П Р О Б Л Е М Ы О Б Р А Б О Т К И И Н Ф О Р М А Ц И И
(Вычислительные системы)
1983 год Выпуск 100

УДК 519.766.24:577

МЕХАНИЗМЫ ОБНАРУЖЕНИЯ СТРУКТУРНЫХ ЗАКОНОМЕРНОСТЕЙ В СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ

В.Д.Гусев

С анализом символьных последовательностей (текстов) приходится иметь дело в теории связи, лингвистике, генетике, информатике и ряде других областей. Для примера можно указать на важные в прикладном отношении задачи сжатия текстов (без потери содержащейся в них информации), автоматического обнаружения и исправления ошибок, закрытия баз данных, идентификации текстов (по способу порождения, времени порождения и т.д.), обнаружения знаков пунктуации в первичных структурах нуклеотидных молекул, восстановления текста по отдельным пересекающимся фрагментам.

Традиционно используемые для анализа символьных последовательностей модели (порождающие грамматики и марковские цепи) описывают ансамбли объектов (текстов). В приложениях же объектом исследования часто выступает отдельный уникальный по своей природе текст (ДНК-молекула конкретного вируса, литературное или музыкальное произведение конкретного автора, файл данных, подлежащий сжатию и т.д.). Каждый из них обладает индивидуальной с т р у к т у р о й, которая в первом приближении может быть охарактеризована с о с т а в о м, к о л и ч е с т в о м и р а с п о л о ж е н и е м внутри текста отдельных специфических фрагментов (или образцов), являющихся подпоследовательностями исходного текста. Выявление (а иногда и изменение) понимаемой таким образом структуры лежит в основе перечисленных выше содержательных задач.

Целью данной работы является установление взаимосвязи и систематизация под указанным углом зрения известных и частично новых фактов, понятий, подходов, алгоритмов, которые в совокупности можно назвать механизмами обнаружения закономерностей в символьных последовательностях.

Изложение будет сконцентрировано на следующих пяти вопросах, упорядоченных по нарастанию их сложности:

I) как определить элементы структуры (проблема представления)?

2) как быстро отыскать их в тексте (проблема поиска)?

3) каким образом могут эволюционировать (варьировать) объекты, описываемые символьными последовательностями? Как определить меру близости двух объектов с учетом класса допустимых вариаций (проблема близости)?

4) как быстро отыскать в тексте элементы структуры с учетом их возможных вариаций (поиск приближенных соответствий)?

5) каким образом по выборке близких в определенном смысле текстов синтезировать текст-обобщение, содержащий варьируемые элементы, заменой которых определенными элементами (или цепочками элементов) алфавита можно получить любой из текстов обучающей выборки (проблема индуктивного синтеза текстов)?

I. Проблема представления. Поиск элементов структуры должен идти от содержательной задачи. В наиболее простых ситуациях они задаются в явном виде. Иллюстрирующим примером может служить задача о разметке генетических текстов (первичных структур НК-молекул), когда по расшифрованной НК-молекуле (геному) требуется определить ее генетическую структуру (кодирующие и некодирующие участки, число и расположение генов, различные знаки пунктуации - инициирующие и терминальные кодоны, рибосомные сайты связывания, промоторы и т.д.). Уровень знаний, накопленный о знаках пунктуации, сейчас таков, что почти для каждого из них может быть выписана обобщенная "формула" (см. вопрос 5). С учетом этого задача о разметке сводится к поиску внутри генома по обобщенным "формулам" всех потенциально возможных знаков пунктуации, анализу их взаимного расположения и устранению "ложных" знаков (иногда за счет привлечения дополнительной информации или после экспериментальной проверки).

Гораздо более типичной является ситуация, когда элементы структуры не заданы. В этом случае можно рекомендовать несколько стандартных заготовок для них, хорошо зарекомендовавших себя при решении различных содержательных задач. Первая из них связана с понятием п о в т о р а - наличием одинаковых фрагментов в различных частях текста. Структура текста в этом случае характеризуется совокупностью всевозможных повторов ("алфавитом" повторов), час-

тотами встречаемости каждого из элементов этого алфавита и (иногда) указанием мест вхождения в текст наиболее характерных элементов алфавита. Приведем примеры, подтверждающие перспективность такого подхода в содержательном плане.

Применительно к естественному языку в терминах повторов можно сформулировать понятие морфемы - элементарной смыслонесущей единицы языка [I]. Аналогом морфемы в музыке выступает интонация [2]. Дерево повторов используется в [3] для индуктивной реконструкции грамматики по текстовой выборке достаточно большого объема, адекватно отражающей определенную языковую подсистему. Дупликация является одной из важнейших элементарных операций, с помощью которых осуществляется эволюция ДНК-молекул. Наконец, большинство последних приближений к оценке сложности последовательностей [4] и изображений [5] апеллируют к понятию повтора.

I.I. Для формального определения элементов структуры введем следующие обозначения: A - алфавит (конечное множество символов); n - мощность алфавита A; текст - конечная последовательность символов из A; $N = |T|$ - длина текста T; T[i] - i-й элемент текста T (или элемент, занимающий i-ю позицию, $1 \le i \le N$); T[i:j] - элементы текста T с i-го по j-й включительно ($1 \le i < j \le N$); T_1T_2 - конкатенация (объединение) текстов T_1 и T_2. Если текст T представлен в виде конкатенации XYZ, то X будем называть префиксом (началом) текста T, Y - словом или цепочкой, Z - суффиксом (концом) текста T.

Назовем l-г р а м м о й связную подпоследовательность текста из l подряд расположенных символов ($l = 1,2,\ldots, N$). Полное число l-грамм в тексте длины N равно $N - l + 1$. Поскольку среди них могут быть повторяющиеся, количество разных l-грамм $M_l \le N - l + 1$. Ч а с т о т н а я х а р а к т е р и с т и к а п о р я д к а l текста T есть совокупность элементов $\Phi_l(T) = \{\Phi_{l1}, \Phi_{l2}, \ldots, \Phi_{lM_l}\}$, где каждый элемент Φ_{li} ($1 \le i \le M_l$) есть пара ⟨i-я l-грамма, частота ее встречаемости в тексте⟩. П о л н ы й ч а с т о т н ы й с п е к т р т е к с т а T - совокупность частотных характеристик $\Phi(T) = \{\Phi_1(T), \Phi_2(T), \ldots, \Phi_{l_{max}}(T), \Phi_{l_{max}+1}(T)\}$, где l_{max} - минимальное значение l, начиная с которого в тексте уже отсутствуют повторяющиеся l-граммы.

Заметим, что по частотной характеристике l-го порядка может быть восстановлена характеристика (l-1)-го порядка (за иск-

лючением, быть может, частот начальной и конечной (l-1)-грамм). По частотной же характеристике $\Phi_{l_{max}+1}(T)$ однозначно могут быть идентифицированы начальная и конечная l-граммы текста, после чего может быть восстановлен сам текст путем выявления пар $(l_{max}+1)$-грамм с совпадающими l_{max}-граммами [6].

Из форм представления частотных характеристик отметим представление с упорядочением l-грамм по убыванию частоты встречаемости, с лексикографическим упорядочением и в виде хеш-таблицы, обеспечивающей максимально быстрый поиск каждой l-граммы. Полный частотный спектр удобно представлять в виде дерева.

Описанные структуры удобно использовать при вычислении энтропийных характеристик текста, построении эффективных кодов для сжатия текста, для оценки по выборке переходных вероятностей в моделях марковского типа, в задачах классификации.

К недостаткам этих структур следует отнести:

а) некоторую их избыточность (в достаточно длинном тексте, как правило, содержится очень много повторов и отнюдь не все из них являются функционально значимыми (в плане интересующей исследователя содержательной задачи));

б) зачастую необходимой бывает и информация о расположении повторов в тексте.

I.2. Естественным шагом к устранению первого недостатка является выделение из частотного спектра лишь тех l-грамм, которые являются функционально значимыми. В связи с этим возникает очень важный для приложений вопрос: какие l-граммы считать ф у н к ц и о н а л ь н о з н а ч и м ы м и?

Рецептов, универсальных для всех приложений, здесь быть не может. Во многих случаях, однако, помогает сопоставление характеристик анализируемого текста с аналогичными характеристиками случайной последовательности такой же длины и с тем же алфавитом, полученной по схеме независимых испытаний. Наличие а н о м а л ь н ы х о т к л о н е н и й от схемы независимых испытаний часто свидетельствует о функциональной значимости наблюдаемого эффекта.

Для отделения "неслучайных" повторов от "случайных" нужно учитывать такие признаки, как длина повтора, его частота, расположение повторов в тексте и возможность их расширения. Как правило, неслучайными оказываются короткие повторы, имеющие аномально низкую частоту встречаемости, или достаточно длинные повторы, ли-

бо повторы хоть и не столь длинные, но расположенные подряд (повторы - п е р и о д и ч н о с т и, имеющие вид VV, где V - слово), либо повторы, которые могут быть расширены до длинных при условии, что допускается небольшое число замен, вставок или устранений символов.

Вероятностные оценки параметров $l_{max}(N, n)$ для случайных текстов дают представление о том, какой длины случайные повторы могут встречаться в тексте длины N. В [7,8] получены явные неравенства для допредельных и предельных распределений параметра $l_{max}(N, n)$, последнее достаточно легко табулируется. Для иллюстрации в табл.I приведены представляющие интерес для генетических приложений ($n = 4$) значения функции распределения $P\{l_{max}(N,4) \le k\}$ для отдельных значений N и k. Значения k, включенные в таб-

Т а б л и ц а I

Численная иллюстрация предельного распределения параметра $l_{max}(N,4)$ и результаты имитационного моделирования

Предельное распределение			Имитационное моделирование		
N	k	$P\{l_{max}(N,4) < k\}$	N	$\bar{l}_{max}(N,4)$	$s^2(l_{max}(N,4)$
50	8	0,985	20	3,42	0,80
50	9	0,996	50	4,88	I,06
I00	9	0,985	I00	5,95	I,06
500	I2	0,994	500	8,I2	0,9I
I000	II	0,9I0	700	8,6	0,75
I000	I2	0,976	I000	9,09	0,83
3000	I4	0,964	2000	I0,2	0,93
3000	I5	0,996	3000	I0,7	I,I2
5000	I3	0,859			

лицу, характеризуют длины повторов, вероятность встречаемости которых в случайных последовательностях длины N весьма невелика (порядка нескольких сотых). Здесь же для сравнения приведены результаты имитационного моделирования*) по оценке математического ожидания и дисперсии параметра l_{max} (усреднения при всех значениях N проводились по I00 реализациям).

*) Эксперимент выполнен Т.Н.Титковой (ИМ СО АН СССР).

Для численных оценок часто достаточно знать лишь порядок величины $l_{max}(N, n)$. Он зависит от вероятностей $p_r (1 \le r \le n)$ элементов алфавита и длины текста следующим образом [8]:

$$l_{max} \sim 2 \ln N / |\ln \sum_{r=1}^{n} p_r^2|. \qquad (1)$$

Аналогичный результат получен в [9].

Ярким примером такого подхода к обнаружению функционально значимых повторов является недавнее сообщение об установлении гомологии (соответствия) между одним из генов вируса SV 40, вызывающего рак у обезьян, и содержащимся в крови человека белком PDGF, играющим роль фактора роста клеток при заживлении ран, что позволило выдвинуть новую гипотезу о механизме возникновения рака. Гомологичные участки имеют вид:

```
PDGF: SLGSLTIAEPAMIAECKTRTEVFEISRRLID
      |||||  ||||||||||||||||||||||
SV40:RSLGSLSVAEPAMIAECKTRTEVFEISRRLIDR
```

Гомология была установлена после расшифровки аминокислотной последовательности белка PDGF (104 аминокислоты) и сопоставления ее с имеющимся банком расшифрованных белковых структур. Длина повтора составляет 24 аминокислоты, что при мощности алфавита n = 20 делает чрезвычайно маловероятной гипотезу о случайном совпадении. К тому же обнаруженный повтор расширяем влево (при наличии двух несовпадений).

Другим примером является так называемый ЦГ-эффект у эукариотов (см. подробнее [10]), заключающийся в аномально низкой (по сравнению с остальными биграммами) частоте встречаемости пары ЦГ в геномах эукариотов. Обнаруженный формальными методами этот эффект получил затем содержательную интерпретацию и существенно прояснил регуляцию процесса транскрипции в организме. Аналогичный пример, но уже из естественного языка связан с обнаружением ошибок в тексте по коротким редким биграммам.

В заключение данного подраздела отметим, что очень длинные повторы в тексте встречаются довольно редко. Чаще они "зашумлены", т.е. отличаются друг от друга, но могут быть переведены один в другой относительно небольшим числом операций из класса допустимых. В разделе 4 мы коснемся вопроса о том, какие критерии существуют в этом случае для отделения случайных "повторов" от неслучайных.

I.3. Возвращаясь ко второму недостатку чисто частотных структур, опишем еще две структуры, куда информация о позициях характерных 1-грамм входит уже в явном виде. Эти структуры суть дерево идентификаторов [II,I2] и суффиксное дерево [I3].

Будем говорить, что слово t идентифицирует позицию i в тексте T, если в позиции i начинается единственное вхождение слова t в T. Для того, чтобы каждая позиция идентифицировалась хотя бы одним словом, текст T обычно дополняют справа произвольным символом $z \notin A$. Кратчайшее из возможных слов, идентифицирующих позицию i в Tz, называется идентификатором позиции i. Обозначим его через $t(i)$.

Множество идентификаторов текста Tz удобно представлять в виде дерева, ребра которого помечены символами из $\{ A \cup z \}$. Дерево имеет $N+1$ листьев, помеченных числами $1,2,\ldots, N+1$, однозначно соответствующими позициями в Tz (в силу чего это дерево иногда называют позиционным [I2]). Идентификатор $t(i)$ определяется последовательностью реберных меток, характеризующих путь из корня в лист с меткой i. Для примера на рис.I изображено дерево идентификаторов для текста Tz = abcababz.

i	t(i)
I	abc
2	bc
3	c
4	aba
5	ba
6	abz
7	bz
8	z

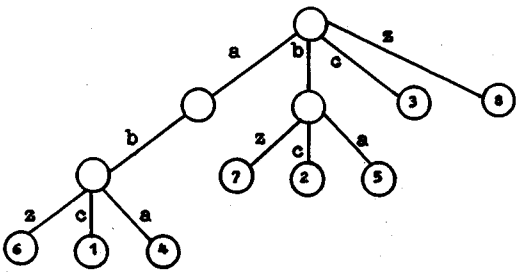

Рис. I.

Пусть $T = a_1 a_2 \ldots a_i a_{i+1} \ldots a_N$, где $a_i \in A$ $(1 \le i \le N)$. Слово $T[i:N] = a_i a_{i+1} \ldots a_N$ назовем i-м суффиксом текста T и будем обозначать его $s(i)$. По аналогии с идентификаторами совокупность всех суффиксов текста удобно представлять в виде суффиксного дерева, где i-му суффиксу соответствует путь из корня в терминальный узел (лист), помеченный числом i. Существование терминального узла для каждого суффикса обеспечивается дополнением текста T элементом $z \notin A$ (в этом случае никакой суффикс не явля-

ется продолжением другого). Линейные (неразветвляющиеся) участки дерева часто заменяют одним ребром, помеченным словом, соответствующим линейному участку (компактная форма суффиксного дерева). На рис.2 изображено суффиксное дерево текста ababa в прямой (а) и компактной (б) форме.

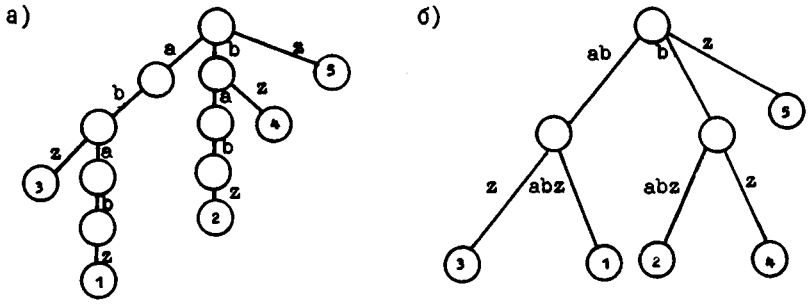

Рис.2.

Две описанные структуры имеют много общих свойств. Можно показать, что и дерево идентификаторов и суффиксное дерево могут иметь $O(N^2)$ узлов. Переход к компактной форме представления снижает в обоих случаях количество узлов до $O(N)$. В наихудшем случае число узлов в дереве идентификаторов не превышает $4N - 2$, а в суффиксном дереве - $2N$.

Эти конструкции можно считать функционально эквивалентными. Имеющиеся различия в реализации (см.раздел 2) не носят принципиального характера. С помощью этих структур можно, например, построить линейные алгоритмы для решения следующих задач:

а) найти все вхождения заданного слова (или группы слов) в текст;

б) найти самое длинное повторяющееся слово в тексте;

в) найти самое длинное слово, начинающееся с i-й позиции текста и встречающееся также в произвольной позиции j, предшествующей i (с подобной задачей приходится сталкиваться при сжатии текстовой информации [I4]);

г) найти самое длинное слово, общее для двух заданных текстов.

2. Проблема поиска. Эта проблема возникает, когда тексты достаточно длинны и содержат большое количество структурных эле-

ментов, либо когда запросы на поиск нужного структурного элемента следуют слишком часто.

Простейшая задача поиска сводится к определению всех вхождений слова в текст. Для ее решения разработаны сублинейные в среднем алгоритмы, требующие меньшего, чем длина самого текста числа попарных сравнений различных символов [I2,I5]. В случае группового запроса трудоемкость поиска составляет $O(N + M)$, где M - суммарная длина образцов, предъявленных для поиска. Образцы при этом упаковываются в дерево, аналогичное представленным на рис.I,2, и поиск организуется так, что каждый символ текста просматривается один раз вне зависимости от числа и длин образцов.

Следующая по сложности задача - вычисление $\Phi_1(T)$ - частотной характеристики 1-го порядка. При небольших значениях 1 задача может быть решена с помощью процедуры хеширования [I6] алгоритмом с линейной трудоемкостью. Продвигаться в сторону больших значений 1 можно, комбинируя более крупные 1-граммы из мелких, например, так, как это сделано в [I7].

Как уже упоминалось выше, при компактном представлении суффиксные деревья и деревья идентификаторов содержат $O(N)$ узлов. Деревья строятся итеративно по i, т.е. на каждом шаге к дереву добавляется новый идентификатор или суффикс. Весьма нетривиальным является факт возможности получения компактного представления этих деревьев непосредственно по исходному тексту за линейное по N число шагов. Это достигается введением вспомогательных структур, облегчающих поиск на дереве. С другой стороны, они же и усложняют логическую структуру алгоритма, что может привести к большой мультипликативной константе в оценке трудоемкости. Поэтому при о г р а н и ч е н н ы х значениях N конкуренцию данному подходу могут составить асимптотически более трудоемкие, но имеющие меньшую мультипликативную константу алгоритмы.

К недостаткам конструкций следует отнести необходимость хранения всего текста в оперативной памяти (ОП) и зависимость конструкций от размера алфавита. К примеру, в приведенном в [I2] алгоритме построения дерева идентификаторов каждому узлу приписывается двоичный вектор длины n. При $n \sim N$ получаем квадратичные затраты по памяти. В [I3] память зависит от n уже логарифмическим образом, но трудоемкость при этом оказывается линейной лишь в среднем.

Для последовательностей, целиком размещающихся в ОП, полный частотный спектр, в принципе, может быть получен с помощью указанных конструкций, если, к примеру, приписывать узлам дерева при его построении метки, соответствующие числу посещений данного узла. Пути, ведущие из корня в узлы 1-го уровня, будут определять 1 -граммы, вошедшие в $\Phi_1(T)$, а метки самих узлов - их частоты.

Если последовательности очень длинны и не помещаются в ОП, их разбивают на части, каждая из которых может быть обработана в ОП, а затем объединяют результаты обработки отдельных частей. Принципы такого разбиения, а также алгоритмы вычисления полного частотного спектра и отыскания наиболее длинных повторов приведены в [18,19]. Алгоритмы основаны на нетрадиционных модификациях процедуры хеширования, использующих специфику задачи. Трудоемкость алгоритма вычисления $\Phi(T)$ для класса случайных последовательностей составляет $O(N^2 \ln N/S)$, где S - объем доступной ОП в битах. Использование в этой ситуации конструкций, описанных в п.I.3, проблематично.

Как уже говорилось выше, повторы - периодичности даже при относительно небольшой длине могут оказаться функционально значимыми. Слисенко А.О. в 1977 г. [20] построил алгоритм для нахождения в "реальное время" всех периодичностей в тексте (для вычислений в ОП). Термин в "реальное время" означает, что время работы алгоритма после поступления очередной буквы текста, но до поступления следующей не превосходит константы. В [21] для этих же целей предложен алгоритм с трудоемкостью $O(N \log N)$ и линейной памятью.

Как показывают приведенные ссылки, для нахождения типовых структурных элементов существуют эффективные алгоритмы даже в случае текстов большой длины.

3. Проблема близости. Выделяемые в текстах структурные элементы могут изменяться от реализации к реализации. К примеру, слова в текстах естественного языка могут быть зашумлены ошибками типа замены одного символа другим, перестановки двух символов, вставки или устранения символа. Эволюция генетических объектов характеризуется такими элементарными операциями, как замена одного нуклеотида другим, вставка или устранение г р у п п ы нуклеотидов, дупликация, инверсия, транспозиция. Не менее широк диапазон вариаций, присущих музыкальным текстам. В общем случае вариации могут привести к изменению не только состава, но и длины текста.

Весьма полезными в плане сопоставления двух зашумленных текстов оказались понятия максимально длинной общей подпоследовательности (МДП) и совместного частотного спектра.

3.I. Назовем U подпоследовательностью V, если существует монотонно возрастающая последовательность целых $r_1, r_2, \ldots, r_{|U|}$ такая, что $U[i] = V[r_i]$ для $1 \le i \le |U|$. Слово U является общей подпоследовательностью последовательностей T_1 и T_2, если U - подпоследовательность как T_1, так и T_2. МДП есть общая подпоследовательность с наибольшим возможным числом элементов. Две последовательности могут иметь несколько МДП, отличающихся как по составу элементов, так и по их расположению (при одинаковом составе). К примеру, последовательности T_1 = abcdb и T_2 = abbd имеют $(МДП)_1$ = abb и $(МДП)_2$ = abd.

Понятие МДП связано с введенным Левенштейном [22] расстоянием между парой текстов, определяемым как минимальное число операций (типа замены одного символа другим, устранения символа и включения символа), переводящих один текст в другой. Позднее операции указанного типа получили название редакционных и этот термин был перенесен на само расстояние [23]. Каждая из этих операций может иметь свою "стоимость". Если принять стоимость второй и третьей операций равной I, а стоимость первой - 2, то длина $p(T_1,T_2)$ МДП будет связана с введенным таким образом расстоянием $d(T_1,T_2)$ соотношением $d(T_1,T_2) = |T_1| + |T_2| - 2p(T_1,T_2)$.

В [24,25] изучается поведение случайной величины p (длины МДП) для последовательностей независимых случайных величин с вероятностями p_r=1/n появления каждого из элементов алфавита. Показано существование $c_n = \lim\limits_{N\to\infty} Ep_{N,n}(T_1,T_2)/N$, где E - знак математического ожидания. Нижние ($\underline{c}_n$) и верхние ($\bar{c}_n$) границы константы c_n уточняются. В табл.2 приведены значения $\underline{c}_n$ из [25] и $\bar{c}_n$ из [24] для $1 \le n \le 15$.

В [25] показано, что для больших значений n константа c_n убывает не быстрее, чем $1/\sqrt{n}$. Этот результат согласуется с работами других авторов по оценке c_n для случайных перестановок (n = N): $Ep_{N,N}(T_1,T_2) = 2\sqrt{N}$.

Приведенные оценки при не слишком различающихся и не слишком малых длинах последовательностей могут служить ориентиром для принятия решения о наличии либо отсутствии сходства между последовательностями.

Таблица 2

Известные нижние и верхние границы константы c_n

n	$\underline{c}_n$	$\bar{c}_n$
2	0,7615	0,8665
3	0,6153	0,7864
4	0,5454	0,7297
5	0,5061	0,6861
6	0,4716	0,6509
7	0,4450	0,6209
8	0,4223	0,5967
9	0,4032	0,5750
10	0,3865	0,5559
11	0,3719	0,5389
12	0,3589	0,5236
13	0,3473	0,5097
14	0,3368	0,4970
15	0,3273	0,4853

3.2.Понятие совместного частотного спектра д в у х текстов вводится по той же схеме, что и понятие полного частотного спектра о д н о г о текста.Назовем с о в м е с т н о й частотной характеристикой 1-го порядка текстов T_1 и T_2 совокупность элементов $\Phi_l(T_1, T_2)=\{\varphi_{l1}(T_1,T_2),\varphi_{l2}(T_1,T_2),\ldots,\varphi_{lM_l}(T_1,T_2)\}$, где $M_l(T_1,T_2)$ - количество различных l-грамм, общих для обоих текстов ($0\le M_l(T_1,T_2) \le \min(M_l(T_1),M_l(T_2))$), а каждый элемент $\varphi_{li}(T_1,T_2)$ ($1 \le i \le M_l(T_1,T_2)$) есть тройка: $\langle$ i-я l-грамма, частота ее встречаемости в T_1 - $F_{li}^{(1)}$, частота ее встречаемости в T_2 - $F_{li}^{(2)}$ $\rangle$.

Совместный частотный спектр текстов T_1 и T_2 есть совокупность совместных частотных характеристик $\Phi(T_1,T_2) = \{\Phi_1(T_1,T_2),\Phi_2(T_1,T_2),\ldots,\Phi_L(T_1,T_2)\}$, где L - максимальное значение l, при котором в текстах T_1 и T_2 еще есть общие l-граммы (т.е. $M_L(T_1,T_2) \neq 0$, а $M_{L+1}(T_1,T_2)=0$).

Проиллюстрируем возможность использования введенного понятия на двух примерах (третий будет представлен в п.5). В [26] предложен способ вычисления расстояния между двумя символьными последовательностями в виде

$$h(u,v) = \max(m_u,m_v) - m_0, \qquad (2)$$

где m_u и m_v - количество признаков, выделяемых из текстов u и v, а m_0 - число совпавших признаков. Если в качестве признаков использовать l-граммы, получаемые сдвигом вдоль текста скользящей рамки, охватывающей l символов, то $m_u=N_u-l+1$, $m_v=N_v-l+1$, с учетом чего (2) можно переписать в виде:

$$h_l(u,v) = \max(N_u-l+1,N_v-l+1) - \sum_{i=1}^{M_l(u,v)} \min\{F_{li}^{(u)},F_{li}^{(v)}\}. \qquad (3)$$

Таким образом, вся информация, необходимая для вычисления (3), содержится в $\Phi_l(u,v)$. Авторы метрики (2) утверждают, что по своим свойствам она близка к редакционному расстоянию.

Второй пример относится к мере близости двух последовательностей, предложенной в [27]:

$$\lambda(u,v) = \frac{\sum\limits_{\alpha} \min\{F(u:\alpha), F(v:\alpha)\}\cdot|\alpha|}{\sum\limits_{\alpha} \max\{F(u:\alpha), F(v:\alpha)\}\cdot|\alpha|}, \qquad (4)$$

где $F(u:\alpha)$ и $F(v:\alpha)$ - частоты встречаемости произвольной 1-граммы α в u и v соответственно, а $|\alpha| = 1$. Авторы, предложившие эту меру, не указывают алгоритма ее вычисления, но нетрудно показать, что информации, содержащейся в $\Phi(u,v)$, достаточно для получения (4). Действительно, из определения $\Phi(u,v)$ следует, что в совместном частотном спектре содержатся только те 1-граммы, для которых $\min\{F(u:\alpha), F(v:\alpha)\} \neq 0$. Именно эти 1-граммы и дают ненулевой вклад в числитель (4), который для краткости обозначим через S_{min}. Таким образом, S_{min} может быть получен линейным просмотром всех $\Phi_1(u,v) \in \Phi(u,v)$.

Для вычисления знаменателя (4) представим каждое слагаемое в виде: $\max\{F(u:\alpha), F(v:\alpha)\} = F(u:\alpha) + F(v:\alpha) - \min\{F(u:\alpha), F(v:\alpha)\}$ и учтем, что для любого текста T $\sum\limits_{\alpha} F(T:\alpha)\cdot|\alpha| = \frac{1}{6} N_T(N_T+1)(N_T+2)$. Тогда (4) запишется в виде:

$$\lambda(u,v) = \frac{S_{min}}{\frac{1}{6}\sum\limits_{T=u,v} N_T(N_T+1)(N_T+2) - S_{min}}, \qquad (5)$$

требующем для своего вычисления лишь знания $\Phi(u,v)$.

Приведенные примеры говорят о достаточной естественности введенного понятия. Отметим, что для достаточно больших и близких значений N_1 и N_2 порядок величины L в классе случайных последовательностей может быть получен из (I) заменой N на N_1+N_2.

4. Поиск приближенных соответствий. Как использовать рассмотренные в п.3 понятия для отыскания в тексте "зашумленных" структурных элементов? Один из возможных подходов заключается в следующем.

Вычислим частотную характеристику текста l-го порядка, где l выберем таким, чтобы количество кратных l-грамм было не слишком велико. Информация о местах вхождения их в текст содержится в информационной ленте - битовой строке длиной N-l+1 символов, каждый разряд которой находится в одном из двух состояний: 0 -

кратная 1-грамма, I - единичная (см.,например, [I8]). Дополнительным просмотром (с хешированием) разобьем 1-граммы, помеченные кодом О, на классы эквивалентности, каждый из которых содержит только одинаковые 1-граммы. Каждому классу при этом просмотре ставится в соответствие список вхождений 1-грамм данного класса в текст.

Применим к каждому классу эквивалентных 1-грамм процедуру левостороннего и правостороннего р а с ш и р е н и й, которая осуществляет удлинение (влево и вправо) всевозможных пар 1-грамм из данного класса с одновременной проверкой (на каждом шаге удлинения) гипотезы о близости расширяемых 1-грамм. Не касаясь всевозможных оптимизационных идей, позволяющих сократить число попарных сравнений, отметим только, что удлиняемые участки на первом этапе целесообразно сравнивать в хемминговой метрике, допуская, что наиболее вероятными эволюционными операциями являлись замены символов. Для случайных последовательностей с $p(a_i)=1/n$ $(1 \le i \le n)$ математическое ожидание числа x совпавших символов при сравнении двух последовательностей длины N составляет $p \cdot N = E(x)$, а дисперсия $\sigma^2(x) = p(1-p) \cdot N$. Неравенство Чебышева дает при этом грубую, но достаточную для наших целей оценку вероятности отклонения от математического ожидания. Скажем, при n = 4 и N = 8 вероятность случайного совпадения пяти из восьми символов не превышает I/9, т.е. 2-3 несовпадения на 8 символов допустимы, чтобы принять гипотезу о близости с вероятностью ошибиться $\sim 0,I$.

Линейная процедура расширения двух 1-грамм не выявит сходства, если имели место вставка или устранение символов в каком-либо из расширяемых участков. В этом случае сравнение расширяемых участков (не обязательно равной длины) следует проводить в смысле МДП или совместного частотного спектра (меры (3),(5)),что требует бо́льших вычислительных затрат (трудоемкость вычисления МДП текстов u и v составляет $O(N_u \cdot N_v)$ [23], а $\Phi(u,v) - O(L \cdot (N_u + N_v))$ в наихудшем случае).

Правило останова для процедуры расширения - невыполнение гипотезы о близости расширяемых участков. Функционально значимыми можно считать расширения с количеством совпадающих элементов,не меньшим, чем это вытекает из (I).

Другой важный механизм получения приближенных соответствий - целенаправленная п е р е к о д и р о в к а элементов алфавита, переводящая близкие (но не совпадающие) фрагменты текста в тождес-

твенные, легко локализуемые затем как повторы в перекодированном тексте. Перекодировки такого типа определяются содержательной стороной задачи. Поиск их и реализация самого алгоритма перекодировки иногда весьма нетривиальны (например, в музыке [2], в естественном языке – переход от словоформ к канонической форме).

Простейший вариант перекодировки – агрегирование элементов алфавита, т.е. замена нескольких элементов алфавита одним, приводящая к уменьшению мощности алфавита (пример – переход от алфавита нуклеотидов ($n = 4$) в генетических текстах к двоичному алфавиту: пурины (А,Г), пиримидины (Ц,Т)). Другая возможность – отождествление не отдельных элементов алфавита, а целых кодовых комбинаций. Пример этому дает переход от нуклеотидного алфавита к алфавиту аминокислот ($n = 20$), при котором каждая тройка нуклеотидов заменяется одной аминокислотой. Так, последовательности $T_1 =$ = ТТ [Ц] ААТГГ [Т] ГЦ [А] и $T_2 =$ ТТ [Т] ААТГГ [Ц] ГЦ [Г], отличающиеся по 3-й, 9-й и 12-й позициям, после перекодировки станут неразличимыми: $T_1 = T_2 =$ Фен Асн Гли Ала, где Фен, Асн, Гли, Ала – элементы алфавита аминокислот.

Приведем более сложный пример. Предположим, что нас интересуют повторы, допускающие несовпадения по каждой третьей позиции (например, вида ЦА[Т]ГГ[Т] и ЦА[Г]ГГ[А] или [А]ГГ[Т]ТТ[Ц]ЦА и [Т]ГГ[А]ТТ[Г]ЦА). Введем $(n+1)$-й элемент алфавита z, представим для определенности, что N кратно 3, и запишем текст $T = a_1a_2a_3a_4\ldots a_N$ в трех формах: $T_1 = za_2a_3za_5a_6\ldots a_N$, $T_2 =$ $= a_1za_3a_4za_6\ldots a_N$, $T_3 = a_1a_2za_4a_5z\ldots z$. Нетрудно видеть, что задача будет решена, если мы образуем конкатенации $T_1 \backslash\!\!\!\!\sqcup T_2$, $T_1 \backslash\!\!\!\!\sqcup T_3$, $T_2 \backslash\!\!\!\!\sqcup T_3$, где разделитель $\backslash\!\!\!\!\sqcup \notin A$, и обнаружим в них все повторы длины 3 и выше, не содержащие разделитель. Таким образом, здесь перекодировка привела к увеличению как длины обрабатываемого текста, так и числа самих текстов.

Аналогичная ситуация возникает и при поиске таких закономерностей, как симметрии или инверсии (в генетических текстах). Они могут быть выявлены при помощи алгоритма вычисления совместного частотного спектра, где роль второго текста играет исходный текст, преобразованный соответствующим образом.

5. Проблема индуктивного синтеза текстов. В разделе I говорилось о двух подходах к выявлению структуры текста. Первый из них предполагал наличие почти исчерпывающей априорной информации об элементах структуры, которые задавались или точно, или в виде

некоторой обобщающей формулы.Другой подход - прямо противоположный - характеризовался полным отсутствием соответствующей априорной информации, и для выявления структуры текста предлагались некоторые стандартные заготовки. В данном разделе рассматривается промежуточная ситуация, когда априорная информация об элементах структуры представлена выборкой образцов по каждому классу элементов. Эта ситуация характерна для распознавания образов.

Каким образом использовать имеющуюся априорную информацию для выявления структурных элементов в незнакомом тексте? Укажем на две возможности, первая из которых связана с построением обобщающей формулы по выборке образцов (сведение к первому случаю), а вторая - с использованием непосредственно обучающей выборки.Более подробно рассмотрим первую из них.

Элементы обучающей выборки в принципе могут иметь разную длину и быть различным образом выравненными (сфазированными) один относительно другого. Желательно выравнивать последовательности по какому-либо характерному элементу или группе элементов, несущих одинаковую функциональную нагрузку. Таким элементом может быть корень в словах естественного языка, инициирующий кодон в выборке из рибосомных сайтов связывания в генетических приложениях и т.д. Фазировка целесообразна в ситуациях, когда предполагается, что другие функционально значимые зоны (а их может быть несколько) находятся у всех элементов выборки на примерно одинаковом расстоянии от точки фазировки. В этом случае в результате фазировки они оказываются друг под другом и могут быть выявлены статистическими средствами, описанными ниже.

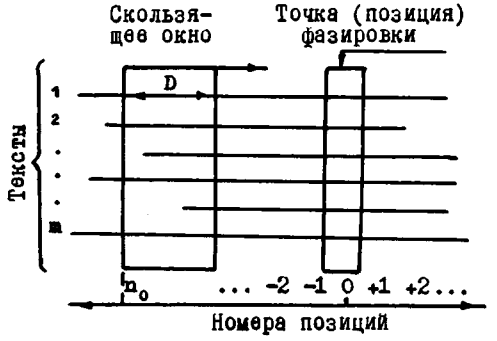

Рис.3.

Пусть тексты $T_1, T_2, T_3, \ldots, T_m$ обучающей выборки сфазированы указанным образом (см. позицию 0 на рис.3). Фиксируем окно анализа шириной в D символов. Положение окна задаем параметром n_0. Окно анализа может скользить вдоль текста слева направо со сдвигом на один (или несколько) символов. Для выделенного ок-

ном участка каждого текста (если он не имеет пустых позиций) вычисляется частотная характеристика 1-го порядка $\Phi_1(T_1)$, элементы (1-граммы) которой у п о р я д о ч и в а ю т с я по убыванию частоты встречаемости. Потенциально возможное число 1-грамм - $|A_1| = n^1$. Тем из них, которые отсутствуют в $\Phi_1(T_1)$, присваивается нулевая частота.

Мерой согласованности (сходства) текстов в пределах заданного окна анализа может служить коэффициент конкордации [28] 1-го порядка

$$W_1 = \frac{12\, V_1}{m^2(|A_1|^3 - |A_1|)}, \qquad (6)$$

где V_1 - сумма квадратов отклонений суммы рангов каждой 1-граммы (по всем m упорядочениям) от среднего значения суммы рангов одной 1-граммы, равного $\frac{1}{2} m(|A_1|+1)$. Для устранения неоднозначности ранжировки групп равночастотных 1-грамм используется процедура усреднения их рангов, что приводит к необходимости введения соответствующей поправки в (6).

Фиксируя l, D и изменяя n_0 с шагом I, получим кривую значений $W_1(n_0, D, m)$ вдоль длины текстов. Для статистики W_1 имеются хорошие аппроксимационные распределения, позволяющие нанести на график зависимости W_1 от n_0 пороговое значение $p(\alpha)$, соответствующее заданному уровню значимости α. Превышение статистикой $W_1(n_0, D, m)$ величины $p(\alpha)$ свидетельствует о статистической значимости полученной оценки. Естественно предполагать, что статистически значимым никам на кривой значений W_1 соответствуют функционально значимые зоны в текстах.

Параметр D рекомендуется изменять с шагом I от значения $D_{min} \geq 1$ до D_{max}, при котором происходит слияние всех локальных максимумов. Вариация параметра n_0 позволяет определить положение функционально значимых зон, а вариация параметра D - оценить их размер. При $l = D = 1$ эта методика допускает возможность получения текста-обобщения для элементов обучающей выборки, который, например, может иметь вид:

```
 АА ..... АГГГГАГ . АА . ТАТГА . ААА .. ТТАА ..
  |     |    ГА|А   |      |      Т      |А    |
  |     |      |    |      |      |      |     |
-20   -15    -10   -5      0     +5     +10   +15
```

Точками здесь обозначены позиции со значениями $W_1(n_0, 1, m) < p(\alpha)$. Наличие двух символов в одной позиции (например, А и Т в (+5)-й) означает, что здесь может встречаться как А, так и Т, причем А имеет бо́льшую вероятность. Отметим, что детерминированный подход к решению подобных задач развивается в [29].

Описанная методика не проходит, когда заранее не известны элементы, по которым можно осуществить выравнивание текстов, либо сильно варьируют расстояния между одинаковыми функциональными элементами в разных текстах. В этом случае можно не пытаться синтезировать текст-обобщение, а использовать непосредственно обучающую выборку для локализации структурных элементов в исследуемом тексте Т.

С этой целью вычисляется с о в м е с т н ы й частотный спектр текстов Т и $T_1 \frown T_2 \frown \ldots T_m$, где тексты T_i $(1 \le i \le m)$ принадлежат обучающей выборке. Из этого спектра выделяются все 1-граммы со значением 1, большим заданного порога, и определяются места вхождения их в текст Т. Каждое вхождение можно интерпрети - ровать точкой на целочисленном отрезке [1 + N-1+1]. Далее решается задача выделения сгущений точек на этом отрезке, для чего используются специально разработанные критерии. Выделенные сгустки точек можно интерпретировать как структурные элементы текста Т. Указанная схема обработки является обобщением метода покрытий, описанного в [30].

6. Заключение. Рассмотрена задача выявления структуры уникальных по своей природе символьных последовательностей. Структура характеризуется составом, количеством и расположением внутри текста отдельных специфических фрагментов, являющихся подпоследовательностями исходной последовательности. Анализируются три уровня задания априорной информации об элементах структуры - от почти полной определенности до полной неопределенности. Предлагаются стратегии действий во всех случаях, основанные на эффективных вычислительных процедурах и учитывающие возможность вариации элементов структуры. Многие из описанных подходов реализованы в рамках пакета прикладных программ СИМВОЛ [31].

Литература

I. СУХОТИН Б.В. Оптимизационные методы исследования языка. Автореф.дис. на соиск.учен.степени доктора филолог.наук. М., 1979 (МГУ).

2. ЗАРИПОВ Р.Х. Анализ и алгоритмизация мелодий с помощью частотных словарей музыкальных интонаций - ДАН СССР,1983, т. 268, № 2, с.303-306.

3. ТИМОФЕЕВА М.К. Индуктивная реконструкция грамматик флек - тивных языков. - В кн.: Методы обнаружения закономерностей с помощью ЭВМ (Вычислительные системы,вып.91). Новосибирск,1981,с.57-67.

4. LEMPEL A., ZIV J. On the complexity of finite sequences.- IEEE Trans.on Inf.Th.,1976, v.IT-22, N 1,p.75-81.

5. ГРИГОРЬЕВА А.Н. Оценки сложности и сокращение описаний изображения. Автореф. дис. на соиск.учен.степени канд. физ.-мат. наук. Л.,1982 (ЛГУ).

6. ГУСЕВ В.Д. Характеристики символьных последовательностей. - В кн.: Машинные методы обнаружения закономерностей (Вычислительные системы,вып.88). Новосибирск,1981, с.112-123.

7. ЗУБКОВ А.М., МИХАЙЛОВ В.Г. О повторениях S-цепочек в последовательности независимых величин. - Теория вероятностей и ее применения, 1979,т.XXIV,№ 2,с.267-279.

8. ЗУБКОВ А.М., МИХАЙЛОВ В.Г. Предельные распределения случайных величин, связанных с длинными повторениями в последовательности независимых испытаний. - Теория вероятностей и ее применения,1974, т.XIX, № 1, с.173-181.

9. New approaches for computer analysis of nucleic sequences /Samuel Karlin, Ghassan Ghandour, Friedemann Ost et al.-Proc. Natl. Acad. Sci. USA, 1983, v.80, N 18,p.5660-5664.

10. ГУСЕВ В.Д., КУЛИЧКОВ В.А., ТИТКОВА Т.Н. Анализ генетических текстов. I. 1-граммные характеристики. - В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы,вып. 83). Новосибирск,1980, с.11-33.

11. WEINER P. Linear pattern matching algorithms. - Conf. Record, IEEE 14th Annual Symposium on Switching and Automata Theory, 1973,p.1-11.

12. АХО А., ХОПКРОФТ Дж., УЛЬМАН Дж. Построение и анализ вычислительных алгоритмов. - М.: Мир, 1979.

13. McCREIGHT E.M. A space-economical suffix tree construction algorithm. - JACM, 1976, v.23, N 2,p.262-272.

14. RODEH M., PRATT V.R., EVEN S. Linear algorithm for data compression via string matching. - JACM, 1981, v.28, N 1,p.16-24.

15. BOYER R.S., MOORE J.S. A fast string searching algorithm. - CACM, 1977, v.20, N 10, p.762-772.

16. КНУТ Д. Искусство программирования для ЭВМ. Т.3. Сортировка и поиск. - М.: Мир, 1978.

17. KARP R.M., MILLER R.E., ROSENBERG A.L. Rapid identification of repeated pattern in strings, trees and arrays.- Proc.4th

Annual ACM Symposium on Theory of Computing. Denver, Colorado, 1972, p.125-136.

18. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. О задаче поиска повторяющихся отрезков текста. - В кн.: Вычислительные системы, вып.62. Ассоциативное кодирование. Новосибирск, 1975, с.49-71.

19. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. Отыскание статистических закономерностей текстов методом ассоциативного кодирования. - Там же, с.72-89.

20. СЛИСЕНКО А.О. Распознавание предиката вхождения в реальное время. - Л., 1977. - 24 с. (Препринт/ ЛОМИ: Р-7-77).

21. APOSTOLICO A., PREPARATA F.P. Optimal off-line detection of repetitions in a string. - Theoretical Computer Science, 1983, v.22, p.297-315.

22. ЛЕВЕНШТЕЙН В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов. - ДАН СССР, 1965, т.163, № 4, с. 845-848.

23. WAGNER R.A., FISCHER M.I. The string-to-string correction problem. - JACM, 1974, v.21, N 1, p.168-173.

24. CHVATAL V., SANKOFF D. Longest common subsequences of two random sequences. - J.Appl.Probability, 1975, v.12, p.306-315.

25. DEKEN I.G. Some limit results for longest common subsequences.- Discrete Mathematics, 1979, v.26, N 1, p.17-31.

26. KOHONEN T., REUHKALA E. A very fast associative method for the recognition and correction of misspelt words, based on redundent hash addressing. - Proc.of the fourth international joint conference on pattern recognition, 1978. Kyoto, Japan, p.807-809.

27. FINDLER N.V., Van LEEUWEN J. A family of similarity measures between two strings. - IEEE Trans. on Pattern Analysis and Machine Intelligence, 1979, v.PAMI-1, N 1, p.116-118.

28. КЕНДЭЛ М. Ранговые корреляции. - М.: Статистика, 1975. - 213 с.

29. ANGLUIN D. Finding patterns common to a set of strings.- Journal of computer and system sciences, 1980, v.21, p.46-62.

30. ТИТКОВА Т.Н. Выявление информативных признаков в задачах распознавания символьных последовательностей (на примере генетических текстов). - В кн.: Машинные методы обнаружения закономерностей (Вычислительные системы, вып.88). Новосибирск, 1981, с.124-127.

31. Пакет прикладных программ для анализа произвольных символьных последовательностей значительной длины (ППП СИМВОЛ)/ Высоцкая Г.С., Гусев В.Д., Косарев Ю.Г. и др. - В кн.: R-технология программирования: Тез.докл. I-й Всесоюз. конф. Ч.2. Опыт применения. Киев, 1983, с.48-50.

Поступила в ред.-изд.отд.
7 декабря 1983 года