КЛАСТЕРИЗАЦИЯ ОРГАНИЗМОВ ПО ХАРАКТЕРИСТИКАМ СТРОЯ ИХ ДНК

Н.Н. Поздниченко
ОмГТУ
e-mail: nick670@yandex.ru

Аннотация

В работе рассмотрены подходы для анализа структуры знаковых последовательностей разной природы. Особо выделен анализ строя цепи, который позволяет исследовать структуру цепи безотносительно к её природе, непосредственно учитывая взаимное расположение элементов.

Даётся краткое описание строя цепи как нового математического объекта – особым образом организованного кортежа на основе данной знаковой последовательности. Определена декомпозиция строя на однородные знаковые цепи. Определены две числовые характеристики строя – средняя удалённость и регулярность элементов цепи.

Приведены вычисленные значения характеристик строя 29 нуклеотидных последовательностей живых организмов от простейших до позвоночных. Будучи упорядоченными по характеристике средней удалённости организмы разделились на три группы – между вышеупомянутыми выделилась группа беспозвоночных.

В дальнейших исследованиях производилась автоматическая кластеризация в двумерном λ-пространстве для следующих пар характеристик: <длина нуклеотидной цепи, средняя удалённость>, <регулярность, средняя удалённость>, <регулярность, нормированная удалённость>. Результаты автоматической кластеризации представлены в табличном и графическом виде и хорошо совпадают с экспертными оценками.

Ключевые слова: нуклеотидная последовательность, строй цепи, интервал, однородные цепи, числовые характеристики строя, кластеризация.

Введение

Исследование биологических текстов является одной из актуальнейших задач современного естествознания. При этом под биологическими текстами понимаются символьные модели нуклеотидных и аминокислотных последовательностей. Символьные последовательности являются классическим объектом математики, а также встречаются как предмет изучения во многих прикладных задачах – от теоретического программирования и теории управления до биологии и лингвистики.

Изучение функциональных, химических, физико-химических и прочих свойств нуклеиновых кислот активно ведется специалистами в различных областях, и одним из важных направлений является изучение нуклеиновых кислот как символьных последовательностей. В настоящей работе предполагается исследовать свойства нуклеиновых кислот, которые определяются взаимным расположением нуклеотидов друг относительно друга в одной изучаемой молекуле; никакие другие факторы – физико-химическое окружение, особенности состава нуклеотидов и т.п. – не рассматриваются. Полагается, что символьные последовательности являются целостно-завершенными объектами и их следует рассматривать целиком. Переход от рассмотрения совокупности фрагментов последовательности к рассмотрению её как таковой дает возможность исследовать её строй, как особое свойство

последовательности, имеющую корреляцию с физико-химическими, биологическими и другими свойствами, определенными взаимным расположением элементов.

Следует отметить, что рассмотрение полных геномов и отдельных генов, с точки зрения системного подхода, позволяет ставить задачу по выявлению «естественных» элементов, из которых состоит данная система. Другая важная проблема, имеющая общую значимость для любых прикладных исследований, предметом которых являются те или иные символьные последовательности, состоит в сравнении двух (или нескольких) символьных последовательностей. Дело в том, что символьные последовательности относятся к такому классу объектов, для которых определение расстояния между ними возможно, однако оно очень "бедно": очень часто формально введённое расстояние никак не отражает близость или, наоборот, существенные различия свойств, приписываемых исследователем изучаемым символьным последовательностям. Здесь имеет смысл рассматривать меру близости двух (или нескольких) последовательностей, и для её построения информационные и статистические методы являются весьма универсальными и продуктивными.

Подходы к анализу структуры знаковых последовательностей

Методы анализа структуры знаковых последовательностей можно разделить на два класса: анализирующие структуру знаковых цепей на основе её состава и методы, которые, кроме состава, косвенно учитывают взаимное расположение элементов цепи. Предлагается анализировать структуру цепи на основе непосредственного учёта взаимного расположения элементов. Оценим отмеченные подходы.

К методам косвенного анализа структуры на основе состава знаковых последовательностей можно отнести критерий Ю. Орлова (для целостно-завершённого текста) и информационную энтропию [1,2]. Методы первого класса обладают очевидным недостатком нечувствительности к порядку расположения элементов в цепи.

К методам косвенного анализа структуры текста на основе строя и состава можно отнести Марковские цепи и графовые модели текста. Методы второго класса опираются на нечёткие понятия структуры и взаимного расположения элементов и поэтому не позволяют получить инвариантные к природе знаковых последовательностей и универсальные числовые описания порядка следования элементов.

Основы аппарата анализа строя цепи представлены А.С. Гуменюком в работе [3]. Подход для непосредственного анализа взаиморасположения элементов опирается на понятие нового абстрактного объекта – строя цепи. В отличие от кортежа или упорядоченного множества, один и тот же строй конечной цепи отображает счётное множество знаковых последовательностей той же длины и одинаковой мощностью состава. *Строем* цепи сообщений (событий, знаков и т.п.) называется кортеж (упорядоченное множество), в котором каждому компоненту данной цепи в соответствие поставлено натуральное число, причем идентичные по выбранному признаку компоненты отображены одним и тем же числом. Самый первый компонент такого кортежа – единица, а все остальные первые встречные разные натуральные числа (представляющие вместе с единицей алфавит строя) возрастают на единицу.

Нуклеотидная цепь 1	T	T	A	T	T	T	C	A	A	A	A	A	G	C	G
Цепь комплементарная цепи 1	A	A	T	A	A	A	G	T	T	T	T	T	C	G	C
Строй цепи 1 и цепи компементарной ей	**1**	**1**	**2**	**1**	**1**	**1**	**3**	**2**	**2**	**2**	**2**	**2**	**4**	**3**	**4**
Нуклеотидная цепь 2	G	C	T	C	C	T	G	C	C	A	A	T	A	G	C
Строй цепи 2	**1**	**2**	**3**	**2**	**2**	**3**	**1**	**2**	**2**	**4**	**4**	**3**	**4**	**1**	**2**

$\Delta_{11} = 6$ $\Delta_{12} = 7$

Для анализа оригинального строя предварительно производится его декомпозиция на однородные цепи, отдельные позиции которых заняты выделенными по определенному правилу компонентами, другие позиции пусты. Если выделенными элементами являются все одинаковые, то такую цепь назовём однородной. Возможна в некотором смысле противоположная декомпозиция по разнородным цепям. Из теории массового обслуживания известно разложение неоднородного потока заявок на однородные потоки заявок (событий); при этом в качестве случайной величины выступает интервал между заявками. В однородной цепи строя в качестве первичного измеримого информационно-значимого элемента выступает интервал между ближайшими соседними выделенными элементами, который представлен натуральным числом не меньшим 1 и обозначается Δ_{ji}. Все числовые характеристики строя и их распределения (в отличие от вероятностно-статистических) получаются перемножением интервалов в однородных цепях и вычислением их средних геометрических значений [3]. На практике удобно использовать эти характеристики в логарифмическом масштабе. В том числе из-за их аддитивности.

Изучение символьных последовательностей, в том числе их статистических и информационных характеристик является центральным местом во многих задачах биоинформатики, лингвистики, некоторых других областях знания. Биоинформатики, занимающиеся изучением различных свойств символьных последовательностей, отмечают особую важность взаимного расположения элементов последовательностях. При решении любых прикладных задач биоинформатики исследователь сталкивается с проблемой учета построения последовательности, для решения которой в большинстве случаев используются Марковские модели, опосредованно учитывающие локальное взаиморасположение элементов. Следует отметить, что, несмотря на признание биоинформатиками особой важности взаимного расположения нуклеотидов и триплетов в ДНК и РНК последовательностях исследования их взаимного расположения как такового не проводились.

Исследования

Для классификации организмов на основе характеристик строя их ДНК использовался разработанный нами программный комплекс, который реализует алгоритм λ-KRAB, описанный Н.Г. Загоруйко в [4].

Исходными данными для исследования были нуклеотидные тексты различных организмов, для их разных представлений вычислялись характеристики строя, после чего проводилась кластеризация в признаковом пространстве этих характеристик.

В таблице 1, а также на рис. 2, 3 представлены имена 29 организмов упорядоченные по характеристике средней удалённости любых соседних одинаковых элементов нуклеотидной цепи [5], которая представлена в виде

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

где m – мощность алфавита, n_j – число вхождений j-го элемента, n – длина цепи, Δ_g – среднегеометрический интервал.

Кроме того, в таблице отмечены курсивом 18 организмов сравнение характеристик которых и осуществлённая на их основе кластеризация были опубликованы ранее в [5]. Легко видеть, что даже по одной этой характеристике строя можно более или менее уверенно судить о принадлежности организма к одному из трёх типов: позвоночным беспозвоночным или простейшим. Предварительно была выполнена экспертная кластеризация, которая подтверждает размещение значений характеристики строя g по трём отмеченным кластерам.

Таблица 1

id	название	g
1	***M.musculus - мышь***	***1,417414581***
2	***C.crocodylus - Крокодил***	***1,418643679***
3	***C.familiaris - Собака***	***1,427959968***
4	***G.gallus - курица***	***1,428402795***
5	Sus scrofa - Кабан	1,440507111
6	***A.calva - рыба***	***1,440937027***
7	***H.s - человек***	***1,442946531***
8	***Th.thermophilus - микроорганизм***	***1,448283347***
9	***Th.thermarum - микроорганизм***	***1,448292643***
10	Gallus gallus - Банкивская джунглевая курица	1,455663685
11	Bos taurus 18S ribosomal RNA gene - Дикий бык	1,460651525
12	Erinaceus europaeus - Обыкновенный ёж	1,463618632
13	Homo sapiens - Человек разумный	1,463758914
14	Mus musculus - Домовая мышь	1,46560513
15	Cricetulus griseus - серый хомячок	1,468230019
16	Rattus norvegicus - Серая крыса	1,470291648
17	Crocodylus niloticus - Нильский крокодил	1,478396611
18	***I.persulcatus - Иксодовые клещи***	***1,483028363***
19	Zebrias zebra - Рыба	1,490709799
20	Kareius bicoloratus - двухцветная камбала	1,49358883
21	***O.moubata - клещи***	***1,495843469***
22	***P.humanus cap - блоха***	***1,503112618***
23	***M.domestica - муха***	***1,510952625***
24	***S.pyogenes - Стрептококк***	***1,51643989***
25	***B.anthracis - Сибирская язва***	***1,522106822***
26	***B.burgdorferi - боррелиоз***	***1,522174205***
27	***Candidatus N.m - бактерия***	***1,523941071***
28	***M.pneumoniae - атипичная пневмония***	***1,532855504***
29	***N.g - гонорея***	***1,535260296***

Для наглядности результатов кластеризации номера всех 29 организмов помечены на оси характеристики средней удалённости – g (рис. 2). Ниже для сравнения в том же виде представлены характеристики исследованных ранее 18 организмов (рис. 3).

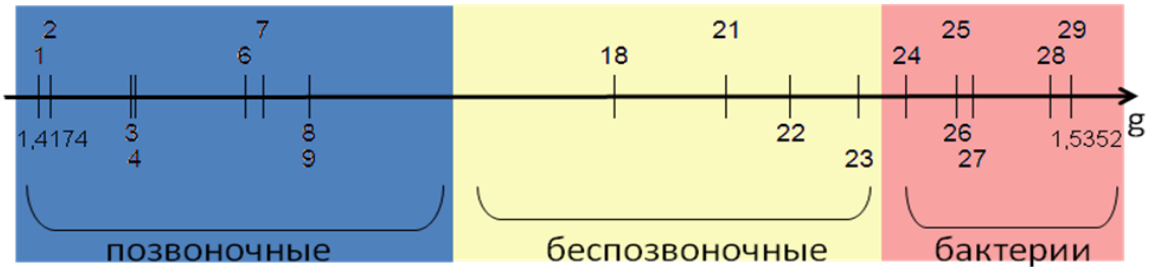

Рисунок 2

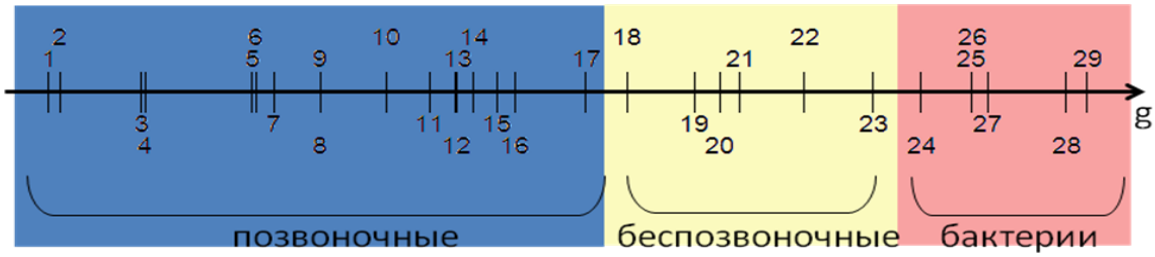

Рисунок 3

В данном исследовании значительная часть добавленных к ранее исследованным 18 цепочкам принадлежит позвоночным, поэтому использование большого значения коэффициента равномощности таксонов алгоритма кластеризации не имело смысла [4]. В таблице 2 и на рис. 4 приведено разбиение всего множества организмов на два таксона при коэффициента равномощности равном 2. Признаковое пространство составляют длинна цепей и удалённость.

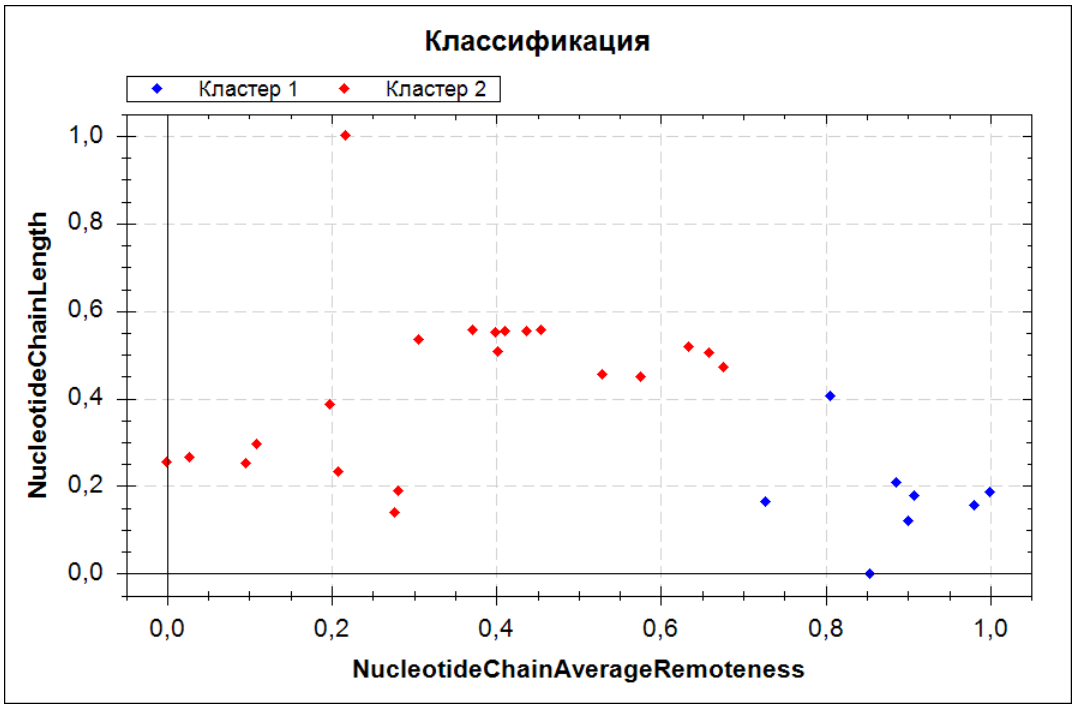

Рисунок 4 Классификация по удалённости и длине

Таблица 2 Результаты кластеризации на 2 группы

id	Название	номер кластера
469	Candidatus N.m - бактерия	1
466	B.burgdorferi - боррелиоз	1
465	B.anthracis - Сибирская язва	1
479	S.pyogenes - Стрептококк	1
478	P.humanus cap - блоха	1
476	N.g гонорея	1
475	M.pneumoniae - Атипичная пневмония	1
473	M.domestica - муха	1
468	C.crocodylus - Крокодил	2
467	C.familiaris - Собака	2
464	A.calva - рыба	2
504	Zebrias zebra - Рыба	2
503	Sus scrofa - Кабан	2
502	Rattus norvegicus - Серая крыса	2
501	Mus musculus - Домовая мышь	2
500	Kareius bicoloratus - двухцветная камбала	2
499	Homo sapiens - Человек разумный	2
498	Gallus gallus - Банкивская джунглевая курица	2
497	Erinaceus europaeus - Обыкновенный ёж	2
496	Crocodylus niloticus - Нильский крокодил	2
495	Cricetulus griseus - серый хомячок	2
494	Bos taurus 18S ribosomal RNA gene - Дикий бык	2
481	Th.thermophilus - микроорганизм	2
480	Th.thermarum - микроорганизм	2
477	O.moubata - клещи	2
474	M.musculus - мышь	2
472	I.persulcatus - Искодовые клещи	2
471	H.s - человек	2
470	G.gallus - курица	2

На рис. 5 и в таблице 3 представлено разбиение на три класса при коэффициенте равномощности равном 1, признаковое пространство составляют характеристики удалённости и регулярности $r = \frac{\Delta_g}{D}$, где $D = \Delta_{g\ max}$ – максимальный среднегеометрический интервал, или число описательных информаций по М. Мазуру [6].

Таблица 3 Результаты кластеризации на 3 группы

id	Название	номер кластера
469	Candidatus N.m - бактерия	1
466	B.burgdorferi - боррелиоз	1
465	B.anthracis - Сибирская язва	1
479	S.pyogenes - Стрептококк	1
478	P.humanus cap - блоха	1
476	N.g гонорея	1
475	M.pneumoniae - Атипичная пневмония	1
473	M.domestica - муха	1
468	C.crocodylus - Крокодил	2
467	C.familiaris - Собака	2
464	A.calva - рыба	2
504	Zebrias zebra - Рыба	2
502	Rattus norvegicus - Серая крыса	2
501	Mus musculus - Домовая мышь	2
500	Kareius bicoloratus - двухцветная камбала	2
499	Homo sapiens - Человек разумный	2
498	Gallus gallus - Банкивская джунглевая курица	2
497	Erinaceus europaeus - Обыкновенный ёж	2
496	Crocodylus niloticus - Нильский крокодил	2
495	Cricetulus griseus - серый хомячок	2
494	Bos taurus 18S ribosomal RNA gene - Дикий бык	2
477	O.moubata - клещи	2
474	M.musculus - мышь	2
472	I.persulcatus - Иксодовые клещи	2
471	H.s - человек	2
470	G.gallus - курица	2
481	Th.thermophilus - микроорганизм	3
480	Th.thermarum - микроорганизм	3

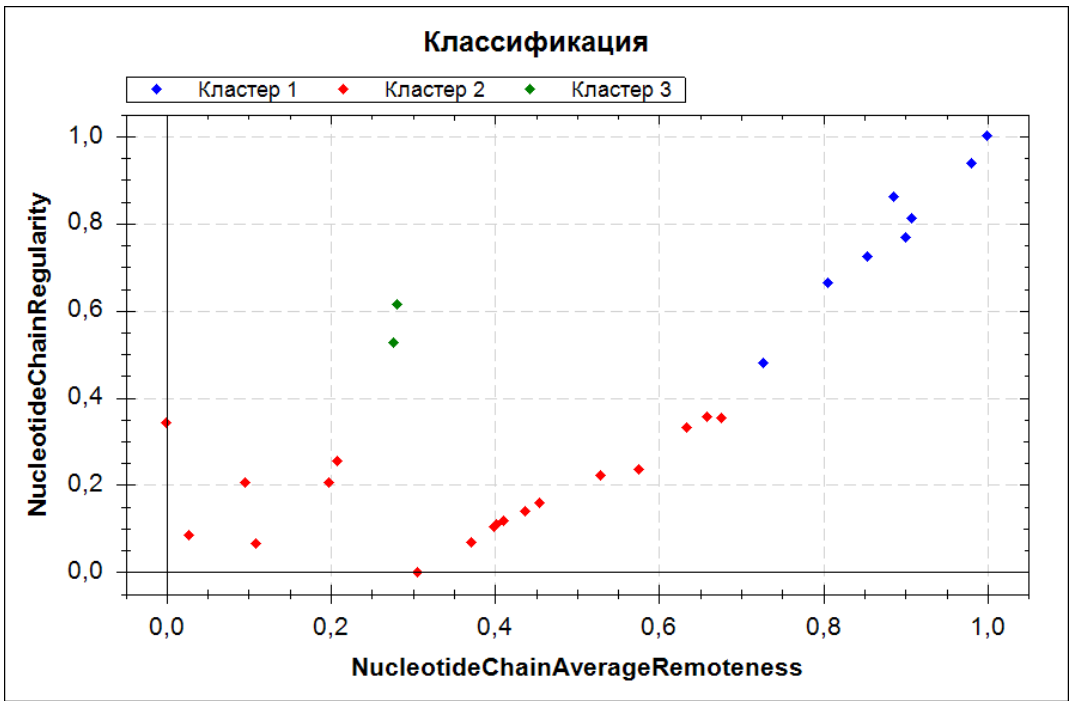

Рисунок 5 Классификация по удалённости и регулярности

В таблице 4 представлено разбиение на три класса по нормированным характеристикам удалённости и регулярности при коэффициенте равномощности равном 1.

Таблица 4 Результаты кластеризации на 3 группы

id	Название	номер кластера
481	Th.thermophilus - микроорганизм	1
480	Th.thermarum - микроорганизм	1
469	Candidatus N.m - бактерия	2
466	B.burgdorferi - боррелиоз	2
465	B.anthracis - Сибирская язва	2
504	Zebrias zebra - Рыба?	2
500	Kareius bicoloratus - двухцветная камбала	2
496	Crocodylus niloticus - Нильский крокодил	2
479	S.pyogenes - Стрептококк	2
478	P.humanus cap - блоха	2
477	O.moubata - клещи	2
476	N.g гонорея	2
475	M.pneumoniae - Атипичная пневмония	2
473	M.domestica - муха	2
472	I.persulcatus - Иксодовые клещи	2
468	C.crocodylus - Крокодил	3
467	C.familiaris - Собака	3
464	A.calva - рыба	3
503	Sus scrofa - Кабан	3
502	Rattus norvegicus - Серая крыса	3

501	Mus musculus - Домовая мышь	3
499	Homo sapiens - Человек разумный	3
498	Gallus gallus - Банкивская джунглевая курица	3
497	Erinaceus europaeus - Обыкновенный ёж	3
495	Cricetulus griseus - серый хомячок	3
494	Bos taurus 18S ribosomal RNA gene - Дикий бык	3
474	M.musculus - мышь	3
471	H.s - человек	3
470	G.gallus - курица	3

Нетрудно видеть, что при правильном выборе коэффициента равномощности типы организмов легко отделяются друг от друга, впрочем, на границе таксонов всё равно наблюдается некоторое смешение, например, клещи (номера 477, 472) оказались в группе позвоночных во всех разбиениях.

Выводы

Проведённые исследования показывают применимость характеристик строя цепи в качестве признакового пространства для кластеризации и классификации организмов по данным полученным только из генетических текстов. Это говорит о возможности в дальнейшем осуществлять машинный анализ и построение классификаций организмов, используя, в первую очередь, информацию об их ДНК или отдельных генах.

Кроме того, данный инструментарий планируется использовать для проверки качества сегментации (разбиения) генетического текста на естественные единицы и построения гипотетического словаря языка генетических текстов.

Литература

[1]. Садовский М.Г. Информационно-статистический анализ нуклеотидных последовательностей. Диссертация на соискание ученой степени доктора физико-математических наук. Красноярск, 2004. - 394 с.

[2]. Орлов Ю.К. Невидимая гармония // Сб. Число и мысль. – М.: Знание, 1980. – 73 с.

[3]. Алгоритмы анализа структуры сигналов и данных: монография / А.С. Гуменюк, Ю.Н. Кликушин, В.Ю. Кобенко, В.Н. Цыганенко; под науч. ред. д-ра техн. наук Ю.Н. Кликушина. – Омск: Изд-во ОмГТУ, 2010. – 272 с.

[4]. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во ин-та математики, 1999. – 270 с.

[5]. Гуменюк А.С., Морозенко Е.В. О формализмах непосредственного анализа строя знаковых цепей. // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ-09) Том 2 – Новосибирск: Изд-во института математики СО РАН, 2009. – С. 83-92.

[6]. Мазур М. Качественная теория информации. – М.: Мир, 1984. – 280 с.