

Поиск регуляторных сайтов в группах ортологичных генов гамма-протеобактерий

Л.В.Данилова*, М.С.Гельфанд**

*Институт проблем передачи информации РАН,
101447, Москва, Большой Каретный переулок, 19, Россия
e-mail: dlv2k@mail.ru, тел. (095) 900-3119, факс: (095) 209-0579

**Integrated Genomics-Moscow
e-mail: gelfand@integratedgenomics.ru

Поступила в редколлегию 06.05.2002

Аннотация—Для подтверждения адекватности алгоритма поиска сигнала, мы применили его к областям перед ортологичными генами из 8 родственных гамма-протеобактерий. Полученные результаты сравнили с набором известных сайтов и обнаружили, что из 311 известных сайтов было найдено 99.

1. ВВЕДЕНИЕ

До сих пор задача выделения регуляторных сигналов в наборах нуклеотидных последовательностей остается одной из актуальных проблем молекулярной биологии, и существуют разные подходы к ее решению. В частности, в качестве исходных данных для поиска сигналов берутся регуляторные области у ортологичных генов родственных геномов [1],[2]. Вполне логично предположить, что гены, образующие ортологичный ряд, регулируются одинаково, т.е. существует общий регулятор, который узнает общий сигнал. Это не всегда верно, но в большинстве случаев происходит именно так, поэтому выборка формировалась, исходя из этого предположения. Целью данной работы было тестирование алгоритма на такого рода данных для организмов родственных *Escherichia coli*, с тем, чтобы применять его для поиска новых сигнала у других наборов родственных организмов.

2. ДАННЫЕ И МЕТОДЫ

Для данного исследования были извлечены фрагменты последовательностей, соответствующих ортологичным рядам, из полных геномов гамма протеобактерий *Escherichia coli*, *Escherichia coli* O157, *Salmonella typhi*, *Salmonella typhimurium*, *Yersinia pestis*, *Vibrio cholerae*, *Haemophilus influenzae*, *Pasteurella multocida*.

Пара генов из двух геномов считались ортологичными, если эти два гена являлись наиболее близкими родственниками друг для друга в этих двух геномах. Далее пары ортологов объединялись в ряды, при этом не требовалась транзитивность и игнорировались малые различия в уровне сходства (так что, вообще говоря, одному гену могли соответствовать более одного ортолога в другом геноме).

Для анализа бралась область длиной 200 нуклеотидов до начала гена или весь межгенный интервал, если он был короче 200 нуклеотидов.

После этого была произведена процедура фильтрации для того, чтобы из каждой группы очень сходных генов оставить только одного представителя (по возможности, из *E. coli*). Это было сделано для того, чтобы поиск консервативных регуляторных сигналов мог происходить без интерференции

со стороны недостаточно дивергировавших последовательностей из близких геномов (штаммов). Критерием похожести являлось совпадение 35 нуклеотидов из 40 подряд идущих.

В тестировании участвовали выборки состоявшие из трёх или более последовательностей, после фильтрации таковых оказалось 1967. Далее из базы данных dpinteract [3] были взяты 345 последовательностей известных регуляторных сайтов *E. coli* длиной от 10 до 22 нуклеотидов и проверены на наличие в выборках, причем сайты рассматривались в обоих направлениях относительно последовательности ДНК. Оказалось, что присутствуют 239 последовательностей, причем 2 из них встретились 4 раза, 5 — 3 раза, 55 — 2 раза и 179 — 1 раз, таким образом, необходимо было найти 311 сайтов. Остальных не было, либо, из-за того, что сайт лежит на расстоянии более, чем 200 нуклеотидов от начала гена или находится в кодирующей области, либо из-за отсутствия ортологов у соответствующего гена.

Была написана программа на основе алгоритма поиска регуляторного сигнала [4],[5], которая на вход получает несколько последовательностей, в данном случае от 3 до 40, длиной от 40 до 200 нуклеотидов, и выдает систему попарно похожих между собой слов по одному из каждой последовательности. Во время работы программы производится оптимизация функции качества, которая определяется через попарное сходство слов. Функция качества также может учитывать особенности самих слов, такие, как палиндромность. В данной работе алгоритм применялся для поиска сигнала длиной 15, 20, 22, а также палиндромного сигнала длины 15, 16 и 22.

3. РЕЗУЛЬТАТЫ

В таблице (см. следующую страницу) представлены результаты поиска сайтов. Из 311 известных сайтов, присутствующих в выборке, нашлось 99, причем найденными считались только сайты либо совпавшие со словом, которое выдал алгоритм, либо полностью входившие в него. Остальные сайты не были найдены или потому, что сигнал был слабым, или потому, что ортологичный ген потерял регуляцию.

Мы благодарны П. Новичкову за помощь в подготовке данных и В.А. Любецкому, научному руководителю Л.В. Даниловой, К.Ю. Горбунову и А.А. Миронову за полезные обсуждения.

СПИСОК ЛИТЕРАТУРЫ

1. McCue Lee Ann, Thompson William, Carmack C. Steven, Ryan Michael P., Liu Jun S., Derbyshire Victoria and Lawrence Charles E. Phylogenetic footprinting of transcription factor binding site in proteobacterial genomes. *Nucleic Acids Research*, 2001, vol. 29, no. 3, pp. 774–782.
2. Terai Goro, Takagi Toshihisa and Nakai Kenta. Prediction of co-regulated genes in *Bacillus subtilis* on the basis of upstream elements conserved across three closely related species. *Genome Biology*, 2001, vol. 2, no. 11, pp. 54–69.
3. Robison K., McGuire A.M., Church G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.*, 1998, vol. 284, pp. 241–254.
4. Выюгин В.В., Горбунов К.Ю., Любецкий В.А. Алгоритмы выявления регуляторного сигнала и построения эволюционных деревьев. *Проблемы управления и моделирования в сложных системах, тезисы конференции*, РАН, 2000, стр. 130–137.
5. Данилова Л.В., Горбунов К.Ю., Гельфанд М.С., Любецкий В.А. Алгоритм выделения регуляторных сигналов в последовательностях ДНК. *Молекулярная биология*, 2001, том 35, № 6, стр. 987–995.

Статью представил к публикации член редколлегии В.А. Любецкий

Регулятор	Количество известных сайтов	Кол-во известных сайтов, присутствующих в выборках	На прямой цепи	На Обратной цепи	Найденные сайты
arcA	14	9	2	7	9
argR	17	20	14	6	3
cpxR	12	6	4	2	2
crp	49	41	26	15	3
cspA	4	6	3	3	1
cynR	2	4	2	2	1
cytR	5	4	4	0	0
deoR	3	1	1	0	0
dnaA	8	4	3	1	1
fadR	7	9	7	2	1
farR	4	8	2	6	4
fnr	14	10	6	4	1
fruR	12	6	3	3	4
fur	9	9	6	3	4
galR	7	4	3	1	2
gcvA	4	1	1	0	1
glpR	13	14	9	5	4
hns	15	11	6	5	4
hu	3	1	1	0	0
iclR	2	1	0	1	1
lacI	3	0	0	0	0
lexA	19	17	13	4	9
malT	10	17	9	8	5
melR	2	4	2	2	0
metJ	15	20	13	7	11
metR	8	10	7	3	1
narL	11	7	4	3	1
narP	8	10	6	4	0
ntrC	5	4	3	1	1
ompR	9	7	5	2	2
pdhR	2	2	2	0	1
purR	22	15	12	3	8
rpoN	6	4	3	1	3
torR	4	12	8	4	6
tyrR	17	13	13	0	5
Всего:	345	311	203	108	99

Таблица. Результаты поиска регуляторного сигнала ортологичных генов родственных бактерий.