АНАЛИЗ ВРЕМЕННЫХ РЯДОВ И СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ
(Вычислительные системы)
1991 год Выпуск 141

УДК 519.764

ПОИСК И КЛАССИФИКАЦИЯ ФРАГМЕНТОВ ТЕКСТА С ОДИНАКОВОЙ СЛОЖНОСТНОЙ СТРУКТУРОЙ

В.Д.Гусев, О.М.Чупахина

В в е д е н и е

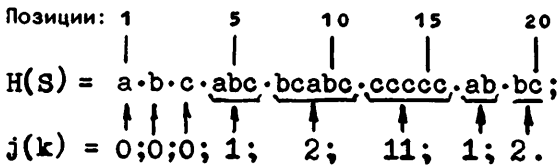
В [1] предложена мера сложности конечной последовательности, которая ассоциируется с числом шагов некоторого гипотетического процесса, порождающего данную (известную заранее) последовательность. Множество порождающих операций определяется спецификой прикладной области (соответствующие примеры приведены в [2]). В простейшем случае оно состоит из операций генерации нового символа и копирования любого фрагмента из предыстории. Если по ходу процесса встречается символ, которого не было ранее, используется первая операция. Если же очередная (еще не порожденная) цепочка символов уже встречалась, ее можно скопировать. При этом для копирования выбирается тот из возможных прототипов, который позволяет максимальным образом удлинить последовательность.

Так, например, последовательность

$$S = abcabcboabccccccabbc$$

может быть синтезирована за 8 шагов. Шагу с номером k, $1 \le k \le 8$, соответствует k-й компонент сложности. Это фрагмент из S, получаемый в результате использования одной из порождающих операций. Разбиение S на компоненты сложности

имеет вид:



Здесь компоненты сложности выделены точками; $C(S)$ - значение сложности; $j(k)$ - указатель копирования: значение $j(k) = 0$, если используется операция генерации нового символа; в противном случае значение $j(k)$ интерпретируется как адрес (позиция первого символа) того фрагмента из предыстории, который послужил прототипом для k-го компонента сложности. Более детальное обсуждение описанной меры сложности содержится в [2].

Будем говорить, что два фрагмента имеют одинаковую сложностную структуру, если: а) их длины совпадают б) вектора длин компонентов сложности (а, следовательно, и сами значения сложности) одинаковы для обоих фрагментов; в) вектора значений указателей копирования совпадают. К примеру, последовательности

$$S_1 = a\cdot b\cdot bb\cdot abb\cdot c\cdot abbc \quad \text{и} \quad S_2 = c\cdot a\cdot aa\cdot caa\cdot b\cdot caab$$

имеют одинаковую сложностную структуру. В то же время последовательность $S_3 = a\cdot b\cdot bb\cdot abb\cdot a\cdot abba$ имеет сложностную структуру, отличную от структур S_1 и S_2, поскольку значение указателя копирования для пятого компонента сложности равно нулю у S_1 и S_2, а у S_3 - единице.

В работе рассматривается задача отыскания в достаточно длинных текстах всевозможных пар фрагментов фиксированной длины с одинаковой сложностной структурой. Иными словами, речь

идет об отыскании повторов на структурном уровне. Разновидностями таких повторов являются повторы в обычном смысле (точно совпадающие фрагменты), симметрии, палиндромно-шпилечные структуры в генетических текстах, секвентные переносы в музыкальных произведениях и т.п.

Тесная связь прослеживается между структурными повторами и понятием "интерпретации" слова, существующим в теории формальных языков [3]. Пусть α - слово над алфавитом Σ. Слово α' той же длины, что и α, называется интерпретацией слова α при выполнении следующего условия: если в слове α i-я буква отличается от j-й, то и в слове α' i-я буква отличается от j-й (предполагается, что буквы в обоих словах пронумерованы одинаково). В простейшем случае интерпретация слова α может быть получена переименованием букв алфавита. Слова α и α' будут иметь тогда одинаковую сложностную структуру. Таким образом, о структурных повторах можно говорить как о повторах фрагментов с точностью до переименования входящих в них элементов.

Повторы играют фундаментальную роль во всех языковых системах. Выявление их на разных иерархических уровнях способствует пониманию принципов организации языковой системы и создает предпосылки для решения многочисленных эволюционных и классификационных задач. Достоинством предлагаемого подхода является его универсальность, проявляющаяся в том, что заранее не фиксируется способ переименования элементов*) (взаимно однозначное отображение $f: \Sigma \to \Sigma$), тем не менее для всевозмож-

*) В каждом таком конкретном случае задача может быть сведена к отысканию повторов в обычном смысле с использованием известной техники (хеширование, префиксные и суффиксные деревья, ациклические графы слов). Однако, если потенциально возможных способов переименования много, то перебор по всем им будет неэффективным.

ных f гарантируется эффективное выявление всех пар фрагментов, образующих повторы в указанном смысле.

Заметим, что направления просмотра последовательности (слева направо и справа налево) в рамках предлагаемого подхода считаются равноправными,что позволяет фиксировать пары фрагментов с одинаковой сложностной структурой, но разными направлениями просмотра. В частности, по такой схеме происходит выявление симметричных конструкций.

Ниже на идейном уровне описаны основные шаги алгоритма, реализующего поиск в тексте всех фрагментов заданной длины,образующих повторы с точностью до переименования (f) входящих в них элементов. Будем называть их для краткости f-повторами (или структурными повторами).

1. Схема алгоритма отыскания f-повторов

Пусть T - анализируемый текст, N - длина текста, $T[i]$ - i-й элемент текста, $T[i:j]$ - фрагмент текста, включающий символы с i-го по j-й, $1 \le i < j \le N$, Σ - алфавит, $|\Sigma|$ - мощность алфавита, D - длина каждого из фрагментов, образующих f-повтор . В принципе любая пара фрагментов из T может образовать f-повтор,поэтому при $D \ll N$ потенциально возможное количество пар, которое требуется просмотреть , имеет порядок N^2. Идея алгоритма состоит в том, чтобы все множество фрагментов из T последовательно дробить на все более мелкие непересекающиеся подмножества таким образом, чтобы элементы каждого подмножества могли образовывать потенциально возможные f-повторы лишь друг с другом, но не с элементами других подмножеств. Это позволяет избежать лишних сравнений и существенно уменьшает порядок трудоемкости.

Разбиение фрагментов из T на подмножества ведется вначале по значениям векторов длин компонентов сложности. Образовавшиеся подмножества затем,в свою очередь,делятся в соответствии со значениями векторов указателей копирования.

Схема алгоритма выглядит следующим образом.

Этап 1. Окно анализа размера D сдвигаем вдоль текста слева направо с шагом в один символ. Положение окна характеризуем номером позиции начального символа. На i-м шаге, $1 \le i \le N-D+1$, выделяем фрагмент текста $T[i:i+D-1]$, осуществляем разбиение его на компоненты сложности и формируем вектор длин компонентов сложности $L_i = (l_1^{(i)}, l_2^{(i)}, \ldots, l_{c(i)}^{(i)})$ и вектор указателей копирования $J_i = (j_1^{(i)}, j_2^{(i)}, \ldots, j_{c(i)}^{(i)})$, где $c(i)$ - сложность i-го фрагмента (число компонентов в разбиении).

Инвертируем исходный текст и получаем текст $T^R = T[N]\,T[N-1]\ldots T[1]$. Осуществляем с ним те же операции, что и с текстом T, и формируем множество векторов $L^R = \{L_i^R\}$ и $J^R = \{J_i^R\}$, $i = 1 \div N-D+1$.

Для реализации данного этапа используется алгоритм вычисления сложностного профиля последовательности [4]. В его основу положена конструкция, называемая деревом префикс-идентификаторов. Отличительной особенностью алгоритма является то, что при сдвиге окна анализа на один символ, т.е. при переходе от $(i-1)$-го фрагмента к i-му вектора L_i и J_i не вычисляются заново, а получаются в результате корректировки уже известных векторов L_{i-1} и J_{i-1}. При этом используется "зацепленность" $(i-1)$-го и i-го фрагментов, следствием которой является сохранение (как правило) большинства компонентов в обоих разбиениях.

Трудоемкость данного этапа составляет $O(N \cdot \log^2 D / |\Sigma|)$ [4], затраты памяти - порядка $N \cdot D / \log D$, где логарифмы берутся по основанию $|\Sigma|$. При размере алфавита, сопоставимом с длиной окна (типичный случай), затраты памяти по порядку близки к ND.

Этап 2. Среди множества векторов L_i и L_i^R, $1 \le i \le N-D+1$, выявляем одинаковые и объединяем их в отдельные кластеры. Для этого в принципе может быть использован алгоритм лексикографической сортировки векторов L_i и L_i^R с трудоемкостью, не превышающей по порядку величины $N \cdot D$ [5]. Затраты по памяти имеют тот же порядок. Для уменьшения последних сортировка может быть заменена упаковкой векторов в дерево аналогично тому, как это делается в алгоритме Ахо-Корасика [6]. Именно этот вариант использован в описываемом алгоритме, причем этап 2 совмещается с этапом 1, т.е. вектора упаковываются в дерево по мере их вычисления. Экономия памяти достигается за счет "склеивания" одинаковых начал у разных векторов.

ПРИМЕР 1. Пусть размер окна анализа $D = 15$. Рассмотрим подмножество множества $L = \{L_i, i = 1 \div N\}$, состоящее из 8 векторов:

$$L_{i_1} = (1,1,2,1,3,1,4,2);$$

$$L_{i_2} = (1,3,1,2,1,3,1,2,1);$$

$$L_{i_3} = (1,1,2,1,5,1,2,2);$$

$$L_{i_4} = (1,2,1,3,1,5,2);$$

$$L_{i_5} = L_{i_6} = L_{i_3};$$

$$L_{i_7} = L_{i_4};$$

$$L_{i_8} = (1,2,1,1,4,6).$$

Здесь i_k, $k = 1\text{-}8$, - номера позиций, определяющих положение окна в каждом из 8 случаев. Соответствующее дерево изображено на рис.1.

Путь от корня дерева к листу, помеченному индексом i_k, указывает последовательность длин компонентов сложности во

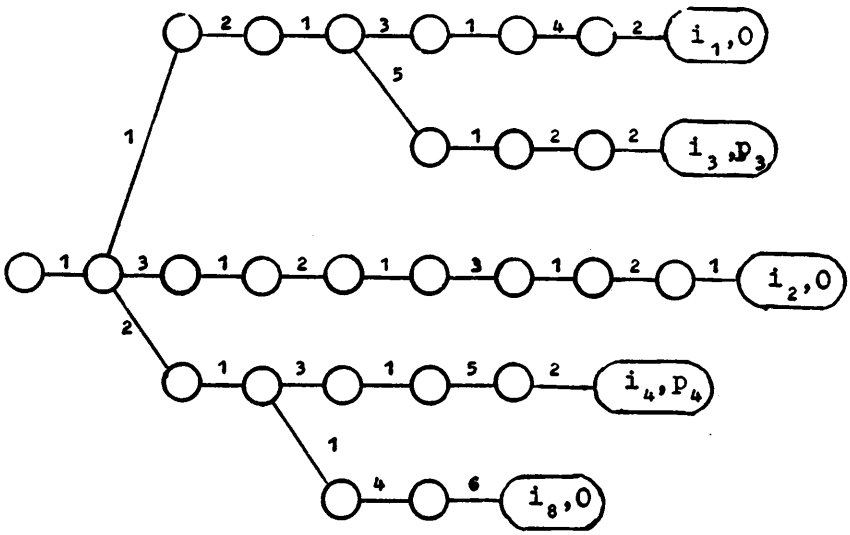


Рис. 1

фрагменте текста $T[i_k : i_k+D-1]$. Если эта последовательность уникальна, т.е. отлична от тех, которыми представлены другие фрагменты, то значение параметра p_k , связываемого с листом i_k , равно нулю. В противном случае $p_k \neq 0$ и интерпретируется как указатель отсылки на продолжение списка фрагментов, обладающих той же самой последовательностью длин компонентов. В рассматриваемом примере, в частности,будут продолжены списки с заголовками i_3 (добавятся элементы i_5 и i_6) и i_4 (добавится элемент i_7). Заметим также, что в соответствии с определением меры сложности первый компонент во всех векторах L_i равен единице, поэтому упаковку в дерево можно начинать сразу со вторых компонентов.

Этап 3. Анализируем списки, сформированные на предыдущем этапе и состоящие более чем из одного элемента. Фрагменты текста, входящие в каждый из списков, характеризуются одинаковыми векторами L_i или L_i^R . Осуществляем дальнейшую их дифферен-

циацию, но уже по значениям векторов указателей копирования J_i и J_i^R.

С этой целью для каждого таксона (списка) строим дерево векторов J_i и J_i^R по рассмотренной выше схеме. Листья таких деревьев, содержащие отсылки на списки из двух и более фрагментов, будут давать решение поставленной задачи. Если, к примеру, список содержит p фрагментов, то все они обладают одинаковой сложностной структурой и служат основой для формирования C_p^2 структурных повторов в указанном выше смысле.

Трудоемкость и затраты памяти на данном этапе мажорируются аналогичными характеристиками для этапа 2 ($O(ND)$ в обоих случаях). Суммарная трудоемкость алгоритма составляет $O(N(\log^2 D/|\Sigma| + D))$, затраты памяти - $O(ND)$.

Корректность алгоритма вытекает их того, что: 1) рассматриваются все фрагменты текста длины D (окно анализа сдвигается каждый раз на один символ); 2) множество фрагментов делится на непересекающиеся подмножества, такие, что элементы разных подмножеств имеют разные последовательности длин компонентов сложности и, следовательно, не могут образовывать структурные повторы, тогда как элементы одного подмножества - могут,но не все; 3) для отсеивания лишних фрагментов неединичные (по числу элементов) подмножества, возникшие на этапе 2, дробятся на еще более мелкие подмножества, каждое из которых содержит те и только те фрагменты текста, которые характеризуются как одинаковой последовательностью длин компонентов, так и одинаковыми значениями указателей копирования. Отсюда следует, что элементы неединичных подмножеств, возникших на этапе 3, и только они удовлетворяют определению структурного повтора.

2. Классификация структурных повторов

Можно выделить несколько классификационных признаков для разбиения множества структурных повторов на отдельные подклассы.

2.1. Направление считывания. Если оба фрагмента, образующих f-повтор, считываются в одном и том же направлении (как правило, слева направо), будем называть такой f-повтор прямым, в противном случае - инвертированным. Первый тип повторов в графической форме удобно представлять в виде пары одинаково ориентированных отрезков (→ →), второй - в виде пары противоположно ориентированных отрезков (→ ← или ← →).

2.2. Локализация. Этот признак характеризует удаленность фрагментов, составляющих f-повтор, друг от друга. В первом приближении рассмотрим всего две градации признака. Если мы имеем дело с прямым f-повтором и составляющие его фрагменты расположены в тексте непосредственно друг за другом, назовем такой повтор f-периодичностью (графическое обозначение: →→); если же фрагменты отделены друг от друга произвольным числом символов, назовем такую структуру разнесенным f-повтором (→ →).

Иногда имеет место частичное наложение фрагментов, образующих f-повтор ($\overrightarrow{\rightarrow}$) . Мы не выделяем этот случай в особую градацию, поскольку фактически он сводится к периодичности (или серии периодичностей) с длиной периода равной величине сдвига между фрагментами. Полное наложение фрагментов возможно лишь при тождественном отображении f, но при этом теряет смысл само понятие повтора.

В случае инвертированных f-повторов может иметь место (и быть трактуемым) полное наложение фрагментов ($\overleftrightarrow{\rightarrow}$), частичное ($\overleftarrow{\rightarrow}$) и непосредственное их следование друг за другом (→←). Повторы, соответствующие всем этим трем случаям, назовем совмещенными. Побудительным мотивом для объединения их в один подкласс служит то обстоятельство, что во всех случаях имеет место "обобщенная симметрия", понимаемая следующим обра-

разом: если $x_i x_{i+1} \dots x_{i+n-1}$ - минимальный сегмент текста, содержащий оба фрагмента (длины D), образующие инвертированный f-повтор, то существует ось симметрии, проходящая посередине сегмента (между символами, если n - четно, и по символу, если n - нечетно), такая, что $x_{i+k} = f(x_{i+n-1-k})$, $k = 0,1,\dots,[n/2]$, $D \le n \le 2D$.

Если в предыдущем определении $n > 2D$, а $k = 0,1,\dots, D-1$, получаем класс f-повторов, которые характеризуем как инвертированные разнесенные (⇒ ⇐).

2.3. Тип отображения f. Взаимно однозначное отображение $f: \Sigma \to \Sigma$ можно интерпретировать как подстановку, которая определяется с помощью таблицы из двух строк. Каждая строка содержит все элементы алфавита Σ, причем элементу x в верхней строке соответствует элемент $f(x)$ в нижней. Переход от произвольного фрагмента $X = x_1 x_2 \dots x_n$ текста T к его f-аналогу осуществляется посимвольно, т.е. $f(X) = f(x_1) f(x_2) \dots f(x_n)$.

Каждую подстановку можно представить в виде суперпозиции циклов. Легко заметить, что если взять произвольный элемент x_0 из верхней строки подстановки и рассмотреть цепочку элементов $x_1 = f(x_0)$, $x_2 = f(x_1)$ из этой же строки, то после конечного числа шагов мы вновь вернемся к элементу x_0, т.е. найдется такое целое $l \ge 1$, что $x_l = f(x_{l-1}) = x_0$. Элементы $x_0, x_1, \dots, x_{l-1}$ образуют цикл длины l. Продолжая аналогичным образом процесс выделения циклов вплоть до исчерпания всех элементов алфавита, получим разложение подстановки на циклы. К примеру, подстановка

$$f = \begin{pmatrix} a & b & c & d & e & f & g \\ g & d & a & b & e & c & f \end{pmatrix}$$

может быть представлена в виде суперпозиции следующих циклов:

$$f = \begin{bmatrix} a & g & f & c \\ g & f & c & a \end{bmatrix} \begin{bmatrix} b & d \\ d & b \end{bmatrix} \begin{bmatrix} \bullet \\ e \end{bmatrix} .$$

Тип подстановки определяется набором ее циклов $(r_1, \ldots, r_m)$, где $m = |\Sigma|$, а r_i, $1 \le i \le m$, - число циклов длины i (очевидно, что $\Sigma i r_i = m$). Указывают, как правило, лишь ненулевые элементы этого набора в виде i^{r_i}. Так, к примеру, рассматривавшаяся выше подстановка из 7 элементов имеет тип $1^1 2^1 4^1$.

В нашей классификации выделим лишь некоторые типы подстановок, имеющих достаточно наглядную интерпретацию и часто встречающихся на практике.

2.3.1. Тождественные подстановки представлены исключительно циклами длины 1 $(x_i = f(x_i), 1 \le i \le m)$. Они выделяют из класса f-повторов подкласс повторов в обычном смысле. Назовем их совершенными, если это прямые повторы, палиндромами, если это инвертированные совмещенные повторы, и симметричными, если это инвертированные разнесенные повторы. Смысл первого термина пояснен ниже, он используется в молекулярно-биологических приложениях; два других термина являются более или менее устоявшимися.

2.3.2. Подстановки с превалирующим числом циклов длины 1 $(m-1 > r_1 \ge m/2)$. Если в переименованном фрагменте содержится достаточно большое количество неизменяемых в процессе подстановки элементов (именно они характеризуются циклами длины 1), повтор отличается от совершенного лишь наличием замен по некоторым позициям. Естественно назвать такие повторы несовершенными и распространить этот термин на палиндромы и симметрии.

С несовершенными повторами часто приходится сталкиваться в теории кодирования, при анализе естественных языков (ошибки

типа замен символов в словах), в молекулярно-биологических приложениях (точечные мутации). Алгоритмы отыскания несовершенных повторов достаточно трудоемки. Возможность отыскания пусть даже части их более простыми методами представляет практический интерес.

2.3.3. Подстановки с циклами длины 2 часто встречаются в различных приложениях. Из множества совмещенных инвертированных f-повторов, в частности, можно выделить подкласс f-палиндромов, которые характеризуются следующими свойствами: а) D - четно; б) фрагменты, образующие повтор, полностью налагаются друг на друга ($\rightleftarrows$); в) отсутствуют тождественно отображаемые элементы (этот случай рассмотрен в п. 2.3.2). Если $x_1x_2\ldots x_D$ - f-повтор с указанными свойствами, имеет место $x_k = f(x_{D-k})$ и $x_{D-k} = f(x_k)$, $k = 0,1,\ldots,D/2$, т.е. элементы x_k и x_{D-k} при любом k образуют цикл длины 2, поскольку связаны нетождественным отображением.

Примером f-палиндромов являются комплементарные палиндромы в генетических текстах, где $f = \begin{pmatrix} A & T & G & C \\ T & A & C & G \end{pmatrix}$ - отношение комплементарности. Комплементарные палиндромы с $D = 4,6$ (CCGG, ATGCAT и т.п.) часто играют роль рестрикционных сайтов - участков разрезания НК-молекул особыми ферментами (рестриктазами). Если фрагменты, образующие повтор подобного типа, разнесены на 5-10 элементов, возникает другая значимая конструкция, называемая шпилечной. Шпильки играют важную роль в формировании вторичной структуры РНК-молекул и в регуляции основных генетических процессов.

2.4. Структура фрагментов, образующих повтор. Очень часто фрагменты, образующие f-повтор, оказываются структурированными, т.е., в свою очередь, состоят из каких-то более мелких

структурных элементов подобного же вида (повторов, серий, симметрий и т.п.). Для описания структуры f-повторов иногда удобно использовать "язык образов" [7]. Образ - это любая конечная цепочка в алфавите из константных (Σ) и переменных (V) символов. Язык образа p есть множество слов $L(p)$, получаемых подстановкой вместо переменных символов цепочек из константных символов (возможно, с некоторыми ограничениями). Если, к примеру, $\Sigma = \{A,T,G,C\}$, $V = \{V_1,V_2,V_3,\ldots\}$, $p = AV_1V_2V_2V_1T$ и имеется ограничение на длину цепочек из константных символов, подставляемых вместо V_i $(|V_i| = 3)$, то применительно к генетическим текстам образ p будет фиксировать наличие симметрии в f-повторе на аминокислотном уровне. Примеры соответствующих закономерностей будут приведены ниже (см. п.3).

Элементы описанной в данном разделе схемы классификации удобно представить в виде дерева, изображенного на рис.2. Нижний ярус этого дерева, отражающий структурное разнообразие f-фрагментов, опущен, поскольку типы структур сильно варьируют для разных предметных областей.

3. Выявление f-повторов в геноме бактериофага λ

Геном бактериофага λ был выбран в качестве объекта для эксперимента, поскольку это один из наиболее длинных (~ 50000 символов) и интересных в регуляторном отношении геномов из числа секвенированных к настоящему времени. Ранее авторы провели анализ сложностных профилей генома фага λ [8], и было бы интересно дополнить результаты этого анализа информацией о наличии и распределении f-повторов в данном геноме.

Размер окна анализа D был выбран равным 15. Фактически были выявлены все f-повторы с $D \geq 15$, поскольку проверялась возможность расширения каждого f-повтора длины 15. Значение $D = 15$ соответствует длине максимального совершенного повтора (или повтора в обычном смысле) в геноме фага λ. По

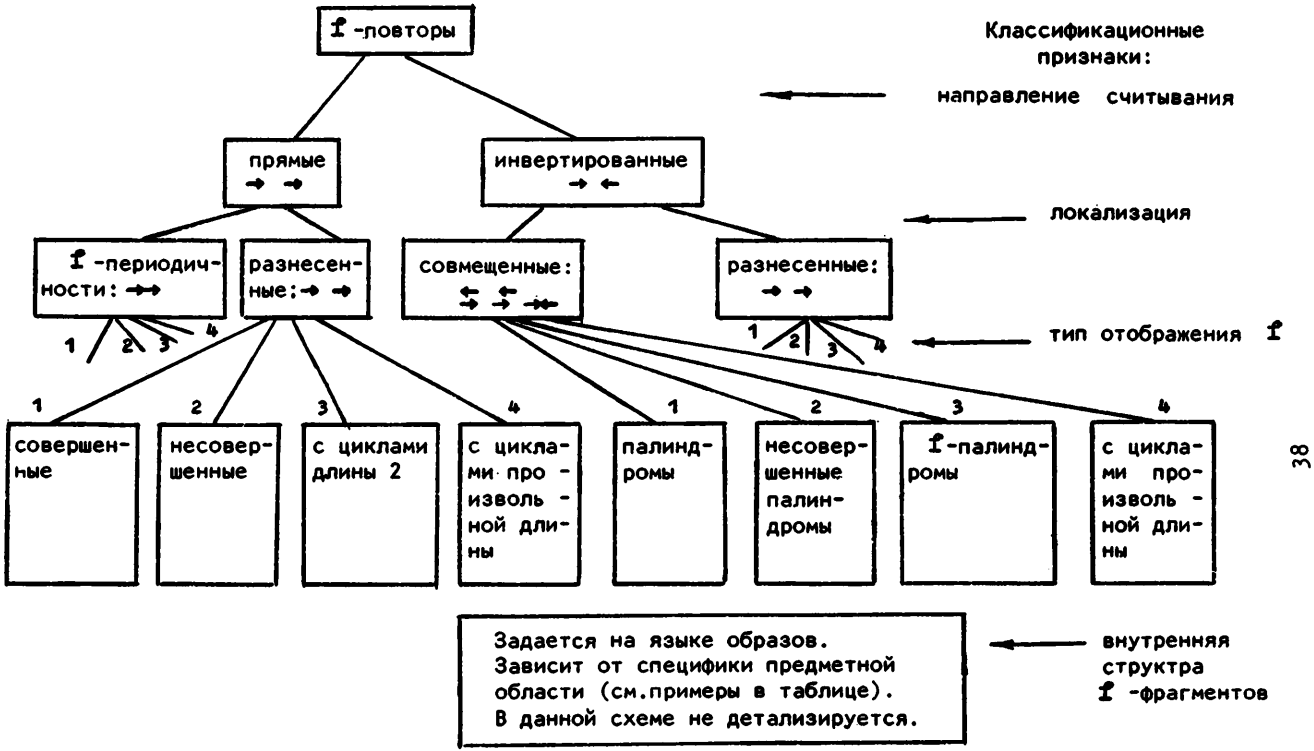


Рис.2. Классификационное дерево f-повторов

Таблица

Примеры f-повторов из генома бактериофага λ

№ п/п	Длина	Фрагменты, составляющие f-повтор	Локализация	Тип f-повтора	Закономерности внутренней структуры	Комментарии
1	2	3	4	5	6	7
1	15	Met Thr Glu Asp Asp CATG ACC GAG GAT GA CG CATG ACG GAG GAT GA TG Met Thr Glu Asp Asp	п.10480, ген T, п.19925, orf-401	Совершенный повтор; f = (AGCT / AGCT)	Повтор ATGACG	Самый длинный совершенный повтор в геноме фага λ. Ему соответствует повтор 5 аминокислот в генах T и orf-401. Три последних - отрицательно заряженные
2	16	A AAA GAA A AA AGA AAA Lys Glu Lys Arg Lys	п.39138, ген O	Палиндром; f = (AGCT / AGCT)	p = XYXZX, X = Lys, Y = Glu, Z = Arg	Пурино-богатый палиндром, расположенный в 3'-конце зоны начала репликации. Все кодируемые им аминокислоты - заряженные
3	16	AGAAAG GAAAC GACAG CTGTC GTTTC CTTTCT	п.109, некодирующий участок, п.151, некодирующий участок	Разнесенный, инвертированный с циклами длины 2, f = (AGCT / TCGA)	Несовершенные периодичности: $(GA^3G)^3$ и $(CT^3C)^3$	Данный f-повтор является ядром мощной шпилечной структуры (см. пояснения в тексте)
4	17	Asn Ser Asp Ile Gly ATT TGA ATC AAT TCC GGAA TTG ATT CAA AT Phe Gln Asn Leu	п.22848, ген bEa 47; компл. п.23807, ген bEa 47; компл.	см.№ 3	TTGAATCAATT AATTGATTCAA	Оба фрагмента локализованы в одном гене вблизи его концов
5	16	CATG T TG CA TG G TGCA 1 2 1 2	п.27162, некодирующий участок	f-палиндром, f= (AGCT / CTAG)	Повторы элементарных комплементарных палиндромов CATG и TGCA	Возникают сдвинутые шпилечные структуры на повторах 1 и 2 CATGTTGCATGGTGCA
6	15	T TTATT CTG TTATT T	п.34050, некодирующий участок	Несовершенный палиндром, f = (AGCT / ACGT)	Повтор симметричного фрагмента TTATT	Пример "усиления" конструкции: повтор короткого палиндрома приводит к образованию более длинного палиндрома

1	2	3	4	5	6	7
7	15	TCT CCATT|CCATT CTC	п.29251, некодирующий участок	Несовершенный палиндром $f = \begin{bmatrix} ATC \\ ACT \end{bmatrix}$	Повтор f-палиндрома CCATT	Аналогия с № 5,6: повтор короткого f-палиндрома способствует образованию более длинного
8	15	AAA CAG AAA GAT AAA Lys Gln Lys Asp Lys	п.42615, orf-204	Несовершенный палиндром; $f = \begin{bmatrix} AGCT \\ AGTC \end{bmatrix}$	p = XYXZX, X = Lys, Y = Gln, Z = Asp (см. № 2)	Расположен внутри аномальной по сложности зоны; все кодируемые им аминокислоты (кроме Gln) - заряженные. Дальнейшие пояснения см. в тексте
9	18	T|GCG GCA GAA|AAC AGC CGC|A Ala Ala Glu Asn Ser Arg	п.5480, ген Nu3	Несовершенный f-палиндром $f = \begin{bmatrix} AGCT \\ ACGT \end{bmatrix}$		Расположен вблизи аномальной по сложности зоны; образует шпилечную структуру на G-,C-элементах
10	15	AA GCG CGG GCG CGT T Ala Arg Ala Arg	п.11593, ген H	Несовершенный палиндром; $f = \begin{bmatrix} AGCT \\ TGCA \end{bmatrix}$	1) Повтор симметричного фрагмента G C G C G; 2) $p = (XY)^2$, X=Ala, Y=Arg	Интересный пример наложения комплементарного палиндрома и симметрии. Фрагмент содержит серии повторяющихся элементов: $A^2(GC)^2G^3(CG)^2T^2$
11	15	|CCT TTT AAA TTT TGG|CTT Arg Lys Phe Lys Pro Lys	п.36527, ген rex A компл.	Несовершенный палиндром; $f = \begin{bmatrix} AGCT \\ ACGT \end{bmatrix}$	1) Повтор T^4; 2) p = XYXZX, X = Lys, Y = Pro, Z = Phe (см.№ 2 и 8)	Фрагмент расположен в аномальной зоне, структурно похож на предыдущий и также составлен из коротких серий: $C^2T^4A^3T^4G^2$ Вновь, как и в № 2,8, характерно вхождение лизина, несущего положительный заряд
12	16	Gln Ala Leu Leu Ala G CAG GCG CTG CTG GCG G CTG GCG CAG CAG GCG Leu Ala Gln Gln Ala	п.9233, ген V, п.12212, ген H	Несовершенный повтор. $f = \begin{bmatrix} AGCT \\ TGCA \end{bmatrix}$	$p = XY^2X$, X = Ala, $Y \in \{Leu, Gln\}$	Оба фрагмента расположены в аномальных по сложности зонах, обладают внутренней симметрией и образуют повтор длины 16 с тремя несовпадениями
13	23	GCT CAT GCT GCC CTG CTG ACG CT Ala His Ala Ala Leu Leu The Leu	п.11952, ген H	Несовершенный f-палиндром; $f = \begin{bmatrix} AGCT \\ ATCG \end{bmatrix}$	$p = X_1Y_1X_1^2 \mid X_2^2Y_2X_2$, X_1=Ala, X_2= Leu, Y_1= His, Y_2= Thr	Самый длинный f-повтор в геноме фага λ. Характеризуется симметричным расположением аминокислот Ala и Leu относительно середины

1	2	3	4	5	6	7
14	15	Val Glu Glu Val Ala |GTG GAA GAG GTG GCG|C |C TCC ACC GCG GCC AC|G Ser Thr Ala Ala Thr	п.19977, orf-401, п.20315, orf-401	Разнесенный инвертированный f-повтор с циклом длины 4, $f = \begin{pmatrix} AGCT \\ GCTA \end{pmatrix}$	$p = XY^2X$; $X \in \{Val, Thr\}$, $Y \in \{Glu, Ala\}$ (см.№ 12)	Оба фрагмента из одного гена (orf-401), обладающего аномальными характеристика-ми
15	15	Met Val Leu Gly Asn ATG GTG CTG GGG AAC A CTT CTG CTT TTA AG Leu Leu Leu Leu	п.6948, ген E, п.34580, ген git	Прямой разнесенный f-повтор с циклом длины 3, $f = \begin{pmatrix} AGCT \\ ATGC \end{pmatrix}$	$p_2 = X^4$; X = Leu	Сдвиг на 1 символ рамки кодирования приводит к резкому различию структур на аминокислотном уровне
16	15	ACG TTC ACG CTT ACG Thr Phe Thr Leu Thr	п.18790, ген I	Несовершенный палиндром; $f = \begin{pmatrix} AGCT \\ GACT \end{pmatrix}$	p = XYXZX, X = Thr, Y = Phe, Z = Leu (см. № 2,8,11)	Расположен в аномальном по сложности фрагменте. Дальнейшие пояснения см. в тексте
17	16	CA CTG AAT GAA TGC AC Leu Asn Glu Cys	п.13451, ген L	Несовершенный палиндром; $f = \begin{pmatrix} AGCT \\ ATCG \end{pmatrix}$	$p = XY^{2,5}X$, X = CAC, Y = TGAA	Расположен в аномальном по сложности фрагменте. Периодичность "нецелой кратности" (TGAA) фланкирована повторами CAC
18	17	AG AG T TG TG GCT TGGCT CATTC CAT TC TC C TG TG	п.24110, некодирующий участок, п.29255, некодирующий участок	Разнесенный инвертированный f-повтор с циклом длины 4; $f = \begin{pmatrix} AGCT \\ GTAC \end{pmatrix}$	Периодичности: $(AG)^2$, $(TG)^2$, $(TGGCT)^2$, $(CATTC)^2$	Пример структурного сходства на нуклеотидном уровне.

данному параметру геном фага λ не отличается значимо от своих случайных аналогов, получаемых путем равномерного перемешивания символов. В связи с этим представлялась возможность проверить, не обнаруживаются ли значимые отличия при расширении класса повторов.

Всего в геноме фага λ было выявлено 77 различных f-повторов с длинами, колеблющимися в диапазоне от 15 до 23. Наиболее интересные из них приведены в таблице. Для f-повторов, выделенных в кодирующих частях (таких большинство), указан их аминокислотный состав в соответствии с информацией о разметке. Все последовательности выписаны в ориентации $5'$-$3'$ (слева направо). Если кодирование происходит по комплементарной цепи ($3'$-$5'$), в графе "локализация" ставится соответствующая пометка (в принципе информацию о кодирующей цепи можно извлечь из сопоставления нуклеотидной и аминокислотной последовательностей). Иногда бывает полезно расширить в обе стороны фрагменты, составляющие f-повтор. В этом случае фрагменты отделяются от контекста вертикальными линиями.

По результатам анализа f-повторов в геноме фага λ можно сделать следующие выводы.

3.1. Из 77 выявленных f-повторов оказалось 20 прямых и 57 инвертированных, 24 совмещенных и 53 разнесенных. Совершенных f-периодичностей длины 15 и выше не обнаружено. Из 24 возможных типов подстановок реализованы практически все. Наиболее часто встречавшиеся подстановки:

$f_1 = \begin{bmatrix} AGCT \\ ACGT \end{bmatrix}$ – 10 раз, $f_2 = \begin{bmatrix} AGCT \\ AGTC \end{bmatrix}$ – 7 раз,

$f_3 = \begin{bmatrix} AGCT \\ TGCA \end{bmatrix}$ – 5 раз, $f_4 = \begin{bmatrix} AGCT \\ GACT \end{bmatrix}$ – 5 раз

$f_5 = \begin{bmatrix} AGCT \\ AGCT \end{bmatrix}$ – 5 раз, $f_6 = \begin{bmatrix} AGCT \\ CTAG \end{bmatrix}$ – 5 раз,

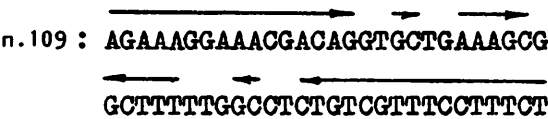
$f_7 = \begin{bmatrix} AGCT \\ TCGA \end{bmatrix}$ – 4 раза, $f_8 = \begin{bmatrix} AGCT \\ CTGA \end{bmatrix}$ – 4 раза.

$f_9 = \begin{bmatrix} AGCT \\ GCTA \end{bmatrix}$ – 4 раза и т.д.

Если предположить, что в каждом f-повторе с одинаковой вероятностью может быть реализована любая из 24 возможных подстановок, то десяти- и семикратное использование подстановок f_1 и f_2 соответственно будет иметь малую вероятность в рамках рассматриваемой схемы ($p_1 \sim 0{,}001$, $p_2 \sim 0{,}03$). Можно предположить, что, по крайней мере, первая из этих подстановок отражает какие-то специфические особенности генетического текста, в частности, наличие комплементарных структур (палиндромов, шпилек), а также поли-А- и поли-Т-участков. Анализ десяти f_1-повторов показал, что среди них преобладают шпилечные структуры, причем стебель шпильки представлен G–, C-элементами, а петля - А–, Т-элементами (см., к примеру, № 9 из таблицы) либо симметрии из А–, Т-элементов, разделенные (см. №6) или фланкированные (см. № 11) С–, G-элементами.

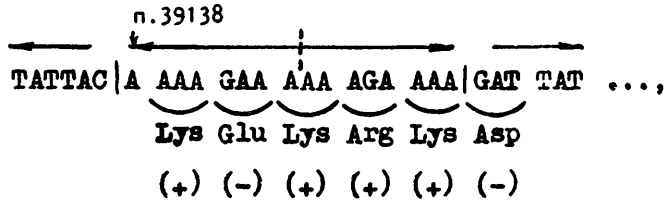
3.2. Длины f-повторов в геноме фага λ значимо не отличаются от аналогичных параметров для случайных (равномерно перемешанных) текстов с тем же нуклеотидным составом. Об этом свидетельствуют и результаты проведенных нами имитационных экспериментов. Из этого не следует, что все f-повторы носят случайный характер. О значимости некоторых из них можно судить по косвенным показателям, фиксирующим: а) позиционную близость фрагментов, составляющих f-повтор; б) возможность расширения f-повтора с сохранением (в той или иной степени) какой-либо характерной структурной особенности; в) возможность выделения данного фрагмента как аномального с помощью другой (по возможности, независимой) методики. Любой из перечисленных показателей играет роль своеобразного "усилителя" закономерности. Проиллюстрируем это на ряде примеров.

ПРИМЕР 2. Фрагменты, составляющие разнесенный инвертированный повтор №3 в таблице, расположены вблизи друг от друга (п.109,151), что позволяет предполагать их функциональную связанность. Анализ контекста показывает, что они составляют ядро мощной шпилечной структуры



расположенной в некодирующей области перед началом гена Nu1 в аномальной по сложности зоне.Как правило,такого рода структуры играют важную роль в регуляции основных генетических процессов.

ПРИМЕР 3. Палиндром №2 (см.таблицу) сохраняет симметричность при расширении его в обе стороны



кодирует кластер заряженных аминокислот и расположен в аномальном по сложности районе в 3'-конце зоны начала репликации.Совокупность перечисленных признаков позволяет предполагать его функциональную значимость.

ПРИМЕР 4. Несовершенный палиндром №8 (см.таблицу), аналогично предыдущему (см.пример 3), также состоит из заряженных аминокислот (за единственным исключением). Это свойство сохраняется при расширении его в обе стороны:

(продолжение примера см.на следующей странице)

CTT AAG ATT CGA AAA
Leu Lys Ile Arg Lys
(+) (+) (+)

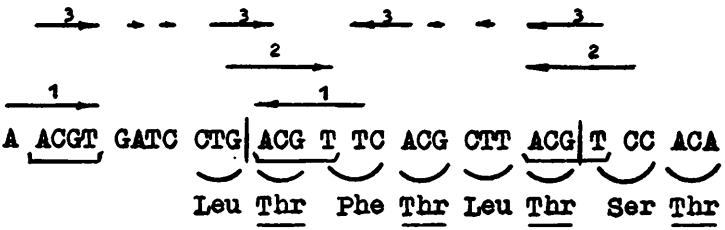
В работе [9], посвященной анализу серий заряженных аминокислот в различных геномах с помощью статистических критериев, данная область выделена в качестве аномальной. При анализе сложностных профилей эта область также попала в число аномальных (она содержит супердлинную серию из 68 (не T)-элементов и множество периодичностей). Выделение данной области в качестве аномальной с помощью различных методик предполагает ее функциональную значимость. В [9] высказана гипотеза, что электростатические взаимодействия, возникающие между кластерами заряженных аминокислот, могут содействовать прикреплению фаговых частиц к соответствующим рецепторам клеток хозяина.

3.3. Закономерности внутренней структуры f-повторов можно трактовать на двух уровнях: нуклеотидном и аминокислотном. Нуклеотидный уровень выступает на первый план, если f-повтор расположен в некодирующей части генома (см. f-повторы № 5-7, 18 из таблицы). Элементами структуры здесь выступают короткие периодичности и палиндромно-шпилечные конструкции. Иногда и кодирующие участки более естественно трактовать на нуклеотидном уровне, а не на аминокислотном. В частности, это касается периодичностей с длиной периода, не кратной трем (см. №17 из таблицы), и ситуаций, связанных с наложением знаковых конструкций на кодирующую область (см. №9).

Структурированность f-повторов на аминокислотном уровне проявляется в характерном расположении и составе аминокислот. Несмотря на ограниченное количество анализировавшихся f-повторов, наблюдается кластеризация их по определенным подтипам. Так, закономерность, описываемая образом p = XYXZX, где X - лизин, а Y и Z - произвольные (часто заряженные) аминокислоты, встречается в f-повторах № 2,8,11 из таб-

лицы; закономерность $p = X^2YX$ встречается 4 раза, но в качестве X в разных f-повторах фигурируют разные аминокислоты; закономерности $p = (XY)^2$ и $p = XY^2X$ встречаются по 2 раза и т.д. Заметим, что вследствие вырожденности генетического кода структурированность на аминокислотном уровне может не проявлять себя на нуклеотидном.

Некоторые f-повторы обнаруживают структурированность как на нуклеотидном, так и на аминокислотном уровнях. В качестве примера рассмотрим f-повтор № 16 из таблицы, расширив его в обе стороны:



Трехкратный повтор элементарного комплементарного палиндрома $\overrightarrow{ACGT}\overleftarrow{}$ приводит к возникновению множественных шпилечных структур (1,2,3), которые могут играть определенную роль на этапе транскрипции. На аминокислотном уровне имеет место уже отмечавшаяся выше закономерность вида $p = XYXZX$, но роль лизина (Lys) играет треонин, регулярно чередующийся с другими аминокислотами.

3.4. Примеры, приведенные выше, позволяют предположить, что методика анализа f-повторов содержит в себе потенциальные возможности обнаружения неожиданных ассоциаций, взаимодействий (типа упоминавшихся электростатических взаимодействий, проявляющих себя в наличии кластеров заряженных аминокислот). В связи с этим интересно было проанализировать наличие и характер прямых f-повторов с циклами длины 2, где $f = \begin{bmatrix} AGCT \\ TCGA \end{bmatrix}$,

имеющих, предположительно, отношение к так называемой "парал - лельной ДНК" [10]. В проведенном эксперименте был выявлен всего один повтор такого типа длины 15, который,по-видимому, следует считать случайным. Возможно, рассматривавшийся нами объект (геном бактериофага λ) в этом смысле нельзя считать показательным.

З а к л ю ч е н и е

Введено понятие f-повтора как пары фрагментов, совпадающих с точностью до переименования входящих в них элементов.Частным случаем f-повторов являются повторы в обычном смысле и симметрии. В общем случае фрагменты, образующие f-повтор,характеризуются одинаковой сложностной структурой, т.е. они могут быть составлены из разных элементов, но повторяемость и чередуемость этих элементов в каждом из фрагментов подчинены одному и тому же закону.

Предложен универсальный и эффективный алгоритм отыскания f -повторов заданной длины в достаточно длинных текстах,гарантирующий выявление всех пар фрагментов, связанных друг с дру - гом любым из допустимых (но не фиксируемым заранее) способов переименования элементов алфавита. Намечены основы классификации f-повторов, в основу которой положены такие признаки,как направление считывания фрагментов, составляющих f-повтор, их локализация, тип отображения f , внутренняя структура f-фрагментов.

Проведен эксперимент по обнаружению f-повторов в геноме бактериофага λ . Их анализ подтвердил возможность выявления формальными методами (по типу превалирующего отображения f) некоторых специфических особенностей предметной области (в частности, наличия комплементарных конструкций и кластеров заряженных аминокислот). Это позволяет рекомендовать описанную ме-

тодику для анализа плохо изученных языковых систем, представ - ленных текстовыми выборками достаточно большого объема.

Литература

1. LEMPEL A., ZIV J. On the complexity of finite sequen - ces //IEEE. Trans. on Inf. Th. - 1976. -Vol.IT-22, N1.- P. 75-81.

2. ГУСЕВ В.Д. Сложностные профили символьных последовательностей //Методы обработки символьных последовательностей и сигналов. - Новосибирск, 1989. -Вып. 132: Вычислительные сис - темы. - С. 35-63.

3. САЛОМАА А. Жемчужины теории формальных языков.-М.: Мир, 1986. - 159 с.

4. ЧУПАХИНА О.М. Алгоритм построения сложностного профи - ля символьной последовательности //Методы обработки символьных последовательностей и сигналов. - Новосибирск, 1989.- Вып. 132: Вычислительные системы. - С. 64-91.

5. АХО А., ХОПКРОФТ Дж., УЛЬМАН Дж. Построение и анализ вычислительных алгоритмов /Под ред. Ю.В.Матиясевича. -М.: Мир, 1979. - 536 с.

6. AHO A.V.,CORASICK M.J. Efficient string matching: an aid to bibliographic search //Communications of the ACM.-1975. - Vol. 18, N 6. -P. 333-340.

7. ANGLUIN D. Inductive inference of formal languages from positive data //Inform. and control. - 1980. -Vol. 45. N. 3, - P. 117-135.

8. ГУСЕВ В.Д., КУЛИЧКОВ В.А., ЧУПАХИНА О.М. Сложностной анализ генетических текстов (на примере фага λ). - Новоси - бирск, 1989. - 48 с. - (Препринт/АН СССР. Сиб. отд-ние.Инсти - тут математики, № 20).

9. KARLIN S., BRENDEL V. Charge configurations in viral proteins //Proc. Natl. Acad. Sci. USA. - 1988. - Vol. 85. - P. 9396-9400.

10. Параллельная ДНК-возможность существования /Н.А.Чури - ков, В.Б.Чернов, Ю.Б.Голова, Ю.Д.Нечипуренко //Докл. АН СССР. - 1988. - Т. 303, № 5. -С. 1254-1258.

Поступила в ред.-изд.отд.

23 июля 1991 года