Метрическая индексация ампликонных библиотек гена 16S рРНК

Ромащенко Николай Сергеевич

научный руководитель: к.б.н. Е. Е. Андронов

СПбАУ РАН

15 июня 2017 г.

Введение

- Метагеномные чтения гена 16S rRNA активно используются для анализа биоразнообразия микробных сообществ
- Число ампликонных библиотек 16S быстро растет
- Большинство библиотек плохо проаннотированы

Введение

Наиболее популярные программные пакеты для анализа биоразнообразия ампликонных библиотек 16S:

- QIIME
- mothur
- R: phyloseq
- . . .

Проблемы:

- Нет готового решения метрической индексации для поиска
- Все пакеты используют метрики, основанные на OTU-picking, то есть кластеризации чтений
- Для плохо изученных микробиомов OTU-picking вычислительно трудоемок

Цель и задачи

Цель:

- Создать готовое решение для метрической индексации и последующего поиска ампликонных библиотек 16S

Задачи:

- Разработать поисковый индекс для библиотек 16S, не зависящий от метрики
- Реализовать метрику, основанную на подсчете k-меров
- Сравнить производительность индексации для метрик, основанных на OTU-picking и основанных на подсчете k-меров

Amquery

Исходный код:
https://github.com/nromashchenko/amquery

- Конфигурирование
- Предобработка
  - Фильтрация ридов
  - Метрико-специфичная предобработка
- Индексация библиотек
- Использование
  - Поиск ближайших по метрике
  - Добавление в индекс
  - Переиндексация

Метрики для индексации

- Weighted UniFrac
  Пользователь должен самостоятельно произвести:
  - OTU-picking
  - Построение филогенетического дерева
- $\sqrt{\text{JSD}}$ (Jensen-Shannon divergence) на распределениях частот k-меров
  - $\frac{1}{2}\sum_i X_i \log(X_i) + Y_i \log(Y_i) - (X_i + Y_i) \log \frac{1}{2}(X_i + Y_i)$

Измерения эффективности индексации и поиска

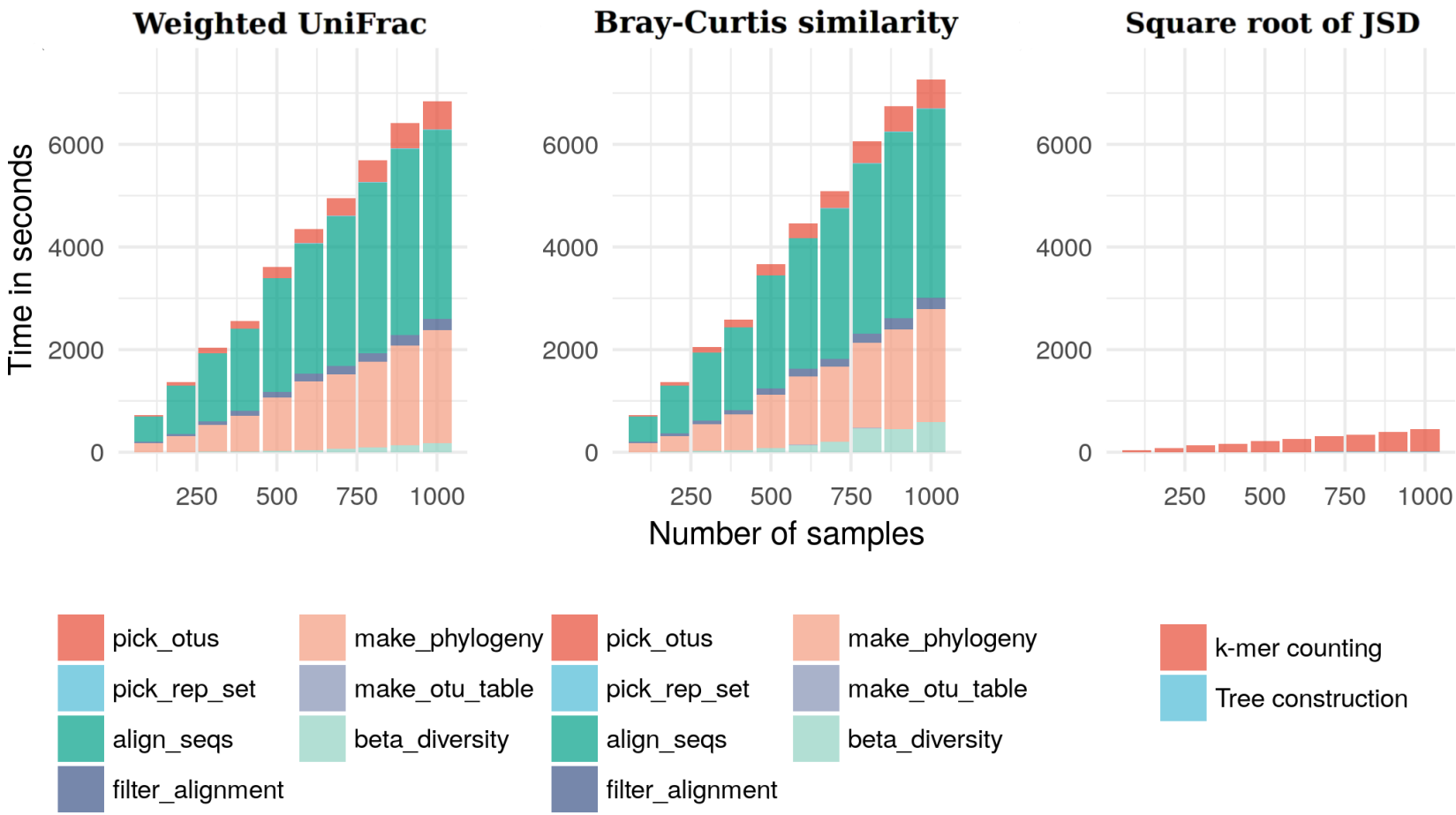
Данные:

- 1978 библиотек микробиоты человека из Sequence Read Archive
- 250+ длина ридов
- 10000+ ридов в каждой библиотеке

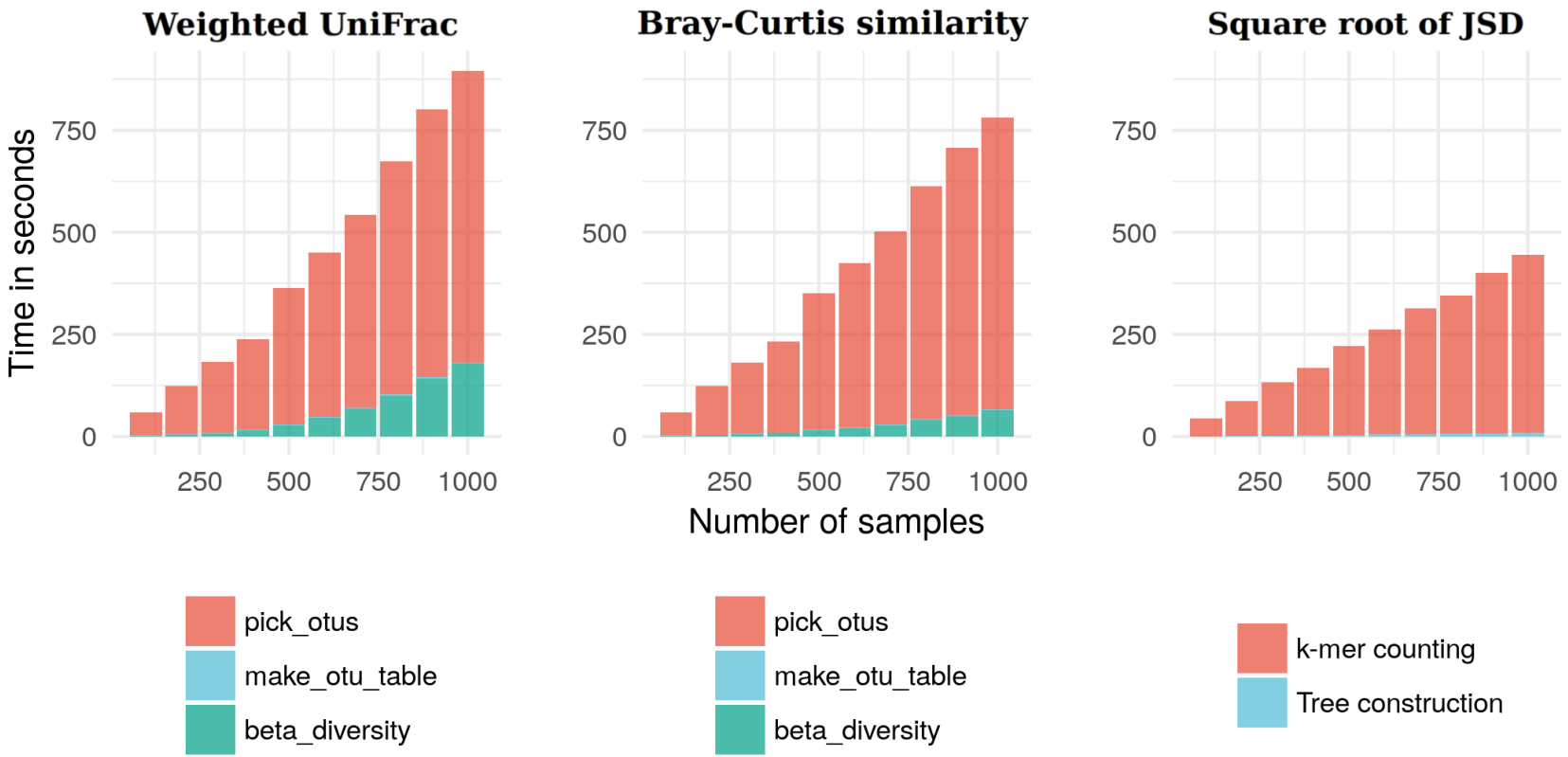
Метрики для сравнения:

- Weighted UniFrac (open-referenced/denovo OTU-picking)
- Bray-Curtis similarity (open-referenced/denovo OTU-picking)

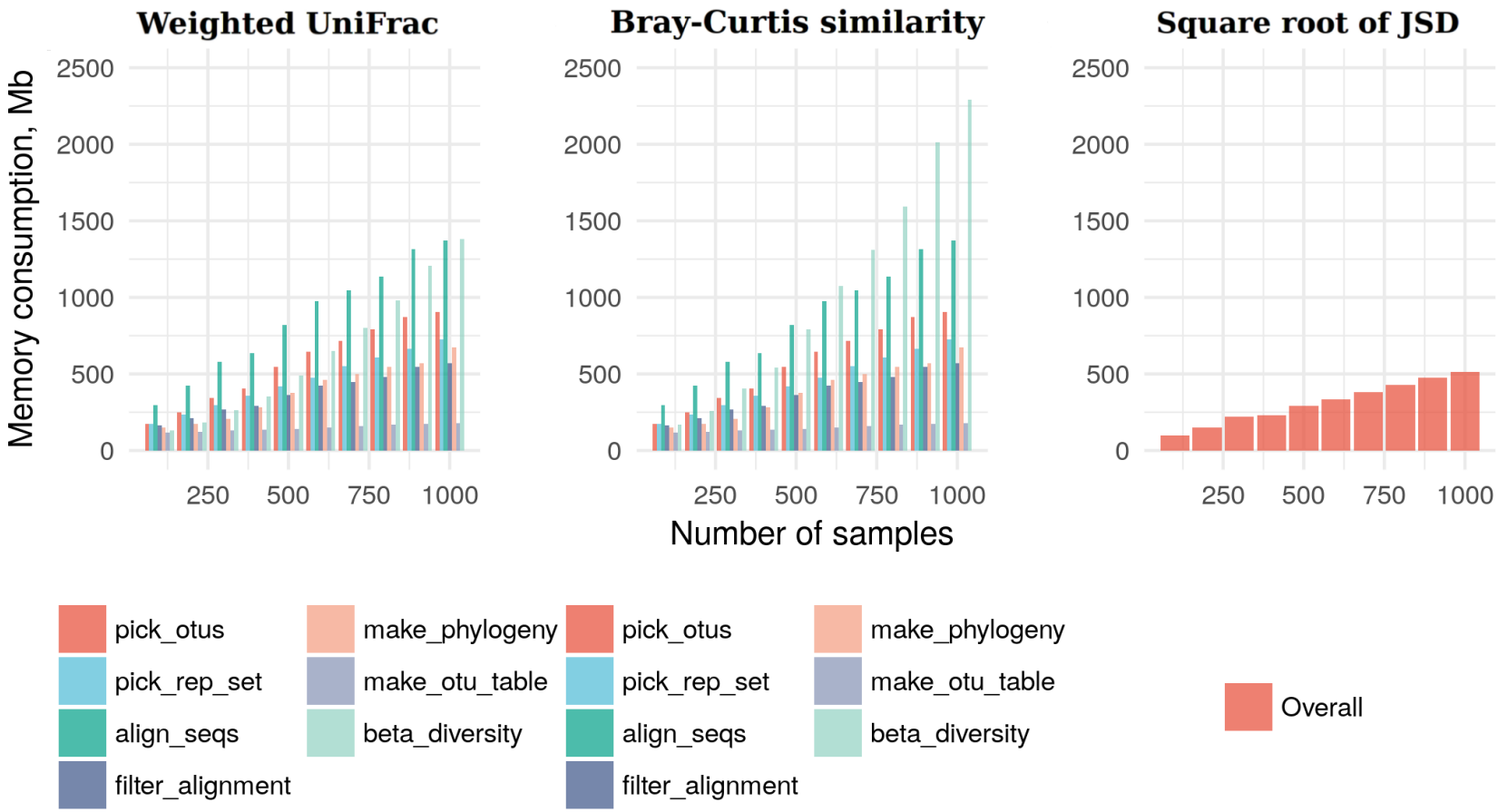
Индексация: время (denovo OTU-picking)



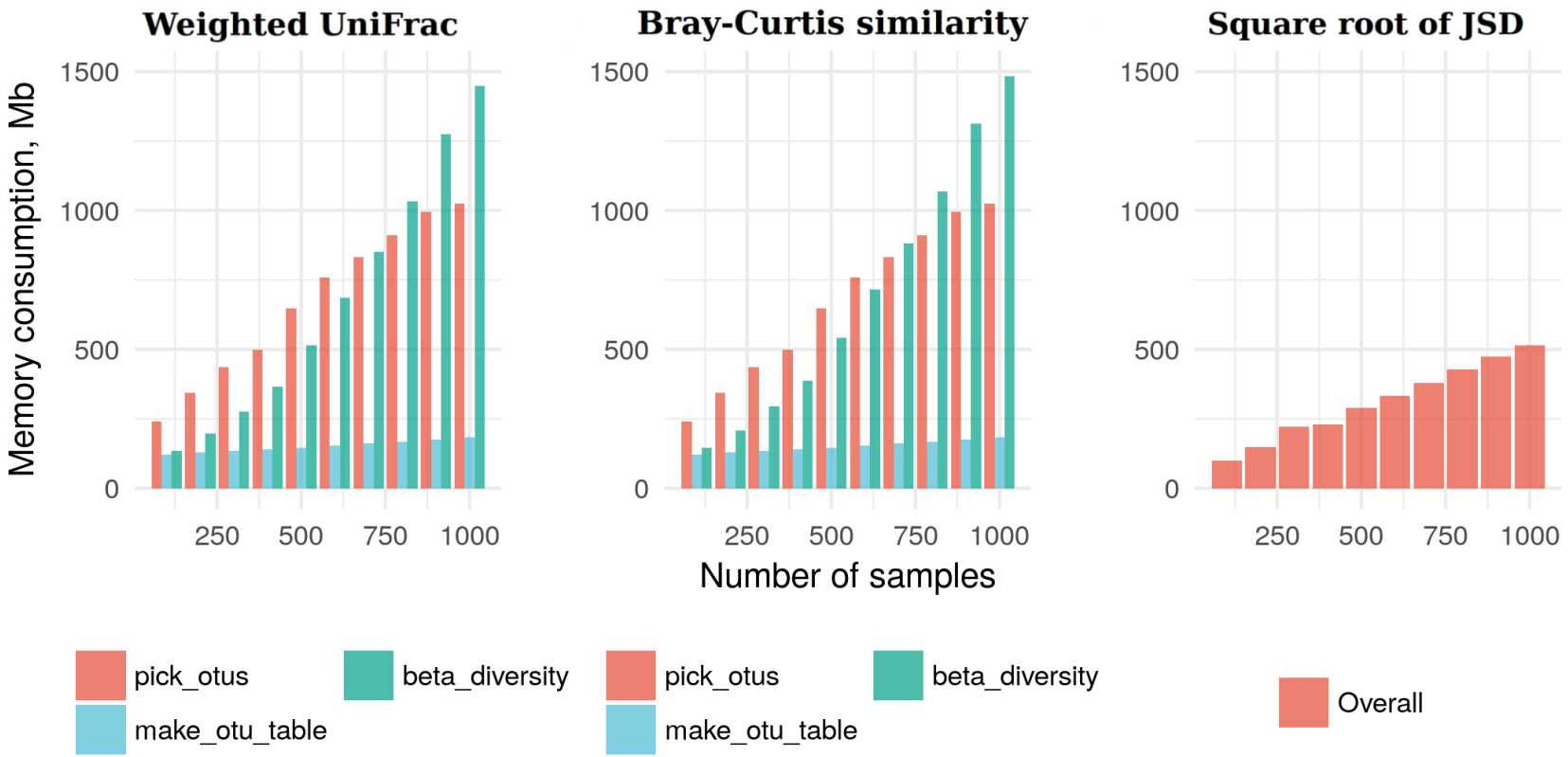
Индексация: время (open-referenced OTU-picking)



Индексация: RAM (denovo OTU-picking)



Индексация: RAM (open-referenced OTU-picking)

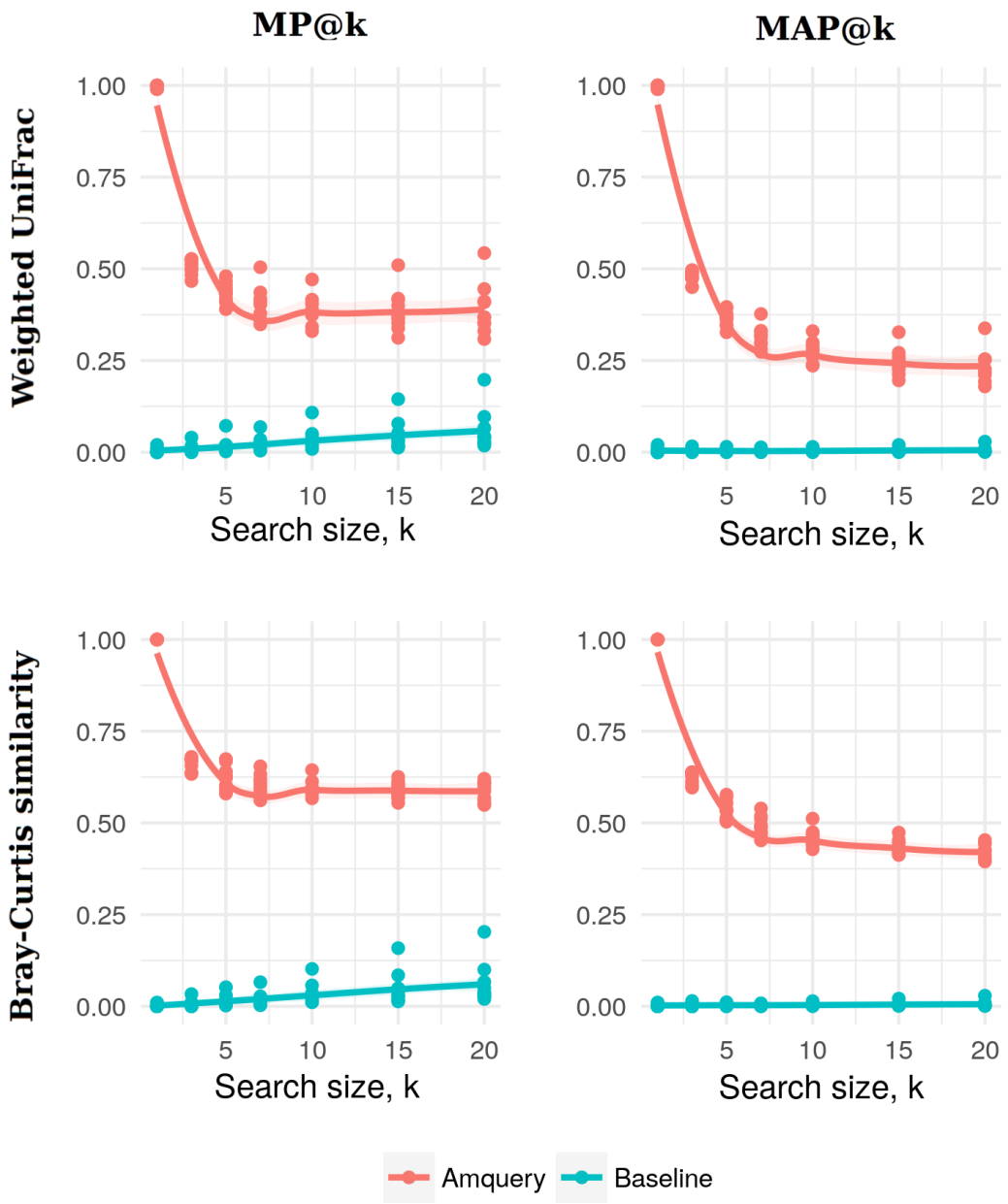


Оценки качества метрики

$I(x_i)$ — индикатор релевантности образца запросу
K — число ближайших образцов

- $MP@K = \frac{1}{N} \sum_{query} \frac{\sum_{i=0}^{K} I(x_i)}{K}$
- $MAP@K = \frac{1}{N} \sum_{query} \frac{\sum_{i=0}^{K} I(x_i) P@i}{K}$

Поиск: P@k, MAP@k, NDCG@k



Результаты

- Разработано масштабируемое решение для метрической индексации и поиска ампликонных библиотек 16S
- Разработана эффективная метрика для сравнения ампликонных библиотек 16S, основанная на подсчете k-меров, позволяющая значительно сократить затраты по времени и памяти при индексации
- По результатам сравнения предложенной метрики с расстояниями, основанными на OTU-picking, точность поиска составляет 0.4-0.7 в пределах небольших запросов