*Информационные процессы, Том 6, № 1, 2006, стр. 33–36.*




# Алгоритм поиска консервативных участков нуклеотидных последовательностей

**А.В. Селиверстов, В.А. Любецкий**

*Институт проблем передачи информации РАН, 127994, Россия, Москва, Большой Каретный переулок, 19,*
*e-mail: slvstv@iitp.ru, lyubetsk@iitp.ru*

Поступила в редколлегию 20.12.2006

**Аннотация**—Дано подробное описание эвристического алгоритма поиска клики в многодольном графе и результаты применения к выравниванию нуклеотидных последовательностей. В частности, описаны THI-рибосвитчи у гипотетического ABC транспортера актинобактерий.

## 1. ВВЕДЕНИЕ

Разработана модификация алгоритма для поиска консервативных подслов в последовательностях нуклеотидов. Алгоритм основан на эвристическом подходе к поиску клики в многодольном графе и является модификацией ранее разработанного авторами алгоритма поиска плотного подграфа [1].

Были исследованы 5'-нетранслируемые области перед генами актинобактерий, кодирующими предполагаемый ABC транспортер, гомологичный гену *ykoE* сенной палочки. Найдены консервативные THI-рибосвитчи, некоторые из которых ранее описаны в [2], а другие открыты нами в секвенированных геномах.

Кроме того, предложена схема регуляции трансляции гена *alr3806* при участии T-бокса в цианобактерии *Nostoc* sp. PCC 7120.

## 2. АЛГОРИТМ

Поиск консервативных фрагментов основан на поиске клики в многодольном графе. По исходному набору невыравненных последовательностей ищется сигнал — набор наиболее попарно похожих слов одинаковой длины не более, чем по одному, из каждой последовательности. Рассматривается граф, в котором вершинам приписаны всевозможные слова этой длины (без повторений) из всех исходных последовательностей, а долей являются все вершины, которым приписаны слова из одной последовательности. Таким образом, долей столько, сколько последовательностей. Ребром соединяются две вершины, если приписанные им слова похожи друг на друга выше некоторого порога, который является параметром алгоритма. Алгоритм ищет в таком многодольном графе клику данного размера $q$ ("$q$-клика"), при чем $q$ может постепенно увеличиваться.

Поиск $q$-клики происходит итеративным исключением вершин графа, которые соединяются с числом долей (хотя бы одним ребром), которое строго меньше $q-1$, и также ребер, принадлежащих меньшему, чем пороговое число "треугольников" (т.е. 3-клик) $q-2$, или меньшему, чем пороговое число "тетраэдров" (т.е. 4-клик) $(q-2)(q-3)/2$. Указанная процедура применяется пока возможно. Если после этого: не осталось ребер, алгоритм завершает работу; если имеется $q$-клика, все вершины которой смежны внутри нее и не смежны вне нее, то алгоритм

выдает эту клику и удаляет ее; если это не так, то одно из ребер, входящих в наименьшее число треугольников (первое в смысле фиксированной исходной нумерации всех ребер). При этом используется процедура, описанная ниже.

**Подсчет числа малых клик, содержащих данное ребро.** Рассмотрим неориентированный граф $G$ с $V$ вершинами и $E$ ребрами. Он определяется своей матрицей смежности, т.е. симметричной порядка $V$ матрицей $A$ из 0 и 1 с нулями на главной диагонали. Арифметическая сложность матричного умножения $O(V^w)$, где $w < 2.376$. Напомним, что для обычного метода $w = 3$ и для алгоритма Штрассена $w = \log_2 7$. Для двух матриц $A$ и $B$ порядка $V$ произведение Адамара $A * B$ есть матрица порядка $V$ из произведений $(A * B)_{ij} = A_{ij}B_{ij}$. Рассмотрим произведение Адамара $A * A^2$. Для любых значений $i < j$ элемент $(A * A^2)_{ij}$ равен числу треугольников, содержащих обе вершины $i$ и $j$. Сложность подсчета числа треугольников в графе $G$ должна составлять $O(V^w)$ арифметических операций, где $w$ равно показателю для матричного умножения, или при малом числе треугольников - $O(VE)$ арифметических операций. Тетраэдр (т.е. 4-клика) определяется парой несмежных ребер. Подсчет числа тетраэдров, содержащих данное ребро, должно выполняться за $O(E)$ арифметических операций.

Возможно построение алгоритмом вспомогательного графа $G'$, который уменьшает число недопредсказаний итоговых клик. Этот граф $G'$ строится следующим образом. Его вершины соответствуют ребрам исходного графа $G$. Две вершины в $G'$ смежные, если соответствующие ребра графа $G$ являются несмежными ребрами хотя бы в каком-то тетраэдре. Треугольник в графе $G'$ соответствует 6-клике в графе $G$. Поэтому подсчет числа 6-клик должен не превышать $O(E^w)$ арифметических операций, где $w$ — показатель для матричного умножения.

Используя поиск клики, легко искать наборы попарно похожих слов фиксированной дины по не более одной из каждой последовательности нуклеотидов, где предполагается наличие сигнала.

Параметрами программы являются: число строк $q$, в которых должен быть сигнал, длина сайтов, максимальное расстояние (по Хэммингу) между любой парой сайтов из сигнала. Состав входных данных программы: файл, содержащий последовательности в формате FASTA и параметры поиска.

## 3. THI-РИБОСВИТЧИ

Геномы бактерий получены из базы данных GenBank (NCBI). В качестве набора последовательностей нами были взяты 5'-нетранслируемые области перед гомологами гена *ykoE* у актинобактерий, перечисленных в таблице 1, а также у *Bifidobacterium longum*, *Corynebacterium jeikeium*, *Corynebacterium efficiens* и *Tropheryma whipplei*, содержащих гомологи гена *ykoE*, но у которых сигнал не был обнаружен.

| Краткое | и полное название вида | Белок |
|---|---|---|
| Blin | *Brevibacterium linens* | ZP_00378910.1 |
| Krad | *Kineococcus radiotolerans* | ZP_00619644.1 |
| Lxx | *Leifsonia xyli* | YP_062345.1 |
| Ppa | *Propionibacterium acnes* | YP_054871.1 |
| Tfus | *Thermobifida fusca* | YP_288648.1 |
| Cdip | *Corynebacterium diphtheriae* | NP_939314.1 |
| Cglu | *Corynebacterium glutamicum* | YP_225369.1 |

**Таблица 1** Краткие обозначения геномов и номера белков, гомологичных YkoE в сенной палочке.

В таблице 2 показано множественное выравнивание, построенное на основе поиска консервативных участков. Всего найдено шесть консервативных участков с длинами десять, двена-

дцать, десять, тринадцать, шесть и шесть нуклеотидов. При этом выявились консервативные спирали, характерные для THI-рибосвитча.

| | |
|---|---|
| Blin | acAGGGgAGCGCCga**********uaggGGCGCUgagagUGCAGa*********ugaagCUGCAgaCCCUc |
| Krad | acAGGGgAGCGCCg**********uggGGCGCUgagagUGCGG**********guuuCCGCAgaCCCUc |
| Lxx | acACGGgAGUCCGGu**********gagCCGGGCUgagagGAAGCUU********auccAAGCUUCgaCCGUc |
| Pacn | acAGGGgAGCAUCg**********ucgGAUGCUgagagUGGGC**********accGCCCAgaCCCUc |
| Tfus | acAGGGgAGCGCcu**********cuaGCGCUgagagUGCGGC**********acaGCCGCAgaCCCUu |
| Cdip | ucACGGguGCUggacGGCAuacguuUGCCacaaAGCugagaCAGGGcgagaagacgugcacguCCCUGaaCCGUu |
| Cglu | acACGGguGCUCGGuga*********aaauCCGGGCugagaucUGGC**********auaGCCAcgaCCGUc |
| | **->1>*->->-2->->->******<-<-<-2<-<-<****->->3->**********<-3<-<-**<1<-* |
| Blin | *gaaCCUgauGCGGcuagcaCCGCcga*AGGaag |
| Krad | *gaaCCUGauCCGGuucagaCCGGcg*UAGGgag |
| Lxx | *gaaCCUGauCUGGgucaugCCAGcg*CAGGgag |
| Pacn | *gaaCCUGaaCCGGuuaggaCCGGcg*UAGGgag |
| Tfus | acuaCCUgauCUGGguaaugCCAGcga*AGGaag |
| Cdip | *gaaCCUGauCCGGguaauaCCGGcgaUAGGaag |
| Cglu | *gaaCCUGauCCGGauaaugCCGGcgaUAGGgag |
| | ****>4->**->5>******<5<-***<6<-*** |

**Таблица 2**. Множественное выравнивание, построенное на основе поиска консервативных слов. Прописными буквами выделены нуклеотиды, входящие в состав спиралей консервативной вторичной структуры РНК.

Мы предполагаем, что найденная консервативная структура связана с регуляцией трансляции. При этом у *Brevibacterium linens*, *Kineococcus radiotolerans*, *Leifsonia xyli*, *Propionibacterium acnes* и *Thermobifida fusca* область связывания рибосомы перекрывается дополнительной короткой спиралью, а у *Corynebacterium diphtheriae* и *Corynebacterium glutamicum* рибосвитч примыкает непосредственно к области связывания рибосомы.

Рассмотренные рибосвитчи у трех актинобактерий *T. fusca*, *C. diphtheriae* и *C. glutamicum* описаны ранее в статье [2].

## 4. ГИПОТЕТИЧЕСКИЙ Т-БОКС ИЗ *NOSTOC*

Добавляя к набору последовательностей, содержащих известный хороший сигнал, новую последовательность, можно предсказывать наличие сигнала в ней. Так предсказана регуляция трансляции с участием Т-бокса в цианобактерии *Nostoc* sp. PCC 7120 гена *alr3806*, который не имеет ортологов ни в каком другом полном геноме из базы NCBI.

Открытая рамка считывания *alr3806* длиной 450 аминокислот из *Nostoc* sp. PCC 7120 предсказана теоретически и не соответствует экспериментально подтверждённому белку. Перед этой открытой рамкой на расстоянии 147 нуклеотидов на комплементарной цепи расположена другая открытая рамка считывания *alr3805*, экспрессия которой также не подтверждена экспериментом. Предполагаемый Т-бокс почти целиком расположен в промежутке между этими рамками, включая старт кодон *alr3805*. Предлагаемая структура имеет слово с правильным консенсусом (собственно Т-бокс) и шпильки, характерные для Т-бокса. Важно, что тРНК стабилизирует такую структуру РНК, которая не препятствует трансляции гена *alr3806*. В противном случае возникает спираль, перекрывающая область связывания рибосомы гена *alr3806*, препятствуя трансляции.

Специфицирующм кодоном, вероятно, является аргининовый кодон agg в выпячивании на 3'-плече соответствующей шпильки. Такое расположение кодона соответствует ранее известным примерам Т-боксов. Таким образом формирование структуры РНК зависит от концентрации аргинина и, следовательно, связанного азота.

Ортологов для гена *alr3806* не найдено, но его N-концевой домен обладает АТФазной активностью. Ближайшие гомологи гена *alr3806* как в той же актинобактерии *Nostoc* sp. PCC 7120 (ген *all4835*), так и других цианобактериях порядка Nostocales — *Anabaena variabilis* ATCC 29413 и *Nostoc punctiforme* PCC 73102 — ортологичны друг другу и принадлежат COG1066: АТФ-зависимые сериновые протеазы.

Сказанное выше о предполагаемой регуляции и гомологах гена *alr3806* позволяет предположить, что соответствующий белок участвует в деградации цианофицина и освобождении аргинина. Однако цианофициновые гранулы характерны для других Nostocales, не имеющих ортолога гена *alr3806*.

Работа частично поддержана грантом МНТЦ (2766).

## СПИСОК ЛИТЕРАТУРЫ

1. В.А. Любецкий, А.В. Селиверстов. Некоторые алгоритмы, связанные с конечными группами. *Информационные процессы*, 2003, 3(1), стр. 39–46

2. D.A. Rodionov, A.G. Vitreschak, A.A. Mironov, M.S. Gelfand. Comparative Genomics of Thiamin Biosynthesis in Procaryotes, *The biological chemistry*, 2002, 277(50), pp. 48949–48959.