Информационные процессы, Том 6, № 3, 2006, стр. 258–263.

Алгоритм поиска мощных вторичных структур в нуклеотидной последовательности и его применение для подсчета числа таких структур в разных областях геномов

Е.В. Любецкая, Л.И. Рубанов, А.В. Селиверстов, В.А. Любецкий

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

Поступила в редколлегию 03.07.2006

Аннотация—Разработан алгоритм для массового поиска шпилек в нуклеотидной последовательности, который применим для массового поиска шпилек большой мощности с заданными параметрами черенка и петли в трейлерных областях и в других типах областей генома у актинобактерий. С помощью предложенного алгоритма подсчитывались в геномных областях разного типа числа шпилек, которые могут образовать крест, т.е. пару шпилек на комплементарных цепях ДНК. В результате в этих областях обнаружены существенно различные количества длинных шпилек, прежде всего, в области между сходящимися генами (конвергонами). Возможно, что найденные шпильки связаны с формированием нетривиальной вторичной структуры ДНК, которая, возможно, служит для снятия конформационного напряжения и для терминации транскрипции.

1. ВВЕДЕНИЕ

В геномах бактерий нередко чередуются гены, расположенные на комплементарных цепях ДНК. Если направленные навстречу друг другу гены (так называемые, *конвергоны*) интенсивно транскрибируются, то можно предполагать присутствие в области между ними терминатора транскрипции. Кроме того, в этом месте происходит частое образование супервитков ДНК. Нами предложен новый алгоритм для массового поиска шпилек в нуклеотидной последовательности, который применим для массового поиска шпилек большой мощности с заданными параметрами её черенка и петли в трейлерных областях и в других типах областей генома у актинобактерий. С помощью предложенного алгоритма подсчитывались в геномных областях разного типа числа шпилек, которые могут образовать крест, т.е. определённую пару шпилек на комплементарных цепях ДНК. В результате в этих областях обнаружены существенно различные количества *длинных* (от 17 пар нуклеотидов в черенке) шпилек, прежде всего, в области между сходящимися генами (конвергонами). Эти шпильки не принадлежат известным типам регуляторных элементов и их количество гораздо большее, чем в аналогичных областях у *Escherichia coli* и *Bacillus subtilis*. Мы предположили, что найденные шпильки связаны с формированием нетривиальной вторичной структуры ДНК, которая, возможно, служит для снятия конформационного напряжения и для терминации транскрипции.

У актинобактерий экспрессия генов часто регулируется на уровне трансляции [1], в то время как для гамма- и альфа- протеобактерий характерна классическая аттенюаторная регуляция на уровне транскрипции [2]. У наиболее изученных бактерий *B. subtilis* и *E. coli* оперoны обычно заканчиваются *классическим транскрипционным терминатором*, т.е. GC-богатой шпилькой с небольшой петлёй и с $5-10$ парами комплементарных оснований в спаренном участке, за которой следует U-богатый участок. У некоторых видов бактерий, включая *Mycobacterium* spp., такие классические терминаторы относительно редки [3, 4].

В данной работе исследуются области, в которых следует ожидать вторичные структуры РНК или ДНК, связанные с терминацией транскрипции. Особый интерес представляют участки после генов тРНК и после генов, кодирующих высоко экспрессируемые белки.

Такое исследование может быть полезно для предсказания границ оперонов и для различения высоко экспрессируемых генов от редко транскрибируемых паралогов. Последнее обсуждалось в [5].

2. АЛГОРИТМ

Геномы бактерий получены из ГенБанка. Рассмотрены актинобактерии и бактерии *E. coli* и *B. subtilis*. В этих геномах последовательно рассматривались участки следующих четырёх *типов*: все лидерные области, все сходящиеся трейлерные области (конвергоны), все кодирующие области, все расходящиеся лидерные области (дивергоны).

В областях каждого из этих типов искалось по одной наиболее мощной шпильке среди всех шпилек со следующими особенностями: шпилька имеет *черенок* с небольшими *выпячиваниями* вплоть до *концевой петли* и в петле имеется спаривание наибольшего возможного числа пар нуклеотидов. Как обычно, мощность шпильки измеряется как сумма: положительное слагаемое за каждое спаривание комплементарных нуклеотидов и отрицательное слагаемое (штраф) за каждый не спаренный нуклеотид только внутри *плеч* черенка, т.е. внутри участков, которые с двух сторон примыкают к концевой петле. Среди таких шпилек одинаковой мощности алгоритм отбирает те, у которых меньше отрезков в петле черенка. Стандартные термины, выделенные выше курсивом, поясняются, например, в [6].

Поиск таких шпилек осуществлялся следующим алгоритмом. Сначала на рассматриваемом участке последовательности выделяются все непродолжаемые спирали, которые потенциально могут выступать в роли черенка искомой шпильки. Для этого отбираются спирали, у которых число пар в спаренном участке не менее данного порога и при этом длина петли не выходит за пределы заданного интервала. Границы интервала и порог для длины являются параметрами алгоритма; в экспериментах использовались значения $[3, 100]$ и 7.

Спаренный участок здесь понимается в расширенном смысле, т.е. может включать выпячивание длиной не более двух нуклеотидов, однако все спаренные сегменты шпильки должны состоять не менее чем из трёх пар нуклеотидов (указанные значения также являются параметрами алгоритма). Мощность черенка измеряется с прибавлением небольшого штрафа (например, -0.1) за каждый нуклеотид из выпячивания. Дополнительное ограничение состоит в том, что петля не должна содержать участка, комплементарного любому из плеч черенка, т.е. черенок складывается с петлей минимально возможной длины. После этого для каждого из отобранных черенков-кандидатов выполняется поиск максимально мощной структуры в его петле, понимая под мощностью суммарную длину всех спаренных участков. При поиске действуют те же, что и выше, ограничения на минимальную длину петли и спаренного сегмента, однако расстояния между соседними сегментами не ограничиваются, т.е. в петле черенка могут образовываться конфигурации с любыми выпячиваниями и внутренними петлями.

Алгоритм поиска структуры максимальной мощности зависит от общего числа непродолжаемых спаренных сегментов, присутствующих внутри петли (разумеется, часть из них будет взаимно исключать друг друга, в том числе и из-за появления псевдоузлов). Вначале в такую структуру включаются все сегменты, которые не противоречат каким-либо другим сегментам внутри петли. Если после этого внутри петли остается $k < 24$ потенциальных сегментов, то проводится полный перебор 2^k всевозможных конфигураций с выбором из них максимальной по мощности. Если в петле остается свыше 24 сегментов, то полный перебор их сочетаний становится слишком громоздким. Тогда применяется один из двух следующих эвристических квазиоптимальных алгоритмов, выбор варианта является параметром общего алгоритма:

1. оставшиеся сегменты перебираются слева направо в порядке позиции начала их левого плеча; очередной сегмент добавляется в структуру, после чего из списка оставшихся сегментов исключаются несовместимые с ним (из-за перекрытия плеч, не допускающего добавления сегмента минимально допустимой длины, или образования псевдоузла); эти действия продолжаются до опустошения списка;
2. оставшиеся сегменты перебираются в порядке убывания их мощности (длины спаренного участка), в остальном действия аналогичны первому варианту.

Как показали компьютерные эксперименты, второй вариант, при котором в квазиоптимальную структуру внутри петли вначале включаются более длинные отрезки, намного чаще первого приводит к построению структуры большей мощности, этот вариант принимается по умолчанию.

После того, как описанным выше образом для всех потенциальных черенков построены структуры максимальной мощности в их петле, найденные решения упорядочиваются по убыванию суммарной мощности шпильки, и в качестве общего результата алгоритма выдается заданное параметром алгоритма число лучших (в смысле мощности) конфигураций.

Изложенный алгоритм реализован в виде программы Destree на языке C, которая содержит около 600 строк исходного текста и имеет интерфейс командной строки. Типичное время обработки последовательности длиной 1500 букв составляет единицы секунд.

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Напомним, что рассматриваются участки четырёх указанных выше *типов*: все лидерные области, все сходящиеся трейлерные области (конвергоны), все кодирующие области, все расходящиеся лидерные области (дивергоны). С помощью описанного выше алгоритма на каждом участке отбиралось не более одной шпильки с самым длинным черенком у неё. При этом отбирались шпильки с черенками длины не меньше порогового значения ℓ, с длиной петли черенка не более 15 нуклеотидов, с односторонними выпячиваниями в черенке, состоящими из не более чем двух нуклеотидов. В столбце, соответствующем конвергонам, после скобок указана частота шпилек в областях данного типа, после которых не имеется участка остатков тимидина, состоящего не менее чем из 7 остатков с не более чем двумя исключениями. Такой участок ищется только в окрестности последнего нуклеотида шпильки, которая состоит из 5 нуклеотидов влево и 12 вправо. Для генов, кодирующих белки, в случае конвергонов типичны длинные шпильки. Максимальные длины шпилек в трейлерных областях достигают 31 у *Mycobacterium tuberculosis*, *M. bovis* и *Streptomyces coelicolor*, 29 у *S. avermitilis*, 27 у *C. diptheriae*. В отличие от актинобактерий, данные по которым приведены в Таблице, у *B. subtilis* и особенно у *E. coli* шпильки с черенками от 17 нуклеотидов встречаются много реже. Если хотя бы один из двух направленных навстречу друг другу генов интенсивно транскрибируется, то после гена у актинобактерий часто обнаруживаются терминатор транскрипции в виде длинной шпильки. Эта шпилька также находится в месте возможного образования супервитков ДНК. Эти шпильки не принадлежат известным типам регуляторных элементов. Мы предполагаем, что они соответствуют формированию двух симметричных шпилек на комплементарных цепях ДНК, образующих *крест*, образованный двумя шпильками на комплементарных цепях ДНК, [7]. Роль так образованной структуры связана с двумя обстоятельствами. Во-первых, изменение конформации ДНК может снимать конформационные напряжения на ДНК, возникающие при интенсивной транскрипции. Обычный механизм, вовлекающий топоизомеразы, менее эффективен и связан с дополнительным риском повреждения ДНК. Во-вторых, изменение конформации ДНК может обеспечивать терминацию транскрипции. Большинство найденных шпилек значительно длиннее обычных регуляторных терминаторов, и, например, за ними не расположены тимидин-богатые участки. Возникающая структура симметрична и может играть роль

терминатора для РНК полимеразы, движущейся по каждой из цепей ДНК. Косвенным свидетельством в пользу такой гипотезы служит большая частота указанных структур именно в конвергонах по сравнению с другими четырьмя типами областей.

Таблица Указана частота p в процентах и в столбце $(2b)$ абсолютные количества шпилек (в скобках без участка остатков тимидина) в зависимости от нижней границы на длину черенка ℓ (столбец (0)), найденных в перечисленных выше четырёх типах областей (столбцы $(1)-(4)$ в том порядке, как они выше перечислены) у актинобактерий и у *B. subtilis* и *E. coli*.

ℓ	Лидерные области	Конвергоны		Кодирующие области	Дивергоны
0	1	$2a$	$2b$	3	4
Corynebacterium efficiens					
25	0.23	0.38	1 (0)	0.00	0.00
23	0.23	0.76	2 (0)	0.00	0.00
20	0.68	6.06	16 (11)	0.00	0.00
17	2.60	13.64	37 (29)	0.17	0.31
15	4.98	21.59	57 (48)	0.43	1.85
10	20.61	45.83	121 (95)	31.79	8.31
Corynebacterium glutamicum					
25	0.00	0.35	1 (1)	0.00	0.00
23	0.00	1.04	3 (3)	0.03	0.00
20	0.18	5.54	16 (5)	0.03	0.00
17	1.11	13.84	40 (11)	0.13	0.00
15	2.68	20.42	59 (19)	0.25	0.00
10	20.41	46.02	133 (53)	19.65	6.86
Corynebacterium diphtheriae					
25	0.00	1.24	2 (1)	0.00	0.00
23	0.13	3.73	6 (4)	0.00	0.00
20	0.26	8.70	14 (7)	0.00	0.00
17	1.85	19.88	32 (17)	0.02	0.00
15	3.17	29.81	48 (28)	0.08	0.00
10	18.10	47.20	76 (43)	5.68	8.21
Mycobacterium bovis					
25	0.16	3.95	12 (12)	0.02	0.18
23	0.24	4.93	15 (15)	0.02	0.18
20	0.32	5.59	17 (17)	0.05	0.18
17	0.95	6.91	21 (21)	0.14	0.36
15	1.35	7.24	22 (22)	0.67	0.53
10	14.91	12.17	37 (35)	36.10	6.57
Mycobacterium leprae					
25	0.62	0.00	0	0.06	0.74
23	0.62	0.00	0	0.06	0.74
20	0.99	1.72	1 (1)	0.12	0.74
17	1.23	1.72	1 (1)	0.37	0.74
15	1.73	1.72	1 (1)	0.82	1.48
10	13.70	6.90	4 (4)	14.69	5.93

Таблица (Продолжение)

ℓ	Лидерные области	Конвергоны		Кодирующие области	Дивергоны
0	1	2*a*	2*b*	3	4
Mycobacterium avium					
25	0.09	0.42	1 (1)	0.00	0.00
23	0.28	0.42	1 (1)	0.00	0.00
20	0.37	1.68	4 (4)	0.00	0.00
17	0.74	3.36	8 (8)	0.18	0.00
15	1.58	7.98	19 (17)	0.55	0.51
10	23.82	19.33	46 (44)	49.95	11.17
Mycobacterium tuberculosis					
25	0.31	3.29	11 (11)	0.07	0.18
23	0.38	4.19	14 (14)	0.07	0.18
20	0.54	5.09	17 (17)	0.07	0.36
17	1.30	5.99	20 (20)	0.14	0.53
15	1.83	6.29	21 (21)	0.57	0.89
10	15.44	10.78	36 (35)	35.83	5.35
Propionibacterium acnes					
25	0.00	0.00	0	0.00	0.00
23	0.00	0.00	0	0.00	0.00
20	0.00	0.96	2 (2)	0.04	0.00
17	0.45	8.17	17 (16)	0.09	0.00
15	1.20	12.98	27 (23)	0.13	0.00
10	15.19	30.29	63 (57)	24.29	6.34
Streptomyces avermitilis					
25	0.10	0.54	5 (5)	0.00	0.00
23	0.33	1.19	11 (11)	0.00	0.26
20	0.60	2.81	26 (26)	0.03	0.34
17	2.13	9.83	91 (90)	0.30	0.60
15	4.23	16.63	154 (152)	0.93	0.69
10	30.38	39.52	366 (359)	49.23	12.35
Streptomyces coelicolor					
25	0.08	0.95	8 (8)	0.00	0.00
23	0.15	2.25	19 (18)	0.01	0.00
20	0.54	3.08	26 (25)	0.02	0.17
17	1.73	9.60	81 (78)	0.09	0.41
15	3.61	14.45	122 (117)	0.35	1.00
10	31.07	37.32	315 (306)	23.09	14.36
Leifsonia xyli					
25	0.11	0.00	0	0.00	0.00
23	0.11	0.00	0	0.00	0.00
20	0.65	2.78	4 (4)	0.05	0.00
17	1.63	4.86	7 (7)	0.34	0.00
15	2.28	7.64	11 (11)	1.01	0.47
10	21.93	22.92	33 (33)	39.28	9.48

Таблица (Продолжение)

ℓ	Лидерные области	Конвергоны		Кодирующие области	Дивергоны
0	1	2*a*	2*b*	3	4
Escherichia coli					
25	0.00	0.00	0	0	0.00
23	0.00	0.00	0	0	0.00
20	0.00	0.00	0	0	0.00
17	1.38	0.22	1 (1)	0.02	0.19
15	4.15	2.91	13 (8)	0.13	0.56
10	77.67	47.65	213 (159)	18.56	11.52
Bacillus subtilis					
25	0.00	0.00	0	0	0.00
23	0.00	0.00	0	0	0.00
20	0.13	0.51	2 (1)	0	0.00
17	0.89	3.55	14(6)	0.02	0.20
15	2.74	13.96	55(12)	0.11	0.20
10	25.86	67.77	267 (97)	18.71	13.64

Длинные ($10 - 25$ пар нуклеотидов в черенке) шпильки обнаружены после генов многих тРНК, которые интенсивно транскрибируются. Например, у *Propionibacterium acnes* длинные шпильки следуют за генами *tRNA-Ala* (*ppa*2421), *tRNA-Arg* (*ppa*2413), *tRNA-Arg* (*ppa*2189), *tRNA-Asn* (*ppa*2422), *tRNA-Glu* (*ppa*2432), *tRNA-Lys* (*ppa*0181), *tRNA-Lys* (*ppa*1961), *tRNA-Met* (*ppa*2423), *tRNA-Phe* (*ppa*2454), *tRNA-Pro* (*ppa*2428), *tRNA-Thr* (*ppa*2412). У *Corynebacterium efficiens* длинные шпильки следуют за *tRNA-Ala, tRNA-Arg, tRNA-Asp, tRNA-Leu, tRNA-Pro, tRNA-Ser*.

Авторы благодарят М.С. Гельфанда за обсуждение и ценные замечания. В создании программного обеспечения принял участие М.А. Ширшин, которому авторы глубоко благодарны. Работа частично поддержана грантом ISTC 2766.

СПИСОК ЛИТЕРАТУРЫ

1. A.V. Seliverstov, H. Putzer, M.S. Gelfand, V.A. Lyubetsky. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiology*, 2005, 5: 54, 14p.

2. A.G. Vitreschak, E.V. Lyubetskaya, M.A. Shirshin, M.S. Gelfand, V.A. Lyubetsky. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis. *FEMS Microbiol Lett.*, 2004, 234, p. 357–70.

3. T. Washio, J. Sasayama, M. Tomita. Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Research*, 1998, 26, p. 5456–5463.

4. S. Unniraman, R. Prakash, V. Nagaraja. Conserved economics of transcription termination in eubacteria. *Nucleic Acids Research*, 2002, 30, p. 675–684.

5. I.M. Ishchukov, V.A. Likhoshvai, Yu.G. Matushkin. A new algorithm for recognizing the operon structure of procaryotes. *Proceedings of the fourth international conference on bioinformatics of genome regulation and structure*. Новосибирск: ред.-изд. отдел ИЦиГ, 2004, p. 73-76.

6. М. Сингер, П. Берг. *Гены и геномы*. Т.1. М.: Мир, 1998.

7. Д.Г. Кнорре, С.Д. Мызина. *Биологическая химия*, М.: Высшая школа, 2003.