ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2018 Управление, вычислительная техника и информатика № 42

ОБРАБОТКА ИНФОРМАЦИИ

УДК 004.8:575
DOI: 10.17223/19988605/42/4

А.С. Гуменюк

СЕРИАЛЬНОЕ И КОМПЛЕКСНОЕ ОПИСАНИЕ СТРОЯ КОМПОНЕНТОВ В МАССИВАХ ДАННЫХ

Отмечено «интервальное» представление строя информационной цепи, числовые характеристики которого показали высокую чувствительность к расположению компонентов в «текстовых» массивах данных. Предложено «сериальное» представление строя цепи, числовые характеристики которого оценивают массивы, состоящие из чередующихся серий одинаковых данных. Вводится комплексное представление массивов, сформированных как отдельно стоящими разными элементами, так и чередующимися сериями одинаковых данных.
Ключевые слова*:* строй цепи; числовые характеристики строя; удаленность сообщения; протяженность серии; межнуклеотидное расстояние.

В работах [1, 2] введено понятие «строй цепи». Представим его компактное описание. Рассмотрим кортежи, представляющие собой комбинации типа «перестановки с повторениями», компонентами которых являются натуральные числа, не превышающие n (длина кортежа). В теоретико-множественном представлении такой кортеж называется вектором. Рассмотрим в множестве таких векторов подмножество векторов, первый компонент которых «единица», каждый следующий компонент с номером $i \le n$ либо представлен натуральным числом, равным компоненту-числу с номером $j < i$, либо этот компонент на единицу больше, чем максимальное компонент-число с номером $k < i$ (где $k \ne j$). Назовем элемент этого подмножества «вектор строя» (далее – строй). Другими словами, строй цепи определен как особого рода кортеж, компонентами которого являются натуральные числа $<a_1, a_2, \ldots, a_k, \ldots, a_n>$, где $a_1 = 1$, $a_n \le n$; если $a_k \notin \{a_1, a_2, \ldots, a_{k-1}\}$, то $a_k = \max\{a_1, a_2, \ldots, a_{k-1}\} + 1$. Заменим в любом кортеже длины n первые встречные разные компоненты (элементы алфавита мощностью $m \le n$) возрастающими на единицу натуральными числами (начиная с единицы). Другие повторяющиеся одинаковые компоненты отметим теми же числами. Очевидно, что в результате любой кортеж преобразуется в вектор строя.

В работах, представляющих средства формального описания и анализа строя цепи сообщений (ФОАС), рассматривались такие длинные кортежи (знаковые последовательности, большие массивы данных), в которых отдельные различные компоненты на протяжении цепи почти всегда чередуются, а серии (одинаковых подряд расположенных) сообщений редкие и короткие. Исследования таких разных по природе цепей, как литературные тексты, нотные записи музыкальных произведений, нуклеотидные цепи (геномов прокариот), показали высокую чувствительность «интервальных» числовых характеристик строя (представленных целым комплексом) к расположению сообщений в цепи. Эффективность средств ФОАС при исследовании больших массивов данных, кроме прочего, обусловлена линейной зависимостью вычислительной сложности от длины последовательностей. Функции числовых характеристик строя эффективно представляют локальное расположение сообщений на разных участках цепи [2]. Заметим, что числовое отображение знаковой цепи позволяет применять разнообразные средства анализа сигналов (математический, спектральный, статистический и корреляционный анализ и др.). Наше особое внимание к исследованию символьных последовательностей объясняется недостатком адекватных средств для их описания и анализа.

Предложенные ранее средства предполагают отображение знаковой цепи ее строем, в результате декомпозиции которого выделяются j-е неполные однородные цепи, в каждой из которых отмечены (натуральным числом) занятые места расположения одинаковых сообщений в количестве n_j (остальные места на позиции всей цепи пустые). В качестве первичной измеримой единицы используется очередной (i-й) интервал между ближайшими занятыми местами, размер которого Δ_{ij} определяется числом пустых мест плюс один. Таким образом, определяется расстояние от одного сообщения до ближайшего такого же (по-английски distance). Заметим, что учет интервалов между однородными заявками (событиями) издавна используется в теории массового обслуживания при моделировании потоков однородных и неоднородных случайных событий [3]. В последние годы [4, 5] учет интервалов находит распространение, в частности, при анализе нуклеотидных цепей (inter-nucleotide distance). Однако использование массивов интервалов знаковых последовательностей осуществляется традиционными математическими средствами, которые не позволяют осуществлять непосредственное описание и анализ расположения компонентов в целостной цепи.

Очевидна необходимость разработки средств описания и анализа строя таких массивов данных, которые преимущественно представлены чередующимися сериями одинаковых сообщений. Это могут быть, например, оцифрованные изображения, последовательности измеряемых величин и т.п.

В работе [2] представлены более подходящие для таких данных числовые характеристики строя. При этом используется декомпозиция знаковой цепи на неполные неоднородные цепи. Вначале считывается неоднородная цепь, содержащая только первые встречные разные знаки. Такая цепь содержит весь алфавит или словарь (мощностью m) рассматриваемого массива. Затем – другая цепь, содержащая вторые встречные разные знаки, и т.д.

1. Сериальное описание строя цепи

Для непосредственного учета чередующихся серий в массиве данных введем необходимые далее понятия и их обозначения. Назовем отрезком или серией упорядоченное множество одинаковых данных, выделенных подряд в составе j-й неполной однородной цепи. Допустим, что на позиции некоторого отрезка можно «считать» несколько серий по числу элементов в этом отрезке. В качестве первичной измеримой единицы используем длину серии, размер которой определяется числом занятых подряд мест в j-й однородной цепи. Наименьшая длина отрезка – единица. Для краткости будем употреблять термин «длина», величину которой обозначим τ_{ij} (i – номер элемента в j-й однородной цепи). Назовем протяженностью данной серии логарифм ее длины $l_{ij} = \log_2 \tau_{ij}$. Очевидно, протяженность отдельно стоящего элемента равна нулю.

Сериальный объем j-й однородной цепи, составленной отрезками, определим в виде:

$$V_{\tau_j} = \prod_{i=1}^{n_j} \tau_j \,, \tag{1}$$

где i – номер элемента в j-й однородной цепи, а n_j – число серий в j-й однородной цепи (при чтении серий начиная с каждого отмеченного в ней сообщения).

Средняя геометрическая длина серий в j-й однородной цепи определяется в виде:

$$\tau_{gj} = \sqrt[n_j]{V_{\tau_j}} \,. \tag{2}$$

При этом сериальный объем j-й однородной цепи можно определить как

$$V_{\tau j} = \tau_{gj}^{n_j} \,. \tag{3}$$

Подставляя (1) в (2) определим τ_{gj} в виде:

$$\tau_{gj} = \sqrt[n_j]{\prod_{i=1}^{n_j} \tau_{ij}} \,.$$

Логарифмируя (1), получим суммарную протяженность (далее – протяжение) всех серий j-й однородной цепи:

$$L_j = \sum_{j=1}^{m} \log_2 \tau_{ij} \,.$$

Определим из (3) протяжение серий в j-й однородной цепи на основе их средней геометрической длины:

$$L_j = n_j \cdot \log_2 \tau_{gj}. \quad (4)$$

Из (4) средняя протяженность серии в данной однородной цепи определится как

$$l_j = \log_2 \tau_{gj}.$$

Суммарная и средняя длины всех серий j-й однородной цепи определяются, соответственно, в виде:

$$s_j = \sum_{j=1}^{n_j} \tau_{ij}, \ \tau_{aj} = s_j / n_j.$$

Отношение средних длин (геометрического и арифметического) определим в виде:

$$\delta\tau_j = \tau_{gj} / \tau_{aj}.$$

Эта величина является средней относительной (нормированной) длиной серий в j-й однородной цепи. В некотором смысле она аналогична длительности ноты в музыке, измеряемой в долях такта. Очевидно, $\delta\tau_j$ максимально и равно единице, когда все отрезки в однородной цепи равны. Вероятно, эту величину можно назвать длительностью серий, или равносерийностью, или просто серийностью j-й однородной цепи.

Сериальный объем полной неоднородной цепи определим произведением всех m j-х сериальных объемов:

$$V_\tau = \prod_{j=1}^{m} V_{\tau_j}. \quad (5)$$

Суммарная и средняя длины всех серий m разных однородных цепей данного массива, соответственно, определяются как

$$s = \sum_{j=1}^{m} s_j, \ \tau_a = s / n,$$

где n – длина всей цепи данных.

Определим среднюю геометрическую длину всех серий данного массива в виде:

$$\tau_g = \sqrt[n]{V_\tau}. \quad (6)$$

Подставляя (1) в (5) и (5) в (6), определим эту величину как

$$\tau_g = \sqrt[n]{\prod_{j=1}^{m} \prod_{i=1}^{n_j} \tau_{ij}}.$$

При этом сериальный объем массива данных можно определить из (6) в виде:

$$V_\tau = \tau_g^n. \quad (7)$$

Отношение средних длин (геометрического и арифметического) назовем средней относительной (нормированной) длиной серий в строе цепи определим в виде:

$$\delta\tau = \tau_g / \tau_a.$$

Эту величину можно назвать длительностью серий, или равносерийностью, или серийностью строя цепи.

Логарифмируя (7) получим суммарную протяженность (протяжение) всех серий в полной неоднородной цепи:

$$L = n \cdot \log_2 \tau_g.$$

Средняя протяженность серий в полной неоднородной цепи определится в виде:

$$l = \log_2 \tau_g.$$

Назовем «полным» сериальным описанием строя цепи, образованного только сериями данных, распределения суммарных и средних протяженностей j-х однородных цепей вида: $\{L_j\}$, $\{<l_j, n_j>\}$,

где $j = 1, m$. При компактном описании сериальных данных применимы разного рода числовые характеристики протяженности серий, а именно средняя протяженность, дисперсия, или СКО, протяженностей и центрированные моменты более высоких порядков.

Наконец, отметим, что строй цепи, содержащий кроме «разряженных» данных также и их серии, следует описывать комплексами числовых характеристик и их распределений, учитывающих как интервалы и удаленности, так и длины серий и их протяженности.

2. Интервально-сериальное описание строя цепи

Рассмотрим массив, содержащий как отдельно стоящие чередующиеся разные сообщения, так и серии идущих подряд одинаковых данных. Для описания строя такой информационной цепи также необходима предварительная ее декомпозиция на однородные цепи.

Неполная j-я однородная цепь, на позиции которой некоторые места заняты (одинаковыми сообщениями), а другие пусты, может быть отображена как «интервальными», так и «сериальными» данными. При интервальном рассмотрении определяется упорядоченное множество интервалов вида:

$$\left\langle \Delta_{1j}, \Delta_{2j}, \ldots, \Delta_{ij}, \ldots, \Delta_{n_j j} \right\rangle . \tag{8}$$

При сериальном отображении получается кортеж длин серий вида:

$$\left\langle \tau_{1j}, \tau_{2j}, \ldots, \tau_{ij}, \ldots, \tau_{n_j j} \right\rangle . \tag{9}$$

Последовательности интервалов и длин j-й однородной цепи (8) и (9) комплементарно дополняют друг друга следующим образом: если $\Delta_{ij} \geq 2$, то $\tau_{ij} = 1$; если $\tau_{ij} \geq 2$, то $\Delta_{ij} = 1$.

При сериальном отображении всех m j-х неполных однородных цепей и целого массива данных получаем **набор сериальных характеристик строя** τ_{ij}, $V_{\tau j}$, V_{τ}, τ_{gj}, τ_g, $l_{ij} = \log_2 \tau_{ij}$, $l_j = \log_2 \tau_{gj}$, $L_j = n_j \cdot \log_2 \tau_{gj}$, $l = \log_2 \tau_g$, $L = n \cdot \log_2 \tau_g$.

При интервальном отображении j-х неполных однородных цепей и полной неоднородной последовательности ранее получен аналогичный набор интервальных характеристик строя со следующими (уточненными здесь) обозначениями и наименованиями:

Δ_{ij} – интервал от i-го до $(i + 1)$-го ближайшего такого же сообщения в j-й однородной цепи;

$V_{\Delta j}$ – интервальный объем j-й однородной цепи;

V_{Δ} – интервальный объем полной неоднородной цепи;

Δ_{gj} – средний геометрический интервал в j-й однородной цепи;

Δ_g – средний геометрический интервал в полной неоднородной цепи;

$g_{ij} = \log_2 \Delta_{ij}$ – удаленность $(i + 1)$-го сообщения относительно i-го в j-й однородной цепи;

$g_j = \log_2 \Delta_{gj}$ – средняя удаленность сообщений в j-й однородной цепи;

$g = \log_2 \Delta_g$ – средняя удаленность сообщений в полной неоднородной цепи;

$G_j = n_j \cdot \log_2 \Delta_{gj}$ – суммарная удаленность, или удаление, сообщений в j-й однородной цепи (ранее – глубина j-й цепи);

$G = n \cdot \log_2 \Delta_g$ – суммарная удаленность, или удаление, полной неоднородной цепи (ранее – глубина цепи).

Отдельно запишем выражение, аналогичное (3), в котором интервальный объем j-й однородной цепи определен через средний геометрический интервал как

$$V_{\Delta j} = \Delta_{gj}^{n_j} .$$

Произведение интервального и сериального объемов j-й однородной цепи назовем (полным) объемом j-й однородной цепи и запишем в виде:

$$V_j = V_{\Delta j} \cdot V_{\tau j} . \tag{10}$$

Используя комплементарную дополнительность интервального (8) и сериального (9) представлений всех n_j сообщений j-й однородной цепи, произведения пар интервалов и длин серий для каждого i-го сообщения запишем в виде:

$$\prod_{i=1}^{n_j}\left(\Delta_{ij}\cdot\tau_{ij}\right)=\prod_{i=1}^{n_j}\Delta_{ij}\cdot\prod_{i=1}^{n_j}\tau_{ij}=V_{\Delta j}\cdot V_{\tau j}. \qquad (11)$$

Из сравнения (10) и (11) видно, что одна и та же комплексная характеристика V_j может быть вычислена либо с раздельным определением интервальной и сериальной характеристик, либо – без раздельной оценки этих свойств j-й однородной цепи.

Назовем емкостью i-го сообщения в j-й цепи произведение вида

$$\upsilon_{ij}=\Delta_{ij}\cdot\tau_{ij}.$$

«Емкостное» представление j-й однородной цепи, определенное кортежем вида

$$\left\langle v_{1j}, v_{2j}, \ldots, v_{ij}, \ldots, v_{n_j j}\right\rangle,$$

не содержит компонентов, равных единице, так как $v_{ij}\geq 2$ для любых компонентов, однако при этом интервальные и сериальные свойства цепи неразличимы.

Для получения емкостных характеристик строя цепи применимы выражения, представленные выше. Введем обозначения и названия для некоторых из них

v_{ij} – емкость i-го сообщения в j-й однородной цепи;

V_j – (емкостной) объем j-й однородной цепи;

V – (емкостной) объем массива данных;

v_{gj} – средняя геометрическая емкость сообщений в j-й однородной цепи;

v_g – средняя геометрическая емкость сообщений в массиве данных;

$c_{ij}=\log_2\upsilon_{ij}$ – размер i-го сообщения в j-й однородной цепи;

$c_j=\log_2\upsilon_{gj}$ – средний размер сообщений j-й однородной цепи;

$c=\log_2\upsilon_g$ – средний размер сообщений в массиве данных;

$C_j=n_j\cdot\log_2 v_{gj}$ – (полный) размер j-й однородной цепи;

$C=n\cdot\log_2 v_g$ – (полный) размер массива данных.

Используя представленные характеристики, легко получить следующие выражения для комплексных характеристик строя цепи:

$$V=V_\Delta\cdot V_\tau,\ \upsilon_g=\Delta_g\cdot\tau_g,\ C=G+L,\ c=g+l, \qquad (12)$$

Формула (12) в более подробном представлении имеет вид:

$$c=\sum_{j=1}^{m}\frac{n_j}{n}\cdot\log_2\Delta_{gj}+\sum_{j=1}^{m}\frac{n_j}{n}\cdot\log_2\tau_{gj}, \qquad (13)$$

Для псевдотекстовых регулярных массивов данных, у которых $\Delta_{gj}=\Delta_{aj}=n/n_j$, первое слагаемое в (13) для $n\to\infty$ совпадает с энтропией; при любых других расположениях символов средняя удаленность сообщений в цепи, представленная первыми слагаемыми в (12) и (13), меньше числа идентифицирующих информаций (по М. Мазуру, [6]), т.е. $g < H$, а средний геометрический интервал меньше числа описательных информаций (по М. Мазуру), т.е. $\Delta_g < D$.

Важно заметить, что так же, как и у интервальных характеристик, время вычисления сериальных и емкостных характеристик линейно зависит от размера входных данных (длины последовательности). И, как следствие, они эффективны для обработки длинных последовательностей (больших массивов данных).

Заключение

Очевидно, что интервальные, сериальные и емкостные характеристики строя цепи попарно дополняют друг друга. Однако емкостные характеристики дают «полное» количественное описание массива данных, которое нельзя получить только интервальными или сериальными характеристиками. Полное

описание строя интервальными характеристиками возможно только для «текстовых» массивов данных, где $\tau_{ij} = 1$. Сериальные характеристики дают полное описание таких массивов, которые практически не имеют «точечных» данных, но состоят из длинных повторов (серий) одинаковых данных, где $\Delta_{ij} = 1$.

Дальнейшая разработка средств описания и анализа строя предполагает исследование свойств введенных здесь характеристик, а также возможную их модификацию. Так как представленные сериальные оценки предусматривают неоднократное чтение серий на позиции некоторого отрезка, в дальнейшем предусматривается разработка средств для однократного чтения отрезков с последующими исследованиями и сравнением чувствительности сериальных и «отрезочных» характеристик строя цепи.

Компьютерная обработка больших «текстовых» массивов данных (проза, стихотворения, нотные записи, нуклеотидные последовательности) показала высокую чувствительность интервальных характеристик строя к оригинальному расположению компонентов (букв, слов, нот и т.п.) в длинных и очень длинных кортежах. Предполагается, что сериальные и емкостные характеристики строя обеспечат такие же свойства для массивов данных, содержащих длинные последовательности повторяющихся компонентов. Наконец, отметим, что разрабатываемые средства формального описания и анализа и большой набор его числовых характеристик отражают разнообразные свойства нового абстрактного объекта – строя информационной цепи.

ЛИТЕРАТУРА

1. Gumenyuk A., Kostyshin A., Simonova S. An approach to the research of the structure of linguistic and musical texts // Glottometrics. 2002. No. 3. P. 61–69.
2. Гуменюк А.С., Поздниченко Н.Н., Шпынов С.Н., Родионов И.Н. О средствах формального анализа строя нуклеотидных цепей // Математическая биология и биоинформатика. 2013. Т. 8, № 1. С. 373–397. URL: http://www.matbio.org/article.php?journ_id=15&id=158 (дата обращения: 15.04.2016).
3. Вентцель Е.С. Исследование операций: задачи, принципы, методология. М. : Наука, 1988. 208 с.
4. Nair A.S., Mahalakshmi T. Visualization of genomic data using inter-nucleotide distance signals // Proc. of IEEE Genomic Signal Processing. Bucharest, 2005. URL:http://www.ece.iit.edu/~biitcomm/research/references/Achuthsankar%20S%20Nair/Visualization%20of%20genomic%20data%20using%20inter-nucleotide%20distance%20signals.pdf (access date: 21.12.17).
5. Afreixo V., Bastos C.A.C., Pinho A.J., Garcia S.P., Ferreira P.J.S.G. Genome analysis with inter-nucleotide distances // Bioinformatics. 2009. V. 25 (23). P. 3064–3070.
6. Мазур М. Качественная теория информации. М. : Мир, 1974. 240 с.

Гуменюк Александр Степанович, канд. техн. наук, доцент. E-mail: gumas45@mail.ru
Омский государственный технический университет

Поступила в редакцию 2 октября 2017 г.

Gumenyuk Alexander S. (Omsk State Technical University, Russian Federation).

Serial and complex description of the order of components in data sets.

Keywords: order; order characteristics; remoteness; spread of series; inter-nucleotide distance.

DOI: 10.17223/19988605/42/4

The paper describes the concept of «order of data sequence», which is defined as a special kind of tuple – «vector of order». The components of order are integer numbers i that are not more than its length n; first encountered different numbers $j \le m \le n$ are increasing by one.

The works representing the means of formal description and analysis of order of data sequences, considered such long tuples (symbolic sequences, data sets), in which separate different components throughout the chain nearly always alternate, and the series (of the same elements arranged in row) are rare and short. Computer processing of large «text» data sets (prose, poems, musical compositions, nucleotide sequences) showed high sensitivity of characteristics of order to the arrangement of components (letters, words, notes, etc.) in long and very long tuples. The proposed means suggest representation of symbolic sequence with its order. The result of decomposition of order is congeneric sequences in each of which places occupied by similar elements (in amount of n_j) are marked with integer number and all other positions are empty. In these congeneric sequences interval between the nearest occupied positions is used as basic measure and calculated as Δ_{ij} – number of empty positions plus one.

This paper discusses means for analysis of ordered data sets, which are mainly represented by alternating series of identical messages. This may be, for example, digitized images, sequences of measured values, etc. These means are represented by a set of «serial» characteristics of order, which are defined, named, marked in a similar way to the previously introduced «interval» characteristics

of order. Series length is used as basic measured value and its size is calculated as a number of occupied places in a row in the *j*-th congeneric sequence. The length of series is marked τ_{ij} (*i* is the number of element in the *n*-th partial congeneric sequence). Below, some of the serial characteristics of the system are given.

Serial volume of the *j*-th congeneric sequence is defined as (1), and serial volume of the complete sequence is defined as (2):

$$V_{\tau_j} = \prod_{i=1}^{n_j} \tau_j \,, \tag{1}$$

$$V_{\tau} = \prod_{j=1}^{m} V_{\tau_j} \,. \tag{2}$$

The total spread of all series of the *j*-th congeneric sequence is defined as (3), and the total spread of all the series in the complete sequence is defined as (4):

$$L_j = n_j \cdot \log_2 \tau_{gj} \,, \tag{3}$$

$$L = n \cdot \log_2 \tau_g \,. \tag{4}$$

where τ_{gj} is the geometric mean length of the series in the *j*-th congeneric seqience; τ_g is the geometric mean length of the series in the complete sequence; $l_j = \log_2\tau_{gj}$ is the average spread of the series in the *j*-th congeneric sequence; $l = \log_2\tau_g$ is the average spread of the series in the complete sequence.

The «complete» serial description of an order is defined by the following distributions: $\{L_j\}$, $\{<l_j, n_j>\}$.

The order of sequence containing, in addition to «sparse» data also a series of elements described by complexes of numerical characteristics and their distributions, taking into account both intervals Δ_{ij} and remoteness $g_{ij} = \log_2\Delta_{ij}$ and length of series τ_{ij} and their spread $l_{ij} = \log_2\tau_{ij}$. Description of order of such sequence also requires preliminary decomposition into congeneric sequences. It is possible to carry out a separate interval and a serial description of an order.

The concept of capacity of *i*-th message $v_{ij} = \Delta_{ij}\,\tau_{ij}$ is introduced. Based on the capacity of messages set of «capacitive» characteristics of an order is defined, among which are following:

$$V = V_{\Delta} \cdot V_{\tau} \,,\ \upsilon_g = \Delta_g \cdot \tau_g \,,\ C = G + L \,,\ c = g + l \,,$$

where *V* and V_Δ are complete and interval volume of order; *G* is the total remoteness (depth) of order; Δ_g is the geometric mean interval and *g* is the average remoteness of messages in a complete sequence; $C = \log_2 V$ is the complete size of the data array; $c = \log_2\upsilon_g$ is the average size of messages in the data array.

Developed tools and a large set of numerical characteristics reflect a variety of properties of the new abstract object – the order of information chain.

REFERENCES

1. Gumenyuk, A., Kostyshin, A. & Simonova, S. (2002) An approach to the research of the structure of linguistic and musical texts. *Glottometrics*. 3. pp. 61–69.
2. Gumenuk, A.S., Pozdnichenko, N.N., Shpynov, S.N. & Rodionov, I.N. (2013) Formal analysis of structures of nucleotide chains. *Matematicheskaya biologiya i bioinformatika – Mathematical Biology and Bioinformatics*. 8(1). pp. 373–397. [Online] Available from: http://www.matbio.org/article.php?journ_id=15&id=158. (Accessed: 15th April 2016). (In Russian).
3. Ventsel, E.S. (1988) *Issledovanie operatsiy: zadachi, printsipy, metodologiya* [Operations research: tasks, principles, methodology]. Moscow: Nauka.
4. Nair, A.S. & Mahalakshmi, T. (2005) Visualization of genomic data using inter-nucleotide distance signals. *Proc. of IEEE Genomic Signal Processing*. [Online] Available from: http://www.ece.iit.edu/~biitcomm/research/references/Achuthsankar%20S%20Nair/Visualization%20of%20genomic%20data%20using%20inter-nucleotide%20distance%20signals.pdf (Accessed: 21st December).
5. Afreixo, V., Bastos, A.C., Pinho, A.J., Garcia, S.P. & Ferreira, P.J.S.G. (2009) Genome analysis with inter-nucleotide distances. *Bioinformatics*. 25(23). pp. 3064–3070. DOI: 10.1093/bioinformatics/btp546
6. Mazur, M. (1974) *Kachestvennaya teoriya informatsii* [Qualitative information theory]. Moscow: Mir.