УДК 004.891.2

МЕТОДЫ ВЫДЕЛЕНИЯ ПРИЗНАКОВ ИЗ МЕТАГЕНОМНЫХ ДАННЫХ ДЛЯ КЛАССИФИКАЦИИ ВОСПАЛИТЕЛЬНЫХ ЗАБОЛЕВАНИЙ КИШЕЧНИКА

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

Проведена разработка методов для выделения признаков (в том числе с использованием k-меров и графа де Брейна) из метагеномных данных пациентов с воспалительными заболеваниями кишечника (ВЗК) с целью выявления специфичных маркеров заболеваний. Предложенные методы применены для обработки данных метагеномного секвенирования из открытых источников и классификации образцов. Полученные результаты показывают высокую точность и могут быть использованы как вспомогательный метод при неинвазивном диагностировании ВЗК.

Введение.

Воспалительные заболевания кишечника (ВЗК), среди которых выделяют язвенный колит (ЯК) и болезнь Крона (БК), активно распространяются в индустриальных странах и поражают до 0,3 % людей по всему миру. Заболевание может развиваться в результате комбинации различных факторов, среди которых выделяют генетическую предрасположенность, образ жизни, состояние микробиома кишечника и другие. Микробиота кишечника человека играет важную роль в регуляции обмена веществ и иммунном ответе организма и изменяется в результате воздействия внешних факторов или заболеваний.

В настоящее время стандартом для диагностирования ВЗК является биопсия при колоноскопии, однако это инвазивная процедура, поэтому ведется поиск методов для неинвазивного диагностирования. Текущие исследования направлены на выявление зависимостей между таксономическим и функциональным составом микробиоты кишечника и диагнозом. Однако полученные ранее результаты не обладают достаточным уровнем точности, что ведет к необходимости поиска новых признаков в метагеноме для более точной диагностики и помощи в клиническом ведении и лечении ВЗК.

Основная часть.

В данной работе проводится анализ коротких прочтений ДНК, полученных в результате секвенирования метагеномных сообществ с помощью секвенаторов второго поколения (next-generation sequencing, NGS). Используется подход, основанный на выделении из прочтений коротких участков (k-меров) и работы непосредственно с ними или объединении их в граф де Брейна. В отличие от таксономических и функциональных методов анализа, данное решение позволяет работать непосредственно с информацией о последовательностях ДНК организмов, населяющих кишечник человека. Преимуществом данного подхода является обработка всего массива полученных данных без потери какой-либо информации. При условии обнаружения специфичных маркеров, которые позволяют достичь высокой точности классификации, возможен их дальнейший анализ для перехода на уровень таксономии и установления биологически-значимых факторов.

Были рассмотрены различные варианты для выделения признаков из метагеномных образцов:

1. С помощью программы MetaFast строится совместный граф де Брейна для всех образцов, который затем разбивается на компоненты. Затем производится оценка содержания каждой компоненты в каждом образце, в результате каждому образцу сопоставляется численный вектор признаков. Далее эти признаки совместно с известными диагнозами используются для построения моделей машинного обучения на основе случайного леса деревьев решений. Полученные модели обучаются на

образцах с известными классами, обученные модели затем могут быть использованы для классификации новых метагеномных образцов.

2. Метагеномные образцы от людей с известными диагнозами разбиваются на категории по заболеваниям. Из каждой категории выделяются k-меры, специфичные только для данной категории и отсутствующие в других категориях. Таким образом, каждому диагнозу сопоставляется набор специфичных для него k-меров. Полученные наборы затем используются для классификации образцов с неизвестным диагнозом.
3. Предыдущее решение улучшено для учета возможности преобладания k-меров в одной из категорий и при этом не полного отсутствия в другой. Для каждого k-мера подсчитывается число образцов из каждой категории, в которых он содержится. Затем применяется масштабирование, и отбираются только k-меры, которые преобладают в одной из категорий. Они используются в дальнейшем для классификации новых образцов.

Все методы были применены к данным метагеномного секвенирования из трех открытых источников. Всего было проанализировано 390 образцов с известными диагнозами (174 ЯК, 93 БК, 123 отсутствие ВЗК). Для оценки точности получаемых результатов были применены следующие метрики: коэффициент корреляции Мэттьюса и площадь под графиком точность-полнота. Полученные результаты значительно превосходят случайные предсказания в задачах двухклассовой классификации и были использованы в международном «Исследовании по метагеномной диагностике при воспалительных заболеваниях кишечника» (Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge, https://www.intervals.science/resources/sbv-improver/medic), где показали лучшие результаты среди всех участников.

Выводы.

Были разработаны и реализованы методы выделения признаков из метагеномных данных для диагностирования воспалительных заболеваний кишечника. Данные методы показали высокую точность при сравнении с другими разрабатываемыми в текущий момент подходами и могут быть использованы для помощи при мониторинге и диагностировании ВЗК.

Иванов А.Б. (автор) _____

Ульянцев В.И. (научный руководитель) _____