МАШИННЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1981 год Выпуск 88

УДК 519.767.6:577

ВЫЯВЛЕНИЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ В ЗАДАЧАХ РАСПОЗНАВАНИЯ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ (на примере генетических текстов)

Т.Н.Титкова

При анализе генетических текстов (первичных структур ДНК-молекул простейших микроорганизмов) возникает задача автоматического выделения внутри текста участков с различным функциональным назначением (генов, знаков пунктуации и т.д.).Роль знаков пунктуации играют подпоследовательности длиной от трех до нескольких десятков символов.

Важным классом знаков пунктуации являются промоторы-участки генома (текста), ответственные за начало процесса транскрипции.Их длина порядка 50-70 нуклеотидов. В настоящее время расшифровано уже несколько десятков промоторов, т.е. установлена их первичная структура и доказано (молекулярно-биологическим анализом), что соответствующие последовательности нуклеотидов выполняют роль инициаторов транскрипции. Совокупность этих последовательностей может рассматриваться как обучающая выборка по промоторам в задаче классификации "промотор-не промотор".Для элементов этой выборки существует реальное опознающее устройство - РНК-полимераза. Обучающая выборка по "не промоторам" может быть получена из текстов уже расшифрованных геномов с исключенными промоторными зонами.

Удачное решение задачи классификации формальными методами (т.е. прямо по виду символьной последовательности) может сильно упростить и ускорить весьма трудоемкую процедуру выявления генетической структуры генома.

В основу построения решающего правила кладется различие в характеристиках элементов первого и второго образов, выявляемое

при анализе обучающих выборок. Исследовалась возможность использования для этой цели 1-граммных характеристик [1,2], показывающих, какие 1-граммы (связные подпоследовательности из 1 символов, $1 = 1,2,3,\ldots$) и в каком количестве входят в анализируемые последовательности. Поскольку промоторы имеют довольно большую длину при малом алфавите ($n = 4$), почти все из n^1 возможных комбинаций нуклеотидов при $1 = 1,2,3$ присутствуют в обучающей выборке промоторов. Естественно было поэтому искать различия между промоторами и "не промоторами" на уровне длинных 1-грамм ($1 \geq 7$).

Назовем 1-грамму *типичной для промоторов*, если она встречается как минимум в двух промоторах и не встречается у "не промоторов". Методика выявления типичных 1-грамм такова. Составим конкатенацию $T_n = T_1 \sqcup T_2 \sqcup T_3 \sqcup \ldots \sqcup T_m$, где T_i $(1 \leq i \leq m)$ - текст, описывающий i-й промотор, m - число промоторов, $\sqcup$ - разделитель между каждой парой промоторов (символ, не содержащийся ни в одном из T_i). Получим полный частотный спектр текста T_n и определим параметр $1_{max}(T_n)$, т.е. минимальное 1, начиная с которого в тексте T_n уже отсутствуют повторяющиеся 1-граммы. Аналогичную операцию проделаем с текстом T_H, содержащим последовательность "не промоторов".

Рассмотрим множество кратных 1-грамм из T_n длины $(1_{max}(T_n) - 1)$, не содержащих разделителя, не встретившихся в T_H и оказавшихся в разных промоторах. Составим таблицу из нулей и единиц, где каждая строка соответствует "типичной" 1-грамме, а каждый столбец - одному из промоторов. На пересечении соответствующей строки и столбца ставится единица, если данная 1-грамма присутствует в рассматриваемом промоторе. Проверяем, обеспечивает ли выделенное множество $(1_{max}-1)$-грамм полное покрытие всех промоторов из обучающей выборки. Если да, то процесс выбора "информативных" для классификации "промотор-не промотор" признаков (1-грамм) заканчивается. В противном случае дополняем уже отобранное множество $(1_{max}(T_n)-1)$-грамм кратными 1-граммами длиной на единицу меньше и удовлетворяющими тем же трём условиям и одному дополнительному: новые 1-граммы не должны быть подпоследовательностями уже выписанных ранее 1-грамм. Исключение делается лишь для тех 1-грамм, которые имеют бо́льшую частоту, чем содержащая их $(1+1)$-грамма. Это означает, что данная 1-грамма встретилась еще в одном или нескольких промоторах.

Процесс понижения значений l продолжается до тех пор, пока не будет сформировано полное покрытие. Если по мере понижения l больше не находится l-грамм, удовлетворяющих всем четырем условиям, некоторые из них могут быть ослаблены (например, можно допустить присутствие данной l-граммы и в T_H, но с частотой существенно меньшей, чем в T_n). Полученное покрытие может оказаться избыточным и при необходимости может быть минимизировано.

В нашем эксперименте обучающая выборка по промоторам содержала 24 последовательности. Обучающая выборка по "не промоторам" состояла из 4 геномов (ФХ174, MS2, G4, SV40) с исключенными промоторными зонами. Длина текста T_H в несколько раз превышала длину текста T_n.

Таблица I

l-грамма	Промоторы, содержащие ее
ТГТТГАЦА	λ-PL, E.coli-TRP, ФХ174D, S, TYPH
АЦАЦТТТ	E.coli-LAC, E.coli-GAL, TPHK-TUP, E.coli K12-ARAC
ГЦГТТГАТА	λ-PL, λ-PR, λ-PRM
АТГТТТАЦА	ФХ174B, SV40
АГГАААТА	ФХ174A, C17
ГАТАЦАААТЦ	FD:G3, T7:A2
ААЦАААААЦ	FD:G1, T7:A3
ЦТГТАТТТГ	λ-P_0, λ-IMM434
АТТГАЦТТА	T7:A1, λ-PR'
ЦГЦТТТ	FD:G2, E.coli K12-ARAC, E coli K12-ARAB

По результатам эксперимента можно отметить следующее.

I. Покрытие для промоторов (см.табл.I) удалось сформировать из l-грамм длиною ≥ 8 (за единственным исключением l= 6), причем все l-граммы удовлетворяли четырем сформулированным выше требованиям. Это означает, что гипотеза о наличии закономерных связей между промоторами в виде достаточно длинных совпадающих подпоследовательностей полностью подтвердилась.

2. Поскольку длина l-граммы, общей для пары промоторов, характеризует силу связи, то все множество промоторов по степени их близости можно разбить на таксоны в этом смысле. Все исследовавшиеся группы промоторов в первом приближении разбиваются на два класса. Первый из них, куда входят группы T7 и λ, характеризуется, за некоторыми исключениями, сильной связью между промоторами внутри каждой группы. Второй класс, куда входят группы ФХ174, E.coli и FD, характеризуется слабой связью.

3. Анализ расположения характерных 1-грамм по позициям показал, что существуют "устойчивые" 1-граммы, расположенные примерно в одних и тех же позициях у разных промоторов. "Устойчивые" 1-граммы предлагается в первую очередь включать в покрытие. Однако существуют характерные 1-граммы, расположенные в промоторах довольно далеко друг от друга. Такая неустойчивость в расположении 1-грамм требует, по-видимому, особой трактовки в каждом отдельном случае.

Литература

1. 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.

2. 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.

Поступила в ред.-изд.отд.
14 апреля 1981 года