

# Адаптация алгоритма множественного выравнивания последовательностей библиотеки SeqAn для работы с глубокими выравниваниями

---

**Олег Яснев**

Руководитель: Prof. Dr. **Knut Reinert**

Санкт-Петербургский Академический Университет

2015

# Множественное выравнивание последовательностей

Приложения:

- Предсказание структуры и функций белка
- Построение филогении и анализ эволюции
- Получение консенсуса ридов

**Глубокое выравнивание** – выравнивание *большого* числа ( $>1000$ ) сравнительно *коротких* ( $<500$  букв) последовательностей

# Существующие решения

Стратегия рафинирования (refinement):

- PASTA (2014)
- MAFFT (2002—2013)
- Clustal Omega (2011)

Стратегия консистентности (consistency):

- T-Coffee (2000—2014)
- SeqAn::T-Coffee (2009)
- MSAProbs (2010)

# Цель и задачи работы

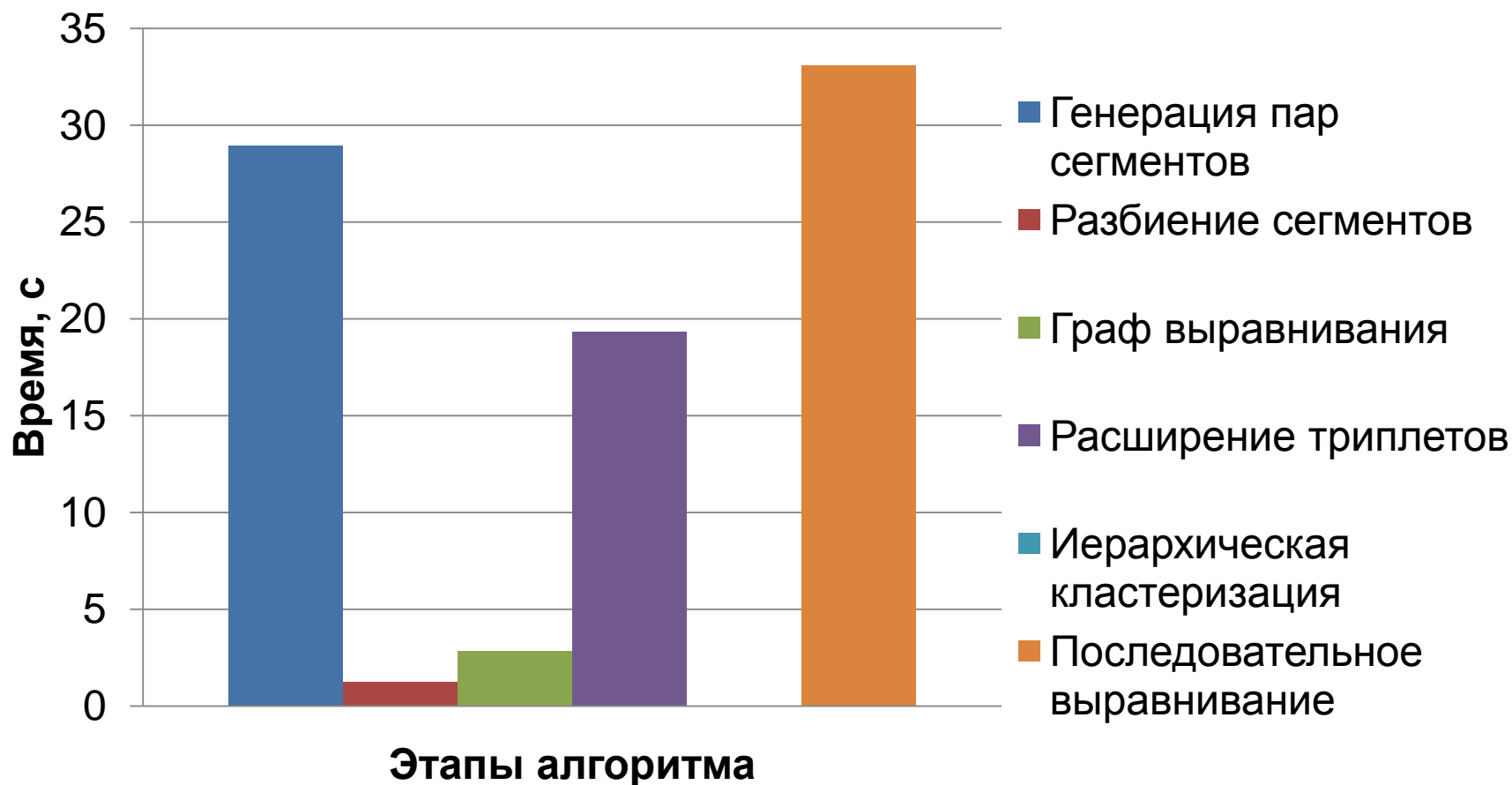
**Цель:** Уменьшить время исполнения SeqAn::T-Coffee при работе с глубокими выравниваниями, сохранив высокое качество

## Задачи:

1. Проанализировать алгоритм SeqAn::T-Coffee и выявить в нем узкие места
2. Предложить и проанализировать стратегии по устранению узких мест
3. Реализовать лучшую стратегию и сравнить новую версию с другими решениями

# Анализ алгоритма и выявление узких мест

## 200 последовательностей



# Генерация пар сегментов

- Исходно: квадратичные алгоритмы глобального и локального выравнивания
- Испытанные стратегии:
  1. Только глобальные выравнивания
  2. Только локальные выравнивания
  3. Окаймленное выравнивание
  4. Совпадающие k-меры
  5. Совпадающие k-меры в сокращенном алфавите

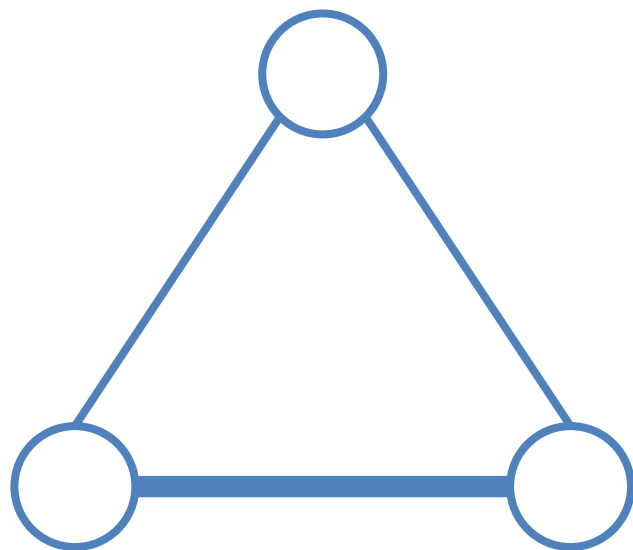
Очень плохо

Плохо

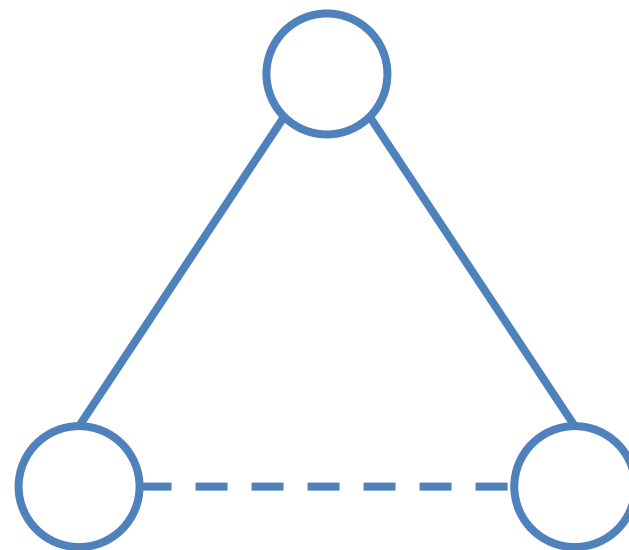
Удовлетворительно

Хорошо

# Расширение триплетов



Увеличиваем вес третьего ребра



Создаем третье ребро

# Расширение триплетов и последовательное выравнивание

Ключевая идея:

1. Уменьшить число обрабатываемых триплетов
2. Уменьшить число создаваемых ребер

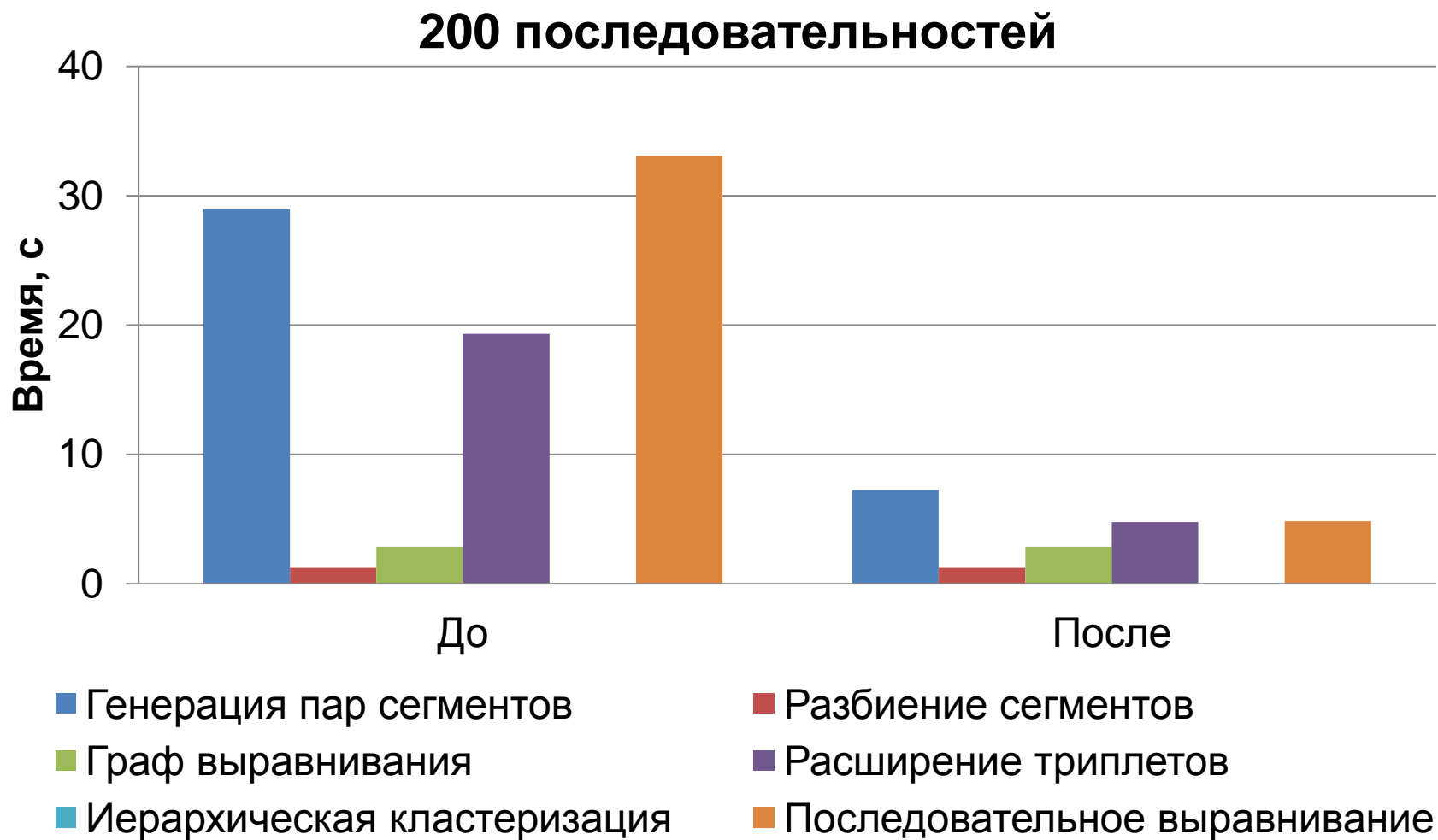
Испытанные стратегии:

1. Ребро максимального веса, вместо минимального
2. Обработать только самые тяжелые ребра
3. Удалить часть ребер
4. Обработать триплеты внутри небольших кластеров
5. Не создавать новые ребра
6. **Комбинация хороших стратегий**

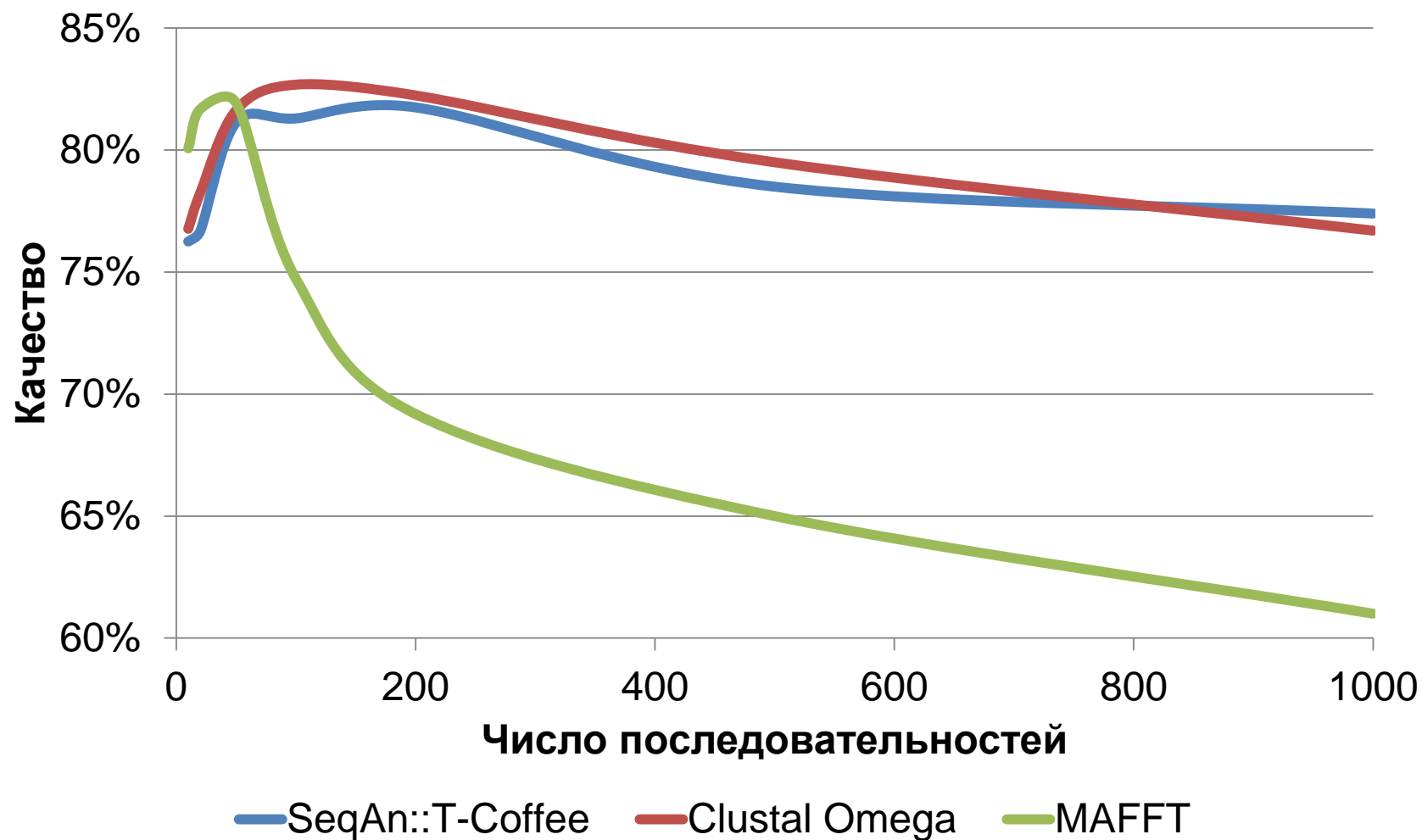
Очень плохо    Плохо    Удовлетворительно    Хорошо  
Очень хорошо



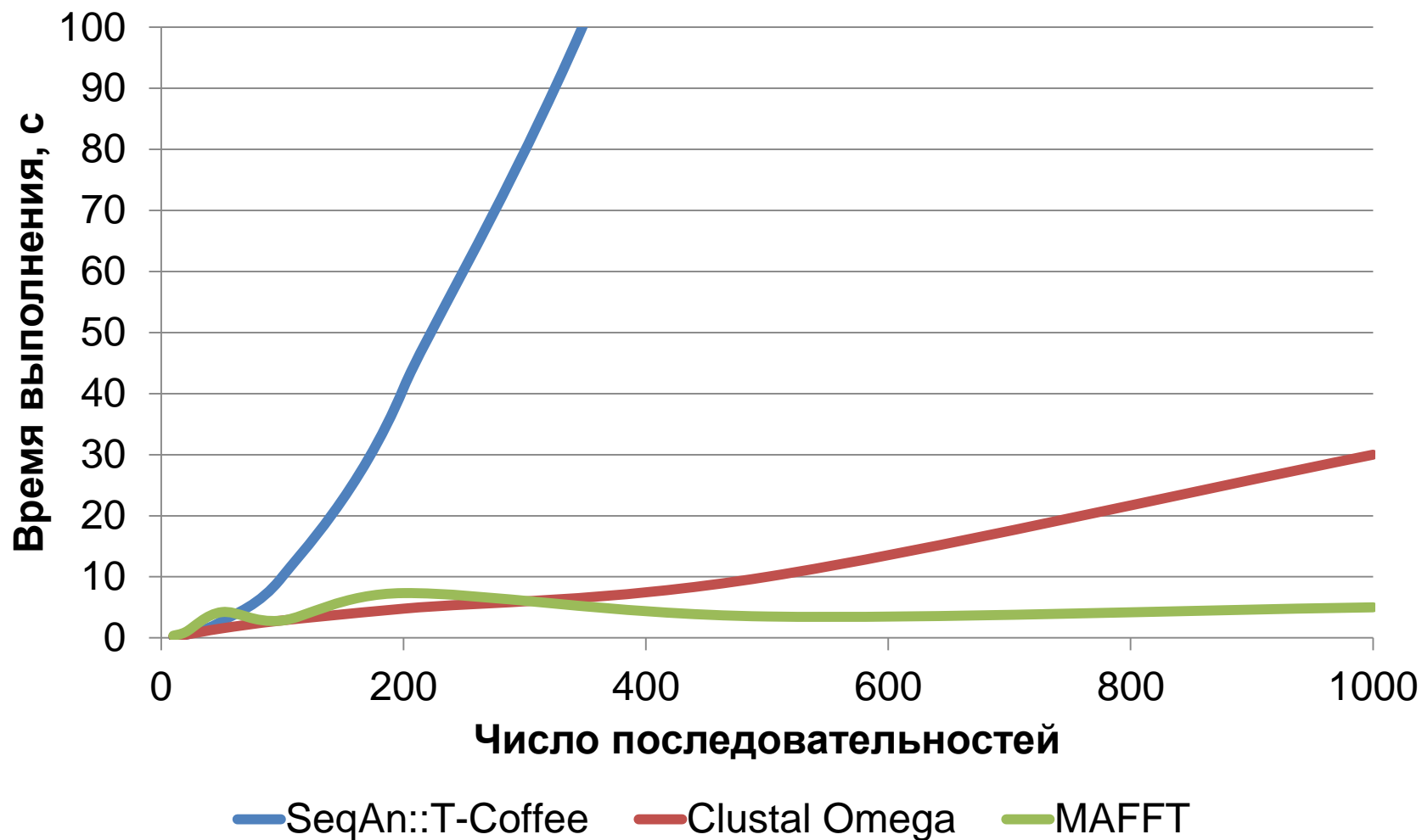
# Результаты



# Сравнение с другими инструментами



# Сравнение с другими инструментами



# Выводы

1. Изучен и проанализирован алгоритм работы SeqAn::T-Coffee
2. Выявлены 3 узких места
3. Для каждого из них предложены и проанализированы различные стратегии
4. Устранены все 3 узких места
5. Общее время работы уменьшилось в 32 раза с падением в качестве на 1%\*
6. При 1000 последовательностей качество выше чем у Clustal Omega на 0.7%

\* для 200 последовательностей, в сравнении с полной стратегией