УДК 006.72 **DOI** 10.15622/sp.18.2.471-503

А.С. Гуменюк, А.А. Скиба, Н.Н. Поздниченко, С.Н. Шпынов

О МЕРАХ СХОДСТВА РАСПОЛОЖЕНИЯ КОМПОНЕНТОВ В МАССИВАХ ЕСТЕСТВЕННО УПОРЯДОЧЕННЫХ ДАННЫХ

Гуменюк А.С., Скиба А.А., Поздниченко Н.Н., Шпынов С.Н. **О мерах сходства расположения компонентов в массивах естественно упорядоченных данных.**

Аннотация. В настоящее время в публикациях специалистов по анализу массивов естественно упорядоченных данных различной природы (в том числе символьных последовательностей) не имеют широкого распространения математические средства, адекватно учитывающие расположение компонентов. Поэтому затруднены или невозможны измерение и сравнение порядка следования сообщений, выделенных в длинных информационных цепях. Основные подходы при сравнении символьных последовательностей используют вероятностные модели и статистический инструментарий, попарное и множественное выравнивание, позволяющее определить степень сходства цепей с помощью мер редакционного расстояния. Отмеченные подходы почти не уделяют внимания исследованию и обнаружению закономерностей конкретного расположения всех знаков, слов, компонентов массивов данных, составляющих отдельную целостную последовательность. Объектом исследования в наших работах является специальным образом организованный числовой кортеж — расположение компонентов (строй) в символьных или числовых последовательностях. При этом в качестве основы для количественного отображения строя цепи используются интервалы между ближайшими одинаковыми ее компонентами. Перемножение всех интервалов или суммирование их логарифмов позволяет получить числа, которые однозначно отображают расположение компонентов в конкретной последовательности. Эти числа, в свою очередь, позволяют получить целый набор нормированных характеристик строя, среди которых средний геометрический интервал и его логарифм. В данной работе представлен подход для количественного сравнения построений массивов естественно упорядоченных данных (информационных цепей) произвольной природы. Предложены меры сходства-расхождения и процедура сравнения строя цепей, основанные на выделении списка совпадающих и сходных по характеристикам строя подпоследовательностей. При этом для быстрого выделения списка совпадающих компонентов используются ранговые распределения. В работе представлен инструментарий для сравнения построений информационных цепей и продемонстрированы некоторые его возможности при исследовании строя нуклеотидных последовательностей.

Ключевые слова: знаковая последовательность, информационная цепь, строй цепи, глубина строя, средняя удаленность, нуклеотидная последовательность, меры сходства-расхождения, матрица сходства, alignment-free genome comparison, межнуклеотидное расстояние.

1. Введение. Уже более 100 лет используются формальные средства для анализа знаковых последовательностей разной природы. В начале прошлого века при зарождении математической лингвистики появились работы по статистическим исследованиям текстов на естественных языках [1]. В 50-60 годы на фоне широкого использования цифровых вычислительных машин отдельные исследователи применили формальный анализ к музыкальным произведениям, и одновременно

стали использоваться математические модели и средства анализа так называемых «генетических текстов», то есть нуклеотидных, аминокислотных последовательностей и тому подобных. Кроме того, начались интенсивные исследования больших массивов естественно упорядоченных данных измерений (данные мониторинга). В таких массивах обычно требуется учитывать оригинальное расположение выделенных компонентов. В процессе анализа подобных последовательностей исследователи пытаются выявить структуру массива данных. При этом непосредственное исследование структуры никакими средствами не осуществляется, так как не определено и само понятие структуры для цепей данных. Обычно исследователи пытаются опереться на определенную природу компонентов таких цепей (слова, нуклеотиды, триплеты, кодоны, аминокислоты, амплитуды сигналов, высоты звучания нот и тому подобное). Это направление исследований структуры породило большое число «тонких формальных техник», применимых к цепям выделенной природы [2-10].

Общие подходы к исследованию структуры знаковых цепей без опоры на материальную природу компонентов представлены двумя основными направлениями: суждения о структуре цепи на основе статистического распределения ее элементного состава и косвенные суждения о структуре цепи с помощью оценки и исследования локального порядка следования компонентов в цепях.

В первом подходе суждение о структуре полной цепи осуществляется на основе статистического распределения (состава) ее компонентов (двоек, троек и в общем случае n-ок) [11]. Так как компоненты в рамках вероятностной модели цепи представляют собой случайные события (но не величины), 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 [12], длина окна для короткого кортежа (n-ки) может быть такой достаточной 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 расположение всех компонентов исходной цепи. Однако такое численное описание структуры достигается путем введения многократной избыточности за счет $(n-1)$-кратного тиражирования цепи. Суждения же на основе обычного статистического распределения компонентов (или блоков компонентов) данной цепи, полученного экспериментально, не претендуют на возможность восстановления

исходной последовательности. Косвенно такие распределения все же являются количественным описанием взаимного расположения элементов, так как исследователю по умолчанию известно, что он взял не случайную выборку данных, а конкретный текст, нуклеотидную последовательность и тому подобное [13]. Это «проклятие априорного неосознаваемого знания» об очевидной упорядоченности цепи широко распространено в математической лингвистике, биоинформатике (математической биологии) и других аналогичных областях науки.

Другое направление исследований и анализа структуры массивов данных в основном использует мощные известные вероятностно-статистические средства и модели — марковские цепи, потоки заявок и теорию очередей, взвешенные графы, с помощью которых удается хоть и громоздко описать локальную структуру знаковых цепей, но не оригинальное расположение компонентов всей цепи [8, 14].

Особо выделим работу выдающегося математика современности В.И. Арнольда, посвященную разработке теории сложности конечных бинарных последовательностей [15].

Кроме того, косвенный анализ структуры осуществляется путем 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», которую, по существу, можно считать обобщенной метрикой Хемминга, представляющей расстояние между словами одинаковой длины [16-19].

В настоящее время в противоположность вычислительно сложным методам, основанным на выравнивании последовательностей, в биоинформатике выделяют группы подходов (в основном статистических), называемых alignment-free sequence analysis, которые позволят сравнивать и описывать нуклеотидные последовательности без применения выравнивания и отличаются высоким быстродействием [20-26].

Среди них выделим методы символической динамики, основанные на подходах систем «динамического хаоса», которые обычно используются для визуализации символьных последовательностей в виде траекторий в пространстве некоторой размерности (обычно 2 или 3). В некоторых работах из таких траекторий далее получают интегральные числовые характеристики [27, 28].

В рамках выделенных направлений, по исследованию структуры знаковых последовательностей широко применяется получение вероятностно-статистических и энтропийно-информационных характеристик и оценок. На протяжении десятилетий

глубокие разработки на основе этих средств и моделей, а также с учетом природы объектов со значительным практическим выходом выполняются в Институте математики им. Соболева сибирского отделения РАН в лаборатории под руководством В.Д. Гусева [2-6].

Приведем утверждение из [12], которое определяет методологический подход для анализа и исследования структуры символьных последовательностей: «Как хранение, так и реализация какой-либо информации напрямую обусловлены тем обстоятельством, что в ходе этих процессов актуальную роль играют символьные последовательности. При этом хорошо известно, что в природе фактически нет процессов, связанных с переработкой либо реализацией той или иной информации, которые бы вовлекали всю такого рода символьную последовательность целиком: чтение и обработка файлов вычислительными машинами происходит малыми порциями (байтами) и последовательно, чтение и переработка письменной информации человеком происходит малыми порциями (словами, либо абзацами) и последовательно, чтение и переработка наследственной информации в биологических системах происходит малыми порциями (кодонами) и последовательно. Это простое обстоятельство, тем не менее, имеет важные последствия. Оно требует перехода от рассмотрения всей символьной последовательности в целом к рассмотрению набора ее фрагментов.». Данное утверждение является своего рода постулатом редукционизма или познавательной установкой по Ю. Шрейдеру для данного методологического подхода [29]. По мнению авторов такое допущение существенно ограничивает возможности теории информации, прикладной информатики, в частности математической лингвистики, информатики нуклеотидных последовательностей и средств анализа массивов естественно упорядоченных данных любой природы.

Для исследования знаковых цепей, текстов разной природы и массивов данных измерений разработаны и используются, как отмечалось выше, большое число специальных подходов, процедур и моделей, которые можно дополнить математическим, спектральным, статистическим, корреляционным, фрактальным и другими анализами. Однако почти не уделяется внимания исследованию и обнаружению закономерностей *конкретного расположения всех знаков, слов, компонентов массивов данных, составляющих отдельную целостную последовательность.* Можно констатировать, что до настоящего времени массивы естественно упорядоченных (текстовых) данных обычно рассматриваются как множества (но не кортежи), в которых не принято учитывать и численно представлять расположение их элементов. Можно высказаться

более категорично — до настоящего времени массивы естественно упорядоченных данных обрабатывались, анализировались и исследовались как «вещество», а не как кортежи, информационные цепи или связные «тексты».

По мнению авторов такое положение в некоторой степени объясняется следующими причинами:

1. Отсутствие английского перевода фундаментальной работы М. Мазура «Качественная теория информации», в которой определяются и особо рассматриваются информационные цепи (массивы данных) в отличие от кодовых цепей [30].

2. Отказ от системного подхода, который учитывал бы тексты и массивы естественно упорядоченных данных как абстрактные объекты, каждый из которых представляет собой единое целое.

3. Очевидность определенного расположения компонентов в естественно упорядоченной последовательности, которое нельзя искажать при обработке данных; очевидное не побуждает к формализации.

4. Отсутствие формализма для особого абстрактного объекта, представляющего расположение компонентов и называемого нами строем или построением цепи [31, 32].

Следует отметить, что разные по природе последовательности событий с одинаковыми статистическими распределениями (в дальнейшем — с равномощными составами) могут иметь одинаковый 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 большинство из них будут иметь разное расположение компонентов.

В работах [9, 10, 33-36] использовались интервалы между ближайшими одинаковыми компонентами (межнуклеотидное расстояние) в качестве основы для исследования и сравнения нуклеотидных последовательностей. Однако в них рассматривались только статистические (ранговые) распределения интервалов [37, 38], а также преобразования Фурье и вейвлет-преобразования [39] нуклеотидных последовательностей, что не позволяло описать отдельную последовательность одним числом. Для таких (обычно гиперболических) распределений не определялись статические характеристики (мат ожидание, СКО и тому подобное).

Наконец, особо отметим открытие Ю. Орловым феномена «целостно-завершенного текста», который обнаруживается только при хорошем совпадении статистического рангового распределения

слов данного текста с законом Ципфа — Мандельброта [40, 41]. Ю. Шрейдер формально доказал, что существование такого «идеального» распределения слов для отдельного текста вытекает из фундаментального принципа «минимума симметрии» для целостной системы [29]. При этом для исследования естественных систем им предложена комбинированная методология поочередного использования системного подхода и редукционизма. В России это направление развивается для исследования техноценозов под руководством Б. Кудрина [42]. Применимость закона Ципфа — Мандельброта при исследовании статистических распределений генетических текстов рассматривалось М. Гельфандом [43]. К сожалению, отмеченные разработки советских и российских ученых до настоящего времени не попали в поле зрения англоязычного научного сообщества.

В наших ранних работах [31] предлагается подход, который предназначен для «формального описания и анализа строя» (ФОАС) отдельного текста любой природы (знаковой цепи), в том числе представляющего нуклеотидную последовательность, или массив данных измерений.

2. Формальное описание строя. *Строй цепи* сообщений (событий, знаков и тому подобных) — 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.

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 $\{1, 2, 3, 4, 5, \ldots, m\}$.

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

3. Первые вхождения элементов алфавита располагаются на позиции строя по возрастанию, начиная с единицы в первой позиции, возможно с пропусками некоторых мест:

`⟨1 2-3-4--5---6 7⟩`.

4. Места на позиции строя, не занятые первыми вхождениями элементов алфавита, заполняются натуральными числами, по значению не превышающими максимального среди всех лежащих слева чисел:

⟨1 2 1 3 2 3 4 4 4 1 5 3 4 5 1 1 1 6 7⟩.

Мощность алфавита строя — это количество различных компонентов в цепи.

Примеры разных последовательностей (кортежей) символов с одинаковым строем приведены на рисунках 1 и 2.

Для сравнения по строю нескольких кортежей реальных сообщений необходимо правильно выполнить однозначное *прямое преобразование* для каждого из них, а затем сравнить полученные строи.

Для кодирования разных знаков при прямом преобразовании цепи сообщений в строй цепи кроме натуральных чисел возможно использовать любой (упорядоченный) алфавит символов достаточно большой мощности. Соответствие между исходной и закодированной таким образом последовательностями называется «совпадением с точностью до переименования». Однако такой алфавит необходимо выбрать или специально построить и самое трудное — сделать его общепринятым. Кроме того, все реальные алфавиты и словари неявно упорядочены натуральными числами для удобства запоминания и использования.

В теоретико-множественном представлении вектором называется кортеж, компонентами которого являются числа. В соответствии с таким определением вектора, назовем специфически сформированный (организованный) кортеж «*вектором строя*».

Таким образом, строй цепи и вектор строя — это синонимы одного и того же абстрактного объекта. Однако на практике следует различать «вектор строя» данной цепи или некоторого их множества и «вектор строя» как элемент множества разных векторов строя.

Заметим, что при несоблюдении ограничений на порядок расположения натуральных чисел мы получим кортеж, точнее вектор, который не представляет собой строй. На рисунке 3 представлен такой вектор.

Рассмотрим отличный от представленного на рисунке 1 строй цепи. Очевидно неоднозначное преобразование данного строя в знаковые последовательности. Для наглядности, пусть они имеют мощность состава элементов такую же, как на рисунке 1. Условимся называть преобразование строя в знаковую цепь «*обратным преобразованием строя*» (рисунок 4).

При одинаковой мощности составов знаковых цепей их частотные распределения одинаковы, то есть инвариантны относительно расположения элементов в цепях, что видно из примеров.

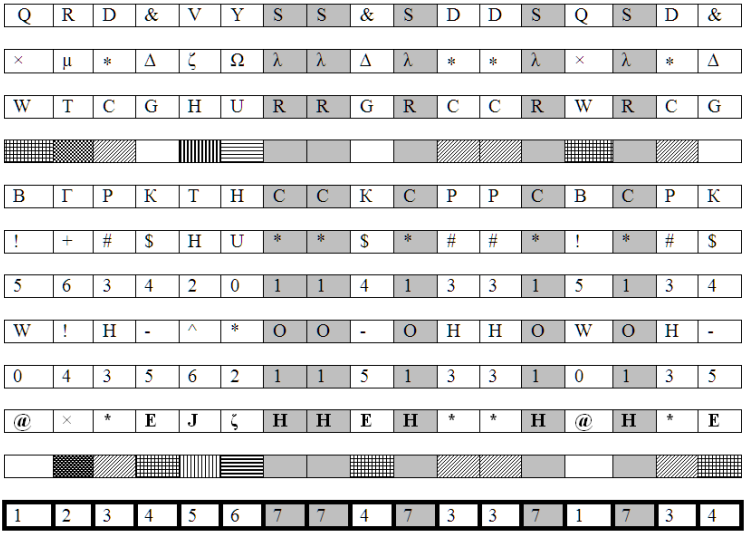

Рис. 1. Пример прямого преобразования 11 разных знаковых цепей, цифровых последовательностей и диаграмм в строй цепи

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 РНК бактерии Candidatus nitrosopumilus maritimus с одинаковым строем

Строй цепи — это идея или план построения некоторого множества кортежей, цепей реальных сообщений, сигналов или событий. *Строй цепи* в определенном смысле соответствует введенному Гете понятию «*архетип*». Этот термин предложил использовать Юлий Шрейдер для обозначения описания структуры таксона (класса объектов) [29]. Другими словами, *строй цепи — это ее архетип*.

Операция выявления в разных по природе информационных цепях одинаковых построений расширяет возможности междисциплинарных исследований. Однако результат такой операции ограничен описанием строя в форме обычного числового кортежа, хотя и имеющего определение «вектор строя». Рассмотрим более удобное для анализа формальное описание строя, которое позволяет получать компактные числовые характеристики (подобные используемым для описания случайных величин), полезные, в частности, при опознавании строев цепей и

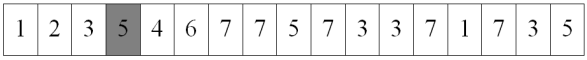

Рис. 3. Вектор натуральных чисел, который не представляет строй цепи

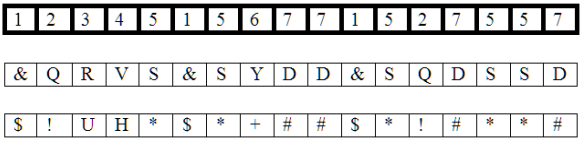

Рис. 4. Обратное преобразование данного строя в две разные знаковые последовательности

определении степени их различия. Для определения такого формализма строя отдельной цепи при обычном (естественном) способе ее чтения «поэлементно» (подряд) введем две нумерации:

– первая нумерует элементы собственного алфавита (словаря) данной знаковой последовательности по порядку их встречи;

– вторая дает сквозную нумерацию всех компонентов кортежа от начала до конца.

Разложим *полную неоднородную* (без пустых мест на позиции) символьную последовательность на m *неполных однородных кортежей*, на позициях которых заняты одинаковыми знаками только некоторые места (рисунок 5). Такое разложение цепи называют *декомпозицией*. Аналогом однородной последовательности является поток однородных событий (заявок) из теории массового обслуживания. Очевидно, что *композиция* всех однородных строев данного полного строя дает полный неоднородный строй, аналогом которого в теории очередей является неоднородный поток событий. Вообще разложение цепи может осуществляться по разным правилам. Декомпозиция строя полной неоднородной знаковой цепи на неполные однородные представлена на рисунке 5.

В приведенных на рисунке 5 примерах используется привязка к концу последовательности, то есть последний интервал считается от последнего вхождения компонента до конца последовательности. Кроме данной привязки могут также использоваться следующие варианты: к началу, к началу и к концу, циклическая, либо отсутствие привязки.

Определим «*интервал*» как расстояние от выделенного в цепи компонента до другого ближайшего, отмеченного в направлении просмотра (рисунок 5); *величина интервала* — это натуральное число, определенное как модуль разности номеров мест двух выделенных

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

компонентов на позиции кортежа. В дальнейшем для краткости будем называть эти понятия одинаково — интервал.

Назовем направление чтения текста или знаковой цепи «подряд слева направо» «обычным способом считывания». Пусть первое считывание текста осуществляется отличным от обычного способом с самого начала до конца таким образом, что выбираются только элементы строя с номером «1»; при этом последний интервал определяется до знака «финиш» (возможен и другой вариант — определение первого интервала от начала текста — «старта»). Интервалы данной однородной последовательности разместим в соответствии с номерами считываемых элементов в первой строке матрицы. Далее, при втором просмотре строя текста аналогично выберем элементы с номером «2» и разместим вектор интервалов, соответствующий другой однородной последовательности, во второй строке матрицы. В каждой следующей строке помещается вектор интервалов «новой» при очередном просмотре однородной последовательности. Одиночные знаки, слова или сообщения будут представлены всего одним интервалом (до финиша), который размещается в крайнем столбце соответствующей строки матрицы. Число столбцов $n_{j\,max}$ в «*матрице интервалов» однородных цепей* равно числу вхождений самого частого знака (или слова) текста. Незанятые интервалами элементы матрицы заполним нулями. Число строк m равно мощности алфавита или словаря текста. Результаты описанных действий представлены на рисунке 5.

В случае правильного выполнения декомпозиций, полученные множества однородных последовательностей (рисунки 6, 7) будут несовместными (так как не содержат занятых мест с одинаковыми номерами на их позициях). *Композиция* или «совмещение» всех

неполных однородных кортежей дает исходную полную неоднородную последовательность.

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. 6. Фрагменты нуклеотидных цепей *Cricetulus griseus* и *Homo sapiens* с одинаковым строем (длина фрагментов 15). Выделены путем просмотра рибосомальной РНК общей длиной 1871и 1559. Совпадение 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. 7. Однородные фрагменты нуклеотидных цепей *Cricetulus griseus* и *Homo sapiens* с одинаковым строем (длина фрагментов 19). Совпадение строя начинается с позиций: 108 для первой цепочки и 1847 — для второй цепочки

Следует также отметить описание знаковых цепей и текстов графами. Обычно это взвешенный граф типа «дерево», узлы которого представляют выделенные по определенным признакам символы или слова, а ребра — это интервалы между ними.

Еще раз подчеркнем, что *для исследования построения реальной информационной цепи вводится формальное понятие — строй знаковой последовательности*, который представляет только определенный порядок следования, расположение различимых и одинаковых его компонентов без учета их обозначений и содержимого. Заметим, что цели и методы исследования строя, если его рассматривать как обычный кортеж, не отличаются от исследований реальных текстов и знаковых цепей и тому подобных. Если рассматривать строй как новый абстрактный объект, отображающий информационную цепь, то открывается возможность исследовать и использовать его особые свойства, в том числе применять новые формулы для подсчета информации в массиве данных [32].

3. О мерах сходства-расхождения цепей на основе числовых характеристик строя. Первые разработки по применению развиваемого здесь подхода для сравнения построений цепей представлены в публикациях [44, 45]. Целью же данной статьи является представление разработанного авторами нового подхода и инструментария для сравнения

символьных последовательностей на основе характеристик строя их частей (подпоследовательностей).

Определим *меры сходства и расхождения* оригинальных построений разных знаковых последовательностей A и B на основе числовых характеристик строя отдельной цепи. В данной работе используются характеристики глубины (G) и средней удаленности (g). В частности, глубины расположения всех одинаковых знаков в однородной цепи и любых одинаковых символов в полной неоднородной цепи, соответственно, определяются в виде:

$$G_j = \sum_{i=1}^{n_j} \log \Delta_{ij} = \sum_{i=1}^{n_j} g_{ij}; \tag{1}$$

$$G = \sum_{j=1}^{m} G_j, \tag{2}$$

где Δ_{ij} — интервал от i-го до $(i+1)$ вхождения j-го символа в однородной цепи (например, в литературных текстах это могут быть интервалы между ближайшими одинаковыми словами или буквами, а в нуклеотидных последовательностях — это интервал между двумя ближайшими одинаковыми нуклеотидами — межнуклеотидное расстояние); g_{ij} — удаленность i-го вхождения j-го символа (текущая удаленность); n_j — число вхождений j-го символа (нуклеотида); m — мощность алфавита в последовательности ($m = 4$ — число различных нуклеотидов {A, T, C, G}).

В свою очередь, средние удаленности одинаковых символов в однородной цепи и полной неоднородной знаковой цепи определяются в виде:

$$g_j = G_j / n_j; \tag{3}$$

$$g = G/n, \tag{4}$$

где n — длина последовательности (гена, генома и тому подобного), определяемая числом компонентов (например, нуклеотидов).

С помощью компьютерных экспериментов было установлено, что числовые характеристики строя цепи с высокой точностью и

однозначно отображают оригинальное построение компонентов в данной последовательности [46].

Для приближенного сравнения при условии равенства мощностей алфавитов (словарей), когда $m_A = m_B = m$, сравниваемых цепей A и B, определим меры их расхождения в виде:

$$\Delta G_1 = |G_A - G_B|; \tag{5}$$

$$\Delta g_1 = |g_A - g_B|, \tag{6}$$

где G_A, G_B, g_A, g_B — соответственно глубины и средние удаленности полных цепей A и B. Формула (6) позволяет сравнивать последовательности разной длины.

Назовем отдельные части и так называемые однородные цепи, составляющие полную неоднородную знаковую цепь (кортеж) A, одинаково — частями (целого). Частями целого текста могут быть, например, его предложения, абзацы, параграфы и тому подобное. В качестве частей полного генома принято выделять гены, регуляторные зоны, рибосомальные и транспортные РНК и другие составляющие аннотации генома [47].

Однородная цепь представляет собой кортеж, алфавит которого составляют два компонента: выделенный элемент алфавита полной последовательности и «пустой» элемент. Выделенный элемент алфавита располагается на позиции так же, как и в полной последовательности, а все остальные места заняты «пустым» элементом. В качестве однородной цепи на основе выделенного слова в данном тексте выступает полная многоместная позиция данного текста, в которой «заняты» выделенным словом только отдельные места; остальные места на позиции пусты.

Для точного сравнения построений отдельных частей в разных цепях A и B необходимы следующие действия:

1. Получить два распределения значений интегральной характеристики строя наборов частей (составляющих A и B) — $\{G_{A_j}\}$ и $\{G_{B_k}\}$. В этих распределениях A_j и B_k соответственно j-ая и k-ая части сравниваемых цепей A и B; $j = 1, 2, \ldots, m_A,\ k = 1, 2, \ldots, m_B$; где m_A, m_B — количество частей в цепях A и B, выбранных исследователями для их сравнения.

2. Построить ранговые распределения значений характеристики строя для двух наборов частей A и B, в которых $G_{A_j} \leqslant G_{A_{j+1}}$ и $G_{B_K} \leqslant G_{B_{k+1}}$ (возможно построение убывающих распределений как на рисунке 8).

При совпадении количества частей в разных последовательностях, когда $m_A = m_B = m$, для более точного сравнения можно использовать меры расхождения построений цепей A и B в виде:

$$\Delta G_2 = \sum_{j=1}^{m} |G_{A_j} - G_{B_j}|; \tag{7}$$

$$\Delta g_2 = \sum_{j=1}^{m} |g_{A_j} - g_{B_j}|. \tag{8}$$

И, наконец, для детального сравнения построений цепей A и B, когда совпадают числа вхождения n_j j-ых компонентов у сравниваемых частей данных цепей, можно использовать поинтервальную меру расхождения, в которой сравниваются текущие удаленности в виде:

$$\Delta g_3 = \sum_{j=1}^{m} \sum_{i=1}^{n_j} |g_{A_{ji}} - g_{B_{ji}}|. \tag{9}$$

В случаях, когда размер наборов частей сравниваемых цепей A и B различается и $m_A \neq m_B$, предлагается использовать меры сходства и расхождения, в которых *приоритетно учитываются подмножества сходных по строю (с определенной точностью) отдельных частей, составляющих сравниваемые цепи* A *и* B ($\{A_i\} \subseteq \{A_j\}$ и $\{B_i\} \subseteq \{B_k\}$).

Для предварительного отбора сходных по расположению компонентов в каждой из сравниваемых частей предлагается использовать интегральные характеристики строя: G_{A_i}, G_{B_i}.

При этом сходной парой назовем сравниваемые части цепей A_i и B_i, для которых относительное расхождение значений интегральной характеристики их строя не превышает некоторую величину $\delta \leqslant 1$. Такой критерий сходства пар частей определим в виде:

$$\delta_{AB_i} = \frac{|G_{A_i} - G_{B_i}| \cdot 2}{G_{A_i} + G_{B_i}} \leqslant \delta, \tag{10}$$

где индекс i представляет номер пары совпадающих частей у сравниваемых цепей; при этом количества таких частей могут быть разными, когда $m_{A_i} \neq m_{B_i}$.

Определим среднее относительное отклонение по множеству совпадающих пар $\{\langle A_i, B_i\rangle\}$ в виде:

$$\delta_{AB} = \frac{1}{M_{AB}} \cdot \sum_{i=1}^{M_{AB}} \delta_{AB_i}, \tag{11}$$

где $M_{AB} = |\{\langle A_i, B_i\rangle\}|$ — мощность множества сходных по строю пар частей.

Кроме отмеченного в пунктах 1) и 2) для сравниваемых ранговых распределений значений характеристики строя отдельных частей $\{G_{A_j}\}$ и $\{G_{B_k}\}$ (у которых $m_A \neq m_B$) необходимо определить *сходные по строю части двух цепей* (в количестве $0 \leqslant m_{AB} \leqslant m_A + m_B$). *Для ранговых распределений сходные части можно выделить за небольшое число проходов (без полного перебора).*

Сходные по строю пары частей из состава цепей A и B определяются в два этапа:

– приближенным отбором с помощью критерия (10), в котором используются пары значений характеристики строя G_{A_i}, G_{B_i};

– дальнейшим разделением списка пар, выделенных на первом этапе, при котором пары попадают в одно из трех подмножеств: совпадающие по строю пары, сходные по строю (с малыми различиями строя), «псевдосходные» — сходные только по значениям характеристики G_{A_i}, G_{B_i}, но не являющиеся при этом гомологичными.

Для тонкой селекции сходных пар можно использовать инструменты ФОАС: локальные и однородные характеристики строя, а в случае нуклеотидных последовательностей — общепринятые имена частей (имена компонентов аннотации) [48]. В такой функции «аргументом» является начальная позиция фрагмента (L-граммы) в последовательности. Значения функции вычисляются «скользящим окном» для множества L-грамм — последовательных, равных по длине и смещенных на один элемент относительно друг друга участков цепи (в примере ниже длина L-грамм — $l = 50$ нуклеотидов). Фактически, функция характеристики строя позволяет поэлементно представлять рассматриваемую знаковую цепь. То есть для каждого компонента последовательности (кроме $l - 1$ конечных компонентов) может быть вычислено соответствующее ему значение «локальной» характеристики.

На практике тонкая селекция осуществлялась непосредственным рассмотрением пар графиков функций характеристик. Возможна автоматизация сравнения пар функций.

После выделения сходных по строю частей последовательностей A и B (при условии их наличия) предлагается использовать следующие меры их сходства и расхождения:

1. Относительное число (доля) сходных по строю частей.
2. Среднее относительное расхождение ранговых распределений.
3. Относительная глубина сходных по строю частей.

Первая мера определяется в виде:

$$\delta_1 = \frac{m_{AB}}{m_A + m_B} \leqslant 1, \tag{12}$$

где $m_{AB} = 2 \cdot m_{A_i}$, если $m_{A_i} < m_{B_i}$;
$m_{AB} = 2 \cdot m_{B_i}$, если $m_{A_i} > m_{B_i}$;
$m_{AB} = m_{A_i} + m_{B_i}$, если $m_{A_i} = m_{B_i}$;
$\delta_1 = 1$, если $m_{A_i} = m_A$ и $m_{B_i} = m_B$.

Вторая мера определяется в виде:

$$\delta_2 = \frac{\delta_{AB}}{\delta_1}. \tag{13}$$

С учетом первой меры видно, что уменьшение доли сходных по строю частей искусственно увеличивает расхождение ранговых распределений цепей A и B.

Третья мера представляет отношение суммарных значений характеристики строя сходных частей к сумме этих характеристик для всех составляющих частей сравниваемых цепей A и B, в виде:

$$\delta_3 = \frac{G_{AB}}{G_A + G_B}, \tag{14}$$

где $G_{AB} = 2 \cdot \sum_{i=1}^{m_{A_i}} G_{A_i}$, если $\sum_{i=1}^{m_{A_i}} G_{A_i} < \sum_{i=1}^{m_{B_i}} G_{B_i}$;

$G_{AB} = 2 \cdot \sum_{i=1}^{m_{B_i}} G_{B_i}$, если $\sum_{i=1}^{m_{A_i}} G_{A_i} > \sum_{i=1}^{m_{B_i}} G_{B_i}$;

$G_{AB} = \sum_{i=1}^{m_{A_i}} G_{A_i} + \sum_{i=1}^{m_{B_i}} G_{B_i}$, если $\sum_{i=1}^{m_{A_i}} G_{B_i} = \sum_{i=1}^{m_{B_i}} G_{B_i}$;

$\delta_3 = 1$, если $\sum_{i=1}^{m_{A_i}} G_{A_i} = G_A$ и $\sum_{i=1}^{m_{B_i}} G_{B_i} = G_B$.

4. О процедуре сравнения построений цепей на основе интегральных числовых характеристик строя. Как отмечалось, в качестве информационных цепей выступают массивы естественно упорядоченных данных разной природы: нуклеотидные последовательности (геномы, плазмиды, гены, и др.), лингвистические тексты, нотные записи, массивы данных измерений, в частности временные ряды. Определим процедуру сравнения построений двух наборов частей, представляющих знаковые последовательности A и B. Так как алгоритм (процедура) сравнения двух неубывающих ранговых распределений (за малое число проходов) очевиден для любого специалиста в области информатики, отметим только некоторые его детали:

1. Для пары сравниваемых последовательностей выделяется набор исследуемых частей. Например, для геномов можно использовать не все компоненты из аннотаций.

2. Выбирается характеристика для сравнения частей.

3. Вычисляется отображение всех выбранных частей каждой целостной знаковой цепи (генома, текста и тому подобных) соответствующими значениями характеристики.

4. Части каждой последовательности упорядочиваются по возрастанию (убыванию) вычисленной характеристики. В результате получаются два неубывающих (невозрастающих) ранговых распределения значений этой характеристики (рисунок 8).

5.

В получившихся распределениях производится поиск совпадающих с заданной точностью частей. Особенностью сравнения двух ранговых распределений является необходимость сопоставления очередного элемента данного распределения, в том числе с теми элементами другого распределения, которые выявлены как сходные для предыдущего элемента данного распределения. При этом предполагается, что как в первом, так и во втором распределениях возможны «последовательности» одинаковых по характеристике частей.

6. При нахождении совпадающей части из второй цепи сохраняем данную пару совпадающих (или схожих) частей.

7. Кроме перебора отдельных частей первой цепи на предмет сходства с частями второй цепи, необходимо выполнить данную процедуру для выявления сходства частей второй цепи относительно первой.

8. После выявления всех пар сходных частей первой и второй цепи, вычисляются значения следующих характеристик:

– среднее относительное отклонение для множества сходных пар по формуле (11);

– мера сходства пары сравниваемых цепей по формуле (12), определяемая долей сходных по строю частей в каждой из цепей по отношению к общему числу составляющих их частей;

– мера расхождения данной пары цепей по формуле (13), в которой искусственно завышается среднее относительное отклонение сходных по строю частей за счет доли несовпадающих частей;

– мера сходства по формуле (14), определяемая долей суммы значений характеристики сходных по строю частей.

5. Матрица сходства-расхождения для набора цепей. На практике возникает необходимость попарного сравнения построений некоторого множества цепей. Это могут быть геномы наборов организмов (возможно сходных), оригинальный текст и множество его изложений, множество текстов с одинаковым содержанием, записанных на разных языках.

Результат такого сравнения может быть представлен в виде квадратной матрицы. Ячейки такой матрицы могут содержать как значения всех мер сходства-расхождения, вычисленных по формулам (12), (13), (14), так и только некоторые из них (рисунок 12). Первая строка матрицы формируется как результат сравнения первой цепи со всеми другими цепями. Затем определяется сходство второй цепи со всеми рассматриваемыми цепями и так далее. При этом каждой ячейке матрицы соответствует множество сходных пар частей сравниваемых цепей.

6. Примеры сравнения плазмид разных штаммов бактерии *Coxiella burnetii* на основе числовых характеристик строя их частей (компонентов аннотаций). Для апробации разработанного инструментария использовались знаковые (нуклеотидные) последовательности плазмид (и их аннотации) семи штаммов *Coxiella burnetii*, взятых из GenBank [49]. Их нумерованный список представлен под рисунком 12.

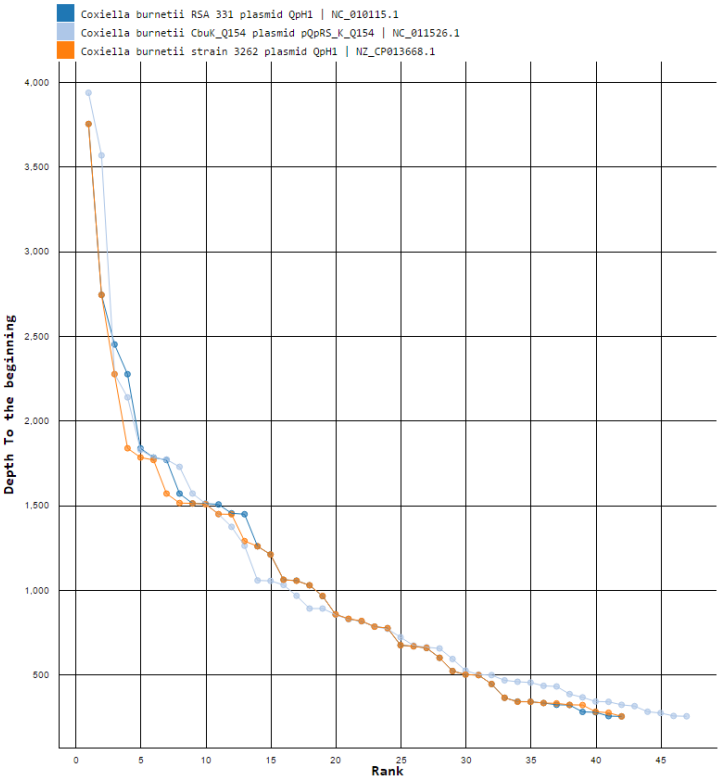

Рис. 8. Ранговые распределения значений характеристики G_j генов и других компонентов аннотаций плазмид штаммов *Coxiella burnetii* с номерами 2, 3, 5

На рисунке 8 для примера представлены ранговые распределения значений характеристики строя G_j сопоставляемых частей (генов и других компонентов аннотаций) плазмид трех микроорганизмов из этого списка. Два распределения представляют 2-ю и 3-ю плазмиды сходные на 76,19% по мере (12). Третье распределение представляет 5-ю плазмиду. По этой мере оно имеет меньшее сходство с ними (2-я с 5-ой на 29,21%, 3-я с 5-ой 24,72%).

Ниже списком представлены некоторые сходные по характеристике G_j пары генов и степень их сходства по мере (10) с порогом $\delta \leqslant 0.1$, вычисленные для пары плазмид штаммов *Coxiella burnetii* с номерами 1 и 3, которые сравнивались по мере (12). На рисунках 9, 10, 11 показаны пары графиков, представляющих функции характеристик строя

сопоставляемых генов.

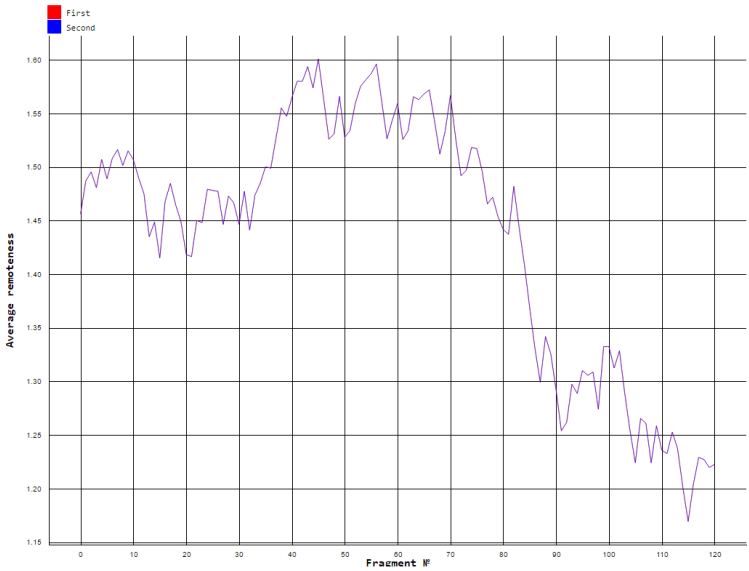

Рис. 9. Графики двух функций характеристики g совпадающих по строю генов 1-ой и 3-ей плазмид (графики полностью совпадают)

Совпадающие по строю гены:
First sequence: Coxiella burnetii RSA 493 plasmid pQpH1 | NC_004704.1
Feature: Coding DNA sequence. Position = 12829. Length = 171.
Product: hypothetical protein. Characteristic value = 252.39884557174787.

Second sequence: Coxiella burnetii RSA 331 plasmid QpH1 | NC_010115.1
Feature: Coding DNA sequence. Position = 1863. Length = 171.
Product: tyrosine recombinase. Characteristic value = 252.39884557174787.

Абсолютное расхождение — 0. Относительное расхождение — 0%.

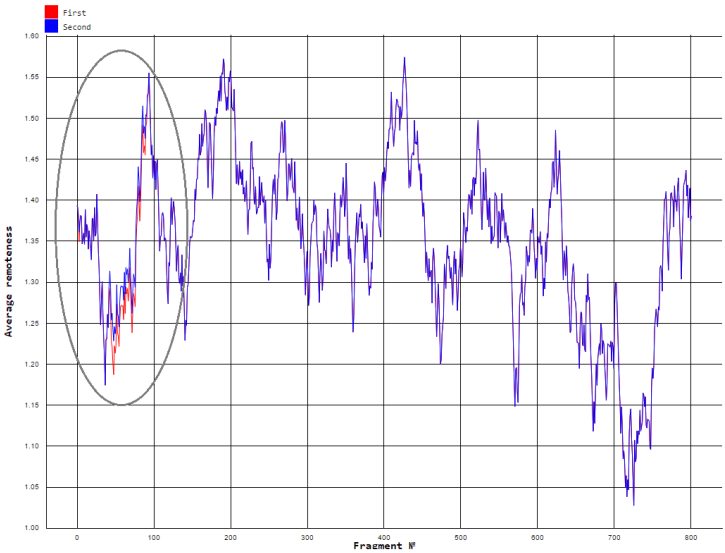

Рис. 10. Графики двух функций характеристики g сходных по строю генов 1-ой и 3-ей плазмид (отмечено место расхождения)

Сходные по строю гены:
First sequence: Coxiella burnetii RSA 493 plasmid pQpH1 | NC_004704.1
Feature: Coding DNA sequence. Position = 19735. Length = 852 Complement.
Product: hypothetical protein. Characteristic value = 1207.4344521538703.

Second sequence: Coxiella burnetii RSA 331 plasmid QpH1 | NC_010115.1
Feature: Coding DNA sequence. Position = 8763. Length = 852. Complement.
Product: hypothetical protein. Characteristic value = 1208.6188767250078.

Абсолютное расхождение — 1.18442457113. Относительное расхождение — 0.098046227455%.

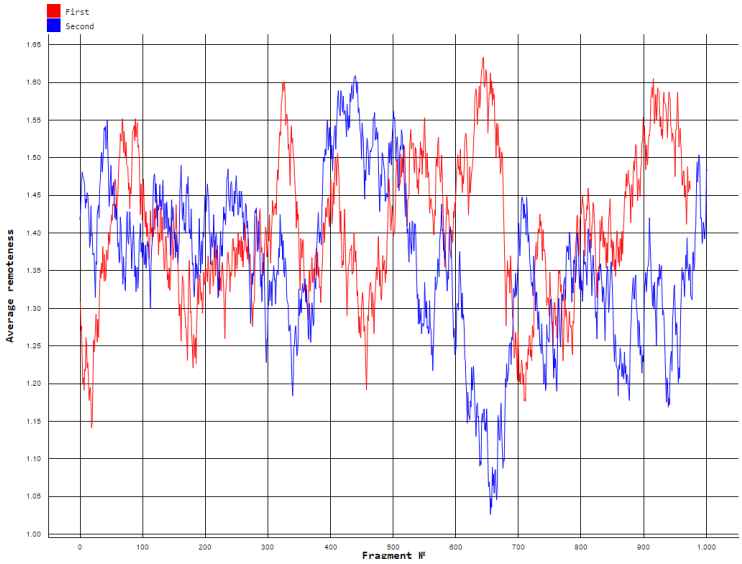

Рис. 11. Графики двух функций характеристики g псевдосходных генов 1-ой и 3-ей плазмид

Псевдосходные гены:
First sequence: Coxiella burnetii RSA 493 plasmid pQpH1 | NC_004704.1
Feature: Coding DNA sequence. Position = 2545. Length = 1026
Product: hypothetical protein. Characteristic value = 1510.194562707094.

Second sequence: Coxiella burnetii RSA 331 plasmid QpH1 | NC_010115.1
Feature: Coding DNA sequence. Position = 9641. Length = 1053. Complement.
Product: hypothetical protein. Characteristic value = 1509.2896095025235.

Абсолютное расхождение — 0.90495320457. Относительное расхождение — 0.059940913941%.

Для демонстрации значений трех мер сходства-расхождения плазмид ниже представлена матрица для всех семи микроорганизмов списка. Рассматриваемые микроорганизмы предварительно упорядочены и пронумерованы на основе одной интегральной характеристики строя — средней удаленности g, вычисленной для каждой плазмиды.

На рисунке 12 по горизонтали и вертикали соответствующими номерами обозначены последовательности:

1 - Coxiella burnetii RSA 493 plasmid pQpH1 | NC_004704.1
2 - Coxiella burnetii strain 3262 plasmid QpH1 | NZ_CP013668.1
3 - Coxiella burnetii RSA 331 plasmid QpH1 | NC_010115.1
4 - Coxiella burnetii str. Namibia plasmid QpRS | NZ_CP007556.1
5 - Coxiella burnetii CbuK_Q154 plasmid pQpRS_K_Q154 | NC_011526.1
6 - Coxiella burnetii MSU Goat Q177 plasmid QpRS | NC_010258.1
7 - Coxiella burnetii Dugway 5J108-111 plasmid pQpDG | NC_009726.1

Рис. 12. Матрица сходства-расхождения плазмид семи штаммов *Coxiella burnetii*

Для проведения исследований сходства цепей разработан комплекс программных средств [50], который позволил построить матрицу сходства для 42 полных геномов организмов семейства Rickettsia.

Кроме отмеченного, компьютерные эксперименты были направлены на исследование однозначности отображения строя конкретной информационной цепи соответствующим значением характеристики строя.

Результаты исследования нуклеотидных последовательностей с использованием характеристики G_j позволили сформулировать следующие предварительные выводы:

– сходные по строю пары частей (компонентов аннотаций) можно предварительно отбирать из состава двух хромосом или плазмид по критерию (10) на основе сравнения значений их интегральных характеристик G_{A_i} и G_{B_i}, различающихся на величину $\delta \leqslant 0,05\%$;

– однако даже при таком малом пороге в списке сходных по строю частей, кроме полностью совпадающих и сходных (с малыми различиями) по строю частей, попадают псевдосходные пары частей, для которых близки только величины G_{A_i} и G_{B_i} (но их построения сильно различаются, как на рисунке 11);

– для формального отсева из списка большинства псевдосходных частей следует подбирать величину порога $\delta < 0,05\%$ для данной пары сравниваемых организмов или некоторого их множества;

– рассмотрение множества пар графиков, представляющих последовательности значений локальной характеристики строя (функции характеристик строя), выявляет, во-первых, некоторое фиксированное подмножество пар частей (в данном исследовании — компонентов аннотаций), строй которых полностью совпадает, если интегральные характеристики равны (G_{A_i} и G_{B_i}); во-вторых, некоторое фиксированное подмножество сходных по строю (с малыми различиями) пар частей; в-третьих, увеличивающееся по мощности (по мере увеличения порога δ) подмножество псевдосходных пар частей, которое частично перемешивается с парами сходных частей по расхождению характеристики;

– в большинстве случаев оригинальное расположение нуклеотидов (компонентов) в данной цепи однозначно отображается уникальной цифровой последовательностью — одним числом интегральной характеристики строя. Однако изредка при равенстве характеристик $G_{A_i} = G_{B_i}$ в принципе могут фиксироваться псевдосовпавшие цепи. Такие совпадения объясняются отличиями в некоторых одинаковых местах цепей, которые не изменяют наборы (межнуклеотидных) интервалов в сравниваемых цепях;

– зачастую несовпадение между частями геномов или плазмид обусловлено низким качеством аннотаций, в частности отсутствием жесткой регламентации (стандартов) при их составлении, не точно или неполно указанной позицией компонента, не полным или некорректным его названием и тому подобным [51].

7. Заключение. В работе отмечено фактическое отсутствие адекватных средств прикладной математики и информатики для исчисления оригинального расположения знаков в символьных последовательностях.

Указан недостаток традиционного методологического подхода — редукционизма, применяемого для описания и анализа массивов естественно упорядоченных данных. Предложено подобные исследования дополнить средствами системного подхода.

Предложены меры сходства и расхождения знаковых цепей на основе числовых характеристик строя.

Представлена процедура сравнения построений цепей на основе ранговых распределений значений числовой характеристики строя, отображающих отдельные части этих цепей. Данная процедура допускает сравнение цепей, мощности алфавитов (словарей) которых могут быть различны.

Сформулированы и рассмотрены возможные исходы при попарном сравнении частей символьных последовательностей по характеристикам строя: полное совпадение; сходство по характеристике с заданной точностью; псевдосходство по заданной характеристике.

Для демонстрации возможностей предложенных мер сходства и процедуры сравнения представлены результаты сравнения плазмид семи штаммов бактерии *Coxiella burnetii* в виде матрицы сходства на основе числовых характеристик строя их частей (генов и других компонентов аннотаций).

Литература

1. *Zipf G., Kingsley G.* Selected Studies of the Principle of Relative Frequency in Language // Harvard University Press. 1932. 128 p.
2. *Гусев В.Д., Косарев Ю.Г., Титкова Т.Н.* Методы поиска и анализ статистических закономерностей в символьных последовательностях // Машинные методы обнаружения закономерностей: материалы всесоюзного симпозиума. 1976. С. 75–84.
3. *Гусев В.Д., Куличков В.А., Никулин А.Е.* Алгоритмы поиска несовершенных повторов в генетических текстах // Анализ символьных последовательностей: вычислительные системы. 1985. Вып. 113. С. 107–122.
4. *Гусев В.Д., Немытикова Л.А.* Векторная мера сложности нуклеотидных последовательностей // Третий сибирский конгресс по прикладной и индустриальной математике (ИНПРИМ-98). 1998. 115 с.
5. *Гусев В.Д., Мирошниченко Л.А., Саломатина Н.В.* Методы выделения структурных единиц в символьных последовательностях. Межъязыковые аналоги // Материалы Всероссийской конференции с международным участием «Знания-Онтологии-Теории». 2009. Т. 2. С. 53–62.
6. *Беликов С.И., Гусев В.Д., Мирошниченко Л.А., Титкова Т.Н.* Сравнительный анализ геномов вирусов клещевого энцефалита: дифференциация по степени вирулентности // Математическая биология и биоинформатика: IV международная конференция. 2012. С. 52–53.
7. *King B.R, Aburdene M., Thompson A., Warres Z.* Application of Discrete Fourier Inter-Coefficient Difference for Assessing Genetic Sequence Similarity // EURASIP Journal on Bioinformatics and Systems Biology. 2014. vol. 2014, no. 1. 8 p.
8. *Srivastava S., Baptista M.S.* Markovian language model of the DNA and its information content // Royal Society open science. 2016. vol. 3. no. 1. pp. 150527.
9. *Nair A.S.S., Mahalakshmi T.* Visualization of genomic data using inter-nucleotide distance signals // Proceedings of IEEE Genomic Signal Processing. 2005. vol. 408.
10. *Afreixo V. et al.* Genome analysis with inter-nucleotide distances // Bioinformatics. 2009. vol. 25(23). pp. 3064-3070.

11. *Jin S. et al.* A Generalized Topological Entropy for Analyzing the Complexity of DNA Sequences // PLoS One. 2014. vol. 9(2). pp. e88519.
12. *Садовский М.Г.* Информационно-статистический анализ нуклеотидных последовательностей: диссертация // Институт биофизики СО РАН. 2004. 394 с.
13. *Amiri S., Dinov I.D.* Comparison of genomic data via statistical distribution // Journal of Theoretical Biology. 2016. vol. 407. pp. 318–327.
14. *Manca V., Bonnici V.* Infogenomics Tools: A Computational Suite for Informational Analyses of Genomes // Journal of Bioinformatics, Proteomics and Imaging Analysis. 2015. vol. 1. no. 1. pp. 7-14.
15. *Арнольд В.И.* Сложность конечных последовательностей нулей и единиц и геометрия конечных функциональных пространств // Публичная лекция. 2006. Т. 13. 14 р.
16. *Kullback S., Leibler R.A.* On information and sufficiency // The Annals of Mathematical Statistics. 1951. vol. 22. no. 1. pp. 79–86.
17. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. 1965. Т. 4. С. 845–848.
18. *Hamming R. W.* Error detecting and error correcting codes // Bell System Technical Journal. 1950. vol. 29(2). pp. 147–160.
19. *Zielezinski A., Vinga S., Almeida J., Karlowski W.M.* Alignment-free sequence comparison: benefits, applications, and tools // Genome Biology. 2017, vol. 18(1):186 p.
20. *Bonham-Carter O, Steele J., Bastola D.* Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis // Briefings in Bioinformatics. 2014. vol. 15(6). pp. 890-905.
21. *Song K. et al.* New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing // Briefings in Bioinformatics. 2014. vol. 15(3). pp. 343-353.
22. *Bernard G. et al.* Alignment-free inference of hierarchical and reticulate phylogenomic relationships // Briefings in Bioinformatics. 2017.
23. *Chan C.X., Ragan M.A.* Next-generation phylogenomics // Biology Direct. 2013. vol. 8(1). pp. 3.
24. *La Rosa M., Fiannaca A., Rizzo R., Urso A.* Alignment-free analysis of barcode sequences by means of compression-based methods // BMC Bioinformatics. 2013. vol. 14(7). pp. S4.
25. *Haubold B.* Alignment-free phylogenetics and population genetics // Briefings in Bioinformatics. 2013. vol. 15(3). pp. 407–418.
26. *Ren J. et al.* Alignment-Free Sequence Analysis and Applications // Annual Review of Biomedical Data Science. 2018. vol. 1. pp. 93–114.
27. *Wang S., Tian F., Feng W., Liu X.* Applications of representation method for DNA sequences based on symbolic dynamics // Journal of Molecular Structure: THEOCHEM. 2009. vol. 909. no. 1-3. pp. 33-42.
28. *Salgado-Garcia R., Ugalde E.* Symbolic Complexity for Nucleotide Sequences: A Sign of the Genome Structure // Journal of Physics A: Mathematical and Theoretical. 2016. vol. 49. no. 44. pp. 445601.
29. *Шрейдер Ю.А., Шаров А.А.* Системы и модели // М.: Радио и связь. 1982. 152 с.
30. *Мазур М.* Качественная теория информации // М.: Мир. 1974. 240 с.
31. *Gumenjuk A., Kostyshin A., Simonova S.* An approach to the research of the structure of linguistic and musical texts // Glottometrics. 2002. vol. 3. pp. 61–89.

32. *Гуменюк А. С., Поздниченко Н. Н., Родионов И. Н., Шпынов С.Н.* О средствах формального анализа строя нуклеотидных цепей // Математическая биология и биоинформатика. 2013. Т. 8. № 1. С. 373-397.
33. *Freitas A., Afreixo V., Cruz S.E.* Mixture models of geometric distributions in genomic analysis of inter-nucleotide distances // Statistics, optimization & information computing Stat. 2013. Vol. 1. no. 1. pp. 8–28.
34. *Wasito I., Veritawati I.* Fractal Dimension Approach for Clustering of DNA Sequences Based on Internucleotide Distance // IEEE 2013 International Conference of Information and Communication Technology (ICoICT). 2013. pp. 82–87.
35. *Tavares A. et al.* Detection of exceptional genomic words: a comparison between species // 22nd International Conference on Computational Statistics (COMPSTAT 2016). 2016.
36. *Zhou L.Q., Li R., Han G.S.* A Method Based on the Improved Inter-Nucleotide Distances of Genomes to Construct Vertebrates Phylogeny Tree // IEEE 2014 7th International Conference on Biomedical Engineering and Informatics. 2014. pp. 776-780.
37. *Kolekar P., Kale M., Kulkarni-Kale U.* Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping // Molecular Phylogenetics and Evolution. 2012. vol. 65. no. 2. pp. 510–522.
38. *Bonnici V., Manca V.* Recurrence Distance Distributions in Computational Genomics // American Journal of Bioinformatics and Computational Biology. 2015. vol. 3. pp. 5-23.
39. *Messaoudi I., Oueslati A.E., Lachiri Z.* Wavelet analysis of frequency chaos game signal: a time-frequency signature of the C. elegans DNA // EURASIP Journal on Bioinformatics and Systems Biology. 2014. vol. 2014(1). pp. 16.
40. *Орлов Ю.К.* Частотные структуры конечных сообщений в некоторых естественных информационных системах: диссертация // Издательство Тбилисского университета. 1974.
41. *Орлов Ю.К.* Невидимая гармония // Число и мысль. 1980. Вып. 3. С. 70-105.
42. *Кудрин Б.И.* Философия техники: основания постнеклассической философии техники // М.: Техника. 2007. Вып. 36. 196 с.
43. *Попова О.В., Гельфанд М.С.* Существует ли аналог закона Ципфа в генетическом языке? // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2000. № 4. С. 19-24.
44. *Волчкова И.А., Гуменюк А.С.* О мерах сходства разноязычных текстов с одинаковым содержанием. // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ-13). 2013. Т. 1. С. 98-105.
45. *Гуменюк А.С., Волчкова И.А.* Использование средств анализа строя знаковой последовательности для формальной оценки качества перевода. // Омский научный вестник. 2013. Т. 3(123). С. 251-256.
46. *Шпынов С.Н., Гуменюк А.С., Поздниченко Н.Н.* Применение числовой характеристики строя нуклеотидов в геномах прокариот для реклассификации внутри рода Rickettsia // Математическая биология и биоинформатика. 2016. Т. 11. № 2. С. 336-350.
47. The DDBJ/ENA/GenBank Feature Table Definition. URL: http://www.insdc.org/files/feature_table.html (дата обращения: 15.04.2018).
48. *Гуменюк А.С., Поздниченко Н.Н., Шпынов С.Н.* Формальный анализ строя локальной структуры нуклеотидных последовательностей // Вестник Томского государственного университета. 2014. Т. 4(29). С. 23–30.
49. GENBANK DataBase. URL: http://www.ncbi.nlm.nih.gov/nuccore/ (дата обращения: 02.03.2018).

50. *Гуменюк А.С., Поздниченко Н.Н., Скиба А.А., Шпынов С.Н.* Матрица сходства нуклеотидных последовательностей по их компонентам. Свидетельство о государственной регистрации программы для ЭВМ. №2017616679. 09.06.2017.

51. *Поздниченко Н.Н., Гуменюк А.С., Шпынов С.Н.* О картографическом представлении множества геномов прокариот с помощью числовых характеристик строя их компонентов. // Новые информационные технологии в исследовании сложных структур: материалы 11-й международной конференции. 2016. С. 84-85.

Гуменюк Александр Степанович — канд. техн. наук, доцент, доцент кафедры информатики и вычислительной техники факультета информационных технологий и компьютерных систем, Омский государственный технический университет (ОмГТУ). Область научных интересов: формальный анализ строя компонентов в массивах упорядоченных данных, формальный анализ строя литературных текстов, нотных записей и нуклеотидных последовательностей, социальные системы информационных коммуникаций, формализация типов общения, теория и организация обучения студентов на основе делового общения, алгебра ментальных событий, физические основания передачи сообщений. Число научных публикаций — около 200. gumas45@mail.ru; пр. Мира, 11, 644050, Омск, Российская Федерация; р.т.: +7(3812) 65-24-98.

Скиба Артемий Андреевич — инженер-программист, ООО "Компания Элмис". Область научных интересов: формальный анализ строя, формальный анализ строя нотных записей и нуклеотидных последовательностей, биоинформатика. Число научных публикаций — 13. skiba.artem@inbox.ru; Маяковского, 14, 644046, Омск, Российская Федерация;+7(3812) 51-06-30.

Поздниченко Николай Николаевич — старший преподаватель, кафедра информатики и вычислительной техники факультета информационных технологий и компьютерных систем, Омский государственный технический университет (ОмГТУ). Область научных интересов: формальный анализ строя, формальный анализ строя нуклеотидных последовательностей, биоинформатика. Число научных публикаций — 45. nick670@yandex.ru; пр. Мира, 11, 644050, Омск, Российская Федерация; р.т.: +7(3812) 65-24-98.

Шпынов Станислав Николаевич — д-р мед. наук, заведующий лабораторией экологии риккетсий, лаборатория экологии риккетсий, Федеральное государственное бюджетное учреждение «Федеральный научно-исследовательский центр эпидемиологии и микробиологии имени почетного академика Н.Ф. Гамалеи». Область научных интересов: микробиология, экология, эпидемиология, молекулярная биология, биоинформатика. Число научных публикаций — около 200. stan63@inbox.ru; Н.Ф. Гамалеи, 18, 123098, Москва, Российская Федерация; р.т.: +7(499) 193-61-85.

DOI 10.15622/sp.18.2.471-503

A.S. GUMENYUK, A.A. SKIBA, N.N. POZDNICHENKO, S.N. SHPYNOV

ABOUT SIMILARITY MEASURES OF COMPONENTS ARRANGEMENT OF NATURALLY ORDERED DATA ARRAYS

Gumenyuk A.S., Skiba A.A., Pozdnichenko N.N., Shpynov S.N. **About Similarity Measures of Components Arrangement of Naturally Ordered Data Arrays.**

Abstract. At present, mathematical tools that adequately take into account the arrangement of components are not widespread in works of specialists in the fields of research of naturally ordered data of different nature. Therefore, it is difficult or impossible to measure and compare the order of messages allocated in long information chains. The main approaches for comparing symbol sequences are using probabilistic models and statistical tools, pairwise and multiple alignment, which makes it possible to determine the degree of similarity of sequences using edit distance measures. The noted approaches almost do not pay attention to the study and detection of the patterns of the specific arrangement of all symbols, words, and components of data sets that constitute a separate sequence. The object of study in our works is a specifically organized numerical tuple — the arrangement of components (order) in symbolic or numerical sequence. The intervals between the closest identical components of the order are used as the basis for the quantitative representation of the chain arrangement. Multiplying all the intervals or summing their logarithms allows one to get numbers that uniquely reflect the arrangement of components in a particular sequence. These numbers, allow us to obtain a whole set of normalized characteristics of the order, among which the geometric mean interval and its logarithm. In this paper, we present an approach for quantitative comparing the arrangement of arrays of naturally ordered data (information chains) of an arbitrary nature. The measures of similarity/distinction and procedure of comparison of the chain order, based on the selection of a list of equal and similar by the order characteristics of the subsequences, are proposed. Rank distributions are used for faster selection of a list of matching components. The paper presents a toolkit for comparing the order of information chains and demonstrates some of its applications for studying the structure of nucleotide sequences.

Keywords: data array, symbolic sequence, information chain, numeric characteristics of order, depth of order, average remoteness, nucleotide sequence, similarity measures, similarity matrix, alignment-free genome comparison, inter-nucleotide distance.

Gumenyuk Alexander Stepanovich — Ph.D., Associate Professor, Associate Professor of Informatics and computer technology Department of Information Technologies and Computer Systems Faculty, Omsk State Technical University (OmSTU). Research interests: formal order analysis of components in ordered data arrays, formal order analysis of literary works, musical scores and nucleotide sequences, social systems of information communications, formalization of communication types, theory and organization of student learning based on professional communication, mental event algebra, physical foundation of messages transmission. The number of publications — about 200. gumas45@mail.ru; 11, pr. Mira, 644050, Omsk, Russian Federation; office phone: +7(3812) 65-24-98.

Skiba Artemiy Andreevich — Software Developer, Company Elmis. Research interests: formal order analysis, formal order analysis of musical scores and nucleotide sequences, bioinformatics. The number of publications — 13. skiba.artem@inbox.ru; 14, Mayakovskogo, 644046, Omsk, Russian Federation; office phone: +7(3812) 51-06-30.

Pozdnichenko Nikolay Nikolaevich — Senior Lecturer, Informatics and computer technology Department of Information Technologies and Computer Systems Faculty, Omsk State Technical University (OmSTU). Research interests: formal order analysis, formal order analysis of nucleotide sequences, bioinformatics. The number of publications — 45. nick670@yandex.ru; 11, pr. Mira, 644050, Omsk, Russian Federation; office phone: +7(3812) 65-24-98.

Shpynov Stanislav Nikolaevich — Ph.D., Dr.Sci., Head of Laboratory, Laboratory of Rickettsia Ecology, N. F. Gamaleya Federal Research Center for Epidemiology & Microbiology. Research interests: microbiology, ecology, epidemiology, molecular biology, bioinformatics. The number of publications — about 200. stan63@inbox.ru; 18, N.F. Gamaleya, 123098, Moscow, Russian Federation; office phone: +7(499) 193-61-85.

References

1. *Zipf G., Kingsley G.* Selected Studies of the Principle of Relative Frequency in Language. Harvard University Press. 1932. 128 p.
2. Gusev V.D., Kosarev Y.G., Titkova T.N. [Methods of search and analysis of statistical regularities in character sequences]. *Mashinnye metody obnaruzheniya zakonomernostej (Materialy vsesoyuznogo simpoziuma)* [Machine methods for detecting regularities: Materials of the All-Union Symposium]. 1976. pp. 75–84. (In Russ.).
3. Gusev V.D., Kulichkov V.A., Nikulin A.E. [Algorithms for searching for imperfect repetitions in genetic texts]. *Analiz simvol'nyh posledovatel'nostej: Vychislitel'nye sistemy – Analysis of symbol sequences: Computing systems.* 1985. vol. 113. pp. 107–122. (In Russ.).
4. Gusev V.D., Nemytikova L.A. [Vector measure of complexity of nucleotide sequences]. *Tretiy sibirskiy kongress po prikladnoy i industrial'noy matematike (INPRIM-98)* [The Third Siberian Congress on Applied and Industrial Mathematics (APLINM-98)]. 1998. 115 p. (In Russ.).
5. Gusev V.D., Miroshnichenko L.A., Salomatina N.V. [Methods for allocating structural units in character sequences. Interlingual analogues]. *Materialy Vserossiyskoy konferencii s mezhdunarodnym uchastiem «Znaniya-Ontologii-Teorii»* [Materials of the All-Russian Conference with International Participation «Knowledge – Ontologies – Theories»]. 2009. Issue 2. pp. 53–62. (In Russ.).
6. Belikov S.I., Gusev V.D., Miroshnichenko L.A., Titkova T.N. [Comparative analysis of genomes of tick-borne encephalitis viruses: differentiation according to the degree of virulence]. *Matematicheskaya biologiya i bioinformatika: IV mezhdunarodnaya konferenciya* [Mathematical biology and bioinformatics: IV international conference]. 2012. pp. 52–53. (In Russ.).
7. King B.R., Aburdene M., Thompson A., Warres Z. Application of Discrete Fourier Inter-Coefficient Difference for Assessing Genetic Sequence Similarity. *EURASIP Journal on Bioinformatics and Systems Biology.* 2014. vol. 2014. no. 1. 8 p.
8. Srivastava S., Baptista M.S. Markovian language model of the DNA and its information content. *Royal Society open science.* 2016. vol. 3. no. 1. pp. 150527.
9. Nair A.S.S., Mahalakshmi T. Visualization of genomic data using inter-nucleotide distance signals. Processings of IEEE Genomic Signal Processing. 2005. vol. 408.
10. Afreixo V. et al. Genome analysis with inter-nucleotide distances. *Bioinformatics*. 2009. vol. 25(23). pp. 3064–3070.
11. Jin S. et al. A Generalized Topological Entropy for Analyzing the Complexity of DNA Sequences. *PLoS One.* 2014. vol. 9(2). pp. e88519.
12. Sadovskii M.G. *Informacionno-statisticheskiy analiz nukleotidnyh posledovatel'nostey: dissertaciya* [Informational-statistical analysis of nucleotide sequences: Ph.D. Thesis]. Institute of Biophysics SB RAS. 2004. 394 p. (In Russ.).

13. Amiri S., Dinov I.D. Comparison of genomic data via statistical distribution. *Journal of Theoretical Biology.* 2016. vol. 407. pp. 318–327.
14. Manca V., Bonnici V. Infogenomics Tools: A Computational Suite for Informational Analyses of Genomes. *Journal of Bioinformatics, Proteomics and Imaging Analysis.* 2015. vol. 1. no. 1. pp. 7-14.
15. Arnold V.I. [The complexity of finite sequences of zeros and ones and the geometry of finite function spaces]. *Publichnaya lekciya* [Public lecture]. 2006. vol. 13. 14 p.
16. Kullback S., Leibler R.A. On information and sufficiency. *The Annals of Mathematical Statistics*. 1951. vol. 22. no. 1. pp. 79–86.
17. Levenshtein V.I. [Binary codes capable of correcting deletions, insertions, and reversals]. *Doklady Akademiy Nauk SSSR - Soviet Physics Reports*. 1965. vol. 4. pp. 845-848. (In Russ.).
18. Hamming R.W. Error detecting and error correcting codes. *Bell System Technical Journal.* 1950. vol. 29(2). pp. 147–160.
19. Zielezinski A., Vinga S., Almeida J., Karlowski W.M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology.* 2017. vol. 18(1). 186 p.
20. Bonham-Carter O., Steele J., Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in Bioinformatics.* 2014. vol. 15(6). pp. 890–905.
21. Song K. et al. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in Bioinformatics.* 2014. vol. 15(3). pp. 343–353.
22. Bernard G. et al. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings in Bioinformatics.* 2017.
23. Chan C.X., Ragan M.A. Next-generation phylogenomics. *Biology Direct.* 2013. vol. 8(1). pp. 3.
24. La Rosa M., Fiannaca A., Rizzo R., Urso A. Alignment-free analysis of barcode sequences by means of compression-based methods. *BMC Bioinformatics.* 2013. vol. 14(7). pp. S4.
25. Haubold B. Alignment-free phylogenetics and population genetics. *Briefings in Bioinformatics.* 2013. vol. 15(3). pp. 407–418.
26. Ren J. et al. Alignment-Free Sequence Analysis and Applications. *Annual Review of Biomedical Data Science.* 2018. vol. 1. pp. 93–114.
27. Wang S., Tian F., Feng W., Liu X. Applications of representation method for DNA sequences based on symbolic dynamics. *Journal of Molecular Structure: THEOCHEM.* 2009. vol. 909. no. 1-3. pp. 33–42.
28. Salgado-Garcia R., Ugalde E. Symbolic Complexity for Nucleotide Sequences: A Sign of the Genome Structure. *Journal of Physics A: Mathematical and Theoretical.* 2016. vol. 49. no. 44. pp. 445601.
29. Shreider Y.A., Sharov A.A. *Sistemy i modeli* [Systems and models]. – M.: Radio i svyaz. 1982. 152 p. (In Russ.).
30. Mazur M. *Kachestvennaya teoriya informacii* [Qualitative information theory]. M.: Mir. 1974. 240 p. (In Russ.).
31. Gumenjuk A., Kostyshin A., Simonova S. An approach to the research of the structure of linguistic and musical texts. *Glottometrics.* 2002. vol. 3. pp. 61–89.
32. Gumenuk A.S., Pozdnichenko N.N., Rodionov I.N., Shpynov S.N. [Formal Analysis of Structures of Nucleotide Chains]. *Matematicheskaya biologiya i bioinformatika - Mathematical biology and bioinformatics.* 2013. Issue 8. vol. 1. pp. 373–397. (In Russ.).
33. Freitas A., Afreixo V., Cruz S.E. Mixture models of geometric distributions in genomic analysis of inter-nucleotide distances. *Statistics, Optimization & Information Computing Stat.* 2013. vol. 1. no. 1. pp. 8–28.

34. Wasito I., Veritawati I. Fractal Dimension Approach for Clustering of DNA Sequences Based on Internucleotide Distance. IEEE 2013 International Conference of Information and Communication Technology (ICoICT). 2013. pp. 82–87.
35. Tavares A. et al. Detection of exceptional genomic words: a comparison between species. 22nd International Conference on Computational Statistics (COMPSTAT 2016). 2016.
36. Zhou L.Q., Li R., Han G.S. A Method Based on the Improved Inter-Nucleotide Distances of Genomes to Construct Vertebrates Phylogeny Tree. IEEE 2014 7th International Conference on Biomedical Engineering and Informatics. 2014. pp. 776–780.
37. Kolekar P., Kale M., Kulkarni-Kale U. Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping. *Molecular Phylogenetics and Evolution.* 2012. vol. 65. no. 2. pp. 510–522.
38. Bonnici V., Manca V. Recurrence Distance Distributions in Computational Genomics. *American Journal of Bioinformatics and Computational Biology.* 2015. vol. 3. pp. 5–23.
39. Messaoudi I., Oueslati A.E., Lachiri Z. Wavelet analysis of frequency chaos game signal: a time-frequency signature of the C. elegans DNA. *EURASIP Journal on Bioinformatics and Systems Biology.* 2014. vol. 2014(1). pp. 16.
40. Orlov Y.K. *Chastotnye struktury konechnyh soobshcheniy v nekotoryh estestvennyh informacionnyh sistemah: dissertaciya* [Frequency structures of finite messages in some natural information systems. Ph.D. Thesis]. Tbilisi University. 1974. (In Russ.).
41. Orlov Y.K. [Invisible harmony]. Number and thought. 1980. vol. 3. pp. 70-105. (In Russ.).
42. Kudrin B.I. *Filosofiya tekhniki: osnovaniya postneklassicheskoy filosofii tekhniki.* [Philosophy of technology: the foundations of the post-non-classical philosophy of technology]. M.: Tehnika. 2007. vol. 36. 196 p. (In Russ.).
43. Popova O.V., Gelfand M.S. [Is there an analogue of the Zipf's law in genetic language?]. *Nauchno-tekhnicheskaya informaciya. Seriya 2: Informacionnye processy i sistemy - Scientific and technological information. Series 2: Information Processes and Systems.* 2000. vol. 4. pp. 19-24. (In Russ.).
44. Volchkova I.A., Gumenjuk A.S. [On measures of similarity of multilingual texts with the same content]. *Materialy Vserossiyskoy konferencii s mezhdunarodnym uchastiem «Znaniya – Ontologii – Teorii» (ZONT-13)* [Materials of the All-Russian Conference with International Participation «Knowledge – Ontologies – Theories» (KONT-13)] 2013. Issue 1. pp. 98-105. (In Russ.).
45. Gumenjuk A.S., Volchkova I.A. [Application of means of formal order analysis of a sign sequences for a formal assessment of the quality of translation]. *Omskiy nauchnyy vestnik. Seriya «Pribory, mashiny i tekhnologii» - The Journal Omsk Scientific Bulletin. Series «Devices, machines and technologies».* 2013. Issue 3(123). pp. 251-256. (In Russ.).
46. Shpynov S.N., Gumenuk A.S., Pozdnichenko N.N. [Application of the Numerical Characteristic of Formal Order Analysis of the Prokaryotic Genomes for Reclassification within the Genus Rickettsia]. *Matematicheskaya biologiya i bioinformatika - Mathematical biology and bioinformatics.* 2016. Issue 11. vol. 2. pp. 336-350. (In Russ.).
47. The DDBJ/ENA/GenBank Feature Table Definition. Available at: http://www.insdc.org/files/feature_table.html (accessed 15.04.2018).
48. Gumenuk A.S., Pozdnichenko N.N., Shpynov S.N. [Formal analysis of order in the local structure of the nucleotide sequences]. *Vestnik Tomskogo gosudarstvennogo universiteta - Tomsk State University Journal.* 2014. Issue 4(29). pp. 23–30. (In Russ.).
49. GENBANK DataBase. Available at: http://www.ncbi.nlm.nih.gov/nuccore/ (accessed 02.03.2018).
50. Gumenuk A.S., Pozdnichenko N.N., Skiba A.A., Shpynov S.N. [Computer Program «Matrix of similarity of nucleotide sequences by their components»]. *Svidetel'stvo o*

gosudarstvennoy registracii programmy dlya EHVM. №2017616679 [Certificate of State Registration of the Computer Program in the Register of Computer Programs № 2017616679]. 09.06.2017. (In Russ.).

51. Pozdnichenko N.N., Gumenuk A.S., Shpynov S.N. [On the cartographic representation of the set of prokaryotic genomes and their components by means of numerical characteristics of order]. *Novye informacionnye tekhnologii v issledovanii slozhnyh struktur: materialy 11-y mezhdunarodnoy konferencii* [Computer-aided technologies in applied mathematics: 11th international conference]. 2016. pp. 84-85. (In Russ.).