

**ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ ИЗ МЕТАГЕНОМНЫХ ДАННЫХ И
КЛАССИФИКАЦИЯ ОБРАЗЦОВ С ПОМОЩЬЮ БИБЛИОТЕКИ METAFX**

Иванов А.Б. (Университет ИТМО)

Научный руководитель – канд. техн. наук Ульянов 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 работы является улучшение сравнительного анализа метагеномных образцов за счет извлечения интерпретируемых признаков и применения методов машинного обучения для классификации.

Библиотека алгоритмов разделяется на две группы. Модули из первой группы разработаны с целью извлечения признаков из метагеномных образцов в виде подграфов графа де Брейна.

Для построения признаков, наиболее релевантных задаче классификации образцов на категории, на этапе извлечения признаков используется информация о принадлежности образцов к разным группам. Реализованы следующие алгоритмы:

1. Алгоритм на основе уникальных k -меров. Для каждой категории образцов извлекаются уникальные k -меры, присутствующие в подавляющем большинстве образцов данной группы. Затем они объединяются в признаки в виде компонент графа де Брейна.
2. Алгоритм на основе статистически значимых k -меров. Для каждой категории образцов извлекаются специфические k -меры с использованием статистических тестов хи-квадрат и Манна-Уитни. Затем они объединяются в признаки в виде компонент графа де Брейна.

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.

Модули из второй группы позволяют обучать модели искусственного интеллекта и строить классификационные модели на основе извлеченных признаков. Поддерживаются алгоритмы случайного леса деревьев решений из библиотеки scikit-learn, бустинга из библиотеки XGBoost и нейронных сетей из библиотеки PyTorch.

Работа с графами и к-мерами реализована на языке программирования Java, а с классификационными моделями – на языке Python3. Реализация библиотеки доступна по адресу <https://github.com/ctlab/metafx>.

Выводы. В данной работе была разработана открытая библиотека алгоритмов извлечения признаков из данных полногеномного секвенирования и построения моделей машинного обучения для классификации образцов на основе извлекаемых признаков. Она показала высокую точность как на тестовых данных, так и на задачах обработки реальных образцов метагеномного секвенирования кишечника человека.

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.

Иванов А.Б. (автор)

Ульянцев В.И. (научный руководитель)