УДК 577.2:519.23

В. А. Кутыркин, М. Б. Чалей

ДЕКОМПОЗИЦИЯ СТРУКТУРЫ ПАТТЕРНА СКРЫТОЙ ПРОФИЛЬНОЙ ПЕРИОДИЧНОСТИ В ПОСЛЕДОВАТЕЛЬНОСТЯХ ДНК

Рассмотрены методы декомпозиции случайных паттернов периодичности кодирующих последовательностей ДНК, в которых наблюдается скрытая профильная периодичность. Наблюдаемое в таких последовательностях явление 3-регулярности, обусловленное триплетным генетическим кодом, позволило значительно повысить эффективность процесса декомпозиции.

E-mail: vkutyrkin@yandex.ru, maramaria@yandex.ru

***Ключевые слова*:** *скрытая профильная периодичность, спектрально-статистический подход, кодирующие районы ДНК, декомпозиция паттерна периодичности.*

Понятие скрытой профильной периодичности (профильности) в последовательностях ДНК введено в работе [1]. Для распознавания наличия профильности в ДНК был разработан специальный спектрально-статистический подход [2, 4]. Согласно этому подходу, последовательность ДНК рассматривается как реализация случайной строки, составленной из независимых случайных букв, каждая из которых задается вероятностным распределением четырех букв алфавита ДНК, соответствующих четырем нуклеотидам (нукл.): A — аденин, T — тимин, G — гуанин, C — цитозин. Такую случайную строку называют периодичной, если ее можно представить в виде последовательного повторения некоторой подстроки, называемой паттерном периодичности. Если вся строка периодична, ее называют случайным тандемным повтором, определяемым таким паттерном. В этом случае паттерн, состоящий из строки случайных букв, определяет мультиполиномиальную схему из N независимых испытаний, где N — длина случайного тандемного повтора и его реализаций (последовательностей ДНК). Если паттерн схематически представлен в виде строки $P = C_1, ..., C_L$, где C_1, ..., C_L — случайные буквы, случайный тандемный повтор имеет вид $P...P\ C_1...C_M$, где $0 \le M < L$. При этом паттерн в виде заданной строки случайных букв определяет профиль мультиполиномиальной схемы из N независимых испытаний. Таким образом, мультиполиномиальная схема получена соответствующим сцеплением полиномиальных схем, каждая из которых определяется одной из случайных букв в составе паттерна.

Оценка периода скрытой профильной периодичности. Разработанный спектрально-статистический подход [2, 4] позволяет оценить не только период скрытой профильной периодичности последовательностей ДНК, но и ее случайный паттерн. При таком подходе статистическим материалом является анализируемая последовательность ДНК, т. е. только одна реализация случайной строки. Поэтому при формальном использовании спектрально-статистического подхода в качестве оценки периода скрытой профильной периодичности может быть получен обертон искомого периода и соответствующий ему паттерн периодичности. На практике длина этого паттерна задает размер алфавита случайных букв тандемного повтора, реализацией которого предположительно является анализируемая последовательность ДНК. При более детальном анализе статистического материала можно улучшить оценку длины паттерна, тем самым сократить размер алфавита случайного тандемного повтора. Оценка паттерна периодичности получается в виде профильной матрицы [2, 4], в столбцах которой стоят вероятностные распределения соответствующих случайных букв этого паттерна. На практике, как правило, столбцы профильной матрицы оценки паттерна попарно различны, т. е. длина паттерна совпадает с размером алфавита случайных букв тандемного повтора. Но среди этих столбцов (на заданном уровне значимости) могут встречаться статистически неразличимые. Отождествив эти столбцы с одной случайной буквой, распределение которой получено усреднением этих столбцов, можно сократить размер алфавита случайного тандемного повтора. В результате получится декомпозиция паттерна скрытой профильной периодичности, ведущая к оптимизации оценки паттерна скрытой профильной периодичности анализируемой последовательности ДНК.

В рассматриваемой работе исследуются случайные паттерны последовательностей ДНК кодирующих районов из генома человека, в которых наблюдается скрытая профильная периодичность. Согласно статье [6], в большинстве последовательностей кодирующих районов ДНК (CDS) наблюдается скрытая профильная периодичность, коррелирующая (в некоторых случаях) с известными структурно-функциональными свойствами кодируемых белков [2, 3]. Кроме того, характерные свойства CDS позволяют значительно упростить процесс декомпозиции их паттернов периодичности. В настоящей работе предложены методы, упрощающие декомпозицию паттернов скрытой периодичности в CDS.

Общие методы декомпозиции паттернов скрытой периодичности CDS. Как упоминалось в [6], практически для всех CDS в характеристических спектрах наблюдается регулярная повторяемость пиков через два нуклеотида. Такое явление в работах [2—4] было названо 3-регулярностью последовательностей ДНК, которая обусловлена триплетной структурой универсального генетического кода. Поэтому

далее CDS пошагово разбивают на подстроки в 3 нуклеотида (далее — нукл.). Затем создают текстовые строки, последовательно составленные из нуклеотидов первой, второй и третьей позиций этих подстрок. В результате CDS сводится к трем производным подпоследовательностям, каждая из которых соответствует первой, второй и третьей позициям триплетного разбиения исходной последовательности. С помощью спектрально-статистического подхода [2, 4] такие подпоследовательности исследуют на наличие скрытой профильной периодичности. Далее рассматривают CDS, для которых во всех производных подпоследовательностях выявляется скрытая профильная периодичность. Для случайного паттерна скрытой периодичности анализируемой подпоследовательности производят его внутреннюю декомпозицию, т. е. статистически неотличимые столбцы профильной матрицы паттерна заменяют усредненным столбцом, определяющим одну случайную букву (полиномиальную схему) в алфавите этого паттерна. В результате такой внутренней декомпозиции сокращается алфавит случайного паттерна периодичности производной подпоследовательности.

Возможна также декомпозиция случайных паттернов производных подпоследовательностей, имеющих одинаковый скрытый период. В случае статистической неотличимости таких паттернов их заменяют единым случайным паттерном, полученным в результате их усреднения, что позволяет сократить алфавит паттерна периодичности исходной анализируемой CDS. Такое сокращение назовем внешней взаимной декомпозицией случайных паттернов производных подпоследовательностей. Кроме того, размер алфавита паттерна исходной анализируемой CDS можно сократить путем статистического сравнения случайных букв паттернов различных производных подпоследовательностей.

Рассмотрим гипотетический пример описанной выше декомпозиции CDS, в которой производная подпоследовательность, соответствующая первой позиции триплетов в разбиении исходной последовательности, имеет скрытый период в одну случайную букву; производная подпоследовательность, соответствующая второй позиции, — в две случайные буквы; производная подпоследовательность, соответствующая третьей позиции, — в три. Пусть случайные паттерны этих последовательностей имеют вид А, АБ и БАВ соответственно, где А, Б, В — некоторые случайные буквы (полиномиальные схемы с четырьмя исходами). Тогда оценка скрытого периода всей анализируемой CDS ДНК будет не менее 18 нукл. Формальная обработка статистического материала может дать значение обертона этого периода (36, 54 и т. д.). Следовательно, формальный подход приведет к оценке размера алфавита случайного паттерна периодичности всей анализируемой CDS не менее чем из 18 случайных букв, несмотря на то, что в действительности алфавит этого паттерна состоит их трех случайных букв.

В общем случае если периоды трех производных подпоследовательностей равны M_1, M_2 и M_3 соответственно, скрытый период всей анализируемой последовательности ДНК M = 3НОК (M_1, M_2 и M_3), где НОК — наименьшее общее кратное. Таким образом, при формальной обработке оценка размера алфавита паттерна периодичности будет не менее чем $3M_1M_2M_3$, если числа M_1 M_2 и M_3 — попарно взаимно простые. Но в этом случае без учета внутренней декомпозиции размер алфавита случайного паттерна не превышает числа $M_1 + M_2 + M_3$, т. е. в несколько раз меньше формальной оценки.

Примеры декомпозиции случайных паттернов скрытой периодичности. Первичная декомпозиция паттерна скрытой профильной периодичности выполнена для более чем 12 000 CDS из генома человека (база данных KEGG [5]), кодирующих белки с исследованными физико-химическими свойствами. В результате компьютерной обработки были получены формальные оценки паттерна периодичности для каждой CDS и трех ее производных подпоследовательностей, соответствующих позициям триплетов нуклеотидов. Согласно спектрально-статистическому подходу, возможность такой декомпозиции обусловлена наличием 3-регулярности (систематическое чередование пиков через 2 нукл.) в характеристических спектрах этих CDS. Более детальную декомпозицию случайных паттернов периодичности рассмотрим на конкретных примерах CDS из базы данных KEGG, используя спектрально-статистический подход [2, 4].

Проведем декомпозицию CDS гена аполипопротеина A-II человека. Согласно спектрально-статистическому подходу [2, 4], спектр D_1 отклонения от однородности (рис. 1, *а*) выявляет неоднородность этой последовательности. Ее характеристический спектр H (рис. 1, *б*) указывает на формальную оценку размера случайного паттерна скрытой профильности в 9 нукл. Однако спектр отклонения от 3-профильности позволяет снизить эту оценку до 3 нукл., что не противоречит характеристическому спектру, поскольку амплитуда пика на 9 нукл. (рис. 1, *в*, *г*) незначительно отличается от других пиков на тестируемых периодах, кратных 3. В трех производных подпоследовательностях этого CDS (рис. 2, *а*—*в*) выявляется их однородность, т. е. скрытая периодичность в 1 нукл.

Оценки профильных матриц паттернов периодичности, состоящих из одного столбца, для трех подпоследовательностей, соответствующих первой, второй и третьей позициям триплетов, приведены на рис. 2, *г*. Фактически эти три столбца образуют матрицу случайного паттерна 3-профильной периодичности CDS. Графическая визуализация столбцов этой матрицы показана на рис. 2, *д*, где можно видеть сходство графического представления случайных букв в первой и третьей позициях паттерна периодичности анализируемых CDS.

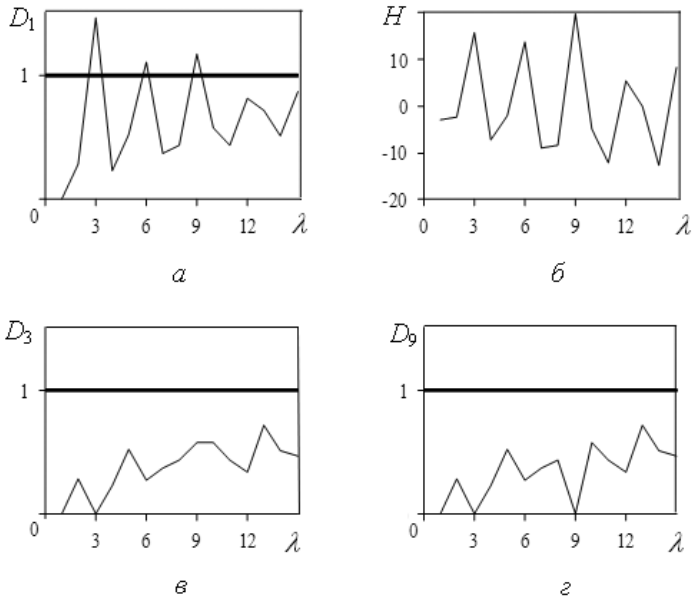


Рис. 1. Спектры D_L отклонения от *L*-профильности для тестируемого периода λ при $L = 1$ (*а*), 3 (*в*), 9 (*г*) и характеристический спектр (*б*) кодирующего района гена аполипопротеина A-II (KEGG, hsa:336, 303 нукл.)

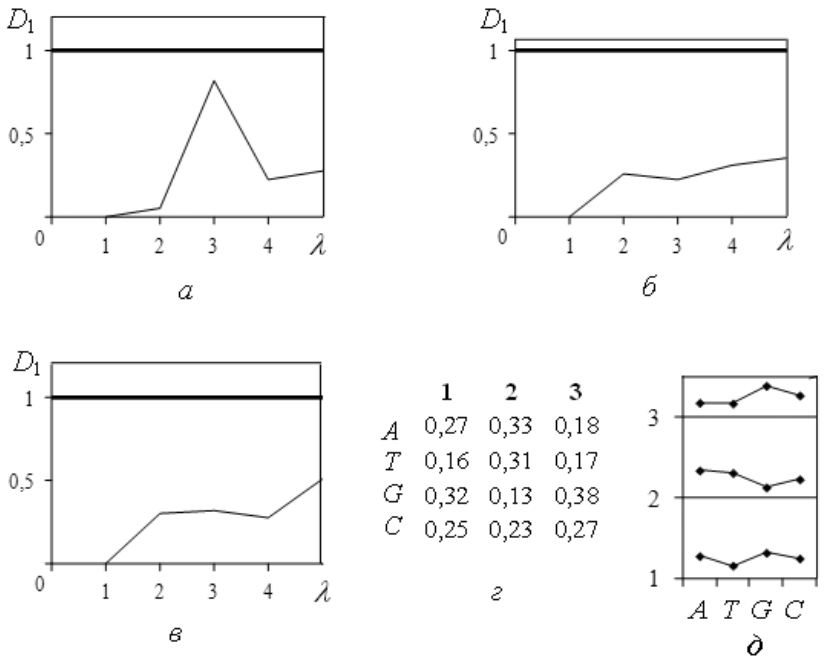


Рис. 2. Спектры D_1 отклонения от 1-профильности (однородности) для тестируемого периода λ производных подпоследовательностей, соответствующих первой (*а*), второй (*б*), третьей (*в*) позициям триплетов гена аполипопротеина A-II (KEGG, hsa:336, 303 нукл.), матрица паттерна 3-профильной периодичности гена аполипопротеина A-II (*г*) и ее визуализация по позициям периода (*д*)

Согласно критерию Пирсона (на уровне значимости 5 %), эти случайные буквы статистически неотличимы. Поэтому их отождествляют с одной случайной буквой C_1, имеющей усредненное распределение $(0{,}23\ 0{,}17\ 0{,}35\ 0{,}26)^{\mathrm{T}}$. Если случайную букву во второй позиции паттерна периодичности CDS обозначить символом C_2, этот паттерн периодичности будет иметь вид $C_1C_2C_1$, т. е. алфавит паттерна скрытой 3-профильности на самом деле состоит из двух букв. Таким образом, декомпозиция паттерна периодичности завершена.

Выполним декомпозицию случайного паттерна периодичности CDS гена аполипопротеина *L* человека (рис. 3). Согласно спектру отклонения от однородности (рис. 3, *а*), эта последовательность является неоднородной и по результатам компьютерной обработки обладает скрытой профильной периодичностью в 42 нукл.

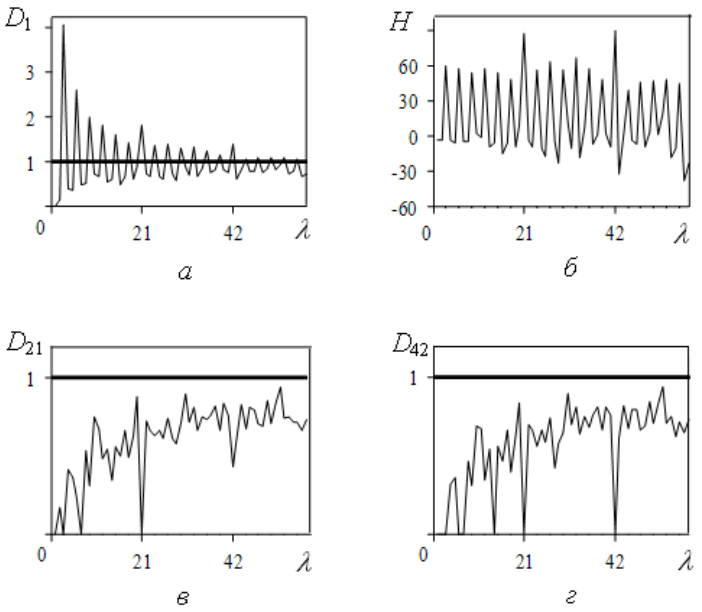


Рис. 3. Спектры D_L отклонения от *L*-профильности при *L* = 1 (*а*), 21 (*в*), 42 (*г*) и характеристический спектр *H* (*б*) гена аполипопротеина *L* (KEGG, hsa:8 542, 1 197 нукл.)

Анализ характеристического спектра (рис. 3, *б*), спектров отклонения от 21-профильности (рис. 3, *в*) и 42-профильности (рис. 3, *г*) позволяет сделать вывод о наличии в этой последовательности скрытой периодичности в 21 нукл. В свою очередь, анализ спектров отклонения от однородности (рис. 4, *а*—*в*) для трех производных подпоследовательностей, соответствующих позициям триплетов генетического кода, показал однородность первой и третьей подпоследовательностей и выявил неоднородность второй подпоследовательности. Следовательно, первая и третья подпоследовательности имеют период в одну случайную букву.

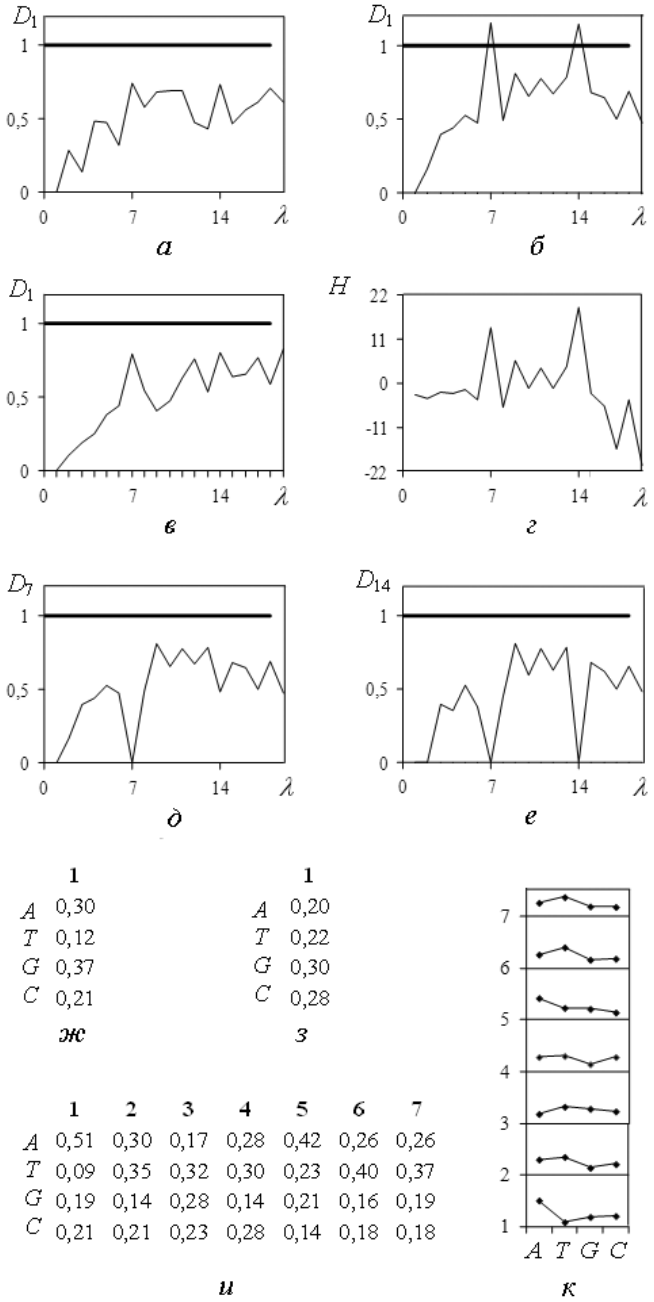


Рис. 4. Спектры D_1 отклонения от 1-профильности для тестируемого периода λ производных подпоследовательностей, соответствующих первой (*а*), второй (*б*), третьей (*в*) позициям триплетов гена аполипопротеина *L* (KEGG, hsa:8542, 1197 нукл.), характеристический спектр *H* (*г*) и спектры D_L отклонения от *L*-профильности для производной подпоследовательности, соответствующей второй позиции при $L = 7$ (*д*), $L = 14$ (*е*), матрицы паттернов периодичности производных подпоследовательностей, соответствующих первой (*ж*), третьей (*з*), второй (*и*) позициям триплетов в гене и визуализация матрицы паттерна (*и*) по позициям периода (*к*)

С формальной точки зрения, согласно характеристическому спектру (рис. 4, *г*) и спектру отклонения от 14-профильности (рис. 4, *е*), во второй подпоследовательности выявляется скрытый период в 14 нукл. Однако характеристический спектр (см. рис. 4, *г*) и спектр отклонения от 7-профильности (см. рис. 4, *д*) указывают на наличие в ней 7-профильной периодичности, т. е. тестируемого периода в 14 нукл., являющегося обертоном периода в 7 нукл.

На рис. 4, *ж*—*и* приведены также и профильные матрицы случайных паттернов периодичности для трех анализируемых производных подпоследовательностей. Критерий Пирсона показывает статистическое отличие случайных букв B_1 и B_3 паттернов периодичности для первой и третьей производных подпоследовательностей.

Для второй подпоследовательности статистическая обработка ее профильной матрицы (рис. 4, *и*), графически представленной на рис. 4, *к*, позволяет представить случайный паттерн периодичности этой последовательности в виде строки случайных букв: $C_1C_2C_3C_4C_5C_2C_2$. Значит, случайный паттерн периодичности полной анализируемой CDS имеет алфавит из семи случайных букв и состоит из 21 случайной буквы. Таким образом, декомпозиция завершена.

Декомпозиция CDS гена аполипопротеина *E* человека дает следующие результаты. Согласно спектрально-статистическому подходу, в этой последовательности наблюдается скрытая 33-профильность (рис. 5).

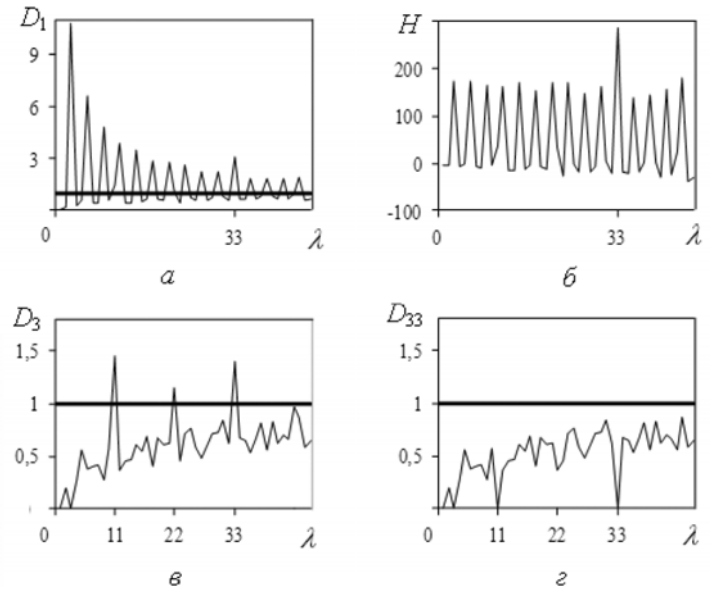


Рис. 5. Спектры D_L отклонения от *L*-профильности для тестируемого периода λ при L = 1 (*а*), 3 (*в*), 33 (*г*) и характеристический спектр H (*б*) кодирующего района гена аполипопротеина *E* (KEGG, hsa:348, 954 нукл.)

Анализ позиций триплетов из разбиения исходной последовательности выявляет наличие скрытой 11-профильности в первых двух позициях и однородность в третьей позиции (рис. 6). Сравнение профильных матриц (рис. 7, *а*, *б*) паттернов периодичности производных подпоследовательностей для первых двух позиций выявляет их статистическое различие (рис. 7, *в*, *г*). Следовательно, случайный паттерн периодичности анализируемых CDS имеет алфавит не более чем из 23 случайных букв.

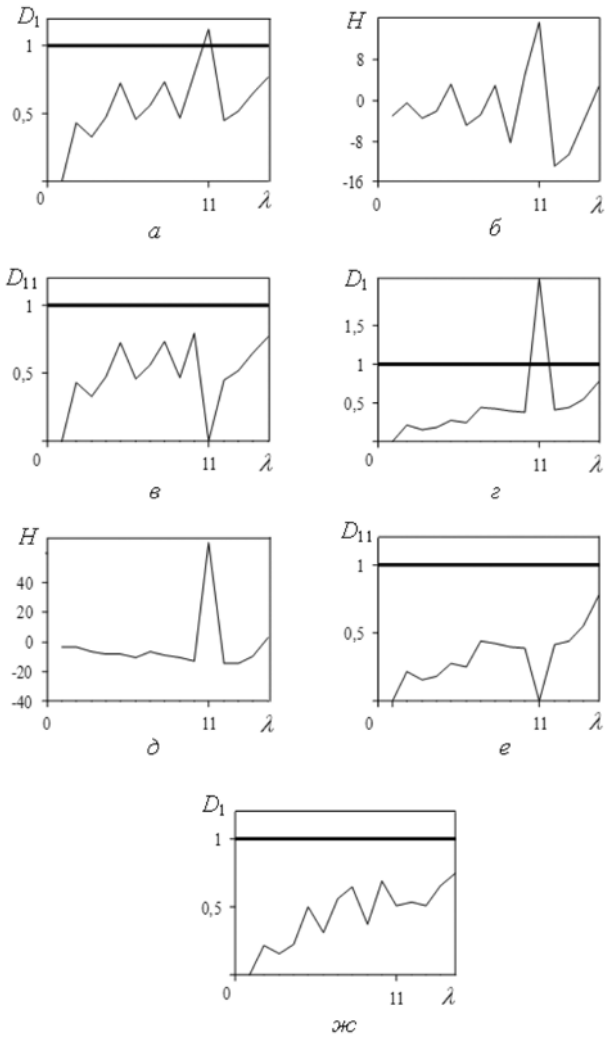


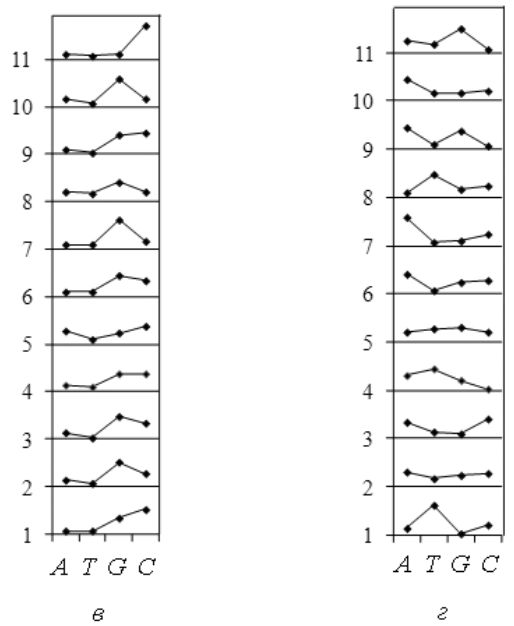
Рис. 6. Спектры D_1 отклонения от 1-профильности для тестируемого периода λ производных подпоследовательностей гена аполипопротеина *E* (KEGG, hsa: 348, 954 нукл.) для первой (*а*), второй (*г*), третьей (*ж*) позиций триплетов гена, характеристические спектры *H* для производных подпоследовательностей, соответствующих первой (*б*) и второй (*д*) позициям триплетов, и спектры D_{11} отклонения этих подпоследовательностей от 11-профильности для первой (*в*) и второй (*е*) позиций

	1	2	3	4	5	6	7	8	9	10	11
A	0,07	0,14	0,14	0,14	0,28	0,10	0,10	0,21	0,10	0,17	0,11
T	0,07	0,07	0,04	0,10	0,10	0,10	0,10	0,17	0,04	0,07	0,07
G	0,34	0,52	0,48	0,38	0,24	0,46	0,63	0,41	0,41	0,59	0,11
C	0,52	0,27	0,34	0,38	0,38	0,34	0,17	0,21	0,45	0,17	0,71

а

	1	2	3	4	5	6	7	8	9	10	11
A	0,14	0,31	0,34	0,31	0,21	0,41	0,59	0,10	0,45	0,45	0,25
T	0,62	0,17	0,14	0,45	0,28	0,07	0,07	0,49	0,10	0,17	0,18
G	0,03	0,24	0,10	0,21	0,30	0,24	0,10	0,17	0,38	0,17	0,50
C	0,21	0,28	0,42	0,03	0,21	0,28	0,24	0,24	0,07	0,21	0,07

б



в *г*

Рис. 7. Матрицы паттернов 11-периодичности по первой (*а*) и второй (*б*) позициям триплетов в гене аполипопротеина *E* (KEGG, hsa:348, 954 нукл.), визуализация матрицы паттерна по первой (*в*) и второй (*г*) позициям

В рамках рассматриваемой работы предложены методы декомпозиции паттернов скрытой периодичности в CDS ДНК. В основе декомпозиции лежит анализ наблюдаемой в таких последовательностях 3-регулярности, обусловленной генетическим триплетным кодом. Явление 3-регулярности позволило значительно упростить процесс декомпозиции. С помощью такой декомпозиции возможно в несколько раз сократить размер алфавита случайного паттерна скрытой профильной периодичности, что было продемонстрировано на ряде примеров из генома человека.

Аналогичные методы можно использовать для распознавания паттерна периодичности в мультиполиномиальных схемах испытаний с одинаковым количеством исходов.

СПИСОК ЛИТЕРАТУРЫ

1. Chaley M., Kutyrkin V. Model of perfect tandem repeat with random pattern and empirical homogeneity testing polycriteria for latent periodicity revelation in biological sequences // Math. Biosci., 2008, N 211. P. 186–204.
2. Chaley M., Kutyrkin V. Profile-statistical periodicity of DNA coding regions // DNA Res., 2011, N 18. P. 353–362.
3. Chaley M. B., Kutyrkin V. A. Structure of proteins and latent periodicity in their genes // Moscow Univ. Biol. Sci. Bull., 2010, N 65. P. 133–135.
4. Кутыркин В. А., Чалей М. Б. Распознавание различных уровней в организации кодирования генетической информации // Вестник МГТУ им. Н.Э. Баумана. Сер. Естественные науки. 2011. Спец. выпуск Математическое моделирование. С. 200–215.
5. Kanehisa M. et al. KEGG for integration and interpretation of large-scale molecular data sets // Nucleic Acids Res. 1–6. 2011 doi:10.1093/nar/gkr988.
6. 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. Спец. выпуск № 3. Математическое моделирование. М., 2012. С. 146–157.

Статья поступила в редакцию 03.07.2012.