ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ
(Вычислительные системы)
2002 год Выпуск 171

УДК 519.764

ВЫЯВЛЕНИЕ АНОМАЛИЙ В РАСПРЕДЕЛЕНИИ СЛОВ ИЛИ СВЯЗНЫХ ЦЕПОЧЕК СИМВОЛОВ ПО ДЛИНЕ ТЕКСТА [1]

В.Д.Гусев, Л.А.Немытикова, Н.В.Саломатина

Введение

Задача автоматического смыслового сжатия текстов давно интересует лингвистов и специалистов в области информационного поиска [1–3]. Важным элементом ее является оценка значимости того или иного слова, предложения или отдельного фрагмента текста. Значимость слов, т.е. единиц низшего уровня, чаще всего оценивается статистически (используется информация о частоте их встречаемости в обрабатываемом тексте). Однако уже неоднократно отмечалось, что очень важной является и *позиционная информация*, т.е. информация о местах вхождения заданного слова в текст. Позиционная информация в явном виде фигурирует в некоторых структурах данных, предназначенных для анализа символьных последовательностей (см. [4, гл. 9]), процедурах их факторизации (сегментации) [5] и количественных характеристиках [6]. Особый интерес представляют слова, демонстрирующие аномалии в позиционном распределении. Обычно они оказываются более значимыми, чем слова, распределенные равномерно (примером последних являются служебные слова в естественном языке).

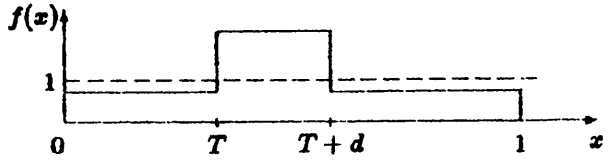
[1]Работа выполнена при финансовой поддержке РФФИ, проект 00-06-80420.

Целью работы является разработка методики *выявления неравномерностей позиционного распределения* с помощью различных статистических критериев. Попутно решается задача *классификации выявляемых аномалий*. Объектом исследования служат не только тексты на естественном языке, но и неструктурированные последовательности (нуклеотидные и аминокислотные). Для последних актуальной является задача формирования словарей "значимых слов", где под словом понимается произвольная цепочка символов [7,8].

Известные алгоритмы выявления неравномерностей в распределении слов или связных цепочек символов по длине текста (см., например, [9–11) ориентированы преимущественно на обнаружение *кластеров* — скоплений однородных элементов в ограниченном участке последовательности. Взятые нами за основу сканирующие статистики [12,13], как будет показано далее, позволяют *при соответствующем подборе параметров* обнаруживать *более широкий спектр закономерностей*, в частности, аномально длинные *"зазоры"* (участки, не содержащие заданного слова), *"изолированные точки"* (слова), *"аналоги разделителей"* (слова, не допускающие слишком большого сближения и (или) разрежения), *периодичности*.

1. Статистики для выявления неравномерностей позиционного распределения

Задача о выявлении неравномерностей позиционного распределения элементов символьной последовательности в непрерывном случае сводится к изучению различных схем расстановки точек на линии. Простейшая постановка выглядит следующим образом. Пусть $x_1, x_2, \ldots x_N$ будет случайный набор точек из единичного интервала $(0, 1]$. Требуется проверить гипотезу о равномерности (H_0) против альтернативы (H_1), связанной с тем или иным типом кластеризации. Обозначим число точек, попавших в интервал $(x, x + d]$ через $n(x, x + d)$. Если отклонение от равномерности локализовано в ограниченной области и может быть описано ступенчатой функцией



$$f(x) = \begin{cases} 1/\{1+(\mu-1)d\}, & 0 < x \le T, \\ \mu/\{1+(\mu-1)d\}, & T < x \le T+d, \\ 1/\{1+(\mu-1)d\}, & T+d < x \le 1, \end{cases}$$

где $\mu > 1$ и T — неизвестны, а d — известно, то обобщенное отношение правдоподобия, используемое для проверки гипотезы, приводит к статистике вида

$$n(d) = \sup_{0 \le a \le 1-d} n(x, x+d), \qquad (1)$$

фиксирующей максимальное число точек, попавших в интервал длины d при всевозможных расположениях этого интервала внутри единичного отрезка [12]. Статистика (1) называется сканирующей, поскольку вычисление ее ведется путем подсчета числа точек, попадающих в окно ширины d, скользящее вдоль отрезка.

Со статистикой $n(d)$ тесно связана другая — $d(n)$, фиксирующая длину минимального интервала, содержащего ровно n точек ($2 \le n \le N$). Связь двух статистик определяется соотношением: $\Pr(d(n) \le p) = \Pr(n(p) \ge n)$. Из чисто алгоритмических соображений мы будем пользоваться статистикой $d(n)$, дополнив ее еще одной — $D(n)$, определяющей длину максимального интервала, содержащего ровно n точек. Последняя статистика удобна для выявления "тэпов" и разрежений в расстановке точек на отрезке.

Распределение статистики $d(n)$ при нулевой гипотезе, т.е. формулы для вычисления $\Pr(d(n) \le p)$ для всех n, N и рациональных p, получены в [14] и частично затабулированы, но они не покрывают широкий диапазон возможных применений сканирующей статистики. Эти формулы достаточно сложны, поэтому для отдельных случаев получены более простые аппроксимации (см., например, [15]).

Ряд статистик предложен для выявления характера распределения точек на плоскости. Они легко могут быть переформулированы для одномерного случая. Упомянем о двух из них.

Статистика Эберхардта [16] оперирует расстояниями до ближайшего соседа. Пусть $O_1, O_2, \ldots, O_N$ — точки, случайно распределенные внутри области S; x_i — расстояние от точки i до ближайшего соседа, $1 \leq i \leq N$. Статистика имеет вид

$$A = N \cdot \frac{\sum\limits_{i=1}^{N} x_i^2}{\left(\sum\limits_{i=1}^{N} x_i\right)^2} \quad . \tag{2}$$

Критические значения для нее затабулированы в [16] ($10 \leq N \leq 1000$). Сопоставление на модельных данных статистики (2) с другими, рекомендованными в литературе, показало ее хорошую чувствительность к обнаружению скоплений точек и регулярностей в их расположении.

Индекс Моришиты — еще одна статистика, используемая для обнаружения кластеризации точек на плоскости [17]. Способ ее реализации напоминает построение серии гистограмм со все более мелким дроблением исходной области. В качестве таковой рассматривается квадрат со стороной L, который последовательно разбивается на 4, 16, 64 и т.д. ячеек. Размер ячейки определяется параметром $S = \frac{L}{\delta}$, $\delta = 1, 2, 4, 8, \ldots$, $Q = \delta^2$ — число ячеек на шаге δ, n_i — число точек, попавших в ячейку с номером i, N — общее число точек.

Индекс Моришиты задается выражением

$$I_\delta = Q_\delta \cdot \frac{\sum\limits_{i} n_i(n_i - 1)}{N(N-1)} \quad . \tag{3}$$

При случайном равномерном распределении точек $I_\delta \to 1$, при регулярном равномерном — $I_\delta < 1$, при кластеризации $I_\delta > 1$.

2. Схема анализа позиционного распределения заданной цепочки по длине текста

Возьмем за основу статистики $d(n)$ и $D(n)$, описанные в предыдущем разделе. Адаптируя их применительно к символьным

последовательностям, будем придерживаться терминологии, используемой в [18]. Пусть x — заданная цепочка (символ) текста, $F(x)$ — частота ее (его) вхождения в текст, n — фиксированное число последовательных вхождений x в текст, $2 \le n \le F(x)$. Будем рассматривать следующие статистики:

$d1(n)$ — длина минимального фрагмента последовательности, содержащего n вхождений цепочки x (первый минимум);

$d2(n)$ — длина следующего по величине фрагмента, содержащего n вхождений цепочки x (второй минимум); по определению $d2(n) > d1(n)$;

$D1(n)$ — длина максимального интервала, начинающегося и заканчивающегося цепочкой x и содержащего ровно n вхождений этой цепочки (первый максимум);

$D2(n)$ — второй максимум ($D2(n) < D1(n)$);

$P(n)$ — длина списка фрагментов, на которых достигается соответствующий минимум или максимум в любой из предыдущих статистик. Этот параметр в непрерывном случае (точки на отрезке) не рассматривается. В дискретной постановке (символьные последовательности) он вполне уместен и трактуем. В частности, весьма информативными оказываются аномально короткие и аномально длинные списки;

$D_{\text{нач}}(x)$ — длина максимального начального фрагмента последовательности, не содержащего цепочку x;

$D_{\text{кон}}(x)$ — длина максимального конечного фрагмента последовательности, не содержащего цепочку x.

Две последние статистики введены как дополнение к $D1(n)$. С помощью статистики $D1(n)$ при $n = 2$ определяется длина максимального "гэпа" между двумя последовательными вхождениями цепочки x в текст. Однако аномально длинными (а, следовательно, информативными) могут оказаться начальный и конечный фрагменты текста, не содержащие x. Формально они не выявляются с помощью статистики $D1(n)$.

Предлагаемая нами схема выявления аномалий в позиционном распределении конкретной цепочки x по длине текста учитывает следующие факторы:

1) аномалия может иметь место при любом значении $2 \le n \le F(x)$, т.е. необходим перебор по параметру n;

2) точные формулы для распределения сканирующей статистики очень сложны для табулирования; многочисленные аппроксимации тоже достаточно сложны и имеют ограниченные области применимости. В частности, для вариантов сканирующих статистик, представленных в [18] и адаптированных применительно к символьным последовательностям, получены при нулевой гипотезе лишь асимптотические распределения ($F(x) \to \infty$), а для ограниченных по длине последовательностей рекомендовано выбирать $n \ll F(x)$;

3) некоторые статистики (типа $P(n)$) никем не исследовались. В этой ситуации для оценки значимости отклонения вычисляемых статистик от гипотезы H_0 (равномерность) целесообразно прибегнуть к имитационному моделированию.

Схема выявления позиционных аномалий выглядит следующим образом.

1. Фиксируем конкретную цепочку x и определяем частоту ее встречаемости в тексте $F(x)$.

2. Для каждого $2 \le n \le F(x)$ вычисляем:

а) значения интересующих нас сканирующих и связанных с ними статистик в анализируемом тексте;

б) с помощью имитационного моделирования для заданных N (длина текста), x, $F(x)$ и n оцениваем распределения этих статистик при гипотезе H_0. С этой целью путем многократного перемешивания исходного текста T формируются m его рандомизированных аналогов со случайным равномерным распределением цепочки x по длине текста. При малых $|x|$ целесообразно проводить перемешивание с сохранением l-граммного состава, $l = 1, 2, \ldots |x|$. По полученной подборке вычисляются оценки минимального, максимального и среднего значений каждой статистики (соответственно $S_{\min}$, $S_{\max}$ и $\overline{S}$), а также характеристики разброса (среднеквадратичные отклонения $\widehat{\sigma}$). Приемлемый диапазон значений m (объем подборки) может колебаться в пределах от 100 до 1000.

3. Выявляются аномальные отклонения наблюдаемых значений статистик ($S_{\text{набл.}}$) от оценок, полученных в имитационном эксперименте. Постулируется, что аномалия имеет место, если

выполняется хотя бы одно из следующих условий:

$$S_{\text{набл.}} \le S_{\min} \quad (\text{соответственно} \quad S_{\text{набл.}} \ge S_{\max}), \tag{4}$$

$$S_{\text{набл.}} \le \overline{S} - 3\widehat{\sigma} \quad (\text{соответственно} \quad S_{\text{набл.}} \ge \overline{S} + 3\widehat{\sigma}). \tag{5}$$

Иногда полезно учитывать обе характеристики разброса сразу, например, использовать правила вида

$$\begin{aligned} &(S_{\text{набл.}} \le S_{\min}) \quad \& \quad (S_{\text{набл.}} \le \overline{S} - 2\widehat{\sigma}), \\ \text{или} \quad &(S_{\text{набл.}} \ge S_{\max}) \quad \& \quad (S_{\text{набл.}} \ge \overline{S} + 2\widehat{\sigma}). \end{aligned} \tag{6}$$

Они хорошо работают, когда распределение сканирующей статистики сильно асимметрично, что обычно имеет место при малых n и значительных $F(x)/N$. Например, при $N = 637$, $F(x) = 55$ и $n = 3$ в одном из экспериментов имели $S_{\text{набл.}} = d1(3) = 3$ (минимально возможное значение). Имитационное моделирование дало следующие результаты: $S_{\min} = 3$, $S_{\max} = 9$, $\overline{S} = 4{,}2$, $\widehat{\sigma} = 1{,}2$, т.е. условие (4) выполняется, а условие (6) — нет. Иными словами, минимум достигается, но вероятность этого события достаточно велика, поэтому оно не может рассматриваться как аномальное и анализ дисперсии это подтверждает. Бывают и обратные ситуации, когда выполнено условие (5), но не выполнено (4). Для них можно сформулировать аналог (6), смягчив условие $S_{\text{набл.}} \le S_{\min}$.

4. Выявленные позиционные аномалии интерпретируются в содержательных терминах ("кластер", "тэп" и т.п.). Правильная интерпретация позиционных аномалий очень важна и тесно связана с классификацией наблюдаемых закономерностей. Эти вопросы рассматриваются в следующем разделе и иллюстрируются конкретными примерами из обработки аминокислотных последовательностей и текстов на естественном языке.

3. Описание экспериментов. Интерпретация результатов.

3.1. Исходные данные. Рассматривались данные трех типов: подборка аминокислотных последовательностей из разных белковых семейств, подборка нуклеотидных последовательностей

(промоторные области генов гормона роста) и тексты на естественном языке (оригинал известной книги Алана А. Милна "Винни-Пух" на английском языке) и два его перевода на русский язык: Б. Заходера (1960 г.) и В. Вебера, Н. Рейн (1999 г.)).

Для первого типа данных анализировалось распределение отдельных аминокислот (каждой из 20) по длине последовательностей. *Для второго типа* данных исследовалось распределение биграмм (двухэлементных цепочек) по длине нуклеотидных последовательностей. Алфавит биграмм включает в себя $4^2 = 16$ элементов. Аномалии в биграммных характеристиках часто связаны с существованием нестандартных форм ДНК. В этих экспериментах использовалось перемешивание с сохранением биграммного состава.

Для третьего типа данных анализировалось распределение отдельных слов по длине текста. Учет словоизменительной парадигмы в русских текстах проводился путем отбрасывания окончаний. Данная процедура, осуществляемая в автоматическом режиме, не гарантирует 100%-й точности из-за омонимии окончаний ("ими", "ми", "и"), а также совпадения окончаний с частью основы ("ет" — как окончание в глагольных формах и как часть основы в слове "привет"). Однако для целей нашего исследования возникающими погрешностями можно пренебречь.

3.2. Описание экспериментов. Целью экспериментов являлась оценка неравномерности позиционного распределения элементарных структурных единиц в текстах различной языковой природы, выявление наиболее характерных аномалий в позиционном распределении и их взаимосвязи, а также выработка рекомендаций по интерпретации результатов обработки.

Были реализованы все три статистики, описанные в разделе 1 (см. (1), (2), (3))[2]. Для оценивания значимости выявляемых аномалий во всех случаях использовалось имитационное моделирование. Число рандомизированных аналогов исходной последовательности в имитационных экспериментах менялось от 100 (основной вариант) до 1000. Наиболее чувствительными к выявлению различных позиционных аномалий, обусловленных в

[2]Две последние были переформулированы для одномерного случая.

общем случае не только кластеризацией, но и другими эффектами (периодичностью, запретами на сближение и т.п.), оказались сканирующие статистики. Причины подобной "универсальности" кроются, по-видимому, во взаимосвязи наблюдаемых аномалий, которая будет проиллюстрирована далее на различных примерах.

По степени предпочтительности на второе место можно поставить статистику (3) (индекс Моришиты), а на третье — статистику Эберхардта. То, что последняя уступает по своим возможностям сканирующим статистикам, видимо, объясняется недоиспользованием имеющейся позиционной информации (учитываются лишь расстояния от каждой точки до ближайшего соседа). Индекс Моришиты по построению ближе к сканирующим статистикам, но в нем отсутствует, как таковой, элемент сканирования, позволяющий точно локализовать (не пропустить) закономерность, к тому же процедура дробления (уменьшения величины интервала) носит более грубый характер, чем в сканирующих статистиках, что также может привести к пропуску аномалии.

По причинам, изложенным выше, основное внимание далее будет уделено сканирующим статистикам.

3.3. Интерпретация результатов. Пусть x — анализируемый объект (символ, цепочка символов, слово, цепочка слов), распределение которого по длине текста T мы исследуем, $l(x)$ — длина цепочки x (в символах или словах), $F(x)$ — частота встречаемости x в T, $l(T)$ —длина текста T (в символах или словах).

1. Пусть n — целое число, не меньшее 2. Если $d1(n) = n \cdot l(x)$ и аномально мало, имеет место *аномально длинная серия* из повторяющихся элементов x. Здесь предполагается использование правила (4) со строгим неравенством либо правила (6) со значением $\widehat{\sigma} \neq 0$.

2. Если $d1(2)$ аномально велико (что, как минимум, означает отсутствие тандемных повторов, т.е. выполнение неравенства $d1(2) > 2 \cdot l(x)$), это можно трактовать как запрет на чрезмерное сближение элементов x. Таким свойством обладают *разделители* между словами (или более крупными структурными элемен-

тами) в естественном языке. Поэтому объекты с указанными свойствами будем называть *"аналогами разделителей"*.

3. Если $D1(2)$ аномально мало, это можно трактовать как *запрет на чрезмерное удаление*.

4. Если имеет место (2) и (3) одновременно, можно говорить о *"сверхравномерном"* распределении пар смежных элементов по длине текста. В общем случае, если выполняются условия $\{d1(n)$ — аномально велико, а $D1(n)$ — аномально мало, $n>2\}$, можно говорить о "сверхравномерном" распределении n-ок из объектов x по длине текста.

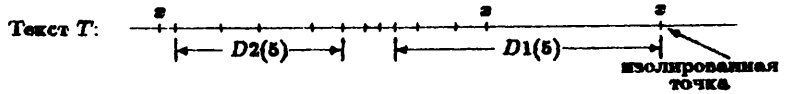
5. Если $D1(2)$ аномально велико, имеет место *"гэп"* — участок текста, не содержащий вхождений элемента x (длина его равна $D1(2) - 2 \cdot l(x)$). Аномально большие значения $D_{\text{нач.}}(x)$ и $D_{\text{кон.}}(x)$ фиксируют "гэпы" в начале и конце текста. Если определить среднее расстояние между вхождениями элемента x в текст как $u(x) = \dfrac{l(T)}{F(x)}$, длину "гэпа" удобно характеризовать безразмерной величиной $\Delta(x) = \dfrac{D_1(2) - 2l(x)}{u(x)}$. С помощью этого показателя можно сравнивать по значимости гэпы в последовательностях разной длины и для разных цепочек x, отличающихся значением частоты $F(x)$.

6. Если $d1(n)$ аномально мало при некотором не слишком малом значении n (например, $n \geq 5$), имеет место кластеризация цепочек x. Как и в случае гэпов, значимость кластеров удобно характеризовать безразмерной величиной $\delta(x) = \dfrac{u(x)}{v(x)}$, где $u(x)$ — введенное выше среднее внутритекстовое расстояние между вхождениями x, а $v(x) = \dfrac{d1(n)}{n}$ — среднее внутрикластерное расстояние между вхождениями x.

Если $n = F(x)$ и $d1(n)$ аномально мало, это означает, что кластеризованы все вхождения x, т.е. значительные фрагменты текста свободны от x (левый конец, правый, или оба вместе).

7. Аномально малые значения $D2(n)$ часто (но не всегда) связаны с наличием *"изолированной точки"* (элемента x, удаленного на значительное расстояние от ближайшего соседа). Малость $D2(n)$ обеспечивают не любые изолированные точки, а лишь те,

что расположены *в начале или конце текста* (см. схематический рисунок ниже).



Здесь точками указаны места вхождения элемента x в текст, а для случая $n = 5$ выделены интервалы, соответствующие первому и второму максимуму. Нетрудно видеть, что при наличии изолированной точки (начальной или концевой) интервал $D1(n)$ включает ее в качестве своего левого или правого конца. Интервал $D2(n)$ в этом случае всегда располагается в области сгущения точек, поэтому часто оказывается аномально малым. Различия между $D1(n)$ и $D2(n)$ могут оказаться очень большими.

Если бы на правом конце были две кластеризованные и удаленные от остальных точки, первый и второй максимумы были бы достаточно велики и близки друг другу, а аномально малым стал бы третий максимум, нами не оцениваемый.

8. Если список фрагментов текста, на которых сканирующая статистика аномальна при фиксированных x и n, слишком велик, т.е. велико $P(n)$, это означает, что вхождения цепочки x в текст, задаваемые списком позиций $Pos(x) = \{i_1, i_2, \dots i_{P(n)}\}$, $1 \le i_k \le l(T) - l(x) + 1$, характеризуются определенной регулярностью. Она проявляется в том, что из списка $Pos(x)$ можно выделить две подпоследовательности длины $P(n)$:

$$Pos^{(1)}(x) = \{j_1^{(1)}, j_2^{(1)}, \dots, j_{P(n)}^{(1)}\}$$

и

$$Pos^{(2)}(x) = \{j_1^{(2)}, j_2^{(2)}, \dots, j_{P(n)}^{(2)}\}$$

такие, что

$$j_k^{(2)} - j_k^{(1)} = d(n) - l(x) \quad (\text{или } D(n) - l(x)) \tag{7}$$

для всех $1 \le k \le P(n)$ и $j_k^{(1)}, j_k^{(2)} \in Pos(x)$. Иными словами, в $Pos^{(1)}(x)$ входят адреса элементов x, с которых начинаются аномальные фрагменты, а в $Pos^{(2)}(x)$ — адреса элементов x, которыми они заканчиваются. Поскольку по определению и минимальный и максимальный фрагменты, содержащие ровно n

вхождений элемента x, начинаются и заканчиваются этим элементом, многократное ($P(n)$ раз) достижение минимума или максимума приводит к появлению указанной регулярности.

Обозначим элемент текста T, стоящий в i-й позиции через $T[i]$, $1 \le i \le l(T)$, а фрагмент текста, расположенный в позициях с i-й по j-ю — через $T[i:j]$. Из сказанного выше следует, что номерам позиций, представленных в $Pos^{(1)}(x)$ и $Pos^{(2)}(x)$, соответствуют в тексте T одинаковые элементы алфавита (символы — для слитного текста или слова — для структурированного), т.е. имеет место $T[j_k^{(1)}] = T[j_l^{(2)}] = x[1]$ для всех $1 \le k,l \le P(n)$, где $x[1]$ — начальный элемент цепочки x. Кроме того из (7) следует, что $j_k^{(1)} - j_{k-1}^{(1)} = j_k^{(2)} - j_{k-1}^{(2)}$, $1 \le k \le P(n)$, откуда вытекает, что $j_{P(n)}^{(1)} - j_1^{(1)} = j_{P(n)}^{(2)} - j_1^{(2)}$. Если выровнять фрагменты $T[j_1^{(1)} : j_{P(n)}^{(1)}]$ и $T[j_1^{(2)} : j_{P(n)}^{(2)}]$ друг под другом:

$$\begin{array}{lll} T[j_1^{(1)}]\ldots & T[j_2^{(1)}]\ldots & T[j_{P(n)}^{(1)}]\ldots \\ \| & \| & \| \\ T[j_1^{(2)}]\ldots & T[j_2^{(2)}]\ldots & T[j_{P(n)}^{(2)}]\ldots, \end{array}$$

то в позициях, помеченных знаком $\|$, будут стоять одинаковые элементы. В промежутках между ними соответствующие друг другу элементы верхней и нижней строки могут как совпадать, так и не совпадать. Если совпадают все элементы, имеет место *совершенный повтор* (часто в тандемном варианте), в противном случае — *несовершенный повтор*. Если число совпадающих элементов во фрагментах $T[j_1^{(1)} : j_{P(n)}^{(1)}]$ и $T[j_1^{(2)} : j_{P(n)}^{(2)}]$ невелико по сравнению с их длиной, уместнее говорить не о повторе, а лишь о наличии общей "разрывной" L-граммы, где $L = P(n)$. Заметим также, что в общем случае фрагменты $T[j_1^{(1)} : j_{P(n)}^{(1)}]$ и $T[j_1^{(2)} : j_{P(n)}^{(2)}]$ могут пересекаться.

4. Примеры позиционных аномалий. Их взаимосвязь

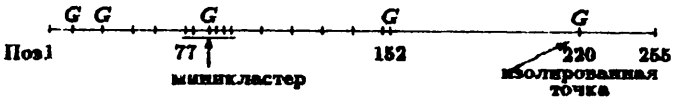
Приведенную в п. 3.3 интерпретацию позиционных аномалий можно одновременно рассматривать как их классификацию.

Ниже будут приведены наиболее характерные примеры позиционных аномалий в последовательностях разной природы.

ПРИМЕР 1. Рассматривается 5′-фланкирующий район гена гормона роста у "Lates calcarifer (barramundi perch)". Длина последовательности — 813 нуклеотидов (заканчивается перед инициирующим кодоном). Анализируется распределение биграммы AC по длине последовательности; $F(x = AC) = 60$. Оно иллюстрирует сильную аномалию типа (1) — см. п. 3.3: $d1(n = 15) = 30$ (−), поз. 632, т.е. минимальный интервал, содержащий 15 вхождений биграммы AC, равен 30 и начинается в позиции 632 — примерно в 180 нк от точки начала трансляции. Этот интервал аномально мал (здесь и далее для облегчения восприятия аномально малые интервалы будем сопровождать знаком (−), а аномально большие – знаком (+)). О силе проявления закономерности свидетельствуют результаты имитационного моделирования: $S_{\min} = 62$; $S_{\max} = 191$; $\overline{S} = 121{,}4$; $\widehat{\sigma} = 20{,}5$. Здесь $S_{\min}$ и $S_{\max}$ определяют диапазон изменения значений статистики $d1(n = 15)$ в имитационном эксперименте, $\overline{S}$ — среднее ее значение , $\widehat{\sigma}$ — среднеквадратичное отклонение. Нетрудно видеть, что наблюдаемая закономерность — *серия из 15 тандемных повторов биграммы* AC — с большим "запасом" классифицируется как аномальная по любому из критериев (4)–(6) (см. раздел 2). Известно,что тандемные повторы, расположенные в межгенных областях и интронах, могут оказывать влияние на активность соседних генов. Некоторые механизмы, реализующие модулирующую функцию тандемных повторов при протекании тех или иных генетических процессов, описаны в [19].

ПРИМЕР 2. Рассматривается аминокислотная последовательность гена HOXD4, поисковый номер p09016 в базе данных Swiss-Prot, длина последовательности $l(T) = 255$. Анализируется позиционное распределение глицина (G), $F(G) = 27$. Оно характеризуется наличием *"гэпа"* (см. подпункт 5 в п. 3.3.), который является аномальным по любому из рассмотренных критериев: $D1(n = 2) = 69$ (+); поз. 152; $S_{\min} = 17$; $S_{\max} = 61$; $\overline{S} = 33{,}7$; $\widehat{\sigma} = 8{,}4$. Длина "гэпа" $= D1(2) - 2 = 67$; $\Delta(G) \approx 7{,}1$. "Гэп" расположен во второй половине последовательности (поз. 153–219) и правее его имеется лишь одно вхождение G (в 220

поз.), которое является *"изолированной точкой"* (см. рисунок ниже).



Формально о наличии изолированной точки сигнализируют аномально малые значения $D2(n = 10, 12\text{–}26)$. Все соответствующие этой статистике интервалы при указанных значениях n лежат левее "гэпа".

Аномалия в виде "гэпа" является довольно сильной. Как следствие, аномально велики и значения $D1(n = 4\text{–}7)$, т.е. максимальные интервалы, содержащие 4, 5, 6 и 7 точек. Все они заканчиваются в 220-й позиции. Значение $D1(n = 3)$ не является аномальным, т.к. из-за наличия биграммы GG в 151-й позиции оно лишь на 1 превышает значение $D1(n = 2)$.

Распределение глицинов в левой части последовательности демонстрирует наличие миникластеров. Один из них имеет вид: $GHGQEPGGPGG$. Его характеристики: $d1(n = 6) = 11$ (−); поз. 77; $S_{min} = 11$; $S_{max} = 32$; $\overline{S} = 22$; $\widehat{\sigma} = 5{,}6$. Это не слишком сильная аномалия, поскольку, например, критерию (5) (раздел 2) она уже не удовлетворяет.

ПРИМЕР 3. Текст T — перевод на русский язык книги Алана А. Милна "WINNIE–THE–POOH", сделанный Б. Заходером. Длина текста $l(T) = 39806$ словоупотреблений. Анализируется распределение по длине текста междометия "АГА", частота встречаемости которого в тексте $F(x = \text{"АГА"}) = 19$. Имеет место очень сильная (при всех $3 \le n \le 16$) *кластеризация* этого слова:

$d1(n = 10) = 190$ (−);	поз.10917;	$S_{min} = 4365$;	$S_{max} = 16960$;
	$\overline{S} = 12110$;	$\widehat{\sigma} = 2388$;	$\delta(x) = 110$;
$d1(n = 16) = 2024$ (−);	поз.10917;	$S_{min} = 13566$;	$S_{max} = 33339$;
	$\overline{S} = 25850$;	$\widehat{\sigma} = 3550$;	$\delta(x) = 16{,}6$.

Больше половины вхождений этого слова (10 из 19) расположены в интервале, содержащем 190 словоупотреблений (на одной странице текста), еще 6 — в той же главе (из них 5 снова образуют кластер). Подобная кластеризация совершенно нетипична

для слов, относящихся к разряду "служебных" (не несущих самостоятельного значения) и исключаемых на этом основании из рассмотрения при автоматическом анализе содержания текста. Объяснение состоит в том, что в данном конкретном случае слово "АГА" используется не по прямому назначению, а иносказательно, как планируемый заранее и несущий в себе скрытую угрозу и предупреждение вариант ответа на вопрос кенгуру о том, куда делся ее детеныш. Таким образом содержательный статус слова "АГА" резко повышается, оно становится значимым элементом плана похищения Крошки Ру и кластеризация отражает эту метаморфозу. Аномально малый интервал, содержащий 10 из 19 вхождений слова "АГА" в текст, четко выделяет границы локального эпизода, связанного с обсуждением плана похищения.

ПРИМЕР 4.

1. Анализируется распределение артикля "the" в английской версии "Винни-Пуха" (оригинал). Длина текста $l(T) = 49269$, $F(x = \text{"the"}) = 1526$. Имеет место сильная аномалия:

$$d1(n=2) = 3\ (+), \quad \text{поз. } 21023, \ldots \text{(всего 8 интервалов)}, \quad S_{\min} = 2,\ S_{\max} = 2,\ \overline{S} = 2,\ \widehat{\sigma} = 0;$$
$$d2(n=2) = 4\ (+), \quad \text{поз. } 171, \ldots \text{(всего 129 интервалов)}, \quad S_{\min} = 3,\ S_{\max} = 3,\ \overline{S} = 3,\ \widehat{\sigma} = 0.$$

Это очевидный *запрет на сближение* (любые последовательные вхождения артикля разделены, как минимум, одним словом). Обращает на себя внимание равенство нулю оценки дисперсии в имитационном эксперименте. Это означает, что при случайном размещении 1526 элементов x по тексту в любой из 100 рандомизированных последовательностей наблюдались случаи тандемного вхождения артикля ($S_{\min} = 2$).

2. В том же тексте анализируется позиционное распределение слова "chapter" (глава), $F(x = \text{"chapter"}) = 20$. Наблюдаются два типа аномалий:

$$D1(n=2) = 3041\ (-); \quad \text{поз. } 15242; \quad S_{\min} = 4078;\ S_{\max} = 16235;\ \overline{S} = 8859;\ \widehat{\sigma} = 2361;$$
$$d2(n=2) = 1275\ (+); \quad \text{поз. } 4275; \quad S_{\min} = 14;\ S_{\max} = 947;\ \overline{S} = 247;\ \widehat{\sigma} = 177{,}3,$$

т.е. максимальное расстояние между парами соседних точек слишком мало, а минимальное — слишком велико. Эта ситуация характеризует *"сверхравномерное" распределение*, что вполне ассоциируется со значением слова "глава" как наименования раздела книги.

ПРИМЕР 5. Исходный текст — перевод "Винни-Пуха", сделанный В. Вебером и Н. Рейн, $l(T) = 43554$ словоупотреблений. Анализируется распределение по длине текста слова "Тигер" (кличка тигра — одного из персонажей книги); $F(x = \text{"Тигер"}) = 183$. Имеем: $D_{\text{нач.}}(x) = 22685$ (+); $S_{\min} = 886$; $S_{\max} = 2412$; $\overline{S} = 1398{,}4$; $\widehat{\sigma} = 293{,}7$, т.е. первая половина текста не содержит вхождений слова "Тигер". Это очень сильная аномалия, следствием которой является кластеризация данного слова во второй половине текста. Выделяются три крупных кластера:

$d1(n=62)=1952\,(-)$;	поз.22686;	$S_{\min} = 8912$;	$S_{\max} = 13330$;
	$\overline{S} = 11756{,}6$;	$\widehat{\sigma} = 937{,}3$;	$\delta(x) = 7{,}55$;
$d1(n=30)=626\,(-)$;	поз.27666;	$S_{\min} = 2847$;	$S_{\max} = 5602$;
	$\overline{S} = 4627{,}4$;	$\widehat{\sigma} = 545$;	$\delta(x) = 11{,}4$
$d1(n=67)=3512\,(-)$;	поз.33750;	$S_{\min} = 10219$;	$S_{\max} = 14737$;
	$\overline{S} = 12886$;	$\widehat{\sigma} = 976{,}7$;	$\delta(x) = 4{,}55$.

Первый кластер практически полностью накрывает гл. 2, ч. II (ее позиционные координаты: 22682–24676), в которой впервые появляется Тигер. Второй кластер — это эпизод с участием Тигера в гл. 4, ч.II. Третий кластер накрывает конец гл. 6, ч. II (эпизод с участием Тигера) и практически полностью гл. 7, ч. II (ее координаты: 34474–37275). Таким образом, аномальные кластеры в данном случае либо совпадают с крупными структурными единицами текста (главы), либо выделяют значимые подразделы в этих структурных единицах (эпизоды в главах). Сильные аномалии в позиционном распределении указывают на значимость данного слова в плане определения содержания текста. Интересно отметить, что стоящее рядом со словом "Тигер" в частотном упорядочении слово "вот" ($F = 183$) вовсе не демонстрирует позиционных аномалий, что подтверждает статус данного слова как "служебного".

ПРИМЕР 6. Исходный текст — аминокислотная последовательность гена TCH3, поисковый индекс p25071 в базе данных

Swiss-Prot, $l(T) = 324$. Анализируется распределение по длине текста аспарагиновой кислоты D; $F(x = D) = 39$. Имеет место: $d1(n = 13) = 90$ (+); поз. 3, 7, ..., 92 (всего 13 интервалов), на которых реализуется указанный минимум; $S_{min} = 44$; $S_{max} = 87$; $\overline{S} = 66$; $\widehat{\sigma} = 10{,}57$. Интерес в данном случае представляет аномально большая длина списка интервалов ($P(n) = 13$) с одинаковой (минимальной) длиной. Как пояснялось в подпункте 8 п. 3.3., это может быть связано с наличием повторов в тексте. Выравнивая друг под другом цепочки из начальных элементов интервалов ($T[3] = T[7] = \ldots = T[92] = D$) и конечных ($T[92] = T[96] = \ldots = T[181] = D$), непосредственной проверкой убеждаемся, что совпадают и цепочки символов, стоящие в обоих упорядочениях между соответствующими друг другу опорными точками, т.е. $T[4 : 6] = T[93 : 95]$ и т.д. В конечном итоге выявляем наличие периодичности с длиной периода $t = 89$. При этом два периода образуют точный повтор, третий зашумлен, четвертый оборван. Это свидетельствует о дупликативном характере эволюции данного гена.

5. Обсуждение результатов. Основные выводы

5.1. Аномалии в позиционном распределении отдельных слов или цепочек символов в текстах различной языковой природы — не случайность, а *закономерность*. Такие аномалии обычно демонстрируют среднечастотные слова или цепочки символов, которые можно выделить из текста путем задания некоторых порогов. В частности, в текстах на естественном языке верхний порог должен отсекать служебную и общеупотребительную лексику, характеризующуюся распределением близким к равномерному, а нижний порог — редко встречающиеся слова, которые обычно составляют очень значительную (свыше половины) долю объема словаря конкретного текста.

Значимость языковых единиц определяется не только частотой их появления в тексте, но и позиционными характеристиками. Если текст структурирован, что имеет место для естественных языков, наибольшую значимость имеют слова, стоящие в заголовках, а также открывающие и завершающие структурно выделенные подразделы текста (главы, параграфы, абзацы)

[2]. Если текст неструктурирован (ДНК- и аминокислотные последовательности, шифротексты, двоичные компрессированные последовательности и т.п.), о значимости той или иной цепочки символов в значительной мере можно судить по характеру ее распределения вдоль текста. Как правило, цепочки, демонстрирующие яркие позиционные аномалии, являются функционально и/или эволюционно значимыми. Этот вывод справедлив и по отношению к структурированным текстам, где позиционные аномалии могут выявлять элементы структуры, промежуточные по отношению к уже существующим (межфразовые единства, описания отдельных сцен, эпизодов и т.п.). Перспективным также представляется использование позиционного анализа для разделения высокочастотной лексики на служебную и тематическую, поскольку сделать это на основании одних лишь частот слов и даже их значений в общем случае невозможно (см. пример 3).

5.2. Наиболее характерны следующие *типы позиционных аномалий*: *кластеры* (сгущения однотипных элементов в ограниченной области); *"гэпы"* (протяженные участки, не содержащие вхождений заданного элемента); *"свободные концы"* (гэпы, охватывающие начальный и конечный фрагменты текста); *ограничения на попарное сближение*, проявляющиеся, как минимум, в отсутствии тандемных повторов; *"изолированные точки"* (элементы, удаленные на значительное расстояние от ближайших соседей); *сверхравномерно распределенные элементы*, похожие по своим свойствам на формальные разделители; *тандемные повторы* значительной длины (т.е. с аномально большой величиной произведения длины периода на кратность повторений).

5.3. Указанные позиционные аномалии удобно выявлять с помощью различного рода *сканирующих статистик*. Преимущество последних перед другими, предназначенными для той же цели, в том, что по ходу анализа проверяются на аномальность все интервалы, содержащие фиксированное число точек (исследуемых объектов), а само это число принимает все допустимые значения. За счет этого сканирующие статистики выигрывают в чувствительности по сравнению с методами, основанными на выявлении ближайшего соседа или на последовательных разбие-

ниях исследуемого интервала на более мелкие части, когда допускается иерархическое дробление интервалов, но не их сдвиг.

5.4. Приведенные в п. 5.2. типы позиционных аномалий *не являются* полностью *независимыми*. Их связь часто имеет форму своего рода "закона сохранения", обусловленного постоянством числа точек на анализируемом интервале. Если на каком-то значительном по длине участке текста отсутствует интересующий нас объект (т.е. имеет место гэп), с большой вероятностью следует ожидать кластеризации этих объектов в других частях текста. Наличие изолированной точки в начале и конце интервала может послужить причиной не только аномально больших значений $D1(n)$, но и аномально малых $D2(n)$. Если гэп расположен в середине интервала и слева от него находится n_1 точек, а справа — n_2, так что $n_1 + n_2 = F(x)$, то при $n > \max(n_1, n_2)$ могут иметь место одновременно аномально большие значения как $d1(n)$, так и $D1(n)$, поскольку каждый из экстремальных интервалов включает гэп. При $n \to F(x)$ и отсутствии изолированных точек статистики $d1(n)$ и $D1(n)$ будут сближаться, поскольку количество перебираемых интервалов, содержащих ровно n точек, уменьшается (оно равно $F(x) - n + 1$), степень пересекаемости интервалов увеличивается, и, как следствие, разница между минимальным и максимальным интервалом нивелируется.

Кроме перечисленных (и ряда других) взаимосвязей между разными статистиками существуют еще "наследственные" взаимосвязи между значениями одной и той же статистики при отличающихся аргументах n. Они возникают вследствие того, что сильная аномалия в значении какой-либо статистики при фиксированном n распространяется обычно и на соседние значения n. Иными словами, сильная аномалия не может "неожиданно" исчезнуть. Так, добавление нескольких точек к крупному кластеру увеличивает длину интервала и внутрикластерное расстояние между точками, но он все еще может идентифицироваться как аномальный, хотя и с меньшим значением $\delta(x)$, характеризующим степень проявления аномальности. Аналогично, если $D1(n = 2)$ аномально велико ("гэп"), то $D1(n = 3, 4$ и т.д.) тоже могут оказаться аномально большими, поскольку соответствующие значениям $n = 3, 4$ и т.д. максимальные интервалы с боль-

шой вероятностью будут включать в себя этот гэп. Чтобы списки выявляемых аномалий не были излишне длинными, наследственные взаимосвязи желательно фильтровать, хотя автоматизировать этот процесс не так просто.

5.5. Простейшая, но весьма эффективная процедура фильтрации относительно слабых аномалий связана с изменением порогов $S_{\min}$, $S_{\max}$ и коэффициента при $\widehat{\sigma}$ в решающих правилах (4)–(6). В частности, увеличение числа рандомизированных последовательностей с $m = 100$ до 1000 в имитационном эксперименте уменьшает $S_{\min}$ и увеличивает $S_{\max}$, т.е. правила отбора становятся более жесткими. К аналогичному эффекту приводит замена схемы перемешивания с сохранением l-граммного состава, $l = 1, 2, \ldots$, схемой, основанной на марковской модели с переходными вероятностями, оцениваемыми по исходной последовательности. Заметим в связи с этим, что схему перемешивания с сохранением l-граммного состава целесообразно применять лишь для небольших значений $l = 1$–3, поскольку с ростом l все труднее становится удовлетворить требованию сохранения l-граммного состава (уменьшается число потенциально возможных последовательностей, имеющих тот же l-граммный состав, что и у наблюдаемой последовательности).

5.6. Сканирующие статистики очень просты в вычислительном отношении (линейные в зависимости от длины текста N алгоритмы). Поэтому для текстов умеренной длины m-кратная рандомизация ($m \sim$ 100–1000) не представляет проблем. Для длинных текстов можно реализовать достаточно простые приближенные алгоритмы выявления кластеров и гэпов, основываясь на текущей (в скользящем окне) оценке введенных выше безразмерных параметров $\delta(x)$ и $\Delta(x)$ и сравнении их с порогами, полученными при обучении. При этом процедура перемешивания используется лишь при обучении, когда устанавливаются пороги аномальности для $\delta(x)$ и $\Delta(x)$, а анализ наблюдаемых последовательностей ведется в реальном масштабе времени.

Мотивацией для такого подхода может служить опыт обработки различных аминокислотных последовательностей (типичный диапазон длин — 10^2–10^3 символов), в которых аномальные гэпы и кластеры в подавляющем большинстве случаев характе-

ризовались значениями $\Delta(x)$ и $\delta(x)$ порядка 5 и выше. Это означает, что длина гэпа не менее чем в 5 раз превышает среднетекстовое расстояние между вхождениями элементов x в текст, соответственно, среднетекстовое расстояние в 5 и более раз превышает среднее расстояние между вхождениями элементов x внутри кластера.

5.7. Косвенным показателем значимости позиционных аномалий может служить степень их устойчивости к искажению (или варьированию) исходной последовательности. Некоторые из рассматриваемых нами типов аномалий, например, гэпы и изолированные точки, очень чувствительны к внесению даже незначительных искажений. Поэтому сохранение позиционных аномалий в эволюционно и/или функционально близких текстах является подтверждением их значимости. Подобную устойчивость в наших экспериментах неоднократно демонстрировали представители отдельных семейств белков, не слишком разошедшиеся в процессе эволюции. Аналогичные эффекты наблюдались и на параллельных переводах "Винни-Пуха", где характер позиционных аномалий позволяет судить о композиционных отличиях двух текстов (пропуск одной из глав у Заходера, перестановка двух других и т.п.).

5.8. Содержательная трактовка позиционных аномалий многообразна и специфична для каждой предметной области. Применительно к естественно-языковым текстам мы попытались ее проиллюстрировать в примерах 3–5. Аномалии в аминокислотных последовательностях могут быть связаны с пространственной структурой белка (интересны в этом плане кластеры из заряженных аминокислот, регулярно встречающиеся у представителей разных белковых семейств). Некоторые аномалии указывают на характер эволюционных преобразований, которые привели к формированию соответствующей аминокислотной последовательности (см. пример 6). "Сверхравномерное" распределение отдельных аминокислот может быть связано с избеганием вхождения в состав элементов, образующих вторичную структуру (α-спирали, β-участки). В кодирующих участках ДНК-последовательностей наличие кластеров из определенных (неудобных для данного организма) кодонов, например, таких как AGG у E.coli,

может привести к существенному снижению уровня экспрессии соответствующего гена [20]. Количество подобных примеров, демонстрирующих связь позиционных аномалий с эволюционными, функциональными, структурными особенностями анализируемых текстов, легко может быть расширено.

Заключение

Предложена методика выявления аномалий в распределении слов или связных цепочек символов по длине текста. В основе ее лежит использование сканирующих статистик и имитационного моделирования для оценки значимости наблюдаемых аномалий. Выделены основные типы аномалий, прослежена их взаимосвязь друг с другом. Эксперименты с текстами различной языковой природы показали, что элементарные языковые единицы часто демонстрируют яркие позиционные аномалии, которые, как правило, имеют содержательную трактовку. Это может служить основанием для широкого использования позиционной информации в задачах сегментации слитных текстов (например, полных геномов), автоматического смыслового сжатия (для текстов на естественном языке), классификации и информационного поиска.

Литература

1. LUHN H.P. The automatic creation of literature abstracts // IBM Journal of Research and Development. – 1958. – Vol. 2, №2. – P. 159–165.

2. ГИНДИН С.И. Позиционные методы автоматического фрагментирования текста, их теоретико-текстовые и психолингвистические предпосылки // Семиотика и информатика. – 1978. – Вып. 10. – М., ВИНИТИ. – С. 32–73.

3. MANI I., MAYBURY M. (Eds.) Advances in automatic text summarization // MIT Press, Cambridge, M.A. – 1999.

4. АХО А., ХОПКРОФТ Дж., УЛЬМАН Дж. Построение и анализ вычислительных алгоритмов // М.: Мир, 1979. – С. 354–403.

5. LEMPEL A., ZIV J. On the complexity of finite sequences // IEEE Trans. on Inf. Th. – 1976. – Vol. IT – 22, №1. – P. 75–81.

6. ГУМЕНЮК А.С. Об исчислении строений лингвистических текстов // Квантитативная лингвистика и семантика. – Сб. научн. тр., Новосибирск: Изд-во НГПУ, 2000. – Вып. 2. – С. 3–11.

7. BRENDEL V., BECKMANN J.S., TRIFONOV E.N. Linguistics of nucleotide sequences: morphology and comparison of vocabularies // J. Biomol. Struct. Dyn. – 1986. – №4. – P. 11–21.

8. PEVZNER P.A., BORODOVSKY M.Yu., MIRONOV A.A. The significance of deviation from mean statistical characteristics and prediction of the frequency of occurrences of words // J. Biomol. Struct. Dym. – 1989. – №6. – P. 1013–1026.

9. ГУСЕВ В.Д. Сложностные профили символьных последовательностей // Методы обработки символьных последовательностей и сигналов. – Новосибирск, 1989. – Вып. 132: Вычислительные системы. – С. 35–63.

10. HANCOCK J.M., ARMSTRONG J.S. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences // Comput. Appl. Biosci. – 1994. – № 10. – P. 67–70.

11. NIKOLAOU C., ALMIRANTIS Y. A Study of the middle-scale nucleotide clustering in DNA sequences of various origin and functionality, by means of method based on modified standard deviation // J. Ther. Biol. – 2002. – Vol. 217. – P. 479–492.

12. NAUS J.I. The distribution of the size of the maximum cluster of points on a line // J. Amer. Statist. Assoc. – 1965. – Vol. 61, № 310. – P. 532–538.

13. NAUS J.I. A power comparison of two tests of nonrandom clustering // Technometrics. – 1966. – №8. – P. 493–517.

14. WALLENSTEIN S.R., NAUS J.I. Probabilities for a k-th nearest neighbor problem on the line // The Annals of Probability. – 1973. – Vol. 1, №1. – P. 188–190.

15. GLAZ J. Approximations and bounds for the distribution of the scan statistic // Journal of the American Statistical Association. – 1989. – Vol. 84, №406. – P. 560–566.

16. HINES W.G.S., HINES R.J.O'Hara. The Eberhardt statistic and the detection of nonrandomness of spatial point distributions // Biometrica. – 1979. – Vol. 66, №1. – P. 73–79.

17. АРЕФЬЕВ С.С., ШЕБАЛИН Н.В. Оценка уровня скученности (кластеризации) землетрясений Кавказа // ДАН СССР. – 1988. – Т. 298, №6. – С. 1349–1352.

18. KARLIN S., MACKEN C. Some statistical problems in the assessment of inhomogeneities of DNA sequence data // Journal of the American Statistical Association. – 1991. – Vol. 86, №413. – P. 27–35.

19. ТРИФОНОВ Э.Н. Генетическое содержание последовательностей ДНК определяется суперпозицией многих кодов // Молекулярная биология. – 1997. – Т. 31, №4. – С. 759–767.

20. SHARP P.M., LI W-H. The codon adaptation index — a measure of directional synonymous codon usage bias, and its potential applications // Nucleic Acids Research. – 1987. – Vol. 15, №3. – P. 1281–1295.

Поступила в редакцию
15 января 2003 года.