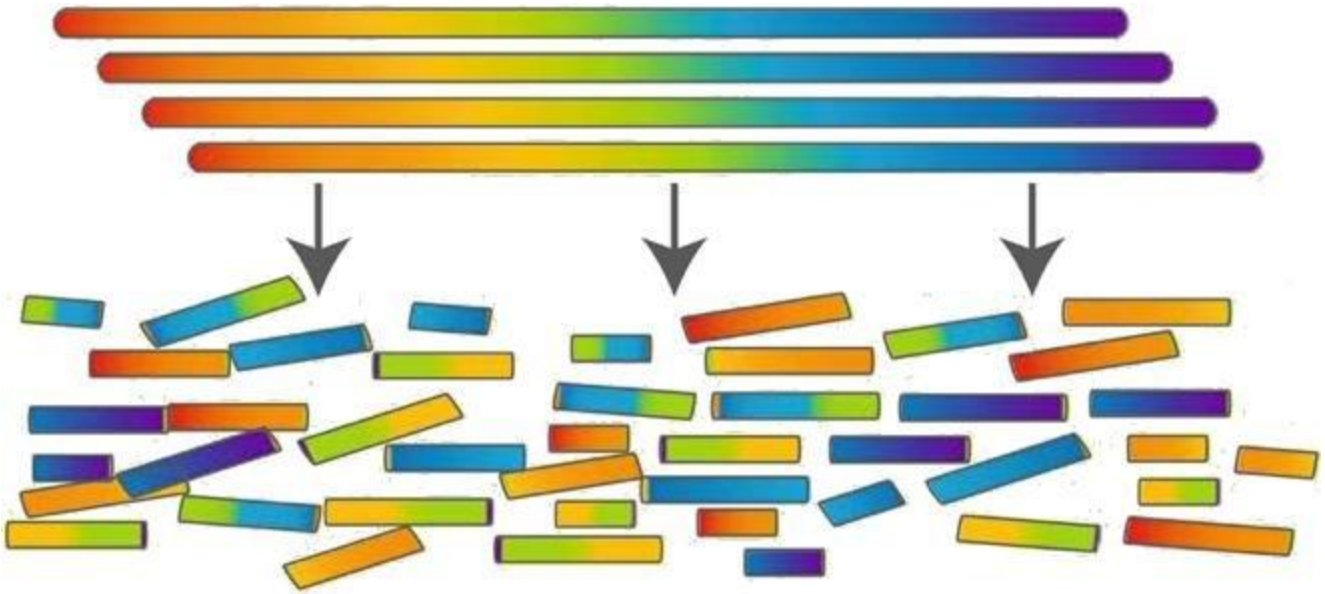
Немного о сборке генома

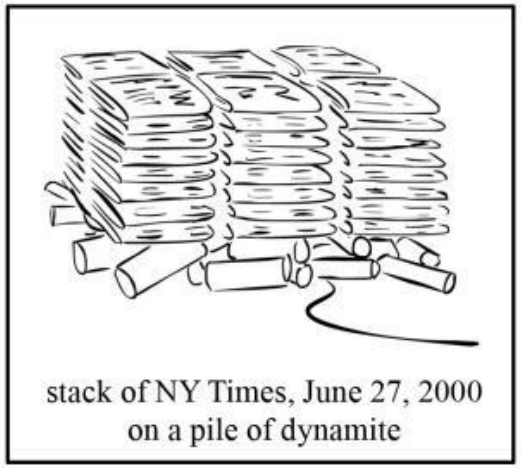
Антон Банкевич
Сергей Нурк
Лаборатория вычислительной биологии
АУ РАН
http://bioinf.spbau.ru

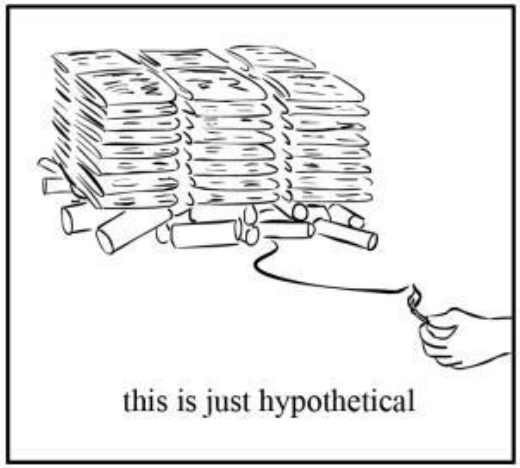
Введение

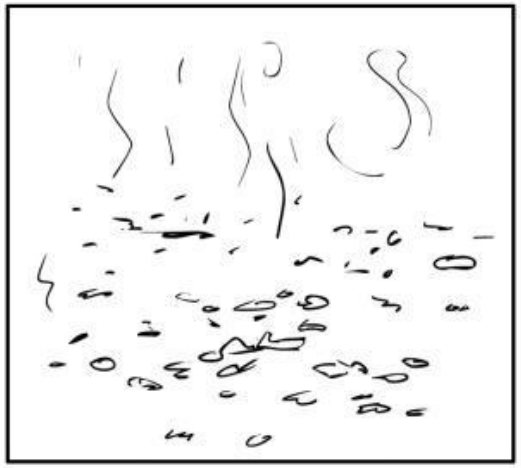
Секвенирование ДНК

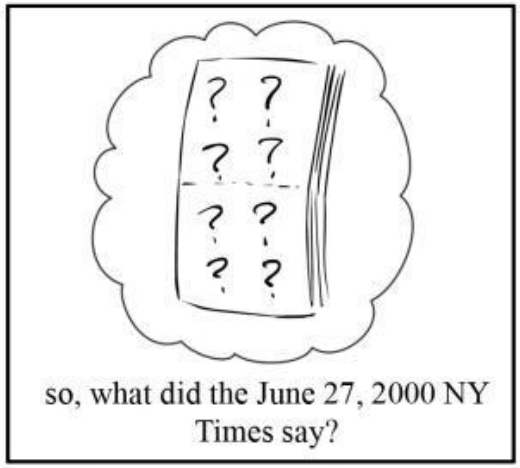


Секвенирование ДНК









Задача сборки

Получить последовательности нуклеотидов (контиги), которые:

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

Как написать ассемблер за выходные

Граф де Брёйна

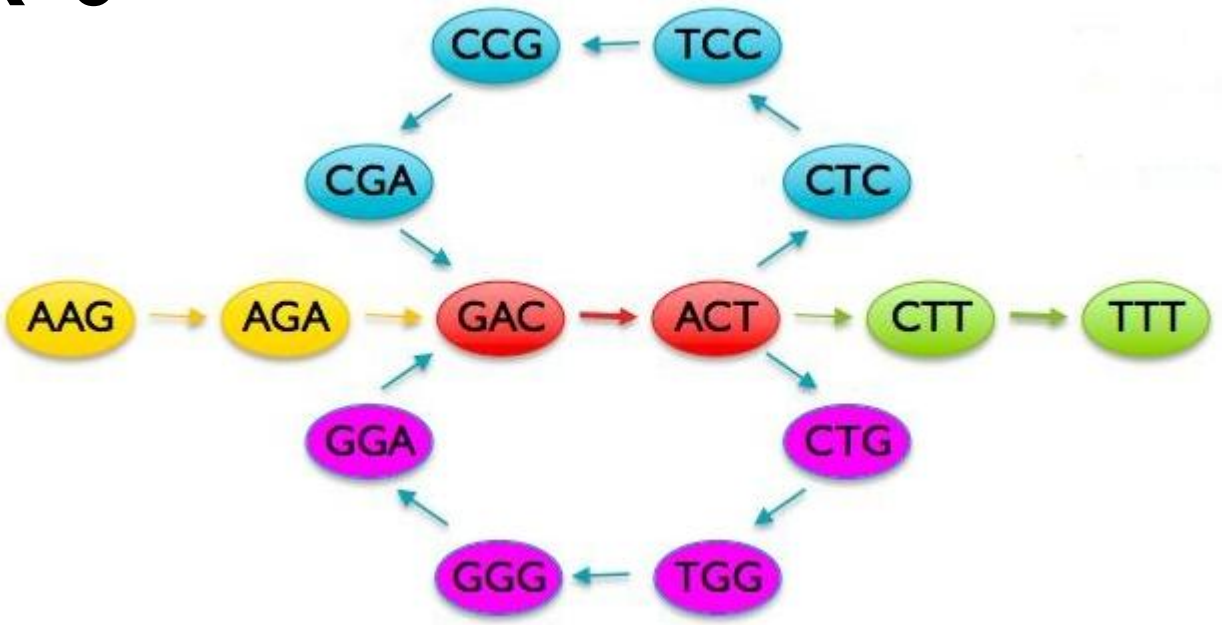
Граф де Брёйна

- k-мер: последовательность из *k* нуклеотидов
- Вершины графа де Брёйна: все *k*-меры
- Рёбра графа де Брёйна: все *(k+1)*-меры
- Ребро *e* соединяет префикс и суффикс *e*

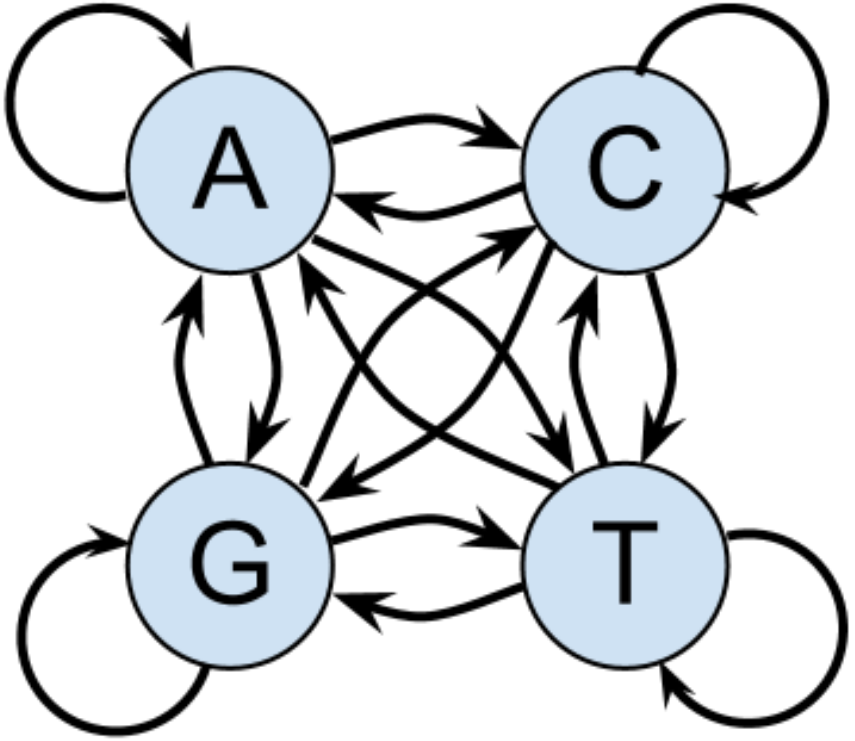
Граф де Брёйна

K=3

AAGACTC
GACTCCG
TCCGACT
GACTGGG
TGGGACT
GGACTTT



K имеет значение!

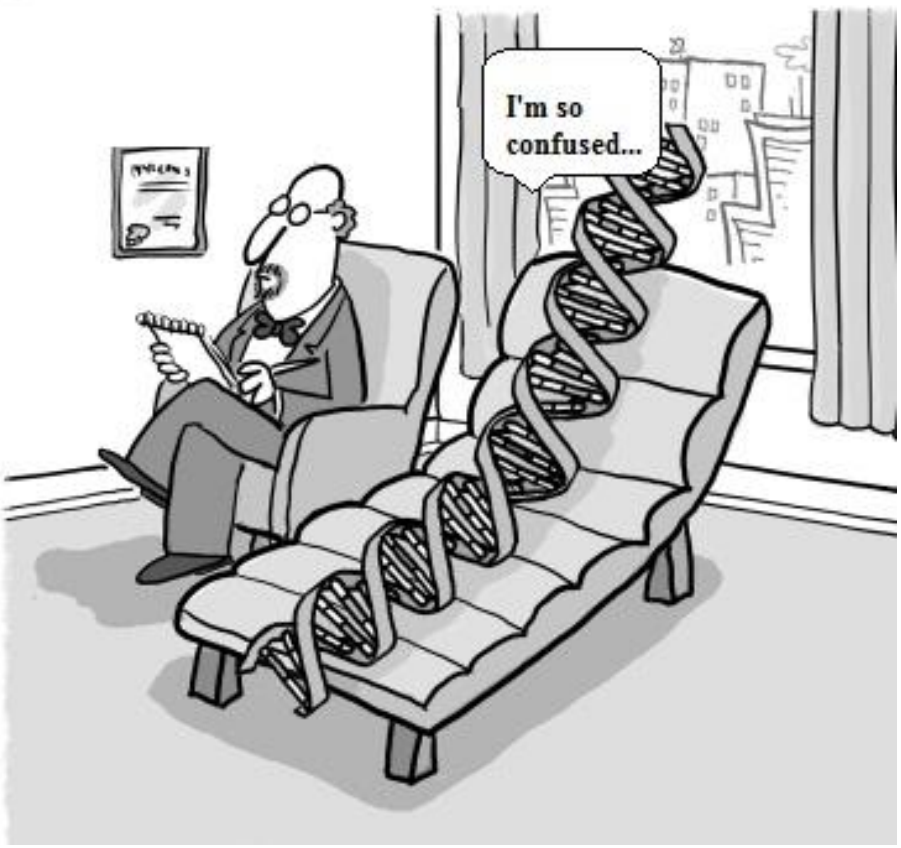


Проблема повторов

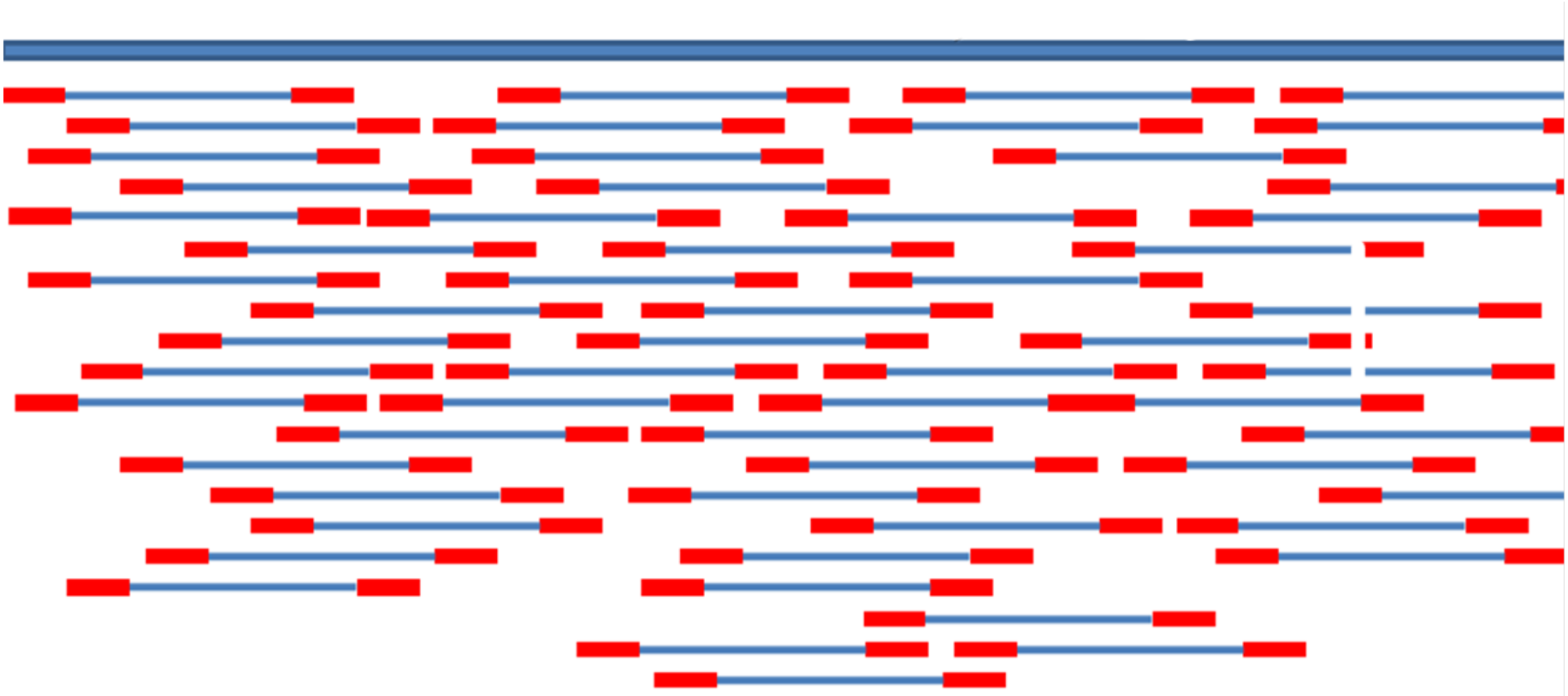
ALU

длина: 300

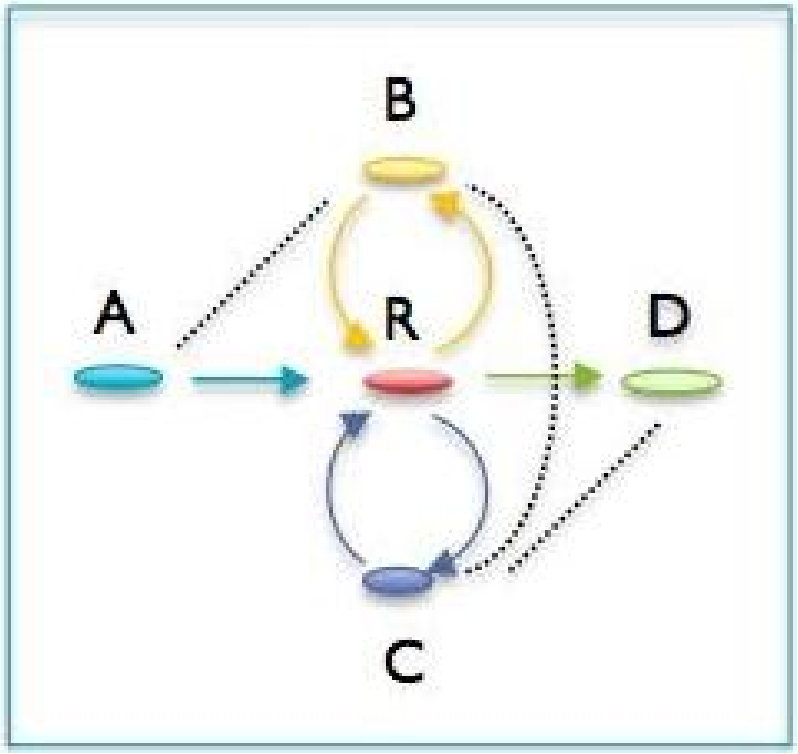
кратность: 1000000

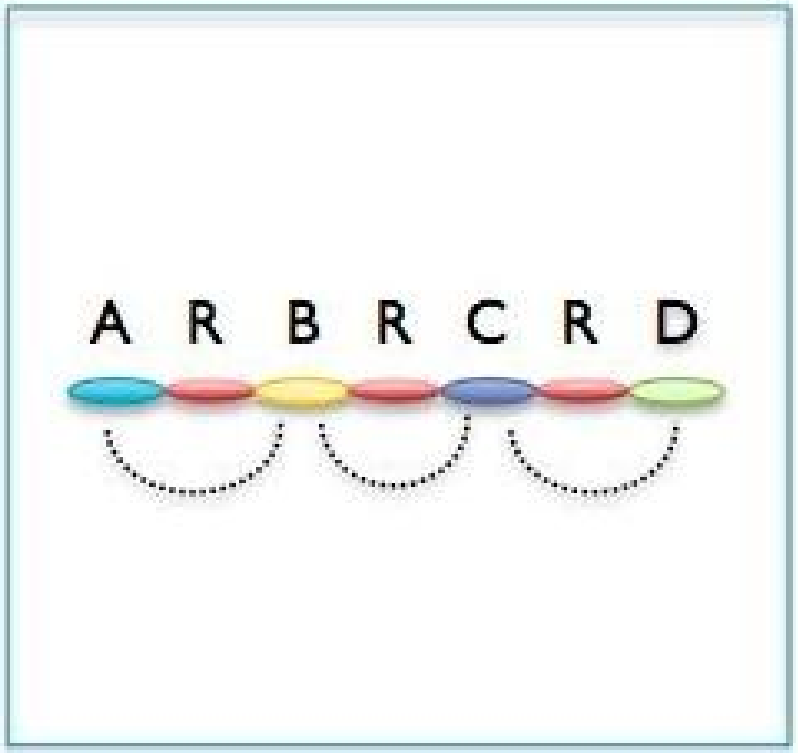


Парные риды



Разрешение повторов

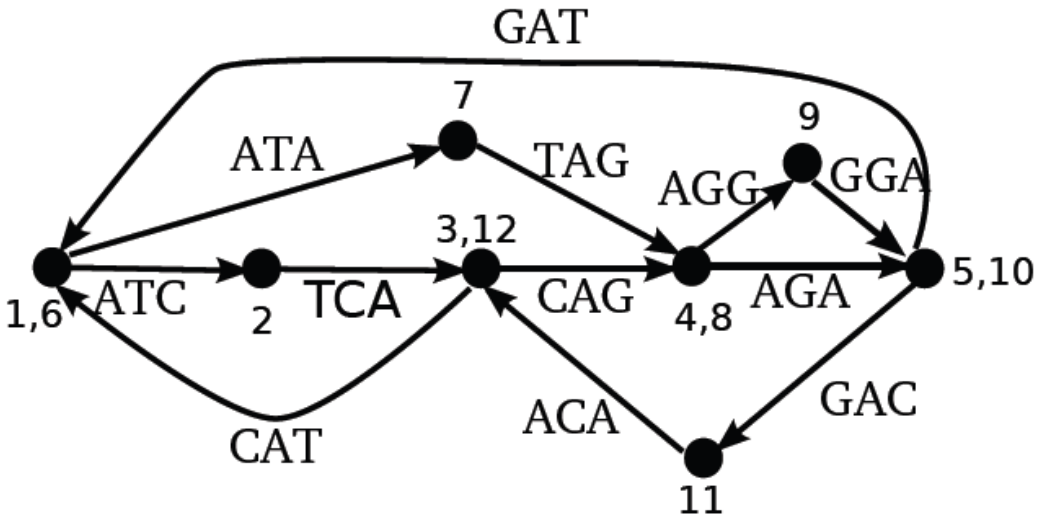




Парный граф де Брёйна

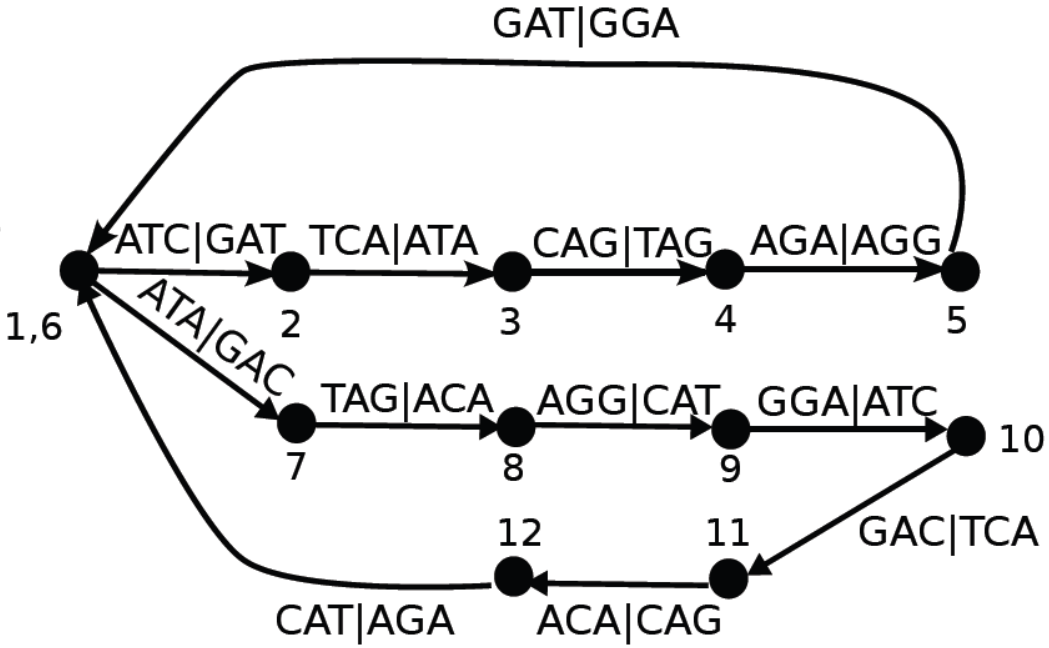
Парный граф де Брёйна

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



Граф де Брюина

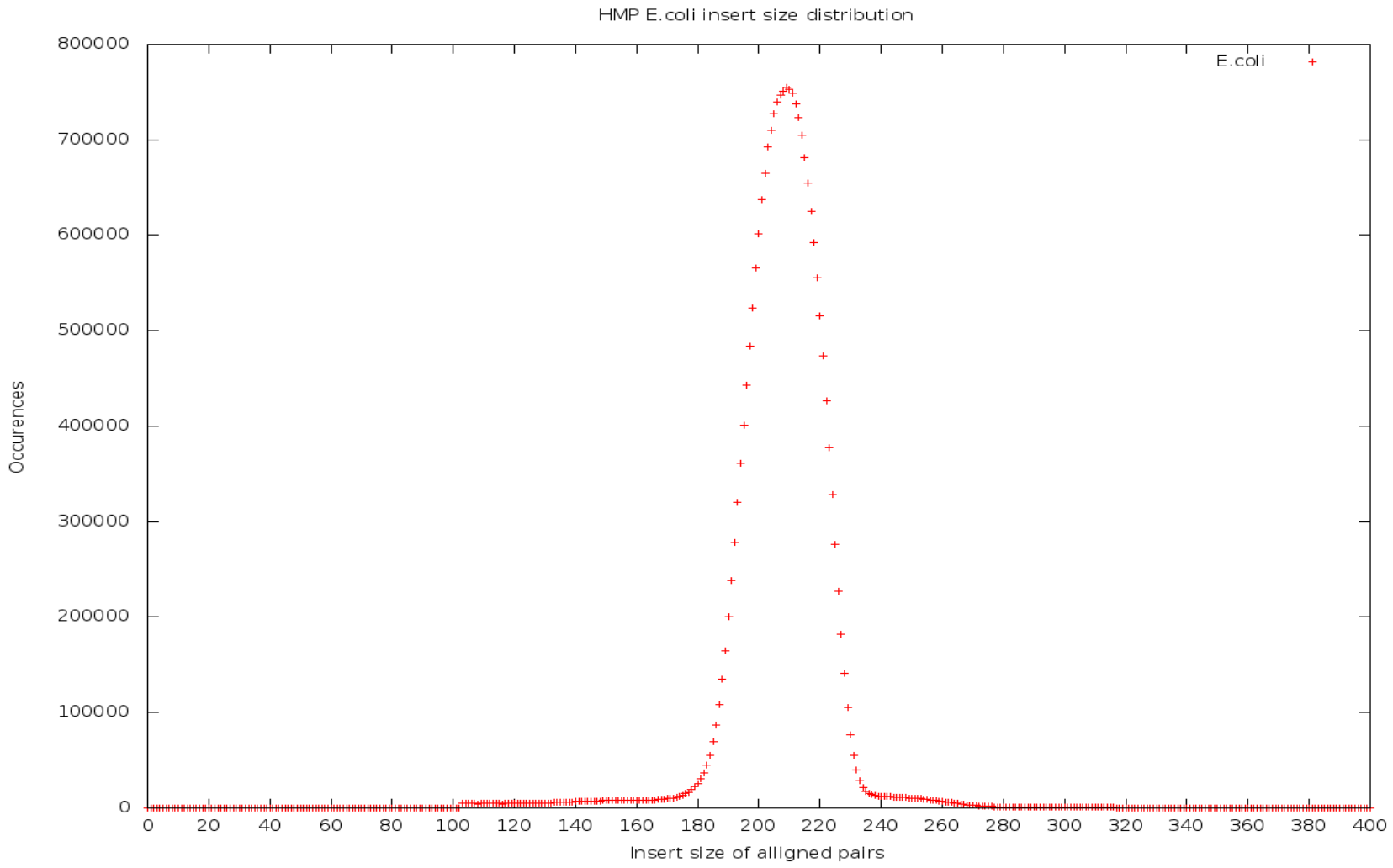
Парный граф де Брюина



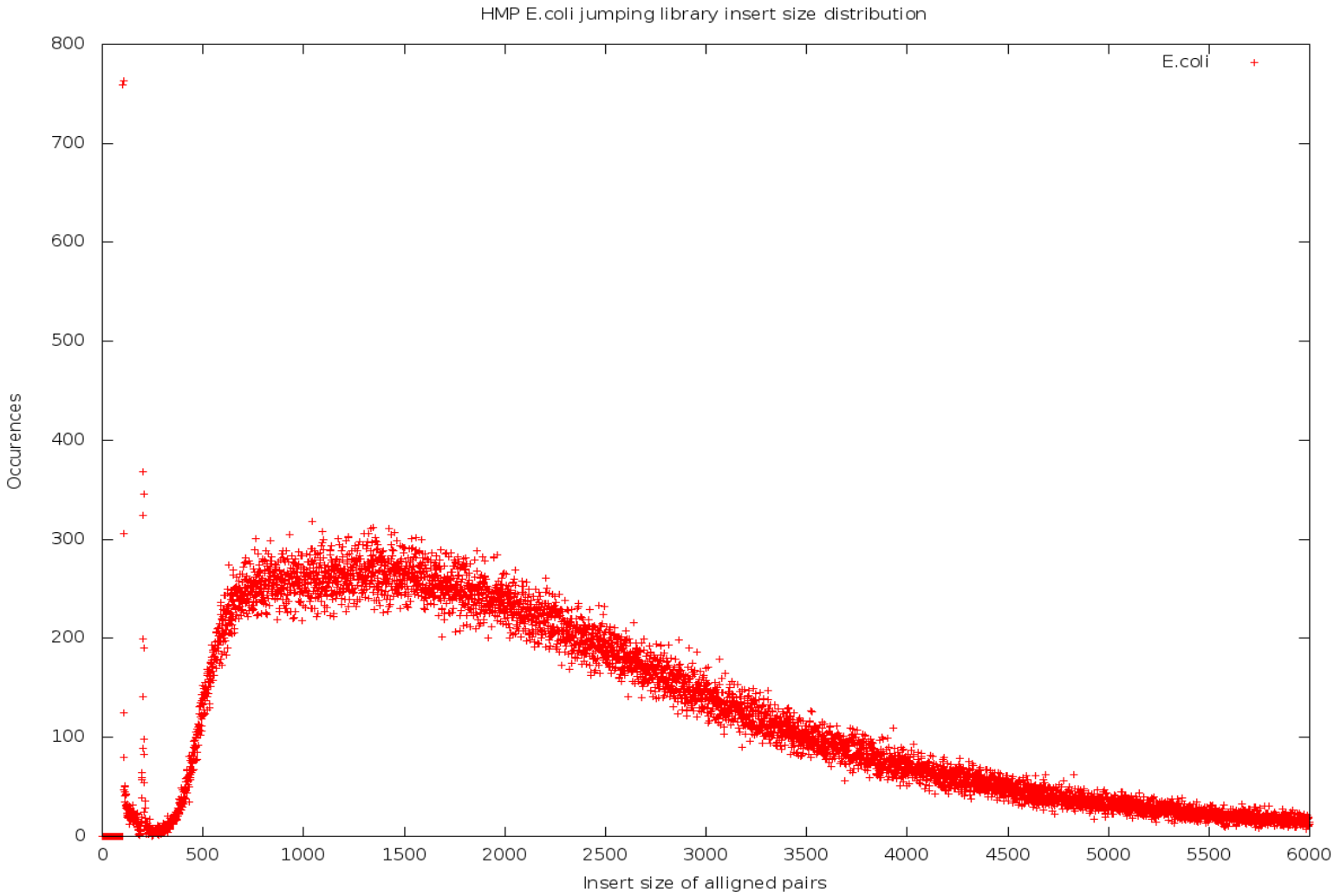
Некоторые проблемы

- Разброс расстояния
- Разрывы в покрытии
- Ошибки секвенирования
- Проблемы с ресурсами
  - память
  - время

Разброс расстояния



Разброс расстояния



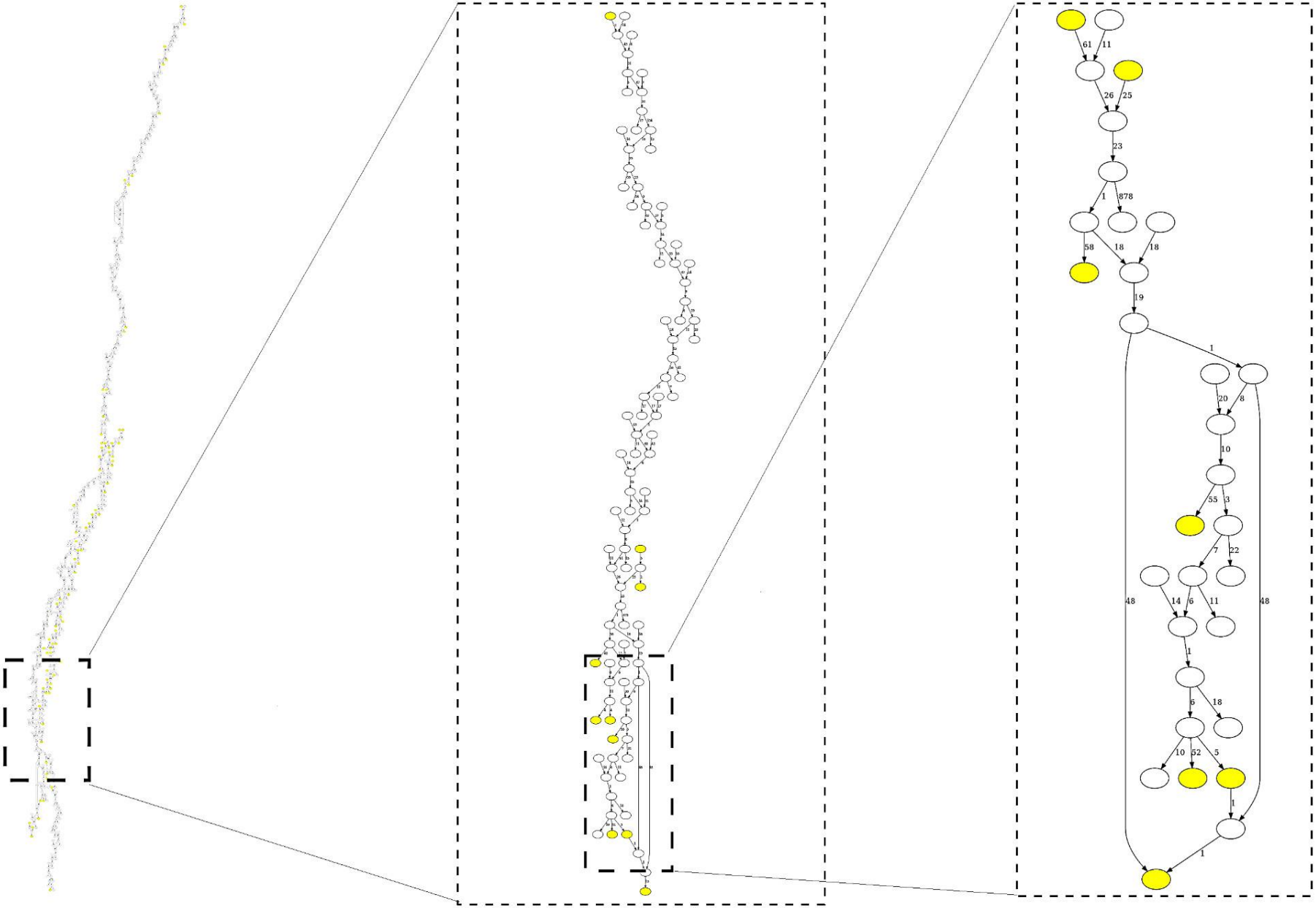
Разрывы в покрытии

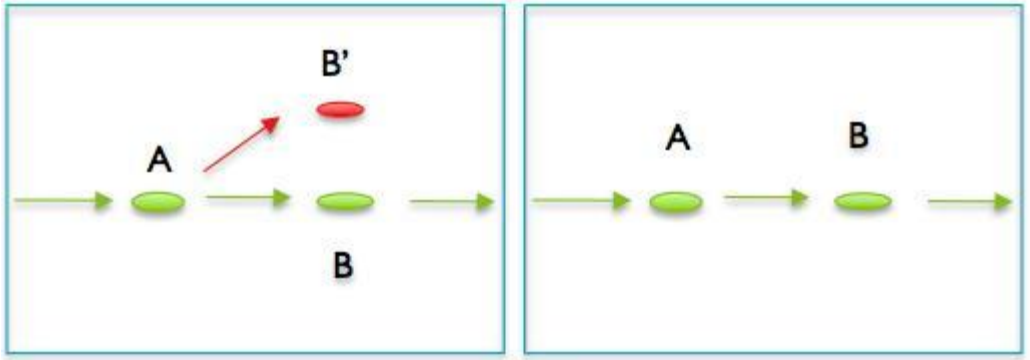
Покрытие конкретого *(k+1)*-мера — случайная величина.

Обычно приходится использовать *k* значительно меньше 100.

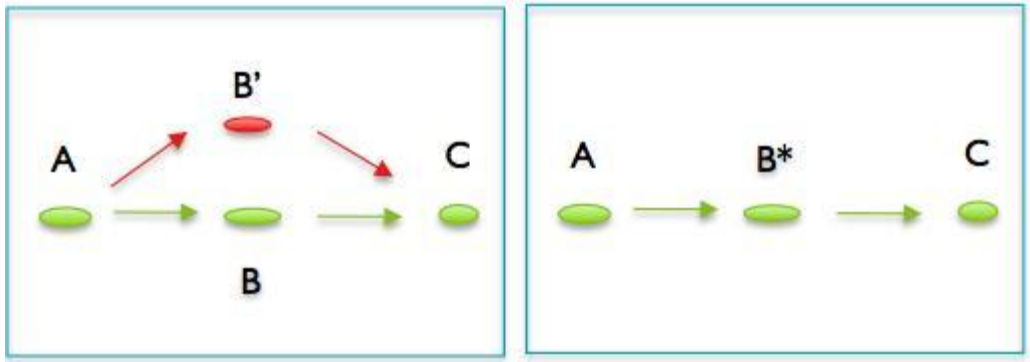
Ошибки секвенирования

- Тип и частота зависят от технологий
- Предобработка ридов: Quake, BayesHammer
- Неисправленные ошибки превращаются в лишние ребра в графе

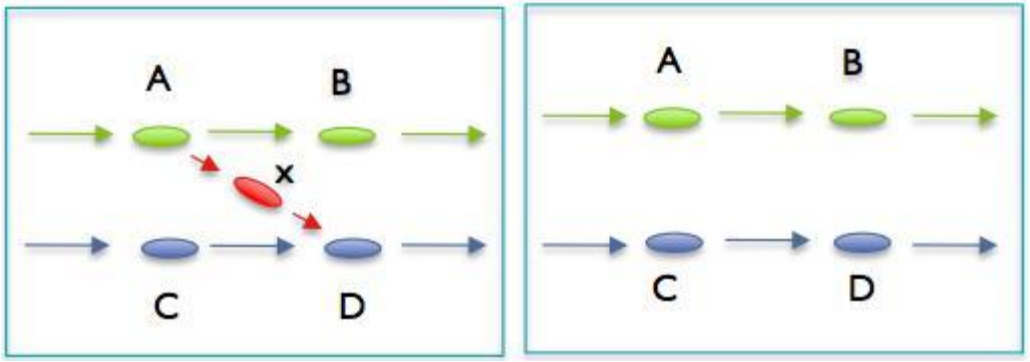

61
11
26
25
23
1
878
58
18
18
19
1
20
8
10
55
3
7
22
48
14
6
11
48
1
6
18
10
52
5
1
1



tip



bulge



chimeric connection

Можно потратить больше времени...

- Velvet
- IDBA
- SOAP-denovo
- Ray
- ABySS
- Allpaths
- EULER
- Minia

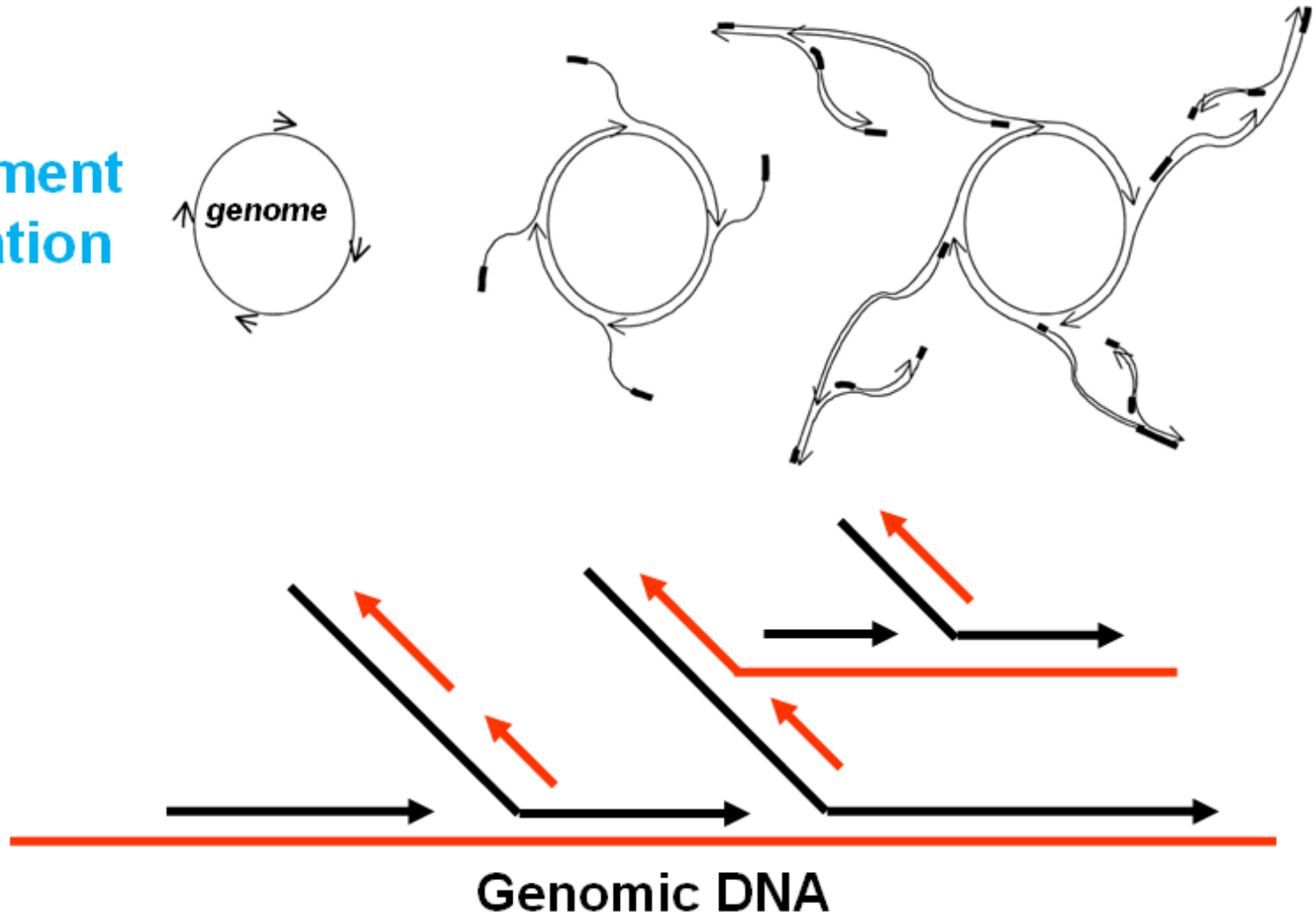
SPAdes

Single-cell секвенирование

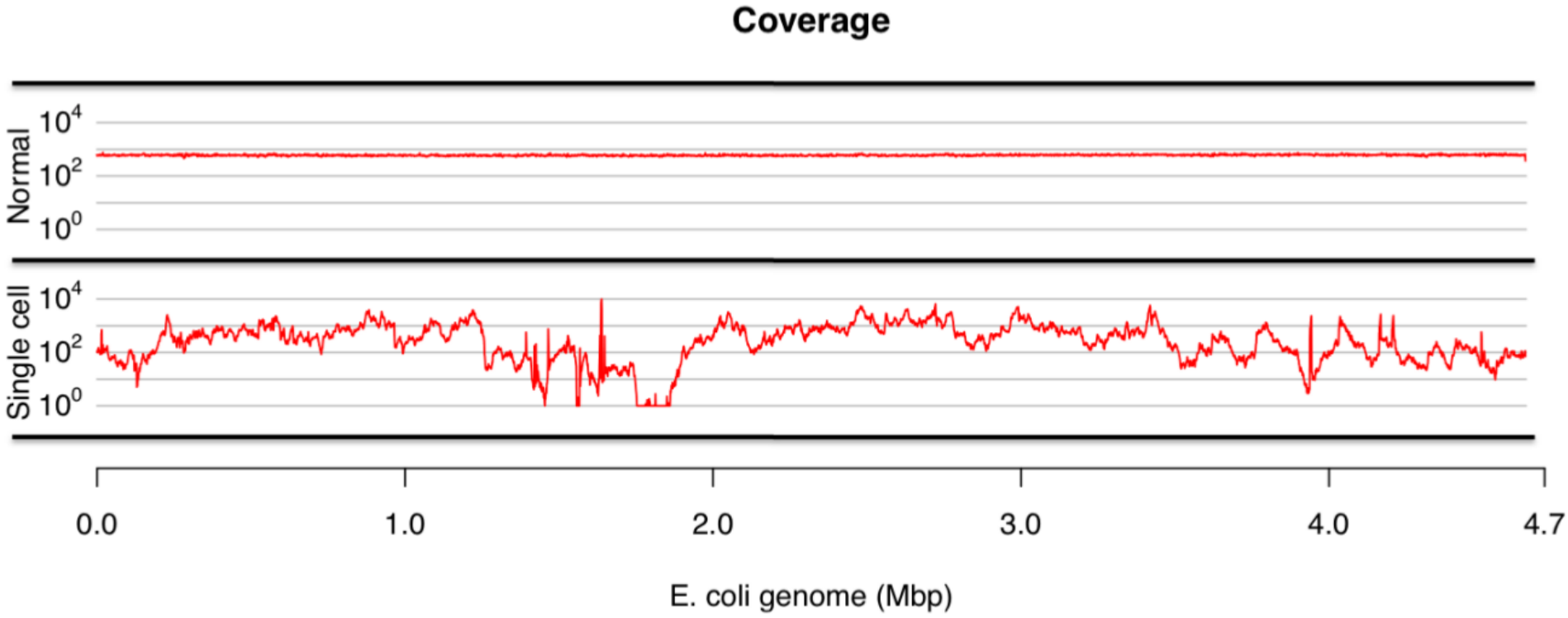
- Для секвенирования бактерии необходимо иметь значительное количество её клонов
- Большинство бактерий невозможно клонировать в лабораторных условиях
- Single-cell секвенирование позволяет увеличить количество ДНК не прибегая к клонированию

MDA

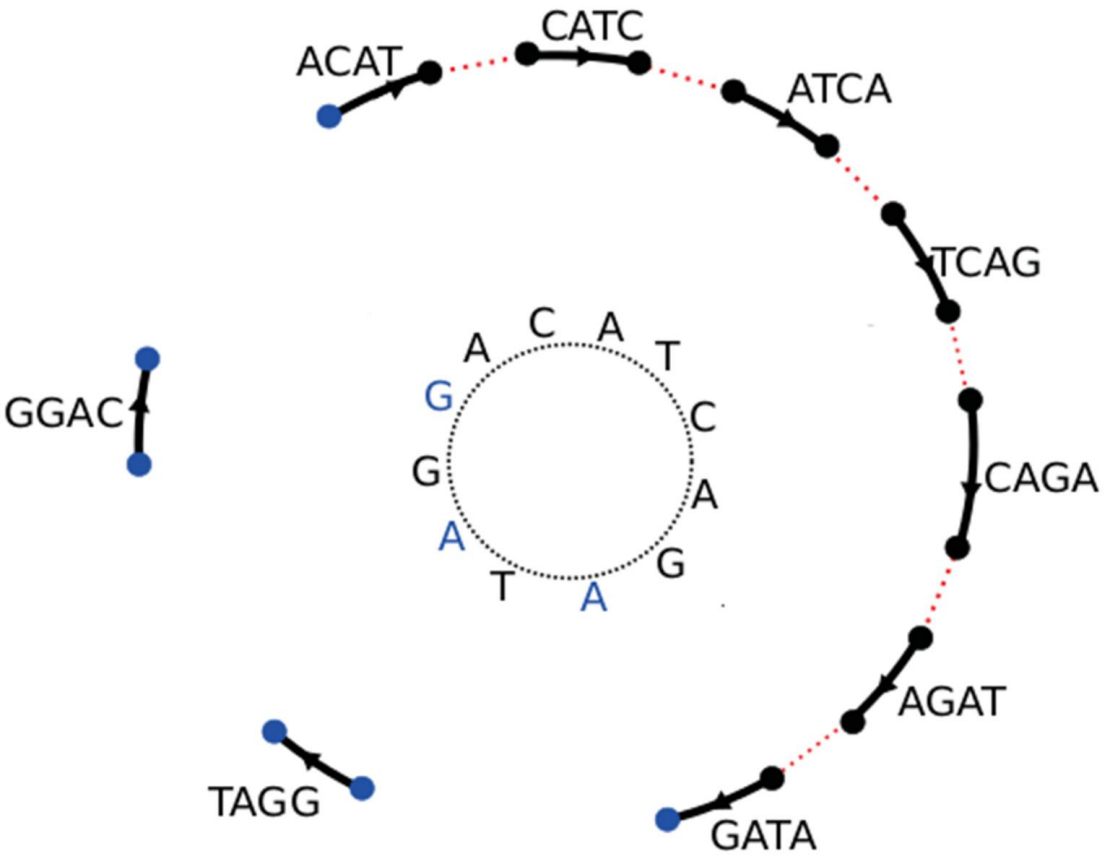
Multiple Displacement Amplification (MDA)

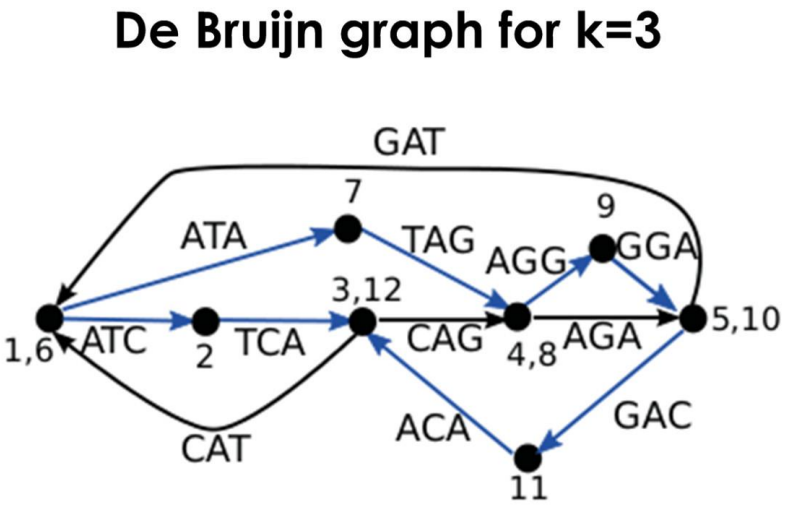


Покрытие генома ридами

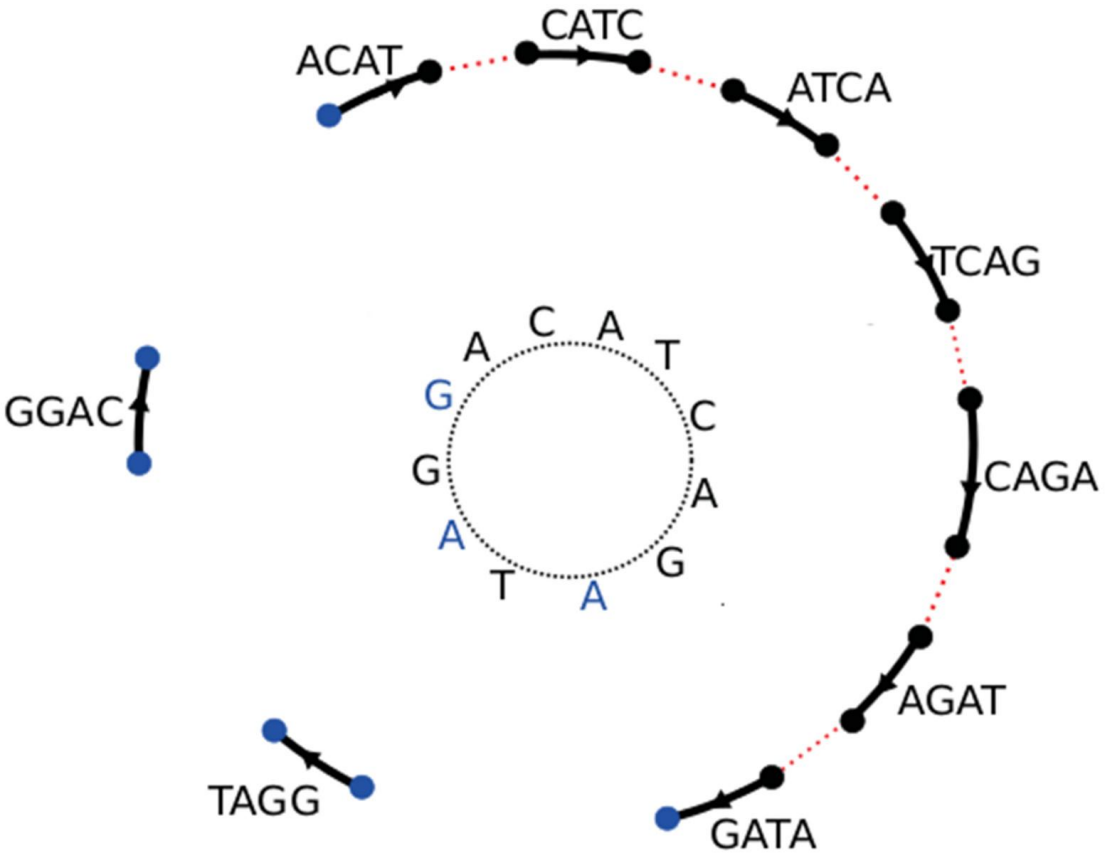


Борьба с разрывами

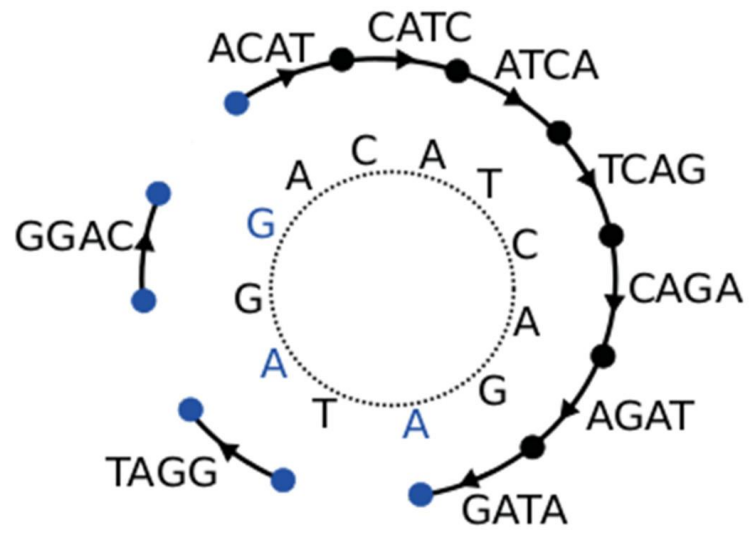


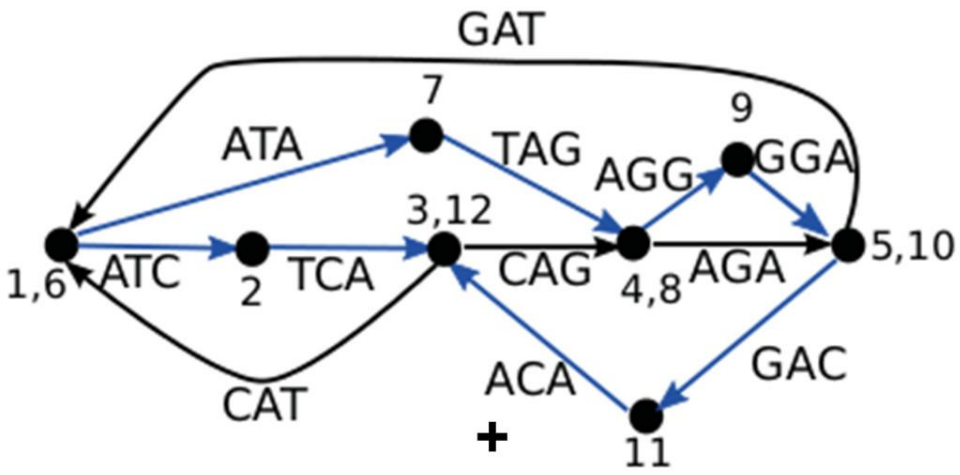


Борьба с разрывами

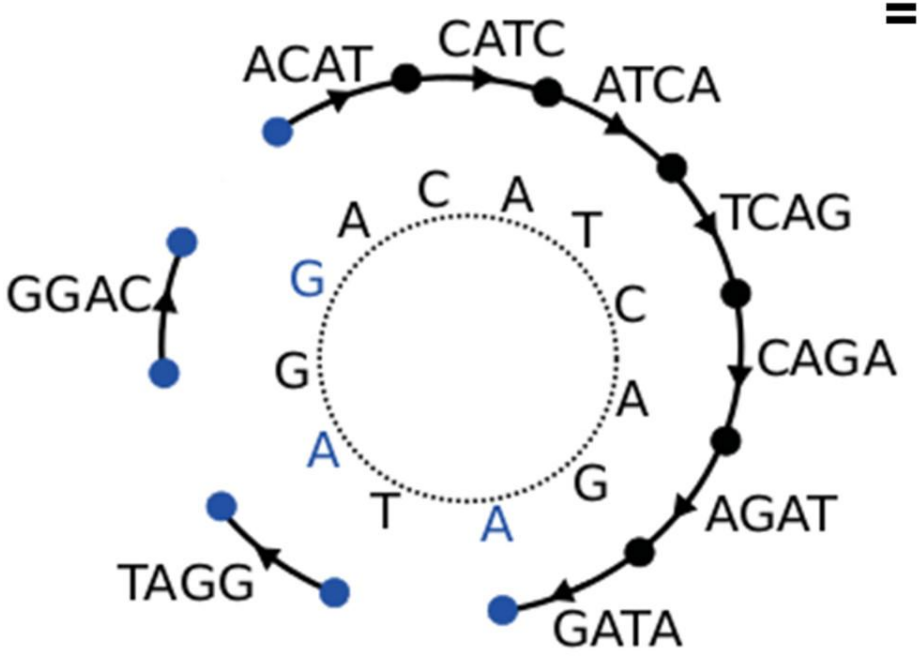


De Bruijn graph for k=4

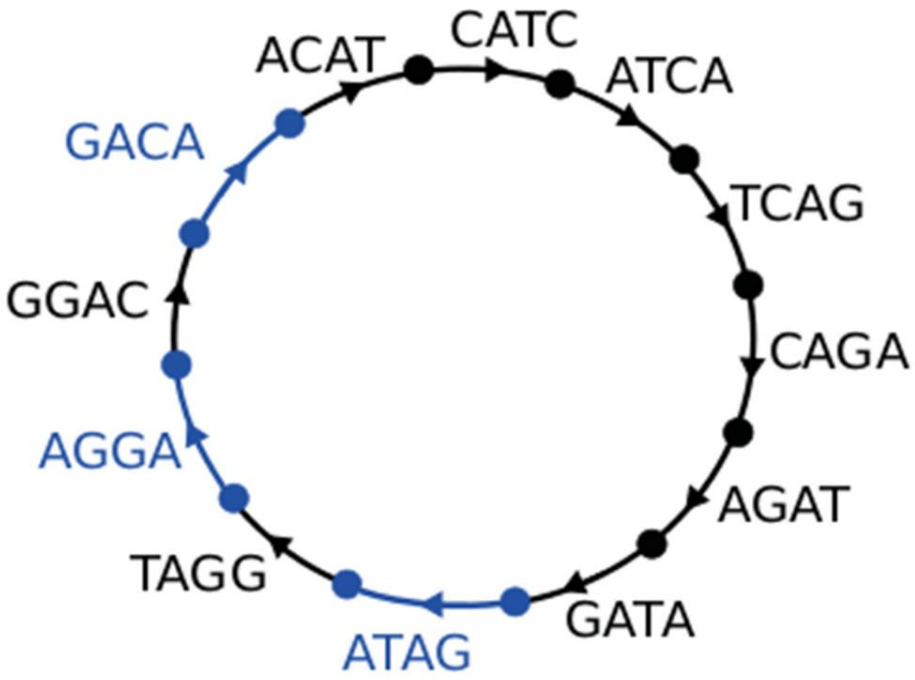




+

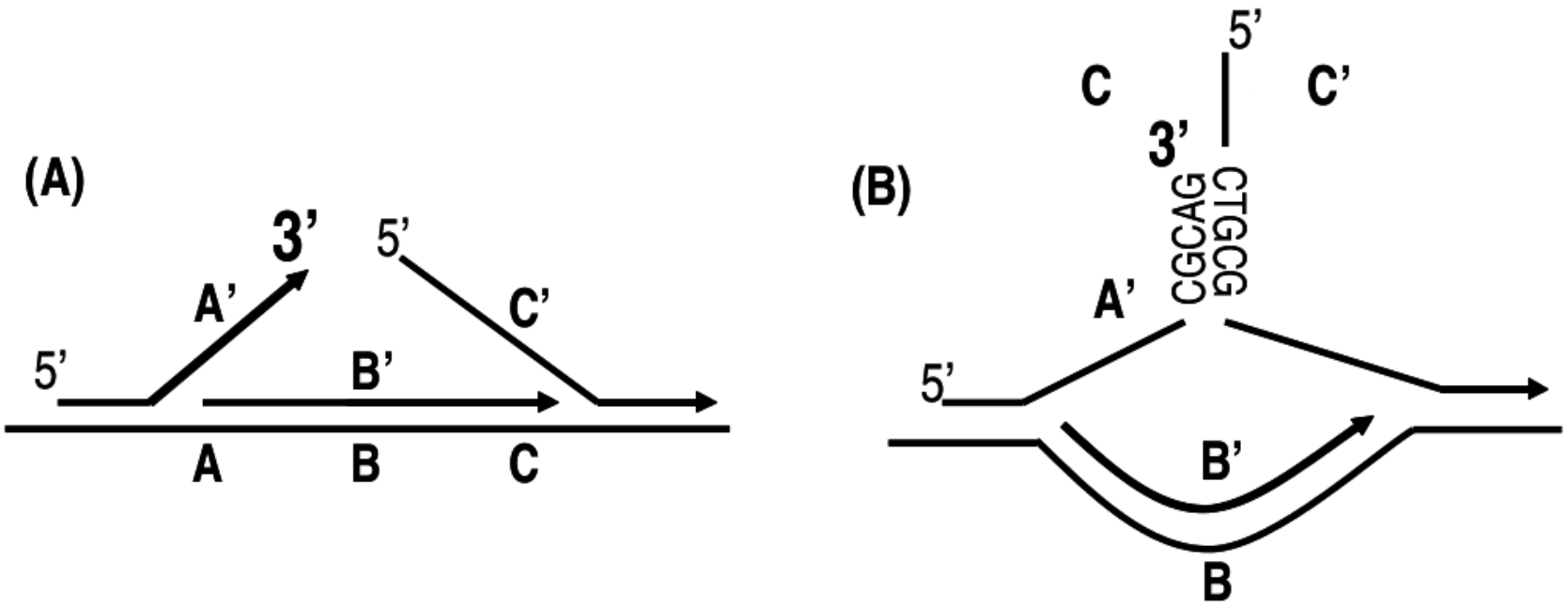


=

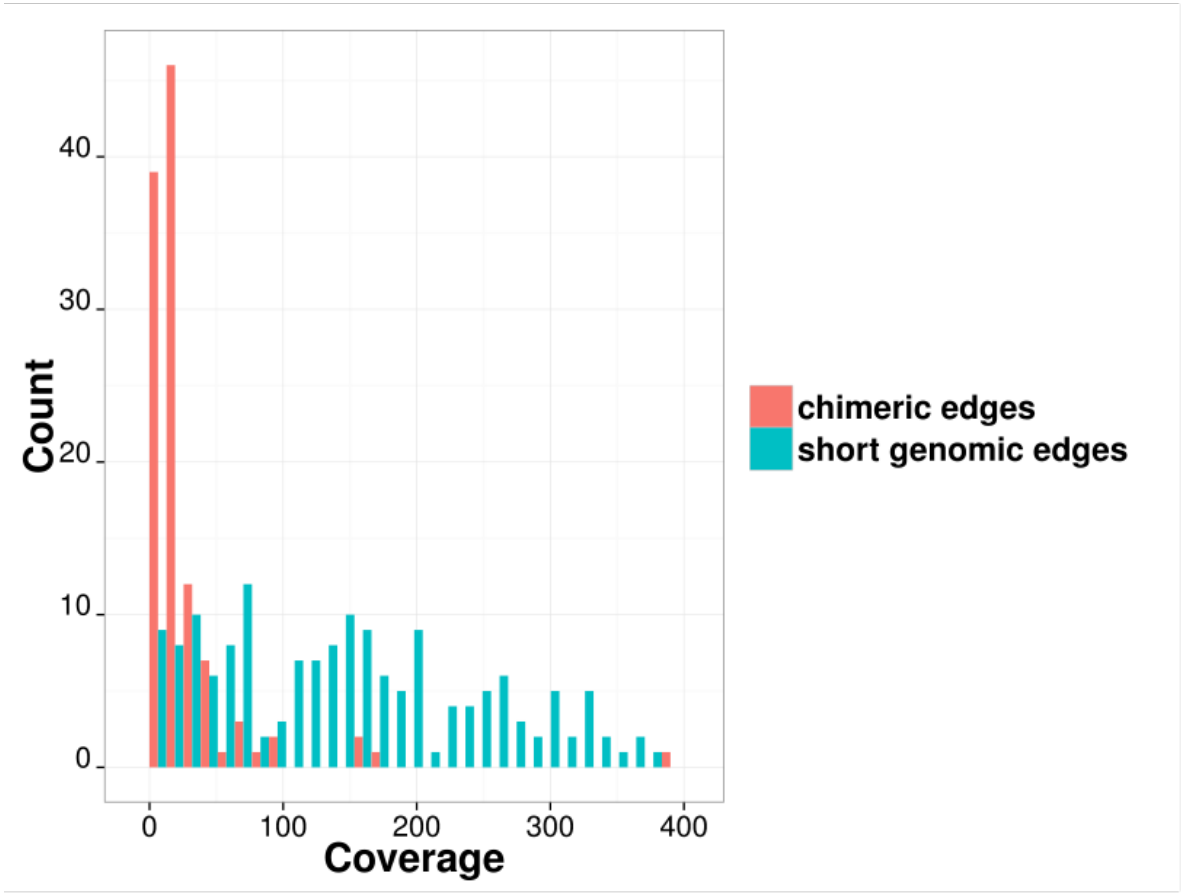


de Bruijn graph for k=3,4

Chimeric connections



Chimeric connections



Представление графа

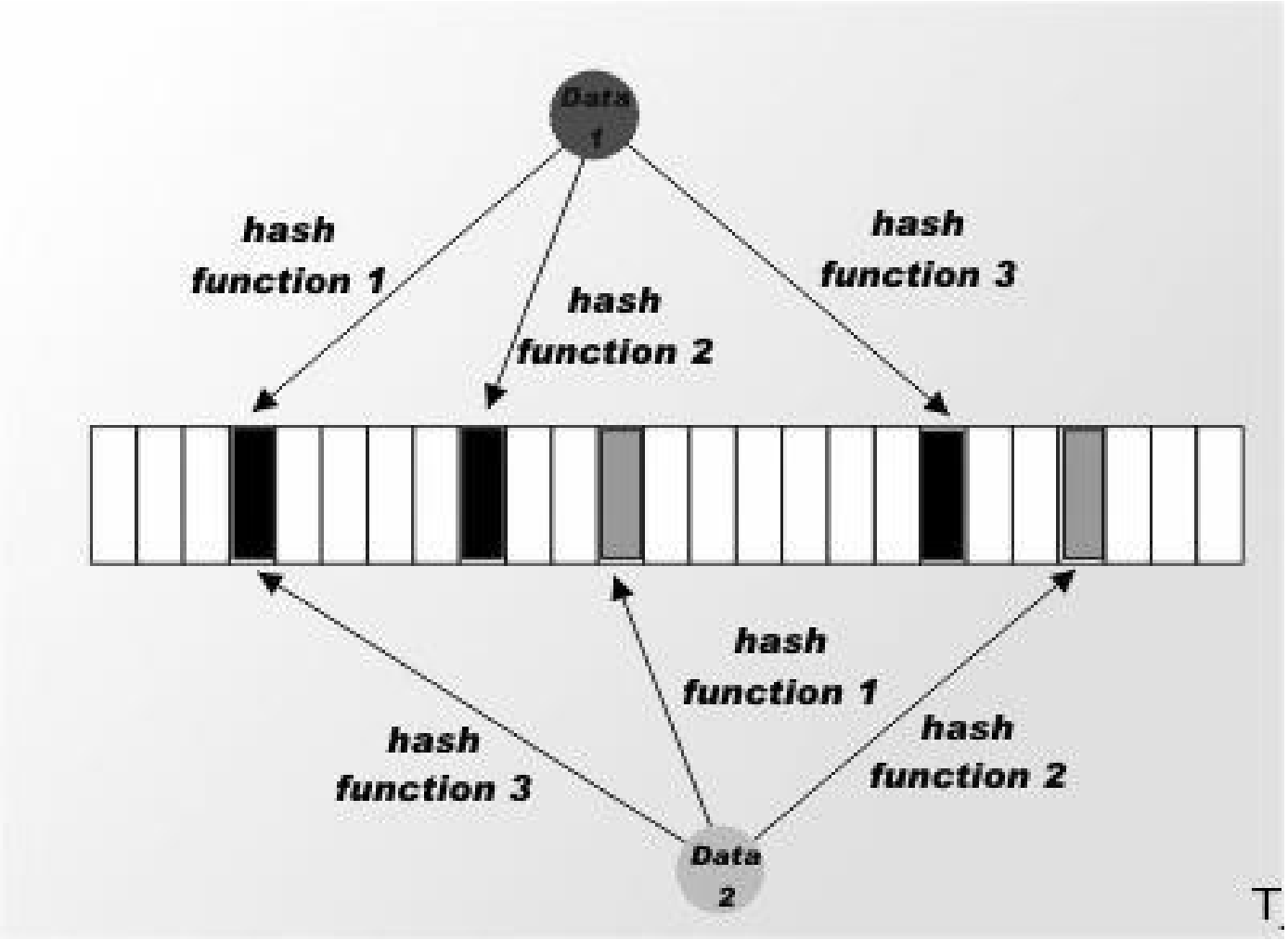
- Память
- Время

Что нужно для построения графа де Брюина?

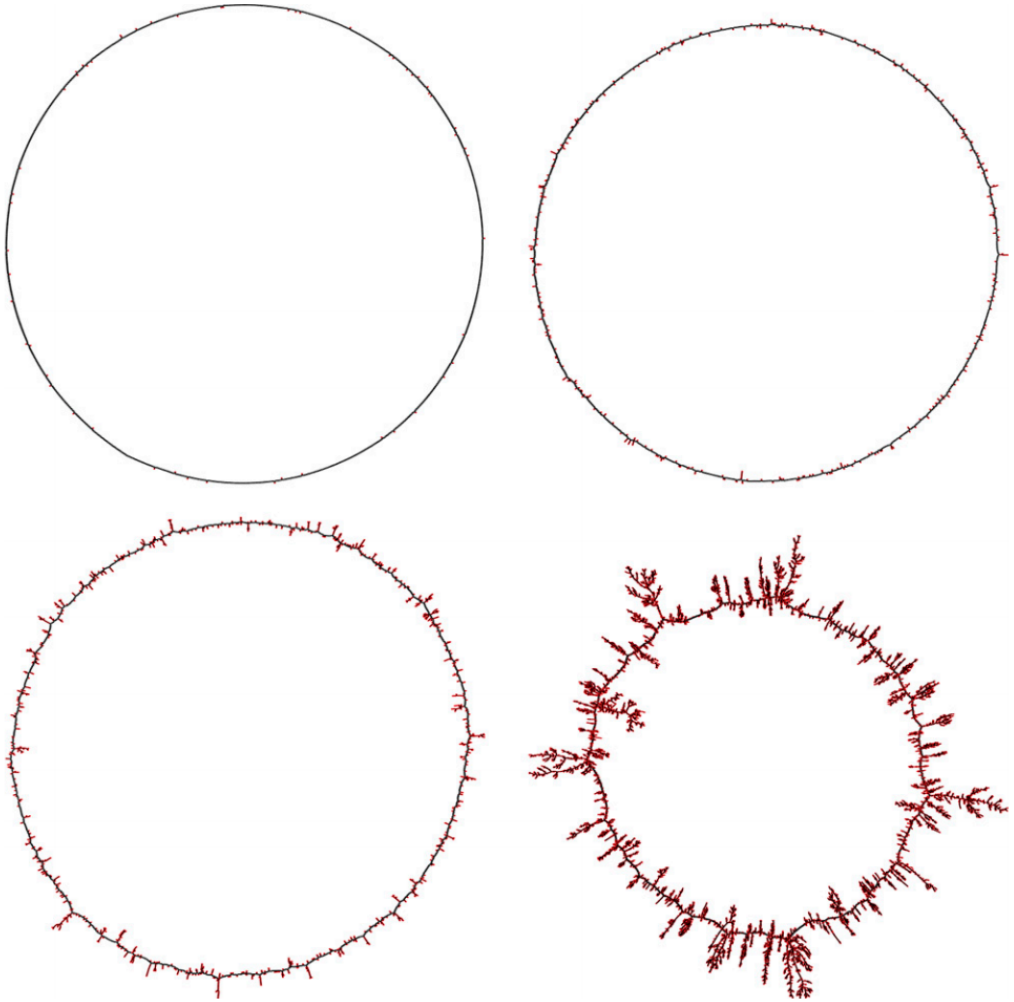
- Возможность перебрать все k-меры
- Возможность найти соседей k-мера

Пример: Множество всех (k+1)-меров

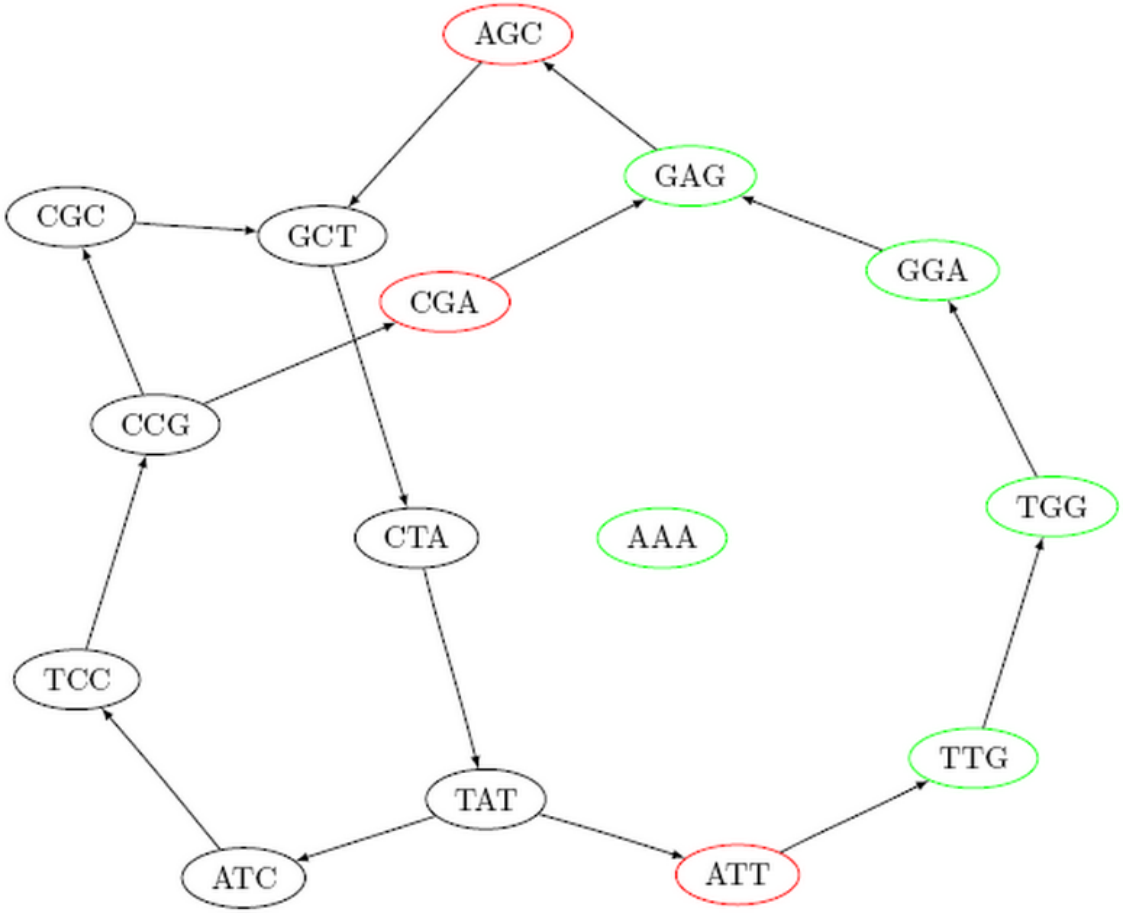
Фильтр Блума



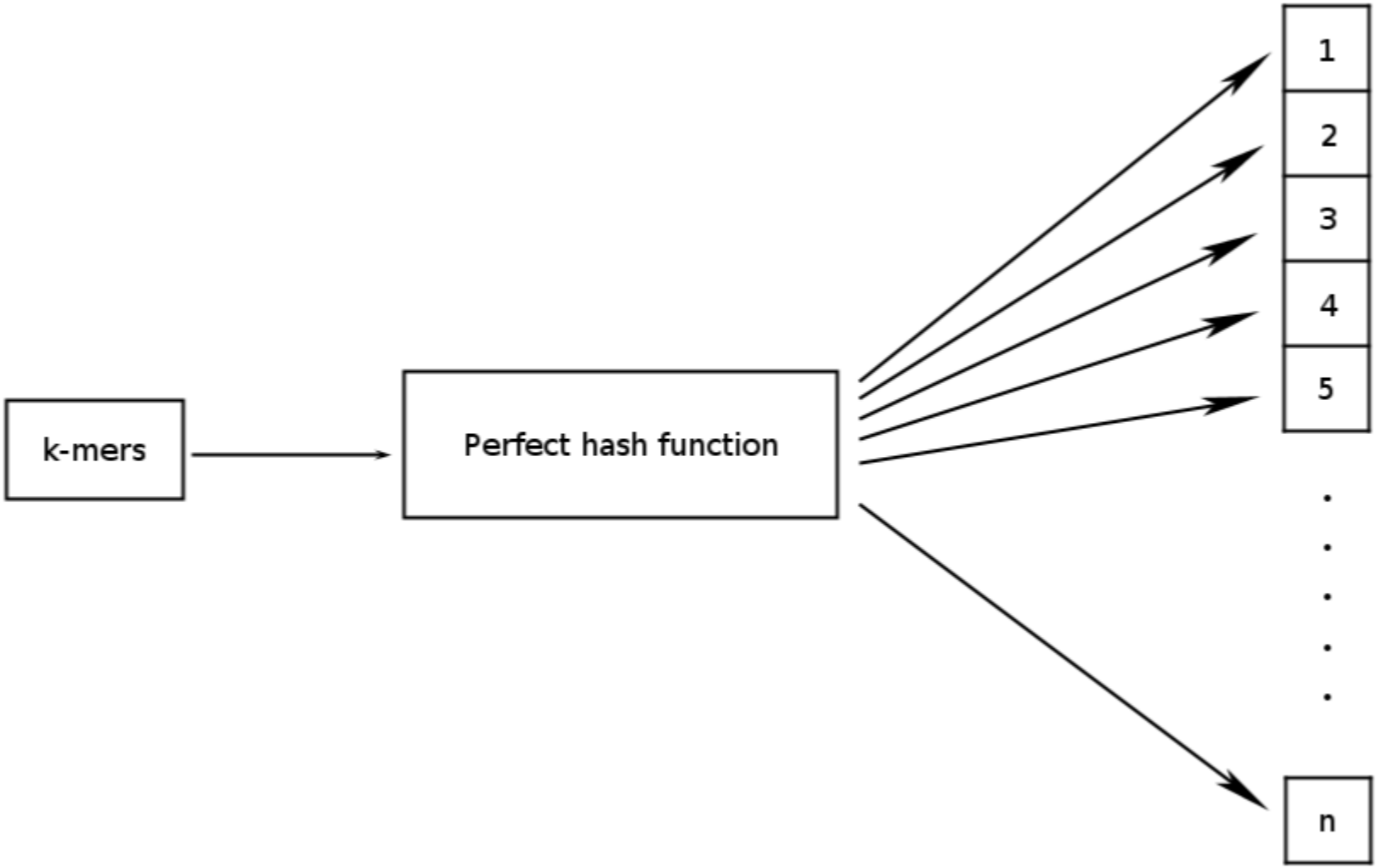
Вероятностный граф де Брюина



Точное представление



Хэширование без коллизий



Хэширование без коллизий

Позволяет:

- Хранить информацию в массиве
- Не хранить ключи

Не позволяет:

- Проверять наличие элемента в множестве

Требует:

- Предварительного нахождения уникальных k-меров

Реализация графа де Брюйна

- В хэш таблице хранятся все k-меры
- Для каждого k-мера хранятся все его соседи (8 бит)

Распределенное хранение

- Позволяет собрать что-то на кластере
- На порядок медленнее
- ABySS, Ray
- K-меры распределяются по нодам в соответствии с некоторым хэшем.

Ссылки

1. "Genome Reconstruction: A Puzzle with a Billion Pieces", P. Compeau, P Pevzner
2. *"De novo* assembly and genotyping of variants using colored de Bruijn graphs", Zamin Iqubal et al.
3. "Scaling metagenome sequence assembly with probabilistic de Bruijn graphs", Jason Pell et al.
4. "Space-efficient and exact de Bruijn graph representation based on a Bloom filter", Rayan Chikhi, Guillaume Rizk
5. "External Perfect Hashing for Very Large Key Sets", Fabiano C. Botelho, Nivio Ziviani
6. http://bioinf.spbau.ru/en/spades