Знания-Онтологии-Теории (ЗОНТ-09)

О формализмах непосредственного анализа строя знаковых цепей

А.С. Гуменюк, Е.В. Морозенко

Омский государственный технический университет, пр. Мира, д. 11, г. Омск, 644050, Россия

gumas45@mail.ru, morozenkoev@rambler.ru

Аннотация. *В настоящее время не используются адекватные математические средства для анализа порядка следования событий в цепях конечной длины. Поэтому затруднены или невозможны измерение и сравнение формы сигналов, рельефа изображений, композиции музыкальных произведений и речевых высказываний, порядка элементов в длинных молекулярных цепях и текстах, взаимного расположения элементов в массивах данных, представляющих цепи событий разной природы. Объектом исследования в данной работе являются знаковые последовательности, с помощью которых представимы также любые упорядоченные данные измерений. Сформулированы выражения для числовых характеристик, которые описывают оригинальный порядок элементов цепи. В данной работе ставилась цель представить разрабатываемый инструментарий и продемонстрировать некоторые его возможности.*

Ключевые слова: строй элементов, интервал, числовые характеристики строя цепи, энтропия, информация

О формализме строя цепи

Для исследования текстов и знаковых цепей, разработаны и используются большое число специальных подходов, процедур и моделей, среди которых можно отметить спектральный, корреляционный, фрактальный, статистический подходы, марковские цепи, потоки заявок. Однако почти не уделяется внимания закономерностям **«конкретного расположения всех знаков или слов, составляющих отдельную символьную последовательность»**. На наш взгляд, такое положение, в некоторой степени, объясняется отсутствием формализма для выделенного абстрактного объекта и названного здесь **«строем или построением цепи»** [1,2]. Следует отметить, что разные по природе последовательности событий с равномощным составом могут иметь один и тот же оригинальный строй. С другой стороны, очевидно, что мультимножество событий некоторой мощности, может быть основой для построения последовательностей, имеющих разный строй. В данной работе рассматривается подход, который предназначен для формального анализа построения отдельного текста, любой знаковой последовательности или цепи сообщений. Подчеркнем, что здесь не рассматриваются методы исследования локальной структуры символьных последовательностей.

Рассмотрим множество знаковых последовательностей конечной длины. Выделим подмножество кортежей с одинаковыми наборами чисел вхождения символов собственных алфавитов. Так как имеется ввиду неоднократное вхождение некоторых компонентов, то все множество кортежей, составленных на основе конкретного алфавита – это комбинации типа «перестановки с повторениями», каждая из которых отличается оригинальным взаимным расположением компонентов, собственной их композицией или построением.

Кортеж, в котором знаки заменены натуральными числами, как это показано ниже, назовем порядком следования элементов, построением или строем цепи [1,2]. В результате таких замен все равномощные по составу знаковые последовательности, имеющие одинаковое взаимное расположение компонентов (композицию), будут отображены одной и той же

последовательностью натуральных чисел, которая представляет оригинальный для данного подмножества текстов **строй цепи** (порядок следования элементов).

Q	***R***	***D***	***&***	***V***	***Y***	***S***	***S***	***&***	***S***	***D***	***D***	***S***	***Q***	***S***	***D***	***&***	знаковые последовательности
W	***T***	***C***	***G***	***H***	***U***	***R***	***R***	***G***	***R***	***C***	***C***	***R***	***W***	***R***	***C***	***G***	
T	***Y***	***M***	***H***	***T***	***O***	***S***	***S***	***H***	***S***	***M***	***M***	***S***	***T***	***S***	***M***	***H***	
U	***I***	***L***	***J***	***W***	***O***	***Y***	***Y***	***J***	***Y***	***L***	***L***	***Y***	***U***	***Y***	***L***	***J***	
V	***N***	***A***	***B***	***J***	***K***	***T***	***T***	***B***	***T***	***A***	***A***	***T***	***V***	***T***	***A***	***B***	
1	**2**	**3**	**4**	**5**	**6**	**7**	**7**	**4**	**7**	**3**	**3**	**7**	**1**	**7**	**3**	**4**	строй цепи

Таким образом, строй знаковой последовательности или информационной цепи сообщений (по Мазуру [3]) характеризуется мощностью ее состава и оригинальной композицией компонентов. Кратко определим: **строй – это план строения или архетип кортежа** [4]. Кортежи могут представлять последовательности акустических элементов речи человека, хроники исторических событий, тексты, последовательности данных о субъективном состоянии человека, цепи сообщений, массивы измеряемых величин.

Разложим полную (без пустых позиций) **неоднородную символьную последовательность** на ***m*** неполных «однородных» кортежей, в которых заняты только некоторые позиции одинаковыми знаками. Аналогом **однородной последовательности** является однородный поток событий СМО. Вообще разложение цепи может осуществляться по разным правилам. Так ниже (для общности) представлены декомпозиции строя неоднородной знаковой цепи на неполные однородные (рис.1) и неполные «разнородные» (рис.2) цепи. В последнем случае позиции заняты только разными знаками, выбираемыми по следующему правилу: при просмотре цепи от ее начала в состав первой разнородной цепи выбираются все первые вхождения каждого знака из алфавита, при втором – все вторые вхождения и т.д..
В случае правильно выполненных декомпозиций полученные множества однородных (разнородных) последовательностей будут несовместными (т.е. не содержат занятых позиций с одинаковыми номерами). Композиция или «сложение» всех неполных однородных (разнородных) кортежей дает исходную полную неоднородную знаковую последовательность.

Определим **«интервал»** числом позиций, начиная с выделенного в цепи компонента, до другого ближайшего, отмеченного в направлении просмотра (рис.1).

Пусть первое считывание текста осуществляется отличным от обычного (чтение подряд) 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 собственного алфавита или словаря текста.

V	***N***	***A***	***B***	***J***	***K***	***T***	***T***	***B***	***T***	***A***	***A***	***T***	***V***	***T***	***A***	***B***	знаковая цепь
1	**2**	**3**	**4**	**5**	**6**	**7**	**7**	**4**	**7**	**3**	**3**	**7**	**1**	**7**	**3**	**4**	строй цепи
1													**1**				однородные цепи и соответствующие им цепи интервалов
	2																
		3								**3**	**3**				**3**		
			4					**4**								**4**	
				5													
					6												
						7	**7**		**7**			**7**		**7**			

матрица интервалов				
0	**0**	**0**	**13**	**4**
0	**0**	**0**	**0**	**16**
0	**8**	**1**	**4**	**2**
0	**0**	**5**	**8**	**1**
0	**0**	**0**	**0**	**13**
0	**0**	**0**	**0**	**12**
1	**2**	**3**	**2**	**3**

Рис. 1. Декомпозиция строя неоднородной знаковой цепи на неполные однородные.

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

V	*N*	*A*	*B*	*J*	*K*	*T*	*T*	*B*	*T*	*A*	*A*	*T*	*V*	*T*	*A*	*B*	знаковая цепь
1	2	3	4	5	6	7	7	4	7	3	3	7	1	7	3	4	строй цепи
1	2	3	4	5	6	7											разнородные цепи и соответствующие им цепи интервалов
							7	4		3			1				
									7		3					4	
												7			3		
														7			

матрица интервалов						
1	1	1	1	1	1	11
0	0	0	1	2	3	4
0	0	0	0	2	5	1
0	0	0	0	0	3	2
0	0	0	0	0	0	3

Рис. 2. Декомпозиция строя неоднородной знаковой цепи на неполные разнородные.

О числовых характеристиках строя цепи

Используем понятие однородной знаковой последовательности и ее векторное отображение в виде соответствующей строки матрицы интервалов для определения числовых характеристик строения текста (таблица 1) [5].

Таблица 1. Числовые характеристики строя на основе однородных цепей.

(1) $V_j = \prod_{i=1}^{n_j} \Delta_{ij}$	(2) $\Delta_{gj} = \sqrt[n_j]{V_j}$	(3) $G_j = \sum_{i=1}^{n_j} \log_2 \Delta_{ij}$
(4) $V = \prod_{j=1}^{m} V_j$	(5) $\Delta_g = \sqrt[n]{V}$	(6) $G = \sum_{j=1}^{m} G_j$
(7) $\Delta_{aj} = \frac{n}{n_j} = \frac{1}{P_j}$	(8) $\tau_j = \frac{\Delta_{gj}}{\Delta_{aj}}$	(9) $r = \frac{\Delta_g}{D} \le 1$
(10) $\Delta_g = \prod_{j=1}^{m} \Delta_{gj}^{\frac{n_j}{n}}$	(11) $g = \sum_{j=1}^{m} \frac{n_j}{n} \log \Delta_{gj}$	(12) $g = \log \Delta_g$
(13) $D = \prod_{j=1}^{m} \Delta_{aj}^{\frac{n_j}{n}}$	(14) $I = \sum_{j=1}^{m} \frac{n_j}{n} \log \Delta_{aj}$	(15) $H = -\sum_{j=1}^{m} P_j \log P_j$

В таблице обозначены:
j – номер знака в алфавите или номер однородной цепи;
Δ_{ij} – интервал между *i*-м и (*i*+1)-м вхождениями знака *j* в однородной цепи;
$\log\Delta_{ij}$ – **удаленность** (*i*+1)-го вхождения знака *j* относительно его *i*-го вхождения;
n_j – число вхождения знака *j* в цепи;
m – мощность алфавита знаков цепи;
n – длина полной знаковой цепи, равная числу всех ее позиций;
Δ_{aj}, Δ_{gj} – соответственно средний арифметический и средний геометрический интервалы строя *j*-ой однородной цепи;
V_j – абсолютный объем строя *j*-ой однородной цепи;
G_j – глубина расположения строя *j*-ой однородной цепи;
$g_j = \log\Delta_{gj}$ – средняя удаленность знака *j* в строе однородной цепи;
V – абсолютный объем строя на основе однородных цепей;
G – глубина расположения строя на основе однородных цепей;
Δ_g – средний геометрический интервал строя на основе однородных цепей;
g – средняя удаленность любого элемента строя на основе однородных цепей;
D – число описательных информаций, используемых (по Мазуру) для определения некоторого знака цепи;

$\boldsymbol{I}$ – число идентифицирующих информаций, используемых (по Мазуру) для дихотомического распознавания отдельного знака;

$\boldsymbol{\tau_j}$ – периодичность (следования знаков) в строе j-ой однородной цепи;

$\boldsymbol{r}$ – регулярность (следования любого элемента) строя на основе однородных цепей.

Множество величин $\{G_j\}$ всех однородных последовательностей является **распределением глубин** однородных последовательностей для строя отдельной цепи. На рисунке 3 представлено ранговое распределение глубин. Множество величин $\{g_j\}$ является **распределением средних удаленностей** однородных последовательностей. При анализе и описании строя данной цепи распределение средних удаленностей дополняет обычное частотное распределение знаков $\{n_j\}$, **комплексным распределением строя** вида $\{<n_j, g_j>\}$.

Легко определяются аналогичные по форме, но отличные по содержанию, числовые характеристики строя данной цепи, с использованием векторов интервалов, представленных строками матрицы для разнородных цепей. В таком случае формулы (1-15) определяются при выделении знаков j в i-ых разнородных цепях и представляют **числовые характеристики строя на основе разнородных цепей**.

О значениях числовых характеристик для «предельных» моделей строя цепи

Для **«регулярной» знаковой цепи**, в которой все интервалы каждой однородной цепи равны $\boldsymbol{\Delta_{ij} = \Delta_{aj} = n/n_j}$, числовые характеристики строя цепи (10) и (11) принимают максимальные значения ($\boldsymbol{\Delta_g = \Delta_{g\,max} = D}$; $\boldsymbol{g = g_{max} = I}$) и записываются формулами Мазура (13) и (14), представляющими соответственно числа описательных и идентифицирующих информаций в цепи сообщений.

Для бесконечной знаковой цепи $\boldsymbol{(n\to\infty)}$ формула Мазура (14), в которой $\boldsymbol{\Delta_{aj} = (n/n_j)\to (1/P_j)}$, принимает вид формулы К.Шеннона (15) для энтропии или количества информации. Такая информация используется только для дихотомической идентификации (но не для описания) отдельных сообщений.

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

$$\boldsymbol{D > \Delta_g,\ \ H > g = \log\Delta_g}.$$

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

Таким образом, **формулы для среднего геометрического интервала и средней удаленности знаковой цепи обобщают формулы Мазура и Шеннона**, так как, в отличие от последних, при описании строя данной цепи учитывают не только мощность состава, но и взаимное расположение ее компонентов (знаков, слов).

Полагаем важным 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.

Пример числовых характеристик строя текста литературных произведений и его модификаций

В таблице 2 представлены результаты обработки **строя** произведения А.С. Пушкина «Капитанская дочка» и его двух модификаций, в виде энтропийных характеристик и числовых характеристик строя текста. Проведенные эксперименты показали, что даже малые изменения строя текста (модификация 2) приводят к изменениям значений числовых характеристик строя: для глубины G во втором знаке после запятой, для удаленности g и регулярности r – в шестом. На рисунке 3, для сравнения, в билогарифмическом масштабе приведены частотно-ранговые распределения слов и ранговые распределения глубин однородных цепей для двух литературных произведений – «Поднятая целина» (Шолохов М.А., книга 1) и «Капитанская дочка» (Пушкин А.С.).

При автоматической сегментации текстов в качестве элементарных единиц использовались слова, нормированные путем отбрасывания окончания. Критерием качества сегментации являлась степень совпадение мощностей составов, полученных автоматически и вручную, для текстов приведенных в работах Орлова Ю.К.[6].

Таблица 2. Энтропийные и числовые характеристики строя повести Пушкин А.С. «Капитанская дочка».

	n	***m***	***Hs***	***G***	***H***	***q***	***r***
1	29312	4789	290585.01	253104.07	9.913517	8.634847	0.412207
2	29312	4789	290585.01	253104.03	9.913517	8.634845	0.412206
3	29312	4789	290585.01	263790.75	9.913517	8.999375	0.530715

Hs - количество информации в тексте; вычисляется в предположении статистической независимости слов путем перемножения *H* и *n*;
1 – строй оригинального текста;
2 – модификация строя с небольшими изменениями (переставлены два смежных слова);
3 – модификация строя с большими изменениями (со случайными перестановками всех слов).

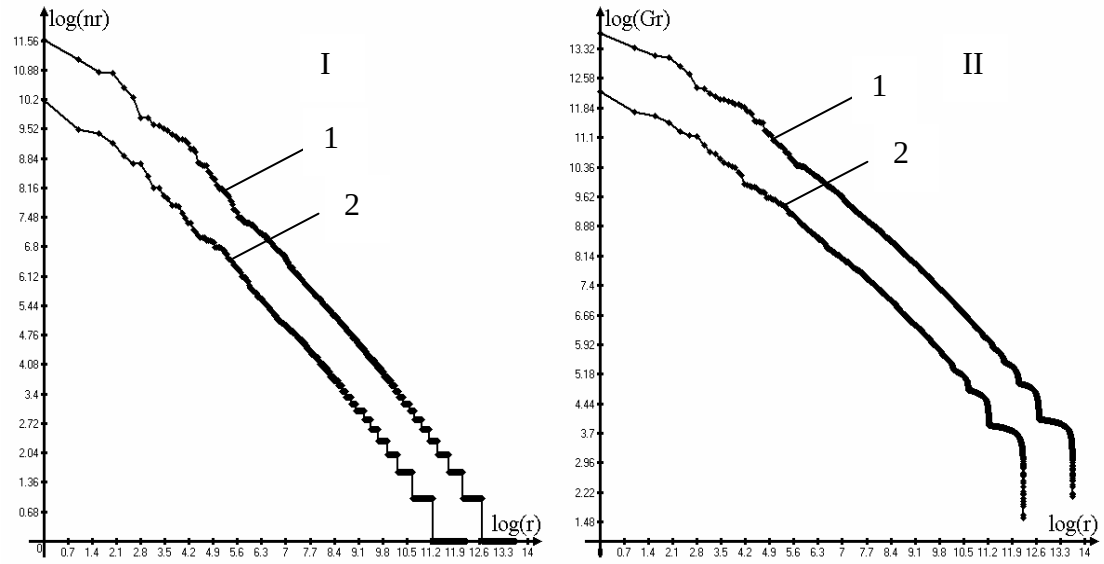


Рис.3. Ранговые распределения частот слов текста (I) и глубин строя (II) для произведений:
1 – Шолохов М.А. «Поднятая целина» (книга 1), 2 – Пушкин А.С. «Капитанская дочка»;
G_r – глубина расположения строя однородной цепи ранга r выделенного слова;
n_r – абсолютная частота слова ранга r;
r – ранг или порядковый номер элемента словаря текста в ранговом распределении.

Пример применения средств анализа для сравнения строя нуклеотидных последовательностей и кластеризации организмов

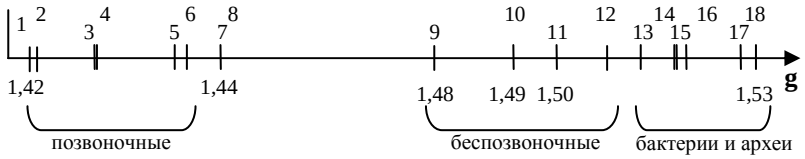
В работе представлена **мера расхождения** построений нуклеотидных последовательностей на основе выше рассмотренных числовых характеристик строя произвольной информационной цепи [7]. Данные формализмы апробированы при компьютерном исследовании рибосомальных РНК 18 организмов, представляющих два царства жизни – прокариота и эукариота. Так же приведены результаты описания и сравнения с использованием только одной числовой характеристики полного неоднородного строя РНК и четырех его характеристик однородных цепей: гуанина, цитазина, тимина, аденина (обозначенных далее соответственно: **G**, **C**, **T**, **A**). Признаковое четырехмерное пространство числовых характеристик, использовалось для построения филогенетического дерева данных организмов (классификации данных генетических текстов), с помощью алгоритма кластеризации в λ-пространстве KRAB, описанного в [8].

В таблице 3 представлены: исследуемые организмы (упорядоченные по возрастанию величины средней удаленности *g*), длины *n* и глубины *G* строя нуклеотидных цепей их малых субъединиц rRNA. Далее, для компактности, названия организмов представлены соответствующими номерами в данной таблице.

Таблица 3. Числовые характеристики строя нуклеотидных цепей.

№	Название организма	*g*	*n*	*G*
1	Mus musculus domesticus	1.4174	1582	2242.3
2	Caiman crocodylus	1.4186	1592	2258.4
3	Canis familiaris	1.4279	1580	2256.1
4	Gallus gallus	1.4284	1620	2314
5	Amia calva	1.4409	1708	2461.1
6	Homo sapiens (человек)	1.4429	1559	2249.5
7	Thermotoga thermarum	1.4482	1471	2130.4
8	Thermus thermophilus	1.4482	1517	2197
9	Ixodes persulcatus	1.4830	1771	2626.4
10	Ornithodorus moubata	1.4958	1790	2677.5
11	Pediculus humanus capitis	1.5031	1493	2244.1
12	Musca domestica	1.5109	1728	2610.9
13	Streptococcus pyogenes	1.5164	1335	2024.4
14	Bacillus anthracis	1.5221	1506	2292.2
15	Borrelia burgdorferi B31	1.5221	1537	2339.5
16	Candidatus nitros. maritimus	1.5239	1452	2212.7
17	Mycoplasma pneumoniae	1.5328	1487	2279.3
18	Neisseria gonorrhoeae	1.5352	1516	2327.4

Ниже на числовой оси организмы расположены в соответствии со значениями их характеристики *g* – средней удаленности элементов в нуклеотидной цепи. Как видно, эукариоты (позвоночные) образовали группу организмов (слева на оси), имеющих наименьшие значения средней удаленности. Прокариоты (бактерии) образовали группу организмов (справа на оси), имеющих наибольшие значения средней удаленности. Эукариоты (беспозвоночные) расположены между этими крайними на оси группами, примыкая к бактериям. Факт выделения из группы прокариот (по характеристике *g*) двух бактерий (7 и 8), значительно отличающихся условиями существования от других, исследуемых в данной работе, требует проведения дальнейших исследований.



Определим **меру расхождения** оригинальных построений разных неоднородных последовательностей A и B, на основе средних удаленностей g_j, составляющих их однородных цепей, в виде

$$r = \sum_{j=1}^{m} \left| g^{A}{}_{j} - g^{B}{}_{j} \right|. \qquad (16)$$

Таблица 4 представляет **матрицу расхождений** *r* строя рибосомальных РНК только некоторых исследуемых организмов (из-за большого размера полной матрицы).

Таблица 4. Расхождения строя рибосомальных РНК

	1	3	6	7	8	10	14	15	18
1	0	0,870	1,266	2,149	2,131	0,575	0,952	0,481	1,169
3	0,870	0	0,454	2,366	2,349	0,792	1,229	1,060	1,443
6	1,266	0,454	0	2,063	2,046	1,066	1,329	1,204	1,543
7	2,149	2,366	2,063	0	0,103	1,574	1,197	1,816	0,980
8	2,131	2,349	2,046	0,103	0	1,557	1,179	1,798	0,962
10	0,575	0,792	1,066	1,574	1,557	0	0,486	0,503	0,700
14	0,952	1,229	1,329	1,197	1,179	0,486	0	0,619	0,217
15	0,481	1,060	1,204	1,816	1,798	0,503	0,619	0	0,836
18	1,169	1,443	1,543	0,980	0,962	0,700	0,217	0,836	0

На основе полученных числовых характеристик однородных цепей выполнено отображение всех РНК в λ-пространство и последующая кластеризация с помощью алгоритма

KRAB. С использованием итерационной кластеризации построено филогенетическое дерево – дендрограмма, размещенная на рис. 4 справа.

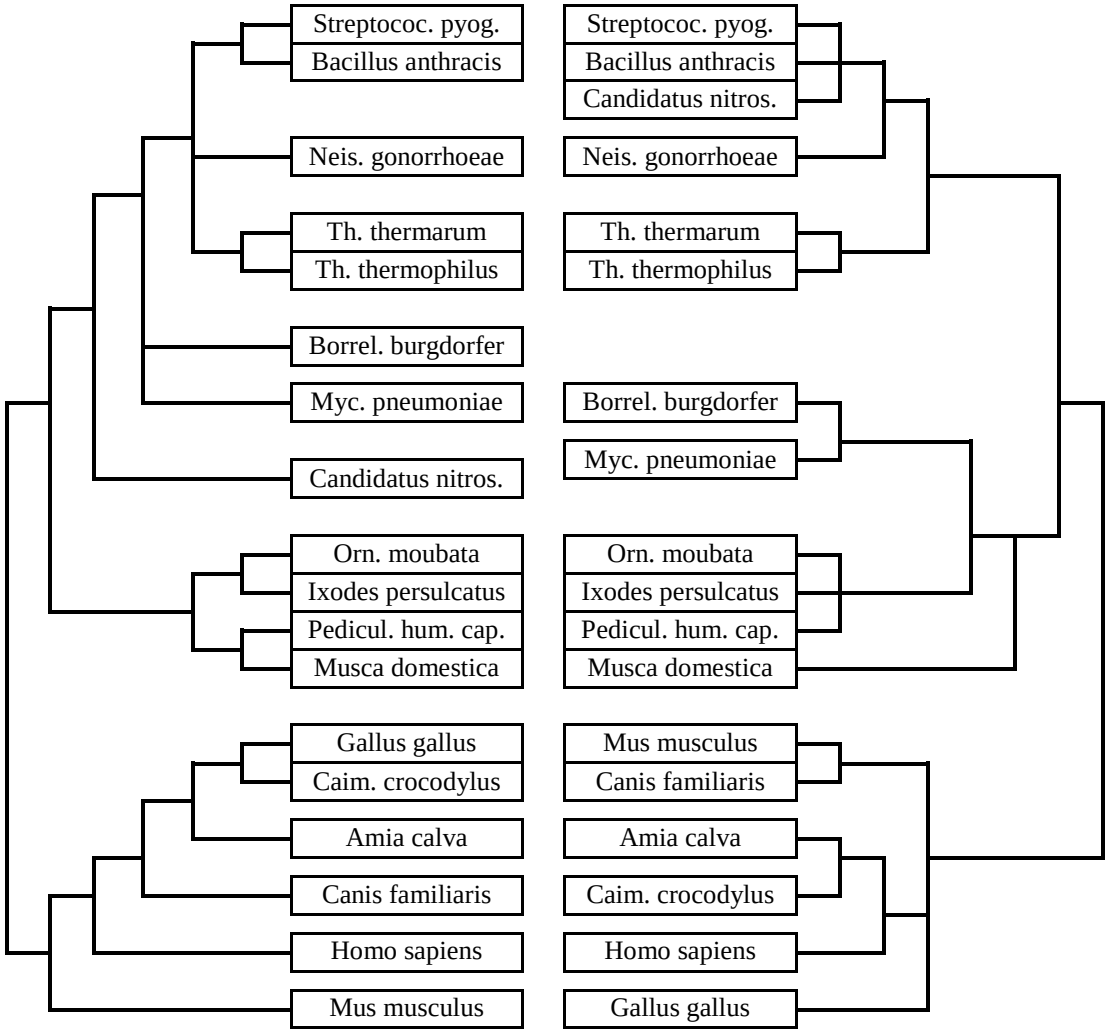


Рис. 4. Филогенетические деревья – дендрограммы.

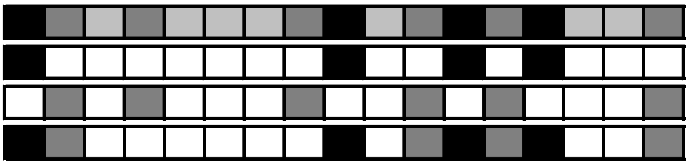
Слева, для сравнения, приведена дендрограмма, полученная с использованием известной процедуры построения филогенетических деревьев. Выравнивание нуклеотидных последовательностей производилось с помощью программы CLUSTAL в оболочке BISANCE (Dessen et al., 1990). Для подсчета процента гомологии нуклеотидных последовательностей по методу Кимура (Kimura, 1980) применялся пакет программ DNADIST (Hitachi Software Engineering America, Brisbane, CA). Выполнение филогенетического анализа проводилось с применением метода ближайшего соседа (neighbour-joining) в программе DNAPARS версии PHYLIP, с использованием метода максимальной вероятности (maximum-likelihood). Для построения филогенетических деревьев (дендрограмм) применялась программа ClustalW в версии 1.8.

На наш взгляд представленные дендрограммы хорошо коррелируют между собой, а числовые характеристики строя адекватно представляют оригинальные построения нуклеотидных цепей. Это дает основания для проведения детальных исследований больших множеств нуклеотидных последовательней с помощью средств интервального анализа.

Об одной характеристике зависимости однородных цепей

В работе с помощью диаграмм представлена одна из введенных в [9], характеристика зависимости пары однородных цепей (*j*-ой и *l*-ой), взятых из состава данной неоднородной цепи. На наш взгляд это позволяет соотнести субъективное суждение о степени «пространственной зависимости» элементов сопоставляемых цепей с формальной числовой

характеристикой, представляющей такую зависимость. Ниже, на диаграмме, в полной неоднородной цепи выделена «бинарно-однородная» цепь (далее бинарная), в состав которой включены только те пары разных знаков *j* и *l* сравниваемых однородных цепей, которые определяются знаками *l*, «ближайшими справа» относительно знаков *j*.



полная неоднородная цепь
j -ая однородная цепь
l -ая однородная цепь
бинарно-однородная цепь

Предполагается, что наличие бинарной цепи свидетельствует о такой причинно-следственной связи цепей, при которой хотя бы некоторые элементы (следствия) *l*-ой цепи расположены вслед за событиями (причинами) *j*-ой цепи на определенных интервалах. Пространственная связь такого типа названа **«причинной зависимостью»** событий одной *l*-ой цепи от другой *j*-ой цепи.

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 $\Delta(l/j)_i$. В примере: $\Delta(l/j)_1 = \Delta(l/j)_3 = 1$; $\Delta(l/j)_2 = 2$; $\Delta(l/j)_4 = 3$. Среднее геометрическое значение всех *n*(*l*/*j*) установленных интервалов (бинарной цепи) $\Delta(l/j)_i$ между «смежными» элементами двух однородных цепей определено в виде

$$\Delta(l/j)_{cp} = \sqrt[n(l/j)]{\prod_{i=1}^{n(l/j)} \Delta(l/j)_i} \ . \qquad (17)$$

Интервалы только между теми знаками выделенной однородной цепи *l*, которые являются «ближайшими справа» относительно знаков цепи *j*, обозначены $\Delta(l_j)_i$. В примере: $\Delta(l_j)_1 = 9$; $\Delta(l_j)_2 = 2$; $\Delta(l_j)_3 = 4$; $\Delta(l_j)_4 = 1$. Среднее геометрическое значение интервалов $\Delta(l_j)_i$ между выделенными элементами l_j в данной однородной цепи определено в виде

$$\Delta(l_j)_{cp} = \sqrt[n(l/j)]{\prod_{i=1}^{n(l/j)} \Delta(l_j)_i} \ . \qquad (18)$$

При условии $\Delta(l_j)_{cp} > \Delta(l/j)_{cp}$, разность вида

$$\upsilon(l_j) = \left(1 - \frac{\Delta(l/j)_{cp}}{\Delta(l_j)_{cp}}\right) \qquad (19)$$

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 (диаграммы 1,4). На основе отмеченного в [9], введен **коэффициент частичной зависимости** *l*-ой однородной цепи (от *j*-ой цепи) в виде

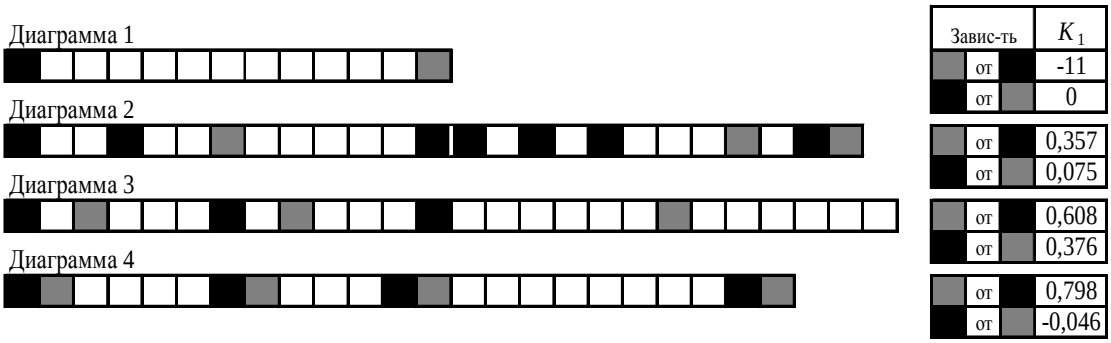
$$K_1(l/j) = \frac{n(l/j)}{n_l} \cdot \upsilon(l_j). \qquad (20)$$

Отметим, что отношение $n(l/j)/n_l$ представляет собой **условную «вероятность»** события**,** состоящего в появления пары знаков *l* и *j*, связанных причинной зависимостью, от появления знаков цепи *l*.

Если все элементы *l*-ой цепи связаны интервалами $\Delta(l/j)_i$ с элементами *j*-ой однородной цепи $n(l/j) = n_l$ (диаграммы 1-4), то оценка частичной зависимости становится оценкой **полной зависимости** и определяется в виде $K_1(l/j) = \upsilon(l_j)$. Если, кроме отмеченного, числа выделенных элементов сравниваемых однородных цепях равны $n_l = n_j$, то такие цепи считаются полностью **взаимозависимыми** (диаграммы 3,4).

Причинная зависимость названа **установленной**, **определенной** или **закономерной**, если размеры всех интервалов $\Delta(l/j)_i$ бинарной цепи задаются определенным операционным преобразованием [3]. В противном случае отмечается **неопределенная причинная зависимость.** Частный случай определенной зависимости назван **фиксированной причинной зависимостью**. При этом «следование за» в бинарной цепи представлено равными интервалами $\Delta(l/j)_i = \Delta(l/j)_{cp} = const$, $i = 1,2,\ldots, n(l/j)$. Предельный случай определенной зависимости назван **непосредственной причинной зависимостью**; при этом бинарная цепь задается единичными интервалами $\Delta(l/j)_i = 1$, $i = 1,2,\ldots, n(l/j)$ (диаграмма 4).

Наконец отметим, что наличие «сильной» причинной зависимости (диаграммы 3, 4), может быть использовано для «сжатия» информации при описании взаимного расположения элементов массива данных.



Для примера, в таблицах 5 и 6 представлены **матрицы причинной зависимости** однородных последовательностей только некоторых букв (из-за большого размера матриц для всего алфавита), рассматриваемых в составе двух повестей А.С. Пушкина («Барышня крестьянка» и «Пиковая дама»). Для каждой пары однородных цепей j и l в таблицах приведены отношение $n(l/j)/n_l$, избыточность $\upsilon(l_j)$ и определяемый ими коэффициент частичной зависимости K_1. События-причины представлены в столбцах, а события-следствия в строках. Максимальные значения K_1 для рассматриваемых букв окрашены.

Таблица 5. «Барышня крестьянка»

	E			M			H			O			P		
	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1
E	0	0	0	0,784	0,839	0,657	0,635	0,785	0,499	0,442	0,765	0,338	0,734	0,837	0,614
M	0,280	0,836	0,234	0	0	0	0,334	0,831	0,278	0,243	0,837	0,204	0,461	0,773	0,356
H	0,462	0,819	0,378	0,681	0,800	0,545	0	0	0	0,411	0,799	0,328	0,633	0,765	0,484
O	0,531	0,776	0,412	0,819	0,855	0,700	0,678	0,783	0,531	0	0	0	0,809	0,816	0,660
P	0,348	0,796	0,277	0,614	0,788	0,484	0,413	0,797	0,329	0,320	0,816	0,261	0	0	0

Таблица 6. «Пиковая дама»

	E			M			H			O			P		
	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1	$\frac{n(l/j)}{n_l}$	$\upsilon(lj)$	K_1
E	0	0	0	0,759	0,858	0,651	0,597	0,782	0,466	0,440	0,764	0,336	0,725	0,846	0,613
M	0,304	0,819	0,249	0	0	0	0,335	0,838	0,281	0,263	0,837	0,220	0,483	0,771	0,372
H	0,501	0,792	0,397	0,702	0,788	0,553	0	0	0	0,439	0,790	0,347	0,640	0,752	0,482
O	0,544	0,773	0,420	0,813	0,855	0,695	0,647	0,780	0,504	0	0	0	0,769	0,812	0,624
P	0,372	0,799	0,298	0,619	0,811	0,502	0,392	0,798	0,313	0,319	0,806	0,257	0	0	0

Компьютерные эксперименты и сравнительный анализ строя знаковых цепей показывают высокую чувствительность предлагаемых характеристик относительно порядка элементов и его изменений в цепи.

Литература

1. Гуменюк А.С. О формализме «строение знаковой последовательности» и его использовании для анализа, сравнения, преобразования и распознавания лингвистических текстов // Четвертый сибирский конгресс по прикладной и индустриальной математике: Тез.докл. – Новосибирск: Изд-во Института математики СО РАН, 2000. – ч.3. – С.84-85.
2. Gumenyuk A., Kostyshin A., Simonova S. An approach to the research of the structure of linguistic and musical texts // Glottometrics. – 3 (2002). Liideschtid RAM – Verl S.61-89.

3. Мазур М. Качественная теория информации, Москва: Мир, 1974.
4. Шрейдер Ю. А., Шаров А. А. Системы и модели. – М.: Радио и связь, 1982. – 152 с.
5. Гуменюк А.С., Богорад Д.М. Об исчислении взаимного расположения элементов в массивах данных // Труды РНТОРЭС им. Попова, серия: научная сессия, посвященная Дню радио. Выпуск:LIX – 2. 2004. – С.114-116.
6. Орлов Ю.К., Частотные структуры конечных сообщений в некоторых естественных информационных системах. Канд. диссертация, Тбилиси: Издательство тбилисского университета, 1974.
7. Гуменюк А.С., Шпынов С.Н., Морозенко Е.В., Родионов И.Н. Пример применения средств анализа для сравнения строя нуклеотидных последовательностей и кластеризации организмов // Нейроинформатика, ее приложения и анализ данных: Материалы XVI Всероссийского семинара / Под ред. А.Н. Горбаня, Е.М. Миркеса; ИВМ СО РАН, - Красноярск; 2008, - с. 166-170.
8. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики, 1999. – 270 с.
9. Гуменюк А.С., Морозенко Е.В. Определение характеристик зависимости цепей сообщений // СИБРЕСУРС-12-2006: Доклады (материалы) 12-й Международной научно-практич. конференции, Тюмень. / Отв. ред. В.Н. Масленников. – Томск: САН ВШ; Изд-во «В-Спектр» – с. 325-328