

# СРАВНЕНИЕ БЛИЗКИХ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

В. Д. Гусев, Л. А. Мирошниченко, Т. Н. Титкова

*Институт математики им. С.Л.Соболева СО РАН, 630090, Новосибирск*

УДК 004.93

Во многих приложениях, связанных с анализом символьных последовательностей (текстов) произвольной языковой природы, встречаются объекты, похожие друг на друга (разные переводы одного и того же произведения, кодирующие последовательности родственных генов, распевы песнопения в разных гласах и др.). Традиционно используемые меры близости оказываются в таких случаях малоинформативными. Представляет интерес не количественная оценка близости, как правило, высокая, а детали, позволяющие различать близкие последовательности. В работе рассматриваются методы выявления таких различий. В качестве признаков, характеризующих определенный класс последовательностей, используются цепочки символов (фрагменты текстов), присутствующие в обучающих последовательностях одного из классов, но отсутствующие в других. Рассматриваются как  $L$ -граммы (фрагменты фиксированной длины  $L$ ), так и фрагменты произвольной длины со специфическими структурными особенностями. Изложение иллюстрируется примерами анализа геномов вируса клещевого энцефалита с близкими последовательностями, относящимися, тем не менее, к разным классам.

**Ключевые слова:** символьные последовательности (тексты), меры близости,  $L$ -граммы, периодичности, фракталоподобные структуры, комбинированные структуры.

## Введение

Символьные последовательности встречаются в различных областях знания: математике, информатике, биологии, лингвистике, музыке. Для сравнения последовательностей в первом приближении можно использовать такие признаки как размер алфавита, частоту встречаемости символов и т.п. В общем случае для сравнения последовательностей и их классификации используются меры близости, выбор которых осуществляется на основе операций, характерных для конкретной предметной области. На уровне слов естественного языка, такими операциями являются замены, вставки, пропуски отдельных символов, приводящие к ошибке или изменению смысла слова. Те же операции характеризуют эволюцию ДНК-последовательностей. На уровне предложений возможны вставки и удаления отдельных слов или их комбинаций (например, причастных оборотов), синонимичные замены, перестановки, часто демонстрирующие «игру слов» («лучше обед без аппетита, чем аппетит без обеда» — М. Жванецкий) и т.д. На хромосомном уровне основными эволюционными событиями являются крупноблочные перестройки: инверсии и транспозиции. Специфическими операциями для музыкальных произведений являются секвентные переносы и заполнение интервалов.

Значительный интерес представляют меры близости, применимые на разных иерархических уровнях. Общий подход для конструирования таких мер, восходящий к работам Колмогорова [1], основан на оценке сложности перевода одной последовательности в другую с использованием фиксированного набора операций, допустимых на конкретном уровне. Примерами такого рода мер являются редакционное расстояние и относительная сложность [2].

В случае близких последовательностей исследователя может интересовать не столько расстояние между ними, сколько детали (признаки), отличающие одну последовательность от другой или один класс последовательностей от другого. К таким признакам можно отнести фрагменты, присутствующие в одном классе текстов и отсутствующие в остальных. В работе рассматриваются фрагменты фиксированной длины ( $L$ -граммы) и произвольной, характеризующиеся скоплениями разных повторов. Подход иллюстрируется на подборке геномов вируса клещевого энцефалита. Отслеживаются две линии классификации этих последовательностей: по принадлежности их к одному из четырех генотипов и по степени вирулентности.

# 1 Сравнение текстов путем выделения «характерных» фрагментов

## 1.1 $L$ -граммный подход к представлению текстов

Пусть  $\Sigma$  — конечное множество символов (алфавит);  $T = t_1 t_2 \dots t_N$  — последовательность символов (текст), составленный из элементов  $\Sigma$  ( $t_i \in \Sigma, 1 \leq i \leq N$ );  $N = |T|$  — длина текста;  $T[i : j] = t_i t_{i+1} \dots t_j$  — фрагмент текста, включающий элементы с  $i$ -го по  $j$ -й ( $1 \leq i \leq j \leq N$ ).  $L$ -граммой называется фрагмент текста  $T[i : i + L - 1]$  ( $1 \leq i \leq N - L + 1$ ), содержащий ровно  $L$  символов. Количество различных  $L$ -грамм в  $T$  обозначим через  $M_L(T)$  ( $M_L(T) \leq N - L + 1$ ), а частоту встречаемости  $k$ -й ( $1 \leq k \leq M_L(T)$ )  $L$ -граммы — через  $F_k$ .

$L$ -граммная частотная характеристика текста  $T$  есть совокупность элементов  $\Phi_L(T) = \{\phi_{L1}, \phi_{L2}, \dots, \phi_{LM_L}\}$ , где каждый элемент  $\phi_{Lk}$  ( $1 \leq k \leq M_L$ ) есть пара: « $k$ -я  $L$ -грамма;  $F_k$ ». Полный  $L$ -граммный спектр текста  $T$  [3] есть множество  $\Phi(T) = \{\Phi_1(T), \Phi_2(T), \dots, \Phi_{L_{\max}}(T)\}$ , т.е. совокупность  $L$ -граммных характеристик для  $1 \leq L \leq L_{\max}(T)$ , где  $L_{\max}(T)$  — длина максимального повтора в тексте.

$L$ -граммная частотная характеристика группы текстов  $\Pi = \{T_1, T_2, \dots, T_m\}$  есть совокупность  $L$ -грамм, представленных в текстах  $\{T_1, T_2, \dots, T_m\}$  с указанием частот их встречаемости и распределения по отдельным текстам:  $\Phi_L(\Pi) = \{\phi_{L,1}(\Pi), \phi_{L,2}(\Pi), \dots, \phi_{L,M_L}(\Pi)\}$ , где каждый элемент  $\phi_{L,i}$  ( $1 \leq i \leq M_L(\Pi)$ ) есть четвёрка: « $i$ -я  $L$ -грамма  $x_i$ ;  $F_{\Pi}(x_i)$  — число текстов из  $\Pi$ , в которых представлена  $x_i$ ;  $F_{\Sigma}(x_i)$  — абсолютная суммарная частота встречаемости  $x_i$  в  $\Pi$ ; вектор числа вхождений  $x_i$  в каждый из текстов подборки  $\Pi$ :  $\bar{F}(x_i) = \{F_1(x_i), F_2(x_i), \dots, F_m(x_i)\}$ ». Совокупность  $L$ -граммных характеристик  $\Phi(\Pi) = \{\Phi_1(\Pi), \Phi_2(\Pi), \dots, \Phi_{L_{\max}}(\Pi)\}$ , где  $L_{\max} = L_{\max}(\Pi)$  — длина максимального межтекстового повтора в  $\Pi$ , будем называть совместным  $L$ -граммным спектром группы текстов  $\Pi$ . Иными словами, мы включаем в  $\Phi(\Pi)$  такой набор  $L$ -граммных частотных характеристик, который содержит полную информацию о связях между текстами в виде общих цепочек с длинами  $L \leq L_{\max}(\Pi)$ . Заметим, что длины максимальных внутритекстовых повторов  $L_{\max}(T_j)$ ,  $j = 1, 2, \dots, m$ , могут быть как меньше, так и больше длины максимального межтекстового повтора  $L_{\max}(\Pi)$ , поскольку механизмы их возникновения различны.

В случае *нескольких* ( $k$ ) *классов текстов*  $K = \{\Pi_1, \Pi_2, \dots, \Pi_k\}$ , где  $(\Pi_i = \{T_{i1}, T_{i2}, \dots, T_{im_i}\}, 1 \leq i \leq k)$ , а  $m_i$  — количество последовательностей в  $i$ -м классе, рассматривается характеристика  $\Phi_L(\Pi_i/K)$  — множество  $L$ -грамм, присутствующих в каждом тексте  $i$ -го класса и отсутствующих во всех остальных текстах. Такие «контрастные»  $L$ -граммы удобно использовать в качестве признаков для классификации новых текстов. Интерес представляет весь диапазон значений  $L$ : от минимальной длины до максимальной. Фрагменты максимальной длины — консервативные фрагменты последовательностей, наиболее устойчивые в пределах одного класса. Однако эта характеристика зачастую зависит от количества текстов в классе.

Для вычисления  $L$ -граммных характеристик, совместных  $L$ -граммных спектров группы текстов и  $\Phi_L(\Pi_i/K)$  используется алгоритм рекуррентного хеширования [4].

## 1.2 $L$ -граммно-позиционный «срез» группы последовательностей

$L$ -граммный подход применим к произвольным текстам и их классам. В случае «близких», предварительно выравненных по длине последовательностей, в качестве классификационных признаков можно использовать  $L$ -граммно-позиционный «срез», т.е. совокупность  $L$ -грамм, привязанных к определенной позиции.

Пусть  $N$  — длина последовательностей после выравнивания,  $T_{i,j}[pos : pos + L - 1]$  —  $L$ -грамма, расположенная в позиции  $pos$   $j$ -го текста  $i$ -го класса. Через  $\Omega_{L,pos}(K) = \{T_{i,j}[pos : pos + L - 1]\}$  обозначим полную совокупность таких  $L$ -грамм.  $L$ -граммная позиционная характеристика  $\Lambda_{L,pos}(\Pi_i)$  — множество различных  $L$ -грамм, расположенных в позиции  $pos$  в текстах класса  $\Pi_i$ .  $M_{L,pos}(\Pi_i)$  — число различных  $L$ -грамм, расположенных в позиции  $pos$  в текстах класса  $\Pi_i$  (мощность множества  $\Lambda_{L,pos}(\Pi_i)$ ).

Если  $M_{L,pos}(\Pi_i) = M_{L,pos}(\Pi_j) = 1$  и  $\Lambda_{L,pos}(\Pi_i) = \Lambda_{L,pos}(\Pi_j)$  ( $1 \leq i, j \leq k$ ), т.е. во всех последовательностях, независимо от класса, в позиции  $pos$  стоит одна и та же  $L$ -грамма, то позиция  $pos$  определяет консервативный фрагмент. Такие фрагменты характеризуют набор текстов в целом.

С точки зрения классификации текстов наиболее интересным является случай, когда во всех последовательностях одного класса в позиции  $pos$  стоит  $L$ -грамма, уникальная для своего класса и присущая только ему, т.е.  $M_{L,pos}(\Pi_i) = M_{L,pos}(\Pi_j) = 1$ , и  $\Lambda_{L,pos}(\Pi_i) \neq \Lambda_{L,pos}(\Pi_j)$  ( $1 \leq i, j \leq k, i \neq j$ ). В общем случае, позиция  $pos$  трактуется как «классифицирующая», если каждая из  $L$ -грамм, расположенных в позиции  $pos$ , встречается только в текстах одного класса, т.е.  $\Lambda_{L,pos}(\Pi_i) \cap \Lambda_{L,pos}(\Pi_j) = \emptyset$ .

Для вычисления  $L$ -граммных позиционных характеристик классов используются  $L$ -граммные деревья, которые строятся отдельно для каждой рассматриваемой позиции  $pos$ .

### 1.3 Фрагменты с регулярной структурой

Наряду с  $L$ -граммами без явных проявлений структуры можно рассматривать фрагменты текста относительно небольшой длины, характеризующиеся скоплением повторов разного типа. Некоторые комбинации этих повторов образуют устойчивые конфигурации. В данной работе рассматриваются три типа таких конфигураций: периодичности, фракталоподобные структуры и комбинированные структуры. Сначала поясним, что мы понимаем под повторами разных типов.

Пусть  $f: \Sigma \rightarrow \Sigma$  — взаимнооднозначное отображение алфавита  $\Sigma$  на себя («переименование» элементов алфавита). Пусть  $u = u_1 u_2 \dots u_L$  — произвольное слово в алфавите  $\Sigma$  (фрагмент текста  $T$ ), тогда результатом применения подстановки  $f$  к слову  $u$  является слово  $f(u) = f(u_1) \dots f(u_L)$ . Под инверсией слова  $u$  будем понимать слово  $u^R = u_L \dots u_1$ . Результатом последовательного применения преобразований  $f$  и  $R$  к слову  $u$  является  $(f(u))^R = f(u_L) f(u_{L-1}) \dots f(u_1)$ . Пусть  $u$  и  $v$  — произвольные фрагменты текста  $T$  одинаковой длины. Будем говорить, что пара  $(u, v)$  образует *прямой повтор*, если  $u = v$ , *симметричный повтор*, если  $v = u^R$ , *прямой  $f$ -повтор*, если  $v = f(u)$ , *симметричный  $f$ -повтор*, если  $v = (f(u))^R$ . Примером прямых  $f$ -повторов являются секвентные переносы в текстах мелодий, симметричных повторов — инверсии дисков политенных хромосом, симметричных  $f$ -повторов — комплементарные инвертированные повторы в ДНК-последовательностях  $(\Sigma = a, c, g, t)$ , где комплементарное соответствие определяется подстановкой  $f(a) = t$ ,  $f(t) = a$ ,  $f(c) = g$ ,  $f(g) = c$ .

Под **периодичностями** мы понимаем тандемно повторяющиеся цепочки символов, например, *tacg tacg tacg ta* (или, для краткости,  $(tacg)^3 ta$ ). Здесь базовую цепочку *tacg* естественно называть периодом. Длину периода будем обозначать через  $p$ , а длину самой периодичности — через  $P$ . Отношение  $k = P/p$  (равное 3, 5 в данном случае) характеризует кратность повторения.

**Локальными фракталами** мы называем фрагменты последовательностей, которые характеризуются проявлениями самоподобия. Повторение обычного симметричного палиндрома приводит к усилению конструкции, т.е. образованию нового палиндрома вдвое большей длины. Таким же свойством обладают  $f$ -палиндромы:  $u(f(u))^R u(f(u))^R = v(f(v))^R$ , если  $f$  обладает свойством симметрии. В частности, повторение комплементарного палиндрома в ДНК-последовательности приводит к образованию аналогичной структуры вдвое большей длины.

При наличии незначительных искажений внутри повторяющихся фрагментов, равно как и вставок между ними, используем термины «**фракталоподобные структуры**» или «несовершенные локальные фракталы». Предполагается, что размеры вставок сопоставимы с длинами повторяющихся фрагментов [5].

**Комбинированными** мы называем структуры, состоящие из двух разнотипных повторов (традиционный прямой плюс симметричный, прямой плюс симметричный  $f$ -повтор и т.п.) с ограничениями снизу на длины повторяющихся цепочек и сверху — на те же длины и расстояния между соседними элементами структуры [6]. Порядок чередования цепочек, образующих повторы разных типов — произвольный, возможные наложения и совпадения цепочек, относящихся к разным повторам.

Перечисленные структуры часто являются хорошими маркерами разных (в том числе близких) классов последовательностей

## 2 Сравнение последовательностей на примере геномов ВКЭ

Анализируется выборка кодирующих частей геномов вируса клещевого энцефалита (ВКЭ) из базы данных Genbank, содержащая 161 РНК-последовательность, каждая из которых имеет длину 10244 символов. Вся выборка разбита на четыре класса. Первые три отнесены к дальневосточному, европейскому и сибирскому генотипам, соответственно. Четвертый класс представлен «группой 886», циркулирующей на территории Восточной Сибири, и «претендующей» на роль отдельного генотипа. Число штаммов, характеризующих эти классы, различается существенным образом. Первый генотип — 80 штаммов, второй — 46, третий — 28, группа 886 включает всего 7 штаммов.

Среди ВКЭ встречаются высоковирулентные и низковирулентные (инаппарантные) штаммы в соответствии с их способностью вызывать заболевание. Эксперименты по выявлению маркеров, разделяющих ВКЭ по степени вирулентности, проводились на ограниченном материале (38 из 80 штаммов дальневосточного генотипа). По степени близости кодирующих частей эту подборку можно разделить на две группы из 22-х и 16 геномов соответственно. Это разбиение в значительной мере коррелирует с априорными оценками степени вирулентности штаммов, однако имеются исключения. В частности, два штамма выделены от умерших людей, но по нуклеотидной последовательности они ближе к инаппарантным штаммам. В то же время три

штамма выделены от людей с инаппарантной формой протекания заболевания, но на нуклеотидном уровне они ближе к высоковирулентным штаммам. Такого рода исключения могут быть обусловлены особенностями иммунной системы инфицированных людей, но подобная информация отсутствует.

## 2.1 $L$ -граммный анализ

Для выяснения близости представителей одного класса и отличия их от представителей других классов выявим  $\Phi_L(\Pi_i/K)$  — множество «контрастных»  $L$ -грамм, присутствующих в каждом тексте  $i$ -го класса и отсутствующих в текстах других классов.

Наибольший «консервативный» фрагмент, общий для всех 7 последовательностей группы 886 имеет длину 491 символов. Кроме него выявлено более 30 фрагментов с длиной  $L > 100$ , общих для всех последовательностей этой группы, но отсутствующих (в неискаженном виде) в последовательностях других классов. В данном случае более информативными оказываются фрагменты минимальной длины, присутствующие только в этой группе. Такими фрагментами являются 7-граммы *cccgcgtg* (поз. 2763 и 5221) и *atggcgc* (поз. 3859 и 5517).

Классы 1, 2 и 3 более многочисленны и вариативны. Длины максимальных внутриклассовых повторов (отсутствующих в текстах других классов) составляют 17, 35 и 35 символов, соответственно. Для класса 1 минимальная длина контрастного фрагмента составляет 6 символов. Эта 6-грамма (*agtaga*) присутствует во всех 80 текстах первого генотипа 182 раза, с частотой встречаемости в одном тексте от 1 до 4.

Во втором генотипе минимальная длина контрастного фрагмента составляет 7 символов. Выделено восемь 7-грамм (*gttattc*, *gcattcg*, ...). Каждая из них встречается в текстах второго класса не более трех раз. И, наконец, минимальная длина общих  $L$ -грамм для текстов третьего генотипа составляет 8 символов. Таких 8-грамм насчитывается 9, 3 из которых встречаются в каждом тексте по одному разу: *gggtggac* в позиции 5029, *ataagccg* (поз. 6562) и *aggagagt* (поз. 8020), а остальные до трех раз.

Минимальная длина «контрастных»  $L$ -грамм, разделяющих штаммы дальневосточного генотипа на низко- и высоковирулентные равна 6. Для низковирулентных геномов выделено 10 таких 6-грамм (*gtccga*, *cggatt* ...), а для высоковирулентных — 16 (*gctaga*, *acttca*, ...).  $L$ -грамма максимальной длины ( $L = 88$ ) обнаружена среди низковирулентных геномов в позиции 8645.

## 2.2 $L$ -граммно-позиционный анализ

$L$ -граммно-позиционный анализ проиллюстрируем на примере 9-граммных «срезов» последовательностей, соответствующих трем триплетам, кодирующим аминокислоты. Для каждой позиции  $pos = 1, 4, 7, \dots, 10240$  формируем совокупность из 161-й нуклеотидной 9-граммы (или аминокислотной 3-граммы). Выделяется 8 «устойчивых» позиций, содержащих 9-граммы, общие для всех последовательностях независимо от генотипа или разбиения на какие-либо другие классы. Пара таких позиций расположена по соседству, т.е. формируется «устойчивая» 12-грамма.

Позиция	Фрагмент РНК	Аминокислотная тройка
55	tcg-aaa-gag	S-K-E
73	aag-acg-cgt	K-T-R
115	ttg-atg-cgc	L-M-R
118	atg-cgc-atg	M-R-M
2269	atg-tcc-atg	M-S-M
2929	ctc-tgg-atg	L-W-M
4621	acg-atg-tgg	T-M-W
9382	aac-ata-aag	N-I-K

Заметим, что в позиции 4618, предшествующей 4621, стоит аминокислота Н (гистидин), кодируемая триплетом *sac* во всех последовательностях основных генотипов, и триплетом *cat* в последовательностях группы 886 («тонкое» различие). Разделение по генотипам (1,2,3,4) наиболее ярко демонстрируют 9-граммы в позициях 3568 и 3571.

3568: (VAV)	1: gtg-gca-gtt	3571: (AVG)	1: gca-gtt-ggg
	2: gtg-gca-gtg		2: gca-gtg-ggg
	3: gtg-gcg-gtg, gtt-gcg-gtg		3: gcg-gtg-ggg
	4: gta-gca-gtc		4: gca-gtc-ggg

Соответствующие этим 9-граммам тройки аминокислот совпадают для всех 4 классов (VAV для 9-грамм в позиции 3568 и AVG для 9-грамм в поз. 3571), однако на уровне РНК наблюдается четкое расслоение на 4 класса (генотипа). Каждый характеризуется своей РНК-цепочкой длины 9 за единственным исключением: генотип 3 в поз. 3568 представлен двумя такими цепочками, ни одна из которых не встречается в последовательностях других генотипов. Такого рода эффекты мы также склонны относить к категории «тонких различий». Другим примером  $L$ -граммно-позиционного разбиения последовательностей на 4 класса является совокупность 9-грамм в позиции 2350:

9-грамма	Аминокислотная тройка	класс	Кол-во текстов
gac-acc-gaa	DTE	1	44 из 80
gac-act-gaa	DTE		34 из 80
gat-act-gaa	DTE		1
gac-ccc-gaa	DPE		1
gac-acg-gaa	DTE	2	45 из 46
gac-aca-gaa	DTE		1
gac-acc-gag	DTE	3	25 из 28
gac-act-gag	DTE		3
gac-aca-gag	DTE	4	7 из 7

Здесь каждому классу соответствует несколько 9-грамм, ни одна из которых не встречается в последовательностях других классов. 9-граммы с малой частотой встречаемости можно отнести к разряду малоинформативных. Кроме указанной позиции, существуют и другие, однозначно разделяющие тексты на 4 класса. Еще больше маркеров можно использовать для деления 38 последовательностей первого генотипа на высоко- и низковирулентные: выделяется около 30 позиций, однозначно разбивающих подборку на два указанных класса. Практически во всех позициях аминокислотные последовательности совпадают, т.е. всё различие проявляется лишь на нуклеотидном уровне.

### 2.3 Фрагменты со специфическими структурными особенностями

**Периодичности.** В каждом геноме выявлялись все периодичности с длиной  $P \geq r$ , где  $r$  — пороговое значение (в нашем случае  $r = 10$ ). Их можно разделить на несколько групп. В первую включены те, которые присутствуют во всех генотипах. Они могут быть использованы для сопоставления ВКЭ с родственными организмами. Ко второй группе отнесены редко встречающиеся (а потому и малоинформативные) периодичности, существующие во всей подборке в одном-двух экземплярах. Оставшиеся периодичности разделим на две группы: А и Б. В группу А включим периодичности, представленные только (или преимущественно) в одном из генотипов. Они представляют наибольший интерес в плане генотипирования. В группу Б включим периодичности, представленные в разных (но не во всех четырех) генотипах. Факт отсутствия их в том или ином генотипе сужает число претендентов при классификации. Кроме того, периодичности из этой группы дают представление о связях между разными генотипами.

К группе А отнесены свыше 30 периодичностей. Для иллюстрации ограничимся лишь одним представителем каждого класса. Периодичность  $(aagag)^2$  выявлена в позиции 3162 у 79 из 80 текстов первого генотипа. В позиции 8144 каждого из 46 текстов второго генотипа находится периодичность  $g^{10}$ . Она присутствует и во всех текстах группы 886, но в позиции 5241. У 25 из 28 текстов третьего генотипа в позиции 7720 обнаружена периодичность  $(ctggc)^2$ , в трех оставшихся она представлена в слегка варьированном виде:  $(tggt)^2$  (позиция 7721). Из 20 наиболее ярких представителей группы Б отметим периодичность  $(gaaag)^2g$  ( $p = 5$ , поз. 21), которая выявлена в 70-ти из 80 текстов первого генотипа, в 27 из 28 — третьего и в каждом тексте группы 886, но не представлена в последовательностях генотипа 2. Другая периодичность  $(aagga)^2$  ( $p = 5$ ) выявлена в позиции 398 у 36 из 46 текстов 2-го генотипа и 22 из 28 — третьего. У оставшихся 10 из 46 текстов второго генотипа фрагмент  $(gaagg)^2$  находится в 397 позиции.

Для классификации по степени вирулентности, также можно указать маркеры — периодичности для каждого класса. Периодичность  $(ggagg)^2$  типична для высоковирулентных представителей ВКЭ первого генотипа, как и  $(ccaaatg)^2$ , но последняя не обнаружена у двух штаммов, отнесенных к исключениям: выделены от инфицированных пациентов с отсутствием симптомов заболевания. В инаппарантных штаммах выявлены периодичности  $(agga)^3$  и  $(aactc)^2a$ . Периодичности разных классов могут отличаться длиной. Так, выявленная в инаппарантных штаммах периодичность  $(accgc)^2acc$ , в вирулентных присутствует в «сокращенном» варианте:  $(accgc)^2$ .

**Фрактальные структуры** — частный случай периодичностей. Их относительно немного, однако каждая из выявленных структур соответствует только (или преимущественно) одному из трех основных генотипов. В геномах группы 886 точные фрактальные структуры отсутствуют. Только в штаммах генотипа 1 в позиции 805 обнаружена фрактальная структура  $(gtg)^4$ , которой соответствует аминокислотная цепочка  $V^4$  (встретилась 10 раз, преимущественно в высоковирулентных штаммах). Аналогично, в позиции 5870 того же генотипа представлена фрактальная структура  $(agga)^4$  (встретилась 25 раз, из них в 22 инаппарантных штаммах). Эти структуры являются слабыми маркерами при генотипировании, поскольку в сумме они встретились меньше чем в половине геномов. Но и при разделении штаммов по степени вирулентности они могут допустить ошибку, поскольку в штамме  $> JQ825147.1shkotovo94$  присутствуют обе структуры.

Только в штаммах генотипа 2 в позиции 9411 выявлена фрактальная структура  $(ggaagg)^2gg$ , которой соответствует, начиная со второй позиции, аминокислотная цепочка  $EGEG$ . Однако один раз эта структура появляется и в генотипе 3. И, наконец, только в генотипе 3 в позиции 6055 встречается фрактальная структура  $ga(agaga)^2a$  (в 20 из 28 штаммов).

Из **фракталоподобных структур** рассматривались лишь такие, в которых отсутствовали искажения в повторяющихся ядрах, но между ними могли быть вставки ограниченной длины. Примером такой структуры, присутствующей в большинстве штаммов (146 из 161) независимо от генотипа, является  $gagctcaaaactggagagctc$ , (ядро — комплементарный палиндром длины 6, подчеркнут). В отличие от предыдущей структуры, демонстрирующей сходство разных генотипов, цепочка  $gttggttgctggttggtg$ , с ядром  $gttg$  присутствует только в генотипе 1 (70 штаммов из 80). Поэтому штаммы, содержащие такую структуру, можно достаточно уверенно относить к первому генотипу. На том же основании штаммы, содержащие в позиции 8471 фракталоподобную структуру  $cggcaccaaccggctcggcgggc$  с ядром  $cggc$ , можно без особого риска относить ко второму генотипу (41 случай из 46 на обучении). Фракталоподобных структур, пригодных для идентификации штаммов третьего генотипа, не обнаружено.

Напомним, что **комбинированными** мы называем структуры, состоящие из двух позиционно сближенных разнотипных повторов. В приводимых ниже примерах выбраны следующие значения параметров: минимальная длина повторов  $r = 7$ , максимальная  $R = 20$ , максимально допустимое расстояние между соседними компонентами структуры  $d = 14$ . Пары фрагментов, образующих симметричный повтор (обычный или комплементарный), будем обозначать  $Xs$  и  $X$ . Пары, образующие обычные прямые повторы, будем обозначать через  $Y$ . Допускаются наложения компонентов структуры друг на друга.

У 66 из 80 штаммов первого генотипа выделена структура:

Позиции :  $Xs$  - 1916;  $X$  - 1927;  $Y$  - 1944;  $Y$  - 1947  
 $gataacaccccaaccccaacaatcgaaaccaatggtggtggttt$   
 $Xs = taacaccc \quad cccacaat = X$   
 $Y = \quad \quad \quad tggtggt$   
 $Y = \quad \quad \quad tggtggt$   
 638: L I T P N P T I E T N G G G

Здесь разнесенный симметричный повтор сопровождается периодичностью  $(tgg)^3$ , которая формально образует повтор длины 7 с наложением. Еще две структуры (поз. 4341 и 6970) встречаются более чем у половины последовательностей первого генотипа. Первая составлена из симметричного и прямого повтора, вторая — из комплементарного палиндрома, комплементарного инвертированного повтора и прямого повтора.

В то же время выделена структура, присутствующую только у инаппарантных штаммов:

$gagtgacacttgacgccacagtgcggaagagagagacggcaccactgtg$   
 $Xs = gtgacac \quad cacagtg = X$   
 $Y = \quad \quad \quad agagaga$   
 $Y = \quad \quad \quad agagaga$   
 $X = cacagtg \quad Xc = cactgtg$   
 G V T L A A T V R K E R D G T T V

В 18 последовательностях из 28 третьего генотипа в позиции 752 две комбинированные структуры налагаются друг на друга, в результате чего возникает структура, состоящая из двух прямых повторов и одного симметричного. В каждой из 7 последовательностей группы 886 присутствуют по две структуры (позиции 851 и 8393) типа «прямой повтор плюс симметричный» и одна структура (поз. 6954) типа «прямой повтор плюс комплементарный инвертированный». Комбинированных структур, позволяющих идентифицировать генотип 2, не обнаружено.

## Заключение

Традиционно используемые меры сходства символьных последовательностей часто оказываются малоинформативными при сравнении очень близких текстов, относящихся, тем не менее, к разным классам. Исследование сходства и различия такого рода «параллельных» текстов обычно проводится путем их предварительного выравнивания, что само по себе представляет самостоятельную задачу. Возможен и другой вариант сравнения близких текстов, не требующий выравнивания. Структура, позволяющая осуществить такое сравнение, названа авторами совместным  $L$ -граммным спектром двух (и большего числа) последовательностей.

В работе показано, что важную информацию о параллельных текстах можно извлечь путем анализа их  $L$ -граммного состава не только при малых, но и при больших значениях  $L$ . Наряду с  $L$ -граммами, общими для сравниваемых текстов (или групп текстов) можно выявить и «контрастные»  $L$ -граммы, встретившиеся только в одном тексте (или группе текстов). Особый интерес представляют фрагменты последовательностей, обладающие регулярной структурой (тандемные повторы, симметрии, фракталоподобные структуры и др.). Они представляют комбинации повторов разного типа и легко интерпретируемы. Многие из них являются хорошими маркерами различных классов последовательностей, причем «тонкие различия» между классами могут проявить себя на уровне числа повторений в тандемной структуре или наличия характерного искажения в повторяющейся единице.

Подход иллюстрируется на двух задачах классификации очень близких геномов вируса клещевого энцефалита. Первая связана с разбиением подборки геномов на разные генотипы, вторая — с определением степени их вирулентности. Получены сильные РНК-маркеры, позволяющие решать оба типа задач.

## Список литературы

- [1] Колмогоров А.Н. Три подхода к понятию «количество информации» // Проблемы передачи информации. — Т.1, вып. 1, 1965. — С. 3–11.
- [2] Gusev V. D., Miroshnichenko L. A Complexity decompositions in problems of comparison of symbolic sequences // Proc. 11th Int. Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013) September 23–28, 2013, Samara, Vol. 1., 94–97.
- [3] Гусев В.Д. Характеристики символьных последовательностей // Машинные методы обнаружения закономерностей. — Новосибирск, ИМ СО РАН, 1981. — Вып. 88: Вычислительные системы. — С. 112–123.
- [4] Гусев В.Д., Титкова Т.Н. Хеширование символьных цепочек в режиме скользящего окна. // Анализ последовательностей и таблиц данных. Новосибирск, 1994. — Вып. 150: Вычислительные системы. С. 94–106.
- [5] Гусев В.Д., Мирошников Л.А., Чужанова Н.А. Выявление фракталоподобных структур в ДНК-последовательностях // Information Science and Computing. International Book Series, № 8: Classification, Forecasting, Data Mining. — ITNEA, Sofia, 2009.— P. 117–123.
- [6] Гусев В.Д., Мирошников Л.А. Поиск комбинированных структур в ДНК-последовательностях // Доклады всероссийской конференции "Математические методы распознавания образов"(ММРО-13), Ленинградская обл., г. Зеленогорск, 30 сентября–6 октября 2007г., М., Макс-Пресс, 473–476.

*Владимир Дмитриевич Гусев — к.т.н., ст. науч.сотр. Института математики им. С.Л.Соболева СО РАН;  
e-mail: gusev@math.nsc.ru;*

*Любовь Александровна Мирошников — к.т.н., ст. науч.сотр. Института математики им. С.Л.Соболева СО РАН;  
e-mail: luba@math.nsc.ru;*

*Татьяна Николаевна Титкова — к.т.н., науч.сотр. Института математики им. С.Л.Соболева СО РАН;  
e-mail: titkova@math.nsc.ru.*

*Дата поступления — 31 мая 2017 г.*