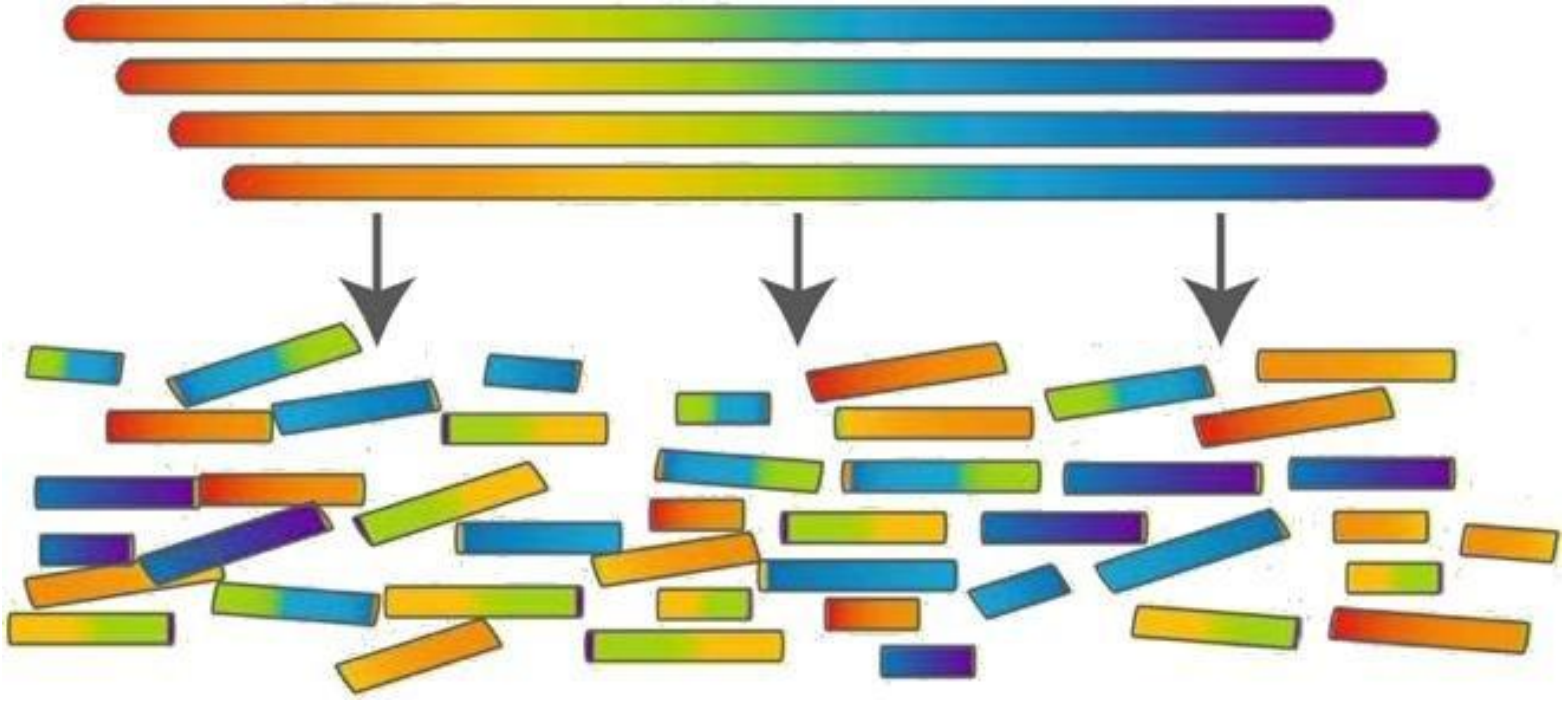
Сборка генома
Часть II

Антон Банкевич
Сергей Нурк
Лаборатория вычислительной биологии
АУ РАН
http://bioinf.spbau.ru

В предыдущей серии

- Секвенирование
- Сборка
- Графы де Брюина
- Борьба с разрывами
- Исправление ошибок
- Способы представления графа

Секвенирование



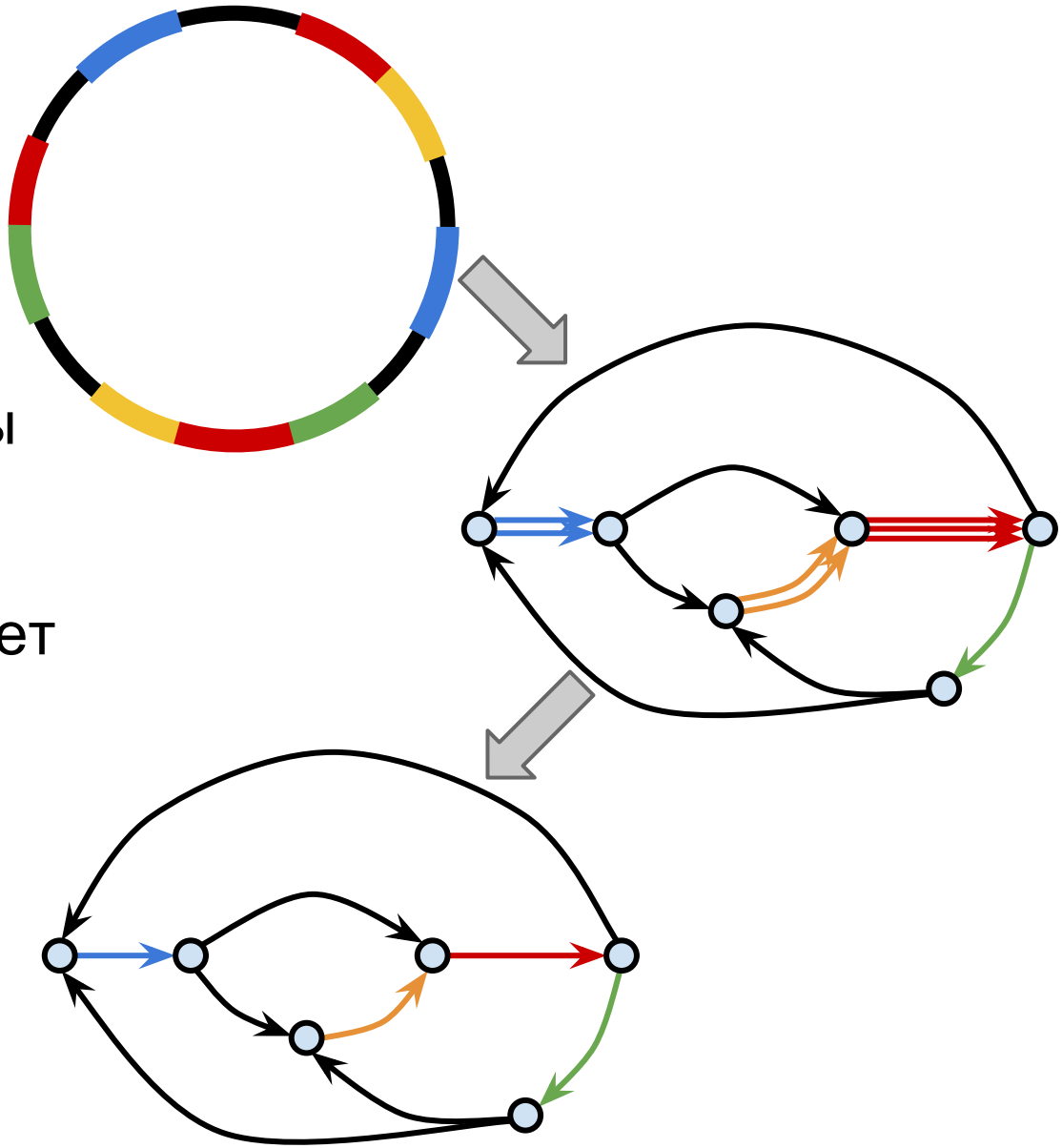
Задача сборки

Получить последовательности нуклеотидов (контиги), которые:

- являются фрагментами генома
- подлиннее
- имеют поменьше перекрытий
- получше покрывают геном

Граф де Брюйна

Заметки про граф де Брюйна



1. Склеивает повторы (длиннее **k**)
2. Геном соответствует циклу в графе
3. **Ребра сжатого графа можно рассматривать как контиги**

Техника

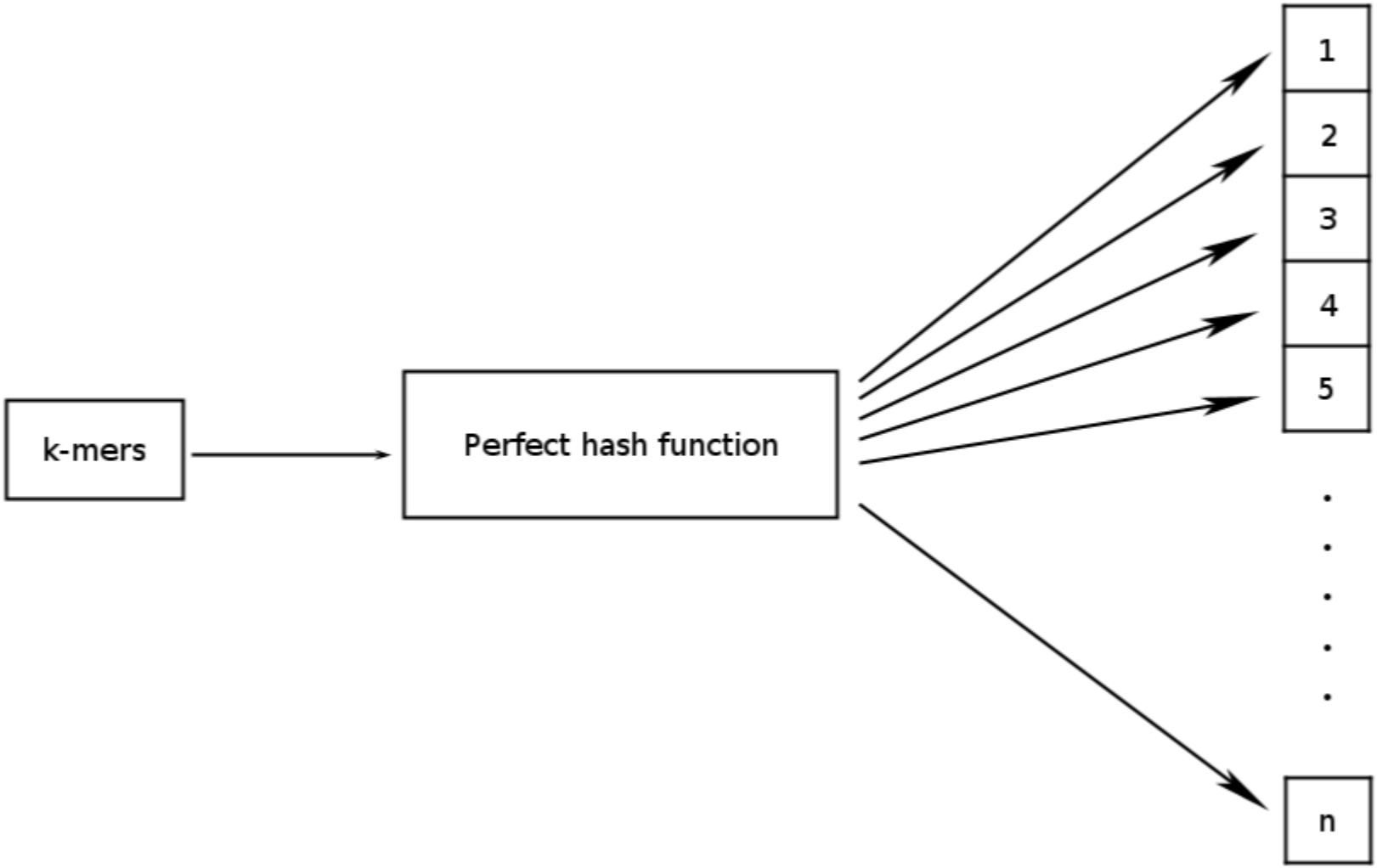
Представление графа

Требования:

- Возможность перебрать все *k*-меры
- Возможность найти соседей *k*-мера

Пример: Множество всех *(k+1)*-меров

Хэширование без коллизий



Хэширование без коллизий

Позволяет:

- Хранить информацию в массиве
- Не хранить ключи

Требует:

- Предварительного нахождения уникальных ключей

Не позволяет:

- Проверять наличие произвольного элемента в множестве

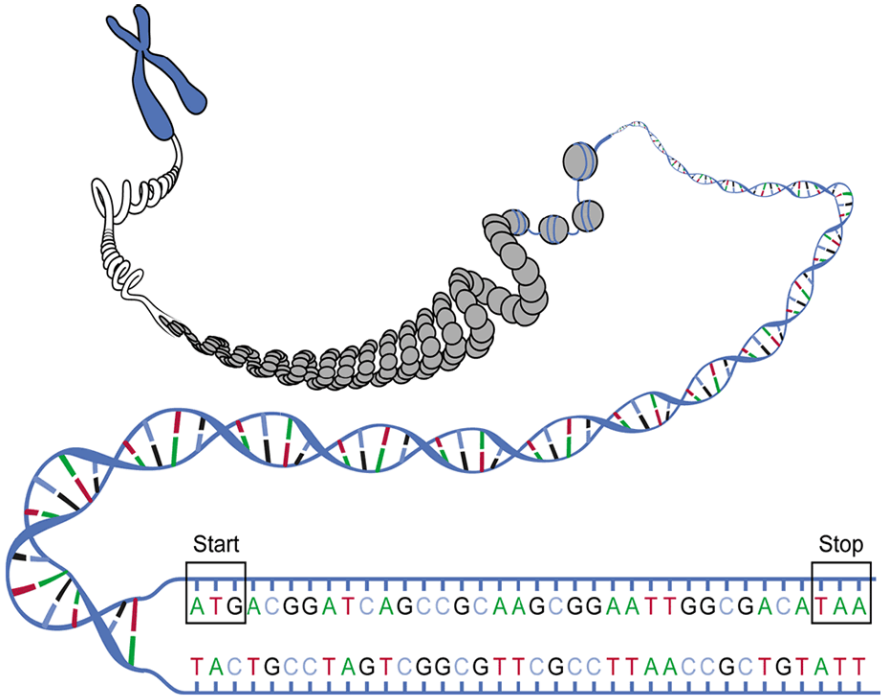
Реализация графа де Брюйна

- Ключи — *k*-меры
- Для каждого k-мера хранятся все его соседи (8 бит)

Распределенное хранение

- Позволяет собрать что-то на кластере
- *k*-меры распределяются по нодам в соответствии с некоторым хэшем
- Чего хочется от хэша?
- На порядок медленнее

Двустрендовость



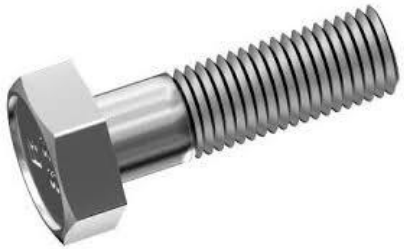
→ CCCAGAACTGAGATCAAT →

← GGGTCTTGACTCTAGTTA ←

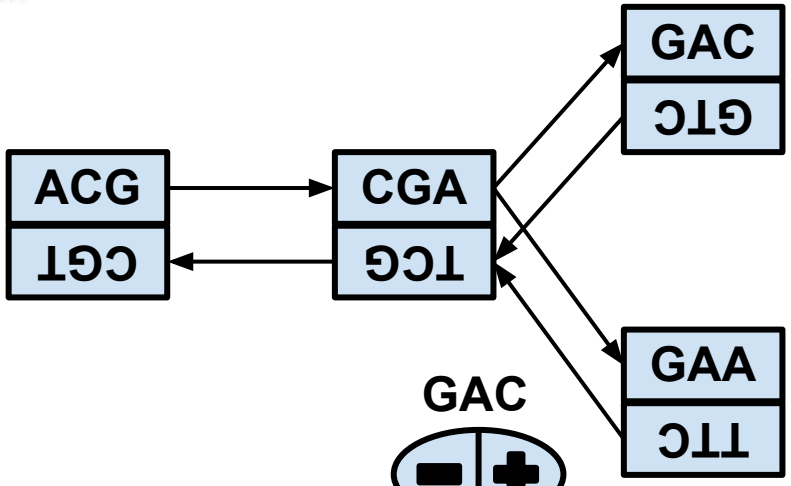
Двустрендовость

Стратегии:

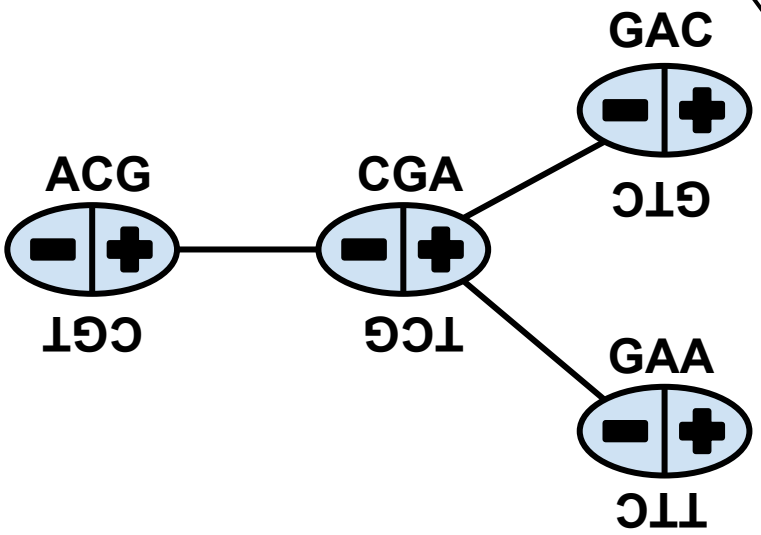
1. Игнорировать



2. Синхронизированный удвоенный граф

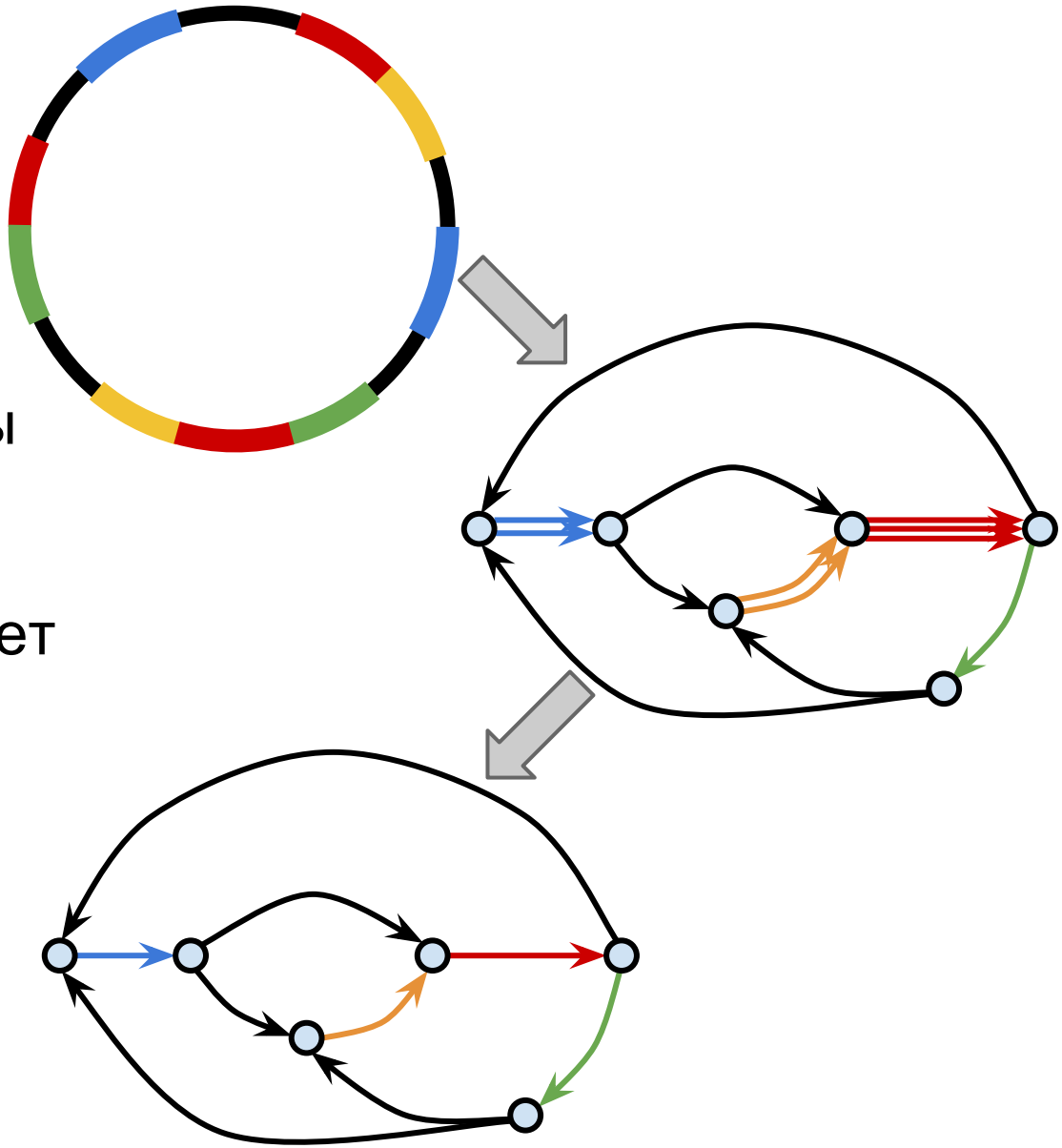


3. Двунаправленный граф



Борьба с повторами

Заметки про граф де Брюйна

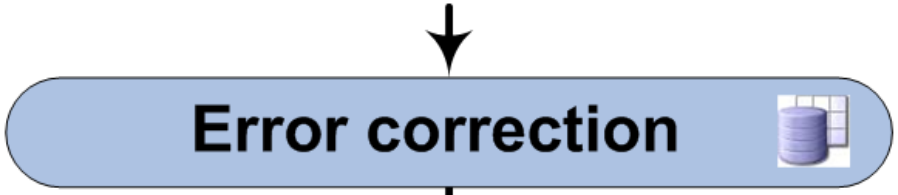


1. Склеивает повторы (длиннее **k**)
2. Геном соответствует циклу в графе
3. **Ребра сжатого графа можно рассматривать как контиги**

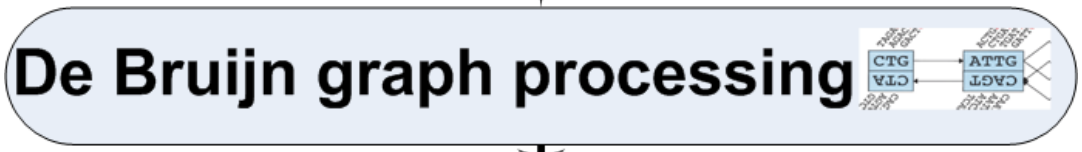
Некоторые типы повторов

- Low-Complexity DNA
- Microsatellite repeats
- Gene Families
- Segmental duplications
- SINE Transposon
- LINE Transposon
- LTR retroposons

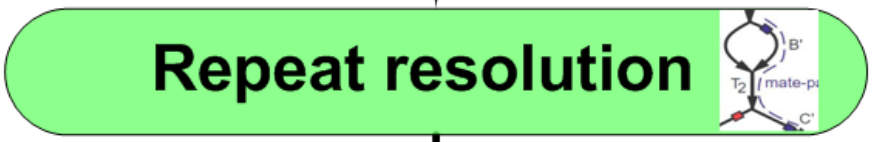
Original reads



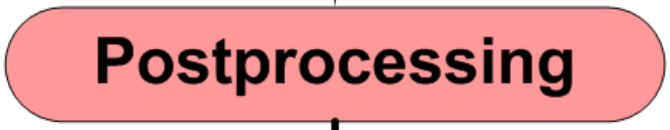
Corrected reads



Simplified DBG

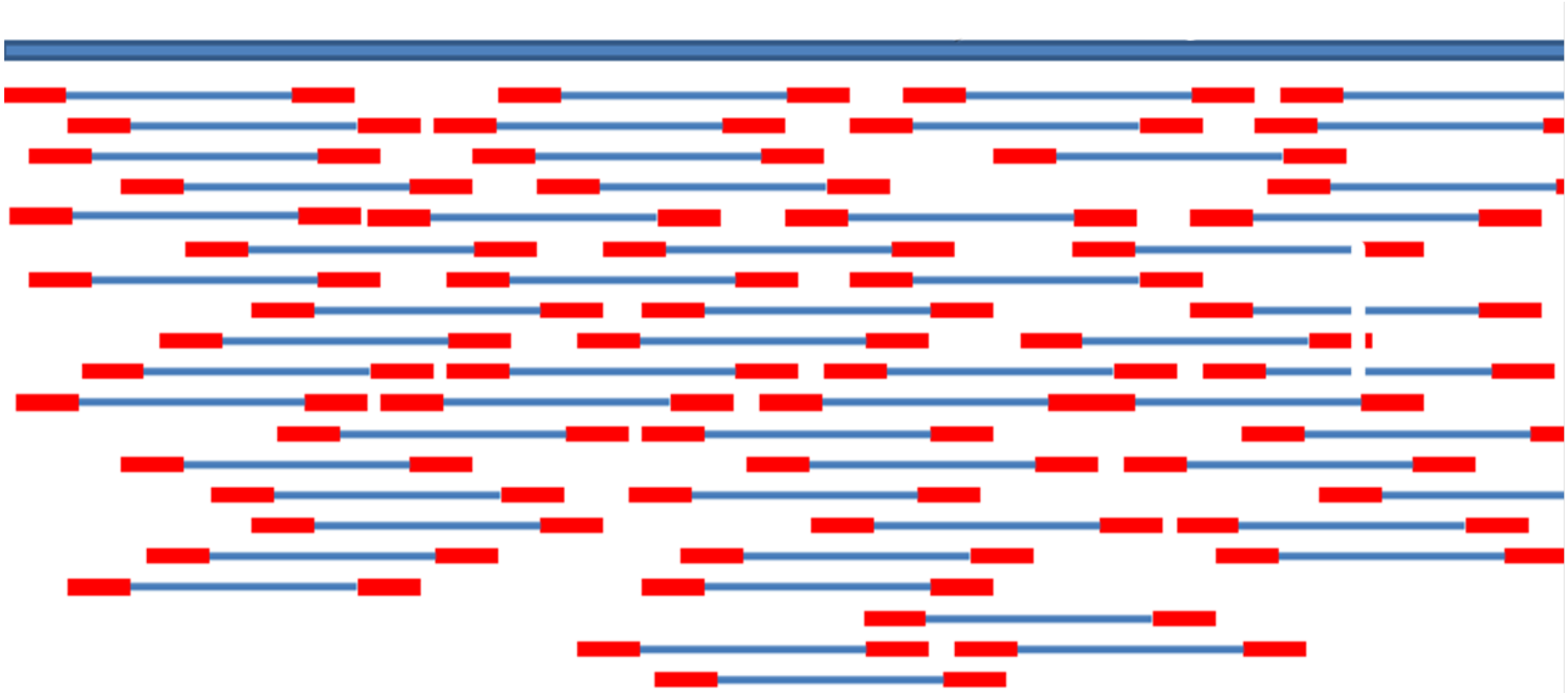


Contigs



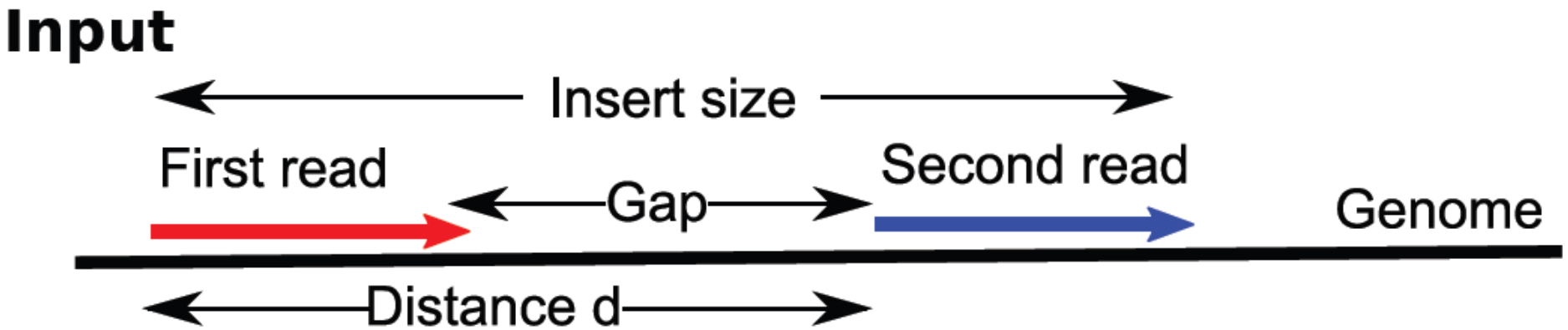
Final contigs

Парные риды



Для техногии Illumina, типичная длина ридов 100-200bp

Параметры парного рида

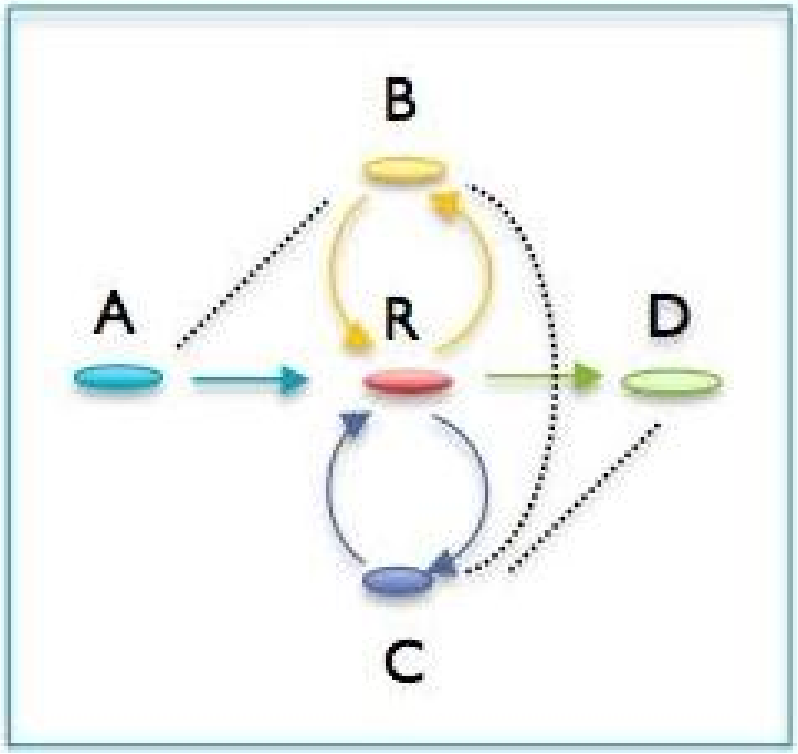


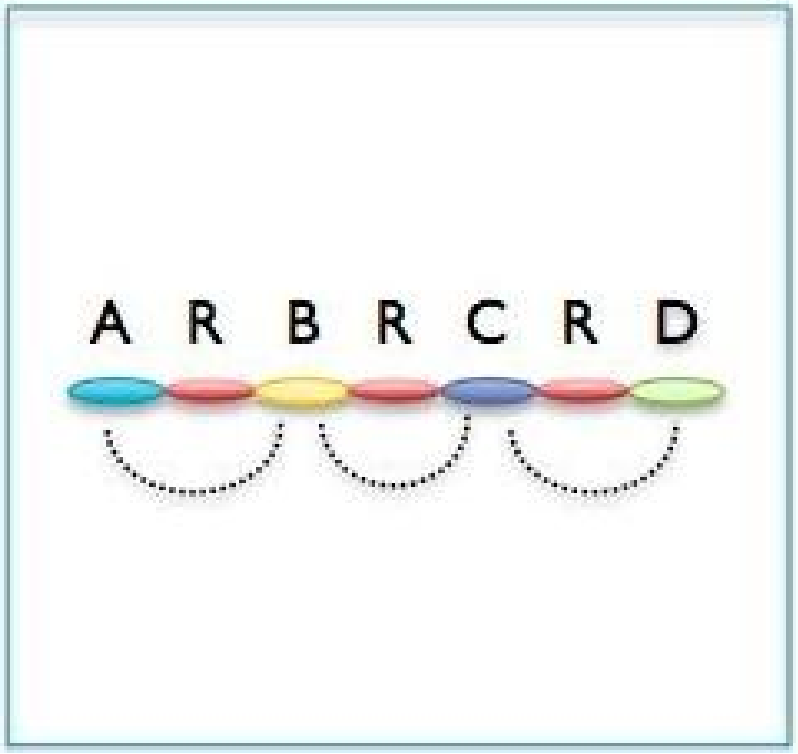
Типы библиотек ридов

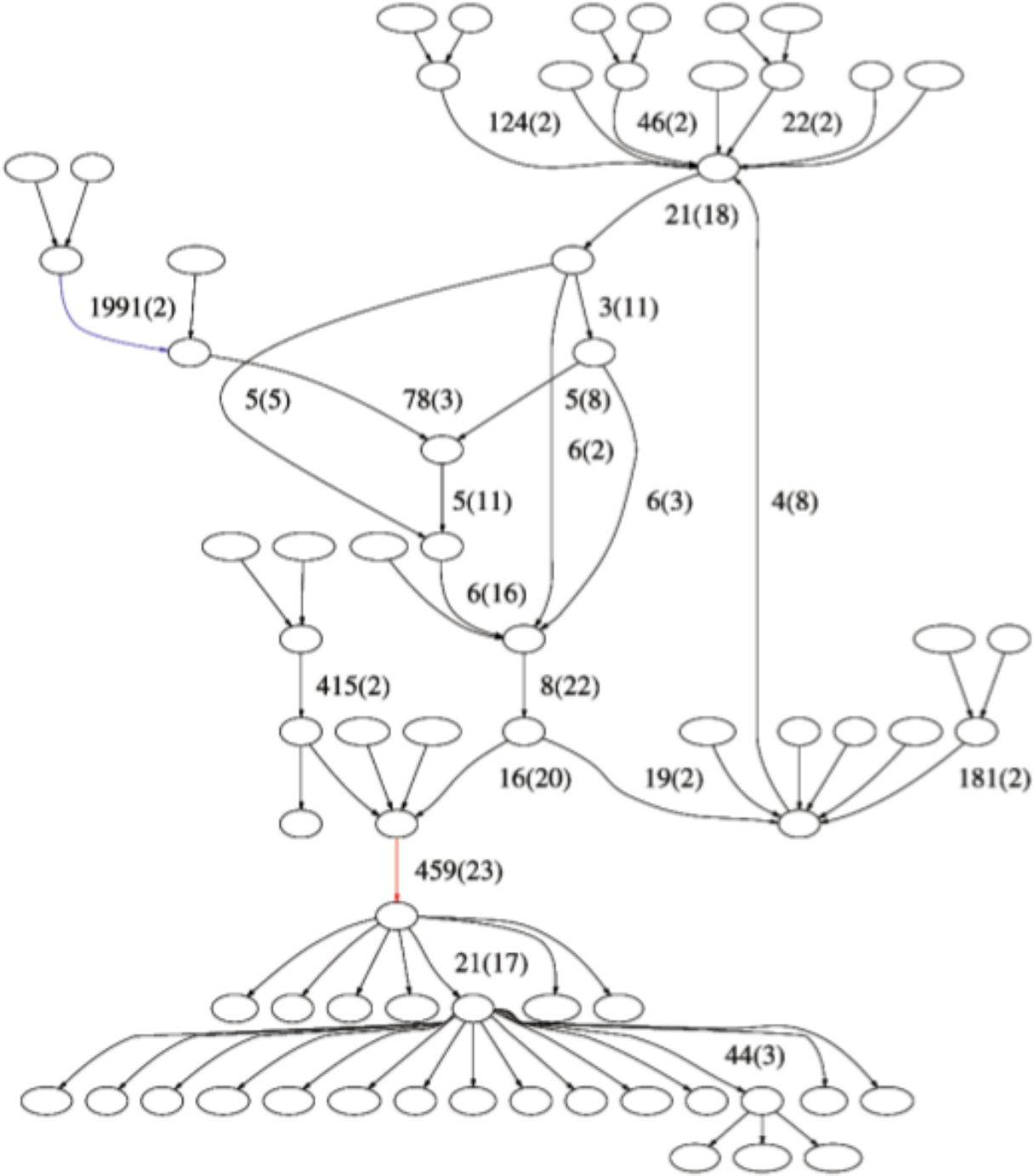
Paired-end (200-500 bp)

Mate pair (2-5 kbp)

Разрешение повторов (наивный подход)






124(2)
46(2)
22(2)
21(18)
1991(2)
3(11)
5(5)
78(3)
5(8)
6(2)
5(11)
6(3)
4(8)
6(16)
415(2)
8(22)
16(20)
19(2)
181(2)
459(23)
21(17)
44(3)

Стратегии использования "парной" информации

1. Преобразование структуры графа
2. Извлечение контигов из графа
3. "Скаффолдинг" (с заполнением промежутков)

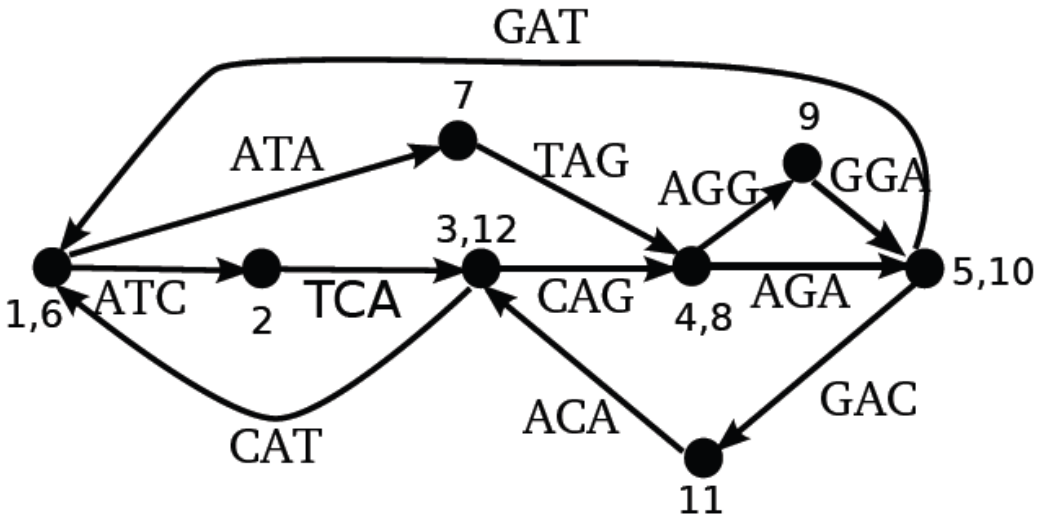
Принципиальное ограничение: повторы, длиннее расстояния вставки с помощью парой информации разрешить не получится

Парный граф де Брюйна

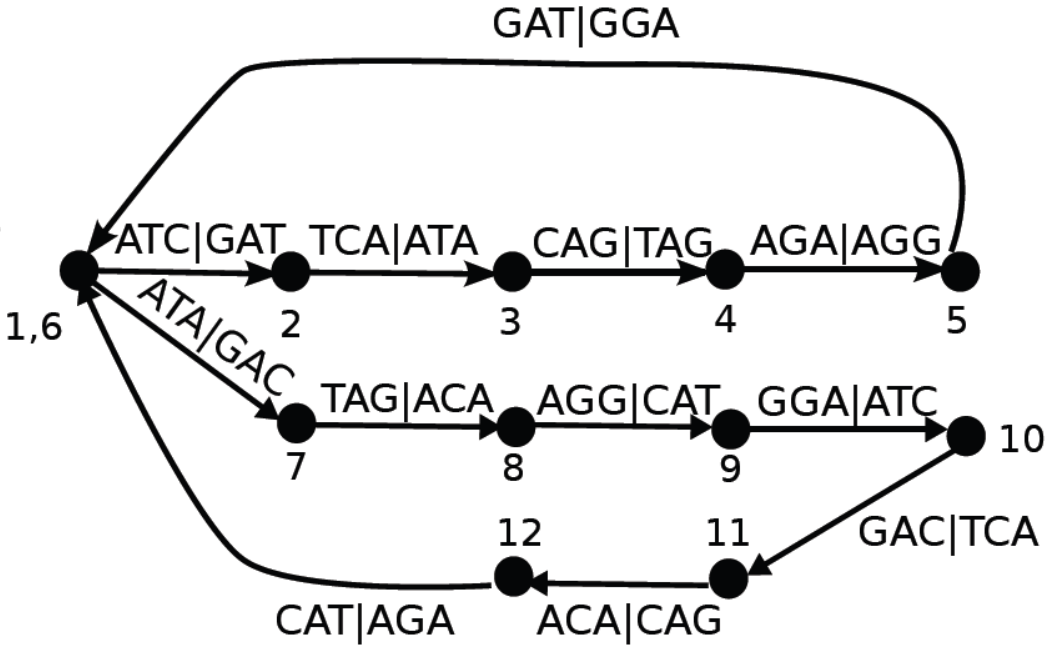
Парный граф де Брюйна

- Вершины парного графа де Брюйна: все пары *k*-меров на фиксированном расстоянии
- Рёбра парного графа де Брюйна: все пары *(k+1)*-меров на фиксированном расстоянии
- Ребро *e* соединяет пару префиксов *e* и пару суффиксов *e*

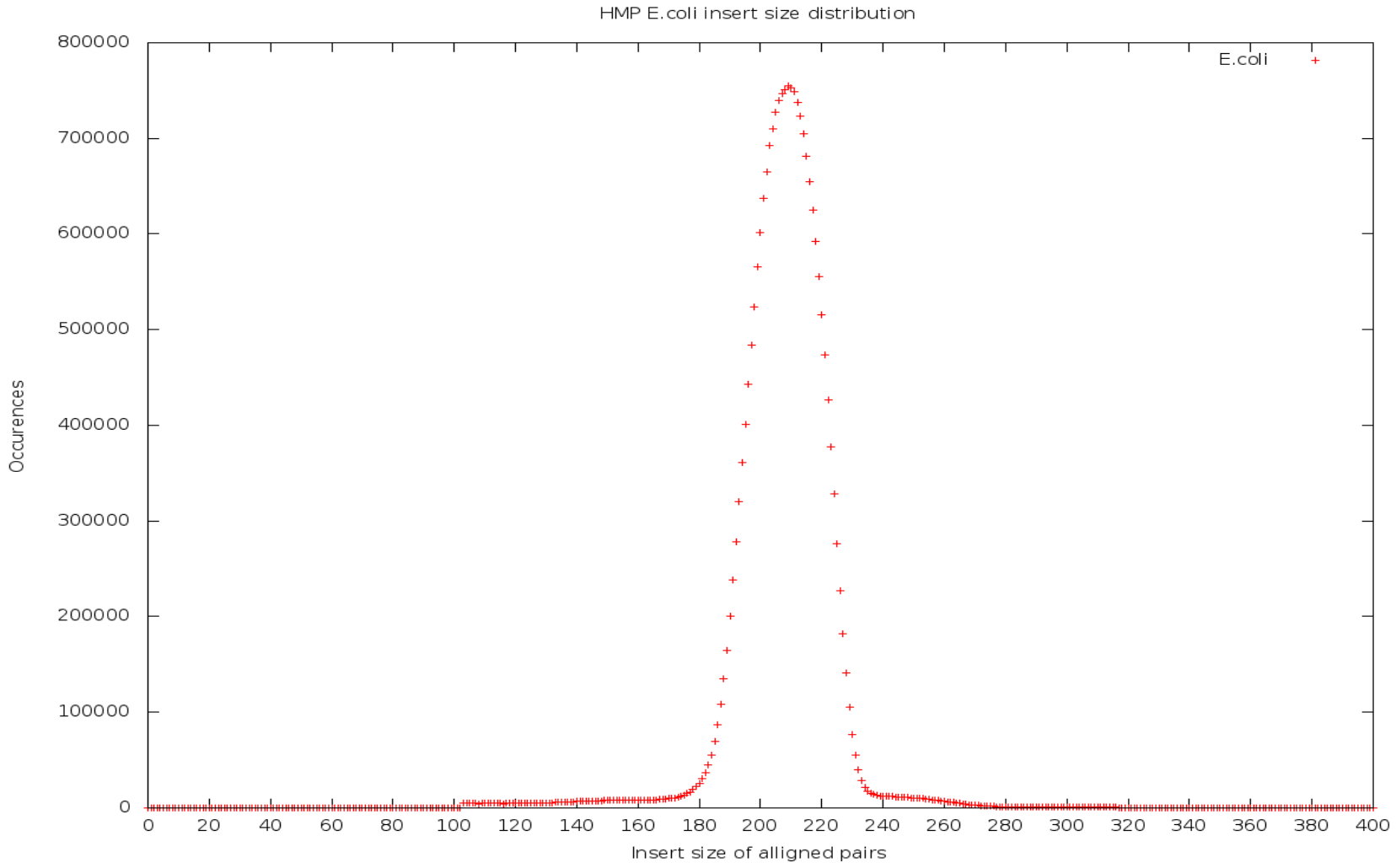
Граф де Брюйна



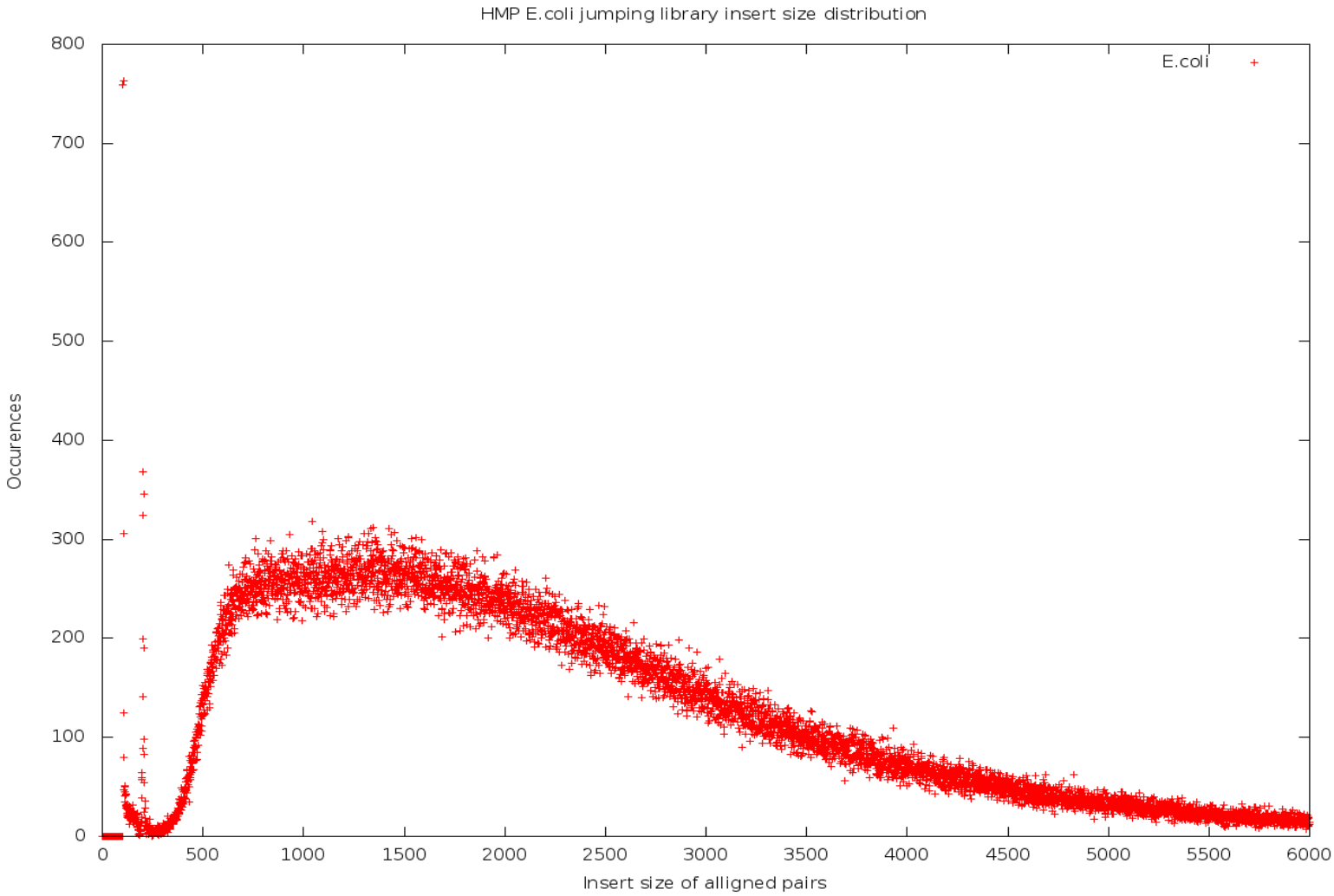
Парный граф де Брюйна



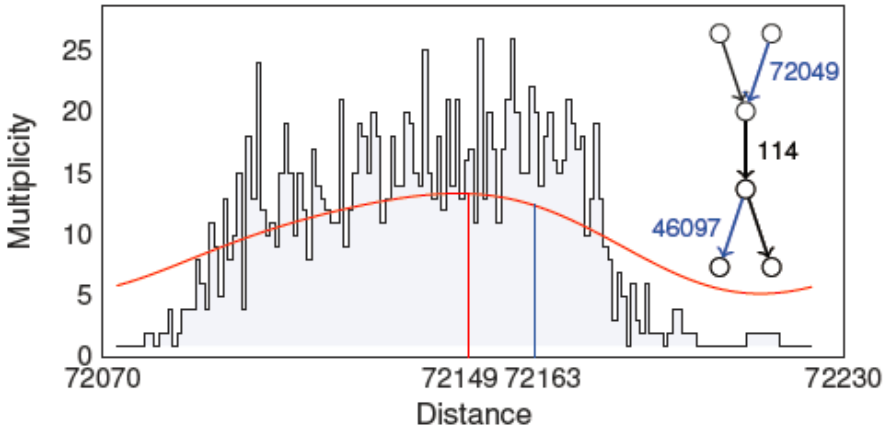
Разброс расстояния

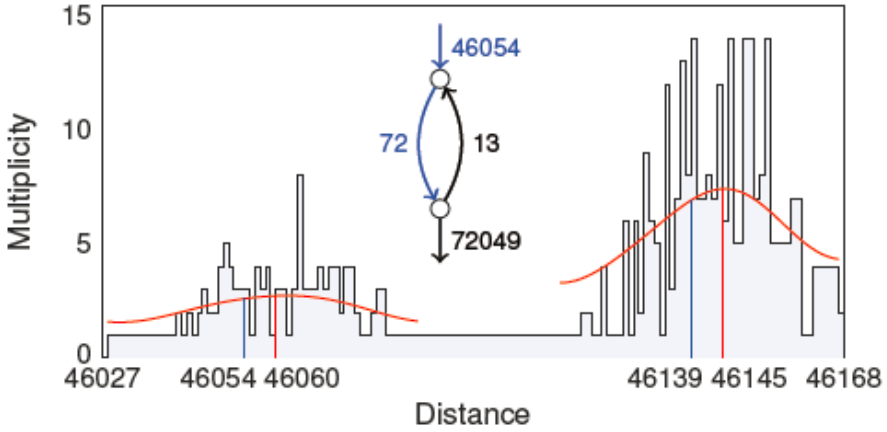


Разброс расстояния

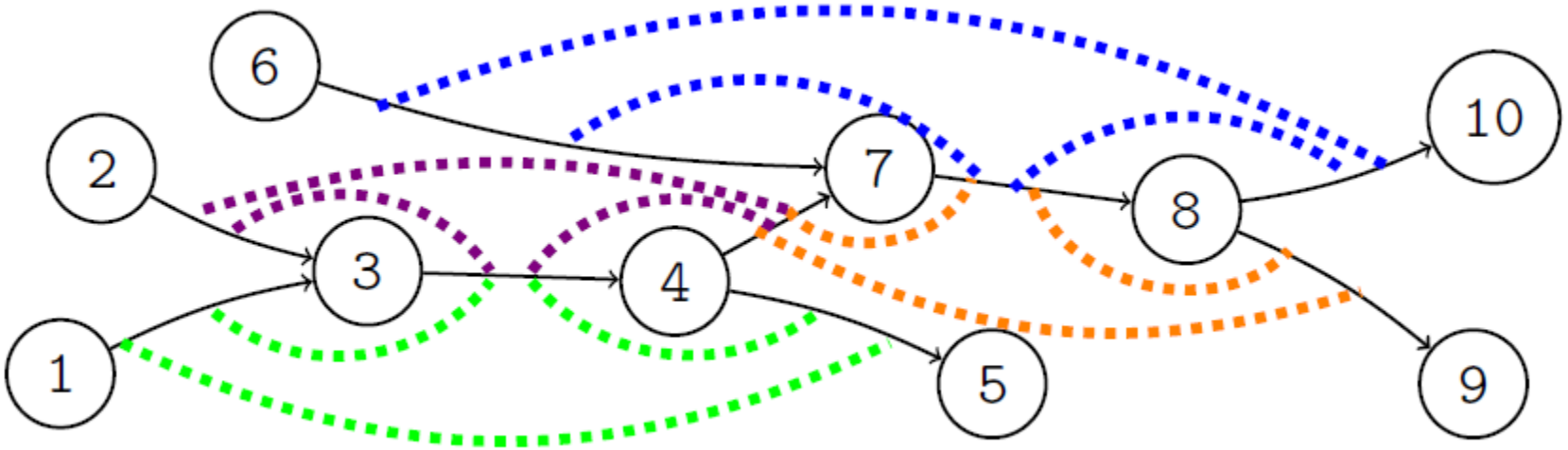


Уточнение расстояния между ребрами

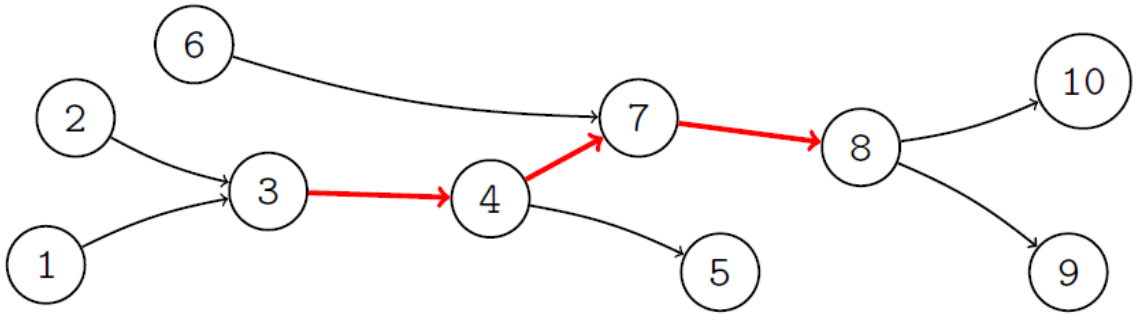


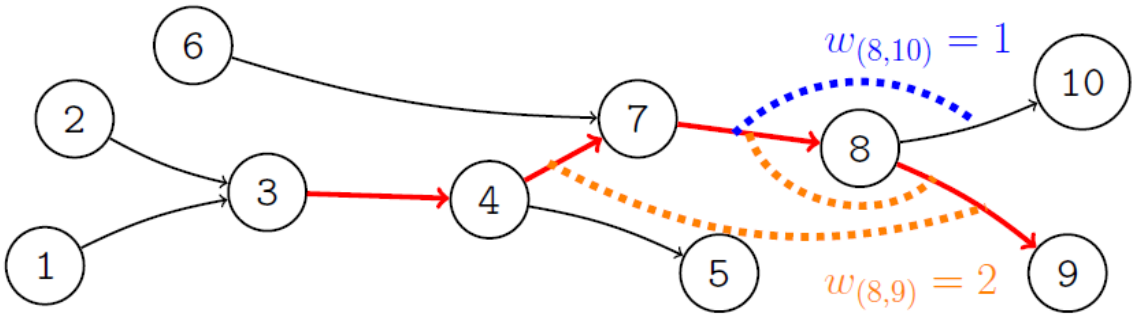


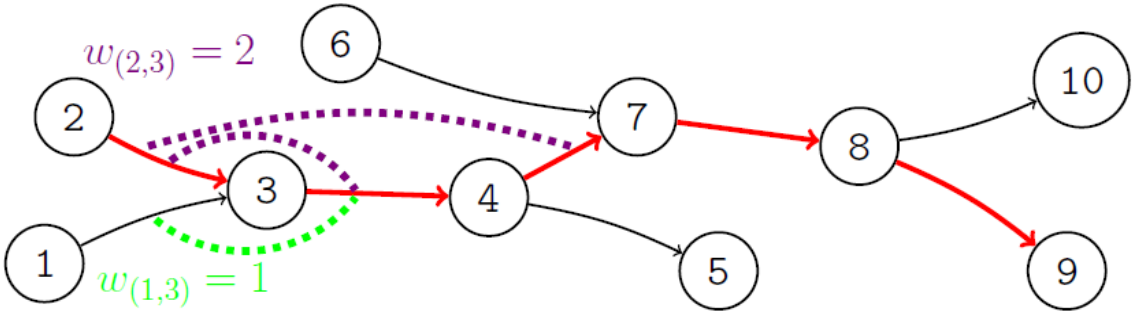
Извлечение контигов из графа



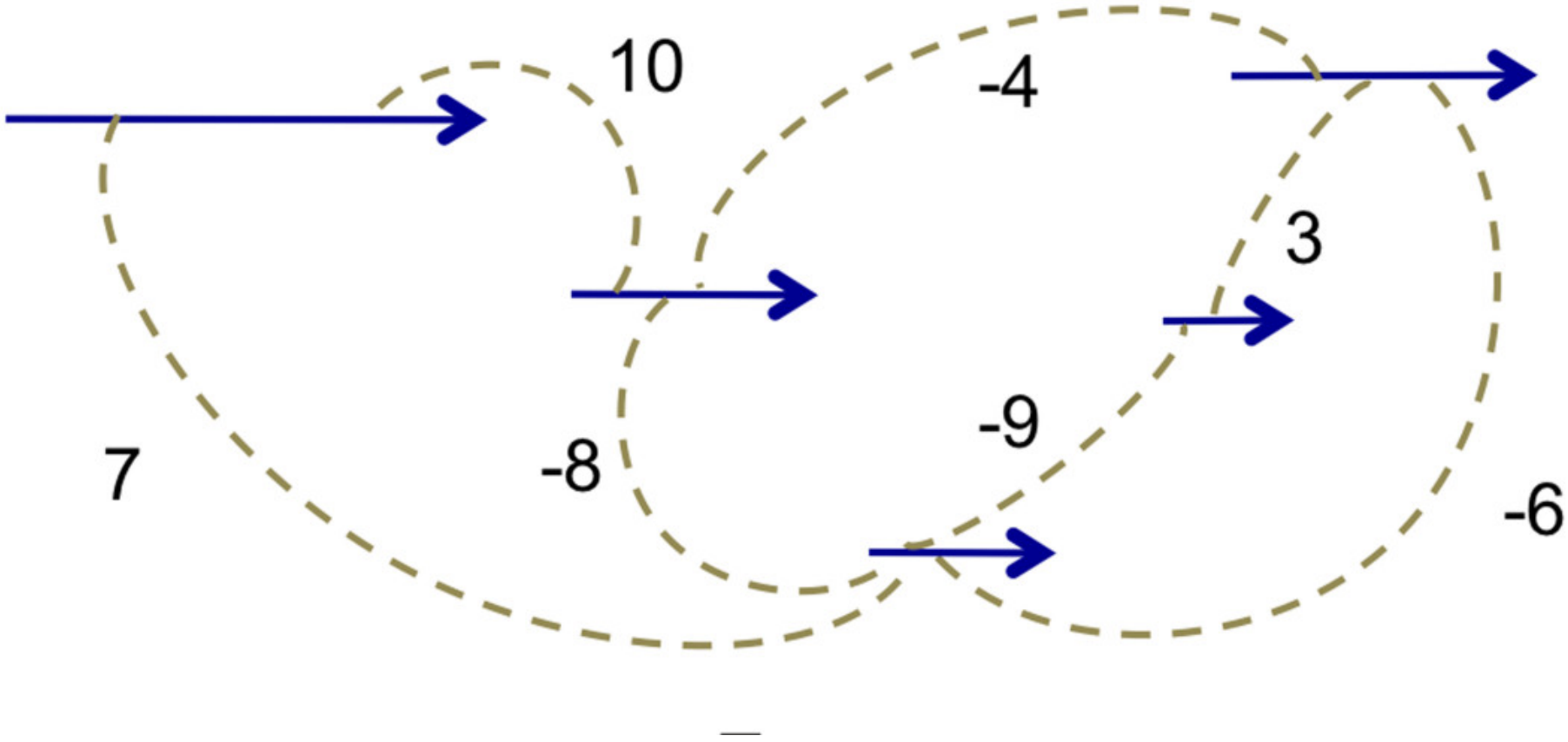
Извлечение контигов из графа







Скаффолдинг



Осталось за кадром

- Информация о покрытии
- Информация о путях ридов
- Использование информации нескольких библиотек
- Проблемы, связанные с использованием mate-pair библиотек

Альтернативные подходы

Недостатки графов де Брюина

Склеиваются даже k-меры из середин ридов (важно в случае длинных ридов)

Из-за ошибок может нарушаться связность в случае низкого покрытия

Жадная стратегия

Ассемблеры: phrap, TIGR, CAP3, SSAKE, VCAKE, SHARCGS

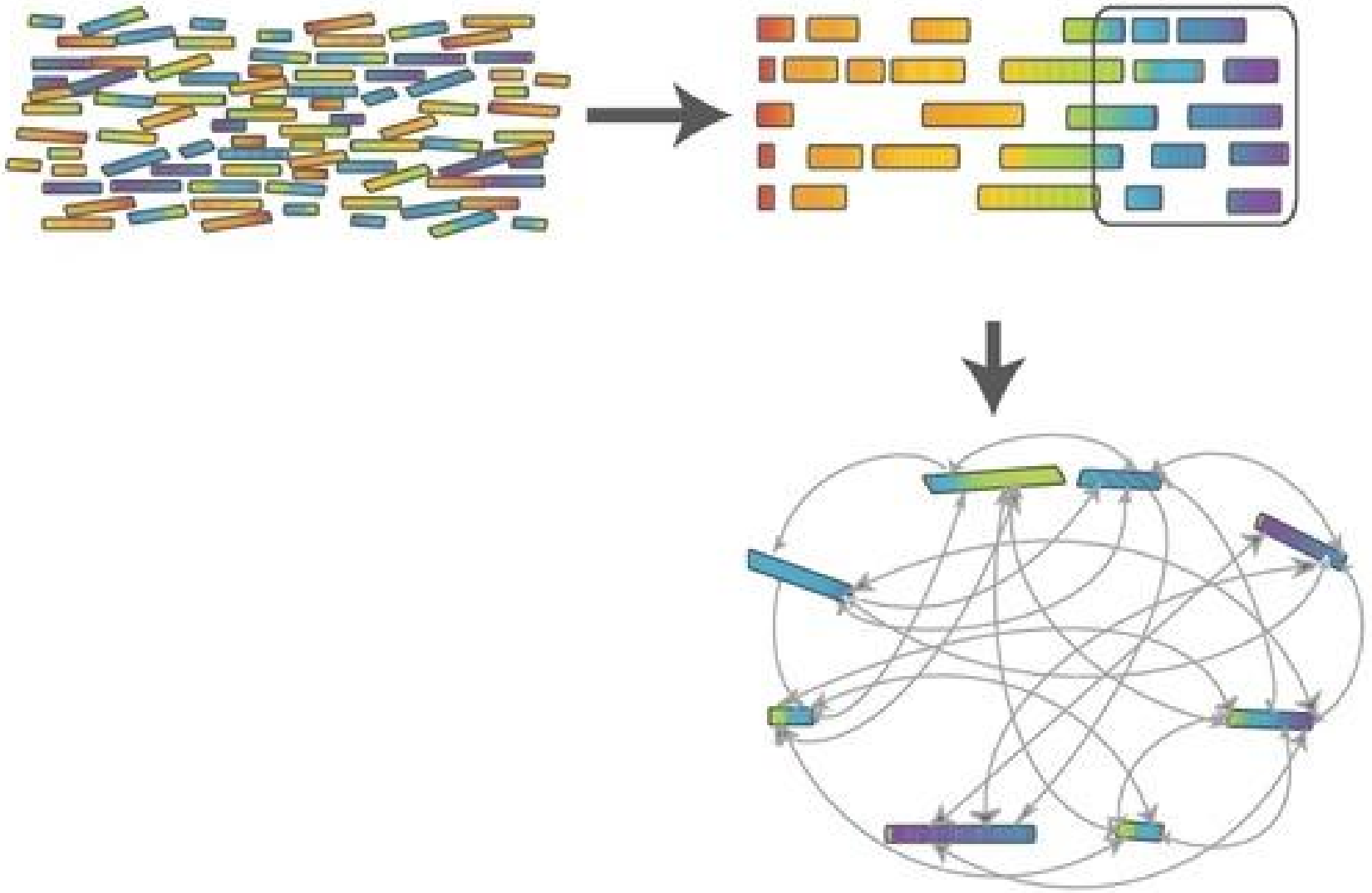
Напоминает эвристику для решения SSP:

- Пока возможно, объединяем риды с наилучшим перекрытием.

Проблема: повторы

Улучшение: остановка, если есть несколько различных продолжений с неплохим перекрытием

Граф перекрытий (OLC)

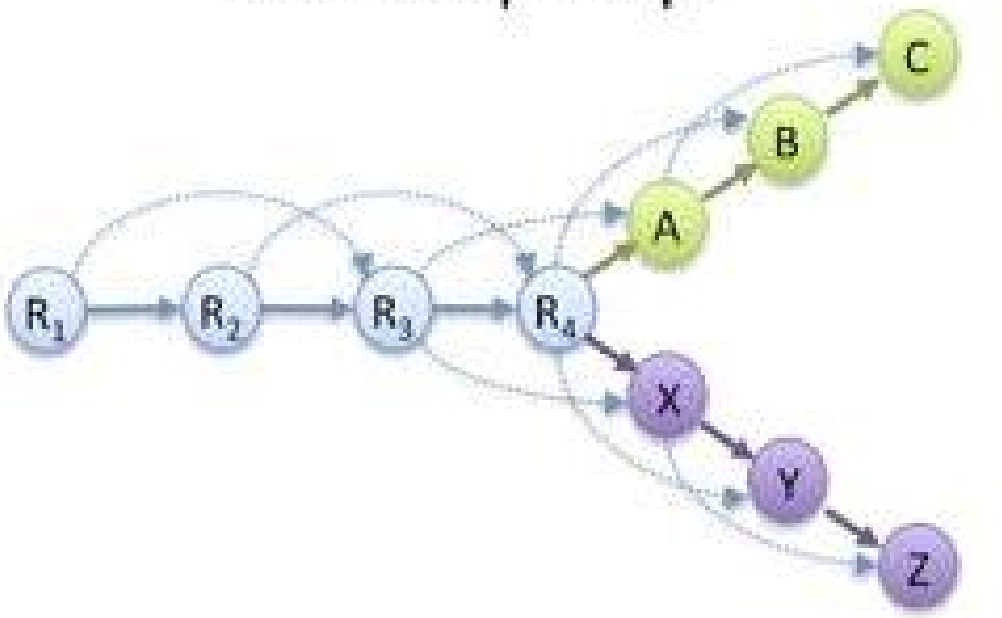


Транзитивные рёбра

A Read Layout

```
R1: GACCTACA
R2:    ACCTACAA
R3:     CCTACAAG
R4:      CTACAAGT
A:        TACAAGTT
B:         ACAAGTTA
C:          CAAGTTAG
X:        TACAAGTC
Y:         ACAAGTCC
Z:          CAAGTCCG
```

B Overlap Graph



Как искать перекрытия

Для каждой пары ридов найти перекрытие

- Квадратично от количества контигов
- Только точные перекрытия
- Можно применить поиск редакционного расстояния, но тогда алгоритм станет квадратичным и по длине рида.

Как искать перекрытия

Для каждой пары ридов найти перекрытие

- Квадратично от количества контигов
- Только точные перекрытия
- Можно применить поиск редакционного расстояния, но тогда алгоритм станет квадратичным и по длине рида.

Для оптимизации строится индекс k-меров

А можно лучше

Ассемблер SGA использует структуру FM-индекса, основанную на суффиксном массиве для поиска перекрытий.

Кроме поиска перекрытий SGA позволяет:

- Игнорировать транзитивные рёбра
- Не хранить саму структуру графа в памяти (поиск входящих/исходящих рёбер происходит на лету)

Гибридные сборки

Использование нескольких типов ридов вместе может улучшить сборку:

- Illumina риды: высокое покрытие, мало ошибок, НО короткие
- PacBio риды: очень длинные, НО вероятность ошибки 15%!!!

Коррекция PacBio

Стратегия: исправить ошибки в PacBio ридах при помощи Illumina ридов.

Реализации:

1. Приложить Illumina к PacBio
2. Приложить PacBio к Illumina

Финишинг

Результат работы сборщика: черновая сборка (draft assembly)

Чтобы получить полный геном проводятся дополнительные эксперименты

Ссылки

1. "External Perfect Hashing for Very Large Key Sets", Fabiano C. Botelho, Nivio Ziviani
2. *"De novo* assembly and genotyping of variants using colored de Bruijn graphs", Z.Iqubal et al.
3. "Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers", P.Medvedev et al.
4. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing", A.Bankevich et al.
5. SOPRA: Scaffolding algorithm for paired reads via statistical optimization, A.Dayarian et al.
6. "Computability of Models for Sequence Assembly", P.Medvedev et al.
7. "Efficient de novo assembly of large genomes using compressed data structures", J.Simpson, R.Durbin
8. "Hybrid error correction and de novo assembly of single-molecule sequencing reads", S.Koren et al.
9. http://bioinf.spbau.ru/en/spades

Вопросы
???