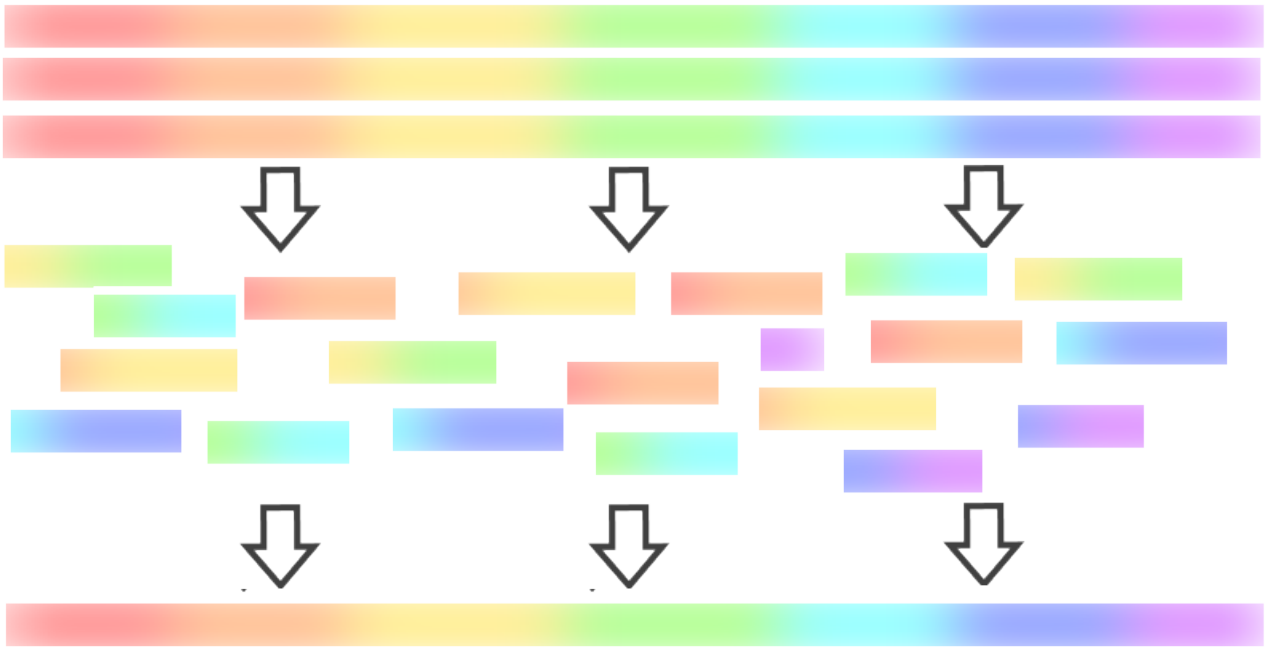
Построение скаффолдов по ридам РНК и визуализация графа связей между контигами

Черникова Ольга
Руководитель: Пржибельский Андрей

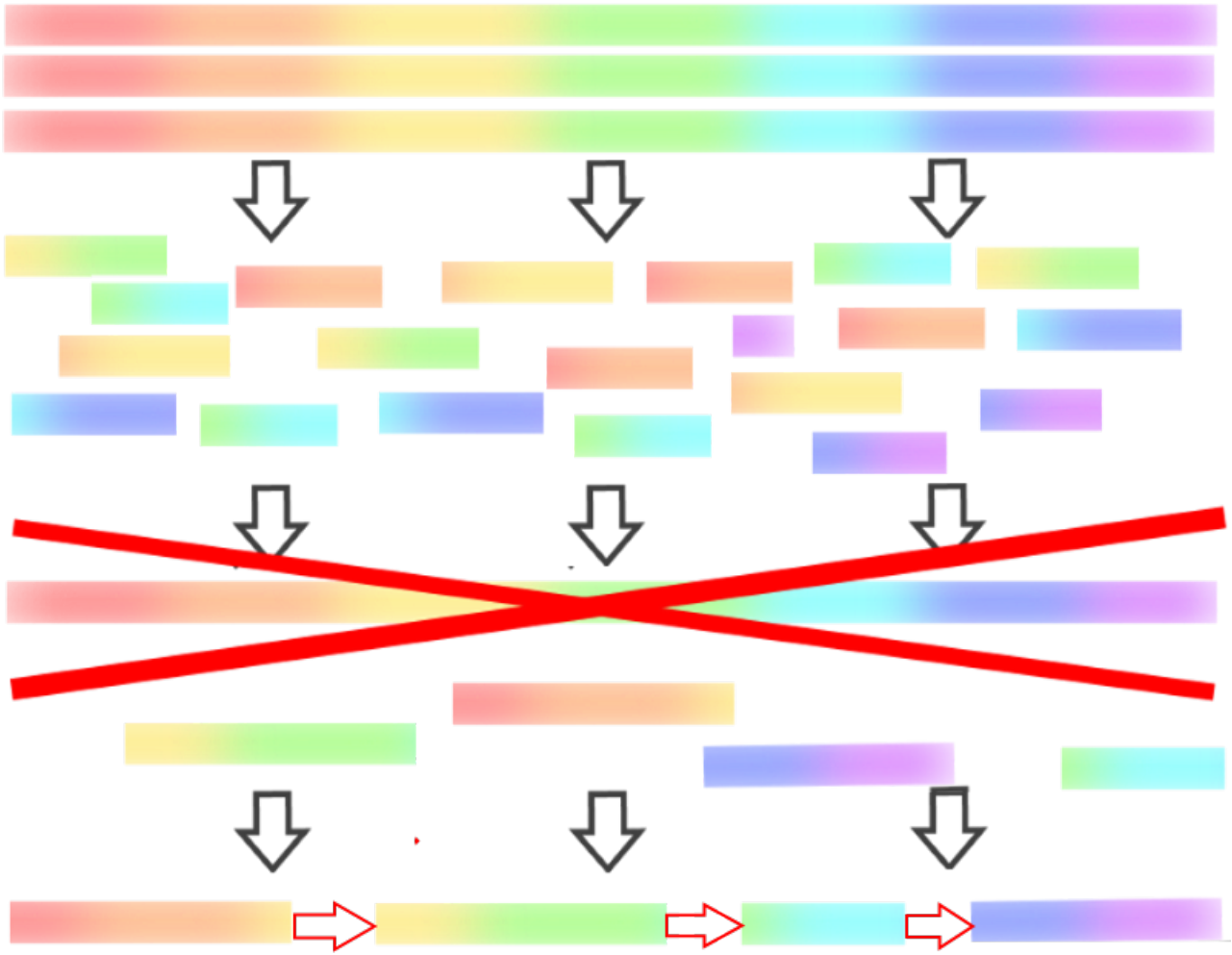
Центр алгоритмических биотехнологий

27.07.2017

Задача сборки генома



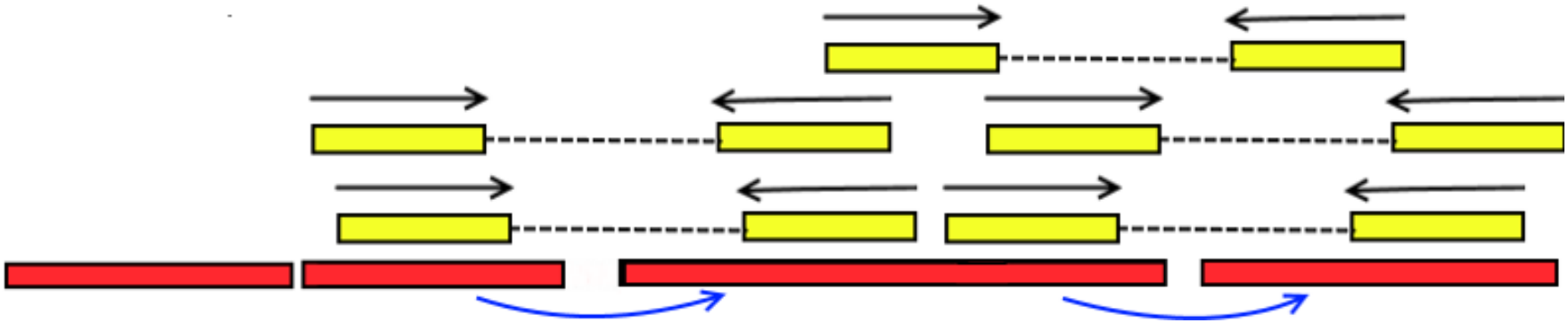
Задача сборки генома



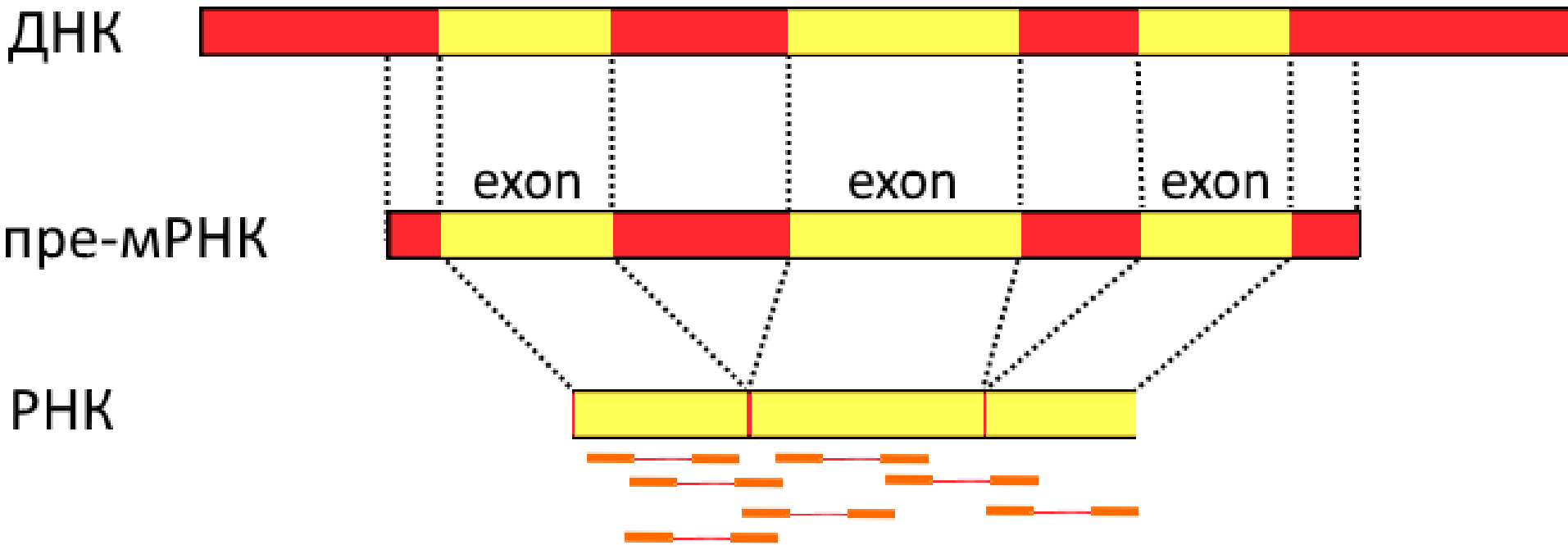
По парным ридам ДНК

- Парные риды:

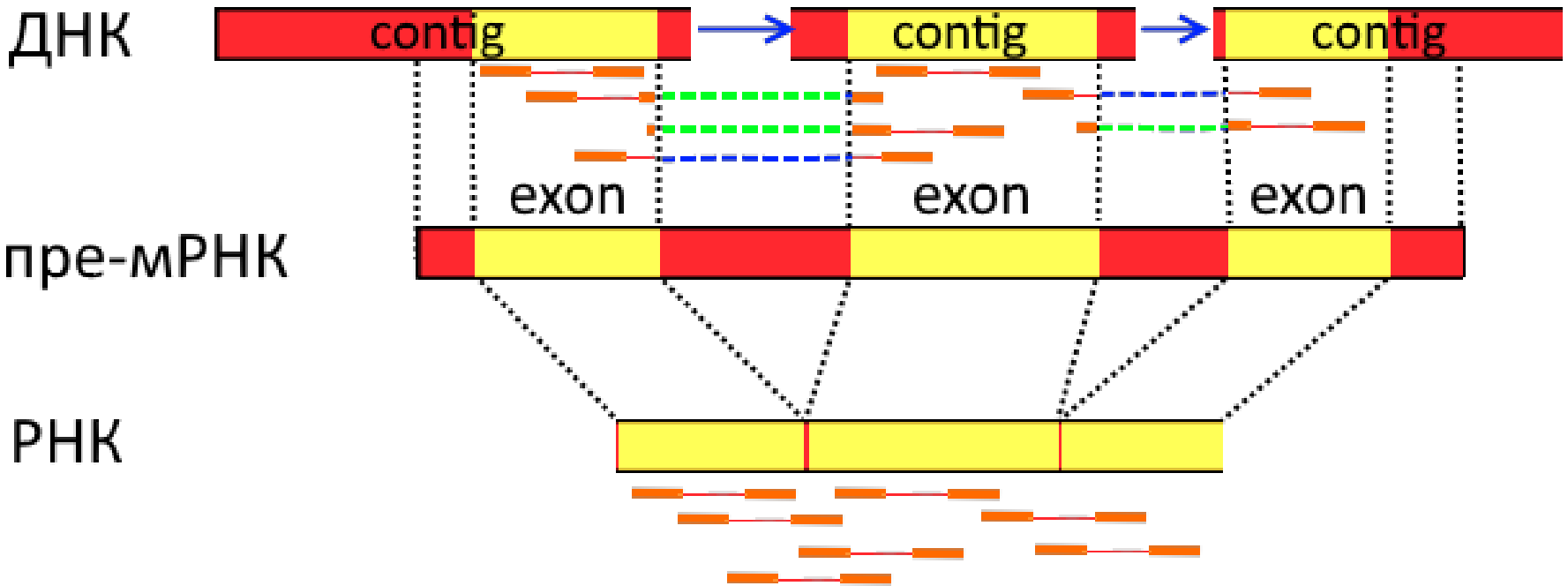
- Нахождение связей с помощью парных ридов:



По ридам РНК



По ридам РНК



Цель и задачи

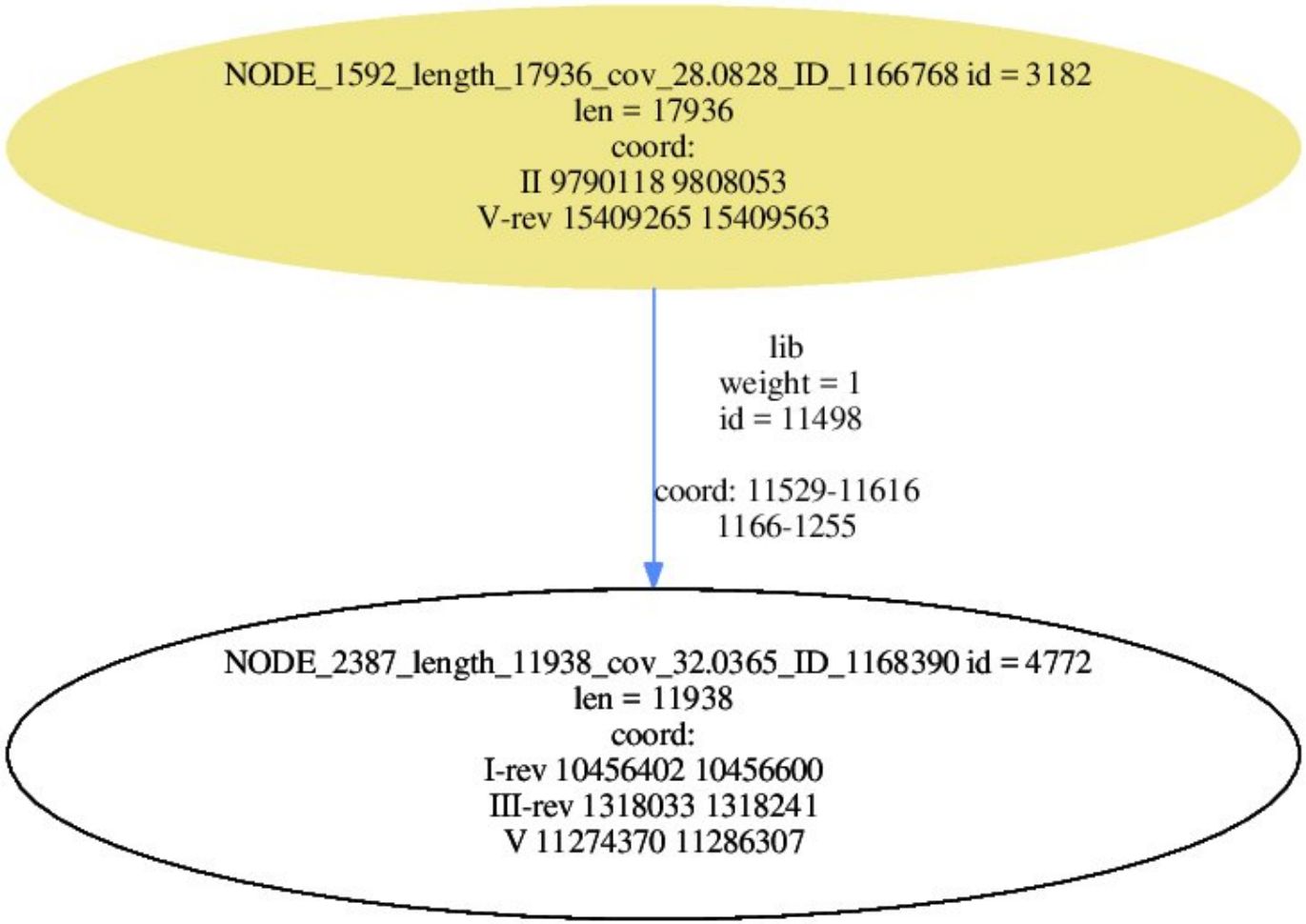
Цель

Построение скаффолдов по ридам РНК

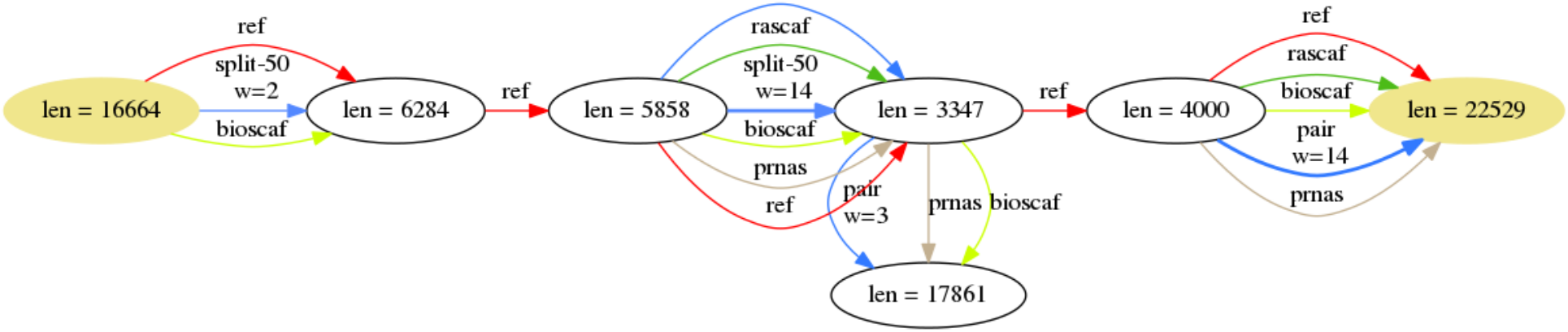
Задачи

- Построение графа связей
- Построение скаффолдов по полученным связям
- Создание инструмента для визуализации графа связей между контигами
- Сравнение получившихся результатов с результатами других инструментов для построения скаффолодов по ридам РНК

Визуализация



Визуализация



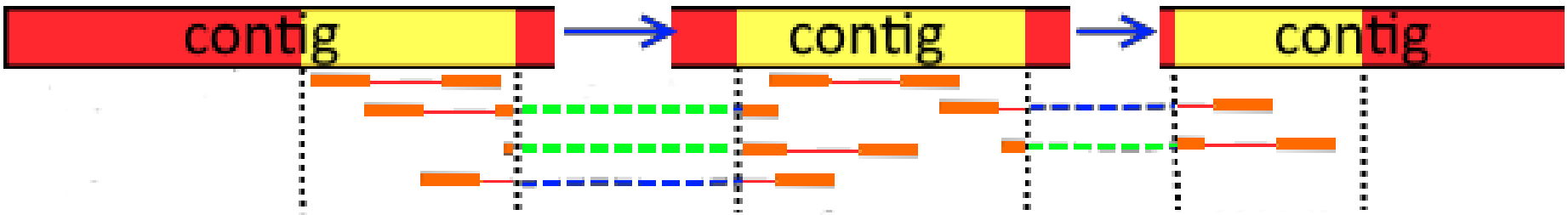
Визуализация

Возможность фильтрации графа:

- по весу ребер и размеров контигов
- вывод только участков с разнецей в двух библиотеках
- только участков, где есть одна библиотека и нет второй
- только участков с ошибочными соединениями
- и т.д.

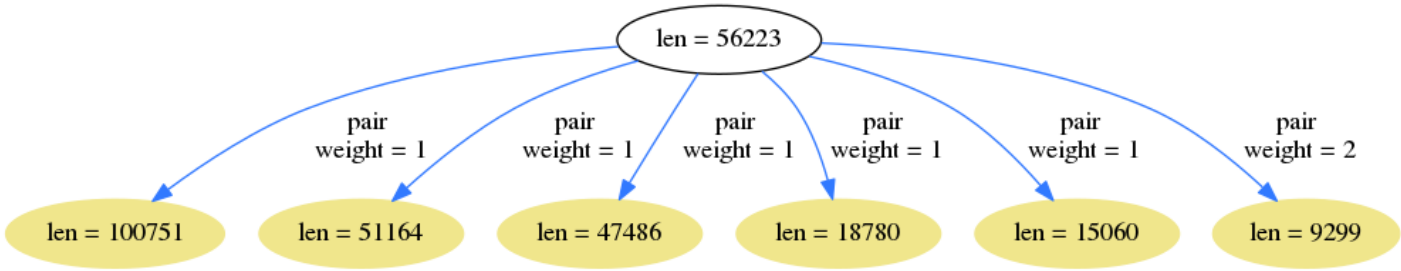
Построение графа связей

- выравнивание парных ридов PHK
- построения графа связей по парным ридам
- разрезание ридов на две части
- выравнивание половин ридов PHK
- построение графа по половинам ридов
- сохранение графа



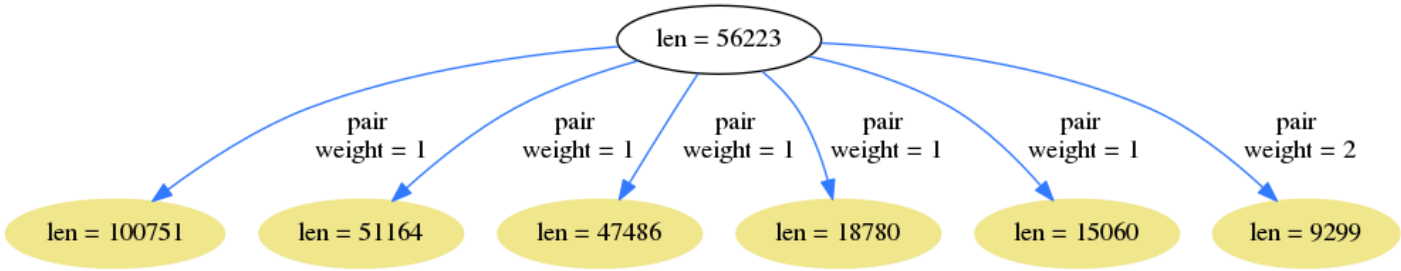
Упрощение графа

- удаление ребер маленького веса

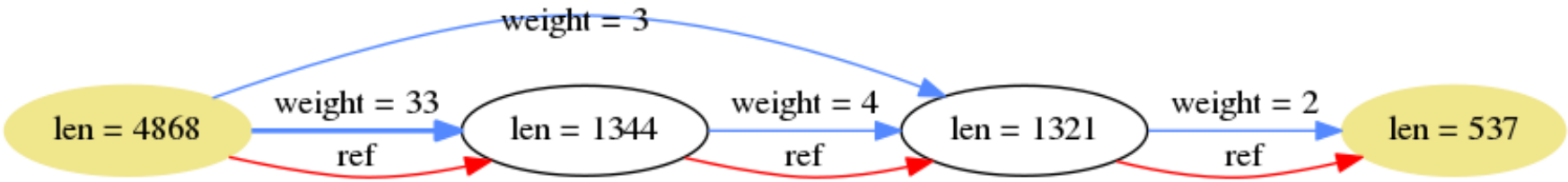


Упрощение графа

- удаление ребер маленького веса

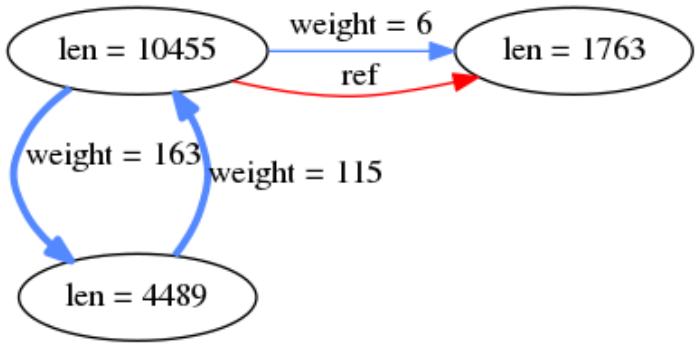


- проекция ребер



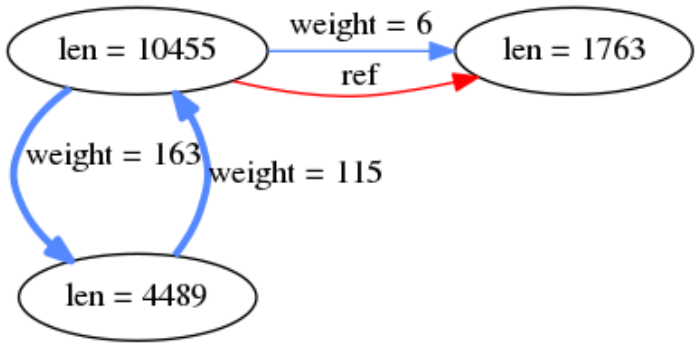
Упрощение графа

- удаление циклов

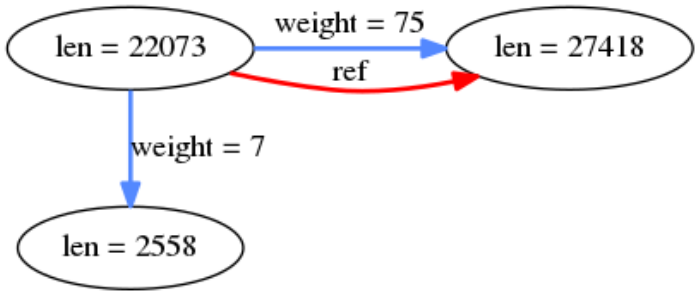


Упрощение графа

- удаление циклов

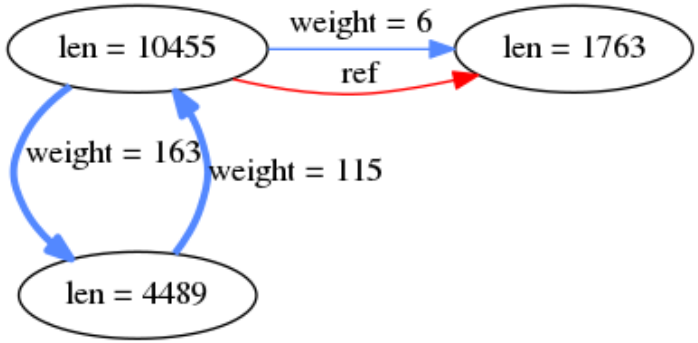


- развилки с большой разницей в весе

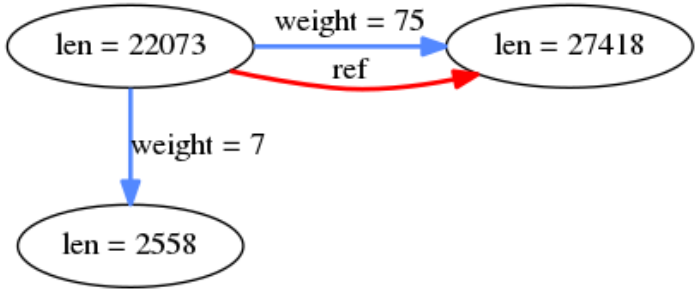


Упрощение графа

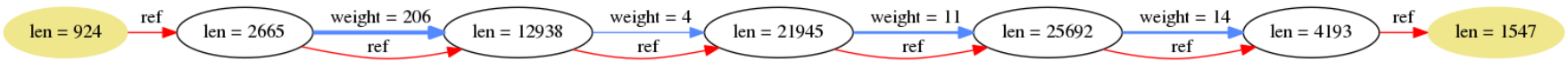
- удаление циклов



- развилки с большой разницей в весе



- соединение простых путей в скаффолды



Сравнение

C.elegans, SRR1560107

	bio_scaffolder	P_RNA_scaffolder	rascaf
NG50	36855	36075	32879
NG75	17299	17188	18395
NGA50	30383	28828	27116
NGA75	12735	12489	11667
LGA50	918	955	995
misassemblies	529	621	521

Использованные инструменты

- Язык разработки - **C++**
- **SeqAn** - библиотека для работы с файлами в SAM/BAM и fasta/fastq форматах.
- **gtest** - библиотека для тестирования.
- Программы для выранивания - **STAR, nucmer, bowtie2**.
- **QUAST** - для анализа качетсва сборки.
- **Tablet** - для визуализации выравненых ридов.

Результаты и планы

Результаты

- Создание программы для построения скаффолдов по данным РНК
- Создание инструмента для визуализации графа сязей

Дальнейшее развитие

- Тестирование и сравнение на большем разнообразии данных
- Ускорение работы приложения
- Реализация новых идей для построения скаффолдов
- Написание документации и удобного интерфейса
- Написание статьи

Спасибо за внимание

Репозиторий: https://github.com/olga24912/bio_scaffolder