Предсказание патогеннности бактерий по данным NGS-секвенирования

Страх и ненависть в биоинформатических базах данных

www.nature.com/scientificreports

OPEN

PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data

Received: 08 July 2016

Carlus Deneke, Robert Rentzsch & Bernhard Y. Renard

Задача бинарной классификации

Version 0.1

- На вход подаем данные NGS-секвенирования бактерии – риды
- На выходе – оценка патогенности бактерии

Задача бинарной классификации

Version 0.2

- На вход подаем данные NGS-секвенирования бактерии – риды
- Обучаем алгоритм классификации отдельных ридов – вероятность того, что рид от патогенной бактерии
- Агрегация результатов классификации отдельных ридов, для классификации бактерии целиком

Данные для обучения

Данные для обучения

NCBI Pathogens

Данные для обучения

NCBI Pathogens

Люди изучают/секвенируют
в основном патогены

Feature selection / feature extraction

- Для классификации нужен набор признаков
- Встречаемость k-mer и другие похожие признаки
- Что-нибудь посложнее
- Признаки можно извлечь из данных самим