Быстрая индексация метагеномов

Николай Ромащенко

Руководители: Илья Корвиго, Евгений Андронов, ГНУ ВНИИСХМ РАСХН

Институт биоинформатики

Санкт-Петербург, 2016

Цель и задачи проекта

Цель: Исследовать возможность построения быстрого индекса метагеномов

Задачи:

1. Подобрать метрику для аппроксимации расстояния Unifrac
2. Построить модели предсказания ближайших соседей по Unifrac
3. Построить поисковый индекс для предложенной модели

Мотивация: На данный момент нет сервиса для глобального поиска ближайших соседей по Unifrac

Unifrac distance

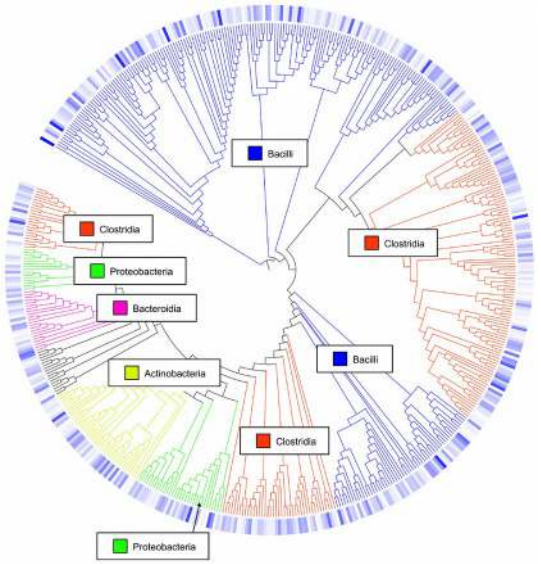
I. Попарное выравнивание

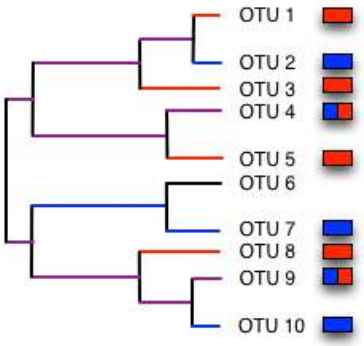
II. OTU-picking

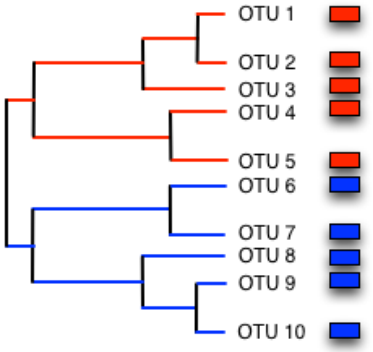
III. Множественное выравнивание

IV. Построение филогенетического дерева

V. Построение матрицы попарных расстояний







https://peerj.com/articles/237/

Unifrac distance

I. Попарное выравнивание $O(n m_{max} S_{max})^p, \quad p \in [1,2]$

II. OTU-picking $O(n m_{max} S_{max})$

III. Множественное выравнивание $O(n m_{max} S_{max})^p$

n – число образцов
m_{max} – максимальное покрытие
S_{max} – максимальная длина рида

IV. Построение филогенетического дерева $O(L N^{1.5} \log N)$

$L = O(S_{max})$ – ширина выравнивания
$N = O(n m_{max})$ – число уникальных ридов

V. Построение матрицы попарных расстояний $O(n^2 S_{max} h(T))$

$h(T) = O(n m_{max})$ – высота дерева

Предлагаемый подход: индексация

I. Построение индекса k-меров для каждого метагеномного образца

$O(n m_{max} S_{max})$

II. Построение вспомогательных попарных расстояний на индексах

- Jensen-Shannon divergence
- Bray-Curtis dissimilarity
- Generalized Jaccard Index

$O(n^2 m_{max})$ *или*
$O(n \log(n)\ m_{max})$

Предлагаемый подход: индексация

III. Построение координатной системы в пространстве индексов образцов:

- **генетический алгоритм** с минимизацией матрицы корреляций расстояний между точками координатной системы
- **жадный алгоритм** с той же целевой функцией

$O(p^2)$

– (нет)

$O(p^3)$

$p \sim 10^2 - 10^3$ – размер координатной системы

Предлагаемый подход: поиск

I. Поиск k ближайших соседей в координатной системе:

vantage-point tree:

- по попарным расстояниям на индексах k-меров
- по расстояниям до координатной системы

II. Обучение бинарного классификатора

$O(n^2 \log(n) S_{max} m_{max})$,
$O(S \log(n) m)$

$O(np \log(n) S_{max} m_{max})$,
$O(S \log(n) m)$

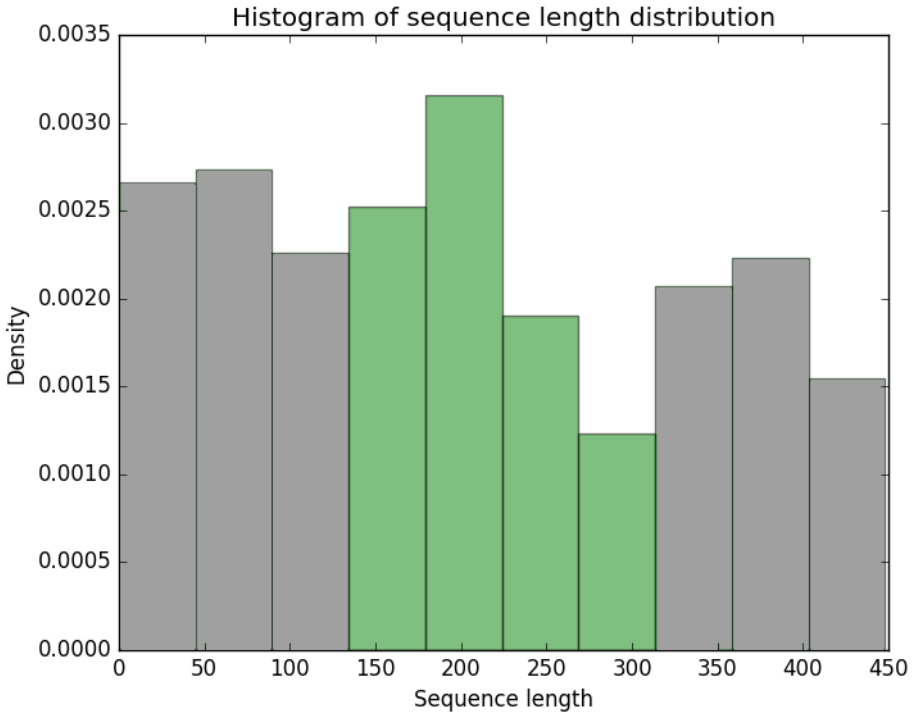
n – число образцов в базе
S – максимальная длина рида в образце
m – покрытие образца
p – размер координатной системы

Неоднородность данных

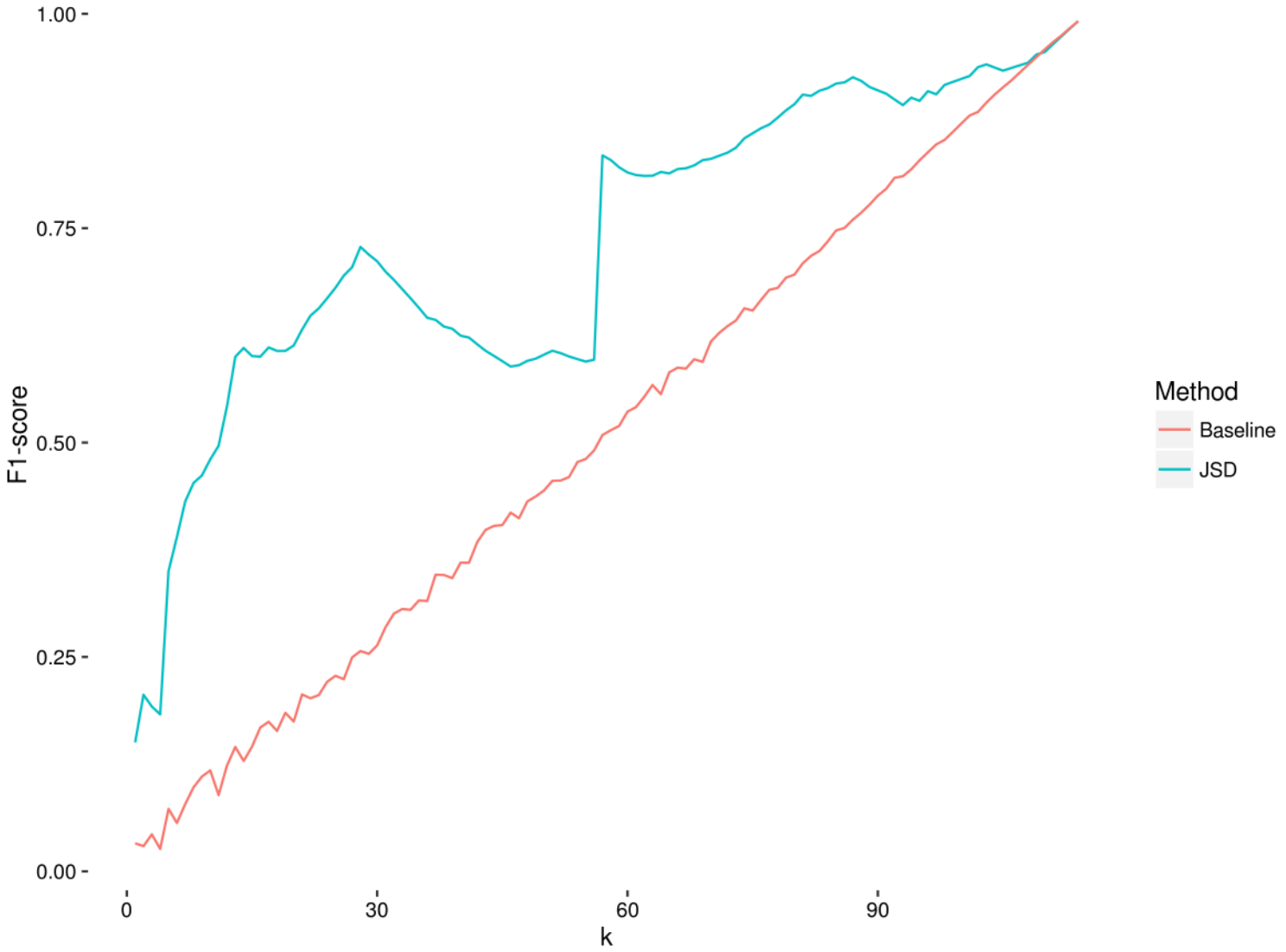
Проблемы:

I. Разные технологии секвенирования
II. Разные праймеры
III. Битые риды

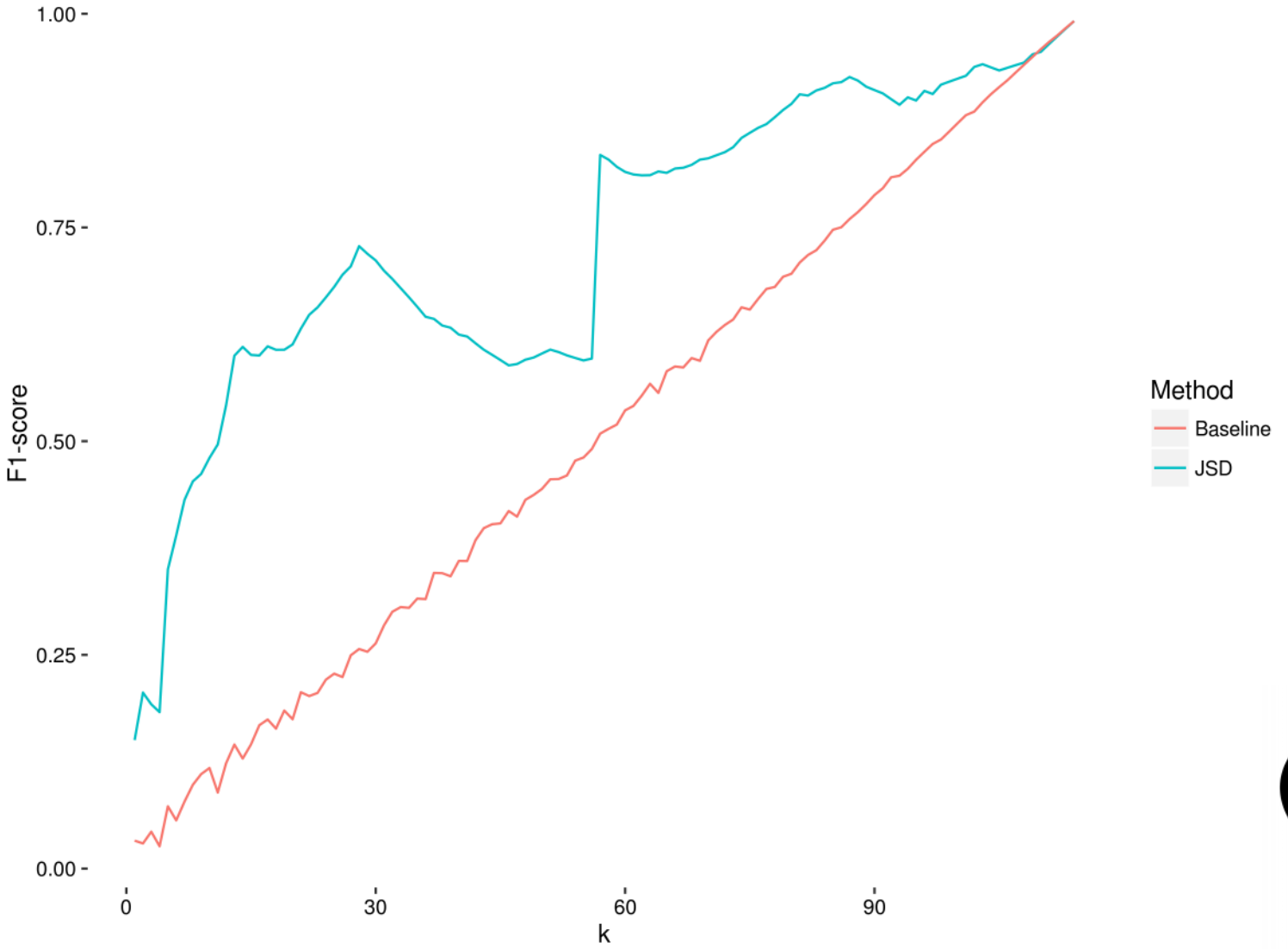
Решение: фильтрация ридов



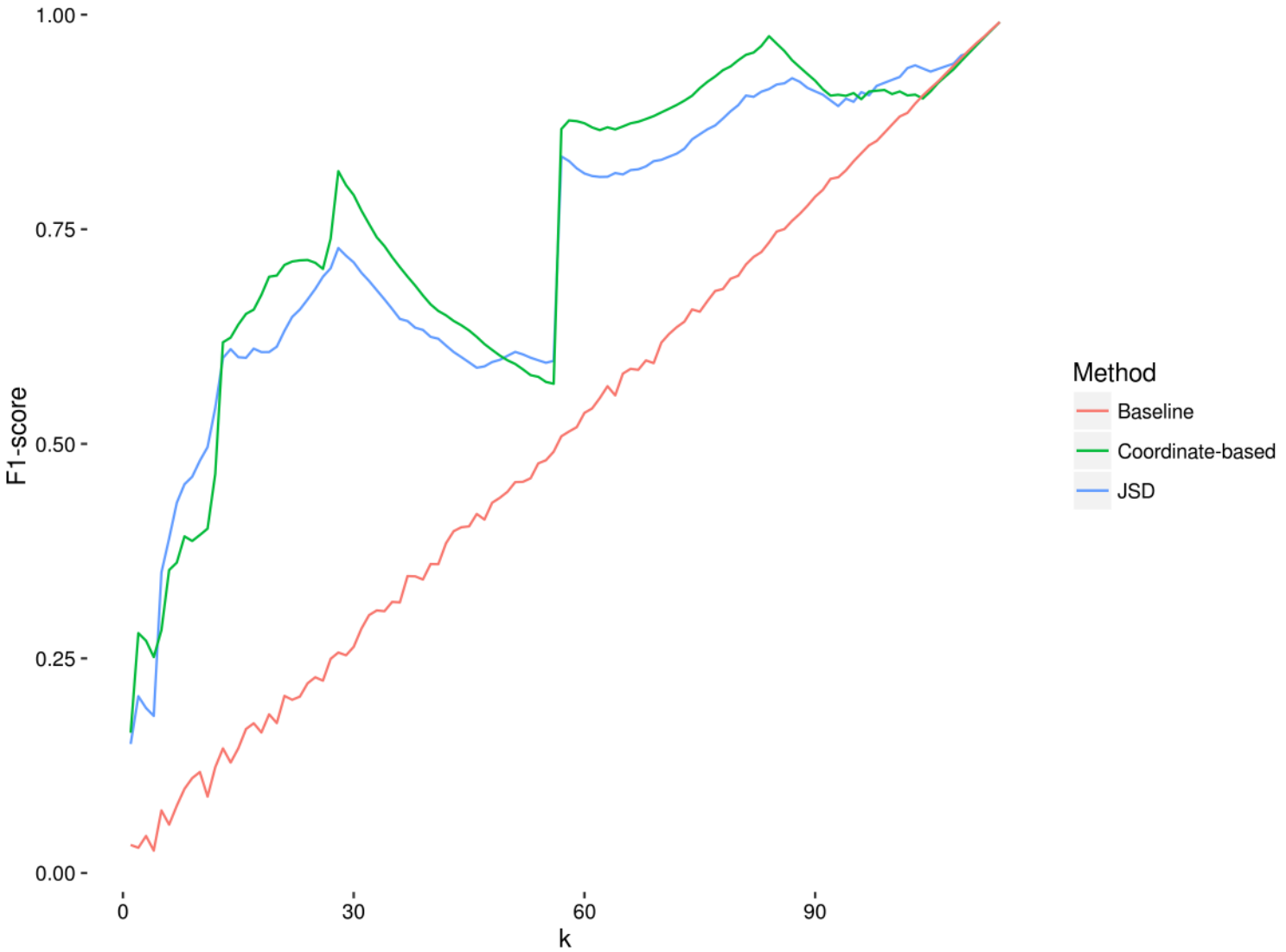
VP-tree performance



VP-tree performance



VP-tree performance

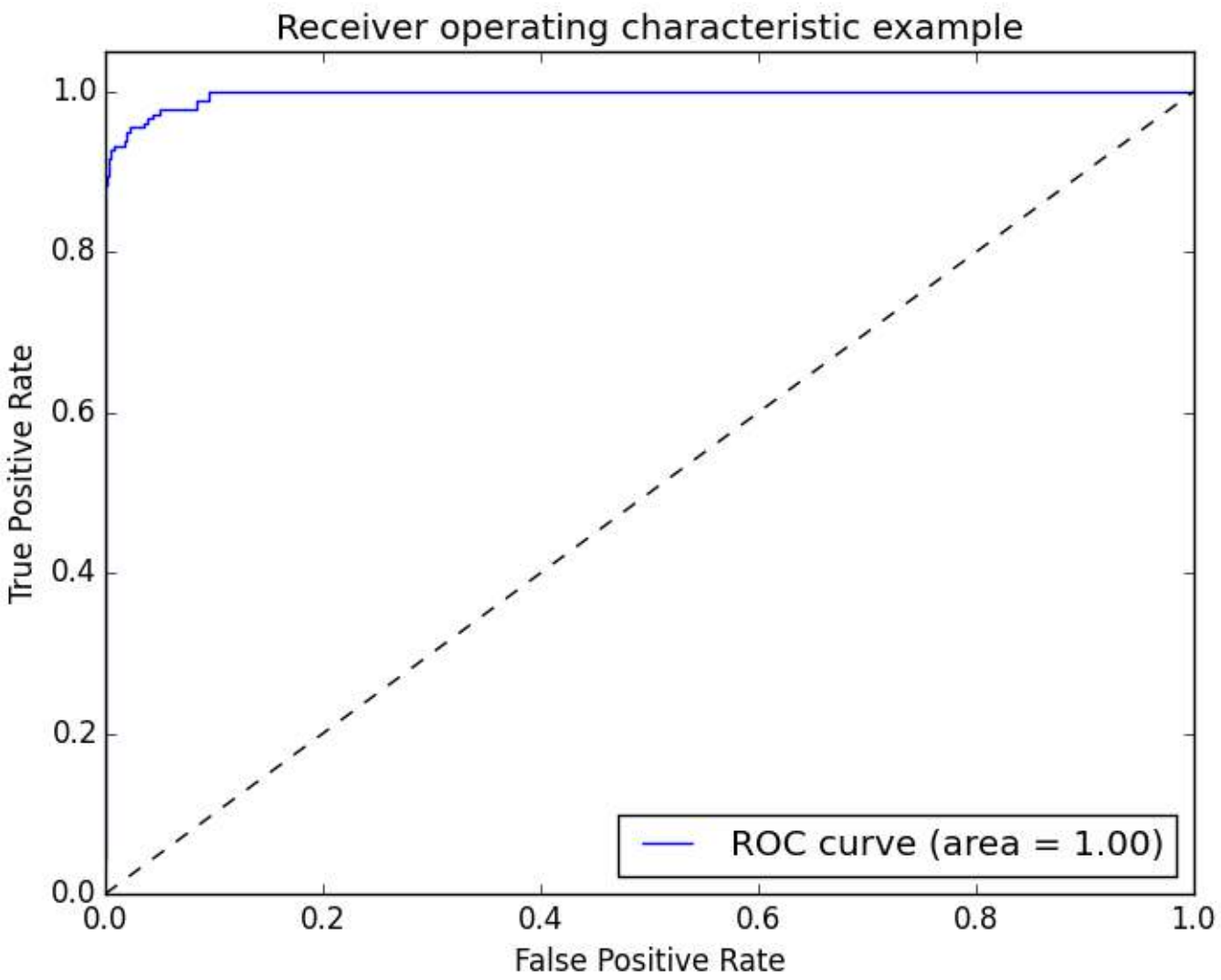


Бинарная классификация

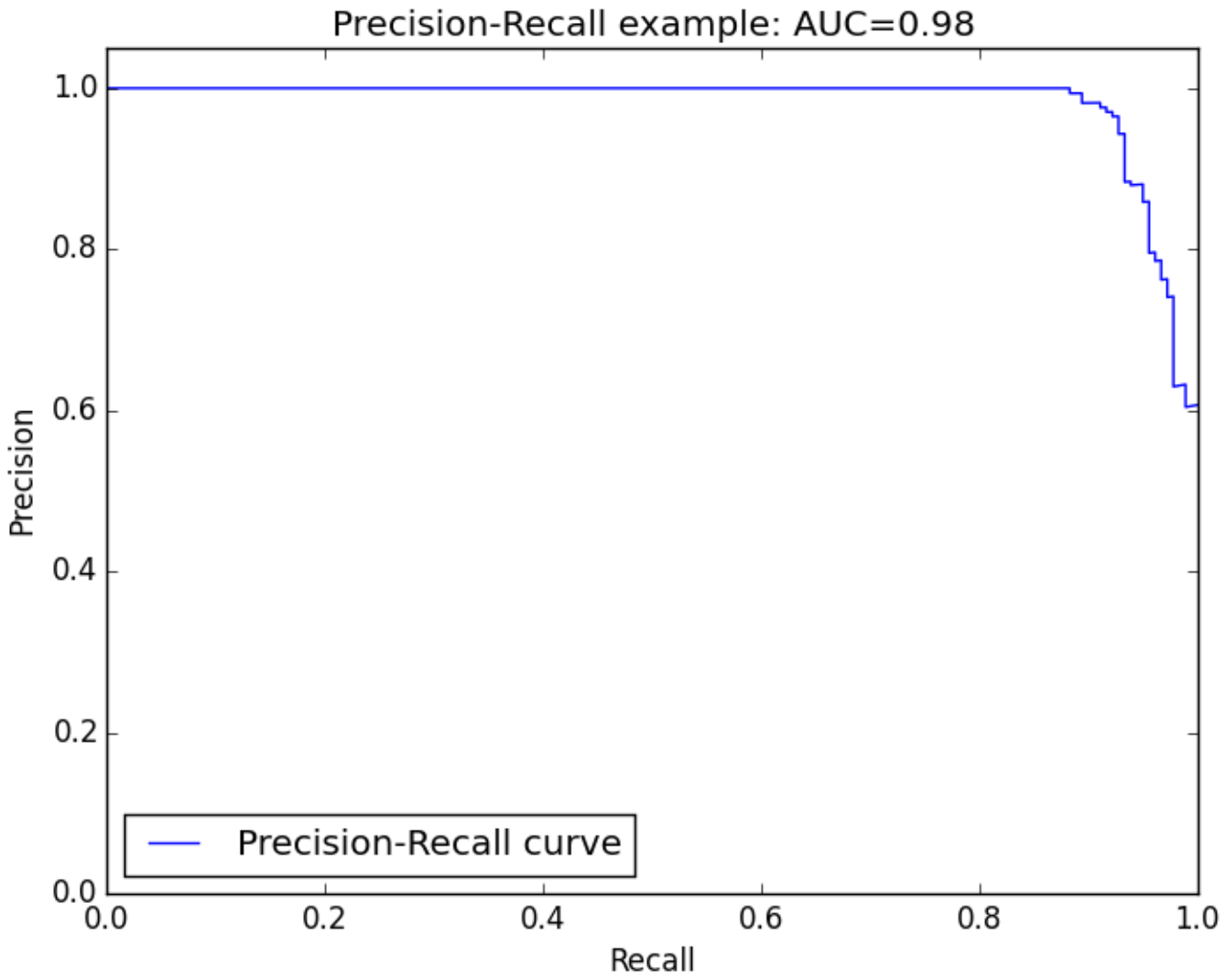
II. Нейросеть с двумя ReLU скрытыми слоями

- Тренировка в mini-batch режиме с дропаутом и нестеровым моментом в течение 10000 эпох
- Сигмоидный выходной слой
- Подбор параметров обучения – генетический алгоритм в кросс-валидации

Nnet performance



Nnet performance



Дальнейшие планы

I. **Исследовать возможности аппроксимации поиска в других условиях**
   - Другие классификаторы
   - Не k-mer-based подходы
   - Крупные датасеты
   - smth else

II. **Публикация в PyPI**

III. **???**

IV. **PROFIT!!!**