УДК 616.314.17-008.1:579.61:004.8

ПРЕДСКАЗАНИЕ ПАРОДОНТИТА НА ОСНОВЕ АГРЕГИРОВАННОЙ ВЫБОРКИ WGS-МЕТАГЕНОМНЫХ ДАННЫХ МИКРОБИОМА СЛЮНЫ ЧЕЛОВЕКА

Добровольский С.К.[1]
Научные руководители - Ракитько А.С.[2], Денисова А.А.[2]
[1]Прикладная геномика, ПИШ ИТМО
[2]Центр Genotek: ИИ в персонализированной медицине, ИТМО, Санкт-Петербург
Serafimin2002@gmail.com

Введение

Пародонтит - распространенное хроническое воспалительное заболевание полости рта. Его патогенез тесно связан с дисбалансом орального микробиома, что делает образцы слюны перспективным источником неинвазивных диагностических маркеров. При этом разработка предиктивных моделей требует отбора и валидации признаков, устойчивых между независимыми исследованиями и различными популяциями [1].

Полногеномное секвенирование (shotgun WGS) обеспечивает более высокое таксономическое разрешение по сравнению с ампликонным 16S рРНК и дополнительно позволяет проводить функциональное профилирование метагенома. Показано, что даже shallow-shotgun WGS с низкой глубиной покрытия воспроизводит основные таксономические паттерны 16S и одновременно предоставляет функциональную информацию о микробиоме [2]. Несмотря на то что публично доступных WGS-датасетов с клинической разметкой на сегодняшний день значительно меньше, чем 16S-данных, именно полногеномный подход открывает более широкие перспективы для построения диагностических моделей. Практическая значимость исследования обусловлена наличием у компании Генотек десятков тысяч метагеномных профилей слюны, представляющих целевую выборку для применения разрабатываемых предсказательных моделей.

Основная часть

Цель работы - построение и валидация предсказательных ML-моделей риска пародонтита на основе агрегированной выборки WGS-метагеномных данных микробиома слюны из публичных репозиториев с последующим применением обученных моделей к накопленным образцам компании Генотек.

Поскольку прямое обучение на данных Генотек невозможно из-за отсутствия верифицированной клинической разметки, принято решение провести метаанализ публичных исследований с клинически подтвержденными диагнозами и перенести обученные модели на данные компании. Приоритет отдан пародонтиту как заболеванию с наибольшим числом WGS-образцов слюны с клинической разметкой в открытом доступе.

Сформирован объединенный датасет из 9 публичных репозиториев (SRA, ENA, CNGB), включающий 367 образцов с примерно равным распределением между больными пародонтитом и здоровыми людьми. Все образцы обрабатывались единым метагеномным пайплайном: техническая очистка ридов, таксономическая и функциональная аннотация. Центральной методологической проблемой является батч-эффект, обусловленный гетерогенностью исходных исследований. Рассматриваются три стратегии его учета: (1) LOSO-валидация без коррекции, при которой каждое исследование поочередно используется только как тестовая выборка; (2) предварительная коррекция батч-эффекта специализированными методами для микробиомных данных; (3) включение принадлежности к исследованию как

дополнительного признака в модель. Выбор оптимальной стратегии коррекции батч-эффекта не является универсальным и определяется характеристиками конкретного набора данных [3], в связи с чем все три подхода будут систематически сопоставлены в рамках настоящего исследования.

В качестве признаков для обучения моделей используются таксономические и функциональные профили как по отдельности, так и в комбинации. Входные данные представляют собой разреженные таблицы относительных представленностей с высокой долей нулевых значений и композиционной природой признаков, что накладывает ограничения на выбор методов анализа и требует специализированных подходов к нормализации. Дополнительным ограничением является относительно небольшой объем публично доступных WGS-образцов с клинической разметкой, обусловленный высокой стоимостью метода.

Выводы

Разрабатываемый подход позволяет строить предсказательные модели пародонтита по WGS-метагеномным данным микробиома слюны, обученные на агрегированной публичной выборке с клинической разметкой и применимые к массиву образцов компании Генотек. Результаты представляют практический интерес для создания диагностического инструмента, выявляющего риск развития пародонтита непосредственно из данных полногеномного секвенирования слюны без дополнительных клинических обследований. Дальнейшей задачей является разработка протокола косвенной валидации предсказаний, включая анализ согласованности с анкетными данными и, при возможности, клиническую верификацию на части образцов.

Литература

1. Geng M., Li M., Li Y. et al. A universal oral microbiome-based signature for periodontitis // iMeta. – 2024. – Vol. 3. – e212. doi: 10.1002/imt2.212

2. Stothart M.R., McLoughlin P.D., Poissant J. Shallow shotgun sequencing of the microbiome recapitulates 16S amplicon results and provides functional insights // Molecular Ecology Resources. – 2023. – Vol. 23. – P. 549–564. doi: 10.1111/1755-0998.13713

3. Lee S., Lee I. Comprehensive assessment of machine learning methods for diagnosing gastrointestinal diseases through whole metagenome sequencing data // Gut Microbes. – 2024. – Vol. 16, № 1. – 2375679. doi: 10.1080/19490976.2024.2375679