Секция «Биоинформатика»

Создание алгоритма на языке Python для поиска видоспецифических последовательностей бактерий

Научный руководитель – Мирошников Константин Анатольевич

Рассказова Полина Михайловна
Студент (магистр)
Московский физико-технический институт, Москва, Россия
E-mail: Polinchen98@mail.ru

Использование систем диагностики на основе ПЦР - один из наиболее удобных и востребованных методов детекции самых разных бактерий: с целью идентификации целевой группы, для детекции патогенов в пробе и многих других применений. Чтобы провести точную видоспецифическую диагностику методом ПЦР-анализа важно подобрать уникальный для таксономического вида участок генома и праймеры, исключающие неспецифическую амплификацию. На данный момент в открытом доступе есть большое количество данных полногеномного секвенирования, что дает возможность найти самый подходящий участок для этой цели. Кроме того, можно автоматизировать весь процесс, что существенно облегчит подбор видоспецифического участка в геноме, а следовательно, и разработку диагностикума.

Для осуществления поиска был разработан пайплайн для командной строки Linux. Основная часть программы реализована на языке Python с помощью специальной библиотеки Biopython, предназначенной для биоинформатических задач. Для работы пайплайна используется также программа BLAST.

Пайплайн содержит в себе 2 этапа:

1) Подготовка данных;

2) Поиск последовательностей.

На первом этапе скачиваются все геномы исследуемого рода с базы данных NCBI и формируются две локальные базы данных: "позитивная" и "негативная". В первой геномы только вида, уникальные последовательности которого нужно найти, во второй все остальные виды исследуемого рода бактерии. На этом же этапе выбирается геном типового штамма из "позитивной" базы и разрезается на более мелкие фрагменты.

На втором этапе осуществляется непосредственный поиск уникальных последовательностей. На вход программе подаются данные с первого этапа: файл с разрезанным геномом типового штамма и "негативная" база данных. Пайплайн производит выравнивание целевого генома против выбранной базы данных и записывает совпавшие последовательности. Затем программа создает 2 файла: hits и no_hits, в которые попадают последовательности, имеющие совпадение и нет с базой данных соответственно. Алгоритм работает так, что дублирующиеся фрагменты удаляются автоматически. Таким образом, получена информация о том, какие последовательности уникальны для типового штамма. Следующий шаг — это сравнение файла no_hits с "позитивной" базой данных. В завершении второго этапа мы получаем файл с последовательностями уникальными только для исследуемого вида. Далее полученную информацию можно использовать для подбора праймеров для ПЦР-диагностики.

Данный пайплайн оптимизирует и ускоряет работу по поиску уникальных для вида последовательностей. В будущем планируется сделать программу удобной для использования любым человеком, то есть исследователи, не обладающие знаниями в биоинформатике, тоже смогут подбирать праймеры для своего объекта изучения.