Анализ групп метагеномных образцов с использованием графовых признаков для решения задач сравнительной метагеномики

Попов Владимир Витальевич, группа 21.М03-мм

Санкт-Петербургский государственный университет
Математическое моделирование, программирование и искусственный интеллект

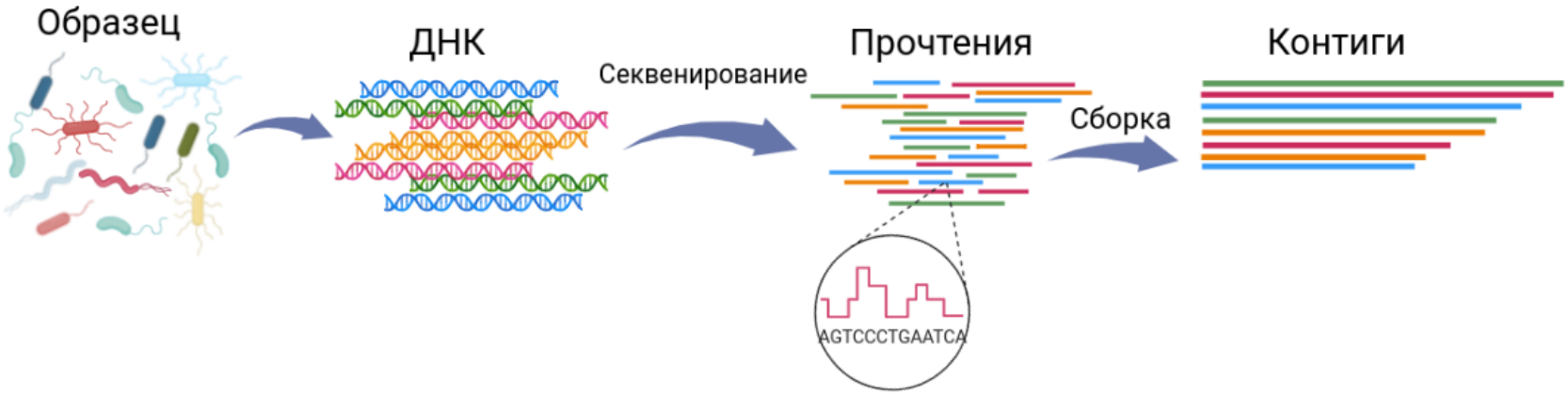
Научный руководитель: к.ф.-м.н., доцент Шпилев П.В.

Консультант: к.ф.-м.н., доцент Коробейников А.И.

Санкт-Петербург
2023 г.

Метагеномика: основные понятия

- Геном – хранилище наследственной информации, строка над алфавитом $\{A, C, G, T\}$.
- Прочтение – предполагаемая подстрока генома полученная в результате секвенирования.
- Контиг – непрерывный участок ДНК собранный из нескольких прочтений.
- Метагеном – набор геномов организмов, населяющих некоторую среду.



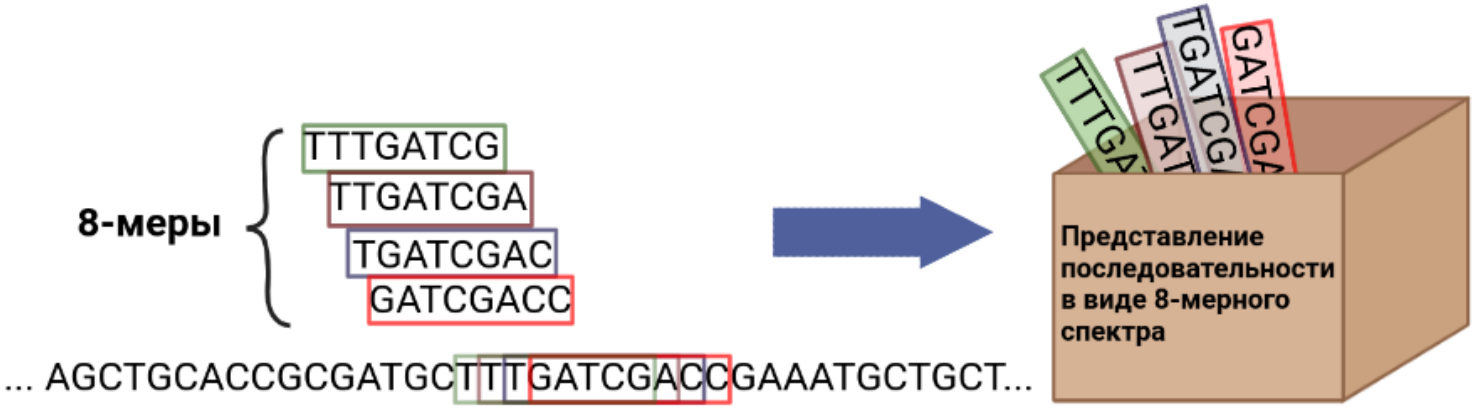
Сравнительный анализ метагеномов

Мотивация: Клинические исследования, анализ изменений в составе микроорганизмов в почве и водоемах.

Методы сравнения групп метагеномных образцов – референсные/безреферересные.

k-мерный поход:

- k-мер – подстрока прочтения длины k.
- Подсчёт k-меров – получение словаря $\{k$-мер : его встречаемость в образце$\}$.



Постановка задачи

Цель:
Улучшение качества методов сравнительного анализа групп метагеномных образцов.

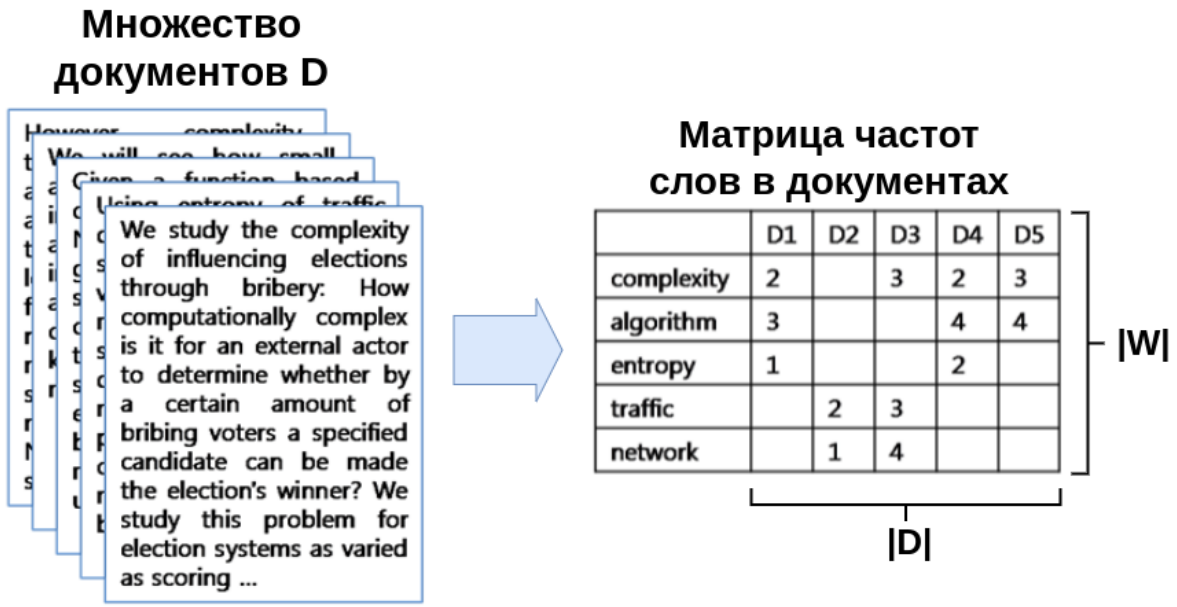
Метод:
Построение признаков на основе подмножества k-меров, описывающих скрытую структуру наборов образцов.

Задачи:

- Разработка и реализация эффективного алгоритма для извлечения наиболее подходящего подмножества k-меров.
- Проведение вычислительных экспериментов по классификации метагеномных образцов методами машинного обучения.

Тематическое моделирование

Дано:



Предположения:

- Тема — некоторое распределение на множестве слов W.
- Каждый документ описывается некоторым распределением на множестве тем.
- Порядок слов в документе не важен.
- Вероятность появления слова в документе зависит только от темы (не зависит от документа).

Термины:

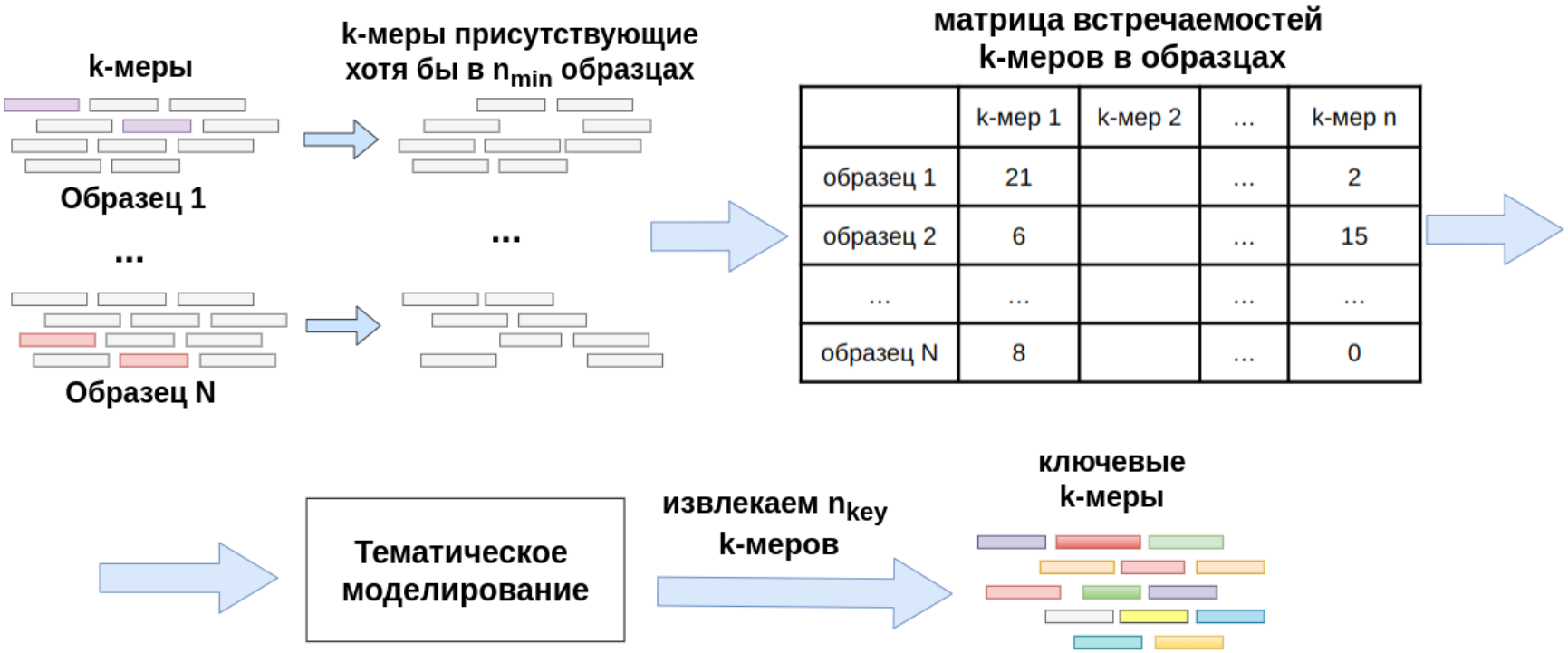
- Документы – метагеномные образцы.
- Слова – k-меры.
- Темы – скрытая структура групп образцов.

Предположения:

- Порядок k-меров в образцах не важен.
- Вероятность появления k-мера в образце d по теме t зависит только от темы.
- Каждый образец имеет несколько смешанных в некоторой пропорции тематик.

Таким образом тематическое моделирование может быть адаптировано для извлечения ключевых k-меров.

Алгоритм извлечения ключевых k-меров



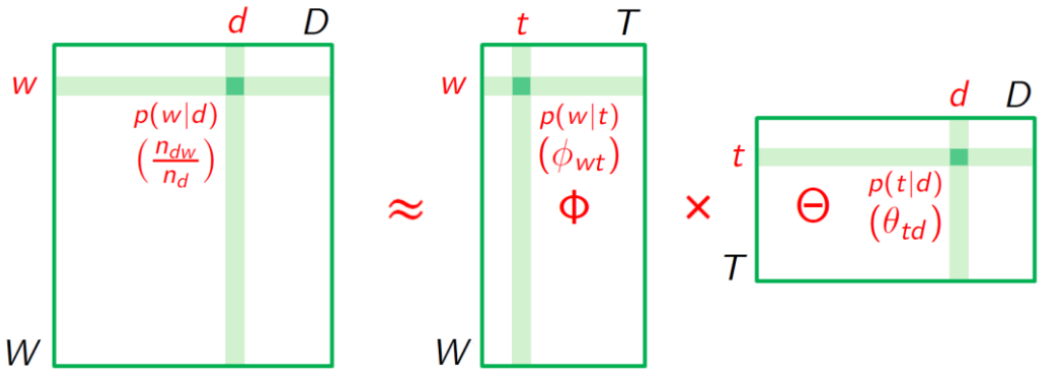
	k-мер 1	k-мер 2	...	k-мер n
образец 1	21		...	2
образец 2	6		...	15
...	...		...	...
образец N	8		...	0

Представленный алгоритм был реализован с помощью языков программирования Java и Python.

Извлечение ключевых слов

Обозначим T – набор тем, $\varphi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$.

Стохастическое матричное разложение: $P \approx \Phi\Theta$



В работе[1] для извлечения ключевых слов предлагается использовать величину «значимости» слов:

$$S(w) = p(w) \sum_{t \in T} p(w|t) \ln \frac{p(w|t)}{p(t)},$$

где $p(t)$ — частное распределение темы, $p(w)$ – частное распределение слова.

[1] Chuang J., Manning C. D., Heer J. Termite : Proceedings of the International Working Conference on Advanced Visual Interfaces. — ACM, 2012.

Программа MetaFast

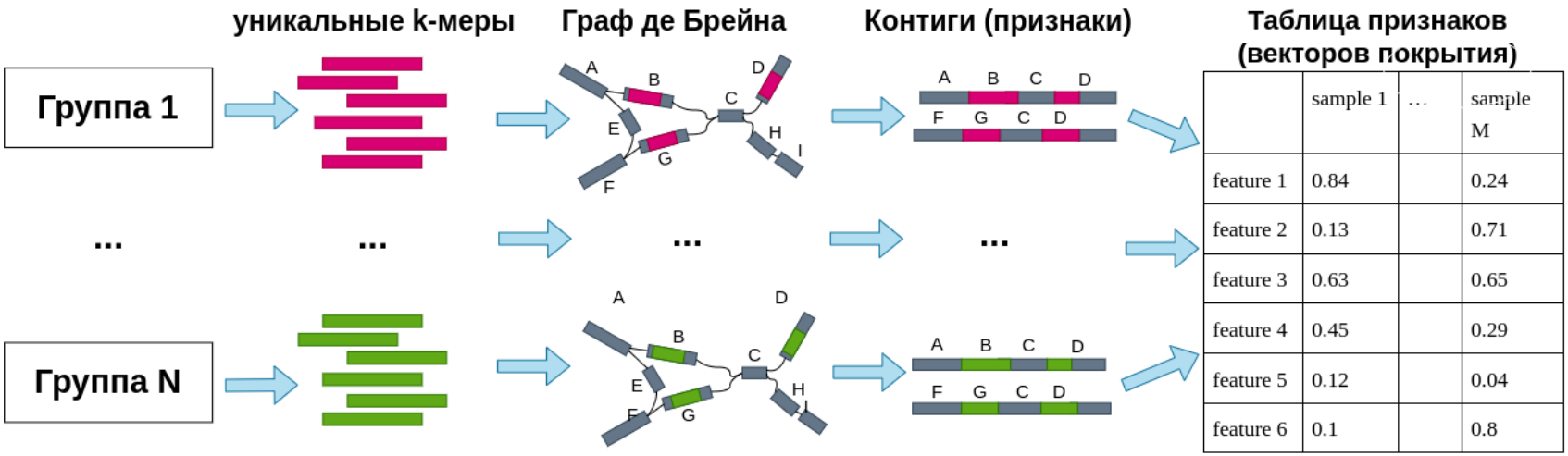
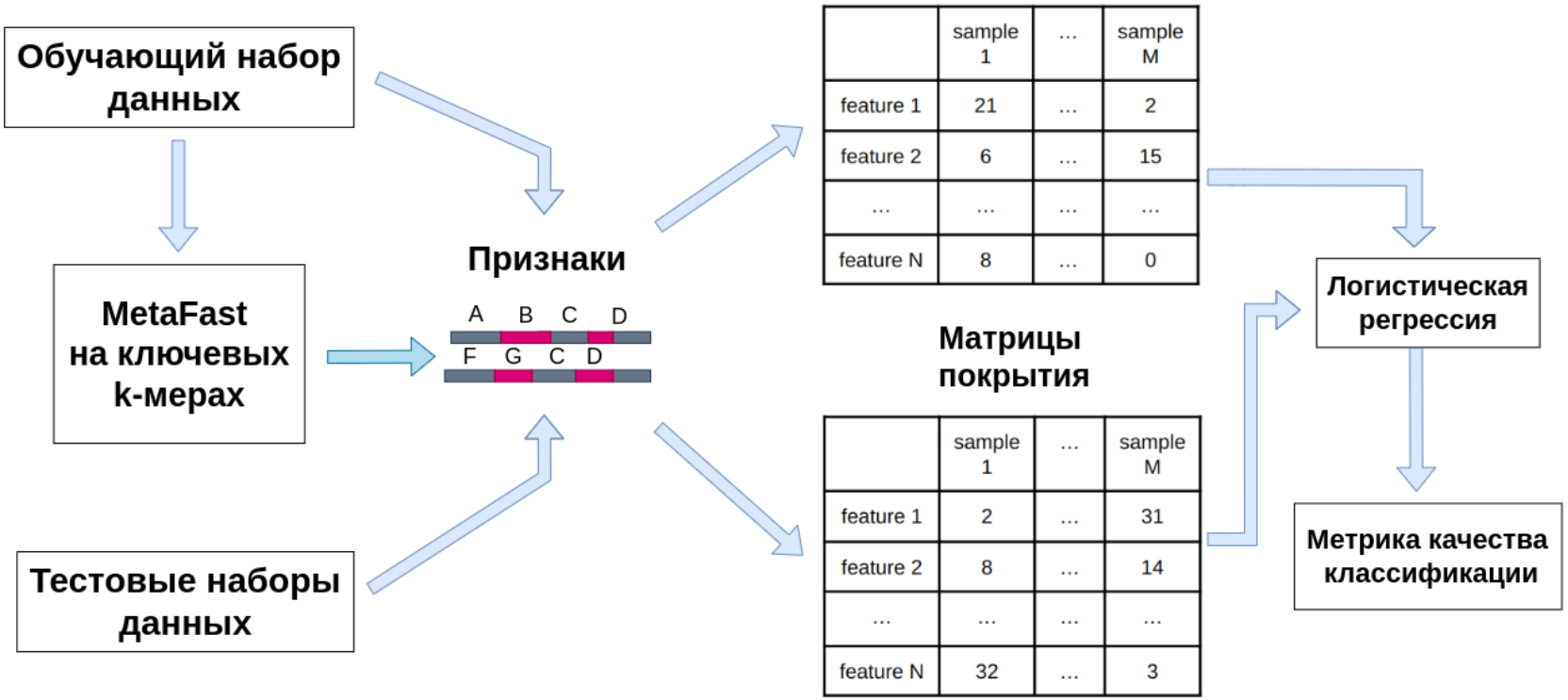


Рис.: Схема работы программы **MetaFast**[2]

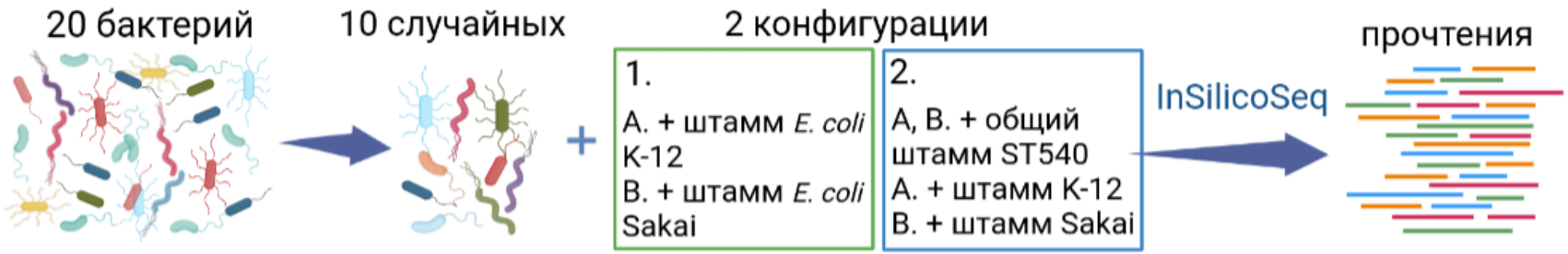
[2]Ulyantsev V.I., Kazakov S.V., Dubinkina V.B., Tyakht A.V., Alexeev D.G. (2016). MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data. Bioinformatics, 32(18), 2760-2767.

План вычислительных экспериментов



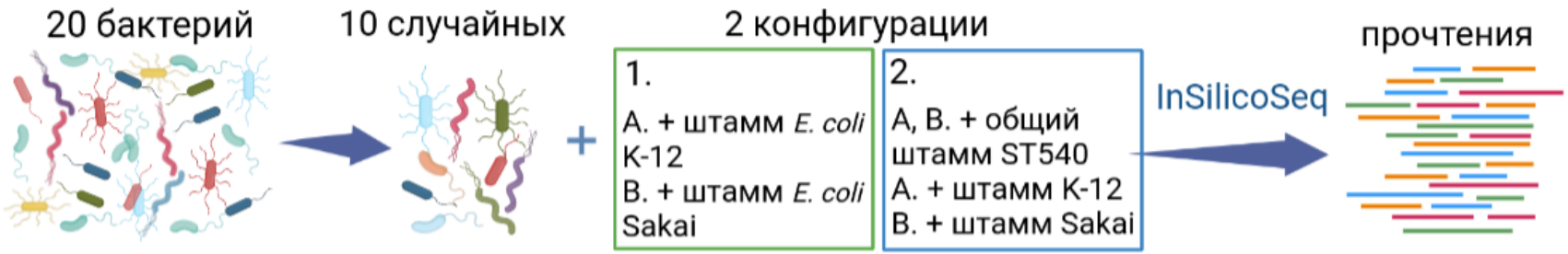
Наборы моделированных метагеномных данных

План моделирования одного образца:

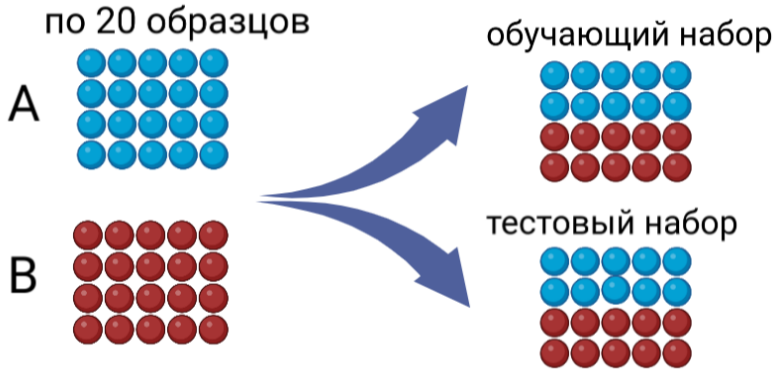


Наборы моделированных метагеномных данных

План моделирования одного образца:

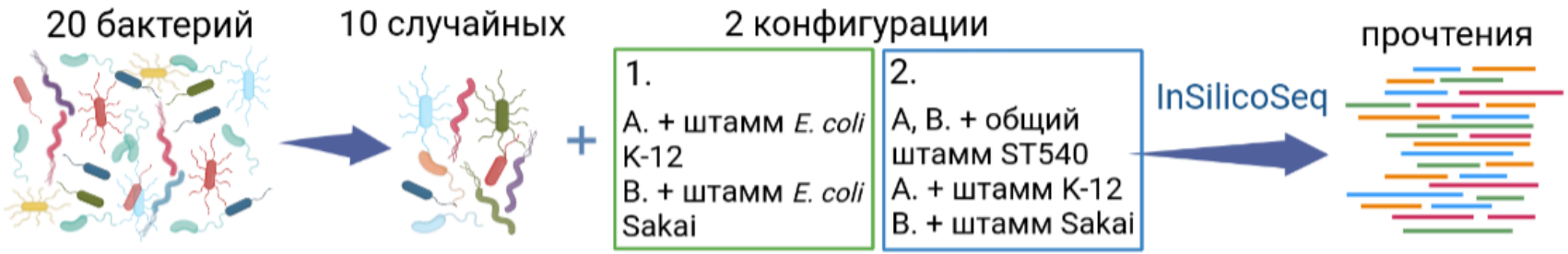


Для каждой конфигурации:

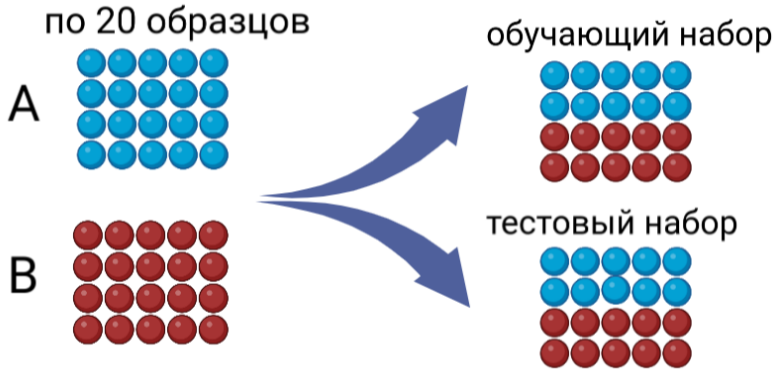


Наборы моделированных метагеномных данных

План моделирования одного образца:



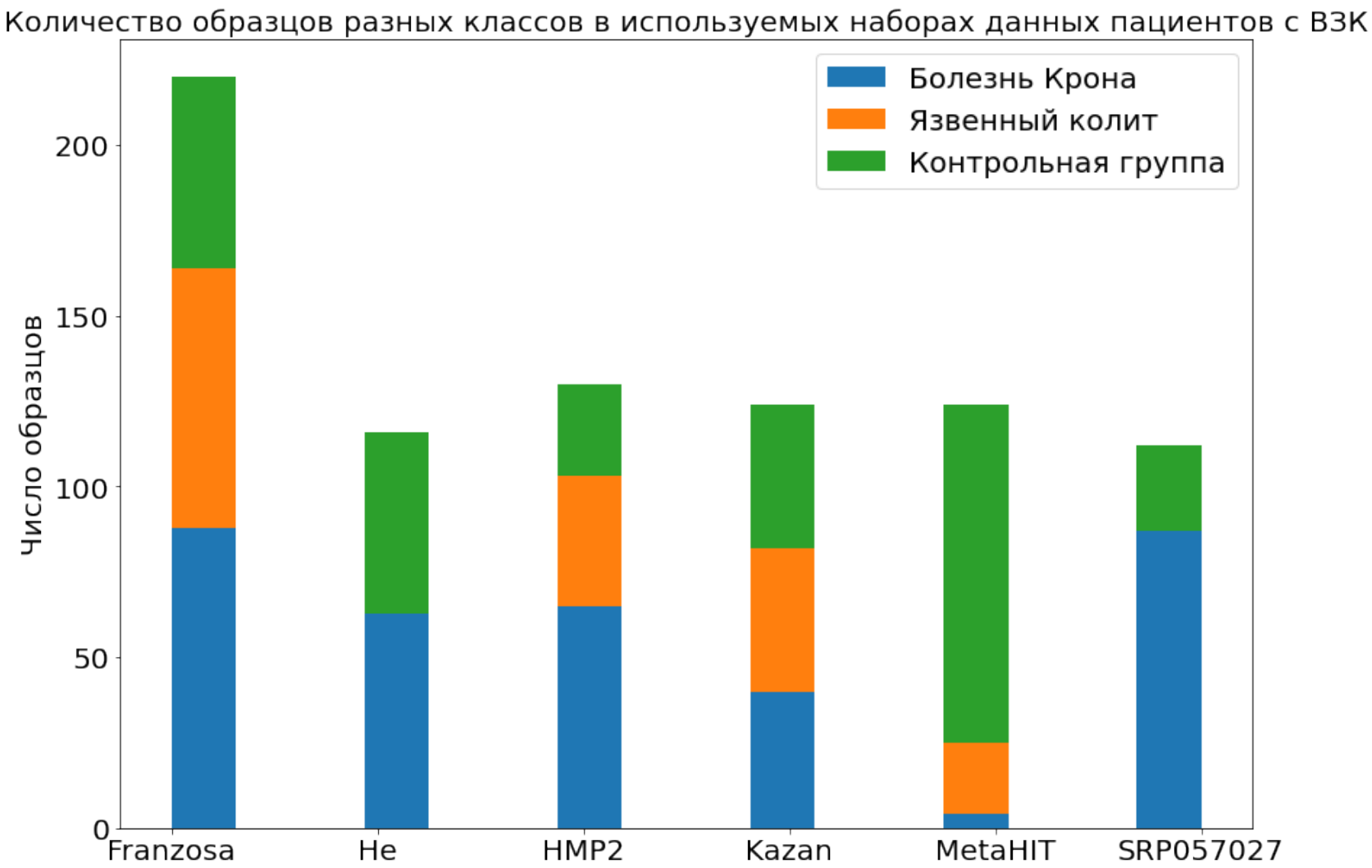
Для каждой конфигурации:



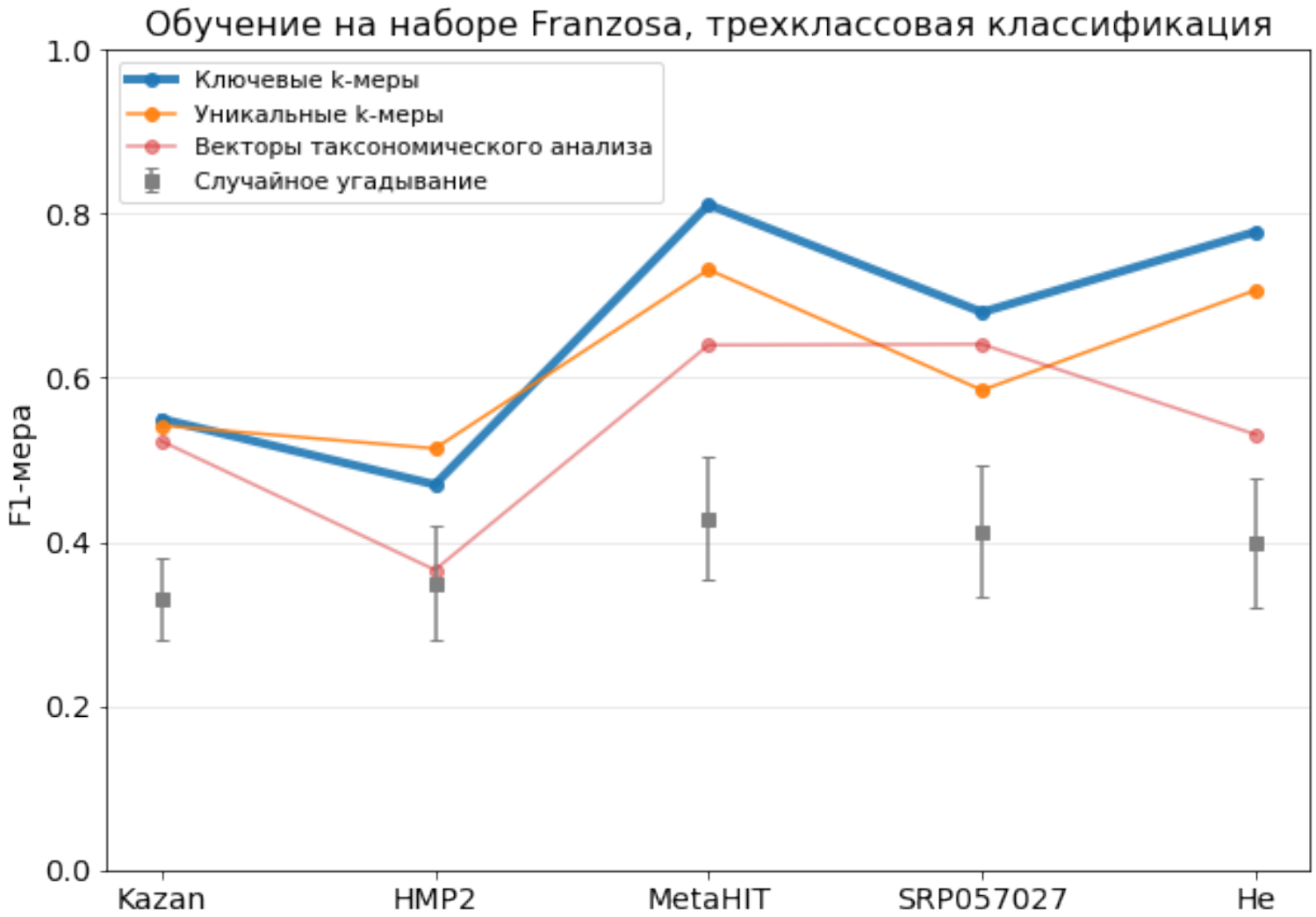
Результаты классификации:

Accuracy на тестовом наборе	1	2
Ключевые *k*-меры	0.95	1.00
Уникальные *k*-меры	0.95	1.00
Таксономическая аннотация	0.7	0.55

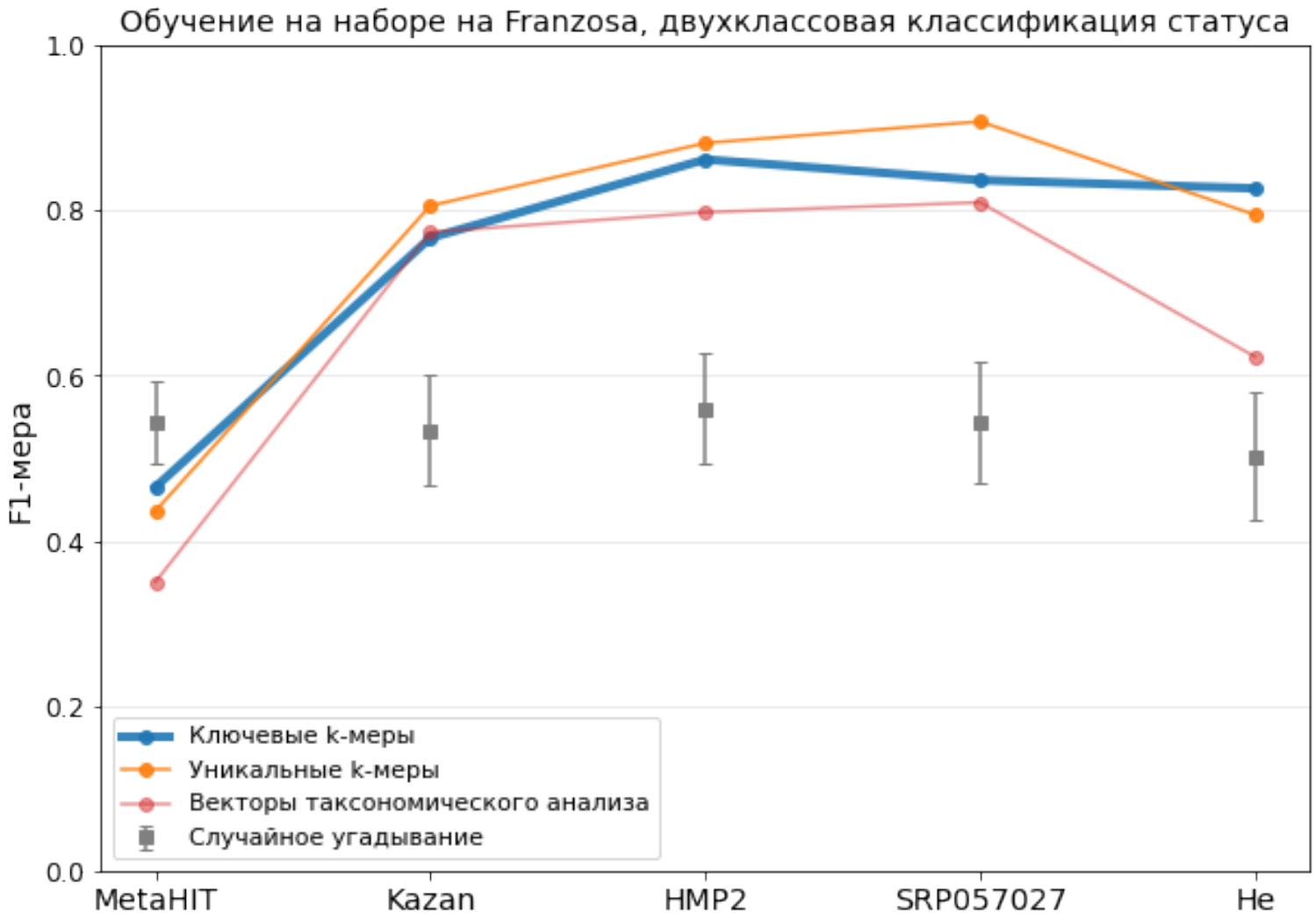
Наборы реальных метагеномных данных



Вычислительные эксперименты: трехклассовая классификация



Вычислительные эксперименты: бинарная классификация статуса пациента



Заключение

Результаты:

- Разработан и реализован метод извлечения признаков из групп метагеномных образцов на основе ключевых k-меров.
- Метод внедрен в набор модулей программы MetaFast, работа алгоритма провалидирована на смоделированных метагеномных данных.
- Проведены вычислительные эксперименты по классификации метагеномных образцов пациентов, результаты показали преимущество разработанного метода в некоторых задачах.