МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2008, том 42, № 4, с. 707–720

МАТЕМАТИЧЕСКАЯ И СИСТЕМНАЯ БИОЛОГИЯ

УДК 577.212.2; 577.214

КЛАССИФИКАЦИЯ ТРИПЛЕТНОЙ ПЕРИОДИЧНОСТИ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК ГЕНОВ, СОБРАННЫХ В БАНКЕ ДАННЫХ KEGG

© 2008 г. Ф. Е. Френкель*, Е. В. Коротков

Центр "Биоинженерия" Российской академии наук, Москва, 117312

Поступила в редакцию 14.12.2007 г.
Принята к печати 04.03.2008 г.

Проведена классификация 472 288 районов триплетной периодичности, найденных нами в 578 868 генах из банка данных KEGG (версия 29). Введено понятие класса триплетной периодичности и меры подобия между классами. Всего удалось создать 2 520 классов, которые содержат 94% от общего количества найденных случаев триплетной периодичности. Одинаковая привязка триплетной периодичности к рамке считывания наблюдается для 92% районов триплетной периодичности, входящих в классы. В остальных районах периодичности обнаружен сдвиг рамки считывания гена относительно рамки считывания, свойственной большинству генов, входящих в данный класс триплетной периодичности. Эти районы периодичности перекодированы в гипотетические аминокислотные последовательности в соответствии с рамкой считывания, построенной по классу триплетной периодичности. С использованием программы Blast показано, что 2 660 гипотетических аминокислотных последовательностей имеют статистически значимое подобие к белкам из банка данных UniProt. Мы предполагаем, что 8% районов триплетной периодичности, вошедших в классы, мутировали вследствие сдвига рамки считывания. Созданные классы триплетной периодичности могут применяться для идентификации кодирующих районов генов, в том числе и для поиска мутаций вследствие сдвига рамки считывания.

***Ключевые слова*: триплетная периодичность, классификация, поиск генов, кодирующие районы, сдвиг открытой рамки считывания.**

CLASSIFICATION OF TRIPLET PERIODICITY IN DNA SEQUENCES OF GENES TAKEN FROM KEGG DATABANK, by *F. E. Frenkel*, E. V. Korotkov* (Bioengineering Center, Russian Academy of Sciences, Moscow, 117312 Russia; *e-mail: felix.frenkel@gmail.com). We conducted classification for 472 288 regions of triplet periodicity found in 578 868 genes from release 29 of KEGG databank. A new concept of triplet periodicity class and a measure of similarity between them are introduced. Totally 2 520 classes were created that contain 94% of found triplet periodicity. For 92% of triplet periodicity regions contained in classes an identical linkage of triplet periodicity to reading frame is observed. For the rest triplet periodicity cases a shift between reading frame of a gene and reading frame common for majority of genes contained in a class of triplet periodicity was observed. These periodicity regions were encoded into hypothetical amino acid sequences in accordance with reading frame built by triplet periodicity class. By BLAST program it was shown that 2660 hypothetical amino acid sequences have statistically significant similarity with proteins from UniProt databank. We suppose that 8% of triplet periodicity regions that joined classes mutated by means of reading frame shift. Created classes of triplet periodicity can be used for identification of coding regions of genes as well as for searching for mutations arisen from reading frame shift.

***Key words*: triplet periodicity, classification, gene finding, coding regions, open reading frame shift.**

ВВЕДЕНИЕ

Триплетная организация последовательностей ДНК, кодирующих белки, является свойством всех известных на настоящее время живых систем [1–9]. Причина этого кроется не только в структуре генетического кода [10, 11], но и в использовании "любимых" триплетов оснований ДНК для кодирования определенных аминокислот [12, 13], а также в насыщенности белков определенными аминокислотами [14, 15]. Также предполагается, что триплетная периодичность могла возникнуть как свойство нуклеотидной последовательности гена, которое может контролировать в нем мутации, возникающие вследствие сдвига рамок считывания [16]. Изучение триплетной периодичности генов может представлять интерес как для развития более мощ-

* Эл. почта: felix.frenkel@gmail.com

Рис. 1. Влияние делеции одного основания на триплетную периодичность нуклеотидной последовательности. Цифры сверху последовательностей S1, S2, S3 и S4 показывают позиции нуклеотидов в рамке считывания. Из последовательности S1 удалено 25-ое основание (подчеркнуто). В результате последовательность S1 можно представить как две последовательности – S2 и S3. В этих последовательностях будет такая же триплетная периодичность (матрицы *M2* и *M3*), но в последовательности S3 она будет сдвинута на одно основание относительно новой рамки считывания, образовавшейся после делеции. Это означает, что первый столбец матрицы *M2* соответствует третьему столбцу матрицы *M3*, второй столбец матрицы *M2* соответствует первому столбцу матрицы *M3,* третий столбец матрицы *M2* соответствует второму столбцу матрицы M3. Как результат делеции мы получаем последовательность S4, которая имеет матрицу триплетной периодичности *M4 = M1 + M2*. В результате такого суммирования первый столбец матрицы *M2* сливается с первым столбцом матрицы *M3* и так далее, что приводит к слиянию неидентичных столбцов и значительно ухудшает статистическую значимость триплетной периодичности в последовательности S4.

ных алгоритмов поиска кодирующих участков ДНК [17], так и для исследования эволюции кодирующих последовательностей ДНК [10, 11]. С целью 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 применяются преобразование Фурье [18, 19], скрытые цепи Маркова [20, 21], нейронные сети [22, 23] и некоторые другие статистические методы, основанные на позиционно-зависимых предпочтениях нуклеотидов в кодирующих последовательностях [17].

При использовании применяемых в настоящее время математических методов как кодирующих потенциалов возникают определенные проблемы. Во-первых, методы, основанные на преобразовании Фурье, не позволяют выявить триплетную периодичность со вставками и делециями символов, а также не позволяют различить триплетную периодичность, найденную в одной последовательности оснований ДНК от триплетной периодичности, найденной в другой последовательности. Это можно сделать, если ввести такое понятие как класс триплетной периодичности. Под классом можно понимать некоторое свойство, которое позволит количественным образом отличить триплетную периодичность, найденную в одной последовательности ДНК от триплетной периодичности, найденной в другом районе ДНК. Методы, основанные на динамическом программировании, позволяют выявлять периодичность с делециями и вставками символов, но не позволяют найти достаточно размытую периодичность в кодирующих районах ДНК из-за использования весовых матриц для нуклеотидных пар [24, 25]. При использовании нейронных сетей создается обучающая выборка [22, 23] из кодирующих последовательностей, где могут присутствовать антагонистические триплетные периодичности, которые ослабляют специфичность триплетной периодичности для выявления кодирующих районов.

Триплетная периодичность важна для разработки математических алгоритмов предсказания кодирующих районов ДНК [17]. Задача состоит в том, чтобы разработать такой способ выявления триплетной периодичности, который выявлял бы ее на наиболее статистически значимом уровне, что позволило бы поднять специфичность обнаружения кодирующих районов ДНК. Однако на этом пути существует несколько проблем. Во-первых, мутационные изменения последовательностей оснований нуклеиновых кислот включают не только замены оснований, но также вставки и делеции символов, из-за чего возможности обнаружения триплетной периодичности в кодирующих последовательностях оснований ДНК уменьшаются (рис. 1). Во-вторых, различные нуклеотидные последовательности могут содержать антагонистические виды триплетной периодичности, которые после их объединения в одно обучающее множество (например, для настройки нейронной сети) обладают пониженной специфичностью выявления кодирующих районов (рис. 2).

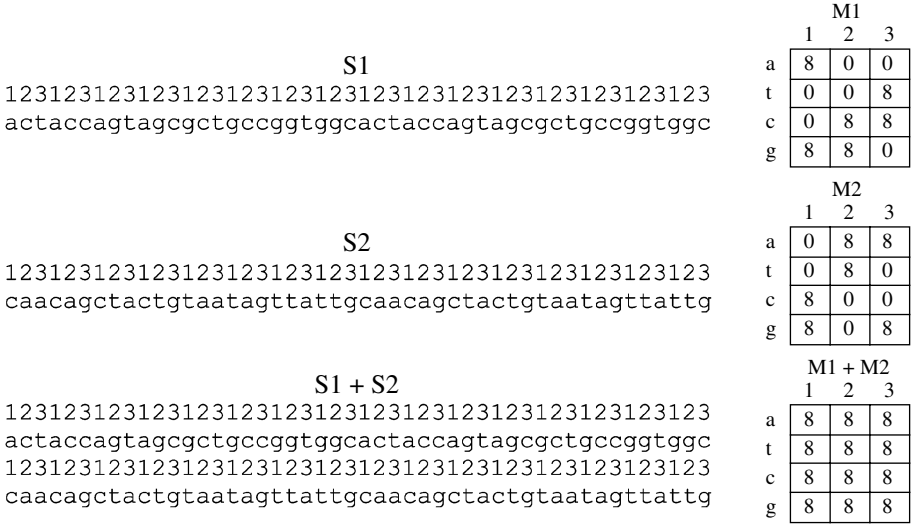

Рис. 2. Пример последовательностей оснований ДНК, обладающих антагонистической триплетной периодичностью. Под антагонистической понимается такая периодичность, которая исчезает при объединении матриц периодичности. Матрицы *M1* и *M2* построены для триплетной периодичности в последовательностях S1 и S2. При использовании последовательностей S1 и S2 в обучающей выборке их триплетные периодичности будут "уничтожать" друг друга. Это можно заметить, если объединить матрицы *M1* и *M2*. После такого объединения (матрица *M3*) триплетная периодичность исчезает.

Этих проблем можно избежать и увеличить мощность триплетной периодичности как кодирующего потенциала, если выделить классы триплетной периодичности, которые будут объединять гены, обладающие близкородственной триплетной периодичностью. В этом случае можно проводить поиск кодирующих последовательностей методами профильного анализа [26], используя в качестве профиля созданные классы триплетной периодичности. На этом пути возможен поиск триплетной периодичности со вставками и делециями нуклеотидов методом профильного анализа, а также изучение связи класса периодичности с рамками считывания, которые наблюдаются у последовательностей, вошедших в класс периодичности. Кроме того, если триплетная периодичность будет иметь корреляцию с рамкой считывания, то это позволит выявлять в гене районы, в которых произошел сдвиг рамки считывания относительно триплетной периодичности, или районы, где была осуществлена инверсия последовательности оснований ДНК. Такие возможности связаны с тем, что периодичность не может исчезнуть в отдельных эволюционных событиях, таких как делеции и вставки нуклеотидов или же инверсии последовательностей оснований ДНК [25].

При создании классов триплетной периодичности подходит метод информационного разложения [24, 25], который позволяет ввести понятие класса триплетной периодичности в виде матрицы размером 3×4. В матрице признаками столбцов являются позиции периода, а признаками строк – нуклеотиды [25]. Введение матриц упрощает объединение похожей триплетной периодичности в классы и поиск сдвигов между рамкой считывания и триплетной периодичностью какого-либо класса, а также поиск триплетной периодичности в инвертированном виде.

В настоящей работе решали три задачи. Во-первых, мы хотели методом информационного разложения найти все кодирующие последовательности, обладающие триплетной периодичностью в генах из банка данных KEGG (http://www.genome.ad.jp/kegg/) версии 29 (более 4×10^5 генов) и затем объединить в классы близкородственные матрицы триплетной периодичности. Поэтому для каждого анализируемого гена построили соответствующую матрицу триплетной периодичности, привязанную к существующей в данном районе ДНК рамке считывания. Это означает, что первый столбец в матрице триплетной периодичности соответствовал первому основанию рамки считывания, существующей в районе ДНК с триплетной периодичностью. В итоге мы создали 2520 классов триплетной периодичности путем объединения близких матриц с учетом возможности циклических сдвигов и инверсий. В эти классы входит 94% всех найденных участков триплетной периодичности в генах. После объединения для каждого класса составлены список матриц, входящих в данный класс, список позиций рамки считывания с наличием (или отсутствием) инверсии, с которой каждая матрица вошла в этот список. Этот список для каждого класса всегда содержал одну рамку считывания, которую можно считать доминирующей для данного класса, так как туда входило более 50% всех матриц триплетной периодичности, объединенных данным классом. Мы назвали эту

рамку считывания рамкой считывания класса триплетной периодичности. В каждом классе триплетной периодичности мы выявляли те гены, где триплетная периодичность класса сдвинута относительно рамки считывания класса, а также гены, где триплетную периодичность наблюдали только для инвертированной последовательности оснований ДНК. Во-вторых, мы хотели проверить, действительно ли гипотетические аминокислотные последовательности, полученные при использовании рамки считывания класса триплетной периодичности, имеют гомологию с последовательностями из банка данных UniProt (http://www.ebi.uniprot.org/). Такую проверку мы делали для генов, у которых наблюдали несовпадение между рамкой считывания данного гена и рамкой считывания триплетного класса. Это может свидетельствовать о сдвиге рамки считывания в анализируемом гене, произошедшем в ходе эволюции гена. Мы подтвердили существование таких сдвигов, так как нашли гомологию между гипотетическими аминокислотными последовательностями и аминокислотными последовательностями из банка данных UniProt. В-третьих, мы хотели проверить возможность существования в белках аминокислотных последовательностей, созданных посредством инверсии фрагмента ДНК. Мы отобрали гены, в которых триплетная периодичность класса наблюдалась только в инвертированном виде. Также определили гипотетические аминокислотные последовательности этих инвертированных последовательностей генов. В данной работе мы показали, что многие из таких гипотетических аминокислотных последовательностей имеют высокую гомологию с аминокислотными последовательностями из банка данных UniProt. В целом, это доказывает, что сдвиги рамки считывания и инверсии нуклеотидных последовательностей достаточно широко распространены в генах, и классы триплетной периодичности могут быть полезны в идентификации подобных событий.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Поиск триплетной периодичности в генах

Триплетную периодичность выявляли в кодирующих последовательностях (CDS) генов, накопленных в выпуске 29 банка данных KEGG, в котором представлены как последовательности эукариотических, так и прокариотических генов. В случае эукариотических генов кодирующие последовательности объединяли все экзоны. Поиск периодичности проводили при помощи метода информационного разложения [24, 25]. С этой целью последовательность оснований ДНК $A(n) = \{a(1)a(2), \ldots, a(n)\}$ каждого гена сравнивали с равной по длине искусственной периодической последовательностью вида: $S(3) = \{s(1)s(2)s(3)s(1)s(2)s(3), \ldots, s(1)s(2)s(3)\}$, где $s(1) \equiv '1'$, $s(2) \equiv '2'$, $s(3) \equiv '3'$. В данной последовательности символы рассматривали как числа. При сравнении последовательностей заполняли матрицу совпадений $M(3 \times 4)$. У этой матрицы признаками столбцов являются символы *'1', '2'* и *'3'*, а признаками строк – символы последовательности оснований ДНК $w(1) \equiv 'a'$, $w(2) \equiv 't'$, $w(3) \equiv 'c'$, $w(4) \equiv 'g'$. Элемент матрицы $m(i,j)$ показывает число совпадений символов $w(i)s(j)$ у двух сравниваемых последовательностей. При заполнении матрицы M первое основание первого кодона всегда соответствовало символу $s(1)$ искусственной периодической последовательности. После заполнения матрицы M рассчитывали взаимную информацию по формуле [27]:

$$I = \sum_{i=1}^{3}\sum_{j=1}^{4} m(i,j)\ln m(i,j) - \sum_{1}^{3} x(j)\ln x(j) - \sum_{1}^{4} y(j)\ln y(j) + n\ln n, \quad (1)$$

где n – длина изучаемой символьной последовательности, $x(i), I = 1, 2, 3$ есть число символов *'1'*, '2' и '3' в искусственной символьной последовательности $S(3)$ (для введенной выше периодической последовательности $x(i) = n/3, i = 1, 2, 3$); $y(j), j = 1, 2, 3, 4$ – число символов $w(j)$ в изучаемой символьной последовательности. После вычисления взаимной информации мы смогли оценить вероятность случайного подобия между последовательностью $S(3)$ и $A(n)$. При этом использован метод Монте-Карло [28] и величина Z, которая рассчитывается как:

$$Z = (I - \bar{I})/\sqrt{D(I)}, \quad (2)$$

где $\bar{I}$ и $D(I)$ показывают среднее значение и дисперсию величины взаимной информации для множества случайных матриц с такими же суммами $x(i)$ и $y(j)$, как и в исходной матрице M. Величина Z имеет распределение, близкое к нормальному, что мы и проверили, сравнив искусственную периодическую последовательность с множеством случайных последовательностей объемом 10^7 символов. Это позволило использовать Z как меру подобия искусственной периодической последовательности и последовательности оснований ДНК. Чем выше значение Z, тем выше подобие последовательностей S и A, более явно выражена периодичность в последовательности A. Кроме того, матрицу M удобно использовать для представления вида периодичности, наблюдаемой в последовательности A. Дело в том, что одни и те же значения Z могут получаться при различных матрицах M.

Мы искали в последовательности A район с максимально выраженной триплетной периодичностью, которую мы назовем максимальной последовательностью. С этой целью рассматривали все возможные позиции левой и правой границ для последовательности $A(i)$, $n \geq i \geq 30$. При этом мы искали такую последовательность, для которой значе-

Таблица 1. Преобразование матрицы триплетной периодичности для различных индексов сдвига относительно центральной матрицы класса

Матрица *M* после сдвига

Индекс сдвига: 1

	1	2	3
a	m_{11}	m_{21}	m_{31}
t	m_{12}	m_{22}	m_{32}
c	m_{13}	m_{23}	m_{33}
g	m_{14}	m_{24}	m_{34}

Индекс сдвига: 2

	1	2	3
a	m_{31}	m_{11}	m_{21}
t	m_{32}	m_{12}	m_{22}
c	m_{33}	m_{13}	m_{23}
g	m_{34}	m_{14}	m_{24}

Индекс сдвига: 3

	1	2	3
a	m_{21}	m_{11}	m_{31}
t	m_{22}	m_{32}	m_{12}
c	m_{23}	m_{33}	m_{13}
g	m_{24}	m_{34}	m_{14}

Индекс сдвига: 4

	1	2	3
a	m_{32}	m_{22}	m_{12}
t	m_{31}	m_{21}	m_{11}
c	m_{34}	m_{24}	m_{14}
g	m_{33}	m_{23}	m_{13}

Индекс сдвига: 5

	1	2	3
a	m_{12}	m_{32}	m_{22}
t	m_{11}	m_{31}	m_{21}
c	m_{14}	m_{34}	m_{24}
g	m_{13}	m_{33}	m_{23}

Индекс сдвига: 6

	1	2	3
a	m_{22}	m_{12}	m_{32}
t	m_{21}	m_{11}	m_{31}
c	m_{24}	m_{14}	m_{34}
g	m_{23}	m_{13}	m_{33}

ние *Z* достигало бы максимума. Последовательности *A* и *S* не менялись и не сдвигались относительно друг друга, при поиске максимальной последовательности сдвигались лишь начальные и конечные координаты последовательности *A* при заполнении матрицы *M*. Сдвиг позиций левой и правой границ происходил на длину, кратную трем основаниям. Это означает, что первая позиция матрицы *M* у максимальной последовательности оснований ДНК всегда соответствовала первому основанию кодона и рамки считывания. Индекс сдвига у матрицы *M* (табл. 1) считали равным 1. В таблице показаны перемещения элементов матрицы триплетной периодичности *M* при шести вариантах сдвигов, с которыми она может войти в класс. Индекс сдвига, равный 1, отражает полное соответствие двух матриц, т.е. отсутствие преобразования матрицы *M*. Это также означает, что рамки считывания в последовательностях ДНК, где были найдены матрицы *M* и центральная матрицы класса (далее ЦМК), совпадают. Индексы сдвига, равные 2 и 3, соответствуют циклическому сдвигу матрицы после преобразования на 1 и 2 основания. Индекс сдвига, равный 4, соответствует инверсии матрицы *M* относительно ЦМК. Это означает, что в матрице *M* строки, соответствующие основаниям *a* и *g*, *t* и *c* меняются местами. После этого столбцы 1 и 3 в матрице также меняются местами. Это преобразование обозначено звездочкой. Индексы сдвига, равные 5 или 6, показывают соответствие матриц, аналогичное соответствию матриц с индексом сдвига равному 4, но с циклическим сдвигом столбцов на 1 или 2 основания.

Если для максимальной последовательности *A* значение *Z* было большим, чем 5.0, то мы считали, что нашли район с триплетной периодичностью. Значение *Z*, большее, чем 5.0, обеспечивает вероятность случайного обнаружения триплетной периодичности в последовательности оснований ДНК менее 10^{-6}. После этого мы запоминали найденную максимальную для данного гена последовательность, ее координаты в данном гене и матрицу периодичности *M*, которая показывает тип найденной триплетной периодичности. Найденные последовательности с сопутствующей информацией сохраняли для дальнейшего создания классов триплетной периодичности.

Мы выбрали пороговый уровень $Z > 5.0$ для поиска триплетной периодичности для того, чтобы число случайно найденных триплетных периодичностей было около 1% от всех районов с триплетной периодичностью, найденных в генах из банка данных KEGG-29. При выборе порогового значения *Z* мы генерировали выборку случайных последовательностей ДНК такого же объема и с таким же распределением длин последовательностей, как у генов из банка данных KEGG. Для $Z > 5.0$ число найденных

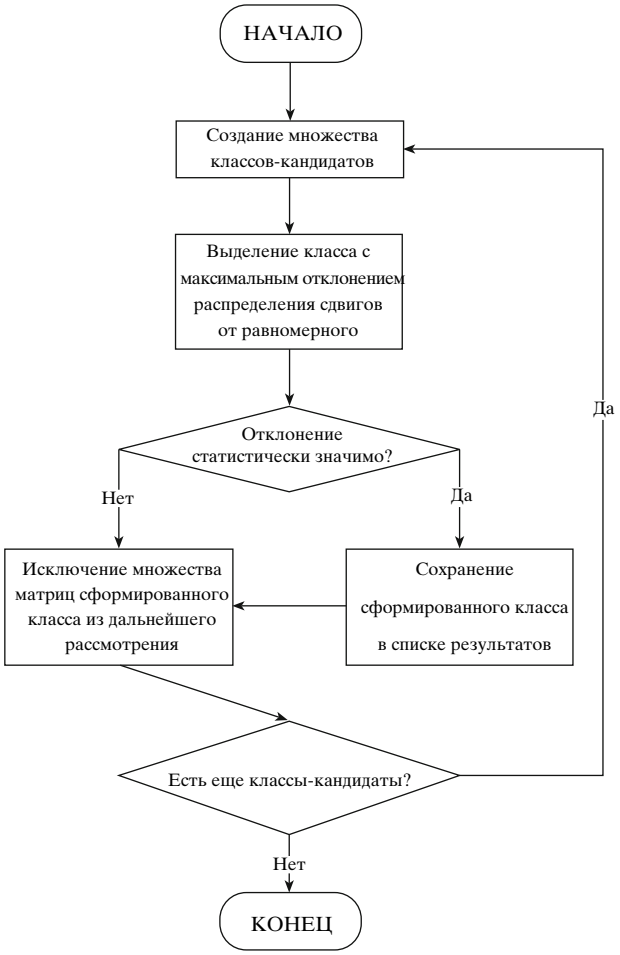

Рис. 3. Алгоритм классификации матриц триплетной периодичности.

последовательностей составляло 7200, т.е. приблизительно 1.5% от найденных участков с триплетной периодичностью (см. ниже). При $Z > 6.0$ таких последовательностей мы нашли 172, а при $Z > 7.0$ таких последовательностей не найдено. Мы сознательно выбрали уровень $Z > 5.0$, чтобы наиболее полно сформировать значимые классы триплетной периодичности, которые присутствуют в различных генах. Мы проверяли устойчивость выделения значимых классов (см. пункт 2.2) путем добавления случайных матриц с триплетной периодичностью в интервале $5.0 < Z < 7.0$ во множество найденных матриц триплетной периодичности. Добавление до 2% случайных матриц не меняет значимые классы триплетной периодичности.

В случае нахождения в гене района с триплетной периодичностью, он исключался из дальнейшего рассмотрения. Если в гене оставались фрагменты, не включенные в найденную максимальную последовательность, то они поступали на повторное рассмотрение с целью дальнейшего поиска последовательностей генов с триплетной периодичностью. Подобное рассмотрение сделано для того, чтобы найти гены, где присутствует две и более последовательности триплетной периодичности, в том числе и с различными матрицами *M*.

Классификация триплетной периодичности

Классификацию матриц триплетной периодичности проводили последовательным формированием классов. Процесс классификации состоял из множества циклически повторяющихся шагов (далее называемых итерациями), на каждом из которых формировался один класс триплетной периодичности. В свою очередь, каждая итерация проводилась в два этапа, описанных ниже.

При каждой итерации на первом этапе объединяли матрицы триплетной периодичности в классы-кандидаты по мере их подобия друг другу с учетом возможности 3-х циклических сдвигов и такого же количества циклических сдвигов в случае инверсии. Одновременно для каждого класса-кандидата мы создавали множество индексов сдвига, с которыми матрицы вошли в этот класс. Затем в каждом классе-кандидате наиболее представленному в нем индексу сдвига присваивали значение, равное 1, а остальные индексы сдвигов циклически менялись.

Например, если наибольшее представительство имел индекс сдвига относительно центральной матрицы, равный 2, то ему присваивали индекс сдвига, равный 1. В этом случае индекс сдвига 1 заменяли на 3, 3 на 2, 4 на 5, 5 на 6, 6 на 4 (см. табл. 1). Рассмотрим также пример, когда доминирующим был индекс сдвига 4 (инверсия без циклического сдвига). При этом мы заменили индекс сдвига 1 на 4, 2 на 5, 3 на 6, 4 на 1, 5 на 2, 6 на 3. В итоге, наиболее представленный в классе-кандидате индекс сдвига относительно центральной матрицы всегда имел значение, равное 1.

На втором этапе итерации формирования класса мы выбирали среди созданных классов-кандидатов наиболее "неслучайный" класс на основе степени однородности множества индексов сдвига, которое было получено для каждого класса-кандидата. Это означает, что классы-кандидаты, где количество каждого индекса сдвига (всего возможно от 1 до 6, см. табл. 1) примерно одинаково, признавали случайными. После этого матрицы, включенные в наиболее "неслучайный" класс-кандидат по индексу сдвига, исключали из рассмотрения, и проводили следующую итерацию для формирования нового класса. Схема построения классов показана на рис. 3. Рассмотрим подробно каждый из этапов итерации.

На первом этапе очередной итерации классификации для каждой матрицы триплетной периодичности мы определяли множество матриц, которые ей подобны. Такую матрицу будем называть матрицей центра класса, а само множество матриц – классом-кандидатом. В качестве меры различия матриц мы

использовали описанный ниже критерий согласия. При сравнении матриц учитывали все возможные циклические сдвиги и инверсии последовательности оснований ДНК, т.е. для пары матриц делали 6 сравнений. При каждом сравнении центральную матрицу фиксировали, а преобразовывали только ту матрицу, которую с ней сравнивали. Из этих 6 пар матриц выбирали именно ту пару, которая давала наименьшее различие. Для каждого класса-кандидата создавали список вошедших в него матриц триплетной периодичности с индексами сдвига матриц (от 1 до 6), соответствующий виду преобразования матрицы. Соответствие индексов сдвига и преобразования матрицы показано в табл. 1. Процесс создания классов-кандидатов проводился нами с каждой из найденных матриц триплетной периодичности в качестве центральной матрицы класса. На втором этапе итерации, выбирали класс-кандидат, имеющий наибольшее отклонение распределения сдвигов матриц внутри него от равномерного. Все матрицы, вошедшие в этот класс, исключали из дальнейшего рассмотрения, после чего формирование следующего класса проводили заново для оставшихся матриц. Процесс классификации продолжали, пока можно было выделить классы-кандидаты, содержащие хотя бы две матрицы (матрицу центра класса и еще одну). В результате такой классификации мы получили множество классов матриц триплетной периодичности и множество матриц, которые не входят в какой либо класс. В каждом классе также сформировано множество индексов сдвигов для матриц триплетной периодичности.

Далее рассмотрим процесс сравнения двух матриц и критерий различия двух матриц. В качестве меры различия двух матриц M^1 и M^2 использована мера W [29], определенная ниже как:

$$W = \sum_i \sum_j t_{ij}. \qquad (3)$$

Определим матрицу $T = \{t_{ij}\}$ как:

$$t_{ij} = \frac{m\left(\frac{m_{ij}^1}{y_j^1} - \frac{m_{ij}^2}{y_j^2}\right)}{\sqrt{p(1-p)\left(\frac{1}{y_j^1} + \frac{1}{y_j^2}\right)}}, \qquad (4)$$

где $y_j^k = \sum_j m_{ij}^k$, $p = 1/3$. Величина t_{ij} имеет близкое к нормальному распределение. Значение W имеет распределение χ^2 с 8 степенями свободы.

При определении класса-кандидата вводили пороговое значение величины $W = W_0 = 3.44$. Мы считали, что матрица принадлежит классу-кандидату, если величина W, рассчитанная по формуле (3), меньше порогового значения W_0. Значение W_0 обеспечивает вероятность случайного объединения матриц в класс, равную 8.22×10^{-4}. Это значение получено методом Монте-Карло путем генерации случайных матриц и отбора среди них матриц с $Z > 5.0$ (см. формулу (2)).

Выбор значения $W_0 = 3.44$ связан с двумя факторами. Во-первых, мы стремились объединить в классах как можно больше матриц М, и насколько можно меньше матриц оставить вне классов. Одновременно с этим хотелось бы, чтобы введенные нами классы были максимально представительными, а число самих классов было бы сравнительно небольшим. Но с другой стороны, желательно, чтобы максимально большое число матриц было объединено в неслучайные классы по однородности индексов сдвига, имеющихся в данном классе. Проведенная нами классификация матриц триплетной периодичности при значениях W_0 выше 3.44 показала, что число матриц, входящих в неслучайные классы по индексу сдвига уменьшается с увеличением W_0. В этом смысле значение $W_0 = 3.44$ оказалось оптимальным, так как оно обеспечивает максимум матриц триплетной периодичности, входящих в неслучайные классы при сравнительно небольшом количестве получающихся классов.

На втором этапе классификации матриц мы отбирали среди найденных классов-кандидатов максимально однородные по индексу сдвигов. С этой целью использовали информационный критерий. Пусть $x_1, x_2.., x_6$ показывают, сколько присутствует в классе индексов сдвига 1, 2, …, 6 при их суммарном количестве, равном N, т.е. $\sum_{i=1}^{6} x_i = N$.

При проверке однородности индексов сдвига мы проверяем гипотезу, что выборка принадлежит полиномиальной популяции $p_1, \ldots, p_6$, $\sum_{i=1}^{6} p_i = 1$, где $p_i = 1/6$. Различающая информация рассчитывается как [27]:

$$I = \sum_{i=1}^{6} x_i \log \frac{x_i}{N p_i} = \sum_{i=1}^{6} x_i \log x_i - \sum_{i=1}^{6} x_i \log p_i - N \log N. \qquad (5)$$

Величина $2I$ имеет распределение χ^2 с 5 степенями свободы [27]. Значение $2I$ равно нулю, если вероятности $f_i = \frac{x_i}{N}$ равны по своему значению вероятностям p_i и значение $2I$ принимает максимальное значение, если значение f_i. для какого-либо i равно 1.0, а остальные значения f_i. равны нулю. Это означает, что максимально неоднородное распределение индексов сдвига в классе дает максимум значения $2I$. В результате среди созданных классов мы отбирали класс, имеющий максимальное значение $2I$. После этого процесс повторялся. Такое повторение происходит до тех пор, пока будет находиться класс, име-

ющий величину $2I$ больше некоторого критического значения $2I_0$. Значение $2I_0$ выбирали таким, чтобы в результате классификации каждого множества случайных матриц сформировать, в среднем, не более одного класса. Это обеспечивается выбором значения $2I_0$, равного 40.5. Мы определили число случайных классов, которые возникают при таком выборе порогового значения $2I_0$ при классификации случайных матриц. С этой целью мы генерировали 30 множеств случайных матриц триплетной периодичности со значением $Z > 5.0$ такого же объема, как и множество матриц, полученное нами при анализе банка данных KEGG. Это означает, что приблизительно только одну из 10^6 сгенерированных случайных матриц отбирали для вхождения в эти множества. Затем вышеописанная классификация из двух этапов проведена для каждого из 30 множеств. Результаты показали, что для этих 30 множеств случайных матриц удалось сформировать только $2I$ класс периодичности, где значение величины $2I > 2I_0$. Это означает, что при такой классификации в среднем мы можем генерировать ~0.7 случайного класса на множество. В этих сформированных случайных классах содержится всего 0.015% от всех случайных матриц, подвергнутых классификации.

Анализ неоднородности распределения индексов сдвига в незначимых классах

Можно ожидать, что не все созданные классы будут обладать $2I > 40.5$. Существуют две причины для такого явления. Во-первых, класс может иметь крайне неоднородное распределение по индексам сдвигов, но значение N (формула (5)) может быть невелико (мало матриц в классе). Во-вторых, распределение индексов сдвигов (x_i в формуле (5)) может быть близко к равномерному. Для разделения этих двух возможностей необходимо ввести меру, которая отражала бы степень неоднородности значений x_i и не зависела от объема выборки N. Для значений $p_i = 1/6$ формулу (5) можно переписать как:

$$I = \sum_{i=1}^{6} x_i \log 6 x_i - N \log N. \quad (6)$$

Если x_i также будут равномерно распределены ($x_i = N/6$), то $I = 0.0$. Максимальное значение взаимной информации наблюдали, когда какое-либо одно значение из x_i будет равно N, а оставшиеся пять значений x_i равны нулю. В этом случае:

$$I_{\max} = N \log 6. \quad (7)$$

$$\frac{I}{I_{\max}} = 1 + \frac{\sum_{i=1}^{6} f_i \log f_i}{\log 6}. \quad (8)$$

Отношение величин не будет зависеть от объема выборки N, где $f_i = x_i/N$. Поэтому для решения вопроса о распределении индексов сдвига в незначимых классах удобно использовать значение $I/I_{\max}$, которое не зависит от объема выборки.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Проанализировано 578 868 генов, накопленных в банке данных KEGG версии 29 (http://www.genome.ad.jp/kegg/). Общее число участков триплетной периодичности в 457333 генах – 472288. Покрыты районами триплетной периодичности более чем на 90% их длины 285190 генов, что составляет 62% от общего числа генов. В 12697 генах найдено два и более участка периодичности. Эти данные показывают, что более чем 79% генов имеют районы с триплетной периодичностью. Результаты согласуются с более ранними работами по обнаружению триплетной периодичности (информационными методами и другими подходами [19–21, 30]). Для каждой последовательности ДНК с триплетной периодичностью рассчитывали соответствующую матрицу совпадений символов M, и затем эти матрицы объединяли в классы. При классификации мы решаем две задачи. Во-первых, мы можем выяснить, насколько разнообразны матрицы триплетной периодичности. Во-вторых, мы можем проверить, существуют ли участки генов, для которых матрицы триплетной периодичности сходны с точностью до сдвига рамки и/или перехода к комплементарной цепи, и насколько равномерно распределены “фазы” матриц внутри класса эквивалентности. Такое исследование может помочь при поиске генов, образовавшихся в результате инверсии или сдвига рамки считывания. В процессе поиска триплетной периодичности методом информационного разложения все первые позиции триплетной матрицы соответствовали первым позициям кодона, и они могли меняться только в ходе объединения матриц в классы, когда был возможен циклический сдвиг или инверсия матрицы. Чтобы изучить эту связь одновременно с матрицей класса триплетной периодичности создавали множество, где содержались все индексы сдвига матриц, вошедших в данный класс. В результате классификации получены 2520 классов (http://victoria.biengi.ac.ru/ancorfs/classes.php). Классы матриц периодичности имеют большой разброс в своих размерах, от 1 до десятков тысяч (рис. 4). В классах с $2I > 40.5$ (формула (5)) содержится 443798 случаев периодичности из 472288 найденных матриц триплетной периодичности. И 8591 матриц не вошли ни в один класс, остались автономными, а 19899 матриц вошли в незначимые классы, для которых $2I \le 40.5$. Как видно, около 94% матриц входят в значимые классы, что показывает существование связи между триплетной периодичностью и рамкой считывания. Классы, имеющие $2I \le 40.5$, могут содержать небольшое количество матриц в сво-

ем составе, и именно малое количество матриц не позволило этим классам получить $2I > 40.5$, тогда как распределение по индексам сдвига в этих классах может быть неоднородным. С целью проверки этой гипотезы мы построили распределение величины I/I_{max} для значимых и незначимых классов (рис. 5*а*,*б*). Видно, что для основной части классов значение I/I_{max} находится в интервале от 0.5 до 1.0. В случае незначимых классов их основное количество лежит в этом же интервале, причем около четырех тысяч классов имеют I/I_{max} близкое или равное единице. Следовательно, и в случае незначимых классов наблюдается связь между рамкой считывания и классом триплетной периодичности. Общий вывод из проделанной классификации состоит в том, что между классом триплетной периодичности и рамкой считывания в гене наблюдается существенная связь. При анализе вхождения матриц периодичности в классы выявлено, что только около 8% от них входят в класс с каким-либо сдвигом рамки считывания или инверсией направления считывания. Это означает, что из 443798 выявленных и входящих в значимые классы случаев периодичности только 36111 имели рамку считывания, отличную от характерной для большинства вошедших в класс матриц.

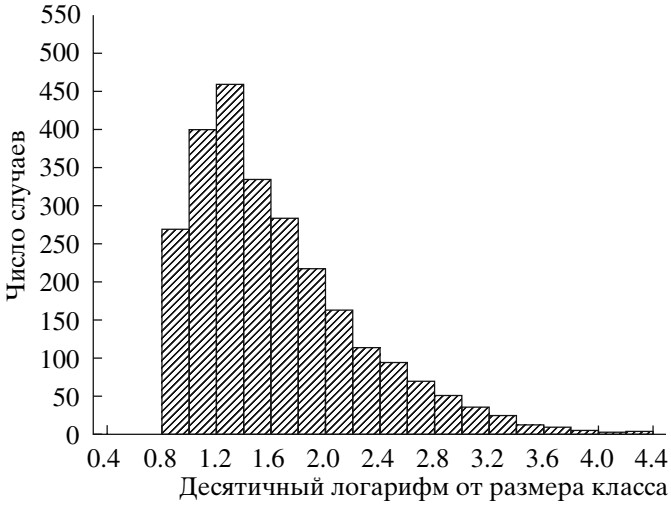

Рис. 4. Распределение классов триплетной периодичности по числу матриц в них. По логарифмической оси X показан размер класса, по оси Y показано число матриц в классе. Больше всего имеется классов с размером от 1 до 50 матриц.

Эти данные позволяют для каждого класса ввести главную рамку считывания, которая присутствует в данном классе (рамка класса). Как ясно из вышеприведенных данных, такой рамкой для всех классов является рамка считывания без сдвига и инверсии (индекс сдвига для матриц равен 1). Далее мы рассмотрели общее количество матриц, вошедших в значимые классы с индексами сдвига 2–6. (табл. 2). Данные таблицы свидетельствуют: наибольшее число матриц с рамкой считывания, отличной от доминирующей в классе, входит в него с индексом сдвига, равным 4. Это соответствует инверсии матрицы без циклического сдвига, т.е. инверсии кодирующей последовательности оснований ДНК с наложением рамок считывания без сдвига (табл. 1). На втором месте по численности стоит случай инверсии матрицы, но с наложением кодонов со сдвигом. Далее мы наблюдаем подобие матриц только с циклическим сдвигом (индексы сдвига 3 и 2), и в конце по численности стоит подобие матриц с инверсией и со сдвигом на 2 основания.

Мы также проверили матрицы периодичности на наличие в них симметричности, что может быть

Таблица 2. Распределение периодичностей по сдвигу относительно рамки считывания класса (для значимых классов)

Сдвиг периодичности относительно рамки считывания класса	Число матриц	Индексы сдвига матрицы
Рамка считывания 1	407687	1
Рамка считывания 2	2558	2
Рамка считывания 3	3162	3
Всего прямых со сдвигом рамки	5720	2 + 3
Всего прямых	413407	1 + 2 + 3
Инверсия, рамка считывания 1	20199	4
Инверсия, рамка считывания 2	7616	5
Инверсия, рамка считывания 3	2576	6
Всего антисмысловых со сдвигом рамка	10192	5 + 6
Всего инвертированных последовательностей	30391	4 + 5 + 6
Всего со сдвигами или инверсией	36111	2 + 3 + 4 + 5 + 6
Всего участков периодичности	443798	1 + 2 + 3 + 4 + 5 + 6

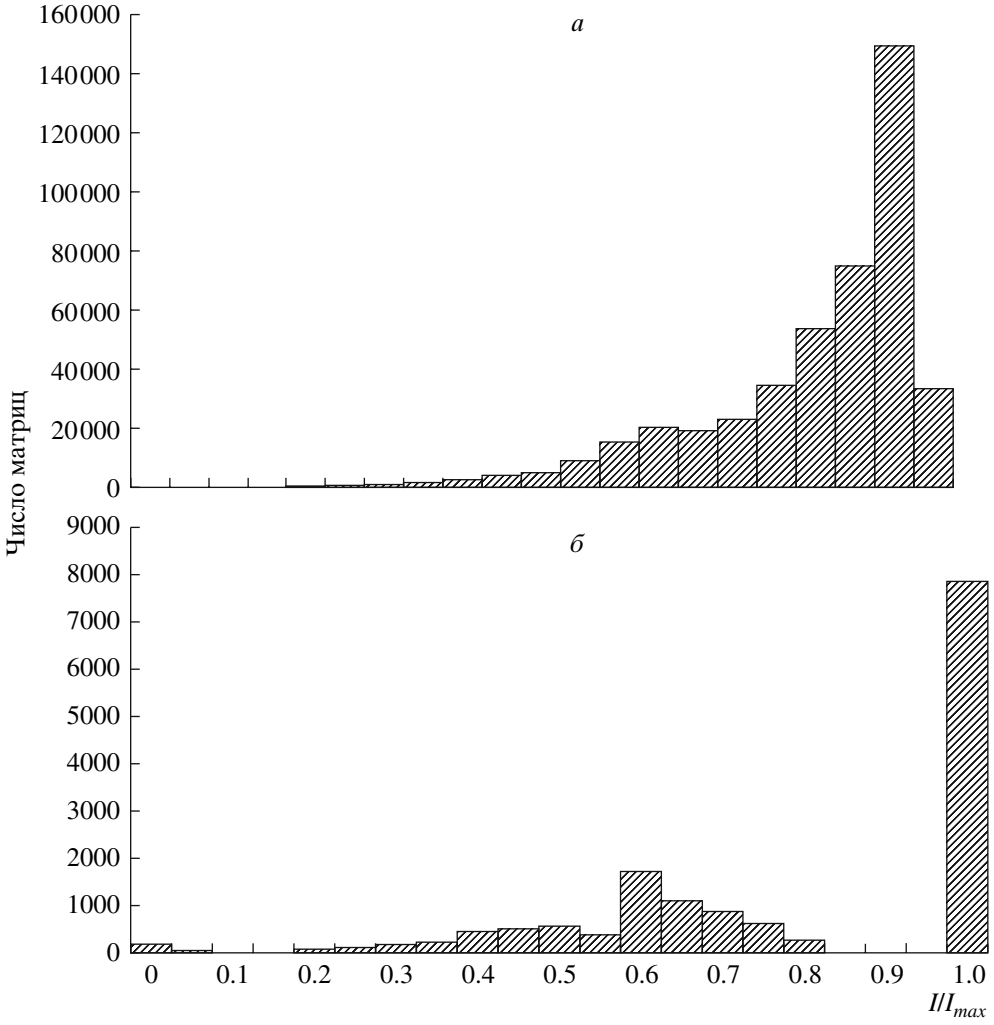

Рис. 5. Распределение значения $I/I_{\max}$ для значимых (*а*) и незначимых (*б*) классов.

причиной найденных сдвигов в классах, для этого мы сравнивали каждую матрицу саму с собой после проведения в ней циклических сдвигов и инверсий. Обнаружено, что у 0.4% матриц, вошедших в классы, наблюдается подобие к одному из их (пяти) вариантов, полученных сдвигом и инверсией. Оценка подобия производилась по тому же критерию сходства матриц периодичности, что и при классификации. Хотя симметричность и наблюдается у некоторых матриц, но существенной роли в формировании сдвигов между матрицами внутри классов, в целом, не играет (0.4% симметричных матриц против 8% матриц со сдвигом).

Данные табл. 2 показывают, что в кодирующих последовательностях ДНК происходят сдвиги рамки считывания и инверсии последовательности ДНК. В силу того, что триплетную периодичность достаточно трудно изменить отдельными мутациями, мы можем наблюдать ее как след существовавшей ранее рамки считывания, измененной в результате делеций, вставок или же инверсий последовательностей ДНК входящих в состав генов [25]. Например, в 6892 генах выявлены два района периодичности, вошедшие в один класс с одинаковыми сдвигами, а для 42 генов оба их района периодичности вошли в один класс, но с различными сдвигами, что указывает на возможные вставки или делеции символов, приведшие к сдвигу рамки считывания внутри этих генов. Однако данная гипотеза нуждается в дополнительной проверке. Такую проверку можно провести, если перекодировать нуклеотидную последовательность, в которой мы обнаружили сдвиги рамок считывания, в аминокислотную по рамке считывания, доминирующей в классе, в который она вошла. То же самое мы можем сделать для последовательностей, где получена матрица триплетной периодичности с индексами сдвига 4, 5 и 6. При этом мы должны также добавить процедуру инвертирования последовательности оснований

Таблица 3. Число подобий для различных случаев сдвига ОРС (значимые классы)

Число подобий					
Вид трансформации кодирующей последовательности	по обеим ОРС	только по ОРС класса	только по ОРС KEGG	индекс сдвига	доля от общего числа периодических последовательностей
Рамка считывания 2	8	868	346	2	0.48
Рамка считывания 3	143	784	558	3	0.47
Антисмысловой, рамка считывания 1	25	443	12991	4	0.67
Антисмысловой, рамка считывания 2	13	355	3236	5	0.47
Антисмысловой, рамка считывания 3	1	39	923	6	0.37
Всего	**190**	**2489**	**18054**		**0.57**

ДНК, т.е. ее переворот на 180° и замену оснований последовательности на комплементарные. После такого преобразования мы получаем гипотетическую аминокислотную последовательность, которая могла бы быть в гене до процесса сдвига рамки считывания или инверсии последовательности (рамка класса). Если мы сможем показать, что у гипотетической аминокислотной последовательности есть гомологичные последовательности в банке данных UniProt, то это будет доказывать, что процессы сдвига рамки считывания или инверсии реально происходили с фрагментом ДНК, для которого матрица триплетной периодичности входила в соответствующий класс с индексами сдвига, отличными от 1. При этом данный факт можно считать очень вероятным, если аминокислотная последовательность, полученная по рамке считывания, представленной в базе данных KEGG (рамка KEGG), также имеет аминокислотные подобия.

Чтобы проверить эту гипотезу, нами проведен поиск гомологов белковых продуктов, кодируемых найденными районами генов. Кодирование проводили как по рамке KEGG, так и по рамке класса. Поиск гомологов осуществляли по банку данных UniProt с помощью программы BLAST. Среди найденных подобий рассматривали только значимые случаи, имеющие вероятность случайного совпадения (e-value) менее 5%. Каждая изучаемая последовательность ДНК, имеющая индекс сдвига от 2 до 6, была перекодирована в пару аминокислотных последовательностей в соответствии с реально существующей и гипотетической (введенной на основе индекса сдвига) рамками считывания. В результате проделанного сканирования получены списки гомологов для 20 733 участков ДНК с триплетной периодичностью. В случае 15 378 участков периодичности ни одного значимого подобия к их белковым продуктам по обеим рамкам считывания найдено не было. В 190 аминокислотных последовательностях наблюдали подобие, как для гипотетической, так и для реальной аминокислотной последовательности (табл. 3).

Пример найденного подобия гипотетической аминокислотной последовательности, полученной при использовании рамки класса, показан на рис. 6 и доступен по адресу http://victoria.biengi.ac.ru/ancorfs/perinfo.php?perid = 355386. Такое подобие найдено для гена с идентификатором “7561” из банка данных KEGG и аминокислотной последовательности из банка данных UniProt с идентификатором “Q8N782” [31]. Ген “7561” имеет длину 1 929 нуклеотидов, триплетная периодичность наиболее выражена в последовательности оснований с 157 по 1911 нуклеотид. Матрица триплетной периодичности этого участка ДНК вошла в соответствующий класс триплетной периодичности с индексом сдвига, равным 2 (рис. 7), что соответствует второй рамке считывания (первая рамка считывания соответствует кодировке аминокислот в данном гене). Класс триплетной периодичности (рис. 8) объединил 2246 матриц триплетной периодичности, причем в данном классе 96.7% (2 172 из 2 246) матриц имели индекс сдвига, равный 1.

По данному гену для рамок считывания 1 и 2 (рис. 7) построены аминокислотные последовательности, и проанализировано подобие этих аминокислотных последовательностей банку данных UniProt программой Blast. В результате такого исследования был найден 31 случай подобия белковой последовательности, кодированной по рамке класса (гипотетическая аминокислотная последовательность), и 1427 случаев подобия при кодировании белка по рамке KEGG. Наилучшее подобие по оригинальной рамке считывания найдено для последовательности “P17017” [32] (подобие с 53 по 637 аминокислоту), где наблюдали полное подобие (100%) между аминокислотными последовательностями (рис. 6). Таким образом, данные табл. 3 показывают, что только для небольшого количества генов, для которых триплетная периодичность вошла в

```
224 KNPTNVKNVVKPSVFSVLFEGIKGLIVERNPMNVKNVEKPSFILQAFEHMXXYTLGLDLI 283
    + P  +  V +PSV       I   I+ERN  +VKNV K S   Q  + +  + +  +
 98 EKPYKLMIVARPSVICQPLHAIVDFILERNLTSVKNVMKLSVSNQVLKDIGEFIMERNCT 157

284 NVKNVGKPSTLLIPVECMKELILEKNHMNVNDVANHSVGPFLFDCMKELILERNLMSVNS 343
    NV +V +PS    P+  + + I+E+N  +V +V   +V   +   +++ ILERN++SV
158 NVMSVARPSVRSYPLPAIVDFIVERNLASVENVTRLTVSNKILKYVRKFILERNVISVMI 217

344 VIKPSVFQVPFENTKQLTLERNPMNVNNVVKPSVFQVPFKDMKGLTMQRNPMNVNSVGKP 403
    V + SV + P        +ERN  NV NV+K SV     KD+   T+ RN   V  V +P
218 VARSSVIRHPLYTIINFIVERNLTNVKNVMKLSVSNQTLKDIGEFTLVRNLTGVIGVARP 277

404 SGVQVIFEFMKGHTLERNPMNVNSVEKFSFVPVPFDCMKEHTLERNPMNVNYAVKPSVFQ 463
    S +      +    LERN  NV +V K S         + E  +ERN  +V    +PSV
278 SVIHHPLHAIIDFILERNLTNVKNVMKLSDTNQILKDIGEFIMERNRTSVMSVARPSVRS 337

    VPFENMKKFTLEISLLSVSNVVRPS 488
         +  F LEI+L+SV +V RPS
    HALHAIIDFILEINLISVMSVARPS 362
```

Рис. 6. Подобие гипотетической аминокислотной последовательности, полученной по второй рамке (F2) считывания (224–488 а.о.), аминокислотной последовательности локуса ZN525_HUMAN базы данных UniProt (98–362 а.о.).

```
KEGG ORF:  [E] D  D  H  R  N  Q  G  K  N  R  R  C  H  M  V  E  R  L  C ...
F1. aggacatt[gaa]gatgaccacagaaaccaggggaaaaatcgaagatgtcatatggttgagagactctgt...
F2. aggacattg[aag]atgaccacagaaaccaggggaaaaatcgaagatgtcatatggttgagagactctgt...
Class ORF:    [K] M  T  T  E  T  R  G  K  I  E  D  V  I  W  L  R  D  S  V...

...F  R  L  H  E  R  T [H]:  KEGG ORF
...ttcgactgcatgaaaggact[cat]atgggagagaaagtctaa ...
...ttcgactgcatgaaagga[ctc]atatgggagagaaagtctaa ...
... F  D  C  M  K  G [L]:  Class ORF
```

Рис. 7. Перекодировка последовательности оснований локуса 7561 из базы данных KEGG29 (нуклеотиды 157–1911 (F1) и 158–1909 (F2) в аминокислотную последовательность. Показаны только начало и конец последовательностей.

	1	2	3
A	97	91	86
T	50	116	77
C	62	73	86
G	123	52	83

Рис. 8. Матрица класса триплетной периодичности, к которому была отнесена триплетная периодичность оснований (нуклеотиды 157–1911) локуса 7561 из базы данных KEGG29.

класс триплетной периодичности со сдвигом рамки считывания, удается одновременно выявить подобие по рамке класса и для рамки KEGG. Вероятно, после сдвига рамки считывания гены накопили большое количество замен и подобие уже невозможно заметить или же данная последовательность вообще не содержит какой-либо подобной последовательности в UniProt. В то же время 2489 гена имеют подобие аминокислотных последовательностей, созданных по рамке класса, но не имеют подобия для аминокислотных последовательностей, созданных по рамке KEGG. Эти данные говорят о том, что, по крайней мере, 2489 последовательностей могли быть образованы посредством сдвигов рамок считывания или инверсий.

По результатам сканирования создан банк данных, содержащий информацию о найденных подобиях для всех рассмотренных участков периодичности (http://victoria.biengi.ac.ru/ancorfs/), матриц классов и генов, входящих в тот или иной класс триплетной периодичности (http://victoria.biengi.ac.ru/ancorfs/classes.php).

Феномен триплетной периодичности кодирующих последовательностей отмечен уже достаточно давно [1–10]. Полученные нами данные свидетельствуют о том, что большую часть (94%) триплетной периодичности, обнаруженной в генах из базы данных KEGG29, можно свести к 2520 классам. Если учесть, что в среднем в каждом классе содержится 86.2% матриц, где первое оснований матрицы *M* при-

ходится на первое основание кодона, и только очень небольшое число кластеров (около 2%) содержат матрицу с индексом сдвига, равным 1, менее 50% (всего 53 из 2 520), то можно утверждать о существовании сильной корреляции между триплетной периодичностью и рамкой считывания в гене. Если бы такой корреляции не было, то формирование классов триплетной периодичности при помощи применяемого алгоритма было бы невозможно. Это показано при классификации 30 случайных множеств матриц триплетной периодичности, полученных при $Z > 5.0$.

Введение классов триплетной периодичности представляется нам важным для разработки более совершенных алгоритмов поиска кодирующих районов ДНК. При проведении классификации мы детально характеризуем триплетную периодичность, а матрицы, являющиеся антагонистическими, выделяются в различные классы и не "уничтожают" друг друга. Под антагонистическими матрицами понимаются такие матрицы, которые после объединения (первый столбец сливается с первым, второй со вторым и т.д.) дают уровень триплетной периодичности меньший, чем у исходных двух матриц (рис. 2). Это позволит на более значимом статистическом уровне выявлять кодирующие последовательности методами HMM или профильным анализом по триплетной периодичности, принадлежащей к антагонистическим классам. Наличие взаимосвязи между триплетной периодичностью и рамкой считывания дает возможность предсказывать для новых нуклеотидных последовательностей не только сам факт принадлежности их к кодирующим участкам, но и определять в них наиболее вероятную рамку считывания.

Присутствие в каждом классе триплетной периодичности матриц, сдвинутых или инвертируемых относительно используемой в соответствующем гене рамки считывания, свидетельствует в пользу двух гипотез. Во-первых, такой сдвиг или инверсия могли быть следствием делеций, вставок нуклеотидов или инверсий последовательностей, после чего образуется новая рамка считывания и новая аминокислотная последовательность [33–34]. Поскольку изучаемая нами скрытая триплетная периодичность возникает как коллективное свойство, то она не может исчезнуть из-за отдельных вставок, делеций или инверсий, а также нуклеотидных замен [25]. Вследствие этого в гене с делецией и вставкой нуклеотидов или инверсией нуклеотидной последовательности в нижележащей последовательности будет существовать триплетная периодичность, привязанная к старой рамке считывания; на этой же последовательности будет образована новая рамка считывания. При обнаружении триплетной периодичности в такой нуклеотидной последовательности и ее последующей классификации она может войти в соответствующий класс со сдвигом или инверсией относительно новой рамки считывания, что мы и обнаружили в данной работе. Частично эту гипотезу подтверждают данные табл. 3 и наличие гомологичных последовательностей в банке данных UniProt, созданных по рамке считывания соответствующего класса.

Во-вторых, нельзя исключить тот факт, что однотипная триплетная периодичность может быть случайно образована в разных генах в различных рамках считывания. В этом случае при классификации такие матрицы триплетной периодичности войдут в соответствующий класс с различными сдвигами или инверсией, что может приводить к обнаружению нами в классе триплетной периодичности некоторого количества матриц, вошедших в класс со сдвигом. Это может быть причиной существования некоторого количества матриц в каждом классе, вошедших в класс со сдвигами или же инверсиями. Трудно оценить долю таких матриц от их общего объема в классе, но вероятно она не слишком велика, так как есть большое количество классов триплетной периодичности, где число матриц со сдвигом не наблюдалось вообще, или их количество не превышало нескольких штук.

СПИСОК ЛИТЕРАТУРЫ

1. Fickett J.W. 1998. Predictive methods using nucleotide sequences. *Methods Biochem. Anal*. **39**, 231–245.
2. Staden R. 1994. Staden: statistical and structural analysis of nucleotide sequences. *Methods Mol. Biol*. **25**, 69–77.
3. Baxevanis A.D. 2001. Predictive methods using DNA sequences. *Methods Biochem. Anal*. **43**, 233–52.
4. Gutierrez G., Oliver J.L., Marin A. 1994. On the origin of the periodicity of three in protein coding DNA sequences. *J. Theoret. Biol*. **167**, 413–414.
5. Gao J., Qi Y., Cao Y., Tung W.W. 2005. Protein Coding Sequence Identification by Simultaneously Characterizing the Periodic and Random Features of DNA Sequences. *J. Biomed. Biotechnol*. **2**, 139–146.
6. Yin C., Yau S.S. 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J. Theor. Biol*. **247**, 687–694.
7. Eskesen S.T., Eskesen F.N. Kinghorn B., Ruvinsky A. 2004. Periodicity of DNA in exons. *BMC Mol. Biol*. **5**, 12.
8. Bibb M.J., Findlay P.R., Johnson M.W. 1984. 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*. **30**, 157–166.
9. Konopka A.K. 1994. Sequences and codes: fundamentals of biomolecular cryptography. In: *Biocomputing: Informatics and genome projects*. Ed. Smith D. San Diego: Acad. Press, pp. 119–174.
10. Trifonov E.N. 1999. Elucidating sequence codes: three codes for evolution. *Annals New York Acad. Sci*. **870**, 330–338.
11. Eigen M., Winkler-Oswatitsch R. 1981. Transfer-RNA: the early adaptor. *Naturwissenschaften*. **68**, 217–228.

12. Zoltowski M. 2007. Is DNA Code Periodicity Only Due to CUF – Codons Usage Frequency? *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **1**, 1383–1386.
13. Antezana M.A., Kreitman M. 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. *J. Mol. Evol.* **49**, 36–43.
14. Karlin S., Bucher P. 1992. Correlation analysis of amino acid usage in protein classes. *Proc. Natl. Acad. Sci.* USA. **89**, 12165–12169.
15. Zhang J. 2005. On the Evolution of Codon Volatility. *Genetics*. **169**, 495–501.
16. Trifonov E.N. 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. *J. Mol. Biol*. **194**, 643–652.
17. Fickett J.W. 1996. The gene identification problem: an overview for developers. *Comput. Chem*. **20**, 103–118.
18. Issac B., Singh H., Kaur H., Raghava G.P.S. 2002. Locating probable genes using Fourier transform approach. *Bioinformatics*. **18**, 196–197.
19. Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., Ramaswamy R. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Bioscie*. **13**, 263–70.
20. Azad R.K., Borodovsky M. 2004. Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. *Briefings Bioinform*. **5**, 118–130.
21. Henderson J., Salzberg S., Fasman K.H. 1997. Finding genes in DNA with a Hidden Markov Model. *J. Comput. Biol*. **4**, 127–141.
22. Snyder E.E., Stormo G.D. 1993. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res*. **21**, 607–613.
23. Thomas A., Skolnick M.H. A probabilistic model for detecting coding regions in DNA sequences. 1994. *IMA J. Math. Appl. Med. Biol*. **11**, 149–160.
24. Korotkov E.V., Korotkova M.A., Kudryashov N.A. 2003. Information decomposition method for analysis of symbolical sequences. *Physics Lett*. A. **312**, 198–310.
25. Коротков Е.В., Короткова М.А., Френкель Ф.Е., Кудряшов Н.А. 2003. Информационная концепция поиска периодичности в символьных последовательностях. *Молекуляр. биология*. **37**, 436–451.
26. Gribskov M., Veretnik S. 1996. Identification of sequence pattern with profile analysis. *Methods Enzymol*. **266**, 198–212.
27. Kullback S. 1978. *Information Theory and Statistics*. Gloucester: Peter Smith. 399 c.
28. Chaley M.B., Korotkov E.V., Skryabin K.G. 1999. Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples. DNA Res. **6**, 153–163.
29. Гмурман В.Е. 2003. *Теория вероятности и математическая статистика*. М.: Высшая школа. 479 с.
30. Grosse I., Buldyrev S.V., Stanley H.E., Holste D., Herzel H. 2000. *Pacific Symposium on Biocomputing*. Hawaii, USA: Abstract book. P. 611.
31. Ota T., Suzuki Y., Nishikawa T., Otsuki T., Sugiyama T., Irie R., Wakamatsu A., Hayashi K., Sato H., Nagai K., Kimura K., Makita H., Sekine M., Obayashi M., Nishi T., Shibahara T., Tanaka T., Ishii S., Yamamoto J., Sugano S. 2004. Complete sequencing and characterization of 21.243 full-length human cDNAs. *Nature Genetics*. **36**, 40–45.
32. Thiesen H.J. 1990. Multiple genes encoding zinc finger domains are expressed in human T cells. *New Biologist*. **2**, 363–374.
33. Raes J., van de Peer Y. 2005. Functional divergence of proteins through frameshift mutations. *Trends Genetics*. **21**, 428–431.
34. Hahn Y., Lee B. 2005. Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics*. **21**, 186–194.