Секция «Биоинформатика»

Могут ли два гена бактерии быть закодированными в одном месте на противоположных цепях ДНК?

Мошенский Денис Михайлович
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: moshenskydenis@gmail.com

Известно, что на одном участке генома вируса могут располагаться два гена, один напротив другого: гены orf-401 с координатами 19650-20855 и orf206b с координатами c20767-20147 в *Enterobacteria phage lambda* [1]. Такое явление встречается и у бактерий, однако в известных нам случаях оно объясняется, скорее всего, ошибками в аннотациях. Например, в геноме *Chlorobaculum tepidum* (идентификатор NCBI NC_002932) напротив гена с идентификатором RefSeq WP_010932651 была найдена открытая рамка считывания. Для гена не было найдено гомологов с помощью blast, однако для открытой рамки было найдено 12 последовательностей с совпадающими позициями $> 60\%$ и E-value $<$ 1E-121.

Цель работы - исследовать, почему на цепи, комплементарной к кодирующей, находятся длинные открытые рамки считывания и найти среди них гены.

Для решения поставленной задачи мы написали программу **ReverseORFs**, которая ищет все длинные открытые рамки считывания на цепи ДНК, комплементарной к кодирующей цепи, и для каждой из них оценивает вероятность появления такой рамки.

ReverseORFs перебирает все гены во входящем геноме, расширяет их координаты (для того, чтобы учесть открытые рамки считывания, которые не полностью перекрываются с кодирующей последовательностью) и удаляет открытые рамки, длина участка перекрывания которых с геном меньше заданной. На выходе программа выдает таблицу с информацией об отобранных рамках. Также для каждой отобранной открытой рамки проводится поиск обоснования существования гена PE (Uniprot Protein Evidence), напротив которого она расположена, и рассчитывается уровень значимости (P-value).

P-value открытой рамки считывания без сдвига, то есть кодоны гена и открытой рамки располагаются напротив друг друга, рассчитывается таким образом:

P = [(1 - (v(TTA) + v(CTA)))^N(L)]*[(1 - v(TCA))^N(S)],

где N - количество лейцина или серина в перекрывающейся части кодирующей последовательности;

v - частота кодонов TTA, CTA или TCA, взятая из частот кодонов для данного организма.

Кодоны TTA, CTA и TCA рассматриваются потому, что напротив них располагается стоп-кодон открытой рамки считывания.

Для открытых рамок со сдвигом рассматриваются пары смежных нуклеотидов для того, чтобы учесть несовпадение в позициях кодонов.

P-value также рассчитывается и для гена, только за исходную последовательность берется открытая рамка считывания, расположенная напротив обсчитываемого гена.

При параметрах: расширение координат гена - **180** нуклеотидов; длина участка перекрывания - **180** нуклеотидов - в 5-ти геномах было найдено 1345 открытые рамки считывания на некодирующей цепи с P-value $<$ 0.0001, из которых 182 имеют P-value $<$ 1E-7.

Для генов с PE 1 и 2 было найдено 18 находок с P-value < 0.001

С помощью blast для открытой рамки считывания, расположенной напротив одного из первых 5 генов с PE 1 и 2, отсортированных в порядке возрастания P-value, была найдена 1 гомологичная последовательность. Для рамок напротив 5 генов с PE 3 и 4 было обнаружено 14 гомологов.

Источники и литература

1) http://www.ncbi.nlm.nih.gov/nuccore/9626243?report=graph