Математическая биология и биоинформатика. 2009. Т. 4. № 2. С. 66-80.
URL: http://www.matbio.org/downloads/Korotkov2009(4_66).pdf

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

УДК: 577.212.2; 577.214

Сдвиг фазы триплетной периодичности в нуклеотидных последовательностях генов

Коротков Е.В., Руденко В.М.*

Центр «Биоинженерия» Российской академии наук, Москва, 117312, Россия

Аннотация: Триплетная периодичность является общеизвестным свойством кодирующих последовательностей оснований ДНК. Однако недавно проведенные исследования показали, что около 20% (122829) последовательностей генов из банка данных KEGG (версии 29) не имеют триплетной периодичности без делеций/вставок на статистически значимом уровне. В данной работе ставилась задача показать, что отсутствие триплетной периодичности в некоторых последовательностях может быть обусловлено сдвигами открытой рамки считывания. Для выявления сдвигов рамки считывания предложен новый математический метод, основанный на вычислении меры сходства между типами триплетной периодичности фрагментов последовательности до и после позиции предполагаемого сдвига рамки считывания. С помощью разработанного подхода было найдено 4724 последовательности, имеющие сдвиг рамки считывания. Мы предположили, что в этих случаях мутации типа вставок или делеций привели к образованию новой рамки считывания и нарушению триплетной периодичности последовательности. Выявленные последовательности были перекодированы в аминокислотные последовательности по существующей рамке считывания и древней рамке считывания с учетом сдвига. Оказалось, что 243 последовательности, построенные по древней рамке, имеют подобие к аминокислотным последовательностям банка данных Swiss-Prot, что подтверждает наше предположение о возможности эволюции генов посредством сдвигов рамки считывания.

Ключевые слова: *последовательность ДНК, триплетная периодичность, открытая рамка считывания, сдвиг открытой рамки считывания.*

1. ВВЕДЕНИЕ

Триплетная организация последовательностей ДНК, кодирующих белки, является общим свойством всех известных в настоящее время живых систем [1-9] и она привязана к рамке считывания, существующей в гене [10]. Причина этого заключается как в структуре генетического кода, который практически одинаков у представителей прокариот и эукариот, так и в насыщенности белков определенными аминокислотами [11-14]. Для выявления триплетной периодичности в настоящее время разработаны методы, использующие регулярность в предпочтении символов в различных позициях триплета в последовательностях ДНК. В качестве математического аппарата в них использовались преобразование Фурье, скрытые цепи Маркова и другие статистические методы, основанные на позиционно-зависимых предпочтениях нуклеотидов в кодирующих последовательностях [15-20]. Позже для поиска

*v.m.rudenko@gmail.com

триплетной периодичности был предложен метод информационного разложения [21, 22], который позволяет обнаруживать триплетную периодичность нуклеотидной последовательности без учета вставок и делеций нуклеотидов и позволяет ввести понятие класса триплетной периодичности в виде матрицы М размером 4×3. В матрице признаками столбцов являются позиции периода, а признаками строк являются нуклеотиды. Этот подход был использован для поиска районов с триплетной периодичность в генах из базы данных KEGG (версии 29) [23] и классификации типов триплетной периодичности в виде матриц [10]. Проведенное исследование позволило обнаружить триплетную периодичность у примерно 80% известных генов из базы данных KEGG-29. Относительно оставшихся примерно 20% генов, где триплетная периодичность была на статистически незначимом уровне, можно выдвинуть две гипотезы. Во-первых, можно предполагать, что эти гены могут содержать вставки или делеции коротких фрагментов ДНК с длиной не кратной трем основаниям, что приводит к циклическому сдвигу триплетной периодичности ниже района вставки и к образованию сдвига рамки считывания в гене. Такой циклический сдвиг может значительно уменьшить статистическую значимость триплетной периодичности [24]. Во-вторых, в качестве альтернативной гипотезы можно рассматривать возможность, что в этих 20% генах триплетная периодичность отсутствует вовсе, и ее невозможно выявить на неслучайном уровне, даже если мы будем учитывать возможность существования вставок и делеций нуклеотидов. В случае справедливости первой гипотезы эти 20% генов могут содержать сдвиги рамки считывания [24]. Выбор между этими двумя гипотезами важен для понимания того, насколько часто в генах могут происходить сдвиги рамки считывания, связанные с делециями и вставками нуклеотидов. Из литературных данных известно всего нескольких сотен [25–27] генов, где сдвиги рамки считывания были идентифицированы. Если даже часть из этих 20% генов содержит сдвиги рамки считывания, образовавшиеся вследствие вставок или же делеций коротких фрагментов нуклеотидов с длиной некратной трем основаниям, то число генов со сдвигами рамки считывания может стать значительно большим, чем несколько сотен.

Для того чтобы выделять триплетную периодичность на фоне вставок и делеций нуклеотидов в генах, необходимо разработать математический метод для их поиска. В качестве сигнала о существовании делеции или вставки и связанного с этим событием сдвига рамки считывания в нуклеотидной последовательности гена, как мы показываем в данной работе, может выступать сдвиг фазы триплетной периодичности (рис.1). Поскольку триплетную периодичность последовательности ДНК трудно существенно изменить посредством сравнительно небольшого числа замен оснований [22], то такой сдвиг может сохраняться сравнительно долго.

В настоящей работе ставились две задачи. Во-первых, мы хотели найти все гены из базы данных KEGG-29, где ранее не была выявлена триплетная периодичность и в которых существует сдвиг фазы триплетной периодичности. Для этой цели в данной работе разработан математический подход по выявлению сдвига фазы триплетной периодичности в нуклеотидной последовательности. Во-вторых, мы хотели проверить, действительно ли гипотетические аминокислотные последовательности, полученные при использовании рамки считывания триплетной периодичности, имеют гомологию с последовательностями из банка данных Swiss-Prot. Такую проверку мы делали для тех районов, у которых наблюдался сдвиг по фазе между триплетной периодичностью и рамкой считывания. Мы подтвердили существование таких сдвигов для части генов, так как нашли гомологию между некоторыми гипотетическими аминокислотными последовательностями и аминокислотными последовательностями из банка данных Swiss-Prot.

2.МЕТОДЫ ИССЛЕДОВАНИЯ

2.1. Определение фазы триплетной периодичности

Будем считать, что задана кодирующая нуклеотидная последовательность $S = \{s(k), k = 1, 2, ..., L\}$, где каждое значения $s(k)$ выбирается из алфавита $A = \{a, t, c, g\}$, L есть длина последовательности S, кратная трем. Введем три рамки считывания в последовательности S и обозначим их как T_1, T_2 и T_3 (рис. 1). 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, M_2 и M_3. 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 встречаются в j позиции кодона (j может быть равно 1, 2 или 3) для рамок считывания T_1, T_2 и T_3 соответственно [22, 23].

```
123123123123123123123123123123123 – рамка считывания T1
312312312312312312312312312312312 – рамка считывания T2
231231231231231231231231231231231 – рамка считывания T3
atgatgatgatgatgatgCatgatgatgatgatg – последовательность S
```

M1(1,18)

	1	2	3
A	6	0	0
T	0	6	0
C	0	0	0
G	0	0	6

M1(19,37)

	1	2	3
A	0	6	0
T	0	0	6
C	1	0	0
G	6	0	0

M2(19,37)

	1	2	3
A	6	0	0
T	0	6	0
C	0	0	1
G	0	0	6

M3(19,37)

	1	2	3
A	0	0	6
T	6	0	0
C	0	1	0
G	6	6	0

Рис. 1. Влияние вставки одного основания на сдвиг фазы триплетной периодичности. Первые три последовательности показывают рамки считывания T_1, T_2 и T_3. После этого показана кодирующая последовательность S, имеющая триплетную периодичность. В этой последовательности произведена вставка нуклеотида **C** в 19 позицию. Явная периодичность этой последовательности выбрана для наглядности. В случае более «размытой» периодичности ситуация будет такой же как на этом рисунке, только зрительно периодичность трудно будет заметить. Затем мы строим матрицы триплетной периодичности $M_1(1, 18)$, $M_1(19, 37)$, $M_2(19, 37)$ и $M_3(19, 37)$. Первая матрица M_1 строится для района ДНК с 1 по 18-ое основание. Элементы этих матриц $m_1(i, j)$, $m_2(i, j)$ и $m_3(i, j)$ показывают число оснований a, t, c и g (индекс i) напротив позиций в триплетах рамок считывания T_1, T_2 и T_3 (индекс j). Если сравнивать матрицу $M_1(1, 18)$ с матрицами $M_1(19, 37)$, $M_2(19, 37)$ и $M_3(19, 37)$, то можно заметить, что она более всего похожа на матрицу $M_2(19, 37)$. Это означает, что мера расхождения U для пары матриц $\{M_1(1, 18), M_2(19, 37)\}$ будет меньше U_0, а для пар матриц $\{M_1(1, 18), M_1(19, 37)\}$ и $\{M_1(1, 18), M_3(19, 37)\}$ она будет больше чем U_0 (пункт 2.1). Начальная фаза матриц M_1, M_2 и M_3 в последовательности S равна 1, 2 и 3, так как основания последовательности S с номерами k равными 1, 2 и 3 являются первыми основаниями триплета в рамках считывания T_1, T_2 и T_3. Поэтому в последовательности S после позиции x = 18 наблюдается сдвиг фазы триплетной периодичности на 1 основание (разница начальных фаз матриц M_2 и M_1).

За начальную фазу матриц M_1, M_2, M_3 возьмем координату k того основания, 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. Пусть также $M_1(i_1, i_2)$, $M_2(i_1, i_2)$ и $M_3(i_1, i_2)$ представляют собой матрицы триплетной

периодичности, 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 $S(i_1, i_2)$.

Далее определим условия, при которых можно считать, что в последовательности S после нуклеотида $s(x)$ существует сдвиг фазы триплетной периодичности. Для этого, во-первых, в последовательности S должна существовать триплетная периодичность. Условия существования триплетной периодичности и количественная мера для выявления триплетной периодичности в последовательности S или ее фрагмента показана ниже в пункте 2.2. Во-вторых, мы должны ввести количественную меру подобия матриц триплетной периодичности. Определим функцию U, и будем считать, что две матрицы триплетной периодичности подобны друг другу, если $U \le U_0$, где U_0 – пороговое значение. В противном случае будем считать их различными. Детальный вид функции U, а также вычисление U_0 рассматривается далее. Будем считать, что после нуклеотида $s(x)$ в последовательности S существует сдвиг фазы триплетной периодичности на 1 основание, если одновременно выполняются условия:

$$\left\{\begin{array}{l} U\{M_1(1,\ x),\ M_2(x+1,\ L)\} \le U_0 \\ U\{M_1(1,\ x),\ M_1(x+1,\ L)\} > U_0 \\ U\{M_1(1,\ x),\ M_3(x+1,\ L)\} > U_0 \end{array}\right\}. \qquad (1)$$

Будем также считать, что после нуклеотида $s(x)$ в последовательности S существует сдвиг фазы триплетной периодичности на 2 основания, если одновременно выполняются условия:

$$\left\{\begin{array}{l} U\{M_1(1,\ x),\ M_3(x+1,\ L)\} \le U_0 \\ U\{M_1(1,\ x),\ M_1(x+1,\ L)\} > U_0 \\ U\{M_1(1,\ x),\ M_2(x+1,\ L)\} > U_0 \end{array}\right\}. \qquad (2)$$

Если же выполняются условия:

$$\left\{\begin{array}{l} U\{M_1(1,\ x),\ M_1(x+1,\ L)\} \le U_0 \\ U\{M_1(1,\ x),\ M_2(x+1,\ L)\} > U_0 \\ U\{M_1(1,\ x),\ M_3(x+1,\ L)\} > U_0 \end{array}\right\}, \qquad (3)$$

то будем считать, что после нуклеотида $s(x)$ фаза триплетной периодичности остается без изменения, т.е. сдвиг фазы триплетной периодичности равен нулю. Сдвиг фазы триплетной периодичности на 1 и 2 основания соответствует вставке после нуклеотида $s(x)$ $1 + 3n$ и $2 + 3n$ (или делеции $2 + 3n$ и $1 + 3n$) оснований соответственно (n = 0, 1, 2, 3,…).

2.2. Триплетная периодичность последовательности *S*

Матрицы триплетной периодичности можно рассматривать как таблицы сопряженности признаков [28]. Далее рассмотрим матрицу M_1, для матриц M_2 и M_3 все выводы будут аналогичны. Признаками строк матрицы M_1 являются основания последовательности S, а признаками столбцов являются позиции оснований в кодонах рамки считывания T_1. Будем считать, что в нуклеотидной последовательности S существует триплетная периодичность, если уровень взаимной информации I между основаниями последовательности S и позициями оснований в кодонах будет больше некоторой величины I_0 [10]. Взаимная информация вычисляется по формуле [28]:

$$I = \sum_{i=1}^{4}\sum_{j=1}^{3} m_1(i,\ j)\ln m_1(i,\ j) - \sum_{i=1}^{4} x(i)\ln x(i) - \sum_{j=1}^{3} y(j)\ln y(j) + L_1 \ln L_1, \qquad (4)$$

где $x(i) = \sum_{j=1}^{3} m_1(i, j)$, $y(j) = \sum_{i=1}^{4} m_1(i, j)$.

Удвоенная взаимная информация $2I$ имеет распределение χ^2 с 6 степенями свободы, что позволяет оценить статистическую значимость найденной периодичности. На практике для оценки статистической значимости удобно пользоваться соотношением, связывающим χ^2 и стандартное нормальное распределение [29]:

$$X = \sqrt{4I} - \sqrt{2n-1} \qquad (5)$$

n – количество степеней свободы. Соответствие $2I$ распределению χ^2 с 6 степенями свободы, а величины X стандартному нормальному распределению, достигается в случае достаточно большого объема статистических данных, т.е. длины последовательности S. Чтобы определить минимальную длину последовательности S, для которой возможно использование функции χ^2 для описания распределения величины $2I$, мы протестировали соответствие $2I$ распределению χ^2 для различных длин последовательностей. Для этого с помощью датчика случайных чисел генерировались множества нуклеотидных последовательностей для каждой длины от 30 до 1000 нуклеотидов. Каждое такое множество содержало 10000 последовательностей. Далее для всех последовательностей из каждого множества рассчитывалась взаимная информация и строилась гистограмма распределения величины $2I$. Эта гистограмма сравнивалась с теоретическим распределением по критерию χ^2. Оказалось, что для последовательностей длиной более 60 п.н. распределение $2I$ соответствует $\chi^2(6)$ с вероятностью не менее 99%. Все последовательности с триплетной периодичностью, анализируемые в настоящем исследовании, были длиннее 60 нуклеотидных пар. Это позволило использовать в данной работе распределение χ^2 для статистических оценок попадания $2I$ в интервал от некоторого порогового значения $2I_0$ до ∞. Мы считали, что фрагмент последовательности $S(x_1, x_2)$ имеет триплетную периодичность, если значение X для него было больше нуля.

2.3. Алгоритм поиска сдвига фазы триплетной периодичности

Пусть x показывает координату основания $s(x)$ в последовательности S и пусть x выбирается как $L_1 + 3n$, где $n = 0, 1, 2, 3,\ldots(L - L_1)/3$, где L_1 кратно трем и находится в интервале от 60 до 600. Рассмотрим фрагмент последовательности $S(x - L_1 + 1, x)$ и для него построим матрицу триплетной периодичности $M_1(x - L_1 + 1, x)$ для рамки считывания T_1 последовательности S. Рассмотрим также фрагменты $S(x + 1, x + L_1)$, $S(x + 2, x + L_1 + 1)$ и $S(x + 3, x + L_1 + 2)$, и для этих фрагментов построим матрицы триплетной периодичности $M_1(x + 1, x + L_1)$, $M_2(x + 2, x + L_1 + 1)$ и $M_3(x + 3, x + L_1 + 2)$ для рамок считывания T_1, T_2 и T_3 последовательности S соответственно. Если сразу же за позицией x в последовательности S произойдет сдвиг рамки считывания на одно или два основания посредством вставки одного или двух нуклеотидов (или делеции или вставки большей длины), то матрица $M_1(x - L_1 + 1, x)$ будет больше похожа с точки зрения функции U на матрицу $M_2(x + 2, x + L_1 + 1)$ или $M_3(x + 3, x + L_1 + 2)$. Если же за позицией x нет вставок нуклеотидов, то матрица $M_1(x - L_1 + 1, x)$ будет больше всего подобна матрице $M_1(x + 1, x + L_1)$. В качестве функции U, которая позволяет сделать вывод о различии двух матриц триплетной периодичности M_k и M_l, была выбрана величина

$$I_{kl} = I_{kl}(1) + I_{kl}(2) + I_{kl}(3), \qquad (6)$$

где:

$$I_{kl}(j) = \sum_{i=1}^{4} m_l(i, j)\ln(m_l(i, j)) + \sum_{i=1}^{4} m_l(i, j)\ln(m_l(i, j)) -$$
$$-\sum_{i=1}^{4}(m_k(i, j) + m_l(i, j))\ln(m_k(i, j) + m_l(i, j)) + \quad (7)$$
$$+(y_k(j) + y_l(j))\ln(y_k(j) + y_l(j)) - y_k(j)\ln y_k(j) - y_l(j)\ln y_l(j).$$

Здесь $m_k(i, j)$ и $m_l(i, j)$ есть элементы двух сравниваемых матриц M_k и M_l, $y_l(j) = \sum_{i=1}^{4} m_l(i, j)$, $y_k(j) = \sum_{i=1}^{4} m_k(i, j)$. Значение $I_{kl}(j)$ показывает различие столбцов с номером j у двух матриц M_k и M_l. Величина $2I_{kl}(j)$ имеет распределение χ^2 с 3 степенями свободы при сравнении двух матриц, построенных для случайных последовательностей [28]. $2I_{kl}$ есть сумма трех χ^2 распределений с 3 степенями свободы. Поэтому закон распределения $2I_{kl}$ будет также χ^2. Поскольку сумма элементов во всех столбцах одинакова, итоговое количество степеней свободы распределения χ^2 будет равно 8. В дальнейшей работе мы использовали I_{11}, I_{12}, I_{13} для поиска сдвига рамки считывания. Если в последовательности S после x отсутствуют вставки или делеции оснований ДНК (с длиной делеции или вставки не кратной трем), то в этом случае $I_{11} < I_{12}$ и $I_{11} < I_{13}$, Если присутствует вставка фрагмента длиной $Q = 3i + 1$ или же делеция фрагмента длиной $Q = 3i + 2$, (i = 1, 2,…), то мы можем говорить о переходе после x от рамки считывания T_1 к рамке считывания T_2. В этом случае $I_{12} < I_{11}$ и $I_{12} < I_{13}$. Если присутствует вставка фрагмента длиной $Q = 3i + 2$ или же делеция фрагмента длиной $Q = 3i + 1$, (i = 1, 2,…), то мы можем говорить о переходе после позиции x от рамки считывания T_1 к рамке считывания T_3. В этом случае $I_{13} < I_{11}$ и $I_{13} < I_{12}$. Обозначим за $P_{11} = P(\chi^2(8) > 2I_{11})$, $P_{12} = P(\chi^2(8) > 2I_{12})$, $P_{13} = P(\chi^2(8) > 2I_{13})$ – вероятности того, что $\chi^2(8)$ превысит величины $2I_{11}$, $2I_{12}$, $2I_{13}$ соответственно.

Для поиска сдвигов фазы триплетной периодичности удобно использовать величины:

$$F_1 = -\log_{10}(P_{11} / P_{12}),$$
$$F_2 = -\log_{10}(P_{11} / P_{13}). \quad (8)$$

В случае присутствия сдвига фазы триплетной периодичности P_{11} будет уменьшаться, что будет приводить к росту величин F_1 или F_2. В тоже время либо P_{12}, либо P_{13} будет увеличиваться, что приведет к тому, что одна их величин F_1 или F_2 будет принимать максимальное значение. Если максимум достигается на величине F_1, то в позиции x присутствует вставка фрагмента длиной $Q = 3i + 1$ или же делеция фрагмента длиной $Q = 3i + 2$, (i = 1, 2,…). Если максимум достигается на величине F_2, то в позиции x присутствует вставка фрагмента длиной $Q = 3i + 2$ или же делеция фрагмента длиной $Q = 3i + 1$, (i = 1, 2,…).

Для каждой координаты x мы варьировали значение L_1. Варьирование проводилось с целью поиска такой длины L_1, которая обеспечила бы максимальное значение для величин F_1, и F_2. Такой поиск необходимо провести, поскольку триплетная периодичность может меняться по длине последовательности, и это изменение влияет на значения функций F_1 и F_2. Мы варьировали L_1 для каждого x в интервале от 60 до 600 оснований. Если же при этом мы доходили до начала или же конца последовательности S, то максимальное значение L_1 при варьировании равнялось минимуму из (600, x) в начале последовательности и минимуму из (600, $L - x$) в конце последовательности.

В результате исследования всех возможных сдвигов триплетной периодичности для последовательности S строился график зависимости максимальных величин F_1, и F_2 от

координаты x, каждая из которых была получена для некоторой длины L_1. По графику определялись координаты локальных максимумов. Мы считали, что в последовательности есть сдвиг фазы триплетной периодичности, если значение F_1 или F_2 в локальном максимуме больше некоторого порогового значения F_0. Пороговое значение F_0 определялось методом Монте-Карло (см. пункт 2.4).

Кроме того, для идентификации сдвига фазы триплетной периодичности в позиции x необходимо убедится, что в последовательностях $S(x - L_1 + 1,\ x)$, $S(x + 1,\ x + L_1)$, $S(x + 2,\ x + L_1 + 1)$ и $S(x + 3,\ x + L_1 + 2)$ существует триплетная периодичность, так как значения I_{11}, I_{12}, I_{13} показывают меру расхождения матриц триплетной периодичности. Эта мера будет тем меньше, чем выше подобие друг другу у матриц триплетной периодичности, что при отсутствии триплетной периодичности может обуславливаться чисто случайными факторами. Для исключения подобия чисто случайных матриц мы брали к рассмотрению только такие последовательности $S(x - L_1 + 1,\ x)$, $S(x + 1,\ x + L_1)$, $S(x + 2,\ x + L_1 + 1)$ и $S(x + 3,\ x + L_1 + 2)$ которые обладают достаточно выраженной триплетной периодичностью (с X больше нуля).

2.4. Определение порогового значения F_0 методом Монте-Карло

Для поиска порогового значения F_0 мы использовали те 20% последовательностей генов, собранные в банке данных KEGG версии 29, где не была найдена триплетная периодичность. Из этого множества были удалены последовательности РНК, после чего количество последовательностей составило 122829. Далее мы создали случайный банк данных путем перемешивания последовательности оснований ДНК исходного множества последовательностей. Это позволяет сохранить такое же распределение длин в случайных последовательностях, как в исходной выборке. Также в случайном множестве последовательностей была сохранена триплетная периодичность. Для этого мы разбивали последовательность S на три подпоследовательности. Первая из них (обозначим ее как C_1) была получена выбором из последовательности S оснований, которые стоят на номерах, равных $i = 1 + 3n$. Вторая последовательность C_2 получается посредством выбора оснований, стоящих на позициях $i = 2 + 3n$, а третья последовательность C_3 получена выбором оснований, стоящих на позициях с номерами $i = 3 + 3n$. При создании последовательностей C_1, C_2 и C_3 n меняется от 0 до $L/3 - 1$.

Каждая из полученных последовательностей C_1, C_2, C_3 перемешивалась случайным образом. Затем они объединялись в последовательность R, в которой на позициях $i = 1 + 3n$ стояли нуклеотиды из последовательности C_1, на позициях $i = 2 + 3n$ стояли нуклеотиды из последовательности C_2 и на позициях $i = 3 + 3n$ стояли нуклеотиды из последовательности C_3. Таким образом, длина последовательности R была равна L, а также в ней был сохранен такой же состав нуклеотидов, как и в последовательности S.

После создания банка случайных последовательностей, мы исследовали его на предмет наличия в последовательностях этого банка сдвигов триплетной периодичности. Было выбрано два уровня F_0, равные 3.0 и 4.5. Затем мы подсчитали число генов для последовательностей и для созданных на их основе случайных последовательностей, которые имеют хотя бы один локальный максимум (как это описано в пункте 2.3) для F_1 и F_2 выше F_0. Для уровня $F_0 = 3.0$ число найденных сдвигов фазы триплетной периодичности в случайных нуклеотидных последовательностях составляет ~25% от числа сдвигов, которые мы нашли для 122829 реальных последовательностей генов. Для уровня $F_0 = 4.5$ число найденных сдвигов фазы триплетной периодичности в случайных нуклеотидных последовательностях составляет ~5% от числа сдвигов, которые мы нашли для 122829 реальных последовательностей генов. Поэтому уровень $F_0 = 3.0$ может быть выбран как показывающий потенциальные сдвиги фазы триплетной периодичности, а уровень $F_0 = 4.5$ как показывающий реально существующие сдвиги фазы триплетной периодичности.

3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

3.1. Поиск генов со сдвигом фазы триплетной периодичности в базе данных KEGG

Для демонстрации работы предложенного метода определения сдвигов фаз триплетной периодичности нами был рассмотрен ген b0004 (ген thrC из генома *E.coli*). На рис. 2 показан вид функций F_1 и F_2 для этого гена. Видно, что значения F_1 и F_2 невелики для всех позиций. То есть сдвигов фаз триплетной периодичности не идентифицировано. Далее мы удалили из последовательности b0004 основание в позиции 600. На рис. 3 показаны графики функций F_1 и F_2 после проведенных изменений. Видно, что внесение делеции приводит к сдвигу фазы триплетной периодичности и к большому значению функции F_2 в позиции 601. Рис. 2 и рис. 3 показывают, что разработанный математический алгоритм может находить делеции и вставки оснований ДНК в реальных генах по сдвигу фазы триплетной периодичности.

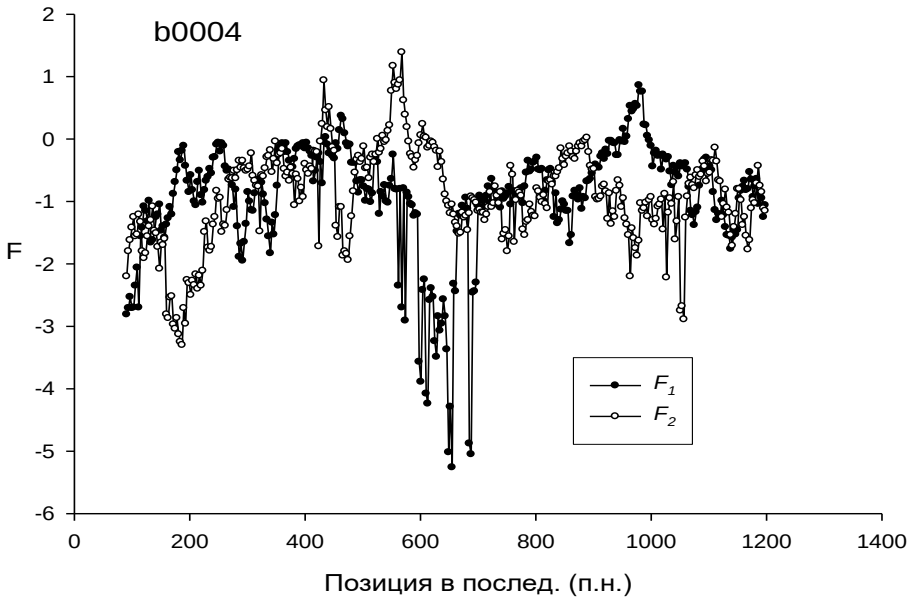

Рис. 2. Значения функций F_1 и F_2 для последовательности b0004 (ген thrC из генома E.coli) из банка данных KEGG. В этом гене отсутствует сдвиг фазы триплетной периодичности, и в этом случае F_1 и F_2 меньше 2.0 для всех позиций в гене.

Затем были проанализированы 122829 генов банка данных KEGG версии 29, в которых ранее триплетная периодичность не была обнаружена. Число генов с потенциальным сдвигом фазы триплетной периодичности (F_0 = 3.0) составило 4724 гена, а число генов с реально существующим сдвигом фазы составило 1382 гена. Примеры генов со сдвигом фазы триплетной периодичности показаны на рис. 4 и рис. 5. Первый пример относится к последовательности VV3117. Этот ген кодирует Protein affecting phage T7 exclusion by the F plasmid (Q7MGV9) из генома *Vibrio vulnificus*. Из рис. 4 видно, что функция F_1 имеет максимум около 380 нуклеотида. Этот максимум показывает, что в районе 380 нуклеотида был сдвиг триплетной периодичности на одно основание вправо от позиции 380. В соответствии с этим мы можем предполагать, что и рамка считывания тоже была сдвинута на один нуклеотид вправо и аминокислотная последовательность после точки сдвига была изменена. Если перекодировать ген по второй рамке считывания, то после 380 основания аминокислотная последовательность не имеет стоп-кодонов, что может косвенно указывать на то, что вторая рамка считывания могла выполнять кодирующую функцию. Второй пример, показанный на рис. 5. относится к гену zfp36, который кодирует белок, индуцируемый фактором роста. Из графика видно, что в этом гене можно выделить два сдвига фазы триплетной

периодичности. Первый сдвиг произошел в районе 300 нуклеотида и он соответствует сдвигу рамки считывания на одно основание вправо. На графике это видно из-за локального максимума функции F_1. Затем в районе 560 основания происходит сдвиг рамки считывания еще на два основания вправо, т. е. после 560 основания рамка считывания вернулась к своему первоначальному местоположению. В районе 560 основания функция F_2 достигает локального максимума со значением, большим 6.0. Если перекодировать последовательность гена zfp36 по второй рамке считывания, то в районе с 300 по 560 основание не встречается стоп-кодонов, что косвенно свидетельствует о том, что с этой рамки считывания могло происходить образование аминокислотной последовательности.

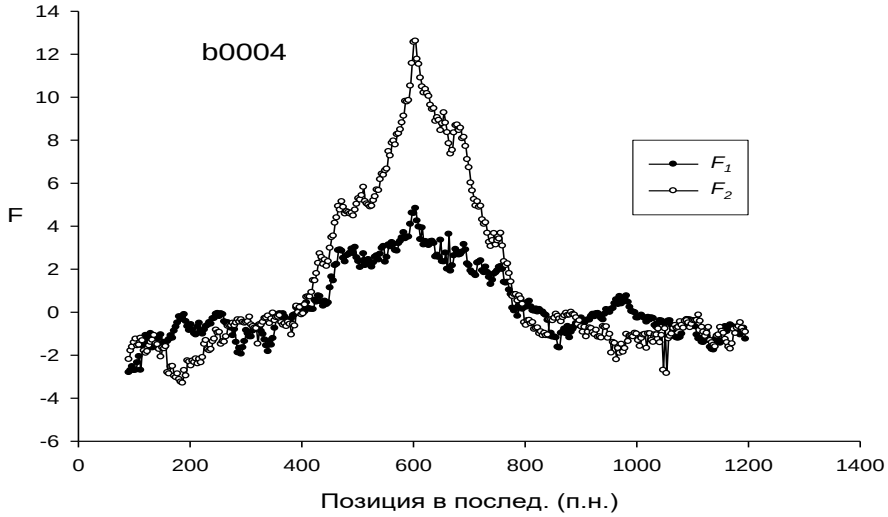

Рис. 3. Значения функций F_1 и F_2 для последовательности b0004 (ген thrC из генома *E.coli*) из банка данных KEGG в которой была сделана искусственная делеция основания в позиции 600. В этом случае присутствует сдвиг фазы триплетной периодичности и его можно заметить по значению функции F_2 в позиции 601, которое больше, чем 12.0.

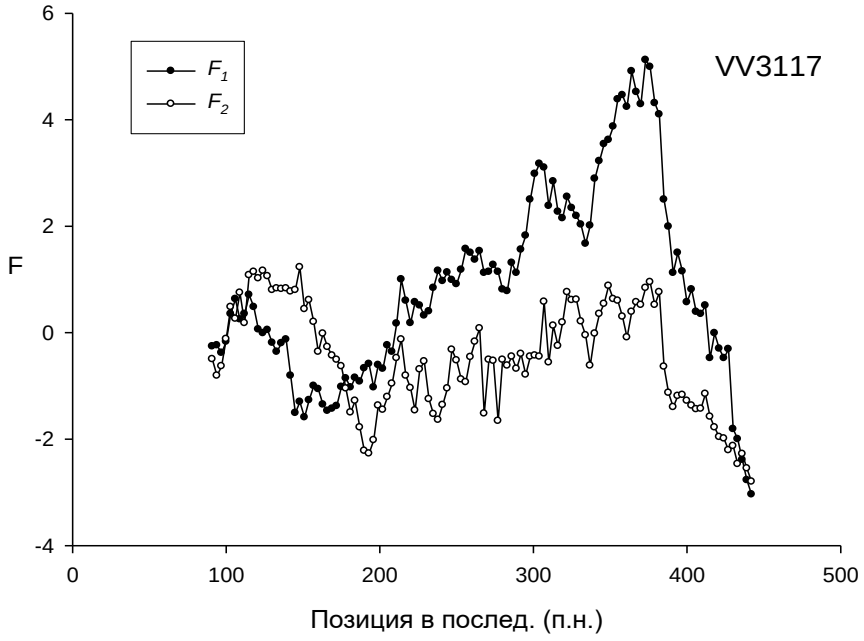

Рис. 4. Значения функций F_1 и F_2 для последовательности VV3117 из банка данных KEGG. Этот ген кодирует Protein affecting phage T7 exclusion by the F plasmid (Q7MGV9) из генома *Vibrio vulnificus*.

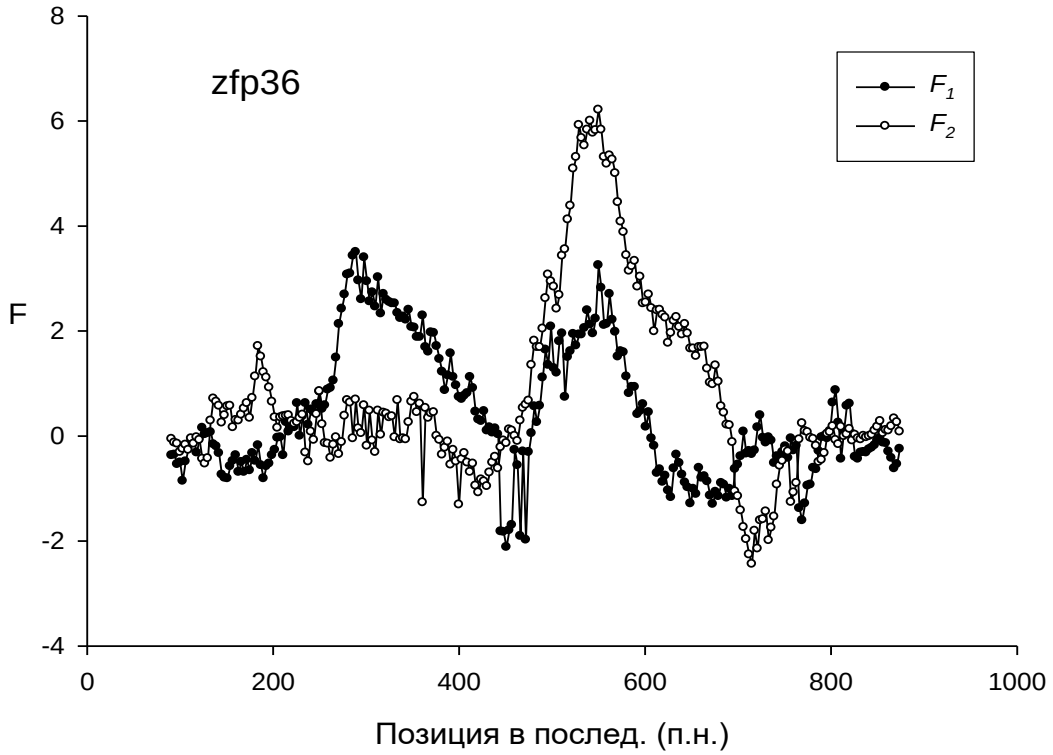

Рис. 5. Значения функций F_1 и F_2 для последовательности Zfp36 из генома Mus musculus из банка данных KEGG. Этот ген кодирует аминокислотную последовательность белка (NUP475), индуцируемого фактором роста (Growth factor-inducible nuclear protein NUP475).

3.2. Поиск белковых подобий для аминокислотных последовательностей

Рассмотрим гены, в которых мы обнаружили сдвиг фазы триплетной периодичности. Обозначим номер нуклеотида i, где произошел сдвиг фазы триплетной периодичности, как x_0. В этом случае для последовательности $s(i)$ для $i < x_0$ мы можем считать, что есть совпадение между триплетной периодичностью и существующей рамкой считывания. В то же время для последовательности $s(i)$ для $i > x_0$ будет наблюдаться сдвиг между триплетной периодичностью и существующей в гене рамкой считывания. Можно предполагать, что триплетная периодичность определяет рамку считывания, которая существовала в гене до мутации посредством сдвига рамки считывания. Будем называть эту рамку считывания древней рамкой считывания. Таким образом, в последовательности $s(i)$ для $i > x_0$ существует две рамки считывания – одна реально существующая в гене, а другая полученная на основе информации о сдвиге фазы триплетной периодичности. Эту рамку мы будем называть древней рамкой считывания. Если мутация посредством сдвига рамки считывания произошла в гене сравнительно недавно, то могут остаться варианты этого гена без мутации. Мы произвели перекодировку последовательностей ДНК со сдвигом фаз триплетной периодичности к аминокислотным последовательностям по двум рамкам считывания. Первая из них будет реально существующая аминокислотная последовательность, а вторая – предполагаемая аминокислотная последовательность, которую назовем древней аминокислотной последовательностью. Если в нуклеотидной последовательности было идентифицировано несколько сдвигов фазы триплетной периодичности, то мы для простоты рассматривали только сдвиг с максимальным значением x_0. Таким образом, мы создали 4724 пары аминокислотных последовательностей, и для них определялось наличие подобий в банке данных Swiss-Prot [30] при помощи программы Blast [31, 32].

```
T1: K  G  V  P  G  F  G  W  A  M  Q  A  A  A  Y  I  F  I  H  R
    aaaggtgttcctggatttggttgggccatgcaggctgctgcctatatcttcattcatagg
T2:  K  V  F  L  D  L  V  G  P  C  R  L  L  P  I  S  S  F  I  G

T1: K  W  K  D  D  K  S  H  F  E  D  M  I  D  Y  F  C  D  I  H
    aaatggaaggatgacaagagccatttcgaagacatgattgattacttttgtgatattcac
T2:  N  G  R  M  T  R  A  I  S  K  T  $  L  I  T  F  V  I  F  T

T1: E  P  L  Q  L  L  I  F  P  E  G  T  D  L  T  E  N  S  K  S
    gaaccacttcaactcctcatattcccagaagggactgatctcacagaaaacagcaagtct
T2:  N  H  F  N  S  S  Y  S  Q  K  G  L  I  S  Q  K  T  A  S  L

T1: R  S  N  A  F  A  E  K  N  G  L  Q  K  Y  E  Y  V  L  H  P
    cgaagtaatgcatttgctgaaaaaaatggacttcagaaatatgaatatgttttacatcca
T2:  E  V  M  H  L  L  K  K  M  D  F  R  N  M  N  M  F  Y  I  Q

T1: R  T  T  G  F  T  F  V  V  D  R  L  R  E  G  R  R  Q  S  E
    agaactacaggctttacttttgtggtagaccgtctaagagaaggcaggaggcagagcgag
T2:  E  L  Q  A  L  L  L  W  $  T  V  $  E  K  A  G  G  R  A  R

T1: G  S  K  G  P  L  Q  G  E  L  Q  I  T  A  Q  G  N  Q  R  G
    ggaagtaaaggaccacttcaaggagaactacaaatcactgctcaaggaaatcagagagga
T2:  E  V  K  D  H  F  K  E  N  Y  K  S  L  L  K  E  I  R  E  D

T1: H  K  Q  M  E  K  H  S  M  L  M  D  W  K  N  Q  Y  R  E  I
    cacaaacaaatggaaaaacactccatgctcatggattggaagaatcaatatcgtgaaatt
T2:  T  N  K  W  K  N  T  P  C  S  W  I  G  R  I  N  I  V  K  L

T1: G  H  T  A  Q  S  N  L  Q  I  Q  C  Y  S  H
    ggccatactgcccaaagcaatttacagattcaatgctattcccattaa
T2:  A  I  L  P  K  A  I  Y  R  F  N  A  I  P  I
```

Подобие аминокислотной последовательности полученной по рамке T_1 (реально существующая рамка) с последовательностью Q6UWP7 (LCLT1_HUMAN) из генома человека кодирующей *Lysocardiolipin acyltransferase 1*

```
Score =  205 bits (522), E-value = 7×10-53
Identities = 108/161 (67%), Positives = 117/161 (72%), Gaps = 17/161 (10%)

Query: 1   KGVPGFGWAMQAAAYIFIHRKWKDDKSHFEDMIDYFCDIHEPLQLLIFPEGTDLTENSKS 60
           KGVPGFGWAMQAAAYIFIHRKWKDDKSHFEDMIDYFCDIHEPLQLLIFPEGTDLTENSKS
Sbjct: 154 KGVPGFGWAMQAAAYIFIHRKWKDDKSHFEDMIDYFCDIHEPLQLLIFPEGTDLTENSKS 213

Query: 61  RSNAFAEKNGLQKYEYVLHPRTTGFTFVVDRLREGRRQSEGSKGPLQGELQITAQGNQRG 120
           RSNAFAEKNGLQKYEYVLHPRTTGFTFVVDRLREG+         L     IT
Sbjct: 214 RSNAFAEKNGLQKYEYVLHPRTTGFTFVVDRLREGKN--------LDAVHDITV-AYPHN 264

Query: 121 HKQMEKHSMLMDWKNQYREIGH--------TAQSNLQIQCY 153
             Q EKH +  D+  +     H        T++ +LQ+ C+
Sbjct: 265 IPQSEKHLLQGDFPREIHFHVHRYPIDTLPTSKEDLQLWCH 305
```

Подобие аминокислотной последовательности полученной по рамке T_2 (древняя рамка) с последовательностью P08548 (LIN1_NYCCO) из генома обезьяны *Nycticebus coucang* кодирующей LINE-1 reverse transcriptase homolog

```
Score = 96.3 bits (238), E-value = 6×10-20
Identities = 42/56 (75%), Positives = 50/56 (89%)

Query: 97  REVKDHFKENYKSLLKEIREDTNKWKNTPCSWIGRINIVKLAILPKAIYRFNAIPI 152
           ++VKD +KENY++L KEI ED NKWKN PCSW+GRINIVK++ILPKAIY FNAIPI
Sbjct: 774 KDVKDLYKENYETLRKEIAEDVNKWKNIPCSWLGRINIVKMSILPKAIYNFNAIPI 829
```

Рис. 6. Аминокислотные последовательности, полученные перекодированием последовательности гена 253558 (геном *H.sapiens*) по рамкам считывания T_1 и T_2, начиная с позиции 460 и до конца последовательности. Обозначение аминокислоты расположено непосредственно над первой позицией кодона для рамки T_1 и под первой позицией кодона для рамки T_2. Символом $ обозначены стоп-кодоны. Ниже приведены подобия этих аминокислотных последовательностей с белками базы данных Swiss-Prot. Координаты выравнивания для последовательности 253558 указаны без учета сдвига на 460 н.п. (или 153 ак).

Пороговое значение E-value [31,32] при использовании программы Blast было выбрано равным 0.001, что дает в среднем порядка ~ 5 случайных подобий при поиске подобий в базе данных Swiss-Prot для 5 тысяч аминокислотных последовательностей.

В результате проведенного сравнения 2555 пар последовательностей не имели подобия к каким-либо белкам Swiss-Prot. Для 1926 пар последовательностей подобие наблюдалось только для аминокислотной последовательности, созданной по рамке считывания, которая присутствует в гене. Для 182 пар последовательностей подобие наблюдалось только для древней аминокислотной последовательности, и для 61 пары последовательностей подобие наблюдалось для обеих аминокислотных последовательностей из пары.

В большинстве случаев, когда подобие было обнаружено для обеих аминокислотных последовательностей – древней и реально существующей, их функции совпадают. Этот факт может говорить о том, что мутация, приведшая к сдвигу рамки считывания, не затронула функциональный центр белка, и он сохранил способность к выполнению своей роли. Однако в некоторых случаях аминокислотные последовательности, полученные перекодированием по древней и существующей рамке считывания, имеют подобия к совершенно различным белкам. На рис.6 представлен фрагмент последовательности гена 253558 из генома человека, для которого в позиции 460 был найден сдвиг фазы триплетной периодичности на 1 нуклеотид и аминокислотные последовательности, полученные перекодирование нуклеотидной последовательности по двум различным рамкам считывания. Ниже приведены фрагменты результатов поиска подобий при помощи программы Blast между полученными аминокислотными последовательностями и белками банка данных Swiss-prot. Белок, кодируемый по древней рамке подобен аминокислотной последовательности LINE-1 reverse transcriptase с весом (Score) 96.3 бит. Значение E-value, показывающее ожидаемое число выравниваний с весом больше либо равным данному, составляет 5×10^{-20}. Белок, кодируемый геном 253558 в настоящее время, имеет высокую степень подобия с *Lysocardiolipin acyltransferase 1* (Score=205 бит , E-value= 2×10^{-50}).

4. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В данной работе удалось показать, что из 122829 генов, где триплетная периодичность не была выявлена на статистически значимом уровне, 4724 генов содержат вставки или делеции нуклеотидов, которые регистрируются посредством сдвига фазы триплетной периодичности. Это число в несколько раз больше количества генов со сдвигами рамки считывания, которые ранее были найдены поиском подобий или при помощи динамического программирования [25–27]. В этом смысле метод поиска сдвига фазы триплетной периодичности оказался в несколько раз более эффективным, чем все применяемые ранее методы.

Найденные 4724 генов со сдвигом фазы триплетной периодичности делают более вероятным гипотезу о том, что большая часть из 122829 генов, где триплетная периодичность не была найдена на статистически значимом уровне [10], вообще не содержат триплетную периодичность. Однако такая сравнительно небольшая часть генов, где потенциально может быть осуществлен сдвиг рамки считывания, может быть объяснена также несколькими другими причинами.

Во-первых, разработанный математический метод позволяет искать только сравнительно небольшие по размеру вставки и делеции символов. Это связано с тем, что протяженная вставка может разрушить как саму триплетную периодичность (формула 4), так и подобие матриц триплетной периодичности, которое мы определяем по формуле (6). Таким образом, данный метод будет пропускать значительную часть

генов содержащих большую (длиной более 50 оснований) вставку или же делецию нуклеотидов, которая может приводить к сдвигу рамки считывания.

Во-вторых, применяемый нами подход хорошо работает при небольшом количестве районов, где были произведены вставки или же делеции нуклеотидов. Если плотность вставок и делеций будет больше, чем одна вставка или делеция (не кратная трем основаниям) на несколько десятков нуклеотидов (~60), то выявить сдвиг фазы триплетной периодичности при помощи применяемого алгоритма будет невозможно. Это приведет к тому, что мы не сможем получить статистически значимое значение F_1 или F_2 для такого гена.

В-третьих, мы задали определенный уровень триплетной периодичности для гарантированного обнаружения сдвигов фазы триплетной периодичности (X больше нуля, формула (5)). Изучение сдвигов фазы триплетной периодичности для более низких значений X, вероятно, может позволить выявить большее количество генов, для которых F_1 или F_2 будет больше порогового уровня.

Применяемый нами подход поиска сдвигов между триплетной периодичностью и рамкой считывания для поиска мутаций в генах посредством рамки считывания при его дальнейшем развитии представляется нам предпочтительнее, чем применение поиска возможных подобий при помощи программы Blast. Это связано с тем, что данный метод для обнаружения в гене мутаций посредством сдвига рамки считывания не нуждается в дополнительной информации о подобиях в банке данных аминокислотных последовательностей. Так как объем банка данных ограничен, то всегда будет существовать вероятность того, что подобия не будут найдены, а в реальности мутация посредством сдвига рамок считывания существует. Мы полагаем, что более полное выявление мутаций посредством сдвига рамок считывания в генах возможно на пути объединения двух подходов. Это означает, что нужно также исследовать те гены, для которых F_1 и $F_2 < F_0$ и считать, что мы нашли мутацию посредством сдвига рамки считывания, если для аминокислотных последовательностей, созданных по рамкам считывания T_2 и T_3, существуют статистически значимые подобия. В этом случае сравнительно небольшой сдвиг фазы триплетной периодичность только указывает на возможность сдвига рамки считывания и факт такой мутации можно считать доказанным только после обнаружения подобий. С другой стороны, совершенствование применяемого в настоящей работе подхода может происходить на пути использования более совершенных алгоритмов поиска триплетной периодичности, например, таких как скрытые марковские модели. В этом случае, вероятно, удастся выявлять сдвиги рамки считывания, вызванные множеством событий вставок и делеций нуклеотидов в различные районы гена.

С функциональной точки зрения мутации посредством сдвига рамок считывания представляются событиями, способными кардинально изменить функцию гена и кодируемого им белка. Их осуществление может вносить большой вклад в образование новых генов посредством копирования известных генов и образования там мутаций посредством сдвига рамки считывания [25–27]. Однако генетический код также должен быть как-то адаптирован для этих событий [33] и новая аминокислотная последовательность должна обладать какой-то биологической функцией. В противном случае перебор мутационных событий для создания новой функции гена в его копии может быть слишком велик и неосуществим за разумное эволюционное время.

В свете этих предположений триплетная периодичность может служить неким тестом по проверке целостности гена в геноме. Если же ген был дуплицирован в геноме, то у новой копии такая проверка может не осуществляться, что открывает возможности для эволюционных изменений копии гена посредством сдвига рамки считывания и в итоге создания нового гена с новой биологической функцией.

СПИСОК ЛИТЕРАТУРЫ

1. Fickett J.W. Predictive methods using nucleotide sequences. *Methods Biochem. Anal.* 1998. V. 39. P. 231–245.
2. Staden R. Staden: statistical and structural analysis of nucleotide sequences. *Methods Mol. Biol.* 1994. V. 25. P. 69–77.
3. Baxevanis A.D. Predictive methods using DNA sequences. *Methods Biochem. Anal.* 2001. V. 43. P. 233–52.
4. Gutierrez G., Oliver J.L., Marin A. On the origin of the periodicity of three in protein coding DNA sequences. *J. Theor. Biol.* 1994. V. 167. № 4. P. 413–41.
5. 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.
6. 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.
7. Eskesen S.T., Eskesen F.N., Kinghorn B., Ruvinsky A. Periodicity of DNA in exons. *BMC Molecular Biology.* 2004. V. 5:12.
8. 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. V. 30. P. 157–166.
9. Konopka A.K. Sequences and codes: fundamentals of biomolecular cryptology. In: *Biocomputing: Informatics and genome projects.* Ed. Smith D. San Diego: Academic Press. P. 119–174.
10. Frenkel F.E., Korotkov E.V. Classification analysis of triplet periodicity in protein-coding regions of genes. *Gene.* 2008. V. 421. P. 52–60.
11. Trifonov E.N. Elucidating sequence codes: three codes for evolution. *Ann NY Acad. Sci.* 1999. V. 870. P. 330–338.
12. Eigen M., Winkler-Oswatitsch R. Transfer-RNA: the early adaptor. *Naturwissenschaften.* 1981. V. 68. P. 217–228.
13. Zoltowski M. Is DNA Code Periodicity Only Due to CUF - Codons Usage Frequency? *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2007. V. 1. P. 1383–1386.
14. 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.
15. Issac B., Singh H., Kaur H., Raghava G. P. S. Locating probable genes using Fourier transform approach. *Bioinformatics*. 2002. V. 18. № 1. P.196–197.
16. Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Biosci.* 1997. V. 13. № 3. P. 263–70.
17. Azad R.K., Borodovsky M. Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. *Briefings in bioinformatics*. 2004. V. 5. № 2. P. 118–130.
18. Henderson J., Salzberg S., Fasman K.H. Finding genes in DNA with a Hidden Markov Model. *J. Comput. Biol.* 1997. V. 4. P. 127–141.
19. Snyder E.E., Stormo G.D. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucl. Acids Res.* 1993. V. 21. P. 607–613.
20. Thomas A., Skolnick M.H. A probabilistic model for detecting coding regions in DNA sequences. *IMA J. Math. Appl. Med. Biol.* 1994. V. 11. № 3. P. 149–160.
21. Korotkov E.V., Korotkova M.A., Frenkel F.E., Kudryashov N.A. Information approach for search of periodicity of symbolical sequences. *Molek. Biol.* 2003. V. 37. P. 372–386.

22. Korotkov E.V., Korotkova M.A., Kudryashov N.A. Information decomposition method for analysis of symbolical sequences. *Physical Letters A*. 2003. V. 312. P. 198–210.
23. Ogata H., Goto S., Sato K., Fujibuchi W., Bono H., Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.* 1999. V. 27. P. 29–34.
24. 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.
25. Okamura K., Feuk L., Marquиs-Bonet T., Navarro A., Scherer S.W. Frequent appearance of novel protein-coding sequences by frameshift translation. *Genomics*. 2006. V. 88. P. 690–697.
26. Raes J., Van de Peer Y. Functional divergence of proteins through frameshift mutations. *Trends Genet.* 2005. V. 21. P. 428–431.
27. Kramer E.M., Huei-Jiun Su, Cheng-Chiang Wu, Jer-Ming Hu. A simplified explanation for the frameshift mutation that created a novel C-terminal motif in the *APETALA*3 gene lineage. *BMC Evolutionary Biology.* 2006. V. 6. № 30.
28. Kullback S. *Information Theory and Statistics.* New York: Wiley, 1959.
29. Hudson D. J. *Statistics: Lectures on Elementary Statistics and Probability*. Geneva: CERN, 1964; Moscow: Mir, 1967.
30. UniProt Consortium. The Universal Protein Resource (UniProt). *Nucl. Acids Res.* 2007. V. 35. P. 193–197.
31. Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 1970. V. 48. № 3. P. 443–453.
32. Altschul S.F. et al. Basic local alignment search tool. *J. Mol. Biol.*. 1990. V. 215. № 3. P. 403–410.
33. Bollenbach T., Vetsigian K., Kishony R. Evolution and multilevel optimization of the genetic code. *Genome Res.* 2007. V. 17. № 4. P. 405–412.

Материал поступил в редакцию 4.08.2009, опубликован 7.09.2009.