METAFAST: ВЫСОКОПРОИЗВОДИТЕЛЬНЫЙ СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТАГЕНОМОВ НА ОСНОВЕ ГРАФА ДЕ БРЕЙНА

Казаков С.В. [a], Ульянцев В.И. [a], Дубинкина В.Б. [b,c], Тяхт А.В. [c], Алексеев Д.Г. [b,c]

[a] *Университет ИТМО, Санкт-Петербург, Россия*
[b] *Московский Физико-Технический Институт (Государственный Университет), Москва, Россия*
[c] *НИИ Физико-Химической Медецины, Москва, Россия*

svkazakov@rain.ifmo.ru

С развитием технологий высокопроизводительного секвенирования был накоплен огромный объем метагеномных данных. Эффективный статистический анализ таких данных требует получения их сжатого представления (извлечение признаков).

Традиционные методы извлечения признаков из данных полногеномного секвенирования микробиоты имеют существенные недостатки. Например, методы, основанные на картировании последовательностей на каталог известных геномов, требуют репрезентативной базы геномов, однако многие виды бактерии до сих пор не изучены и не имеют референсного генома. Методы, основанные на *de novo* сборке, требуют больших вычислительных ресурсов и трудноприменимы в случае метагеномов со сложным бактериальным составом (таких, как микробиота кишечника человека).

Нами было разработано программное средство *MetaFast* для сжатого представления метагеномов, не использующее априорное знание о микроорганизмах, которые могут содержаться в изучаемой среде. Преимущества подхода по сравнению с указанными аналогами – гибкость, скорость работы и экономия памяти.

Алгоритм состоит из следующих этапов:

1. Выделение коротких геномных последовательностей из ридов для каждого метагенома на основе анализа графа де Брейна.

2. Объединение последовательностей по всем метагеномам в один граф де Брейна, выделение компонент связности в нем. При этом для больших компонент связности производится их итеративное разделение на подкомпоненты. В дальнейшем каждая компонента используется как единичный признак.

3. Вычисление вектора признаков для каждого метагенома основанного на покрытии k-мерами каждой компоненты.

4. Попарное сравнение метагеномов путем расчета матрицы расстояния между ними на основе полученных векторов признаков с использованием индекса Брея-Кёртиса.

Были произведены экспериментальные исследования предлагаемого подхода как на искусственных данных, так и на реальных. Результаты исследований показывают высокую корреляцию матриц расстояний, полученных *MetaFast*'ом и методом, основанным на картировании последовательностей на каталог известных геномов (корреляция Спирмена r=0.96, p-value=0.001).

Исходный код и исполняемый пакет *MetaFast* доступен по адресу https://github.com/ctlab/metafast.