Структурные аналогии в символьных последовательностях различной языковой природы*

В. Д. Гусев, Л. А. Мирошниченко, Н. В. Саломатина
gusev@math.nsc.ru, luba@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН

Изучение структуры символьных последовательностей (текстов) играет важную роль при решении многоплановых задач анализа данных, возникающих в биологии, лингвистике и других областях знания. При всем многообразии текстов их объединяет наличие повторов как элементарных структурообразующих единиц. Целью работы является систематизация повторов и их комбинаций, т.е. *структурных единиц* более высокого уровня. Для их выделения используются сложностные профили последовательности (введены авторами) и аппарат сканирующих статистик (адаптирован для текстов на естественном языке). По итогам обработки текстов различной языковой природы выделены и описаны структурные единицы, характеризующиеся «межъязыковой общностью», что и является отличительной особенностью работы.

Ключевые слова: *символьные последовательности; структурные единицы; разнотипные повторы; профили сложности; профили кластеризуемости*

Structural analogies in symbolic sequences of different nature*

V. D. Gusev, L. A. Miroshnichenko, N. V. Salomatina
Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

Symbolic sequences (words, strings, texts) as an object of study are encountered in various areas of knowledge: informatics, biology, linguistics, music. The notion of integrated repeats as elementary structure-forming units is general on conceptual level for all symbolic sequences, despite of diversity of alphabets, lengths, and nature of the texts. The purpose of this work is the systematization of elementary repeats and their combinations, i. e., structural units of higher level. Their function in different language systems is discussed.

Too low complexity of fragments of the text is usually correlated with existence of too long repeats or their high concentration. Thus, a basis of all methods of research is a complexity profile construction and the analysis of complexity decomposition of the text in the sliding window mode. Such analysis gives a conception of the most typical structural units which can be found in texts. In the natural language texts, where repeatability is less expressed, also the profile of clustering can be used.

DNA sequences of different organisms, texts in a natural language, and also neume himns are a source material for investigation. Systematization of structural units is the result of the complexity analysis of a huge number of texts of various nature. The interlanguage community principle is a reason for selection of the illustrating structures.

The approach stated in this work and the algorithms realizing it have rather universal character in respect of its applicability to various language systems. The interlanguage analo-

*Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00400.

gies described at the level of structures can extend to formulation of substantial problems and selection of tools of their solving.

Keywords: *symbolic sequences; structural units; repeats; complexity profile; clustering profile*

Введение

Символьные последовательности (тексты) как объект исследования встречаются во многих областях знания: математике, информатике, биологии, лингвистике, музыке. Примерами могут служить тексты на естественном языке, биологические последовательности (ДНК, РНК, аминокислотные, последовательности генов и др.), знаменные песнопения (знаковые последовательности, использовавшиеся для записи древнерусского церковного пения) и др.

Изучение структуры последовательностей в целом и отдельных их фрагментов является основой для решения многочисленных задач классификации (родо-видовой, жанровой, тематической и т. п.). Так, сходство первичных структур ДНК-последовательностей обычно предполагает и функциональную близость кодируемых ими белков. Специфические тандемные повторы используются для проведения ДНК-дактилоскопии. Этот же тип структур помогает идентифицировать в текстах знаменных песнопений фрагменты, характеризующиеся нестандартным («зашифрованным») распевом (так называемые «лица»). Аномально длинные тандемные или разнесенные повторы разного типа (см. далее) в последовательностях, вырабатываемых датчиками «случайных» чисел, или в шифротекстах сигнализируют о несовершенстве схем генерации чисел и шифрования. Длинная цепочка существительных в родительном падеже (серия в частеречном алфавите) трактуется как стилистическая погрешность и т. д.

При всем многообразии языковых систем и характеризующих их текстов объединяющим началом для них является понятие *повтора в широком смысле*, выступающего в качестве *основного структурообразующего элемента* текста. В частности, аномально длинные, а также аномально частые или редкие повторы различных типов уже могут рассматриваться как потенциально возможные структурные единицы. Аномальность определяется в сопоставлении со значениями, ожидаемыми для указанных параметров в предполагаемой модели порождения последовательности. Различают повторы прямые и симметричные, следующие друг за другом (тандемные) и разнесенные, совершенные (точные) и несовершенные (с искажениями). Последние связаны с проявлениями вариативности языковых единиц, присущей всем эволюционирующим языковым системам. Возможны также повторы с переименованием элементов алфавита или повторы с точностью до фиксированного агрегирования элементов алфавита.

Комбинации позиционно близких элементарных повторов разного типа порождают структурные единицы более высокого уровня. *Целью работы* является *систематизация* такого рода структурных единиц, характерных (общих) для различных языковых систем. Обсуждается их функциональная нагрузка. Иллюстрируются возможности использования *межъязыковых аналогий* при постановке содержательных задач и выработке подходов к их решению.

Заметим, что эволюционный фактор не позволяет говорить о полноте систематизации в каком-либо строгом смысле этого слова. В процессе эволюции меняется даже алфавит обсуждаемых языковых систем. Поэтому акцент сделан не столько на полноту охвата структур, присутствующих в реальных текстах, сколько на проявления межъязыковой

общности. Мы не апеллируем к схемам порождения текста, основанным на использовании формальных грамматик, не учитывающих возможность появления искажений в процессе эволюции и больше подходящих для языков программирования. Подразумеваемая нами схема порождения основана на использовании операций копирования (см., например, [1]), фиксирующих разноплановые проявления повторности в текстах.

Используемый подход. Сложностные разложения

Некоторые типы структур, представленные в статье (например, фракталоподобные, комбинированные и др.), выявлены в реальных текстах и даже поименованы авторами данной работы. Часть структур (например, кумулятивные) была описана и систематизирована ранее другими авторами, но применительно к какой-либо одной языковой системе. Наша роль в этом случае сводилась к поиску аналогов этих структур в других языковых системах (хотя бы в одной). В связи с этим работа носит и частично обзорный характер. Для выявления структур разными авторами использовались разные подходы, в том числе не алгоритмические (см., к примеру, [2, 3]). Наш подход, кратко описанный ниже, основан на идеях А. Н. Колмогорова относительно определения понятия «количество информации» [4]. Объективным индикатором насыщенности символьной последовательности повторами может служить ее сложность. А. Н. Колмогоров предложил оценивать сложность объекта (в данном случае последовательности S) длиной кратчайшего описания $K(S)$, по которому этот объект можно восстановить однозначно. Известно, однако, что колмогоровская сложность не является вычислимой функцией. Из возможных конструктивных приближений к оцениванию $K(S)$ мы опираемся на меру сложности конечной символьной последовательности, предложенную Лемпелем и Зивом [1]. Она в явном виде *апеллирует к понятию повтора* в традиционном его понимании (прямой совершенный) и легко обобщается на случай фиксированной совокупности *разнотипных повторов*, характерных для конкретной языковой системы [5].

Пусть Σ — конечный алфавит; $|\Sigma|$ — размер алфавита; S — конечная последовательность, составленная из элементов Σ (текст); $N = |S|$ — длина текста S; $S[i]$ — элемент S, стоящий в i-й позиции ($1 \leqslant i \leqslant N$); $S[i : j]$ — фрагмент S, включающий элементы с i-го по j-й ($1 \leqslant i \leqslant j \leqslant N$); x^m — m-кратное повторение символа (символьной цепочки) x; $S = S_1S_2$ — конкатенация (сцепление) последовательностей S_1 и S_2.

Лемпель и Зив определили сложность конечной символьной последовательности S как число шагов гипотетического процесса синтеза S с использованием двух допустимых операций: «порождение нового символа», и «копирование *максимально длинного* («готового») прототипа из предыстории» , т. е. из уже синтезированной части текста. Последовательность фрагментов, отражающих процесс синтеза, мы называем «*сложностным разложением*» S:

$$H(S) = S[1 : i_1]S[i_1 + 1 : i_2] \dots S[i_{k-1} + 1 : i_k] \dots S[i_{c-1} + 1 : N],$$

где $S[i_{k-1} + 1 : i_k]$ — фрагмент S, добавляемый на k-м шаге, а c — число шагов процесса (*сложность* S). В этом разложении операция порождения символа задействована не более чем $|\Sigma|$ раз. Подавляющая же часть компонентов получена с применением операции копирования. Таким образом, сложностное разложение — это представление текста в виде конкатенации повторяющихся фрагментов в том смысле, что каждому компоненту в $H(S)$ (за исключением порождаемых) соответствует свой прототип (повтор) в предыстории. Возможны случаи наложения (со сдвигом) компонента на прототип, сигнализирующие о наличии тандемной повторности.

Проиллюстрируем, например. как выглядит сложностное разложение последовательности $S = caa(ccatgat)^5at$ $(N = 40)$, содержащей достаточно длинную периодичность:

k	1	2	3	4	5	6	7	8	9	10	
$H(S) =$	$c\cdot$	$a\cdot$	$a\cdot$	$c\cdot$	$ca\cdot$	$t\cdot$	$g\cdot$	$at\cdot$	$(ccatgat)^4\cdot$	$at;$	$c(S) = 10$
q_k	1	2	3	4	5	7	8	9	11	39	
j_k	0	0	2	1	1	0	0	6	4	37	
l_k	1	1	1	1	2	1	1	2	28	2	

Здесь компоненты разложения разделены точками, k — номер компонента, l_k — его длина, q_k — начальная позиция k-го компонента, а j_k — начальная позиция прототипа для k-го компонента (в случае, если символ встретился впервые и применяется операция порождения, полагаем $j_k = 0$). Нетрудно видеть, что прототипом для девятого компонента $S[11 : 38]$ служит фрагмент $S[4 : 31]$, т. е. имеет место наложение компонента на прототип, сигнализирующее о наличии периодичности. Первые 7 символов компонента №9 копируются из предыстории, а все последующие — с элементов, синтезированных на текущем (еще не завершенном) шаге. Следует заметить, что если бы кратность повторения периода *ccatgat* была выше, то длина S могла бы возрасти в разы, но число компонентов в разложении осталось бы прежним.Формально, о наличии периодичности свидетельствует выполнение условия $j_k + l_k \geqslant q_k$. Легко показать при этом, что длина периода $p = q_k - j_k$, а кратность повторений не меньше, чем $l_k/p + 1$. На этом свойстве и основан алгоритм обнаружения периодичностей [8].

Сложность последовательности можно оценивать как в целом, так и в окне заданного размера W, которое скользит вдоль нее. В последнем случае речь идет об отслеживании «*локальной сложности*». Кривую изменения локальной сложности вдоль последовательности S мы называем *сложностным профилем* S и обозначаем $P(S, W)$. Формально $P(S, W) = c_1 c_2 \ldots c_i \ldots c_{N-W+1}$, где c_i — сложность фрагмента из S, выделяемого окном на i -м шаге, т.е. включающего в себя позиции $i, i + 1, \ldots, i + W - 1$.

Фрагменты текста, которым соответствуют *аномально низкие* значения локальной сложности, характеризуются высокой концентрацией повторов, т. е. *высокой степенью структурированности*. Именно эти фрагменты дают представление о наиболее характерных структурах, представленных в последовательности. Параметр W при этом может меняться в широких пределах , что позволяет выделять структуры, соответствующие разным иерархическим уровням. Для иллюстрации на рис. 1 и 2 приведены сложностные профили бактериального генома микоплазмы «Mycoplasma gallisepticum str. R(low)» (AE015450) для $W = 60$ и $W = 1000$ соответственно. Средние значения сложности $c_{\exp} =$ $= 20{,}72$ при $W = 60$ и $c_{\exp} = 186{,}81$ при $W = 1000$. Пунктиром обозначены линии, соответствующие уровням $c_{\exp} \pm 3\sigma$. К аномальным можно отнести все пики, лежащие ниже нижней пунктирной линии. При малых размерах окна ($W = 60$ в нашем случае) фрагменты с низкими значениями сложности зачастую содержат достаточно длинные периодичности с небольшой длиной периода, комплементарные палиндромы, либо комбинированные структуры (см. ниже). Главный минимум на рис. 1 обусловлен наличием в позиции 497 461 периодичности $(aga)^{27}$. При больших размерах окна (см. рис. 2) низкие значения сложности обычно обусловлены крупными повторами разных типов (тандемными или разнесенными), а также их комбинациями. Главный минимум объясняется наличием в позиции 925 156 фрагмента *gttttagcactgtacaatacttgtgtaagcaataac*, регулярно (с равными промежутками) повторяющегося более 30 раз (см. ниже раздел «Периодичности со сложной струк-

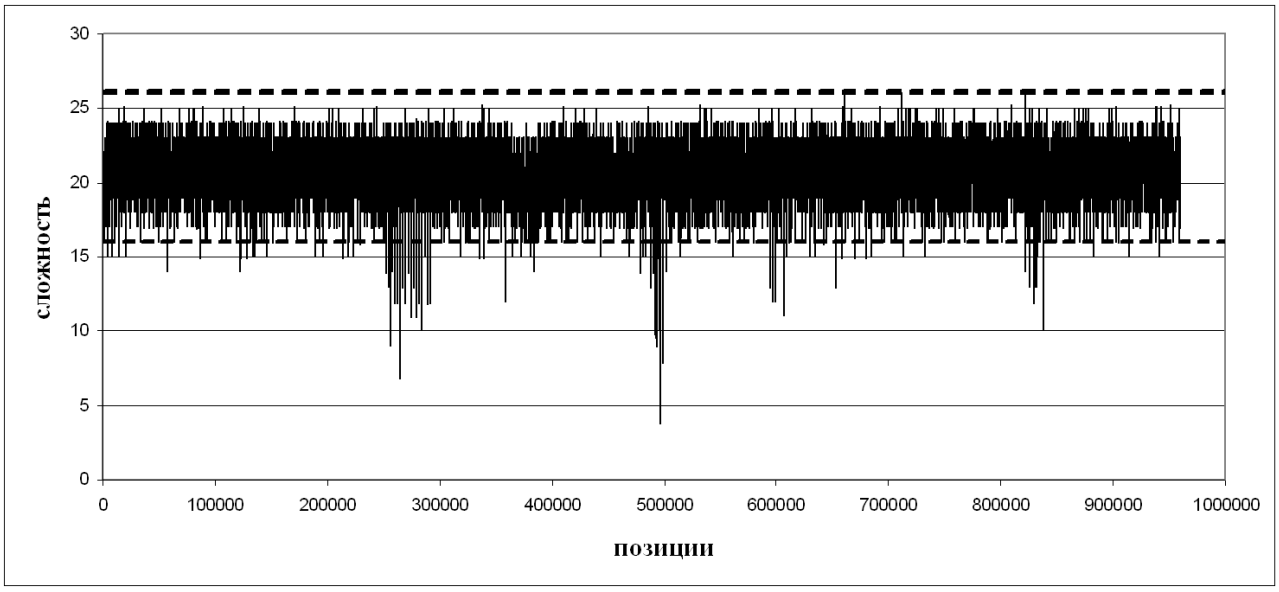


Рис. 1. Профиль сложности генома микоплазмы «R» при размере окна $W = 60$

турой») Распределение пиков на обоих кривых демонстрирует существенные различия в числе, размерах и расположении аномальных по сложности зон в геноме.

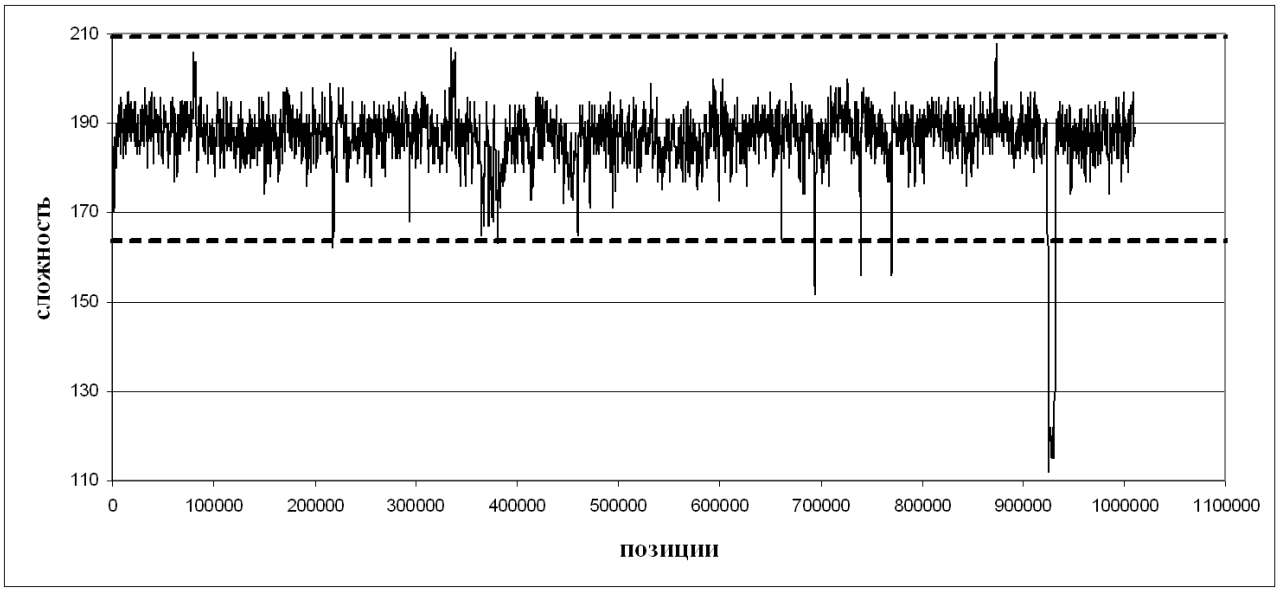


Рис. 2. Профиль сложности генома микоплазмы «R» при $W = 1000$

Обобщение подхода Лемпеля и Зива

Для конкретной языковой системы, содержащей специфические типы повторов, подход Лемпеля и Зива можно модифицировать, введя дополнительные операции копирования. Продемонстрируем это на примере ДНК-последовательностей [5]. А именно, наряду с прямым копированием, фиксирующим повторы в обычном смысле ($\ldots atcgag \ldots atcgag \ldots$), допускается симметричное копирование, выявляющее инверсии ($\ldots atcgag \ldots gagcta \ldots$), а также прямое и симметричное копирование с точностью до

подстановки $(a \leftrightarrow t)$, $(c \leftrightarrow g)$, реализующей известное отношение комплементарности на элементах ДНК-алфавита. Прямому комплементарному копированию соответствуют повторы вида $(\ldots atcgag \ldots tagctc \ldots)$, где второй фрагмент получен из первого заменой a на t, t на a, c на g, g на c. Симметричному комплементарному копированию соответствуют повторы вида $(\ldots atcgag \ldots ctcgat \ldots)$, где второй фрагмент совпадает с первым при прочтении его в обратном направлении и тех же заменах.

Расширение спектра допустимых операций копирования обусловило: (а) некоторое усложнение алгоритма вычисления $H(S)$ (добавляется перебор, связанный с выбором операции копирования, которая максимально удлиняет синтезируемую последовательность); (б) выявление значительного количества комбинированных структур, представленных разнотипными повторами (см. ниже); (в) появление «эффекта маскировки», связанного с возможностью наложения структур в анализируемом фрагменте текста [6]. Следует отметить, что сложностные разложения аномальных фрагментов лишь фиксируют потенциальное многообразие локальных структур. Для получения количественных оценок по каждому типу структур приходится строить специальные алгоритмы (см, например, [7, 8]).

В рассмотренной выше мере сложности операции копирования, использующие комплементарные подстановки, делали эту меру ДНК-ориентированной, т. е. применимой лишь к конкретной языковой системе.

В общем случае (произвольный алфавит и априори неизвестная подстановка $f : \Sigma \to \Sigma$) также нельзя исключать наличия в тексте аномально длинных f-повторов. Они могут быть прямыми и симметричными. Всего имеем $2|\Sigma|!$ типов f- повторов. При построении сложностного разложения выбор прототипа максимальной длины на каждом шаге можно проводить по всем $|\Sigma|!$ подстановкам на элементах алфавита и использовать копирование в обоих направлениях. Прямой перебор по подстановкам возможен лишь при небольших размерах алфавита. В [5] авторами предложен структурный инвариант, использование которого позволяет обойти проблему факториального перебора и выявлять произвольные аномально длинные f-повторы, если таковые присутствуют в тексте. Ориентиром для выработки критерия аномальности могут служить оценки длин максимальных f-повторов в случайных последовательностях, представленные в [9]. Важно отметить, что описанная в данном абзаце мера сложности C_f вновь приобретает свой «универсальный характер» в смысле возможности применения ее к текстам любой природы, в частности, к музыкальным текстам, где f-повторам соответствуют секвентные переносы фрагментов мелодии (звуковысотные сдвиги). Однако платой за такую универсальность будет увеличение трудоемкости алгоритма.

Завершая раздел об используемом аппарате, отметим, что:

- сложностное разложение как способ представления текста в терминах повторов применимо к текстам любой длины и произвольной языковой природы, в связи с чем широко используется для сжатия данных. Нас интересует не сжатие, как таковое, а аномально длинные компоненты в разложении всего текста и фрагменты с аномально низким значением сложности, выявляемые с помощью сложностного профиля;
- спектр и специфика повторов более ярко просматривается на неструктурированных (без разделителей) текстах с малым размером алфавита. При обработке текстов со значительным размером алфавита полезным может оказаться осмысленное агрегирование алфавита (например, переход к частеречным значениям для текстов на естественном языке);
- нами разработаны модификации сложностного подхода, ориентированные на сравнение пар и групп текстов (разложение одного текста по другому или одной группы тек-

стов по другой). При малоповторности отдельных текстов значимыми могут оказаться межтекстовые повторы

Исходные данные

При выявлении межъязыковых аналогий в качестве базовых рассматривались три языковые системы, представленные: (а) биологическими текстами; (б) текстами на естественном языке; (в) знаменными песнопениями. Обработка их производилась в разные годы в связи с решением конкретных прикладных задач, связанных с формализацией выделения структурных единиц в символьных последовательностях и многоплановой классификацией.

Биологические тексты нижнего уровня – это последовательности ДНК и РНК. Авторы обрабатывали полные геномы простейших микроорганизмов (бактериофагов $\varphi x174$, $g4$, λ [10], вирусов гриппа, клещевого энцефалита, бактерий из семейства микоплазм и др.) с длинами в диапазоне от 10^3 до 10^7 символов, а также фрагменты более крупных геномов вплоть до генома пшеницы. Последний был представлен данными частичного секвенирования хромосомы 5B (свыше 100 тысяч фрагментов длиной от 50 до 600 символов по длинному плечу и вдвое меньше по короткому [11]).

Биологические тексты верхнего уровня — это последовательности генов в геномах и еще более крупных единиц — дисков в политенных хромосомах двукрылых. Эволюция на хромосомном уровне идет уже не путем точечных замен или вставок, но путем более редких крупноблочных операций: транспозиций и инверсий. Уникальная коллекция последовательностей дисков в хромосомных плечах видов рода Chironomus (комары-звонцы) собрана в Институте цитологии и генетики СО РАН. Эти данные являют собой пример бесповторных последовательностей (перестановки длиной до 150 символов), переводимых одна в другую конечным числом инверсий. Инверсия предполагает изменение порядка следования символов на обратный в выделенном фрагменте. При сравнении двух перестановок мы использовали представление одной из них в виде конкатенации минимально возможного числа фрагментов (прямых или инвертированных) из другой. Построенная нами матрица попарного сходства последовательностей дисков из упомянутой коллекции позволила на геномном уровне уточнить филогенетические связи между видами [12].

Тексты на естественном языке уже частично структурированы (разбивка на слова, предложения, абзацы...), однако часто возникает потребность в выделении промежуточных уровней иерархии, фиксирующих, например, устойчивые словосочетания или сверхфразовые единства. Первые доминируют в терминологических словарях различных предметных областей, вторые могут быть использованы для построения квазирефератов текста. Для выделения устойчивых словосочетаний могут быть использованы сложностные разложения значительных по объему предварительно нормализованных текстов с традиционной операцией прямого копирования. Для выделения сверхфразовых единств приходится строить аналог сложностного профиля (см. ниже) и использовать другую технику выявления скоплений значимых языковых единиц [13]. Материалом для указанных разработок послужили весьма объемные труды конференций по компьютерной лингвистике на русском языке (Диалог-2002 и др.) и по катализу на английском языке (EuropaCat-2005 и др.).

Уникальный материал по знаменным песнопениям собран авторами данной работы. Это певческие книги конца XVII – начала XVIII вв., в которых песнопения представлены параллельно в знаменной и нотолинейной форме (так называемые двознаменники — своего рода билингвы знаменного распева, положенные нами в основу дешифровки знаменной

нотации [14, 15, 16]). Кодирование материала проводилось вручную, поскольку программ распознавания рукописного знаменного текста не существует. На данный момент закодированы и анализируются четыре двознаменника, примерно по 200 песнопений в каждом. Длины песнопений колеблются в диапазоне от нескольких десятков до нескольких сот знамен, каждое знамя интерпретируется цепочкой от 1 до 6 нотных знаков. Текст песнопения (старославянский) синхронизован со знаменным и нотолинейным.

Систематизация локальных структур

Как уже упоминалось во введении, фиксируются локальные структуры, встречающиеся во многих эволюционирующих языковых системах. Некоторые из этих структур могут быть описаны и на «языке образцов» [17].

Совершенные периодичности (тандемные повторы) Под ними мы понимаем фрагменты текста, представимые в виде $P = x^m$, где x — произвольная цепочка символов из Σ (период), $|x| \geqslant 1$ — длина периода, $m \geqslant 2$ — кратность повторения. Этот класс структур чрезвычайно распространен в ДНК-последовательностях. Они достаточно детально описаны и систематизированы. Механизмы их возникновения известны. Длины периодов могут меняться от 1 до 10^3 и более символов, а кратность повторений доходит до сотен раз и выше. Считается, что насыщенность ДНК-последовательностей повторами способствует повышению помехоустойчивости генетического языка. Разнесенные повторы, в частности, симметричные комплементарные, часто образуют палиндромно-шпилечные конструкции, играющие важную роль в регуляции генетических процессов. Ниже приведена для иллюстрации структурная единица шпилечного типа, выявленная в позиции 27 724 генома фага λ (выделена сходящимися стрелками):

$$\ldots \overrightarrow{gctttttta}\overrightarrow{tactaagttggcat}\overleftarrow{tataaaaaagc}\ldots$$

Она, предположительно, участвует в связывании int-белка с ДНК фага λ. Легко видеть, что основу выделенных повторов составляют периодичности t^6 и a^6.

Совершенные тандемные повторы в повествовательных текстах естественного языка встречаются редко и обычно сигнализируют об ошибке редактирования (повтор слова, строки и т. д.). Повтор осмысленный представляет собой специальную конструкцию, которую хорошо проиллюстрировал Б. Заходер: «В *чаще чаще* меньше пищи, значит в *чаще чаще* чище». Здесь тандем «чаще чаще» — это омографы — слова, которые пишутся и звучат одинаково только в определенной форме (числе, падеже, времени, лице).

В стихотворных текстах тандемные повторы слов и строк встречаются довольно часто и далеко не всегда с целью усиления значения, а скорее как элемент формообразования («Однажды *вечером, вечером, вечером*, когда пилотам, прямо скажем, делать нечего. . . »). Здесь имеет место своего рода заполнение повторами стихотворной строки. Применительно к музыкальным текстам термин «заполнение интервала» используется в ситуациях, когда значительный звуковысотный скачок, например на 4 ступени, заменяется серией шагов на одну ступень, образующих в интервальном представлении тандемный повтор с длиной периода 1 и кратностью 4 ($4 = 1 + 1 + 1 + 1$).

Тандемные повторы в знаменных песнопениях претендуют, по нашему мнению [16], на роль самостоятельных структурных единиц. Они важны в плане дешифровки, поскольку тандемному повтору на знаменном уровне не всегда соответствует нотолинейный повтор и наоборот. Их функциональная нагрузка разнообразна. Серии «стопиц» $\left(\text{Г}\right)^m$ соответствуют речитативным участкам. По насыщенности ими песнопений можно судить о датировке певческих рукописей [14]. Тандемы вида $\left(\text{Ґ}\ \text{Ґ}\right)^2$, $\left(\text{Ґ}\ \text{Ґ}\ \text{Г}\right)^2$, $\left(\text{Г}\ \text{Ґ}\ \text{Ґ}\right)$

 $)^2$, ...с длиной периода от 2 до 4 и кратностью повторения 2 или 3, составленные из простых по распеву высокочастотных знамен, обычно встречаются на стыках попевок (основных структурных единиц знаменного распева), регулируя расстояния между ними путем изменения длины периода и кратности повторений. Здесь явная аналогия с заполнением интервалов, но уже не на высотном, а на позиционном уровне. Тандемы «статей» в попевках или цепочек знамен, заканчивающихся статьей, например,  $)^2$,  $)^2$,  $)^2$ и др. усиливают кадансовую структуру попевок. И, наконец, короткие тандемы из достаточно сложных по распеву и редко используемых знамен являются индикаторами начертаний лиц и фит — наиболее ярких и нестандартных структурных единиц знаменного распева, служащих для его украшения. Такие индикаторы помогают вычленять эти структурные единицы (в первую очередь, лица) из текстов песнопений. К сожалению, этот признак носит факультативный характер: не все лица и фиты снабжены им.

Несовершенные периодичности (тандемные повторы с искажениями). Практически во всех эволюционирующих языковых системах наряду с совершенными повторами встречаются и несовершенные. Доля последних как минимум сопоставима с совершенными повторами, а чаще всего превалирует. Характер искажений на нижних уровнях языковой иерархии в большинстве случаев точечный: одиночные замены, короткие вставки и делеции (устранения) символов. На более высоких уровнях те же операции приобретают «блочный» характер, т. е. применяются к цепочкам символов (примером в русском языке может служить замена корней слов при сохранении аффиксального окружения [18]). Характерной для верхних уровней является операция транспозиции, связанная с переносом языковых форм из одного места текста в другое или (в музыкальных текстах) с одного звуковысотного уровня на другой (секвентные переносы). Последовательности дисков в политенных хромосомах искажаются путем инверсий [12] и т. д.

Единого определения, как уже отмечалось, и, соответственно, алгоритма отыскания несовершенной периодичности не существует. В немалой степени это обусловлено многообразием возможных способов искажения последовательностей. Для простейшей модели порождения ДНК-последовательностей, путем дупликации фрагментов произвольной длины с последующими их точечными искажениями, предложено несколько достаточно эффективных алгоритмов [19, 20], дающих близкие, но не тождественные результаты.

В сложностных разложениях несовершенные периодичности проявляют себя при наличии в периодах «совершенных» ядер, что имеет место достаточно часто. Важно отметить, что точечные мутации , накладывающиеся на совершенную периодичность, зачастую способствуют формированию регуляторных структур. Так, приводимый ниже фрагмент генома бактериофага λ содержит несовершенную периодичность с периодом длины 13 (выделен скобками):

$$\begin{array}{c} \text{поз. } 6112 \downarrow \qquad \text{I} \qquad\quad RBS \qquad \text{II} \quad \text{начало гена } E \\ \ldots(g\ g\,c\,t\,t\,t\,t\,t\,t\,t\,a\,c\ \underline{g)(g\,g\,a}\ t\,t\,t\,t\,t\,t\,t\ \overrightarrow{\underline{a\,t\,g)}}\,t\,c\,g\ldots \end{array}$$

Две замены, которыми II отличается от I формируют рибосомный сайт связывания RBS (подчеркнут) и инициирующий кодон atg, т.е. элементы, обеспечивающие начало трансляции гена E. В [10] приводятся и другие примеры на эту тему.

Интересным частным случаем совершенных и несовершенных периодичностей являются фрактальные и фракталоподобные структуры. Так мы называем периодичности, образованные повторением палиндрома, например ($aga\,aga\,aga\,\ldots$) или комплементарного палиндрома ($acgt\,acgt\,acgt\ldots$). Применительно к совершенным периодичностям такого

рода используется термин фрактальная структура, а к несовершенным — фракталоподобная. Это связано с проявлениями самоподобия в том смысле, что повторение палиндрома любого типа приводит к образованию аналогичной структуры вдвое большей длины, т.е. имеет место «усиление закономерности» (см. примеры «а» и «б»)

$$\text{(а)}\ \ldots \overbrace{tac\,cat}\,\overbrace{tac\,cat}\ldots;\ \text{б)}\ \overrightarrow{actg}\,\overleftarrow{cagt}\,\overrightarrow{actg}\,\overleftarrow{cagt}.$$

Здесь расходящиеся стрелки соответствуют палиндромам (случай «а»), а сходящиеся — комплементарным палиндромам (случай «б»).

В [8] авторами описан алгоритм отыскания фрактальных и фракталоподобных структур в режиме скользящего окна для случая, когда искажение самих палиндромов не допускается, но возможны вставки между ними, размер которых не превышает заданного порога r. Фрагмент *agagaagactagattcaagatcaga*, например, при $r = 4$ относится к категории фракталоподобных структур с повторяющимся (базовым) палиндромом «aga» .

Периодичности со сложной структурой . Многие периодичности (как совершенные, так и несовершенные) имеют иерархическую структуру в том смысле, что внутри большого периода могут, в свою очередь, присутствовать периодичности с меньшей длиной периода или другие регулярности. Так, в [21, 22], например, рассматриваются структуры вида $(Xx^n)^m$, где X и x — цепочки символов, $X \neq x$, n и m — целые, большие 1, а также структуры с переменным значением $n : x^{n_0}Xx^{n_1}X\ldots x^{n_{l-1}}Xx^{n_l}$, где $l > 2$, $n_i \geqslant 1$ для $i = 1, \ldots, l-1$ и хотя бы одно из $n_i \geqslant 2$. Допускается наличие ограниченных искажений в X и x. Фактически речь идет об обнаружении тандемных повторений цепочки x, прерываемых одиночными вставками цепочки X, причем количество таких вставок должно быть не менее трех. Представляет интерес то, что расстояния между вставками X регулируются количеством и длиной тандемных повторов x. О таком способе разнесения значимых структурных единиц (в данном случае цепочек X) на «нужное расстояние» мы упоминали в связи с обсуждением функциональной нагрузки тандемных повторов в знаменных песнопениях.

Другой интересный случай вставок теперь уже неидентичных цепочек в периодическую структуру выявлен нами с помощью сложностного разложения в геноме микоплазмы «Mycoplasma synoviae 53» (ID AE017245) . Соответствующий фрагмент текста (начальная позиция 690229) представлен в виде выравнивания:

$$\begin{array}{cc} X & Y \\ gttttggggttgtacaattattttgttaagtaaaac & aaatgataaataacgcttaacttgcttact \\ gttttggggttgtacaattattttgttaagtaaaac & cctataaacaaatcaggattatatgtacta \\ gttttggggttgtacaattattttgttaagtaaaac & ttaagtcaagatttttaataccagggtgca \\ gttttggggttgtacaattattttgttaagtaaaac & tccatatttttccttactattactatgct \text{ и т. д.} \end{array}$$

Структура имеет вид $XY_1XY_2XY_3\ldots$ (свыше 10 повторений), где $X = gt^4g^4t^2gtaca^2t^2at^4gt^2a^2gta^4c$ ($|X| = 36$) — регулярно повторяющийся фрагмент, а вставки Y_i, $i = 1, 2\ldots$, $|Y_i| \approx 30$, не обнаруживают значимого сходства. Информация об этой структуре в разметке генома отсутствует. Аналогичная, но более сильная структура указана в разметке другого представителя этого семейства «Mycoplasma gallisepticum str. R(low)» (ID AE015450). По-видимому, она выявлена с помощью инструмента CRISPFinder [23].

Подводя итог, отметим, что граница между совершенными и несовершенными периодичностями может быть размытой, если искажения носят регулярный характер. Так, приведенная в [21] структура $((cagta)(cagca)(cagta)(caaca))^3$, представленная здесь как совершенная периодичность с длиной периода 20, может рассматриваться и как несовершенная

периодичность с длиной периода 5, отличающаяся от $(cagta)^{12}$ девятью заменами, и как несовершенная периодичность с длиной периода 10, отличающаяся от $((cagta)(cagca))^6$ тремя заменами. Структуры, допускающие многозначную трактовку, выделены авторами [21] в отдельный класс.

Кумулятивные структуры. Это интересный класс структур, характеризующийся специфическими проявлениями повторности. Длина повторяющейся единицы нарастает или убывает на каждом шаге. Схема формирования кумулятивной структуры пошаговая. Например, в известном переводе С. Маршака английской песенки «The house that Jack built» эта схема имеет вид: $A + BA + CBA + \cdots$

$A =$ Вот дом, который построил Джек.
$BA =$ А это пшеница, которая в темном чулане хранится в доме, который построил Джек.
$CBA =$ А это веселая птица — синица, которая часто ворует пшеницу, которая в темном чулане хранится в доме, который построил Джек ...

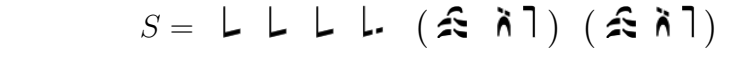
Подобные схемы встречаются и в ДНК-последовательностях. Приведем одну из них, в которой дуплицируются уже последовательно уменьшающиеся фрагменты из генома фага λ, представленные для наглядности в виде выравнивания:

```
Позиции
21479:  g a t t t g - - g g g a c g a a a a c c a c a  t  c g t c g t t t
21509:  g a t t a c - - g g g a c g a a a a c a a c a - g g c a g t t t c
21539:  g a t t a c - - g g c a c c a a a t c g a c g
21560:  a a t a a c a  c g g g g
```

В [10] подобного рода схемы названы «иерархической дупликацией», а в музыке — «вариационностью типа прорастания» [2], характеризующейся тем, что тематическое ядро при повторении получает новое продолжение. И, наконец, стоит упомянуть целый пласт детских сказок («Репка», «Колобок» и др.), классифицируемых как цепочные, кумулятивные или сказки с «нанизыванием», т. е. присоединением новой информации при повторении уже известной [3]. Считается, что такие сказки способствуют развитию устной речи у детей.

Комбинированные структуры. К ним мы относим позиционно сближенные (или даже налагающиеся одна на другую) комбинации элементарных структур (разнотипных повторов) и периодичностей. Простейшим примером могут служить «компаунды» — примыкающие друг к другу периодичности, отличающиеся составом элементов в периодах или даже длиной периода. Так, в упомянутой выше подборке секвенированных фрагментов генома пшеницы мы наблюдали компаунды $(ca)^{19}(at)^{14}$; $(ag)^7(agat)^{10}$; $(tc)^{11}(ta)^{12}(ac)^7$ и много других. Обращает на себя внимание наличие общего элемента в разнотипных периодах каждого компаунда: чаще всего на нем и осуществляется стыковка соседних периодичностей. На данный момент не совсем понятно, несет ли какую-то смысловую нагрузку позиционное сближение периодичностей, образующих компаунд. Применительно к естественному языку этот вопрос звучал бы так: образует ли какая-то пара соседних слов в тексте устойчивое словосочетание или нет?

Компаунды встречаются и в других языковых системах. Песня «В темном лесе» представляет собой совершенный компаунд. Приведем один пример из знаменных песнопений:



Здесь компаунд образован серией «стопиц» $\left(\text{┗}\right)^3$ и тандемным повторением цепочек длины 3 (в скобках). За компаундом следует «статья мрачная» $\left(\text{}\right)$, которая интерпретируется целой нотой, обычно завершающей структурные единицы знаменного распева. С большой вероятностью фрагмент $S[4:11]$ относится к категории «лиц», что подтверждается индикатором в виде тандемного повтора и ритмическим остановом, реализуемым статьей. Распевам лиц и фит нередко предшествует речитативный участок, называемый «разбегом стопиц». Здесь он представлен серией из трех стопиц. Таким образом, позиционное сближение двух видов тандемных повторов выглядит закономерным.

Пример наложения разнотипных ДНК-повторов иллюстрирует фрагмент эукариотического промотора (регуляторной структуры, ответственной за начало транскрипции) [7]:

$$\ldots ag\,\overset{1}{\overline{g}}\,c\,\overrightarrow{cgggc}^{\,1}\,g\,\overleftarrow{ccgcct}^{\,2}\,\overrightarrow{tccgcc}^{\,2}\,t\,\overleftarrow{gcccg}^{\,1}\,c\,\overset{1}{\overline{c}}\,t\ldots$$

Здесь симметричный повтор (2) фланкирован разнесенными симметричными комплементарными повторами (1).

Межъязыковые аналогии на уровне содержательных задач и подходов к их решению

Описанные выше образцы структурного сходства, обнаруживаемого в текстах различных языковых систем, позволяют предположить, что межъязыковые аналогии распространяются также на постановку содержательных задач и выработку подходов к их решению. Возможны переносы идей и технологий из одной языковой системы в другую. Упомянем в связи с этим сходные по постановкам и используемым алгоритмам задачи выявления гомологий в ДНК-последовательностях, неосознанных заимствований в музыкальных произведениях, плагиатов в научных текстах. Другим примером может служить формализация понятия «структурная единица» и разработка алгоритмов автоматического выделения структурных единиц из текстов конкретных языковых систем. В частности, сходные подходы использованы для выделения: (а) морфем из текста на русском языке, записанного без пробелов и знаков препинания [24]; (б) попевок — основных структурных единиц знаменного распева [15]; (в) устойчивых словосочетаний (коллокаций) из русскоязычных текстов [25].

Чуть подробнее осветим возможность переноса понятийного аппарата из одной языковой системы в другую. Так, весьма плодотворным в плане выявления локальных структур в биологических текстах оказалось понятие «сложностного профиля». Возникает вопрос, что могло бы служить его аналогом для текстов на естественном языке, где проявления повторности не так заметны? Напомним, что аномальные по сложности фрагменты в биологических текстах — это участки с высокой концентрацией разнотипных повторов. Поэтому в текстах на естественном языке следует обратить внимание на фрагменты, характеризующиеся аномально высокой (по сравнению с оставшейся частью текста) концентрацией вхождений какой-либо лексической единицы (слова или словосочетания). Такого рода закономерности будем называть позиционной кластеризацией лексических

единиц. Выделенные фрагменты, соответствующие разным лексическим единицам, могут пересекаться, оказаться вложенными один в другой или разнесенными. Одно и то же предложение текста может покрываться разными фрагментами. Имея эту информацию, можно определить профиль кластеризуемости лексических единиц в тексте как ступенчатую функцию, аргументом которой является порядковый номер предложения в тексте, а значение равно числу различных фрагментов, включающих в себя данное предложение. Поскольку каждый фрагмент связан с конкретной лексической единицей, то в каждой точке профиля фиксируется совокупность лексических единиц, определяющих локальное содержание данного участка текста.

В основе построения профиля кластеризуемости лежит отбор лексических единиц, демонстрирующих аномалии в позиционном распределении. Для этого используется аппарат сканирующих статистик . В частности, статистика $d(n, x)$ фиксирует длину минимального интервала, содержащего ровно n последовательных вхождений лексической единицы x $(2 \leqslant n \leqslant F(x))$, где $F(x)$ — частота встречаемости x в тексте. Распределение этой статистики при случайной расстановке x вдоль текста известно. Если наблюдаемое значение $d(n, x)$ при каком-то значении n аномально мало по сравнению с ожидаемым при равномерном распределении, фиксируется наличие позиционного кластера. В [13] описанный подход рассмотрен более детально. Пикам профиля кластеризуемости сопоставлены структурные единицы более высокого уровня, чем предложение — так называемые «сверхфразовые единства», определяющие макроструктуру текста. Предложен способ построения квазиреферата текста на основе профиля кластеризуемости.

Заключение

Основными структурообразующими элементами в текстах, представляющих различные эволюционирующие языковые системы, являются повторы. Номенклатура повторов очень широка. Существуют повторы, типичные для всех языковых систем и присущие лишь отдельным языковым системам. Повторяющиеся цепочки символов отличаются своей длиной, составом элементов алфавита, частотой встречаемости в тексте и характером распределения по длине текста. Наиболее значимыми считаются цепочки, распределенные неравномерно. Одним из наиболее типичных проявлений неравномерности позиционного распределения является высокая концентрация повторов разного типа в ограниченном участке текста. В работе проведена типизация таких участков, сопровождаемая примерами из различных языковых систем (биологические последовательности, тексты на естественном языке, знаменные песнопения). Обсуждается их функциональная нагрузка в разных языковых системах.

Особое внимание уделено межъязыковым аналогиям. Предполагается, что структурам, наблюдаемым в одной языковой системе, могут быть найдены аналоги и в других языковых системах. Более того, межъязыковые аналогии могут распространяться и на постановку различных содержательных задач, а также отыскание способов их решения.

Нами разработаны в рамках сложностного подхода и апробированы на реальном материале эффективные (квазилинейные) алгоритмы отыскания описанных выше структур (см. [5, 7, 8, 11, 12, 13, 16, 25]). Из решенных (и решаемых) прикладных задач отметим разработку формальных методов для выделения (и дешифровки) структурных единиц знаменного распева [15, 16], сверхфразовых единств в текстах на естественном языке, используемых для автоматического построения квазирефератов текста [13], фракталоподобных и комбинированных структур в частично секвенированном геноме пшеницы [11]. Последние могут выполнять роль регуляторов основных генетических процессов — тран-

скрипции, трансляции и др. Разработаны модификации метода для выявления сходства (различия) пар или групп текстов, допускающих нестандартные редакционные операции, в частности, инверсии (в перестановках) и транспозиции. Они использованы для структурного (и филогенетического) анализа уникальных данных — последовательностей дисков политенных хромосом [12]. Эти же методы применимы для установления авторства литературного или музыкального произведения, но результаты существенно зависят от объема предоставленного материала. В качестве удачного примера можно указать на работу [26], близкую в идейном плане к описываемому подходу.

Литература

[1] *Lempel A., Ziv J.* On the complexity of finite sequences // *IEEE Trans. Inf. Theor.*, 1976. Vol. IT-22, no. 1. P. 75–81.

[2] *Протопопов В.* Вариационные процессы в музыкальной форме. М.: Музыка, 1967. 150 с.

[3] *Пропп В. Я.* Кумулятивная сказка // *Фольклор и действительность. Избр. статьи.* М.: Наука, 1976. С. 242–249.

[4] *Колмогоров А. Н.* Три подхода к определению понятия «количество информации» // *Проблемы передачи информации*, 1965. Т. 1, № 1. С. 3–11.

[5] *Gusev V. D., Nemytikova L. A., Chuzhanova N. A.* On the complexity measures of genetic sequences // *Bioinformatics*, 1999. Vol. 15, no. 12. P. 994–999.

[6] *Гусев В. Д., Мирошниченко Л. А.* Использование сложностных разложений в задачах анализа символьных последовательностей // *Докл. 8-й Междунар. конф. «Интеллектуализация обработки информации» (ИОИ-2010).* Кипр, Пафос, 2010. С. 469–472.

[7] *Гусев В. Д., Мирошниченко Л. А.* Поиск комбинированных структур в ДНК-последовательностях // *Докл. Всеросс. конф. ММРО-13 «Математические методы распознавания образов».* М.: Макс-Пресс, 2007. С. 473–476.

[8] *Гусев В. Д., Мирошниченко Л. А., Чужанова Н. А.* Выявление фракталоподобных структур в ДНК-последовательностях // *Classification, forecasting, data mining. Information science and computing international book ser.* Sofia: ITHEA, 2009. No. 8. P. 117–123.

[9] *Михайлов В. Г., Шойтов А. М.* О числах множеств эквивалентных цепочек в последовательности независимых случайных величин // *Математические вопросы криптографии*, 2013. Т. 4. № 1. С. 77–86.

[10] *Гусев В. Д., Куличков В. А., Чупахина О. М.* Сложностной анализ генетических текстов (на примере фага λ). Новосибирск, 1989. Препринт № 20 ИМ СО РАН. 41 с.

[11] *Sergeeva E. M., Afonnikov D. A. , Koltunova M. K. , Gusev V. D., Miroshnichenko L. A., Vrána J., Kubaláková M., Poncet C. , Sourdille P. , Feuillet C., Doležel J., Salina E. A.* Common wheat chromosome 5B composition analysis using low-coverage 454 sequencing // *Plant Genome*, 2014. Vol. 7, no. 2. doi: 10.3835/plantgenome2013.10.0031.

[12] *Gunderina L. I., Kiknadze I. I., Istomina A. G., Gusev V. D., Miroshnichenko L. A.* Divergence of polytene chromosome band sequences as a reflection of evolutionary reorganization of the linear structure of the genome // *Rus. J. Genet.*, 2005. Vol. 41, no. 2. P. 130–137.

[13] *Гусев В. Д., Мирошниченко Л. А., Саломатина Н. В.* Тематический анализ и квазиреферирование текста с использованием сканирующих статистик // *Компьютерная лингвистика и интеллектуальные технологии. Тр. Междунар. конф. Диалог'2005.* М.: Наука, 2005. С. 121–125.

[14] *Бражников М. В.* Пути развития и задачи расшифровки знаменного распева XII–VIII веков. Л., М.: Гос. муз. изд., 1949. 103 с.

[15] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* L-граммные азбуки для дешифровки знаменных песнопений // *Сибирский журнал индустриальной математики*, 1998. Т. 1, № 2. С. 51–66.

[16] *Бахмутова И. В., Гусев В. Д., Мирошниченко Л. А., Титкова Т. Н.* Тандемные повторы в знаменных песнопениях // *Анализ структурных закономерностей: Вычислительные системы*, 2005. Вып. 174. С. 13–28.

[17] *Matescu A., Salomaa A.* Aspects of classical laguage theory // *Handbook of formal languages*, 1996. Vol. 1. P. 230–242.

[18] *Саломатина Н. В.* Количественные исследования морфемной структуры слов русского языка (на базе электронного словаря Д. Уорта) // *Обнаружение эмпирических закономерностей: Вычислительные системы*, 1999. Вып. 166. С. 104–118.

[19] *Benson G.* Tandem repeats finder: A program to analyze DNA sequences // *Nucleic Acids Res.*, 1999. Vol. 27, no. 2. P. 573–580.

[20] *Sokol D., Benson G., Tojeira J.* Tandem repeats over the edit distance // *Bioinformatics*, 2007. Vol. 23, no. 2. P. e30–e35.

[21] *Hauth A. M, Joseph D. A.* Beyond tandem repeats: Complex pattern structures and distant regions of similarity // *Bioinformatics*, 2002. Vol. 18. Suppl. 1. P. s31–s37.

[22] *Matroud A. A., Hendy M. D, Tuffley C. P.* NTRFinder: a software tool to find nested tandem repeats // *Nucleic Acids Res.*, 2012. Vol. 40, no. 3. P. e17.

[23] *Grissa I., Vergnaud G., Pourcel C.* CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats // *Nucl. Acids Res.*, 2007. Vol. 35. Suppl. 2. P. W52–W57.

[24] *Сухотин Б. В.* Морфологический анализ текста без пробелов // *Оптимизационные методы исследования языка.* М.: Наука, 1976. С. 73–169.

[25] *Гусев В. Д, Саломатина Н. В.* Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) // *Компьютерная лингвистика и интеллектуальные технологии. Тр. Междунар. конф. Диалог'2004.* М.: Наука, 2004. С. 530–535.

[26] *Кукушкина О В., Поликарпов А. А., Хмелев Д. В.* Опреление авторства текста с использованием буквенной и грамматической информации // *Проблемы передачи информации*, 2001. Т. 37. Вып. 2. С. 96–109.

References

[1] *Lempel A., Ziv J.* 1976. On the complexity of finite sequences. *IEEE Trans. Inf. Theor.* IT-22(1):75–81.

[2] *Protopopov V.* 1967. Variation processes in a musical form. Moscow: Music. 150 p.

[3] *Propp V. Ia.* 1976. Cumulative fairy tale. *Folklore and Reality.* Moscow. 242–249.

[4] *Kolmogorov A. N.* 1965. Three approaches to the quantitative definition of information *Problemy Peredachi Informatsii [Problems of Information Transmission]* 1(1):3–11.

[5] *Gusev V. D., Nemytikova L. A., Chuzhanova N. A.* 1999. On the complexity measures of genetic sequences. *Bioinformatics* 15(12):994–999.

[6] *Gusev V. D., Miroshnichenko L. A.* 2010. Complexity decompositions in problems of symbolic sequences analysis. *8th Conference (International) "Intelligent Information Processing" (IIP-2010).* Cyprus, Paphos. 469–472.

[7] *Gusev V. D., Miroshnichenko L. A.* 2007. Detection of the combined structures in DNA sequences. *13th All-Russian Conference "Mathematical methods of pattern recognition".* Zelenogorsk (Leningrad Region). 473–476.

[8] *Gusev V. D., Miroshnichenko L. A., Chuzhanova N. A.* 2009. Identification of the fractal-like structures in DNA sequences. *Classification, forecasting, data mining.* Information science and computing international book ser. Sofia: ITHEA. 8:117–123.

[9] *Mikhailov V. G., Shoitov A. M.* 2013. About numbers of sets of equivalent chains in sequence of independent random variables. *Mathematical Questions Cryptography* 4(1):77–86.

[10] *Gusev V. D., Kulichkov V. A., Chupakhina O. M.* 1989. *Complexity analysis of genetic texts (on the example of a phage λ).* Novosibirsk. Preprint No. 20 IM SB RAS. 41 p.

[11] *Sergeeva E. M., Afonnikov D. A., Koltunova M. K., Gusev V. D., Miroshnichenko L. A., Vrána J., Kubaláková M., Poncet C. , Sourdille P. , Feuillet C., Doležel J., and Salina E. A.* 2014. Common wheat chromosome 5B composition analysis using low-coverage 454 sequencing. *Plant Genome* 7(2). doi: 10.3835/plantgenome2013.10.0031.

[12] *Gunderina L. I., Kiknadze I. I., Istomina A. G., Gusev V. D., Miroshnichenko L. A.* 2005. Divergence of polytene chromosome band sequences as a reflection of evolutionary reorganization of the linear structure of the genome. *Rus. J. Genet.* 41(2):130–137.

[13] *Gusev V. D., Miroshnichenko L. A., SalomatinaN. V.* 2005. The thematic analysis and quasiabstracting of the text with the scan statistics using. *Conference (International) "Computational Linguistics and Intellectual Technologies" (Dialogue-2005).* Moscow. 121–125.

[14] *Brazhnikov M. V.* 1949. *Puti razvitiia i zadachi rasshifrovki znamennogo rospeva XII-XVIII vekov.* Leningrad; Moscow: Gos. Muz. Izd-vo. 103 p. (In Russian).

[15] *Bakhmutova I. V., Gusev V. D., Titkova T. N.* 1998. *L*-gramm alphabet for deciphering the neume hymns. *J. Appl. Ind. Math.* 1(2):51–66.

[16] *Bakhmutova I. V., Gusev V. D., Miroshnichenko L. A., Titkova T. N.* 2005. Tandem repeats in the neume hymns. *Computing Syst. Analysis of structural regularities* 174:13–28.

[17] *Matescu A., and Salomaa A.* 1996. Aspects of classical laguage theory. Handbook of Formal Languages. 1:230–242.

[18] *Salomatina N. V.* 1999. Quantitative researches of morphemic structure of words of Russian (on the basis of the electronic dictionary of D. Worth) *Computing systems. Detection of empirical regularities* 166:104–118.

[19] *Benson G.* 1999. Tandem repeats finder: A program to analyze DNA sequences *Nucleic Acids Res.* 27(2):573–580.

[20] *Sokol D., Benson G., Tojeira J.* 2007. Tandem repeats over the edit distance. *Bioinformatics* 23(2):e30–e35.

[21] *Hauth A. M., Joseph D. A.* 2002. Beyond tandem repeats: Complex pattern structures and distant regions of similarity. *Bioinformatics* 18(1):s31–s37.

[22] *Matroud A. A., Hendy M. D., Tuffley C. P.* 2012. NTRFinder: A software tool to find nested tandem repeats. *Nucleic Acids Res.* 40(3):e17.

[23] *Grissa I., Vergnaud G., Pourcel C.* 2007. CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucl. Acids Res.* 35(2):W52–W57.

[24] *Sukhotin B. V.* 1976. Morphological analysis of the text without gaps. *Optimization methods in language research.* Moscow: Nauka. 73–169.

[25] *Gusev V. D., Salomatina N. V.* 2004. Algorithm of identification of stable word combination taking into account their variability (morphological and combinatory). *Conference (International) "Computational Linguistics and Intellectual Technologies" (Dialogue-2004).* Moscow. 530–535.

[26] *Kukushkina O V., Polikarpov A. A., Khmelev D. V.* 2001. Using literal and grammatical statistics for authorship attribution. *Problems Information Transmission* 37(2):172–184.