На правах рукописи

Суворова Юлия Максимовна

ИЗУЧЕНИЕ ТОЧЕК РАЗЛАДКИ ТРИПЛЕТНОЙ ПЕРИОДИЧНОСТИ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК, КОДИРУЮЩИХ БЕЛКИ

03.01.09 - математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2015

Работа выполнена в Федеральном государственном бюджетном научном учреждении Центр “Биоинженерия” Российской Академии Наук.

Научный руководитель: доктор биологических наук, профессор
Коротков Евгений Вадимович

Официальные оппоненты: **Макеев Всеволод Юрьевич,** доктор физико-математических наук, Федеральное государственное бюджетное учреждение науки «Институт общей генетики им. Н.И. Вавилова» РАН, научный руководитель отдела, заведующий лабораторией.

Иванисенко Владимир Александрович, кандидат биологических наук, Федеральное государственное бюджетное учреждение науки «Федеральный исследовательский центр Институт цитологии и генетики» СО РАН, заведующий лабораторией.

Ведущая организация: Федеральное государственное бюджетное учреждение науки «Институт молекулярной биологии им. Энгельгардта» РАН.

Защита состоится 19 ноября 2015 года в 11 часов на заседании диссертационного совета Д 001.010.01 при Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» по адресу: 119121, Москва, ул. Погодинская, д. 10, стр. 8.

С диссертацией можно ознакомиться в библиотеке ИБМХ и на сайте www.ibmc.msk.ru

Автореферат разослан _____ 2015 г.

Учёный секретарь диссертационного совета,
кандидат химических наук  Карпова Е.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы

Вопрос о происхождении современных генов и белков исследуется уже много лет и является одним из фундаментальных вопросов. В процессе эволюции геном в целом и кодирующие последовательности в частности подвергаются различным типам мутаций: таким как вставки и делеции (как отдельных символов, так и довольно протяженных участков), а также замены одних символов на другие. В случае, если изменения затрагивают кодирующие участки, это может привести к различным исходам: мутация может оказаться незначительной и не изменит функцию белка; или напротив мутация может оказаться летальной для данного гена – белок окажется неспособным к выполнению своей функции; возникновение стоп-кодонов в последовательности приведет к образованию, так называемых, псевдогенов. И наконец – наиболее интересный вариант – появление в результате мутации белка с новыми свойствами.

Считается что, на определенном этапе эволюции дальнейшее усложнение происходит не за счет создания новых, а за счет комбинации более мелких существующих элементов. Различные виды хромосомных перестроек могут приводить к образованию так называемых гибридных генов, состоящих из частей ранее независимых генов. Гибридные гены могут образоваться в результате объединения (склейки) двух ранее независимых генов или их частей (далее будем называть такие гены склеенными) или посредством вставки одного гена или его фрагмента внутрь другого.

Изучение нуклеотидных последовательностей в настоящее время очень актуально, так как дает возможность получить принципиально новую информацию. В качестве такой новой информации могут выступать данные о местах склеек генов или их фрагментов, которые были осуществлены в ходе эволюции. Программы, основанные на выравниваниях, которые в настоящее время используются в качестве основного инструмента для компьютерного предсказания склеек, имеют свои ограничения. Эти ограничения, прежде всего, связаны с поиском предковых последовательностей в банках данных, из которых могли образоваться гибридные гены. Последовательности предшественников могут отсутствовать в базе данных либо потому, что еще не были секвенированы, либо потому, что были утеряны в процессе

эволюции. Кроме того, они могли настолько измениться в процессе эволюции, что не распознаются существующими программами. Потому дополнительные методы, которые могли бы предсказывать места склеек генов или фрагментов генов без использования дополнительной информации в виде баз данных, имеют большое значение.

В качестве метода поиска мест склеек можно предложить метод обнаружения точек разладки триплетной периодичности в нуклеотидных последовательностях. Задача о разладке состоит в нахождении точки изменения статистических свойств последовательности. Такие изменения отражают внутренние изменения исследуемого процесса, детектируемые математическими методами. Впервые они были разработаны для контроля качества на производстве. Позже некоторые из этих методов были применены к последовательностям ДНК. Примером тому может быть выделение изохор, отделение кодирующих участков от некодирующих и т.д. Все эти методы используют статистические свойства последовательностей без использования выравниваний. Большинство работ по поиску точек разладки посвящены изучению неоднородности последовательностей ДНК на уровне геномов. В то же время существование неоднородностей возможно и на уровне отдельных генов, так как в генах существует так называемая триплетная периодичность. Она отсутствует в некодирующих областях генома и интронах. Триплетная периодичность характеризуются неравномерным распределением нуклеотидов в различных позициях кодонов.

Точки разладки триплетной периодичности в последовательности, могут отражать эволюционные изменения, приведшие к формированию данной последовательности. Исследование, посвященное классификации триплетной периодичности, показало, что триплетная периодичность большинства генов может быть отнесена к одному из примерно 2500 классов. Это позволяет предположить, что если некоторый ген был сформирован, в результате вставки или склейки двух последовательностей, триплетная периодичность которых значительно отличалась, то на границе будет присутствовать точка разладки триплетной периодичности. В этом месте статистические свойства триплетной периодичности будут резко меняться. Это означает, что можно разработать математические алгоритмы по поиску таких точек разладки в генах. Найденные таким образом точки разладки триплетной периодичности будут предсказывать

существование в местах разладки склеек генов или их фрагментов. Эти данные могут помочь в выстраивании картины эволюционной изменчивости генов. Полученные результаты могут быть применены для создания искусственных гибридных генов.

Степень научной разработанности проблемы

Уже несколько десятилетий известно, что последовательности ДНК, кодирующие белки, обладают свойством триплетной периодичности. За это время разработано большое число методов для ее определения, такие как корреляционные методы, методы, основанные на динамическом программировании, преобразовании Фурье, вэйвлет-преобразованиях, информационные методы и некоторые другие. Триплетная периодичность нашла свое применение в компьютерных программах анализа последовательностей ДНК, таких как программы для поиска генов эукариотических и прокариотических геномов. Все разработанные методы для поиска точек разладки в последовательностях ДНК направлены на установлении факта наличия или отсутствия триплетной периодичности на данном участке последовательности, а не на определении различия триплетной периодичности двух участков. Математические методы по сравнению триплетной периодичности отдельных районов генов ранее не были разработаны. Поэтому для поиска точек разладки требуется разработать метод, позволяющий изучать изменения триплетной периодичности вдоль некоторой последовательности. Также ранее не изучался вопрос о существовании точек разладки триплетной периодичности в реальных последовательностях генов.

Цель работы

Целью работы является разработка новых математических алгоритмов для поиска точек разладки в нуклеотидных последовательностях генов и проведение поиска и изучения точек разладки триплетной периодичности в кодирующих последовательностях различных генов.

Задачи исследования

Для достижения поставленной цели был определен следующий список задач:

Разработка математического метода для поиска точек разладки триплетной периодичности в кодирующих последовательностях ДНК с учетом возможного сдвига рамки считывания

Разработка и тестирование программного обеспечения, реализующего метод поиска точек разладки триплетной периодичности в кодирующих последовательностях ДНК с учетом возможного сдвига рамки считывания.

Обработка при помощи разработанного программного обеспечения кодирующих последовательностей банка данных KEGG.

Изучение найденных случаев точек разладки триплетной периодичности посредством поиска подобий в банке данных Swiss-Prot.

Разработка метода поиска парных точек разладки триплетной периодичности в кодирующих последовательностях ДНК с учетом возможных сдвигов рамки считывания.

Разработка и тестирование программного обеспечения, реализующего метод поиска парных точек разладки триплетной периодичности в кодирующих последовательностях ДНК с учетом возможных сдвигов рамки считывания.

Изучение при помощи разработанного метода поиска парных точек разладки триплетной периодичности кодирующих последовательностей 17 бактериальных геномов.

Исследование распределения триплетной периодичности генов внутри одного генома и генов, принадлежащих разным геномам. С целью оценки того, какой процент генов при склейке генов или их фрагментов может привести к появлению точки разладки триплетной периодичности.

Личный вклад автора

1. Разработка алгоритмов поиска одинарных и парных точек разладки триплетной периодичности в кодирующих последовательностях ДНК.

2. Разработка и тестирование программного обеспечения, реализующего алгоритмы поиска точек разладки в кодирующих последовательностях ДНК.

3. Отладка параметров работы программ поиска точек разладки триплетной периодичности при помощи имитационного моделирования.

4. Создание версии программы поиска одинарных точек разладки триплетной периодичности в кодирующих последовательностях ДНК для параллельной обработки на вычислительном кластере. Обработка реальных биологических последовательностей из банка данных KEGG.

5. Анализ связи найденных случаев точек разладки триплетной периодичности с различными биологическими причинами с использованием известных программ.

6. Разработка и тестирование программного комплекса для сравнения триплетной периодичности генов, принадлежащих одному геному, и генов, принадлежащих разным геномам.

Научная новизна

Данная работа обладает научной новизной, так как в ее рамках:

Впервые разработан математический метод поиска одинарных точек разладки триплетной периодичности в кодирующих последовательностях ДНК с учетом возможного сдвига рамки считывания и реализующее его программное обеспечение.

Впервые разработан алгоритм поиска парных точек разладки в кодирующих последовательностях ДНК и реализующее его программное обеспечение.

Впервые показано существование множества одинарных и парных точек разладки триплетной периодичности как в генах прокариот, так и в генах эукариот.

Впервые проведено исследование распределения триплетной периодичности на множестве генов, принадлежащих одному геному, и разным геномам.

Теоретическое и практическое значение работы

Теоретическое значение работы состоит в демонстрации того факта, что триплетная периодичность белок-кодирующих последовательностей ДНК неоднородна на протяжении одной последовательности и может содержать статистически значимые точки разладки, как одинарные, так и парные. Эти точки разладки указывают на возможность существования склейки фрагментов ДНК в данном месте гена. Это означает, что разработанные методы имеют предсказательную функцию.

Практическое значение созданных алгоритмов, программного обеспечения и полученных результатов состоит в следующем. Разработаны методы поиска одинарных и парных точек разладки триплетной периодичности в кодирующих последовательностей ДНК и реализующее их оригинальное программное обеспечение. Исследование точек разладки триплетной периодичности в кодирующих последовательностях ДНК может быть использовано для поиска генов, образованных в результате вставки или склейки. Исследование таких последовательностей может дать ключ к более глубокому пониманию эволюции генов. Новый метод поиска склеенных генов может дать информацию о том, как формируются новые белки в процессе

эволюции и дать возможность для разработки новых методов создания искусственных ферментов. Возникает возможность объединения фрагментов генов в тех позициях, которые были уже использованы в ходе эволюции для создания гибридных генов. Специфичность триплетной периодичности к определенному геному может быть использована для поиска генов, появившихся в геноме в результате горизонтального переноса.

Положения, выносимые на защиту

Метод поиска одинарных точек разладки триплетной периодичности с учетом возможного сдвига рамки считывания в белок-кодирующих последовательностях ДНК.

Метод поиска парных точек разладки, на основании мер различия и подобия матриц триплетной периодичности с учетом возможного сдвига рамки считывания в белок-кодирующих последовательностях ДНК.

Исследование однородности триплетной периодичности генов, принадлежащих одному геному и генов, принадлежащих разным геномам.

Достоверность научных результатов

Достоверность полученных результатов основана на использовании методов математической статистики и тестировании разработанных алгоритмов с использованием модельных объектов с заранее известными свойствами. Достоверность также проверялась путем сравнения полученных результатов с результатами, полученными ранее как теоретическими, так и экспериментальными методами.

Методология и методы исследования.

Теоретические основы исследования составили научные труды широкого круга отечественных и зарубежных ученых в области анализа символьных последовательностей. В диссертационной работе для решения поставленных задач использовались следующие методы исследования: методы теории вероятностей, математической статистики, методы теории информации и методы математического моделирования.

Апробация работы

Основные результаты, представленные в данной диссертационной работе, докладывались на следующих конференциях: международной конференции “Новые информационные технологии в медицине, биологии, фармакологии и экологии”, Гурзуф, Украина. 2010; III и IV международных конференциях “Математическая

биология и биоинформатика”, Пущино, в 2010 и 2012 годах; Четвертой международной конференции для молодых ученых “Молекулярная биология: достижения и перспективы” Киев, Украина, 2011; Школе-конференции молодых ученых “Фундаментальная наука для биотехнологии и медицины-2011” Москва, Россия; Конференции «Методы математической физики и математическое моделирование физических процессов», проводимой в рамках «Научной сессии НИЯУ МИФИ-2012» Москва, 2012; Средиземноморской конференции по встроенным вычислениям (MECO 2012), Бар, Черногория, 2012; Конференции, посвященной сложности генома, проводимой в рамках Европейской конференции по сложным системам, Брюссель, Бельгия, 2012 и межлабораторном семинаре Центра «Биоинженерия» РАН, Москва, 2015.

Публикации по теме диссертации

По материалам диссертации опубликовано 15 печатных работ, из них 5 работ - в рецензируемых научных изданиях и 10 - в материалах научных конференций.

Структура и объем диссертации

Диссертация состоит из введения, трех глав, выводов и списка литературы. Общий объем работы составляет 135 страниц, в том числе 28 рисунков, 7 таблиц и список литературы из 142 наименований.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Основной объект исследования данной работы - это последовательности ДНК, кодирующие белки. В случае прокариот представлены гены, а в случае эукариотических геномов кодирующие последовательности представлены объединением соответствующих экзонов (без интронов). Основной интерес представляют последовательности, содержащие точки разладки триплетной периодичности. Будем говорить, что последовательность содержит точки разладки триплетной периодичности, если ее можно разделить на однородные участки таким образом, что внутри участков триплетная периодичность похожа, а между ними триплетная периодичность значимо отличается (границы участков являются точками разладки триплетной периодичности).

Поиск одинарных точек разладки триплетной периодичности

Рассмотрим кодирующую последовательность S состоящую из символов четырехбуквенного алфавита $\mathrm{A} = \{A, C, T, G\}$. Обозначим $s(x_k)$ – символ, стоящий в последовательности S в позиции k, $S(x, x_k)$ – участок последовательности S. Задача первой части исследования – разработать алгоритм для определения точек, разделяющих участки последовательности, между которыми триплетная периодичность значимо отличается. Такие точки будем называть точками разладки триплетной периодичности. Матрицей триплетной периодичности назовем частотную матрицу размером 4×3: строки такой матрицы соответствуют символам алфавита A (i=1для 'A', i=2; для 'T', i=3, для 'G' и i=4 для 'C'); столбцы – трем позициям кодона. Таким образом, элемент матрицы m_{ij} соответствует числу нуклеотидов типа i, стоящих на позиции кодона j рассматриваемой последовательности. Матрицу триплетной периодичности, построенную для участка последовательности от позиции x_1 до позиции x_2, будем обозначать $M(x_1, x_2)$.

Рассмотрим позицию x в последовательности S и два смежных участка равной длины l справа и слева от x. Построим матрицы триплетной периодичности для этих участков: $N = N(x,l) = M(x-l+1, x)$ и $M_1 = M_1(x,l) = M(x+1, x+l)$. Чтобы учесть возможный сдвиг рамки считывания рассмотрим две дополнительные матрицы для второго участка: $M_2 = M_2(x,l) = M(x+2, x+l+1)$ и $M_3 = M_3(x,l) = M(x+3, x+l+2)$. Для того чтобы избежать влияния обогащенности участка определенными символами на конечную меру, каждая рассматриваемая матрица приводится к нормальному виду при помощи следующего поэлементного преобразования:

$$w(i,j) = \frac{m(i,j) - lp(i,j)}{\sqrt{lp(i,j)(1-p(i,j))}} \qquad (1)$$

Где

$$p(i,j) = (x(i)y(j)) \Big/ l^2 \qquad (2)$$

Где $m(i,j)$- элемент одной из матриц M_1, M_2 или M_3, $x(i)$ и $y(j)$ – соответствующие маргинальные суммы. В результате такого преобразования $n(i,j) \sim N(0,1)$ для всех значений i и j. Обозначим преобразованную матрицу левого участка V, а матрицы правого участка - W_1, W_2 и W_3, соответственно.

Различие между матрицей V и каждой из матриц W_k рассчитывается по следующей формуле:

$$D_k(x,l) = \sum_{i=1}^{4}\sum_{j=1}^{3}\left(\frac{v(i,j) - w_k(i,j)}{\sqrt{2}}\right)^2 (k = 1,2,3) \qquad (3)$$

С учетом возможного сдвига рамки считывания, для каждой позиции x выбирается минимум $D_{\min} = \min_{k=1,2,3}(D_k(x,l))$. Итоговая мера определялась по формуле: $F = -\log_{10}(p(D_{\min} > X))$.

Перемещая скользящий указатель x вдоль последовательности (с шагом в три основания) и синхронно варьируя границы участков справа и слева от x от 60 до 600 нуклеотидов с шагом кратным трем, производится поиск максимума итоговой величины различия. Найденный максимум является потенциальной точкой разладки триплетной периодичности. Для того, чтобы определить его значимость требуется выбрать пороговый уровень на множестве случайных последовательностей. Если значение максимума превосходит выбранный порог, точка разладки считается значимой.

При выборе порогового значения генерация случайных последовательностей, осуществлялась путем перемешивания последовательностей генов с сохранением триплетной периодичности. Полученные таким образом последовательности имеют ту же длину, частоты символов и уровень триплетной периодичности, что и исходная последовательность. При этом перемешивание должно выровнять все имеющиеся неоднородности триплетной периодичности исходного гена.

Поиск парных точек разладки триплетной периодичности

Во второй части исследования основное внимание было уделено поиску генов, содержащих парные точки разладки триплетной периодичности. Будем говорить, что последовательность содержит парные точки разладки триплетной периодичности, если ее можно разделить на три последовательных участка, таких что первый и третий участки обладают похожей триплетной периодичностью, а триплетная периодичность второго участка отличается от них. Таким образом, одна точка разладки разделяет первый и второй участок, другая – второй и третий участки. Схема такой последовательности приведена на рисунке ниже (Рисунок 1).

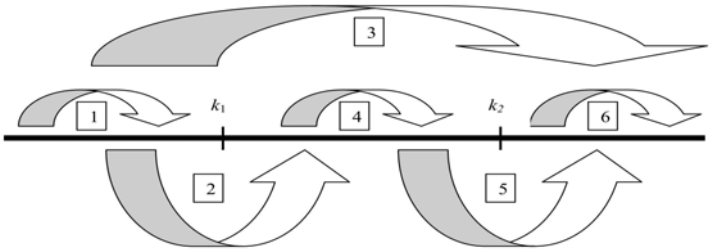

Рисунок 1. Схема последовательности с парными точками разладки триплетной периодичности с указанием соответствующих элементов формулы (7). Снизу стрелками показано различие триплетной периодичности участков, сверху – подобие.

Для поиска парных точек разладки требуется мера подобия триплетной периодичности. Для введения меры подобия рассмотрим последовательность S длины L. Выберем две координаты x_1 и x_2 ($0 < x_1 < x_2 < L - l$) и выделим два участка длины l: $[x_1, x_1 + l)$ *и* $[x_2, x_2 + l)$. Для каждого из них построим частотные матрицы $M_1 = M(x_1, l) = [m_1(\mathrm{i}, \mathrm{j})]_{4\times3}$ и $M_2 = M(x_2, l) = [m_2(\mathrm{i}, \mathrm{j})]_{4\times3}$

Рассмотрим нулевую гипотезу о том, что матрицы M_1 и M_2 являются случайными, нескоррелированными матрицами. Приведем матрицы к нормальному виду, используя поэлементное преобразование по формулам (1) и (2). Полученные в результате такого преобразования матрицы, обозначим N_1 и N_2. И построим новую матрицу $Z = [z(i, j)]_{4\times3}$, перемножая соответствующие элементы матриц N_1 и N_2

$$z(i, j) = n_1(i, j) \cdot n_2(i, j) \tag{4}$$

Так как случайная величина, являющаяся произведением двух стандартных нормально распределенных случайных величин, имеет плотность распределения [Craig, 1936] $f(z) = \pi^{-1} K_0(|z|)$ (K_0 – модифицированная функция Бесселя второго рода). То для любого $z(i, j)$ можно рассчитать вероятность $P(z > z(i, j))$. Используя обратную функцию нормального распределения, можно найти соответствующую величину $y(i, j)$, удовлетворяющую условию $P(y > y(i, j)) = P(z > z(i, j))$. После чего, суммируя значения, рассчитанные для отдельных ячеек, получим

$$S(x_1, x_2) = \sum_{i=1}^{4} \sum_{j=1}^{3} y(i, j) \tag{5}$$

Таким образом, в рамках нулевой гипотезы $S(x_1,x_2)\sim N(0,6)$, а величина $P(N(0,6)>S(x_1,x_2))$ есть вероятность того, что сходство двух матриц обусловлено случайными совпадениями. При больших значениях $S(x_1,x_2)$ эта вероятность позволяет отвергнуть нулевую гипотезу.

Алгоритм поиска парных точек разладки

Последовательность S разобьем точками $x_k = step\cdot(k-1)+1;\ k=1,2\ldots K$. Для каждой позиции x_k заполним матрицы $M(x_k,l)$. Всего $K=\lfloor (L-l)/step \rfloor+1$ матриц должно быть рассчитано (размер рассматриваемых участков l = 60, величина шага $step=9$). Затем, сравнивая K частотных матриц каждую с каждой, заполняем две большие матрицы $Sim=[sim(i,j)]_{K\times K}$ и $Dif=[dif(i,j)]_{K\times K}$

$$\begin{aligned} sim(i,j) &= S(x_i,x_j) \\ dif(i,j) &= D(x_i,x_j) \end{aligned} \qquad (6)$$

Элементы матрицы *Sim* рассчитываются по формуле (5) и отражают подобие, а элементы матрицы *Dif* - различие соответствующих участков аналогичное тому, что использовалось в первой части работы. Далее для всех k_1 и k_2 $(1\le k_1<k_2\le K)$ вычислялись значения

$$\begin{aligned} W_1(k_1,k_2) = &\sum_{1\le i<k_1}\sum_{1\le j<k_2} sim(i,j) + r\sum_{1\le i<k_1}\sum_{k_1\le j\le k_2} dif(i,j) + \sum_{1\le i<k_1}\sum_{k_2<j\le K} sim(i,j) \\ &+ \sum_{k_1\le i\le k_2}\sum_{k_1\le j\le k_2} sim(i,j) + r\sum_{k_1\le i\le k_2}\sum_{k_2<j\le K} dif(i,j) + \sum_{k_2<j\le K}\sum_{k_2<j\le K} sim(i,j) \end{aligned} \qquad (7)$$

Чтобы проиллюстрировать смысл выражения (7), рассмотрим случай, когда рассматриваемая последовательность содержит вставку кратную трем между позициями, соответствующими k_1 и k_2 (случай не кратной трем вставки будет рассмотрен ниже). Тогда первое, четвертое и седьмое слагаемые уравнения (7) отражают подобие внутри интервалов $(1,x_{k_1}),(x_{k_1},x_{k_2}),(x_{k_2},x_K)$. Второе и пятое слагаемые отражают различие триплетной периодичности между участками $(1,x_{k_1})-(x_{k_1},x_{k_2})$ и $(x_{k_1},x_{k_2})-(x_{k_2},x_K)$ соответственно. А третье слагаемое – подобие

триплетной периодичности между участками $(1, x_{k_1})$ и (x_{k_2}, x_K). Схема, иллюстрирующая формулу представлена на рисунке (Рисунок 1).

Также возможен сдвиг рамки считывания на один или два символа после второй точки разладки (что соответствует вставке, длина которой не кратна трем). Для того, чтобы учесть такую возможность, в формуле (7) в третьем слагаемом нужно заменить матрицу первой рамки на втором участке, на матрицу, соответствующую второй или третьей рамке.

Гены, содержащие одинарные точки разладки, также могут приводить к превышению порога итоговой величины для данной последовательности. Следовательно, каждый ген, в котором было найдено итоговое значение, превышающее соответствующий порог должен быть дополнительно исследован на предмет наличия одинарной точки разладки. Процесс поиска одинарных точек разладки похож на поиск парных, описанный выше, но здесь рассматривается только одна координата k_1 и вычисление величины $W_1(k_1, k_2)$, заменяется величиной

$$W_1(k_1) = \sum_{1 \le i < k_1} \sum_{1 \le j < k_1} sim(i, j) + r \sum_{1 \le i < k_1} \sum_{k_1 \le j \le K} dif(i, j) + \sum_{k_1 \le i < K} \sum_{k_1 < j \le K} sim(i, j) \qquad (8)$$

Итоговую величину обозначим V. Таким образом, если итоговое значение, полученное для одинарной точки разладки превосходит значение, соответствующее парной точке разладки, то данное событие рассматривается как парная точка разладки, иначе как одинарная.

Исследование распределения триплетной периодичности внутри геномов

Для того чтобы выяснить, насколько триплетная периодичность однородна внутри генома и проверить какой процент генов в результате склейки может привести к образованию точек разладки, мы попарно сравнили матрицы триплетной периодичности генов, принадлежащих одному геному. И построили распределение величины различия триплетной периодичности на множестве генов внутри одного генома с использованием меры различия матриц триплетной периодичности (формула (3)). Аналогичные распределения были построены на модельных множествах. Были созданы два типа модельных последовательностей, соответствующие двум

альтернативным гипотезам распределения триплетной периодичности на некотором множестве последовательностей.

Первая гипотеза состоит в однородности триплетной периодичности на некотором множестве. В рамках этой гипотезы мы предполагаем, что все последовательности множества произошли от одного типа периодичности и соответственно различие триплетной периодичности последовательностей такого множества должно быть невелико. Соответствующие этой гипотезе модельные последовательности строятся из совершенной периодической последовательности (повтор периода три) для всего множества используется периодичность одного типа. Для каждой рассматриваемой последовательности *S* строится совершенная периодическая последовательность такой же длины. Затем уровень триплетной периодичности в периодической последовательности постепенно размывается путем случайных перестановок, до уровня триплетной периодичности исходной последовательности. Вторая гипотеза - “случайная”. Она состоит в том, что тип триплетной периодичности последовательностей внутри некоторого множества генов случаен и независим от остальных. При моделировании, соответствующем второй гипотезе, использовались случайные последовательности. Для создания одной последовательности такого типа, символы рассматриваемой кодирующей последовательности *S* были перемешаны случайным образом (без сохранения периодичности). Затем, в полученной последовательности путем случайных перестановок символов, восстанавливался уровень триплетной периодичности до уровня исходной последовательности.

Программная реализация используемых алгоритмов

Программная реализация основных алгоритмов поиска точек разладки триплетной периодичности и анализа распределения триплетной периодичности внутри геномов и между геномами была выполнена на языке программирования C++ как приложения для операционной системы Linux. Для сокращения времени расчетов была разработана параллельная версия программы, с использованием библиотеки MPI и подхода, основанного на параллелизме данных. Вычисления проводились с использованием вычислительных ресурсов Центра «Биоинженерия» РАН и МСЦ РАН.

Программы для дальнейшего анализа результатов, сравнение с другими банками данных реализованы на языке Python.

РЕЗУЛЬТАТЫ

Результат поиска одинарных точек разладки триплетной периодичности

В результате моделирования с использованием множества искусственных последовательностей был выбран пороговый уровень, соответствующий 5% вероятности ошибки первого рода. С учетом выбранного уровня были проанализированы кодирующие последовательности банка данных KEGG/Genes-48 (около 4 миллионов генов). В результате был найден 311 221 ген с точкой разладки триплетной периодичности. Пример последовательности с точкой разладки триплетной периодичности приведен на рисунке (Рисунок 2).

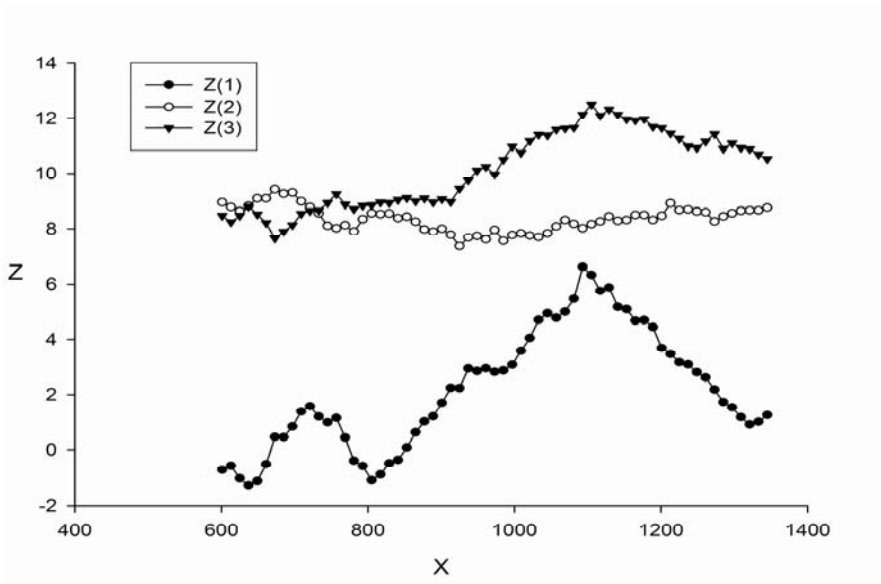

Рисунок 2. Профиль изменения триплетной периодичности по всем трем рамкам считывания в аргументах нормального распределения для последовательности гена ECP_0691 генома *E.coli* (*L*=600 п.н.). Точке разладки триплетной периодичности соответствует позиция x = 1100 п.н..

Другой пример последовательности с точкой разладки – это ген B0879 генома *E.coli* известный в литературе, как ген, образованный в результате события склейки [Enright и др., 1999; Serres, Riley, 2005]. В этой последовательности также удалось обнаружить точку разладки триплетной периодичности, соответствующую позиции

склейки – 630 п.н.. График различия триплетной периодичности в аргументах нормального распределения для этой последовательности представлен на рисунке (Рисунок 3).

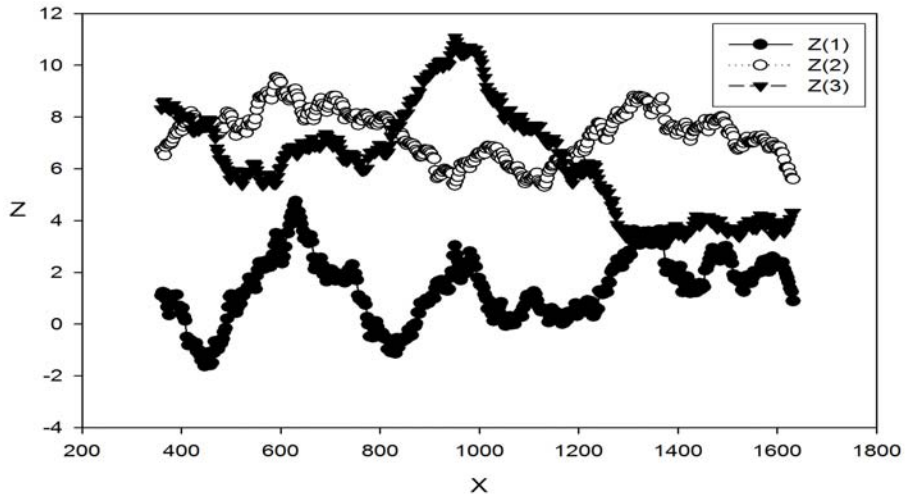

Рисунок 3. Профиль изменения триплетной периодичности по всем трем рамкам считывания в аргументах нормального распределения для последовательности B0879 генома *E.coli* K-12 (размер окна *L*=360 п.н.).

Известно, что аминокислотные последовательности, содержат различные повторы, которые могут влиять на уровень триплетной периодичности соответствующего участка и, следовательно, на разработанную меру. Проведенный анализ показал, что около 7% найденных случаев точек разладки триплетной периодичности могут быть объяснены влиянием аминокислотных повторов.

Для оценки того, какая часть найденных генов, содержащих точку разладки, могла произойти в результате события склейки, был проведен поиск возможных предковых последовательностей с использованием поиска подобий. Рассмотрим последовательность гена, содержащего точку разладки триплетной периодичности в позиции *x*. Если событие склейки имело место не так давно, то возможно будет найти независимые предковые последовательности, сформировавшие данный ген в других или в том же геноме, если они сохранились в процессе эволюции и находятся в банке данных. Таким образом, могут существовать независимые кодирующие последовательности для левой и правой части (от точки *x*) или хотя бы для одной из них. Поиск подобий проведен с использованием программы BLASTx в банке данных

Swiss-Prot. В результате для 131 323 исследуемых последовательностей с точками разладки было найдено подходящее подобие. Среди них 54 406 последовательностей, для которых подобие было найдено только для левого участка (до точки *x*), 60 333 гена получили подобие для правого участка (после позиции *x*). И наконец, для 16 584 случаев подходящее выравнивание было найдено для обоих участков.

Результаты поиска парных точек разладки триплетной периодичности

`В результате обработки 17 бактериальных геномов (общий объем 69 936 генов) разработанным алгоритмом, с учетом выбранного на множестве случайных последовательностей порогового уровня, было найдено 6 459 случаев одинарных точек разладки и 2 700 случаев парных точек разладки триплетной периодичности. Подробная статистика для каждого генома приведена в таблице (Таблица 1).

Таблица 1 – Статистика парных и одинарных точек разладки триплетной периодичности в 17 бактериальных геномах

Геном	**Число генов с одиночной ТР**	**Число генов с ПТР**			**Всего**
		без сдвига	**сдвиг 1**	**сдвиг 2**	
A.butzleri	227	50	23	20	320
A.vinelandii_Ent	477	92	71	64	704
B.avium	232	72	32	14	350
B.mallei	847	150	94	87	1178
B.subtilis	444	114	49	20	627
E.coli	357	70	35	32	494
L.fermentum	170	41	15	19	245
M.capsulatus	281	78	25	26	410
P.aeruginosa	635	142	96	98	971
S.aureus_COL	221	51	17	18	307
S.enterica_Choleraesuis	417	88	60	33	598
S.pneumoniae	150	29	13	8	200
S.sonnei	396	71	35	30	532
S.typhimurium	392	95	50	43	580
V.cholerae	246	48	31	17	342
X.campestris	604	91	63	33	791
Y.pseudotuberculosis_YPIII	363	80	48	19	508
Всего	**6459**	**1362**	**757**	**581**	**9159**

В качестве примера на рисунке (Рисунок 4) представлен контурный график различия триплетной периодичности последовательности гена BSU02140. Более темные области соответствуют большему различию триплетной периодичности соответствующих участков.

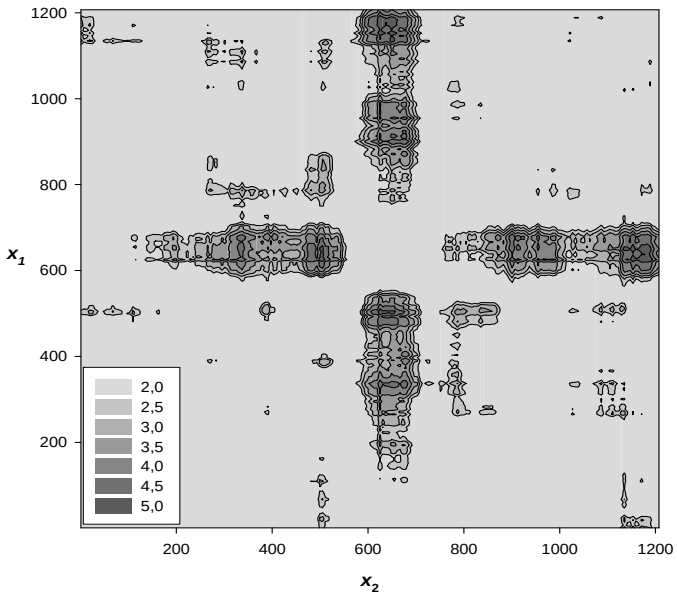

Рисунок 4. Контурная диаграмма различия триплетной периодичности для гена BSU02140. Координаты точек разладки составляют около ~600 и ~700 нуклеотидов соответственно.

Результаты исследования распределения триплетной периодичности внутри геномов и между геномами.

На рисунке (Рисунок 5, А), представлены распределения значений различия триплетной периодичности на множестве генов генома *E.coli* и модельных множествах. Рисунок показывает, что распределение различий между матрицами генов (Real) ближе к распределению, построенному на основании гипотезы об общей триплетной периодичности генома (Perf), чем к распределению, построенному на основании гипотезы о случайном распределении триплетной периодичности (Rand). Лишь немногие последовательности, порядка 15% пар, имеют значительно различающуюся триплетную периодичность. Это именно те последовательности, которые в результате события склейки могут привести к появлению значимой точки разладки триплетной периодичности.

Эта тенденция сохраняется и для других геномов. На рисунке (Рисунок 5, C) представлены только медианы распределений различия для 45 прокариотических геномов, видно, что результат сравнения реальных последовательностей, ближе к результату сравнения последовательностей, полученных из одного типа триплетной периодичности, чем из случайного типа триплетной периодичности.

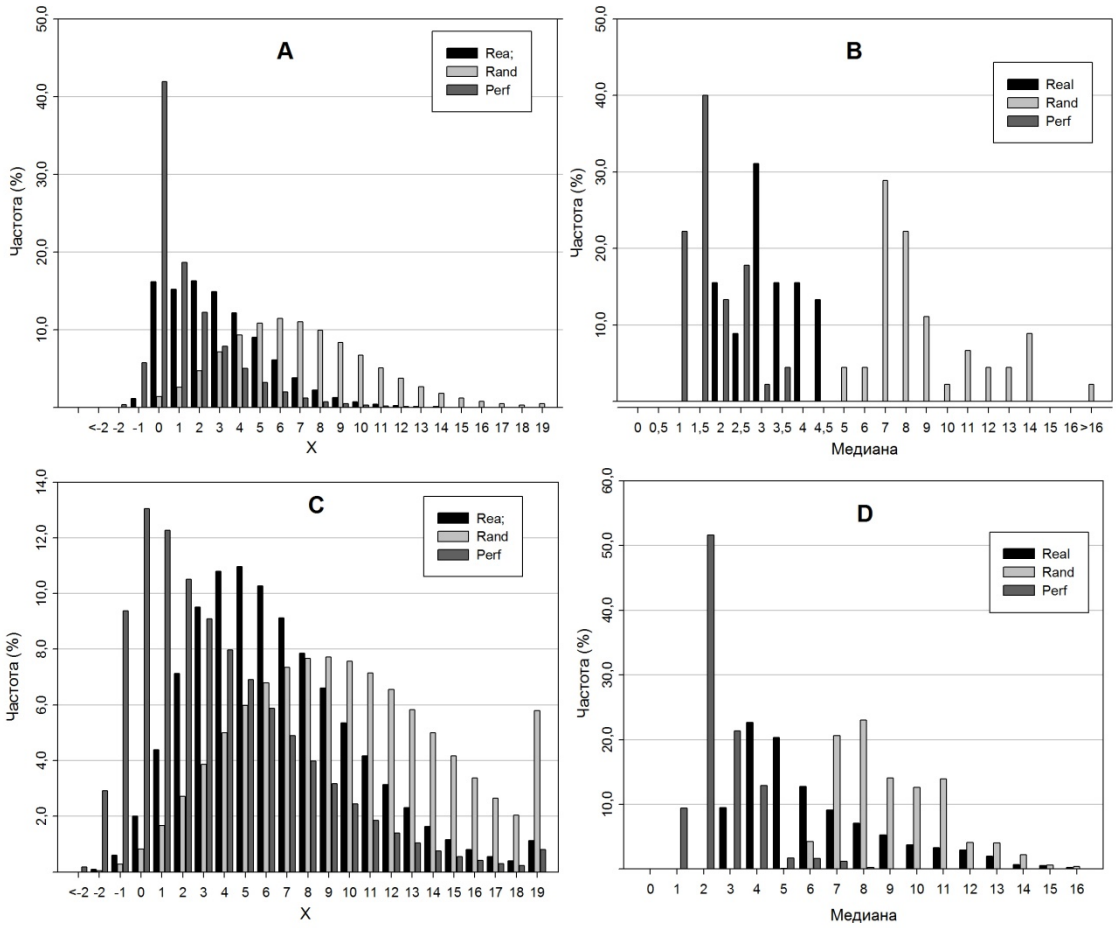

Рисунок 5. (A) Распределения различия триплетной периодичности реальных и модельных последовательностей генома *E.coli*. Real – различие между матрицами кодирующих последовательностей Rand – распределение различия между модельными последовательностями, полученными из случайного типа триплетной периодичности, Perf – различие между модельными последовательностями одного типа триплетной периодичности. (B) Распределение различия триплетной периодичности реальных и модельных последовательностей геномов *E.coli* и *B.avium*. (C) Распределение медиан различий реальных и модельных последовательностей 45 геномов. (D) Распределение медиан различий триплетной периодичности реальных и модельных последовательностей, принадлежащих разным геномам.

При сравнении матриц триплетной периодичности кодирующих последовательностей, принадлежащих двум разным геномам (на примере геномов,

E.coli и *B.avium*, Рисунок 5 (В)) реальное распределение сдвигается вправо (в сторону последовательностей со случайным типом триплетной периодичности), а положение модельных распределений не изменяется. Это говорит о том, что триплетная периодичность между геномами различается сильнее, чем внутри геномов.

Это означает, что триплетная периодичность достаточно однородна внутри геномов, при этом от 15 до 20% (в зависимости от генома) последовательностей внутри генома обладают достаточным различием триплетной периодичности, чтобы в результате склейки привести к появлению значимой точки разладки триплетной периодичности**.**

ЗАКЛЮЧЕНИЕ

В рамках данной диссертационной работы проведено исследование феномена точек разладки триплетной периодичности. Для этого разработаны математические алгоритмы для поиска одинарных и парных точек разладки триплетной периодичности в последовательностях ДНК, кодирующих белки, с учетом возможного сдвига рамки считывания. Алгоритмы основаны на сравнении матриц триплетной периодичности соответствующих участков последовательностей. Создано реализующее эти алгоритмы программное обеспечение. Разработанным программным обеспечением обработаны кодирующие последовательности банка данных KEGG-48. В результате показано, что более 7,5% последовательностей содержат статистически значимые точки разладки триплетной периодичности. Дальнейший анализ показал, что более 40% найденных случаев могут быть объяснены происхождением данной последовательности в результате объединения предковых последовательностей, а около 7% результатов объясняются влиянием аминокислотных повторов. Проведен более детальный анализ кодирующих последовательностей 17 бактериальных геномов, который показал, что 9% кодирующих последовательностей этих геномов содержат одинарные точки разладки триплетной периодичности, а в 4% последовательностей были найдены парные точки разладки триплетной периодичности. Для 45 прокариотических геномов проведено исследование распределения триплетной периодичности на множестве генов, принадлежащих одному геному и генов, принадлежащих разным геномам. Проведена классификация генов для пар геномов на основании матриц триплетной периодичности.

ВЫВОДЫ

1. Разработан метод поиска одинарных точек разладки триплетной периодичности с учетом возможного сдвига рамки считывания в последовательностях ДНК, кодирующих белки. Создано соответствующее программное обеспечение. Обработаны 4 013 150 белок-кодирующих последовательностей банка данных KEGG-48. В результате найдено 311 221 последовательностей, содержащих точки разладки триплетной периодичности. Полученное значение составляет примерно 7,5% от общего числа проанализированных генов

2. Для найденных последовательностей с точками разладки триплетной периодичности проведен поиск подобий в банке данных Swiss-Prot. В результате этого исследования установлено, что 131 323 случая точек разладки триплетной периодичности могли быть вызваны событиями склейки генов (более 40% от общего числа генов с точками разладки триплетной периодичности). Кроме того, около 7% случаев точек разладки триплетной периодичности могут быть отнесены на счет повторяющихся последовательностей и последовательностей низкой сложности.

3. Разработана мера подобия триплетной периодичности, основанная на сходстве соответствующих частотных матриц. На основании этой меры создан алгоритм поиска парных точек разладки триплетной периодичности и реализующее его программное обеспечение. Разработанным методом были исследованы кодирующие последовательности 17 бактериальных геномов (всего 69 936 генов). В результате было найдено 6 459 генов, содержащих одинарные точки разладки триплетной периодичности, и 2 700 генов, содержащих парные точки разладки триплетной периодичности (как без сдвига, так и со сдвигом рамки считывания). Что составляет соответственно ~9% и ~4% от общего числа проанализированных генов.

4. Проведено исследование распределения триплетной периодичности генов, принадлежащих как одному геному, так и различным геномам. Показано, что в большинстве случаев триплетная периодичность однородна внутри генома и специфична по отношению к геному. При этом от 15 до 20% (в зависимости от генома) последовательностей внутри одного генома обладают достаточными различиями в триплетной периодичности, чтобы в результате события склейки привести к появлению значимой точки разладки триплетной периодичности.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в рецензируемых научных журналах

1. Suvorova Y. M., Rudenko V. M., Korotkov E.V. Detection change points of triplet periodicity of gene // Gene. 2012. V. 491, P. 58-64.

2. Суворова Ю. М., Короткова М. А., Коротков Е. В. Изучение точек разладки триплетной периодичности в нуклеотидных последовательностях генов // Прикладная информатика. 2012. Т. 5. С. 75-89.

3. Суворова Ю. М., Коротков Е. В. Изучение геномной специфичности триплетной периодичности генов прокариот // Вестник НИЯУ МИФИ. 2014. Т. 3(2). С. 232-239.

4. Suvorova, Y. M., Korotkova M.A., Korotkov E. V. Study of the Paired Change Points in Bacterial Genes // IEEE/ACM Trans Comput Biol Bioinform. 2014. V. 11(5) P 955 – 964.

5. Suvorova Y. M., Korotkov E. V. Study of triplet periodicity differences inside and between genomes. // Statistical Applications in Genetics and Molecular Biology. 2015. Vol. 14 №2. С. 113-123.

Публикации в трудах конференций:

6. Суворова Ю.М., Коротков Е. В. Поиск склеенных генов в банке данных KEGG // Сборник трудов международной конференции “Новые информационные технологии в медицине, биологии, фармакологии и экологии”. Гурзуф. 2010. С 138-139.

7. Суворова Ю. М., Коротков Е. В. Изучение склеенных генов в банке данных KEGG // Сборник трудов III международной конференции “Математическая биология и биоинформатика”. Пущино. 2010. С 133-134.

8. Suvorova Y. M., Korotkov E. V. Detecting genes with triplet periodicity splicing // Proceed. The International Moscow Conference On Computational Molecular Biology (MCCMB’11). Москва. 2011 С. 358-359.

9. Suvorova Y.M., Korotkov E.V. Splicing of the triplet periodicity in genes from different species. // Proc. of the 6th International Symposium on Health Informatics and Bioinformatics (HIBIT 2011). Измир 2011. С. 245-249.

10. Suvorova Y. M., Korotkov E. V. Changes of triplet periodicity in coding sequences // Abstr. The 4th international IMBG Conference For Young Scientists "Molecular Biology: Advances And Perspectives", Киев. 2011. С. 197.

11. Суворова Ю. М., Коротков Е. В. Метод поиска точек разладки в последовательностях генов. // Научная сессия НИЯУ МИФИ-2012 Москва. 2012. Т. 3. С. 146.

12. Суворова Ю. М., Коротков Е. В. Анализ распределения триплетной периодичности между генами одного генома // IV Международная конференция “Математическая биология и биоинформатика”, Пущино. 2012. С. 61-62.

13. Коротков Е. В., Суворова Ю. М. Изучение одиночных и парных точек разладки в кодирующих последовательностях ДНК // V съезд биофизиков России, Нижний Новгород. 2012. Том. 1 С. 383.

14. Suvorova Y., Korotkov E. Change points in DNA coding sequences // Mediterranean Conference on Embedded Computing. Бар. 2012. С. 251 – 254.

15. Suvorova Y. M., Korotkova M.A., Korotkov E.V. Search of Possible Insertions in Bacterial Genes. // International Conference on Bioinformatics Models, Method and Agorithms. Анже. 2014. С.99-108.