периодичности, обобщающая модель совершенных текстовых тандемных повторов. В этой модели конкретным эталоном скрытой профильной периодичности является профильная строка, являющаяся специальной случайной строкой из независимых случайных букв, исходный упорядоченный текстовый алфавит которых имеет вид $A = \langle a_1, \dots, a_K \rangle$. В последовательностях ДНК размер текстового алфавита $K = 4$, и этот алфавит имеет вид $A = \langle a_1, \dots, a_4 \rangle = \langle a, t, g, c \rangle$.

В общем случае случайную строку длиной n из независимых случайных букв можно рассматривать как схему из n независимых испытаний случайных величин с K исходами в виде букв алфавита A . Каждая из этих случайных величин, называемая случайной буквой, обозначается выражением $Chr(\mathbf{p})$ и определяется столбцом вероятностей $\mathbf{p} = (p^1, \dots, p^K)^T$, где p^i – вероятность появления i -той буквы алфавита A в рассматриваемом испытании. Следовательно, такую схему из n независимых испытаний можно представлять формальной строкой $Str_n(\boldsymbol{\pi}) = Chr(\mathbf{p}_1) \dots Chr(\mathbf{p}_n)$, являющейся n -мерной случайной величиной, где $Chr(\mathbf{p}_j)$ – случайная буква, характеризующая j -тое испытание в рассматриваемой схеме из n независимых испытаний. Такая случайная строка однозначно индуцируется матрицей $\boldsymbol{\pi} = (\mathbf{p}_1, \dots, \mathbf{p}_n) = (\pi_{ij}^K)$, называемой n -профильной матрицей или профильной матрицей строки $Str_n(\boldsymbol{\pi})$. Для этой случайной строки, согласно работам [5, 6, 9], любое целое число L из диапазона $1, \dots, L_{\max} \approx \frac{n}{5K}$ ($L_{\max} \leq \frac{n}{5K}$) называется её тест-периодом.

Пусть L – тест-период строки $Str = Str_n(\boldsymbol{\pi})$, $0 \leq M < L$ и $Str_n(\boldsymbol{\pi}) = Str_L(\boldsymbol{\pi}_1) \dots Str_L(\boldsymbol{\pi}_m) Str_M(\boldsymbol{\pi}_{m+1})$ – разложение строки Str на подстроки длины L , т.е. $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_m, \boldsymbol{\pi}_{m+1})$. Тогда, если $M = 0$ ($\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_m)$ и строка $Str_M(\boldsymbol{\pi}_{m+1})$ – пустая),

матрица $\Pi_{Str}(L) = \frac{1}{m} \sum_{i=1}^m \boldsymbol{\pi}_i$ называется L -профильной матрицей строки $Str = Str_n(\boldsymbol{\pi})$.

Если $M \neq 0$, то в матрицу $\Pi_{Str}(L)$ вносятся соответствующие поправки. Таким образом, для строки $Str = Str_n(\boldsymbol{\pi})$ вводится профильно-матричный спектр Π_{Str} , определённый на каждом тест-периоде. Если $\boldsymbol{\pi}_1 = \dots = \boldsymbol{\pi}_m = \boldsymbol{\pi}_0$ и $\boldsymbol{\pi}_0 = (\boldsymbol{\pi}_{m+1}, \boldsymbol{\pi}_{01})$, то строка $Str_n(\boldsymbol{\pi})$ называется L -профильной строкой со случайным паттерном периодичности $Str_L(\boldsymbol{\pi}_0)$. В этом случае для строки $Str_n(\boldsymbol{\pi})$ 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. Кроме того, число L называется периодом этой строки (предполагается, что число L не является обертоном).

Если $L = 1$, то профильная строка $Tdm_1(\boldsymbol{\pi}_0, n) = Tdm_1(\mathbf{p}, n) = \underbrace{Chr(\mathbf{p}) \dots Chr(\mathbf{p})}_{n \text{ раз}}$ будет

называться *однородной строкой*, так как её период равен единице. Однородную строку длиной n можно рассматривать как схему из n независимых испытаний единственной случайной величины с K исходами в виде букв алфавита $A = \langle a_1, \dots, a_K \rangle$, т.е. как полиномиальную схему с K исходами.

Замечание 2.1. Букву $a_i \in A$ можно отождествить со случайной буквой, все компоненты столбца вероятностей (частот) которой – нулевые, за исключением i -

той единичной компоненты. Такая случайная буква будет называться текстовой буквой. Следовательно, любую текстовую строку в алфавите A можно отождествлять с соответствующей специальной случайной строкой той же длины. Такая случайная строка также будет называться текстовой строкой. В частности, такая текстовая строка $Tdm_1(p, n)$ состоит только из единственной текстовой буквы алфавита A , повторенной в этой строке n раз. Таким образом, в 2С-подходе эталонная модель совершенного текстового тандемного повтора является частным случаем эталонной модели в виде совершенного случайного тандемного повтора (профильной строки), паттерн которого состоит из независимых случайных букв в исходном алфавите текстового тандемного повтора. Кроме того, совершенный текстовый тандемный повтор, согласно сделанному выше отождествлению, является профильной строкой. Поэтому, согласно такому отождествлению, допустимо выражение: пусть профильная строка является текстовой. ►

Замечание 2.2. Как и для любой случайной величины, для профильной строки $Str = Tdm_L(\pi_0, n)$ можно провести n соответствующих её схеме испытаний. В результате этих испытаний будет получена тестовая строка, называемая реализацией строки $Str = Tdm_L(\pi_0, n)$. В частности, при таком подходе нечёткий тандемный повтор можно рассматривать как результат реализации профильной строки, главная профильная матрица которой близка к главной матрице совершенного текстового тандемного повтора с тем же периодом и с компонентами из нулей и единиц. В общем случае в результат реализации строки $Str = Tdm_L(\pi_0, n)$ может добавляться малое количество вставок и делеций букв алфавита A . Если при реализации строки $Str = Tdm_L(\pi_0, n)$ получена тестовая строка str , то для строки str можно поставить вопрос о наличии в ней скрытой профильной периодичности. При достаточно большой длине строк $Str = Tdm_L(\pi_0, n)$ и str ($L < L_{\max}$, $L_{\max} \leq \frac{n}{5K}$) их профильно-матричные спектры с большой долей вероятности будут статистически близки. Это свойство используется в 2С-подходе для распознавания скрытой периодичности в текстовых строках (последовательностях ДНК). Если же профильная строка $Str = Tdm_L(\pi_0, n)$, согласно ранее сделанному в Замечании 2.1 отождествлению, – близка к текстовому совершенному тандемному повтору, то, согласно принятой в литературе терминологии, её реализация называется размытым тандемным повтором. ►

Подводя итоги настоящего подраздела, отметим, что, согласно 2С-подходу, для распознавания в тестовой строке скрытой профильной периодичности необходимо найти такую профильную строку, для которой текстовая строка может рассматриваться как реализация найденной профильной строки. Такой подход является естественным продолжением методов распознавания размытых тандемных повторов.

2.3. Использование статистических спектров текстовых строк для оценки периода скрытой профильной периодичности

Пусть $Str = Str_n(\pi^*)$ – случайная строка из n независимых случайных букв в исходном алфавите $A = \langle a_1, \dots, a_K \rangle$, индуцированная её n -профильной матрицей $\pi^* = (p_1, \dots, p_n)$, и $p = (p^1, \dots, p^K)^T = \frac{1}{n} \sum_{j=1}^n p_j = \Pi_{Str}(1)$ – вектор вероятностей (частот) встречаемости соответствующих букв алфавита A в строке $Str = Str_n(\pi^*)$. Тогда для

каждого тест-периода λ строки Str λ -профильная матрица $\Pi_{Str}(\lambda) = (\pi_j^i)_\lambda^K$ определяет число $\Psi_t(\lambda)$, для которого:

$$\Psi_t(\lambda) = \Psi_t(\Pi_{Str}(\lambda), \Pi_{Str}(1), n) = \frac{n}{\lambda} \sum_{j=1}^{\lambda} \sum_{i=1}^K (\pi_j^i - p^i)^2 / p^i. \quad (1)$$

Таким образом, для строки $Str = Str_n(\pi^*)$ водится функция Ψ_t , определённая на тест-периодах этой строки и называемая её главным спектром.

Для неоднородной профильной строки $Str = Tdm_L(\pi_0, n)$ периода L , в частности, для текстового тандемного повтора с длиной периода L ($L \neq 1$), математически строго доказывается следующее утверждение.

Теорема 2.1. *Главный спектр Ψ_t , определённый формулой (1), для неоднородной профильной строки имеет период, равный длине паттерна периодичности этой строки, и максимальные значения этого спектра Ψ_t достигаются только на тест-периодах, кратных размеру L ($L \neq 1$) паттерна периодичности строки $Str = Tdm_L(\pi_0, n)$. Для однородной строки, согласно формуле (1), её главный спектр принимает только нулевые значения. ►*

Для наглядной иллюстрации теоремы 2.1 на рис. 1 показаны графики главных спектров текстового совершенного тандемного повтора (рис. 1,а) и профильной строки (рис. 1,б) в исходном алфавите последовательностей ДНК, где профильная строка такова, что практически все её реализации не являются размытыми тандемными повторами.

По аналогии с формулой (1) для текстовой строки str длиной n вводится главный спектр Ψ , который на тест-периоде $\lambda < L_{\max} \leq \frac{n}{5K}$ принимает значение:

$$\Psi(\lambda) = \Psi(\Pi_{str}(\lambda), \Pi_{str}(1), n) = \frac{n}{\lambda} \sum_{j=1}^{\lambda} \sum_{i=1}^K (\pi_j^i - p^i)^2 / p^i, \quad (2)$$

где $\Pi_{str}(\lambda) = (\pi_j^i)_\lambda^K$ — λ -профильная матрица строки str и $\Pi_{str}(1) = (p^1, \dots, p^K)^T$. Для реализаций однородной профильной строки длиной n , в соответствии с критерием согласия Пирсона [14, С. 483], справедливо соотношение: $\Psi(\lambda) = \Psi_h(\lambda) \sim \chi_{(K-1)(\lambda-1)}^2$ (распределение $\Psi_h(\lambda)$ статистически эквивалентно распределению $\chi_{(K-1)(\lambda-1)}^2$), где χ_N^2 — распределение χ^2 с N степенями свободы.

При построении графика главного спектра Ψ текстовой строки str , полученной при реализации профильной строки $Str = Tdm_L(\pi_0, n)$, идеальный характер графика главного спектра Ψ_t строки $Str = Tdm_L(\pi_0, n)$ будет искажён. Для иллюстрации такого искажения на рис. 1 показаны графики главных спектров для реализаций однородной (1-профильной) строки (рис. 1,в) и 9-профильной строки (рис. 1,г). Отметим, что график главного спектра исходной 9-профильной строки приведён на рис. 1,б и частоты букв в исходной однородной строке совпадают с частотами букв этой 9-профильной строки. При этом рассматриваемая реализация 9-профильной строки является последовательностью кДНК генома человека из базы данных KEGG [15]. Кроме того, жирной линией на рисунках 1,в и 1,г показан график зависимости правого критического значения $\chi_{crit}^2(N, \alpha)$ для χ_N^2 -распределения на уровне значимости $\alpha = 0.05$ от тест-периода λ , где $N = (K-1)(\lambda-1)$. Демонстрация этого графика на рисунках 1,в и 1,г обусловлена тем, что для однородной (homogeneous) строки длиной n значения главного спектра Ψ_h её реализаций на тест-периоде $\lambda < L_{\max}$, где

$L_{\max} \leq \frac{n}{5K}$, имеют, как упоминалось выше, χ^2_N -распределение с $N = (K-1)(\lambda-1)$ степенями свободы. Поэтому в 2С-подходе [1, 5–7, 16] для проверки гипотезы об однородности строки str на уровне значимости $\alpha = 0.05$ (на $\alpha \cdot 100\% = 5\%$ -ом уровне значимости) используется спектр D_1 , который на тест-периоде λ (см. формулу (2)) принимает значение:

$$D_1(\lambda) = \Psi(\lambda) / \chi^2_{crit}((K-1)(\lambda-1), \alpha). \quad (3)$$

Если значение $D_1(\lambda) > 1$, то, в соответствии с критерием согласия [14, С. 483], на тест-периоде λ в анализируемой строке проявляется неоднородность. Поэтому для текстовой строки спектр D_1 называется её спектром отклонения от однородности.

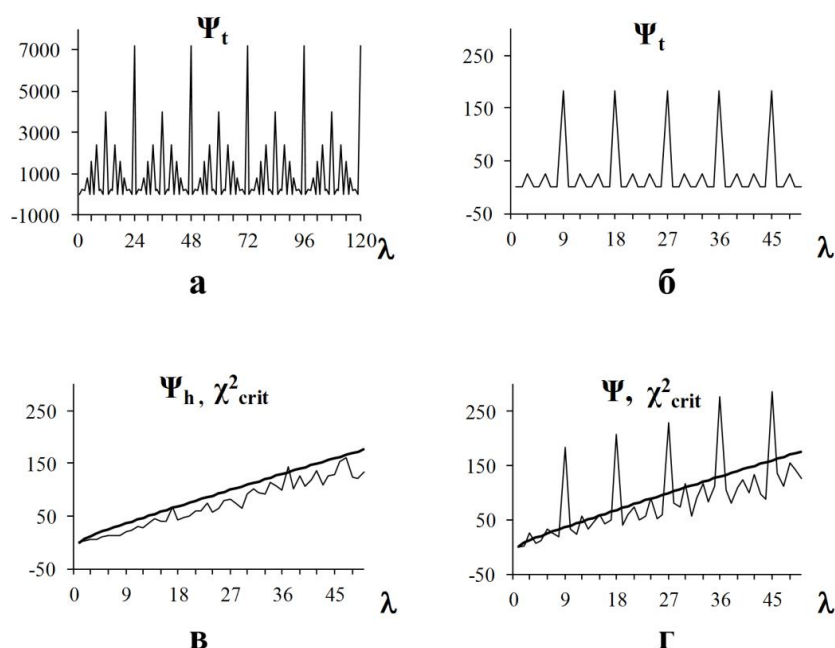


Рис. 1. Главные спектры профильных строк: **а)** совершенный тандемный повтор из 100 копий текстового паттерна «atgcaattggcсааatttggggccс»; **б)** 9-профильная строка, главная профильная матрица которой оценивается по её реализации (последовательности нуклеотидов, совпадающими с соответствующими частотами в последовательности кДНК (KEGG, hsa:338872); **в)** однородная строка с частотами встречаемости нуклеотидов, совпадающими с соответствующими частотами в последовательности кДНК (KEGG, hsa:338872, 1002 нукл.). Вместе с главными спектрами профильных строк на рисунках 1, в и 1, г жирной линией показан график зависимости значений χ^2_{crit} от тест-периода λ , для которых вероятность $P(\chi^2_N \geq \chi^2_{crit}(N, 0.05)) = 0.05$, где $N = (K-1)(\lambda-1)$ и $K = 4$.

Замечание 2.3. Для неоднородной профильной строки длиной n вероятностное распределение значений главного спектра её реализаций на тест-периоде λ не совпадает с χ^2 -распределением, имеющим $N = (K-1)(\lambda-1)$ степеней свободы. По сравнению с этим χ^2 -распределением, реальное распределение значений главного спектра реализаций неоднородной строки индуцирует существенно большую, чем $\alpha = 0.05$, вероятность превышения критического уровня $\chi^2_{crit}((K-1)(\lambda-1), \alpha)$. Поэтому в спектрах D_1 реализаций неоднородной строки будут наблюдаться тест-

периоды, на которых значения спектра D_1 превышают единицу. В этом случае текстовые строки-реализации будут называться неоднородными. ►

На рис. 2,а показан спектр D_1 отклонения от однородности, полученный из главного спектра Ψ , показанного на рис. 1,г. Согласно сделанному относительно спектра D_1 Замечанию 2.3, последовательность кДНК из генома человека (KEGG, hsa:338872) рассматривается как неоднородная.

На рис. 2 ещё раз приведены графики главных спектров профильной строки (рис. 2,г) и её реализации (рис. 2,в), являющейся последовательностью кДНК (KEGG, hsa:338872). Из этих рисунков следует, что отличие главных спектров профильной строки и её реализации, практически, имеет вид функции, линейно зависящей от тест-периодов этих строк. По аналогии с $\chi^2_{(K-1)(\lambda-1)}$ -распределением с увеличением λ возрастает число степеней свободы в вероятностном распределении значений главного спектра Ψ реализаций исходной профильной строки. Чтобы нивелировать такой рост для реализации str вводится спектр C , для которого:

$$C(\lambda) = \Psi(\lambda) - M(\chi^2_{(K-1)(\lambda-1)}) = \Psi(\lambda) - (K-1)(\lambda-1), \quad (4)$$

где $M(\chi^2_N) = (K-1)(\lambda-1)$ – математическое ожидание χ^2 -распределения с N степенями свободы. В дальнейшем для анализируемой текстовой строки этот спектр будет называться *характеристическим*. Для анализируемой реализации str график этого спектра показан на рис. 2,б.

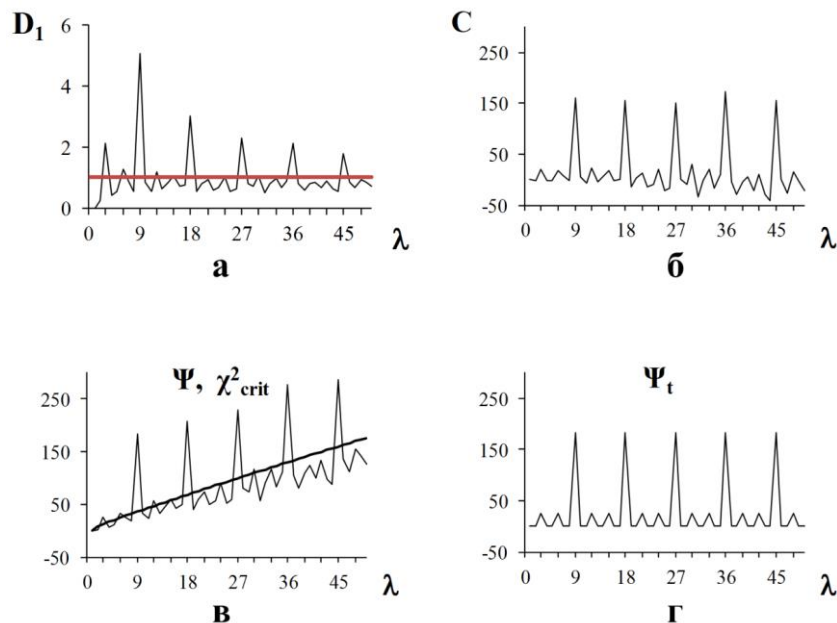


Рис. 2. Спектры 2С-подхода для последовательности кДНК белка – фактора некроза опухоли из генома человека (KEGG, hsa:338872, 1002 нукл.): **а)** спектр отклонения от однородности (см. формулу (3)); **б)** характеристический спектр (см. формулу (4)); **в)** главный спектр (см. формулу (2)) представлен тонкой линией, график $\chi^2_{crit}((K-1)(\lambda-1), 0.05)$ – жирной линией ($P(\chi^2_N \geq \chi^2_{crit}(N, 0.05)) = 0.05$); **г)** главный спектр профильной строки, реализацией которой можно считать анализируемую последовательность кДНК.

Сравнение характеристического спектра (рис. 2,б) реализации исходной 9-профильной строки и главного спектра (рис. 2,г) этой 9-профильной строки демонстрирует их наглядное сходство. Проведённое исследование показало, что такое сходство характерно для реализаций профильных строк, близких к текстовым

совершенным тандемным повторам. Поэтому на такое сходство и опирается 2С-подход при распознавании скрытой профильной периодичности в текстовых строках. При этом, как правило, вследствие своего построения, для неоднородных текстовых строк-реализаций максимальное значение характеристического спектра достигается, с учётом небольших случайных погрешностей, на периоде скрытой профильной периодичности. Такие свойства характеристического спектра используются в 2С-подходе для оценки периода скрытой профильной периодичности в текстовых строках, в частности, в последовательностях ДНК.

Замечание 2.4. Для оценки периода скрытой профильной периодичности в анализируемой текстовой строке предлагается использовать следующее правило. В диапазоне допустимых тест-периодов строки выделяется тот, на котором достигается первое ярко выраженное максимальное значение её характеристического спектра C . Этот выделенный тест-период рассматривается в качестве оценки скрытого периода профильной периодичности только в том случае, когда в диапазоне тест-периодов строки значения спектра отклонения от однородности D_1 превышают единицу на этом выделенном тест-периоде и его обертонах, т.е. когда анализируемая строка признаётся неоднородной. ►

На рис. 2 приведены спектр D_1 отклонения от однородности (рис. 2,а), полученный из главного спектра Ψ , показанного на рис. 2,в (рис. 1,г), и соответствующий ему характеристический спектр C (рис. 2,б). Согласно принятому в Замечании 2.4 правилу, для анализируемой последовательности кДНК из генома человека (KEGG, hsa:338872) предлагается оценка скрытого периода профильной периодичности в 9 нукл.

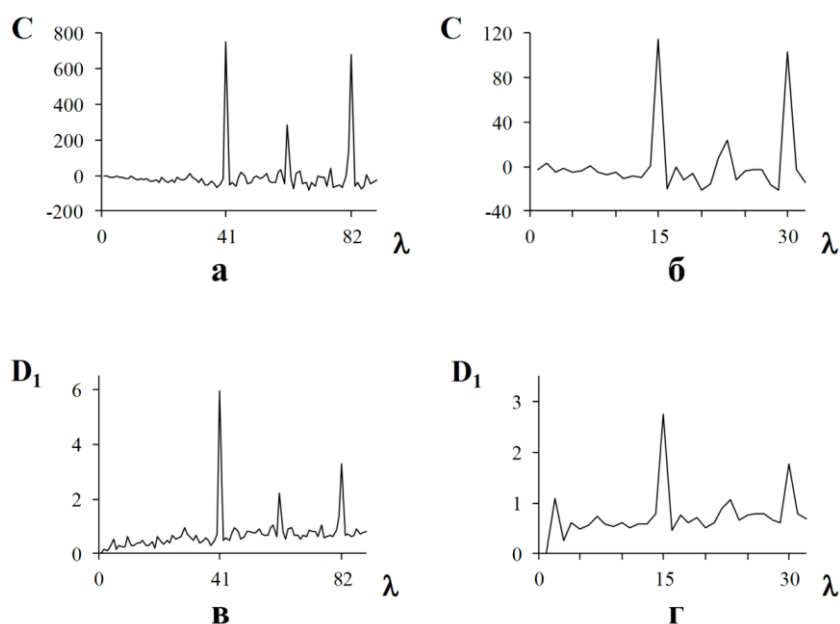


Рис. 3. Характеристический спектр (а) и спектр отклонения от однородности (в) размытого тандемного повтора на хромосоме I из генома человека (TRDB, индексы: 2956 – 4750, размер паттерна 41 нукл., % несовпадений между копиями – 13, % вставок и делеций между копиями – 5). Характеристический спектр (б) и спектр отклонения от однородности (г) размытого тандемного повтора на хромосоме XI из генома мыши *M. musculus* (TRDB, индексы: 6611441 – 6612079, размер паттерна 15 нукл., % несовпадений между копиями – 15, % вставок и делеций между копиями – 5).

Эффективность правила, сформулированного в Замечании 2.4, была проверена на размытых тандемных повторах из базы TRDB [10]. На рис. 3 приведены характеристические спектры размытых тандемных повторов с периодом 41 нукл. (рис. 3,а) и с периодом 15 нукл. (рис. 3,б). Эти спектры, вместе с соответствующими

спектрами отклонения от однородности D_1 (см. рис. 3, в и г), согласно этому сформулированному правилу, позволяют получить правильные оценки скрытых периодов для рассматриваемых tandemных повторов.

В работах [1, 5, 6] была подтверждена эффективность указанного выше правила для оценки периода скрытой профильной периодичности в неоднородных последовательностях ДНК, не являющихся размытыми tandemными повторами. На рис. 4 показаны примеры характеристических спектров и спектров отклонения от однородности для таких последовательностей. Далее будет показано, что в этих последовательностях выявляется скрытая периодичность с периодами $L = 10$ (рис. 4, а), $L = 84$ (рис. 4, в) и $L = 9$ (рис. 4, д) нуклеотидов, оценки которых соответствуют правилу из замечания 2.4.

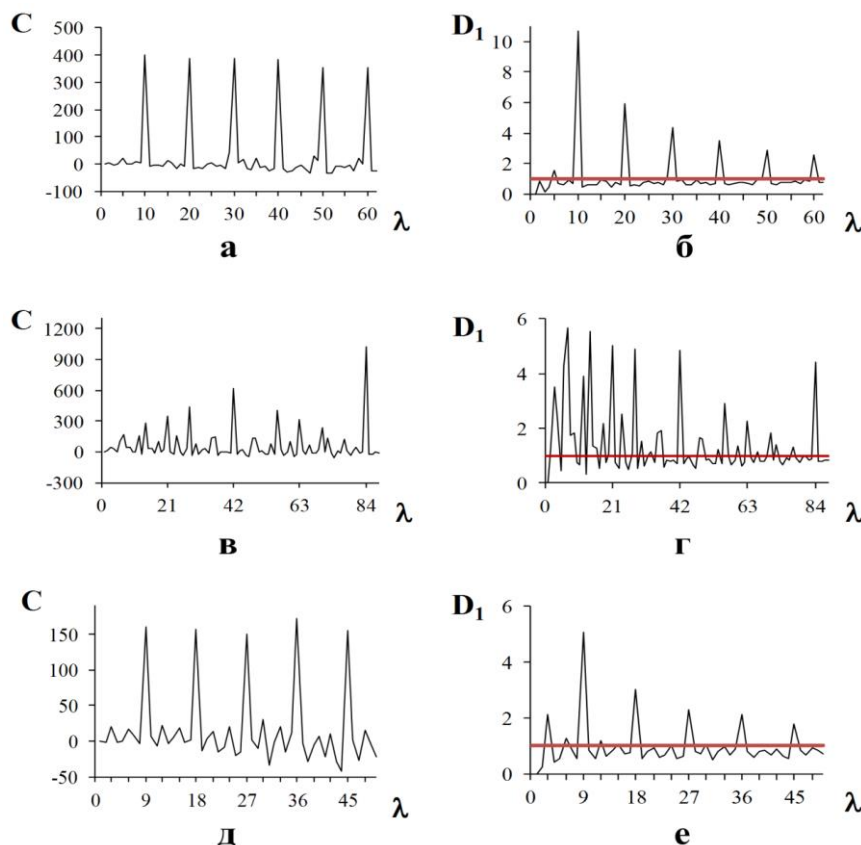


Рис. 4. Характеристические спектры и спектры отклонения от однородности для последовательностей ДНК, не являющихся размытыми tandemными повторами: фрагмент ДНК хромосомы III генома *C. elegans* (HeteroGenome, индексы: 307381–308580, 1200 нукл.) (а,б); кДНК из генома человека для белка 285А «цинковых пальцев» (KEGG, hsa: 26974, 1773 нукл.) (в,г); кДНК из генома человека белка – фактора некроза опухоли (KEGG, hsa:338872, 1002 нукл.) (д,е).

2.4. Оценки паттерна эталона скрытой профильной периодичности на основе критерия согласия

Пусть, согласно правилу, сформулированному в замечании 2.4, для текстовой строки *str* на основании её спектров C (см. формулу (4)) и D_1 (см. формулу (3)) получена оценка скрытого периода профильной периодичности L ($L > 1$). Тогда, по аналогии с главным спектром (см. формулу (2)), для того чтобы проверить является ли тест-период L периодом скрытой профильной периодичности, используется статистика:

$$\Psi_L(\lambda) = \Psi_L(\Pi_{str}(\lambda), \Pi_{Tdm_L}(\lambda), n) = \frac{n}{\lambda} \sum_{j=1}^{\lambda} \sum_{i=1}^K (\pi_j^{*i} - \pi_j^i)^2 / \pi_j^i \sim \chi_{(K-1)(\lambda-1)}^2, \quad (5)$$

где $\Pi_{str}(\lambda) = (\pi_j^{*i})_{\lambda}^K$ и $\Pi_{Tdm_L}(\lambda) = (\pi_j^i)_{\lambda}^K$ – λ -профильные матрицы текстовой строки str и L -профильной строки $Tdm_L = Tdm_L(\Pi_{str}(L), n)$, соответственно. Кроме того, в формуле (5) указано, что вводимая статистика имеет χ^2 -распределение с $(K-1)(\lambda-1)$ степенями свободы при $\frac{n}{\lambda} > 5K$, если строку str можно рассматривать как реализацию строки Tdm_L . В этом случае используются критерии согласия Пирсона [14, С. 483].

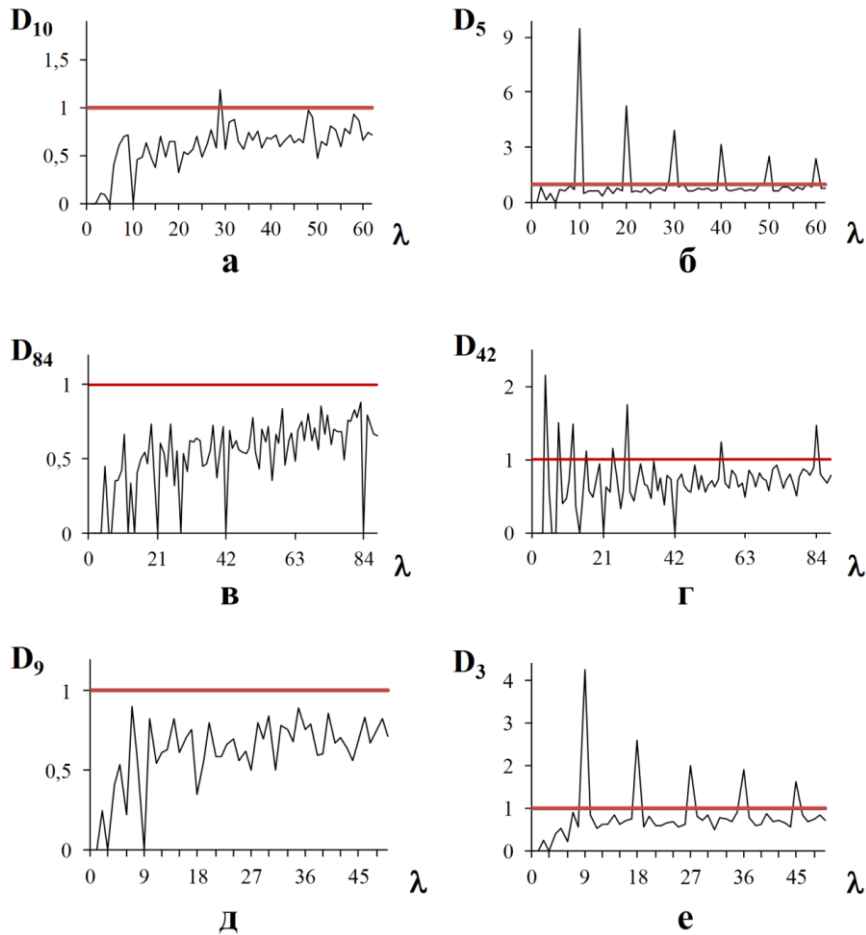


Рис. 5. Спектры отклонения от λ -профильности ($\lambda = 10, 5, 84, 42, 9, 3$) для последовательностей ДНК, не являющихся размытыми тандемными повторами, характеристические спектры и спектры отклонения от однородности для которых показаны на рис. 4. Фрагмент ДНК хромосомы III генома *C. elegans* (HeteroGenome, индексы: 307381–308580, 1200 нукл.) (а,б). кДНК из генома человека для белка 285А «цинковых пальцев» (KEGG, hsa: 26974, 1773 нукл.) (в,г). кДНК из генома человека для белка – фактора некроза опухоли (KEGG, hsa:338872, 1002 нукл.) (д,е).

Используя статистику (5), по аналогии со спектром отклонения от однородности D_1 (см. формулу (3)), вводится спектр D_L отклонения строки str от L -профильности, принимающий на тест-периоде $\lambda < L_{\max} \leq \frac{n}{5K}$ значение:

$$D_L(\lambda) = \Psi_L(\lambda) / \chi_{crit}^2((K-1)(\lambda-1), \alpha), \quad (6)$$

где $\chi_{crit}^2(N, \alpha)$ – критическое значение χ_N^2 -распределения с N степенями свободы на уровне значимости $\alpha = 0.05$. Поэтому, согласно формулам (5) и (6), спектр $\mathbf{D}_L$ может быть использован для проверки гипотезы о существовании в анализируемой текстовой строке скрытой L -профильности, в соответствии со следующим правилом.

Замечание 2.5. Если в диапазоне допустимых тест-периодов анализируемой строки доля тест-периодов, на которых спектр $\mathbf{D}_L$ более единицы, не превышает величины 0.05, то принимается гипотеза о том, что в строке существует скрытая L -профильность, т.е. её период профильной периодичности равен L ($L \neq 1$). В противном случае, эта гипотеза отвергается. Такой критерий согласия является достаточно жёстким и обеспечивает уровень его значимости не более 0.05. ►

Приведём пример использования сформулированного в замечании 2.5 правила выявления скрытой профильной периодичности. Выше, согласно характеристическим спектрам и спектрам отклонения от однородности, показанным на рис. 4 для трёх последовательностей ДНК, не являющихся размытыми тандемными повторами, были предложены оценки скрытых периодов профильной периодичности в 10, 84 и 9 нукл. Эти оценки подтверждаются с помощью спектров отклонения от соответствующей профильности, показанных на рис. 5. Рис. 5 иллюстрирует работоспособность предложенного 2С-подхода, демонстрируя спектры отклонения $\mathbf{D}_L$ от различных испытываемых λ -профильностей ($\lambda = 10, 5, 84, 42, 9, 3$) для анализируемых последовательностей. Согласно правилу из замечания 2.5, из рисунков 5, а, в и д следует, что в анализируемых последовательностях существуют 10-, 84- и 9-профильности, соответственно.

Замечание 2.6. Результаты анализа текстовой строки str , где, согласно предложенному 2С-подходу, выявлена скрытая L -профильная периодичность, позволяют предложить в качестве оценки паттерна этой периодичности строку $Str_L(\Pi_{str}(L))$ из независимых случайных букв, которая однозначно характеризуется L -профильной матрицей $\Pi_{str}(L)$ этой анализируемой текстовой строки. Таким образом, принимается гипотеза о статистической близости (на уровне значимости $\alpha = 0.05$) текстовой строки str и профильной строки $Tdm_L(\Pi_{str}(L), n)$, являющейся выявленным эталоном профильной периодичности для строки str . ►

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

Пусть, согласно замечанию 2.6, для текстовой строки str принята гипотеза о наличии в ней скрытой L -профильности ($L \neq 1$). Следовательно, строку str можно рассматривать, как реализацию эталонной L -профильной строки $Tdm_L = Tdm_L(\Pi_{str}(L), n)$.

При создании эталона профильной периодичности $Tdm_L = Tdm_L(\Pi_{str}(L), n)$ для анализируемой строки str , согласно Замечанию 2.5, использовался критерий согласия, основанный на статистике $\mathbf{D}_L$ отклонения строки str от L -профильности (6). Но для полученной оценки паттерна скрытой профильной периодичности можно получить дополнительное подтверждение. В работах [1, 5, 6] в качестве такого подтверждения был предложен метод сравнения спектров $\mathbf{D}_1$ и $\mathbf{Th}_L$ отклонения от однородности строк str и Tdm_L , соответственно. Спектр $\mathbf{D}_1$ (см. формулу (3)) выбран как наиболее информативный по отношению к статистической структуре строки str . По аналогии с формулой (3) для спектра $\mathbf{D}_1$, в качестве значения спектра отклонения от

однородности $\mathbf{Th}_L$ в случайной строке Tdm_L на тест-периоде λ , согласно формуле (1), используется формула:

$$Th_L(\lambda) = \Psi_t(\Pi_{Tdm_L}(\lambda), \Pi_{Tdm_1}(\lambda), n) / \chi_{crit}^2((K-1)(\lambda-1), \alpha). \quad (7)$$

Фактически, такой спектр $\mathbf{Th}_L$ из формулы (7) является теоретической реконструкцией спектра $\mathbf{D}_1$ строки str и, в случае корректной оценки паттерна эталона скрытой периодичности для строки str , спектр $\mathbf{Th}_L$ имеет наглядное сходство со спектром $\mathbf{D}_1$. На рис. 6,г приведён спектр теоретической реконструкции спектра отклонения от однородности для последовательности кДНК из генома человека. Наблюдаемое наглядное сходство этой реконструкции с исходным спектром отклонения от однородности $\mathbf{D}_1$ последовательности кДНК (рис. 6,б) подтверждает выявленную скрытую профильную периодичность.

Для подтверждения выявленной скрытой L -профильной периодичности можно использовать и статистическую реконструкцию спектра $\mathbf{D}_1$ 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 $\mathbf{St}_L$. При достаточном количестве повторов периода, в случае корректности оценки паттерна эталона скрытой периодичности для строки str , спектр $\mathbf{St}_L$ имеет наглядное сходство со спектром $\mathbf{D}_1$, что проиллюстрировано на рис. 6,з и рис. 6,е.

2.6. Понятие 3-регулярности в кодирующих районах последовательностей ДНК

Анализ на основе спектров 2С-подхода более 17000 кодирующих последовательностей (кДНК) из генома человека в базе данных KEGG [15] показал, что, с точностью до статистической погрешности, эти последовательности являются не только неоднородными, но и обладают следующим свойством регулярности. Пики в характеристических спектрах этих последовательностей практически регулярно повторяются на тест-периодах, кратных трём нуклеотидам. Для описания такого свойства в работах [5, 7] было введено понятие 3-регулярности (триплетной регулярности) в последовательности ДНК. В работе [16] был разработан количественный критерий выявления 3-регулярности, использующий вычисление соответствующего индекса 3-регулярности. Оказалось, что, согласно разработанному критерию, практически все кДНК (с точностью до статистической погрешности) являются 3-регулярными. При распознавании в этих последовательностях скрытой профильной периодичности, оказалось, что скрытый период практически всегда равен или трём нукл. (~60%), или кратен трём нукл. (~12%). Анализ более 277000 интронов показал, что практически все (с точностью до статистической погрешности) интроны не обладают свойством 3-регулярности. Такие результаты позволили сделать вывод о том, что свойство 3-регулярности обусловлено свойством триплетного кодирования аминокислот. Кроме того, наблюдаемая 3-регулярность практически всех характеристических спектров кДНК не является обязательным признаком наличия в этих последовательностях скрытой профильной периодичности, в том числе и триплетной. Поэтому в работе [16] для 3-регулярных кДНК, где распознаётся профильность, отличающаяся от триплетной, говорится о наличии в таких последовательностях двух уровней кодирования. На рис. 4,д приведён характеристический спектр кДНК с двухуровневой организацией кодирования. Анализ спектров на рисунках 4,д и е, а также спектров для этой же последовательности на

рисунках 5,д и е, показывает, что в ней выявляется только скрытая 9-профильность, хотя, согласно характеристическому спектру на рис. 4,д, эта последовательность кДНК является 3-регулярной.

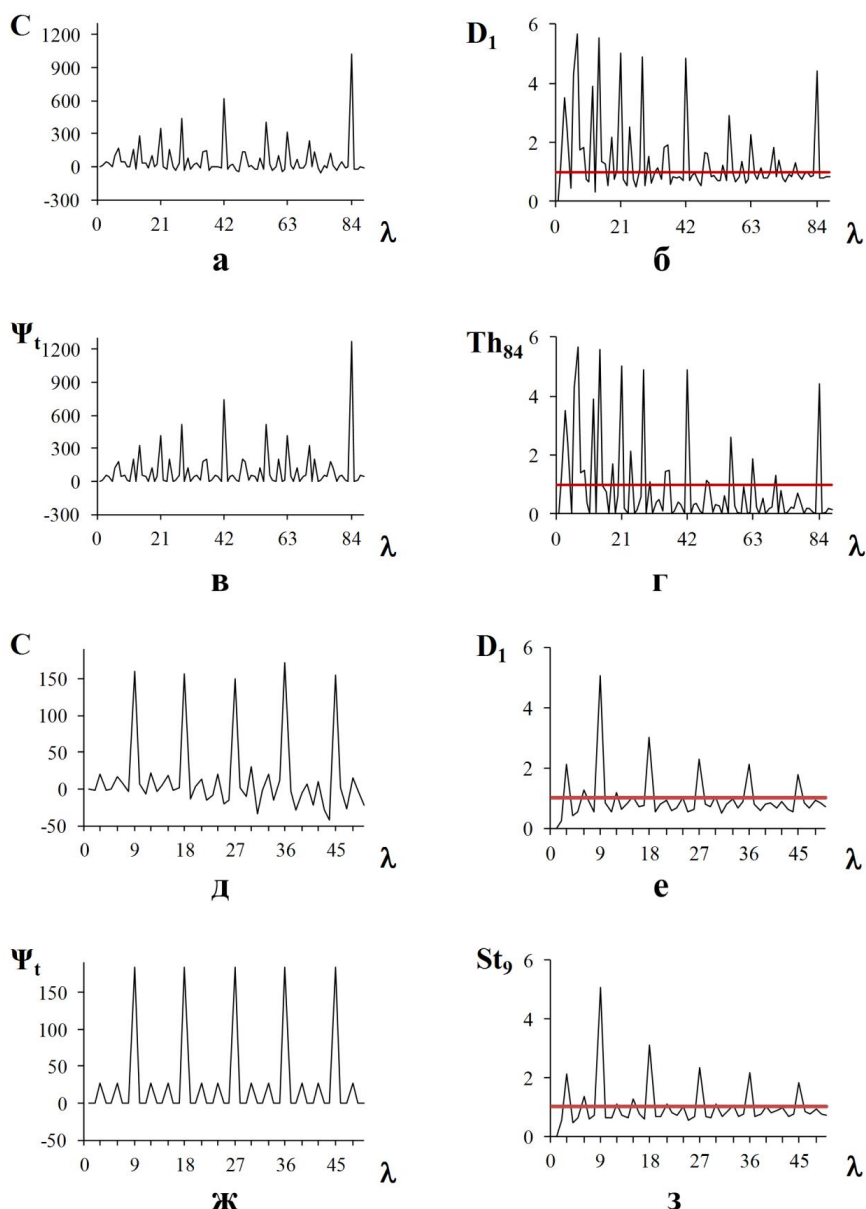


Рис. 6. Подтверждение оценки паттерна эталона скрытой профильной периодичности с помощью спектров 2С-подхода для последовательностей ДНК. Пример для кДНК из генома человека для белка 285А «цинковых пальцев» (KEGG, hsa: 26974, 1773 нукл.) (а-г). Для кДНК из генома человека для белка – фактора некроза опухоли (KEGG, hsa:338872, 1002 нукл.) (д-з).

2.7. Статистический спектр уровня сохранности буквы для оценки периода в последовательностях ДНК, близких к размытым тандемным повторам

Для получения оценки скрытого периода в текстовых строках (последовательностях ДНК), несущественно отличающихся (во фланкирующих районах) от размытых тандемных повторов, в работах [9, 11, 12] было предложено использовать ещё один статистический спектр pl , названный уровнем сохранности букв паттерна.

Профильная матрица $\Pi_{str}(\lambda) = (\pi_{ij}^*)_{\lambda}^K$ текстовой строки str на испытываемом периоде λ определяет значение $pl(\lambda)$ уровня сохранности буквы в виде:

$$pl(\lambda) = \frac{1}{\lambda} \sum_{j=1}^{\lambda} \max\{\pi_j^* : i = 1, \dots, K\}. \quad (8)$$

На рис. 7,а показан спектр **pl** уровня сохранности букв паттерна для совершенного тандемного повтора, график главного спектра Ψ_t которого представлен на рис.1,а. Для сравнения обоих спектров, график спектра Ψ_t повторно приведён на рис. 7,б. Рисунки 7,а и 7,б демонстрируют практическое совпадение структур этих спектров. Следовательно, для размытого тандемного повтора вместо характеристического спектра можно использовать спектр **pl**. Спектр уровня сохранности букв паттерна **pl** имеет преимущество перед характеристическим спектром **C**, поскольку для спектра **pl** область его определения (диапазон допустимых периодов) может быть значительно расширена.

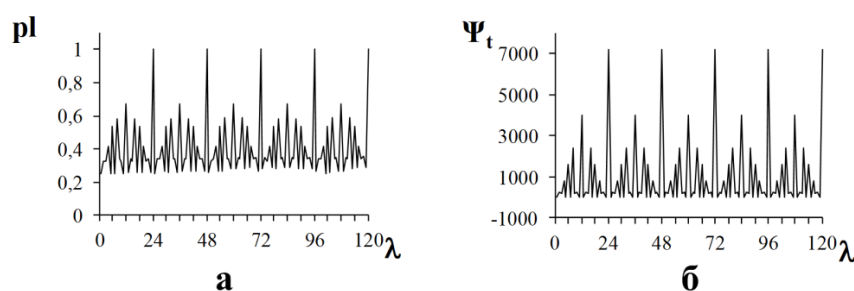


Рис. 7. а) Спектр уровня сохранности букв паттерна (см. формулу (8)) и **б)** главный спектр (см. формулу (1)) последовательности совершенного тандемного повтора из 100 копий текстового паттерна «atgcaattggccaaatttggccccc».

Расширение области определения спектра **pl** обусловлено тем, что этот спектр проверяет совпадение (отличие) букв в позициях испытуемого периода независимо от вида конкретной буквы алфавита повтора. Следовательно, проверку такого совпадения можно рассматривать как биномиальную схему. Поэтому для спектра **pl** максимальное

значение $L_{\max} \leq \frac{n}{5K}$ для допустимых тест-периодов характеристического спектра

можно существенно увеличить вплоть до половины длины анализируемой текстовой строки. Эта возможность используется в базе данных HeteroGenome [13] для выявления последовательностей ДНК, близких к размытым тандемным повторам. При анализе последовательности ДНК в качестве оценки размера паттерна размытого тандемного повтора выбирается тест-период, на котором достигается первое максимальное значение спектра **pl** этой последовательности при условии, что это значение достаточно велико (~ 0.5). Во многих случаях отличие для таких последовательностей от размытого тандемного повтора наглядно проявляется только во фланкирующих районах. Специальное визуальное представление последовательности в базе HeteroGenome [13] помогает уточнить границы размытого тандемного повтора.

Для иллюстрации эффективности использования в базе HeteroGenome [13] введённого формулой (8) спектра **pl** на рис. 8 показаны спектры уровня сохранности буквы и отклонения от однородности для различных размытых тандемных повторов. Оценки размеров паттернов для рассматриваемых размытых тандемных повторов, полученные на основе этих спектров, совпадают с данными базы TRDB [10]. Кроме того, для наглядной иллюстрации достоверности получаемых результатов, на рис. 8 также приведены главные спектры Ψ_t совершенных тандемных повторов, полученных с помощью выравнивания этих размытых повторов.

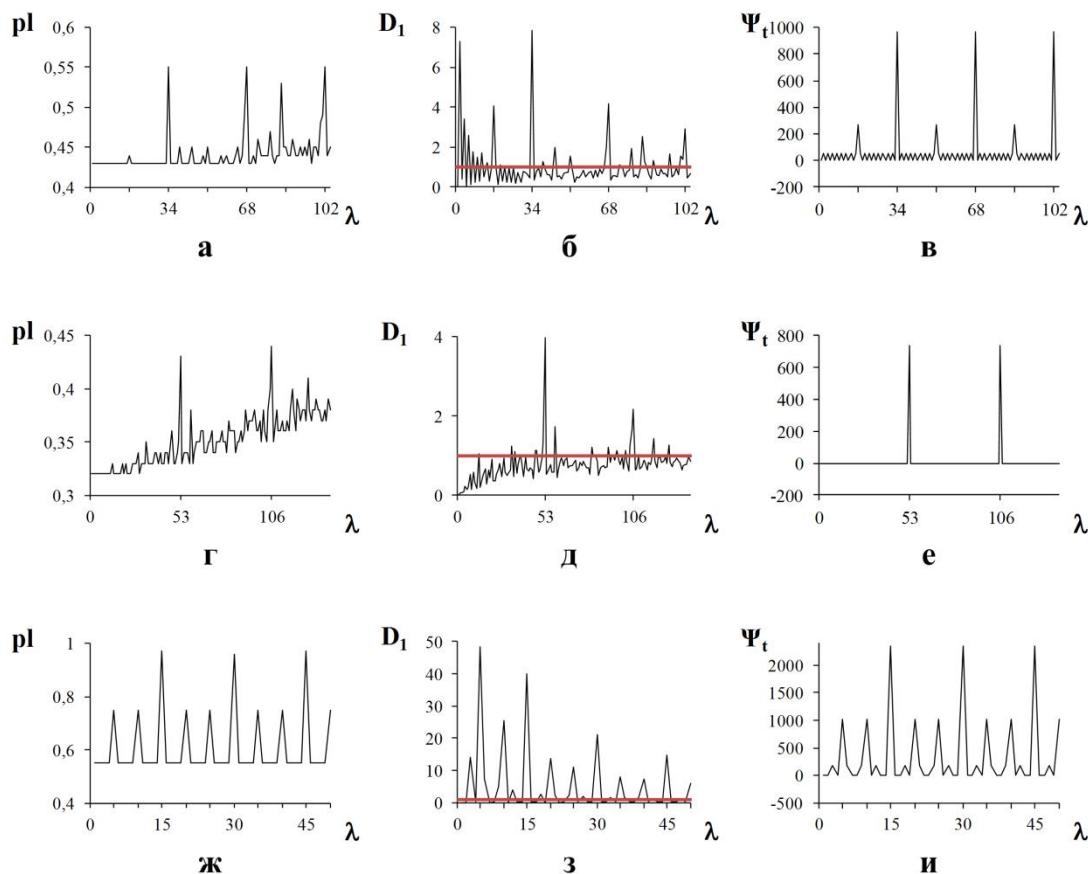


Рис. 8. Спектры 2С-подхода для размытых tandemных повторов. Tandемный повтор на хромосоме 10 из генома мыши *M. musculus* (TRDB, индексы: 25163299–25165343, размер паттерна 34 нукл., % несовпадений между копиями – 4, % вставок и делеций между копиями – 1) (а-в). Tandемный повтор на хромосоме 7 из генома шимпанзе *P. troglodytes* (TRDB, индексы: 159304996–159307732, размер паттерна 53 нукл., % несовпадений между копиями – 6, % вставок и делеций между копиями – 1) (г-е). Tandемный повтор на хромосоме 1 из генома пчелы *A. mellifera* (TRDB, индексы: 2290729–2291645, размер паттерна 15 нукл., % несовпадений между копиями – 4, % вставок и делеций между копиями – 0) (ж-и).

3. СРАВНЕНИЕ СПЕКТРАЛЬНО-СТАТИСТИЧЕСКОГО ПОДХОДА С ДРУГИМИ МЕТОДАМИ ОЦЕНКИ СКРЫТОГО ПЕРИОДА В ПОСЛЕДОВАТЕЛЬНОСТЯХ ДНК

Спектрально-статистический подход (2С-подход) выявления скрытой профильной периодичности полностью удовлетворяет методологии теории распознавания образов. Он позволяет не только получить оценку паттерна скрытой профильной периодичности, но и подтвердить её с помощью определяемого совершенного эталона профильной периодичности, к которому близка, в статистическом смысле, анализируемая последовательность ДНК. Такой подход к распознаванию скрытой профильной периодичности аналогичен подходу, используемому в методе TRF [8] для распознавания размытых tandemных повторов.

Как правило, существующие в настоящее время методы оценки периода скрытой периодичности в последовательностях ДНК, не являющимися размытыми tandemными повторами, нельзя отнести к методам распознавания. Такие методы, как отмечалось во введении, используют косвенные показатели, которые не гарантируют в общем случае существование скрытой периодичности в анализируемых последовательностях. Предлагая ту или иную оценку скрытого периода, эти методы не обосновывают тип скрытой периодичности, соответствующий этой оценке. Далее, с использованием 2С-

подхода, на примере двух методов, использующих косвенные показатели, проводится анализ достоверности оценок скрытых периодов, полученных этими методами.

3.1. Анализ Фурье-спектров

Все приводимые в настоящем разделе Фурье-спектры были получены с помощью программы Genscan [17] на FTG-сервере [18].

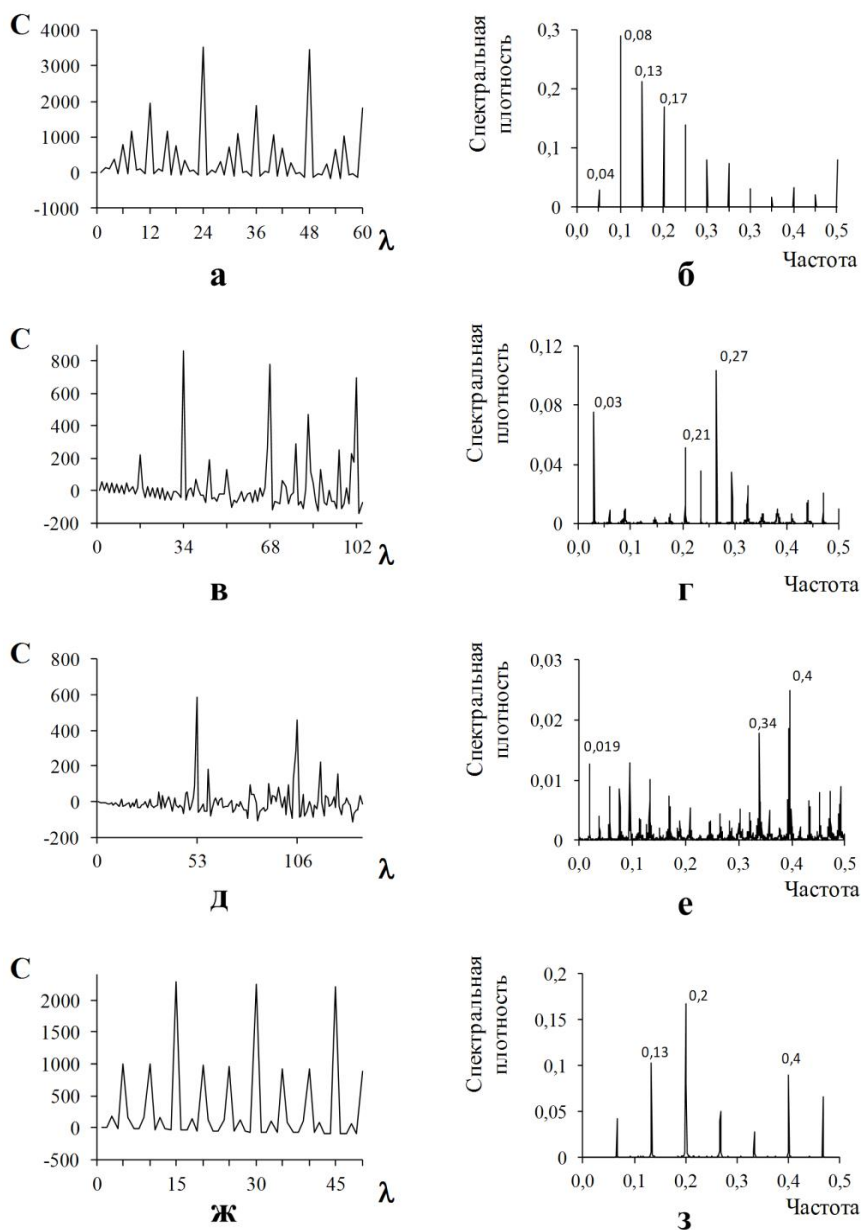


Рис. 9. Характеристические спектры и спектры Фурье для последовательностей tandemных повторов. На спектрах Фурье указаны частоты для первых трёх пиков наибольшей плотности и пиков, соответствующих размеру периода. **а), б)** Совершенный tandemный повтор из 50 копий текстового паттерна «atgcaattggccaaattgggcc». **в), г)** Tandemный повтор на хромосоме 10 из генома мыши *M. musculus* (TRDB, индексы: 25163299–25165343, размер паттерна 34 нукл., % несовпадений между копиями – 4, % вставок и делеций между копиями – 1). **д), е)** Tandemный повтор на хромосоме 7 из генома шимпанзе *P. troglodytes* (TRDB, индексы: 159304996–159307732, размер паттерна 53 нукл., % несовпадений между копиями – 6, % вставок и делеций между копиями – 1). **ж), з)** Tandemный повтор на хромосоме 1 из генома пчелы *A. mellifera* (TRDB, индексы: 2290729–2291645, размер паттерна 15 нукл., % несовпадений между копиями – 4, % вставок и делеций между копиями – 0).

На рисунках в настоящем разделе приведены характеристические спектры и Фурье-спектры для размытых tandemных повторов (рис. 9) и для последовательностей ДНК, не являющихся размытыми tandemными повторами (рис. 10), в которых, как было показано выше, распознаётся профильная периодичность.

Из рис. 9 следует, что те тест-периоды, на которых наблюдаются максимальные пики Фурье-спектров, не соответствуют скрытым периодам в размытых tandemных повторах. Необходимо особо подчеркнуть, что характеристический спектр на рис. 9,а получен для совершенного tandemного повтора с периодом 24 нукл., в то время как максимальный пик Фурье-спектра соответствует тест-периоду в 12 нукл., а пик, соответствующий тест-периоду в 24 нукл. является одним из минимальных пиков.

Как видно на рис.10, оценки периодов для последовательностей ДНК со скрытой профильной периодичностью на основе характеристических спектров и Фурье-спектров совпадают для двух и различаются для одной из рассматриваемых последовательностей.

Приведённый сравнительный анализ показывает, что оценки скрытых периодов, основанные на Фурье-спектрах, требуют дополнительного подтверждения, без которого они могут оказаться недостоверными. Так как Фурье-спектр можно получить для любой последовательности ДНК, то, в общем случае, на его основе, без дополнительного анализа, нельзя делать вывод о наличии скрытой периодичности в анализируемой последовательности.

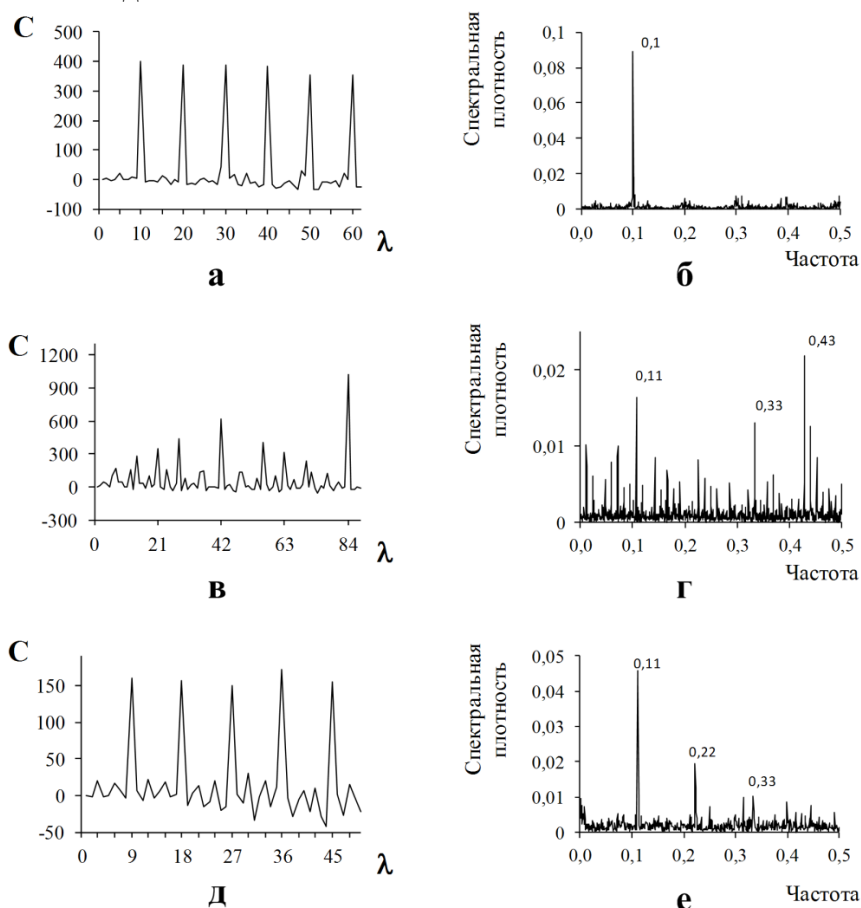


Рис. 10. Характеристические спектры (С) и спектры Фурье для последовательностей ДНК, не являющихся размытыми tandemными повторами. На спектрах Фурье указаны частоты для первых трёх пиков наибольшей плотности и пиков, соответствующих размеру периода скрытой профильной периодичности. а), б) Фрагмент ДНК хромосомы III генома нематоды *C. elegans* (HeteroGenome, индексы: 307381–308580, 1200 нукл.). в), г) кДНК из генома человека для белка 285А «цинковых пральцев» (KEGG, hsa: 26974, 1773 нукл.). д), е) кДНК из генома человека для белка – фактора некроза опухоли (KEGG, hsa:338872, 1002 нукл.).

3.2. Анализ метода информационного разложения (ИР-метода)

При анализе на основе 2С-подхода ИР-метода, предложенного в работах [2, 3] для оценки скрытого периода в последовательностях ДНК, будем использовать терминологию, принятую в настоящей работе.

ИР-метод даёт оценку скрытого периода на основе косвенного показателя, не гарантирующего наличие скрытой периодичности. Этот метод использует введённую ранее Кульбаком стандартную информационную статистику (см. [19], С. 170, формула (2.5)), которая для анализируемой текстовой строки str длиной n в алфавите $A = \langle a_1, \dots, a_K \rangle$ на тест-периоде размера λ имеет вид

$$I(\lambda) = \sum_{j=1}^{\lambda} \sum_{i=1}^K f_{ij} \ln(f_{ij}) - \sum_{j=1}^{\lambda} f_{\cdot j} \ln(f_{\cdot j}) - \sum_{i=1}^K f_{i \cdot} \ln(f_{i \cdot}) + n \ln(n), \quad (9)$$

где f_{ij} - количество буквы a_i из алфавита A в j -ых позициях тест-периода λ при разбиении строки str на подстроки длиной λ , $f_{i \cdot} = \sum_{j=1}^{\lambda} f_{ij}$ для $i = 1, \dots, K$ и $f_{\cdot j} = \sum_{i=1}^K f_{ij}$ для $j = 1, \dots, \lambda$.

Неотрицательные целые числа f_{ij} образуют компоненты матрицы $\mathbf{M}$ размера $K \times \lambda$, т.е.

$$\mathbf{M} = (f_{ij})_{K \times \lambda}. \quad (10)$$

В монографии [19] (см. формулу (2.7), стр. 171) показано, что для реализаций однородной профильной строки длиной n на её тест-периоде λ для формулы (9) справедливо соотношение:

$$\Phi(\lambda) = 2I(\lambda) \sim \chi_{(K-1)(\lambda-1)}^2. \quad (11)$$

В работах [9, 12] формула (11), с учетом (9), была представлена в виде

$$\Phi(\lambda) = 2 \frac{n}{\lambda} \sum_{j=1}^{\lambda} \sum_{i=1}^K (\pi_j^i \ln(\pi_j^i) - p^i \ln(p^i)), \quad (12)$$

где $\frac{n}{\lambda} > 20$, $\boldsymbol{\pi} = (\pi_j^i)_L^K$ - λ -профильная матрица и $p^i = \frac{1}{\lambda} \sum_{j=1}^{\lambda} \pi_j^i$ ($i = 1, \dots, K$) - частота

встречаемости i -той буквы алфавита A в текстовой строке str . Вид (12) более наглядно, чем представление (9), демонстрирует статистическое информационное отличие на тест-периоде λ строки str от однородной строки с теми же частотами встречаемости букв. Поэтому, согласно формулам (11) и (12), спектр Φ для реализаций однородной профильной строки статистически эквивалентен главному спектру Ψ строки str , вид которого (2) используется в спектрах 2С-подхода (см. формулы (3), (4)). Следовательно, спектр Φ , как и эквивалентный ему спектр Ψ , может использоваться для выявления неоднородности в анализируемой текстовой строке.

В ИР-методе [2, 3] для анализируемой текстовой строки str длиной n вместо статистики $\Phi(\lambda) = 2I(\lambda)$ (см. (9) и (11)) используется статистика J , принимающая на тест-периоде λ значение:

$$J(\lambda) = 2I(\lambda) - (K-1)(\lambda-1). \quad (13)$$

Далее, согласно [2, 3], на тест-периоде λ с помощью метода Монте-Карло вычисляется статистика J для каждой из случайно выбранных целочисленных матриц (с неотрицательными компонентами) размера $K \times \lambda$, сумма компонент которых в каждом столбце и каждой строке совпадает с такими же суммами в соответствующих столбцах и строках исходной матрицы $\mathbf{M}$, введённой в обозначениях формулы (9) видом (10). Затем, на основе полученных данных вычисляется среднее значение $\bar{J}(\lambda)$ и

выборочная дисперсия $d(\lambda)$ статистики $\mathbf{J}$ для таких случайно выбранных матриц. В результате создаётся статистика $\mathbf{Z}$, которая для текстовой строки str на тест-периоде λ принимает значение:

$$Z(\lambda) = \frac{J(\lambda) - \bar{J}(\lambda)}{\sqrt{d(\lambda)}}. \quad (14)$$

На основании эмпирических данных в [2, 3] относительно статистики $\mathbf{Z}$ выдвигается предположение о том, что, при выявлении скрытой периодичности, её значения для различных тест-периодов имеют единое пороговое значение. В качестве такого порогового значения выбрано число 7. Для анализируемой последовательности ДНК среди всех испытываемых периодов, на которых значения статистики $\mathbf{Z}$ превышают пороговое значение, в качестве «наиболее значимого» периода скрытой периодичности выбирается испытываемый период с максимальным значением $\mathbf{Z}$. Следовательно, без какого-либо теоретического обоснования для статистики $\mathbf{Z}$ предполагается статистическое равноправие всех испытываемых периодов. Кроме того, в методе Монте-Карло для тест-периодов не описан и не обоснован способ случайного выбора соответствующих матриц, необходимых для вычисления статистики $\mathbf{Z}$. Без такого описания и обоснования нельзя гарантировать свойства статистики $\mathbf{Z}$ и, следовательно, статистика $\mathbf{Z}$ становится невоспроизводимой для других исследователей. Кроме того, если в анализируемой последовательности str длиной n испытываемый период λ имеет малое число повторов ($\frac{n}{\lambda} < 20$), то исходная матрица

$\mathbf{M}$ (см. формулы (9) и (10)) теряет статистическую устойчивость, т.е. репрезентативность. Поэтому, область определения для спектра значений статистик $\mathbf{I}$ (см. формулы (9), (11) и (12)) и $\mathbf{Z}$ (см. формулы (13), (14)) будет совпадать с областью определения главного спектра Ψ (или Φ) строки str . Однако, в работах [2, 3] область определения для спектра значений статистики $\mathbf{Z}$ необоснованно расширена вплоть до двух повторов. Таким образом, для анализируемой текстовой строки str , в общем случае, установление единого порога значений спектра статистики $\mathbf{Z}$ также становится проблематичным, что может привести к некорректности предлагаемых ИР-методом оценок длины периода скрытой периодичности.

В работе [2] для анализируемой строки str отмечалось подобие спектра статистики $\mathbf{Z}$ и Фурье-спектра. Поэтому оценка скрытого периода на основе спектра статистики $\mathbf{Z}$, как и на основе Фурье-спектра, требует дальнейшего подтверждения.

Приведём примеры анализа результатов применения ИР-метода для последовательностей ДНК из работы [2] с позиций 2С-подхода который, как отмечалось ранее, в отличие от ИР-метода, является одним из методов распознавания скрытой периодичности в последовательностях ДНК. Ранее в [1], на основе 2С-подхода, такой анализ был выполнен без учёта выравнивания, предлагаемого в таблице 4 работы [2]. В настоящей работе, этот анализ проводится с учётом этого выравнивания.

Сначала рассмотрим три из четырёх последовательностей ДНК, для которых результаты их анализа на основе 2С-подхода не совпадают с оценками периодов скрытой периодичности, указанными в работе [2].

В работе [2] на основе выравнивания и последующего использования статистики $\mathbf{Z}$ утверждается, что последовательность ДНК, спектры которой приведены на рис. 11, обладает скрытой периодичностью в 2 нукл. Согласно 2С-подходу, из анализа спектров на рис. 11,а для исходной (без выравнивания) последовательности ДНК нельзя сделать вывод о наличии в этой последовательности скрытого профильного периода в 2 нукл. Этот вывод нельзя принимать и для выравненной последовательности, что иллюстрируют спектры на рис. 11,б. В обоих случаях (для исходной и для выравненной

последовательностей) характеристические спектры ($\mathbf{C}$) обладают свойством 2-регулярности и спектр отклонения от однородности $\mathbf{D}_1$ фиксирует неоднородность анализируемой последовательности. Однако, спектр отклонения от 2-профильности ($\mathbf{D}_2$), как и от 8-профильности ($\mathbf{D}_8$), согласно Замечанию 2.5, не позволяют сделать вывод о наличии в этой последовательности как скрытой 2-профильности, так и скрытой 8-профильности. Сразу отметим, что вывод об отсутствии в анализируемой последовательности скрытой 2-профильности, не исключает наличие в ней скрытой периодичности в 2 нукл. какого-то другого типа. Но для такого альтернативного случая необходимо иметь метод распознавания, основанный на математической модели периодичности, отличной от профильной. Однако при использовании спектра статистики $\mathbf{Z}$ никакой строгой математической модели периодичности не предполагается.

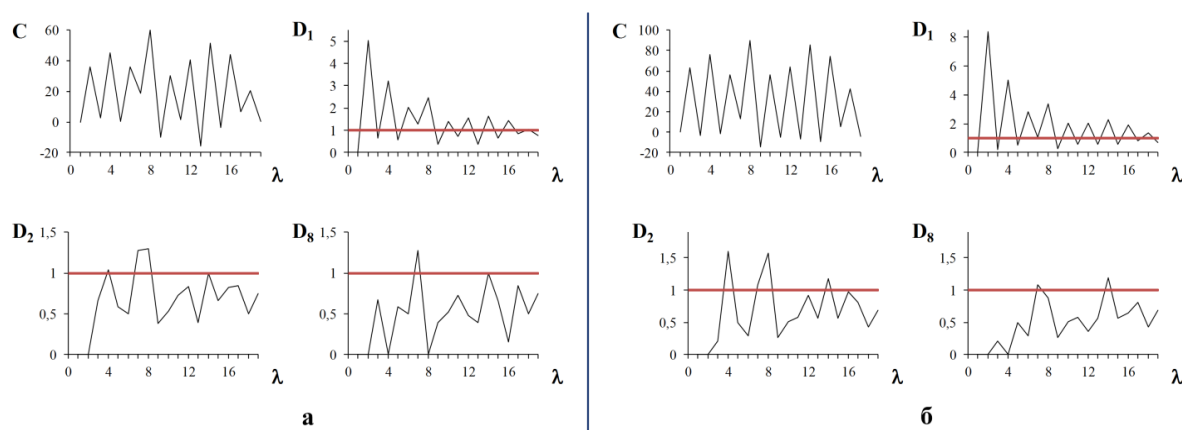


Рис. 11. а) Иллюстрация анализа последовательности ДНК из генома бактерии *X. campestris* (GenBank AE012486, GI:21114930, индексы: 6393 – 6682) с помощью спектров 2С-подхода. б) Иллюстрация анализа этой последовательности с учётом её выравнивания, приведённого в табл. 4 работы [2].

Анализ другой последовательности ДНК из работы [2], выполненный на основе 2С-подхода, проиллюстрирован на рис. 12. На рис. 12,а приведены спектры 2С-подхода для исходной последовательности и на рис. 12,б – аналогичные спектры этой последовательности с учётом её выравнивания. Согласно Замечанию 2.5, из анализа рис. 12,а следует, что в исходной последовательности распознаётся скрытая профильная периодичность в 3 нукл., которая подтверждается спектром теоретической реконструкции $\mathbf{Th}_3$ (показан жирной линией) для её спектра $\mathbf{D}_1$. Анализ той же последовательности ДНК, но с учётом её выравнивания, проиллюстрирован на рис. 12,б. Согласно этому рисунку, в этой выравненной последовательности распознаётся скрытая профильная периодичность в 6 нукл., что выявляет спектр $\mathbf{D}_6$ и подтверждает спектр $\mathbf{Th}_6$ (выделен на рисунке жирной линией), являющийся теоретической реконструкцией спектра $\mathbf{D}_1$. Однако, в работе [2] утверждается, что в рассматриваемой последовательности ДНК наблюдается скрытая периодичность в 2 нукл. Такой вывод плохо согласуется с тем, что анализируемая исходная последовательность ДНК принадлежит кодирующему району гена.

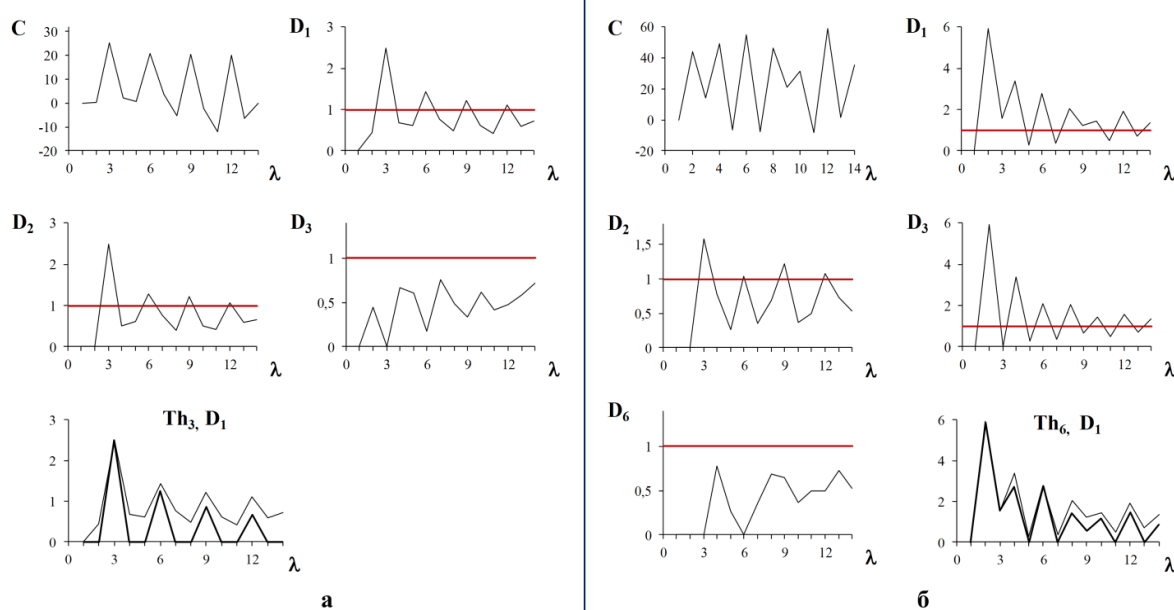


Рис. 12. а) Иллюстрация анализа последовательности кДНК из генома бактерии *B. cerasia* (GenBank AF453480, GI:18182536, индексы: 4166–4368) с помощью спектров 2С-подхода. **б)** Иллюстрация анализа этой последовательности с учётом её выравнивания, приведённого в табл. 4 работы [2].

Анализ на основе 2С-подхода ещё одной последовательности ДНК из работы [2], проиллюстрирован на рис. 13. Из этого рисунка следует, что на уровне значимости $\alpha = 0.05$, как исходная (рис. 13,а), так и выравненная (рис. 13,б) последовательности, неотличимы от однородной последовательности (период в 1 нукл.).

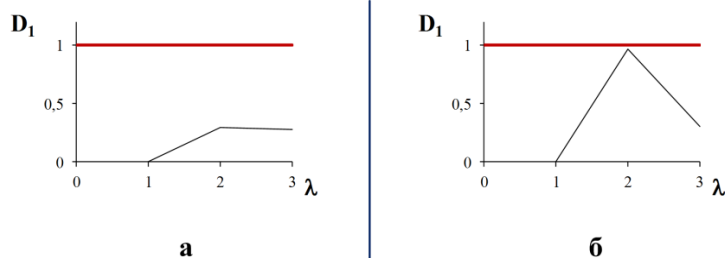


Рис. 13. а) Иллюстрация анализа исходного фрагмента ДНК из генома бактерии *C. jejuni* (GenBank AL111168, GI:3047139, индексы: 165728 – 165785) с помощью спектра отклонения от однородности D_1 2С-подхода. **б)** Аналогичный анализ этого фрагмента с учётом выравнивания, приведённого в табл. 4 [2].

Отметим, что в работе [2] для последовательности ДНК, спектры фрагментов которой показаны на рисунках 13 и 14, был указан несуществующий на сегодняшний день локус CJ11168X1. В настоящее время этому локусу в GenBank соответствует локус AL111168.

Рисунок 14 иллюстрирует наличие в выравненной четвёртой последовательности ДНК из таблицы 4 работы [2] скрытой профильной периодичности в 2 нукл., распознанной согласно изложенному в настоящей работе 2С-подходу.

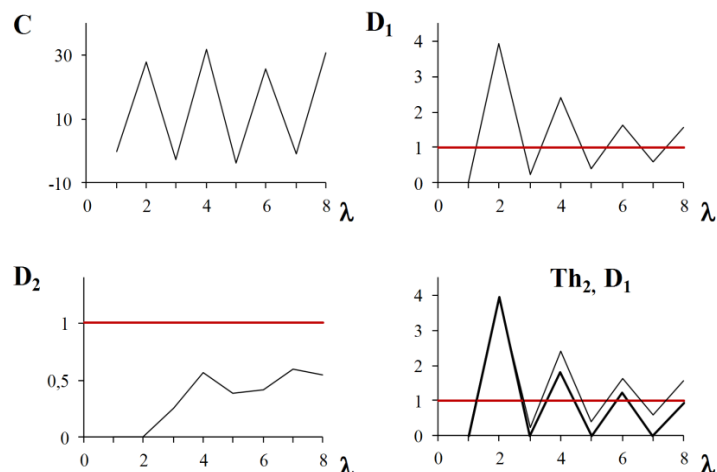


Рис. 14. Иллюстрация анализа выравненного фрагмента ДНК из генома бактерии *C. jejuni* (GenBank AL111168, GI:3047139, индексы: 176412–176535) с помощью спектров 2С-подхода.

Таким образом, оценка скрытого периода анализируемой последовательности ДНК на основе спектра статистики $\mathbf{Z}$, как и на основе Фурье-спектра, не всегда корректна и требует дополнительного подтверждения. Кроме того, анализ только одного спектра статистики $\mathbf{Z}$ нельзя рассматривать как метод распознавания скрытой периодичности в последовательности ДНК.

4. НЕСОСТОЯТЕЛЬНОСТЬ КРИТИКИ СПЕКТРАЛЬНО-СТАТИСТИЧЕСКОГО ПОДХОДА (2С-ПОДХОДА)

В настоящем разделе показывается, что критика 2С-подхода, которой посвящена публикация [4], выявляет непонимание авторами этой публикации основ теории распознавания образов, стандартных методов математической статистики и понятий математического анализа. В результате такая критика превращается в целый ряд математически ошибочных заключений о некорректности 2С-подхода. Эта критика порождена неверными представлениями о том, что при распознавании скрытой периодичности в последовательностях ДНК можно опираться на одни только косвенные признаки, которые, как было показано выше (разделы 3.1 и 3.2), не гарантируют обнаружения скрытой периодичности в этих последовательностях. Также необходимо отметить некоторые искажения в публикации [4] традиционной научной терминологии. Например, во введении на стр.529 написано: «...для поиска скрытой периодичности нуклеиновых последовательностей...». Видимо, авторы публикации имели в виду скрытую периодичность в нуклеотидных последовательностях? Далее, на стр. 531, в подписи к рис. 3 вместо термина «равновероятного распределения букв» используется ошибочный термин «равномерного содержания букв». Кроме того, авторы [4] смешивают математические понятия максимума и максимального значения (см., например, начало второго абзаца на стр. 533) при анализе поведения спектральных зависимостей в области их определения.

Далее, отвечая на критику 2С-подхода в [4], при нумерации подразделов настоящего раздела будем придерживаться, за исключением подраздела 4.1, соответствующей нумерации разделов публикации [4].

4.1. О наличии различных скрытых периодичностей в последовательности ДНК

В критической публикации [4] (см., например, начала разделов 4 и 5), без каких-либо обоснований, многократно используется произвольное предположение о

существовании в отдельно взятой последовательности ДНК нескольких различных скрытых периодов какого-то неизвестного типа периодичности. Несостоятельность такого предположения иллюстрирует характеристический спектр (как и Фурье-спектр) совершенного текстового тандемного повтора, показанный на рис. 9,а (рис. 9,б). Поскольку паттерн периодичности этого тандемного повтора имеет вид «atgcaattggccaaatttgggccc», то, несмотря на наличие в этом спектре многих значимых максимумов, нельзя говорить о существовании в этом тандемном повторе «скрытых периодов», отличных от периода 24 нукл. Согласно теории распознавания образов, ссылки на наличие нескольких значимых максимумов в таких косвенных показателях, как статистика Z , Фурье-спектр, характеристический спектр и спектр отклонения от однородности, не могут служить достоверной аргументацией для произвольного предположения о наличии в последовательности ДНК нескольких различных скрытых периодов какого-то неизвестного типа периодичности. Поэтому, многократно повторённая в публикации [4] критика 2С-подхода, использующая такое произвольное предположение, вследствие его несостоятельности, лишена каких-либо оснований.

4.2. О поиске скрытой периодичности во фрагментах последовательностей ДНК

В разделе 2 публикации [4] отмечалось, что 2С-подход используется для распознавания скрытой периодичности в последовательностях ДНК из работы [2] без учёта выполненной в [2] процедуры выравнивания. В конце раздела 3.2 настоящей работы распознавание с помощью 2С-подхода скрытой периодичности было выполнено с учётом такого выравнивания (см. рис. 11–13,б и рис. 14). Было показано, что для трёх из четырёх, рассмотренных в [2] последовательностей, оценки периода скрытой периодичности, полученные на основе ИР-метода, нельзя признать корректными. Корректность оценки периода скрытой периодичности для четвёртой последовательности подтверждена только благодаря 2С-подходу, поскольку, как неоднократно отмечалось выше, ИР-метод не является методом распознавания скрытой периодичности в последовательностях ДНК.

4.3. О проблеме малой выборки в спектрально-статистическом подходе

Замечание в адрес 2С-подхода по проблеме малой выборки, выдвинутое в разделе 3 публикации [4], может вызвать, по меньшей мере, недоумение. В работе [1] в первом абзаце раздела 3.1 указано ограничение в виде минимального порога для числа анализируемых периодов (повторов периода). Согласно этому ограничению (см. также текст после формулы (1) из работы [1]), для последовательностей ДНК такое число должно быть больше 20. Все тест-периоды анализируемых в [1] последовательностей ДНК удовлетворяют этому ограничению. В работе [1] это требование выражается в определении допустимого диапазона тест-периодов, ограниченного сверху максимальным периодом $L_{\max} = \lambda_{\max} \leq n/5K$, где n – длина анализируемой последовательности и $K = 4$ – размер алфавита последовательностей ДНК. Поэтому критический пассаж о проблеме малой выборки в разделе 3 публикации [4] не имеет никакого отношения к использованию 2С-подхода для распознавания скрытой профильной периодичности в последовательностях ДНК. Недоумение авторов тем более оправдано, что авторы публикации [4] на стр. 534 раздела 5 сами используют введённый авторами настоящей работы в публикации [1] верхний порог $\lambda_{\max}$ для тест-периодов анализируемых последовательностей, чтобы с помощью некорректных рассуждений указать на ограниченность 2С-подхода. Более того, достаточность такого порога, обеспечивающего для последовательности ДНК 20-ти повторов периода, авторы критики сами демонстрируют на рис. 3 в публикации [4]. Замечание авторов публикации [4] в начале раздела 3 о том, что «это явление достаточно широко известно

при статистическом анализе данных», можно отнести, видимо, к авторам публикации [4], которые, наверное, недавно познакомились с этим явлением. Иначе, никак нельзя объяснить результаты проведенного авторами [4] численного эксперимента (см. рис. 3 в [4]), демонстрирующего это широко известное явление.

4.4. О поиске «наиболее статистически значимого периода» для выявления скрытой периодичности в последовательностях ДНК

Несостоятельность произвольного предположения авторов публикации [4] о наличии в отдельно взятой последовательности ДНК нескольких различных скрытых периодов неизвестного типа периодичности рассмотрена в разделе 4.1 настоящей работы. Также следует отметить, что утверждение авторов публикации [4], в начале раздела 4, о том, что информационный спектр $2I$ (см. формулы (9-12) настоящей работы) «...был получен нами в работах...», не соответствует действительности. Как уже отмечалось выше (см. текст к формулам (9-12) настоящей работы), этот спектр и его свойства были описаны ранее другим автором в монографии [19].

Критика авторами публикации [4] использования в 2С-подходе спектра отклонения от однородности D_1 (см. формулу (3)) и характеристического спектра C (см. формулу (4)) обусловлена непониманием этими авторами стандартных статистических критериев согласия.

Согласно 2С-подходу, спектр D_1 может использоваться на тест-периоде λ для проверки однородности текстовой строки. Если на тест-периоде λ рассматриваются значения главного спектра Ψ текстовых строк, являющихся реализациями однородной профильной строки, то для проверки однородности этих реализаций критерий согласия Пирсона на уровне значимости $\alpha = 0.05$ предполагает справедливость соотношения: вероятность $P(\Psi(\lambda) > \chi_{crit}^2((K-1)(\lambda-1), \alpha)) = \alpha$. Поскольку, согласно формуле (3), $D_1(\lambda) = \Psi(\lambda) / \chi_{crit}^2((K-1)(\lambda-1), \alpha)$, отсюда следует, что для реализаций однородной строки справедливо утверждение: вероятность $P(D_1(\lambda) > 1) = \alpha$ не зависит от тест-периода λ . Но, согласно теории вероятности, в этом случае также справедлив и вывод о зависимости вероятности $P(D_1(\lambda) > 1.5)$ от тест-периода λ , проиллюстрированной в публикации [4] на рис. 5. Таким образом, эта зависимость никак не влияет на использование критерия согласия при выбранном уровне значимости. Отсюда следует, что авторы публикации [4] не понимают смысла критериев согласия с фиксированным уровнем значимости. Вследствие такого заблуждения, в первом абзаце на стр. 533 публикации [4] появляются ошибочные заключения о якобы некорректном использовании в работе [1] спектра D_1 .

Непонимание авторами публикации [4] стандартных статистических критериев проявляется также при критике используемого в 2С-подходе характеристического спектра C (см. формулу (4)). В 2С-подходе, согласно Замечанию 2.4, при оценке скрытого периода профильной периодичности в анализируемой текстовой строке, спектр C используется только в том случае, когда на основании анализа спектра D_1 строка признана неоднородной. Следовательно, для главного спектра Ψ реализаций неоднородных случайных строк на тест-периоде λ нельзя использовать предположение: $\Psi(\lambda) \sim \chi_{(K-1)(\lambda-1)}^2$, т.е. предположение о том, что распределение значений $\Psi(\lambda)$ имеет χ^2 -распределение с $N = (K-1)(\lambda-1)$ степенями свободы. Поэтому для главного спектра реализаций случайных неоднородных строк, тем более неоднородных профильных строк, нельзя основываться на предположении: $\Psi(\lambda) \sim \chi_{(K-1)(\lambda-1)}^2$ (см. текст после формулы (2)), что показано на рис. 1,в и рис. 1,г.

Следовательно, проиллюстрированная на рис. 4 критической публикации [4] зависимость $C(\lambda) = \Psi(\lambda) - (K-1)(\lambda-1)$ (см. формулу (4) настоящей работы), построенная в предположении: $\Psi(\lambda) \sim \chi^2_{(K-1)(\lambda-1)}$, не может служить основанием для критики правила получения оценки периода скрытой периодичности, сформулированного в Замечании 2.4. Поэтому вывод публикации [4] в конце страницы 533 о неверности оценки скрытого периода профильной периодичности в работах [1, 5, 16] математически ошибочен. По той же причине опровержение результатов работы [16] в публикации [4] (см. конец стр. 533 и начало стр. 534) математически ошибочно и несостоятельно и, кроме того, лишний раз демонстрирует непонимание авторами критической публикации [4] использования стандартных методов математической статистики.

В начале третьего абзаца раздела 4 публикации [4] утверждается, что в работе [1] допущена ошибка при использовании правила оценки скрытого периода, проиллюстрированного в настоящей работе в Замечании 2.4. Это утверждение базируется на непонимании авторами публикации [4] различия между понятиями максимума (локальной характеристики) и максимального значения (глобальной характеристики) для графика функции в области её определения. Напомним, что для оценки скрытого периода рассматриваемое правило ориентируется на выбор наименьшего тест-периода, на котором достигается максимальное значение характеристического спектра. Для демонстрации отмеченного непонимания приведём цитату из указанного третьего абзаца, где, из-за допущенной небрежности, ссылку на рис. 2,б необходимо заменить ссылкой на рис. 4,б (в настоящей работе этот рисунок представлен на рис. 4,в):

«В работе [1] допущена также ошибка на рис. 2,б. Как видно, авторы применяют свое правило, которое мы цитировали выше в этом пункте, и из всех периодов, видимых на рис 4,б, они берут такой, который имеет максимально значение $C(\lambda)$, т. е. пропускают все λ до $\lambda = 84$, где наблюдается максимум $C(\lambda)$ Почему в данном случае авторы противоречат своему правилу, и почему они считают, что это правило позволяет получить «правильные оценки скрытых периодов», остается загадкой. Авторы публикации [1] не приводят никаких доводов и расчетов в доказательство справедливости своего правила».

4.5. О понятии профильной периодичности

В разделе 5 авторы публикации [4] на основе указанных выше ошибочных математических утверждений критикуют понятие скрытой профильной периодичности в текстовых строках. Если последовать принципам этой критики профильной периодичности, то следовало бы говорить и о неинформативности понятия совершенного тандемного повтора, естественным стохастическим обобщением которого является понятие профильной периодичности. Поэтому авторы настоящей работы не могут признать правомерность таких принципов критики, основанной к тому же, как уже было показано в разделе 4.1 настоящей работы, на произвольном предположении о наличии в каждой из анализируемых последовательностях ДНК работ [1, 5, 16] нескольких различных скрытых периодов какого-то неизвестного типа периодичности.

В настоящей работе в разделах 2 и 3 приведено обоснование достоверности и эффективности использования понятия профильной периодичности для последовательностей ДНК при распознавании скрытой периодичности в последовательностях ДНК.

4.6. О двухуровневой организации скрытой периодичности в генах

Ранее [5, 7, 16], использование 2С-подхода для анализа последовательностей ДНК из генома человека позволило авторам настоящей работы сделать вывод о наличии в последовательностях кДНК двухуровневой организации кодирования. Первый уровень, обусловленный генетическим кодом, индуцирующим 3-регулярность характеристического спектра последовательности, не всегда приводит к наличию в ней скрытой триплетной профильной периодичности. В этом случае в последовательности ДНК может проявляться второй уровень организации кодирования в виде скрытой профильной периодичности с периодом, кратным трём. В настоящей работе этот вопрос уже обсуждался разделе 2.6.

Авторы публикации [4] в разделе 6 утверждают, что наличие двухуровневой организации кодирования в последовательностях ДНК была отмечена ими ранее в работе [20]. Сразу заметим, что ссылка в этом разделе на работу [20] ошибочна, на самом деле её следует заменить ссылкой на работу [3]. Говоря в работе [3] о повторяющихся через 2 нукл. максимумах в анализируемых ими спектрах информационного разложения, авторы работы [3] такое явление сравнили с «амплитудной модуляцией радиосигналов», когда «скрытая периодичность, кратная трём основаниям, как бы модулирует периодичность, равную трём основаниям» (цитата из раздела 6 публикации [4]). На этом основании в разделе 6 публикации [4] сделано заявление, цитируем: «Авторы работ [4,8] поэтому не самостоятельно заметили этот факт, а всего лишь копируют выводы работы [7], в данном случае, без ссылок на первоисточник». Цитируемые ссылки на работы «[4]», «[8]» и «[7]» соответствуют ссылкам [5], [16] и [20] настоящей работы с учётом того, что, как отмечено выше, ссылку на работу [20] следует заменить ссылкой на публикацию [3] в настоящей работе. Из приведённой цитаты видно, что авторы публикации [4] не понимают научной ошибочности своих обвинений. В работе [16] экспериментально обоснован вывод о том, что триплетная природа генетического кода в последовательностях кДНК, вызывающая 3-регулярность характеристических спектров, обуславливает профильную периодичность, кратную трём. Авторы публикации [4] ошибочно отождествляют этот вывод с противоположным необоснованным тезисом о формальном наведении («амплитудной модуляцией») триплетной регулярности спектров информационного разложения в последовательностях ДНК скрытой периодичностью, имеющей период, кратный трём. Таким образом, искажая выводы работы [16], авторы критической публикации [4] приписывают эти выводы своей публикации [3]. К этому следует добавить, что триплетная регулярность максимумов в различных спектрах последовательностей ДНК была замечена многими другими авторами и, как правило, это явление ассоциировалось с триплетной природой генетического кода.

4.7. О преимуществах спектрально-статистического подхода

В разделе 7 публикации [4] говорится о преимуществе ИР-метода для поиска скрытой периодичности в нуклеотидных последовательностях перед 2С-подходом, изложенном авторами настоящей работы в публикации [1]. В разделе 3.2 настоящей работы продемонстрирована ошибочность такого вывода. Как было показано, по сравнению с 2С-подходом, являющимся методом распознавания скрытой профильной периодичности, основным недостатком ИР-метода и Фурье-анализа является то, что эти методы ориентируются на косвенные показатели, не гарантирующие наличие скрытой периодичности в нуклеотидных последовательностях. Поэтому, строго говоря, согласно теории распознавания образов, ИР-метод и Фурье-анализ не являются методами распознавания скрытой периодичности в нуклеотидных последовательностях.

4.8. Об ошибках первого и второго рода

Для обсуждения раздела 8 критической публикации [4] необходимо отметить ещё одну неправильную ссылку на работу [1] авторов настоящей работы. На самом деле это ссылка [16] настоящей работы и ссылка [8] публикации [4].

Как в настоящей работе, так и в работах [5, 6, 16], для распознавания скрытой профильной периодичности используются стандартные статистические критерии согласия на уровне значимости $\alpha = 0.05$. В этих статистических критериях уровень значимости и является ошибкой I-го рода. Утверждение в разделе 8 из критической публикации [4] о том, что в перечисленных работах не определена ошибка I-го рода, можно объяснить только тем, что авторы критической публикации [4] не знают этого факта. К этому следует добавить, что авторы критической публикации [4], используя ИР-метод в своих предшествующих работах, не приводят для этого метода строгого определения не только ошибок II-го рода, но и ошибок I-го рода.

Критерии согласия, используемые в 2С-подходе, не предполагают определения ошибок II-го рода, поскольку для рассматриваемых феноменов, как и других явлений, широко распространенных в практике использования математической статистики, невозможно предложить математически строгие альтернативные гипотезы. Таким образом, предъявляя для методов распознавания скрытой периодичности требования определения не только ошибок I-го рода, но и ошибок II-го рода, авторы публикации [4] показывают своё недостаточное знакомство со спецификой применения статистических методов в прикладных исследованиях. К этому следует добавить, что при использовании 2С-подхода для распознавания скрытой профильной периодичности оценки, полученные на основе соответствующих критериев согласия, подтверждались с помощью дополнительных методов, изложенных в разделе 2.5 настоящей работы.

5. ВЫВОДЫ

В настоящей работе показано, что авторы критической публикации [4] не владеют основными понятиями теории распознавания образов, методами математической статистики и математического анализа. Представленная в публикации [4] критика, помимо небрежного изложения, построена на математически ошибочных аргументах, произвольных выводах и неверных интерпретациях утверждений из цитируемых работ, что и продемонстрировано в настоящей работе. Подобная, основанная на заблуждениях, манера критики публикации [4] не может служить основанием для отрицания значимости результатов применения спектрально-статистического подхода в критикуемых публикациях.

В работе на основе спектрально-статистического подхода продемонстрировано, что оценки периодов скрытой периодичности в последовательностях ДНК, основанные на косвенных показателях методов Фурье-анализа и информационного разложения, могут оказаться недостоверными и требуют дополнительного подтверждения.

Авторы настоящей работы надеются, что, более развёрнутое, чем ранее, изложение спектрально-статистического подхода, приведённое в настоящей работе, позволит прояснить значимость этого подхода, позволяющего распознавать новый тип скрытой периодичности в последовательностях ДНК.

СПИСОК ЛИТЕРАТУРЫ

1. Чалей М.Б., Кутыркин В.А. Распознавание скрытой периодичности в последовательностях ДНК. *Математическая биология и биоинформатика*. 2013. Т. 8. № 2. С. 502–512. URL: http://www.matbio.org/2013/Chaley_8_502.pdf (дата обращения 30.12.2013).

2. Shelenkov A., Skryabin K., Korotkov E. Search and classification of potential minisatellite sequences from bacterial genomes. *DNA Res.* 2006. V. 13. P. 89–102.
3. 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.
4. Коротков Е.В., Шеленков А.А., Короткова М.А. К вопросу о распознавании скрытой периодичности в последовательностях ДНК. *Математическая биология и биоинформатика.* 2013. Т. 8. № 2. С. 529–536. URL: http://www.matbio.org/2013/Korotkov_8_529.pdf (дата обращения 30.12.2013).
5. Chaley M., Kutyrkin V. Profile-Statistical Periodicity of DNA Coding Regions. *DNA Res.* 2011. V. 18. P. 353–362.
6. 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.
7. 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.
8. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res.* 1999. V. 27. P. 573–580.
9. 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.
10. 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 30.12.2013).
11. 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.
12. 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](http://www.matbio.org/downloads/Chaley2007(2_20).pdf) (дата обращения 30.12.2013).
13. 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 30.12.2013).
14. Крамер Г. *Математические методы статистики.* М.: Мир, 1975. 648 с.
15. Kanehisa M., Goto S., Sato Y., Furumichi M., and Tanabe M.; KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.* 2012. V. 40. D109–D114.
16. 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.
17. Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. *Computer Applications in Biosciences.* 1997. V. 13. № 3. P. 263–270.
18. FTG: Fast Fourier Transform based GENE Prediction Server. URL: <http://www.imtech.res.in/raghava/ftg/> (дата обращения 30.12.2013)
19. Кульбак С. *Теория информации и статистика.* М.: Наука, 1967. 408 с.

20. 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.

Материал поступил в редакцию 29.12.2013, опубликован 29.01.2014.