А Н А Л И З Т Е К С Т О В И С И Г Н А Л О В
(Вычислительные системы)
1987 год Выпуск 123

УДК 519.237.8:519.764

АЛГОРИТМЫ ОБНАРУЖЕНИЯ ЗНАКОВ ПУНКТУАЦИИ В ГЕНЕТИЧЕСКИХ ТЕКСТАХ

В.Д.Гусев, В.А.Куличков, Н.А.Чужанова

В в е д е н и е

Под генетическим текстом будем понимать представление молекул нерегулярных полимеров (ДНК, РНК, белков) в виде последовательности мономеров (нуклеотидов или аминокислот). Наиболее длинные из полностью расшифрованных (секвенированных) к настоящему времени НК-молекул (геномов) содержат порядка 10^5 элементов. Анализ текстов такой длины без привлечения ЭВМ затруднителен.

Целью анализа на начальном этапе является установление важнейших структурных особенностей генома, как-то: кодирующих и некодирующих участков, числа генов, их локализации, знаков пунктуации, т.е. фрагментов, отвечающих за регуляцию основных генетических процессов – редупликации, транскрипции, трансляции. Последние, подобно знакам пунктуации в естественных языках, осуществляют иерархическое структурирование генетических текстов. Но если в естественных языках знаки пунктуации – это специально выделенные (а потому легко опознаваемые) элементы алфавита, то в генетических текстах – это фрагменты (длиной до нескольких десятков символов), составленные из тех же элементов, что и весь текст. Более того, знаки каждого типа сильно варьируют, что усложняет задачу их обнаружения.

Локализация знаков пунктуации – довольно трудоемкий процесс, включающий в себя проведение многочисленных химико-биологических экспериментов. Попытка решения этой задачи методами распознавания образов представляется актуальной по следующим соображениям: а) сужается область возможного поиска для экспериментаторов;

б) наряду с реально функционирующими знаками выделяются потенциально возможные претенденты, которые могут начать функционировать как знаки при изменении внешних условий; в) на этапе обучения в том или ином виде вырабатывается формальное определение знака, которое может оказаться полезным в ряде ситуаций (например, при синтезе новых молекул, оценке эффективности (силы) отдельных знаков).

Авторы (совместно с Т.Н.Титковой и Г.С.Высоцкой) в течение ряда лет занимались разработкой алгоритмов обнаружения знаков пунктуации*). Цель данной работы - систематизация и сопоставление (на одном и том же материале) разработанных алгоритмов и выработка общей методологии обнаружения знаков. Суть ее сводится: а) к учету специфики объектов распознавания (наиболее характерных свойств знаков); б) разработке систем описания, апеллирующих к фундаментальному для генетических текстов понятию повтора; в) к использованию коллектива простейших (пороговых и таксономических) решающих правил для принятия решения.

К настоящему времени сложились два основных подхода к обнаружению знаков пунктуации. Первый из них связан с построением "консенсуса" (или формулы знака) на основе анализа обучающей выборки [1-4]. Процедура распознавания сводится в этом случае к поиску в тексте всех фрагментов, удовлетворяющих консенсусу.

Второй подход связан с учетом значимости ("веса") каждого нуклеотида в пределах окна анализа, примерно равного длине знака [5-7]. Если суммарный вес всех нуклеотидов превышает заданный порог, принимается решение о наличии знака. Соответствующие веса и величина порога определяются в процессе обучения. Различные модификации этого подхода отличаются способами оценки значимости нуклеотидов и вариантами нормировки результирующей статистики. Разработанная нами версия фигурирует далее под названием "метод весовых функций".

Ни один из подходов не претендует на универсальность и имеет свою (подчас довольно узкую) область применения. Указанное обстоятельство мотивирует разработку новых алгоритмов, чему способствует также выявление экспериментаторами новых классов знаков пунк-

*) Первые результаты в этом направлении изложены в отчете Института математики СО АН СССР "Исследование возможностей 1-грамного представления текста для решения классификационных задач генетики", Новосибирск, 1980 г., 103 с.

туации со специфическими свойствами (таких, например, как "энхансеры" - усилители транскрипции).

Среди элементов новизны отметим разработку способов фазировки и алгоритмов обнаружения таксономического типа, а также применение в качестве систем описания общих подпоследовательностей и языка образов.

§I. Специфические особенности знаков пунктуации

I.I. Знаки не имеют формального определения. Представление о каждом классе знаков мы получаем в виде совокупности примеров (обучающей выборки), объем которой, как правило, невелик. На молекулярном уровне для каждого типа знаков существует свое опознающее "устройство" (РНК-полимераза для инициаторов транскрипции, рибосома для инициаторов трансляции и т.д.). Механизм опознавания в деталях неизвестен, поэтому информация о нем почти не используется при конструировании алгоритмов обнаружения. Выявляемые формальными методами "знаки" верифицируются лишь экспериментальным путем.

I.2. Полезная информация распределена по длине знака неравномерно. Обычно в знаке имеется несколько функционально значимых зон, разделенных малоинформативными (по состоянию наших знаний на сегодняшний день) фрагментами. Использование этих фрагментов (например, при вычислении весовых функций) может уменьшить надежность распознавания. Актуальной задачей поэтому является разработка формальных методов локализации функционально значимых зон.

Формула знака дает обобщенное представление о его информативных зонах. Вид формулы зависит от методики выявления значимости соответствующих позиций, объема обучающей выборки и способа фазирования (выравнивания) элементов обучающей выборки относительно друг друга. Приведем примеры "канонических" (без излишней детализации) формул для наиболее изученных классов знаков.

а) Рибосомный сайт связывания (инициатор трансляции у прокариотов) в первом приближении содержит две функционально значимые зоны (рис.I). С первой из них (GGAGG) в процессе трансляции комплементарно связывается 3'-конец I6S рибосомальной РНК, содержащий фрагмент AUUCCUCCACUA.

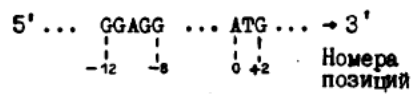

Рис.I

Вторая зона представлена инициирующим кодоном (ATG) и его ближайшим окружением (преимущественно справа). В позициях, помеченных точками, закономерность выражена не столь явно или не проглядывает вовсе. Более слабые функциональные зоны наблюдаются как левее (-I2)-й позиции, так и правее (+2)-й, т.е. в кодирующей области. Их наличие свидетельствует о том, что инициация трансляции обусловлена не только связыванием 3'-конца I6S рибосомальной РНК с первой зоной, но и рядом побочных факторов, трактовка которых не всегда ясна и выходит за рамки данной работы.

б) Каноническая форма промотора E.coli (инициатора транскрипции) представлена на рис.2. Нулевая позиция соответствует нача-

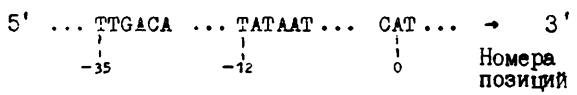

Рис. 2

лу транскрипции. Функционально значимые зоны ("боксы") связывают с фамилиями их первооткрывателей. Зона начала транскрипции (CAT) варьирует весьма сильно и не по всем методикам проходит как функционально значимая.

в) Терминатор транскрипции обычно содержит поли-Т участок, которому предшествуют две G,C-богатые (и сильно варьирующие) области, потенциально способные к образованию шпилечной структуры в РНК-транскрипте. Анализ выборки из 30 терминаторов [7], выравненных по общим 1-граммам (см.ниже), приводит к формуле знака, представленной на рис.3. Закономерности здесь выражены менее ярко, чем в случаях "а" и "б", поэтому степень детализации выше.

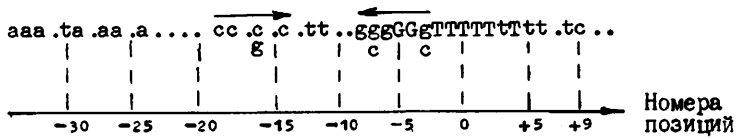

Рис. 3

Большими буквами помечены элементы, превалирующие в данной позиции (с относительной частотой $f \geq 0,6$), малыми - элементы с $0,4 \leq f < 0,6$. Во второй строке указаны альтернативные символы. Участки, потенциально способные к формированию шпильки, помечены стрелками. Окончание транскрипции приходится примерно на (+7)-ю и (+8)-ю позиции.

I.3. Обучающие выборки по каждому типу знаков весьма разнородны. Даже у одного организма знаки фиксированного типа варьируются весьма сильно. К примеру, промоторы бактериофага FD, анализировавшиеся в указанном выше отчете, имеют более слабые связи друг с другом (на уровне общих 1-грамм), чем с промоторами фага T7. Значения весовой функции у рибосомных сайтов связывания имеют бо́льшую вариацию, чем у "не сайтов". Практически в каждой выборке знаков встречаются аномальные элементы, имеющие мало общего с приведенными выше каноническими представлениями. На рис.4 показаны примеры аномальных знаков пунктуации: а) рибосомный сайт связывания гена белка оболочки фага Qβ [4]; б) промотор E.coli gal P1 [3]; в) терминатор транскрипции E rho att [7]. Позиции, помеченные нулем, указывают начало (или окончание) соответствующего процесса (трансляции, транскрипции). Функционально значимые зоны подчеркнуты. Фрагменты, выделенные скобками, комментируются ниже.

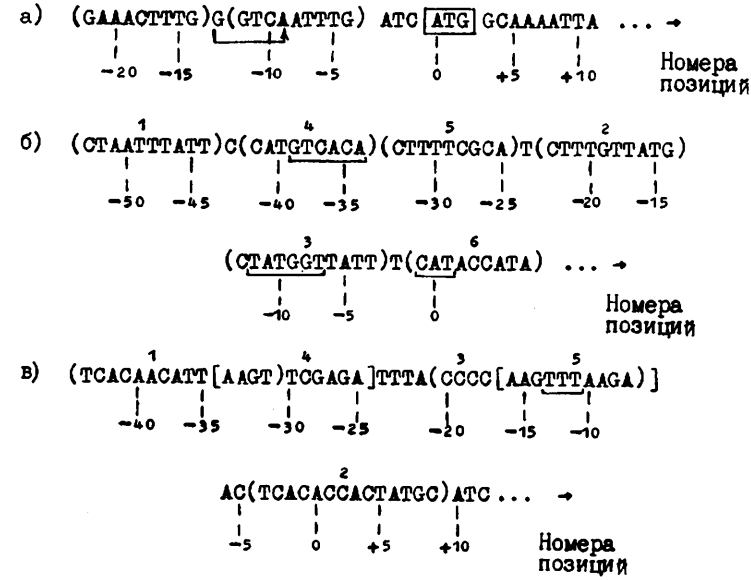

Рис. 4

I.4. Существенную роль в образовании и эволюции знаков играют одиночные замены, короткие вставки и делеции, а также дупликации.

Замены приводят к искажению функционально значимых зон и влияют на эффективность знака. Гомология функционально значимых зон в реальных знаках с их каноническими представлениями может составлять порядка 50 (и даже менее) процентов.

Длины допустимых вставок и делеций - одного порядка с вариацией расстояний между значимыми зонами. Последние довольно жестко закреплены за своими позициями (сдвиги влево или вправо обычно не превышают двух-трех позиций). В противном случае эффективность знака падает или он вовсе перестает функционировать.

Особый интерес (в плане исследования возможных механизмов возникновения знаков) представляют дупликации фрагментов, сопоставимых по размеру и нуклеотидному составу с функционально значимыми зонами в знаках. Чем выше кратность повтора, тем больше вероятность возникновения знака. Многократное тиражирование фрагмента способствует тому, что: I) в любой из его копий в результате одиночных замен может усилиться сходство с функционально значимой зоной; 2) произойдет сближение одной функциональной зоны с другой, и они вместе начнут функционировать как знак. Поскольку разные функциональные зоны в пределах одного знака не всегда являются абсолютно непохожими (к примеру, фрагменты (TTGACA) и (TATAAT) в промоторах уже обладают 30%-й гомологией), то даже однократное повторение фрагмента с последующими благоприятными заменами может обусловить возникновение знака. Вновь возвращаясь к рис.4, отметим в качестве примера, что дупликации (фрагменты, заключенные в скобки) сыграли, по-видимому, значительную роль в формировании изображенных там знаков. В первую очередь это касается промотора (случай "б"), целиком составленного из двух групп повторов: в одну входят фрагменты I,2,3, ответственные за образование бокса Прибнова, в другую - фрагменты 4,5,6 (бокс Гильберта и начало транскрипции).

I.5. Немаловажное влияние на функционирование знаков оказывают и такие типы повторов, как инверсии, палиндромы, симметрии. Некоторые из них (например, инвертированные повторы в терминаторах) являются составной частью самого знака; другие (палиндромы, симметрии) влияют на эффективность протекания процесса (например, замедляют транскрипцию).

Указанные выше свойства знаков служат в определенной степени обоснованием излагаемой ниже методики обнаружения. В частности, неравномерность распределения информации по длине знака (п.I.2) обусловила разработку алгоритма выявления функционально значимых зон. Разнородность обучающих выборок (п.I.3) стимулировала интерес к алгоритмам таксономического типа. Язык образов, используемый для описания таксонов, перспективен в плане развития понятия консенсуса. Повторяемость некоторых свойств у знаков обучающей выборки, а также отдельных фрагментов внутри самих знаков (п.I.4) объясняет появление таких систем описаний, как совместный частотный спектр группы текстов, спектр общих подпоследовательностей и т.д.

§2. Системы описания текстов в терминах "повторов"

Определим широко используемые далее понятия l-граммы и подпоследовательности. Пусть u - цепочка символов, $|u|$ - ее длина, $u[i]$ - i-й элемент цепочки ($1 \le i \le |u|$), $u[i:j]$ - фрагмент цепочки с i-го по j-й символ включительно. Назовем цепочку u подпоследовательностью цепочки v, если существует монотонно возрастающая последовательность целых $r_1, r_2, \ldots, r_{|u|}$ такая, что $u[i] = v[r_i]$ для $1 \le i \le |u|$. Связную подпоследовательность длины l будем называть l-граммой. Множество всех l-грамм цепочки является подмножеством множества подпоследовательностей длины l. Различие между ними хорошо просматривается на комбинаторном уровне: число элементов первого множества равно $|u| - l + 1$, тогда как второго - $C^l_{|u|}$ (число сочетаний из $|u|$ по l). Среди элементов каждого множества могут быть повторяющиеся. Поскольку число подпоследовательностей в тексте большой длины велико, их целесообразно использовать лишь для описания очень коротких фрагментов (например, функциональных зон в знаках пунктуации).

Связь двух текстов T_1 и T_2 характеризуется множеством l-грамм (или подпоследовательностей), общих для обоих текстов. Применительно к группе текстов $\overline{T} = (T_1, T_2, \ldots, T_m)$ можно рассматривать множество l-грамм (или подпоследовательностей), общих, как минимум, для t текстов группы ($2 \le t \le m$).

Для представления функционально значимых зон знаков пунктуации использовались следующие характеристики:

$\Phi_1(T)$ - частотная характеристика 1-го порядка текста T. Это совокупность всех 1-грамм текста (с учетом их кратности), упорядоченная по убыванию частот встречаемости;

$U_i(T)$ - аналогичная характеристика для подпоследовательностей длины 1 (спектр подпоследовательностей длины 1);

$\Phi_{1,t}(\overline{T})$ - совместная частотная характеристика группы текстов $\overline{T} = (T_1, T_2, \ldots, T_m)$. Это совокупность общих, как минимум, для t текстов 1-грамм с учетом их кратностей и мест вхождений;

$U_{1,t}(\overline{T})$ - аналогичная характеристика, построенная на общих (как минимум, для t текстов) подпоследовательностях длины 1 (спектр общих подпоследовательностей длины 1);

$p(\overline{T})$ - образ, порождающий группу текстов $\overline{T}$.

Первые четыре характеристики описаны в [8] (с некоторыми деталями, касающимися реализации). Остановимся подробнее на последней (пятой) характеристике, тесно связанной с предыдущей, т.е. с $U_{1,t}(\overline{T})$. Если положить t = m, т.е. использовать лишь подпоследовательности, общие для всех текстов выборки, и наложить некоторые ограничения на фрагменты текстов, не вошедшие в общие подпоследовательности, получим описание, которое в теории формальных грамматик ассоциируется с термином "образ" [9]. Интерес распознавателей к данной системе описания обусловлен возможностью индуктивного восстановления грамматики для различных классов языков образов по достаточно представительному (в пределе совпадающему с языком) множеству положительных примеров (в нашем случае - выборке из знаков пунктуации).

Пусть Σ - конечный алфавит, содержащий хотя бы два символа, Σ^* - множество всех слов в алфавите Σ , $X = \{x_1, x_2, \ldots\}$ - счетное множество символов. Элементы алфавита Σ будем называть константными символами, а множества X - переменными ($\Sigma \cap X = \emptyset$). Образ - это любая конечная цепочка в алфавите $\Sigma \cup X$. Количество различных символов из X, встретившихся в образе p, назовем числом переменных (например, если $\Sigma = \{0,1\}$, то $p = x_1 10 x_1 x_2$ - образ с двумя переменными). Язык образа p есть множество слов L(p) , получаемых подстановкой непустых цепочек $s_i \in \Sigma^*$ вместо каждого вхождения x_i в p (например, если $\Sigma = \{0,1,2,3\}$ и $p = 33x_1 2x_2 x_1$, то L(p) содержит слова 330210, 331120 11 и т.п., но не содержит слов 331200, 332 и т.п.).

В приложениях интерес представляют языки, минимальные в том смысле, что не существует других языков, включающих конечную вы -

борку $S \subset \Sigma^*$ и являющихся собственным подмножеством минимального языка. В [10] показано, что проблема построения образа, порождающего минимальный язык (проблема MINL), NP-полна в общем случае. Можно выделить, однако, некоторый вариант проблемы - 1-MINL, сводящийся к поиску образа максимальной длины. Существуют нетривиальные подклассы языков образов, для которых проблема 1-MINL имеет полиномиальную сложность, а именно:

- регулярные языки, порождаемые образом с k переменными, каждая из которых встречается в нем только один раз [11] (сложность $O(m \cdot n^2)$, где m - число цепочек в S, n - длина самой короткой цепочки из S);

- расширенные регулярные, т.е. регулярные, но допускающие пустые подстановки вместо переменных [11] (сложность $O(m \cdot n^4)$);

- языки образов с одной переменной [10] (сложность $O(m \cdot n^3 \cdot \log n)$);

- языки образов с двумя переменными инверсионного типа [12] (сложность $O(m \cdot n^3 \log n)$).

§3. Выявление информативных зон

Метод выявления границ знаков и информативных зон [13] применим к классам знаков, для которых: а) синонимичным функциональным зонам соответствуют "близкие" цепочки символов; б) вектор расстояний $(d_1, d_2, \ldots)$ между последовательно расположенными функциональными зонами $(z_1, z_2, z_3, \ldots)$ мало варьирует при переходе от знака к знаку, т.е. вставки и делеции немногочисленны и имеют малую длину.

При выполнении условия "б" знаки обучающей выборки могут быть сфазированы (выравнены путем сдвигов) таким образом, чтобы синонимичные зоны оказались расположенными друг под другом. Выявление этих зон проводится путем сканирования выборки окном переменной длины и оценки меры согласованности (значимости) фрагментов разных знаков в пределах окна. Мерой согласованности при выполнении условия "а" может служить значение коэффициента конкордации 1-го порядка $W_1(n_0, D, m)$, где n_0 - параметр, задающий положение окна, D - ширина окна, m - число непустых фрагментов в пределах окна, 1 - порядок частотных характеристик, которыми описываются фрагменты, попавшие в окно (см. [13]).

Вариация параметра n_0 позволяет определить положение функционально значимых зон, т.е. фрагментов, для которых $W_1(n_0, D, m) > p(\alpha)$, где $p(\alpha)$ - пороговое значение статистики W_1, соответст-

вующее заданному уровню значимости α. Вариация параметра D (при фиксированных значениях l, n_0 и m) позволяет оценить размер зоны, т.е. величину $D^* = \arg\max W_1(n_0, D, m)$.

Несколько подробнее остановимся на процедуре фазирования знаков обучающей выборки, не рассматривавшейся нами в [13]. Фазирование - желательный (а иногда необходимый) этап предобработки во многих алгоритмах обнаружения знаков (например, в методе весовых функций). В отдельных случаях фазировка бывает очевидной (например, по инициирующему кодону в рибосомных сайтах связывания). Не всегда, однако, начало или окончание процесса локализуются точно или однозначно. В этих случаях целесообразно использовать алгоритм фазирования. В общем случае задача фазирования связана с отысканием максимально длинной общей подпоследовательности группы текстов. Эта задача относится к классу NP-полных проблем. На практике удовлетворительное решение часто дает разработанный нами алгоритм фазировки, основанный на выявлении общих l-грамм в знаках обучающей выборки.

Шаг I. Вычисляем набор совместных частотных характеристик $\{\Phi_{l^*,2}(S), \Phi_{l^*+1,2}(S), \ldots, \Phi_{L,2}(S)\}$, где $S = (T_1, T_2, \ldots, T_m)$ - обучающая выборка, l^* - пороговое значение, задаваемое обычно из вероятностных соображений, L - максимум длин общих l-грамм. Выбираем начальное значение $l = L$.

Шаг 2. Просматриваем поочередно все l-граммы из $\Phi_{l,2}(S)$. Знаки обучающей выборки, содержащие l-грамму $x \in \Phi_{l,2}(S)$, объединяем в таксон K(x) и выравниваем по указанной l-грамме. Знаки, не имеющие общих фрагментов такой длины с другими текстами, рассматриваем как одноэлементные таксоны.

Шаг 3. Анализируем попарно полученные таксоны. Если пересечение двух таксонов непусто, объединяем их и выравниваем одну группу текстов относительно другой по одному из общих текстов. Процесс укрупнения таксонов продолжается до тех пор, пока не получим один таксон или множество попарно-непересекающихся таксонов.

При реализации шага 3 (а иногда и шага 2) возможны конфликтные ситуации, когда допустимы (с одинаковыми основаниями) различные варианты выравнивания. Для разрешения конфликтов используются дополнительные ограничения (например, на величину допустимого сдвига) и вспомогательные критерии, характеризующие степень согласованности текстов при каждом варианте выравнивания.

Шаг 4. Если по завершении шагов I–3 имеем более одного таксона, а $l > l^*$, уменьшаем значение l на единицу и повторяем шаги 2 и 3. Если $l = l^*$, итерации прекращаются, а оставшиеся несфазированными (аномальные) тексты анализируются отдельно (например, с помощью аппарата (l,k)-повторов [I4] или коэффициента конкордации, позволяющего определить оптимальную в определенном смысле величину сдвига аномального знака относительно группы уже сфазированных знаков).

Метод проверялся на выборке из 30 терминаторов транскрипции, сфазированной авторами работы [7] по точкам окончания транскрипции. Мы заново сфазировали терминаторы с помощью вышеописанного алгоритма. Отметим следующие положительные аспекты применения алгоритма фазирования:

а) понижается уровень требуемых априорных сведений о знаке (в частности, о точках инициации и терминации);

б) 90% всех терминаторов фазируются по достаточно длинным общим l-граммам ($l \geq 8$). Длина связей коррелирует с "силой" знака. Аномальные по этому параметру терминаторы (E rho att и E lac I), которые фазируются при минимальных значениях l (6 и 7 соответственно), классифицируются как слабые и по значению весовой функции [7];

в) мера согласованности текстов в пределах основной (поли-Т) области повышается с 0,4 до 0,6 (при нормировке k1);

г) устраняются некоторые нелогичности в исходном выравнивании. Так, терминаторы E leu att и S leu att, имеющие общую 2I-грамму, сфазированы в [7] со сдвигом на 3 позиции между вхождениями этой l-граммы в соответствующие тексты. Более того, в терминаторе S leu att точно не установлена точка окончания транскрипции, хотя она естественным образом прогнозируется по терминатору E leu att, так как лежит внутри общей 2I-граммы.

§4. Алгоритмы таксономии знаков

Объекты обучающей выборки достаточно разнородны даже в пределах информативных зон. Попытка описать их единой закономерностью (например, на языке образов) приводит к формулам вида $p(S) = x$ – "все возможно". Этим и мотивируется разработка алгоритмов обнаружения таксономического типа.

Известны две основные схемы таксономии: "снизу вверх", когда происходит последовательное укрупнение таксонов, как в алгоритме

фазирования, и "сверху вниз" – последовательное дробление. Мы использовали первую схему с коэффициентом конкордации в качестве меры близости.

Шаг 0. Задаем параметры: n_0 – положение информативной зоны, по которой осуществляется таксономия; D – размер зоны; l – порядок частотной характеристики (обычно l = 2 или 3); ε_T – пороговое значение меры близости текстов, объединяемых в таксон. Полагаем, что каждый текст обучающей выборки, представленный своей информативной зоной, является одноэлементным таксоном.

Шаг I. Для каждой информативной зоны $T_i[n_0 : n_0+D-1], 1 \le i \le m$, вычисляем характеристику $\Phi_l(T_i)$, элементы которой ранжируем по убыванию частоты встречаемости. Однократные l-граммы могут быть дополнительно ранжированы в соответствии с порядком их следования в тексте (альтернатива процедуре усреднения рангов в [I5]).

Шаг 2. Вычисляем значения мер близостей для всех пар таксонов. Под мерой близости таксонов X и Y понимаем величину коэффициента конкордации усредненных частотных характеристик каждого таксона ($\Phi_l(X)$ и $\Phi_l(Y)$).

Шаг 3. Если текущее число таксонов $d > 1$ и

$$\max_{i,j} W_l(X_i, X_j) \ge \varepsilon_T, \quad 1 \le i,j \le d, \quad i \ne j, \qquad (1)$$

где $W_l(X_i, X_j)$ – мера близости между таксонами X_i и X_j, то объединяем два таксона с максимальным значением меры близости. Затем возвращаемся к шагу 2 и корректируем ту часть матрицы сходства, которая была связана с объединяемыми таксонами. Размерность матрицы уменьшается на единицу.

Итерации заканчиваются при d = 1 или невыполнении условия (I). Полученное разбиение не зависит от порядка предъявления текстов и характеризуется тем, что мера близости элементов каждого таксона (если он не одноэлементный) не ниже ε_T.

Описанную схему таксономии можно использовать и с другими мерами близости, в частности, основанными на общих подпоследовательностях (см. характеристику $U_{l,t}(\overline{T})$ из §2). Пусть $S = (T_1, T_2, \ldots, T_m)$ – совокупность фрагментов длины D, описывающих информативную зону в знаках обучающей выборки; $B_{l,t}(S)$ – число элементов множества $U_{l,t}(S)$ (параметр t выбирается на этапе обучения). Сопоставим фрагменту x вектор F(x), элемент $F_j(x)$ которого есть число вхождений j-й общей подпоследовательности из $U_{l,t}(S)$

в x (напомним, что во фрагменте длины D содержится C_D^l подпоследовательностей длины l , среди которых могут быть повторяющиеся по составу и порядку букв). Меру близости фрагментов x и y по множеству $U_{l,t}(S)$ можно определить в виде

$$\beta_U(x,y) = \Big(\sum_{j=1}^{B_{l,t}(S)} \min(F_j(x),F_j(y)) \Big)/C_D^l \ . \qquad (2)$$

Аналогичным образом вводится мера близости и между двумя группами фрагментов. Минимум определяется по всем фрагментам обеих групп. Каждый таксон описывается списком общих подпоследовательностей, внесших ненулевой вклад в значение меры близости на этапе формирования таксона. Одноэлементные таксоны представлены теми из своих подпоследовательностей, которые вошли в $U_{l,t}(S)$.

Задача вычисления $U_{l,t}(S)$ может иметь экспоненциальную сложность. В нашем случае отметим два облегчающих фактора: а) размер информативной зоны невелик (D ~ 6–10 символов); б) обучающая выборка довольно представительна (m ~ 10^2 знаков). Добавление новых знаков эквивалентно введению дополнительных ограничений на возможность существования длинных общих подпоследовательностей. Идея алгоритма состоит в направленной генерации возможных подпоследовательностей, начиная с l = 1 - в проверке наличия их в каждом из текстов выборки и отсечении (вместе с продолжениями) тех из них, которые встречаются менее чем в t текстах.

§5. Алгоритмы обнаружения знаков

В §2 рассматривались системы описания текстов на языке l-грамм, общих подпоследовательностей и образов (i = 1,2,3 соответственно). При синтезе алгоритмов обнаружения использовались три типа решающих правил: пороговые, таксономические и грамматические (j = 1,2,3 соответственно). Различные (осмысленные) комбинации перечисленных систем описания и решающих правил образуют совокупность алгоритмов распознавания A_{ij} (i,j = 1,2,3). Конкретизируем вид решающих правил в отдельных алгоритмах.

В алгоритме A_{11} с каждым текстом связывается статистика (весовая функция)

$$I_l^{11}(T) = \sum_k f(T[k: k+l-1]) - I_0 \ , \qquad (3)$$

где T[k: k+l-1]= $x_l(k)$ – l-грамма, начинающаяся в k-й позиции

текста T ; $f(x_1(k))$ - относительная частота встречаемости 1-граммы x_1 в k-й позиции предварительно сфазированных знаков обучающей выборки ("вес" 1-граммы); I_0 - добавка, центрирующая относительно нуля статистику (3) для "не знаков". Суммирование по k проводится в пределах информативных зон, выявленных на этапе предобработки. Статистика (3) характеризует интегрально близость текста T к знакам обучающей выборки по 1-граммному составу.

Решение о принадлежности текста T к классу знаков (Z) или "не знаков" ($\bar{Z}$) имеет вид:

$$\text{если } I_1^{11}(T) \geq \Pi_1 \text{ , то } T \in Z, \text{ иначе } T \in \bar{Z}. \qquad (4)$$

Параметр 1 и порог Π_1 выбираются на основе анализа обучающих выборок из условия минимума суммы ошибок первого и второго рода. Под ошибкой первого рода понимается отождествление знака с "не знаком", а под ошибкой второго рода - отождествление "не знака" со знаком.

В алгоритме A_{12}, реализующем решающее правило с 1-граммным описанием, выборка знаков разбита на совокупность непересекающихся таксонов ($S = \bigcup_k S_k, 1 \leq k \leq K$). Мера близости текста T с таксоном S_k - $\beta_1^{12}(T, S_k)$ есть коэффициент конкордации двух упорядочений: частотной характеристики реализации T и усредненной частотной характеристики таксона S_k. Решающее правило имеет вид:

$$\text{если } \max_k \beta_1^{12}(T, S_k) \geq \varepsilon_1 \text{ , то } T \in Z, \text{ иначе } T \in \bar{Z} . \qquad (5)$$

Параметр 1 и порог ε_1 выбираются по той же схеме, что и в предыдущем случае (ε_1 и ε_T (см. (I)) могут не совпадать).

В алгоритме A_{21} строится основанная на общих подпоследовательностях весовая функция

$$I_1^{21}(T) = \Big(\sum_{i=1}^{C_D^1} k_i \Big) / m \ , \qquad (6)$$

где суммирование ведется по всем подпоследовательностям длины 1, входящим в состав информативной (быть может, условно) зоны текста T; D - длина зоны, k_i - число знаков обучающей выборки, содержащих i-ю подпоследовательность текста T в своей информативной зоне, m - объем обучающей выборки. Решающее правило имеет вид, аналогичный (4).

В алгоритме A_{22} (таксономия по общим подпоследовательностям) мера близости между реализацией T и таксоном S_k определяется в виде

$$\beta^{22}(T,S_k) = \Big(\sum_{j=1}^{B_{l,t}(S)} \min(F_j(T),F_j(S_k)) \Big) / C_D^l ,$$

где l и D имеют тот же смысл, что и в (6), суммирование проводится по всем общим подпоследовательностям из $U_{l,t}(S)$, $F_j(T)$ - число вхождений j-й общей подпоследовательности в информативную зону текста T, $F_j(S_k) = \min\limits_{s \in S_k} F_j(s)$. Параметры t и l определяются в процессе обучения. Решающее правило имеет вид, аналогичный (5).

В алгоритме A_{33} (правило грамматического типа, основанное на языке образов) проверяется принадлежность реализации T языку $L(p_k)$, где p_k - образ, описывающий таксон S_k.

Правило имеет вид:

если $(\exists k)(T \in l(p_k))$, то $T \in Z$, иначе $T \in \bar{Z}$.

§6. Сравнительный анализ алгоритмов

Экспериментальная проверка и сравнительный анализ алгоритмов проводились на задаче обнаружения прокариотических рибосомных сайтов связывания. Обучающая выборка S была представлена 86 реализациями фаговых (φX174, G4, FD, MS2, R17, λ, T7, Qβ) и бактериальных (преимущественно E.coli) сайтов (большая часть их приведена в [4]), сфазированных по инициирующему кодону (ATG или GTG). Выборка $\bar{S}$ ("не сайты") была составлена из фрагментов тех же фагов, содержащих потенциальный инициирующий кодон, не находящийся в фазе с рамкой считывания. Объемы выборок S и $\bar{S}$ и длины реализаций совпадали.

В алгоритме A_{11} суммировались веса с (-15)-й позиции по (+13)-ю. Параметр l менялся от 1 до 4. В остальных алгоритмах использовалась лишь информативная зона, расположенная в позициях с (-13)-й по (-7)-ю. Поскольку размер зоны слишком мал для принятия надежного решения, привлекались дополнительные признаки: а) наличие инициирующего кодона (ATG или GTG) в нулевой позиции; б) отсутствие среди первых 10 кодонов, находящихся в фазе с инициирующим, терминальных и кодирующих цистеин. Указанные закономерности выполнялись для всех сайтов из S. При контрольном распознавании они помогали сразу отсеять до 30% претендентов.

Длина общих подпоследовательностей в алгоритмах A_{21} и A_{22} выбиралась равной 5 (D = 7). Меньшая длина дает слишком грубые результаты, большая - делает алгоритм близким к дешифратору. Порог отбора общих подпоследовательностей по частоте встречаемости определялся в результате сопоставления характеристик $U_{1,2}(S)$ и $U_{1,2}(\bar{S})$. Самая высокочастотная общая подпоследовательность из S (AAGGA) встретилась 2I раз. Для "не сайтов" этот параметр равен 5. Таким образом,порог t целесообразно выбрать равным 4 или 5.В первом случае (t = 4) три сайта из S (MS2LYS,ECTL22 и LC2) не покрываются общими подпоследовательностями из $U_{1,t}(S)$, т.е. при обучении допускается ошибка первого рода,равная 3/86. Во втором случае к указанным сайтам добавляются еще три (φXI74C,G4C и FD7).

Порог ε_T в алгоритме A_{12} задавался равным 0,75. При больших значениях ε_T выделяется много мелких таксонов и разбиение неустойчиво к изменению состава обучающей выборки. При $\varepsilon_T < 0{,}75$ получается слишком грубое разбиение. Порог ε_1 в (5), как правило,превосходил ε_T (например, при l = 2 $\varepsilon_1 = 0{,}84$), т.е. одноэлементные таксоны классифицировались как ошибки первого рода (мера близости их с любым сайтом из S меньше ε_T и уж тем более - ε_1).Однако отбрасывание единичных таксонов существенно уменьшает ошибку второго рода, за счет чего снижается и суммарная ошибка.

Были проведены две серии экспериментов. В первой обучение проводилось на полных выборках S и $\bar{S}$. На контроль предъявлялись они же и все ATG- и GTG-содержащие фрагменты геномов φXI74, G4, FD и MS2 (в количестве свыше 600). Поскольку подавляющая часть последних - "не сайты", мы уже имели возможность получить из данной серии экспериментов оценку ошибки второго рода.

Вторая серия экспериментов проводилась в режиме "скользящего контроля" и служила для оценки ошибки первого рода. Из S удалялись сразу группы сайтов, соответствующие одному из геномов (φXI74, G4 , FD , MS2). После обучения на контроль предъявлялся геном, сайты которого были удалены. Устранение из обучения сразу группы сайтов ставило нас в более жесткие условия по сравнению с традиционной схемой "скользящего контроля", когда удаляется по одному объекту, и, возможно, привело к некоторому завышению числа ошибок первого рода.

Результаты обоих экспериментов сведены в табл.I. В экспериментах с геномами ошибка второго рода разделена на две части: фс→с (число участков, находящихся в фазе с рамкой считывания и

Таблица I

Результаты экспериментов по обнаружению рибосомных сайтов с использованием алгоритмов $A_{11}(D=29,l=2)$, $A_{12}(D=7,l=2)$, A_{21}, $A_{22}(D=7,l=5)$ и A_{33} $(D=7,l=3)$

Тип эксперимента	Текст	Число фрагментов	Количество ошибок в экспериментах													
			A_{11}			A_{12}			A_{21}			A_{22}			A_{33}	
			с→нс	фс→с	нс→с	с→нс	фс→с	нс→с	с→нс	фс→с	нс→с	с→нс	фс→с	нс→с	с→нс	нс→с
Полная обучающая выборка	Обучающая выборка	86с; 86нс	2	-	9	8	-	7	5	-	I2	3	-	I3	9	-
	φXI74	IIс; I95нс	0	I2	I4	0	7	I7	0	6	I5	0	3	I2	0	I3
	G4	IIс; I76нс	0	I2	7	0	7	II	0	9	I2	0	5	I0	I	4
	FD	I0с; I75нс	0	4	7	2	3	2	I	5	7	0	0	3	3	8
	MS2	4с; I09нс	0	2	3	I	6	4	I	6	3	I	6	7	I	6
Скользящий контроль	φXI74	IIс; I95нс	I	8	I3	3	7	I0	I	4	I5	2	3	I0	2	29
	G4	IIс; I76нс	0	I3	I0	I	II	9	I	I0	II	2	6	I3	I	I4
	FD	I0с; I75нс	5	2	2	6	3	I	2	4	7	5	0	2	5	8
	MS2	4с; I09нс	I	I	I	I	6	4	I	6	3	I	6	7	2	5
Суммарная ошибка по каждому A_{ij} на контроле			57			62			65			57			76	

распознанных как сайты) и нс→с (число участков, не находящихся в фазе с рамкой считывания и распознанных как сайты). Прочерки стоят там, где соответствующие эксперименты не проводились (например, в алгоритме A_{33} для обучения использовались лишь "положительные" примеры - только сайты).

По результатам проведенных экспериментов можно сделать следующие выводы.

I. Средняя (по всем алгоритмам) ошибка первого рода (эксперимент "скользящий контроль") составляет (20-25%), а ошибка второго рода не превышает (8-I0)%. С учетом того, что в большинстве случаев распознавание проводилось лишь по одной и довольно короткой (7 символов) информативной зоне, результаты следует признать обнадеживающими.

Сопоставление с опубликованными алгоритмами затруднено, поскольку большинство результатов относится к распознаванию обучающих выборок. Контроль носит характер иллюстрации методики на одном-двух новых примерах. Для контрольных выборок большего объема [6] факт неустойчивости решающего правила, дающего близкую к нулевой ошибку на обучении, становится очевиден (ошибка первого рода - порядка 40% и выше). В этом плане использование процедур типа "скользящий контроль" носит принципиальный характер.

Превалирование ошибки первого рода объясняется тем, что выборка из сайтов в пространстве используемых нами статистик выглядит более размытой, чем из "не сайтов". Если приписать ошибкам первого рода больший вес, число их может быть сокращено,но увеличится число ошибок второго рода.

2. Лучшие результаты на контроле (по суммарной ошибке) получены с помощью алгоритма A_{11}, использующего для принятия решения самую широкую зону (D = 29). Принципиальным является выбор параметра l . Чем больше l , тем лучше разделяются обучающие выборки. При l = 4 имеем всего две ошибки типа "нс→с" на обучении. Однако устойчивость решающего правила падает с увеличением l , что проявляется в большой вариации порогов $П_1$ при изменении объема или состава обучающей выборки и росте числа ошибок первого рода на контроле. Диапазон значений l = 2,3 является предпочтительным. "Веса", вырабатываемые по итогам обучения для каждой l-граммы в каждой позиции, являются аналогами оценок коэффициентов линейной функции в алгоритмах перцептронного типа [6], но вычисляются они гораздо проще.

3. Результаты контрольного обнаружения с помощью алгоритма A_{22} по суммарной ошибке не хуже, чем у A_{11} (при окне анализа, меньшем в 4 раза), но несколько выше ошибка первого рода. Смысл использования общих подпоследовательностей - добиться относительной устойчивости к коротким вставкам, делециям и одиночным заменам (ценою повышения трудоемкости, в основном, на этапе обучения). Резерв алгоритма - в повышении качества таксономии путем более полного учета информации о частотах общих подпоследовательностей.

4. Алгоритм A_{21} характеризуется наиболее низкой ошибкой первого рода на контроле (даже на довольно "тяжелом" для других алгоритмов тексте FD). Несколько отличный вариант нормировки позволяет снизить суммарную ошибку до 53 при незначительном повышении ошибки первого рода.

5. Алгоритм A_{12} характеризуется наибольшей устойчивостью: порог ε_1 практически не меняется при разных вариантах усечения обучающей выборки (по-видимому, это эффект достаточно "грубой" таксономии ($\varepsilon_T = 0{,}75$)). Резерв алгоритма - в дифференцированном выборе ε_1 для разных таксонов (чем меньше элементов в таксоне, тем больше должен быть порог ε_1).

6. Алгоритм A_{33} несколько уступает предыдущим по результатам контроля. Связано это, по-видимому, с недостаточностью объема обучающей выборки сайтов (выборка $\bar{S}$ не используется). В других алгоритмах этот недостаток частично компенсируется наличием $\bar{S}$. Алгоритм достаточно критичен к выбору метода таксономии, который должен быть согласован с используемым классом образов.

7. Поскольку по результатам контрольного обнаружения ни один из алгоритмов не доминирует явно над остальными, целесообразно использовать их в комплексе (коллектив решающих правил). Известно, что при определенных условиях надежность коллективного решения (по большинству голосов) выше надежности любого из правил коллектива (даже для зависимых классификаторов [16]). В табл. 2 приведены результаты обнаружения с помощью коллектива из четырех решающих правил (A_{11}, A_{12}, A_{21}, A_{22}). Анализировались (в режиме "скользящий контроль") геномы φX174 и FD. Поскольку число правил в коллективе четное, в отдельную графу вынесены ситуации, когда голоса разделились поровну: с→? (предъявлен сайт, но имеются 2 голоса против) и нс→? (предъявлен "не сайт", но по двум правилам он классифицируется как сайт).

Таблица 2

Текст	Число фрагментов	Число ошибок и неопределенных ситуаций					
		с→нс	с→?	нс→с	нс→?	фс→с	фс→?
φX174	II с; I95 нс	0	I	6	7	0	5
FD	I0 с; I75 нс	3	3	0	3	0	2

Анализ табл.2 показывает, что решающие правила, составившие коллектив, не слишком коррелированы (в отношении "не сайтов"). Из I5-20 "не сайтов" генома φX174, классифицировавшихся как сайты каждым из алгоритмов (см. табл.I), лишь в отношении шести мнения сходятся,в отношении I2 - разделяются,27 - не проходят по большинству голосов. Доля неопределенных ситуаций весьма высока .Вычленение этой категории из общего числа ошибок - положительный эффект применения коллектива решающих правил. Фильтрующие и дифференцирующие возможности коллектива растут с увеличением числа его членов.

Заключение

Рассмотрены специфические особенности знаков пунктуации и способы учета их в алгоритмах обнаружения знаков (фазирование, выделение информативных зон, информативных признаков, таксономия обучающей выборки). Предложены достаточно простые алгоритмы обнаружения таксономического и порогового типов. Проведена апробация и сопоставление алгоритмов на задаче обнаружения рибосомных сайтов связывания в геномах прокариотов.

Довольно высокий процент ошибок первого рода - следствие наличия значительного числа (до 20%) аномальных элементов, имеющих мало общего с консенсусом, и ограниченности (в целях упрощения процедуры принятия решения) окна анализа (использовалась,в основном, одна информативная зона).Аномальность знака может быть обусловлена такими обстоятельствами, как накопление "знаковых" свойств путем дупликаций слабых закономерностей (см. п.I.4), совмещение нескольких функций в одном знаке, функционирование знака в обоих направлениях,и другими факторами.Обнаружение аномальных знаков требует привлечения информации более высокого уровня,в частности построения достаточно общих моделей регуляции.

Многие ошибки второго рода можно отнести к категории условных: они характеризуют знаки, которые функционируют лишь в специфических условиях (характерные примеры по терминаторам приведены в [I7]). В связи с этим корректнее было бы вместо задачи обнаружения знаков пунктуации решать задачу упорядочения фрагментов по степени проявления свойства "быть знаком". Описанные в данной работе алгоритмы в той или иной степени обеспечивают указанную возможность.

Литература

1. SCHERER Y.F.E.,WALKINSHAW M.D.,ARNOTT S. A computer aided oligonucleotide analysis provides a model seguence for RNA-polymerase-promoter recognition in E.coli//Nucl.Acids Res.- 1978.-Vol.5, N 10.- P.3759-3773.

2. The ribosome binding sites recognized by E.coli ribosomes have regions with signal character in both the leader and protein coding segments / Y.F.E.Scherer, M.D.Walkinshaw, S.Arnott, D.J.Morre// Nucl.Acids Res.- 1980.- Vol.8,N 17.-P.3895-3907.

3. HAWLEY D.K., Mc CLURE W.R. Compilation and analysis of Escherichia coli promoter DNA sequences// Nucl.Acids Res.- 1983.- Vol.11,N 8.-P.2237-2255.

4. STORMO Y.D., SCHNEIDER T.D., GOLD L.M. Characterization of translational initiation sites in E.coli// Nucl.Acids Res. - 1982.-Vol.10,N 9.- P.2971-2996.

5. STADEN R. Computer Methods to locate signals in nucleic acid sequences// Nucl.Acids Res.- 1984.- Vol.12, N 1, part 2. - P.505-519.

6. Use of the "perceptron" algorithm to distinguish translational initiation sites in E.coli / Y.D.Stormo, T.D.Schneider, L. Gold, A.Ehrenfeucht// Nucl.Acids Res.- 1982. -Vol.10,N 9.-P.2997-3011.

7. BRENDEL V., TRIFONOV E.N. A computer algorithm for testing potential prokaryotic terminators// Nucl.Acids Res.- 1984. - Vol.12,N 10.-P.4411-4427.

8. Средства анализа генетических текстов в рамках ППП "Сим - вол"/Г.С.Высоцкая, В.Д.Гусев, Ю.Г.Косарев и др. // Теоретические исследования и банки данных по молекулярной биологии и генетике: Сб.научн.тр. - Новосибирск, 1986. -С.48-53.

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

10. ANGLUIN D. Finding patterns common to set of strings// J. comput.and syst.sci.-1980.-Vol.21,N 1.-P.46-62.

11. SHINOHARA T. Polinomial time inference of extended regular pattern languages// Lect.notes in comput.sci.-1983.- N 147. - P.115-127.

12. ЧУЖАНОВА Н.А. Об одном способе генерации языковых процессоров //Структурный анализ символьных последовательностей.-Новосибирск. - 1984. -Вып.101: Вычислительные системы. -С.44-55.

I3. ВЫСОЦКАЯ Г.С., ГУСЕВ В.Д., КУЛИЧКОВ В.А. Метод выявления информативных зон в генетических знаках пунктуации //Теоретические исследования и банки данных по молекулярной биологии и генетике: Сб. научн. тр. -Новосибирск, I986. -С. 54-58.

I4. ГУСЕВ В.Д., КУЛИЧКОВ В.А., НИКУЛИН А.Е. Алгоритмы поиска несовершенных повторов в генетических текстах //Анализ символьных последовательностей.-Новосибирск, I985. -Вып.II3: Вычислительные системы. -С. I07-I22.

I5. КЕНДЭЛ М. Ранговые корреляции. -М.: Статистика, I975.

I6. ЗУЕВ Ю.А. О статистических свойствах принятия решения большинством голосов в задачах классификации //Докл. АН СССР.- I986. - Т.288. №2. -С. 320-322.

I7. BRENDEL V. Mapping of transcription terminators of bacteriophages ФX174 and G4 by sequence analysis// J.of virology. - 1985.-Vol.53,N 1.-P.340-342.

Поступила в ред.-изд.отд.
28 сентября I987 года