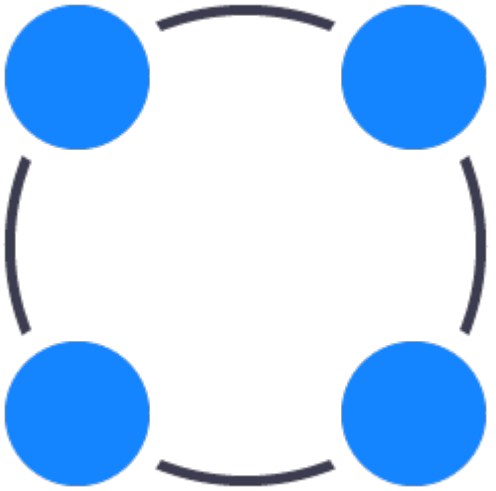
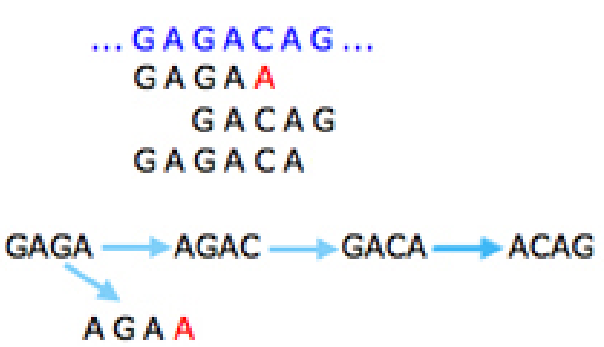


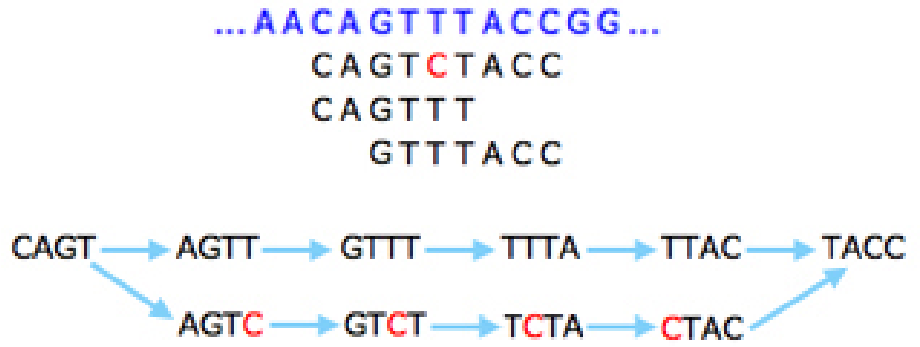
Упрощение графа Де Брёйна с помощью идеального хэширования

Выступающий: **Сергей Чернов**
Руководитель: **Антон Банкевич**

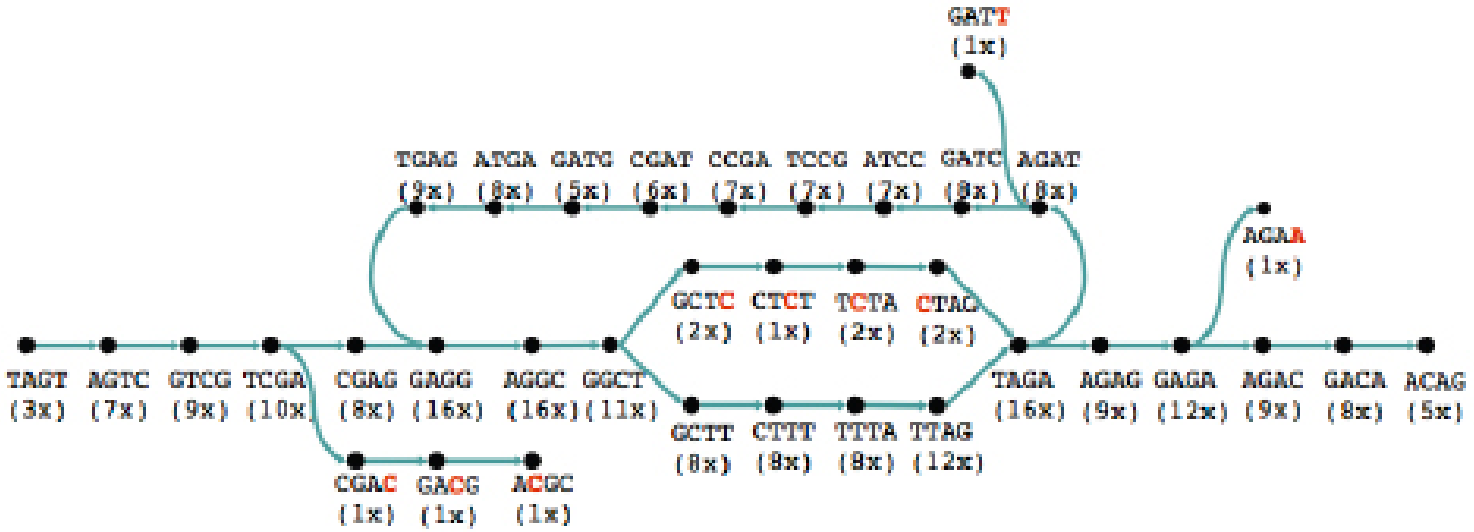
Проблемы сборки генома



Tip



Buldge



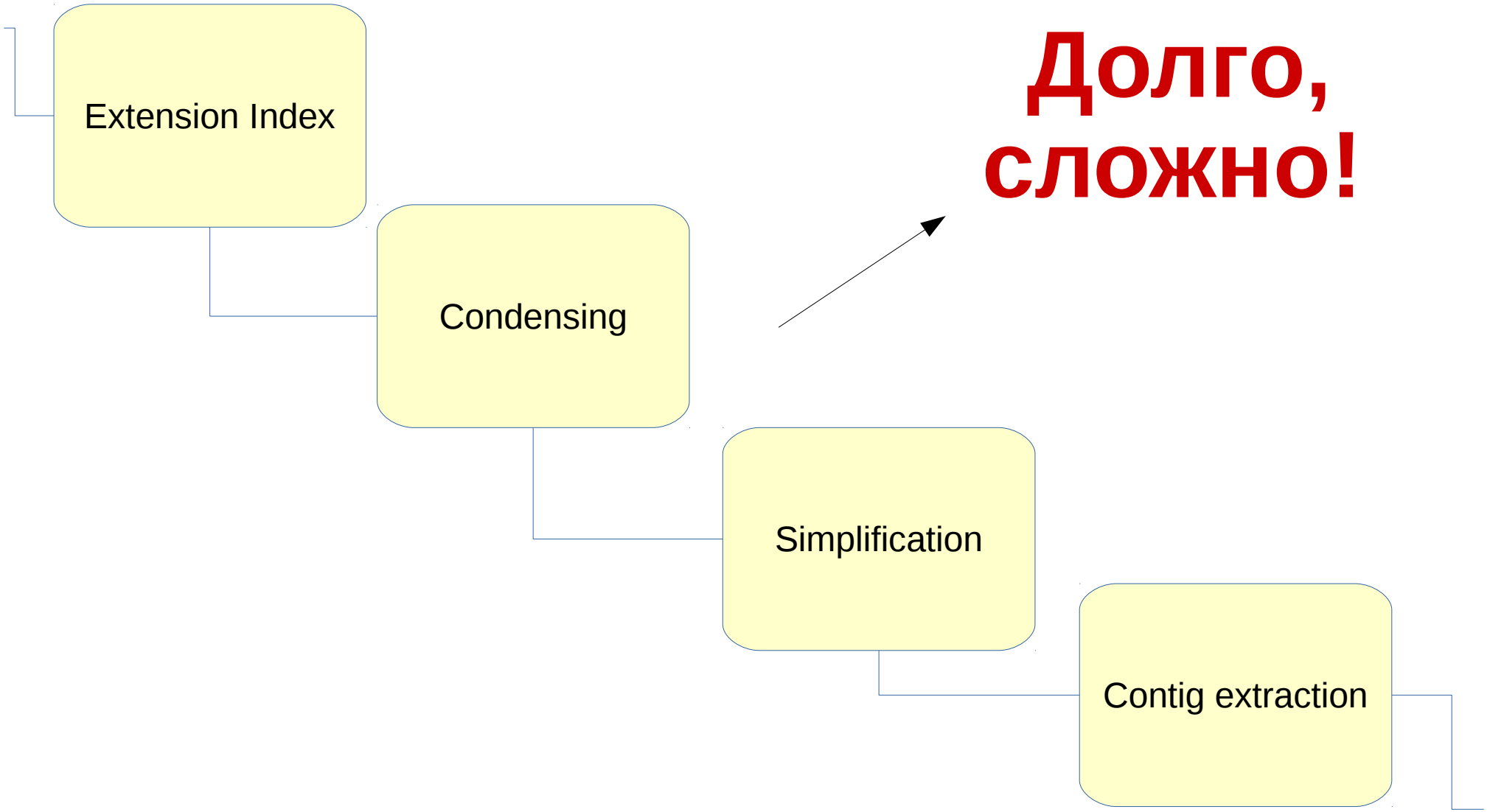
Проблемы сборки генома

- «Лишние» К-меры отнимают ресурсы:
  - Память
  - Время
- Меньше ошибок при сборке

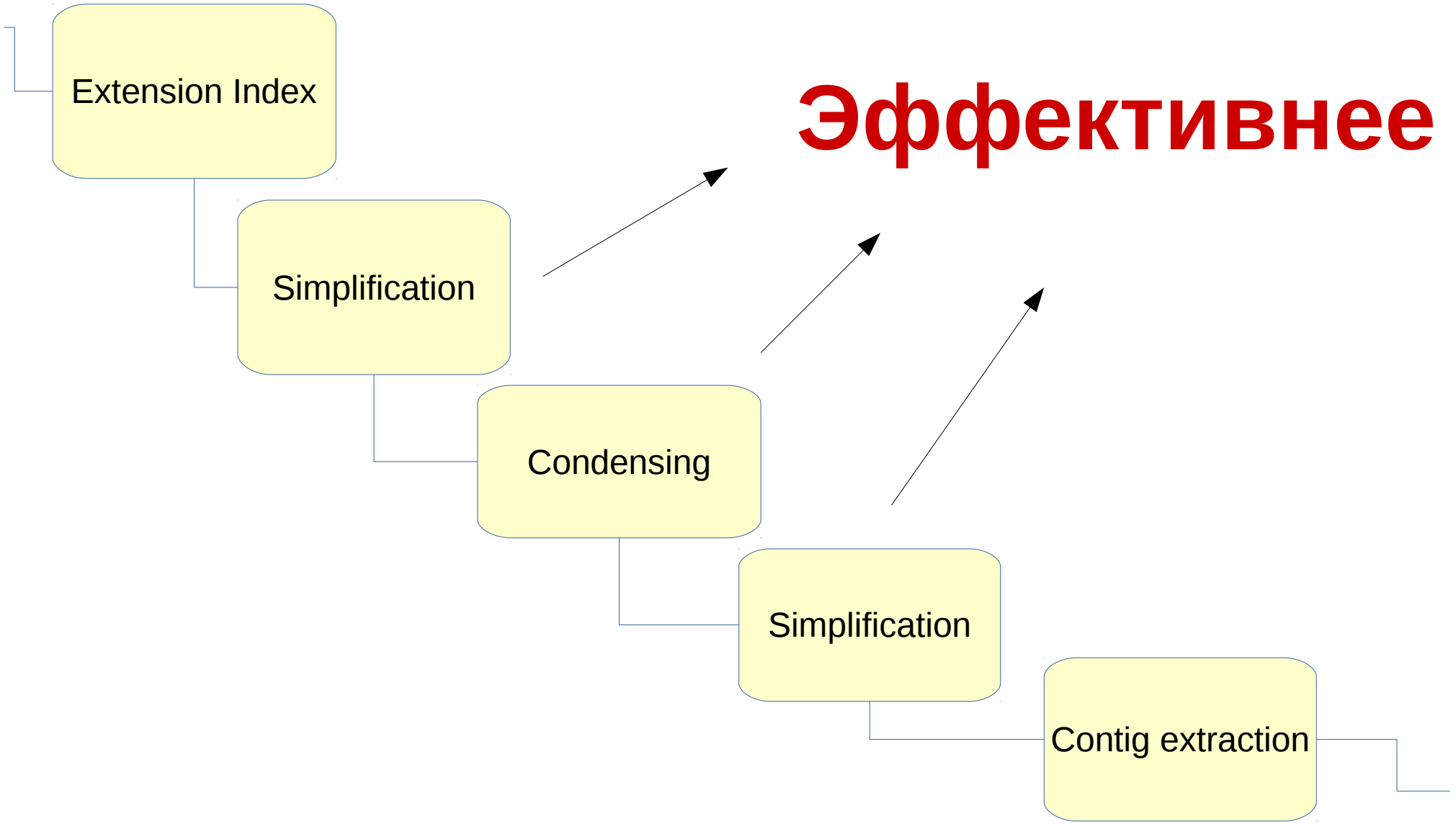
Задача

- Изучить внутреннюю структуру SPAdes
- Разработать структуру данных для вычисления покрытия K-меров
- Реализовать алгоритм удаления K-меров с малым покрытием **до** построения сжатого графа Де Брёйна
- Распараллелить

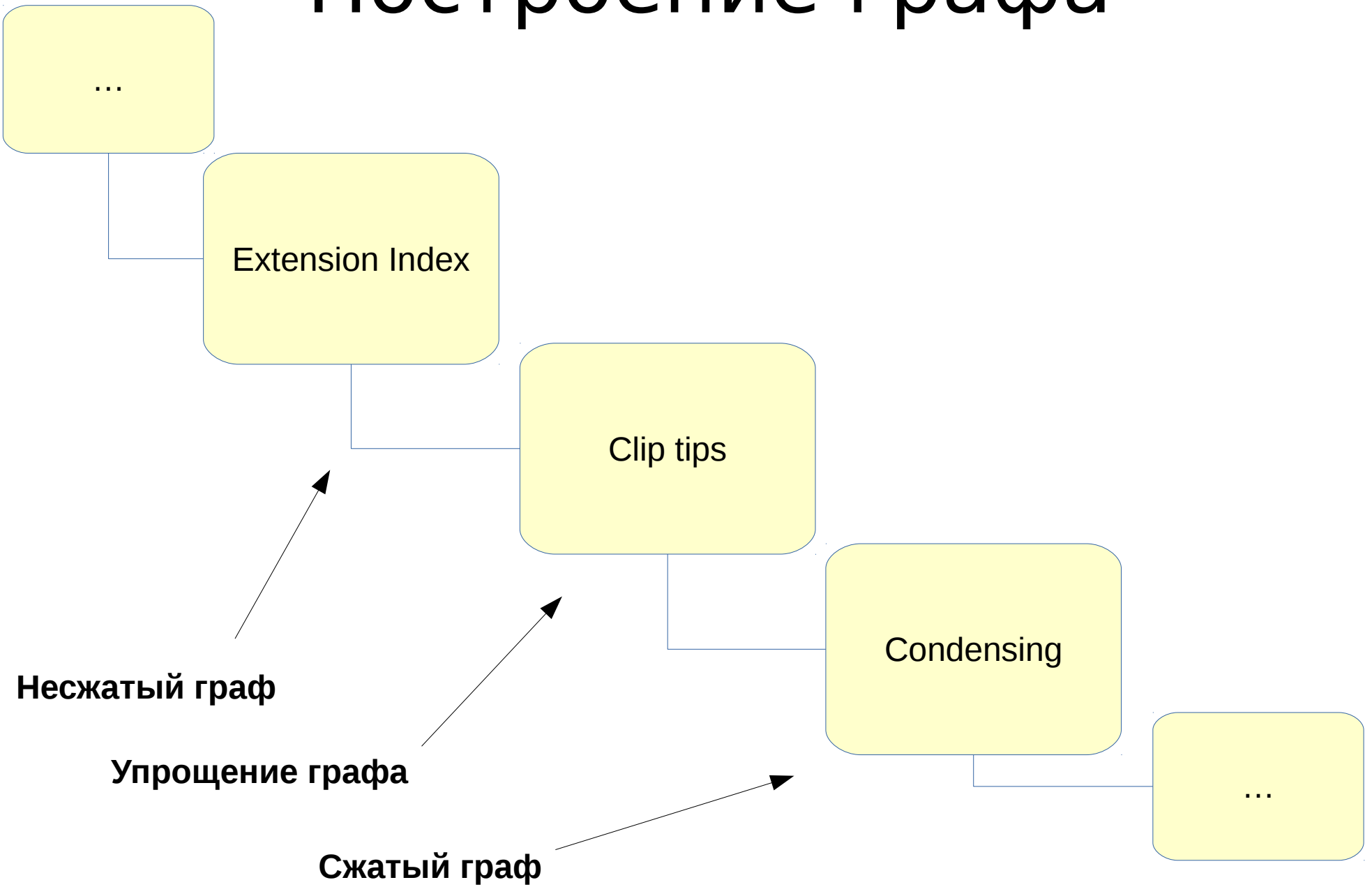
Построение графа



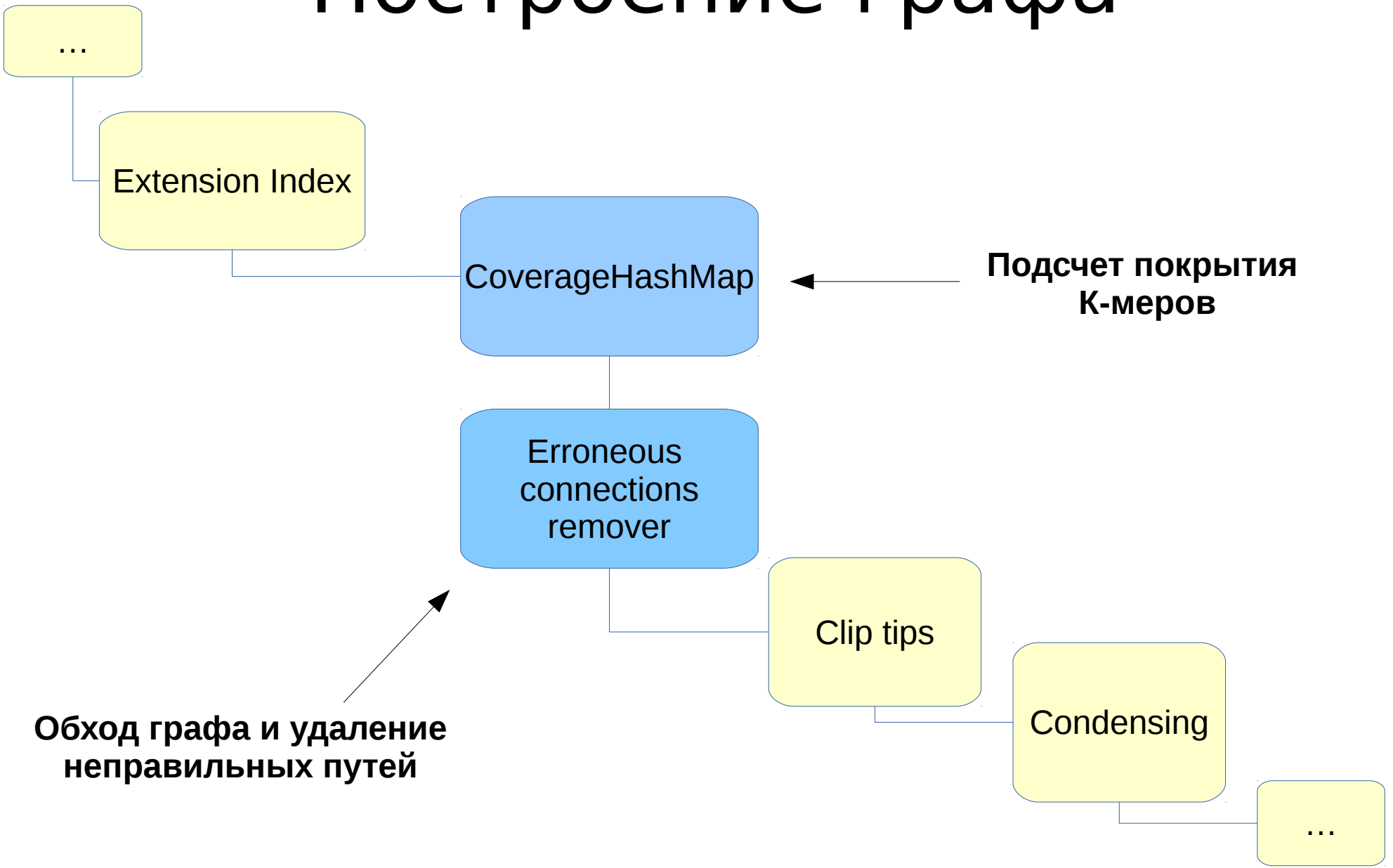
Построение графа



Построение графа



Построение графа



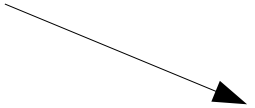
E.coli

K = 21

Total: 186 056 584

Erroneous: 33 069 123

Tips: 82 728 718

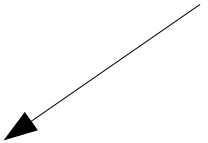


K = 33

Total: 209 211 365

Erroneous: 18 857 871

Tips: 108 025 415

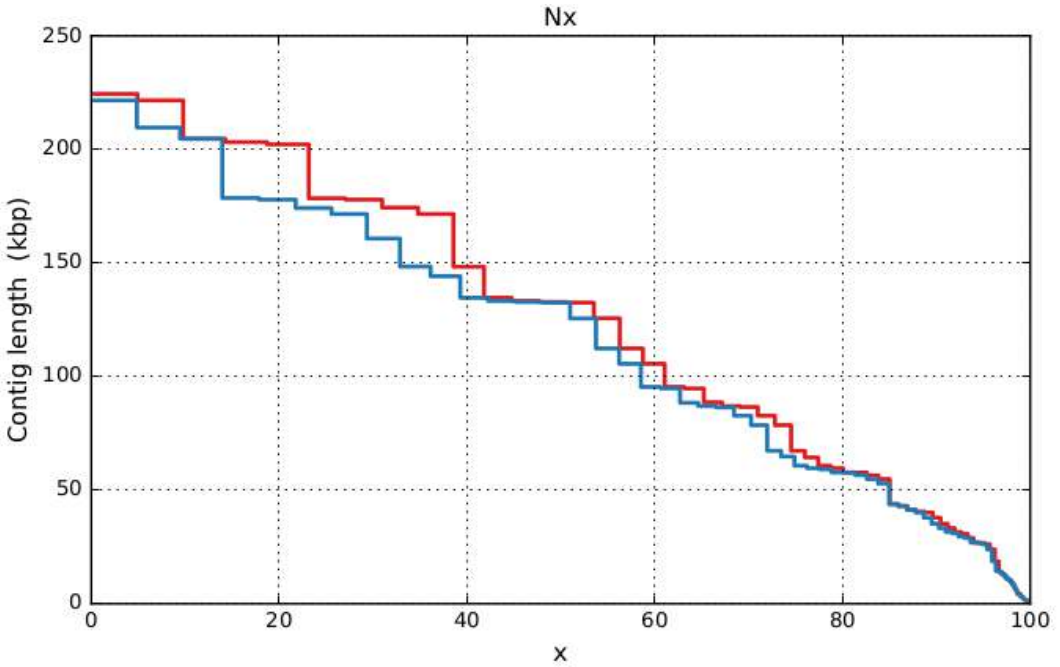


K = 55

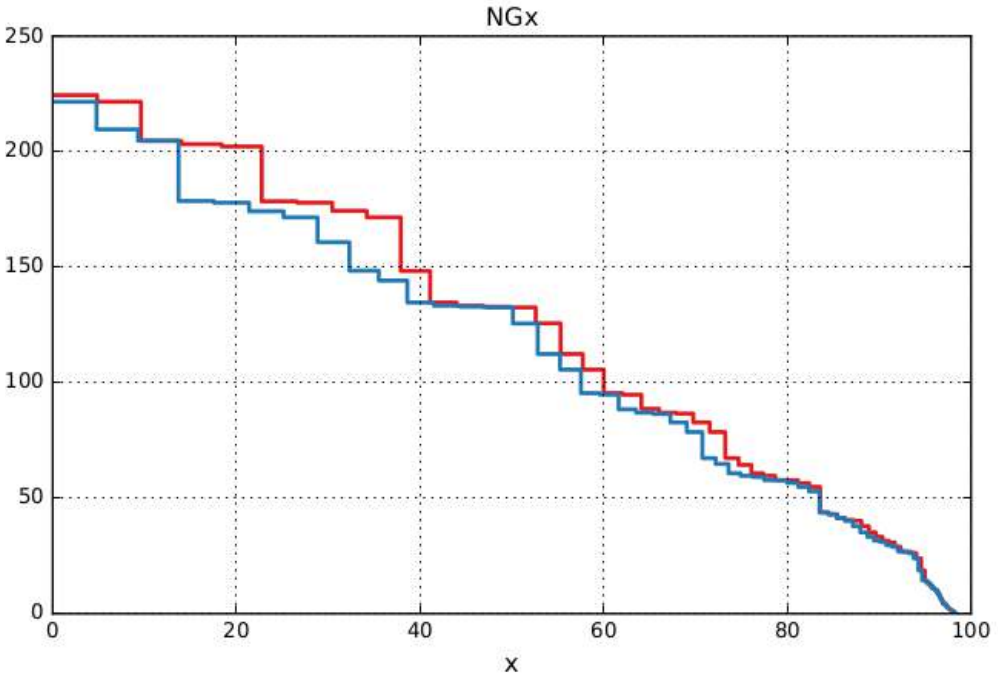
Total: 205 537 046

Erroneous: 6 638 389

Результаты QUAST



До После



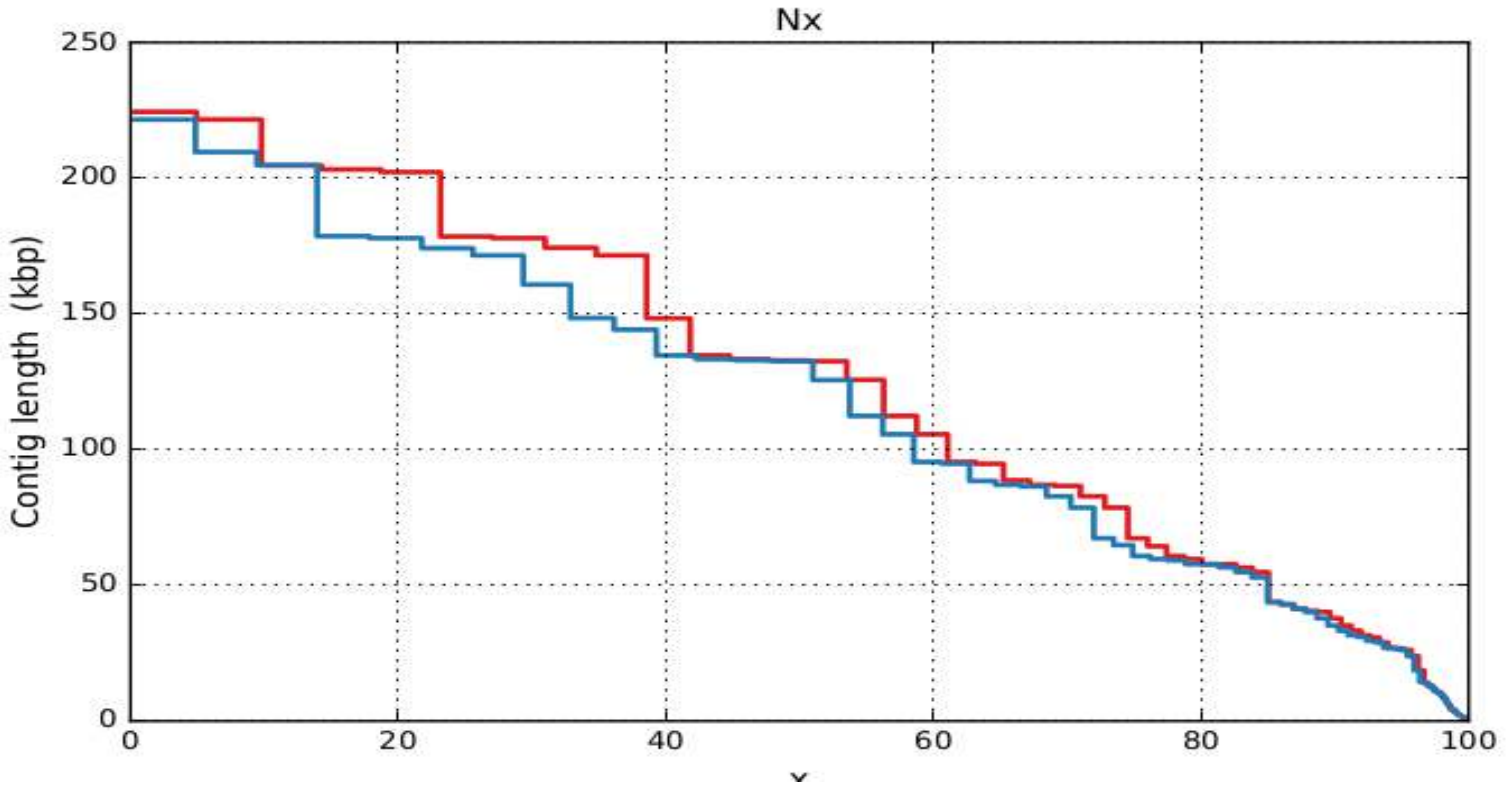
До После

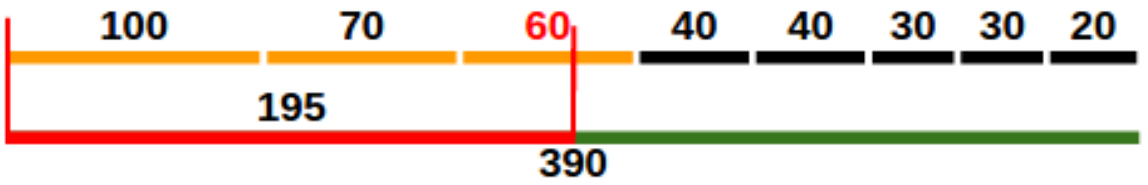
Итог

- Решены все поставленные задачи
  - Изучение кода SPAdes
  - Расширение пайплайна сборки
  - Распараллеливание

- В планах
  - еще больше сократить время сборки

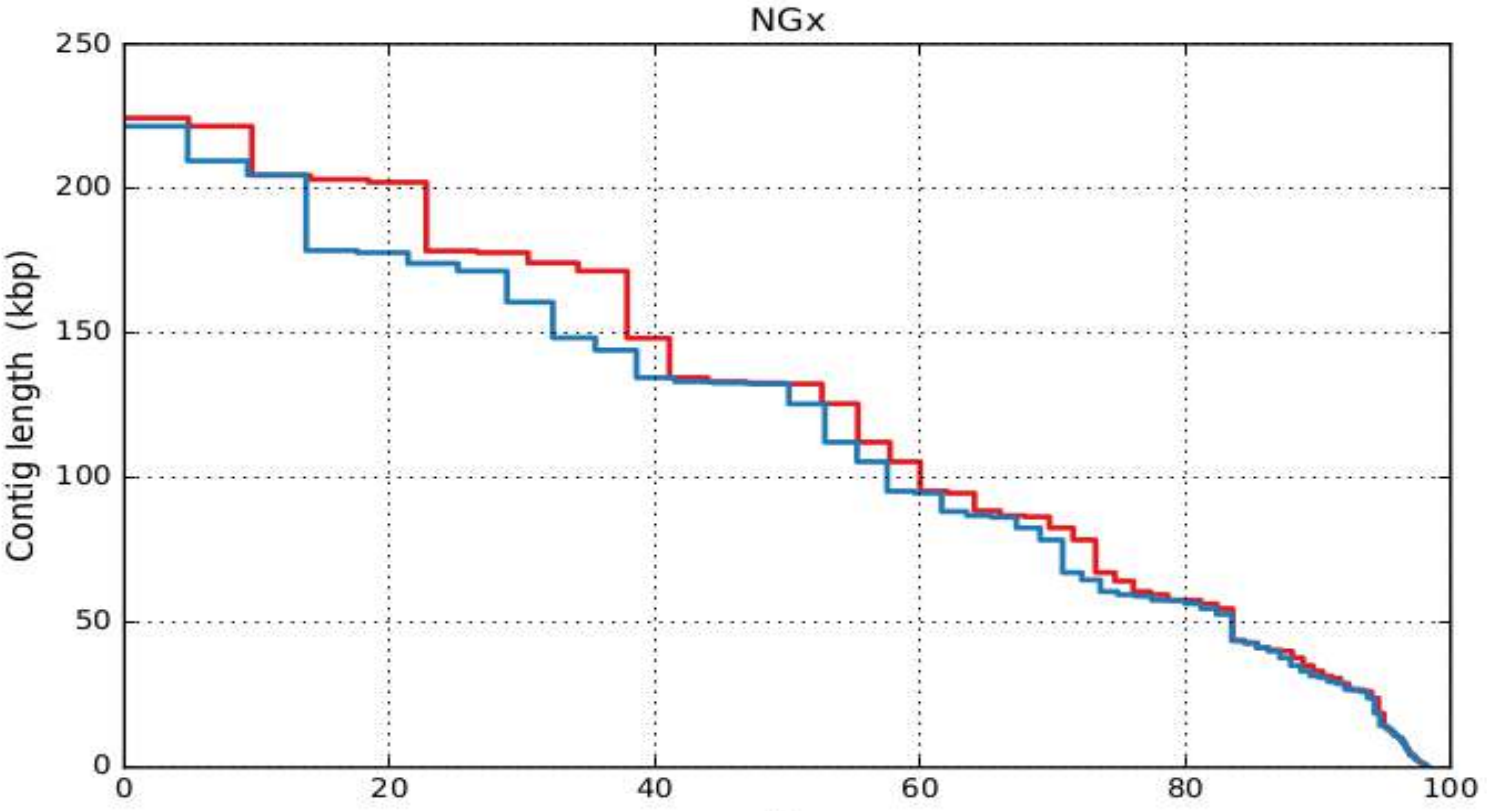
Что такое N50

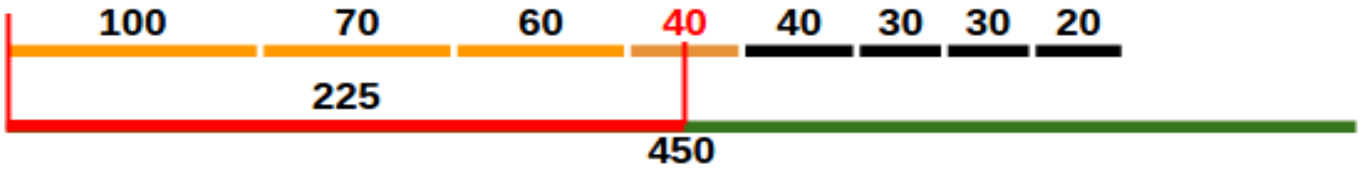




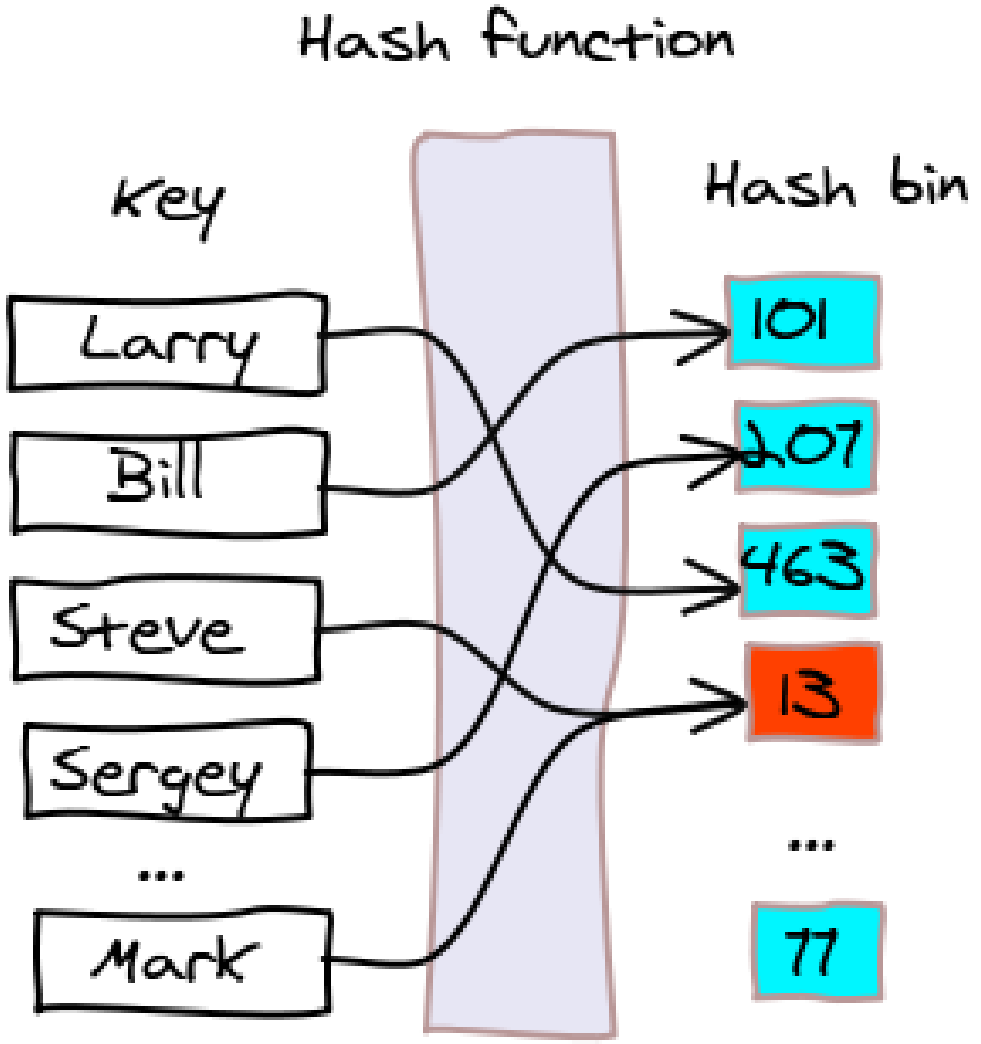
The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50**% of the assembly

Что такое NG50

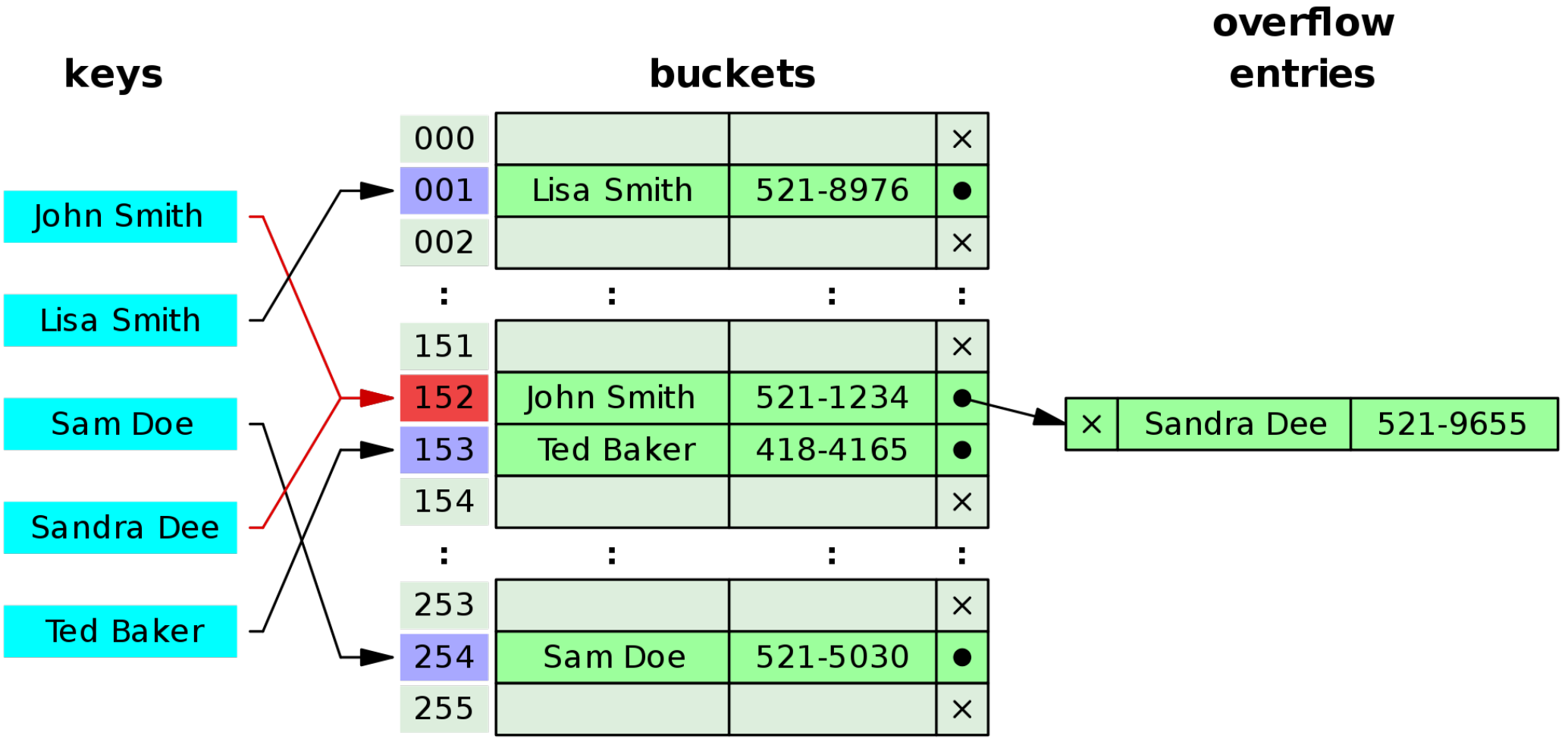




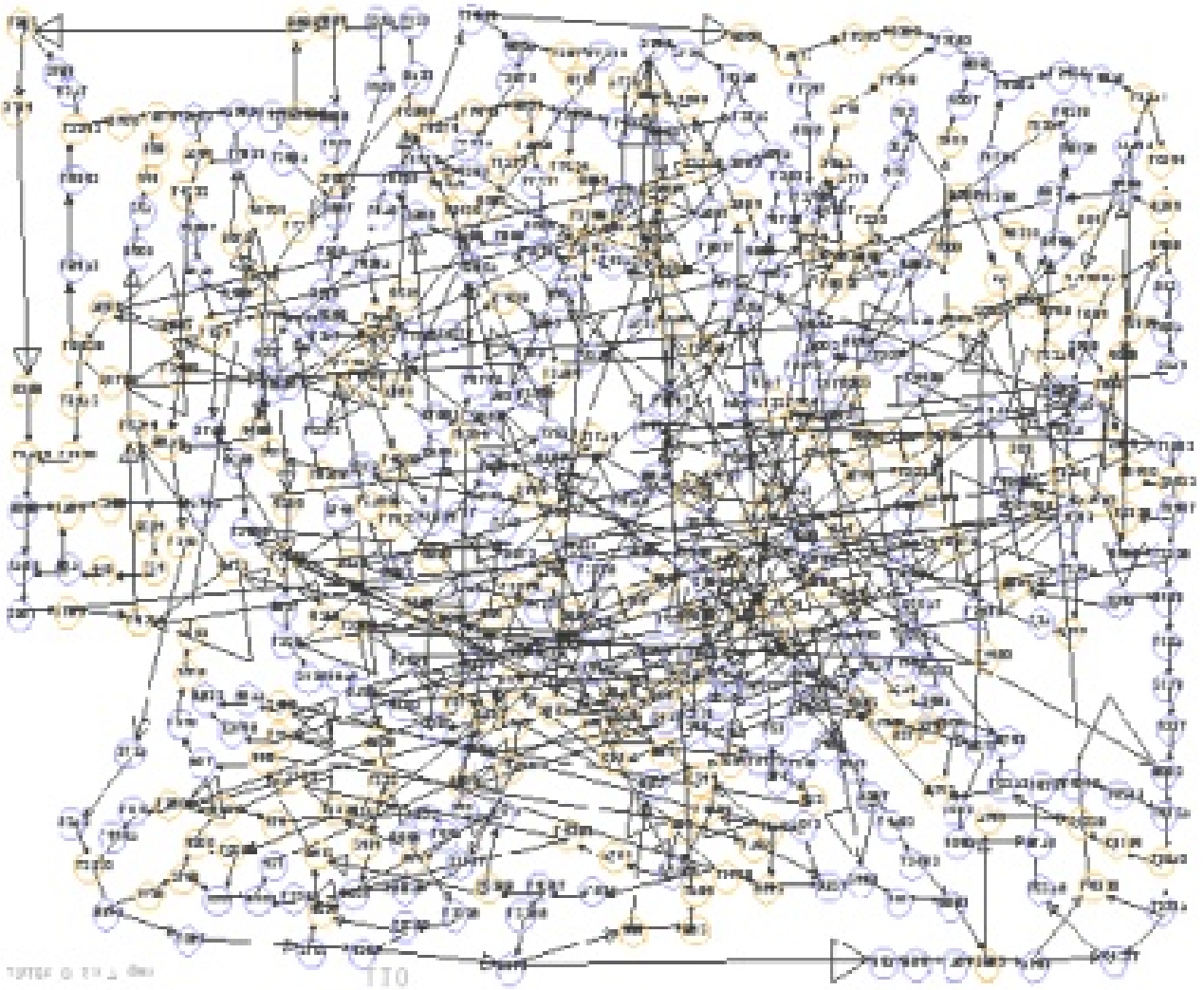
НЕИдеальное хэширование



НЕИдеальное хэширование



A more realistic graph



http://image.slidesharecdn.com/assemblingngsdata-torstenseemann-imb.jpg