Математическая биология и биоинформатика
2016. Т. 11. № 1. С. 24–45. doi: 10.17537/2016.11.24

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

УДК: 577.2:519.23

Модель организации кодирования в прокариотических организмах

Кутыркин В.А.[*1], Чалей М.Б.[2]**

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

Аннотация. На основе спектрально-статистического подхода (2С-подхода) выполнен анализ скрытой периодичности в CDS геномов прокариотических организмов. Показано, что его результаты совпадают с результатами проведённого ранее анализа на скрытую периодичность в CDS геномов эукариотических организмов. Впервые для CDS геномов как низших, так и высших организмов определён тип скрытой триплетной периодичности, гипотеза о существовании которой широко обсуждалась в последние десятилетия. Показано, что CDS геномов характеризуются новым типом скрытой периодичности, названным профильной периодичностью. Для объяснения существования в CDS скрытой (триплетной) профильной периодичности рассматривается стохастическая модель однородной организации кодирования (SHOC-модель) в текстовых строках. Модель объясняет наличие скрытой профильной периодичности и регулярности в последовательностях ДНК. Характерные проявления SHOC-модели демонстрируются в численных экспериментах с бинарно перекодированным литературным текстом.

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

1. ВВЕДЕНИЕ

Хорошо известно, что кодирование белков в последовательностях ДНК основано на триплетном генетическом коде. Вследствие чего, благодаря различным косвенным результатам анализа генов и кодирующих последовательностей ДНК (CDS), широкое распространение получила гипотеза о наличии в них периодичности в 3 основания. Многие математические методы выявления генов и кодирующих районов ДНК пытались использовать эту гипотезу [1–5]. Появляются предположения об особенностях структуры кодирующих последовательностей ДНК, являющихся причиной скрытой триплетной периодичности, например, [6–8]. Однако, существующие достоверные методы распознавания скрытой периодичности, как правило, опираются на модель скрытой периодичности, основанную на понятии размытых тандемных повторов [9, 10]. Но такой тип скрытой периодичности, как правило, не имеет распространения по всей длине значительной части кодирующих районов. Методы, основанные на модели периодичности, можно отнести к прямым методам распознавания скрытой периодичности. Наряду с прямыми методами

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

распознавания скрытой периодичности широкое применение нашли методы, которые можно было бы назвать косвенными, поскольку они не опираются на какую-либо модель периодичности. Например, к таким методам относятся автокорреляционные функции и методы Фурье-анализа, в частности, Spectral-envelope [11]. Доминирующие пики или наличие периодограммы в этих методах являются косвенным подтверждением в пользу существования скрытой периодичности. Строго говоря, полученные с помощью этих методов оценки периода скрытой периодичности требуют дополнительного подтверждения. Кроме того, эти методы не опираются на какой-либо тип периодичности.

Ранее, в работе [12] было предложено новое понятие скрытой периодичности, названное скрытой профильной периодичностью. Это новое понятие основывается на модели профильной периодичности (профильности), позволяющей обобщить понятие размытого тандемного повтора [9]. Благодаря этой модели в работах [13–15] был разработан спектрально-статистический подход (2С-подход) к достоверному распознаванию в последовательности ДНК скрытой профильной периодичности. Поскольку новый тип скрытой периодичности обобщает понятие размытого тандемного повтора, можно было бы предположить, что такой тип периодичности распознаётся в большинстве кодирующих районов ДНК. Это предположение в настоящей работе подтверждается на примере анализа CDS геномов прокариотических организмов. Ранее, в работе [16], справедливость этого предположения была показана для CDS геномов эукариотических организмов, в том числе для генома человека.

В настоящей работе, как и в [16], показывается, что наряду со скрытой триплетной периодичностью, практически во всех CDS из рассматриваемых геномов наблюдается феномен, который в рамках 2С-подхода [13–15] получил название триплетной регулярности. Кроме того, в [16] было показано, что в интронах эукариотических организмов феномены триплетной периодичности и регулярности отсутствуют.

Результаты исследования скрытой профильной периодичности позволяют выдвинуть гипотезу о наличии двухуровневой организации кодирования в CDS геномов прокариотических и эукариотических организмов. Первый уровень организации кодирования обусловлен генетическим кодом, проявлениями которого являются триплетная регулярность и скрытая триплетная профильность. При наличии феномена триплетной регулярности и отсутствии скрытой триплетной профильности, проявлением второго, более высокого, уровня организации кодирования является наличие скрытого профильного периода более трёх оснований.

В настоящей работе рассматривается стохастическая модель однородной организации кодирования (Stochastic Homogeneous Organization of Coding – SHOC-model) в последовательностях ДНК. Эта модель раскрывает возможное смысловое содержание существования в ДНК районов скрытой профильной периодичности. В рамках модели кодирование биологической информации осуществляется кодонами одного размера. Например, триплетный генетический код может являться естественным примером такой модели. В общем случае, согласно рассматриваемой модели кодирования, размер её кодонов совпадает с размером выявляемого в последовательности ДНК скрытого профильного периода. В этом случае особый интерес представляют последовательности ДНК, в которых распознаётся скрытая профильная периодичность, с периодом, превышающим три основания. Например, ранее рассматривались такие последовательности ДНК, в которых размер скрытого профильного периода соответствовал размерам участков, кодирующих повторяющиеся структурные домены белков [13, 14].

Для демонстрации проявления статистических свойств (скрытой периодичности и регулярности) рассматриваемой модели кодирования в настоящей работе приведены результаты модельных экспериментов, использующих перекодированный (с помощью

бинарных кодонов размера пять) текст классического литературного произведения на английском языке “Three Men in a Boat”[17].

2. МАТЕРИАЛЫ И МЕТОДЫ

В работе [12] рассматривалось новое понятие скрытой профильной периодичности (скрытой профильности). Для выявления скрытой профильности в ДНК был разработан 2С-подход [14, 15], который опирается на исследование нескольких статистических спектров, полученных при количественном анализе текстовых строк, в частности, последовательностей ДНК. В настоящей работе этот подход применялся к анализу кодирующих последовательностей ДНК (CDS) десяти прокариотических организмов (табл. 1) из базы данных KEGG версии 54.1 [18]. Полученные результаты сравнивались с аналогичными результатами для CDS генома человека [19, 20] и других эукариотических организмов [16].

Таблица 1. Анализируемые прокариотические организмы

Код организма в базе KEGG	Полное название организма	Сокращённое название организма	Таксономическая принадлежность
bbz	*Borrelia burgdorferi* strain ZS7	*B. burgdorferi*	Bacteria; Spirochaetes
bsu	*Bacillus subtilis* subsp. subtilis 168	*B. subtilis*	Bacteria; Firmicutes
buc	*Buchnera aphidicola* APS (Acyrthosiphon pisum)	*Buchnera*	Bacteria; Proteobacteria
eco	*Escherichia coli* K-12 MG1655	*E. coli*	Bacteria; Proteobacteria
eta	*Erwinia tasmaniensis* Et1/99	*E. tasmaniensis*	Bacteria; Proteobacteria
hin	*Haemophilus influenzae* Rd KW20 (serotype d)	*H. influenzae*	Bacteria; Proteobacteria
hpy	*Helicobacter pylori* 26695	*H. pylori*	Bacteria; Proteobacteria
mmy	*Mycoplasma mycoides* subsp. mycoides SC PG1	*M. mycoides*	Bacteria; Tenericutes
sty	*Salmonella enterica* subsp. enterica serovar Typhi CT18 (Salmonella typhi CT18)	*S. typhi*	Bacteria; Proteobacteria
ype	*Yersinia pestis* CO92 (biovar Orientalis)	*Y. pestis*	Bacteria; Proteobacteria

Две выборки из исходных данных базы KEGG использовались в работе. В первой выборке из анализа исключались последовательности, в описании которых было указано, что они содержат tRNA, rRNA, ncRNA, misc_RNA или tmRNA, и также гены или CDS белков, характеризуемых как «hypothetical», «similar to», «putative», «predicted», «uncharacterized», «with sequence similarity» и «pseudogene». После исключения различного рода предсказанных CDS, от соответствующих исходных записей базы KEGG оставалась, в среднем, половина. Такая выборка, общим числом 16340 CDS, далее будет называться Выборкой 1. При формировании второй выборки из исходных записей базы KEGG исключались только CDS для tRNA, rRNA, ncRNA, misc_RNA или tmRNA. В этом случае для дальнейшего анализа оставалось около 95% от исходных CDS для каждого организма. Полученная таким образом расширенная выборка из 27083 CDS далее называется Выборкой 2. В таблицах 2 и 3 приведены данные о количестве CDS каждого из рассматриваемых организмов в соответствующих выборках.

При введении количественного порога индекса 3-регулярности последовательностей CDS также использовались полученные ранее результаты анализа [19, 20] данных базы KEGG для 17652 CDS генома человека с подтверждённой

функциональной активностью и данные базы EID [21] для 277477 последовательностей интронов генома человека с длинами того же порядка, что и у CDS.

Для модельных численных экспериментов использовался текст на английском языке [17]. Бинарно перекодированные (с размером кодонов букв английского алфавита и знаков пунктуации равным пяти) абзацы текста этой книги рассматривались как своего рода аналоги CDS в анализируемых геномах.

В настоящей работе рассматривается оригинальная стохастическая модель однородной организации кодирования (SHOC-модель) в последовательностях ДНК [16], которая объясняет наличие в этих последовательностях скрытой профильной периодичности, выявляемой с помощью спектрально-статистического подхода. Поэтому прежде описания SHOC-модели приведем краткое описание 2C-подхода, основанного на модели профильной периодичности.

2.1. Модель профильной периодичности

Основу спектрально-статистического подхода составляет модель профильной периодичности. Модель профильной периодичности использует понятие случайной буквы $Chr(\mathbf{p})$ со столбцом частот $\mathbf{p}=(p^1,p^2,...,p^K)^T$. Эта случайная буква является случайной величиной, принимающей с вероятностью p^i значение i-той буквы алфавита $A=\langle a_1,a_2,...,a_K\rangle$. Из n таких независимых случайных букв формируется случайная строка $Str=Str_n(\boldsymbol{\pi})=Chr(\mathbf{p}_1)Chr(\mathbf{p}_2)...Chr(\mathbf{p}_n)$, которая однозначно определяется матрицей $\boldsymbol{\pi}=(\mathbf{p}_1,\mathbf{p}_2,...,\mathbf{p}_n)=(\pi_j^i)_n^K$, называемой n-профильной матрицей строки Str. Фактически, эта случайная строка описывает вероятностную схему из n независимых испытаний случайных величин, в роли которых выступают указанные случайные буквы. Следовательно, столбец $\mathbf{p}_j$ $(j=\overline{1,n})$ описывает соответствующие вероятности появления букв алфавита $A=\{a_1,a_2,...,a_K\}$ в j-ом испытании этой схемы, которое однозначно определяет случайная буква $Chr(\mathbf{p}_j)$. Любое натуральное число L из диапазона $1,2,...,L_{\max}$, где $L_{\max}\sim\frac{n}{5K}$, называется тест-периодом такой строки Str. Букву $a_i\in A$ можно также отождествить со случайной буквой, у которой все компоненты столбца частот – нулевые, за исключением i-той единичной компоненты. Поэтому любую текстовую строку в алфавите A возможно отождествлять с соответствующей случайной строкой той же длины. Как правило, для обозначения тестовой строки будет использоваться символ str, для случайной строки из независимых случайных букв – символ Str. При отождествлении текстовой строки str со случайной строкой допускается использование равенства $str=Str$.

Пусть L – тест-период указанной выше строки $Str=Str_n(\boldsymbol{\pi})=Str_L(\boldsymbol{\pi}_1)Str_L(\boldsymbol{\pi}_2)...Str_L(\boldsymbol{\pi}_m)Str_M(\boldsymbol{\pi}_{m+1})$, которая представлена в виде последовательной записи подстрок длины L и $0\le M<L$. В этом сокращённом представлении $Str_L(\boldsymbol{\pi}_k)=Chr(\mathbf{p}_{1+(k-1)L})Chr(\mathbf{p}_{2+(k-1)L})...Chr(\mathbf{p}_{L+(k-1)L})$ для $k=\overline{1,m}$. Если $M=0$, матрица $\boldsymbol{\Pi}_{Str}(L)=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\pi}_i$ называется L-профильной матрицей строки Str. Следовательно, j-ый столбец $(j=\overline{1,L})$ матрицы $\boldsymbol{\Pi}_{Str}(L)$ описывает соответствующие усреднённые (по числу m – тест-периодов размера L, участвующих в записи строки Str) вероятности появления букв алфавита $A=\{a_1,a_2,...,a_K\}$ в j-ой позиции тест-

периода L. Если $M \neq 0$, то в матрицу $\mathbf{\Pi}_{Str}(L)$ вносятся необходимые поправки. Таким образом, для строки Str вводится профильно-матричный спектр $\mathbf{\Pi}_{Str}$, определённый в диапазоне её тест-периодов.

Приведем пример вычисления 4-профильной матрицы $\mathbf{\Pi}_{str}(4)$ профильно-матричного спектра $\mathbf{\Pi}_{str}$ текстовой строки str, имеющей вид:

$$str = gtttcacagcagcagcctcccttgtgtccttggcttgggccctagcctcctaccagaagg =$$
$$= gttt\ caca\ gcag\ cagc\ ctcc\ cttg\ tgtc\ cttg\ gctt\ gggc\ ccta\ gcct\ ccta\ ccag\ aagg,$$
$$str = Str = Str_4(\boldsymbol{\pi}_1)\,Str_4(\boldsymbol{\pi}_2)\ \ldots\ Str_4(\boldsymbol{\pi}_{14})\,Str_4(\boldsymbol{\pi}_{15}),$$

где, согласно принятому отождествлению текстовой буквы с соответствующей случайной буквой:

$$Str_4(\boldsymbol{\pi}_1) = Chr(\mathbf{p}_g)\,Chr(\mathbf{p}_t)\,Chr(\mathbf{p}_t)\,Chr(\mathbf{p}_t), \quad \mathbf{p}_g = (0,0,1,0)^T, \mathbf{p}_t = (0,1,0,0)^T;$$
$$Str_4(\boldsymbol{\pi}_2) = Chr(\mathbf{p}_c)\,Chr(\mathbf{p}_a)\,Chr(\mathbf{p}_c)\,Chr(\mathbf{p}_a), \quad \mathbf{p}_c = (0,0,0,1)^T, \mathbf{p}_a = (1,0,0,0)^T;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$Str_4(\boldsymbol{\pi}_{14}) = Chr(\mathbf{p}_c)\,Chr(\mathbf{p}_c)\,Chr(\mathbf{p}_a)\,Chr(\mathbf{p}_g);$$
$$Str_4(\boldsymbol{\pi}_{15}) = Chr(\mathbf{p}_a)\,Chr(\mathbf{p}_a)\,Chr(\mathbf{p}_g)\,Chr(\mathbf{p}_g).$$

Матрицы $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_2$, …, $\boldsymbol{\pi}_{14}$, $\boldsymbol{\pi}_{15}$ приведены на рисунке 1. В общем случае столбец

$str = gtttcacagcagcagcctcccttgtgtccttggcttgggccctagcctcctaccagaagg$

$\boldsymbol{\pi}_1$

	1	2	3	4
a	0	0	0	0
t	0	1	1	1
g	1	0	0	0
c	0	0	0	0

$\boldsymbol{\pi}_2$

	1	2	3	4
a	0	1	0	1
t	0	0	0	0
g	0	0	0	0
c	1	0	1	0

$\boldsymbol{\pi}_3$

	1	2	3	4
a	0	0	1	0
t	0	0	0	0
g	1	0	0	1
c	0	1	0	0

$\boldsymbol{\pi}_4$

	1	2	3	4
a	0	1	0	0
t	0	0	0	0
g	0	0	1	0
c	1	0	0	1

$\boldsymbol{\pi}_5$

	1	2	3	4
a	0	0	0	0
t	0	1	0	0
g	0	0	0	0
c	1	0	1	1

$\boldsymbol{\pi}_6$

	1	2	3	4
a	0	0	0	0
t	0	1	1	0
g	0	0	0	1
c	1	0	0	0

$\boldsymbol{\pi}_7$

	1	2	3	4
a	0	0	0	0
t	1	0	1	0
g	0	1	0	0
c	0	0	0	1

$\boldsymbol{\pi}_8$

	1	2	3	4
a	0	0	0	0
t	0	1	1	0
g	0	0	0	1
c	1	0	0	0

$\boldsymbol{\pi}_9$

	1	2	3	4
a	0	0	0	0
t	0	0	1	1
g	1	0	0	0
c	0	1	0	0

$\boldsymbol{\pi}_{10}$

	1	2	3	4
a	0	0	0	0
t	0	0	0	0
g	1	1	1	0
c	0	0	0	1

$\boldsymbol{\pi}_{11}$

	1	2	3	4
a	0	0	0	1
t	0	0	1	0
g	0	0	0	0
c	1	1	0	0

$\boldsymbol{\pi}_{12}$

	1	2	3	4
a	0	0	0	0
t	0	0	0	1
g	1	0	0	0
c	0	1	1	0

$\boldsymbol{\pi}_{13}$

	1	2	3	4
a	0	0	0	1
t	0	0	1	0
g	0	0	0	0
c	1	1	0	0

$\boldsymbol{\pi}_{14}$

	1	2	3	4
a	0	0	1	0
t	0	0	0	0
g	0	0	0	1
c	1	1	0	0

$\boldsymbol{\pi}_{15}$

	1	2	3	4
a	1	1	0	0
t	0	0	0	0
g	0	0	1	1
c	0	0	0	0

Рис. 1. 4-профильные матрицы подстрок в последовательном разложении нуклеотидной строки str на подстроки длиной четыре нуклеотида.

вероятностей случайных букв может иметь произвольный вид распределения вероятностей букв алфавита. В соответствии с матрицами на рисунке 1, получаем значение $\mathbf{\Pi}_{str}(4)$ профильно-матричного спектра $\mathbf{\Pi}_{str}$ строки str на тест-периоде 4:

$$\mathbf{\Pi}_{str}(4)=\frac{1}{15}\sum_{m=1}^{15}\boldsymbol{\pi}_m=\begin{pmatrix}0.07 & 0.20 & 0.13 & 0.20\\ 0.07 & 0.27 & 0.47 & 0.20\\ 0.33 & 0.13 & 0.20 & 0.33\\ 0.53 & 0.40 & 0.20 & 0.27\end{pmatrix}.$$

Аналогичным образом вычисляются остальные матрицы профильно-матричного спектра для рассматриваемой нуклеотидной строки str. В результате получаем профильно-матричный спектр $\mathbf{\Pi}_{str}=\left(\mathbf{\Pi}_{str}(1),\mathbf{\Pi}_{str}(2),\mathbf{\Pi}_{str}(3),\mathbf{\Pi}_{str}(4)\right)$, матрицы которого показаны на рисунке 2.

$\Pi_{str}(1)$

	1
a	0.15
t	0.25
g	0.25
c	0.35

$\Pi_{str}(2)$

	1	2
a	0.10	0.20
t	0.27	0.23
g	0.27	0.23
c	0.36	0.34

$\Pi_{str}(3)$

	1	2	3
a	0.15	0.20	0.10
t	0.25	0.15	0.35
g	0.10	0.25	0.40
c	0.50	0.40	0.15

$\Pi_{str}(4)$

	1	2	3	4
a	0.07	0.20	0.13	0.20
t	0.07	0.27	0.47	0.20
g	0.33	0.13	0.20	0.33
c	0.53	0.40	0.20	0.27

Рис. 2. Матрицы профильно-матричного спектра строки str, показанной на рисунке 1.

Пусть последовательное разбиение на подстроки длины L случайной строки Str из независимых случайных букв имеет вид

$$Str=Str_n(\boldsymbol{\pi})=\underbrace{Str_L(\boldsymbol{\pi}_0)\,Str_L(\boldsymbol{\pi}_0)\ldots\,Str_L(\boldsymbol{\pi}_0)}_{q\text{ раз}}Str_M(\boldsymbol{\pi}_1),$$

где $q>2$, $M<L$, $Str_M(\boldsymbol{\pi}_1)$ – пустая строка, если $M=0$, и $\boldsymbol{\pi}_0=(\boldsymbol{\pi}_1,\boldsymbol{\pi}_{01})$, если $M\neq 0$. Кроме того, строка $Str_L(\boldsymbol{\pi}_0)$ не представима в виде $Str_L(\boldsymbol{\pi}_0)=\underbrace{Str^1\ldots Str^1}_{\tau\text{ раз}}$, где Str^1 – случайная строка длиной $k<L$. Тогда строка Str называется L-профильной строкой со случайным паттерном периодичности $Ptn_L(\boldsymbol{\pi}_0)=Str_L(\boldsymbol{\pi}_0)$. 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 определяет весь профильно-матричный спектр этой строки. Следовательно, если n кратно L и $n=qL$, то

$$Tdm_L(\boldsymbol{\pi}_0,n)=\underbrace{Str_L(\boldsymbol{\pi}_0)Str_L(\boldsymbol{\pi}_0)\ldots\,Str_L(\boldsymbol{\pi}_0)}_{q\text{ раз}}.$$

Если $L=1$, т.е.

$$Tdm_1(\mathbf{p},n)=\underbrace{Chr(\mathbf{p})\,Chr(\mathbf{p})\ldots\,Chr(\mathbf{p})}_{n\text{ раз}},$$

то 1-профильная строка называется однородной строкой, поскольку её паттерн состоит из одной случайной буквы.

2.2. Спектрально-статистический подход к выявлению скрытой профильной периодичности

Спектрально-статистический подход к выявлению скрытой профильной периодичности в текстовых строках, частным случаем которых являются последовательности ДНК, опирается на рассмотренную выше модель профильной периодичности.

Математическая биология и биоинформатика. 2016. Т. 11. № 1. doi: 10.17537/2016.11.24

Пусть Str – анализируемая строка длины n в алфавите A. Строка Str может быть как случайной, так и текстовой строкой. Для оценки её предполагаемого периода Λ используется спектр $\boldsymbol{\psi}_\Lambda$ статистики, которая для строки Str на тест-периоде λ принимает значения:

$$\psi_\Lambda(\lambda) = \frac{n}{\lambda}\sum_{j=1}^{\lambda}\sum_{i=1}^{K}(\pi^*{}_j^i - \pi_j^i)^2 \big/ \pi_j^i\,, \qquad (1)$$

где $(\pi^*{}_j^i)_\lambda^K = \boldsymbol{\Pi}_{Str}(\lambda)$ – λ-профильная матрица строки Str, $(\pi_j^i)_\lambda^K = \boldsymbol{\Pi}_{Tdm_\Lambda}(\lambda)$ – λ-профильная матрица Λ-профильной строки $Tdm_\Lambda = Tdm_\Lambda(\boldsymbol{\Pi}_{Str}(\Lambda), n)$.

Величина $\psi_\Lambda(\lambda)$ (1) описывает стандартную статистику Пирсона [22], с помощью которой проверяют гипотезу о том, что строку Str на тест-периоде λ можно рассматривать как реализацию Λ-профильной строки Tdm_Λ. Следовательно, спектр $\boldsymbol{\psi}_\Lambda$ является спектром сравнения анализируемой строки Str с Λ-профильной строкой, главная матрица которой совпадает с Λ-профильной матрицей анализируемой строки. В частности, если $\Lambda = 1$, то $\boldsymbol{\psi}_1$ является спектром сравнения анализируемой строки Str с однородной (1-профильной) строкой, вероятности букв текстового алфавита которой совпадают с частотами (вероятностями) этих букв в анализируемой строке. Спектр $\boldsymbol{\psi}_1$ называется главным спектром анализируемой строки Str.

На рисунке 3 приведены главные спектры 12-профильной строки (рис. 3,а) и её реализации (рис. 3,б). Из рисунков следует, что на каждом тест-периоде λ отличие выборочного главного спектра реализации от главного спектра 12-профильной строки имеет вид реализации случайной величины, сходной со случайной величиной,

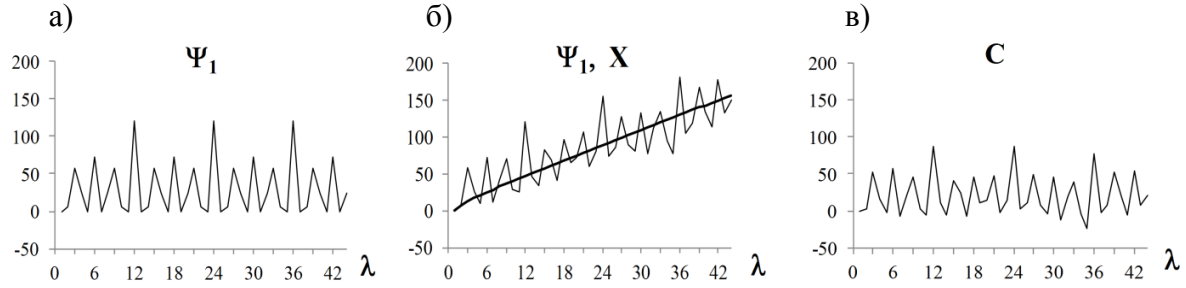


Рис. 3. Оценка периода в профильных и текстовых строках. а) Главный спектр 12-профильной строки. б) Главный (тонкая линия) спектр текстовой «реализации» (CDS; KEGG, eco:b0813, 888 нукл.) этой профильной строки и спектр (жирная линия) значений правой критической границы χ^2-распределения с $N = 3(\lambda - 1)$ степенями свободы на уровне значимости $\alpha = 0.05$. в) Характеристический спектр рассматриваемой CDS (KEGG, eco:b0813, 888 нукл.).

имеющей χ^2-распределение с $N = (\lambda - 1)(K - 1)$ степенями свободы. Поэтому на рисунке 3,б показан дополнительный график спектра $\mathbf{X}$, который на тест-периоде λ принимает значение $X(\lambda) = \chi^2_{crit}(N, \alpha)$ правой критической границы χ^2-распределения с $N = (\lambda - 1)(K - 1)$ степенями свободы на уровне значимости $\alpha = 0.05$. Чтобы устранить основное отличие главных спектров профильной строки и её реализации, для анализируемой текстовой строки $str = Str$ с помощью спектра $\boldsymbol{\psi}_1$ (см. (1), где $\Lambda = 1$) вводится характеристический спектр $\mathbf{C}$ (см. рис. 3,в), принимающий на тест-периоде λ значение:

$$C(\lambda) = \psi_1(\lambda) - M(\chi^2_{(K-1)(\lambda-1)})\,, \qquad (2)$$

где $M(\chi^2_N)$ – математическое ожидание χ^2-распределения с N степенями свободы. Поэтому первый тест-период L с ярко выраженным максимальным значением характеристического спектра $\mathbf{C}$ служит оценкой скрытого периода в строке str, если эта строка признана неоднородной. На рисунке 3,в приведён пример характеристического спектра одной из CDS генома *E. coli*, который сходен с главным спектром 12-профильной строки, показанным на рисунке 3,а.

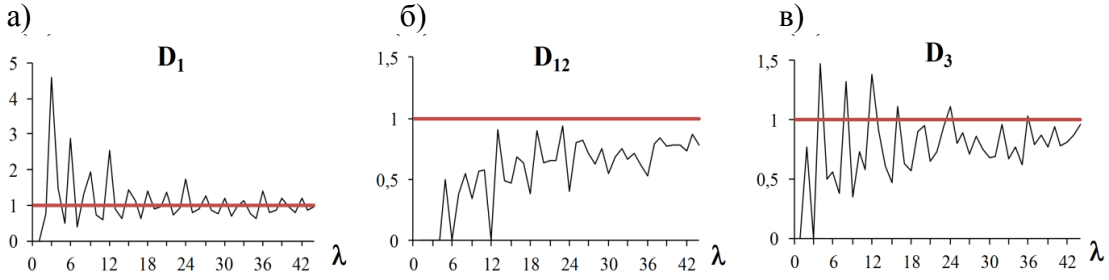
Для реализаций Λ-профильной строки на тест-периоде λ справедливо соотношение:

$$\psi_\Lambda(\lambda) \sim \chi^2_{(K-1)(\lambda-1)}, \tag{3}$$

где $\chi^2_N - \chi^2$-распределение с N степенями свободы. Для анализируемой текстовой строки $str = Str$, используя соотношение (3), вводится спектр $\mathbf{D_1}$ отклонения от однородности, принимающий на тест-периоде λ значение:

$$D_1(\lambda) = \psi_1(\lambda) / \chi^2_{crit}((K-1)(\lambda-1), \alpha), \quad \alpha = 0.05. \tag{4}$$

Строка str признаётся *неоднородной текстовой строкой*, если в диапазоне её тест-периодов условие $\mathbf{D_1} > 1$ справедливо более чем для 5 % тест-периодов. В противном случае строка str признаётся *однородной текстовой строкой*. Например, согласно такому правилу, CDS из генома бактерии *E. coli* признаётся неоднородной (см. рис. 4,а).



г) $\Pi_{str}(12)$

	1	2	3	4	5	6	7	8	9	10	11	12
a	0.16	0.19	0.19	0.18	0.09	0.20	0.14	0.08	0.16	0.31	0.15	0.28
t	0.15	0.43	0.18	0.20	0.43	0.31	0.20	0.34	0.35	0.23	0.41	0.22
g	0.31	0.11	0.46	0.36	0.20	0.26	0.51	0.31	0.30	0.19	0.07	0.36
c	0.38	0.27	0.18	0.26	0.27	0.23	0.15	0.27	0.19	0.27	0.38	0.14

Рис. 4. Спектры 2С-подхода для CDS из генома бактерии *E. coli* (KEGG, eco:b0813, 888 нукл.). а) Спектр отклонения от однородности. б) Спектр отклонения от 12-профильной периодичности. в) Спектр отклонения от 3-профильной периодичности. г) Главная матрица $\mathbf{\Pi}_{str}(12)$ выборочного профильно-матричного спектра анализируемой CDS, рассматриваемой в качестве реализации профильной строки $Tdm_{12}(\mathbf{\Pi}_{str}(12), n)$, где $n = 888$.

Для текстовой строки str оценка $L > 1$ признаётся размером периода скрытой профильной периодичности, если строка str статистически неотличима от L-профильной строки $Tdm_L(\mathbf{\Pi}_{str}(L), n)$. Для этого, согласно соотношению (3), используется спектр $\mathbf{D}_L$ отклонения строки str от L-профильности, принимающий на тест-периоде λ значение:

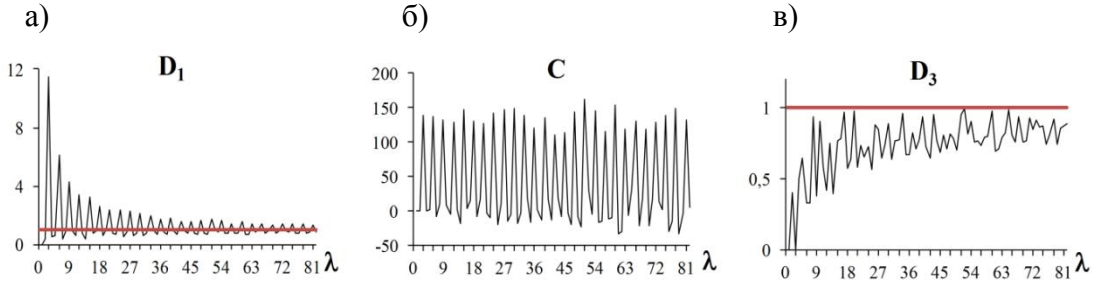
$$D_L(\lambda) = \psi_L(\lambda) / \chi^2_{crit}((K-1)(\lambda-1), \alpha), \quad \alpha = 0.05. \tag{5}$$

Если строка *str* признана неоднородной текстовой строкой и в диапазоне тест-периодов строки условие $D_L(\lambda) < 1$ справедливо более чем для 95 % тест-периодов, тогда принимается гипотеза о существовании в строке *str* скрытой L-профильной периодичности и паттерн строки $Tdm_L(\mathbf{\Pi}_{str}(L), n)$ служит оценкой паттерна скрытой L-профильной периодичности. В противном случае эта гипотеза отвергается.

Приведём пример выявления скрытой 12-профильности для CDS, характеристический спектр которой показан на рисунке 3,в. Поскольку эта последовательность признаётся неоднородной (см. рис. 4,а), её характеристический спектр (рис. 3,в) указывает на оценку скрытого периода в 12 нуклеотидов. Спектр отклонения от 12-профильности (рис. 4,б) подтверждает эту оценку. Отсутствие в этой последовательности скрытой 3-профильности демонстрирует спектр отклонения от 3-профильности на рисунке 4,в. На рисунке 4,г приведена главная матрица профильно-матричного спектра анализируемой CDS, которая является оценкой паттерна 3-профильной строки, в качестве реализации которой рассматривается эта CDS.

2.3. Регулярность в текстовых строках

Ранее [14], для характеристических спектров неоднородных кодирующих последовательностей ДНК отмечалось регулярное повторение пиков на тест-периодах, кратных трём (см. рис 3,в, рис. 5,б и рис. 6,б). В общем случае, это явление, в отличие от скрытой профильности, было названо 3-регулярностью последовательностей ДНК. Феномен 3-регулярности может наблюдаться как в последовательностях обладающих 3-профильностью, так и в последовательностях, в которых 3-профильность отсутствует. Так, например, на рисунке 5 приведёны спектры 2С-подхода, согласно которым в CDS бактерии *B. burgdorferi* наблюдается скрытая профильная триплетная периодичность.



г) $\Pi_{str}(3)$

	1	2	3
a	0.36	0.41	0.36
t	0.18	0.34	0.43
g	0.36	0.14	0.14
c	0.11	0.11	0.07

Рис. 5. Спектры 2С-подхода для CDS из генома бактерии *B. burgdorferi* (KEGG, bbz:BbuZS7_0127, 1656 нукл.). а) Спектр отклонения от однородности. б) Характеристический спектр. в) Спектр отклонения от 3-профильной периодичности. г) Главная матрица $\mathbf{\Pi}_{str}(3)$ выборочного профильно-матричного спектра анализируемой CDS, рассматриваемой в качестве реализации профильной строки $Tdm_3(\mathbf{\Pi}_{str}(3), n)$, где $n = 1656$.

Согласно спектру $\mathbf{D_1}$ (рис. 5,а) последовательность CDS признаётся неоднородной. Характеристический спектр $\mathbf{C}$ (рис. 5,б) позволяет предложить в качестве оценки длины скрытого периода три нуклеотида. Спектр отклонения от 3-профильности $\mathbf{D_3}$ подтверждает справедливость этой оценки. На рисунке 5,г показана главная матрица

профильно-матричного спектра анализируемой CDS, которая представляет собой оценку паттерна профильной строки, в качестве реализации которой рассматривается эта CDS. Однако, согласно аналогичным спектрам, показанным на рисунке 6, в CDS генома бактерии *H. pylori* можно утверждать всего лишь о наличии феномена 3-регулярности в отсутствии скрытой профильной триплетной периодичности, поскольку спектр $\mathbf{D}_3$ отклонения от 3-профильности не подтверждает оценку в три нуклеотида, на которую указывает характеристический спектр $\mathbf{C}$ (рис. 6,б) этой CDS. Таким образом, характерное чередование пиков в каком-либо спектре анализируемой CDS не является достаточным признаком наличия в этой CDS скрытой триплетной периодичности.

Ранее [14], проявление 3-регулярности объяснялось триплетной природой генетического кода. В работах [16, 19, 20] было показано, что практически все CDS генома человека и других эукариотических организмов являются 3-регулярными. Далее будет продемонстрировано, что к аналогичному результату приводит анализ CDS геномов прокариот, рассматриваемых в работе.

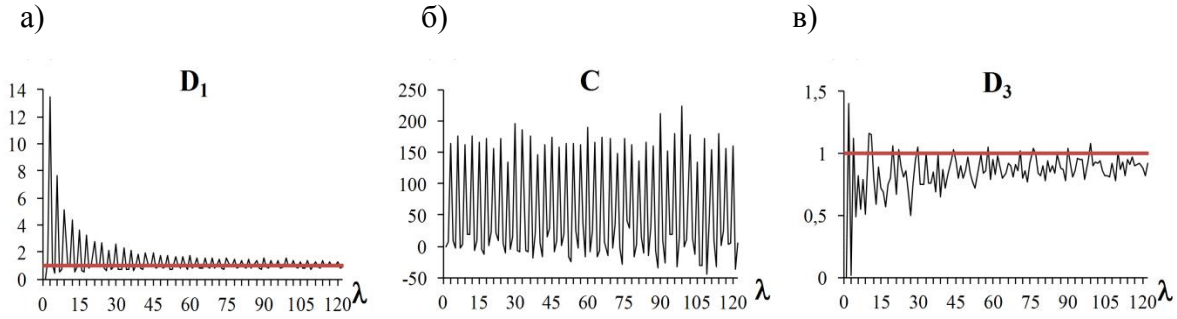


Рис. 6. Спектры 2С-подхода для CDS из генома бактерии *H. pylori* (KEGG, hpy:HP1403, 2454 нукл.). а) Спектр отклонения от однородности. б) Характеристический спектр. в) Спектр отклонения от 3-профильной периодичности.

В работе [14] был приведён пример кодирующей 3-регулярной последовательности ДНК из гена CYA бактерии *Bordetella pertussis* (GenBank, Y00545, 981–6101 нукл.), в которой отсутствовала скрытая профильная периодичность. Однако, скрытая профильность с периодом, кратным трём, была выявлена в нескольких локальных районах этой последовательности. Следовательно, глобальную 3-регулярность последовательности можно наблюдать при последовательном сцеплении районов, в которых выявляется скрытый профильный период, кратный трём. В этом случае 3-регулярность последовательности обусловлена следующими причинами. Характеристические спектры её локальных районов являются 3-регулярными, но паттерны скрытой периодичности этих районов различны. Следовательно, глобальная 3-регулярность может быть связана с локальной скрытой профильностью в кодирующем районе ДНК. Также возможно, что 3-регулярность последовательности отражает проявление статистически ослабленной 3-профильной периодичности. Поэтому в спектрально-статистическом подходе наряду с выявлением скрытой профильной периодичности проводится анализ наличия регулярности в текстовых строках.

В общем случае, в характеристических спектрах неоднородных текстовых строк может наблюдаться регулярная повторяемость пиков на тест-периодах, кратных натуральному числу $R \neq 1$. Такое явление будет называться R-регулярностью текстовых строк.

Введём критерий наличия R-регулярности в неоднородной текстовой строке. Для этого разобьём диапазон тест-периодов характеристического спектра анализируемой текстовой строки на последовательные группы из R тест-периодов. Для каждой такой группы тест-периоду с максимальным значением характеристического спектра ставится

в соответствие единица и нули – остальным тест-периодам группы. В результате образуется бинарная строка из нулей и единиц, т.е. текстовая строка str в алфавите $A=\{0,1\}$ размера $K=2$. Эта строка сравнивается с совершенной периодической строкой той же длины, паттерн периодичности которой имеет вид: $\underbrace{0\ldots0}_{R-1\text{ раз}}1$. Индексом I_R – индексом R-регулярности анализируемой последовательности называется число, равное отношению количества совпадений компонент совершенной периодической строки и бинарной строки str к длине анализируемой строки. При исследовании конкретной R-регулярности вводится соответствующее пороговое значение I_R^*, обоснование которого будет приведено далее. Если для анализируемой текстовой строки $I_R \geq I_R^*$, то для этой строки принимается гипотеза о наличии в ней R-регулярности. В противном случае гипотеза о наличии R-регулярности отвергается.

2.4. Скрытая профильная периодичность и стохастическая модель однородной организации кодирования (SHOC-модель) в текстовых строках

В настоящем разделе рассматривается стохастическая модель организации кодирования в текстовых строках, названная стохастической моделью однородной организации кодирования (Stochastic Homogeneous Organization of Coding – SHOC-модель). Эта модель, вследствие наличия в текстовых строках некоторых семантических единиц, приводит к проявлению в них скрытой профильной периодичности и регулярности. Проведённые в рамках SHOC-модели исследования генетических и перекодированных литературных текстов показали, что скрытая в них профильная периодичность и регулярность действительно обусловлена некоторыми семантическими единицами, на основе которых сконструированы эти тексты.

В рассматриваемой стохастической модели однородной организации кодирования (SHOC-модели) предполагается, что кодирование осуществляется строками (кодонами) одинаковой длины $\Lambda \geq 1$, состоящими из букв алфавита $A=\{a_1,a_2,...,a_K\}$. На совокупности $W_\Lambda^1(A)$ таких кодонов задано вероятностное распределение P_1^W. Таким образом, вводится случайный кодон Cdn, для которого $P_1^W(w)$ – вероятность реализации кодона $w \in W_\Lambda^1(A)$. В качестве SHOC-модели организации кодирования предлагается схема из r независимых испытаний случайного кодона Cdn, которая обозначается списком $(Cdn_1, Cdn_2, \ldots, Cdn_r)$, где $Cdn_m \sim Cdn$ – случайный кодон m-го испытания, тождественный кодону Cdn для $m=\overline{1,r}$. Если в m-ом испытании схемы реализуется кодон w_m ($m=\overline{1,r}$), то, согласно предлагаемой SHOC-модели, реализуется текстовая строка $str = w_1 w_2 ... w_r \in W_\Lambda^r(A)$, где $W_\Lambda^r(A)$ – совокупность текстовых строк длины $r\cdot\Lambda$ в алфавите $A=\{a_1,a_2,...,a_K\}$. Кроме того, согласно SHOC-модели, $P_r^W(str) = P_1^W(w_1)\cdot P_1^W(w_2)\cdot ...\cdot P_1^W(w_r)$ – вероятность появления текстовой строки str из совокупности $W_\Lambda^r(A)$, где P_r^W – вероятностное распределение текстовых строк из совокупности $W_\Lambda^r(A)$.

Рассмотрим введенный выше случайный кодон Cdn. Согласно распределению P_1^W, в каждой позиции кодона $j=\overline{1,\Lambda}$ индуцируется случайная буква $Chr(\mathbf{p}_j)$ со значениями букв из алфавита A, где $\mathbf{p}_j = (p_j^1,..., p_j^K)^T$ и p_j^i – вероятность того, что при реализации кодона из совокупности $W_\Lambda^1(A)$ в его j-той позиции будет находиться

Математическая биология и биоинформатика. 2016. Т. 11. № 1. doi: 10.17537/2016.11.24

буква a_i из алфавита A для $i=\overline{1,K}$. Поэтому случайный кодон Cdn можно отождествить с введённой в разделе 2.1 случайной строкой $Str_\Lambda(\boldsymbol{\pi}) = Chr(\mathbf{p}_1)Chr(\mathbf{p}_2)\ldots Chr(\mathbf{p}_\Lambda)$, где $\boldsymbol{\pi} = (\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_\Lambda) = (\pi_j^i)_\Lambda^K$ – профильная матрица строки $Str_\Lambda(\boldsymbol{\pi})$. Такое отождествление оправдывается тем, что не рассматриваются корреляционные связи между случайными буквами внутри строки $Str_\Lambda(\boldsymbol{\pi})$. Для строки $Str_\Lambda(\boldsymbol{\pi})$ возможны два варианта её представления. В первом варианте, $Str_\Lambda(\boldsymbol{\pi}) = Tdm_L(\boldsymbol{\pi}_0, \Lambda)$ – профильная строка с паттерном $Ptn_L(\boldsymbol{\pi}_0)$, для которого $Str_\Lambda(\boldsymbol{\pi}) = \underbrace{Ptn_L(\boldsymbol{\pi}_0)\ldots Ptn_L(\boldsymbol{\pi}_0)}_{q\text{ раз}}$, $\Lambda = Lq$ и $q>1$. Во втором варианте строку $Str_\Lambda(\boldsymbol{\pi})$ нельзя представить в виде последовательного разложения на одинаковые, более короткие, случайные подстроки. В этом случае вводим обозначения: $q=1$, $L=\Lambda$ $(\Lambda = Lq)$, $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$ и $Ptn_L(\boldsymbol{\pi}_0) = Str_\Lambda(\boldsymbol{\pi})$. Таким образом, и в первом, и во втором варианте схему из r независимых испытаний кодона Cdn, являющуюся моделью организации кодирования строки длины $n = rqL$ в алфавите A, можно отождествить с введённой в разделе 2.1 L-профильной строкой $Tdm_L(\boldsymbol{\pi}_0, n)$ с паттерном $Ptn_L(\boldsymbol{\pi}_0)$, где $Tdm_L(\boldsymbol{\pi}_0, n) = \underbrace{Ptn_L(\boldsymbol{\pi}_0)\ldots Ptn_L(\boldsymbol{\pi}_0)}_{rq\text{ раз}}$. Таким образом, SHOC-модель организации кодирования в рамках спектрально-статистического подхода (2С-подхода) допустимо отождествлять с соответствующей профильной строкой. Поэтому, если в анализируемой текстовой строке выявляется скрытая L-профильная периодичность, то такая периодичность может быть обусловлена проявлением SHOC-модели с кодонами, длина которых совпадает с длиной скрытого периода или кратна ему.

В настоящей работе предлагается гипотеза об организации структурного кодирования функциональных районов последовательностей ДНК согласно введённой выше SHOC-модели. Эта модель описывает тот случай, когда кодирование семантических единиц осуществляется кодонами одного размера. Далее, будет показано, что в рамках SHOC-модели размер кодонов L в анализируемых текстовых строках объясняет наличие в них L-профильности и L-регулярности.

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В настоящем разделе приведены результаты применения спектрально-статистического подхода к CDS с подтверждаемой функциональной активностью из базы данных KEGG [18], а также к последовательностям интронов из базы данных EID [21] и к набору всех абзацев бинарно перекодированного художественного текста на английском языке.

3.1. Сравнительный анализ наличия 3-регулярности в CDS и интронах генома человека

Согласно разделу 2.3 для выявления 3-регулярности в последовательности ДНК необходимо установить пороговое значение для индекса 3-регулярности. Только те последовательности, в которых индекс 3-регулярности превышает пороговое значение, признаются 3-регулярными.

На рисунке 7 приведены процентные распределения встречаемости значений индекса 3-регулярности для неоднородных последовательностей интронов (см. рис. 7,а) и для CDS из генома человека (см. рис. 7,б). Из анализа этих распределений был выбран критический уровень индекса 3-регулярности, равный 0.7. Следовательно, если

индекс 3-регулярности $I_3 > 0.7$, то признаётся, что анализируемая последовательность является 3-регулярной.

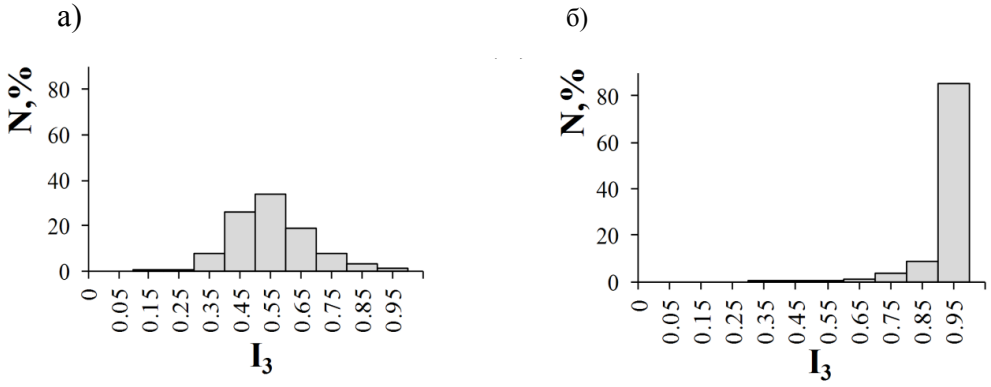


Рис. 7. Процентные распределения встречаемости значений индекса 3-регулярности I_3 среди 70966 неоднородных интронов (а) и среди 17158 неоднородных CDS генома человека (б).

3.2. Результаты применения спектрально-статистического подхода к последовательностям ДНК генома человека

Ранее [19, 20] нами анализировались CDS генома человека из базы KEGG [18] и последовательности интронов генома человека из базы EID [21]. На рисунке 8 представлены обобщённые результаты этого анализа.

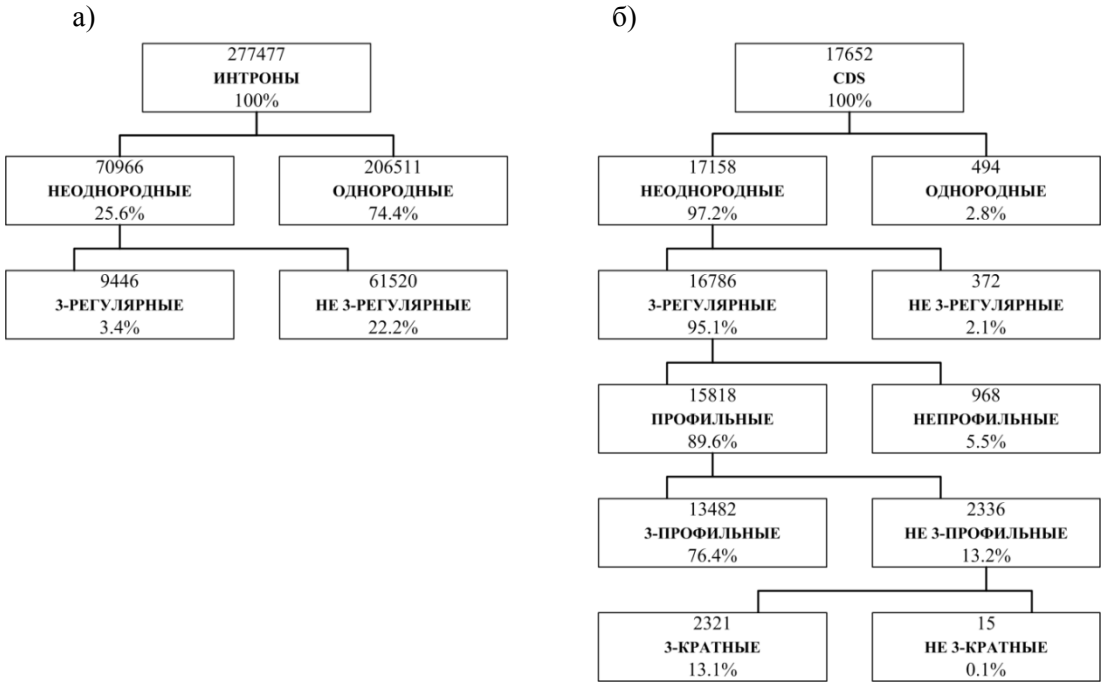


Рис. 8. Дендрограммы, представляющие количественный и структурный состав анализируемых интронов в геноме человека из базы данных EID (а) и CDS генома человека из базы данных KEGG (б).

Из рисунка 8,б следует, что практически все CDS, с точностью до статистической погрешности, являются 3-регулярными последовательностями. В отличие от CDS, практически все последовательности интронов не являются 3-регулярными (см. рис. 8,а). В большей части CDS (~ 90 %) выявляется скрытая профильная периодичность с периодом, кратным трём. При этом в 76 % CDS выявляется скрытая

триплетная профильная периодичность. Кроме того, в 13 % CDS выявляется скрытая не триплетная профильная периодичность с периодом, кратным трём.

3.3. Результаты применения спектрально-статистического подхода к CDS геномов прокариот

Анализ неоднородности, 3-регулярности и скрытой профильной периодичности был выполнен для CDS десяти геномов бактерий (табл. 1), являющихся патогенами для человека и/или животных. Количественные результаты этого анализа приведены в таблицах 2 и 3, структура которых соответствует структуре дендрограммы для CDS генома человека на рисунке 8,б. Таблица 2 отражает анализ CDS с достоверно известной функциональной активностью (Выборка 1). В таблице 3 приведены результаты анализа функционально активных CDS совместно с различного рода предсказанными или предполагаемыми CDS (Выборка 2) в рассматриваемых геномах прокариот.

Таблица 2. Количественное распределение CDS Выборки 1 для десяти прокариотических организмов (соответствующие кодам названия организмов см. в таблице 1) согласно выявленным структурно-статистическим свойствам

Код организма	bbz	bsu	buc	eco	eta	hin	hpy	mmy	sty	ype
CDS	703	2138	513	3308	2323	1185	940	605	2512	2113
неоднородные	684	2079	507	3184	2282	1168	916	604	2437	2056
однородные	19	59	6	124	41	17	24	1	75	57
3-регулярные	676	2053	503	3120	2262	1164	887	604	2407	2024
не 3-регулярные	8	26	4	64	20	4	29	0	30	32
профильные	624	1911	478	2841	2086	1077	805	574	2217	1860
непрофильные	52	142	25	279	176	87	82	30	190	164
3-профильные	537	1617	436	2388	1766	922	661	514	1874	1601
не 3-профильные	87	294	42	453	320	155	144	60	343	259
3-кратные	87	291	42	445	319	154	139	60	341	247
не 3-кратные	0	3	0	8	1	1	5	0	2	12

Таблица 3. Количественное распределение CDS Выборки 2 для десяти прокариотических организмов (соответствующие кодам названия организмов см. в таблице 1) согласно выявленным структурно-статистическим свойствам

Код организма	bbz	bsu	buc	eco	eta	hin	hpy	mmy	sty	ype
CDS	1374	4243	582	4309	3695	1708	1708	1016	4224	4224
неоднородные	1264	4001	572	4118	3516	1645	1645	998	3965	3965
однородные	110	242	10	191	179	63	63	18	259	259
3-регулярные	1232	3893	568	4012	3436	1623	1623	992	3827	3827
не 3-регулярные	32	108	4	106	80	22	22	6	138	138
профильные	1132	3579	535	3653	3139	1483	1483	931	3492	3492
непрофильные	100	314	33	359	297	140	140	61	335	335
3-профильные	979	3051	489	3076	2653	1274	1274	814	2976	2976
не 3-профильные	153	528	46	577	486	209	209	117	516	516
3-кратные	150	515	46	565	476	205	205	117	492	492
не 3-кратные	3	13	0	12	10	4	4	0	24	24

Обобщённые представления результатов из таблиц 2 и 3 показаны на рисунках 9,а и 9,б, соответственно. Содержание дендрограммы на рисунке 9,а, отражающей структурно-статистические свойства прокариотических CDS, практически совпадает с содержанием дендрограммы для CDS генома человека (см. рис. 8,б). Наблюдаемые отклонения параметров дендрограммы на рисунке 9,б (для Выборки 2) от параметров дендрограммы на рисунке 9,а (для Выборки 1) возможно обусловлены неточностью предсказания кодирующих последовательностей, добавленных в Выборку 1.

Из результатов анализа CDS геномов прокариот, генома человека и других геномов эукариот [16] следует, что практически все CDS, с точностью до статистической погрешности, являются 3-регулярными последовательностями. В большей части CDS (~ 90 %) выявляется скрытая профильная периодичность с периодом, кратным трём. При этом в 75 % CDS выявляется скрытая триплетная профильная периодичность. Кроме того, в 13 % CDS выявляется скрытая нетриплетная профильная периодичность с периодом, кратным трём. Таким образом, можно считать, что широко известная в литературе гипотеза о триплетной периодичности в кодирующих районах ДНК получает подтверждение на основе понятия скрытой профильной периодичности.

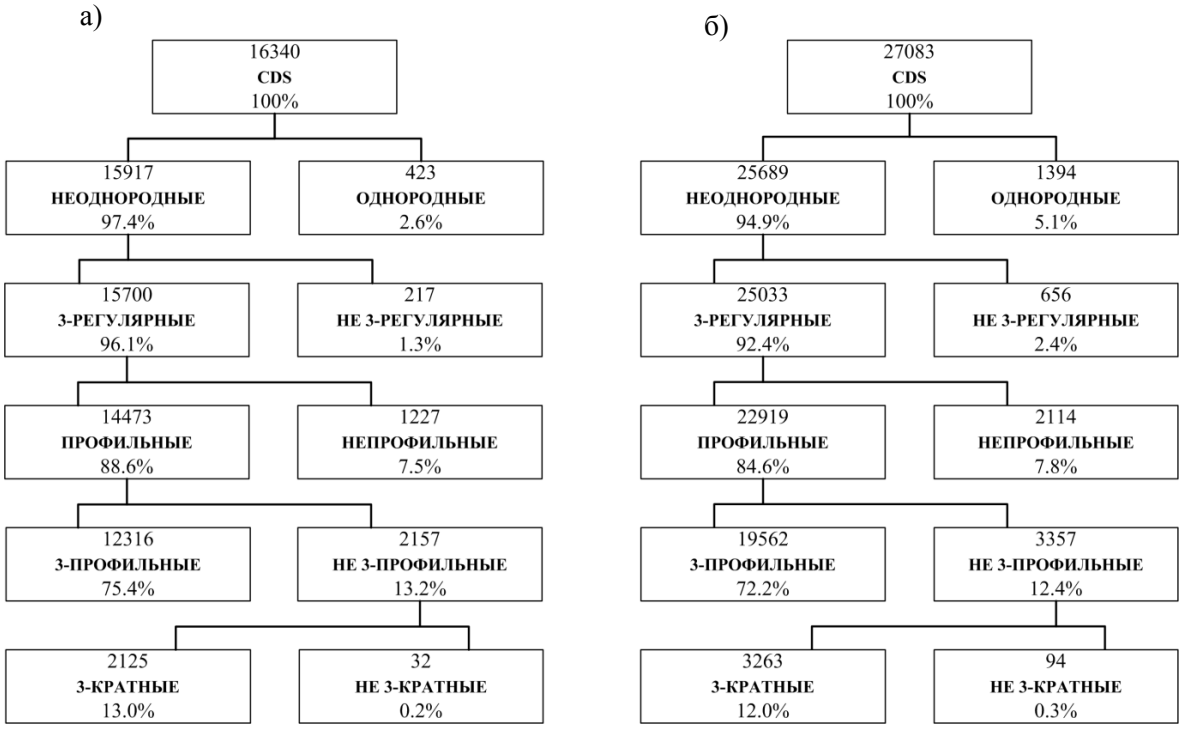


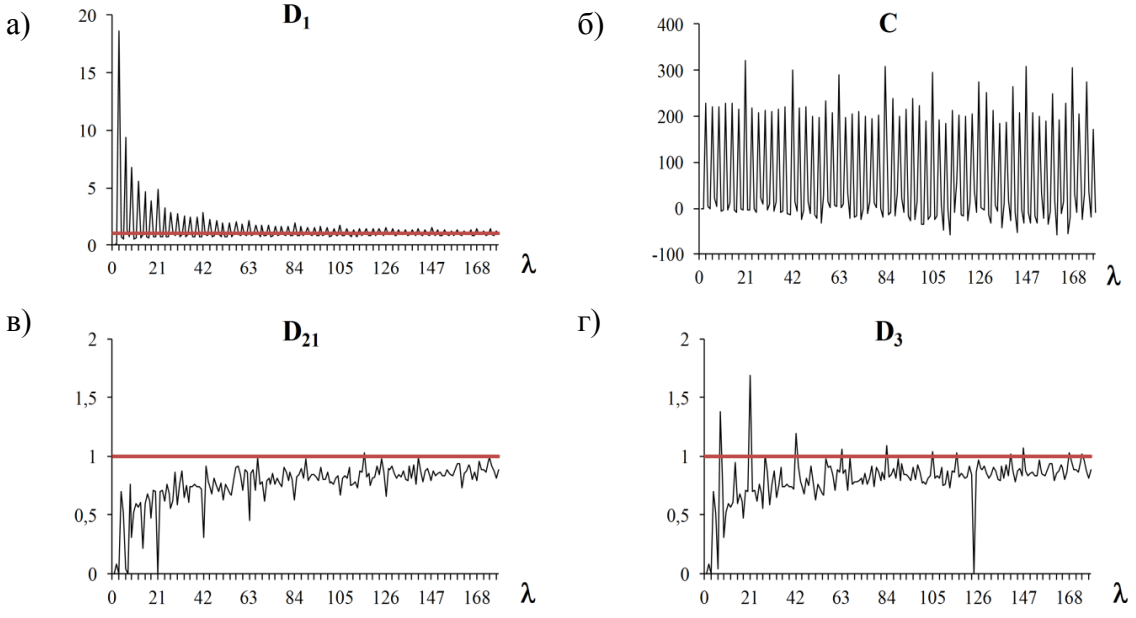
Рис. 9. Дендрограммы, обобщающие распределение структурно-статистических свойств CDS для десяти анализируемых геномов прокариот из базы данных KEGG. (а) Выборка CDS белков с подтверждённой функциональной активностью (Выборка 1, см. табл. 2). (б) Расширенная выборка CDS белков, включающая различные предсказанные CDS (Выборка 2, см. табл. 3).

3.4. Двухуровневая организация кодирования в CDS

В дендрограммах, приведенных на рисунке 8,б и рисунке 9, были выделены 3-регулярные CDS, скрытая профильная периодичность которых была кратна, но не равна трём. Для таких CDS можно говорить о двухуровневой организации кодирования. Ранее CDS с двухуровневой организацией кодирования уже выявлялись [13, 14, 23] с помощью 2С-подхода. Первый уровень организации кодирования проявляется в 3-регулярности последовательности CDS. В этом случае в последовательности может проявляться второй уровень организации кодирования, если

длина скрытого периода профильной периодичности кратна, но не равна трём. Как было показано [13, 14, 23], такая скрытая профильность в кодирующих районах ДНК может коррелировать с особенностями структуры кодируемых белков. Например, наличие скрытой профильности в 33 нукл. (33-профильности) выявляется в кодирующих районах многих генов семейства аполипопротеинов PF01442 из базы белковых семейств Pfam (http://pfam.sanger.ac.uk/) [13, 14]. Как известно, вторичная структура этого семейства содержит несколько пар α-спиралей, состоящих из 11 и 22 аминокислотных остатков, что коррелирует с профильной периодичностью генов аполипопротеинов в 33 нуклеотида. В работе [16] была выявлена скрытая профильная периодичность с паттерном 84 нуклеотида в 203 CDS белков семейства «цинковых пальцев» генома человека. Размер паттерна такой скрытой периодичности соответствует размеру повторяющегося домена «цинкового пальца», что составляет примерно 20 аминокислотных остатков.

Приведём пример скрытой профильной периодичности с паттерном длиной 21 нуклеотида в CDS белков из прокариотических организмов (см. рис. 10). Скрытая 21-профильность была выявлена в 258-ми CDS различных белков в геномах анализируемых прокариот. Рисунок 10 демонстрирует пример одной из 15 CDS метилируемых белков хемотаксиса, в которой распознаётся 21-профильная периодичность.



д) $\Pi_{str}(21)$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
a	0.23	0.51	0.35	0.29	0.45	0.28	0.32	0.30	0.35	0.31	0.45	0.32	0.30	0.43	0.27	0.25	0.38	0.27	0.26	0.40	0.25
t	0.14	0.18	0.23	0.11	0.25	0.28	0.13	0.47	0.21	0.16	0.20	0.21	0.08	0.21	0.21	0.12	0.25	0.21	0.20	0.37	0.33
g	0.46	0.11	0.23	0.39	0.14	0.27	0.26	0.14	0.26	0.38	0.14	0.26	0.36	0.15	0.25	0.39	0.13	0.32	0.30	0.12	0.24
c	0.18	0.21	0.19	0.21	0.16	0.17	0.28	0.09	0.17	0.14	0.21	0.21	0.26	0.21	0.27	0.24	0.23	0.20	0.25	0.12	0.18

Рис. 10. Выявление скрытой профильной периодичности с паттерном длиной 21 нукл. в CDS генома бактерии *B. subtilis* (KEGG, bsu:BSU15940, 3561 нукл.). а) Спектр отклонения от однородности (см. (4)). б) Характеристический спектр. в) и г) Спектры отклонения от 21- и 3-профильности, соответственно (см. (5)). д) Главная матрица $\mathbf{\Pi}_{str}(21)$ выборочного профильно-матричного спектра анализируемой CDS, рассматриваемой в качестве реализации профильной строки $Tdm_{21}(\mathbf{\Pi}_{str}(21), n)$, где $n = 3561$.

Индекс 3-регулярности характеристического спектра (см. рис. 10,б) $I_3 = 0.99$. Согласно рисунку 10,г, в этой последовательности отсутствует 3-профильная периодичность.

Спектр отклонения от 21-профильности (см. рис. 10,в) подтверждает наличие в ней 21-профильности. На рисунке 10,д показана главная матрица профильно-матричного спектра анализируемой CDS, которая представляет собой оценку паттерна профильной строки, в качестве реализации которой рассматривается эта CDS.

3.5. Исследование SHOC-модели на примере бинарного перекодирования художественного текста

Для демонстрации спектрально-статистических свойств SHOC-модели используется текст книги на английском языке «Three Men in a Boat» [17]. Естественными семантическими единицами в этом тексте являются буквы латинского алфавита и общие знаки пунктуации, которые далее называются «характерами». По аналогии с кодированием аминокислот кодонами генетического кода, эти характеры кодировались бинарными строками длиной пять, как показано в таблице 4. При перекодировании текстов не принимались во внимание пробелы между словами. В таблице 4 приведены также частоты встречаемости характеров в оригинальных текстах. Распределения частот встречаемости 287268 характеров в анализируемом литературном произведении показано на рисунке 11.

Таблица 4. Соответствие букв латинского алфавита и символов пунктуации (характеров) кодам длины 5 в бинарном алфавите {1, 0} и частоты встречаемости характеров тексте книги “Three Men in a Boat” [17]

N_C	Характер C	Бинарный код B_C	Частота характера F_C	N_C	Характер C	Бинарный код B_C	Частота характера F_C
1	A a	00000	0.07953	17	Q q	10101	0.00087
2	B b	10000	0.01471	18	R r	01011	0.05232
3	C c	01000	0.01908	19	S s	01101	0.05537
4	D d	00100	0.04674	20	T t	11010	0.09220
5	E e	00010	0.11612	21	U u	10110	0.02965
6	F f	00001	0.01946	22	V V	01110	0.00867
7	G g	11000	0.02417	23	W w	11100	0.02647
8	H h	01100	0.06032	24	X x	11001	0.00127
9	I i	00110	0.06268	25	Y y	10011	0.01833
10	J j	00011	0.00118	26	Z z	00111	0.00038
11	K k	10001	0.00885	27	–	11110	0.00390
12	L l	01001	0.03691	28	’ "	11101	0.00724
13	M m	00101	0.02163	29	,	11011	0.01915
14	N n	10010	0.06724	30	.	10111	0.01018
15	O o	10100	0.07456	31	! ?	01111	0.00147
16	P p	01010	0.01592	32	иные	11111	0.00343

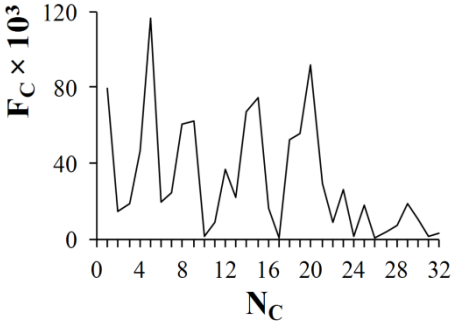


Рис. 11. Распределения частот F_C встречаемости характеров исходного латинского алфавита в анализируемом произведении «Three Men in a Boat» [17]. N_C – номер характера в в таблице 4.

Текст каждого абзаца в исходном литературном произведении, записанный в бинарном алфавите, рассматривался как отдельная бинарная последовательность. Количество абзацев в книге «Three Men in a Boat» [17] составляет 1381. Таким образом, проводилась аналогия этих абзацев с CDS из геномов. Как и для CDS, к анализу бинарно перекодированных абзацев применялся 2С-подход.

Результаты применения 2С-подхода к перекодированным абзацам представлены на рисунке 12. Анализировались только неоднородные перекодированные абзацы, составляющих основную часть (около 80 %) от всех абзацев. Неучтённые перекодированные абзацы, признанные однородными, как правило, имели недостаточную длину для статистического анализа. При выявлении 5-регулярности в неоднородных бинарных текстах перекодированных абзацев, как и в CDS генома человека, для порогового значения индекса 5-регулярности I_5 было выбрано значение, равное 0.7.

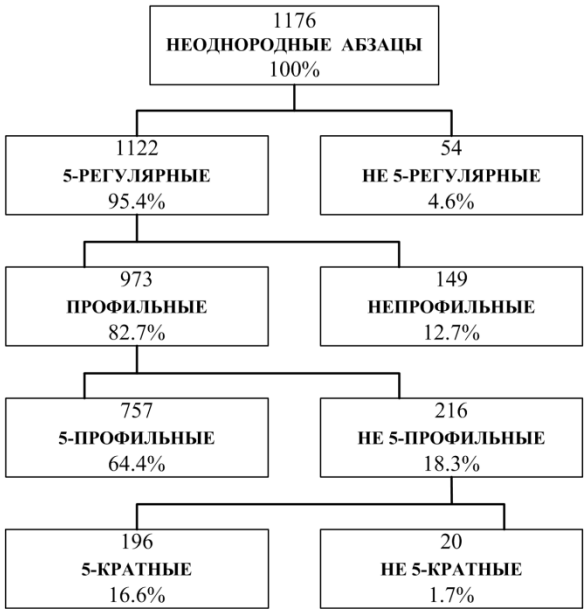


Рис. 12. Дендрограмма, демонстрирующая выявленные структурные свойства в неоднородных бинарно перекодированных (длина бинарного кода 5 символов, см. таблицу 4) абзацах из английского литературного текста «Three Men in a Boat» [17].

Дендрограмма на рисунке 12, представляющая результаты применения 2С-подхода к анализу перекодированного литературного текста, аналогична дендрограммам на рисунке 8,б и рисунке 9. Главное их отличие состоит в том, что CDS из геномов человека и прокариотических организмов являются 3-регулярными, в то время как в перекодированных абзацах литературного текста наблюдается 5-регулярность, обусловленная размером бинарных кодонов, состоящих из пяти символов. Вследствие

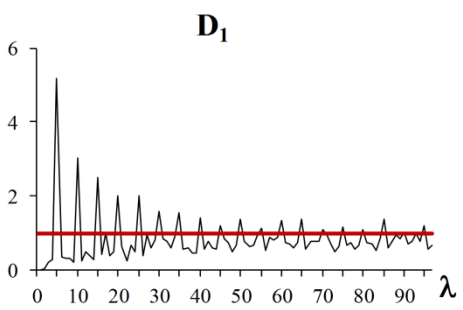
выбранной организации кодирования литературного текста, можно считать, что для каждого перекодированного и неоднородного абзаца справедлива SHOC-модель с бинарными кодонами размера пять. Сходство дендрограмм на рисунке 8,б, рисунке 9 и рисунке 12 совместно с результатами работы [16] позволяет предположить справедливость SHOC-модели с кодонами размера три в алфавите $\{a,t,g,c\}$ для CDS геномов прокариот, человека и других эукариотических организмов. Согласно дендрограмме на рисунке 12, как и в CDS геномов, в перекодированных абзацах литературного текста встречается двухуровневая организация кодирования.

Для дополнительной демонстрации аналогии между перекодированными абзацами литературного текста и CDS из геномов прокариот, человека и других эукариот приведём примеры применения спектрально-статистического подхода к перекодированным абзацам.

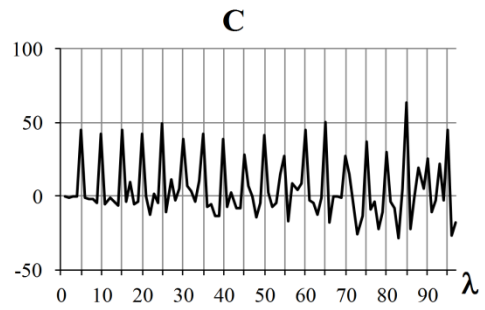
а)

Ittooktwoportersaswellasthedrivertoholdhiminatthe
station;andIdonotthinktheywouldhavedoneit,eventhen,
hadnotoneofthemenhadthepresenceofmindtoputa
handkerchiefoverhisnose,andtolightabitofbrownpaper.

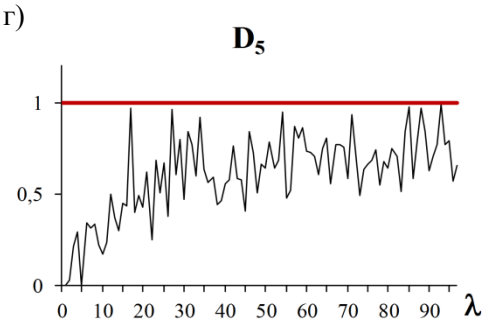
б)



в)



г)



д)

$\Pi_{str}(5)$

	1	2	3	4	5
0	0.59	0.59	0.57	0.47	0.81
1	0.41	0.41	0.43	0.53	0.19

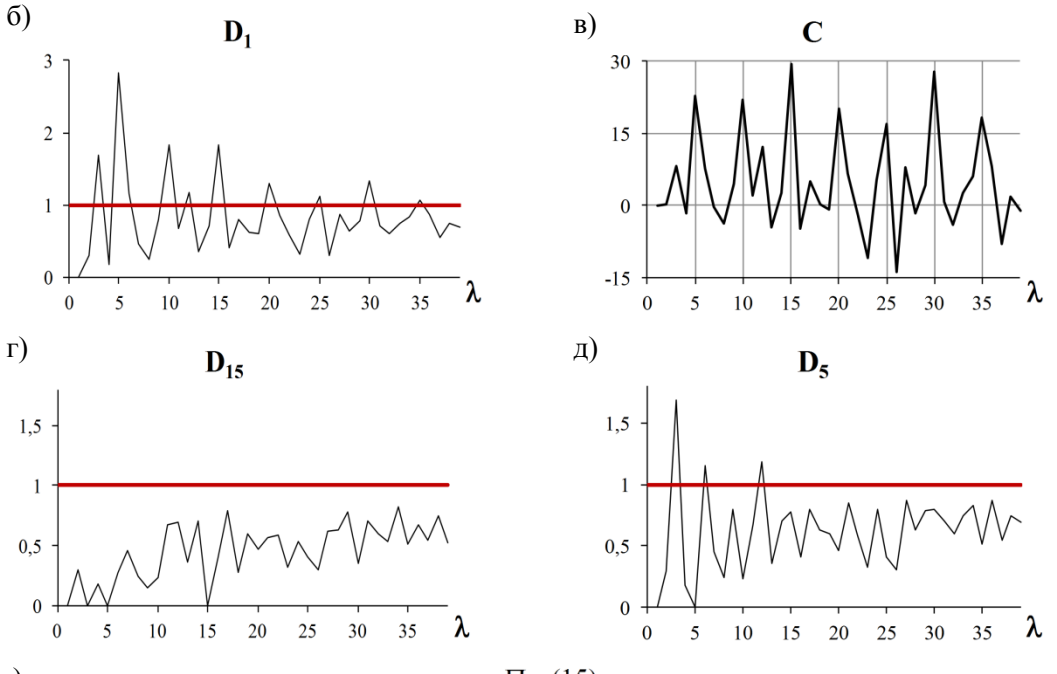
Рис. 13. а) Исходный английский текст 14-го абзаца 4-й главы книги «Three Men in a Boat» [17] и (б–г) спектры этого бинарно перекодированного абзаца. Индекс 5-регулярности $I_5 = 0.99$, длина бинарного текста после перекодирования абзаца – 970 символов. д) Главная матрица $\mathbf{\Pi}_{str}(5)$ выборочного профильно-матричного спектра анализируемого бинарного текста, рассматриваемого в качестве реализации профильной строки $Tdm_5(\mathbf{\Pi}_{str}(5), n)$, где $n = 970$.

На рисунках 13,а и 14,а представлены исходные тексты анализируемых абзацев. Кроме того, для их закодированных бинарных аналогов приводятся спектры 2C-подхода, выявляющие 5-регулярность (рис. 13,в и рис. 14,в) и соответствующую скрытую профильную периодичность (рис. 13,г и рис. 14,г). На рисунках 13,д и 14,е показаны главные матрицы профильно-матричных спектров анализируемых бинарно перекодированных абзацев, которые являются оценками паттернов 5-профильной и 15-

Математическая биология и биоинформатика. 2016. Т. 11. № 1. doi: 10.17537/2016.11.24

профильной строк, в качестве соответствующих реализаций которых рассматриваются эти абзацы.

а) Andthenheadded,withatouchofsadnessinhisvoice:
'Iwishhehadn'tbeencarvingthatpie.'



е) $\Pi_{str}(15)$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0.67	0.41	0.44	0.67	0.81	0.73	0.77	0.46	0.42	0.96	0.58	0.50	0.54	0.46	0.73
1	0.33	0.59	0.56	0.33	0.19	0.27	0.23	0.54	0.58	0.04	0.42	0.50	0.46	0.54	0.27

Рис. 14. а) Исходный английский текст 115-го абзаца 13-той главы книги «Three Men in a Boat» [17] и б)–д) спектры этого бинарного перекодированного (табл. 4) абзаца. Индекс 5-регулярности $I_5 = 0.97$, длина бинарного текста после перекодирования абзаца – 395 символов. е) Главная матрица $\mathbf{\Pi}_{str}(15)$ выборочного профильно-матричного спектра анализируемого бинарного текста, рассматриваемого в качестве реализации профильной строки $Tdm_{15}(\mathbf{\Pi}_{str}(15), n)$, где $n = 395$.

Спектры рисунка 14 демонстрируют наличие в перекодированном абзаце двухуровневой организации кодирования, которая отражена в дендрограмме на рисунке 12, как профильность с периодом, кратным пяти. В отсутствии 5-профильности (рис. 14,д), на фоне 5-регулярности характеристического спектра (рис. 14,в) выявляется скрытая 15-профильность (рис. 14,г).

4. ЗАКЛЮЧЕНИЕ

В настоящей работе дано обоснование широко распространённой в литературе гипотезы о скрытой триплетной периодичности в кодирующих районах ДНК. На основе результатов анализа CDS из геномов прокариотических организмов, геномов человека и других эукариотических организмов впервые показано, что эта скрытая периодичность относится к типу профильной периодичности.

В работе была предложена стохастическая модель однородной организации кодирования (SHOC-модель) в последовательностях ДНК, объясняющая проявление в CDS скрытой профильной триплетной периодичности. В рамках этой стохастической модели кодирование семантических единиц осуществляется последовательно, кодонами одного размера. Вследствие такой организации кодирования в

последовательностях ДНК может проявляться скрытая профильная периодичность и регулярность. Проявление таких свойств модели в текстовых строках продемонстрировано в численных экспериментах с бинарно перекодированными абзацами английского литературного текста, рассматриваемых в качестве аналогов CDS.

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

СПИСОК ЛИТЕРАТУРЫ

1. Issac B., Singh H., Kaur H., Raghava G.P.S. Locating probable genes using Fourier transform approach. *Bioinformatics*. 2002. V. 18. P. 196–197.
2. Yin C., Yau S.S.-T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J. Theor. Biol.* 2007. V. 247. P. 687–694. doi: 10.1016/j.jtbi.2007.03.038/.
3. Wang L., Stein L.D. Localizing triplet periodicity in DNA and cDNA sequences. *BMC Bioinformatics*. 2010. V. 11. P. 550. doi: 10.1186/1471-2105-11-550.
4. Marhon S.A., Kremer S.C. Gene prediction based on DNA spectral analysis: a literature review. *J. Comput. Biol*. 2011. V. 18. P. 639–676. doi: 10.1089/cmb.2010.0184.
5. Rivard S.R., Mailloux J.G., Beguenane R., Bui H.T. Design of high-performance parallelized gene predictors in MATLAB. *BMC Res. Notes*. 2012. V. 5. P. 183. doi: 10.1186/1756-0500-5-183.
6. Sánchez J. 3-base periodicity in coding DNA is affected by intercodon dinucleotides. *Bioinformation*. 2011. V 6. P. 327–329.
7. Shah K., Krishnamachari A. On the origin of the three base periodicity in genomes. *Biosystems*. 2012. V. 107. P. 142–144. doi: 10.1016/j.biosystems.2011.006.
8. Howe E.D., Song J.S. Categorial spectral analysis of periodicity in human and viral genomes. *Nucleic Acids Res*. 2013. V. 41. P. 1395–1405. doi: 10.1093/nar/gks1261.
9. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999. V. 27. P. 573–580.
10. Sokol D., Benson G., Tojeira J. Tandem repeats over the edit distance. *Bioinformatics*. 2007. V. 23. P. e30–e35. doi: 10.1093/bioinformatics/btl309.
11. Stoffer D.S., Tyler D.E., Wendt D.A. The spectral envelope and its applications. *Statistical Science*. 2000. V. 15. P. 224–253.
12. Chaley M., Kutyrkin V. Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences. *Math. Biosci*. 2008. V. 211. P. 186–204. doi: 10.1016/j.mbs.2007.10.008.
13. Chaley M.B., Kutyrkin V.A. Structure of proteins and latent periodicity in their genes. *Moscow Univ. Biol. Sci. Bull*. 2010. V. 65, № 4. P. 133–135. doi: 10.3103/S0096392510040012.
14. Chaley M., Kutyrkin V. Profile-Statistical Periodicity of DNA Coding Regions. *DNA Res*. 2011. V. 18. P. 353–362. doi: 10.1093/dnares/dsr023.
15. Кутыркин В.А., Чалей М.Б. Спектрально-статистический подход к распознаванию скрытой профильной периодичности в последовательностях ДНК. *Математическая биология и биоинформатика.* 2014. Т. 9. № 1. С. 33–62. doi: 10.17537/2014.9.33.
16. Chaley M., Kutyrkin V. Stochastic model of homogeneous coding and latent periodicity in DNA sequences. *J. Theor. Biol.* 2016. V. 390. P. 106–116. doi: 10.1016/j.jtbi.2015.11.014.
17. Jerom K.J. Three Men in a Boat. URL: http://www.bibliomania.com/2/-/frameset.html (дата обращения: 20.10.2015).

18. Kanehisa M., Goto S., Sato Y., Furumichi M., Tanabe M. KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.* 2012. V. 40(D1). P. D109–D114. doi: 10.1093/nar/gkr988.
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. С. 146–157. (Спец. выпуск № 3 «Математическое моделирование»).
20. Кутыркин В.А., Чалей М.Б. Структурно-статистические свойства кодирующих районов ДНК. *Математическая биология и биоинформатика*. 2015. Т. 10. № 1. С. 387–397. doi: 10.17537/2015.10.387.
21. Shepelev V., Fedorov A. Advances in the Exon-Intron Database. *Briefings in Bioinformatics*. 2006. V. 7. P. 178–185. doi: 10.1093/bib/bbl003.
22. Крамер Г. *Математические методы статистики*. М.: Мир, 1975. 648 с.
23. Кутыркин В.А., Чалей М.Б. Распознавание различных уровней в организации кодирования генетической информации. *Вестник МГТУ им. Н.Э.Баумана. Сер.* «*Естественные науки»*. 2011. С. 200–215. (спец. выпуск «Математическое моделирование»).

Рукопись поступила в редакцию 03.11.2015, переработанный вариант поступил 13.01.2016.
Дата опубликования 09.03.2016.