ЭМПИРИЧЕСКОЕ ПРЕДСКАЗАНИЕ И РАСПОЗНАВАНИЕ ОБРАЗОВ
(Вычислительные системы)
1980 год Выпуск 83

УДК 519.767.6:577

АНАЛИЗ ГЕНЕТИЧЕСКИХ ТЕКСТОВ. I. 1-ГРАММНЫЕ ХАРАКТЕРИСТИКИ

В.Д.Гусев, В.А.Куличков, Т.Н.Титкова

Назовем текстом конечную упорядоченную символьную последовательность, составленную из элементов конечного алфавита. Пусть N – длина текста (в символах), $n = |A|$ – число элементов (мощность) алфавита A. Подпоследовательность текста, состоящую из 1 расположенных подряд символов, будем называть 1-граммой ($l = 1,2,..,N$), а анализ текста, основанный на подсчете частот встречаемости в нем различных 1-грамм, – 1-граммным анализом.

Под генетическим текстом будем понимать представление молекул нерегулярных полимеров (ДНК, РНК, белков) в виде упорядоченной последовательности мономеров (нуклеотидов или аминокислот). Важность исследования данного класса молекул обусловлена тем, что они определяют протекание наиболее существенных генетических процессов в организмах. Алфавит нуклеотидов состоит из 4 символов (А,Т,Г, Ц для ДНК и А,У,Г,Ц для РНК), алфавит аминокислот – из 20 символов.

В последние годы благодаря совершенствованию техники определения нуклеотидного состава молекул были расшифрованы и опубликованы первые полные нуклеотидные последовательности (геномы) простейших микроорганизмов (ΦX174, G4, SV40, PBR322, FD, MS2 и ряда других). Длины этих геномов составляют порядка (10^3 – 10^4) символов. Весьма актуальной задачей становится создание и пополнение банка данных по микробиологическим объектам и пакетов прикладных программ по исследованию различных свойств этих объектов.

Цель данной работы – исследование 1-граммных характеристик текстов, кодирующих наследственную информацию простейших микроорганизмов. Указан ряд содержательных задач, для решения которых представляется перспективным использовать 1-граммное описание текста. Ряд примеров иллюстрирует возможность выявления по 1-грам-

мным характеристикам функционально значимых фрагментов (или особенностей) текста.

Отметим, что наряду с информацией о частотах встречаемости различных 1-грамм в тексте очень важна и информация о местах вхождения каждой конкретной 1-граммы в текст. Эта информация зачастую дает возможность отличать случайные совпадения 1-грамм от неслучайных, функционально значимых.

I. Структура геномов вирусных частиц [I]. Из шести исследовавшихся геномов (см.выше) пять первых представляют собой кольцевые молекулы ДНК. Поскольку двухцепочечные молекулы ДНК построены по принципу комплементарности (т.е. любая цепь однозначно получается из другой повсеместной заменой нуклеотидов A,T,Г,Ц нуклеотидами T,A,Ц,Г соответственно), анализу подвергалась лишь одна из них. В тех случаях, когда это было возможно, анализировались цепи, соответствующие структуре м-РНК (ФX174, G4, FD, MS2). Шестой геном (MS2) описывает одноцепочечную молекулу РНК.

Структура геномов фагов и вирусов определяется основными генетическими процессами, обеспечивающими воспроизводство этих организмов. К таким процессам относятся редупликация (копирование ДНК или РНК), транскрипция (синтез РНК на ДНК) и трансляция (синтез белка на РНК). Части текста, отвечающие отдельным актам редупликации, транскрипции и трансляции, называются соответственно р е п л и к о н о м, с к р и п т о н о м и ц и с т р о н о м. Каждая из этих функциональных единиц имеет свои знаки начала и конца - и н и ц и а т о р ы и т е р м и н а т о р ы, представляющие собой нуклеотидные последовательности длиной от трех до нескольких десятков символов.

Между перечисленными функциональными единицами существует, как правило, отношение иерархического вложения: геном - это совокупность репликонов, репликон - последовательность скриптонов, скриптон,в свою очередь, разбивается на цистроны, каждый из которых уже кодирует определенный белок. Текст, отвечающий цистрону, разбивается при этом на последовательность триплетов (кодонов),кодирующих в соответствии с генетическим кодом определенные амино - кислоты. Отдельные цистроны могут располагаться последовательно, не пересекаясь друг с другом, могут частично пересекаться, могут быть даже целиком вложенными один в другой. Различие в синтезируемых ими белках достигается в последнем случае за счет использования отличающихся по фазе триплетных рамок считывания, либо (при совпадающих фазах) за счет несовпадения длин кодирующих последовательностей.

Приведенное выше описание структуры генома весьма схематично, но достаточно для наших целей. Следует отметить лишь, что наряду с указанными функциональными единицами, назначение которых в той или иной степени выяснено, существуют некоторые другие, функциональное назначение которых пока непонятно (нетранскрибируемые участки ДНК, межцистронные интервалы, интроны). Исследование различных характеристик этих участков (в том числе и 1-граммных) и сопоставление их с аналогичными характеристиками известных функциональных единиц может дать толчок к выявлению их функционального назначения.

2. Примеры содержательных задач. Укажем несколько задач из области анализа нуклеотидных последовательностей, при решении которых существенную пользу может принести знание 1-граммных характеристик этих последовательностей.

Следует сразу заметить, что 1-граммные характеристики расшифровываемых нуклеотидных текстов (или отдельных участков этих текстов) всегда интересовали исследователей. Однако интерес этот в подавляющем большинстве случаев был односторонним. А именно анализировались лишь участки геномов, соответствующие цистронам (т.е. участки, кодирующие белки), причем анализ последних ограничивался преимущественно подсчетом частот использования различных кодонов (т.е. триграмм) и исследованием распределения нуклеотидов по различным позициям кодонов.

Интерес к характеристикам третьего порядка обусловлен, естественно, триплетной структурой генетического кода. Нет никаких оснований предполагать, что они будут играть основную роль и при исследовании некодирующих частей геномов. Ниже будет показано, что важная информация содержится не только в характеристиках третьего порядка, но и в характеристиках меньшего ($l = 1,2$) и большего ($l = 4,5,\ldots, l_{max}$) порядков. Параметр l_{max} здесь соответствует тому минимальному значению l, начиная с которого в тексте уже отсутствуют повторяющиеся l-граммы, т.е. для любой l-граммы текста x частота ее встречаемости $F_l(x) \equiv 1$ при $l_{max} \le l \le N$.

2.1. Задача разметки текста сводится к построению алгоритмов автоматического выделения функциональных единиц геномов (см. п.1) и соответствующих им знаков пунктуации (инициаторов и терминаторов). Эта задача разбивается на ряд отдельных подзадач, многие из которых в формальной постановке сводятся к задаче распознавания.

Укажем для примера на задачу автоматического выделения промоторов - знаков пунктуации, ответственных за начало процесса транскрипции. Каждый промотор представляет собой последовательность длиной до 50 символов. В настоящее время расшифровано уже несколько десятков промоторов, т.е. установлена их первичная структура и доказано (молекулярно-биологическим анализом), что соответствующие последовательности нуклеотидов выполняют функции инициаторов транскрипции. Совокупность этих последовательностей может рассматриваться как обучающая выборка по промоторам в задаче классификации "промотор-не промотор". Для элементов этой выборки существует реальное опознающее устройство - РНК-полимераза. Обучающая выборка по "не промоторам" может быть получена из текстов уже расшифрованных геномов с исключенными промоторными зонами.

Целью обучения является построение решающего правила, моделирующего принцип действия РНК-полимеразы, т.е. позволяющего отнести произвольную последовательность к одному из двух указанных классов. Результат классификации может быть подтвержден (либо опровергнут) весьма трудоемким молекулярно-биологическим анализом. Удачное решение задачи классификации формальными методами (т.е. прямо по виду символьной последовательности) может существенно упростить и ускорить процедуру выявления генетической структуры генома.

В основу построения решающего правила кладется различие в характеристиках элементов первого и второго образов, выявляемое при анализе обучающих выборок. Нами исследовалась возможность использования для этой цели 1-граммных характеристик. А именно была выдвинута и экспериментально подтверждена гипотеза о том, что различие между промоторами и "не промоторами" наиболее ярко проявляется на уровне длинных 1-грамм ($l \geq 7$). Было показано, что существуют достаточно длинные 1-граммы, типичные только для промоторов, т.е. 1-граммы, каждая из которых встречается как минимум в двух промоторах, но не встречается у "не промоторов".

Примерами таких 1-грамм являются: ТГТТГАЦА (в промоторах λ-PL, E.COLI-TRP, ΦX174D, S.ТУРН-TRP), АЦАЦТТ (в промоторах E.COLI-LAC, E.COLI-GAL, TRHK-ТИР, E.COLIK12-ARAC), ГЦГТТГАТА (в промоторах λ-PL, λ-PR, λ-PRM) и ряд других. Анализ позиций, занимаемых этими 1-граммами внутри промоторов, подтверждает, что совпадение 1-грамм у разных промоторов носит неслучайный характер. Каждой из указанных 1-грамм внутри разных промоторов соот-

ветствуют одинаковые или очень близкие позиции. Таким образом, сочетание 1-граммного и позиционного анализов позволяет в данном случае выявить функционально важные зоны внутри промоторов и установить возможные варианты нуклеотидных последовательностей для этих зон.

Выделенное множество характерных (в упомянутом выше смысле) 1-грамм естественно назвать "покрытием" обучающей выборки промоторов в случае, если каждый промотор из этой выборки содержит хотя бы одну характерную 1-грамму. В качестве примера укажем, что для обучающей выборки из 24 промоторов удалось получить покрытие из 10 характерных 1-грамм ($6 \le l \le 10$). Элементы покрытия в сочетании с информацией о наиболее вероятных зонах их расположения внутри промоторов представляют основу для выработки различных стратегий распознавания.

2.2. Задача вычисления эволюционного расстояния между геномами является составным элементом более общей задачи построения филогенетических древ для различных семейств геномов. Можно выделить следующие элементарные преобразования, посредством которых осуществляется эволюция геномов:

1) замена одного основания другим;

2) вставка цепочки нуклеотидов в исходную последовательность;

3) устранение цепочки нуклеотидов из исходной последовательности;

4) инверсия (замена цепочки нуклеотидов в исходной последовательности элементами комплементарной цепочки, прочитанными в обратном порядке). Например, цепочка ААТЦГТТЦ исходной последовательности будет заменена инвертированной ГАЦЦГАТТ, которая получается из комплементарной к исходной (ТТАГЦЦАГ) прочтением ее справа налево;

5) дупликация (повторение цепочки символов);

6) транспозиция (перенос цепочки символов из одного места последовательности в другое).

Естественно определить меру различия между двумя геномами как минимальное число осмысленных с биологической точки зрения преобразований типа 1-6, переводящих один геном в другой. Вычисление этой меры представляет нетривиальную комбинаторную задачу, пока нерешенную в общем виде. Для частного случая, когда допустимыми являются лишь преобразования 1-3, причем два последних при -

менимы только к цепочкам длины I, расстояние вычисляется с помощью метода динамического программирования. При этом существует однозначная связь между расстоянием и длиной максимально длинной общей подпоследовательности двух последовательностей (геномов) [2].

Отметим, что преобразования 2-6 носят "групповой" характер, т.е. применимы к цепочкам, состоящим из многих (иногда до нескольких десятков) символов. Для вычисления расстояния необходимо уметь отыскивать в текстах участки, соответствующие вставкам (устранениям), инверсиям, дупликациям и транспозициям. Основу для этого дает вычисление и сопоставление полных частотных спектров обоих текстов.

Так, например, различие в частотах встречаемости какой-либо достаточно длинной*) 1-граммы в двух родственных текстах, как правило, свидетельствует о том, что в процессе эволюции имела место дупликация данной 1-граммы. Сопоставление результатов частотного и позиционного анализов зачастую позволяет выявить дупликацию даже при наличии одного текста. Об этом сигнализирует близость (а при многократной дупликации - периодичность) расположения повторяющихся 1-грамм.

Для выявления инверсий необходимо сравнить частотные спектры двух текстов: любого из исходных и инвертированного оставшегося. Наличие достаточно длинных (например, по отношению к схеме независимого порождения) совпадающих 1-грамм в обоих текстах говорит о возможности инверсий. Если в результате позиционного анализа выявляется, что номера позиций, занимаемых интересующей нас 1-граммой, в обоих исходных текстах совпадают или близки друг к другу, то предположение о наличии инверсии подтверждается, в противном случае возможна транспозиция с инверсией, либо случайное совпадение.

Соображения, аналогичные вышеизложенным, могут быть положены и в основу отыскания транспозиций. О наличии вставок (устранений) сигнализирует различие в длинах анализируемых текстов. Для локализации вставок целесообразно осуществить предварительную синхронизацию обоих текстов по выявленным характерным участкам, в качестве которых могут выступать знаки пунктуации, совпадающие 1-грам-

*) При малых значениях l частоты одинаковых 1-грамм в родственных геномах будут отличаться почти всегда, поскольку степень искажения частотных характеристик под влиянием любого из преобразований I-6 обычно тем больше, чем меньше l.

мы, инверсии, дупликации. Различие длин последовательностей, заключенных между идентичными знаками синхронизации в обоих геномах, свидетельствует о наличии вставки (устранения).

2.3. Восстановление текста по фрагментам. Определение первичной структуры длинных последовательностей (например, целых геномов) в настоящее время технически трудно осуществимо. С помощью существующих методик удается восстанавливать первичную структуру лишь относительно коротких участков (~ 20-300 нуклеотидов). Поэтому для определения первичной структуры всего генома его предварительно разрезают на более короткие участки (фрагменты) специальными ферментами-рестриктазами, определяют первичную структуру каждого участка, а затем, соединяя нужным образом расшифрованные куски, восстанавливают последовательность, соответствующую геному в целом.

Определение порядка следования фрагментов является нетривиальной задачей. Для восстановления текста используют наборы фрагментов, соответствующие разным рестриктазам. Поскольку каждая рестриктаза разрезает геном лишь в характерных для нее участках (например, рестриктаза НаеШ реагирует только на сочетание ГГЦЦ, разрывая связи между Г и Ц), фрагменты, получаемые от разных рестриктаз, оказываются перекрывающимися. Это свойство и используется для упорядочения фрагментов.

С формальной точки зрения задача восстановления текста по набору фрагментов, полностью его покрывающих, сводится к исследованию вопроса о существовании единственного (с учетом перекрываемости) варианта упорядочения фрагментов, полученных при фиксации определенного набора стратегий расчленения текста (набора рестриктаз). Предложено несколько способов сведения данной задачи к задаче отыскания эйлеровых цепей на графе. Структура соответствующего графа определяется типом текста и используемым набором рестриктаз.

Поскольку с каждой рестриктазой ассоциируется определенная 1-грамма, то количество таких 1-грамм в тексте и их расположение будут определять соответствующий набор фрагментов. Зная 1-граммные и позиционные характеристики родственных текстов, можно целенаправленно формировать набор рестриктаз, так чтобы добавление каждой новой рестриктазы в максимальной степени уменьшало неоднозначность, возникающую при восстановлении текста по уже имеющемуся набору фрагментов.

3. Частотные характеристики генетических текстов. В табл.1 приведены частотные характеристики текстов ФХ174, G4, FD, SV40, PBR322 и M82 для значений l = 1,2,3. Все l-граммы упорядочены по убыванию частоты встречаемости их в тексте. Порядковый номер r каждой l-граммы будем называть ее рангом. Параметр r изменяется в диапазоне от 1 до M_l, где M_l - количество различных l-грамм в тексте.

Поскольку с увеличением l объем таблиц возрастает очень быстро (при малых l, как правило, $M_l = n^l$; при $l \to l_{max}$ значение M_l ограничено величиной $N-l+1$), для значений $l > 3$ выборочно приведены лишь интегральные частотные характеристики E_l^k и S_m (табл.2). Здесь E_l^k - количество различных l-грамм текста, каждая из которых встречается ровно k раз; $S_m = \sum_{i=1}^{m} k(i) \cdot E_l^{k(i)}$ - текущая сумма числа l-грамм, соответствующих m первым частотным градациям, упорядоченным по убыванию.

Наиболее длинные повторяющиеся l-граммы, встретившиеся в каждом из текстов, приведены в табл. 3.

Табл. 4 содержит энтропийные характеристики текстов [4] (значения $\hat{H}_l$ и $\hat{R}_l$ для l = 2,...6). Ввиду ограниченности длины текстов энтропийные оценки с увеличением l становятся статистически недостоверными. Формально это выражается в том, что с ростом l величины $\hat{H}_l \to 0$, а $\hat{R}_l \to 1$. Ввиду отсутствия эффективных подходов к определению порогового значения $l(N)$, определяющего статистически достоверные результаты от недостоверных, мы можем воспользоваться для этой цели лишь некоторыми косвенными соображениями. Не обсуждая их детально, укажем, что в данном случае можно, по-видимому, в качестве оценки для $l(N)$ взять значение $\hat{l}(N) = 4$.

В табл.6 приведены величины коэффициентов ранговой корреляции Спирмэна между каждой парой текстов для значений l = 2, 3, 4. При больших значениях l уже не все из 4^l возможных комбинаций нуклеотидов присутствуют в тексте, поэтому использовать данную меру сходства без соответствующей модификации нецелесообразно. Нуклеотиды T (в ДНК-содержащих геномах) и У (в РНК-содержащих) при вычислении ранговых мер сходства отождествлялись.

4. Сравнительный анализ l-граммных характеристик генетических текстов. Ниже обсуждаются наиболее интересные, на наш взгляд, закономерности, выявленные при l-граммном анализе текстов. Далеко

не для всех из них удается предложить соответствующую интерпретацию. И даже там, где это делается, интерпретация, как правило, носит характер гипотезы. Укажем вначале на два интересных свойства, вытекающих из анализа характеристик первого и второго порядков.

4.I. К л а с с и ф и к а ц и я п о Т, А- и л и Ц, Г - п р е о б л а д а н и ю. Сопоставление характеристик первого порядка показывает, что все тексты можно разделить на две группы в зависимости от ранжировки нуклеотидов по частоте встречаемости. В первую группу входят тексты ФX174, G4, FD и SV40 ,в которых преобладают нуклеотиды Т и А (Т, А - богатые тексты), во вторую - PBR322 и MS2 (Ц, Г - богатые). Промежуточные варианты, когда Т и А перемежаются с Ц и Г при частотном упорядочении, отсутствуют.

Данная классификация,по-видимому, характеризует вторичную структуру молекул РНК, кодируемых этими геномами, т.е. их пространственную конфигурацию. Известно, что бактериофаг MS2 в отличие от объектов первой группы имеет сложную вторичную структуру.Ее отличительным признаком является наличие большого количества длинных "шпилек" - участков самокомплементарности одноцепочечной молекулы MS2 . В [3] показано, что шпильки формируются преимущественно из кодонов, содержащих нуклеотиды Ц и Г в третьей позиции. Поскольку комплементарная пара Ц ≡ Г обладает тремя водородными связями, а пара А = У лишь двумя, преимущественное использование Ц и Г в третьей позиции (а только здесь и допустимы вариации) должно обеспечить бóльшую устойчивость шпилек.

Возможность подтверждения либо опровержения предложенной интерпретации содержится в анализе пока еще не установленной вторичной структуры РНК, транскрибируемых с генома плазмиды PBR322 .Он существенно отличается от генома MS2 тем, что является ДНК-содержащим и двухцепочечным, аналогично объектам из первой группы.Если его РНК обладают слабо выраженной вторичной структурой, то гипотеза о связи ранжировки нуклеотидов в частотной характеристике первого порядка со сложностью и стабильностью вторичной структуры соответствующих молекул отпадает. В противном случае эта гипотеза получает сильное подтверждение.

4.2. ЦГ - э ф ф е к т у SV40. Анализ частотных характеристик второго порядка (табл.I) обнаруживает очень сильный аномальный эффект в частоте встречаемости биграммы ЦГ у вируса SV40. Эта биграмма встречается в тексте всего лишь 27 раз (r = I6). Для

Таблица I

Частотные характеристики генетических текстов для значений
l = I,2,3

l	r	ΦX174		G4		FD		SV40		PBR322		M32	
		l-грамма	$P_l(r)$	l-грамма	$P_l(r)$	l-грамма	$P_l(r)$	l-грамма	$P_l(r)$	l-грамма	$P_l(r)$	l-грамма	$P_l(r)$
l = 1	I	Т	I684	А	I5I9	Т	22I0	Т	I582	Ц	I208	Ц	932
	2	А	I29I	Т	I5I0	А	I577	А	I5I8	Г	II34	Г	928
	3	Г	I254	Ц	I446	Г	I32I	Ц	I094	Т	I036	У	875
	4	Ц	II57	Г	II02	Ц	I299	Г	I033	А	984	А	834
l = 2	I	ТТ	572	АА	554	ТТ	8I5	ТТ	575	ГЦ	378	ЦГ	266
	2	ТГ	480	ЦГ	503	АА	544	АА	533	ЦГ	329	УЦ	263
	3	ЦГ	404	ТЦ	443	АТ	5I5	ЦА	422	ТЦ	306	ГТ	250
	4	АА	395	АТ	392	ЦГ	483	ЦГ	398	ЦА	300	ЦУ	242
	5	АТ	383	ТГ	39I	ТГ	48I	ТГ	389	ЦЦ	292	ГЦ	23I
	6	ГЦ	347	АЦ	386	ТА	469	АТ	352	ЦГ	287	ГУ	230
	7	ГА	327	ЦА	352	ТЦ	445	АГ	345	ГТ	285	ЦЦ	223
	8	ГТ	325	ТТ	342	ГТ	397	ТА	326	ТГ	28I	АА	222
	9	ТЦ	320	ТА	334	ГТ	322	ТЦ	292	ТТ	259	ГА	2I7
	I0	ТА	3I2	ГЦ	3I3	ГЦ	3I6	АЦ	288	АА	258	АЦ	2I5
	II	ЦГ	267	ЦЦ	304	ГА	286	ГТ	272	АТ	254	УУ	2II
	I2	АЦ	260	ЦГ	287	ЦА	278	ГЦ	266	АГ	239	УГ	207
	I3	ЦА	257	ГА	279	ЦГ	274	ГГ	257	ГА	236	АГ	204
	I4	ГГ	255	ГГ	272	АЦ	274	ЦЦ	247	ГГ	235	ЦА	20I

	I5	АГ	252	ГТ	238	ЦЦ	264	ГА	237	АЦ	232	УА	I94
	I6	ЦД	229	АГ	I86	АГ	244	ЦГ	27	ТА	I90	АУ	I92
	I	ТТТ	I78	ААА	I97	ТТТ	273	ТТТ	238	ЦГЦ	II2	УЦГ	89
	2	ГЦГ	I49	ЦАА	I57	АТТ	223	ААА	2II	ГЦГ	I05	ЦУЦ	80
	3	ТТГ	I42	ТЦГ	I52	ААА	20I	ЦТГ	I34	ГТЦ	I0I	ГТУ	75
	4	АТГ	I40	ЦТЦ	I44	ТТА	I88	ЦАА	I32	ГЦГ	99	ГЦУ	74
	5	ГТТ	I40	ААЦ	I35	ТТЦ	I86	АЦА	I29	ЦТГ	95	ЦГА	73
	6	ТГА	I39	ААГ	I35	ААГ	I78	АТТ	I29	ЦАГ	93	ЦГУ	70
	7	ААА	I33	ЦТГ	I32	ГТТ	I77	ТТА	I22	ТЦА	9I	УГГ	69
	8	ЦТГ	I32	ГЦГ	I29	ТТГ	I68	ЦАГ	II2	АТЦ	90	УЦЦ	68
	9	ТТЦ	I32	АЦГ	I26	ТГТ	I64	ЦТГ	III	ГЦА	90	УУЦ	68
	I0	ТГЦ	I30	ТЦА	II9	ТАА	I64	ААГ	III	ЦТГ	89	ГТГ	64
	II	ТГТ	I30	ТАГ	II9	ТАГ	I58	ТТЦ	I09	ЦЦГ	88	ЦГТ	63
	I2	ЦТГ	I27	АТГ	II8	ТГА	I54	ТАА	I08	ААА	88	ГГЦ	63
	I3	АТТ	I27	АТЦ	II8	ЦТГ	I42	ТТГ	I06	ТЦЦ	87	ГУУ	63
	I4	ТТА	I20	ЦТГ	II6	ГТГ	I36	ААГ	I06	ГЦЦ	84	УАЦ	63
	I5	ТАГ	II5	ТГЦ	II6	ЦГГ	I29	ААЦ	I05	ЦАГ	84	ЦАА	62
	I6	ТЦГ	I04	ЦГА	III	ТЦГ	I22	ЦАГ	I05	ЦГЦ	80	ЦЦУ	62
	I7	ААГ	I03	ТТЦ	I07	ЦГЦ	I2I	ТЦЦ	I04	ТЦГ	80	АЦГ	62
	I8	ГАГ	I0I	ТГГ	I06	ГЦГ	II9	ТГГ	I03	АГЦ	78	ГУГ	6I
	I9	ТАА	I0I	ТГА	I03	ТГГ	II8	ТЦА	I02	ЦТГ	78	ААЦ	6I
	20	ГТГ	I00	ТАА	I0I	ТЦА	II3	ГЦГ	I02	АЦЦ	76	ЦГЦ	6I
	2I	ААГ	92	ГАА	99	АТА	III	АЦГ	I02	ТТЦ	76	ЦЦГ	60

1 = 3

Продолжение таблицы I

1	2	[illegible]		64		FD		8740		PBR322		M82	
		1-грамма	F_1(г)	1-грамма	F_1(г)	1-грамма	F_1(г)	1-грамма	F_1(г)	1-грамма	F_1(г)	1-грамма	F_1(г)
	22	ЦГЦ	92	ТАЦ	98	ГАТ	110	ЦЦТ	102	ТТТ	75	АУЦ	59
	23	ЦГТ	92	ТТА	98	ЦЦТ	109	АГЦ	97	АТТ	73	ААА	58
	24	ТЦА	91	АЦА	96	АТТ	108	ЦЦА	97	ЦЦА	71	ЦУУ	57
	25	ЦАА	91	ЦГТ	96	ЦАА	103	ГТТ	97	ГТТ	70	ГАГ	57
	26	АЦТ	89	ЦЦТ	96	ЦТА	91	АГТ	95	ТТТ	69	УЦУ	56
	27	ГАЦ	82	ТЦГ	92	АЦТ	91	ТЦЦ	94	ЦАЦ	68	АЦЦ	56
	28	ТТТ	81	ЦГЦ	92	ТЦЦ	90	ГЦА	94	ЦГТ	67	ГУЦ	56
	29	ЦТА	78	ГАЦ	90	ТАЦ	89	ТЦГ	92	ГТА	67	АГА	55
	30	ГАГ	74	АТТ	90	ТГЦ	87	ТТТ	92	ЦЦГ	67	АГЦ	55
	31	АГТ	74	АЦЦ	89	ЦГТ	86	ТГА	90	ТЦЦ	66	ГЦГ	55
	32	АГА	73	ААГ	86	ААЦ	84	ТАТ	87	ГАА	66	ГАЦ	54
	33	ТАЦ	72	ЦАТ	83	ГТЦ	83	АГТ	87	ЦЦЦ	66	ААГ	54
	34	ГТЦ	72	ТЦЦ	80	ААГ	81	АТГ	83	ТГА	65	УУУ	54
	35	ГЦЦ	71	ЦЦГ	77	ГТА	79	ГАА	82	ААЦ	64	ГАА	54
	36	ЦЦГ	70	АЦГ	75	ЦГЦ	78	ЦТЦ	80	ТТТ	64	ЦУА	54
	37	ГАА	70	ГТТ	74	ТЦГ	78	ГГА	79	ТАТ	63	ГЦА	54
	38	ЦГЦ	67	ГТЦ	74	АГГ	77	АТА	74	ГГТ	63	УГЦ	53
	39	ГЦГ	67	ГТА	73	ГТГ	76	ГТТ	73	ТТГ	62	АУГ	53
	40	ТЦГ	67	ГЦЦ	72	ГАА	76	ЦАЦ	73	ГАТ	61	АГГ	53
	41	ААЦ	66	ТТГ	71	АТЦ	73	ЦТА	73	АГГ	61	ГАУ	52

1 = 3	42	ГТГ	66	ГТТ	70	ГЦГ	71	ТАГ	67	ГТЦ	60	ЦАГ	52
	43	ГТЦ	65	ГТГ	69	ЦАТ	69	АГА	66	ААГ	59	ЦГГ	51
	44	ЦАГ	64	ГЦА	69	ЦЦГ	68	АТЦ	66	ТТА	57	ЦЦЦ	51
	45	АЦГ	63	ЦЦА	68	АГЦ	68	ГТГ	66	ГТГ	57	ГУА	50
	46	ГТА	63	ГТЦ	67	ЦАГ	67	ТАЦ	64	ГАГ	56	АЦУ	50
	47	ЦАГ	62	АТА	66	ГТЦ	65	ГТТ	62	АЦГ	56	ЦЦА	50
	48	ЦЦГ	62	ТТТ	66	АЦЦ	63	ГАГ	60	ЦАА	55	УЦА	50
	49	АЦЦ	61	ТТГ	66	ГЦА	63	ГЦЦ	59	ЦГА	55	ААУ	49
	50	АТА	60	АГГ	64	АЦГ	63	ГТА	57	ТГТ	54	ЦАУ	49
	51	ГЦА	60	ЦАЦ	63	ГЦЦ	63	ГТЦ	57	ГТГ	54	ГЦЦ	48
	52	ЦЦА	59	ЦЦЦ	63	ЦЦГ	62	АЦЦ	52	ГАЦ	53	УАА	48
	53	ТТЦ	58	ГТА	59	ГАЦ	61	ГАТ	49	АЦГ	52	ГТА	48
	54	АТЦ	56	ГАТ	55	ГТГ	59	ГАЦ	46	АГГ	51	АЦА	47
	55	ГТА	54	ЦГА	55	ТАГ	58	ЦЦЦ	42	АГА	49	УУА	47
	56	АГЦ	53	ЦАГ	49	АГА	57	ГТЦ	37	ТАА	49	УГУ	44
	57	ЦГА	52	АГА	48	АГГ	53	ЦГТ	12	АЦА	48	АУА	43
	58	АГГ	52	ЦГТ	44	ЦЦЦ	48	ГЦГ	11	ГТА	48	УАУ	42
	59	АЦА	47	ГЦГ	43	АЦА	46	ЦГЦ	8	ТАЦ	47	ХУГ	42
	60	ЦАЦ	40	АГЦ	38	ЦЦА	45	ЦЦГ	6	АТГ	47	УАГ	41
	61	ЦЦЦ	38	АГГ	36	ГТА	44	АЦГ	5	ААТ	46	АГУ	41
	62	ЦГГ	31	ГАГ	35	ЦГА	42	ЦГТ	5	АТА	44	УГА	41
	63	ТАГ	24	ГГГ	24	ЦАЦ	39	ТЦГ	4	ЦГА	41	ЦАЦ	37
	64	ГГГ	20	ТАГ	16	ГАГ	38	ЦГА	2	ТАГ	31	АУУ	37

Таблица 2

Образцы интегральных частотных характеристик генетических текстов

l	m	ФX174			G4			FD			SV40			PBR322			MS2		
		K	E_l^k	S_m	K	E_l^k	S_m	K	E_l^k	S_m	K	E_l^k	S_m	K	E_l^k	S_m	K	E_l^k	S_m
l = 6	1	12	1	12	14	1	14	15	1	15	14	1	14	9	1	9	7	1	7
	2	11	1	23	11	1	25	13	1	28	13	1	27	7	2	23	5	11	62
	3	10	1	33	10	2	45	12	3	64	12	2	51	6	12	95	4	51	266
	4	9	8	105	9	4	81	11	6	130	11	2	73	5	23	210	3	186	824
	5	8	10	185	8	7	137	10	5	180	10	7	143	4	113	662	2	628	2080
	6	7	25	360	7	23	298	9	18	342	9	5	188	3	256	1430	1	1484	3564
	7	6	48	648	6	50	598	8	26	550	8	11	276	2	741	2912			
	8	5	82	1058	5	106	1128	7	42	844	7	25	451	1	1445	4357			
	9	4	198	1850	4	197	1916	6	74	1288	6	51	757						
	10	3	339	2867	3	337	2927	5	157	2073	5	96	1237						
	11	2	669	4205	2	746	4419	4	223	2965	4	166	1901						
	12	1	1176	5381	1	1153	5572	3	378	4099	3	364	2993						
	13							2	629	5357	2	628	4249						
	14							1	1046	6403	1	973	5222						
l = 10	1	2	47	94	2	36	72	7	5	35	3	2	6	2	17	34	2	2	4
	2	1	5238	5377	1	5496	5568	5	1	40	2	102	210	1	4319	4353	1	3556	3560
	3							4	8	72	1	5008	5218						
	4							3	10	102									
	5							2	65	232									
	6							1	6167	6399									
l = 18	1							3	21	63	2	42	84						
	2							2	7	77	1	5127	5211						
	3							1	6314	6391									

сравнения укажем, что частоты встречаемости биграмм с рангом I6 во всех остальных текстах в 7-9 раз выше.

Отметим, что исследователи, занимавшиеся изучением кодирующих частей генома SV40 и родственных вирусов, уже обращали внимание на чрезвычайно малое количество кодонов, содержащих биграмму ЦГ в позициях I-2 или 2-3. Наш анализ показывает, что данный эффект возникает уже на уровне биграмм, т.е. аномалии в использовании кодонов - его следствие, а не причина. Эффект наблюдается для генома в целом, как для кодирующих частей (при любых фазах считывания), так и для некодирующих. Выявление данного и ему подобных эффектов не требует предварительной разметки текстов и может быть автоматизировано с помощью достаточно развитой в настоящее время техники обнаружения аномальных наблюдений.

Мы не можем пока указать, какую функциональную нагрузку несет данный эффект. В [6] предпринята попытка объяснения данного эффекта тем, что в геномах вирусов эукариотов существует механизм перехода ЦГ в ТГ (5-метил-цитозин часто мутирует в Т, причем Ц метилируется именно в составе биграммы ЦГ). Этот процесс должен был бы приводить к соответствующему избытку ТГ. Проверка данной гипотезы на схеме независимого порождения (см.п.4.5) показала, однако, что некоторый избыток ТГ имеет место, но он далеко не компенсирует недостатка ЦГ.

Отметим, что из шести анализировавшихся микроорганизмов SV40 является единственным, который паразитирует на эукариотах, остальные - на прокариотах. Если данный эффект подтвердится и на других вирусах, паразитирующих на эукариотах, он может быть положен в основу алгоритма классификации текстов по признаку: эукариотический-прокариотический.

4.3. Наиболее длинные повторения. Значения параметра l_{max} составляют: для текста SV40 - 59, FD - 33, G4 и PBR322 - 14, ФX174 - 13, M52 - 11. При $l = l_{max} - 1$ в каждом из текстов уже появляются повторяющиеся l-граммы (см. табл.3). Анализ их расположения в текстах проливает некоторый свет на их происхождение.

Самый длинный повтор ($l = 58$) наблюдается в тексте SV40 (l-грамма №14 из табл.3). Эта l-грамма начинается и заканчивается триграммой ТГТ. В тексте обе 58-граммы расположены подряд так, что последние три символа первой l-граммы накладываются на первые три символа второй. Таким образом, наличие данного повтора можно

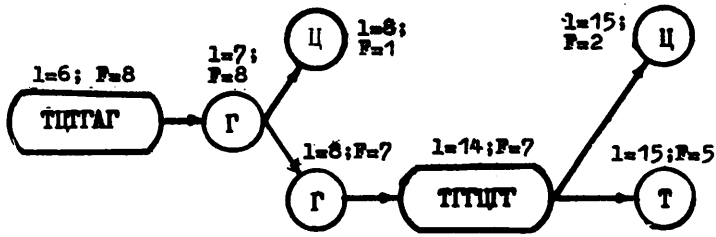
Таблица 3

Наиболее длинные повторяющиеся l-граммы генетических текстов

Текст	l	F_1	l-грамма	№ l-граммы
ϕX174	I2	2	ЦТТЦААГТАЦТГ	I
	I2	2	ЦТТЦТГЦЦГТТТ	2
	II	2	ТТЦТГТТГАТТ	3
	II	2	ЦТТЦТТГААГТ	4
G4	I3	2	ГТТГТЦГЦТАТТЦ	5
	II	2	ГЦТТТТАТТТЦ	6
	II	2	ЦТАТЦТЦТГГТЦ	7
	II	2	АЦЦЦЦТТАААГ	8
FD	32	2	ГТЦТЦТГАТТТТТГТЦГТТТЦТГАГТТТТГТЦГТ	9
	32	2	ТЦТТГАГТТТТТГЦТТЦТЦТГАГТТТГТГЦГТТТЦ	I0
	20	2	ГЦГТТТТЦТГАГТТТГТГЦГТТ	II
	I3	2	ААГТТААТТЦААА	I2
	I2	2	ЦГЦААТТТЦТТТ	I3
SV40	58	2	ТГТТТГТЦТГАЦТААТТГАГАТГЦАТГЦТТТТГЦ АТАЦТТЦТГТЦТГТГЦТТТТТАТГЦЦТТТ	I4
	2I	2	ТГТТГЦТГАТТТАТТТГЦТТГА	I5
	I3	2	ГАГТАЦАЦАГАГТ	I6
	I3	2	АААААЦЦАГАААГ	I7
	II	2	ЦТТЦАТТТТТТ	I8
PBR322	I3	2	АГЦАААААГТЦЦАГ	I9
	II	2	ТТГЦТГТЦГТТ	20
	II	2	ЦТТТТЦЦЦГЦЦ	2I
	II	2	ААААААТГАТЦТ	22
MS2	I0	2	АГАЦГЦЦГТЦ	23
	I0	2	АУУЦЦЦУЦАГ	24
	8	4	АГТУГЦЦУ	25
	8*	3	ГУУУУАЦА	26

было бы объяснить дупликацией соответствующей 55-граммы без наложения (но при этом маловероятно появление еще одного ТТГ после такой композиции) либо дупликацией всей 58-граммы с наложением (одним из возможных объяснений механизма наложения может быть "неравный" кроссинговер). Наличие данного повтора уже отмечалось. Функциональное назначение его пока не выяснено, известно лишь, что в этом районе инициируется репликация генома SV40. С интервалом в три символа от данного участка находится еще одна дупликация (дважды повторяется 21-грамма №15 из табл. 3).

Интересный пример дупликаций наблюдается в тексте FD . Ядро этой дупликации составляет 14-грамма X = ТЦТГАГТТТТГЦТ. Она легко выделяется при анализе дерева всевозможных продолжений 6-граммы ТЦТГАГ:



Из приведенного рисунка видно, что при изменении l от 8 до 13 дерево не имеет разветвлений, т.е. продолжается формирование какого-то сообщения (с неясной пока семантикой). При l = 14 формирование заканчивается, и далее дерево расщепляется. Три 15-граммы, заканчивающиеся на Т, следуют в тексте друг за другом, образуя тройную дупликацию, которой предшествует триграмма ГТЦ(ГТЦХТХТХТ). На расстоянии порядка 400 символов от этого участка расположен другой, образованный четырехкратным (с интервалом в 1 символ) повторением базовой 1-граммы и заканчивающийся триграммой ТТЦ(ХЦХТХЦХТТЦ). Обладая внутренней периодичностью, два этих участка содержат в себе и более длинные (по сравнению с базовой) повторяющиеся последовательности (1-граммы № 9, 10 и 11 из табл. 3).

Интересно отметить, что расположение базовых 1-грамм синхронизовано относительно рамки считывания при трансляции так, что каждая из них кодирует идентичные фрагменты белка. Это указывает на то, что в данном случае, в отличие от предыдущего, семантическую нагрузку дупликаций следует искать на уровне функционирования кодируемого ими полипептида.

Примером дупликаций являются также 1-граммы №5 (самый длинный повтор в G4) и №13 (PBR322) . Интересно отметить, что последняя 1-грамма, аналогично 58-грамме из SV40, имеет конец,совпадающий с началом (биграмма АГ). Механизм дупликации вновь улавливает это и осуществляет дупликацию с наложением. Более короткие 1 -граммы, например №1 (из ФХ174), №23 (из MS2), встречающиеся каждая по два раза, уже разнесены внутри своих текстов, т.е. наличие таких повторов может трактоваться и как случайный фактор.

Таблица 4

Оценки энтропии и избыточности генетических текстов

Текст	Энтропия и избыточность	1				
		2	3	4	5	6
ФХ174	$\hat{H}_1$	I,968	I,932	I,900	I,798	I,453
	$\hat{R}_1$	0,0I5	0,033	0,049	0,I00	0,273
G4	$\hat{H}_1$	I,963	I,937	I,903	I,8I5	I,479
	$\hat{R}_1$	0,0I8	0,03I	0,048	0,092	0,260
FD	$\hat{H}_1$	I,963	I,928	I,890	I,798	I,493
	$\hat{R}_1$	0,0I8	0,036	0,055	0,I0I	0,254
SV40	$\hat{H}_1$	I,906	I,980	I,862	I,8I0	I,45I
	$\hat{R}_1$	0,047	0,055	0,069	0,095	0,274
PBR322	$\hat{H}_1$	I,989	I,974	I,944	I,855	I,389
	$\hat{R}_1$	0,006	0,0I3	0,028	0,073	0,305
MS2	$\hat{H}_1$	I,997	I,985	I,956	I,823	I,284
	$\hat{R}_1$	0,00I	0,007	0,02I	0,088	0,357

4.4. Низкая избыточность. С учетом замечания о достоверности энтропийных оценок, сделанного в п.3,избыточность генетических текстов, как это следует из табл.4, очень невелика (не превышает 0,I). Для сравнения укажем, что оценки избыточности современных естественных языков (русского,английского, немецкого) дают значение $R \approx 0,7$.

Отсутствие избыточности в языке означает, что любая комбинация символов в нем является осмысленной. Именно по такому принципу построен генетический код, использующий все 64 возможные комбинации троек нуклеотидов. Анализ некодирующих частей показывает, что и там на уровне l = 3 нет запрещенных комбинаций.

В порядке нарастания избыточности тексты ранжируются следующим образом: MS2, PBR322, G4 и ФX174, FD, SV40 . Приведенная ранжировка, возможно, отражает многообразие и степень перекрывания различных функциональных единиц в геномах (в некотором смысле сложность организации геномов). У генома MS2 , избыточность которого минимальна, гены практически не перекрываются друг с другом. В геномах G4, ФX174, SV40 имеются зоны перекрытия, иногда весьма обширные. Наложение различных функциональных единиц (гена на ген, промотора на ген и т.д.) увеличивает количество ограничений, которым должна удовлетворять символьная последовательность, т.е. повышает избыточность языка. Аналогии подобного рода проводил К.Шеннон [4], обсуждая связь между избыточностью языка и возможностью построения кроссвордов.

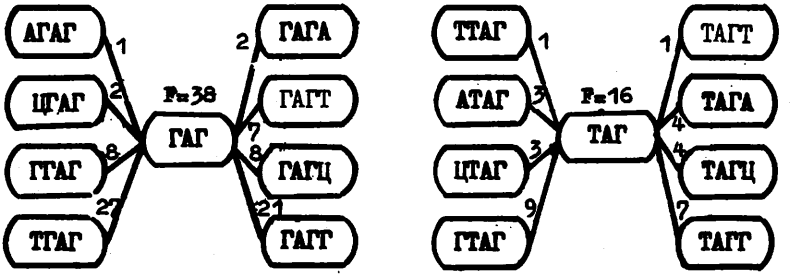
Низкая избыточность находит свое отражение в слабом наклоне частотных характеристик соответствующего порядка. Наклон характеристики 1-го порядка может быть определен как МНК - оценка параметра γ_1 в выражении $F_1(r) = C_1 r^{-\gamma_1}$, используемом для аппроксимации частотной кривой. Значения оценок этого параметра довольно сильно отличаются для текстов двух выделенных групп,но не превышают 0,5(l = 2,3). Это говорит о том, что закон Ципфа (близость параметра γ к единице для частотных словарей естественного языка) не выполняется для l-граммных характеристик генетических текстов. Более того, различия в значениях этого параметра могут быть положены в основу классификации генетических текстов.

4.5. Схема независимого порождения как грубая модель текста. Низкие значения избыточности R для l = 2,3,4 наводят на мысль о том,что, по крайней мере, в данном диапазоне значений l можно воспользоваться в качестве грубой модели текста схемой независимых испытаний с вероятностями порождения каждого символа α равными $F(\alpha)/N$, где $\alpha \in \{A,T,Г,Ц\}$, а $F(\alpha)$ - частота встречаемости символа α в соответствующем тексте. Основанием для этого служит тот факт, что для последовательности, полученной в соответствии со схемой независимых испытаний, $H_1 = H_2 = \ldots = H_l$ для любого l . Анализ табл.4 показывает, что для l = 2,3,4 мы делаем не слишком большую натяж-

ку, допуская эту гипотезу. В наилучшем соответствии с ней находятся тексты с наименьшей избыточностью (MS2, PBR322). Это подтверждается и прямым сопоставлением частот 1-грамм, наблюдаемых фактически и вычисляемых по схеме независимых испытаний.

Схема независимого порождения представляет интерес с точки зрения оценки многих интересующих нас параметров, например, таких как l_{max}, M_1, E_1^1 , определяющих трудоемкость алгоритмов обработки символьных последовательностей. С биологической же точки зрения несомненный интерес представляет анализ очень сильных отклонений фактических частот от модельных. Зачастую такие отклонения являются функционально значимыми.

Рассмотренный выше ЦГ-эффект у SV40 можно трактовать, с одной стороны, как аномалию в частотной характеристике, с другой стороны, - как сильное отклонение от схемы независимых испытаний ($F_{мод}$(ЦГ) = 217, $F_{эксп}$(ЦГ) = 27). Пример дупликации в тексте FD , описанный в п.4.3, можно также трактовать как аномальное отклонение от схемы независимых испытаний, поскольку при увеличении l от 8 до 14 частота соответствующей 1-граммы не должна была оставаться постоянной, а должна была уменьшаться на каждом шаге за счет разветвления дерева на 4 ветви с вероятностями $P(\alpha)$ ($\alpha \in \{A, T, Г, Ц\}$) для каждой ветви. Другой пример аномального ветвления - лево- и правосторонние расширения 1-грамм ГАГ(из FD) и ТАГ(из G4):



В табл. 5 приведен еще ряд примеров аномальных отклонений экспериментально наблюдаемых частот ($F_{эксп}$) от вычисленных в соответствии с моделью независимого порождения ($F_{мод}$). Отметим, что во всех текстах фактически наблюдаемые частоты встречаемости 1-грамм AA, AAA выше, а триграммы ТАГ (терминальный кодон) существенно ниже, чем это следует из схемы независимых испытаний.

Таблица 5

Аномальные отклонения от схемы независимого порождения

Текст	1-граммы	$F_{мод}$	$F_{эксп}$	1-граммы	$F_{мод}$	$F_{эксп}$	Коммен-тарий
ФХ174	АА	310	395	ААА	74	133	
	ТА	403	312	АТГ	94	140	иниц.
	ТГ	392	480	ЦГТ	63	31	
				ГТГ	68	20	
				ТАГ	94	24	терм.
G4	АГ	300	186	ААА	112	197	
	АА	413	554	АГТ	81	36	
	ЦТ	391	503	АТА	112	66	
				ТАГ	81	16	терм.
				ТТТ	111	66	
FD	АА	388	544	ААА	95	201	
	АГ	325	244	ЦАТ	110	69	
				ТАГ	112	58	терм.
				АГТ	112	53	
				ГАГ	67	38	
SV40	ЦГ	217	27	Все ЦГ-содержащие триплеты			
	ТА	459	325	ААА	128	212	
	ЦА	318	424	АТА	133	74	
				ЦАГ	63	113	
				ЦТГ	65	132	
				ГТГ	40	73	
				ТТТ	144	238	
PBR322				ААА	50	88	
				ЦЦЦ	92	66	
				АТЦ	65	90	
				ТАГ	61	31	терм.
MS2	УЦ	228	263	ЦАЦ	57	37	
				УЦГ	59	89	

Поскольку схема независимых испытаний является весьма грубым приближением к описанию генетических текстов, предпринимаются попытки использовать для описания отдельных частей геномов более сложные модели, например, в виде марковских цепей невысокого (первого и второго) порядка [5]. Оценки переходных вероятностей получаются при этом на основе 1-граммных характеристик соответствующих частей геномов. Такого рода кусочные представления могут быть использованы для решения задач разметки и генерации текстов.

4.6. Количественные оценки близости текстов. Простейшими мерами близости текстов являются коэффициенты ранговой корреляции ρ_l, вычисленные с использованием частотных характеристик l-го порядка (см. табл. 6, $l = 2,3,4$). Значения ρ_l в целом согласуются с тем разбиением текстов на 2 класса (по ТА- и ЦГ-преобладанию), которое было приведено в п. 4.1, поскольку коэффициенты ранговой корреляции между представителями разных классов близки к нулю или отрицательны. Представляет интерес проследить степень близости между текстами одного класса.

Таблица 6

Коэффициенты ранговой корреляции

Пара текст/текст	l		
	2	3	4
ФX174/SV40	0,55	0,43	0,37
ФX174/G4	0,57	0,51	0,58
ФX174/MS2	-0,08	0,01	
ФX174/FD	0,86	0,77	0,67
ФX174/PBR	0,10	0,13	0,21
SV40/G4	0,49	0,41	0,34
SV40/MS2	-0,30	-0,32	-0,15
SV40/FD	0,67	0,55	0,48
SV40/PBR	0,22	-0,04	-0,03
G4/MS2	-0,11	0,14	0,11
G4/FD	0,69	0,56	0,50
G4/PBR	0,24	0,15	0,18
MS2/FD	0,02	0,01	0,10
MS2/PBR	0,62	0,32	0,19
FD/PBR	0,04	-0,02	0,01

Весьма близки тексты ФX174 и G4 ($\rho_l > 0.5$ для всех l), что неудивительно, так как эти тексты кодируют родственные вирусы. Неожиданным оказывается факт близости этих текстов с текстом FD, поскольку внутренняя организация вирусов ФX174 и G4 сильно отличается от структуры FD. Объяснение может заключаться либо в том, что при различной внутренней организации похожими могут оказаться белки, кодируемые этими текстами, либо в том, что ранговые меры

близости не отражают адекватным образом различия во внутренней организации геномов. Последнее может объясняться тем, что ранговые меры близости не учитывают информации о расположении 1-грамм в тексте. Те же меры, которые это учитывают, весьма сложны для вычисления.

Тексты PBR322 и MS2 , будучи весьма близкими по биграммным характеристикам ($\rho_2 = 0.62$), с увеличением 1 становятся менее похожими ($\rho_4 = 0.18$). Так как характеристики четвертого порядка учитывают бо́льшую информацию, чем характеристики второго порядка, эти тексты, видимо, не следует считать слишком близкими.

5. Заключительные замечания. Проведенный анализ затронул всего лишь 6 текстов, поэтому многие выводы, сделанные выше, носят предварительный характер и подлежат дальнейшему уточнению по мере накопления материала. Однако уже сейчас можно сказать, что 1-граммный способ описания генетических текстов, сводящийся фактически к поиску всех периодичностей в символьной последовательности, является в сочетании с позиционным анализом адекватным средством для решения многих содержательных генетических задач.

Обобщением 1-граммного способа представления символьных последовательностей является представление, допускающее объединение в один класс не только совпадающих 1-грамм, но и 1-грамм, отличающихся от них в пределах заданной меры различия. Такое описание представляет интерес и с биологической точки зрения, например, в задаче поиска гомологичных участков в родственных геномах.

Литература

1. РАТНЕР В.А. Молекулярно-генетические системы управления.- Новосибирск: Наука, 1975. - 286 с.

2. WAGNER R.A., FISHER M.I. The String-to-String Correction Problem.-J.Assoc.Comput.Machinery,1974,v.21,N 1,p.168-173.

3. MASAMI Hasegawa, TERUO Yasunaga,TAKASHI Miyata.Secondary structure of MS2 phage RNA and Bias in code word usage. - Nucl. Acid.Res.,1979,v.7, N 7,p.2073-2079.

4. ШЕННОН К. Математическая теория связи. -В кн.: Работы по теории информации и кибернетике. М., 1963, с. 243-332.

5. ERICKSON J.W., ALTMAN J.J. A Search for Patterns in the Nucleotide Sequence of the MS2 Genome.- J.Math.Biology,1979, v.7, p.219-230.

6. BIRD A.P. DNA methylation and the frequency of CpJ in animal DNA. - Nucl.Acid.Res.,1980, v.8, N 7,p.1499-1505.

Поступила в ред.-изд.отд.
5 ноября 1980 года