Математическая биология и биоинформатика. 2012. Т. 7. № 1. С. 461–475.
URL: http://www.matbio.org/2012/Pugacheva_7_461.pdf

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

УДК: 577.212.2

Поиск парных точек сдвига фазы триплетной периодичности в генах из 17 бактериальных геномов

©2012 Пугачева В.М.[*1], Коротков А.Е., Коротков Е.В.[1]**

[1]*Центр «Биоинженерия», Российская академия наук, Москва, 117312, Россия*

Аннотация. В настоящей работе был разработан математический метод для поиска парных сдвигов фазы триплетной периодичности, которые могут представлять собой потенциальные сдвиги рамки считывания в генах, возникающие при вставках сравнительно больших фрагментов ДНК. Был разработан программный комплекс на основе предложенного математического метода и проверено присутствие парных точек сдвига фазы триплетной периодичности в генах из 17 бактериальных геномов. Наши результаты показывают, что примерно 1% бактериальных генов в 17 изученных геномах имеет такой парный сдвиг фазы триплетной периодичности. В статье разработан метод визуализации парных сдвигов фазы триплетной периодичности и приведены примеры таких парных сдвигов. Результаты работы нашли частичное подтверждение при поиске подобий аминокислотных последовательностей, созданных по альтернативным рамкам считывания. Обсуждается связь парных сдвигов фазы триплетной периодичности со сдвигами рамки считывания в генах.

Ключевые слова: *триплетная периодичность, рамка считывания, сдвиги, фаза.*

ВВЕДЕНИЕ

Небольшие вставки фрагментов ДНК могут сравнительно часто осуществляться в генах [1,2]. Если длины таких вставок не кратны трем основаниям ДНК, то такие события приводят также к сдвигу рамки считывания после окончания района вставки. Такие вставки могут значительным образом изменить аминокислотную последовательность гена, и важно понять их вклад в осуществление сдвигов рамки считывания [3–5]. В настоящее время используемые математические методы для поиска сдвигов рамки считывания можно разделить на две группы. Обе эти группы объединяет то, что, кроме анализируемой последовательности оснований ДНК, требуется еще и некоторая дополнительная информация. Первая группа методов использует внешние данные в виде банка данных аминокислотных последовательностей и программные комплексы для поиска подобий [6–9]. В этих алгоритмах создаются аминокислотные последовательности, соответствующие альтернативным рамкам считывания, и для них производится поиск подобий в базе данных. Если такое подобие будет найдено, то можно говорить о том, что в анализируемом гене был сдвиг рамки считывания. В этой группе методов роль необходимой дополнительной информации играет банк данных аминокислотных последовательностей. Вторая группа методов использует нуклеотидную последовательность анализируемого гена для поиска сдвигов рамки считывания. В качестве дополнительной информации выступает выборка генов, для которых уже известно, что в них существует сдвиг рамки считывания [10–14]. В результате в

[*] virentis@gmail.com
[**] genekorotkov@gmail.com

анализируемом гене ищутся некие обобщенные свойства тех районов ДНК, где такие сдвиги уже были найдены. В качестве способов описания таких коллективных свойств могут выступать весовые матрицы, частоты *k*-tuples, HMM модели, нейронные сети и некоторые другие математические подходы [10–14].

Однако используемые методы имеют определенные недостатки, которые ограничивают их применение, и они связаны с необходимостью использовать дополнительную информацию. Для методов первой группы эти ограничения состоят в том, что банк данных аминокислотных последовательностей должен содержать последовательность, которая могла существовать до сдвига рамки считывания в гене и которая должна иметь значимое подобие с аминокислотными последовательностями, созданными по альтернативным рамкам считывания. Очень часто такая последовательность может отсутствовать в банке данных, или она может не иметь значимого подобия. В силу этого поиск сдвигов рамки считывания данным методом становится невозможен. Методы второй группы имеют ограничения другого свойства. Они связаны с тем, что при поиске сдвигов рамки считывания используется идея создания неких общих статистических свойств районов генов, где уже наблюдается сдвиг рамки считывания. На основе этих свойств создаются некие разрешающие математические правила, которые применяются для поиска сдвигов рамки считывания в других генах. Однако, как это было показано ранее [15], статистические свойства последовательностей генов могут быть различны, что приводит к тому, что гены обладают различными классами триплетной периодичности. Объединение разных генов с известными сдвигами рамки считывания может приводить к тому, что многие статистические особенности последовательностей становятся слабо выраженными, что может значительно ухудшать мощность распознавания сдвигов рамки считывания.

Ранее был использован подход по поиску сдвига фазы триплетной периодичности для поиска потенциальных сдвигов рамки считывания в генах [16,17]. Этот подход имеет одно существенное отличие от перечисленных выше используемых методов, которое состоит в том, что для его работы не требуется дополнительная информация в виде банка данных аминокислотных последовательностей (методы первой группы) или знания последовательностей оснований ДНК с доказанными сдвигами рамки считывания (методы второй группы). Для идентификации сдвигов рамки считывания используется понятие триплетной периодичности генов и сдвига фазы триплетной периодичности [15–17]. Триплетная организация последовательностей ДНК, кодирующих белки, является общим свойством всех известных на настоящее время живых систем [18–27], и она привязана к рамке считывания, существующей в гене [15]. Причина этого заключается как в структуре генетического кода, который практически одинаков как у представителей прокариот, так и эукариот, так и в насыщенности белков определенными аминокислотами [28–30]. Если на фоне триплетной периодичности в гене произойдет сдвиг фазы, то это можно будет заметить, так как произойдет сдвиг между триплетной периодичностью и рамкой считывания. Поскольку триплетную периодичность последовательности ДНК достаточно трудно существенно изменить посредством сравнительно небольшого числа замен оснований [31], то такой сдвиг может сохраняться сравнительно долго. Присутствие такого сдвига между триплетной периодичностью нуклеотидной последовательности и рамкой считывания может служить указанием на сдвиг рамки считывания в анализируемом гене [17].

Однако предложенный математический аппарат, позволяющий обнаруживать сдвиг фазы триплетной периодичности, несвободен от ряда недостатков. Главный из них состоит в том, что разработанный способ может находить сдвиг фазы рамки считывания, созданный вставкой сравнительно коротких последовательностей (не кратных трем основаниям), с длиной меньше нескольких десятков нуклеотидов. Если происходит вставка более длинного фрагмента, то такая вставка может существенно

менять триплетную периодичность около района сдвига фазы триплетной периодичности, что сильно затрудняет ее обнаружение данным методом.

В настоящей работе ставились две задачи. Во-первых, мы хотели улучшить разработанный ранее математический метод для поиска потенциальных сдвигов фазы триплетной периодичности [17] с целью учета возможных сдвигов рамки считывания, возникающих при вставках сравнительно больших фрагментов ДНК (> 100 оснований ДНК). Во-вторых, мы хотели проверить улучшенным алгоритмом присутствие парных точек сдвига фазы триплетной периодичности, образование которых может быть связано с длинными вставками фрагментов ДНК в ген. Наши результаты показывают, что примерно 1% бактериальных генов из 17 изученных геномов имеет парный сдвиг фазы триплетной периодичности, который может быть обусловлен вставкой в ген сравнительно длинного фрагмента ДНК.

МЕТОД ПОИСКА ПАРНЫХ ТОЧЕК СДВИГА ФАЗЫ ТРИПЛЕТНОЙ ПЕРИОДИЧНОСТИ В ГЕНАХ

Мы предполагаем, что имеем нуклеотидную последовательность $S = \{s(k), k = 1,2,\ldots l\}$, где каждое основание $s(k)$ выбирается из алфавита $A = \{a,t,c,g\}$, l есть длина последовательности S. Введем три рамки считывания в последовательности S и обозначим их как T_1, T_2 и T_3. Основание $s(1)$ последовательности S является первым, третьим и вторым основанием кодона для рамки считывания T_1, T_2 и T_3 соответственно. Рамка считывания T_1 реально существует в последовательности S, а рамки считывания T_2 и T_3 можно рассматривать как гипотетические.

Определим также три матрицы триплетной периодичности $M_1(i_1,i_2)$, $M_2(i_1,i_2)$ и $M_3(i_1,i_2)$, определенные для рамок считывания T_1, T_2 и T_3 для фрагмента последовательности S в координатах от i_1 до i_2. Обозначим этот фрагмент как $S(i_1,i_2)$. Элементы матриц $m_1(i,j)$, $m_2(i,j)$ и $m_3(i,j)$ показывают число оснований типа i в последовательности S ($i=1$ для a, $i=2$ для t, $i=3$ для c, $i=4$ для g), которые встречается в j позиции кодона (j может быть равно 1, 2 или 3) для рамок считывания T_1, T_2 и T_3 соответственно [15]. За фазу матриц триплетной периодичности M_1, M_2, M_3 возьмем координату k того основания из $s(1)$, $s(2)$ и $s(3)$, которое входит в первую позицию триплета рамок считывания T_1, T_2 и T_3, соответственно. Для матриц M_1, M_2, M_3 начальная фаза равна 1, 2 и 3 соответственно.

Для того чтобы найти парные точки сдвига фазы триплетной периодичности в символьной последовательности $S(x)$, нужно определить меру подобия двух матриц триплетной периодичности.

1. Мера подобия матриц триплетной периодичности

Для поиска парных точек сдвига фазы нам необходимо ввести меру подобия матриц триплетной периодичности U. В случае отсутствия сдвига фазы триплетной периодичности в точке x должно выполняться условие:

$$\begin{cases} U(M_1(x-L,x), M_1(x+1,x+1+L)) \ge U_0 \\ U(M_1(x-L,x), M_2(x+2,x+2+L)) < U_0 \\ U(M_1(x-L,x), M_3(x+3,x+3+L)) < U_0 \end{cases} \quad (1)$$

Это условие означает, что матрицы триплетной периодичности слева и справа от точки x, определенные по рамке считывания T_1, подобны друг другу. Введем количественную меру подобия матриц триплетной периодичности. Для этого мы

каждую из триплетных матриц в формуле (1) преобразуем к матрицам, где каждый элемент будет представлять собой аргумент нормального распределения. Для этого воспользуемся аппроксимацией биномиального распределения по формуле:

$$n(i,j) = \frac{m(i,j) - Lp(i,j)}{\sqrt{Lp(i,j)(1-p(i,j))}}, \tag{2}$$

$$p(i,j) = \frac{x(i)y(j)}{L^2}, \tag{3}$$

где $m(i,j)$ есть элемент матрицы $M_1(x-L,x)$, $M_1(x+1,x+1+L)$, $M_1(x+2,x+2+L)$ и $M_1(x+3,x+3+L)$, $n(i,j)$ – нормально распределенная величина, $x(i,j) = \sum_{i=1}^{4} m(i,j)$, $y(i,j) = \sum_{j=1}^{3} m(i,j)$. В результате мы получаем для каждой из матриц $M_1(x-L,x)$, $M_1(x+1,x+1+L)$, $M_1(x+2,x+2+L)$ и $M_1(x+3,x+3+L)$ соответствующие им матрицы V_1, W_1, W_2 и W_3. Затем для каждого элемента матриц V_1 и W_k (k = 1, 2, 3) мы можем рассчитать величину:

$$z_{1k}(i,j) = v_1(i,j)w_k(i,j). \tag{4}$$

Произведение двух независимых случайных величин, распределенных по нормальному закону, имеет плотность распределения:

$$f(z) = p^{-1}K_0(|z|), \tag{5}$$

где K_0 есть модифицированная функция Бесселя второго рода (функция Макдональда). Это позволяет рассчитать вероятность того, что $P(z > z_{1k}(i,j))$. Затем мы проводили вычисление обратной функции нормального распределения вероятностей и рассчитывали такой аргумент нормального распределения $x_{1k}(i,j)$, для которого $P(x > x_{1k}(i,j)) = P(z > z_{1k}(i,j))$. Затем мы рассчитывали величину:

$$D(1,k) = x_{1k}(i,j). \tag{6}$$

Если справедлива гипотеза о том, что матрицы V_1 и W_k (k = 1, 2, 3) являются случайными некоррелированными друг с другом матрицами, то в этом случае $D(1,k)$ имеет приблизительно нормальное распределение с математическим ожиданием, равным нулю, и дисперсией, равной 6.0. Таким образом, вероятность того, что $P(X > D(1,k)), X \sim N(0,6)$, будет показывать вероятность того, что подобие матриц обусловлено случайными факторами. Если $D(1, k)$ будет достаточно велика, то вероятность того, что две матрицы подобны друг другу случайным образом можно будет отвергнуть. Можно считать, что в качестве меры U формулы (1) при сравнении двух матриц мы берем $D(1, k)$.

2. Метод поиска парных точек сдвига фазы триплетной периодичности в генах

Вначале выделяли позиции x в последовательности S равные $9n + 1$, где $n = 0,1,2,3\ldots$ Для каждой позиции x в интервале от x до $x + 59$ мы строили матрицу триплетной периодичности M_n. Таким образом, всего было создано $K = (L–60)/9 + 1$ матриц. Затем мы сравнивали каждую матрицу с каждой и рассчитывали три матрицы $Sim(1,1)$, $Sim(1,2)$, $Sim(1,3)$ согласно формуле (6). При этом вторая матрица триплетной периодичности бралась без сдвига, со сдвигом на одно или два основания, соответственно. Каждая матрица имеет размерность (K×K). Матрицы $Sim(1,1)$, $Sim(1,2)$, $Sim(1,3)$ показывают подобие матрицы с индексом i матрице с индексом j (взятой без сдвига или со сдвигом на одно или два основания соответственно), определенное по

формуле (6). Затем в интервале от 1 до K мы выделяли две точки k_1 и k_2, $k_1 \neq k_2$. Мы рассчитывали следующие суммы:

$$W_1 = \sum_{1\le i<k_1}\sum_{1\le j<k_1} Sim_{ij}(1,1) + \sum_{k_1\le i\le k_2}\sum_{k_1\le j\le k_2} Sim_{ij}(1,1) + \sum_{k_2<i\le K}\sum_{k_2<j\le K} Sim_{ij}(1,1), \quad (7)$$

$$W_{11} = \sum_{1\le i<k_1}\sum_{k_1\le j\le k_2} Sim_{ij}(1,2) + \sum_{1\le i<k_1}\sum_{k_2<j\le K} Sim_{ij}(1,3) + \sum_{k_1\le i\le k_2}\sum_{k_2<j\le K} Sim_{ij}(1,2), \quad (8)$$

$$W_{12} = \sum_{1\le i<k_1}\sum_{k_1\le j\le k_2} Sim_{ij}(1,2) + \sum_{1\le i<k_1}\sum_{k_2<j\le K} Sim_{ij}(1,1) + \sum_{k_1\le i\le k_2}\sum_{k_2<j\le K} Sim_{ij}(1,3), \quad (9)$$

$$W_{21} = \sum_{1\le i<k_1}\sum_{k_1\le j\le k_2} Sim_{ij}(1,3) + \sum_{1\le i<k_1}\sum_{k_2<j\le K} Sim_{ij}(1,1) + \sum_{k_1\le i\le k_2}\sum_{k_2<j\le K} Sim_{ij}(1,2), \quad (10)$$

$$W_{22} = \sum_{1\le i<k_1}\sum_{k_1\le j\le k_2} Sim_{ij}(1,3) + \sum_{1\le i<k_1}\sum_{k_2<j\le K} Sim_{ij}(1,2) + \sum_{k_1\le i\le k_2}\sum_{k_2<j\le K} Sim_{ij}(1,3), \quad (11)$$

$$W_3 = \sum_{1\le i<K}\sum_{1\le j<K} Sim_{ij}(1,1). \quad (12)$$

Затем рассчитывались 4 итоговые суммы V_{11}, V_{12}, V_{21}, V_{22}:

$$V_{11} = W_1 + W_{11} - W_3, \quad (13)$$

$$V_{12} = W_1 + W_{12} - W_3, \quad (14)$$

$$V_{21} = W_1 + W_{21} - W_3, \quad (15)$$

$$V_{22} = W_1 + W_{22} - W_3. \quad (16)$$

Сумма V_{11} соответствует ситуации, когда в точках k_1 и k_2 мы имеем вставку одного основания ДНК. Сумма V_{12} соответствует ситуации, когда в точке k_1 мы имеем вставку одного основания, а в точке k_2 мы имеем вставку двух оснований ДНК. Сумма V_{21} соответствует ситуации, когда в точке k_1 мы имеем вставку двух оснований, а в точке k_2 мы имеем вставку одного основания ДНК. И сумма V_{22} соответствует ситуации, когда в точках k_1 и k_2 мы имеем вставку двух оснований ДНК. Итоговое значение сумм V_{11}, V_{12}, V_{21}, V_{22} было уменьшено на величину W_3.

Такое суммирование было проделано для всех значений k_1 и k_2, и для сумм V_{11}, V_{12}, V_{21}, V_{22} находились такие координаты $k_1^{\max}$ и $k_2^{\max}$, которые имели максимальное значение $V_{11}^{\max}$, $V_{12}^{\max}$, $V_{21}^{\max}$, $V_{22}^{\max}$.

Для определения статистической значимости величин $V_{11}^{\max}$, $V_{12}^{\max}$, $V_{21}^{\max}$, $V_{22}^{\max}$ мы генерировали случайные последовательности с такой же длиной, как и анализируемый ген, и с таким же уровнем триплетной периодичности. Для этого последовательность гена разбивалась на три подпоследовательности. Первая из них, обозначим ее как C_1, получена выбором из последовательности оснований, которые стоят на номерах $i = 1 + 3n$. Вторая последовательность C_2 получается посредством выбора оснований, стоящим на позициях $i = 2 + 3n$, а третья последовательность C_3 получена выбором оснований, стоящих на позициях с номерами $i = 3 + 3n$.

Далее мы создавали датчиком случайных чисел последовательности R_1, R_2 и R_3, которые имеют такую же длину, как и последовательности C_1, C_2, C_3. Затем мы упорядочивали последовательности R_1, R_2 и R_3 по возрастанию и запоминали порядок сделанных перестановок в каждой последовательности. После этого мы переставляли нуклеотиды в последовательностях C_1, C_2 и C_3 так же, как это мы сделали при упорядочивании последовательностей R_1, R_2 и R_3, по возрастанию. После такого перемешивания последовательностей R_1, R_2 и R_3 мы создавали случайную последовательность R. В последовательности R на позициях $i = 1 + 3n$ стояли нуклеотиды из последовательности R_1, на позициях $i = 2 + 3n$ стояли нуклеотиды из последовательности R_2 и на позициях $i = 3 + 3n$ стояли нуклеотиды из

последовательности R_3. Длина последовательности R была равна длине исходного гена, также в ней был сохранен такой же состав нуклеотидов, как и в исходном гене.

Для каждого гена мы генерировали 500 последовательностей R. Для каждой последовательности R мы определяли соответствующие значения $V_{11}^{\max}$, $V_{12}^{\max}$, $V_{21}^{\max}$, $V_{22}^{\max}$ и затем рассчитывали величину:

$$Z_k = \frac{V_k^{\max} - \overline{V_k^{\max}}}{\sqrt{D\left(V_k^{\max}\right)}}. \qquad (17)$$

Здесь k принимает значения 11, 12, 21 или 22. Величины $\overline{V_k^{\max}}$ и $D\left(V_k^{\max}\right)$ для каждого гена определялись на множестве последовательностей R.

3. Использование метода Монте-Карло для определения порогового значения Z_k

Для поиска пороговых значений для Z_k мы использовали последовательности генов, собранные в 17 бактериальных геномах. Всего в этих геномах содержится 66936 генов. Мы создали случайные последовательности путем перемешивания последовательности оснований каждого гена. Это позволяет сохранить такое же распределение длин случайных последовательностей и то же распределение частот оснований, как в банке данных Kegg. Для сохранения триплетной периодичности в случайной последовательности перемешивание производилось таким же образом, как это описано в пункте 2. Такой способ перемешивания позволяет сохранить триплетную периодичность в последовательности, одновременно с этим в последовательности остаются только те сдвиги фазы триплетной периодичности, которые были связаны со случайными факторами. После создания банка случайных последовательностей мы определяли уровни Z_{11}, Z_{12}, Z_{21}, Z_{22}, для которых число генов с парными сдвигами фазы было около 18% от того, что мы нашли для последовательностей генов из 17 генов бактерий. Обозначим эти уровни как Z_{11}^0, Z_{12}^0, Z_{21}^0, Z_{22}^0. Эти уровни оказались равными 3.1, 2.4, 2.5 и 2.6 соответственно. Уровень в 18% был выбран для того, чтобы можно было сравнить полученные в настоящей работе результаты с результатами, полученными ранее [17].

4. Поиск одиночных сдвигов фазы триплетной периодичности

Одиночные сдвиги фазы триплетной периодичности могут приводить к значению Z_k больше, чем введенная пороговая величина. Поэтому все найденные гены с парными сдвигами фазы триплетной периодичности испытывались на существование в них одиночных сдвигов фазы триплетной периодичности. Поиск одиночных точек сдвига фазы проводился аналогично пункту 2, только в качестве величины W_k бралась величина:

$$W_k = \sum_{1\le i<k_1}\sum_{1\le j<k_1} Sim_{ij}(1,1) + \sum_{1\le i<k_1}\sum_{k_1\le j\le K} Sim_{ij}(1,k) + \sum_{k_1\le i\le K}\sum_{k_1\le j\le K} Sim_{ij}(1,1). \qquad (18)$$

В данном случае k было равно 1 или 2, что соответствует вставке в точке с координатой k одного или двух оснований ДНК. Затем рассчитывалась величина:

$$V_k = W_k - W_3. \qquad (19)$$

Здесь k принимало значения 1 или 2. Для этих величин рассчитывались соответствующие значения Z_k по формуле (17). Пороговый уровень для величины Z_k (пункт 3) оказался равным 3.8. Если значения $V_{11}^{\max}$, $V_{12}^{\max}$, $V_{21}^{\max}$, $V_{22}^{\max}$ были большими, чем значения Z_1 и Z_2, то мы считали, что в последовательности S существуют парные сдвиги фазы триплетной периодичности. В противном случае делался вывод о существовании одиночных точек сдвигов фазы триплетной периодичности.

5. Построение графиков $I_{11}(x_1, x_2)$, $I_{12}(x_1, x_2)$ и $I_{13}(x_1, x_2)$

Для иллюстрации найденных парных точек сдвига фазы мы также построили контурные графики для значений $D(1,k)$, k = 1,2,3 по формуле (6). Эти графики строились для двух подпоследовательностей S_1 и S_2, которые имеют длину равную $L_1 = 120$ нуклеотидам и которые начинаются в последовательности S в позициях x_1 и x_2. В этом случае последовательности S_1 и S_2 не следуют друг за другом, а выделяются в координатах $(x_1, x_1 + L_1 - 1)$ и $(x_2, x_2 + L_1 - 1)$. Затем рассчитываются матрицы $M_1(x_1, x_1 + L_1 - 1)$, $M_1(x_2, x_2 + L_1 - 1)$, $M_2(x_2, x_2 + L_1 - 1)$ и $M_3(x_2, x_2 + L_1 - 1)$ и затем по формуле (6) рассчитывалась величина $D(1,k)$. Координаты x_1 и x_2 меняются независимо друг от друга от 1 до $L - L_1 + 1$ с шагом в три основания. Это означает, что $x_1 = 1 + 3i$, $i = 0,1,2,3,\ldots$, а $x_2 = 1 + 3j$, $j = 0,1,2,3,\ldots$, где i и j – натуральные числа. В результате мы получаем три контурных графика. Первый график $I_{11}(x_1,x_2)$ показывает подобие матриц триплетной периодичности в рамке T_1 в разных позициях последовательности S. На этом графике районы со сдвигом фазы будут иметь низкое подобие с остальными последовательностями гена. Второй график $I_{12}(x_1,x_2)$ показывает подобие между матрицами триплетной периодичности последовательности S, рассчитанными по рамкам считывания T_1 и T_2. Третий график $I_{13}(x_1,x_2)$ показывает подобие между матрицами триплетной периодичности последовательности S, рассчитанных по рамкам считывания T_1 и T_3. Данные контурные графики позволяют увидеть районы сдвигов фазы триплетной периодичности в последовательности S.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В данной работе мы провели поиск парных сдвигов фазы триплетной периодичности в генах из бактериальных геномов. Мы изучали правые сдвиги типа 1+1, 1+2, 2+1 и 2+2. Первая и вторая цифра этой нумерации показывает на сколько оснований вправо сдвигается рамка считывания после $k_1^{\max}$ и после $k_2^{\max}$. Таким образом, после сдвигов типа 1+2 и 2+1 фаза триплетной периодичности (и рамка считывания) восстанавливается в гене, а после сдвигов 1+1 и 2+2 фаза триплетной периодичности сдвигается на два и одно основание вправо относительно фазы триплетной периодичности, существующей в гене до точек $k_1^{\max}$ и $k_2^{\max}$.

Для иллюстрации метода мы провели поиск парных сдвигов фазы у искусственной периодической последовательности. Для этого мы взяли периодическую последовательность $(atcgtaggt)_{134}$ и проанализировали ее разработанным алгоритмом. В этом случае не удалось выявить одиночные и парные сдвиги фазы триплетной периодичности. Затем в этой последовательности после 540 основания вставили *a*, которое сдвинуло фазу триплетной периодичности на одно основание, а после 721 основания вставили *aa,* что сдвинуло фазу триплетной периодичности еще на два основания и вернуло ее к первоначальному состоянию. Контурные графики $I_{11}(x_1,x_2)$ и $I_{12}(x_1,x_2)$ для этой последовательности показаны на рисунке 1. На рисунке 1А видно, что подобие между матрицами триплетной периодичности $M_1(x_1, x_1 + L_1 - 1)$ и $M_1(x_2, x_2 + L_1 - 1)$ отсутствует, если координата x_1 или x_2 попадает в район сдвига фазы триплетной периодичности. График для $I_{12}(x_1,x_2)$ на рисунке 1Б показывает, что подобие между матрицами триплетной периодичности $M_1(x_1, x_1 + L_1 - 1)$ и $M_2(x_2, x_2 + L_1 - 1)$ наблюдается, только если одна из координат x_1 или x_2 попадает в район сдвига фазы триплетной периодичности. График $I_{13}(x_1,x_2)$ аналогичен графику $I_{12}(x_1,x_2)$ и получается из него симметричным отображением относительно главной диагонали, поэтому мы его не показываем.

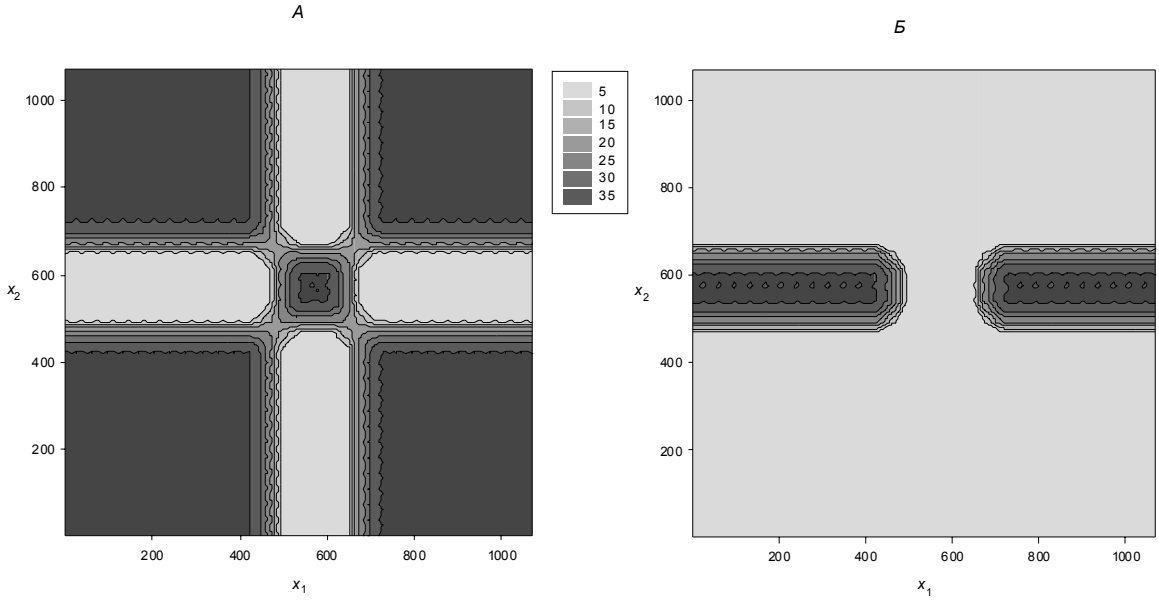

Рис. 1. А. График $I_{11}(x_1,x_2)$ для последовательности $(atcgtaggt)_{60}a(atcgtaggt)_{20}aa(atcgtaggt)_{53}$, где после 540 и 721 основания присуствуют сдвиги фазы триплетной периодичности на одно и два основания ДНК. **Б.** График для $I_{12}(x_1,x_2)$ для той же последовательности. Район между сдвигами фазы триплетной периодичности выделяется темным цветом.

Затем мы взяли последовательность гена, кодирующего chitosanase из генома *B. subtilis* (BSU26890 из банка данных Kegg). В ней также не было обнаружено парных сдвигов фазы триплетной периодичности. Затем после 600 основания ДНК вставили *a,* а после 781 основания вставили *aa*. Для этой последовательности был построен контурный график $I_{11}(x_1,x_2)$ до и после произведенных сдвигов фазы триплетной периодичности (рис. 2А и 2Б), а также график $I_{12}(x_1,x_2)$ (рис. 2В). Район между двумя координатами $k_1^{\max}$ и $k_2^{\max}$ виден на рисунке 2В как горизонтальная темная полоса, так же как и на рисунке 1Б. Эти примеры показывают, что разработанный математический подход позволяет обнаруживать сдвиги фазы в триплетной периодичности гена.

Затем мы проанализировали 17 бактериальных геномов, которые содержат 66936 генов. Список бактерий, гены которых были проанализированы в настоящей работе, показан в таблице 1. Всего на значимом статистическом уровне нами было найдено 2119 генов, которые содержат одиночные и парные сдвиги фазы триплетной периодичности генов, что составляет около 3% от изученных нами генов. Эти данные согласуются с полученными нами ранее результатами [17]. Мы сравнили список одиночных сдвигов фазы триплетной периодичности, которые мы нашли ранее, со списком генов с одиночными и парными сдвигами фазы триплетной периодичности, обнаруженными в настоящей работе. Перекрытие в этих данных составляет около 50%. Такое различие может быть вызвано следующими причинами. Ранее для идентификации сдвигов фазы триплетной периодичности была использована мера различия между матрицами триплетной периодичности. Это могло привести к тому, что часть сдвигов фазы триплетной периодичности была идентифицирована как точка разладки [32]. Также часть одиночных сдвигов фазы триплетной периодичности могла быть пропущена. В данном исследовании мы применили меру подобия матриц триплетной периодичности, и это, как нам кажется, позволило получить более адекватные результаты по поиску одиночных и парных сдвигов фазы. Применение меры подобия матриц триплетной периодичности позволило более точно локализовать координату сдвига в последовательности гена, чем это было возможно ранее [17].

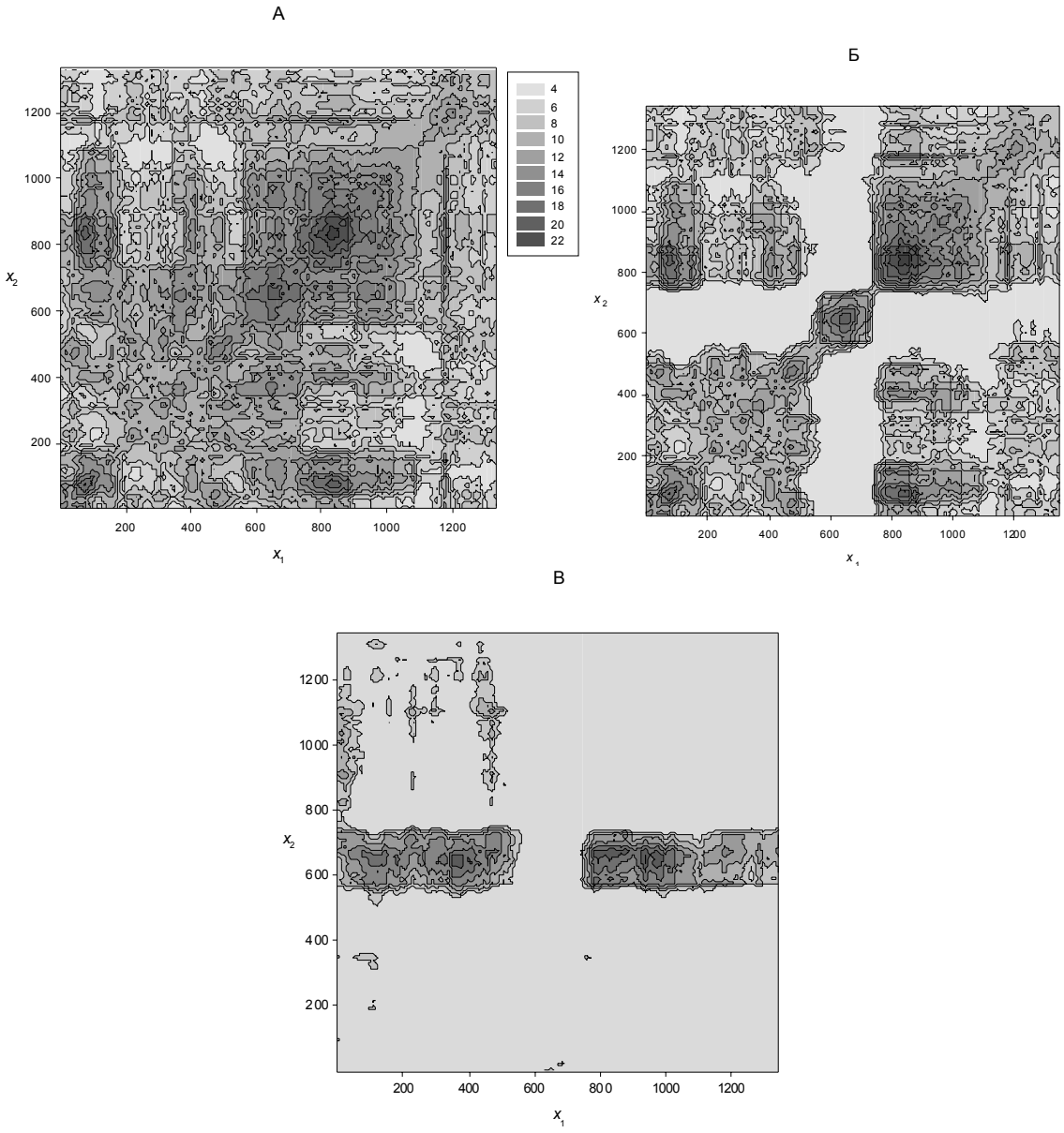

Рис. 2. А. График $I_{11}(x_1,x_2)$ для последовательности гена, кодирующего chitosanase из генома *B. subtilis* (BSU00090 из банка данных Kegg). Видно, что триплетная периодичность равномерно представлена по последовательности гена. **Б.** График $I_{11}(x_1,x_2)$ для последовательности гена, кодирующего chitosanase из генома *B. subtilis* (BSU00090 из банка данных Kegg), где сделана вставка основания a после 600 нуклеотида и вставлены основания aa после 781 нуклеотида. **В.** График для $I_{12}(x_1,x_2)$ для последовательности гена, кодирующего chitosanase из генома *B. subtilis* (BSU00090 из банка данных Kegg), где сделаны вставки. Район между сдвигами фазы триплетной периодичности выделяется темным цветом.

Таблица 1. Показаны бактерии, геномы которых были проанализированы. В последнем столбце показано число генов с двойными сдвигами фазы триплетной периодичности в каждом изученном геноме

N	Бактерия	Число генов с единичным и парным сдвигом фазы триплетной периодичности сдвигом фазы
1	*A. butzleri*	32
2	*A. vinelandii*	111
3	*B. avium*	88
4	*B. mallei*	204
5	*B. subtilis*	102
6	*E. coli*	149
7	*L. fermentum*	56
8	*M. capsulatus*	117
9	*P. aeruginosa*	83
10	*S. aureus col*	70
11	*S. enterica choleraesuis*	202
12	*S. pneumoniae*	74
13	*S. sonnei*	221
14	*S. typhimurium*	150
15	*V. cholerae*	164
16	*X. campestris*	163
17	*Y. pseudotuberculosis ypiii*	133

Доля парных сдвигов фазы составляет, как это видно из таблицы 2, около 1% от числа изученных генов. Примеры парных сдвигов фазы триплетной периодичности показаны на рисунках 3 и 4. На этих рисунках видно, что в последовательностях генов присутствуют районы ДНК, где триплетная периодичность имеет сдвиг фазы. Число парных сдвигов превышает то число, которое можно было бы ожидать на основе того, что сдвиги фазы триплетной периодичности образуются случайно. Если учесть, что одиночные сдвиги были выявлены в данной работе примерно в 2% генов, то можно было бы ожидать присутствие примерно 0.04% генов, которые имели бы парные сдвиги фаз, если бы сдвиги фазы возникали бы случайно. Число парных сдвигов превышает примерно в 25 раз это число, рассчитанное по модели, где они происходят случайным образом. Такое отличие может свидетельствовать о существовании ферментов, которые пытаются исправить сдвиг рамки считывания в генах [33].

Таблица 2. Показано число одиночных и парных сдвигов фазы триплетной периодичности в генах из геномов 17 бактерий

Число одиночных сдвигов фазы триплетной периодичности	Число двойных сдвигов фазы триплетной периодичности типа 1+1	Число двойных сдвигов фазы триплетной периодичности типа 1+2	Число двойных сдвигов фазы триплетной периодичности типа 2+1	Число двойных сдвигов фазы триплетной периодичности типа 2+2
1374	185	239	142	179

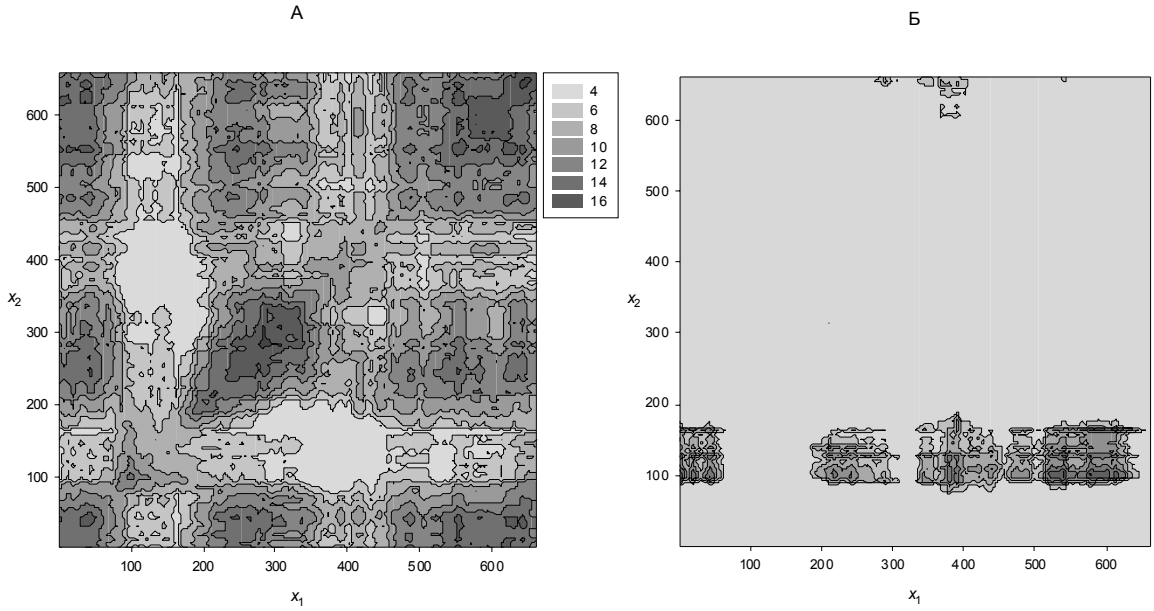

Рис. 3. А. График $I_{11}(x_1,x_2)$ для последовательности гена, кодирующего enoyl-CoA hydratase из генома *B. avium* (BAV1386 из банка данных Kegg). Из рисунка видно, что данный ген содержит сдвиг фазы триплетной периодичности на одно основание ДНК после ~100 основания и сдвиг фазы на 2 основания после ~200 основания. **Б.** График для $I_{12}(x_1,x_2)$ для последовательности гена enoyl-CoA hydratase из генома *B. avium*. Район между сдвигами фазы триплетной периодичности выделяется темным цветом.

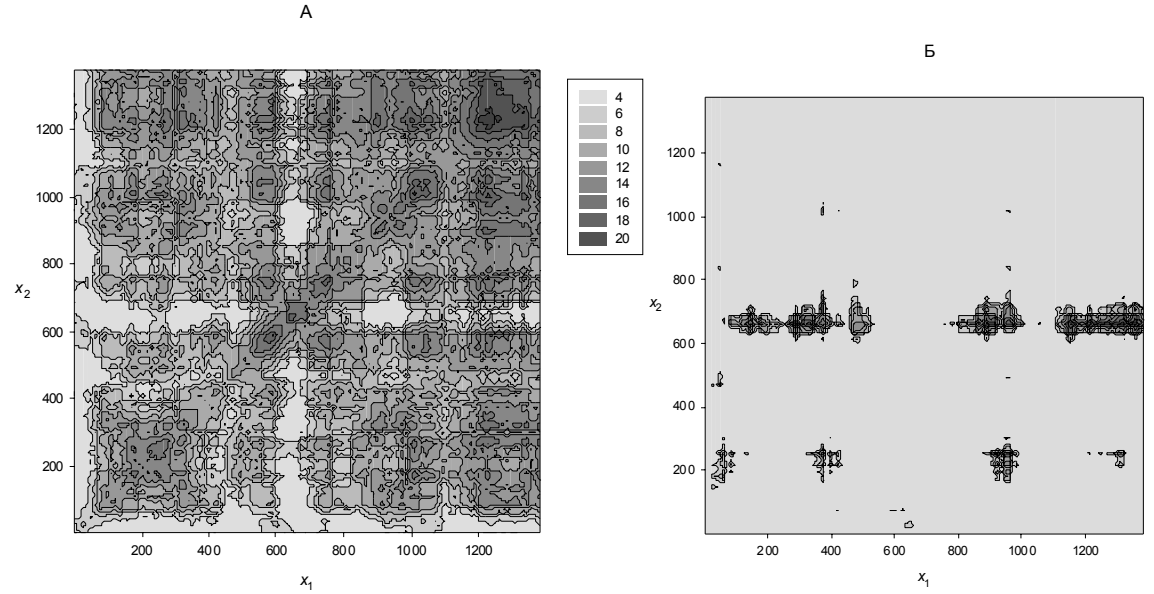

Рис. 4. А. График $I_{11}(x_1,x_2)$ для последовательности для гена, кодирующего белок семейства citrate synthase из генома *B .mallei* (ген BMA2258 из Kegg базы данных). Из рисунка видно, что данный ген содержит сдвиг фазы триплетной периодичности на одно основание ДНК после ~600 основания и сдвиг фазы на 2 основания после ~700 основания. **Б.** График для $I_{12}(x_1,x_2)$ для последовательности гена, кодирующего белок семейства citrate synthase из генома *B. mallei*. Район между сдвигами фазы триплетной периодичности выделяется темным цветом.

Мы также изучили подобие аминокислотных последовательностей, которые были созданы для генов, которые имеют одиночный сдвиг фазы триплетной периодичности для трех рамок считывания. Среди найденных подобий мы выделяли только те, которые заканчиваются или начинаются в точке сдвига фазы (точка А). Из таблицы (табл. 3) видно, что примерно 15% генов, где имеется сдвиг фазы триплетной периодичности, имеют подобие только с аминокислотными последовательностями, созданными для фрагмента S_1 и S_2 по альтернативным рамкам считывания. Причем это подобие заканчивается (для S_1) или начинается (для S_2) в точке сдвига фазы (точка x на

рисунке 5А. То же самое было проделано с аминокислотными последовательностями, которые были созданы по альтернативным рамкам считывания для фрагментов S_1, S_2 и S_3 в случае парных сдвигов. Эти данные показывают, что существуют аминокислотные последовательности, в которых отсутствует парный сдвиг рамки считывания, и эти аминокислотные последовательности имеют функциональное значение (табл. 4).

Таблица 3. Показано количество выравниваний, найденных с помощью программы BLAST, для участков S_1 и S_2 в нуклеотидных последовательностях генов со сдвигами фазы триплетной периодичности последовательности, перекодированными в трёх рамках считывания. Выделение участков S_1 и S_2 показано на рисунке (Рис. 5А)

	Рамка считывания		
	1	2	3
S_1	442	29	16
S_2	535	81	135

Таблица 4. Показано количество выравниваний, найденных с помощью программы BLAST, для участков S_1, S_2 и S_3 в нуклеотидных последовательностях генов со сдвигами фазы триплетной периодичности последовательности, перекодированными в трёх рамках считывания. Выделение участков S_1, S_2 и S_3 показано на рисунке (Рис. 5Б)

	Рамка считывания		
	1	2	3
S_1	168	5	3
S_2	68	9	8
S_3	138	9	12

С помощью BLAST мы провели поиск выравниваний для аминокислотных последовательностей генов, для которых найдены сдвиги фазы триплетной периодичности. На рисунке (рис. 6) показан пример выравниваний для участков S_2 и S_3 гена кодирующего белок внутренней мембраны III типа секреции SctQ (BMAA1629 в банке данных Kegg). В данном гене был найден парный сдвиг: первый сдвиг на одно основание, второй сдвиг на два основания. После первого сдвига около координаты $x = 129$ произошел переход с первой рамки считывания на вторую рамку. После второго сдвига около координаты $y = 201$ – со второй обратно на первую рамку считывания. Найдены выравнивания для второго и третьего участков.

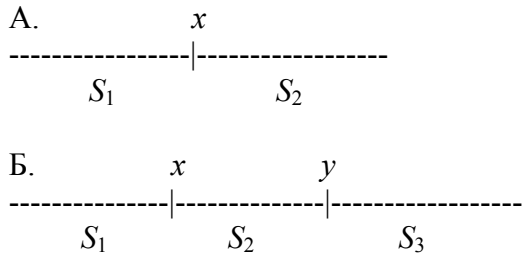

Рис. 5. А. На схеме показано выделение двух участков S_1 и S_2 в случае если ген имеет один сдвиг фазы. Первый участок находится в гене до точки сдвига x, а второй после точки сдвига фазы триплетной периодичности x. **Б.** На схеме показано выделение трех участков S_1, S_2 и S_3 в случае если ген имеет два сдвига фазы триплетной периодичности x и y.

```
А.
Query: 137 LPLDAALLPDTVLHLVLSRFDLRLRFETPDAATRHLLCTHGATLHARLDTLL 188
           L LD  +LP+T L L LS   L LRFE     +RHLL  HG TL  R+  LL
Sbjct: 135 LDLDPKILPETRLTLRLSPHTLSLRFEAGHPRSRHLLSEHGDTLRQRIHALL 186

Б.
Query: 569 LGEVALPVHVEIDTLSLSIAELAALRPGYVLELPLAARDVPVRLVAYGQAIGGGRLVAVG 628
           L ++ + V  E+    L    L +L PG +++L     D  VRL+A G+ +G GRLV +
Sbjct: 233 LNQLPVQVSFEVGRQILDWHTLTSLEPGSLIDLTTPV-DGEVRLLANGRLLGHGRLVEIQ 291

Query: 629 AHLGVRIDRMA 639
             LGVRI+R+
Sbjct: 292 GRLGVRIERLT 302
```

Рис. 6. А. Выравнивание участка S_2 (137-188 аминокислота) с учетом сдвига рамки считывания на одно основание вправо и последовательности sp|P35651|HPAP_RALSO HRP-associated protein OS = *Ralstonia solanacearum*. E value = 0.005. **Б.** Выравнивание участка S_3 (569-639 аминокислота) с учетом сдвига рамки считывания на два основания вправо и последовательности sp|P40296|YSCQ_YERPS Yop proteins translocation protein Q OS=*Yersinia pseudotuberculosis*. E value = 0.004.

Интересен также вопрос о биологическом значении найденных парных сдвигов рамки считывания. В бактериальных генах, в отличие от клеток эукариот, нет интронов, и парные сдвиги рамки считывания могут представлять собой способ изменения последовательности внутри гена. Если ген не разбит на интроны, то трудно внести существенное изменение в середине белка. Однако парные сдвиги рамки считывания могут с успехом выполнять эту функцию, что может обеспечивать эволюционную приспособляемость бактерий к изменяющейся внешней среде.

Данные этой работы согласуются с данными работы [17] о числе генов со сдвигом рамки считывания. Ранее было показано, что около 4% генов имеют, по крайней мере, один сдвиг рамки считывания, тогда как в данной работе это число чуть более 3%. Небольшая разница возникает из-за использования различных мер при сравнении матриц триплетных периодичностей, а также из-за того, что ранее анализировались только гены длиннее 1200 оснований, а в данной работе анализируются гены длиннее 60 нуклеотидов. Вполне резонно считать, что в более длинных генах вероятность появления мутации типа сдвиг рамки считывания будет больше. Значительно большее расхождение получено относительно числа двойных сдвигов по сравнению с работой [34]. Однако в этой работе была использована математическая мера, которая позволила часть вставок фрагментов ДНК в гены рассматривать как двойные сдвиги фазы. Именно этим объясняется расхождение в числе двойных сдвигов фазы примерно в 4–5 раз.

В данной работе мы нашли только те парные сдвиги фазы, расстояние между которыми имеет сравнительно протяженную длину (более 60 оснований ДНК). Также триплетная периодичность должна быть достаточно выражена, чтобы заметить в гене сдвиги фазы триплетной периодичности. Сейчас трудно сказать, изменилась ли функция белка после таких событий и привели ли такие события к созданию новых генов и новых биологических функций у кодируемых белков. На этот интересный вопрос можно будет ответить после проведения экспериментальных работ.

СПИСОК ЛИТЕРАТУРЫ

1. *DNA Repair, Genetic Instability, and Cancer.* Eds. Wei Q., Li L., Chen D.J. World Scientific, 2007.
2. Watson J.D., Levine M., Baker T.A., Gann A., Bell S.P. *Molecular Biology of the Gene.* Benjamin-Cummings Pub. Corp., 2007.

3. Okamura K., Feuk L., Marquès-Bonet T., Navarro A., Scherer S.W. Frequent appearance of novel protein-coding sequences by frameshift translation. *Genomics*. 2006. V. 88. P. 690–697.
4. Raes J., Van de Peer Y. Functional divergence of proteins through frameshift mutations. *Trends Genet*. 2005. V. 21. P. 428–431.
5. Kramer E.M., Su H.-J., Wu C.-C., Hu J.-M. A simplified explanation for the frameshift mutation that created a novel C-terminal motif in the APETALA3 gene lineageBMC. *Evolutionary Biology*. 2006. V. 6 P. 30.
6. States D.J., Botstein D. Molecular sequence accuracy and the analysis of protein coding regions. *Proc. Natl. Acad. Sci., USA*. 1991. V. 88. P. 5518–5522.
7. Pearson W.R., Wood T., Zhang Z., Miller W. Comparison of DNA sequences with protein sequences. *Genomics*. 1997. V. 46. P. 24–36.
8. Birney E., Thompson J., Gibson T. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucl. Acids Res*. 1996. V. 24. P 2730–2739.
9. Guan X., Uberbacher E.C. Alignments of DNA and protein sequences containing frameshift errors. *Comput. Appl. Biosci*. 1996. V. 12. P. 31–40.
10. Antonov I., Borodovsky M. Genetack: frameshift identification in protein-coding sequences by the Viterbi algorithm. *J. Bioinform Comput Biol*. 2010 Jun. V. 8. № 3. P. 535–51.
11. Kislyuk A., Lomsadze A., Lapidus A.L., Borodovsky M. Frameshift detection in prokaryotic genomic sequences. *Int J Bioinform Res Appl*. 2009. V. 5. № 4. P. 458–477.
12. Fichant G.A., Quentin Y. A frameshift error detection algorithm for DNA sequencing projects. *Nucleic Acids Res*. 1995. V. 23. |P. 2900–2908.
13. Médigue C., Rose M., Viari A., Danchin A. Detecting and analyzing DNA sequencing errors: toward a higher quality of the Bacillus subtilis genome sequence. *Genome Res*. 1999. V. 9. P. 1116–1127.
14. Schiex T., Gouzy J., Moisan A., Oliveira Y.D. FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res*. 2003. V. 31. P. 3738–3741.
15. Frenkel F.E., Korotkov E.V. Classification analysis of triplet periodicity in protein-coding regions of genes. *Gene*. 2008. V. 421. P. 52–60.
16. Frenkel F.E., Korotkov E.V. Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. *DNA Res*. 2009. V. 16. P. 105–114.
17. Korotkov E.V., Korotkova M.A. Study of the triplet periodicity phase shifts in genes. *Journal of Integrative Bioinformatics*. 2010. V. 7. P.131–141.
18. Fickett J.W. Predictive methods using nucleotide sequences. *Methods Biochem Anal*. 1998. V. 39. P. 231–245.
19. Staden R. Statistical and structural analysis of nucleotide sequences. *Methods Mol Biol*. 1994. V. 25. P. 69–77.
20. Baxevanis A.D. Predictive methods using DNA sequences. *Methods Biochem Anal*. 2001. V. 43. P.233–252.
21. Gutiérrez G., Oliver J.L., Marín A. On the origin of the periodicity of three in protein coding DNA sequences. *J Theor Biol*. 1994 Apr 21. V. 167. № 4. P. 413–414.
22. Gao J., Qi Y., Cao Y., Tung W.-W. Protein Coding Sequence Identification by Simultaneously Characterizing the Periodic and Random Features of DNA Sequences. *Journal of Biomedicine and Biotechnology*. 2005. V. 2. P. 139–146.
23. Yin C., Yau S.S. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *Journal of Theoretical Biology*. 2007. V. 247. P. 687–694.
24. Eskesen S.T., Eskesen F.N., Kinghorn B., Ruvinsky A. Periodicity of DNA in exons. *BMC Molecular Biology*. 2004. V. 5. P. 12.

25. Bibb M.J., Findlay P.R., Johnson M.W. The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene*. 1984 Oct. V. 30. № 1-3. P. 157–166.
26. Konopka A.K. *S*equences and codes: fundamentals of biomolecular cryptology. In: *Biocomputing: Informatics and genome projects*. Ed. Smith D. San Diego, CA: Academic Press, 1994. P. 119–174.
27. Trifonov E.N. Elucidating sequence codes: three codes for evolution. *Ann NY Acad Sci*. 1999. V. 870. P. 330–338.
28. Eigen M., Winkler-Oswatitsch R. Transfer-RNA: the early adaptor. *Naturwissenschaften*. 1981. V. 68. P. 217–228.
29. Zoltowski M. Is DNA Code Periodicity Only Due to CUF - Codons Usage Frequency? In: *Conf Proc IEEE Eng Med Biol Soc*. 2007. V. 1. P. 1383–1386.
30. Antezana M.A., Kreitman M. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. *J Mol Evol*. 1999. V. 49. № 1. P. 36–43.
31. Korotkov E.V., Korotkova M.A., Frenkel F.E., Kudryashov N.A. The Informational Concept of Searching for Periodicity in Symbol Sequences. *Molecular Biology*. 2003. V. 37. P. 372–386.
32. Suvorova Y.M., Rudenko V.M., Korotkov E.V. Detection change points of triplet periodicity of gene. *Gene*. 2012. V. 491. P. 58–64.
33. Strauss B.S. Frameshift mutation, microsatellites and mismatch repair. *Mutation Research*. 1999. V. 437. P. 195–203.
34. Korotkova M.A., Kudryashov N.A., Korotkov E.V. An approach for searching insertions in bacterial genes leading to the phase shift of triplet periodicity. *Genomics Proteomics Bioinformatics*. 2011. V. 9. P. 158–170.

Материал поступил в редакцию 11.05.2012, опубликован 07.08.2012.