***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e52. doi: 10.17537/icmbb18.8*

Распознавание скрытой периодичности в последовательностях ДНК на основе стохастических моделей кодирования

Кутыркин В.А.[1], Чалей М.Б.[2]

[1]*МГТУ им. Н.Э. Баумана*
[2]*ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН*

vkutyrkin@yandex.ru

В настоящей работе было расширено понятие скрытой профильной периодичности. Ранее это понятие основывалось на модели профильной строки, состоящей из независимых случайных букв. В этом случае текстовая строка рассматривалась как реализация этой профильной строки. Однако модель профильной строки не подходит для реалистичного описания структурно-статических свойств текстовых строк. В работе предложены новые, более реалистичные стохастические модели, объясняющие проявление различных типов скрытой периодичности и регулярности, в частности, присущие кодирующим районам ДНК из геномов, как прокариот, так и эукариот. Благодаря введённому в работе понятию профильной эквивалентности случайных строк, предложенные модели обладают, с одной стороны, достаточно широкой степенью общности и, с другой стороны, позволяют объяснить характерные структурно-статистические свойства, встречающиеся в последовательностях ДНК, в частности, в кодирующих районах эукариот и прокариот. Правомерность предложенных стохастических моделей демонстрируется в численных экспериментах с бинарно перекодированными абзацами литературных текстов на двух языках (английском и итальянском).

Ключевые слова: *профиль случайной строки*, *скрытая профильная периодичность*, *случайный кодон*, *профильно-эквивалентные строки*, SPOC-*модель*.

Recognition of Latent Periodicity in DNA Sequences Based on Stochastic Models of Encoding

Kutyrkin V.A.[1], Chaley M.B.[2]

[1]*Moscow State Technical University n.a. N.E. Bauman*
[2]*IMPB RAS – Branch of KIAM RAS*

Concept of latent profile periodicity has been expanded in the present work. Earlier this concept was based on model of profile string consisting of independent random characters. In such case textual string was considered as realization of this profile string. However, the model of profile string is not appropriate for realistic description of structural-statistical properties of the textual strings. So, new, more realistic stochastic models are proposed in the work to explain manifestations of various types of latent periodicity and regularity which, particularly, are peculiar to coding DNA regions in the genomes of both prokaryotes and eukaryotes. Due to a concept of profile equivalency of random strings, that is introduced in the work, the models proposed have on the one hand sufficiently wide generality and on the other hand they allow explaining the characteristic structural-statistical properties which occur in DNA sequences, particularly, in the coding regions of eukaryotes and prokaryotes. Legitimacy of the stochastic models proposed is demonstrated in the numerical experiments with binary reencoded paragraphs of literary texts in the two languages (English and Italian).

Key words: profile of random string, latent profile periodicity, random codon, profile-equivalent strings, SPOC-*model*.

1. Введение

Поиску скрытых периодичностей, основанных на выявлении паттернов различного типа, посвящено большое количество работ [1–4]. В последнее время было введено новое понятие скрытой профильной периодичности (профильности), обобщающее понятие размытого тандемного повтора [5]. Для распознавания этого типа скрытой периодичности был разработан спектрально-статистический подход (2S-подход)

[5]. На основе этого подхода было показано, что практически во всех кодирующих районах геномов ряда прокариотических и эукариотических организмов наблюдается скрытая профильная периодичность, не описываемая размытыми тандемными повторами [5]. Для описания скрытой профильной периодичности в последовательности ДНК была предложена стохастическая модель в виде специальных случайных строк из независимых случайных букв [5]. Однако такая модель не может служить для реалистичного описания статистической структуры последовательностей ДНК. Следовательно, для объяснения наличия скрытой профильной периодичности в ДНК требуются более общие и реалистичные модели, отражающие связи между нуклеотидами в последовательности. В настоящей работе предложены такие модели, объясняющие проявление скрытой профильной периодичности в ДНК, в частности, скрытой триплетной профильной периодичности в кодирующих районах. В соответствии с этими моделями 2S-подход для распознавания скрытой профильной периодичности получил дальнейшее развитие.

В настоящей работе для демонстрации работоспособности предложенных стохастических моделей приводятся результаты численных экспериментов с бинарно перекодированными абзацами литературных текстов (английского и итальянского). В этих экспериментах буквы и знаки кодируются бинарными кодонами длиной пять. Таким образом, перекодированные абзацы литературных текстов служат аналогами кодирующих районов ДНК. Показано, что в бинарно перекодированных абзацах распознаётся скрытая профильная периодичность с размером периода пять и размером периода, кратным пяти.

2. Модели профильной периодичности для случайных и текстовых строк в спектрально-статистическом подходе

Спектрально-статистический подход (2S-подход) опирается на стохастические модели организации кодирования для текстовых строк. При таком подходе анализируемая текстовая строка, в частности, последовательность ДНК, рассматривается как реализация соответствующей случайной строки в алфавите рассматриваемых текстовых строк, например, в алфавите последовательностей ДНК. В 2S-подходе информация о случайной строке представлена в виде специальной случайной строки, состоящей из независимых случайных букв. Такая специальная случайная строка называется профильной строкой. Таким образом, происходит свёртка информации о случайной строке в виде соответствующей профильной строки, называемой профилем исходной случайной строки. Оказывается, что такой профиль сохраняет основные параметры статистической структуры исходной случайной строки, с которой получен этот профиль. 2S-подход использует спектрально-статистические характеристики не самой строки, а её профиля. Эти характеристики имеют вид функциональных зависимостей, аргументами которых являются тест-периоды профиля. Под тест-периодом строки понимают длину подстрок, на которые последовательно разбивается анализируемая строка.

2.1. Модель профильной периодичности для случайных строк. Паттерн профильной периодичности

В настоящем разделе предлагается стохастическая модель профильной периодичности в случайных строках. Частный случай такой модели, представленной профильной случайной периодической строкой, позволил нам [5] ввести новое понятие скрытой периодичности в последовательностях ДНК (текстовых строках), названное скрытой профильной периодичностью (профильностью). Это новое понятие обобщает понятие размытого тандемного повтора. Далее среди случайных строк в заданном (текстовом) алфавите выделяются более общие случайные строки, обладающие профильной периодичностью. В настоящей работе для реализаций таких случайных строк также вводится понятие скрытой профильной периодичности и предлагаются методы её распознавания в последовательностях ДНК (текстовых строках).

Опишем структуру случайных строк, обладающих профильной периодичностью.

Случайная строка $STR(n, A, \mathbf{p})$ определяется своей длиной n, текстовым алфавитом $A = \langle a_1, a_2, ..., a_K \rangle$ (где a_i – i-я буква алфавита для $i \in \overline{1, K}$, K – размер алфавита) и дискретным вероятностным распределением $\mathbf{p}$ на совокупности текстовых строк $W_n(A)$ длиной n в алфавите A. Следовательно, если текстовая строка $str \in W_n(A)$, то $\mathbf{p}(str)$ – вероятность реализации строки str для случайной строки $STR(n, A, \mathbf{p})$. В частности, если $n = 1$, то случайная строка $STR(1, A, \mathbf{p})$ называется случайной буквой в алфавите A, и для неё используется обозначение $Chr(A, \mathbf{p})$. Случайная буква $Chr(A, \mathbf{p})$ характеризуется вероятностным распределением $\mathbf{p}$ в виде столбца $\mathbf{p} = (p^1, p^2, \ldots, p^K)^T$, где $\mathbf{p}(a_i) = p^i$ – вероятность (частота) реализации буквы (строки длиной 1) $a_i \in A$ ($a_i \in W_1(A)$) и $\sum_{i=1}^{K} p^i = 1$.

Текстовую букву $a_1 \in A$ можно отождествлять со случайной буквой $Chr(A, \mathbf{p})$, где $\mathbf{p} = (1, 0, \ldots, 0)^T$. Аналогичные отождествления возможны и для остальных букв алфавита A. В этом случае случайная буква, служащая аналогом текстовой буквы, называется сосредоточенной случайной буквой.

Если алфавит A зафиксирован, то для случайной строки $STR(n, A, \mathbf{p})$ и случайной буквы $Chr(A, \mathbf{p})$ используются более краткие обозначения $STR(n, \mathbf{p})$ и $Chr(\mathbf{p})$ соответственно.

Пусть для каждого $j \in \overline{1,m}$ определена случайная строка (подстрока) $STR(n_j, A, \mathbf{p}_j)$ Тогда выражение

$$STR(n_1, A, \mathbf{p}_1)STR(n_2, A, \mathbf{p}_2)\ldots STR(n_m, A, \mathbf{p}_m) = STR$$

обозначает специальную случайную строку из перечисленных в указанном порядке независимых случайных подстрок. Другими словами, такую специальную случайную строку STR можно рассматривать как схему из m независимых испытаний, где в j-том испытании осуществляется реализация случайной подстроки $STR(n_j, A, \mathbf{p}_j)$. Если все указанные подстроки являются подстроками единичной длины, т.е. случайными буквами, то такая случайная строка STR называется профильной строкой.

Для обозначения профильных строк, в отличие от общих случайных строк STR, будет использоваться выражение Str.

Текстовую строку можно отождествлять с профильной строкой, где случайные сосредоточенные буквы отождествлены с соответствующими буквами этой текстовой строки.

Случайной строке $STR(n, \mathbf{p})$ следующим образом ставится в соответствие единственная профильная строка. Пусть a_i – i-ая буква алфавита A, $r = \overline{1,n}$ и $W_n(A,i,r) \subset W_n(A)$ – подмножество строк, в которых r-ую позицию занимает буква a_i. Тогда $p_r^i = P\{str \in W_n(A,i,r)\}$ – вероятность того, что в реализации str случайной строки $STR(n, \mathbf{p})$ в r-ой позиции находится буква $a_i \in A$. Это позволяет определить случайную букву $Chr(\mathbf{p}_r)$, где $\mathbf{p}_r = (p_r^1, p_r^2, ..., p_r^K)^T$, и профильную строку $Str = Chr(\mathbf{p}_1)Chr(\mathbf{p}_2)\ldots Chr(\mathbf{p}_n)$, называемую профилем случайной строки $STR(n, \mathbf{p})$. Для обозначения такой профильной строки Str используется выражение $Str_n(\boldsymbol{\pi})$, где $\boldsymbol{\pi} = (\mathbf{p}_1, ..., \mathbf{p}_n) = (\pi_j^i)_n^K$ – матрица из n указанных столбцов вероятностей случайных букв профильной строки $Str = Str_n(\boldsymbol{\pi})$. Введённая таким образом матрица $\boldsymbol{\pi}$ будет называться профильной матрицей строк $STR(n, \mathbf{p})$ и $Str_n(\boldsymbol{\pi})$.

Профильная строка $Str = Str_n(\boldsymbol{\pi})$ называется паттерном профильной периодичности, если её нельзя представить в виде $Str = \underbrace{Str^* Str^* \ldots Str^*}_{q-\text{раз}}$, где $q > 1$ и Str^* – некоторая другая профильная строка.

В свою очередь, понятие паттерна профильной периодичности позволяет выделить случайные строки, обладающие профильной периодичностью. Случайная строка $STR(n, \mathbf{p})$ обладает L-профильной периодичностью (L-профильностью), если её профиль $Str = Str_n(\boldsymbol{\pi})$ имеет вид:

$$Str = Str_L(\boldsymbol{\pi}_0)Str_L(\boldsymbol{\pi}_0)\ldots Str_L(\boldsymbol{\pi}_0),$$

где профильная строка $Str_L(\boldsymbol{\pi}_0)$ является паттерном. В этом случае профильная строка Str называется (стохастическим) профильным тандемным повтором и для её обозначения используется выражение $Str = Tdm_L(\boldsymbol{\pi}_0, n)$. Кроме того, профильная строка $Str_L(\boldsymbol{\pi}_0)$ называется паттерном профильной периодичности строк $STR(n, \mathbf{p})$ и $Str = Str_n(\boldsymbol{\pi})$; матрица $\boldsymbol{\pi}_0$ называется матрицей паттерна профильной периодичности строк $STR(n, \mathbf{p})$ и $Tdm_L(\boldsymbol{\pi}_0, n) = Str_n(\boldsymbol{\pi})$.

Таким образом, случайная строка обладает профильной периодичностью, если её профиль является периодической случайной строкой, индуцированной соответствующим паттерном профильной периодичности.

Если профиль случайной строки является 1-профильной строкой, то эту случайную строку и её профиль будем называть профильно-однородными строками.

3. Распознавание скрытой профильной периодичности в реализациях случайных строк с помощью спектрально-статистического подхода

Ранее [5] 2S-подход использовался для распознавания скрытой периодичности в реализациях профильных случайных строк. В настоящей работе для каждой случайной строки создаётся её профиль в виде соответствующей профильной строки. Согласно статистике, подтверждаемой большим объёмом экспериментов, для реализаций случайной строки, обладающей профильной периодичностью, и реализаций её профиля результаты применения 2S-подхода статистически неразличимы. Поэтому 2S-подход применим и к анализу реализаций произвольных случайных строк в заданном текстовом алфавите. В результате применения 2S-подхода в реализации случайной строки, обладающей скрытой профильной периодичностью, определяется длина периода и оценивается паттерн скрытой профильной периодичности. Таким образом, фактически, 2S-подход применяется к профилю случайной строки. Следовательно, методы анализа статистических спектров 2S-подхода, введённые в работе [5], применимы к распознаванию скрытой профильной периодичности в реализациях не только профильных, но и общих случайных строк. Аналитический вид этих спектров подробно описан ранее [5]. Поэтому в настоящей работе будет использоваться только анализ графических представлений статистических спектров 2S-подхода.

Статистические спектры для анализируемой строки в 2S-подходе имеют вид функциональных зависимостей, аргументы которых принадлежат фиксированному диапазону тест-периодов этой строки. Такие функциональные зависимости называются статистическими спектрами, поскольку их значения являются статистиками. Поэтому в 2S-

подходе для распознавания скрытой профильной периодичности в последовательностях ДНК (текстовых строках) используются статистические критерии.

3.1. Оценка длины периода скрытой профильной периодичности

Для каждого тест-периода L случайной строки длиной n и её профиля в 2S-подходе рассматривается спектр $\mathbf{\Psi}_L$ сравнения этой строки с соответствующей случайной L-профильной строкой. Спектр $\mathbf{\Psi}_1$ называется главным спектром анализируемой случайной строки. В работе [5] показано, что характеристическое свойство главного спектра L-профильной случайной строки состоит в том, что максимальное значение этого спектра достигается только на тест-периодах, кратных числу L (см. рис. 1,**а**, где $L = 10$). Аналогом такого главного спектра для реализации случайной строки служит характеристический спектр **C** этой реализации (см. рис. 1,**б**). Как было показано в [5], для случайной L-профильной строки характеристические спектры её реализаций, с точностью до статистической погрешности, фактически, повторяют главный спектр этой случайной строки (см. рис. 1,**а** и рис. 1,**б**). На основе такого свойства в работе [5] для анализируемой текстовой строки было сформулировано следующее правило оценки размера периода L скрытой профильной периодичности (профильности). Минимальный тест-период, на котором достигается максимальное значение (с учётом статистической погрешности) характеристического спектра **C** (см. рис. 1,**б**, где $L = 10$) анализируемой текстовой строки, рассматривается в качестве оценки длины периода скрытой профильной периодичности.

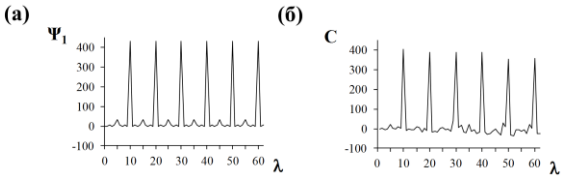

Рис. 1. (**а**) Главный спектр (спектр сравнения с однородной строкой) случайной 10-профильной строки с профилем $Tdm_{10}(\boldsymbol{\pi}_0, n)$. (**б**) Характеристический спектр текстовой строки *str* (последовательности ДНК из генома *C. elegans*, хромосомы III, индексы: 307381–308580), являющейся реализацией 10-профильной строки $Tdm_{10}(\boldsymbol{\pi}_0, n)$.

Предложенные в 2S-подходе статистические критерии позволяют проверить достоверность оценки длины профильного периода $L > 1$, полученной на основе анализа только характеристического спектра рассматриваемой текстовой строки.

4. Свойство 3-регулярности в последовательностях ДНК

На рисунке 2 показан характеристический спектр для CDS белка транспортера D-рибозы (D-ribose transporter ATP-binding protein) из генома бактерии *Erwinia tasmaniensis*. В этом спектре практически все локальные максимумы наблюдаются на тест-периодах, кратных трём. Кроме того, практически на всех этих тест-периодах наблюдаются отклонения от однородности [5]. Однако, согласно методам 2S-подхода в этой последовательности отсутствует 3-профильная периодичность (3-профильность). Такое свойство последовательности ДНК было названо в работе [5] свойством 3-регулярности последовательности ДНК.

В общем случае, наличие свойства 3-регулярности не гарантирует существования в строке какой-либо скрытой профильности. Однако, практически во всех последовательностях ДНК, обладающих скрытой профильностью с периодом, кратным трём, проявляется свойство 3-регулярности.

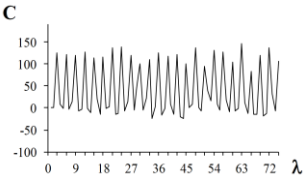

Рис. 2. Характеристический спектр 3-регулярной CDS белка транспортера D-рибозы (D-ribose transporter ATP-binding protein [EC:3.6.3.17]) генома бактерии *E. tasmaniensis*.

5. Сравнение структурно-статистических свойств CDSs из геномов различных организмов и бинарно перекодированных абзацев литературных текстов

Для численных экспериментов использовались тексты книги Марка Твена «Приключения Гекльберри Финна» (“Huckleberry Finn” by Mark Twain, http://www.bibliomania.com/) на английском языке и книги Умберто Эко «Имя розы» (“Il nome della rosa” by Umberto Eco, http://royallib.com) на итальянском языке. Естественными семантическими единицами в этих текстах 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 (пентаплетами), как показано в работе [5].

Текст каждого абзаца в исходном литературном тексте, записанный в бинарном алфавите, рассматривался как отдельная бинарная последовательность. Анализировались неоднород-ные перекодированные абзацы. Таким образом,

проводилась аналогия таких абзацев с CDSs из геномов прокариот и эукариот, рассмотренных в работе [5]. Как и для CDSs, к анализу бинарно перекодированных абзацев применялся 2S-подход.

Сравнение структурно-статистических свойств CDSs эукариот и прокариот и бинарно перекодированных абзацев литературных текстов показало их несомненную аналогию. Единственное отличие состоит в том, что в CDSs геномов наблюдаются свойства 3-регулярности и 3-профильности, в то время как в бинарно перекодированных абзацах фиксируется наличие свойств 5-регулярности и 5-профильности.

6. Модели регулярной стохастической организации кодирования

В настоящем разделе рассматривается кодирование букв текстового алфавита на основе равномерного кода, в котором кодоны имеют одинаковый размер. В частности, в CDSs для кодирования аминокислот используется триплетный генетический код.

Поскольку в работе исследуется профильная периодичность, для случайных строк вводится понятие профильной эквивалентности. Две случайные строки, обладающие одинаковым профилем, называются профильно-эквивалентными случайными строками.

Далее предлагаются стохастические модели, отражающие структурно-статистические свойства кодирующих текстов с естественным смысловым содержанием. Эти модели объясняют проявление в кодирующих текстах скрытой профильной периодичности и свойства регулярности. В основу этих моделей положен класс специальных случайных строк, в реализациях которых 2S-подход выявляет скрытую профильную периодичность и регулярность. Поскольку класс таких случайных строк весьма обширен, в нём выделяется подкласс, обладающий следующим свойством. Любая строка класса профильно-эквивалентна некоторой строке из этого подкласса. Следовательно, строки подкласса представляют в общем виде стохастические модели регулярной организации кодирования для случайных строк, в реализациях которых выявляется скрытая профильная периодичность и регулярность. Кроме того, модели вводимого подкласса, в отличие от периодических профильных строк, более реалистично отражают структурно-статистические свойства рассматриваемых кодирующих текстов.

6.1. Модель стохастической профильной организации кодирования (SPOC-model)

SPOC-model является специальной случайной строкой, в реализациях которой проявляется скрытая профильная периодичность. В основе описания такой строки лежит понятие стохастического (случайного) кодона $Cdn = STR(L, \mathbf{p})$, являющегося случайной строкой длиной L, где $\mathbf{p}$ – вероятностное распределение на текстовых строках длины L в алфавите $A = \langle a_1, a_2, ..., a_K \rangle$. Кроме того, профиль $Str_L(\boldsymbol{\pi})$ кодона Cdn должен быть паттерном (см. раздел 2.1). Следовательно, главный спектр кодона достигает своего максимального значения только на тест-периде L.

Пусть $Cdn_1, Cdn_2, ..., Cdn_m$ – такие случайные кодоны размера Λ в алфавите $A = \langle a_1, a_2, ..., a_K \rangle$, что профиль $Ptn = Str_L(\boldsymbol{\pi})$ случайной строки $STR_0 = Cdn_1 Cdn_2 ... Cdn_m$ является паттерном. Тогда случайную строку $STR = \underbrace{STR_0 STR_0 \ldots STR_0}_{q-\text{раз}}$, где $q > 5K$, будем рассматривать в качестве SPOC-модели со списком (профилем) из m случайных кодонов размера Λ в алфавите A. В этом случае профиль Str случайной строки STR имеет вид $Str = \underbrace{Ptn\, Ptn \ldots Ptn}_{q-\text{раз}}$. Следовательно, этот профиль является L-профильной строкой $Tdm_L(\boldsymbol{\pi}, qL) = Str$ (см. раздел 2.1), где $L = m\Lambda$. Вследствие значительного количества повторов ($q > 5K$) паттерна профильности Ptn, статистический анализ (с помощью 2S-подхода) реализаций строки STR будет приводить к тем же результатам, что и статистический анализ реализаций L-профильной строки Str. Таким образом, практически во всех реализациях строки STR будет распознаваться L-профильная периодичность. Кроме того, если Λ – простое число, то в реализациях строки STR может проявляться исключительно Λ-регулярность. С точки зрения авторов настоящей работы, такая SPOC-модель, где $\Lambda = 3$ и $m > 1$, объясняет наличие свойства 3-регулярности в CDSs, в которых распознаётся L-профильная периодичность с размером периода, кратным трём. Аналогичное явление наблюдается для $\Lambda = 5$ и в бинарно перекодированных абзацах литературных текстов. В общем случае такое явление было названо двухуровневой организацией кодирования [5]. Первый уровень обусловлен наличием свойства Λ-регулярности. Второй уровень связан с распознаванием скрытой профильной периодичности, размер периода которой кратен, но не равен Λ.

Частным случаем SPOC-модели является модель стохастической однородной организации кодирования (SHOC-модель), введённая в работе [5]. SHOC-модель имеет вид случайной строки $STR = \underbrace{Cdn\, Cdn \ldots Cdn}_{q-\text{раз}}$, где $Cdn = STR(L, \mathbf{p})$ – случайный кодон размера L (L – простое число) и $q > 5K$. В этом случае в реализациях строки STR проявляется скрытая L-профильность.

Рассмотренные в настоящем разделе модели с размером кодонов $L = 3$ схематично иллюстрирует рисунок 3, где показаны главные спектры случайных строк, представляющих эти модели. В реализациях этих строк проявляются скрытая

триплетная периодичность (рис. 3,**а**), скрытая 9-профильность на фоне 3-регулярности (рис. 3,**б**), только свойство 3-регулярности (рис. 3,**в**).

7. Заключение

Ранее было показано, что в большинстве CDSs целого ряда прокариотических и эукариотических организмов распознаётся новый тип скрытой периодичности, названный скрытой профильной периодичностью. Кроме того, было показано, что размер паттерна скрытой профильной периодичности CDSs этих геномов кратен трём, и практически во всех CDSs этих геномов выявлялось свойство 3-регулярности. Для объяснения этих проявлений скрытой профильной периодичности в работе [5] была предложена стохастическая модель профильных строк, состоящих из независимых случайных букв. Такая модель вряд ли подходит для реалистичного описания статистической структуры последовательностей ДНК. По этой причине в настоящей работе для описания статистической структуры в последовательности ДНК предложены новые, более общие и реалистичные модели. Эти модели позволяют на основе дальнейшего развития 2S-подхода распознавать в последовательностях ДНК скрытую профильную периодичность.

Правомерность предложенных в настоящей работе стохастических моделей была продемонстрирована в численных экспериментах с бинарно перекодированными абзацами двух литературных текстов на английском (M. Twain) и итальянском (U. Eco) языках.

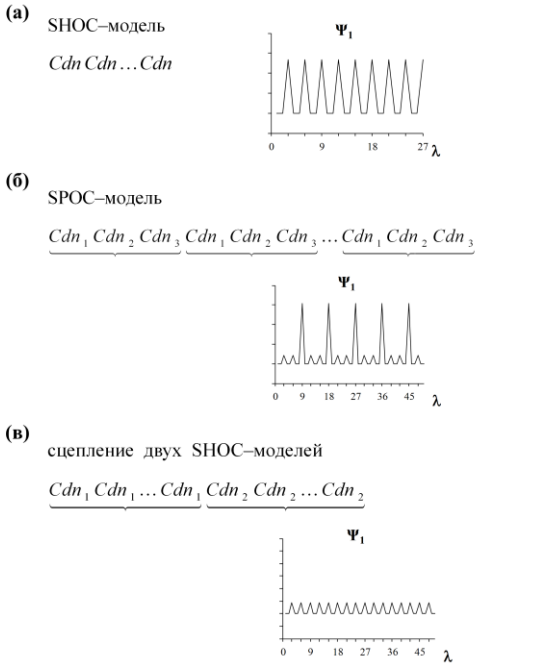

Рис. 3. Главные спектры случайных строк, представляющих: (**а**) SHOC-модель, (**б**) SPOC-модель и (**в**) сцепление SHOC-моделей со случайными кодонами размера три в алфавите ДНК.

8. Список литературы

1. Benson G. Tandem repeat finder: a program to analyse DNA sequences. *Nucleic Acids Res*. 1999. V. 27. P. 573–580. doi: 10.1093/nar/27.2.573.
2. Landau G.M., Schmidt J.P., Sokol D. An algorithm for approximate tandem repeats. *J. Comput. Biol*. 2001. V. 8. P. 1–18. doi: 10.1089/106652701300099038.
3. Grover A., Aishwarya V., Sharma P.C. Searching microsatellites in DNA sequences: approaches used and tools developed. *Physiol. Mol. Biol. Plants*. 2012. V. 18. P. 11–19. doi: 10.1007/s12298-011-0098-y.
4. Sirisha G., Shashi M., Raju G.V.P. Periodic Pattern Mining-Algorithms and Applications. *Global Journal of Computer Science and Technology*. 2013. V. 13. P. 19–28. GJCST – C Classification: D.2.11.
5. 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.