УДК 004.891.2

РАЗРАБОТКА БИБЛИОТЕКИ АЛГОРИТМОВ ИЗВЛЕЧЕНИЯ ПРИЗНАКОВ ИЗ ГРУПП ПОЛНОГЕНОМНОГО СЕКВЕНИРОВАНИЯ МЕТАГЕНОМНЫХ ОБРАЗЦОВ

Иванов А.Б. (Университет ИТМО)
Научный руководитель – канд. техн. наук Ульянцев В.И. (Университет ИТМО)

Введение.

Микробные сообщества населяют различные ниши окружающего мира, такие как водоемы, почва и организм человека. В человеке бактерии участвуют в переработке и усвоении питательных веществ, и регуляции иммунного ответа. Изучением всех микробов в совокупности из одного образца занимается метагеномика. Развитие современных методов полногеномного метагеномного секвенирования позволило извлекать всю информацию из метагеномного образца и изучать все микробное разнообразие и их взаимодействия, в то время как культивирование отдельных бактерий в лаборатории является нерешенной задачей для многих видов.

Изучение микробиоты кишечника является актуальной задачей, поскольку она играет важную роль в организме человека. Последние исследования показывают, что микробиом кишечника влияет на успешность иммунотерапии раковых заболеваний [1], развитие воспалительных заболеваний кишечника [2] и другие заболевания.

Для проведения сравнительного анализа метагеномных образцов существуют различные методы. Одним из них является извлечение из образцов информации о том, какие известные микроорганизмы содержатся в них (таксономическая аннотация) и какие метаболические функции они выполняют (функциональная аннотация) [3]. Недостатком данных подходов является возможность анализировать только ту часть образца, информация о которой находится в базах данных, что может приводить к потере существенного объема данных и пропуску ключевых биологических факторов. Другая группа методов позволяет анализировать образцы на основе коротких последовательностей k-меров [4]. Это позволяет исследовать весь объем данных, однако получаемые результаты плохо поддаются биологической интерпретации. Кроме того, большинство классических методов не используют вспомогательные метаданные, такие как разбиение метагеномных образцов по группам на основе заболевания или других признаков, в то время как эта информация может помочь более качественно сравнить исследуемые метагеномы.

Основная часть.

Целью данной работы является повышение качества сравнительного анализа групп метагеномных образцов. Ранее нами были предложены несколько методов безреференсного анализа метагеномных образцов. Они были реализованы в виде вычислительных алгоритмов в программном средстве MetaFast (https://github.com/ctlab/metafast/).

В данной работе производится разработка открытой библиотеки для анализа групп метагеномных образцов полногеномного секвенирования. Библиотека будет состоять из трех типов алгоритмов и позволит проводить полный цикл анализа метагеномов от сырых прочтений до интерпретируемых признаков и результатов сравнительного анализа.

Первая часть библиотеки содержит алгоритмы для извлечения признаков и проведения сравнительного анализа групп метагеномных образцов без вспомогательных метаданных. На первом этапе планируется поддержка двух алгоритмов.

1. Классический алгоритм извлечения компонент из графа де Брейна, реализованный в программе MetaFast [5]. Затем рассчитывается покрытие компонент к-мерами из образцов и строится матрица расстояний между образцами.
2. Метагеномная сборка контигов с помощью программы metaSpades [6]. Затем контиги используются в качестве отдельных компонент в программе MetaFast.

Вторая часть библиотеки содержит алгоритмы для извлечения признаков из метагеномных образцов, которые значимо различаются между разными категориями образцов. При этом

для извлечения данного типа признаков алгоритмы используют информацию о группах образцов из метаданных. Планируется поддержка следующих алгоритмов.

1. Извлечение уникальных к-меров для каждой группы метагеномных образцов из набора данных. Затем вокруг каждой группы уникальных к-меров строится локальный граф де Брейна и извлекаются признаки.
2. Извлечение к-меров со значимым различием во встречаемости между группами образцов с использованием статистических тестов хи-квадрат и Манна-Уитни. Затем вокруг каждой группы статистически значимых к-меров строится локальный граф де Брейна и извлекаются признаки.
3. Подсчет встречаемости каждого к-мера в образцах различных групп и построение общего графа де Брейна. Разбиение графа де Брейна на компоненты с использованием раскраски вершин на основании частоты встречаемости к-меров.

Третья часть библиотеки содержит алгоритмы и пайплайны для проведения кластеризации и классификации метагеномных образцов с использованием методов машинного обучения на основе извлеченных признаков.

Разработанная библиотека будет распространяться в открытом доступе и будет удобна для проведения полного цикла вычислительного сравнительного анализа метагеномных образцов. Она может быть применена не только к данным микробиоты кишечника человека, но и к любым образцам из других сред, полученным с помощью полногеномного метагеномного секвенирования коротких прочтений.

Выводы. В данной работе была разработана открытая библиотека алгоритмов извлечения признаков из данных полногеномного секвенирования. В дальнейшем планируется ее использование для анализа массива данных из открытых источников как для извлечения новых знаний из данных, так и для улучшения и доработки библиотеки.

Список использованных источников:

1. Frankel A. E. et al. Metagenomic shotgun sequencing and unbiased metabolomic profiling identify specific human gut microbiota and metabolites associated with immune checkpoint therapy efficacy in melanoma patients //Neoplasia. – 2017. – Т. 19. – №. 10. – С. 848-855.
2. Franzosa E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease //Nature microbiology. – 2019. – Т. 4. – №. 2. – С. 293-305.
3. Truong D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling //Nature methods. – 2015. – Т. 12. – №. 10. – С. 902-903.
4. Wang Y. et al. KmerGO: a tool to identify group-specific sequences with k-mers //Frontiers in microbiology. – 2020. – Т. 11. – С. 2067.
5. Ulyantsev V. I. et al. MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data //Bioinformatics. – 2016. – Т. 32. – №. 18. – С. 2760-2767.
6. Nurk S. et al. metaSPAdes: a new versatile metagenomic assembler //Genome research. – 2017. – Т. 27. – №. 5. – С. 824-834.