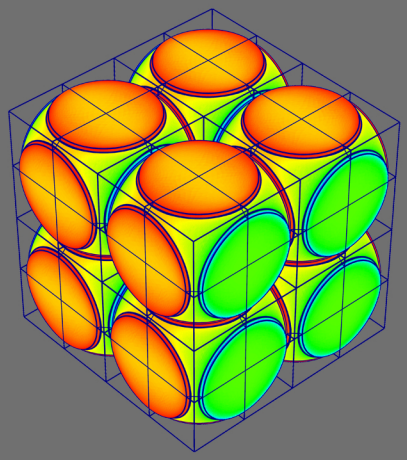
ISSN 2309-3684



Математическое моделирование и численные методы

Кутыркин В.А., Чалей М.Б. Стохастические модели кодирования и распозна- вание структурно-статистических характеристик кодирующих последовательно- стей. Математическое моделирование и численные методы, 2017, No 3, с. 119–138.

Источник: https://mmcm.bmstu.ru/articles/145/

Параметры загрузки:
IP: 216.73.216.187
14.11.2026 05:47:43

УДК 519.25 DOI 10.18698/2309-3684-2017-3-119138

Стохастические модели кодирования и распознавание структурно-статистических характеристик кодирующих последовательностей

© В.А. Кутыркин[1], М.Б. Чалей[2]

[1]МГТУ им. Н.Э. Баумана, Москва, 105005, Россия
[2]ИМПБ РАН — филиал ИПМ им. М.В. Келдыша РАН,
г. Пущино, Московская обл., 142290, Россия

Предложены стохастические модели, объясняющие реальные характерные закономерности кодирующих районов из геномов различных организмов. Вследствие нарастающего объема данных по секвенированным геномам возникает проблема их автоматизированного анализа. С использованием этих моделей разработаны методы распознавания структурно-статистических свойств геномных последовательностей ДНК, которые могут быть использованы для разработки алгоритмов и компьютерных программ для автоматизированной обработки большого объема данных. Свойства предложенных стохастических моделей кодирования продемонстрированы в численных экспериментах с бинарно перекодированными абзацами литературных произведений на английском и итальянском языках.

Ключевые слова: *профиль случайной строки, профильная периодичность, паттерн профильной периодичности, стохастический кодон, мультиполиномиальная модель*

Введение. Исследование структурно-статистических свойств и способов кодирования в геномах различных организмов — актуальная и комплексная проблема анализа быстро нарастающих данных о секвенированных последовательностях геномов. Для разных аспектов этой проблемы постоянно требуется разрабатывать новые алгоритмы и создавать на их основе программы, обеспечивающие автоматизированную обработку большого объема данных. Такие комплексные проблемы характерны и для других областей науки [1, 2].

Настоящее исследование направлено на разработку стохастических моделей кодирования, объясняющих основные статистические закономерности, характерные для кодирующих текстов с некоторым смысловым содержанием. При этом предполагается, что при создании этих текстов используются кодоны одного размера в фиксированном текстовом алфавите, полученные с помощью равномерного кода. На основе таких стохастических моделей в настоящей статье предложены статистические методы и алгоритмы для распознавания в кодирующих последовательностях структурно-статистических свойств, информацию о которых можно получать при автоматизированной обработке большого объема данных.

Ранее [3–5] в численных экспериментах были выявлены характерные структурно-статистические свойства кодирующих районов

нуклеотидных последовательностей ДНК из геномов 10 разных организмов (в том числе из генома человека) и бинарно перекодированных абзацев литературных произведений на английском и итальянском языках.

Согласно генетическому коду, кодирующие районы последовательностей ДНК впоследствии транслируются с помощью равномерного кода в белковые последовательности, состоящие из аминокислотных остатков. При такой трансляции каждая аминокислота представлена в кодирующей последовательности ДНК кодоном из трех нуклеотидов. Для кодирования белковых последовательностей используется равномерный код (генетический код) с кодонами размера три в алфавите ДНК ‹*a*, *t*, *g*, *c*› из четырех нуклеотидов.

В численных экспериментах с литературными текстами для кодирования букв латинского алфавита и синтаксических знаков абзацев текстов использовались кодоны размера пять в бинарном алфавите ‹0, 1›.

В результате анализа численных экспериментов с кодирующими районами последовательностей ДНК в этих районах были выявлены характерные структурно-статистические свойства, которые не наблюдаются в некодирующих районах (интронах) последовательностей ДНК [3]. В подавляющем большинстве кодирующих районов распознавался недавно введенный тип скрытой периодичности, названный *скрытой профильной периодичностью* (*скрытой профильностью*) [6–8]. При этом размер ее периода был равен или кратен трем, что соответствует длине кодона размера три в генетическом коде. Кроме того, практически во всех кодирующих районах было обнаружено так называемое свойство 3-регулярности [3, 4, 8].

В численных экспериментах с бинарно перекодированными абзацами литературных текстов были выявлены аналогичные результатам анализа кодирующих районов последовательностей ДНК статистические закономерности [3–5]. Единственное отличие состояло в том, что при перекодировании с помощью бинарных кодонов размера пять в кодирующих текстах наблюдалась скрытая профильность с размером периода, равным или кратным пяти. Соответственно свойство 3-регулярности сменилось на свойство 5-регулярности.

Для распознавания скрытой профильной периодичности и регулярности в последовательностях ДНК в работах [7–9] были предложены статистические методы и критерии. Опора на такие методы обусловлена тем, что для описания профильной периодичности использовалась стохастическая модель в виде *профильной строки*, состоящей из независимых случайных букв со значениями в буквах фиксированного текстового алфавита. Ранее [10–12] для описания скрытой периодичности в последовательностях ДНК использовалась модель совершенного текстового тандемного повтора, состоящего из

последовательно повторяющейся текстовой строки. На ее основе было введено понятие скрытой периодичности в виде размытого тандемного повтора, где возможны небольшие искажения (~20 %) по сравнению с совершенным тандемным повтором. Однако оказалось, что размытые тандемные повторы занимают достаточно небольшую часть (~10 %) в кодирующих районах последовательностей ДНК. Косвенные методы [13–15] определения размера периода скрытой периодичности (анализ Фурье и т. п.), не опирающиеся на какую-либо модель для описания скрытой периодичности, как показано в работе [7], могут приводить к недостоверным оценкам.

Однако предложенная ранее [6, 16] модель периодической профильной строки, состоящей из независимых случайных букв, вряд ли отражает реальную статистическую структуру кодирующих районов ДНК. Более полное отражение наличия скрытой профильной периодичности в последовательности ДНК требует объяснения с помощью других моделей, которые предложены в настоящей работе. Методы и основанные на них алгоритмы распознавания скрытой профильной периодичности опираются на эти стохастические модели. При таком подходе анализируемая текстовая строка рассматривается как реализация соответствующей случайной строки (стохастической модели) в алфавите исследуемых текстовых строк. Информация об этой стохастической модели представляется в виде соответствующей профильной строки, состоящей из независимых случайных букв. При этом происходит свертка информации о стохастической модели в виде профильной строки, называемой профилем стохастической модели (исходной случайной строки), сохраняющим изучаемые структурно-статистические характеристики исходной случайной строки, с которой получен этот профиль. Эти характеристики имеют вид функциональных зависимостей, аргументами которых являются тест-периоды профиля. Под тест-периодом профильной строки понимают длину подстрок, на которые последовательно разбивается анализируемая строка.

В настоящей статье правомерность предлагаемых стохастических моделей кодирования демонстрируется в численных экспериментах с бинарно перекодированными абзацами литературных текстов.

Модели профильной периодичности в случайных строках. Предлагается стохастическая модель профильной периодичности в случайных строках, частный случай которой позволил ранее [6, 16] ввести новое понятие скрытой периодичности в последовательностях ДНК (текстовых строках), названной скрытой профильной периодичностью (скрытой профильностью). После этого среди случайных строк в заданном текстовом алфавите выделяются более общие случайные строки (стохастические модели), обладающие профильной периодичностью. В настоящей работе для реализаций таких случай-

ных строк введено понятие скрытой профильной периодичности и предложены алгоритмы для ее распознавания в текстовых строках (последовательностях ДНК).

Опишем структуру случайных строк, обладающих профильной периодичностью.

Случайная строка $STR(n, \boldsymbol{A}, \mathbf{p})$ определяется своей длиной n, текстовым алфавитом $\boldsymbol{A} = \langle a_1, a_2, ..., a_K \rangle$ и дискретным вероятностным распределением $\mathbf{p}$ на совокупности $W_n(\boldsymbol{A})$ текстовых строк длиной n в алфавите $\boldsymbol{A}$. Следовательно, если $w \in W_n(\boldsymbol{A})$, то $\mathbf{p}(w)$ — вероятность реализации строки w для случайной строки $STR(n, \boldsymbol{A}, \mathbf{p})$. В частности, если $n = 1$, то случайная строка $STR(1, \boldsymbol{A}, \mathbf{p})$ называется *случайной буквой* в алфавите $\boldsymbol{A}$ и для ее обозначения используется $Chr(\boldsymbol{A}, \mathbf{p})$. Случайная буква $Chr(\boldsymbol{A}, \mathbf{p})$ характеризуется вероятностным распределением $\mathbf{p}$ в виде столбца $\mathbf{p} = (p^1, p^2, ..., p^K)^T$, где $\mathbf{p}(a_i) = p^i$ — вероятность реализации буквы $a_i \in \boldsymbol{A}$ для $i = \overline{1, K}$.

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

Если алфавит $\boldsymbol{A}$ зафиксирован в контексте, для случайной строки $STR(n, \boldsymbol{A}, \mathbf{p})$ и случайной буквы $Chr(\boldsymbol{A}, \mathbf{p})$ 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) $STR(n_j, \boldsymbol{A}, \mathbf{p}_j)$. Тогда выражение $STR(n_1, \boldsymbol{A}, \mathbf{p}_1) \times \times STR(n_2, \boldsymbol{A}, \mathbf{p}_2) \ldots STR(n_m, \boldsymbol{A}, \mathbf{p}_m) = STR$ 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-м испытании осуществляется реализация случайной подстроки $STR(n_j, \boldsymbol{A}, \mathbf{p}_j)$. 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 используем выражение Str.

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

Случайной строке $STR(n, \mathbf{p})$ ставится в соответствие единственная профильная строка. Пусть a_i — i-я буква алфавита $\boldsymbol{A}$, $r = \overline{1, n}$, $W_n(\boldsymbol{A}, i, r) \subset W_n(\boldsymbol{A})$ — подмножество строк длины n, в которых r-ю позицию занимает буква a_i. Тогда $p_r^i = P\{w \in W_n(\boldsymbol{A}, i, r)\}$ — вероятность того, что в реализации $w \in W_n(\boldsymbol{A})$ случайной строки $STR(n, \mathbf{p})$ в r-й позиции находится буква $a_i \in \boldsymbol{A}$. 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 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 строки $Str = Str_n(\boldsymbol{\pi})$. Матрицу $\boldsymbol{\pi} = (\mathbf{p}_1, ..., \mathbf{p}_n)$ назовем профильной матрицей строки $STR(n, \mathbf{p})$.

Профильная строка $Str = Str_n(\boldsymbol{\pi})$ называется *паттерном профильной периодичности*, если ее нельзя представить в виде

$$Str = \underbrace{Str^* Str^* \ldots Str^*}_{q\text{-}times},$$

где $q > 1$, Str^* — некоторая другая профильная строка.

В свою очередь, понятие паттерна профильной периодичности позволяет выделить случайные строки, обладающие профильной периодичностью.

Случайная строка $STR(n, \mathbf{p})$ называется L*-профильной строкой*, если ее профиль $Str = Str_n(\boldsymbol{\pi})$ имеет вид

$$Str = Str_L(\boldsymbol{\pi}_0)Str_L(\boldsymbol{\pi}_0)\ldots Str_L(\boldsymbol{\pi}_0)Str_k(\boldsymbol{\pi}_1).$$

Здесь профильная строка $Str_L(\boldsymbol{\pi}_0)$ — паттерн; $k < L$, $Str_k(\boldsymbol{\pi}_1)$ — подстрока строки $Str_L(\boldsymbol{\pi}_0)$ (пустая строка, если $k = 0$). В этом случае профильная строка $Str = Str_n(\boldsymbol{\pi})$ называется (*стохастическим*) *профильным тандемным повтором*. Для ее 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 $\boldsymbol{\pi}_0$ — *матрицей паттерна профильной периодичности строк* $STR(n, \mathbf{p})$ и $Tdm_L(\boldsymbol{\pi}_0, n) = Str_n(\boldsymbol{\pi})$.

Если профиль случайной строки является 1-профильной строкой, то такую случайную строку и ее профиль будем называть *профильно-*

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

Основные структурно-статистические свойства случайной строки и ее реализаций индуцируются ее профильно-матричным спектром, который определяется далее.

Профильно-матричные спектры случайных и текстовых строк. Основу методов распознавания скрытой профильной периодичности в текстовых строках составляют профильно-матричные спектры случайных и текстовых строк, которые однозначно вычисляются для анализируемой случайной или текстовой строки. В следующем разделе на основе профильно-матричных спектров будут введены другие спектры, анализ которых позволяет создать достоверные статистические методы и критерии для распознавания скрытой профильной периодичности в текстовых строках.

Введем профильно-матричные спектры случайных и текстовых строк. Для случайной строки $STR(n, \mathbf{p})$ в алфавите $\boldsymbol{A} = \langle a_1, a_2, ..., a_K \rangle$ создается ее профиль в виде профильной строки $Str = Str_n(\boldsymbol{\pi})$. *Профильно-матричный спектр случайной строки* $STR(n, \mathbf{p})$ совпадает с профильно-матричным спектром $\boldsymbol{\Pi}_{Str}$ ее профиля Str, который создается следующим образом. Для каждого тест-периода λ профиль Str случайной строки $STR(n, \mathbf{p})$ последовательно разбивается на подстроки длиной λ (последняя подстрока может быть неполной). Для наглядности предположим, что профиль Str последовательно разбит на $m = n/\lambda$ подстрок. Каждая из таких m подстрок характеризуется своей профильной матрицей размера $K \times \lambda$. Таким образом, для тест-периода λ формируется список из m профильных матриц $(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_m)$, который определяет матрицу

$$\boldsymbol{\Pi}_{Str}(\lambda) = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\pi}_i .$$

Матрицы с диапазоном тест-периодов от 1 до $L_{\max} \sim n/5K$ образуют профильно-матричный спектр $\boldsymbol{\Pi}_{Str}$ случайной строки $STR(n, \mathbf{p})$ и ее профиля Str. Если длина n нацело не делится на тест-период λ, то при определении матрицы $\boldsymbol{\Pi}_{Str}(\lambda)$ вносятся соответствующие поправки.

Пусть случайная строка имеет периодический профиль $Tdm_L(\boldsymbol{\pi}_0, n)$, тогда ее профильно-матричный спектр однозначно индуцируется профильной матрицей $\boldsymbol{\pi}_0$ паттерна $Str_L(\boldsymbol{\pi}_0)$ этого про-

филя. Следовательно, матрица $\boldsymbol{\pi}_0$ служит оценкой паттерна профильной периодичности рассматриваемой случайной строки и содержит информацию об основных статистических зависимостях в реализациях рассматриваемой случайной строки.

Для анализируемой текстовой строки *str* длиной n в том же алфавите $\boldsymbol{A}$ *выборочный профильно-матричный спектр* $\boldsymbol{\Pi}_{Str}$ совпадает с профильно-матричным спектром профильной строки *Str*, полученной при отождествлении букв текстовой строки *str* с соответствующими случайными сосредоточенными буквами. Для текстовой строки предполагается, что спектр $\boldsymbol{\Pi}_{Str}$ определен в *диапазоне* $(1, 2, ..., L_{\max} \sim n/5K)$ *тест-периодов текстовой строки* *str*.

Алгоритмы распознавания скрытой профильной периодичности в реализациях случайных строк. В предлагаемом подходе проверяется гипотеза о том, что текстовая строка является реализацией соответствующей случайной строки, обладающей профильной периодичностью. С помощью разработанных статистических методов и критериев распознается наличие скрытой профильной периодичности в анализируемой текстовой строке. Кроме того, дается оценка размера паттерна профильной периодичности и его вида для соответствующей случайной строки.

Если рассматривать случайную строку, обладающую профильной периодичностью с достаточно большим количеством повторов периода, то для ее реализаций и реализаций ее профиля результаты используемого подхода будут статистически неразличимы. Таким образом, фактически предлагаемый подход применяется к профилю случайной строки. Данный подход основан на количественной обработке и анализе статистических спектров, вычисляемых по заданной анализируемой строке и имеющих вид функциональных зависимостей, аргументы которых принадлежат фиксированному диапазону тест-периодов этой строки. Такие функциональные зависимости называются статистическими спектрами, поскольку их значения являются статистиками. По этой причине в предлагаемом спектрально-статистическом подходе для распознавания скрытой профильной периодичности используются статистические методы и критерии.

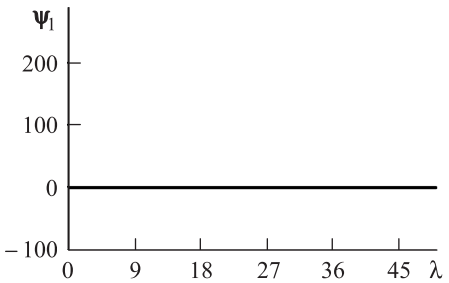
Определим соответствующие спектры, необходимые для распознавания скрытой профильной периодичности (профильности) в анализируемой строке. Для профиля *Str* случайной строки $STR(n, \mathbf{p})$ в алфавите $\boldsymbol{A} = \langle a_1, a_2, ..., a_K \rangle$ введем *спектр* $\boldsymbol{\psi}_1$ *сравнения с профильно-однородной (1-профильной) строкой* $Str^* = \underbrace{Chr(\mathbf{p})Chr(\mathbf{p})\ldots Chr(\mathbf{p})}_{n\text{-}times}$, где

$\mathbf{p} = (p^1, ..., p^K)^T = \mathbf{\Pi}_{Str}(1)$. Для тест-периода λ из диапазона от 1 до $L_{\max} \sim n/5K$ профильные матрицы $\mathbf{\Pi}_{Str^*}(\lambda) = \boldsymbol{\pi}^* = (\pi_j^{*i})_\lambda^K =$ $= (\underbrace{\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}}_{\lambda\text{-}times})$, $\mathbf{\Pi}_{Str}(\lambda) = (\pi_j^i)_\lambda^K$ и $\mathbf{\Pi}_{Str}(1) = \mathbf{p} = (p^1, ..., p^K)^T$ определяют значение $\psi_1(\lambda)$ спектра $\boldsymbol{\psi}_1$ строки *Str* в виде статистики Пирсона

$$\psi_1(\lambda) = \frac{n}{\lambda}\sum_{j=1}^{\lambda}\sum_{i=1}^{K}\frac{(\pi_j^i - \pi_j^{*i})^2}{\pi_j^{*i}} = \frac{n}{\lambda}\sum_{j=1}^{\lambda}\sum_{i=1}^{K}\frac{(\pi_j^i - p^i)^2}{p^i}. \qquad (1)$$

Для спектра $\boldsymbol{\psi}_1$ L-профильной строки справедливо следующее утверждение [3, 4]: *спектр* $\boldsymbol{\psi}_1$ *профильной строки* Str, *обладающей профильной периодичностью с длиной периода* L, *периодичен с тем же периодом* L *и, кроме того, максимальное значение этого спектра* $\boldsymbol{\psi}_1$ *достигается только на тест-периоде* L *и его обертонах.*

Для профиля *Str* случайной строки $STR(n, \mathbf{p})$ спектр $\boldsymbol{\psi}_1$ (см. формулу (1)) будет называться *спектром первого порядка* этих *строк*. На рис. 1, *а* показан спектр $\boldsymbol{\psi}_1$ профильно-однородной строки $Tdm_1(\mathbf{p}, n)$, где $n = 1002$, и на рис. 1, *б* представлена ее матрица паттерна 1-профильной периодичности (1-профильности). Для периодических профильных строк примеры аналогичных спектров первого порядка и их паттернов профильной периодичности показаны на рис. 2.



а

	1
a	0.27
t	0.19
g	0.32
c	0.22

б

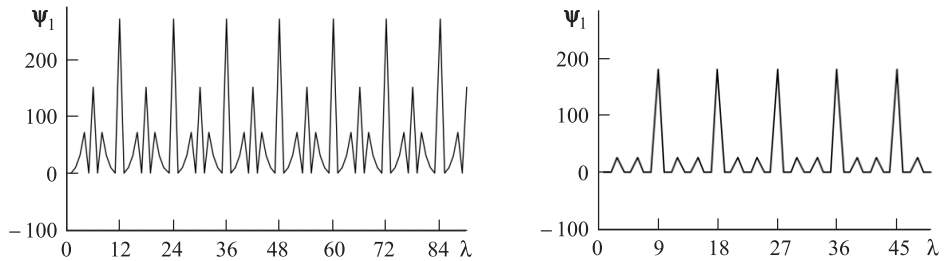
Рис. 1. Спектр первого порядка профильно-однородной строки (*а*) и матрица паттерна профильной периодичности этой строки (*б*)

Пусть $\mathbf{p} = (p^1, ..., p^K)^T = \mathbf{\Pi}_{str}(1)$ — столбец частот встречаемости букв алфавита $\boldsymbol{A}$ в текстовой строке *str* длиной n и λ — ее тест-

период из диапазона $1...L_{\max}$. Тогда по аналогии с формулой (1) для текстовой строки str введем *выборочный спектр первого порядка* $\boldsymbol{\psi}_1$ сравнения с однородной профильной строкой

$$Str^* = \underbrace{Chr(\mathbf{p})Chr(\mathbf{p})\ldots Chr(\mathbf{p})}_{n\text{-}times}.$$

В этом случае в формуле (1) для тест-периода λ запишем матрицу $\boldsymbol{\pi} = (\pi_j^i)_\lambda^K = \boldsymbol{\Pi}_{str}(\lambda)$.



	1	2	3	4	5	6	7	8	9	10	11	12
a	1.0				1.0	1.0						
t		1.0					1.0	1.0				
g			1.0						1.0	1.0		
c				1.0							1.0	1.0

в

	1	2	3	4	5	6	7	8	9
a	0.38	0.32	0.24	0.14	0.16	0.28	0.30	0.39	0.20
t	0.14	0.22	0.28	0.06	0.15	0.23	0.15	0.25	0.20
g	0.21	0.17	0.29	0.68	0.60	0.30	0.32	0.10	0.24
c	0.27	0.29	0.20	0.11	0.08	0.19	0.23	0.26	0.36

г

Рис. 2. Спектры первого порядка для совершенного текстового тандемного повтора (*а*) и для периодической 9-профильной строки (*б*), матрицы паттернов профильной периодичности совершенного текстового тандемного повтора (*в*) и 9-профильной строки (*г*)

На рис. 3, *а* приведен пример выборочного спектра первого порядка для реализации 9-профильной строки с паттерном профильной периодичности, показанным на рис. 2, *г*. Из сравнения графиков спектров первого порядка следует, что выборочный спектр первого порядка (рис. 3, *а*) отличается от спектра первого порядка 9-профильной строки (см. рис. 2, *б*) на некоторую линейную функцию. Аналогично отличие на ту же самую линейную функцию наблюдается между выборочным спектром первого порядка (см. рис. 3, *б*) реализации профильно-однородной строки и ее спектром первого порядка (см. рис. 1, *а*).

Как показано в работах [1, 2], для алфавита $\boldsymbol{A}$ размера K такую линейную функцию можно аппроксимировать зависимостью

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

где $E(\chi^2_N)$ — математическое ожидание χ^2-распределения с N степенями свободы. Исходя из этого, вместо выборочного спектра первого порядка ψ_1 текстовой строки *str* в работах [1, 2] введен характеристический спектр $\mathbf{C}$, который на тест-периоде λ имеет вид

$$C(\lambda) = \psi_1(\lambda) - M(\lambda) = \psi_1(\lambda) - E(\chi^2{}_{(K-1)(\lambda-1)}) = \psi_1(\lambda) - (K-1)(\lambda-1). \quad (3)$$

Характеристический спектр реализации 9-профильной строки показан на рис. 3, *в*, спектр первого порядка показан на рис. 2, *б*. Сравнение рис. 2, *б* и рис. 3, *в* демонстрирует наглядное сходство спектра первого порядка 9-профильной строки и характеристического спектра ее реализации. Такое сходство наблюдается для подавляющего числа реализаций периодических профильных строк.

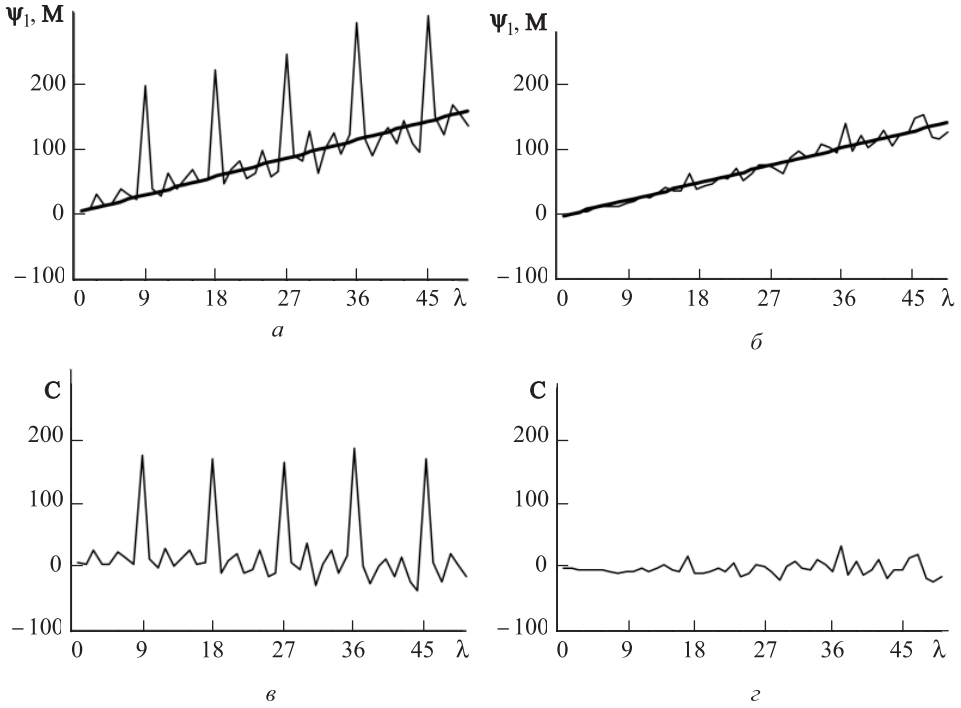


Рис. 3. Графики линейной зависимости $\mathbf{M}$ и выборочного спектра первого порядка ψ_1 реализации 9-профильной строки (*а*) и графики линейной зависимости $\mathbf{M}$ и выборочного спектра первого порядка ψ_1 реализации профильно-однородной строки (*б*) и соответствующие им характеристические спектры (*в*), (*г*)

Такие свойства характеристического спектра анализируемой текстовой строки позволяют оценить длину периода скрытой профильной периодичности следующим образом [3, 4]. *Минимальный тест-период, на котором достигается максимальное значение характерис-*

тического спектра $\mathbf{C}$ (с учетом статистической погрешности) *анализируемой текстовой строки, рассматривается в качестве оценки длины периода скрытой профильной периодичности*. Например, в качестве реализации 9-профильной строки, спектр первого порядка которой показан на рис. 2, *б*, был рассмотрен кодирующий район для белка (фактора некроза опухоли) из генома человека (KEGG, hsa:338872, 1002 bp) [17]. Согласно сформулированному правилу, из анализа характеристического спектра этой последовательности (см. рис. 3, *в*) тест-период 9 выбирается в качестве оценки длины периода скрытой профильной периодичности.

Опишем статистические критерии проверки корректности подобных оценок. Пусть L $(L < L_{\max})$ — оценка длины периода скрытой профильной периодичности, полученная из характеристического спектра анализируемой текстовой строки *str* длиной *n*. В этом случае рассмотрим эту последовательность как реализацию L-профильной случайной строки, профиль которой имеет вид $Str = Tdm_L\left(\mathbf{\Pi}_{str}(L), n\right)$. Для тест-периода λ из диапазона тест-периодов $1...L_{\max} \sim n/5K$ профильные матрицы $\mathbf{\Pi}_{str}(\lambda) = (\pi_j^i)_\lambda^K$ и $\mathbf{\Pi}_{Str}(\lambda) = \boldsymbol{\pi}^* = (\pi^{*i}_j)_\lambda^K$ определяют значение $\psi_L(\lambda)$ выборочного спектра $\boldsymbol{\psi}_L$ строки *str* в виде статистики Пирсона

$$\psi_L(\lambda) = \frac{n}{\lambda}\sum_{j=1}^{\lambda}\sum_{i=1}^{K}\frac{(\pi_j^i - \pi_j^{*i})^2}{\pi_j^{*i}}. \qquad (4)$$

Спектр $\boldsymbol{\psi}_L$ позволяет сравнивать строку *str* (анализируемую последовательность ДНК) с L-профильной строкой. Его также называют *спектром* L*-го порядка строки* *str*. Для реализаций L-профильной строки $Tdm_L(\boldsymbol{\pi}_0, n)$ справедливо соотношение

$$\psi_L(\lambda) \sim \chi^2_{(K-1)(\lambda-1)}. \qquad (5)$$

В связи с этим для проверки существования в последовательности ДНК L-профильной периодичности, согласно статистике Пирсона (4), используется *спектр* $\mathbf{D}_L$ *отклонения* анализируемой текстовой *строки* *str* *от* L*-профильной периодичности* (L*-профильности*). Учитывая соотношение (5), для каждого тест-периода λ из диапазона тест-периодов $1...L_{\max} \sim n/5K$ (где n — длина строки *str* в алфавите $\boldsymbol{A}$ размера K) значение $D_L(\lambda)$ этого спектра полагаем равным

$$D_L(\lambda) = \psi_L(\lambda)\big/\chi^2_{crit}\left((K-1)(\lambda-1), \alpha\right),\ \alpha = 0.05, \qquad (6)$$

где $\chi^2_{crit}(N,\alpha)$ — правое критическое значение χ^2-распределения с N степенями свободы (χ^2_N) на уровне значимости $\alpha = 0.05$, т. е. вероятность $P\{\chi^2_N \geq \chi^2_{crit}(N,\alpha)\} = \alpha$.

Алгоритм проверки корректности оценки длины периода L скрытой профильной периодичности в текстовой строке *str* начинается с анализа спектра $\mathbf{D}_1$ отклонения строки *str* от профильно-однородной строки $Str = Tdm_1\left(\mathbf{\Pi}_{str}(1), n\right)$. Согласно статистике Пирсона (4), где $L = 1$, и соотношениям (5) и (6), если значение спектра $\mathbf{D}_1$ на тест-периоде L и его обертонах превышает единицу, то анализируемая текстовая строка *str* признается неоднородной. На рис. 4, *а* показан спектр $\mathbf{D}_1$ отклонения от профильной однородности (1-профильности) для кодирующего района белка — фактора некроза опухоли из генома человека (KEGG, hsa:338872, 1002 bp) [17]. Согласно принятому правилу, этот район признается неоднородным.

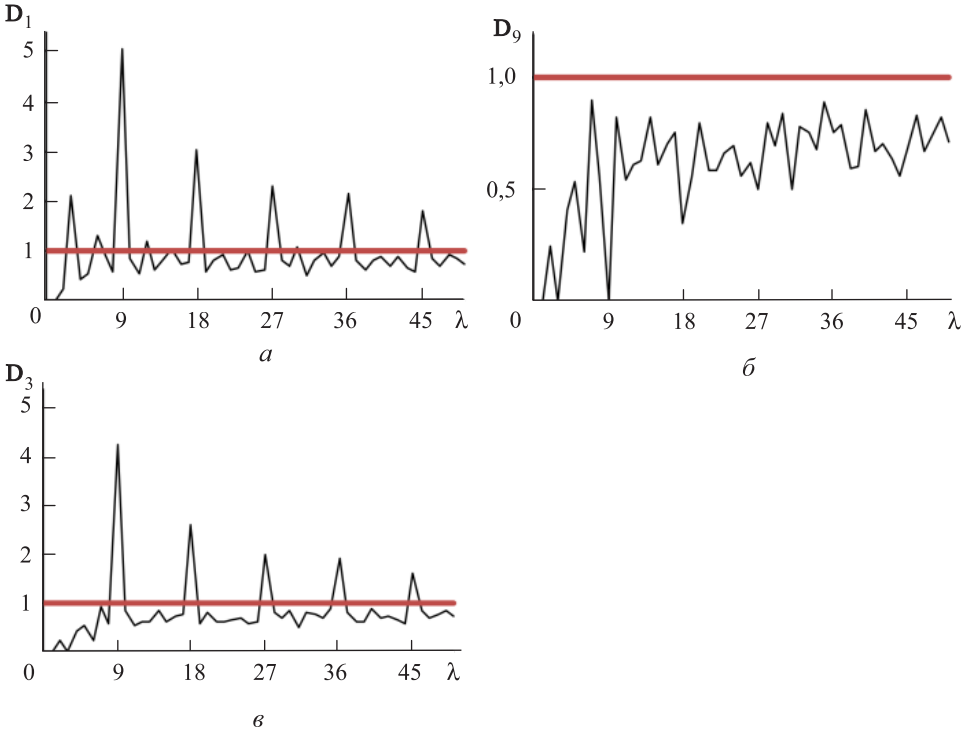


Рис. 4. Спектры отклонения от профильной однородности (1-профильности) (*а*), 9-профильности (*б*) и 3-профильности (*в*) для кодирующего района фактора некроза опухоли из генома человека (KEGG, hsa:338872, 1002 нукл.) [17]

Пусть получена оценка $L>1$ длины периода скрытой профильной периодичности, и строка *str* признана неоднородной. В этом случае для подтверждения достоверности оценки используется спектр $\mathbf{D}_L$ (см. формулу (6)) отклонения анализируемой текстовой строки *str* от L-профильности. Тогда (см. формулы (5) и (6)), если значения спектра $\mathbf{D}_L$ меньше единицы на 95 % тест-периодов из диапазона $1...L_{\max}$, признается гипотеза о том, что в строке *str* распознается скрытая L-профильная периодичность (*L-профильность*). На рис. 4, *б* приведен спектр $\mathbf{D}_9$ для кодирующего района последовательности ДНК (фактора некроза опухоли) из генома человека (KEGG, hsa:338872, 1002 нукл.) [17]. Согласно принятым статистическим критериям, в этой последовательности распознается скрытая 9-профильная периодичность.

Свойство регулярности в кодирующих текстах. На рис. 4, *в* показан спектр $\mathbf{D}_3$ отклонения от 3-профильности в кодирующем районе последовательности ДНК (фактора некроза опухоли) из генома человека. Согласно принятому статистическому критерию, в этой последовательности отсутствует 3-профильная периодичность (3-профильность). Однако заметим, что в характеристическом спектре этого кодирующего района практически все локальные максимумы наблюдаются на тест-периодах, кратных трем (см. рис. 3, *в*). Как правило, на таких тест-периодах также наблюдаются отклонения от однородности (см. рис. 4, *а*). Такое свойство последовательности ДНК в работах [3, 4, 8] введено как *свойство* 3*-регулярности последовательности ДНК*. В общем случае наличие свойства 3-регулярности не гарантирует существования в строке какой-либо скрытой профильности. Однако практически во всех кодирующих районах последовательностей ДНК из исследованных геномов различных организмов, обладающих скрытой профильностью с периодом, кратным трем, проявляется свойство 3-регулярности [3, 4, 8]. При отсутствии скрытой профильности наличие свойства 3-регулярности можно было бы назвать «*размытой триплетной периодичностью*» (или «*размытой 3-профильностью*»).

В работах [3, 4], исходя из анализа максимумов в характеристических спектрах кодирующих районов последовательностей ДНК из геномов различных организмов, на основе введенного достаточно высокого порогового значения индекса 3-регулярности был выработан критерий наличия в последовательности свойства 3-регулярности. Оказалось, что согласно выработанному критерию, практически все кодирующие районы были признаны 3-регулярными. Однако численные эксперименты с некодирующими районами (интронами) последовательностей ДНК из генома человека показали практическое отсутствие в них свойства 3-регулярности и скрытой профильной периодичности с размером периода, кратного трем [3].

Результаты численных экспериментов. Рассмотрим результаты распознавания структурно-статистических свойств в кодирующих районах последовательностей ДНК из геномов различных организмов [3, 4, 8]. Для сравнения приведем результаты численных экспериментов по аналогичному распознаванию для бинарно перекодированных абзацев двух литературных произведений в латинском алфавите на английском (Jerom K. Jerom “Three Men in a Boat”) и итальянском (Carlo Collodi “Le avventure di Pinocchio”) языках [3, 4]. Бинарные кодоны размера 5 синтаксических символов и букв латинского алфавита (знаков) приведены в таблице (пробелы не учитывались). Удовлетворительность предложенного введения кодонов для этих знаков иллюстрирует рис. 5, на котором показана схожесть частотного распределения знаков в произведениях на двух разных языках (английском и итальянском).

Соответствие букв латинского алфавита и символов пунктуации (знаков) бинарным кодонам размера 5 в алфавите {1, 0}

Номер знака, $\mathbf{N_c}$	Знак **C**	Кодон $\mathbf{B_C}$	Номер знака, $\mathbf{N_c}$	Знак **C**	Кодон $\mathbf{B_C}$
1	A a	00000	17	Q q	10101
2	B b	10000	18	R r	01011
3	C c	01000	19	S s	01101
4	D d	00100	20	T t	11010
5	E e	00010	21	U u	10110
6	F f	00001	22	V v	01110
7	G g	11000	23	W w	11100
8	H h	01100	24	X x	11001
9	I i	00110	25	Y y	10011
10	J j	00011	26	Z z	00111
11	K k	10001	27	-	11110
12	L l	01001	28	’ "	11101
13	M m	00101	29	,	11011
14	N n	10010	30	.	10111
15	O o	10100	31	! ?	01111
16	P p	01010	32	Other	11111

Численные эксперименты с бинарным перекодированием показали [3, 4] количественную аналогию структурно-статистических свойств кодирующих районов последовательностей ДНК и бинарно перекодированных абзацев литературных произведений. Скрытая профильная периодичность с периодом, кратным трем, обнаружилась в ~90 % кодирующих районах последовательностей ДНК (в ~76 % —

триплетная периодичность, в ~14 % — периодичность с периодом, кратным трем). Практически во всех кодирующих районах было выявлено свойство 3-регулярности. Аналогичные закономерности наблюдались и в бинарно перекодированных абзацах литературных произведений. Единственное существенное отличие состояло в том, что в бинарно перекодированных абзацах была выявлена профильная периодичность с длиной периода, кратной или равной пяти, и свойство 5-регулярности [3, 4].

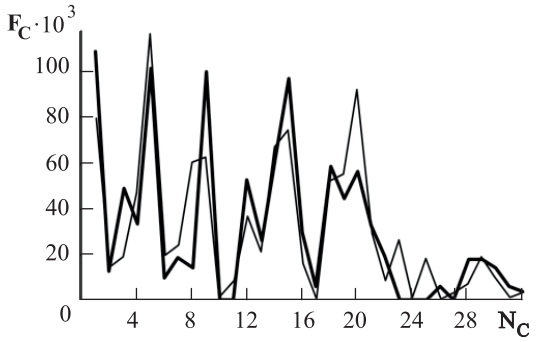


Рис. 5. Распределения частот $\mathbf{F_C}$ встречаемости знаков латинского алфавита в анализируемых литературных произведениях:
—— — Jerom K. Jerom “Three Men in a Boat”; —— — Carlo Collodi “Le avventure di Pinocchio”

Стохастические модели кодирования, обеспечивающие проявление скрытой профильной периодичности в кодирующих текстах. Приведем стохастические модели, отражающие статистическую организацию кодирования в текстах с некоторым смысловым содержанием, объясняющие проявление в них скрытой профильной периодичности и свойства регулярности.

Опишем наиболее общую из предлагаемых моделей. В ее основе лежит *понятие стохастического кодона* $Cdn = STR(L, \mathbf{p})$, профиль которого является стохастическим паттерном и, следовательно, случайной профильной строкой длиной L в текстовом алфавите $\boldsymbol{A} = \langle a_1, a_2, ..., a_K \rangle$.

Пусть $Cdn_1, Cdn_2, ..., Cdn_m$ — такие случайные кодоны размера Λ в алфавите $\boldsymbol{A} = \langle a_1, a_2, ..., a_K \rangle$, что профиль $Ptn = Str_{m\Lambda}(\boldsymbol{\pi})$ случайной строки $STR_0 = Cdn_1 Cdn_2 ... Cdn_m$ является стохастическим паттерном. Тогда случайную строку $STR = \underbrace{STR_0 STR_0 \ldots STR_0}_{q\text{-}times}$, где $q / 5K > 1$, будем рассматривать в качестве *стохастической кодонной мультиполиномиальной модели* (*СКмП-модели*) со случайными кодо-

нами размера Λ в алфавите $\boldsymbol{A}$. В этом случае профиль такой случайной строки STR имеет вид

$$Str = \underbrace{PtnPtn\ldots Ptn}_{q\text{-}times}.$$

Следовательно, этот профиль является L-профильной строкой $Tdm_L(\boldsymbol{\pi},\, qL) = Str$, где $L = m\Lambda$.

Такую модель можно отождествить с мультиполиномиальной схемой независимых испытаний, где в первом испытании происходит реализация кодона Cdn_1, во втором — кодона Cdn_2, …, в m-м испытании — кодона Cdn_m. Затем блок из таких m испытаний повторяется q раз. Как и ранее, вследствие значительного количества повторов ($q > 5K$) паттерна профильности Ptn, статистический анализ реализаций такой случайной строки STR будет приводить к таким же результатам, что и статистический анализ реализаций L-профильной строки Str, т. е. ее профиля. Таким образом, практически во всех реализациях строки (СКмП-модели) STR будет распознаваться L-профильная периодичность. Кроме того, если Λ — простое число, то в реализациях строки (СКмП-модели) STR будет проявляться исключительно Λ-регулярность. Такая СКмП-модель, где $\Lambda = 3$ и $m > 1$, объясняет наличие свойства 3-регулярности в значительном количестве кодирующих районов последовательностей ДНК, в которых распознается L-профильная периодичность с длиной периода, кратной, но не равной трем.

Как было отмечено выше, аналогичное явление для $\Lambda = 5$ наблюдается и в бинарно перекодированных абзацах литературных текстов. В общем случае такое явление было названо двухуровневой организацией кодирования [8, 16]. Первый уровень обусловлен наличием свойства Λ-регулярности. Второй уровень связан с распознаванием скрытой профильной периодичности, длина периода которой кратна, но не равна Λ. В работах [8, 16] наличие такой двухуровневой организации кодирования в кодирующих районах последовательностей ДНК продемонстрировано для аполипопротеинов ($L = 33$), цинковых «пальцев» ($L = 84$) и др. В этих случаях размер паттерна скрытой профильной периодичности в кодирующих районах коррелировал с размером повторяющихся функциональных доменов в кодируемом белке. Следует отметить, что экспериментальное выявление таких доменов — весьма сложная задача.

Когда $m = 1$, предлагаемая стохастическая модель состоит из единственного последовательно повторяющегося стохастического кодона Cdn_1 размера Λ. Следовательно, в реализациях такой модели

будет проявляться Λ-профильность и Λ-регулярность, значит такая модель не служит объяснением наблюдаемой в кодирующих текстах двухуровневой организации кодирования. Такую модель можно назвать *стохастической кодонной полиномиальной моделью* (*СКП-моделью*) [1, 2].

В работах [8, 16] приведены примеры кодирующих районов в последовательностях ДНК, в которых распознавалась различная локальная скрытая профильная периодичность, но на всей длине этого района проявлялось только свойство 3-регулярности. Такое явление в рамках предложенных моделей должно проявляться в том случае, когда весь район является реализацией последовательно соединенных различных последовательных СКмП-моделей со стохастическими кодонами одного размера.

Заключение. Предложены стохастические модели кодирования, описывающие и объясняющие проявление характерных структурно-статистических свойств (профильной периодичности и регулярности), которые присущи кодирующим последовательностям, транслируемым в тексты с некоторым смысловым содержанием. В качестве таких последовательностей рассмотрены кодирующие районы последовательностей ДНК из целой серии геномов различных организмов. Свойства предложенных стохастических моделей продемонстрированы в численных экспериментах с бинарно перекодированными абзацами литературных произведений на английском и итальянском языках.

Полученные результаты численных экспериментов позволяют выдвинуть гипотезу о том, что при кодировании текста, имеющего смысловое содержание, и использовании равномерного кода кодирующим последовательностям присуща скрытая профильная периодичность с длиной периода, равной или кратной размеру кодонов этого кода.

В рамках рассмотренных стохастических моделей предложены методы распознавания скрытой профильной периодичности и регулярности в кодирующих текстах. Эти методы могут быть использованы для разработки алгоритмов и создания на их основе автоматизированного программного обеспечения для распознавания структурно-статистических свойств в последовательностях ДНК. Актуальность его создания обусловлена быстро возрастающим объемом секвенированных геномов различных организмов, для которых необходимо проводить предварительный анализ структуры.

ЛИТЕРАТУРА

[1] Александров А.А., Димитриенко Ю.И. Математическое и компьютерное моделирование — основа современных инженерных наук. *Математическое моделирование и численные методы*, 2014, № 1 (1), с. 3–4. DOI 10.18698/2309-3684-2014-1-None

[2] Зарубин В.С., Кувыркин Г.Н. Особенности математического моделирования технических устройств. *Математическое моделирование и численные методы*, 2014, № 1 (1), с. 5–17. DOI 10.18698/2309-3684-2014-1-517

[3] Chaley M., Kutyrkin V. Stochastic model of homogeneous coding and latent periodicity in DNA sequences. *Journal of Theoretical Biology*, 2016, vol. 390, pp. 106–116.

[4] Кутыркин В.А., Чалей М.Б. Модель организации кодирования в прокариотических организмах. *Математическая биология и биоинформатика*, 2016, т. 11, № 1, с. 24–45. DOI 10.17537/2016.11.24

[5] Chaley M., Kutyrkin V. Spectral-statistical approach for revealing latent regular structures in DNA sequence. *Data Mining Techniques for the Life Sciences*. New York, Springer Science+Business Media, 2016, pp. 315–340.

[6] 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. *Mathematical Biosciences*, 2008, vol. 211, iss. 1, pp. 186–204. DOI 10.1016/j.mbs.2007.10.008

[7] Chaley M.B., Kutyrkin V.A. Structure of proteins and latent periodicity in their genes. *Moscow University Biological Sciences Bulletin*, 2010, vol. 65, iss. 4, pp. 133–135.

[8] Chaley M., Kutyrkin V. Profile-statistical periodicity of DNA coding regions. *DNA Research*, 2011, vol. 18, iss. 5, pp. 353–362. DOI 10.1093/dnares/dsr023

[9] Кутыркин В.А., Чалей М.Б. Спектрально-статистический подход к распознаванию скрытой профильной периодичности в последовательностях ДНК. *Математическая биология и биоинформатика*, 2014, т. 9, вып. 1, с. 33–62. DOI 10.17537/2014.9.33

[10] Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 1999, vol. 27, pp. 573–580.

[11] Sánchez J. 3-base periodicity in coding DNA is affected by intercodon dinucleotides. *Bioinformation*, 2011, vol. 6, pp. 327–329.

[12] Sokol D., Benson G., Tojeira J. Tandem repeats over the edit distance. *Bioinformatics*, 2007, vol. 23, pp. 30–35. DOI 10.1093/bioinformatics/btl309

[13] Marhon S.A., Kremer S.C. Gene prediction based on DNA spectral analysis: a literature review. *Journal Computational Biology*, 2010, vol. 18, pp. 639–676. DOI 10.1089/cmb.2010.0184

[14] Issac B., Singh H., Kaur H., Raghava G.P.S. Locating probable genes using Fourier transform approach. *Bioinformatics*, 2002, 18, pp. 196–197.

[15] Howe E.D., Song J.S. Categorial spectral analysis of periodicity in human and viral genomes. *Nucleic Acids Research*, 2013, vol. 41, pp. 1395–1405. DOI 10.1093/nar/gks1261

[16] Кутыркин В.А., Чалей М.Б. Структурные различия кодирующих и некодирующих районов последовательностей ДНК генома человека. *Инженерный журнал: наука и инновации*, 2012, № 2. DOI 10.18698/2308-6033-2012-2-46

[17] KEGG. Kyoto encyclopedia of genes and genomes. URL: http://www.kegg.jp (дата обращения 23.11.2017).

Статья поступила в редакцию 14.06.2017

Ссылку на эту статью просим оформлять следующим образом:

Кутыркин В.А., Чалей М.Б. Стохастические модели кодирования и распознавание структурно-статистических характеристик кодирующих последовательностей. *Математическое моделирование и численные методы*, 2017, № 3, с. 119–138.

Кутыркин Владимир Андреевич — канд. физ.-мат. наук, доцент кафедры «Вычислительная математика и математическая физика» МГТУ им. Н.Э. Баумана. Автор более 50 печатных работ. Область научных интересов: численные методы, биоинформатика, математическое моделирование. e-mail: vkutyrkin@yandex.ru

Чалей Мария Борисовна — канд. биол. наук, доцент, старший научный сотрудник ИМПБ РАН — филиала ИПМ им. М.В. Келдыша РАН. Автор более 50 печатных работ. Область научных интересов: биоинформатика, математические методы компьютерного анализа генетических текстов. e-mail: maramaria@yandex.ru

Stochastic coding models and recognition of structural and statistical characteristics of coding sequences

© V.A. Kutyrkin[1], M.B. Chalei[2]

[1]Bauman Moscow State Technical University, Moscow, 105005, Russia
[2]Institute of Mathematical Problems of Biology, RAS, branch of Keldysh Institute of Applied Mathematics, RAS (IMPB RAS — Branch of KIAM RAS), Puschino, Moscow Region, 142290, Russia

The paper introduces stochastic models explaining real characteristic regularities of coding regions from genomes of various organisms. Due to the growing volume of data on sequenced genomes, there arises a problem of its computer-aided analysis. By using these models, we developed methods for recognizing the structural and statistical properties of genomic DNA sequences, which can be used to find algorithms and computer programs for the automated processing of large amounts of data. The properties of the proposed stochastic coding models are demonstrated in numerical experiments with binary recoded paragraphs of literary works in English and Italian.

Keywords: *profile line of a random string, profile periodic behaviour, periodic behaviour pattern, stochastic codon, multi-polynomial model*

REFERENCES

[1] Aleksandrov A.A., Dimitrienko Yu.I. *Matematicheskoe modelirovanie i chislennye metody — Mathematical Modeling and Computational Methods*, 2014, no. 1 (1), pp. 3–4. DOI 10.18698/2309-3684-2014-1-None

[2] Zarubin V.S., Kuvyrkin G.N. *Matematicheskoe modelirovanie i chislennye metody — Mathematical Modeling and Computational Methods*, 2014, no. 1 (1), pp. 5–17. DOI 10.18698/2309-3684-2014-1-517

[3] Chaley M., Kutyrkin V. *Journal of Theoretical Biology*, 2016, vol. 390, pp. 106–116.

[4] Kutyrkin V.A., Chaley M.B. *Matematicheskaya biologiya i bioinformatika — Mathematical Biology and Bioinformatics*, 2016, vol. 11, no. 1, pp. 24–45. DOI 10.17537/2016.11.24

[5] Chaley M., Kutyrkin V. Spectral-statistical approach for revealing latent regular structures in DNA sequence. *Data Mining Techniques for the Life Sciences*. New York, Springer Science+Business Media, 2016, pp. 315–340.

[6] Chaley M., Kutyrkin V. *Mathematical Biosciences*, 2008, vol. 211, no. 1, pp. 186–204. DOI 10.1016/j.mbs.2007.10.008

[7] Chaley M.B., Kutyrkin V.A. *Moscow University Biological Sciences Bulletin*, 2010, vol. 65, no. 4, pp. 133–135.

[8] Chaley M., Kutyrkin V. *DNA Research*, 2011, vol. 18, no. 5, pp. 353–362. DOI 10.1093/dnares/dsr023
[9] Kutyrkin V.A., Chaley M.B. *Matematicheskaya biologiya i bioinformatika — Mathematical Biology and Bioinformatics*, 2014, vol. 9, no. 1, pp. 33–62. DOI 10.17537/2014.9.33
[10] Benson G. *Nucleic Acids Research*, 1999, vol. 27, pp. 573–580.
[11] Sánchez J. *Bioinformation*, 2011, vol. 6, pp. 327–329.
[12] Sokol D., Benson G., Tojeira J. *Bioinformatics*, 2007, vol. 23, pp. 30–35. DOI 10.1093/bioinformatics/btl309
[13] Marhon S.A., Kremer S.C. *Journal Computational Biology*, 2010, vol. 18, pp. 639–676. DOI 10.1089/cmb.2010.0184
[14] Issac B., Singh H., Kaur H., Raghava G.P.S. *Bioinformatics*, 2002, 18, pp. 196–197.
[15] Howe E.D., Song J.S. *Nucleic Acids Research*, 2013, vol. 41, pp. 1395–1405. DOI 10.1093/nar/gks1261
[16] Kutyrkin V.A., Chaley M.B. *Inzhenerny zhurnal: nauka i innovatsii — Engineering Journal: Science and Innovation*, 2012, no 2. DOI 10.18698/2308-6033-2012-2-46
[17] *KEGG. Kyoto encyclopedia of genes and genomes*. Available at: http://www.kegg.jp (accessed November 23, 2017).

Kutyrkin V.A., Cand. Sc. (Phys.-Math.), Assoc. Professor, Department of Computational Mathematics and Mathematical Physics, Bauman Moscow State Technical University. Author of over 50 scientific publications. Science research interests include numerical methods, bioinformatics, mathematical simulation. e-mail: vkutyrkin@yandex.ru

Chaley M.B., Cand. Sc. (Bio.), Assoc. Professor, Senior Research Fellow, Institute of Mathematical Problems of Biology, RAS, branch of Keldysh Institute of Applied Mathematics, RAS (IMPB RAS — Branch of KIAM RAS). Author of over 50 scientific publications. Research interests include bioinformatics, mathematical methods of computer analysis of genetic texts. e-mail: maramaria@yandex.ru