ISSN 2071-8632

1, 2010

РОССИЙСКАЯ АКАДЕМИЯ НАУК

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ

ГЛАВНЫЙ РЕДАКТОР С.В. ЕМЕЛЬЯНОВ

БИОИНФОРМАТИКА И МЕДИЦИНА

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

ИНТЕРНЕТ-ТЕХНОЛОГИИ

POLY PRINT SERVICE

Биоинформатика и поиск сдвигов рамки считывания в генах

М.А. Короткова, Е.В. Коротков

Аннотация. Введено понятие сдвига фазы триплетной периодичности (ТП) нуклеотидной последовательности. Разработан математический алгоритм для выявления сдвига фазы триплетной периодичности в нуклеотидных последовательностях. Были получены последовательности генов из банка данных Kegg-46 и выявлены 27417 тысяч генов (0,7% от общего их числа в базе данных Kegg), где был обнаружен сдвиг фазы триплетной периодичности. Мы предполагаем, что сдвиг фазы триплетной периодичности связан со сдвигами рамки считывания в генах.

Ключевые слова: триплетная периодичность, открытая рамка считывания, сдвиги.

Введение

В настоящее время проводится широкомасштабное определение последовательностей различных геномов. В результате этих работ исследователям стали доступны огромные объемы нуклеотидных последовательностей из различных геномов, представляющих собой текст из четырех символов различной длины. На сегодняшний день полностью известна нуклеотидная последовательность генома человека (3 миллиарда нуклеотидов) и нуклеотидные последовательности многих других эукариотических геномов. В нуклеотидных последовательностях присутствуют гены, которые составляют доли процента от длины эукариотического генома [1]. В настоящее время число известных генов, собранных в Киотской энциклопедии генов и геномов (Kegg) [2], превышает 8 миллионов. Задача состоит в том, чтобы научиться извлекать биологическую информацию из нуклеотидных последовательностей. Этой задачей занимается сравнительно новое направление исследований, называемое биоинформатикой.

Рассмотрим очень кратко задачи, которыми в настоящее время занимается биоинформатика. К числу задач биоинформатики можно отнести следующие проблемы. Во-первых, это проблема анализа и извлечения биологической информации из нуклеотидных и аминокислотных последовательностей [3-4], а также некоторых других временных рядов, возникающих из применения биочипов к изучению активности генов (микроэррей анализ) [5-6]. В данном направлении наиболее актуальной является задача по разработке математических методов поиска генов в геномах эукариот, предсказания альтернативного спалайсинга генов и аннотации нуклеотидных последовательностей. Дело в том, что в геноме клеток эукариот гены разбиты некодирующими вставками (интронами) на части, которые называются экзонами [7-8]. Удаление интронов и сшивание экзонов происходит уже на уровне РНК и этот процесс получил название сплайсинга [7-8]. Оказалось также, что возможен альтернативный сплайсинг, когда интроны вырезаются в альтернативных комбинациях (может меняться как состав вырезаемых интронов, так и их границы), что приводит к тому, что экзоны могут входить в РНК в различных комбинациях [9-10]. Этот процесс позволяет одному гену производить различные белки, число которых может достигать нескольких сотен [9-10]. Наличие интронов и

альтернативного сплайсинга значительно усложняет разработку математических методов поиска генов в эукариотических последовательностях математическими методами. Разработанное в настоящее время программное обеспечение обладает чувствительностью до 70% (т.е. позволяет предсказывать до 70% генов из общего числа в геноме) и специфичностью от 40% до 70% [11-12]. Это означает, что гены, предсказанные разработанными математическими методами, содержат от 30% до 60% неправильных предсказаний.

Второй важной задачей данного направления является задача определения биологической функции нуклеотидных и аминокислотных последовательностей, т.е. задача их аннотации. Для генов в большинстве случаев это означает определение биологической функции того белка, который они кодируют. Даже для геномов бактерий, в генах которых отсутствуют интроны, удается аннотировать разработанными математическими методами в среднем не более 50% генов [13-14]. Кроме того, разрабатываются математические методы для аннотации регуляторных последовательностей, которые управляют активностью генов. К числу таких последовательностей относятся в первую очередь промоторные последовательности, которые определяют транскрипцию данного гена [15]. Промоторные последовательности вызывают наибольший интерес для разработки математических методов поиска в геномах изучения из-за того, что их правильное предсказание и возможная классификация последовательностей промоторов позволит более точно находить гены и перейти к конструированию динамических моделей генетических сетей только на основе знаний о нуклеотидных последовательностях. На сегодняшний день разработано несколько математических подходов к поиску промоторов, однако разработанные методы дают вероятность ошибочного предсказания не менее $5x10^{-5}$ [16-17]. Так как длина генома человека составляет порядка $3 x10^{9}$ нуклеотидов и в нем содержится около $3x10^{4}$ генов, средняя длина геномной ДНК, содержащей один ген (соответсвенно, по крайней мере с одним промотором), составляет ~ 10^{5}. На этой длине будет сделано несколько ошибочных предсказаний, что делает достаточно проблематичным обнаружение истинных промоторов. Разработанные подходы могут сравнительно хорошо применяться на бактериальных геномах, где средняя длина гена составляет примерно 10^{3} нуклеотидов.

Еще одной проблемой биоинформатики является создание разнообразных биологических баз данных, чтобы иметь возможность быстрого доступа и получения информации из огромных массивов данных, которыми располагает современная биология [18]. К таким данным следует отнести, в первую очередь, все известные нуклеотидные последовательности, объем которых в банке данных Genbank составляет 100 гигабайт [19]. В ближайшей перспективе объем такого типа данных может значительно увеличиться, так как станет возможным определение полной нуклеотидной последовательности генома каждого человека (3 миллиарда нуклеотидов) за сутки или мне того и за сравнительно небольшую плату (несколько десятков тысяч рублей) [20]. Предполагается, что эти данные могут быть использованы в медицине для определения предрасположенности к различным заболеваниям и для определения индивидуальной чувствительности к различным лекарственным препаратам [21-23].

Третьей проблемой биоинформатики является также разработка динамических моделей генетических сетей, что позволит проектировать генетические сети на компьютере. Задачей таких моделей становится динамическое моделирование взаимодействия как генов, так и различных белков [24-25]. Фактически решается так называемая обратная задача и восстанавливается взимодействие генов по данным изменения активности генов во времени [5-6]. Развитие этого направления позволит поднять уровень современной биоинженерии на новую высоту, когда внедрение генетических конструкций будет проводиться вполне осмысленно. Это означает, что можно будет проверить работу искусственной генетической конструкции на компьютере в динамическом режиме, а затем уже можно будет ее реально создать и внедрить в клетку. В конечном счете, данное направление позволит сконструировать искусственный геном и искусственную клетку. На начальном

этапе это может быть геном, который позволит работать так называемой минимальной клетке [26-27), т.е. клетке, способной осуществлять только функции поддержания собственной жизни и размножения. По оценкам, для этого будет достаточно несколько сотен генов. По мере развития этого направления можно будет сконструировать и новые искусственные биологические виды на компьютере с последующей экспериментальной реализацией такого нового генома методами биоинженерии.

В четвертых, биоинформатика также занимается моделированием и изучением трехмерных структур белков и нуклеотидных последовательностей [28]. Создан банк данных трехмерных структур белков [29]. Разработаны различные математические подходы для сравнения пространственных структур белков [30]. Это направление тесно связано с моделированием лекарственных препаратов, так как механизм действия многих лекарственных препаратов связан с взаимодействием с белковыми молекулами [31].

Мы перечислили только некоторые наиболее актуальные задачи биоинформатики на сегодняшний день. В данной работе мы предлагаем новый математический метод поиска потенциальных сдвигов рамки считывания. Под сдвигом рамки считывание понимается изменение «нарезки кодирующей последовательности ДНК на триплеты после делеции или же вставки некратного трем количества оснований ДНК (Рис.1)[1]. Последовательность оснований ДНК генома не является неизменной, так как в ходе существования живых организмов в генетическом тексте накапливаются определенные изменения, т.е. мутации. Мутации в последовательностях оснований ДНК осуществляются посредством замен оснований, а также путем делеций и вставок как отдельных нуклеотидов, так и целых фрагментов ДНК [32]. На уровне аминокислотных последовательностей белков замены оснований ДНК могут приводить к заменам аминокислот, т.е. от одной замены нуклеотида может измениться только одна аминокислота в аминокислотной последовательности кодируемого геном белка (Рис 2.). Достаточно часто такие аминокислотные замены могут оказывать серьезное влияние на способность белка выполнять биологическую функцию [1]. Однако при делециях и вставках оснований может произойти изменение протяженного участка аминокислотной последовательности из-за сдвига рамки считывания (Рис 1.). В силу этого делеции и вставки оснований можно рассматривать как более существенные эволюционные события для белков, чем замены оснований. Влияние сдвигов рамки считывания в генах на структуру и функцию белка сравнительно мало изучено. Это связано с трудностью регистрации сдвигов рамки считывания, хотя изучение влияния сдвигов рамок считывания в генах на структуру белков представляет большой интерес. Если при замене одного основания ДНК в белке может измениться только одна аминокислота, то при сдвиге рамки считывания (не кратном 3 основаниям) изменяются все аминокислоты, которые лежат ниже точки сдвига по последовательности гена. Если после этого белок не потерял своей функции, то можно предполагать, что сдвиг рамки произошел в функционально малозначимом месте или же что сдвиг рамки считывания порождает аминокислотные последовательности, имеющие похожие функции. Очень интересно определить закономерности изменения аминокислот в последовательности, которые позволяют создать новую аминокислотную последовательность с таким же функциональным значением как у исходной последовательности. Если же после сдвига рамки считывания белок изменил свою функцию, то очень интересно понять, какие изменения аминокислотной последовательности приводят к созданию новой функции белка. Эти результаты могут быть использованы для проектирования искусственных белков с заданной биологической функцией.

Задача данной работы состоит в разработке математического метода, который позволил бы возможно более полно найти сдвиги рамок считывания в существующих генах. В настоящее время основным методом поиска сдвигов рамки считывания и инверсий является поиск подобий между аминокислотными последовательностями при помощи программы Blast или ей подобных [33-34]. При этой процедуре поиска сдви-

[1] Все рисунки и таблицы даны в приложении в конце статьи.

гов рамки считывания нужно каким-либо способом выделить участок гена, где мы предполагаем сдвиг рамки считывания. Затем этот участок нуклеотидной последовательности нужно перекодировать в аминокислотную последовательность по новой рамке считывания и получить гипотетическую аминокислотную последовательность. После этого проводится поиск подобий для гипотетической аминокислотной последовательности в банке данных аминокислотных последовательностей Swiss-prot. Если будут найдены статистически значимые подобия, то достаточно уверенно можно утверждать, что в данном гене был сдвиг рамки считывания и что последовательности, родственные гипотетической аминокислотной последовательности, существует на самом деле. Этим способом в настоящее время удалось найти несколько сотен генов, где с большой уверенностью можно предполагать присутствие сдвига рамки считывания [33-35].

В этой схеме поиска сдвигов рамок считывания есть определенные ограничения. Во-первых, нужно по каким-то признакам выбрать ген, где предполагается сдвиг рамки считывания, затем найти в нем возможное место сдвига рамки считывания. Общий поиск сдвига рамок считывания по всем генам может потребовать достаточно больших компьютерных ресурсов. Во-вторых, даже если мы и решим первую задачу, необходимо, чтобы банк данных Swiss-prot содержал аминокислотную последовательность, имеющую статистически значимое подобие с гипотетической аминокислотной последовательностью. Однако такая последовательность может отсутствовать ввиду ограниченности объема банка данных Swiss-prot или из-за слишком больших эволюционных различий, накопленных между аминокислотными последовательностями. В силу этого используемый подход может выявить только некоторую часть сдвигов рамок считывания, накопленных в существующих генах к настоящему времени.

Для того чтобы более уверенно выделять сдвиги рамки считывания в генах, необходимо разработать другой способ поиска сдвигов рамок считывания вместо поиска подобий между гипотетическими и реальными аминокислотными последовательностями. В качестве сигнала о существовании сдвига рамки считывания в нуклеотидной последовательности гена, как мы показываем в данной работе, может выступать сдвиг фазы триплетной периодичности. Триплетная организация последовательностей ДНК, кодирующих белки, является общим свойством всех известных в настоящее время живых систем [36-43] и она привязана к рамке считывания, существующей в гене [44]. Причина этого заключается как в структуре генетического кода, который практически одинаков как у представителей прокариот, так и у эукариот, так и в насыщенности белков определенными аминокислотами [45-48]. Если на фоне триплетной периодичности в гене произойдет сдвиг рамки считывания, то это можно будет заметить, так как произойдет сдвиг между триплетной периодичностью и рамкой считывания (Рис.3). Поскольку триплетную периодичность последовательности ДНК достаточно трудно существенно изменить посредством сравнительно небольшого числа замен оснований [49], то такой сдвиг может сохраняться сравнительно долго. Присутствие такого сдвига между триплетной периодичностью нуклеотидной последовательности и рамкой считывания может служить указанием на сдвиг рамки считывания в анализируемом гене.

Для выявления триплетной периодичности в настоящее время разработаны методы, использующие регулярность в предпочтении символов в различных позициях триплета в последовательностях ДНК. В качестве математического аппарата в них использовались преобразование Фурье, скрытые цепи Маркова и другие статистические методы, основанные на позиционно-зависимых предпочтениях нуклеотидов в кодирующих последовательностях [50-55]. Применяемые методы использовались для выявления кодирующих последовательностей ДНК и их отделения от некодирующих участков. Позже для поиска триплетной периодичности был предложен метод информационного разложения [49,56], который позволяет ввести понятие класса триплетной периодичности в виде матрицы размерности 4x3. В матрице признаками столбцов являются позиции периода, а признаками строк являются нуклеотиды.

В работе ставились две задачи. Во-первых, мы хотели найти все гены в базе данных Kegg, где существует сдвиг фазы триплетной периодичности. Для этой цели в нами разработан математический подход к выявлению сдвига фазы триплетной периодичности в нуклеотидной последовательности. Для выявления сдвига фазы триплетной периодичности выделяются две следующие друг за другом последовательности длиной от 60 до 600 нуклеотидов (длина кратна трем нуклеотидам). Первое основание первой и второй последовательностей всегда соответствует первому основанию кодона. Затем строятся 4 матрицы триплетной периодичности [49,56]. Первая матрица соответствует рамке считывания в первой последовательности, вторая соответствует рамке считывания во второй последовательности. Две оставшиеся матрицы строятся по второй последовательности со сдвигом на одно и два основания, что соответствует двум альтернативным рамкам считывания. Можно говорить, что между двумя тандемными последовательностями наблюдается сдвиг фазы триплетной периодичности, если первая матрица триплетной периодичности более похожа на третью или четвертую матрицу, а не на вторую матрицу (Рис. 3). На основе разработанного математического подхода был проанализирован банк данных Kegg-46 [2] и на статистически значимом уровне было найдено 27417 случаев сдвига фазы триплетной периодичности, что может указывать на наличие мутаций в гене, приводящих к сдвигу рамки считывания.

Во-вторых, мы хотели проверить, действительно ли гипотетические аминокислотные последовательности, полученные при использовании рамки считывания триплетной периодичности, имеют гомологию с последовательностями из банка данных Swiss-prot. Такую проверку мы делали для тех районов, у которых наблюдался сдвиг по фазе между триплетной периодичностью и рамкой считывания. Мы подтвердили существование таких сдвигов для части генов, так как нашли гомологию между некоторыми гипотетическими аминокислотными последовательностями и аминокислотными последовательностями из банка данных Swiss-prot.

1. Методы исследования

1.1. Определение фазы триплетной периодичности

Будем считать, что задана кодирующая нуклеотидная последовательность $S=\{s(k), k=1,2,\ldots,L\}$, где каждое значения $s(k)$ выбирается из алфавита $A=\{a,t,c,g\}$, L есть длина последовательности S, она кратна трем. Введем три рамки считывания в последовательности S и обозначим их как T_1, T_2 и T_3 (Рис.3). Основание $s(1)$ последовательности S является первым, третьим и вторым основанием кодона для рамки считывания T_1, T_2 и T_3 соответственно. Рамка считывания T_1 реально существует в последовательности S, а рамки считывания T_2 и T_3 можно рассматривать как гипотетические. Определим также три матрицы триплетной периодичности $M_1(i_1, i_2)$, $M_2(i_1, i_2)$ и $M_3(i_1, i_2)$ для рамок считывания T_1, T_2 и T_3 для фрагмента последовательности S в координатах от i_1 до i_2. Обозначим этот фрагмент как $S(i_1, i_2)$. Элементы матриц $m_1(i,j)$ $m_2(i,j)$ и $m_3(i,j)$ показывают число оснований типа i в последовательности S (i=1 для a, i=2 для t, i=3 для c, i=4 для g,), которые встречается в j позиции кодона (j может быть равно 1, 2 или 3) для рамок считывания T_1, T_2 и T_3 соответственно. За начальную фазу матриц M_1, M_2, M_3 примем координату k того основания из $s(1)$, $s(2)$ и $s(3)$, которое входит в первую позицию триплета рамок считывания T_1, T_2 и T_3. Для матриц M_1, M_2, M_3 начальная фаза, соответственно равна 1, 2 и 3.

Далее определим условия, при которых можно считать, что в последовательности S после нуклеотида $s(x)$ существует сдвиг фазы триплетной периодичности. Для этого, во-первых, в последовательности S должна существовать триплетная периодичность. Условия существования триплетной периодичности и количественная мера для выявления триплетной периодичности в последовательности S или ее фрагмента определены в пункте 1.2. Во-вторых, необходимо ввести количественную меру различия матриц триплетной периодичности. Введем некую функцию U и будем считать, что две матрицы триплетной периодичности подобны друг другу, если $U \le U_0$. В противном случае будем считать их различными. Детальный вид функции U рассматривается в пункте 1.3. Будем считать, что после нуклеотида

s(x) в последовательности U существует сдвиг фазы триплетной периодичности на 1 основание, если одновременно выполняются условия:

$$\left\{\begin{array}{ll} U\{M_1(1,\ x),\ M_2(x+1,\ L)\} & \le U_0 \\ U\{M_1(1,\ x),\ M_1(x+1,\ L)\} & > U_0 \\ U\{M_1(1,\ x),\ M_3(x+1,\ L)\} & > U_0 \end{array}\right\} \quad (1)$$

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

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

Если же выполняются условия:

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

то будем считать, что после нуклеотида *s(x)* фаза триплетной периодичности остается без изменения, т.е. сдвиг фазы триплетной периодичности равен нулю. Сдвиг фазы будем определять как разницу начальных фаз матриц для фрагментов $S(x+1,L)$ и $S(1,x)$, входящих в первое уравнение условий (1), (2), (3). Сдвиг фазы триплетной периодичности на 1 и 2 основания соответствует вставке $1+3n$ и $2+3n$ (или делеции 2+3n и 1+3n) оснований (n=0,1,2,3,...) после нуклеотида *s(x)*.

1.2. Триплетная периодичность последовательности S

Матрицы триплетной периодичности можно рассматривать как таблицы сопряженности признаков [57]. Далее рассмотрим матрицу M_1, для матриц M_2 и M_3 все выводы будут аналогичны. Признаками строк матрицы M_1 являются основания последовательности S, а признаками столбцов - позиции оснований в кодонах рамки считывания T_1. Будем считать, что в нуклеотидной последовательности S существует триплетная периодичность, если уровень взаимной информации I_3 между основаниями последовательности S и позициями оснований в кодонах будет больше некоторой величины I_0 [44]. Взаимная информация вычисляется по формуле [57]:

$$I_3 = \sum_{i=1}^{4}\sum_{j=1}^{3} m_1(i,j)\ln m1(i,j) - \sum_{i=1}^{4} x(i)\ln x(i) - \sum_{j=1}^{3} y(j)\ln y(j) + L_1 \ln L_1 \quad (4)$$

где $x(i) = \sum_{i=1}^{3} m_1(i,j)$, $y(j) = \sum_{i=1}^{4} m_1(i,j)$. Удвоенная взаимная информация $2I_3$ имеет распределение χ^2 с 6-ю степенями свободы, что позволяет оценить статистическую значимость найденной периодичности. Величину I_3 можно также привести к стандартному нормальному распределению [58]:

$$X_3 = \sqrt{4I_3} - \sqrt{2n-1} \quad (5)$$

Соответствие $2I_3$ распределению χ^2 с 6-ю степенями свободы, а величины X_3 - стандартному нормальному распределению достигается в случае достаточно большого объема статистических данных, т.е. достаточно большого значения длины последовательности S. Чтобы определить минимальную длину последовательности S, которая позволяет использовать функцию χ^2 для описания распределения величины $2I_3$, мы протестировали соответствие $2I_3$ распределению χ^2 для различных длин последовательностей. С помощью датчика случайных чисел генерировались множества нуклеотидных последовательностей для каждой длины от 30 до 1000 нуклеотидов. Каждое такое множество содержало 10000 последовательностей. После этого для всех последовательностей из каждого множества рассчитывалась взаимная информация и для каждого множество строилась гистограмма, показывающая распределение величины $2I_3$. Эта гистограмма сравнивалась с теоретическим распределением по критерию χ^2. Оказалось, что для последовательностей длиной более 60 нуклеотидных пар распределение $2I_3$ соответствует χ^2 с вероятностью не менее 99%. Все последовательности с триплетной периодичностью, анализируемые в настоящем исследовании, были длиннее 60 н.п. Это позволяет использовать в данной работе распределение χ^2 для статистических оценок попадания $2I_3$ в интервал от некоторого порогового значения $2I_0$ до ∞. Мы искали сдвиг фазы триплетной периодичности в последова-

тельности S, если для обоих фрагментов $S(1,x)$ и $S(x+1,L)$ значение X_3 было больше 4.0. Вероятность того, что данная триплетная периодичность будет образована чисто случайными факторами, составляет менее чем $4\text{x}10^{-4}$.

1.3 Алгоритм поиска сдвига фазы триплетной периодичности

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

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

где:

$$I_{kl}(j) =$$

$$= \sum_{i=1}^{4} m_l(i,j)\ln(m_l(i,j)) + \sum_{i=1}^{4} m_l(i,j)\ln(m_l(i,j)) -$$

$$- \sum_{i=1}^{4} (m_k(i,j) + m_l(i,j))\ln(m_k(i,j) + m_l(i,j)) +$$

$$+ (y_k(j) + y_l(j))\ln(y_k(j) + y_l(j)) - y_k(j)\ln y_k(j) -$$

$$- y_l(j)\ln y_l(j) \qquad (7)$$

Здесь $m_k(i,j)$ и $m_l(i,j)$ есть элементы двух сравниваемых матриц M_k и M_l, $y_l(j) = \sum_{i=1}^{4} m_l(i,j)$, $y_k(j) = \sum_{i=1}^{4} m_k(i,j)$. Значение $I_{kl}(j)$ показывает различие столбцов с номером j у двух матриц M_k и M_l. Значение $2I_{kl}(j)$ распределено как χ^2 с 3 степенями свободы при сравнении двух матриц, построенных для случайных последовательностей [28], в силу этого значение I_{kl} распределено как χ^2 не более, чем с 9-ю степенями свободы. В дальнейшем мы использовали I_{11}, I_{12}, I_{13} для поиска сдвига рамки считывания. Если в последовательности S после x отсутствует вставки или делеции оснований ДНК (с длиной делеции или вставки, не кратной трем основаниям), то в этом случае $I_{11}<I_{12}$ и $I_{11}<I_{13}$. Если присутствует вставка фрагмента длиной $Q=3i+1$ или же делеция фрагмента длиной $Q=3i+2$,(i=1,2,...), то можно говорить о переходе после x от рамки считывания T_1 к рамке считывания T_2. В этом случае $I_{12}<I_{11}$ и $I_{12}<I_{13}$. Если присутствует вставка фрагмента длиной $Q=3i+2$ или же делеция фрагмента длиной $Q=3i+1$,(i=1,2,...), то можно говорить о переходе после позиции x от рамки считывания T_1 к рамке считывания T_3. В этом случае $I_{13}<I_{11}$ и $I_{13}<I_{12}$. Удобно для поиска сдвигов рамки считывания взять величины $F_1=I_{11}I_{12}$ и $F_2=I_{11}I_{13}$. В случае присутствия сдвига фазы триплетной периодичности одна из этих величин будет принимать достаточно большие значения. Величины F_1 и F_2 имеют распределение Фишера в том случае, если они построены для случайных последовательностей [58].

Для каждой координаты x мы варьировали значение L_1. Варьирование проводилось с целью поиска такой длины L_1, которая задавала бы максимальное значение для величин F_1 и F_2. Такой поиск необходимо провести, так как триплетная периодичность по длине последовательности может меняться, и это изменение может влиять на значения F_1 и F_2. Мы варьировали L_1 для каждого x, кратного 3-м, в интервале от 60 до 600 оснований.

В итоге для последовательности S был построен график зависимости максимальных величин F_1 и F_2 от координаты x, каждая из которых была получена для некоторой длины L_1. На этом графике мы отбирали координаты локальных максимумов. Будем считать, что в последовательности есть сдвиг фазы триплетной пе-

риодичности, если значение F_1 или F_2 в локальном максимуме больше некоторого порогового значения F_0. Пороговое значение F_0 определялось методом Монте-Карло (п. 1.4).

Кроме того, для идентификации сдвига фазы триплетной периодичности в позиции x необходимо убедится, что в последовательностях $S(x\text{-}L_1+1,\ x)$, $S(x+1,\ x+L_1)$, $S(x+2,\ x+L_1+1)$ и $S(x+3,\ x+L_1+2)$ триплетная периодичность существует, так как значения I_{11}, I_{12}, I_{13} показывают меру расхождения матриц триплетной периодичности. Эта мера будет тем меньше, чем выше подобие друг другу у матриц триплетной периодичности, что при отсутствии последней может быть обусловлено чисто случайными факторами. Для исключения подобия чисто случайных матриц мы брали к рассмотрению только такие последовательности $S(x\text{-}L_1+1,\ x)$, $S(x+1,\ x+L_1)$, $S(x+2,\ x+L_1+1)$ и $S(x+3,\ x+L_1+2)$, которые обладают достаточно выраженной триплетной периодичностью (п. 1.2).

1.4. Применение метода Монте-Карло для определения порогового значения F_0

Для поиска порогового значения F_0 были использованы последовательности генов, собранные в банке данных Kegg версии 46. Всего в этой версии банка данных содержится 3.318.628 генов. Мы создали случайный банк данных путем перемешивания последовательности оснований каждого гена. С помощью метода Монте-Карло это позволило сохранить такое же распределение длин случайных последовательностей, как в банке данных Kegg, и аналогичное распределение оснований. Для сохранения триплетной периодичности в случайной последовательности на таком же уровне, как и в реальном гене, последовательность S разбивалась на три подпоследовательности. Первая из них (обозначим ее как C_1) получена выбором из последовательности S оснований, которые стоят на номерах, равных i=1+3n. Вторая последовательность C_2 получена путем выбора оснований, стоящих на позициях i=2+3n, а третья последовательность C_3 - выбором оснований, стоящих на позициях с номерами i=3+3n. При создании последовательностей C_1, C_2 и C_3 n меняется от 0 до $L/3$-1.

Далее датчиком случайных чисел создавались последовательности R_1, R_2 и R_3 длиной $L/3$. Затем мы упорядочивали последовательности R_1, R_2 и R_3, по возрастанию и запоминали порядок сделанных перестановок в каждой последовательности. После этого нуклеотиды в последовательностях C_1, C_2 и C_3 переставлялись так, как это было сделано при упорядочивании последовательностей R_1, R_2 и R_3, по возрастанию. После такого перемешивания последовательностей R_1, R_2 и R_3 создавалась случайная последовательность R. В последовательности R на позициях i=1+3n стояли нуклеотиды из последовательности R_1, на позициях i=2+3n - нуклеотиды из последовательности R_2 и на позициях i=3+3n - нуклеотиды из последовательности R_3. Длина последовательности R была равна L, и в ней был сохранен такой же состав нуклеотидов, как и в последовательности S.

После создания банка случайных последовательностей, имеющего такой же обьем, длину и триплетную периодичность каждой последовательности, как и в банке данных Kegg, мы выбрали уровень F_0 равным 2.5 и подсчитали число генов, которые имеют хотя бы один локальный максимум (как это описано в п. 1.3) для F_1 и F_2 выше F_0 для последовательностей из банка данных Kegg-46 и для случайных последовательностей. Для уровня F_0=2.5 число найденных сдвигов фазы триплетной периодичности в случайных нуклеотидных последовательностях составляет ~1,5% от числа сдвигов, которые мы нашли в банке данных Kegg-46. Поэтому данный уровень может быть выбран как пороговый, так как примесь сдвигов триплетной периодичности из-за чисто случайных факторов можно считать незначительной.

1.5. Построение графика $I_{12}(x_1,x_2)$

Был построен контурный график(contour plot) для значений $I_{11}(x_1,x_2)$, где x_1 и x_2 показывают координаты начала последовательностей S_1 и S_2 в последовательности S. В этом случае последовательности S_1 и S_2 уже не следуют друг за другом, а выделяются в координатах (x_1, x_1+L1-1) и (x_2, x_2+L1-1). Координаты x_1 и x_2 меняются независимо друг от друга от 1 до L-$L1$+1 с шагом в три основания. Это означает, что x_1=1+3i, i=0,1,2,3,..., а x_2=1+3j, j=0,1,2,3,...,

где i и j - натуральные числа. Данный контурный график будет симметричен относительно главной диагонали и позволяет увидеть районы сдвигов фазы триплетной периодичности в последовательности S.

2. Результаты и обсуждение

2.1. Анализ искусственной периодической последовательности

Для начала мы изучили сдвиг фазы триплетной периодичности у искусственной периодической последовательности. Для этого в периодическую последовательность $(atg)_{159}$ после позиции 477 основания мы вставили нуклеотиды at, после чего добавили справа последовательность $(atg)_{200}$ и в итоге получили последовательность $(atg)_{159}at(atg)_{200}$. Данная последовательность содержит вставку двух оснований после позиции 477. Как отмечалось выше, в этом случае то $I_{13}<I_{11}$ и $I_{13}<I_{12}$.и следует ожидать, что значения $F_2(477)$ будет достигать максимальных значений, тогда как $F_1(x)$ будет меняться незначительно. На Рис.4 показаны значения $F_2(x)$ и видно, что $F_2(477)>10^4$, тогда как $F_1(x)<1.0$ для всех x.

Для этой же последовательности был построен контурный график (Рис.5) и на нем темным цветом выделены районы последовательности, где триплетные периодичности, выделенные в последовательностях S_1 и S_2 (п. 1.4) отличаются друг от друга. На этом графике четко выделяются два участка, где триплетная периодичность имеет сдвиг фазы. Это районы от 1 до 477 нуклеотида и от 480 до 1080 нуклеотида. Этот тестовый пример показывает, что разработанный математический подход позволяет обнаруживать сдвиги фазы в триплетной периодичности гена.

2.2. Поиск генов со сдвигом фазы триплетной периодичности в базе данных Kegg

Всего было проанализировано 3.318.628 генов, накопленных в банке данных KEGG версии 46 (http://www.genome.ad.jp/kegg/). Общее число генов со сдвигом фазы триплетной периодичности составило 27417 (F_0=2.5). Число генов с одним сдвигом фазы триплетной периодичности составляет 90% от общего количество генов со сдвигами фазы триплетной периодичности. Оставшиеся 10% приходятся на гены, где наблюдается более одного сдвига. В случайной выборке такого же объема мы нашли 411 сдвигов триплетной периодичности, что составляет ~1.5% от числа последних в банке данных Kegg-46. Это сравнение показывает, что почти все найденные сдвиги фазы триплетной периодичности носят неслучайный характер.

На Рис.6 показан пример гена с одним сдвигом фазы триплетной периодичности. Это ген cytochrome C из G.sulfurreducens (имя в Swiss-Prot Q74A96_GEOSL). Из Рис. 6 видно, что в последовательности в районе 880 нуклеотида есть пик для значений F_1, тогда как значения F_2 не превышают 1.0. Это означает, что триплетная периодичность после 880 основания сдвинулась на одно основание к началу гена. Если учесть, что между триплетной периодичностью и рамкой считывания с 1 по 880 основание отсутствует сдвиг фазы, то это означает, что после 880 основания возникает сдвиг фазы между рамкой считывания и триплетной периодичностью. Этот сдвиг соответствует делеции одного основания или же вставки двух оснований около 880 нуклеотида. Не исключена также возможность делеции фрагмента длиной $Q=3i+1$ или же вставка фрагмента длиной $Q=3i+2$, где $i=1,2,\ldots$ Может быть, именно поэтому верхняя часть пика на Рис.6 оказалась несколько сглаженной. Контурный график для этого гена показан на Рис.7.

Второй пример гена со сдвигами фазы триплетной периодичности показан на Рис.8. На этом рисунке приведены зависимости F_1 и F_2 от координаты x для последовательности F15A2.6 из Kegg-46 из генома C.elegans. Этот ген кодирует BR serine/threonine kinase. Из Рис. 8 видно, что в этой последовательности существует не менее пяти сдвигов триплетной периодичности. Можно выделить пять позиций с координатами 1195, 1285, 1816, 1951 и 2326 нуклеотид в последовательности гена. В первой, второй и пятой позициях возможна делеция фрагментов длиной $Q=3i+1$ или же вставка фрагмента длиной $Q=3i+2$, а в третьей и четвертой позициях возможна делеция фрагментов длиной $Q=3i+2$ или же вставка фрагмента длиной $Q=3i+1$, где $i=0,1,2,\ldots$. Рис.8 показывает, что сдвиг фазы

триплетной периодичности хорошо выражен и значения F_1 и F_2 значительно больше, чем 2.5. Для гена с множеством сдвигов фазы триплетной периодичности мы не строили контурного графика, так как в этом случае он не имеет наглядности.

Интересно посмотреть распределение найденных генов со сдвигами фазы триплетной периодичности по биологическим функциям. Список функций, для которых было найдено больше всего генов со сдвигами фазы триплетной периодичности, показан в Табл. 1. Полностью список функций можно скачать из http://victoria.biengi.ac.ru/. Из Табл. 1 видно, что наибольшее число сдвигов фазы триплетной периодичности имеют псевдогены, число которых составляет 4610. Такая высокая частота сдвига в триплетной периодичности в пседогенах не удивительна, так как псевдогены лишены функционального значения и мутации в них являются нейтральным событием для генома и могут сравнительно быстро накапливаться [2, 32].

Второе по классу множество генов со сдвигом фазы триплетной периодичности составляют гены, в которых ранее уже была замечена мутация сдвига рамки считывания. Мы выявили 568 таких генов. Этот факт дополнительно подтверждает, что сдвиг рамки считывания в гене может быть выявлен посредством поиска сдвига фазы триплетной периодичности.

2.3. Поиск белковых подобий для аминокислотных последовательностей

Рассмотрим гены, где мы обнаружили сдвиг фазы триплетной периодичности. Обозначим номер нуклеотида i, где произошел сдвиг фазы триплетной периодичности, как x_0. В этом случае для последовательности $s(i)$ для $i<x_0$ мы можем считать, что есть совпадение между триплетной периодичностью и рамкой считывания. В то же время для последовательности $s(i)$ для $i>x_0$ будет наблюдаться сдвиг между триплетной периодичностью и существующей в гене рамкой считывания. Можно предполагать, что триплетная периодичность определяет рамку считывания, которая существовала в гене до мутации посредством сдвига рамки считывания. Будем называть эту рамку считывания древней рамкой считывания. Таким образом, в последовательности $s(i)$ для $i>x_0$ существует две рамки считывания – одна реально существующая в гене, а другая предполагаемая на основе триплетной периодичности или же древняя рамка считывания. Если мутация посредством сдвига рамки считывания произошла в гене сравнительно недавно, то могут остаться варианты этого гена без мутации. Если перекодировать нуклеотидную последовательность для $i>x_0$ в аминокислотную по двум рамкам считывания, то можно получить две аминокислотные последовательности. Первая из них - это реально существующая аминокислотная последовательность, а вторая - предполагаемая аминокислотная последовательность, которую можно назвать древней аминокислотной последовательностью. Для каждой нуклеотидной последовательности, где был найден сдвиг фазы триплетной периодичности, мы создали две такие аминокислотные последовательности. Если в нуклеотидной последовательности было несколько сдвигов фазы триплетной периодичности, то мы брали последний справа сдвиг. Это означает, что мы не делали «реставрацию» всех сдвигов в последовательности с числом сдвигов фазы триплетной периодичности больше единицы. Мы считали маловероятным, что такая полная реставрация может дать значимое подобие с аминокислотными последовательностями в банке данных.

Таким образом, мы создали 27417 пар аминокислотных последовательностей, и все они программой Blast [59-60] были сравнены с последовательностями базы данных Swiss-prot [61]. В результате сравнения 10705 пар последовательностей не имели значимого подобия. Для 12096 пар последовательностей подобие наблюдалось только для аминокислотной последовательности, созданной по рамке считывания, которая присутствует в гене. Для 4198 пар последовательностей подобие наблюдалось только для древней аминокислотной последовательности и для 488 пар последовательностей подобие наблюдалось для обеих аминокислотных последовательностей из пары. Значение E при использовании программы Blast было выбрано равным 10^{-3}, что дает в среднем порядка 50 случайных подобий при сканировании более чем 54 тысяч аминокислотных последователь-

ностей. Эти цифры говорят о том, что доля случайных подобий от общего числа найденных подобий составляет менее 1%.

Рассмотрим пример, где подобие было найдено для аминокислотных последовательностей, созданных по существующей и древней рамке считывания (локус XOO3621 базы данных Kegg). Введем для этого гена три рамки считывания T_1, T_2 и T_3. Рамки считывания T_1, T_2 и T_3 выделяют триплеты оснований начиная с первого, второго и третьего основания последовательности XOO3621, соответственно. Локус XOO3621 содержит ген extracellular protease из X.oryzae и он имеет длину в 1281 н.п. В нем выделяется два сдвига фазы триплетной периодичности (Рис 9). Первый сдвиг наблюдается в районе 478 основания и он виден по функции F_1, что соответствует переходу от рамки считывания T_1 к рамке считывания T_2. Второй сдвиг фазы триплетной периодичности наблюдается в районе 937 основания и он выявлен по функции F_2, что соответствует переходу рамки считывания от T_2 обратно к T_1. Это означает, что белок, который кодирует данный ген (Q5GWP6_XANOR), будет иметь измененную аминокислотную последовательность только с 160 аминокислоты по 313 аминокислоту (с 478 по 937 нуклеотид в гене). Мы создали две аминокислотные последовательности после 478 нуклеотида до конца гена по рамкам считывания T_1 и T_2. Назовем их W_1 и W_2. Проведем нумерацию аминокислот последовательностей W_1 и W_2 с первой аминокислоты рамок считывания T_1 и T_2 последовательности гена из локуса XOO3621. Сравнение программой Blast последовательности W_1 с банком данных Swiss-prot показало, что она подобна фрагменту аминокислотной последовательности EXPR_XANCP (№1, Табл. 2). Последовательность W_2 также показала наличие подобия с последовательностью EXPR_XANCP (№2, Табл. 2). Затем мы сравнили полностью аминокислотную последовательность Q5GWP6_XANOR (W_1 ее часть со 160-ой аминокислоты) с аминокислотной последовательностью EXPR_XANCP и выявили два участка подобия, которые показаны на Рис. 10 и обозначены как *A* и *C*. Подобие, найденное для последовательности W_1 (№1, Табл. 2), соответствует подобию, обозначенному на Рис.8 как *C*. Подобие последовательности W_2 (№2 в Табл. 2) показывает, что аминокислотная последовательность W_2 (созданная по рамке считывания T_2 для района с 478 по 937 нуклеотид локуса XOO3621) присутствует в последовательности EXPR_XANCP в районе *B* (Рис. 8). Одновременно с этим мы не обнаружили сдвигов фазы триплетной периодичности в нуклеотидной последовательности из локуса XCC0851 (база данных Kegg), где содержится ген, кодирующий аминокислотную последовательность EXPR_XANCP. Таким образом, можно сделать вывод, что последовательность гена из локуса XCC0851 не имеет двух сдвигов рамок считывания и фрагмент последовательности W_2 со 170 по 286 аминокислоту представлен в последовательности EXPR_XANCP в участке *B*. Можно предположить, что последовательности генов XCC0851 и XOO3621 произошли от одного гена тем или иным способом. Затем в результате двух сдвигов рамок считывания в гене XOO3621 возникла новая кодирующая последовательность, которая создала новую аминокислотную последовательность с 160 по 313 аминокислоту. Так как биологическая функция генов не изменилась, то новая аминокислотная последовательность, возникшая в результате такой двойной мутации, сохранила свою биологическую функцию.

2.4. Обсуждение

В данной работе удалось показать, что изучение сдвигов между триплетной периодичностью и рамкой считывания может находить возможные мутации посредством сдвига рамки считывания в гене. Удалось найти 27417 таких генов, где существовало два участка однотипной триплетной периодичности, разделенные делециями или же вставками нуклеотидов, что составляет приблизительно 0.7% от общего числа проанализированных генов. Мы предполагаем, что ранее в этих районах рамка считывания T_1 и триплетная периодичность были однозначно связаны между собой и только после делеций и вставок нуклеотидов между ними образовался сдвиг [62]. Такой сравнительно небольшой процент генов, где мог быть осуществлен сдвиг рамки считывания, может быть объяснен несколькими причинами. Во-первых,

разработанный математический метод позволяет искать только сравнительно небольшие по размеру вставки и делеции символов. Это связано с тем, что протяженная вставка может разрушить как саму триплетную периодичность (4), так и подобие матриц триплетной периодичности (6). Таким образом, данный метод будет пропускать значительную часть генов, содержащих большую (>50 оснований) вставку или же делецию нуклеотидов, которая может приводить к сдвигу рамки считывания. Во-вторых, применяемый нами подход хорошо работает при небольшом количестве районов, где были произведены вставки или же делеции нуклеотидов. Если плотность вставок и делеций будет больше, чем одна вставка или же делеция на несколько десятков нуклеотидов (~50), то точно расставить делеции и вставки при помощи применяемого алгоритма будет не всегда возможно. Это приведет к тому, что мы не сможем получить статистически значимое значение F_1 или F_2 для такого гена. В-третьих, мы задали достаточно высокий уровень триплетной периодичности для гарантированного обнаружения сдвигов фазы триплетной периодичности X_3 (5), которые существует далеко не в каждом гене из базы данных Kegg-49. Изучение сдвигов фазы триплетной периодичности для более низких значений X_3, вероятно, может позволить выявить большее количество генов для которых F_1 или F_2 будет больше порогового уровня.

В целом можно считать, что в данной работе мы обнаружили нижнюю границу числа генов, где возможен сдвиг между рамкой считывания и триплетной периодичностью. В реальности число таких генов может быть значительно большим. Об этом свидетельствует тот факт, что всего в банке данных Kegg-46 содержится только 2069 генов с известным сдвигом рамки считывания, тогда как нам удалось выявить 568 таких генов, что происходило из-за отмеченных выше ограничений применяемого метода. Однако применяемый подход обнаружил еще дополнительно около 27 тысяч генов со сдвигом фазы триплетной периодичности, для которых можно предполагать наличие сдвига рамки считывания. Можно предполагать, что генов со сдвигом рамки считывания будет по крайней мере в 4 раза большим и может составлять несколько процентов.

Применяемый нами подход поиска сдвигов между триплетной периодичностью и рамкой считывания для поиска мутаций в генах посредством рамки считывания при его дальнейшем развитии представляется нам предпочтительнее, чем применение поиска возможных подобий при помощи типа Blast. Это связано с тем, что данный метод для обнаружения в гене мутаций посредством сдвига рамки считывания не требует привлечения дополнительных данных, основывается только на исходной последовательности гена. Так как объем банка данных аминокислотных последовательностей ограничен, то всегда будет существовать вероятность того, что подобия не будут найдены, а в реальности мутация посредством сдвига рамок считывания в гене существует. Мы полагаем, что более полное выявление мутаций посредством сдвига рамок считывания в генах возможно на пути объединения двух подходов. Это означает, что нужно также исследовать те гены, для которых $F_1<F_0$ или $F_2<F_0$ и считать, что мы нашли мутацию посредством сдвига рамки считывания, если для аминокислотных последовательностей, созданных по рамкам считывания *T2* и *T3*, существуют статистически значимые подобия. В этом случае сравнительно небольшое увеличение F_1 и F_2 только указывает на возможность сдвига рамки считывания и факт такой мутации можно считать доказанным только после обнаружения подобий. С другой стороны, улучшение применяемого в настоящей работе подхода может происходить на пути использования более совершенных алгоритмов поиска триплетной периодичности, например, таких как скрытые Марковские модели. В этом случае, вероятно, удастся выявлять сдвиги рамки считывания, вызванные множеством событий вставок и делеций нуклеотидов в различные районы гена.

С функциональной точки зрения мутации посредством сдвига рамок считывания представляются событиями, способными кардинально изменить функцию гена и кодируемого им белка. Это может объяснять сравнительное небольшое число таких событий, найденных в ранее выполненных исследованиях и в настоя-

щей работе [2, 32-33]. Их осуществление может вносить большой вклад в образование новых генов посредством копирования известных генов и образования там мутаций посредством сдвига рамки считывания [33-35]. Однако генетический код также должен быть как-то адаптирован для этих событий [45, 63], и новая аминокислотная последовательность должна обладать какой-то биологической функцией. В противном случае перебор мутационных событий для создания новой функции гена в его копии может быть слишком велик и неосуществим за разумное эволюционное время.

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

Приложение

```
H2A1_HUMAN  MSGRGKQGGKA**RA**--KAK**T**RSSRAGLQFPVGR**V**HRLLRKGNYA**E**RVGAGAPVYLA  1-53
H2A_CAEEL   MSGRGK-GGKA**KT**GGKAK**S**RSSRAGLQFPVGR**L**HRILRKGNYA**Q**RVGAGAPVYLA  1-54

H2A1_HUMAN  AVLEYL**T**AE**I**LELAGNAARDNKKTRI**I**PRHLQLA**I**RNDEELNKLL**GK**VTIAQGGV  54-108
H2A_CAEEL   AVLEYL**A**AE**V**LELAGNAARDNKKTRI**A**PRHLQLA**V**RNDEELNKLL**AG**VTIAQGGV  55-109

H2A1_HUMAN  LPNIQAVLLPKKT  109-121
H2A_CAEEL  LPNIQAVLLPKKT  110-122
```

Рис.1. Сравнение нуклеотидных последовательностей гистонов **H2A** из генома человека (H2A1_HUMAN) и нематоды (H2A_CAEEL)

Несовпадающие аминокислоты выделены жирным шрифтом. Знак пробела показывает делеции аминокислот в сравниваемых последовательностях.

А.

aac tat gcc gag cgg gtc ggg gcc ggc gcg ccg gtg tat ctg gca gcg gtg ctg gag tac ctg ac**C gcc gag atc ctg gaa ctg gcg ggc aac gcg gcc cgc gac aac aag aag**
aac tat gcc gag cgg gtc ggg gcc ggc gcg ccg gtg tat ctg gca gcg gtg ctg gag tac ctg acg **ccg aga tcc tgg aac tgg cgg gca acg cgg ccc gcg aca aca aga ag**

Б.

NYAERVGAGAPVYLAAVLEYLT**AEILELAGNAARDNKK**
NYAERVGAGAPVYLAAVLEYLT**PRSWNWRATRPATTR**

Рис.2. Влияние делеции одного основания в нуклеотидной последовательности гена на аминокислотную последовательность

А – сравнение по триплетам двух нуклеотидных последовательностей. Триплеты оснований кодируют соответствующие им аминокислоты в соответствии с генетическим кодом. Вторая нуклеотидная последовательность получена из первой путем делеции основания С на 66 позиции последовательности. Такая делеция приводит к тому, что после 66 позиции все триплеты меняются. Что приводит к изменению соответствующей аминокислотной последовательности.
Б – аминокислотные последовательности, полученные до делеции оснований С в 66 позиции и после такой делеции. Видно, что после 22 аминокислоты подобие между двумя последовательностями отсутствует. Неподобные части аминокислотных последовательностей выделены жирным шрифтом.

```
1231231231231231231231231231231231231 - рамка считывания T1
3123123123123123123123123123123123123 - рамка считывания T2
2312312312312312312312312312312312312 - рамка считывания T3
atgatgatgatgatgatgCatgatgatgatgatgatg - последовательность S
```

$M_1(1,18)$

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

$M_1(19,37)$ Длина не кратна 3!!

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

$M_2(19,37)$

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

$M_3(19,37)$

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

Рис 3. Влияние вставки одного основания на сдвиг фазы триплетной периодичности

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

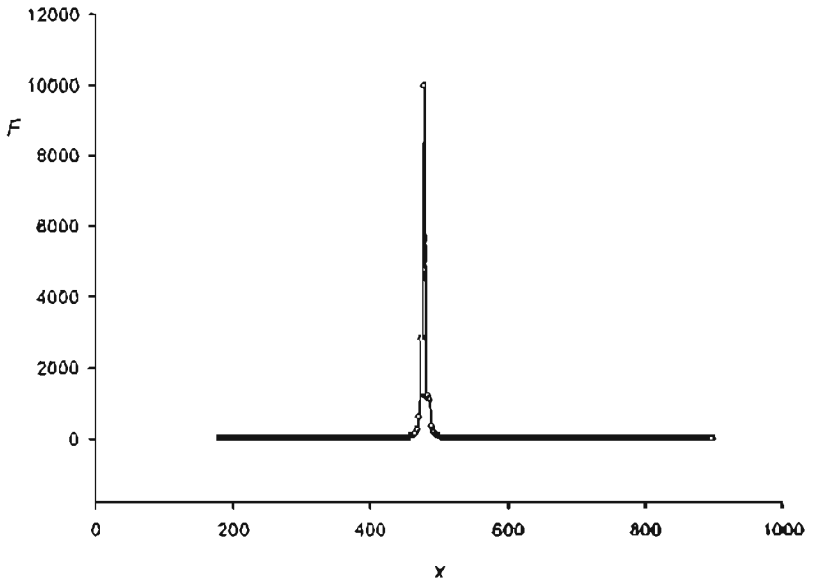

Рис. 4. Значения функции $F_3(x)$ в случае явной триплетной периодичности со вставкой двух оснований

Последовательность S имеет вид $(atg)_{155}at(atg)_{200}$. После позиции x=477 вставлены нуклеотиды *at*.

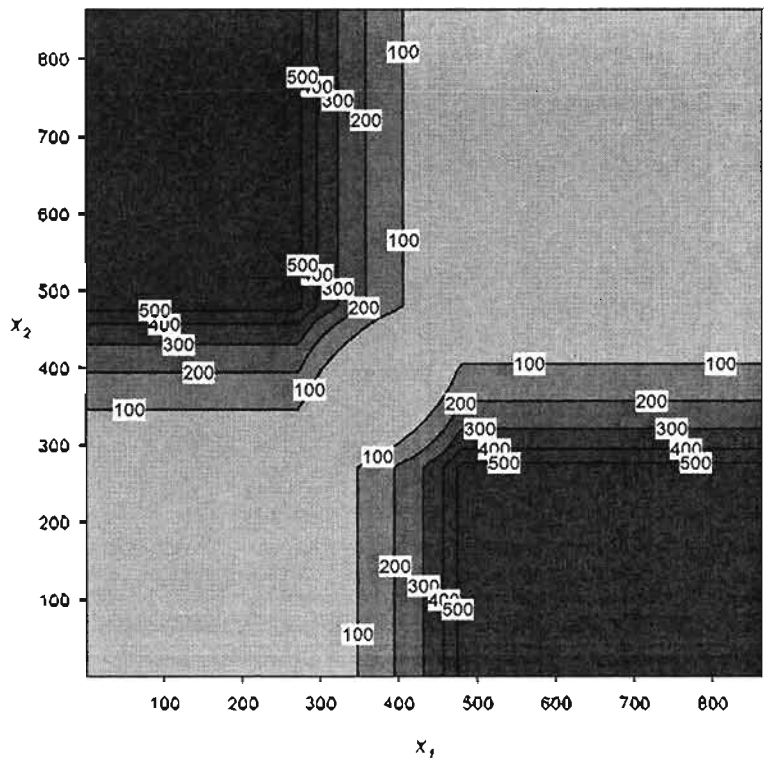

Рис.5. Контурный график для функции $I_{11}(x_1, x_2)$. x_1 и x_2 показывают координаты начала последовательностей S_1 и S_2 в последовательности S (пункт 1.5)

Последовательность S имеет вид $(atg)_{155}at(atg)_{700}$. После позиции x=477 вставлены нуклеотиды *at*

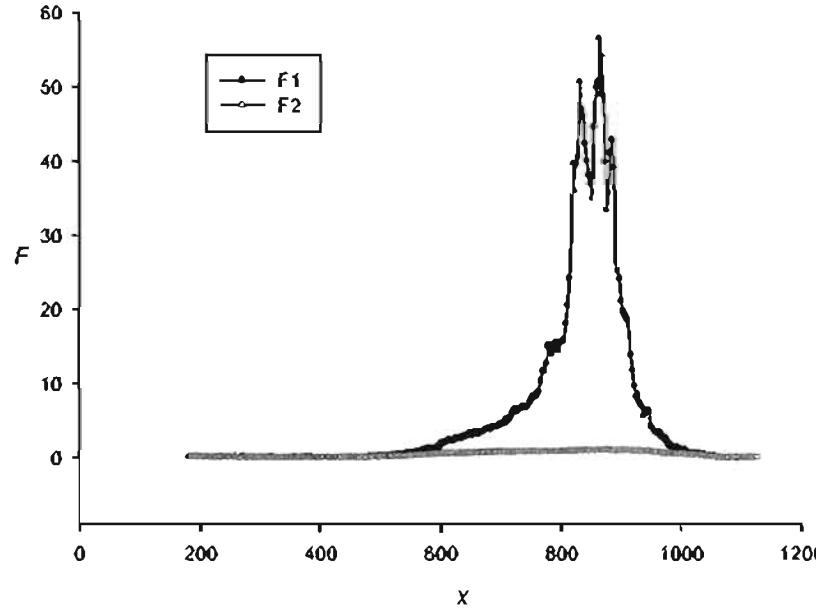

Рис. 6. Значения функций F_1 и F_2 для последовательности GSU2494 из банка данных Kegg
Этот ген кодирует аминокислотную последовательность cytochrome C из генома G.sulfurreducens

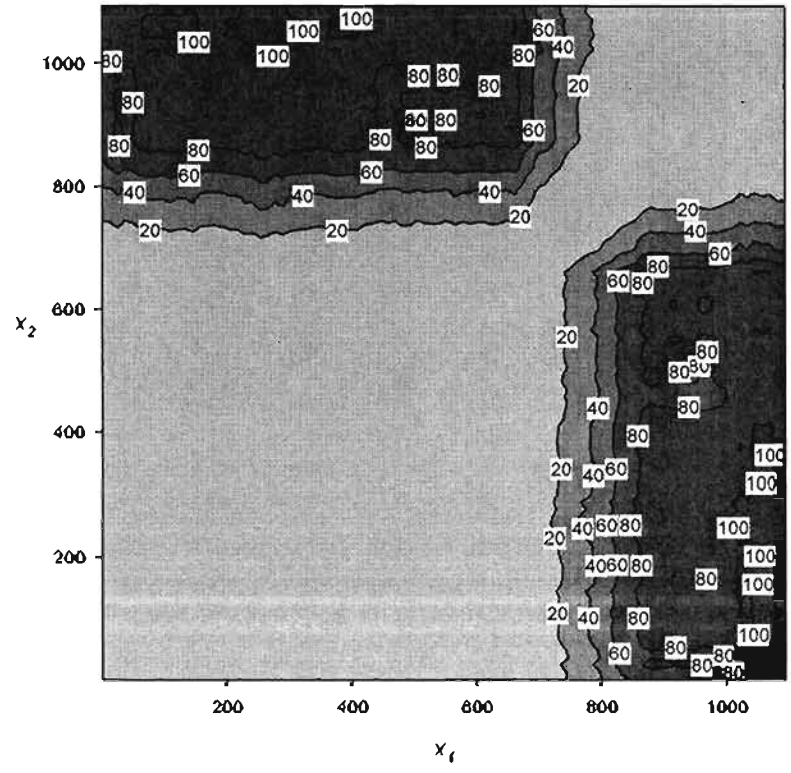

Рис.7. Контурный график для функции $I_1(x_1,x_2)$, где x_1 и x_2 показывают координаты начала последовательностей S_1 и S_2 в последовательности GSU2494 (пункт 1.5)

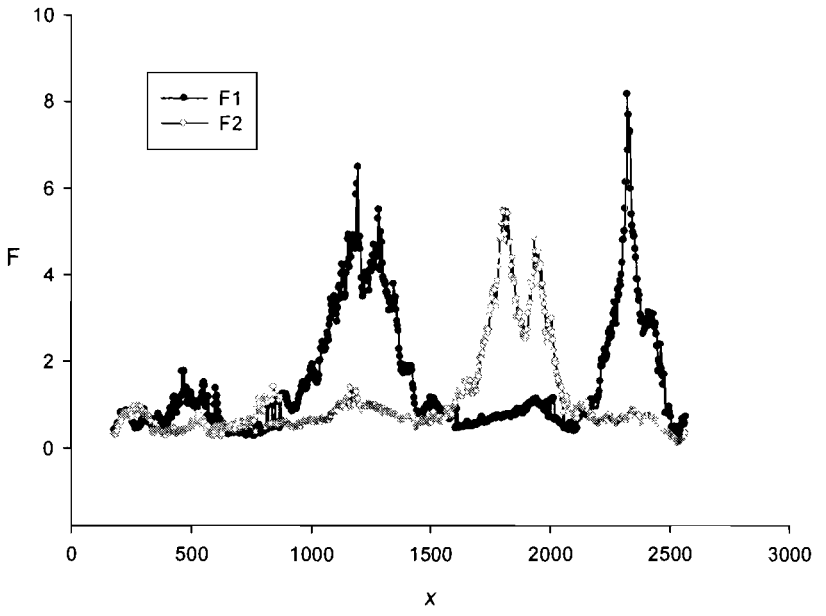

Рис. 8. Значения функций F_1 и F_2 для последовательности F15A2.6 из банка данных Kegg

Этот ген кодирует аминокислотную последовательность. Этот ген кодирует BR serine/threonine kinase. из генома C.elegans

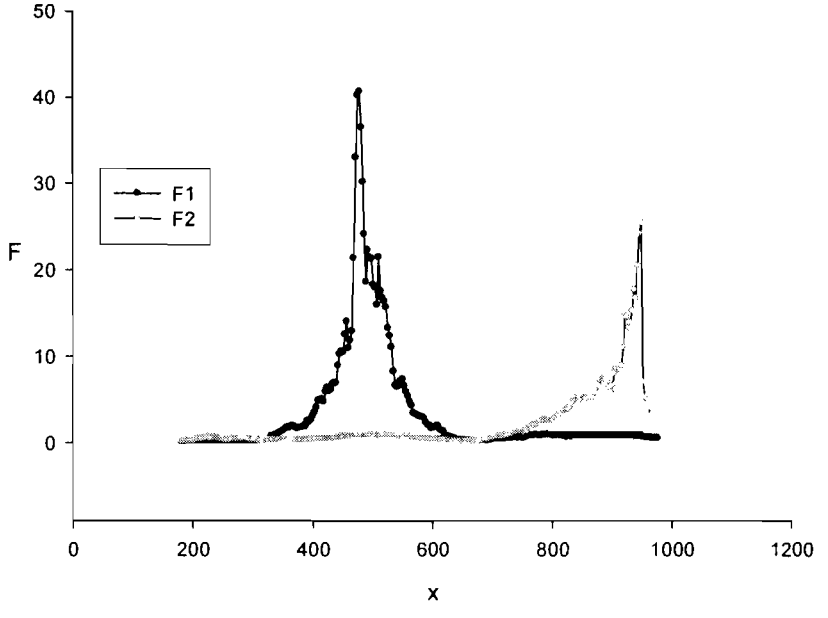

Рис. 9. Значения функций F_1 и F_2 для последовательности XOO3621 из банка данных Kegg

Этот ген кодирует аминокислотную последовательность extracellular protease из генома X.oryzae.

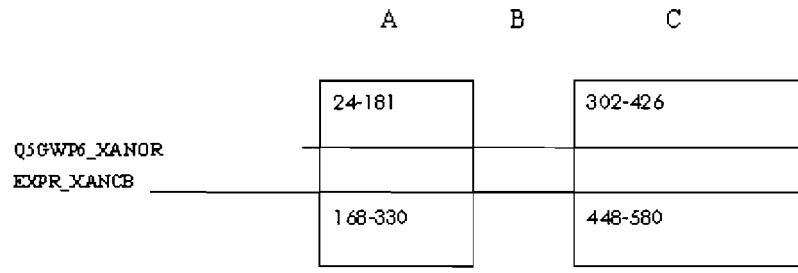

Рис.10. Выравнивание аминокислотной последовательности Q5GWP6_XANOR и аминокислотной последовательности EXPR_XANCB

Подобие между ними было найдено в районе *A* и в районе *C*, которые выделены рамками.

Табл. 1. Число генов со сдвигами фазы триплетной периодичности в группах, созданных по идентичному описанию биологической функции в базе данных Kegg

№	Number of genes	Definition
1	4610	Pseudogene
2	568	Frameshift
3	111	Translation initiation factor IF-2
4	48	PPE family protein
5	41	Ribonuclease E
6	33	PE-PGRS family protein
7	28	50S ribosomal protein L4
8	26	Protein kinase, putative
9	22	Ribonuclease, Rne/Rng family
10	22	Ankyrin repeat protein, putative
11	20	Serine/threonine protein kinase
12	20	Exodeoxyribonuclease VII large subunit
13	20	IS element 1477
14	19	ATP-dependent RNA helicase
15	18	High-affinity nickel-transporter
16	17	Protein kinase
17	17	Erythrocyte membrane protein 1 (PfEMP1)
18	16	60 kDa inner membrane insertion protein
19	16	Protein kinase domain containing protein
20	16	IS1404 transposase

Табл. 2. Выравнивание, найденное для аминокислотных последовательностей W_1 и W_2, созданных по рамкам считывания T_1 и T_2

K_1 и K_2 показывают левые и правые границы для найденного подобия. Последовательность, выделенная жирным шрифтом, показывает совпадающие аминокислоты в найденных выравниваниях

№	Sequence	Score and *E*	K_1	K_2	Alignment
1.	*W1*, frame T_1, gene XOO3621 EXPR_XANCP	88.2 5e-17	339 494	426 580	AALSDSLYYQVNVPAGTRSLKVTLAVGSGNADLSVRAGALPTDAAYSCRSMLPGNGDSCTLAAPAAGVYYVRLKATLGFSGVSVTAAY **AA L Y + VPAG+ +L VT + GSG+ADL VRAG+ PTD+AY+CR GN ++CT+ AP +G YYVRLKA FSGV++ A+Y** AATGAELNYTITVPAGSGTLTVTTSGGSGDADLYVRAGSAPTDSAYTCRPYRSGNAETCTITAP-SGTYYVRLKAYSTFSGVTLRASY
2.	*W2*, frame *T2*, gene XOO3621 EXPR_XANCP	139 1e-32	174 Frame *T2* 320-	311 Frame *T2* 457	QNAINSAVSRGTNVVVACIS-AANVSGLLPANCANVIAVAATTSAGAKASYSNFGAEIDVSAPGSGILSTLNSGTTTPGTPSYASXNGTSMAVPHVAGVVALMQ **QNAIN AVSRGT VVVA + A+NVSG LPANCANVIAVAATTSAGAKASYSNFG IDVSAPGS ILSTLNSGTTTPG+ SYAS NGTSMA PHVAGVVAL+Q** QNAINGAVSRGTTVVVAAGNDASNVSGSLPANCANVIAVAATTSAGAKASYSNFGTGIDVSAPGSSILSTLNSGTTTPGSASYASYNGTSMASPHVAGVVALVQ SVALNPLTPATVKGLLKASARPLLVACTQGGAGQG **SVA LTPA V+ LLK +AR L AC+ GG G G** SVAPTALTPAAVETLLKNTARALPGACS-GGCGAG

Литература

1. Watson, J.D., Baker, T.A., Bell, S.P., Gann, A., Levine, M. and Losick, R. 2007. Molecular Biology of the Gene, 6th Edition. CSHL Press & Benjamin Cummings, San Francisco, CA.
2. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27, 29-34.
3. Baxevanis, A.D. and Ouellette, B.F.F., eds. 2005. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Third edition. Wiley, 2005.
4. Durbin, R., Eddy. S., Krogh A. and Mitchison G.1998. Biological sequence analysis. Cambridge University Press.
5. Statisticasl analysis of gene expression microarray data. Edited s by Terry Speed. Chapman & HALL/CRC. Boca Raton-London-new-York. 2003.
6. Kohane I.S., Kho A. and Butte A.J. 2003. Microarrays for an Integrative Genomics. The MIT Press. Cambridge, Massachusetts, London.
7. Lodish H., Berk A., Zipursky S. L., Matsudaira P., Baltimore D., Darnell J.E. 1999. Molecular Cell Biology. New York: W. H. Freeman & Co.
8. Hartl D.L.Jones E.W. 2005. Genetics: Analysis of Genes and Genomes. Jones & Bartlett Publishers.
9. Black D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. Annual Reviews of Biochemistry 72, 291–336.
10. Matlin A.J., Clark F., Smith C.W.J. 2005. Understanding alternative splicing: towards a cellular code. Nature Reviews. 6, 386–398.
11. Do J.H., Choi D.K. 2006. Computational approaches to gene prediction. J Microbiol. 44, 137-44.
12. Mathé C., Sagot, M.F., Schiex, T. and Rouzé, P. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res., 30, 4103–4117.
13. Ashurst J.L. and Collins J.E. 2003. Gene annotation. Prediciton and testing. Annual Review of Genomics and Human Genetics 4, 69-88.
14. Brent M.R. 2005. Genome annotation past, present, and future: How to define an ORF at each locus. Genome Res. 2005. 15, 1777-1786.
15. Smale T., Kadonaga T. (2003). The RNA polymerase II core promoter. Annual review of biochemistry 72, 449–479.
16. Knudsen S. 1999. Promoter 2.0: for the recognition of PolII promoter sequences. Bioinformatics. 15, 356-361.
17. Hannenhalli S, Levy S. 2001. Promoter prediction in the human genome. Bioinformatics. 17, Suppl 1:S90-96.
18. Bioinformatics: Sequence, Structure and Databanks. 2000. Edited by Des Higgins and Willie Taylor. Oxford University Press.
19. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. 2008. Genbank. Nucleic Acids Res. 36(Database issue):D25-30.
20. Mardis E.R. 2008. Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics. 9: 387-402.
21. Next-Generation Genome Sequencing: Towards Personalized Medicine. 2008. Michal Janitz (Editor). Wiley VCH Verlag GnbH &Co. KGaA, Weinheim.
22. Mardis E.R and Lunshof J.E. 2009. A focus on personal genomics. Personalized Medicine. 6, 603-606.
23. Wagner M.J. 2009. Pharmacogenetics and personal genomes. Pers. Med.6, 643–652.
24. Hidde de Jong. 2002. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. Journal of Computational Biology. 9, 67-103.
25. Trey Ideker T., Galitski T., and Hood L. 2001. A new approach to decoding life: systems biology. Annual Review of Genomics and Human Genetics. 2, 343-372.
26. Luisi P.L., Ferri F., Stano P. 2006. Approaches to semi-synthetic minimal cells: a review. Naturwissenschaften. 93, 1-13.
27. Forster A.C. and Church G.M. 2006. Towards synthesis of a minimal cell. Mol Syst Biol. 2006; 2, 45.
28. Wheeler D.L. et.al. 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 36, D13–D21.
29. Sussman J. L., Lin D., Jiang J., Manning N. O., Prilusky J., Ritter O. and Abola E. E. 1998. Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules. Acta Cryst. D54, 1078-1084
30. Koehl P. 2001. Protein structure similarities. Current opinion in structural biology, 11, 348-353.
31. Kitchen D.B., Decornez H., Furr J.R., Bajorath J. 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. 3, 935-949.
32. DNA REPAIR, GENETIC INSTABILITY, AND CANCER edited by Qingyi Wei , Lei Li & David J Chen. World Scientific 2007.
33. Okamura, K., Feuk, L., Marquès-Bonet, T., Navarro, A. and Scherer, S.W. 2006, Frequent appearance of novel protein-coding sequences by frameshift translation, Genomics, 88, 690–697.
34. Raes J., van de Peer Y. 2005. Functional divergence of proteins through frameshift mutations. Trends Genetics. 21, 428–431.
35. Kramer E.M.,, Su, H-J., Wu, C.C., and Hu J.M. 2006. A simplified explanation for the frameshift mutation that created a novel C-terminal motif in the APETALA3 gene lineage. BMC Evolutionary Biology. 6, 30-36
36. Fickett, J.W. 1998. Predictive methods using nucleotide sequences. Methods Biochem Anal. 39, 231-245.
37. Staden R. 1994. Staden: statistical and structural analysis of nucleotide sequences. Methods Mol. Biol. 25, 69–77.
38. Baxevanis A.D. Predictive methods using DNA sequences. Methods Biochem Anal. 43, 233-252.
39. Gutiérrez G, Oliver J.L, Marín A. 1994. On the origin of the periodicity of three in protein coding DNA sequences. J. Theor. Biol. 167, 413-414.

40. Gao J., Qi Y. and Cao Y., Tung W.W. 2005. Protein Coding Sequence Identification by Simultaneously Characterizing the Periodic and Random Features of DNA Sequences. J. Biomed. Biotechnol. 2, 139–146.
41. Yin C. and Yau S.S. 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J. Theor. Biol. 247, 687–694.
42. Eskesen S.T., Eskesen F.N., Kinghorn I B. and Ruvinsky A. 2004. Periodicity of DNA in exons. BMC Molecular Biology. 5,12.
43. Bibb, M.J., Findlay, P.R. and Johnson, M.W. 1984. The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. Gene. 30, 157-166.
44. Frenkel, F.E. and Korotkov, E.V. 2008. Classification analysis of triplet periodicity in protein-coding regions of genes. Gene. 421, 52-60.
45. Trifonov, E.N. 1999. Elucidating sequence codes: three codes for evolution. Ann NY Acad Sci. 870, 330–338.
46. Eigen, M. and Winkler-Oswatitsch, R. 1981. Transfer-RNA: the early adaptor. Naturwissenschaften. 68, 217–228.
47. Zoltowski M. Is DNA Code Periodicity Only Due to CUF - Codons Usage Frequency? Conf Proc IEEE Eng Med Biol Soc. 2007;1:1383-1386.
48. Antezana, M.A. and Kreitman, M. 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J. Mol. Evol. 49, 36-43.
49. Korotkov, E.V., Korotkova, M.A., Frenkel, F.E. and Kudryashov, N.A. 2003. Information approach for search of periodicity of symbolical sequences. Molek. Biol. (Russian) 37, 372-386.
50. Issac, B., Singh, H., Kaur, H. and Raghava, G. P. S. 2002. Locating probable genes using Fourier transform approach. Bioinformatics, 18, 196–197.
51. Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S. and, Ramaswamy R. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Comput. Appl. Bioscie. 13, 263–70.
52. Azad, R.K. and Borodovsky, M. 2004. Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. Briefings in Bioinformatics. 5, 118-130.
53. Henderson, J., Salzberg, S. and Fasman, K.H. 1997. Finding genes in DNA with a Hidden Markov Model. J. Comput. Biol. 4, 127-141.
54. Snyder E.E. and Stormo G.D. 1993. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. Nucleic Acids Res. 21, 607–613.
55. Thomas A. and Skolnick M.H. A probabilistic model for detecting coding regions in DNA sequences. 1994. IMA J. Math. Appl. Med. Biol. 11, 149–160.
56. Korotkov E.V., Korotkova M.A. and Kudryashov N.A. 2003. Information decomposition method for analysis of symbolical sequences. Physics Lett. A. 312, 198–310.
57. Kullback S, Information Theory and Statistics. New York: Wiley, 1959.
58. Hudson, D.J. 1964. Statistics. Lectures on Elementary Statistics and Probability, CERN, Geneva.
59. Needleman, S.B, Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol.Biol. 48, 443-453.
60. Altschul, S.F., Gish W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403-410.
61. UniProt Consortium. 2007. The Universal Protein Resource. Nucl. Acids Res. 35, 193-197.
62. Frenkel F.E., Korotkov E.V. 2009. Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. DNA Res. 16, 105-114.
63. Bollenbach, T., Vetsigian, K. and Kishony, R. 2007. Evolution and multilevel optimization of the genetic code. Genome Res. 17, 405-412.

Коротков Евгений Вадимович. Руководитель научной группы по биоинформатике, ведущий научный сотрудник Центра «Биоинженерия» РАН, профессор МИФИ. Окончил МИФИ в 1974 году. Доктор биологических наук, профессор. Область научных интересов: биоинформатика, изучение символьных последовательностей. Email: genekorotkov@gmail.com

Короткова Мария Александровна. Доцент МИФИ. Окончила МИФИ в 1976 году. Кандидат технических наук. Область научных интересов: теория алгоритмов, математические модели в биологии. Email: bioinf@rambler.ru.