на правах рукописи

Френкель Феликс Ефимович

КЛАССИФИКАЦИЯ ТРИПЛЕТНОЙ ПЕРИОДИЧНОСТИ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ГЕНОВ ИЗ БАЗЫ ДАННЫХ KEGG-29

03.00.28 – биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата биологических наук

Москва – 2008

Работа выполнена в Центре «Биоинженерия» РАН

Научный руководитель: доктор биологических наук, профессор
Коротков Евгений Вадимович

Официальные оппоненты: доктор физико-математических наук
Ефремов Роман Гербертович

кандидат биологических наук
Иванисенко Владимир Александрович

Ведущая организация: Институт молекулярной биологии
им. В. А. Энгельгардта Российской Академии Наук

Защита диссертации состоится 24 апреля 2008 г. в 12-30 на заседании Диссертационного совета Д 001.010.01 при ГУ НИИ БМХ РАМН по адресу: Москва, ул. Погодинская, д. 10

С диссертацией можно ознакомиться в библиотеке ГУ НИИ БМХ РАМН

Автореферат разослан ___ марта 2008 г.

Ученый секретарь Диссертационного совета

кандидат химических наук Карпова Елена Анатольевна

Общая характеристика работы

Актуальность проблемы

Предметом исследования в данной работе являются последовательности ДНК генов. Последние 25 лет наблюдается экспоненциальный рост объема генетических последовательностей: размер крупнейшего банка данных GenBank удваивается каждые 18 месяцев. Такой массив первичных данных позволяет выявлять свойства генетических последовательностей, присущие разнообразным видам организмов. В рамках данной работе изучались последовательности ДНК, входящие в состав генов. В настоящее время основное внимание уделяется поиску генов во вновь получаемых, неаннотированных последовательностях. В данной работе акцент сделан на классификацию такого свойства кодирующих последовательностей ДНК как их триплетная периодичность. При использовании для классификации триплетной периодичности появляется возможность проследить эволюцию кодирующих последовательностей уже не на уровне подобия символьных последовательностей, а на уровне их производных характеристик, имеющих бо́льшую устойчивость к мутациям, чем сами последовательности. К тому же привязка триплетной периодичности к рамке считывания генов позволяет проводить поиск возможных мутаций, приведших к сдвигу рамки считывания в генах, или инверсии последовательности гена. Ранее сдвиги рамок считывания выявлялись только прямым сравнением кодирующих последовательностей и транслированных по ним аминокислотных последовательностей между собой, что уменьшало чувствительность поиска и не давало возможности ввести точку отсчета для сдвигов рамки считывания. В данном исследовании такая точка отсчета вводится за счет наличия в формируемых классах доминирующей фазы триплетной периодичности. В дополнение к этому поиск возможных мутаций в кодирующих последовательностях генов осуществляется использованием найденных районов непрерывной триплетной периодичности в качестве профилей для поиска в этих же генах триплетной периодичности уже с учетом возможных вставок и делеций нуклеотидов. В данной работе это задача решается методом модифицированного профильного анализа.

Триплетная организация последовательностей ДНК, кодирующих белки, является свойством всех известных на настоящее время живых систем. В данном исследовании с помощью метода информационного разложения символьных последовательностей она была обнаружена в 79% кодирующих районов ДНК. Причинами периодичности являются избыточность генетического кода, предпочтения в использовании специфичных кодонов для кодирования аминокислот, а также насыщенность белков определенными аминокислотами. Существует также гипотеза, что триплетная периодичность может возникнуть вследствие необходимости контроля за мутациями посредством сдвига рамок считывания. Как показано в данной работе, триплетная периодичность имеет корреляцию с рамкой считывания в гене, т.е. фаза периодичности привязана к первой позиции кодонов, кодирующих аминокислоты. Отметим также, что, как показано автором, триплетная периодичность не может исчезнуть в отдельных эволюционных событиях, таких как замены нуклеотидов или же инверсии последовательностей оснований ДНК. При делециях и вставках периодическая последовательность разбивается на подпоследовательности, также обладающие периодичностью, но с различными фазами. Поэтому триплетная периодичность может служить эффек-

тивным инструментом для выявления в генах районов, в которых произошел эволюционный сдвиг рамки считывания или инверсия последовательности оснований ДНК.

При классификации триплетной периодичности формируются множества кодирующих последовательностей генов, имеющих близкий тип периодичности. Для этого было введено понятие типа триплетной периодичности и мера подобия между различными типами периодичности. С помощью разработанного алгоритма классификации было получено 2 520 классов, которые включили в себя 93% всех найденных последовательностей с триплетной периодичностью. Для поиска сдвигов рамок считывания, сравнение типов периодичности проводилось с учетом всех возможных циклических сдвигов и инверсии в периодических последовательностях. Таким образом, классы содержали периодические последовательности с различными сдвигами или инверсиями. Затем в каждом классе периодичности выделялся доминирующий в нем вариант сдвига. Доля периодических последовательностей, которые вошли в классы с доминирующим в нем сдвигом, составила 92%. Это указывает на наличие привязки триплетной периодичности к рамке считывания гена. Для последовательностей с отличными от доминирующего варианта сдвига в классе была выдвинута гипотеза, что такой сдвиг фазы триплетной периодичности или ее инверсия является следом мутации в гене, в результате которой произошел сдвиг рамки считывания в гене, либо инверсии этой последовательности. Для подтверждения этой гипотезы была произведена перекодировка (трансляция) периодических последовательностей ДНК в аминокислотные последовательности как по рамке считывания, указанной в БД KEGG, так и по предполагаемой гипотетической (древней) рамке считывания, соответствующей доминирующему варианту сдвига в классе. Для полученных аминокислотных последовательностей были найдены все имеющиеся подобия к известным белкам из БД UniProt. Таким образом, было найдено 2 679 районов генов, которые имели подобия к известным белкам при кодировании по гипотетической рамке считывания, доминирующей в классе периодичности.

Кроме генов, кодирующих белки, существенный научный интерес представляют гены транспортных РНК (тРНК). В данном направлении достигнуты высокие результаты при поиске действующих генов тРНК. В то же время, не были выявлены их сильно измененные копии, являющихся элементами некоторых структур в последовательностях ДНК. В частности, многие мобильные генетические элементы (например, SINE-повторы) содержат тРНК-подобные последовательности. Поиск и анализ мобильных элементов имеет высокую значимость, так как они составляют существенную часть генома эукариот (особенно высших) и тесно ассоциированы с их эволюционным развитием. Наряду с тем, что тРНК – это одна из древнейших структур генетического аппарата клетки, тРНК-подобные последовательности являются частью некоторых семейств мобильных элементов (повторов). Известно, что у эукариот (в особенности у высших) мобильные элементы составляют существенную часть генома и тесно ассоциируются с их высокой изменчивостью. В этой части работы были выявлены все последовательности в БД GenBank, родственные существующим генам тРНК. Для этого был применен метод модифицированного профильного анализа, выявляющий сильно измененные копии заданных в профиле исходных последовательностей. Данный метод обладает большей чувствительностью, чем такие широко используемые методы поиска подобий как BLAST. По результатам применения метода показано существование большого количества

тРНК-подобных последовательностей, в т.ч. и в составе повторов, не ассоциированных с ними раннее. По результатам исследования создан банк данных тРНК-подобных последовательностей.

Цель и задачи исследования

Целью представленной работы является классификация триплетной периодичности в последовательностях ДНК генов, кодирующих белки. По данным классификации в генах выявляются возможные мутации, приведшие к сдвигу рамки считывания или к инверсии кодирующих последовательностей.

В части анализа генов тРНК задачей данного исследования является поиск последовательностей ДНК, имеющих отдаленное подобие к известным генам тРНК.

В результате проведенного анализа области исследования был определен состав задач, решаемый настоящей работой:

1. Поиск методом информационного разложения последовательностей ДНК со скрытой триплетной периодичностью в кодирующих областях генов из БД KEGG.
2. Построение меры подобия между типами скрытой триплетной периодичности. Разработка алгоритма классификации триплетной периодичности по построенной мере подобия между ее типами. Классификация триплетной периодичности, найденной в кодирующих районах генов из БД KEGG.
3. Поиск мутаций в кодирующих районах генов, приведших к сдвигу рамки считывания или инверсии последовательности. Поиск подобий для аминокислотных последовательностей, полученных при трансляции кодирующих районов генов, имеющих мутацию, как по действующей рамке считывания гена, так и по найденной гипотетической (древней) рамке считывания, к последовательностям известных белков из БД UniProt с использованием программы BLAST.
4. Разработка банка данных, содержащего информацию о классах триплетной периодичности, найденной в кодирующих районах ДНК, и возможных мутациях, приведших к сдвигу рамки считывания в гене или его инверсии.
5. Разработка алгоритма оптимизации параметров поиска слабовыраженного подобия последовательностей ДНК к заданному семейству последовательностей методом модифицированного профильного анализа.
6. Поиск методом модифицированного профильного анализа возможных мутаций, приведших к сдвигу рамки считывания в областях кодирующих последовательностей, смежных с районами непрерывной триплетной периодичности.
7. Разработка банка данных, содержащего информацию о возможных мутациях в кодирующих последовательностях генов, приведших к сдвигу рамки считывания в областях, смежных с районом непрерывной триплетной периодичности.
8. Поиск методом модифицированного профильного анализа последовательностей из БД GenBank, подобных изотипическим семействам генов тРНК и их классификация.
9. Разработка банка данных, содержащего информацию о найденных тРНК-подобных последовательностях.

Научная новизна

1. Создана система классификации триплетной периодичности. В рамках данной системы введено понятие типа триплетной периодичности, меры подобия между типами и разработан алгоритм классификации триплетной периодичности. Введенные классы позволяют на основе сходства типов триплетной периодичности выявлять потенциальные случаи мутаций, приведших к сдвигу рамки считывания генов или инверсии кодирующей последовательности гена.
2. Разработан метод поиска сдвигов рамки считывания в кодирующих областях генов и инверсий кодирующих последовательностей генов с использованием созданных классов триплетной периодичности. Найдено множество случаев возможных мутаций, приведших к сдвигу рамки считывания и инверсии кодирующих последовательностей. Для существенной части кодирующих последовательностей со сдвигами и инверсиями рамки считывания найдены подобия между аминокислотными последовательностями, кодированными по предполагаемой древней рамке считывания, к известным белкам из БД UniProt.
3. Разработан алгоритм оптимизации параметров поиска слабовыраженного подобия последовательностей ДНК к заданному семейству последовательностей методом модифицированного профильного анализа.
4. При поиске подобия к семействам последовательностей генов тРНК в последовательностях ДНК из БД GenBank было обнаружено множество тРНК-подобных последовательностей, не идентифицированных ранее другими методами. тРНК-подобные последовательности были найдены в различных районах ДНК, например, в кодирующих последовательностях генов и повторах нескольких типов, в т.ч. в повторах, не ассоциированных ранее с тРНК-подобными последовательностями. Предложены варианты происхождения найденных тРНК-подобных последовательностей и их возможные функции. Построена классификация тРНК-подобных последовательностей.

Практическое значение работы

В рамках настоящей работы созданы алгоритмы и реализующее их программное обеспечение для классификации триплетной периодичности в последовательностях ДНК генов, кодирующих белки:

1. Создан алгоритм и программа классификации триплетной периодичности и банк данных, содержащий информацию о сформированных классах триплетной периодичности, найденной в белок-кодирующих районах ДНК с информацией об этих районах. Банк данных представляет интерес при исследовании эволюции кодирующих последовательностей. В настоящей работе этот банк данных применен при поиске возможных мутаций, приведших к сдвигу рамки считывания гена. Также созданные классы триплетной периодичности могут использоваться для разработки новых методов поиска белок-кодирующих последовательностей, т.к. классы являются по своей сути профилями для поиска районов триплетной периодичности.
2. Создан банк данных, содержащий информацию о возможных мутациях, приведших к сдвигу рамки считывания гена. В банке данных также показаны аминокислотные последовательности, полученные при трансляции кодирующих районов генов, имеющих мутацию, по действующей рамке считывания гена и по найденной гипотетической (древней) рамке считывания. Дополнительно приведены все случаи подобия этих аминокислотных последовательностей к

известным белкам из БД UniProt. Информация в банке данных представляет ценность при исследовании эволюции кодирующих последовательностей ДНК.

3. Разработан алгоритм оптимизации параметров поиска методом модифицированного профильного анализа слабовыраженного подобия последовательностей ДНК к заданному семейству последовательностей. Алгоритм может применяться для автоматической настройки параметров метода модифицированного профильного анализа для увеличения его чувствительности метода при поиске подобий к любому задаваемому профилю. Также алгоритм может применяться для поиска периодических последовательностей по ее матрице частот с учетом вставок и делеций нуклеотидов.
4. Создан банк данных, содержащий информацию о возможных мутациях, приведших к сдвигу рамки считывания в областях кодирующих последовательностей, смежных с районами непрерывной триплетной периодичности. В банке представлены аминокислотные последовательности, полученные трансляцией кодирующих последовательностей по действующей и гипотетической (построенной с учетом найденных мутаций) рамкам считывания в гене, и случаи подобия этих последовательностей к известным белкам из БД UniProt.
5. Создан банк данных с информацией о тРНК-подобных последовательностях. Банк данных может использоваться при исследованиях в области эволюции генов тРНК и мобильных элементов.

Все созданные банки данных имеют открытый доступ из сети Интернет.

Положения, выносимые на защиту

1. Система классификации скрытой триплетной периодичности в последовательностях ДНК.
2. Метод поиска возможных мутаций, приведших к сдвигу рамки считывания в гене, по созданным классам триплетной периодичности.
3. Метод поиска возможных мутаций, приведших к сдвигу рамки считывания в областях кодирующих последовательностей, смежных с районами непрерывной триплетной периодичности
4. Алгоритм оптимизации параметров поиска слабовыраженного подобия последовательностей ДНК к заданному семейству последовательностей методом модифицированного профильного анализа.
5. Банк данных, содержащий информацию о сформированных классах триплетной периодичности.
6. Банк данных, содержащий информацию о возможных мутациях, приведших к сдвигу рамки считывания гена.
7. Банк данных, содержащий информацию о возможных мутациях, приведших к сдвигу рамки считывания в областях кодирующих последовательностей, смежных с районами непрерывной триплетной периодичности.
8. Банк данных с информацией о найденных тРНК-подобных последовательностях.

Апробация работы

Результаты, представленные в данной диссертационной работе, опубликованы в [1-9] и устно докладывались на международной конференции BGRS’2002 (Новосибирск, 2002), на Международной школе-конференции «Системная биология и биоинженерия» (Звенигород, 2005) и на межлабораторном семинаре Центра «Биоинженерия» РАН (Москва, 2007).

Структура и объем диссертации

Диссертация состоит из введения, четырех глав, заключения, 7 приложений и списка литературы из 106 наименований. Общий объем диссертации составляет 100 страниц, из них 90 страниц – основной текст, который содержит 34 рисунка, 2 схемы и 24 таблицы.

Материалы и методы

Классификация триплетной периодичности последовательностей ДНК генов, собранных в банке данных KEGG

Поиск триплетной периодичности в генах

Триплетная периодичности выявлялась 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 их кодирующие последовательности объединяли все их экзоны. Поиск периодичности проводился при помощи метода информационного разложения. Для этого последовательность оснований ДНК *A(n)={a(1)a(2),…,a(n)}* каждого гена сравнивалась с равной по длине искусственной периодической последовательностью вида: *S(3)={s(1)s(2)s(3)s(1)s(2)s(3),…,s(1)s(2)s(3)}*, где *s(1)≡'1'*, *s(2)≡'2'*, *s(3)≡'3'*. В данной последовательности символы рассматривались как числа. Для сравнения последовательностей заполнялась матрица совпадений *M(3x4)*. У этой матрицы признаками столбцов являются символы *'1'*, *'2'* и *'3',* а признаками строк являются символы последовательности оснований ДНК *w(1)≡'a'*, *w(2)≡'t'*, *w(3)≡'c'*, *w(4)≡'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* рассчитывалась взаимная информация по формуле:

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

Формула 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 можем оценить вероятность случайного подобия между последовательностью *S(3)* и *A(n).* Для этого мы использовали метод Монте-Карло и величину *Z*, которая рассчитывалась как:

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

Формула 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*. Вели-

чина *Z* имеет распределение, близкое к нормальному, что мы проверили, сравнив искусственную периодическую последовательность с множеством случайных последовательностей объемом 10^7 символов. 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}.

Алгоритм поиска обеспечивает привязку районов триплетной периодичности к первой позиции рамки считывания гена. Это достигается смещением границ окна сканирования на длины, кратные трем символам, и существующей привязке кодирующих последовательностей ДНК в банке данных KEGG к первой позиции рамки считывания. 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 т.ч. и с различными матрицами *M*.

Классификация триплетной периодичности

Классификация матриц триплетной периодичности проводилась алгоритмом классификации с порогом качества (QT-clustering), при котором циклически формируется множество классов-кандидатов по установленному порогу подобия, а затем среди них выбирается класс с максимальным или минимальным значением целевой функции. Таким образом, 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 для всех матриц триплетной периодичности мы определяли множество матриц, которые им подобны с учетом возможности 3-х циклических сдвигов и 3-х циклических сдвигов в случае инверсии (Таблица 1).

Таблица 1. Преобразование матрицы периодичности для различных индексов сдвига

Индекс сдвига

Матрица *M* после 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}

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}

3

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

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}

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}

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}

Циклические сдвиги соответствуют возможным сдвигам триплетной периодичности относительно друг друга. Множество подобных матриц назовем классом-кандидатом. А матрицу, относительно которой был построен класс-кандидат, назовем центральной матрицей класса. При сравнении матриц учитывались все возможные циклические сдвиги и инверсия последовательности оснований ДНК, т.е. для пары матриц делалось 6 сравнений. При каждом сравнении центральная матрица класса фиксировалась, а преобразовывалась только та матрица, которая с ней сравнивалась. Для определения степени различия матриц была введена следующая количественная мера. Пусть даны две матрицы M^1 и M^2. Определим меру W как:

$$W = \sum_i \sum_j t_{ij}$$

Формула 3

Здесь матрица $T=\{t_{ij}\}$ равна:

$$t_{ij} = \left(\frac{m_{ij}^1}{y_j^1} - \frac{m_{ij}^2}{y_j^2} \right) \Bigg/ \sqrt{p(1-p)\left(\frac{1}{y_j^1} + \frac{1}{y_j^2}\right)}$$

Формула 4

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

Для определения множества матриц, входящих в класс-кандидат, вводилось пороговое значение величины $W=W_0$=3.44. Мы считали, что матрица принадлежит классу-кандидату, если рассчитанная величина W (Формула 3) была меньше порогового значения W_0. Значение W_0 обеспечивает

вероятность случайного объединения матриц в класс, равную 8.22×10^{-4}. Оно было получено методом Монте-Карло путем генерации случайных матриц и отбора среди них матриц, представляющих триплетную периодичность с уровнем значимости, большим или равным порогу значимости периодичности, установленному при поиске классифицируемых районов периодичности.

Выбор значения W_0=3.44 связан с двумя факторами. Во-первых, мы хотели объединить в классах основную часть матриц *M* и, соответственно, оставить вне классов незначительное число матриц. Одновременно с этим хотелось бы, чтобы введенные нами классы были максимально представительными (объединяли в себе существенное число случаев периодичности), а число самих классов было бы сравнительно небольшим. Проведенная нами классификация матриц триплетной периодичности при значениях W_0 выше 3.44 показала, что число матриц, входящих в классы уменьшается с увеличением W_0. Поэтому значение W_0=3.44 оказалось оптимальным, так как оно обеспечивает высокую долю матриц триплетной периодичности, входящих в классы при сравнительно небольшом количестве получающихся классов.

На втором этапе каждой итерации классификации матриц мы выбираем из классов-кандидатов тот, для которого целевая функция максимальна. Целевая функция должна быть мерой неоднородности распределения индексов сдвигов в классе. Т.е. мы создаем классы, имеющие в своем составе значительную долю одного из индексов сдвигов, что является признаком наличия связи между рамкой считывания гена и триплетной периодичности. Для этого мы использовали информационный критерий.

Пусть $x_1, x_2 .., x_6$ показывают, сколько в классе присутствует 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 $p_1,..,p_6$, $\sum_{i=1}^{6} p_i = 1$, где p_i=1/6. Различающая информация рассчитывается как:

$$I = \sum_{i=1}^{6} x_i \log \frac{x_i}{N p_i} = \sum_{i=1}^{6} x_i \log x_i - \sum_{i=1}^{6} x_i \log p_i - N \log N$$

Формула 5

Величина *2I* имеет распределение χ^2 с 5 степенями свободы. Значение 2*I* равно нулю, если вероятности $f_i = \frac{x_i}{N}$ равны по своему значению вероятностям p_i (в классе нет доминирующих индексов сдвига) и значение 2*I* принимает максимальное значение, если значение f_i. для какого-либо *i* равно 1.0, а остальные значения f_i. равны нулю (в классе представлен только один индекс сдвига). Это означает, что максимально неоднородное распределение индексов сдвига в классе дает максимум значения 2*I*. В результате среди созданных классов мы отбирали класс, который имеет максимальное значение *2I* и после этого итерация, включая оба описанных этапа, повторяется. Такое повторение происходит до тех пор, пока будет создаваться хотя бы один класс-кандидат, содержащий не менее одной матрицы, не считая центральную матрицу класса.

Для оценки статистической значимости значений целевой функции сформированных классов классификация также была проведена на 30 множествах случайных матриц той же мощности, представляющих триплетную периодичность с уровнем значимости, большим или равным порогу значимости периодичности, установленному при поиске классифицируемых районов периодичности. Затем, для каждого класса, сформированного при классификации, вычислялась мера *X* стандартного отклонения целевой функции *I* от его среднего для множеств случайных матриц:

$$X = \frac{I - \overline{I_{rnd}}}{\sigma(I_{rnd})}$$

Формула 6

, где I_{rnd} – значения *I* для классов, сформированных на множествах случайных матриц. Значения I_{rnd} выбирались для классов, сформированных при том же размере множества матриц, при котором был сформирован искомый класс периодичности. Это связано с зависимостью распределения максимума целевой функции от размера множества матриц, на которой она вычисляется. Показано, что *X* будет иметь приближенно нормальное распределение. Поэтому, исходя из числа сформированных классов и необходимой степени разделения исследуемого множества матриц триплетной периодичности и множества случайных матриц, было выбрано пороговое значение величины *X*>=6.0. Если для класса значение *X* было меньше выбранного порога, то класс считался имеющим статистически незначимую величину целевой функции.

Корректность выбранного порогового значения *X* демонстрируют результаты основной и контрольной классификации. На множестве матриц триплетной периодичности было сформировано 2 520 значимых классов, содержащих 94% исходных матриц. На 30 контрольных множествах случайных матриц было сформировано в общей сложности 21 значимый класс (т.е. менее 1 класса на множество), которые содержат 0.015% матриц.

В каждом классе производилось нормирование сдвигов, т.е. наиболее представленному в нем индексу сдвига присваивалось значение 1, а индексы сдвигов остальных матриц циклически менялись. Например, если наибольшее представительство имел индекс сдвига относительно центральной матрицы, равный 2, то ему присваивался индекс сдвига, равный 1. В этом случае индекс сдвига 1 был заменен на 3, 3 на 2, 4 на 5, 5 на 6, 6 на 4 (Таблица 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 индекс сдвига относительно центральной матрицы всегда имел значение, равное 1.

Поиск возможных мутаций, приведших к сдвигу рамки считывания в областях кодирующих последовательностей, смежных с районами непрерывной триплетной периодичности

Построение оптимального выравнивания искомой последовательности относительно профиля триплетной периодичности

Поиск триплетной периодичности со вставками и делециями символов происходит по заданной матрице распределения символов по позициям периода. Поэтому данную задачу можно свести к

задаче поиска по профилю, эффективное решение которой приведено выше. Решение состоит в циклическом замыкании координат внутри периода.

Пусть символьные последовательности определены на алфавите a_j,j=1,n и для периода длины k задана матрица M={m_{ij}}, такая что m_{ij} – количество символов a_i в позиции j периода. Переопределим операции сложения и вычитания координат внутри периода соответствующими операциями по модулю k:

$$i-1=(i+k-1)\ mod\ k$$
$$i+1=(i+1)\ mod\ k$$
$$i-x=(i-(x-1))-1,\ x>1, x\in\mathbb{N}$$
$$i+x=(i+(x-1))+1,\ x>1, x\in\mathbb{N}$$

Формула 7

После переопределения операций с координатами внутри периода, его матрицу можно рассматривать как профиль и использовать описанный выше метод модифицированного профильного анализа для построения оптимального локального выравнивания между заданной последовательностью и профилем периода.

Алгоритм выбора оптимальных параметров поиска отдаленного подобия к профилю последовательностей методом модифицированного профильного анализа

На задаваемые при поиске профили последовательностей или матрицы периодичности не накладывается никаких ограничений на распределение частот символов по позициям профиля/матрицы. При использовании метода модифицированного профильного анализа алгоритма возникает вопрос о точной настройке его параметров для эффективного решения поставленной задачи. Здесь можно выделить две проблемы:

- Введение коэффициента для компенсации роста суммарного веса случайной последовательности при увеличении ее длины, вызванного положительным значением суммы элементов весовой матрицы.
- Вычисление значения веса вставок и делеций символов, обеспечивающего необходимую чувствительность и избирательность алгоритма.

Для увеличения чувствительности алгоритма в рамках данной задачи описанные выше параметры выбирались из следующих принципов:

- Находимый на контрольных последовательностях район подобия должен как можно точнее совпадать с искомой последовательностью или районом периодичности.
- При внесении в район поиска вставок или делеций, они должны быть найдены в заданном количестве.

Пусть *pkof* – параметр, компенсирующий рост веса последовательности при увеличении ее длины. Тогда изменим формулу вычисления весовой матрицы как:

$$w_{ij} = \log \frac{m_{ij} / \sum_{i} m_{ij}}{\sum_{j} m_{ij} / \sum_{ij} m_{ij}} - pkof * M(w)$$

$$M(w) = \frac{\sum_{ij} \log \frac{m_{ij} / \sum_{i} m_{ij}}{\sum_{j} m_{ij} / \sum_{ij} m_{ij}}}{nk}$$

Формула 8

т.е. уменьшим каждый элемент весовой матрицы на взвешенное коэффициентом *pkof* среднее по этой матрице. Таким образом, варьируя коэффициент *pkof*, мы можем компенсировать ожидаемый рост суммарного веса выравнивания на случайной последовательности.

Коэффициент *pkof* для матрицы периодичности *M* подбирался по следующему алгоритму. Пусть дана последовательность S длины L, имеющая периодичность в районе [$X_1..X_2$] и матрица периодичности *M* для района [$X_1..X_2$]. Тогда:

- Создавалось множество {RS_i} (из 100) искусственных последовательностей длины *L*. Искусственные последовательности RS_i содержали в неизмененном виде и искомый район $S[X_1..X_2]$, а фланкирующие участки слева и справа (при их наличии) были случайным образом перемешаны. Тем самым создавалась обучающая выборка для подбора значения коэффициента *pkof*, максимально точно выделяющего искомый район.
- Для всех значений *pkof* из заданного диапазона производилось сканирование исходной матрицей профиля *M* по созданным искусственным последовательностям RS_i. Результаты сканирования сохранялись в виде троек $<RX_1^i, RX_2^i, RZ^i>$, где RX_1^i, RX_2^i - координаты найденного оптимального выравнивания в последовательности RS_i, а RZ^i – значение *Z* для него.
- Среди всех значений *pkof* выбиралось те, для которых в множестве искусственных последовательностей {RS_i} было найдено наибольшее число статистически значимых выравниваний.
- Затем среди них выбирались те, для которых суммарное отклонение координат найденных выравниваний [$RX_1^i .. RX_2^i$] от координат искомого района [$X_1..X_2$] было минимальным, т.е. значение

  $$DX = \sum_{i} \left(\left(RX_1^i - X_1 \right)^2 + \left(RX_2^i - X_2 \right)^2 \right)$$

  было наименьшим.

Среди оставшихся значений *pkof* выбиралось имеющее наибольшее среднее значение *Z*.

Подбор коэффициента v^{do} для матрицы периодичности *M* проводился по сходному с настройкой *pkof* алгоритму. Пусть дана последовательность S длины L, имеющая искомый район район последовательности [$X_1..X_2$] и матрица профиля *M* для района [$X_1..X_2$]. Тогда:

- Создавалось множество {RS_i} (из 100) искусственных последовательностей длины *L*. Искусственные последовательности RS_i содержали искомый район $S[X_1..X_2]$, у которого в позициях

$\{X_j^{ins}\}, X_j^{ins} \in (X_1..X_2)$ были произведены вставки случайных символов, а фланкирующие участки слева и справа (при их наличии) были случайным образом перемешаны. Тем самым создавалась обучающая выборка для подбора значения коэффициента v^{do}, максимально точно выделяющего искомый район последовательности и вставки и делеции символов в нем.

- Для всех значений v^{do} из заданного диапазона производилось сканирование исходной матрицей *M* по созданным искусственным последовательностям RS_i. Результаты сканирования сохранялись в виде четверок $<RX_1^i, RX_2^i, \{RX_j^{ins}\}, RZ^i>$, где RX_1^i, RX_2^i - координаты найденного оптимального выравнивания в последовательности RS_i, $\{RX_j^{ins}\}$ – множество координат вставок символов в найденном выравнивании, а RZ^i – значение *Z* для него.
- Среди всех значений v^{do} выбиралось те, для которых в множестве искусственных последовательностей {RSi} было найдено наибольшее число статистически значимых выравниваний.
- Затем среди них выбирались те, для которых суммарное отклонение координат найденных выравниваний [RX1i.. RX2i] от координат искомого района [X1..X2] было минимальным, т.е. значение

$$DX = \sum_i \left(\left(RX_1^i - X_1 \right)^2 + \left(RX_2^i - X_2 \right)^2 \right)$$

было наименьшим.

Далее среди оставшихся значений v^{do} выбирались то, для которого суммарное отклонение координат вставок символов в найденных выравниваниях $\{RX_j^{ins}\}$ от координат заданных $\{X_j^{ins}\}$ при создании искусственных последовательностей вставок было минимальным, т.е. значение

$$DX^{ins} = \sum_i \left(RX_j^{ins} - X_j^{ins} \right)^2$$

было наименьшим. При наличии нескольких вариантов с минимальным значением DX^{ins} выбиралось значение v^{do}, имеющее наибольшее среднее значение Z.

Метод поиска возможных мутаций, приведших к сдвигу рамки считывания в областях кодирующих последовательностей, смежных с районами непрерывной триплетной периодичности

При поиске мутаций, приведших к сдвигу рамки считывания генов путем сравнения фаз триплетной периодичности, найденной в них, мы можем найти, в основном, мутации, произошедшие в начале генов, когда сохраняются протяженные участки непрерывной триплетной периодичности. Для того, чтобы расширить область поиска сдвигов рамки считывания за счет учета мутаций в середине и в конце гена, а также нескольких мутаций в гене, мы провели поиск триплетной периодичности со вставками и делециями символов, расширяя найденные ранее районы непрерывной триплетной периодичности. Поиск был проведен в кодирующих последовательностей генов, в которых были найдены участки с непрерывной периодичностью, не покрывающие, в то же время, всей длины исследуемой последовательности. Для этих последовательностей проверялась гипотеза о наличии в них вставок и делеций символов, приведших к сдвигу фазы периодичности, в ре-

зультате которого непрерывная триплетная периодичность могла быть обнаружена не на всем протяжении последовательности. Пример исчезновения триплетной периодичности за счет делеции нуклеотида показан ниже (Рисунок 1).

Рисунок 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 было 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 3 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.

Для проверки гипотезы о наличии в кодирующих последовательностях генов вставок и делеций символов, приведших к сдвигу фазы периодичности, поиск периодичности со вставками и делециями символов был проведен не только на анализируемых последовательностях, но и на множестве контрольных последовательностей. Контрольные последовательности были получены из исходных, в которых был сохранен участок с найденной ранее непрерывной триплетной периодичностью, а фланкирующие его районы слева и справа были заменены случайными последовательностями с тем же символьным составом. Таким образом, по разнице в весах выравнивания анализируемого участка триплетной периодичности с весами выравниваний, полученных по контрольным последовательностям, делался вывод о достоверности проверяемой гипотезы.

Опишем алгоритм обнаружения вставок и делеций символов в районе с непрерывной триплетной периодичности подробнее. Пусть дана последовательность *S* длины *L*, имеющая непрерывную триплетную периодичность в районе $[X_1..X_2]$ и матрица периодичности (встречаемости символов по позициям периода) *M* для района $[X_1..X_2]$. Тогда с помощью алгоритма поиска триплетной пе-

риодичности со вставками и делециями символов (см. Раздел 0) построим оптимальное локальное выравнивание последовательности *S* относительно матрицы *M.* Пусть найденное выравнивание имеет координаты [$XB_1..XB_2$]. Тогда построим глобальное выравнивание последовательности *S* относительно матрицы *M*. Пусть W_s и WB_s – веса глобального выравнивания для участка непрерывной триплетной периодичности ([$X_1..X_2$]) и района найденного оптимального локального выравнивания ([$XB_1..XB_2$]) соответственно. В соответствии с алгоритмом построения матрицы выравнивания имеем:

$$W(S)=AM(X_2,\mathrm{mod}(X_2\text{-}1,k)+1)\text{-}AM(X_1,\mathrm{mod}(X_1\text{-}1,k)+1)$$

$$WB(S)=AM(XB_2,\mathrm{mod}(XB_2\text{-}1,k)+1)\text{-}AM(XB_1,\mathrm{mod}(XB_1\text{-}1,k)+1),$$

Формула 9

где mod – функция остатка от деления, *k* – длина периода (профиля).

В случае *WB(S)* > *W(S)* мы можем предполагать, что в последовательности *S* произошла вставка или делеция символа, приведшее к сдвигу фазы периодичности и ее размытию. Для проверки статистической значимости найденной разницы между *WB(S)* и *W(S)* для каждой последовательности *S* были сгенерированы множества случайных последовательностей {RS_i}, такое, что RS_i имели длину, равную длине *S,* и содержали в неизмененном виде и положении периодический участок *S*[$X_1..X_2$], а символы во фланкирующих участках слева и справа были случайным образом перемешаны. Для каждой случайной последовательности RS_i было построено глобальное выравнивание относительно матрицы *M* и найдены веса $WB(RS_i)$ и $W(RS_i)$ аналогично *WB(S)* и *W(S)* (Формула 9). В случае, когда

$$\max_i \{WB(RS_i)-W(RS_i)\} < (WB(S)-W(S))$$

Формула 10

гипотеза о наличии вставки или делеции символов в последовательности, приведшей к размытию непрерывной триплетной периодичности, считалась подтвержденной. Размер $|\{RS_i\}|$ множества случайных последовательностей {RS_i} выбирался исходя из ограничений на оценку ошибку первого рода в 5%. Пусть *NS* – число исходных последовательностей *S*, *KS* – число исходных последовательностей *S*, для которых подтвердилась проверяемая гипотеза (выполнена Формула 10). Тогда оценка ошибки первого рода при проверке гипотезы будет равна

$$p_{err}=\frac{NS}{|\{RS_i\}|KS}$$

Формула 11

и должна быть менее 0.05.

Поиск белковых подобий для аминокислотных последовательностей, транслированных по действующей и древней рамкам считывания гена

Рассмотрим те гены, в которых непрерывный район периодичности был расширен при поиске периодичности с учетом вставок и делеций символов и, в частности, те их районы, в которых наблю-

дается несовпадение действующей рамки считывания генов и триплетной периодичности из-за наличия вставки и/или делеции символов перед ними и внутри них. В связи с тем, что триплетная периодичность не может исчезнуть в отдельных эволюционных событиях, таких как делеции и вставки нуклеотидов, то возникает вопрос о биологической интерпретации таких событий. Если предположить, что обнаруженная триплетная периодичность со вставками и делециями нуклеотидов является следом некоторой древней рамки считывания генов, то можно ожидать, что для части таких случаев можно обнаружить подобие их белковых продуктов, полученных по древней рамке считывания, следы которой обнаружены, к существующим в современных банках данных белкам. Такое исследование было проведено в рамках данной работы.

Все найденные районы периодичности, в которых наблюдается несовпадение действующей рамки считывания генов и триплетной периодичности были транслированы в аминокислотные последовательности в соответствии с обеими альтернативным рамками считывания. Пример такой трансляции по выравниванию показан ниже (Рисунок 2):

Последовательность гена	gcggcaccattgtcaatgt*c**tcctccacgcaggccgcaaggtctatccgcaccatgacg**
Консенсус периодичности	tcgtcgtcgtcgtcgtcgtc**gtcgtcgtcgtcgtcgtcgtcgtcgtcgtcgtcgtcg**
Существующая рамка считывания	231231231231231231233**123123123123123123123123123123123**
Древняя рамка считывания	23123123123123123123**123123123123123123123123123123123123**
Трансляция по существующей рамке считывания	**SSTQAARSIRTMT**
Трансляция по древней рамке считывания	**VLHAGRKVYPHHD**

Рисунок 2. Трансляция районов триплетной периодичности с несовпадающими рамками считЫвания

Для всех полученные таким способ аминокислотных последовательностей с помощью программы BLAST был проведен поиск подобий с последовательностями из банка данных по белкам UniProt.

Поиск тРНК-подобных последовательностей в геномах различных видов

В данной работе мы провели поиск и анализ локализации в различных геномах еще не известных тРНК-подобных последовательностей, в том числе и их очень древних копий. Поиск производился методом модифицированного профильного анализа по последовательностям ДНК из БД GenBank. Локализация тРНК-подобных последовательностей рассматривалась в соответствии со стандартным описанием последовательностей в GenBank. Выявленные тРНК-подобные последовательности были классифицированы на основании распределения их нуклеотидов по позициям выравнивания относительно консенсуса тРНК. Для каждого класса был проанализирован изотипический состав тРНК-подобных последовательностей. Идентификация районов, содержащих тРНК-

подобные последовательности, как повторов или их фрагментов была выполнена с использованием поискового сервиса CENSOR по коллекции повторов RepBase.

Классификация тРНК-подобных последовательностей

Для классификации найденных тРНК-подобных последовательностей по их подобию между собой в качестве меры подобия была выбрана взаимная информация между множествами последовательностей. Если взаимная информация превышала определенное пороговое значение, то последовательности объединялись в общий класс. Нами был реализован следующий алгоритм классификации:

Каждая последовательность из анализируемого множества исходно была представлена в виде матрицы *F(i,j)* - распределения символов *S(i)* = {A,T(U),G,C,-} в выравнивании относительно консенсуса изотипических тРНК, длиной L. В зависимости от наличия или отсутствия конкретного символа в данной позиции консенсуса, соответствующему элементу матрицы *F(i,j)* присваивалось значение «1» либо «0». Полевые элементы матрицы *F* вводились как:

$$X(i) = \sum_j F(i, j),\ Y(j) = \sum_i F(i, j)\ ,\ i=1,\ldots,5,\ j=1,\ldots,L$$

Формула 12

Осуществлялся переход к матрице распределения частот оснований в позициях выравнивания:

$$P(i, j) = F(i, j) / Y(j)$$

Формула 13

Матрицы распределения частот оснований попарно сравнивались друг с другом, при этом заполнялась матрица частот парных соответствий *M(k,l), k,l=1,…5*:

$$M(k,l) = \sum_{j=1}^{L} P^r(k, j) \cdot P^f(l, j)$$

Формула 14

$$X(k) = \sum_l M(k,l),\ Y(l) = \sum_k M(k,l)$$

Формула 15

Мера подобия матриц P^r и P^f определялась как взаимная информация между ними:

$$I = \sum_k \sum_l M(k.l) \ln M(k,l) - \sum_k X(k) \ln X(k) - \sum_l Y(l) \ln Y(l) + Len \ln Len$$

Формула 16

$$Len = \sum_k X(k) = \sum_l Y(l)$$

Формула 17

Известно, что величина удвоенной взаимной информации имеет распределение χ^2 с *(k-1)×(l-1)*=16 числом степеней свободы. Поэтому значение взаимной информации между частотными матрицами *P*, определяющее случайную вероятность для χ^2 на уровне 0.05, было выбрано как пороговое значение I_{cut}. В случае $I \geq I_{cut}$ матрицы P^r и P^f объединялись в один класс простым сложением их соответствующих элементов. После чего новая объединенная матрица возвращалась в сравниваемое множество частотных матриц, а две исходные матрицы удалялись из этого множества. Процесс сравнения пар частотных матриц повторялся до тех пор, пока в анализируемом множестве не находилось ни одной пары матриц со значением взаимной информации более I_{cut}. Все оставшиеся матрицы представляли собой набор классов подобных последовательностей.

Результаты и обсуждение

Результаты классификации триплетной периодичности и поиска возможных мутаций, приведших к сдвигу рамки считывания в кодирующих последовательностях генов

Всего было проанализировано 578 868 генов, накопленных в банке данных KEGG версии 29. Общее число участков триплетной периодичности составило 472 288, которые были найдены в 457 333 генах (79% генов имеют районы с триплетной периодичностью). Такие результаты согласуются с более ранними работами по обнаружению триплетной периодичности. Для каждой последовательности ДНК с триплетной периодичностью рассчитывалась соответствующая матрица совпадений символов *M* и затем эти матрицы объединялись в классы. При классификации мы решили две задачи. Во-первых, мы выяснили, насколько разнообразны матрицы триплетной периодичности. Во-вторых, мы определили, существует ли и насколько постоянна взаимосвязь между первым основанием кодона и первой позицией триплетной периодичности. В процессе поиска триплетной периодичности первые позиции триплетной матрицы соответствовали первым позициям кодона и могли меняться только в ходе объединения матриц в классы, когда был возможен циклический сдвиг либо инверсия матрицы. Для изучения этой связи одновременно с матрицей класса триплетной периодичности создавалось множество, где содержались все индексы сдвига матриц, вошедших в данный класс. В результате классификации были получены 2 520 классов (http://victoria.biengi.ac.ru/ancorfs/classes.php). Классы матриц периодичности имеют большое разнообразие в своих размерах, от 1 до десятков тысяч. В классах с $X \geq 6.0$ (Формула 6) содержится 443 798 случаев периодичности из 472 288 найденных матриц триплетной периодичности. 8 591 матриц не вошли ни в один класс и остались автономными. 19 899 матриц вошли в незначимые классы, для которых $X < 6.0$. Как видно, около 94% матриц входят в значимые классы, что показывает существование связи между триплетной периодичностью и рамкой считывания. Классы, которые имеют $X < 6.0$, могут иметь небольшое количество матриц в своём составе и именно малое количество матриц не позволило этим классам получить $X \geq 6.0$, 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. 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 было выявлено, что только около 8 % от них входят в класс с каким-либо сдвигом рамки считывания или инверсией направления считывания. Это означает, что из 443 798 выявленных случаев периодичности, входящих в значимые классы, только 36 111 имели рамку считывания, отличную от характерной для большинства матриц, вошедших в класс.

Для каждого класса можно ввести главную рамку считывания, которая присутствует в данном классе (рамку 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, которые вошли в значимые классы с индексами сдвига 2-6 (Таблица 2).

Таблица 2. Распределение периодичностей по сдвигу относительно рамки считывания класса

Сдвиг периодичности относительно ОРС класса	**Число матриц**	**Индексы сдвига матрицы**
Рамка считывания 1	407 687	1
Рамка считывания 2	2 558	2
Рамка считывания 3	3 162	3
Всего прямых со сдвигом рамки	5 720	2+3
Всего прямых	413 407	1+2+3
Инверсия, рамка считывания 1	20 199	4
Инверсия, рамка считывания 2	7 616	5
Инверсия, рамка считывания 3	2 576	6
Всего антисмысловых со сдвигом рамка	10 192	5+6
Всего инвертированных последовательностей	30 391	4+5+6
Всего со сдвигами или инверсией	36 111	2+3+4+5+6
Всего участков периодичности	443 798	1+2+3+4+5+6

Мы также проверили матрицы периодичности на наличие в них симметричности, что может быть причиной найденных сдвигов в классах. Для этого мы сравнивали каждую матрицу саму с собой после проведения в ней циклических сдвигов и инверсий. Было обнаружено, что у 0.4% матриц, вошедших в классы, наблюдается подобие к одному из их (пяти) вариантов, полученных сдвигом и инверсией. Оценка подобия производилась тем же критерием сходства матриц периодичности, что использовался при классификации. Как видно, хотя симметричность и наблюдается у некоторых матриц, но, в целом, существенной роли в формировании сдвигов между матрицами внутри классов не играет (0.4% симметричных матриц против 8% матриц со сдвигом).

Данные (Таблица 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, то мы можем наблюдать ее как след существовавшей ранее рамки считывания, измененной в результате делеций, вставок или же инверсий последовательностей ДНК входящих в состав генов. Однако данная гипотеза нуждается в

дополнительной проверке. Такая проверка была сделана перекодированием нуклеотидной последовательности, в которой мы обнаружили сдвиги рамок считывания, в аминокислотную по рамке считывания, доминирующей в классе, в который она вошла. 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 у гипотетической аминокислотной последовательности есть гомологичные последовательности в банке данных UniProt, то это доказывает, 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.

Для проверки этой гипотезы нами был проведен поиск гомологов белковых продуктов, кодируемых найденными районами генов. Кодирование проводилось как по рамке KEGG, так и по рамке класса. Поиск гомологов проводился по БД UniProt с помощью программы BLAST. Среди найденных подобий рассматривались только значимые случаи, имеющие вероятность случайного совпадения (e-value) менее 5%. В результате проведенного сканирования были получены списки гомологов для 20 733 участков ДНК с триплетной периодичностью. Для 15 378 участков периодичности ни одного значимого подобия к их белковым продуктам по обеим рамкам считывания найдено не было. В 190 аминокислотных последовательностях наблюдалось подобие как для гипотетической, так и для реальной аминокислотной последовательности (Таблица 3).

Таблица 3. Число подобий для различных случаев сдвига ОРС

Вид трансформации кодирующей последовательности	**Число подобий**			**Индекс сдвига**	**Доля от общего числа районов периодичности**
	по обеим ОРС	**только по ОРС класса**	**только по ОРС KEGG**		
Рамка считывания 2	8	868	346	2	0.48
Рамка считывания 3	143	784	558	3	0.47
Антисмысловой, рамка считывания 1	25	443	12 991	4	0.67
Антисмысловой, рамка считывания 2	13	355	3 236	5	0.47
Антисмысловой, рамка считывания 3	1	39	923	6	0.37
Всего	**190**	**2 489**	**18 054**		**0.57**

Таким образом, только для небольшого количества генов, для которых триплетная периодичность вошла в 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, 2 489 генов 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, 2 679 последовательности могли быть образованы посредством сдвигов рамок считывания или инверсий.

Результаты поиска тРНК-подобных последовательностей

Для 22 позиционно-специфичных весовых матриц изотипических тРНК из базы данных Sprinzl методом модифицированного профильного анализа выявлено 455 457 тРНК-подобных последовательностей в 9 основных разделах GenBank. В соответствии с описанием в GenBank и RepBase из множества найденных последовательностей были удалены известные гены тРНК, тРНК-подобные последовательности, а также SINE-повторы как содержащие тРНК-производную часть. Районы, отмеченные в GenBank как “repeat_region” и “repeat_unit” без указания конкретного типа повтора, и содержащие найденные тРНК-подобные последовательности, дополнительно были проанализированы CENSOR-сервером с целью установления типа повтора. В случае идентификации этих районов как известных SINE-повторов, тРНК-подобные последовательности, которые им соответствовали, также были исключены из общего списка. После чего осталось 305 321 ранее не известных тРНК-подобных последовательностей. Мы также применили модифицированного профильного анализа для более точного исключения возможных тРНК-подобных районов MIR и Alu повторов. После исключения MIR и Alu повторов осталось 175 901 ранее не известных тРНК-подобных последовательностей.

При поиске прямых повторов, фланкирующих районы, в которых были найдены тРНК-подобные последовательности, было найдено 22 189 таких случаев. Средняя длина фланкированных районов равнялась 330 основаниям, варьируя от 100 до 680 оснований. При сравнении последовательностей этих районов с коллекцией повторов базы данных RepBase 12 301, т.е. не менее половины из них, были идентифицированы как известные повторы (мобильные элементы).

Банки данных

Для представления результатов исследований по каждому разделу диссертационной работы было создано четыре банка данных с доступом из сети Интернет. В качестве СУБД использованы программные комплексы Firebird (http://www.firebirdsql.org/) и PostgreSQL (http://www.postgresql.org/). Пользовательский интерфейс реализован на языке PHP (http://www.php.net/). Web-сервер Apache (http://www.apache.org/) установлен на сервере Вычислительного Комплекса Центра «Биоинженерия» РАН под управлением ОС SUSE Linux (http://www.opensuse.org/).

Банк данных районов скрытой периодичности последовательностей ДНК (MRFGS, http://victoria.biengi.ac.ru/mrgfnps/) содержит результаты анализа нуклеотидных последовательностей из банка данных GenBank-116. В нем представлены все районы со статистически значимой периодичностью, в т.ч. и с различной длиной периода. Для каждого случая периодичности наглядно представлена информация по локусу, в котором он был найден, и ее основные характеристики. Банк данных районов скрытой периодичности последовательностей белков (MRFPS, http://victoria.biengi.ac.ru/mrgfnps/) аналогичен описанному выше банку MRFGS, но содержит информацию по скрытой периодичности в аминокислотных последовательностях банка данных

SwissProt-38. Оба описанных банка данных были созданы для иллюстрации феномена скрытой периодичности в генетических последовательностях.

При классификации триплетной периодичности в белок-кодирующих районах генов были получены результаты, легшие в основу двух банков данных. Первый из них (http://victoria.biengi.ac.ru/ancorfs/) содержит результаты классификации периодичности и случаи возможных мутаций, приведших к сдвигу рамки считывания в гене, полученные на основании данных классификации. По всем случаям сдвига рамки считывания представлена также информация о найденных случаях подобия белковых продуктов, транслированных по действующей и предполагаемой древней рамкам считывания, к известным белкам из БД UniProt.

Данные по поиску возможных мутаций, приведших к сдвигу рамки считывания в областях кодирующих последовательностей, смежных с районами непрерывной триплетной периодичности легли в основу другого созданного банка данных (http://victoria.biengi.ac.ru/pertails/). Он содержит информацию по найденным случаям вставок и делеций, расширяющих район триплетной периодичности. Для каждого такого случая приведено построенное выравнивание последовательности относительно консенсуса периодичности, указаны районы с несовпадением действующей рамки считывания кодирующей последовательности и предполагаемой древней рамки считывания, привязанной к найденной триплетной периодичности.

Для организации публичного доступа к результатам, полученным при поиске тРНК-подобных последовательностей в геномах различных видов, был разработан соответствующий банк данных с доступом из сети Интернет (http://victoria.biengi.ac.ru/trnalikes/). Данный банк данных разбит на разделы в соответствии с оригинальной публикацией по этой теме (см. п. 6 списка публикаций).

Выводы

В рамках настоящей работы:

1. Введена мера подобия между типами скрытой триплетной периодичности, на ее базе разработан алгоритм классификации районов триплетной периодичности. Проведена классификация 472 288 районов триплетной периодичности, найденных в 578 868 кодирующих последовательностях генов из БД KEGG. Получены 2 520 статистически значимых классов, охватывающие около 94% найденных районов триплетной периодичности.
2. Создан банк данных, содержащий информацию о сформированных классах триплетной периодичности.
3. Найдено 36 111 случаев возможных мутаций в кодирующих районах генов, приведшие к сдвигу рамки считывания или инверсии последовательности. Для 2 660 возможных мутаций найдены подобия аминокислотных последовательностей кодированных, по найденной гипотетической (древней) рамке считывания, к последовательностям известных белков из БД UniProt. Для 190 из них найдены также белковые подобия и при трансляции по действующей рамке считывания гена.
4. Создан банк данных, содержащий информацию о возможных мутациях, приведших к сдвигу рамки считывания гена.
5. Найдены 1 135 случаев возможных мутаций в кодирующих районах генов, приведшие к сдвигу рамки считывания в областях кодирующих последовательностей, смежных с районами непрерывной триплетной периодичности. Для 150 обнаруженных возможных мутаций найдены подобия аминокислотных последовательностей, транслированных по предполагаемой древней рамке считывания, к известным белкам. По полученным результатам создан банк данных.
6. Создан банк данных, содержащий информацию о возможных мутациях, приведших к сдвигу рамки считывания в областях кодирующих последовательностей, смежных с районами непрерывной триплетной периодичности.
7. Проведен поиск тРНК-подобных последовательностей в БД GenBank, найдено 175 901 ранее не известных тРНК-подобных последовательностей. Построена классификация найденных последовательностей. По полученным результатам создан банк данных.
8. Создан банк данных с информацией о найденных тРНК-подобных последовательностях.

Публикации по теме диссертации

1. Frenkel F.E., Korotkov E.V. Revealing and functional analysis of tRNA-like sequences in various genomes. // Proceedings of the 2nd international conference on bioinformatics of genome regulations and structure. – Новосибирск. – 2002. – P. 23-26.
2. Френкель Ф.Е., Чалей М.Б., Коротков Е.В. Поиск и функциональный анализ тРНК-подобных последовательностей. // Материалы 7-й Пущинской школы-конференции молодых ученых «БИОЛОГИЯ - НАУКА XXI ВЕКА» . – Пущино. – 2003. – С. 251.
3. Коротков Е.В., Короткова М.А., Френкель Ф.Е., Кудряшов Н.А. Информационная концепция поиска периодичности в символьных последовательностях. // Молекулярная биология. – 2003. – Том 37(3) – С. 436-451.
4. Френкель Ф.Е., Чалей М.Б., Коротков Е.В., Скрябин К.Г. Эволюция тРНК-подобных последовательностей и изменчивость генома. // Материалы 2-ого Московского международного конгресса «Биотехнология: состояние и перспективы развития». – Москва. – 2003. – С. 26-27.
5. Frenkel F.E., Chaley M.B., Korotkov E.V., Skryabin K.G. Informational aspects of the latent triplet periodicity analysis. // Proceedings of the 3rd international conference on bioinformatics of genome regulations and structure. – Новосибирск. - 2004. – P. 58-59.
6. Frenkel F.E., Chaley M.B., Korotkov E.V., Skryabin K.G. Revealing and functional analysis of tRNA-like sequences in various genomes. // In “Bioinformatics Of Genome Regulation And Structure” .- Kluwer. - 2004. – P. 39-46.
7. Frenkel F.E., Chaley M.B., Korotkov E.V., Skryabin K.G. Evolution of tRNA-like sequences and genome variability // Gene. - 2004. - Vol. 335 – P. 57-71.
8. Frenkel F.E., Korotkov E.V. Fuzzy triplet periodicity as a footprint of coding regions evolution. // Proceedings of 2nd FEBS Advanced Lecture Course on Systems Biology: From Molecules to Life. - Austria. - 2007. – P. 176.