Journal of Siberian Federal University. Mathematics & Physics 2012, 5(4), 507–514

УДК 519.95

О структурах, выделяемых в символьных последовательностях

Евгений Ю. Бушмелёв*
Институт фундаментальной биологии и биотехнологии,
Сибирский федеральный университет,
Свободный, 79, Красноярск, 660041,
Россия

Евгений М. Миркес†
Красноярский институт инженеров железнодорожного транспорта,
Толстого, 25, Красноярск, 660028;
ИКиТ СФУ, Киренского, 26, Красноярск, 660074,
Россия

Михаил Г. Садовский‡
Институт вычислительного моделирования СО РАН,
Академгородок, Красноярск, 660036;
ИУБПЭ СФУ, Киренского, 26, Красноярск, 660074,
Россия

Получена 29.02.2012, окончательный вариант 29.03.2012, принята к печати 05.06.2012

В работе рассмотрена проблема поиска и выделения структур в символьных последовательностях на примере генетических текстов. Показано существование нестрогой периодичности и асимметрии в распределении слов различной длины вдоль по последовательности, обсуждены возможные механизмы формирования такого рода структур.

Ключевые слова: триплет, порядок, частота, распределение, марковская модель.

Введение

Символьные последовательности являются традиционным объектом математических исследований. Области приложений такого рода исследований выходят далеко за пределы чистой математики и переплетаются с дисциплинами самого разного уровня: от биологии до лингвистики, не говоря уже о прикладной математике, информатике и физике. Дискретный характер символьных последовательностей (из конечного алфавита $\aleph$) позволяет эффективно применять многие методы, использование которых невозможно для непрерывных объектов.

Поиск и выделение структур в символьных последовательностях, по-видимому, центральная задача в исследованиях символьных последовательностей. Результат такого рода поиска будет существенно зависеть от того, что именно понимается под структурой. Здесь возможны два в некотором смысле противоположных подхода:

– поиск и выделение по образцу;

*hoochie_cool@mail.ru
†mirkes@bk.ru
‡msad@icm.krasn.ru

– поиск и выделение структурных единиц *per se*, посредством построения той или кластеризации объектов. Здесь под объектами понимаются некоторые внутренние элементы, в нашем случае — цепочки символов заданного вида.

В настоящей работе представлены предварительные результаты исследования структур, выделяемых в нуклеотидных последовательностях. Под структурой здесь будет пониматься функция распределения триплетов вдоль по нуклеотидной последовательности, при этом никаких априорных предположений о структуре исходной символьной последовательности делаться не будет, однако выявление структур, очевидным образом, потребует сравнения наблюдаемых особенностей поведения функции распределения с аналогичными, наблюдаемыми на тех или иных модельных последовательностях. К этим последним относятся символьные последовательности, являющиеся реализацией того или иного случайного процесса: мы будем сравнивать реальные символьные последовательности с реализациями тех или иных бернуллиевских и марковских процессов.

Основные определения и понятия

Будем называть символьную последовательность из конечного алфавита $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ **генетическим текстом** (ГТ). Число символов в последовательности N будем называть длиной ГТ. Характерная длина реальных последовательностей, которые мы будем рассматривать в настоящей работе, составляет $N \sim 10^8$. Будем считать, что ГТ не содержат пробелов либо иных символов. **Словом** $\omega_{l_1 l_2 l_3 \ldots l_{q-1} l_q}$ (длины q) будем называть любую связную подпоследовательность этой длины. **Расстоянием** до ближайшего соседа между словами $\omega^{(1)}$ и $\omega^{(2)}$ будем называть число l символов между этими словами при том условии, что эта цепочка из K символов не содержит слова $\omega^{(2)}$.

Функцией распределения до ближайшего соседа $F_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ будем называть число ближайших соседей $\langle\omega^{(1)},\omega^{(2)}\rangle$, встретившихся на расстоянии l друг от друга. Подчеркнем, что так определенная функция принимает целые неотрицательные значения и определена также на множестве натуральных чисел. Условимся считать, что $l \geq 1$; при этом $l = 1$ означает, что два слова $\omega^{(1)}$ и $\omega^{(2)}$ пересекаются, и пересечение это имеет длину $q-1$. Функция $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ является частотой встреч соответствующих пар на расстоянии l друг от друга и обладает естественной нормировкой:

$$\sum_{l\in\widetilde{L}} f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l) = 1\,. \tag{1}$$

Здесь $\widetilde{L}$ означает область определения данной конкретной функции $F_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$. Наша задача — выяснить характер поведения функции (1) для различных ГТ. Теоретически $\widetilde{L} = \mathbb{N}$, однако на практике областью определения является конечное множество $1 \leqslant L \leqslant M$, где M определяется исследователем; обычно $M = 10^4$.

Из общих соображений теоретико-вероятностного характера можно утверждать, что функция $F_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ и функция $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$, соответственно, будут убывающими. Нас будет интересовать как характер такого убывания, так и (значительные) нарушения монотонности убывания, которые мы будем интерпретировать как наличие некоторой структурированости в изучаемой последовательности. Существенным здесь становится вопрос о том, какое поведение функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ следует считать «образцовым»; иными словами, с чем следует сравнивать немонотонность в поведении данной функции, чтобы можно было надежно утверждать о существовании порядка и структурированности в последовательности.

Один из естественных вариантов ответа на этот вопрос — сравнение реальной последовательности с модельными, статистические свойства которых достаточно хорошо изучены.

Очевидно, что сравнивать следует не сами последовательности, а функцию $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$, наблюдаемую на таких последовательностях. Наиболее естественными кандидатами на роль таких «опорных» последовательностей выступают случайные последовательности, являющиеся реализациями бернуллиевского либо марковского процессов того или иного порядка.

Состояние проблемы

Изучение распределения слов фиксированной длины вдоль по символьным последовательностям представляет собой один из простейших подходов к исследованию упорядоченности и структур в них. Несмотря на такую простоту, должного внимания этой проблеме не уделялось. К настоящему времени нет сколько-нибудь исчерпывающего списка работ, в которых данная проблема подвергалась бы анализу.

В работах [1–3] предприняты попытки изучения возможных закономерностей описанного типа — появления ансамбля цепочек заданной структуры — в случайных символьных последовательностях. В работе [1] изучалось распределение вероятностей появления серии единиц заданной длины в случайной последовательности; *в некотором смысле, эта работает задачу, двойственную к той,* которую необходимо решить для ответа на вопрос о распределении ближайших соседей, — частично указанная работа отвечает на вопрос о том, как устроена «середина» между двумя заданными словами. К этой тематике примыкает и работа [3], в которой рассмотрена связь между свойствами случайных последовательностей, порожденных процессом Бернулли и возможными перестановками блоков в таких последовательностях.

Работа [2] посвящена близкой проблеме — изучению распределения слов заданной длины L в случайной бинарной последовательности таких, что первые k символов в этих словах равны 0, а оставшиеся $L-k$ символов — единицам. Эта работа также решает двойственную задачу по отношению к той, которая требуется для ответа на вопрос о распределении ближайших соседей.

Работа [4] также посвящена близкой тематике: в ней рассмотрена задача построения распределения вероятностей обнаружения специфических объектов в символьных последовательностях — мотивов. Все полученные в работе результаты базируются на использовании Марковских цепей. Тем не менее, в этой работе получена оценка вероятности встречи мотива заданной структуры (фактически — ближайшего соседа в терминологии нашей работы) на заданном расстоянии и оценка расстояния между двумя последовательными вхождениями такого мотива. Наконец, в работе [5] рассмотрена асимптотами распределения слов заданной структуры на больших расстояниях. Для теории чисел подобная задача рассмотрена в работе [6].

Поскольку ближайшим естественным объектом, для которого символьные последовательности наиболее подходящий математический аналог, являются генетические тексты, постольку можно ожидать, что именно в этой области были достигнуты какие-либо результаты. Последовательный анализ имеющейся литературы однако не дает никаких содержательных результатов. В некоторых работах (см. работы [7–10]) изложены варианты анализа распределения подпоследовательностей заданной структуры вдоль по последовательностям; однако все эти работы базируются на изучении модельных последовательностей, т. е. реальные последовательности заменены случайными с той или иной заранее заданной структурой; как правило, это марковские последовательности сравнительно низкого порядка.

Материалы

Все проанализированные в работе ГТ были взяты из EMBL–банка *www.ebi.ac.uk/genomes*. Реальные последовательности не всегда свободны от "лишних" символов; поэтому при по-

строении функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ было принято соглашение о лишних символах. Не вдаваясь в подробности появления таких символов, скажем лишь, что при построении функции распределения они игнорировались, а исходная последовательность подвергалась конкатенации: фрагменты, состоящие из лишних символов (любой длины), исключались из изучаемой последовательности, а получившиеся фрагменты «сшивались», образуя связную последовательность.

В настоящей работе представлены некоторые предварительные результаты изучения функции распределения слов длины 3 до ближайшего соседа на примере нуклеотидных последовательностей $22^{\text{й}}$ хромосомы шимпанзе (номер доступа BA000046 в EMBL–банке) и $20^{\text{й}}$ хромосомы шимпанзе (номер доступа CM000096). Длина первого ГТ составляет 32799110 символов, второго — 55268282 символов.

Результаты

Мы исследовали поведение функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ для слов длины три, т. н. триплетов. Это связано с тем, что триплеты играют чрезвычайно важную роль в хранении и реализации генетической информации, содержащейся в ГТ. Общее число триплетов для указанного выше алфавита $\aleph$ составляет $\|\aleph\|^3 = 4^3 = 64$; соответственно, общее число пар триплетов, для которых мыслимо построение функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$, составляет $64^2 = 4096$.

Для каждого реального ГТ строились две суррогатные последовательности той же длины, с тем же составом символов; первая строилась с помощью случайного бернуллиевского потока, вторая — с помощью марковского процесса третьего порядка. Такой порядок был выбран постольку, поскольку изучалось распределение слов длины 3 (триплетов).

На рис. 1 приведен пример поведения функции $F_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$, определенной для пары триплетов (т. е. слов длины 3), для ГТ, представленного $22^{\text{й}}$ хромосомой шимпанзе. Поскольку поведение функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ требует сравнения с поведением этой функции, определяемой для тех или иных «реперных» последовательностей, постольку на этом рисунке представлены три кривые:

- кривая изменения функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ для реального ГТ (черная линия);
- кривая изменения указанной функции для последовательности, полученной реализацией марковского процесса третьего порядка (серая кривая),
- кривая изменения указанной функции для последовательности, полученной реализацией бернуллиевского процесса.

Понятно, что частоты символов в случайных последовательностях совпадали с таковыми для реального ГТ; для случая марковской модели ГТ совпадали частоты слов длины 3, наблюдаемые в реальном ГТ и в модельной последовательности.

Кривая изменения функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$, наблюдаемая для случайной последовательности, полученной реализацией бернуллиевского процесса, выглядит на рис. 1 почти как идеальная экспонента, а ее начальное значение заметно превышает аналогичные значения, наблюдаемые для реального ГТ и марковской случайной последовательности. Следует остановиться на выборе соответствующей марковской последовательности, используемой в качестве модельной. Хорошо известно, что для любой конечной символьной последовательности всегда можно выбрать такой порядок марковского процесса, который обеспечит **абсолютно точное** воспроизведение исходной символьной последовательности [11,12]; очевидно, такой процесс будет по-своему вырожденным: все переходные вероятности там будут равны либо 0, либо 1, однако абсолютно точное восстановление возможно, и такой процесс построить можно всегда.

В качестве марковской модели мы выбрали последовательность, порожденную процессом второго порядка; для избегания терминологических разночтений скажем, что данный процесс восстанавливал (предсказывал) появления троек символов по их паросочетаниям.

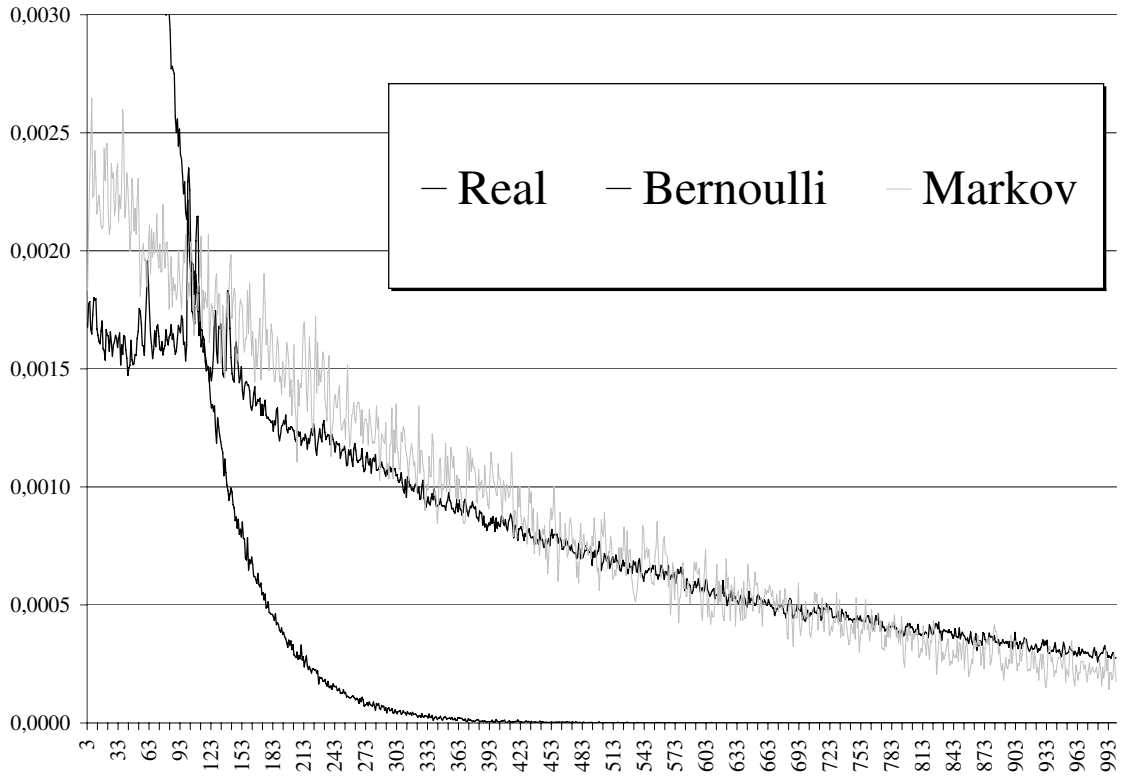

Рис. 1. Функция распределения $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ для пары триплетов $\mathsf{AAA} \div \mathsf{CGA}$ для 22[й] хромосомы шимпанзе. По оси абсцисс отложено расстояние l, по оси ординат — значения функции $f(l)$

Выбор такого порядка обусловлен тем, что мы исследовали распределение триплетов вдоль по последовательности, и тем, что в реальных ГТ триплеты играют очень важную роль.

Ограниченность объема статьи не позволяет изложить все полученные данные, однако можно с уверенностью утверждать, что фактически для любых ГТ наблюдаются следующие эффекты:

- затухание функции распределения $f(l)$ на больших расстояниях носит экспоненциальный характер;
- любое распределение для реального ГТ заметно отличается от аналогичных, получаемых на модельных последовательностях, построенных с помощью бернуллиевского либо марковского процесса;
- во всех проверенных ГТ (более 300 последовательностей) наблюдается «микропериодичность» — дисперсия абсолютных значений разностей двух последовательных значений $|f(l+1) - f(l)|$, $l = 1, 2, \ldots$ выше, чем для модельных последовательностей, построенных с помощью бернуллиевского либо марковского процесса;
- существует эффект, противоположный «микропериодичности», — для значительного количества триплетов на реальных ГТ наблюдаются «двухчастичные» взаимодействия: доля пар с нулевым расстоянием (т. е. с непосредственно примыкающими друг к другу триплетами) заметно превосходит ожидаемую; наконец,
- характер изменения функции $f(l)$ с ростом расстояния показывает, что это распределение не похоже ни на одно стандартное распределение, например Пуассона.

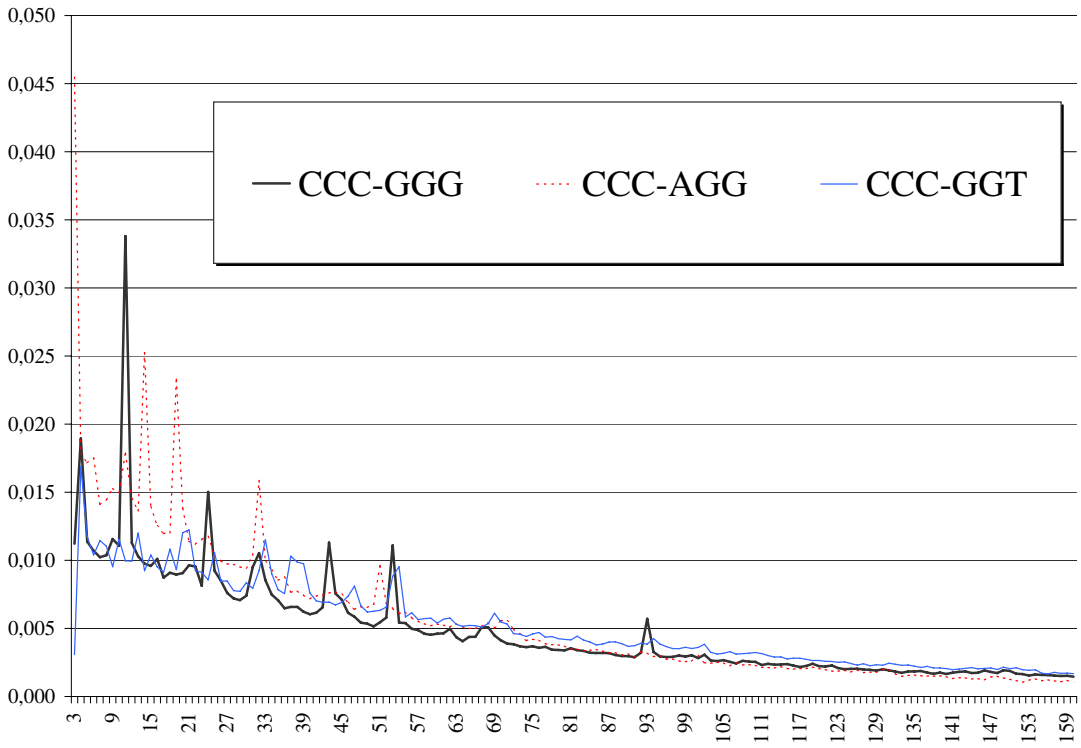

Рис. 2. Корреляции между близкими парами

Обсуждение

Анализ распределения частот расстояний «до ближайшего соседа» выявляет существование структурированности в нуклеотидных последовательностях. Одной из возможных причин существования такого рода структурированности в ГТ считаются длинные повторы (см., напр., [9]). Однако проведенный нами анализ не дал результатов, свидетельствующих в пользу этой гипотезы. Действительно, если бы ведущей причиной возникновения такой структурированности были бы длинные повторы, то тогда очень близкий паттерн изменения функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ должен был бы наблюдаться не только на тех триплетах, для которых четко выделяется такая скрытая периодичность, но и для ближайших с ними пересечений.

Например, для пары триплетов $\mathsf{CCC} \div \mathsf{GGG}$ наблюдается четко выраженная скрытая периодичность; при этом для двух пар — «внутренней» $\mathsf{CCC} \div \mathsf{AGG}$ и «внешней» $\mathsf{CCC} \div \mathsf{GGT}$ — наблюдаемые пики у функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ весьма плохо скоррелированы, что косвенно опровергает гипотезу о том, что в данном случае длинные повторы являются причиной такой скрытой периодичности (см. рис. 2).

Сформулируем кратко основные свойства выявляемых закономерностей.

"Нечеткий порядок". Это означает, что все закономерности проявляются лишь в среднем: их можно наблюдать на достаточно большом ансамбле цепочек, являющихся вложениями в рассматриваемую символьную последовательность.

Наблюдаемый порядок весьма специфический. Это означает, что даже близкие (например, в смысле расстояния Хэмминга) слова, для которых ищется распределение $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$ до ближайшего соседа, является весьма специфичным. При этом такая специфичность характерна и для исследуемой последовательности, и для изучаемой пары триплетов. Иными

словами, одна и та же пара триплетов на разных ГТ дает зачастую радикально различную картину поведения функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$. Одновременно, две пары триплетов, пересекающиеся по словам длины 2, также могут давать весьма различающуюся картину поведения функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$. Поясним на примере этот факт.

Рассмотрим функцию $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$, определенную для того или иного ГТ и для двух пар триплетов таких, что они пересекаются по подслову длины 2, например $\mathsf{ACC} \div \mathsf{TGT}$ и $\mathsf{CCT} \div \mathsf{ATG}$. Можно ожидать, что кривые изменения функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$, построенной для первой и второй пар триплетов, будут достаточно близкими. Как показывает практика, это далеко не так (см. рис. 2).

"Асимметрия". Это означает, что характер поведения двух функций распределения до ближайшего соседа $f_{\langle\nu_1\nu_2\nu_3,\mu_1\mu_2\mu_3\rangle}(l)$ и $f_{\langle\mu_1\mu_2\mu_3,\nu_1\nu_2\nu_3\rangle}(l)$, определенных для одной и той же пары триплетов $\nu_1\nu_2\nu_3$ и $\mu_1\mu_2\mu_3$, будет различным. По-видимому, можно утверждать, что такое различие в поведении двух функций будет наблюдаться и для более длинных слов. По нашим наблюдениям, такая асимметрия наблюдается практически для любых ГТ.

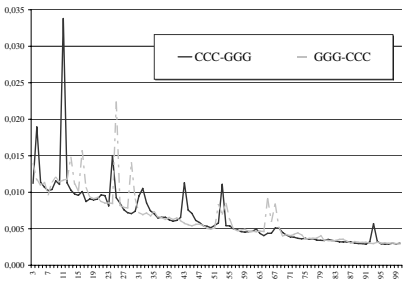

Рис. 3. Пример асимметрии

Строго говоря, факт асимметрии для ГТ хорошо известен: так, например, соответствующие молекулы ДНК химически различаются с противоположных концов. Тем не менее, нам представляется, что факт такой статистической асимметрии является фундаментальным и заслуживает отдельного детального изучения. На рис. 3 показан пример такого рода асимметрии: функция $f(l)$ для «прямой» пары $\mathsf{CCC} \div \mathsf{GGG}$ и для «обратной» ($\mathsf{GGG} \div \mathsf{CCC}$) заметным образом различаются. Более того, заметно различаются и абсолютные значения соответствующих пар: если для первой пары максимальное значение функции $F(12) = 14857$, то для второй — $F(16) = 6887$; при этом $F(12) = 5193$ для этой второй пары. Такого рода различия гораздо заметнее выражены для других пар.

Обнаруженная в настоящей работе сложность в распределении частот «до ближайшего соседа» между двумя цепочками заданной структуры позволяет сформулировать гипотезу о существовании своего рода **гидродинамики генетических текстов**. Распределение, наблюдаемое для функции $f_{\langle\omega^{(1)},\omega^{(2)}\rangle}(l)$, для ГТ весьма напоминает аналогичные распределения, например, энергии в некоторых задачах гидродинамики. Аналогия здесь состоит в следующем: пусть первоначально ГТ представляет собой полностью упорядоченную структуру. Затем над ним производятся некоторые преобразования, в результате которых структура ГТ начинает выглядеть как суперпозиция нескольких регулярных и нескольких случайных (либо сложных до степени неотличимости от случайных). Ключевой вопрос здесь: что это за преобразование? Ответ следует искать среди того класса, который будет удовлетворять принципам теории отбора [13], однако обсуждение этого вопроса выходит за рамки настоящей статьи.

Список литературы

[1] K.Sinha, B.P.Sinha, On the distribution of runs of ones in binary strings, *Comput. Math. Appl.*, **58**(2009), №9, 1816–1829.

[2] F.S.Makri, On occurrences of F–S strings in linearly and circularly ordered binary sequences, *J. Appl. Probab.*, bf 47(2010), №1, 157–178.

[3] J.Sethuraman, S.Sethuraman, Connections between Bernoulli strings and random permutations, The legacy of Alladi Ramakrishnan in the mathematical sciences, Springer, New York, 2010, 389–399.

[4] V.T.Stefanov, Occurrence of patterns and motifs in random strings, *Scan statistics*, pp. 351 – 367, Stat. Ind. Technol., Birkhäuser Boston, Inc., Boston, MA, 2009.

[5] M.,Giraud, Asymptotic behavior of the numbers of runs and microruns, *Information and Computation*, **207**(2009), 1221–1228.

[6] P.Pollack, Long gaps between deficient numbers, *Acta Arith.*, **146**(2011), №1, 33–42.

[7] W.Ebeling, Th.Poschel, K.-F.Albrecht, Entropy, transinformation and word distribution of information-carrying sequences, *Int. J. of Bifurcation &Chaos*, **5**(1995) №1, 51–61.

[8] E.E.Kuruoglu, J.Zerubia, Skewed alpha-stable distributions for modelling textures, *Pattern Recognition Letters*, **24**(2005), №- 3, 339–348.

[9] S.Subramanian, V.M.Madgula, R.George, R.K.Mishra, M.W.Pandit, Ch.S.Kumar, L.Singh, Triplet repeats in human genome: distribution and their association with genes and other genomic regions, *Bioinformatics*, **19**(2003), №5,549–552.

[10] V.Baldazzi, S.Bradde, S.Cocco, E.Marinari, R.Monasson, Inferring DNA sequences from mechanical unzipping data: the large-bandwidth case, *Physical Review* **E**, **75**(2007), 011904.

[11] М.Г.Садовский, Т.Г.Попова, Интроны отличаются от экзонов по своей избыточности, *Генетика*, **31**(1995), № 10, 1365–1369.

[12] М.Г.Садовский, Об информационной емкости символьных последовательностей, *Вычислительные технологии*, **10**(2005), № 4, 82–89.

[13] A.N.Gorban, Selection Theorem for Systems with Inheritance, *Math. Model. Nat. Phenom.*, **2**(2007), №4, 1–45.

On the Structures Revealed from Symbol Sequences

Eugeny Yu. Bushmelev
Eugeny M. Mirkes
Michael G. Sadovsky

The problem of structure search and revealing in symbol sequences is discussed, based on a study of nucleotide sequences. A fuzzy periodicity is proven, as well as the asymmetry in the distribution of various strings alongside a sequence.

Keywords: order, fuzzy periodicity, triplet, frequency, distribution, Markov chain.