Математическая биология и биоинформатика. 2011. Т. 6. № 1. С. 79-91.
URL: http://www.matbio.org/2011/Rudenko2011(6_79).pdf

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

УДК: 577.212.2; 577.214

Применение метода Монте-Карло для поиска потенциальных сдвигов рамки считывания в генах

Руденко В.М.*[1,2], Коротков Е.В.[1,2]

[1]*Центр «Биоинженерия», Российская академия наук, Москва, 117312, Россия*
[2]*НИЯУ МИФИ, Москва, 115409, Россия*

Аннотация. В статье предложен метод поиска потенциальных сдвигов рамки считывания в генах, основанный на обнаружении точек разладки в распределении частот триплетов. Оценка статистической значимости разладки проводилась с помощью метода Монте-Карло. Корректность разработанного метода была продемонстрирована на последовательностях ДНК, содержащих искусственно внесенные в последовательность вставки. На предмет присутствия точек разладки были проанализированы последовательности банка данных *KEGG GENES*. На уровне значимости 6% было обнаружено, что более 140 тыс. последовательностей содержат точки разладки. Была проведена классификация последовательностей, имеющих точки разладки, по полю *description* в банке данных *KEGG GENES*. Оказалось, что большое число этих последовательностей являются псевдогенами, а во многих из них ранее были выявлены сдвиги рамки считывания. Наряду с этими последовательностями точки разладки были идентифицированы в генах, кодирующих *PE-PGRS, cation channel family protein, PPE family protein* и многие другие белки. Обсуждается связь между наличием в последовательности ДНК точки разладки и сдвигом рамки считывания.

Ключевые слова: *последовательность ДНК, рамка считывания, сдвиг рамки считывания, точка разладки, метод Монте-Карло.*

ВВЕДЕНИЕ

Мутации в генах осуществляются посредством замен оснований ДНК, а также посредством вставок или делеций как отдельных оснований ДНК, так и фрагментов ДНК различной длины. Нуклеотидные замены могут приводить к аминокислотным заменам в кодируемых белках. Достаточно часто аминокислотные замены могут оказывать серьезное влияние на способность белка выполнять биологическую функцию [1]. Другой возможный вид мутаций – вставки и делеции нуклеотидов. Они происходят в генах реже замен оснований, однако являются более значимыми эволюционными событиями. При делециях или вставках оснований длины, некратной трем, происходит изменение протяженного участка аминокислотной последовательности из-за сдвига рамки считывания. Вследствие этого аминокислотная последовательность полностью меняется после позиции сдвига рамки считывания.

Влияние сдвигов рамки считывания в генах на структуру и функцию белка сравнительно мало изучено. В ряде случаев они могут являться причиной наследственных болезней, так как белок, кодируемый измененной последовательностью, утрачивает свои функции. Например, было показано, что

* v.m.rudenko@gmail.com

делеция в гене *LAMA*2 приводит к врожденной мышечной дистрофии [2]. Мутация в гене транскрипционного фактора *NKX*2.5, вызываемая сдвигом рамки считывания, ассоциируется с врожденным пороком сердца [3]. По этой причине важной задачей является выявление мутаций в генах, связанных со сдвигом рамки, у потенциальных родителей. Выявление таких мутаций у родителей позволяет сделать вероятностные прогнозы о развитии того или иного наследственного заболевания их детей.

Необходимость поиска сдвигов рамки считывания в генах возникает также при секвенировании новых последовательностей ДНК. Возможные ошибки секвенирования (ошибочные вставки или делеции нуклеотидов в последовательности ДНК) приводят к тому, что аминокислотная последовательность гена определяется неправильно, что затрудняет ее корректную аннотацию.

Перечисленные выше задачи стимулировали разработку математических методов и компьютерных программ для определения возможных сдвигов рамки считывания в последовательностях оснований ДНК, кодирующих белки. Ранее был предложен ряд методов поиска сдвигов рамки считывания в генах. В качестве математического аппарата в них были использованы динамическое программирование [4–7], марковские модели [8, 9], а также некоторые статистические методы [10].

Первые методы, направленные на поиск сдвигов рамок считывания в генах, были связаны с началом многомасштабного секвенирования геномов, и соответственно, возникающими в ходе этих работ задачами выявления ошибок секвенирования и аннотации новых последовательностей. Эти методы были основаны на поиске гомологичных последовательностей в банке данных *Swiss-Prot* [4, 5]. Методы, основанные на поиске гомологий, обладают рядом ограничений. Во-первых, требуется по каким-то признакам выбрать ген, где предполагается сдвиг рамки считывания, затем найти в нем возможное место сдвига рамки считывания. Общий поиск сдвига рамок считывания по всем генам может потребовать достаточно больших компьютерных ресурсов. Во-вторых, необходимо, чтобы банк данных *Swiss-Prot* содержал аминокислотную последовательность, имеющую статистически значимое подобие с исследуемой аминокислотной последовательностью. Однако такая последовательность может отсутствовать ввиду ограниченности объема банка данных *Swiss-Prot* или из-за слишком больших эволюционных различий, накопленных между аминокислотными последовательностями. В силу этого используемый подход может выявить только некоторую часть сдвигов рамок считывания, накопленных в существующих генах к настоящему времени. Действительно, с помощью методов, основанных на поиске гомологий, удалось обнаружить только несколько сотен генов, имеющих сдвиг рамки считывания.

В работе [10] была предпринята попытка идентификации сдвигов рамки считывания без поиска подобий в банке данных *Swiss-Prot*. Для этой цели применялись кодирующие потенциалы, которые обычно используются для поиска кодирующих участков в последовательностях ДНК. Идея авторов статьи достаточно проста. В исследуемой последовательности выявляются кодирующие рамки считывания, и те позиции, где эта рамка меняется, идентифицируются как позиции сдвига. Авторами были рассмотрены несколько кодирующих потенциалов, по которым можно определить, какая из возможных рамок считывания является кодирующей. Наилучшим для поиска сдвигов оказались меры, учитывающие отклонения частот олигонуклеотидов различной длины (длиной 3 или 6 нуклеотидов) от частот, которые должны наблюдаться в кодирующей рамке исследуемого генома. Однако ожидаемые частоты олигонуклеотидов могут различаться даже в пределах одного генома. В связи с этим оказывается невозможным создание универсальной меры для всех последовательностей генома. Поэтому при использовании данного подхода необходимо производить настройку параметров метода в зависимости от того, к какой

категории принадлежит изучаемая последовательность, например, с низким или высоким значением показателя *CBI* (*codon bias index* [11]). Поэтому, несмотря на относительное разнообразие созданных к настоящему времени методов и программ для поиска сдвигов рамки считывания [4–10], до сих пор не существует универсального метода, хорошо работающего для любых, в том числе неизвестных последовательностей, для которых не задана никакая априорная информация.

Целью настоящей работы являлась разработка нового метода поиска потенциальных сдвигов рамки считывания в генах. Предложенный метод использует для поиска потенциальных сдвигов рамки считывания только нуклеотидную последовательность гена и не требует никакой дополнительной информации. Основная идея метода состоит в том, что частоты триплетов для различных рамок считывания в гене различаются. Сдвиг рамки считывания может создавать позицию в гене, в которой наблюдается точка разладки (*change-point*) [12] в распределении частот триплетов. Нами была введена мера разладки распределения частот триплетов, а также предложен метод для оценки статистической значимости разладки. Осуществлялся поиск точек разладки в последовательностях генов из банка данных *KEGG GENES Release 46* (далее: *Kegg-46*) [13]. Для того, чтобы показать связь между наличием точек разладки и присутствием сдвигов рамки считывания в генах, производилась классификация последовательностей, имеющих точки разладки, по полю *'description'* банка данных *Kegg-46*.

МЕТОДЫ

1. Определение точки разладки в распределении частот триплетов

В предыдущей работе [14] нами было отмечено, что разница в типах триплетной периодичности во фрагментах фиксированной длины w справа и слева от некоторой позиции k последовательности может служить показателем сдвига рамки считывания. Поскольку триплетная периодичность объясняется, в частности, предпочтениями клетки в использовании различных триплетов для кодирования определенных аминокислот и вырожденностью генетического кода [15–18], то можно предположить, что кодирующая последовательность без сдвига будет иметь однородное распределение триплетов вдоль всей своей длины. Напротив, в случае присутствия сдвига рамки в некоторой позиции k последовательности, в точке k будет наблюдаться разладка (*change-point*) в распределении частот триплетов.

Пусть $S = \{s(k), k = 1, 2, ..., L\}$ – последовательность оснований ДНК, где $s(k)$ принадлежит множеству символов $\{a, t, c, g\}$. Кодирование аминокислотной последовательности может происходить тремя способами, когда первые позиции триплетов расположены в $1 + 3n$, в $2 + 3n$ или в $3 + 3n$, $n = 1,2\ldots$ позициях последовательности. Эти способы перекодировки, назовем их *T1*, *T2*, *T3*, представляют собой три различные рамки считывания в гене (рис. 1).

```
DNA sequence:     ...atggcgagagaggtgcctatagagaaattg...
T1:               ...123123123123123123123123123123...
Am.acid seq1:     ...M  A  R  E  V  P  I  E  K  L  ...
T2:               ...312312312312312312312312312...
Am.acid seq2:     ....W  R  E  R  C  L  $  R  N  .....
T3:               ...231231231231231231231231231...
Am.acid seq3:     .....G  E  R  G  A  Y  R  E  I  ....
```

Рис.1. Три возможные рамки считывания в гене ($ - стоп-кодон).

Пусть требуется определить, присутствует ли разладка в позиции k последовательности. Для простоты будем рассматривать k, кратные 3. Рассмотрим фрагмент последовательности с координатами ($k - w + 1$, $k + w$), w – ширина окна до и после тестируемой позиции k, w кратно 3. Согласно нашему предположению, если разладка отсутствует, частоты триплетов в районе от $k - w + 1$ до k и в районе от $k + 1$ до $k + w$ будут подобными. При наличии точки разладки, возникающей после вставки или делеции не кратного 3 числа символов в позиции k, подобие частот триплетов должно наблюдаться для первой половины фрагмента, перекодированной по рамке *T1*, и для второй половины фрагмента, перекодированной по рамке *T2* или *T3*.

Будем проверять однородность частот триплетов между фрагментом последовательности с координатами ($k - w + 1$, k) в рамке *T1* и фрагментами ($k + j$, $k + w + j - 1$), перекодированными по рамкам *Tj*, $j = 1…3$, используя формулу [19]:

$$I_j = \sum_{i=1}^{64} f_i \log f_i + \sum_{i=1}^{64} v_i^j \log v_i^j - \sum_{i=1}^{64} (f_i + v_i^j) \log(f_i + v_i^j) + (N_1 + N_2) \log(N_1 + N_2) - \\ - N_1 \log N_1 - N_2 \log N_2 \qquad (1)$$

f_i - частоты триплетов для фрагмента последовательности с координатами ($k - w + 1$, k); v_i^j - частоты триплетов для фрагмента ($k + j$, $k + w + j - 1$); $N_1 = N_2 = w/3$ – количество триплетов. Гипотеза $H1$ состоит в том, что две выборки f_i и v_i^j принадлежат к различным популяциям, а гипотеза $H2$, в том, что f_i и v_i^j принадлежат к одной популяции. Как было показано ранее [19], в случае справедливости гипотезы $H2$ значение $2I_j$ будет распределено как χ^2 с 63 степенями свободы. Это означает, что в случае однородности частот триплетов f_i и v_i^j значения I_j будут невелики. Это также означает, что если в позиции k точки разладки нет, то минимум I_j будет достигаться для $j = 1$. Если I_2 или I_3 принимает малое значение, а I_1 велико, то позиция k представляет собой точку разладки. Согласно [20], распределение $2I_j$ хорошо согласуется со стандартным χ^2 распределением с 63 степенями свободы распределением, только если все значения f_i и v_i^j больше 10. При небольших длинах w (от 150 до 600 оснований) данное требование не выполняется. Поэтому для оценки статистической значимости разладки, под которой мы понимаем вероятность принятия гипотезы $H2$: $P_j = P(2I_j \geq 2I_j^0)$, где $2I_j^0$ – наблюдаемое значение функции $2I_j$ для данной последовательности, использовался метод Монте-Карло.

2. Оценка вероятности $P(2I_j \geq 2I_j^0)$ методом Монте-Карло

Для оценки вероятности $P(2I_j \geq 2I_j^0)$, $j = 1..3$, мы создавали последовательности $S_j = S(k-w+1,k) \,||\, S(k+j,k+w+j-1)$, где $S(i, j)$ – фрагмент последовательности S, начиная с $s(i)$ и заканчивая $s(j)$ символом, || показывает операцию конкатенации. Таким образом, S_1 совпадает с исходной последовательностью $S(k - w + 1,\ k + w)$, S_2 – последовательность, полученная из S_1 в результате делеции одного символа после позиции k, S_3 – последовательность, полученная из S_1 в результате делеции двух символов после позиции k.

Далее на основании последовательностей S_j генерировались три множества случайных последовательностей с такими же частотами триплетов, как в последовательностях S_1, S_2, S_3. Назовем эти множества как Q_1, Q_2 и Q_3. Случайные последовательности во множествах Q_1, Q_2 и Q_3 получались из последовательностей S_1, S_2, S_3 путем перемешивания триплетов. Перемешивание производилось с

использованием датчика случайных чисел по следующему алгоритму. Сначала каждому триплету последовательности ставилось в соответствие случайное число от 1 до 10000. Затем эти числа упорядочивались в порядке возрастания. Вместе с числами аналогичным образом упорядочивались и соответствующие им триплеты.

Количество случайных последовательностей в каждом множестве Q_1, Q_2 и Q_3. бралось равным 1000, что обеспечивает достаточную точность вычисления $P(2I_j \geq 2I_j^0)$. Для каждой случайной последовательности вычислялась величина $2I_j^n$ по формуле (1), $j = 1…3$ – номер множества случайных последовательностей, $n = 1…1000$ – номер последовательности во множестве. Далее для каждого множества Q_1, Q_2 и Q_3. определялись средние значения $2\bar{I}_j$ и дисперсии σ_j^2 случайных величин $2I_j^n$. Под I_j^0, будем понимать значение, рассчитанное по формуле (1) для исходной последовательности S_j. Введем в рассмотрение величины:

$$Z_j = \frac{2I_j^0 - \overline{2I}_j}{\sigma_j}. \qquad (2)$$

Z_j имеют приблизительно стандартное нормальное распределение [21]. Этот факт дает нам возможность использовать для оценки статистической значимости разладки вместо вероятности $P(2I_j \geq 2I_j^0)$ вероятность $p_j = P(N(0,1) > Z_j)$ того, что нормально распределенная с параметрами (0,1) случайная величина примет значение, большее либо равное Z_j.

Если в какой-либо позиции k последовательности имеется точка разладки, обусловленная сдвигом кодирующей рамки *T1* к рамке *T2*, то $Z_1 > Z_2$ и $Z_3 > Z_2$. При сдвиге от рамки *T1* к рамке *T3* будет наблюдаться $Z_1 > Z_3$ и $Z_2 > Z_3$.

Для поиска точек разладки в последовательности S удобно ввести величины F_2 и F_3:

$$F_2 = \log(p_2 / p_1),$$

$$F_3 = \log(p_3 / p_1). \qquad (3)$$

Точка локального максимума функции F_2 в позиции k, в которой выполняются условия: $F_2 > F_3$ и $F_2 > F_0$, где F_0 пороговое значение, будет свидетельствовать о том, что в позиции k наблюдается вставка $1 + 3n$ нуклеотидов, или же делеция $2 + 3n$ нуклеотидов. Аналогичным образом, при $F_3 > F_2$ и $F_3 > F_0$ можно говорить, что в позиции k имеет место вставка $2 + 3n$ нуклеотидов, или же делеция $1 + 3n$ нуклеотидов. Метод определения порогового уровня F_0 приведен ниже.

3. Алгоритм поиска точек разладки распределения частот триплетов в последовательности *S*

Для поиска точки разладки в последовательности S рассматривались все возможные позиции k кратные 3, начиная с $k = 150$. Для этих позиций вычислялись функции F_2 и F_3 для равной ширины окна w справа и слева от позиции k. Ширина окна w варьировалась от 150 до 600 п.н. с шагом 30. Считалось, что сдвиг в позиции k последовательности обнаружен, если для какой-либо ширины окна одна из функций F_2 или F_3 превышала пороговое значение F_0.

Алгоритм поиска сдвигов рамки считывания был реализован программно на языке *C++* и использованием библиотеки параллельного программирования MPI. Поскольку расчеты по методу Монте-Карло требуют больших вычислительных затрат, изучение последовательностей генов из банка данных *Kegg-46* проводилось на компьютерном кластере Центра Биоинженерия РАН, состоящего из 110 процессоров P4.

РЕЗУЛЬТАТЫ

1. Тестирование метода на искусственных последовательностях со вставками символов

Сначала мы изучили разработанным нами методом тестовые последовательности. Тестовая последовательность представляла собой последовательность ДНК из банка данных *Kegg-46*, в которой точки разладки не обнаруживались. Затем в случайную позицию тестовой последовательности k от 150 до $L - 150$, где L – длина последовательности, добавляли один или два нуклеотида с целью создания искусственного сдвига рамки считывания. Модифицированная таким образом последовательность анализировались снова, и в большинстве случаев точка разладки выявлялась вблизи позиции, куда были добавлены символы, на достаточно высоком уровне значимости. Один из примеров тестовой последовательности – последовательность с идентификатором BC0013. Она относится к геному *Bacillus cereus* и кодирует *inosine 5'-monophosphate dehydrogenase*. Для этой последовательности были построены графики функций F_2, F_3 в зависимости от позиции предполагаемого сдвига k при ширине окна $w = 210$ (рис. 2а). Распределение триплетов в последовательности BC0013 однородно по всей ее длине. Об этом говорят малые значения функций F_2 и F_3, их значения не превышают −1.0. Графики F_2 и F_3 после добавления в позицию 600 одного нуклеотида показаны на рис. 2б.

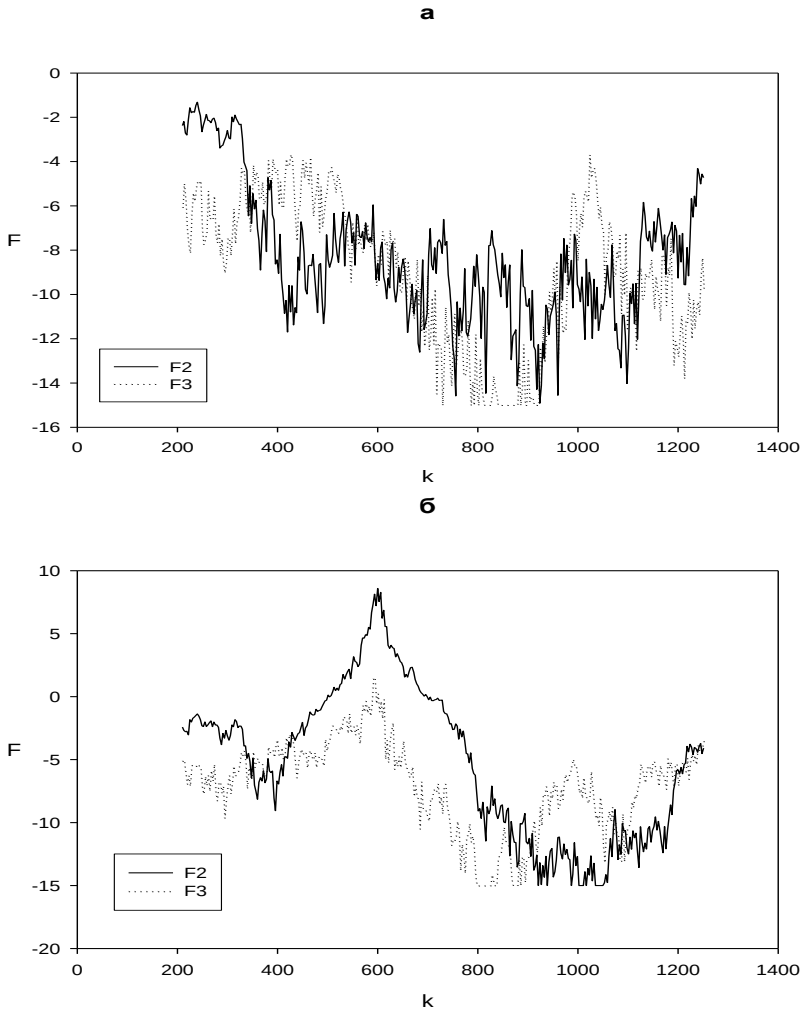

Рис. 2. Графики функций F_2 и F_3 для последовательности а) *BC0013* из генома *Bacillus cereus*, б) та же последовательность со вставкой символа 'c' в позицию 600.

Как видно, в позиции, близкой к 600 наблюдается резкий рост значений функции $F_2(600) = 8.59$, что соответствует переходу от рамки считывания *T1* к *T2*. Функция F_3 по-прежнему сохраняет небольшие значения, $F_3(600) = 0.25$. Аналогичные кривые также были получены для многих других тестовых последовательностей. Это результаты подтверждают наше предположение о том, что неоднородность распределения триплетов слева и справа от позиции k в последовательности S может служить показателем наличия сдвига рамки считывания в гене.

2. Применение метода Монте-Карло для поиска порогового значения F_0

Чтобы найти пороговое значение F_0 для функций F_2, F_3, описанный выше алгоритм применялся для анализа последовательностей из случайного множества, имеющего то же распределение длин последовательностей и частот триплетов, как банк данных генов *Kegg-46*. Последовательности случайного множества были получены из последовательностей банка данных *Kegg-46* путем случайного перемешивания триплетов. Уровень F_0 выбирался таким образом, чтобы число последовательностей, определенных как имеющих точки разладки, в случайном множестве составлял достаточно малую величину от количества последовательностей с точками разладками, обнаруженных в банке данных *Kegg-46*. Эта величина характеризует число ошибок первого рода и при уровне $F_0 = 5.0$ составляет 5.9%. В расчетах мы использовали уровень $F_0 = 5.0$, а также уровень $F_0 = 3.75$, которому соответствует количество ошибок первого рода 18.0%.

3. Поиск последовательностей, имеющих точки разладки в распределении частот триплетов, в банке данных *Kegg-46*

Поиск последовательностей, содержащих точки разладки, был проведен для генов, собранных в банке данных *Kegg-46*. Для исследования выбирались последовательности с длиной более 300 н.п. Их число банке *Kegg-46* составило 2941437. В качестве порогового уровня использовалось $F_0 = 5.0$. На этом уровне было обнаружено 140138 последовательностей с точками разладки, что составляет 4.8% от всего числа исследуемых последовательностей. Из них 81096 (58%) точек разладки ассоциировано с переходом от рамки *T1* к рамке *T2* (F_2 велико). Остальные случаи соответствуют переходу от рамки *T1* к *T3*. Для порогового уровня $F_0 = 3.75$ было обнаружено 225803 последовательностей с точками разладки.

Некоторые примеры последовательностей, имеющих точки разладки в распределении частот триплетов, приведены ниже. На рис. 3 изображены графики функций F_2 и F_3 в зависимости от позиции k для последовательности *VCA0563*. Последовательность *VCA0563* принадлежит геному бактерии *Vibrio cholerae*, кодирует *NAD(P)*. Из рисунка видно, что в позиции k, близкой к 1100 основанию наблюдается максимум функции $F_3(1092) = 6.04$, в то время как $F_2(1092) = 0.97$. Это может свидетельствовать о делеции $1 + 3n$ или вставке $2 + 3n$ символов в последовательности в позиции 1092.

Пример, приведенный на рис. 4, демонстрирует наличие двух точек разладки в последовательности *MCA1493*, кодирующей *cellulose-binding domain protein*. Последовательность относится к геному *Methylococcus capsulatus*. Первая точка разладки наблюдается для координаты $k = 519$, $F_2(519) = -3.33$, $F_3(519) = 5.88$, она соответствует переходу от рамки *T1* к *T3*. Вторая разладка произошла в позиции $k = 765$, $F_2(765) = 10.05$, $F_3(765) = -14.98$, что отвечает переходу от рамки *T1* к *T2*. Таким образом, после двух точек разладки рамка считывания «вернулась» в свое первоначальное положение.

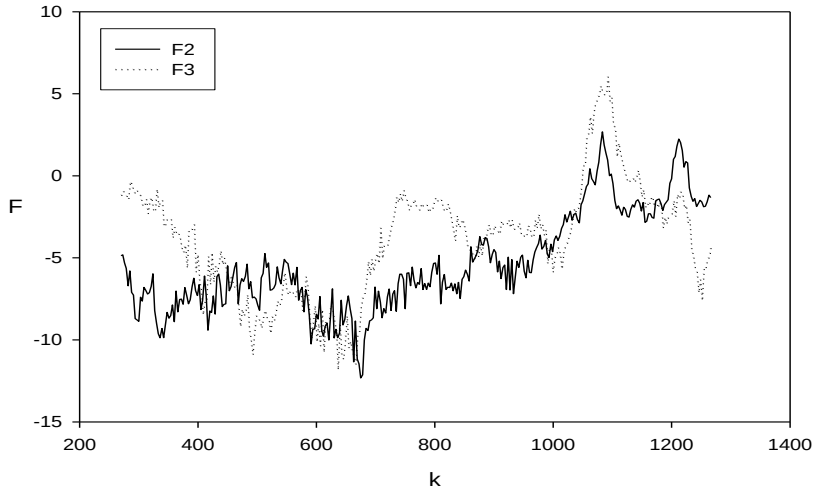

Рис. 3. Графики функций F_2 и F_3 для последовательности *VCA0563* из генома *Vibrio cholerae*.

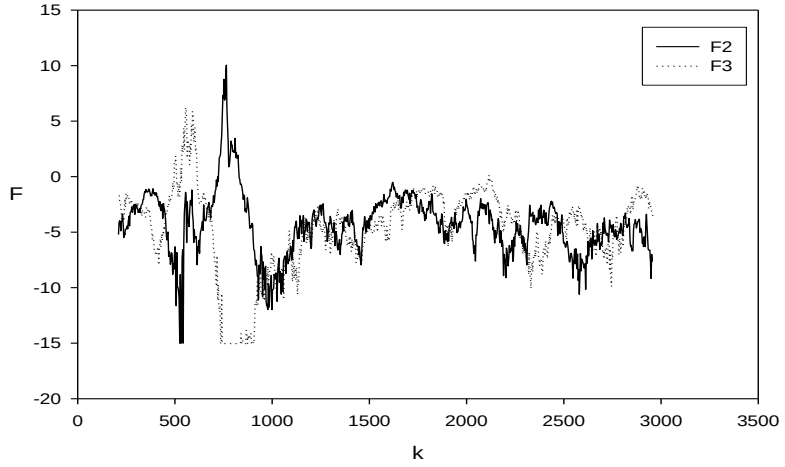

Рис.4. Графики функций F_2 и F_3 для последовательности *MCA1493* из генома *Methylococcus capsulatus*.

Нами была проведена классификация последовательностей, содержащих точки разладки в распределении частот триплетов в соответствии с их описанием (поле *description*) в банке данных *Kegg-46* (табл. 1).

Оказалось, что среди выявленных нами последовательностей, 842 последовательности были ранее аннотированы как гены, имеющие сдвиги рамки считывания (*frameshift*). Это составляет 41% от общего количества генов в *Kegg-46*, аннотированных как *frameshift*. Данный результат служит подтверждением в пользу нашего предположения об однородности частот триплетов по всей длине гена и показывает корреляцию между наличием в последовательности точки разладки и сдвигом рамки считывания. Среди последовательностей, идентифицированных нами как последовательности с точками разладками, достаточно много псевдогенов – 6073.

Интересным фактом является то, что кроме уже известных случаев сдвигов рамки считывания, нам удалось обнаружить большое число точек разладки в генах, кодирующих *transposases*, *PE-PGRS proteins*, *translation initiation factors IF-2*, *protein kinases*. Наличие точек разладки в однотипных белках наводит на мысль о том, что в последовательности, являющейся их общим предком, произошла мутация типа вставки/делеции, которая в дальнейшем была закреплена в процессе эволюции.

Таблица 1. Классификация генов, имеющих точки разладки, по описанию в банке данных *Kegg-46*

№	Описание последовательности (поле 'description')	Кол-во посл. с разладками	Число посл. в *Kegg-46*	Отношение посл. с разладками к общему числу (%)
1	Pseudogene	6073	29048	21
2	Zinc finger	3090	7441	42
3	Protein kinase	2174	10042	22
4	ABC transporter related	1011	50550	2
5	Frameshift	842	2050	41
6	Transposase	802	27440	3
7	Lipoprotein	547	17018	3
8	Translation initiation factor IF-2	359	1414	25
9	PE-PGRS	288	584	49
10	Mucin-associated surface protein (MASP)	249	965	26
11	Cation channel family protein	244	528	46
12	Cyclic nucleotide-binding domain containing protein	172	1141	15
13	Exodeoxyribonuclease	171	2320	7
14	Trans-sialidase	157	756	21
15	PPE family protein	152	485	31

4. Изучение количества стоп-кодонов в последовательностях ДНК, содержащих точки разладки

Рассмотрим последовательность ДНК S длины L, в позиции k которой была обнаружена точка разладки, которая может указывать на наличие сдвига рамки считывания. Мы предполагаем, что вплоть до позиции сдвига, то есть для фрагмента последовательности $S(1, k)$, кодирующей рамкой являлась рамка *T1*. Для второй части последовательности $S(k + 1, L)$ действующая кодирующая рамка также является *T1*. Но перед тем как произошла мутация, образовавшая предполагаемый сдвиг рамки считывания, кодирующая рамка была другая (*T2* или *T3*). Мы будем называть эту рамку считывания древней. Таким образом, каждому гену, в котором была обнаружена точка разладки, можно сопоставить две рамки считывания. Первая рамка (*T1*) реально существует в анализируемом гене, вторая рамка является гипотетической, ее можно восстановить, исходя из информации о найденных разладках. Если сдвиг рамки считывания произошел не очень давно, то похожие нуклеотидные последовательности без сдвига могут присутствовать в геномах других видов. Для координат $i > k$ мы можем получить две аминокислотные последовательности для двух вариантов рамки считывания. Первая последовательность представляет собой реально существующую аминокислотную последовательность, вторая представляет собой гипотетическую аминокислотную последовательность, которую назовем древней аминокислотной последовательностью.

Таким образом, мы использовали две аминокислотные последовательности для каждого гена, в котором была обнаружена точка разладки. Если в какой-либо последовательности присутствовало несколько точек разладки, то было создано несколько аминокислотных последовательностей соответствующих фрагментам исходной последовательности ДНК от позиции k первой точки разладки до позиции k второй точки разладки, от позиции k второй точки разладки до позиции k третьей точки разладки и т. д. Если после какой-либо точки разладки соответствующий фрагмент

последовательности оказывался в рамке *T1*, он не добавлялся во множество древних аминокислотных последовательностей. Это означает, что была произведена полная реконструкция предполагаемых сдвигов рамки считывания в гене на основе найденных точек разладки.

Последовательность белка транслируется из последовательности ДНК в соответствии с таблицей генетического кода, начиная с определенного инициирующего триплета *atg*, и заканчивая одним из стоп-кодонов: *tag, taa* или *tga*. Таким образом, если действительно имел место сдвиг рамки считывания, и древняя рамка кодировала белок, можно ожидать, что количество стоп-кодонов в древней аминокислотной последовательности будет значительно меньшим, чем в случайной последовательности ДНК с тем же распределением нуклеотидов по позициям триплетов [21]. Для того чтобы проверить данное предположение, мы рассчитывали отклонения наблюдаемого числа стоп-кодонов от ожидаемого в последовательностях ДНК, где были обнаружены точки разладки. Позиции триплетов соответствовали древней рамке считывания. Мы определяли в исследуемых последовательностях позиционно-специфические частоты $c(i,j)$, где i – позиция триплета, $i = 1,2,3$, j – нуклеотид, $j = a,t,c,g$. Далее на основании $c(i,j)$ были рассчитаны вероятности стоп-кодонов по формуле:

$$p_{stop} = \frac{27}{L^3} \times \{c(1,t)c(2,a)c(3,g) + c(1,t)c(2,a)c(3,a) + c(1,t)c(2,g)c(3,a)\} \quad (4)$$

Пусть $N = L/3$ – общее число кодонов в древней рамке считывания последовательности S. Ожидаемое количество стоп-кодонов в последовательности S оценивается величиной Np_{stop}, а дисперсия соответственно $-Np_{stop}(1-p_{stop})$. Отклонение наблюдаемого числа стоп-кодонов в последовательности S от ожидаемого можно рассчитать по формуле:

$$X = (N_{stop} - Np_{stop}) / \sqrt{Np_{stop}(1-p_{stop})} \quad (5)$$

Было построено распределение X для множества всех древних последовательностей ДНК, восстановленных с учетом найденных разладок. Далее было сгенерировано множество случайных последовательностей. Случайные последовательности получались из исходных путем перемешивания их символов. Причем отдельно перемешивались символы, расположенные в первых позициях триплетов, вторых и третьих позициях триплетов. Таким образом, случайные последовательности имели те же распределения длин и позиционно-специфических частот $c(i, j)$, как и исходное множество древних последовательностей ДНК. Для случайного множества также было рассчитано распределение статистики X. Оба распределения приведены на рис. 5. По оси абсцисс – значение X, умноженное на 10, по оси ординат – число последовательностей с заданным X. Из графиков можно видеть, что последовательности ДНК, восстановленные по древним рамкам считывания, имеют значительно меньшее число стоп-кодонов, по сравнению со случайными последовательностями. Эти наблюдения работают в пользу того факта, что наличие точек разладки в распределении частот триплетов в последовательности указывают на сдвиги рамки считывания в анализируемых нами генах.

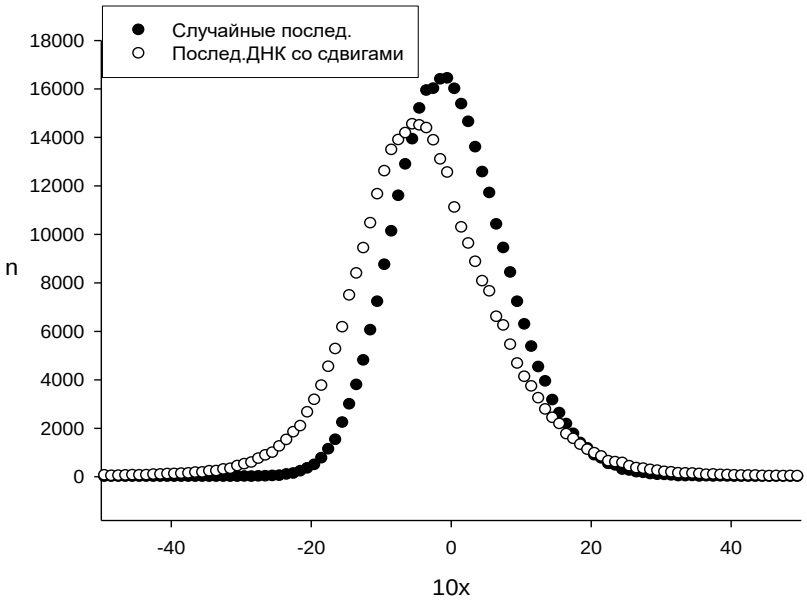

Рис. 5. Распределение X для ° – последовательностей ДНК, восстановленных по древним рамкам считывания и • – случайных последовательностей.

ОБСУЖДЕНИЕ

В данной работе нами был разработан математический аппарат и программное обеспечение для определения точек разладки в распределении частот триплетов в последовательностях генов. Исследовалось присутствие точек разладки в последовательностях генов из банка данных *Kegg-46*. Удалось показать, что при количестве ошибок первого рода около 6% более 140 тыс. генов содержат точки разладки. Мы полагаем, что это число является нижней оценкой числа генов, содержащих точки разладки. В реальности число таких генов может быть значительно бо́льшим. Сравнительно небольшое число точек разладки можно объяснить тем, что для вычисления функции I_j используются достаточно протяженные фрагменты последовательности длины w, минимально 150 нуклеотидов слева и справа от тестируемой позиции k. Таким образом, точки разладки внутри фрагментов последовательности ДНК длины менее 300 нуклеотидов, а также парные точки разладки, между которыми менее 150 нуклеотидов, не выявляются предложенным методом.

Точки разладки в коротких последовательностях можно находить, если сравнивать не частоты триплетов слева и справа от позиции k, а если сравнить триплетные матрицы, так как это было сделано в работе [23]. В [23] производился поиск сдвигов фазы триплетной периодичности в генах. Позиция сдвига по сути дела является точкой разладки в распределении позиционно-специфических частот нуклеотидов. Если сравнивать два метода – предложенный в [23] и представленный в данной работе, то можно отметить, что последний из них определяет большее число точек разладки (140 тыс.) в последовательностях длины более 300 п.н, чем метод поиска сдвигов фазы триплетной периодичности (112 тыс.). Данные приведены для уровня значимости 6%. Таким образом, нам удалось увеличить количество выявляемых последовательностей, имеющих точки разладки, на 25%. Это объясняется тем, что оценка статистической значимости в данной работе проводилась методом Монте-Карло. Статистика Z_j строилась, исходя из набора триплетов для конкретной последовательности, поэтому оценка $P_j = P(2I_j \geq 2I_j^0)$ точнее, чем определяемая на основании формул для стандартных распределений.

Мы считаем, что для обнаружения всех точек разладки в последовательностях ДНК целесообразно применять оба математических подхода, в зависимости от длины сканирующего окна *w*.

Интересно рассмотреть возможные причины возникновения точек разладки в генах. Присутствие точек разладки в большом числе последовательностей ДНК, аннотированных как содержащие frameshift, указывают на то, что одной из причин разладок являются сдвиги рамки считывания, произошедшие вследствие вставок и делеций нуклеотидов с длиной не кратной трем основаниям. Такие вставки или же делеции, приводящие к сдвигам рамки считывания, могут быть одним из эволюционных механизмов возникновения качественно новых белков.

Данные этой работы и полученные ранее результаты [23] показывают достаточно большое количество генов, которые имеют точки разладки. Если точка разладки указывает на сдвиг рамки считывания, то белки как-то должны «выживать» после такого значительного изменения своей последовательности. Можно предполагать, что такие мутации в большинстве случаев либо не влияют на выполняемую им функцию, либо не затрагивают активный центр белка. В противном случае аминокислотные последовательности белков и генетический код должны обеспечивать создание функциональных последовательностей после сдвига рамки считывания, чтобы можно было бы объяснить существование такого большого числа сдвигов рамок считывания в существующих генах.

Можно также предположить, что значительная часть генов «теряется» после образования сдвига рамки считывания, так как белок перестает транслироваться с мутированного гена из-за наличия стоп-кодонов в альтернативной рамке считывания. Сдвиг рамки считывания может быть также неким переключателем между активным и пассивным состоянием гена или между несколькими функциями одной и той же нуклеотидной последовательности.

Нам представляется маловероятным, что наличие точек разладки связано со вторичной структурой белков. Дело в том, что для основной массы генов, где присутствуют почти все всевозможные комбинации α-спиралей и β-слоев, наблюдается однородность в частотах триплетов оснований ДНК слева и справа от позиции k в генах. В последовательностях этих генов точки разладки отсутствуют. Существование такой однородности у почти 90% известных генов представляется достаточно интересным явлением. Такая однородность может заключать в себе некие функции проверки целостности генов в смысле отсутствия сдвигов рамки считывания.

Практическая ценность разработанного в работе метода состоит в том, что он позволяет выявлять в последовательностях ДНК мутации – вставки и делеции длины, некратной трем. На основании этих данных можно «восстановить» древнюю аминокислотную последовательность белка, кодируемого геном, что может представлять интерес для изучения эволюционных процессов в генах.

Работа выполнена при финансовой поддержке ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009-2013гг.

СПИСОК ЛИТЕРАТУРЫ

1. Watson J.D., Levine M., Baker T.A., Gann A., Bell S.P. *Molecular Biology of the Gene.* Benjamin-Cummings Pub Corp., 2007.
2. Salem I.H., Kamoun F. et al. Mutations in LAMA2 and CAPN3 genes associated with genetic and phenotypic heterogeneities within a single consanguineous family involving both congenital and progressive muscular dystrophies. *Bioscience Reports.* URL: http://journals.academia.edu/BioscienceReports (дата обращения: 09.03.2011).

3. Stallmeyer B., Fenge H., Nowak-Gottl U., Schulze-Bahr E. Mutational spectrum in the cardiac transcription factor gene NKX2.5 (CSX) associated with congenital heart disease. *Clin Genet*. 2010. V. 78. № 6. P. 533–540.
4. Posfai J., Roberts R.J. Finding errors in DNA sequences. *Proc. Natl. Acad. Sci. USA.* 1992. V. 89. P. 4698–4702.
5. Claverie J.-M. Detecting frame shifts by amino acid sequence comparison. *J. Mol. Biol*. 1993. V. 234. № 4. P. 1140–1157.
6. 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.
7. Raes J., Van de Peer Y. Functional divergence of proteins through frameshift mutations. *Trends Genet.* 2005. V. 21. P. 428–431.
8. Schiex Th., Gouzy J., Moisan A., Oliveira Y. FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *NAR*. 2003. V. 31. № 13. P. 3738–3741.
9. Kislyuk A., Lomsadze A., Lapidus A.L., Borodovsky M. Frameshift detection in prokaryotic genomic sequences. *Int. J. Bioinformatics Research and Applications*. 2009. V. 5. № 4. P. 458–477.
10. Fichant G.A., Quentini Y. A frameshift error detection algorithm for DNA sequencing projects. *NAR.* 1995. V. 23. № 15. P. 2900–2908.
11. Bennetzen J.L., Hall B.D. Codon selection in yeast. *J. Biol. Chem*. 1982. V. 257. P. 3026–3031.
12. *Change-point problems.* Ed. Carlstein E., Muller H.-G., Siegmund D. Institute of mathematical statistics, 1994. V. 23. (Lecture notes – monograph series).
13. Kanehisa M., Goto S., Kawashima S., Okuno Y., Hattori M. The KEGG resources for deciphering the genome. *Nucleic Acids Res.* 2004. V. 32. P. 277-280.
14. Коротков Е.В., Руденко В.М. Сдвиг фазы триплетной периодичности в нуклеотидных последовательностях генов. *Математическая биология и биоинформатика*. 2009. Т. 4. № 2. С. 66–80.
15. Trifonov E.N. Elucidating sequence codes: three codes for evolution. *Ann. NY Acad. Sci.* 1999. V. 870. P. 330–338.
16. Eigen M., Winkler-Oswatitsch R. Transfer-RNA: the early adaptor. *Naturwissenschaften*. 1981. V. 68. P. 217–228.
17. 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.
18. 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.
19. Kullback S. *Information Theory and Statistics.* New York: Wiley, 1959.
20. Filina M.V., Zubkov A.M. Exact computation of Pearson statistics distribution and some experimental results. *Austr. J. Statist*. 2008. V. 37. № 1. P. 129–135.
21. Sprinthall R.C. *Basic Statistical Analysis: Seventh Edition*. Boston: Pearson Education Group, 2003.
22. Carpena P., Bernaola-Galván P., Román-Roldán R., Oliver J. A simple and species-independent coding measure. *Gene*. 2002. V. 300. № 1-2. P. 97–104.
23. 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.

Материал поступил в редакцию 29.03.2011, опубликован 16.05.2011.