S04-08
METAFAST – ПРОГРАММНОЕ СРЕДСТВО ДЛЯ ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СРАВНИТЕЛЬНОГО АНАЛИЗА МЕТАГЕНОМОВ

Ульянцев В.И.[1], Казаков С.В.[1], Дубинкина В.Б.[2], Тяхт А.В.[3], Алексеев Д.Г.[3]
[1]Университет ИТМО. Санкт-Петербург, Россия, [2]Московский Физико-Технический Институт (Государственный Университет), [3]НИИ Физико-Химической Медицины

С развитием технологий высокопроизводительного секвенирования был накоплен огромный объем метагеномных данных. В методах метагеномного анализа ключевую роль играет получение сжатого представления (извлечение признаков) для эффективного сравнения метагеномов.

Для данных по полногеномному секвенированию микробиоты традиционные методы извлечения признаков используют либо картирование последовательностей, либо сборку *de novo*, которые требуют больших вычислительных затрат 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 и экономия памяти. Алгоритм состоит из следующих этапов.

Выделение коротких геномных последовательностей из ридов для каждого метагенома на основе анализа графа де Брейна. Выделенные последовательности являются аналогами контигов, получаемых в результате сборки *de novo*, однако последовательности в разы короче.

Объединение последовательностей по всем метагеномам и выделение компонент связности в обобщенном графе де Брейна. При этом для больших компонент связности производится их итеративное разделение на подкомпоненты.

Построение вектора признаков для каждого метагенома путем подсчета суммарного числа вхождений *k*-меров из каждой компоненты в метагеном и объединения этих значений от всех компонент в вектор.

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

MetaFast способен за несколько часов работы кластера (6 TFLOPS) проанализировать и построить матрицу расстояний и кладограмму по ней между 150 метагеномами микробиоты кишечника человека (суммарно 700 Гб ридов), что на порядок быстрее методов картирования на референсный каталог геномов и сборки метагеномов *de novo*, а также требует меньше оперативной памяти, чем последний метод. Исходный код и исполняемый пакет *MetaFast* доступен по адресу https://github.com/ulyantsev/metafast.