Секция «Биоинформатика»

Создание пайплайна для автоматизации поиска О-антигена грамотрицательных бактерий

Научный руководитель – Комиссаров Алексей Сергеевич

Чеснокова П.Д.[1], *Зилов Д.С.*[2]

1 - Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, *E-mail: tniapp@yandex.ru*; 2 - Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, *E-mail: zilov@scamt-itmo.ru*

Соматический О-антиген является компонентом наружной мембраны грамотрицательных бактерий. По своей природе это полисахарид, имеющий определенную специфичность, которая способствует идентификации бактерий. Разнообразие О-антигенных форм лежит в основе классификации путем выделения О-серогрупп микроорганизмов. Кроме того, О-антиген содержит специфические сайты связывания с антителами организма-хозяина - эпитопы, что обусловливает развитие иммунных реакций в ответ на инфицирование.

Гены, вовлеченные в биосинтез О-антигена, зачастую объединены в кластер. Несмотря на относительную простоту организации этого кластера его выявление в геноме микроорганизмов до сих пор остается сложным, а серотипирование основано на поиске нескольких генов из этого кластера, что значительно снижает точность. При этом большинство доступных инструментов органичены только геномом *Escherichia coli*. Наша цель - создание пайплайна, позволяющего автоматизировать поиск, сборку и аннотацию кластера генов, вовлеченных в синтез О-антигена.

Для создания этой программы была использована система управления рабочими процессами Snakemake. Мы выбрали этот инструмент, так как, с одной стороны, в него интегрирован менеджер пакетов Conda, который содержит в себе множество биоинформатических инструментов, необходимых в работе, а с другой стороны, можно использовать свои инструменты и скрипты.

На вход разработанному пайплайну подаются сырые данные секвенирования или геномная сборка. Программа автоматически оценивает качество исходных данных, убирая адаптеры и оптические дупликаты. Далее запускается сборка генома с использованием SPAdes и Unicycler. После этого геномная сборка проверяется на наличие контаминаций и плазмид. Очищенная от контаминаций сборка аннотируется с помощью Prokka и eggNOG. На основе полученной сборки создается полная функциональная аннотация генома, из которой вычленяются известные гены, вовлеченные в биосинтез О-антигена. Используя полученную аннотацию и совпадение с базой данных STRING, пайплайн аннотирует опероны, содержащие гены биосинтеза О-антигена, с учетом возможных ошибок сборки и разрывов контигов.

В настоящий момент мы проводим аннотацию О-антигенов всех грамотрицательных бактерий, доступных в базе данных NCBI. Полученная в результате информация будет использоваться для улучшения существующего пайплайна.

Созданная программа позволяет существенно экономить время при классификации грамотрицательных бактерий с идентификацией их серогрупп, а также при выявлении компонентов наружной мембраны, связанных как с вирулентностью микроорганизмов, так и с симбиотическими взаимоотношениями между бактериями и растениями.