Математическая биология и биоинформатика. 2013. Т. 8. № 1. С. 373–397.
URL: http://www.matbio.org/2013/Gumenuk_8_373.pdf

========== МАТЕРИАЛЫ IV МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ==========
========= «МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА» ==========

УДК: 004.8: 575

О средствах формального анализа строя нуклеотидных цепей

Гуменюк А.С.[*1], Поздниченко Н.Н.[1], Родионов И.Н.[***1], Шпынов С.Н.[****2]**

[1]*Омский государственный технический университет, Омск, 644050, Россия*
[2]*ФБУН «Омский НИИ природно-очаговых инфекций» Роспотребнадзора, Омск, 644080, Россия*

Аннотация. Объектом исследования в данной работе является числовая последовательность, отображающая отдельную нуклеотидную цепь, названная строем и формально представляющая взаимное расположение компонентов последовательности ДНК. Сформулированы выражения для числовых характеристик, которые характеризуют порядок элементов цепи. Продемонстрированы некоторые возможности применения этих числовых характеристик для решения задач таксономии организмов, их сравнения, сегментации ДНК и отображения её локальной структуры.

Ключевые слова: *количество информации, энтропия, строй цепи элементов, расположение, интервал, числовые характеристики строя цепи, удаленность, регулярность, нуклеотидная цепь, строй, строй нуклеотидной цепи, рибонуклеиновая кислота, рибосомальная РНК, митохондриальная РНК.*

ВВЕДЕНИЕ

Уже более 100 лет используются формальные средства для анализа знаковых последовательностей разной природы. В начале прошлого века при зарождении математической лингвистики начались статистические исследования текстов на естественных языках [1]. В 50–60-е годы на фоне широкого использования вычислительных машин отдельные исследователи применили формальный анализ к музыкальным произведениям и одновременно стали использоваться математические модели и средства анализа генетических текстов, т. е. нуклеотидных цепей, аминокислотных последовательностей и т. п. Кроме того, начались интенсивные исследования больших массивов данных измерений. В таких массивах бывает важно учитывать взаимное расположение выделенных значений. В процессе анализа подобных последовательностей исследователи пытаются выявить структуру цепи событий. При этом из-за неопределённости самого понятия структуры для цепей данных исследователи пытаются опереться на природу компонентов таких последовательностей (слова, нуклеотиды, триплеты, кодоны, аминокислоты, амплитуды сигналов, высоты звучания нот, и т. п.). Это направление породило большое число тонких формальных техник, применимых к цепям специфической природы.

[*]gumas45@mail.ru
[**]nick670@yandex.ru
[***]goruha@gmail.com
[****]stan63@inbox.ru

Общие подходы для исследования структуры знаковых цепей представлены двумя направлениями: это анализ структуры, основанный на элементном составе последовательностей, и оценке локального порядка следования событий в цепях.

В первом подходе суждение о структуре цепи осуществляется на основе статистического распределения (состава) её компонентов (двоек, троек и, в общем случае, *n*-ок). Так как компоненты в рамках вероятностной модели цепи представляют собой случайные события, в лучшем случае, возможно построение статистических распределений, частотно-ранговых распределений или *H*-статистик. В предельном случае, как это показано в докторской диссертации М.Г. Садовского [2], длина окна для короткого кортежа (*n*-ки) может быть такой величины, что при считывании конкретной нуклеотидной цепи *L*-граммами со сдвигом на один элемент мы получим некоторое конечное множество (алфавит или словарь) *L*-грамм (*n*-ок), на основе которого возможно однозначно восстановить взаимное расположение всех компонентов исходной цепи. Такое численное описание структуры достигается путём введения многократной избыточности за счёт $(n-1)$-кратного тиражирования цепи. Суждения же на основе обычного статистического распределения компонентов данной цепи, полученного экспериментально, не претендуют на возможность восстановления исходной последовательности. Такие распределения являются количественным описанием взаимного расположения элементов, так, как исследователю по умолчанию известно, что он взял не случайную выборку данных, а конкретный текст, нуклеотидную цепь и т. п. Это «проклятие априорного неосознаваемого знания» об упорядоченности цепи широко распространено в математической лингвистике, статистических исследованиях генетических текстов и т. п.

Другое направление исследований и анализа структуры массива данных, в основном, использует марковские цепи, потоки заявок и теорию очередей, взвешенные графы, с помощью которых удаётся, хоть и громоздко, описать локальную структуру знаковых цепей, но не конкретное взаимное расположение компонентов всей цепи. Кроме того, косвенный анализ структуры осуществляется путём сравнения пар цепей, одна из которых может быть эталонной. Для этого используются разные меры сходства (различия), среди которых широко используется несимметричная статистическая мера Кульбака-Лейблера, а также мера Левенштейна в форме «редакционного расстояния» [3, 4]. Для исследования структуры знаковых последовательностей широко применяется получение вероятностно-статистических и энтропийно-информационных характеристик и оценок. На протяжении десятилетий такие разработки выполняются в Институте математики им. С.Л. Соболева СО РАН [5-9].

Подходы, используемые для исследования знаковых цепей, текстов разной природы и массивов данных можно дополнить математическим, спектральным, статистическим, корреляционным, фрактальным и др. анализами. Однако в таких методиках не уделяется внимания исследованию и обнаружению закономерностей конкретного расположения знаков, слов, компонентов массивов данных, составляющих отдельную целостную последовательность. На наш взгляд, такое положение, в некоторой степени, объясняется отсутствием формализма для выделенного абстрактного объекта, называемого здесь «строем или построением цепи» [10]. Следует отметить, что разные по природе последовательности событий с одинаковыми статистическими распределениями (в дальнейшем – с равномощными составами) могут иметь один и тот же оригинальный строй. С другой стороны, очевидно, что множество, которое содержит повторяющиеся элементы (мультимножество), может быть основой для построения различных комбинаций типа «перестановки с повторениями». При этом многие из них будут иметь разное взаимное расположение компонентов. В данной работе рассматривается подход, который предназначен для формального анализа построения отдельного текста произвольной природы, любой знаковой

последовательности, в том числе представляющей нуклеотидную цепь, а также массивов данных, полученных путём измерений.

ОБ ИНФОРМАЦИОННОЙ СУЩНОСТИ НУКЛЕОТИДНОЙ ЦЕПИ

Общепринятое представление нуклеотидной цепи в форме генетического текста (а текст «очевидно» содержит информацию) является гипотезой, которая для своего утверждения требует дополнительных аргументов [11]. Насколько известно, в настоящее время нет общепринятых формальных определений понятиям «информация» и «текст».

В наших работах, в том числе в данной статье, нуклеотидные, триплетные и аминокислотные последовательности рассматриваются только с точки зрения теории информации, и моделируются как информационные цепи (в математике такие объекты называются упорядоченными множествами или кортежами). Таким образом, используется гипотеза о том, что нуклеотидные цепи содержат некоторую информацию (вероятно о материальной структуре и динамике развития отдельного организма или его частей). Принимая эту гипотезу можно сказать, что отдельная нуклеотидная цепь представима двумя информационными цепями, компоненты которых комплементарны. Дополнительное кодирование нуклеотидов одной цепи комплементарными нуклеотидами другой цепи используется для надёжной передачи и хранения генетической информации в двухцепочечной ДНК, а также – при делении клетки, транскрипции ДНК в РНК, при синтезе белков, кодируемых ДНК и в процессах репарации ДНК при её повреждении. С точки зрения теории информации феномен комплементарности представляет дополнительное подтверждение в пользу гипотезы об информационной природе нуклеотидных цепей. Ниже понятие «информация», в соответствии с теорией Мазура [12], не рассматривается как синоним понятия «смысл».

Если для нуклеотидной цепи в качестве элементарных сообщений различают 4 нуклеиновых кислоты ($m = 4$), то в информатике в качестве элементарных сообщений используют всего 2 различимых состояния ($m = 2$), из которых, в свою очередь, строятся двоичные кодовые слова и целые массивы данных. В результате развития, также как в естественных языках, кодовые слова, управляющие действиями машин (бывшие изначально одинаковой длины), оказались разной длины и в настоящее время составляют «словарь компьютерного языка». Подобно этому в генетических текстах допустимо предположить существование «нуклеотидных слов» разной длины, составляющих «словарь генетических текстов». Вероятно, нуклеотидные слова информационной цепи (ДНК) запускают механизмы формирования и функционирования организма.

Отмеченное выше позволяет в дальнейшем, вместо двух комплементарных цепей, рассматривать одну, которая отображена строем цепи сообщений (в частности – строем нуклеотидной цепи). Сформулированный здесь подход при исследовании нуклеотидных цепей, рассматриваемых как информационные цепи, не требует учёта пространственной структуры молекул цепи ДНК.

ВЫБОР ДАННЫХ ДЛЯ ИССЛЕДОВАНИЯ

16S рибосомальная РНК (16S рРНК) является важным маркером эволюционных событий. Поскольку этот панбактериальный ген присутствует у всех прокариот, определение его первичной структуры позволяет изучать степень гомологии микроорганизмов, строить филогенетические деревья и изучать их эволюционные связи.

Применение рибосомальной РНК (16S рРНК у прокариот и 18S рРНК у эукариот) для решения проблем систематики организмов стало возможным по нескольким причинам. Во-первых, эти молекулы обнаружены у всех клеточных форм жизни, что

указывает на их древнейшее происхождение. Во-вторых, их функции всегда одинаковы; первичная структура в целом характеризуется высокой консервативностью. В-третьих, генетический материал рРНК находится вне сферы действия отбора, поэтому данные молекулы эволюционируют в результате спонтанных мутаций, происходящих с постоянной скоростью, и накопление таких мутаций зависит только от времени. Таким образом, мерой эволюционного расстояния между организмами служит количество и локализация нуклеотидных замен в молекулах сравниваемых рРНК.

Изучение структуры генов 16S и 18S рРНК позволило провести разделение клеточных форм жизни на три царства – Archaea, Eukaria и Bacteria.

ФОРМАЛИЗМЫ СТРОЯ ЦЕПИ

Строй цепи сообщений (событий, знаков и т. п.) – это кортеж (упорядоченное множество), в котором каждому компоненту цепи поставлено в соответствие натуральное число, причем идентичные по выбранному признаку компоненты отображены одним и тем же числом. Первый компонент кортежа – единица, каждый следующий компонент цепи, отличный от всех предыдущих, обозначается натуральным числом, которое на единицу больше максимального из расположенных ранее в кортеже. При этом *алфавит строя* [10] дополняется этим числом; *мощность алфавита строя* – это количество различных компонентов в цепи.

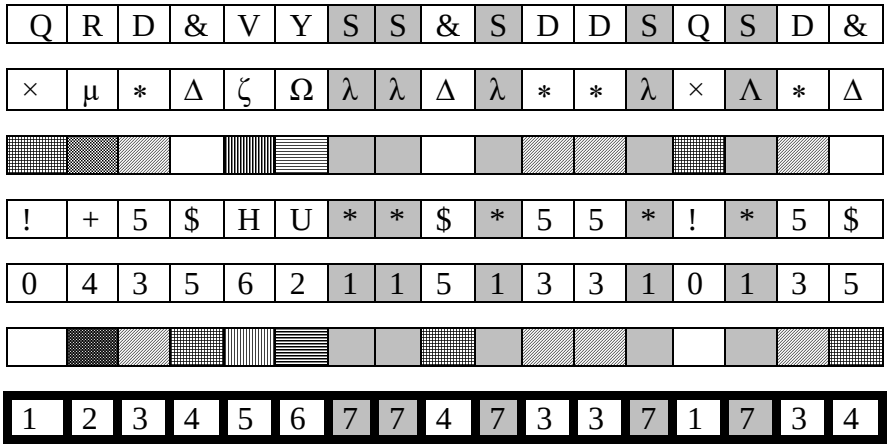

Рис. 1. Пример прямого преобразования 6-ти разных знаковых цепей, цифровых последовательностей и диаграмм в строй цепи с одинаковым порядком.

A	C	C	T	G	A	C	T	G	C	T	A	T	C	G	G	A	T	T	G	A	T	A	C
T	G	G	A	C	T	G	A	C	G	A	T	A	G	C	C	T	A	A	C	T	A	T	G
1	2	2	3	4	1	2	3	4	2	3	1	3	2	4	4	1	3	3	4	1	3	1	2

Рис. 2. Фрагмент двух комплементарных цепочек бактерии *Candidatus nitrosopumilus maritimus* (№ 26 в табл. 1) с одинаковым строем.

Строем назовём упорядоченное множество (кортеж), составленное из элементов алфавита строя с ограничениями, принятыми для строя цепи. В соответствии с определением для формирования строя необходимо учитывать следующие ограничения:

1. Алфавит строя – это множество всех натуральных чисел из диапазона от 1 до m. {1, 2, 3, 4, 5, …, m}

2. Мощность алфавита m всегда не больше длины строя $m <= n$ (предельный случай, когда длина строя равна размеру алфавита ($m = n$) и все элементы (числа) встречаются в строе один раз)

3. Расположение первых вхождений элементов алфавита в строе. Элементы алфавита располагаются в позициях строя по возрастанию, начиная с единицы в первой позиции, возможно, с пропусками некоторых мест.

<1 2 - 3 - - 4 - - - - 5 - - - - - - 6 7>

4. Не занятые первыми вхождениями элементов алфавита позиции строя заполняются натуральными числами, по значению не превышающими максимального натурального числа среди всех лежащих слева чисел.

<1 2 1 3 2 3 4 4 4 1 5 3 4 5 1 1 1 6 7 >

На рис. 1 и 2 приведены примеры разных последовательностей (кортежей) символов имеющих одинаковые строи. При сравнении по строю нескольких кортежей реальных сообщений, необходимо корректно выполнить однозначное *прямое преобразование* для каждого из них, а затем сравнить полученные строи [13].

T	**T**	**G**	**G**	**G**	**T**	**T**	**C**	**C**	**G**	**G**	**G**	**G**	**G**	**G**	*Cricetulus griseus*
G	**G**	**A**	**A**	**A**	**G**	**G**	**T**	**T**	**A**	**A**	**A**	**A**	**A**	**A**	*Homo sapiens*
1	**1**	**2**	**2**	**2**	**1**	**1**	**3**	**3**	**2**	**2**	**2**	**2**	**2**	**2**	строй, общий для обоих фрагментов

Рис. 3. Фрагменты нуклеотидных цепей *Cricetulus griseus* (№ 15 в табл. 1) и *Homo sapiens* (№ 12 в табл. 1) с одинаковым строем (длина фрагментов 15). Выделены путём просмотра рибосомальной РНК общей длиной 1871и 1559, Совпадение порядков строя фрагментов начинается с позиций: 1157 для первой цепочки и 778 для второй цепочки.

G	-	-	-	-	G	G	-	-	-	-	-	-	-	G	G	-	-	G
A	-	-	-	-	A	A	-	-	-	-	-	-	-	A	A	-	-	A
1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	1

Рис. 4. Однородные фрагменты нуклеотидных цепей *Cricetulus griseus* (№ 15 в табл. 1) и *Homo sapiens* (№ 12 в табл. 1) с одинаковым строем (длина фрагментов 19). Совпадение порядков строя начинается с позиций: 108 для первой цепочки и 1847 для второй цепочки.

Заметим, что при несоблюдении ограничений на порядок расположения натуральных чисел мы получим кортеж, который не представляет собой строй. Для примера ниже представлен такой кортеж.

1	2	3	5	4	6	7	7	5	7	3	3	7	1	7	3	5	Вектор натуральных чисел, который не является строем цепи

Разложим полную неоднородную символьную последовательность на m неполных однородных кортежей, у которых одинаковыми знаками заняты только некоторые позиции (рис. 5). Такое разложение цепи называют *декомпозицией*. Аналогом однородной последовательности является поток однородных заявок (событий), определенный в теории массового обслуживания. Очевидно, что композиция всех однородных строев данного полного строя даёт полный неоднородный строй, аналогом которого в теории очередей является поток разнородных событий [14]. Вообще разложение цепи может осуществляться по разным правилам. Ниже представлена декомпозиция строя неоднородной знаковой цепи на неполные однородные (рис. 5) и неполные разнородные (рис. 6) цепи. В последнем случае те позиции, которые заняты разными знаками, заполняются по следующему правилу: при просмотре цепи от ее начала в состав первой разнородной цепи выбираются все первые вхождения каждого элемента алфавита, при втором – все вторые вхождения и т. д.

Определим *интервал* как расстояние от выделенного в цепи компонента до ближайшего, отмеченного в направлении просмотра (рис. 5); величина интервала – это натуральное число, определённое, как модуль разности номеров позиций двух выделенных компонентов кортежа.

Заметим, что в теории массового обслуживания в потоке заявок интервал (времени) между событиями является случайной величиной.

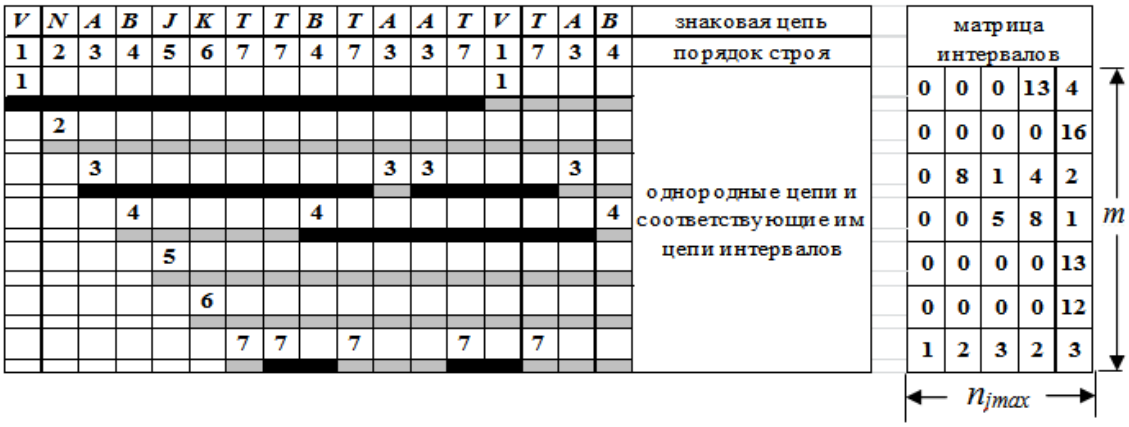

Рис. 5. Декомпозиция строя неоднородной знаковой цепи на неполные однородные цепи и матрица их интервалов.

Пусть считывание текста осуществляется по разнородным цепям. В результате получим *матрицу интервалов разнородных цепей*, в которой число столбцов равно m, а число строк – $n_{j\,max}$.

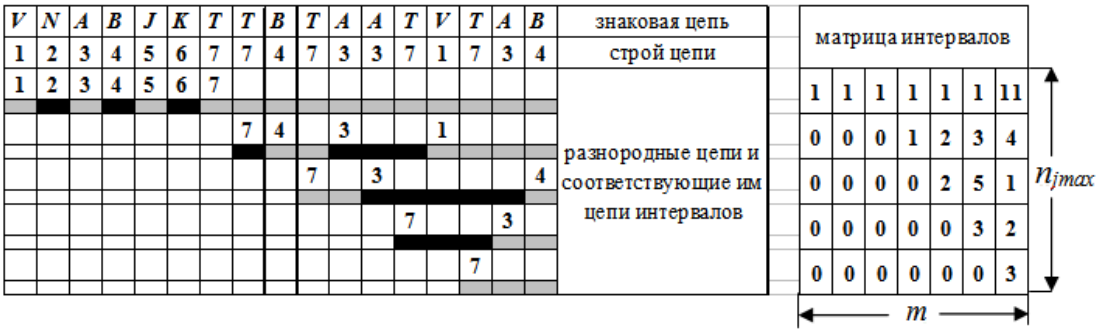

Рис. 6. Декомпозиция строя неоднородной знаковой цепи на неполные разнородные цепи и матрица их интервалов.

ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СТРОЯ

Используем понятие однородной знаковой последовательности и векторное отображение её порядка строя в виде соответствующей строки матрицы интервалов для определения числовых характеристик строя текста.

Перемножением всех интервалов выделенной j-ой однородной последовательности (элементов соответствующей ей строки матрицы, кроме нулевых) определим *абсолютный объем строя j-ой однородной цепи* в виде

$$V_j = \prod_{i=1}^{n_j} \Delta_{ij} \,, \qquad (1)$$

где Δ_{ij} – интервал от i-го до $(i + 1)$-го вхождения j-го символа, n_j – число вхождений j-го символа.

Средний геометрический интервал между занятыми местами на позиции строя однородной цепи определяется в виде:

$$\Delta_{gj} = \sqrt[n_j]{V_j}\,, \tag{2}$$

абсолютный объем строя текста – как произведение абсолютных объемов всех строев *j*-ых однородных последовательностей:

$$V = \prod_{j=1}^{m} V_j\,, \tag{3}$$

при подстановке (1) в (3) получим

$$V = \prod_{j=1}^{m}\prod_{i=1}^{n_j} \Delta_{ij}, \tag{4}$$

где m – мощность алфавита (собственного словаря текста).

Средний геометрический интервал строя цепи на множестве всех однородных цепей текста определяется в виде:

$$\Delta_g = \sqrt[n]{V}, \tag{5}$$

где n – длина текста, равная числу мест для размещения всех слов (знаков) на его позиции.

Средний арифметический интервал строя *j*-ой однородной цепи определяется в виде:

$$\Delta_{aj} = \frac{n}{n_j} = \frac{1}{P_j}, \tag{6}$$

где P_j – частота вхождения или статистическая вероятность вхождения *j*-го элемента в цепи.

Периодичность (следования одинаковых элементов) строя *j*-ой однородной цепи τ_j определим из (2) и (6) отношением среднего геометрического и среднего арифметического интервалов в виде $\tau_j = \dfrac{\Delta_{gj}}{\Delta_{aj}}$.

Регулярность (следования одинаковых элементов) строя полной неоднородной цепи определим из (5) и формулы для числа описательных информации D (по М. Мазуру [12]) отношением вида: $r = \dfrac{\Delta_g}{D}$, где $\Delta_g \le D$, как будет показано ниже в (15).

Логарифмирование представленных величин дает набор удобных для практики компактных аддитивных информационных характеристик строя цепи. При этом интервал соответствует *удаленности определенного j-го символа* *i*-го вхождения относительно его (i + 1)-го вхождения в виде $g_{ij} = \log_2 \Delta_{ij}$, объем строя однородной последовательности выделенного *j*-го символа – *глубине расположения строя j-ой однородной цепи* в виде: $G_j = \log_2 V_j$, при подстановке (1) получим

$$G_j = \sum_{i=1}^{n_j} \log_2 \Delta_{ij}. \tag{7}$$

Объем строя текста соответствует *глубине расположения строя всей цепи* в виде:

$$G = \log_2 V, \tag{8}$$

где при подстановке (4) в (8) получим

$$G = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta_{ij}. \quad (9)$$

При сравнении построений разных знаковых цепей и текстов могут быть полезны оценки относительных глубин расположения $\delta G_j = G_j / G,$ а также оценки средних удаленностей соседних одинаковых элементов в строе *j*-ой однородной цепи, которые отображают средние геометрические интервалы. Из выражений (2) и (7) *средняя удаленность выделенного j-го символа* в строе однородной последовательности определяется в виде:

$$g_j = \log_2 \Delta_{gj} = \frac{1}{n_j} \sum_{i=1}^{n_j} \log_2 \Delta_{ij}, \quad (10)$$

из выражений (5) и (9) *средняя удаленность* любого символа в строе данной цепи или текста определяется в виде:

$$g = \log_2 \Delta_g = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta_{ij} = \sum_{j=1}^{m} \frac{n_j}{n} \log_2 \Delta_{gj}, \quad (11)$$

Отношение средней удаленности выделенного *j*-го символа к средней удаленности отдельного символа $\delta g_j = g_j / g,$ дает характеристику строя, которая дополняет частоту вхождения $P_j = n_j / n$ и названа *относительной удаленностью j-го символа*. Из (7), (10) и (11) следует, что $G_j = n_j \cdot g_j,$ а из (9) и (11) следует, что $G = n \cdot g.$

Очевидно использование числовых характеристик строя, аналогичных по форме моментам разного порядка случайной величины [14]. В табл. 1 приведены различные числовые характеристики строя рибосомальных РНК нескольких организмов, взятых из GenBank [15]. В таблице количество информации обозначено *Hs* и вычисляется в предположении статистической независимости элементов путем перемножения *H* и *n*; не отмеченная в таблице мощность алфавита (*m*) для всех нуклеотидных цепей равна 4.

Выборка нуклеотидных последовательностей организмов была сформирована произвольным образом из представителей трёх царств – Archaea, Eukaria и Bacteria [16], представляющих клеточную форму жизни. Нуклеотидные последовательности были получены из GenBank [15]. В представленную выборку вошли нуклеотидные последовательности 16S рибосомальной РНК прокариот и архей, 18S рибосомальная РНК и митохондриальная рибосомальная РНК эукариот.

На наш взгляд близкие по значениям характеристики удалённости и регулярности демонстрируют схожую основу построения цепочек РНК на уровне нуклеотидов. Данный уровень в нашей работе полагается элементарным. Кроме того, как будет показано ниже, эти характеристики практически инвариантны к размеру исследуемого фрагмента и его местоположению в геноме. Все исследованные и представленные здесь организмы различимы на основе данных характеристик, значения которых отличаются уже во втором-третьем знаках после запятой.

Таблица 1. Числовые характеристики фрагментов нуклеотидных последовательностей

№	название организма/название последовательности/идентификатор Genbank/описание фрагмента/координаты фрагмента	*n*	*Hs*	*G*	*H*	***g***	*r*
1	*Mus musculus domesticus* (мышь домовая)/митохондриальный геном/FJ374665.1/ген 16S рРНК/1094-2675	1582	3037	2241	1,920	**1,4165**	0,705
2	*Caiman crocodilus* (кайман крокодиловый)/митохондриальный геном/AJ404872.2/ген 16S рРНК/1058-2649	1592	3098	2260	1,946	**1,4197**	0,694
3	*Canis lupus familiaris* (собака)/митохондриальный геном/ EU789780.1/ген 16S рРНК/1091-2670	1580	3069	2256	1,943	**1,4276**	0,699
4	*Gallus gallus* (курица)/митохондриальный геном /GU261702.1/ген 16S рРНК/2351-3970	1620	3171	2315	1,958	**1,4292**	0,693
5	*Amia calva* (ильная рыба)/митохондриальный геном/AY442347.1/ген 16S рРНК/1162-2869	1708	3339	2459	1,955	**1,4397**	0,699
6	*Homo sapiens* (человек разумный)/митохондриальный геном/HQ384215.1/ген 16S рРНК/1672-3230	1559	3043	2246	1,952	**1,4409**	0,702
7	*Sus scrofa* (кабан)/ген 18S рРНК/AY265350	2302	4532	3319	1,969	**1,4418**	0,694
8	*Thermotoga thermarum*/частичная последовательность 16S рРНК/NR_024751.1	1471	2847	2131	1,936	**1,4489**	0,714
9	*Thermus thermophilus*/частичная последовательность 16S рРНК/AY788091.1	1517	2924	2199	1,928	**1,4493**	0,717
10	*Gallus gallus* (курица)/ген 18S рРНК/DQ018752.1	1851	3677	2688	1,987	**1,4523**	0,690
11	*Bos taurus* (бык)/частичная последовательность рРНК/DQ222453.1/ген 18S рРНК/1-1873	1873	3723	2735	1,988	**1,4600**	0,693
12	*Homo sapiens* (человек)/ген 18S рРНК/NR_003286.2	1869	3715	2735	1,988	**1,4631**	0,695
13	*Erinaceus europaeus* (ёж обыкновенный)/частичная последовательность 18S рРНК /AJ311675.1	1825	3628	2671	1,988	**1,4636**	0,695
14	*Mus musculus domesticus* (мышь домовая)/частичная последовательность рРНК/BK000964.1/ген 18S рРНК/4008-5877	1870	3717	2739	1,988	**1,4645**	0,696
15	*Cricetulus griseus* (хомячок китайский)/частичная последовательность рРНК/DQ235090.1/ген 18S рРНК/11629-13499	1871	3721	2746	1,989	**1,4676**	0,697
16	*Rattus norvegicus* (серая крыса)/ген 18S рРНК/NR_046237.1	1874	3729	2754	1,990	**1,4697**	0,697
17	*Crocodylus niloticus* (крокодил Нильский)/частичная последовательность 18S рРНК/AJ311672.1	1776	3539	2626	1,993	**1,4784**	0,700
18	*Ixodes persulcatus* (клещ таёжный)/частичная последовательность 18S рРНК/AY274888.1	1771	3536	2628	1,997	**1,4839**	0,701
19	*Zebrias zebra* (рыба)/ген 18S рРНК/EF126044.1	1837	3664	2738	1,995	**1,4906**	0,705
20	*Kareius bicoloratus* (камбала двухцветная)/ген 18S рРНК/EU637036.1	1823	3638	2723	1,996	**1,4935**	0,706
21	*Ornithodoros moubata* (клещ)/ген 18S рРНК/L76355.1	1790	3576	2677	1,998	**1,4956**	0,706
22	*Pediculus humanus capitis* (вошь головная)/ген 18S рРНК/AY236410.1	1493	2975	2242	1,993	**1,5017**	0,712
23	*Musca domestica* (муха домашняя)/частичная последовательность 18S рРНК/DQ133074.1	1728	3430	2611	1,985	**1,5108**	0,720
24	*Streptococcus pyogenes* (стрептококк пиогенный)/полный геном/NC_002737.1/ген 16S рРНК/23068-24617	1335	2651	2025	1,986	**1,5165**	0,722
25	*Borrelia burgdorferi* (возбудитель болезни Лайма)/частичная последовательность 16S рРНК/U03396.1/ген 16S рРНК/841-2377	1537	3040	2337	1,978	**1,5202**	0,728
26	*Candidatus Nitrosopumilus maritimus* (бактерия хемолитотроф)/частичная последовательность 16S рРНК/DQ085097.1	1452	2886	2210	1,988	**1,5220**	0,724
27	*Bacillus anthracis* (возбудитель сибирской язвы)/частичная последовательность 16S рРНК/AF155950.1	1506	2987	2293	1,984	**1,5228**	0,726
28	*Mycoplasma pneumoniae* (возбудитель атипичной пневмонии)/ген 16S рРНК/NC_016807.1/118315-119827	1487	2947	2277	1,982	**1,5313**	0,732
29	*Neisseria gonorrhoeae* (возбудитель гонореи)/ген 16S рРНК/GU395612.1	1516	3000	2325	1,979	**1,5336**	0,734

ПРЕДЕЛЬНЫЕ ЗНАЧЕНИЯ ХАРАКТЕРИСТИК СТРОЯ

Для *регулярной знаковой цепи*, в которой все интервалы каждой однородной цепи равны $\Delta_{ij} = \Delta_{aj} = n/n_j = const$, числовые характеристики строя цепи (5) и (11) принимают максимальные значения: $\Delta_g = \Delta_{g\,max} = D$, $g = g_{max} = J$, записываются формулами Мазура (12) и (13), представляющими соответственно числа описательных и идентифицирующих информаций сообщения в информационной цепи в виде [12]

$$D = \prod_{j=1}^{m} (n / n_j)^{\frac{n_j}{n}}, \tag{12}$$

$$J = \sum_{j}^{m} \frac{n_j}{n} \log \frac{n}{n_j}. \tag{13}$$

Для бесконечной знаковой цепи ($n \to \infty$) формула Мазура (13), в которой $\Delta_{aj} = (n/n_j) \to (1/P_j)$, принимает вид формулы К. Шеннона (14) для энтропии или количества информации. Такая информация используется только для дихотомической идентификации (но не для описания) отдельных сообщений.

$$J = H = -\sum_{j=1}^{m} P_j \log P_j. \tag{14},$$

Для текстов и других нерегулярных последовательностей формулы Мазура и Шеннона дают оценку строя только «сверху», так как в этих случаях

$$D > \Delta_g, \quad H > g = \log \Delta_g, \tag{15), (16}$$

Соответственно, числовые характеристики строя на основе однородных цепей принимают минимальные значения для *сплошных последовательностей*, в которых все одинаковые элементы расположены подряд.

Таким образом, предлагаемые формулы для среднего геометрического интервала и средней удаленности знаковой цепи обобщают формулы Мазура и Шеннона, так как при описании строя данной цепи учитывают не только мощность состава, но и взаимное расположение ее компонентов.

Важно отметить взаимосвязь характеристик строя цепи, полученных на основе однородных и разнородных цепей. Так, для рассмотренных выше моделей регулярной и сплошной последовательностей, характеристики строя на основе разнородных цепей принимают соответственно минимальное и максимальное значения.

РАСПРЕДЕЛЕНИЯ ЧИСЛОВЫХ ХАРАКТЕРИСТИК ОДНОРОДНЫХ ЦЕПЕЙ

Для более полного описания строя цепи следует использовать ранговые распределения однородных цепей, а также их распределения $\{<n_j , \Delta_{gj}>\}$, $\{<n_j , G_j>\}$, $\{<n_j , g_j>\}$, где $j = 1,2, \ldots, m$. Такие распределения, в отличие от статистических распределений, связывают мощность состава с взаимным расположением компонентов цепи.

На рис. 7-10 приведены ранговые распределения рибосомальных РНК нескольких организмов разбитых на слова (описание разбиения приводится ниже). Из сравнения ранговых распределений на рис. 7 и 8 видно, что глубинно-ранговое распределение даёт более однозначное отображение строя цепи. Интегральные и усреднённые характеристики строя (см. табл. 2) также показывают заметную чувствительность к перестановкам слов в цепи, в то время как статистические распределения и энтропия не изменяются.

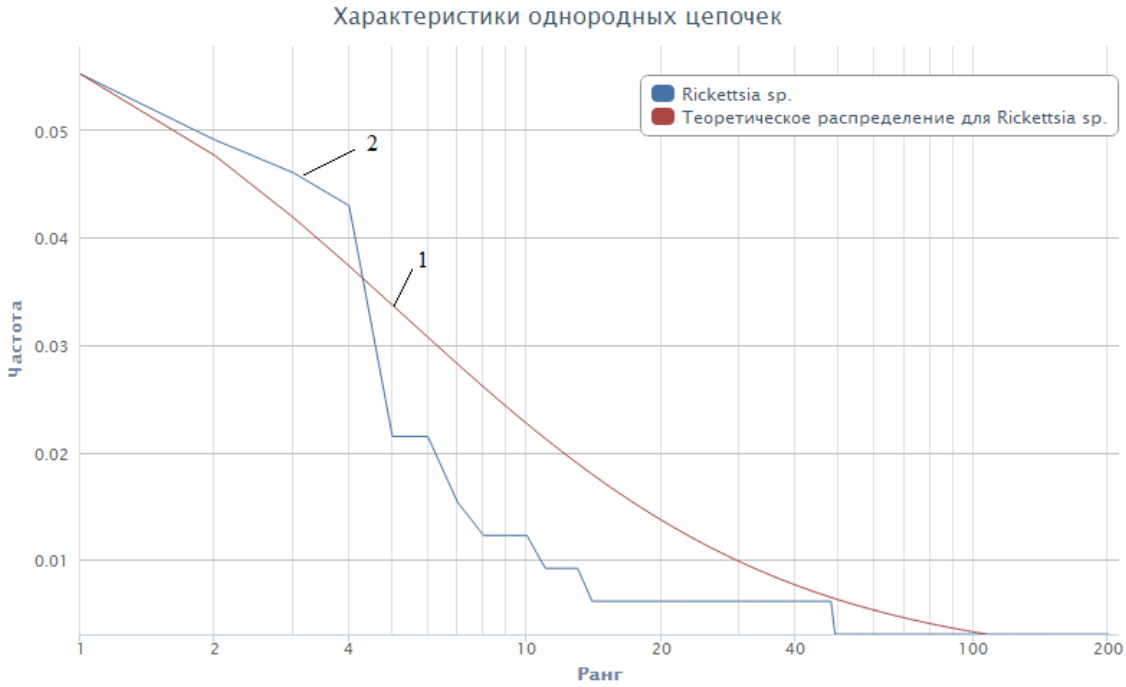

Рис. 7. Частотно-ранговые распределения слов Rickettsia sp./ген 16S рРНК/ L28944: 1 – теоретическое частотно-ранговое распределение слов (в соответствии с законом Ципфа-Мандельброта (17)); 2 – фактическое частотно-ранговое распределение слов.

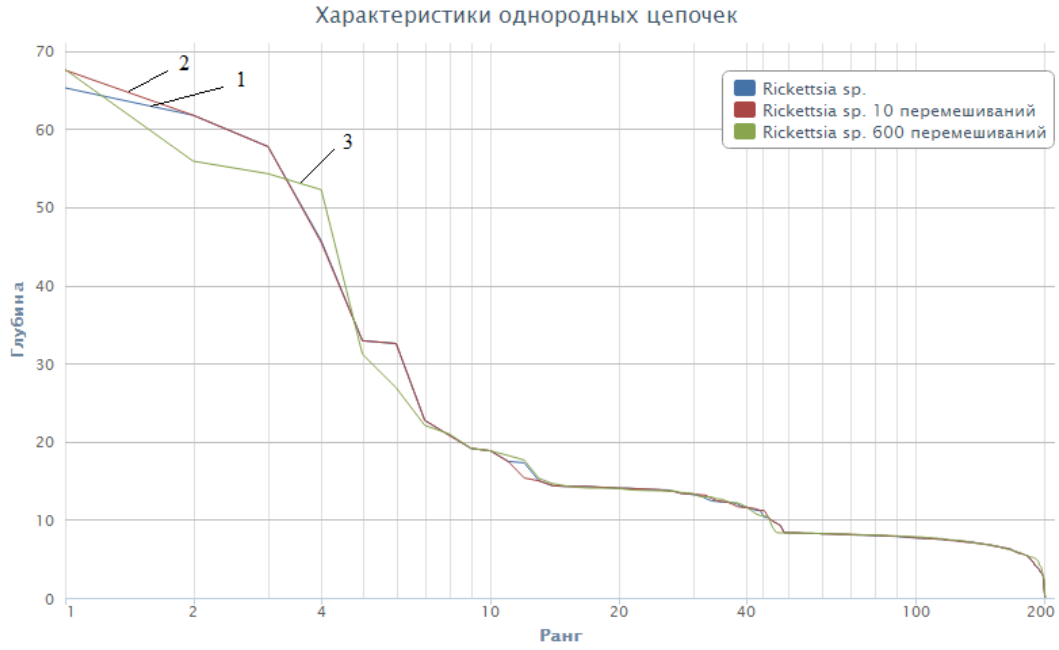

Рис. 8. Глубинно-ранговые распределения однородных цепей слов Rickettsia sp./ген 16S рРНК/ L28944 и её перемешанных вариантов: 1 – ранговое распределение глубин строя однородных цепей слов Rickettsia sp./ген 16S рРНК/ L28944; 2 – ранговое распределение глубин для модификации строя с небольшими изменениями (со случайными перестановками нескольких слов); 3 – ранговое распределение глубин для модификации строя с большими изменениями (со случайными перестановками всех слов).

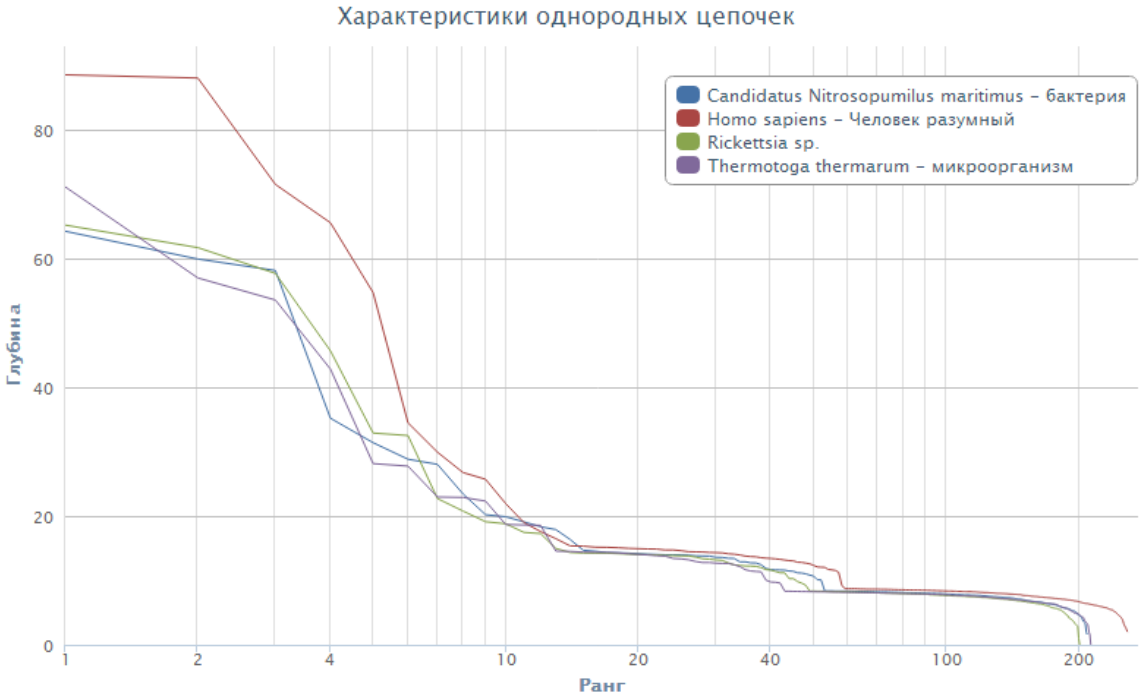

Рис. 9. Глубинно-ранговые распределения однородных цепей слов рибосомальных РНК отдельных организмов.

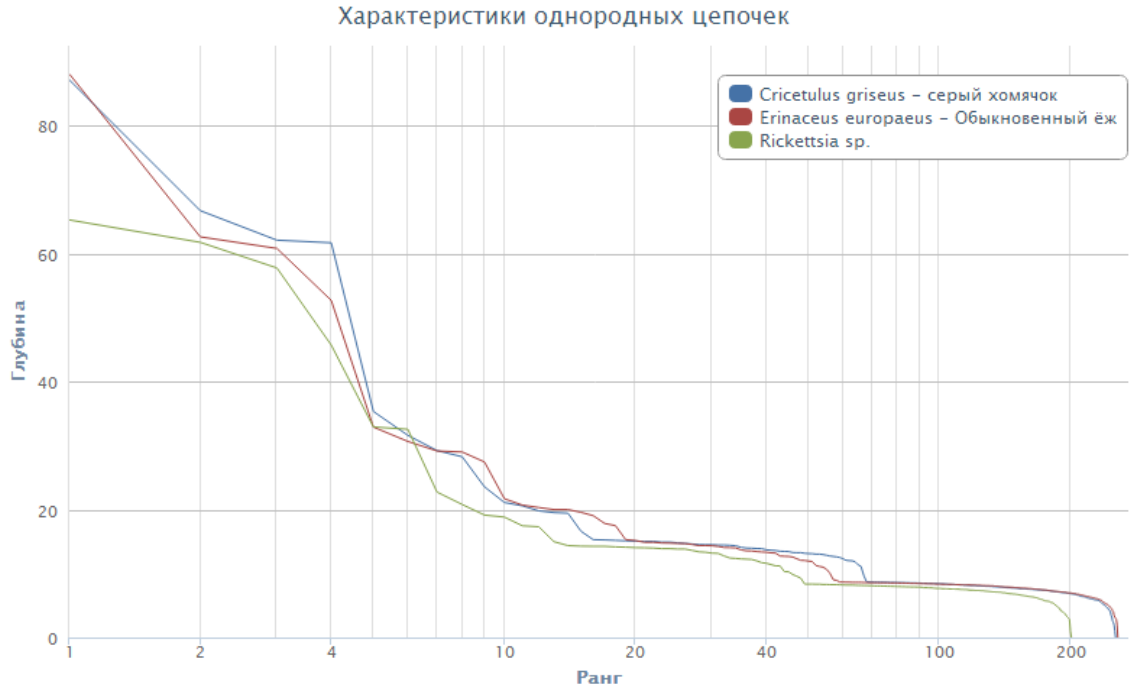

Рис. 10. Глубинно-ранговые распределения однородных цепей слов рибосомальных РНК отдельных организмов.

Таблица 2. Интегральные характеристики исходной цепочки и деформированных цепочек

Название	G	Hs	g	H	r
Rickettsia sp.	2410,70	2907,15	5,87976	7,09062	0,43201
Rickettsia sp. 10 перемешиваний	2413,50	2907,15	5,88658	7,09062	0,43406
Rickettsia sp. 600 перемешиваний	2418,25	2907,15	5,89818	7,09062	0,43756

У всех представленных в табл. 2 цепочек мощность алфавита $m = 254$, а длина $n = 410$.

На рис. 11-13 приведены частотные распределения разбитой на слова последовательности (Rickettsia sp./ген 16S рРНК/L28944).

Рис. 11. Распределение характеристик однородных цепочек по частоте и среднегеометрическому интервалу.

Рис. 12. Распределение характеристик однородных цепочек по частоте и удалённости.

Рис. 13. Распределение характеристик однородных цепочек по частоте и глубине.

Распределения частот, представленные на рис. 11–13 внешне похожи на статистическое распределение дискретной случайной величины, перечень значений которой заранее известен. Однако, размещённые по оси абсцисс величины (на рис. 12 – удалённость, на рис. 13 – глубина) не являются возможными значениями дискретной случайной величины; они представляют определённые значения характеристик расположения элементов однородных цепей, составляющих полную рибосомальную РНК бактерии Rickettsia sp.

Представленная здесь возможность однозначного отображения любой знаковой цепи (массива данных) строем цепи, его порядком, множеством однородных цепей, наконец, упорядоченным множеством интервалов и числовыми характеристиками строя, делает возможным использование разнообразных методов анализа функций и числовых последовательностей.

ХАРАКТЕРИСТИКИ ЗАВИСИМОСТЕЙ ОДНОРОДНЫХ ЦЕПЕЙ

В работе [17] на основе наглядных графических образов (диаграмм) представлены, введенные в [18], модифицированные Е.В. Морозенко, авторами данной работы и А.В. Верховиным [19], характеристики зависимости пары однородных цепей (j-ой и l-ой), взятых из состава данной неоднородной цепи. Это позволило соотнести субъективное суждение о степени «пространственной зависимости» элементов сопоставляемых цепей с формальной числовой характеристикой, представляющей такую зависимость.

Для наглядности в первой строке на рис. 14 представлен строй полной неоднородной цепи, в составе которого выделены две однородные j-ая и l-ая (в строках 2 и 3). На основе этих цепей составлена бинарно-однородная (далее бинарная) цепь (строка 4), в состав которой включены только те пары разных знаков j и l сравниваемых однородных цепей, которые определяются знаками l, «ближайшими справа» относительно знаков j. Предполагается, что наличие бинарной цепи свидетельствует о такой «причинно-следственной» связи цепей, при которой хотя бы некоторые компоненты (следствия) l-ой цепи расположены вслед за компонентами (причинами) j-ой цепи на определенных интервалах.

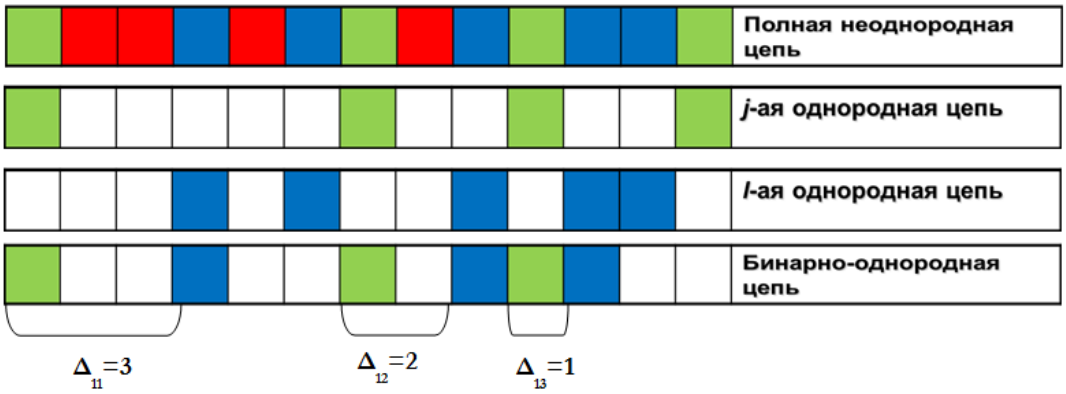

Рис. 14. Пример формирования бинарно-однородной цепи.

Связь такого типа названа *правосторонней пространственной зависимостью* событий одной *l*-ой цепи от другой *j*-ой цепи и обозначаемой (l/j). Соответственно левостороннюю пространственную зависимость обозначим (j/l).

Обозначим: $\Delta(l/j)_i$ – интервал между *i*-ми ближайшими справа знаками l по отношению к знакам *j;* $n(l/j)$ – число пар знаков *j* и *l*, связанных интервалами $\Delta(l/j)_i$. На рис. 14: $\Delta(l/j)_1 = 3;\ \Delta(l/j)_2 = 2;\ \Delta(l/j)_3 = 1$. Среднее геометрическое значение всех $n(l/j)$ установленных интервалов (бинарной цепи) $\Delta(l/j)_i$ между «смежными» элементами двух однородных цепей определено в виде $\Delta(l/j)_{cp} = \sqrt[n(l/j)]{\prod_{i=1}^{n(l/j)} \Delta\ (l/j)_i}$.

Интервалы только между теми знаками выделенной однородной цепи *l*, которые являются «ближайшими справа» относительно знаков цепи *j*, обозначены $\Delta(l_j)_i$. В примере: $\Delta(l_j)_1 = 5; \Delta(l_j)_2 = 2; \Delta(l_j)_3 = 3$. Среднее геометрическое значение интервалов $\Delta(l_j)_i$ между выделенными элементами l_j в данной однородной цепи определено в виде $\Delta(l_j)_{cp} = \sqrt[n(l/j)]{\prod_{i=1}^{n(l/j)} \Delta(l_j)_i}$.

При условии $\Delta(l_j)_{cp} > \Delta(l/j)_{cp}$, разность вида $\upsilon(l_j) = \left(1 - \frac{\Delta(l/j)_{cp}}{\Delta(l_j)_{cp}}\right)$, представляет *избыточность l-ой цепи, зависимой от j-ой*. Некоторые элементы такой цепи «связаны справа» с элементами *j*-ой цепи. В противном случае, при $\Delta(l_j)_{cp} \le \Delta(l/j)_{cp}$, данная разность свидетельствует об отсутствии избыточности. На основе отмеченного в [17] введен *коэффициент частичной зависимости* *l*-ой однородной цепи (от *j*-ой цепи) в виде $K_1(l/j) = (n\ (l/j)/n_l) \cdot \upsilon\ (l_j)$. Отметим, что соотношение $n(l/j)/n_l$ представляет собой условную вероятность события, состоящего в появления пары знаков l и j, связанных пространственной зависимостью, от появления знаков цепи l. В частном случае, когда $n(l/j) = n_l$, имеем полную зависимость этих цепей.

Предельный случай зависимости назван *непосредственной пространственной зависимостью*; при этом бинарная цепь задается единичными интервалами $\Delta\ (l/j)_i = 1 \quad \forall\, i = 1, 2, \ldots, n\ (l/j)$. Степень зависимости одной цепи от другой, с учетом «полноты её участия» в составе обеих однородных цепей определена в виде $K_2(l/j) = \frac{2n(l/j)}{n_j + n_l} \cdot \upsilon(l_j)$.

Если не требуется отдельно учитывать правостороннюю или левостороннюю зависимость пары однородных цепей, то вычисляется (средний) *коэффициент взаимной зависимости* в виде $K_3(j,l) = \sqrt{K_2(l/j) \cdot K_2(j/l)}$.

В случае если подкоренное выражение отрицательно (один из $K_2 < 0$) цепи считаются взаимно независимыми и коэффициент K_3 искусственно приравнивается к 0.

Удобно использовать нормированный коэффициент частичной зависимости, учитывающий частоту встреч тех или иных пар – $n(l/j)/(n/2)$, определяемый в виде $K_{1n}(l/j) = (2n^2(l/j)/n_l \cdot n) \cdot \upsilon(l_j)$. Полученные числовые характеристики зависимостей позволяют легче отделить редкие пары зависимых компонентов цепи от более частых.

На основе представленных выше коэффициентов зависимостей строятся *матрицы зависимостей* всех пар однородных цепей данной полной цепи. Такие матрицы в некотором смысле аналогичны корреляционным матрицам для систем случайных величин.

Таблица 3. Полная матрица межнуклеотидных зависимостей *Thermotoga thermarum* (№ 8 в табл. 1)

j* \ *l	**G**	**T**	**A**	**C**
G	0	**0,1294**	0,1096	0,102
T	0,0553	0	0,0642	0,0483
A	0,0646	0,082	0	0,075
C	0,0797	0,0895	0,0885	0

Ниже отмечены наиболее зависимые нуклеотиды в данной цепи:

GGTGAACGCTGGCGGCGTGCCTAACACATGCAAGTCGAGCGGGTACCCG
CAAGGGTACCAGCGGCGAACGGGTGAGTAACGCGTGGGTAACCTACCCCTC
AGAGGGGGATAACCAGGGGAAACCCTGGCTAATACCCCATACGATCCAGTGA
CGAAGGTCACTGGATGAAAGGGGCAACTGCCCCGCTGAGGGATGGGCCCGC
GTCCCATCAGGTAGTTGGTGGGGTAACGGCCCACCAAGCCTACGACGGGTAG
CCGGCCTGAGAGGGTGGTCGGCCACAGGGGCACTGAGACACGGGCCCCACT
CCTACGGGAGGCAGCAGTGGGGAATCTTGGACAATGGGCGAAAGCCTGATC
CAGCGACGCCGCGTGGGGGATGAAGCCCTTCGGGGTGTAAACCCCTGTTGC
GAGGGACGAATAAGGTGCGGAGTGGAATGCCGCACCGATGACGGTACCTCG
CGAGAAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGG
GCGAGCGTTACCCGGATTTACTGGGCGTAAAGGGTGCGTAGGCGGCCTGGTA
AGTCGGGTGTGAAATCCCACGGCTCAACCGTGGAATTGCGCCCGAAACTGCC
AGGCTTGGGGACGGTAGAGGGAGACGGAACTGCCGGTGTAGGGGTGAAATC
CGTAGATATCGGCAGGAACGCCGGTGGGGAAGCCGGTCTCCTGGGCCGCTC
CCGACGCTGAGGCACGAAAGCTAGGGGAGCAAAC…

Таблица 4. Фрагмент матрицы межтриплетных зависимостей *Thermotoga thermarum* (№ 8 в табл. 1)

	GCC	**CCG**	**GAC**	**AGG**	**CAC**
GGG	**0,0296**	0,0273	0,026	0,0272	0,0185
GGT	0,0202	0,0157	0,0155	0,0166	0,025

Ниже отмечены наиболее зависимые триплеты в данной цепи:

… GGG GCA ACT ***GCC*** CCG CTG AGG GAT GGG CCC GCG TCC CAT CAG
GTA GTT GGT GGG GTA ACG ***GCC*** CAC CAA GCC TAC GAC GGG TAG CCG
GCC TGA GAG GGT GGT CGG CCA CAG GGG CAC TGA GAC ACG GGC CCC
ACT CCT ACG GGA GGC AGC AGT GGG GAA TCT TGG ACA ATG GGC GAA
AGC CTG ATC CAG CGA CGC CGC GTG GGG GAT GAA ***GCC*** … GGG ACG
GTA GAG GGA GAC GGA ACT ***GCC*** GGT GTA GGG GTG AAA TCC GTA GAT
ATC GGC AGG AAC ***GCC*** GGT GGG GAA ***GCC*** … GGG GAG TAC ***GCC*** … GGG
TTA AGT CCC GCA ACG AGC GCA ACC CCT ***GCC*** CCT AGT TGC CAG CGG
TTC GGC CGG GCA CTC TAG GGG GAC TGC CGG CGA CGA ***GCC*** GGA GGA
AGG AGG GGA CGA CGT CAG GTA CTC GTG CCC CTT ATG CCC TGG GCT
ACA CAC GCG CTA CAA TGG GTG GTA CAG TGG GTT GCG ATC CCG CGA
GGG GGA GCT AAT CCC TAA AAC CAC CCC CAG TTC GGA TCG CAG GCT
GCA ACC CGC CTG CGT GAA ***GCC*** GGA ATC GCT AGT AAT CGC GGA TCA
GCC ACG CCG CGG TGA ATA CGT TCC CGG GGT TTG CAC ACA CCG CCC
GTC AAG CCA CCC GAG CTG GGG GCA CCT GAA GAC ***GCC*** …

АНАЛИЗ ЛОКАЛЬНОЙ СТРУКТУРЫ ГЕНЕТИЧЕСКИХ ТЕКСТОВ

Очевидно использование рассмотренных выше характеристик строя (описывающих целую генетическую последовательность) также и для анализа локальной структуры такой цепи, путем просмотра её окном разного размера в форме блоков или *L*-грамм (см. рис. 15, 16), с последующим применением разнообразных методов анализа функций и числовых последовательностей.

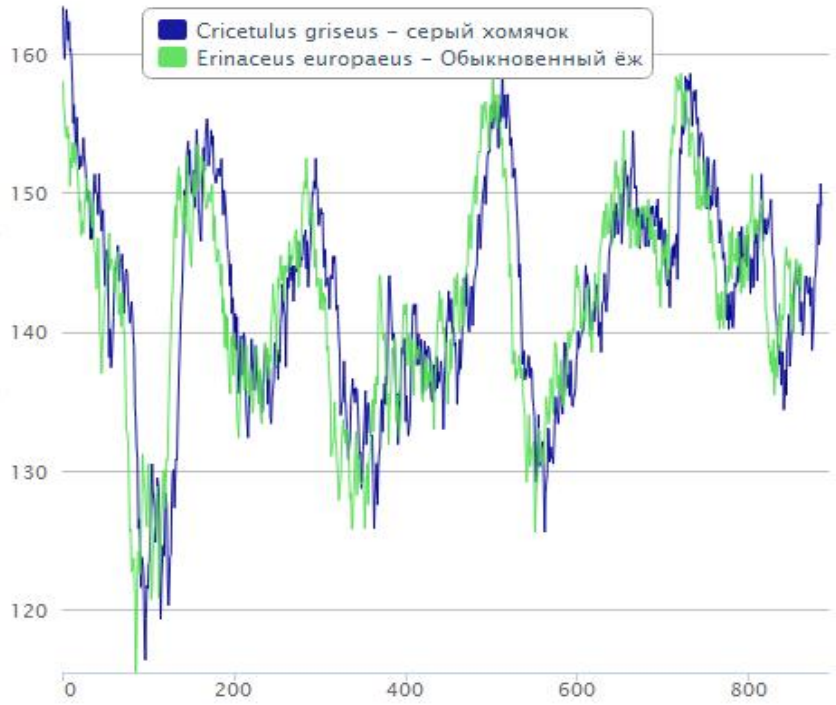

Рис. 15. Изменение глубины рибосомальных РНК двух близкородственных организмов в рамках *L*-граммы размером 50 нуклеотидов на протяжении данных цепей.

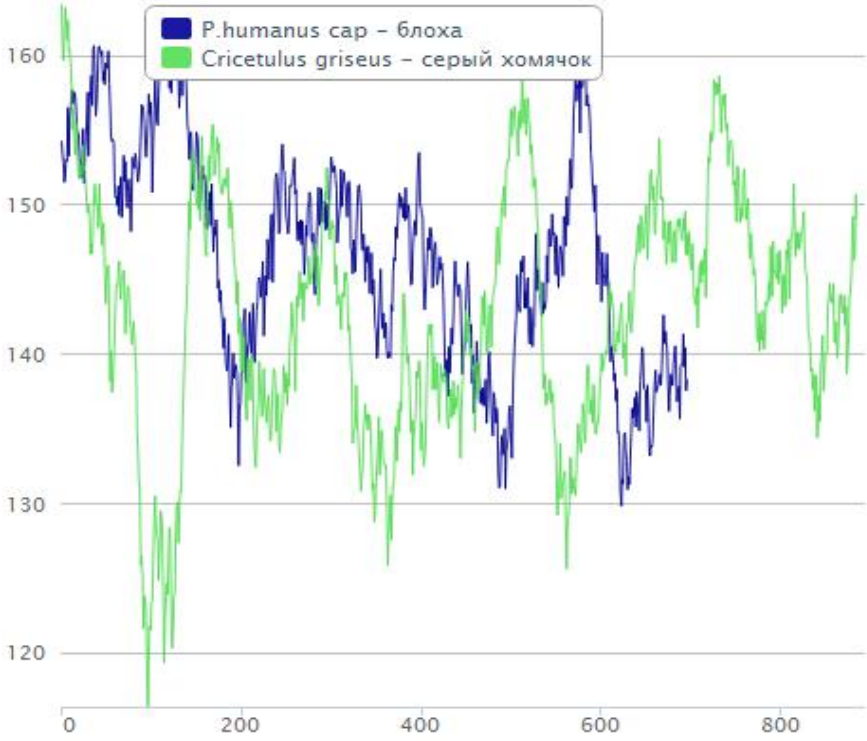

Рис. 16. Изменение глубины рибосомальных РНК двух разных организмов в рамках *L*-граммы размером 50 нуклеотидов на протяжении данных цепей.

О СЕГМЕНТАЦИИ ГЕНЕТИЧЕСКИХ ЦЕПЕЙ НА СЛОВА

Известно, что любой анализ (в том числе – неформальный, экспертный) становится практически невозможным и весьма субъективным, когда затруднено выделение «естественных» структурных единиц в сложно-организованных объектах, которые часто представимы знаковыми последовательностями. Существенный прорыв в исследовании структуры текста произошел в семидесятых годах XX века, когда советский ученый-кибернетик Ю.К. Орлов [20] с помощью средств вероятностно-статистического и энтропийного подходов открыл феномен целостности художественного литературного и музыкального произведений, который проявлялся в хорошем совпадении реального частотно-рангового распределения слов завершенного текста с законом Ципфа-Мандельброта, представляемым в виде

$$p_i = \frac{K}{(B+i)^{\gamma}}, \quad K, B, \gamma - const \tag{17}$$

$$K = \frac{1}{\ln F_1}, \quad B = \frac{K}{p_1} - 1, \tag{18), (19}$$

$$v_T = KZ - B, \tag{20}$$

где

p_i , p_1 - частоты, соответственно, *i*-ого, и 1-ого по рангу слов,

F_1 – число вхождений самого частого слова,

v_T – теоретический объем словаря текста длиной *Z*.

В биологии аналогичные статистические распределения известны, как квазигиперболы или правило Д.К. Виллиса [21].

Набор признаков, на основании которых производилась формальная идентификация текста, назван критерием Орлова [22]. В состав данного критерия включены следующие по приоритету факторы:

– точность совпадения теоретического и фактического объемов алфавита элементов;

– степень совпадения фактического рангового распределения с законом Ципфа-Мандельброта.

Таким образом, Ю. Орлов, считая очевидными слова текста, поставил и решил *прямую задачу* – определение (идентификация) целостности текста. В наших же разработках была поставлена *обратная задача* – полагая наблюдаемые знаковые цепи целостно-завершенными, требовалось выделить в них слова (естественные компоненты), распределение которых удовлетворяет критерию Ю. Орлова.

Представим алгоритм выявления естественных структурных единиц знаковой последовательности или ее сегментация на слова [22]. Данный алгоритм, использует метод, называемый далее базовым [23], который считает выбранное псевдослово (короткую подпоследовательность) словом, если оно имеет неожиданно большую или малую частоту появления в тексте, по сравнению с теоретической частотой, определенной, исходя из предположения, что текст является марковской цепью. Степень различия фактической и теоретической частот задается некоторым пороговым значением. Представляемый здесь алгоритм, использовалась ранее [22, 24] и дополняет базовый следующими операциями. После первой сегментации исходной знаковой цепи с помощью базового алгоритма определяется статистическое частотно-ранговое распределение первой версии словаря, которое затем сравнивается с теоретическим распределением Ципфа-Мандельброта. Если расхождение этих распределений превышает некоторый установленный порог, то изменяется пороговое значение для выделения слов, и базовый алгоритм повторно сегментирует исходную последовательность. Такая процедура сравнения теоретического и фактического распределений потенциальных словарей с последующим изменением порога для выделения слов сопровождается неоднократной сегментацией исходной знаковой цепи вплоть до установленной степени совпадения распределений.

В результате эксперимента по сегментации аминокислотной последовательности *Takifugu rubripes* с помощью модифицированного алгоритма удалось добиться совпадения мощностей фактического и теоретического алфавитов $v = v_T$ и соответствия фактического и расчетного распределений. Характерно то, что алгоритм начал склеивать аминокислоты в слова, группами по три и меньше.

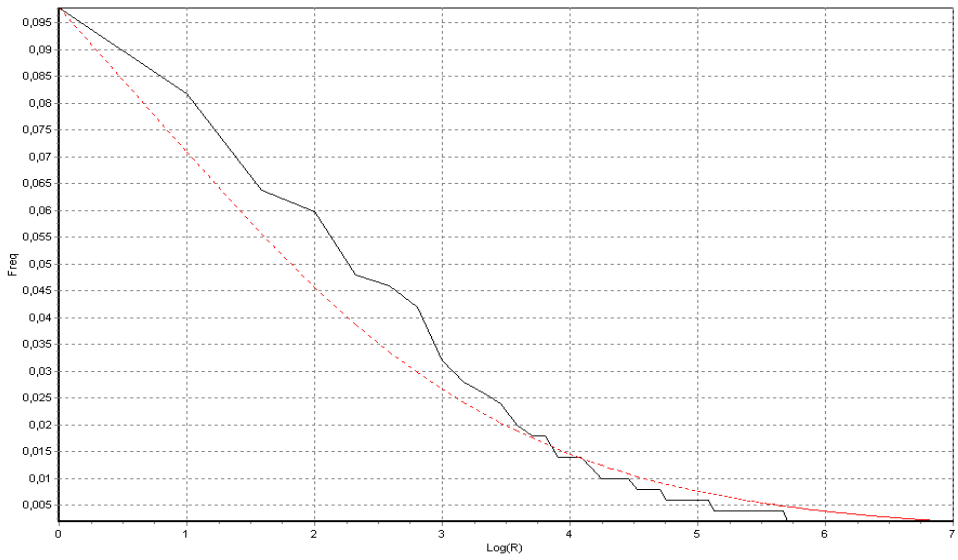

Рис. 17. Фактическое и расчетное частотно-ранговые распределения аминокислотных слов белка *Takifugu rubripes*. По оси абсцисс ранг представлен в логарифмическом масштабе.

Таблица 5. Характеристики аминокислотной цепи *Takifugu rubripes* (AB197152.2)

Z	F_1	p_1	v	v_T
501	49	0,098	128	128

В табл. 5 v – число разных аминокислотных слов в данном разбиении (мощность фактического словаря); остальные обозначения см. выше.

Из факта совпадения мощностей фактического и расчетного словарей данного белка видно, что разбиение на элементы, по критерию Орлова, прошло верно, и расхождения с законом Ципфа-Мандельброта незначительны. Это же демонстрирует рис. 17, представляющий распределения.

Знаковая цепь, отображающая последовательность белка и разбитого на аминокислотные слова в соответствии с критерием Орлова, представлена ниже. В дальнейшем предполагается провести исследования особенностей построения аминокислотных и нуклеотидных цепей, разбитых на естественные единицы, с помощью средств формального анализа строя цепей [25].

M A **A** L **S** S **AES** P **P** P **V** L **N** G **D** AAE **R** DPG **RER** G **L** E **E** L **D** S **G** FNS **A** C **TT** R **V** P **G** GAQ **N** E **EIW** NI **KQM** IKL **TQE** H **LEA** L **L** D **KF** G **G** E **H** N **P** P **TIY** LEA **YE** E **YTS** K **L** D **A** L **Q** Q **R** E **Q** Q **L** LEA **MGN** G **T** D **FPC** S **P** S **PMP** A **L** L **E** VK **M** G **G** C **V** P **G** V **GAQ** APN **S** LA **V** L **Q** TP **T** D **G** S **R** V **N** P **RS** PQK **P** I **V** R **V** F **L** PNK **Q** R **TV** V **S** A **R** CGM **TV** RD **S** LKK **A** L **TMR** G **L** IPE **C** C **A** V **YRM** Q **D** GEK **K** P **IGW** DTD **ISW** L **T** L **E** E **L** HVE **V** L **E** N **VPL** TT **H** NFV **R** K **TFF** T **LA** F **C** DF **CRK** LLF **Q** G **F** R **CQ** TCG **Y** KF **HQ** R **CST** E **VPL** M **C** V **NYD** Q **L** D **L** LLV **S** KF **F** E **H** HPF **TQE** E **V** SSE **G** TT **P** V **S** E **V** C **P** S **L** P **P** S **D** S **T** G **SI** CQ **S** TV **S** P **S** K **SI** P **I** P **P** S **F** R **SSE** E **D** HRN **Q** F **G** Q **RD** RS **S** S **APN** V **HI** NTI **E** P **V** NI **D** DLI **R** VQG **L** P **RS** D **G** G **S** TT **G** L **S** A **TP** P **A** S **L** P **G** S **LTN** VK **V** PQK **S** P **CQ** Q **RER** K **S** S **S** SSE **D** RS **K** MLG **R** RD **S** S **DDW** E **IPE** G **QI** TLG **Q** RIG **S** G **S** F **G** TV **F** KGK **WHG** DVA **VK** MLN **V** T **A** P **TP** Q **Q** L **Q** AF **K N** E **V** G **V** LRK **TRH** V **NI** LLF **M** G **Y** TT **K** P **Q** LA **I** VTQ **WCE** G **S** S **L** YHH **L** HI **IET** KF **E** M **IKL** IDI **ARQ** TAQ **GMD** Y **L** HAK **SI** IH **R** DLK **S** N **NI** FLH **E** D **L** TV **KIG** DF **G** LA **T** VK **SRW** S **G** S **HQ** F **E** Q **L** S **G** S **ILW** M **APE** VIR **L** Q **D** KN **PY** S **F** Q **S** DVY **AF** GIV **L** YE **L** MSG **V** L **PY** S **NI** N **N** RD **Q** IIF **MVG** RGY **L** S **P** D **L** S **K** V **RS** N **C** P **KAM** KRL **MAD** C **LKK** KRE **ERP** LFP **QI** LA **SI** E **L** LA **RS** L **PKI** H **RS** A **S** E **P** S **L** NRA **G** F **QTE** DF **S** L **YC** A **S** PKT **P** IQA **G** G **YGE** F **S** AF **K**

МЕРА РАСХОЖДЕНИЯ НУКЛЕОТИДНЫХ ЦЕПЕЙ

Определим меру расхождения [25] оригинальных построений разных неоднородных последовательностей A и B, на основе средних удаленностей g_j, составляющих их однородных цепей, в виде $r = \sum_{j=1}^{m} \left| g^A{}_j - g^B{}_j \right|$, табл. 6 представляет матрицу расхождений строя рибосомальных РНК некоторых из организмов, представленных в табл. 1.

Таблица 6. Матрица расхождений r строев ряда последовательностей. Номера в столбцах соответствуют номерам организмов из табл. 1, m = 4

	23	**3**	**6**	**21**	**27**	**29**
23	0	0,870	1,266	0,575	0,952	1,169
3	0,870	0	0,454	0,792	1,229	1,443
6	1,266	0,454	0	1,066	1,329	1,543
21	0,575	0,792	1,066	0	0,486	0,700
27	0,952	1,229	1,329	0,486	0	0,217
29	1,169	1,443	1,543	0,700	0,217	0

ПРИМЕНЕНИЕ ХАРАКТЕРИСТИК СТРОЯ ДЛЯ ТАКСОНОМИИ ОРГАНИЗМОВ

Ниже (на рис. 18) приведено распределение удалённостей для 29 организмов из табл. 1. Также здесь обозначено экспертное разбиение данных организмов.

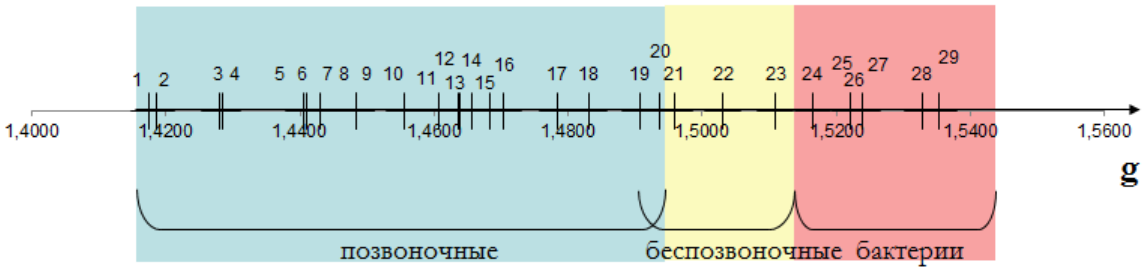

Рис. 18. Распределение по удалённости рРНК 29 организмов. Номера подписей соответствуют номерам организмов из табл. 1.

Были выполнены разбиения на таксоны с опорой на нуклеотиды и с опорой на триплеты выборки из 29 организмов. Для таксономии организмов по их ДНК использовался алгоритм кластеризации λ-KRAB, описанный в [11], отличительной особенностью которого является использование нелинейного нормированного λ-пространства, учитывающего не только расстояние между объектами, но и характеристику локальной плотности расположения объектов в этом пространстве.

Таблица 7. Результаты таксономии в табличном виде (опыт 1 соответствует рис. 19, опыт 2 соответствует рис. 20)

1 опыт	Название объекта	2 опыт
2	*Candidatus Nitrosopumilus maritimus* (бактерия хемолитотроф)	1
2	*Bacillus anthracis* (сибирская язва)	1
1	*Thermus thermophilus* (термофильный микроорганизм)	1
1	*Thermotoga thermarum* (термофильный микроорганизм)	1
2	*Streptococcus pyogenes* (стрептококк пиогенный)	1
2	*Pediculus humanus capitis* (вошь головная)	1
2	*Neisseria gonorrhoeae* (возбудитель гонореи)	1
2	*Mycoplasma pneumoniae* (возбудитель атипичной пневмонии)	1
3	*Mus musculus domesticus* (мышь домовая)	1
2	*Ixodes persulcatus* (клещ таёжный)	2
3	*Homo sapiens* (человек разумный)	2
3	*Gallus gallus* (курица домашняя)	2
3	*Caiman crocodilus* (кайман крокодиловый)	2
3	*Canis lupus familiaris* (собака)	2
2	*Borrelia burgdorferi* (возбудитель болезни Лайма)	2
3	*Amia calva* (ильная рыба)	2
2	*Zebrias zebra* (рыба)	2
3	*Sus scrofa* (кабан)	2
3	*Rattus norvegicus* (серая крыса)	2
3	*Mus musculus domesticus* (мышь домовая) 2	2
2	*Kareius bicoloratus* (камбала двухцветная)	2
3	*Homo sapiens* (человек разумный) 2	2
3	*Gallus gallus* (курица домашняя) 2	2
3	*Erinaceus europaeus* (ёж обыкновенный)	2
2	*Crocodylus niloticus* (крокодил Нильский)	2
3	*Cricetulus griseus* (хомячок китайский)	2
3	*Bos taurus* (бык дикий)	2
2	*Ornithodoros moubata* (клещ)	2
2	*Musca domestica* (муха домашняя)	2

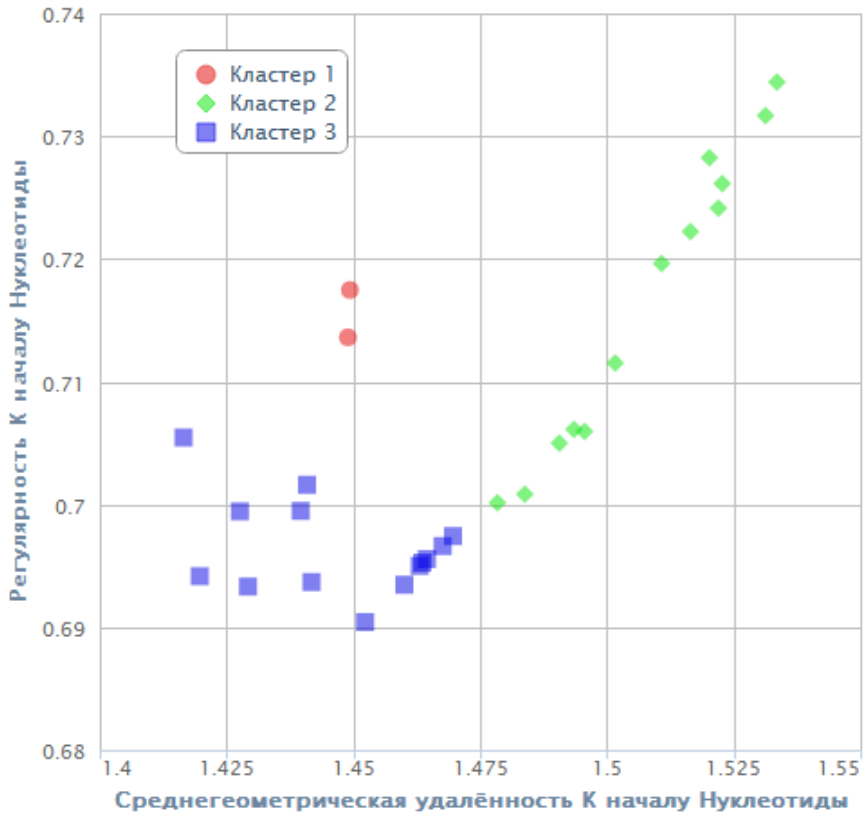

Рис. 19. Результаты автоматической таксономии 29 организмов по характеристикам удалённости и регулярности с опорой на нуклеотиды.

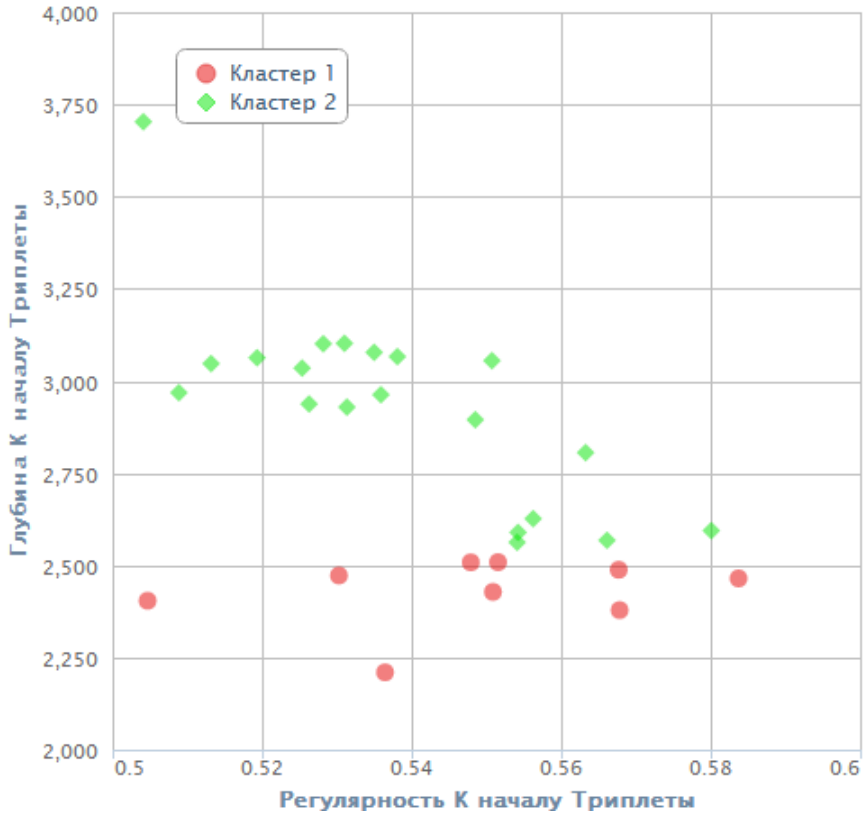

Рис. 20. Проекция на плоскость двух характеристик результатов автоматической таксономии 29 организмов по характеристикам удалённости, регулярности и глубины с опорой на триплеты.

На рис. 20 приведена проекция таксономии по трём характеристикам: удалённости, регулярности и глубине. Характеристики триплетного представления, хотя и дают другую картину разбиения, но при этом представляют дополнительную информацию о свойствах организмов и взаимосвязях между ними.

Предварительные исследования позволяют сделать вывод о том, что классификация с опорой на разные структурные единицы даёт разные представления таксономии. По нашему мнению, для получения правильной классификации, необходимо выбрать адекватные структурные единицы естественных цепей. Открытым остаётся вопрос поиска таких структурных единиц (естественных слов) для генетических текстов.

ЗАКЛЮЧЕНИЕ

В работе представлены средства, позволяющие осуществлять формальный анализ строя нуклеотидных цепей. Определён абстрактный объект – строй цепи, представляющий взаимное расположение компонентов в конкретной последовательности. Сформулированы выражения для числовых характеристик, описывающих порядок строя цепи произвольной природы, а также формулы для определения зависимостей однородных цепей. Установлена связь характеристик строя с общепринятыми статистическими и энтропийными характеристиками. Для более полного описания строя цепей предложены распределения характеристик по однородным цепям. Отмечена возможность использования числового анализа локальной структуры цепей. Представленные средства однозначно отображают знаковые цепи разной природы числовыми последовательностями с возможностью последующего применения разнообразных методов анализа функций и упорядоченных числовых данных.

Описан алгоритм сегментации для выделения естественных единиц – слов в нуклеотидной цепи.

Получены значения характеристик строя нуклеотидных цепей для выборки организмов трёх царств. Графически и таблично представлены распределения этих характеристик.

Показана возможность использования характеристик строя нуклеотидных цепей для автоматической таксономии организмов.

Представленный инструментарий позволяет выявлять конструктивные шаблоны, используемые при формировании строя нуклеотидных цепей.

СПИСОК ЛИТЕРАТУРЫ

1. Zipf G.K. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, 1932.
2. Садовский М.Г. *Информационно-статистический анализ нуклеотидных последовательностей*: диссертация на соискание учёной степени доктора физико-математических наук по специальности 03.00.02 – биофизика. Красноярск, 2004. 394 с.
3. Kullback S., Leibler R.A. On information and sufficiency. In: *The Annals of Mathematical Statistics*. 1951. V. 22. № 1. P. 79–86.
4. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады Академии Наук СССР*. 1965. Т. 4. С. 845–848.
5. Гусев В.Д., Косарев Ю.Г., Титкова Т.Н. Методы поиска и анализ статистических закономерностей в символьных последовательностях. В: *Машинные методы обнаружения закономерностей*: материалы всесоюзного симпозиума. Новосибирск, 1976. С. 75–84.

6. Гусев В.Д., Куличков В.А., Никулин А.Е. Алгоритмы поиска несовершенных повторов в генетических текстах. В: *Анализ символьных последовательностей*. Новосибирск, 1985. С. 107–122. *(Вычислительные системы, вып. 113).*
7. Гусев В.Д., Немытикова Л.А. Векторная мера сложности нуклеотидных последовательностей. В: *Третий сибирский конгресс по прикладной и индустриальной математике (ИНПРИМ-98)*. Новосибирск: Изд-во Института математики СО РАН, 1998. С. 115.
8. Гусев В.Д., Мирошниченко Л.А., Саломатина Н.В. Методы выделения структурных единиц в символьных последовательностях. Межъязыковые аналоги. В: *Материалы Всероссийской конференции с международным участием «Знания-Онтологии-Теории»*. Новосибирск: Изд-во института математики СО РАН, 2009. Т. 2. С. 53–62.
9. Беликов С.И., Гусев В.Д., Мирошниченко Л.А., Титкова Т.Н. Сравнительный анализ геномов вирусов клещевого энцефалита: дифференциация по степени вирулентности. В: *Математическая биология и биоинформатика: IV международная конференция.* Пущино: изд-во ООО «МАКС Пресс», 2012. С 52–53.
10. Гуменюк А.С., Кликушин Ю.Н., Кобенко В.Ю., Циганенко В.Н. *Алгоритмы анализа структуры сигналов и данных*. Омск, 2010. 272 с.
11. Загоруйко Н.Г. *Прикладные методы анализа данных и знаний*. Новосибирск, 1999. 270 с.
12. Мазур М. *Качественная теория информации*. Москва: Мир, 1974. 240 с.
13. Gumenjuk A., Kostyshin A., Simonova S. An approach to the research of the structure of linguistic and musical texts. *Glottometrics.* 2002. № 3. P. 61–89.
14. Вентцель Е.С. *Теория вероятностей*. М.: Наука, 1969. 576 с.
15. *GENBANK DataBase*. URL: http://www.ncbi.nlm.nih.gov/nuccore/ (дата обращения: 13.02.2013).
16. Woese C.R., Kandler O., Wheelis M.L. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA.* 1990. № 87. P. 4576–4579.
17. Гуменюк А.С., Морозенко Е.В., Родионов И.Н. Формализация анализа строя знаковых цепей. *Вестник Томского государственного университета управления, вычислительной техники и информатики*. 2011. № 2. С. 15–24.
18. Гуменюк А.С. О средствах анализа взаимного расположения компонентов знаковой последовательности. В: *Военная техника, вооружение и технологии двойного применения:* материалы III Международного технологического конгресса. Омск: ОмГТУ, 2005. Ч. 2. С. 48–52.
19. Гуменюк А.С., Поздниченко Н.Н. Определение характеристик пространственной зависимости компонентов нуклеотидных цепей. В: *Математическая биология и биоинформатика: IV международная конференция*. Пущино: МАКС Пресс, 2012. С 56–57.
20. Орлов Ю.К. Невидимая гармония. В: *Число и мысль*. М.: Знание, 1980. № 3. С. 70–105.
21. Чайковский Ю.В. *Активный связный мир. Опыт теории эволюции жизни*. М.: Товарищество научных изданий КМК, 2008. 726 с.
22. Гуменюк А.С., Костышин А.С. О компьютерном анализе текстов и одном формализме сегментации стихотворений русской литературы на сочетания фонем. В: *Квантитативная лингвистика и семантика. Сб. науч. тр.* Новосибирск: Изд-во НГПУ, 2000. № 1. С. 3–18.
23. Бородовский М.Ю., Певзнер П.А. Статистические методы анализа генетических текстов. В: *Компьютерный анализ генетических текстов*. М.: Наука, 1990. С. 33–80.

24. Gumenjuk A., Kostyshin A., Borisov K., Salnikova O. On the acoustic elements of a poem and the formal procedures of their segmentation. In: *Glottometrics.* Lüdenscheid: RAM-Verl., 2004. № 8. P. 42–67.
25. Гуменюк А.С., Шпынов С.Н., Морозенко Е.В., Родионов И.Н. О средствах анализа строя нуклеотидных цепей. В: *Математическая биология и биоинформатика: II международная конференция.* Пущино: МАКС Пресс, 2008. С. 123–125.

Материал поступил в редакцию 21.02.2013, опубликован 30.06.2013.