УДК 004.891.2

СТАТИСТИЧЕСКИЙ ОТБОР ПРИЗНАКОВ ДЛЯ КЛАССИФИКАЦИИ ГРУПП МЕТАГЕНОМНЫХ ОБРАЗЦОВ В ПРОГРАММЕ METAFAST.

Попов В.В. (СПбГУ) **Научный руководитель – к.т.н. Ульянцев В.И.**
(Университет ИТМО)

Аннотация.
Данная работа посвящена решению задачи сравнительного анализа метагеномов на примере классификации групп метагеномных образцов микробиоты кишечника. В качестве основного метода для сравнения метагеномов взята программа MetaFast. Предложена модификация MetaFast для извлечения метагеномных признаков с использованием статистических методов. Полученная программа протестирована на различных реальных метагеномных данных.

Введение.

Метагеномика — раздел науки, который занимается изучением микробных сообществ, населяющих различные ниши окружающей среды. Примером является кишечник человека, который населяет огромное число бактерий, играющих важную роль в организме человека. Воспалительные заболевания кишечника (ВЗК) поражают до 0.3% людей, поэтому необходимо развитие методов для диагностирования и понимания причин развития заболеваний. Перспективным методом может быть анализ данных, полученных из микробиоты кишечника. Одной из подходящих программ для осуществления такого анализа является MetaFast. Работа метода устроена следующим образом: для каждого образца подсчитываются k-меры, и извлекаются только те, которые присутствуют в некотором заданном числе образцов из одной категории и отсутствуют в всех других. Вокруг выделенных k-мер для каждой группы строится граф де Брейна. Далее графы разбиваются на компоненты, и для каждого образца подсчитывается покрытие каждой компоненты k-мерами. Таким образом получается информация о представленности компонент, которая может быть использована для сравнения образцов между собой. В работе предложена модификация алгоритма MetaFast с использованием методов статистического анализа для более качественного выделения признаков.

Основная часть.

Программа MetaFast протестирована на реальных данных, показана перспективность ее развития. Тем не менее, в текущей реализации алгоритма выделения уникальных k-меров введен полный запрет на присутствие k-меров в образцах других категорий, что является недостатком программы. В образцах могут быть незначительные загрязнения, попавшие туда по ошибке, которые, однако, не влияют на общие функции и признаки сообщества. Возникает идея выделения специфичных k-меров, для которых допускается возможность попадания в другие классы в небольших пропорциях.

Предполагаемая реализация может выглядеть следующим образом: программа будет принимать на вход все метагеномные образцы с известными классами и для каждого k-мера выдавать набор чисел, соответствующий числу образцов из категории, в которых встречается данный k-мер. На следующем этапе можно выделить специфичные для каждой категории k-меры, взяв, например, только те k-меры, распределения частот которых статистически значимо отличаются в зависимости от группы. В качестве параметра можно взять необходимый уровень значимости для признания k-мера специфичным.

Выводы.

В представленной работе был реализован алгоритм извлечения статистически значимых k-мер для групп метагеномов как улучшение программы MetaFast и проверен на тестовых наборах данных. Полученный алгоритм позволяет получать высококачественные признаки, которые в дальнейшем могут быть использованы для классификации метагеномных образцов, в частности, для диагностики заболеваний.

Попов В.В. (автор) Подпись

к.т.н. Ульянцев В.И. Подпись