

Сборка генома и графы де Брюина

Сергей Нурк

Лаборатория алгоритмической биологии

АУ РАН

<http://bioinf.spbau.ru>

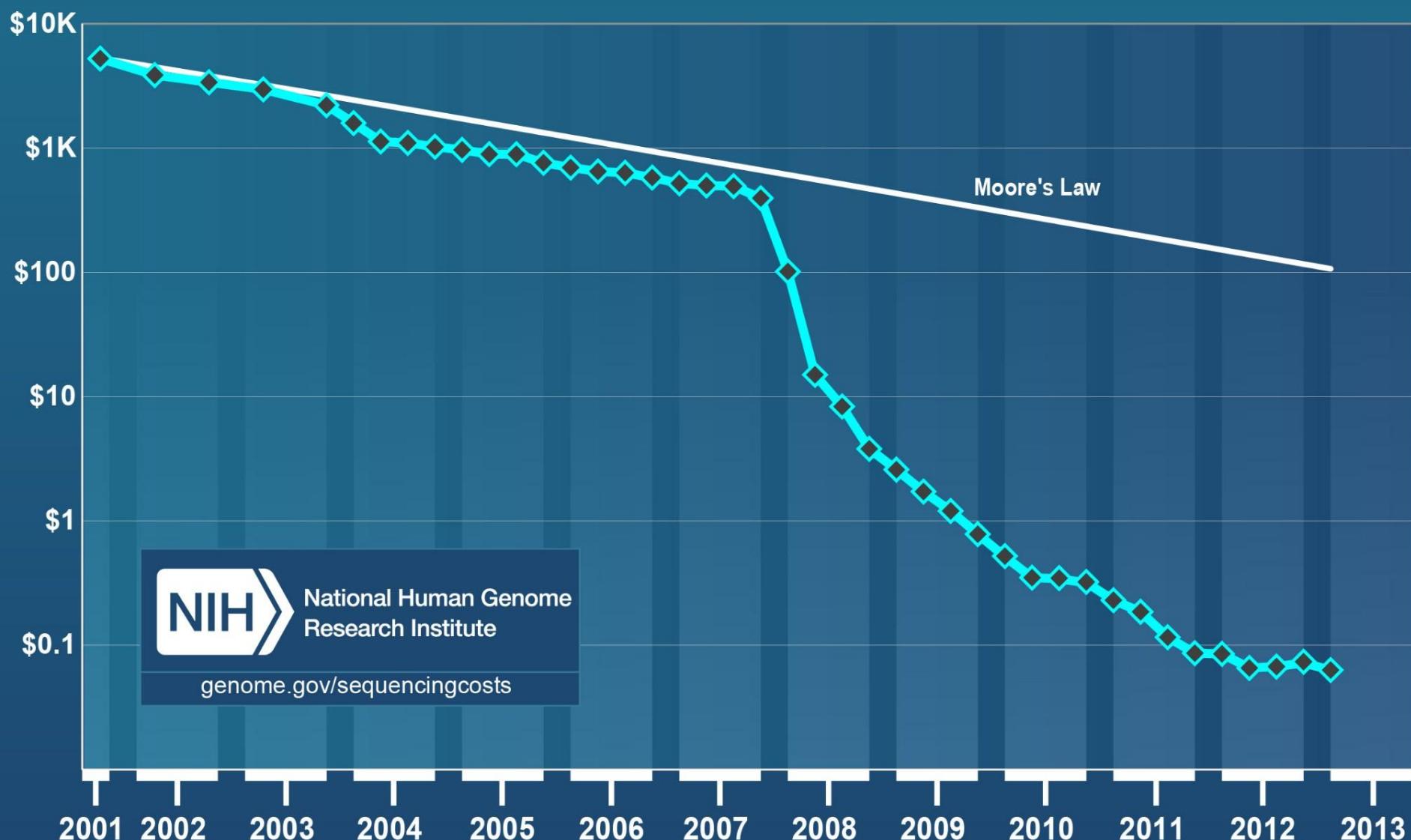
Введение

NGS революция

Начало 2000-х: первые NGS технологий

Вместо длинных, но дорогих фрагментов
секвенаторы выдают много коротких
фрагментов по низкой цене.

Cost per Raw Megabase of DNA Sequence

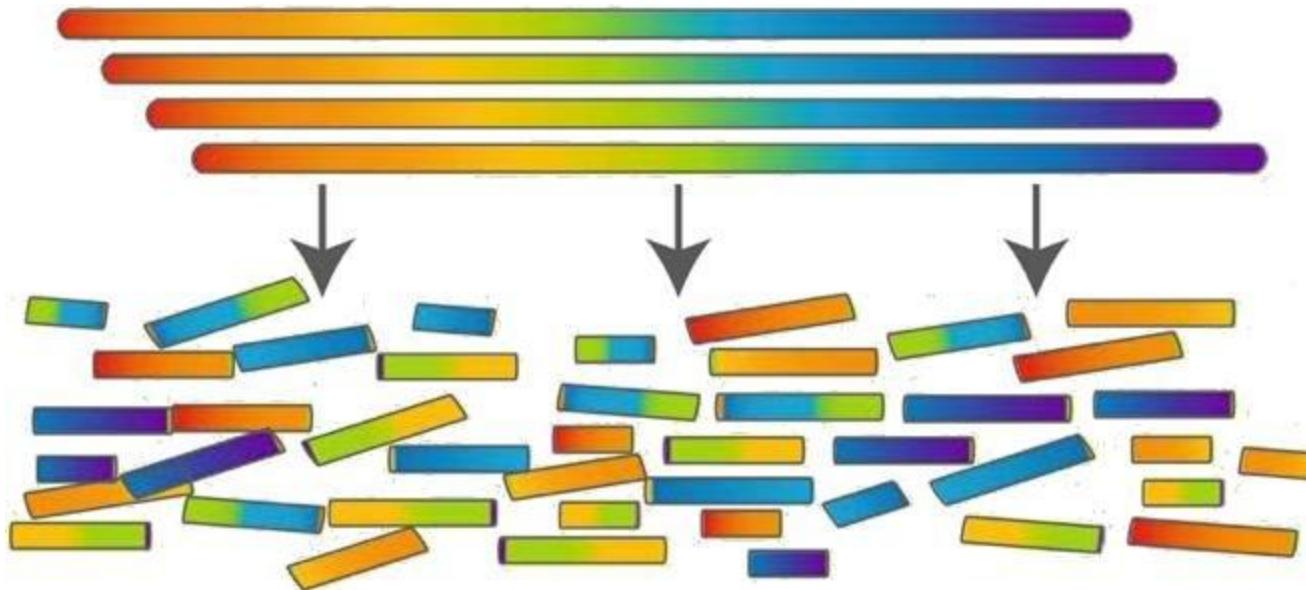


National Human Genome
Research Institute

genome.gov/sequencingcosts

Сборка

Whole genome shotgun sequencing



Сборка (assembly) -- восстановление
участков изначальной последовательности

Задача сборки



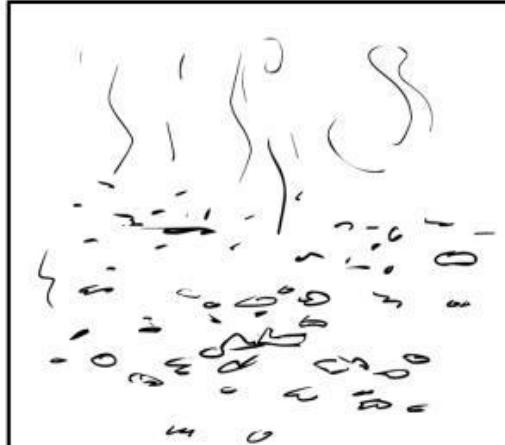
stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000
on a pile of dynamite



this is just hypothetical



so, what did the June 27, 2000 NY
Times say?

SSP

Дано: множество строк S_i

Найти: кратчайшую строку S , содержащую
все S_i

Задача NP-полная

Основная проблема: решение не имеет
отношения к реальности!

Задача сборки

Получить последовательности нуклеотидов (контиги), которые:

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

NGS Ассемблеры

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

Графы де Бройна

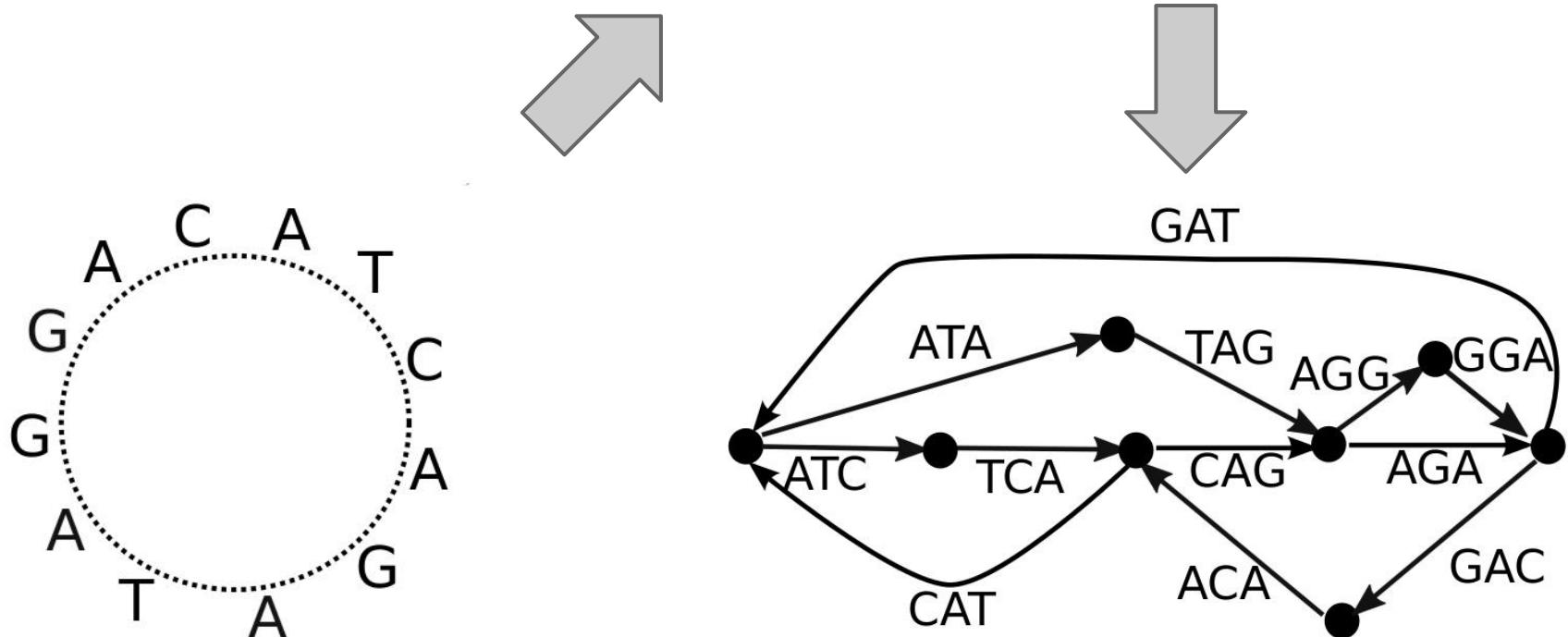


Графы де Брюйна

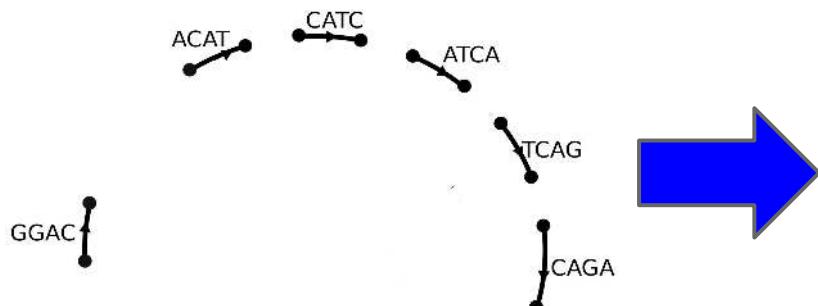
- k -мер: последовательность из k нуклеотидов
- Вершины графа де Брюйна: все k -меры
- Рёбра графа де Брюйна: все $(k+1)$ -меры
- Ребро e соединяет префикс и суффикс e

Графы де Брёйна

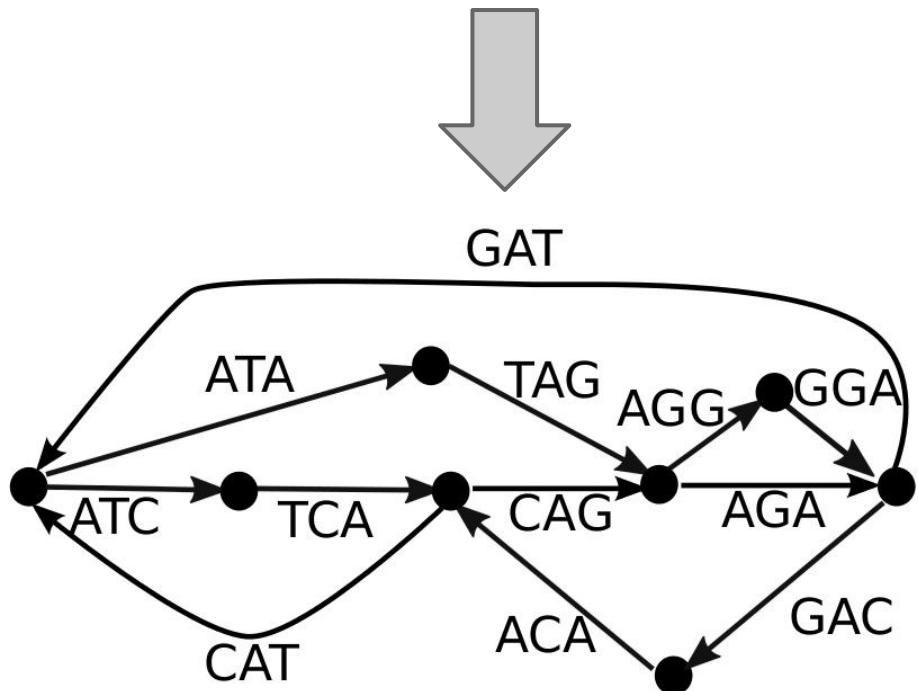
Вершины: k-меры из генома
Рёбра:(k+1)-меры из генома
k=2: 3-мер ACG даёт AC -> CG



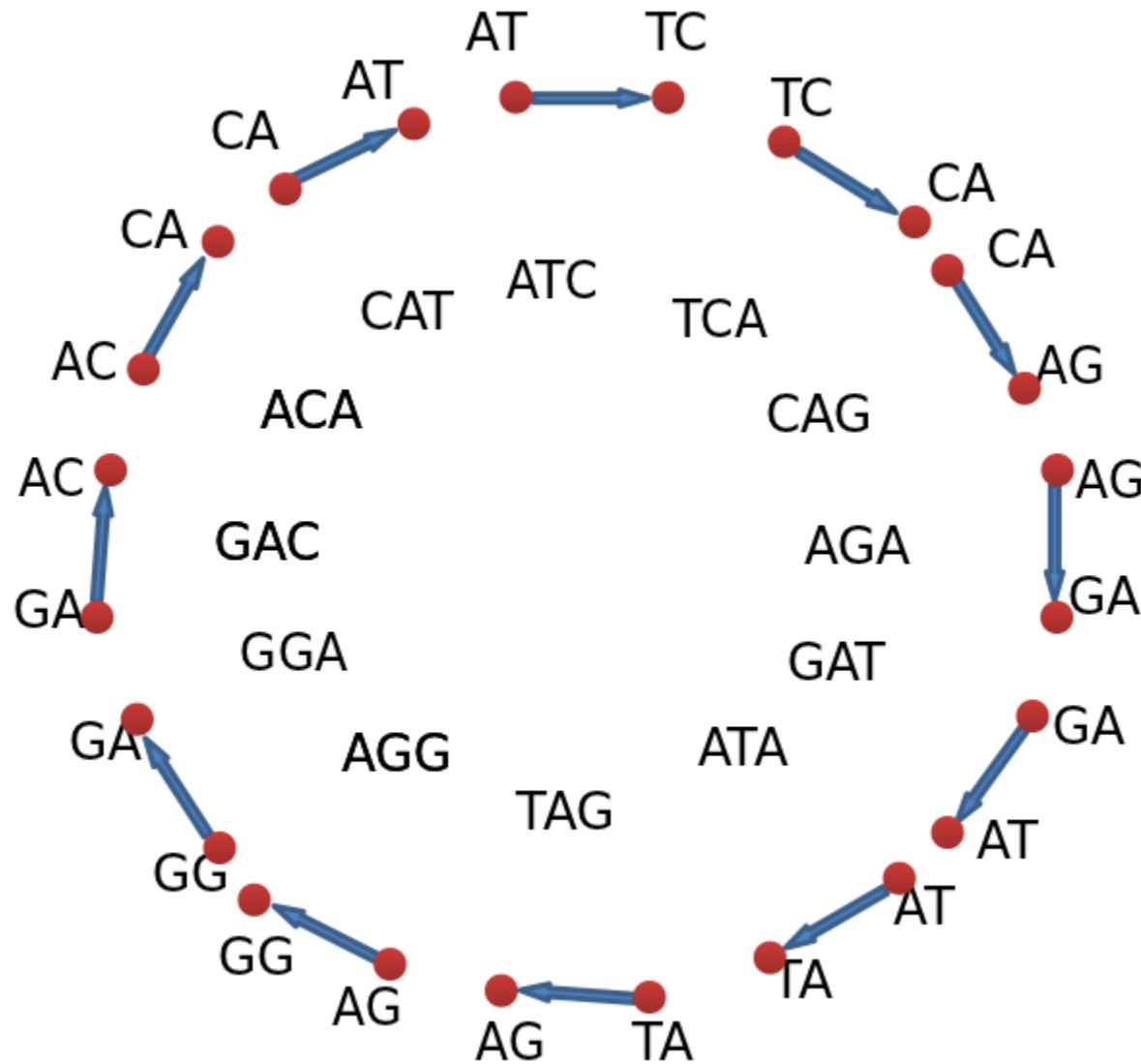
Графы де Брюйна



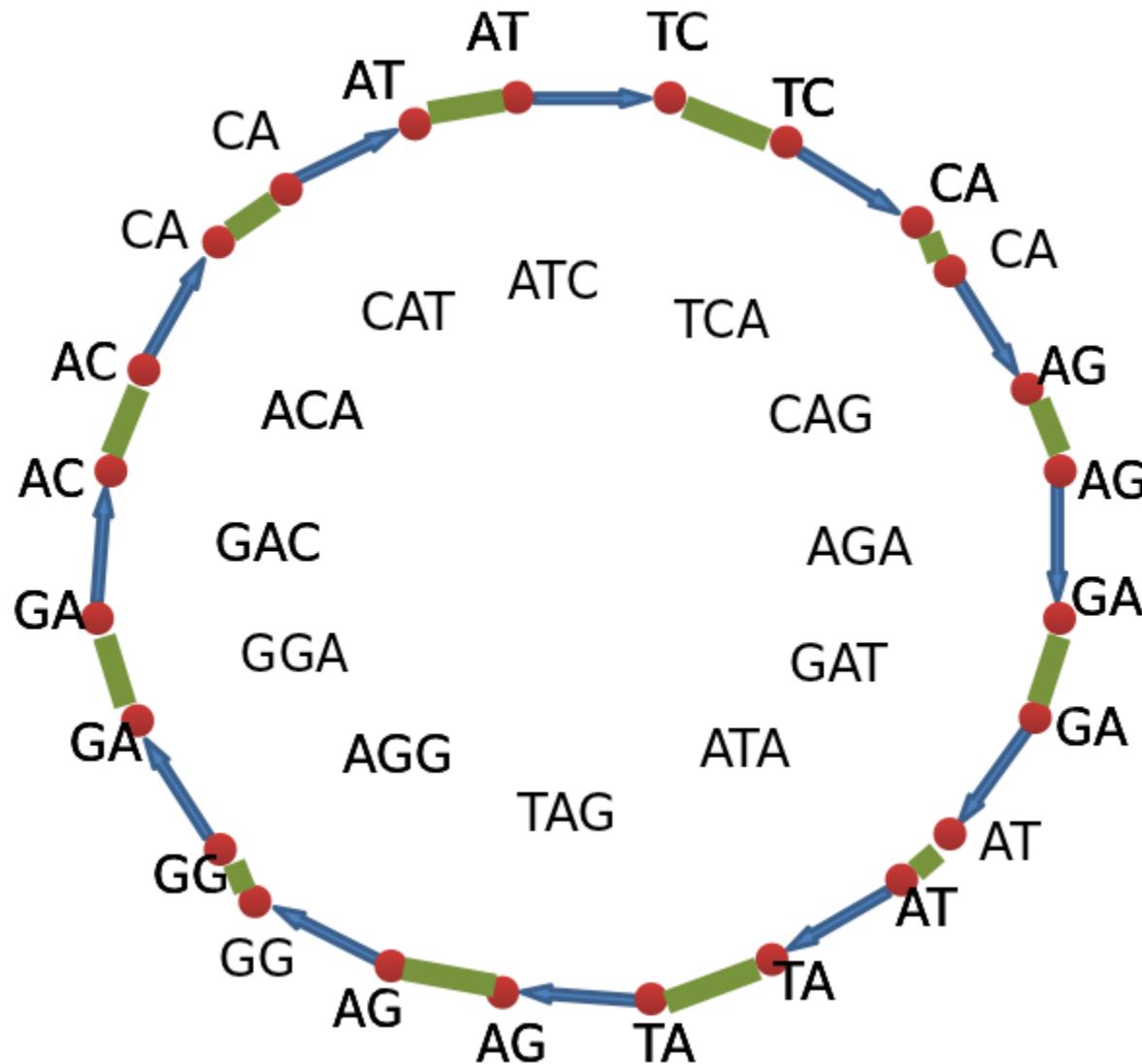
Вершины: k-меры из **ридов**
Рёбра:(k+1)-меры из **ридов**
k=2: 3-мер ACG даёт AC -> CG



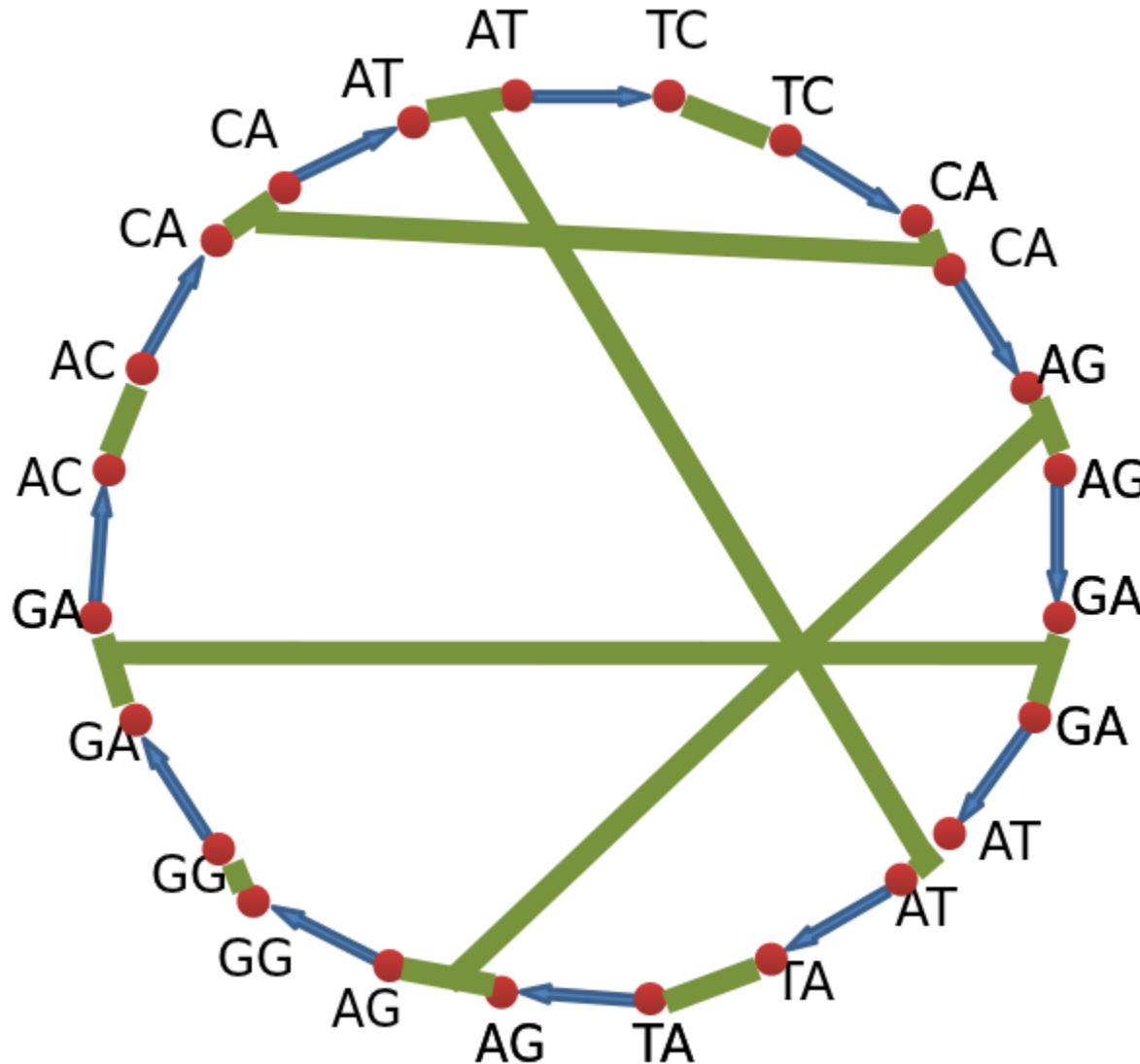
Графы де Брюйна



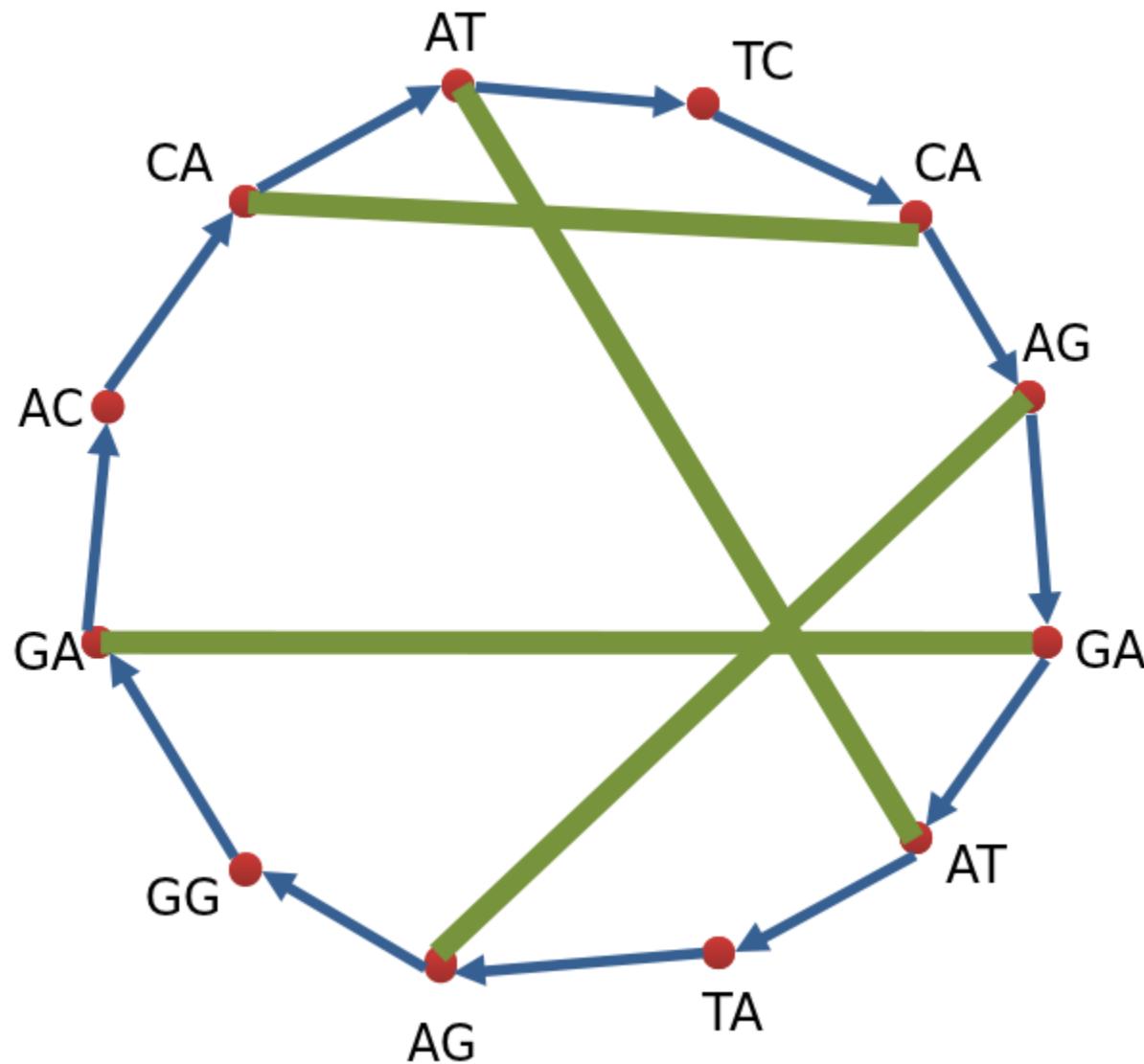
Графы де Брюйна



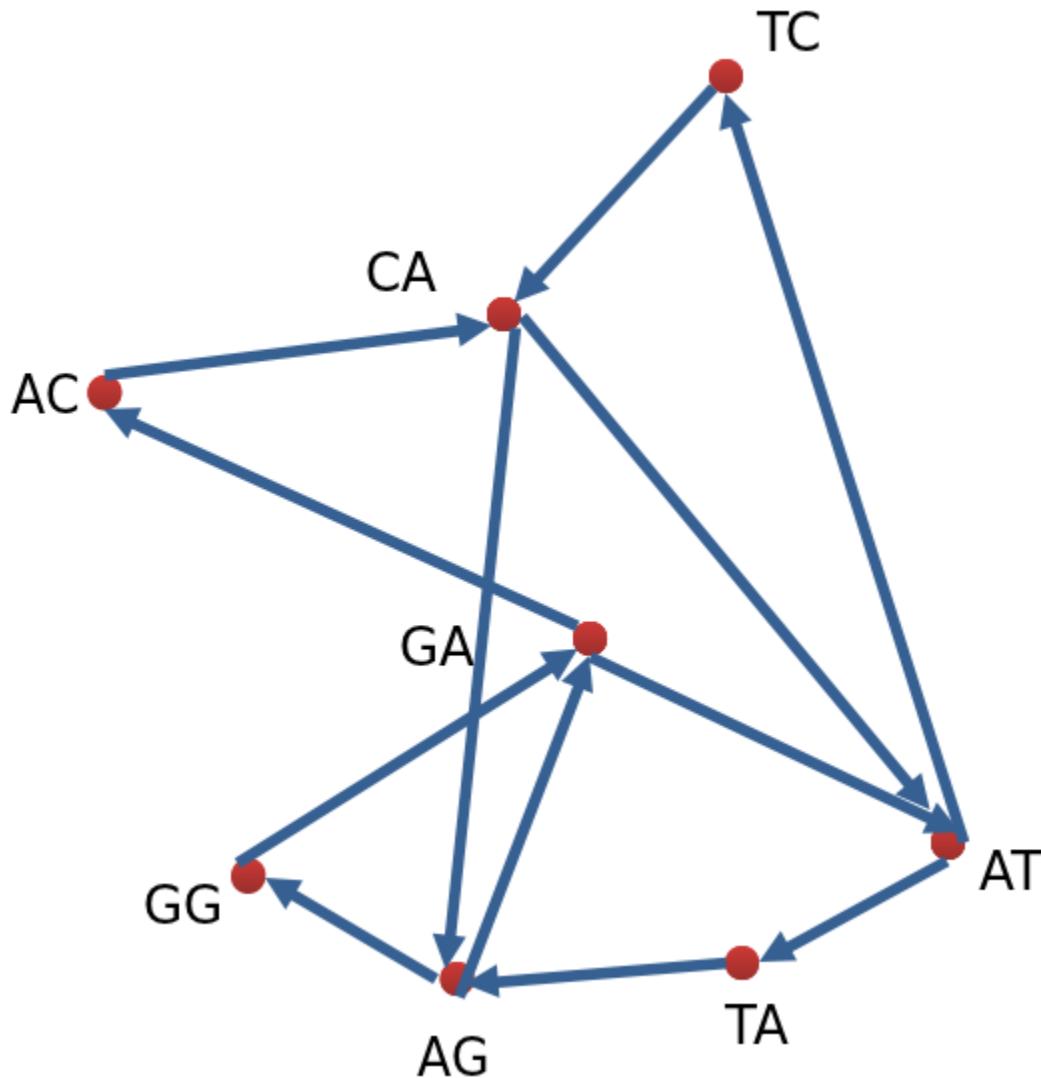
Графы де Брюйна



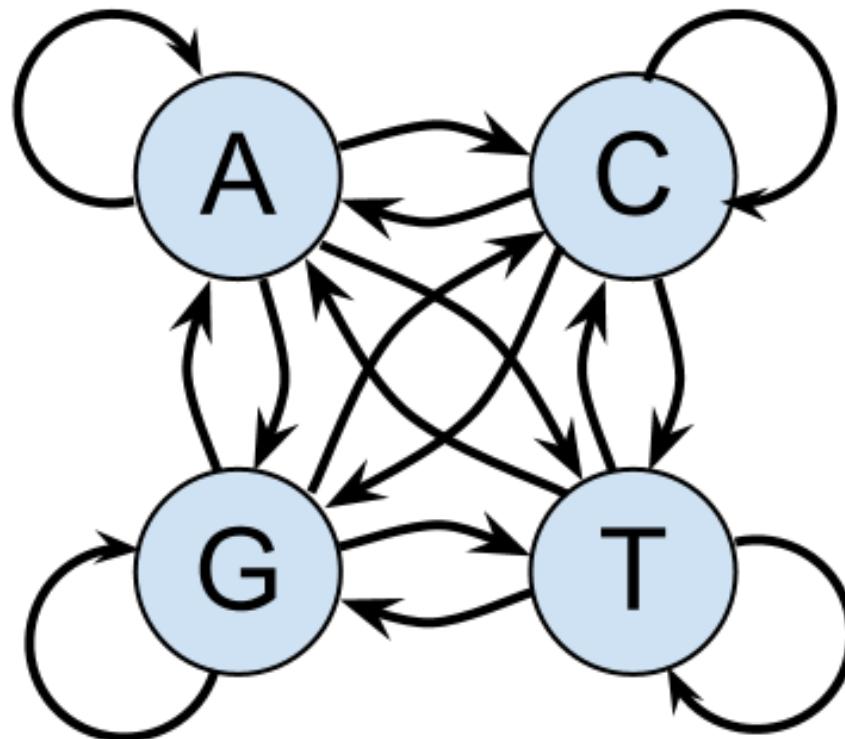
Графы де Брюйна



Графы де Брюйна



К имеет значение!

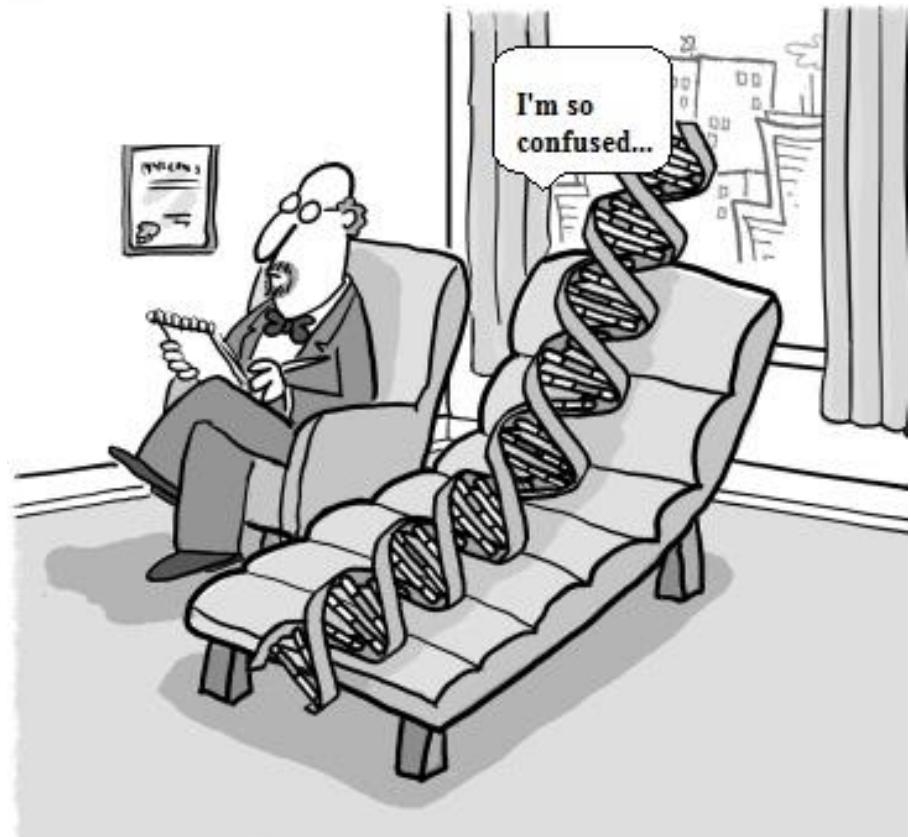


Проблема повторов

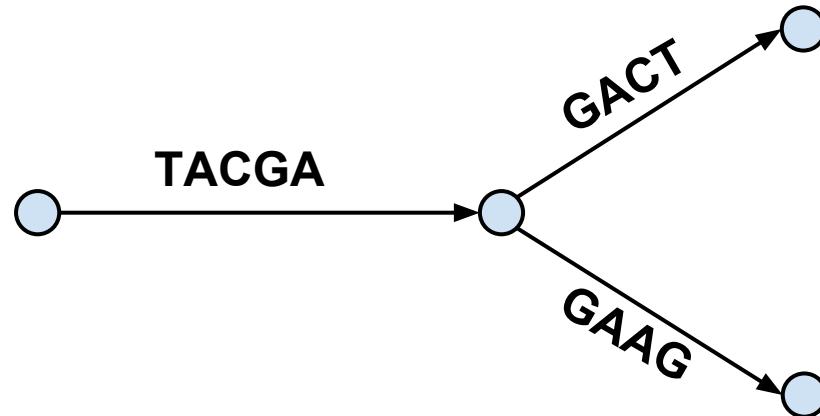
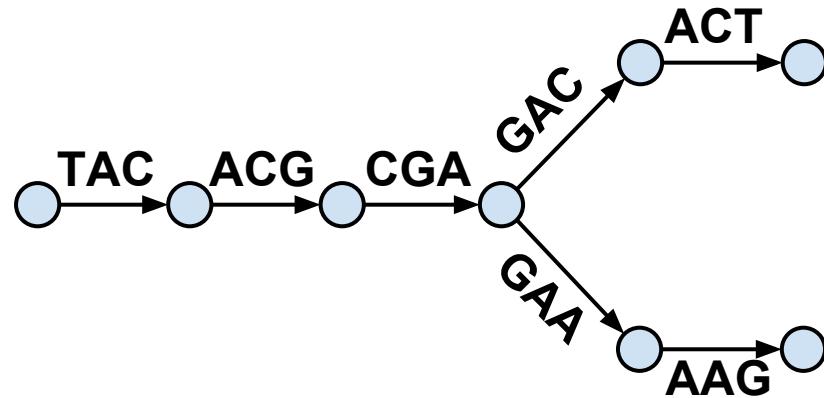
ALU

длина: 300

кратность: 1000000

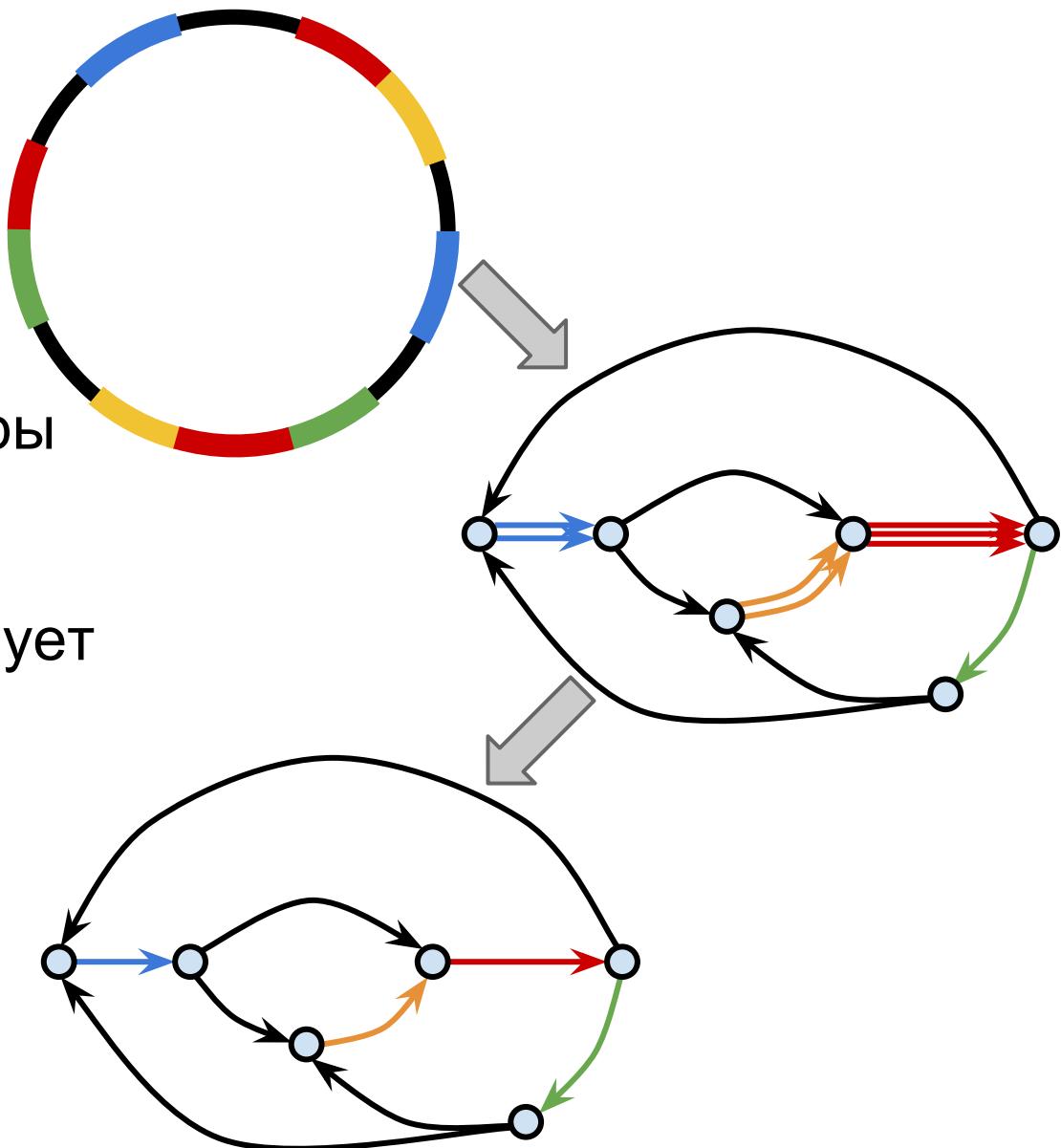


Сжатый граф

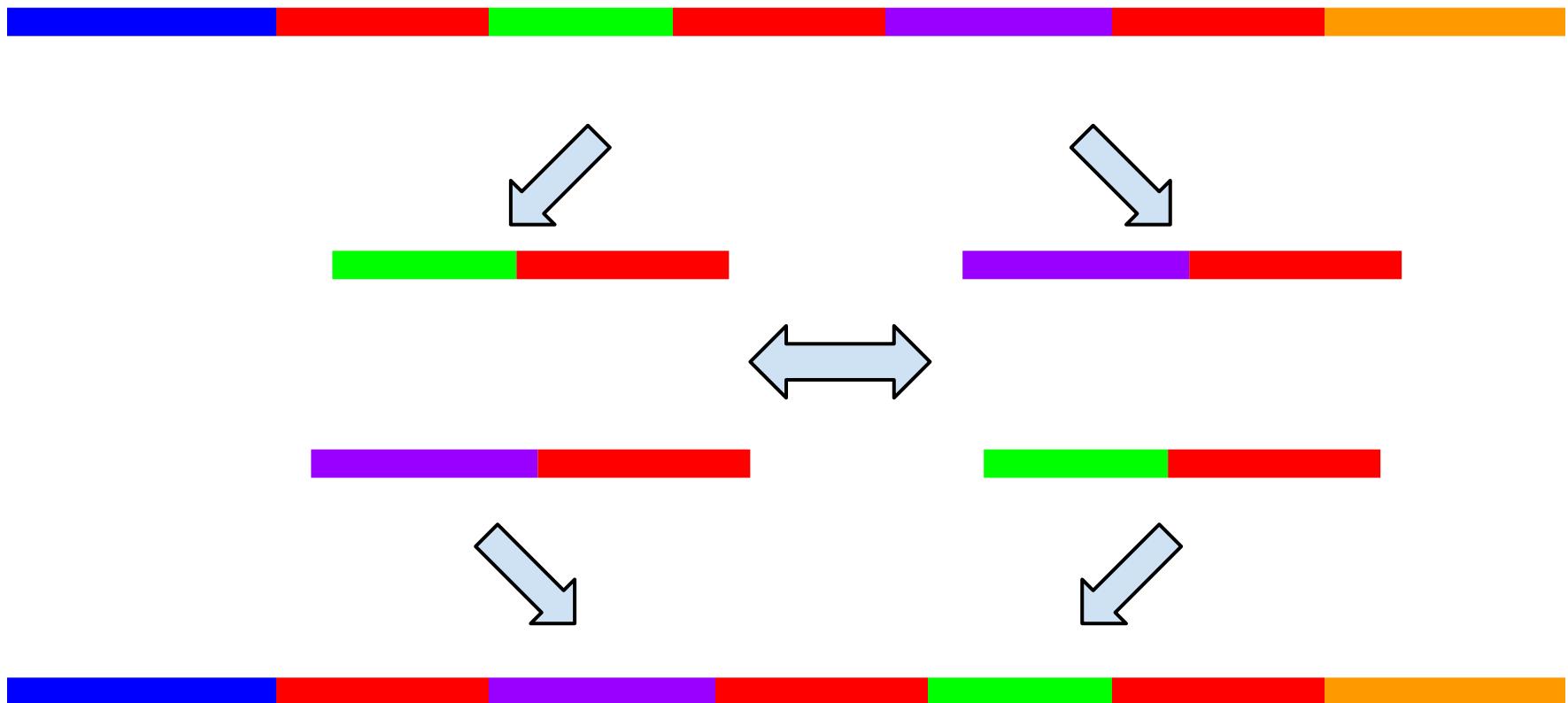


Заметки про граф де Брюйна

1. Склейивает повторы
(длиннее k)
2. Геном соответствует
циклу в графе

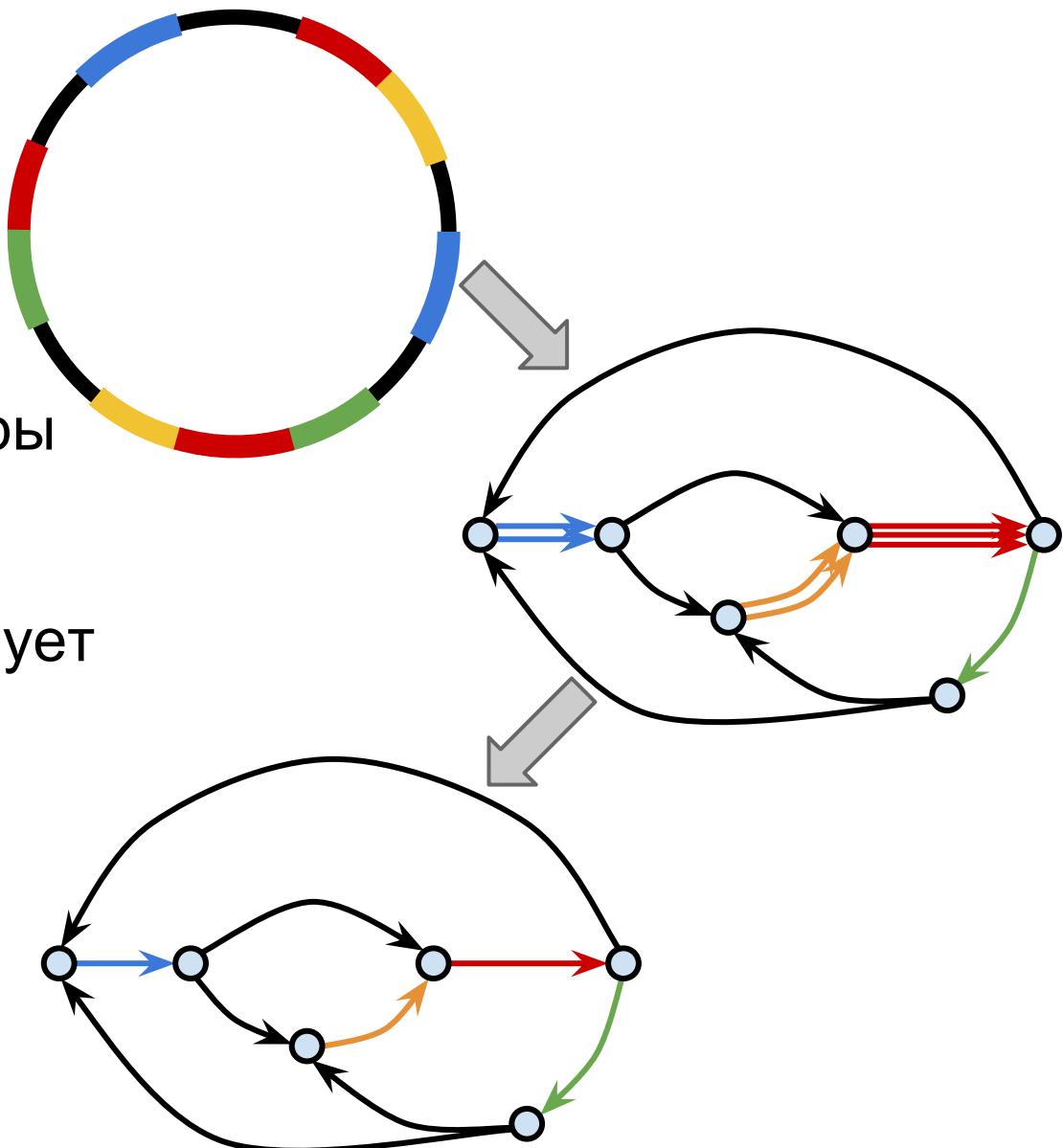


Проблема повторов



Заметки про граф де Брюйна

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



Некоторые проблемы

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

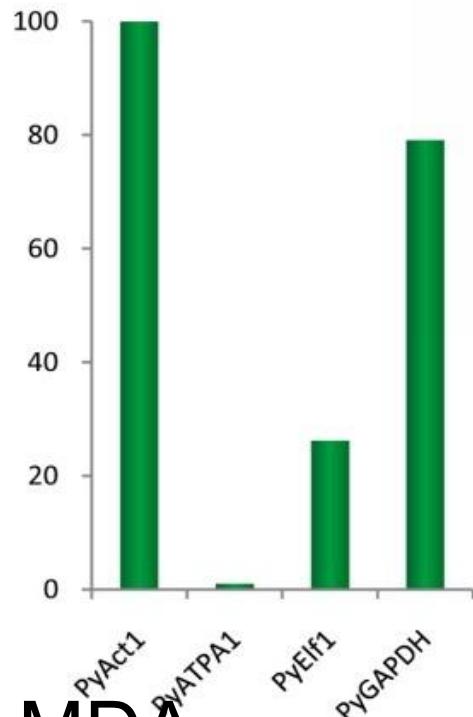
Разрывы в покрытии

Покрытие конкретного $(k+1)$ -мера —
случайная величина

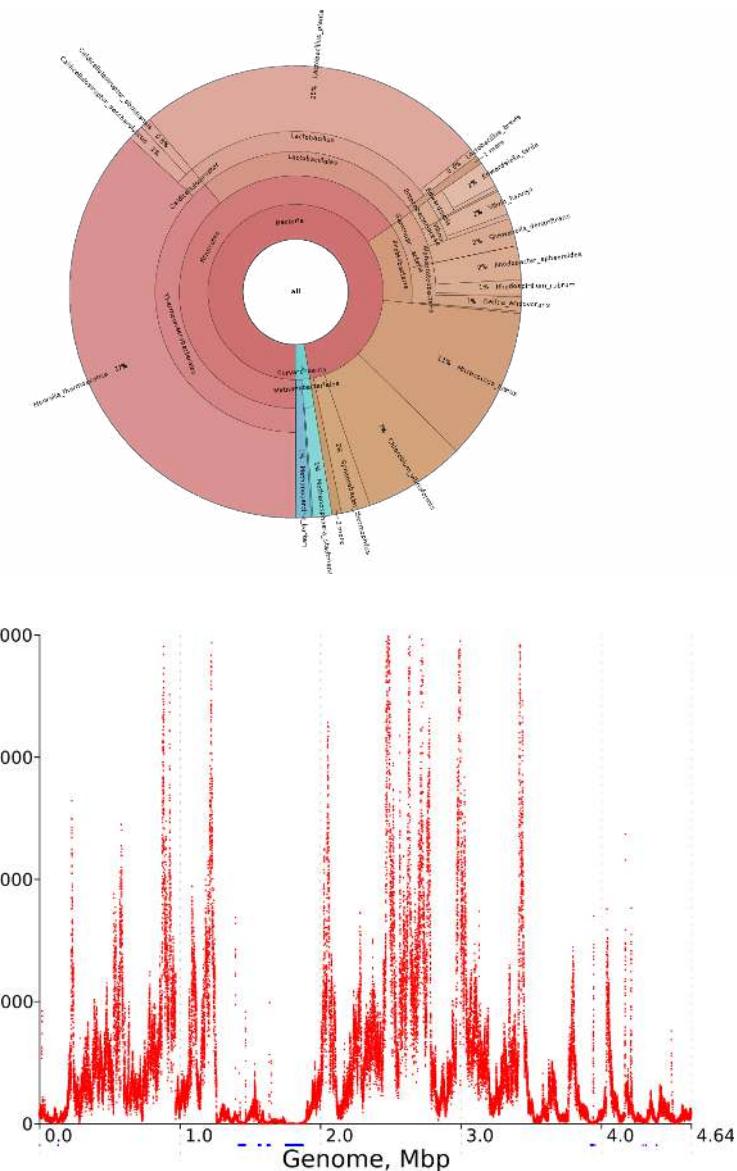
Чтобы снизить вероятность разрыва,
приходится использовать k значительно
меньше длины рида

Неравномерное покрытие

1. Метагеномные данные

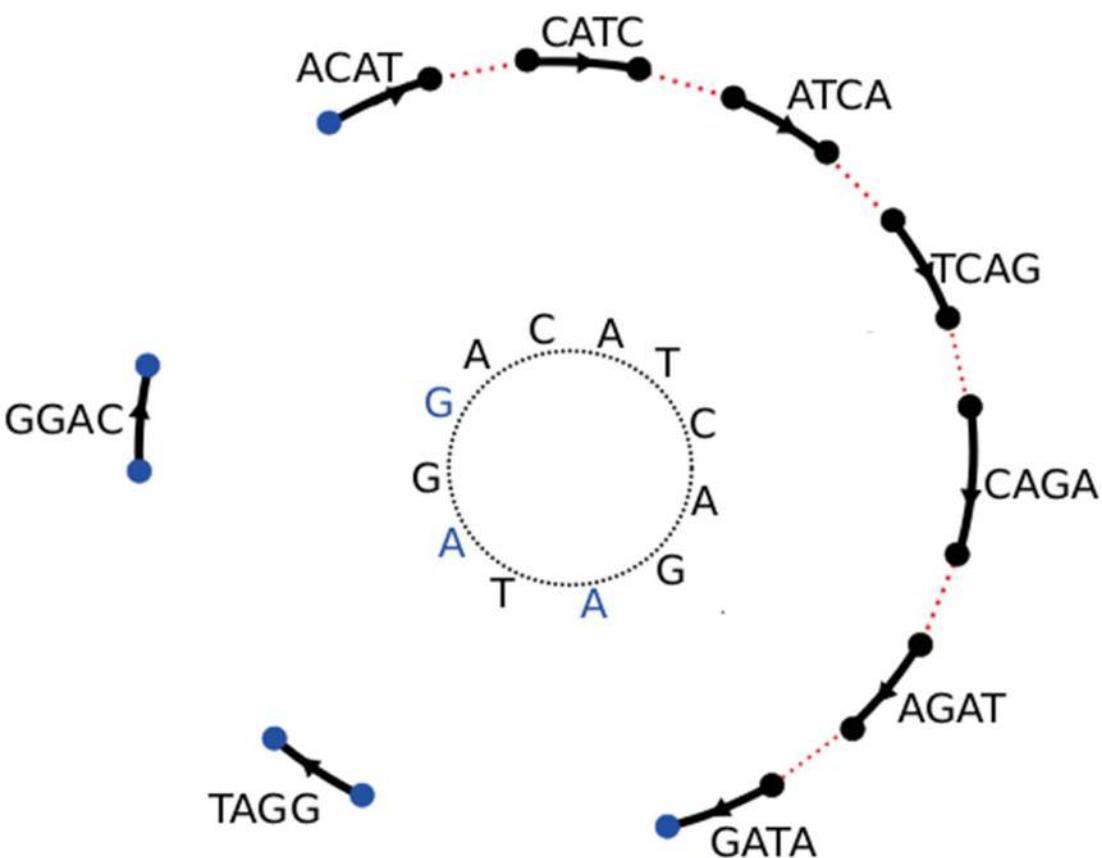


2. RNA-seq

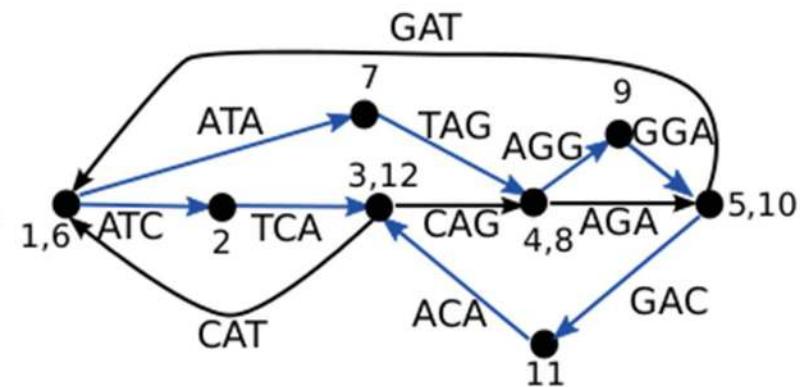


3. Single-cell MDA

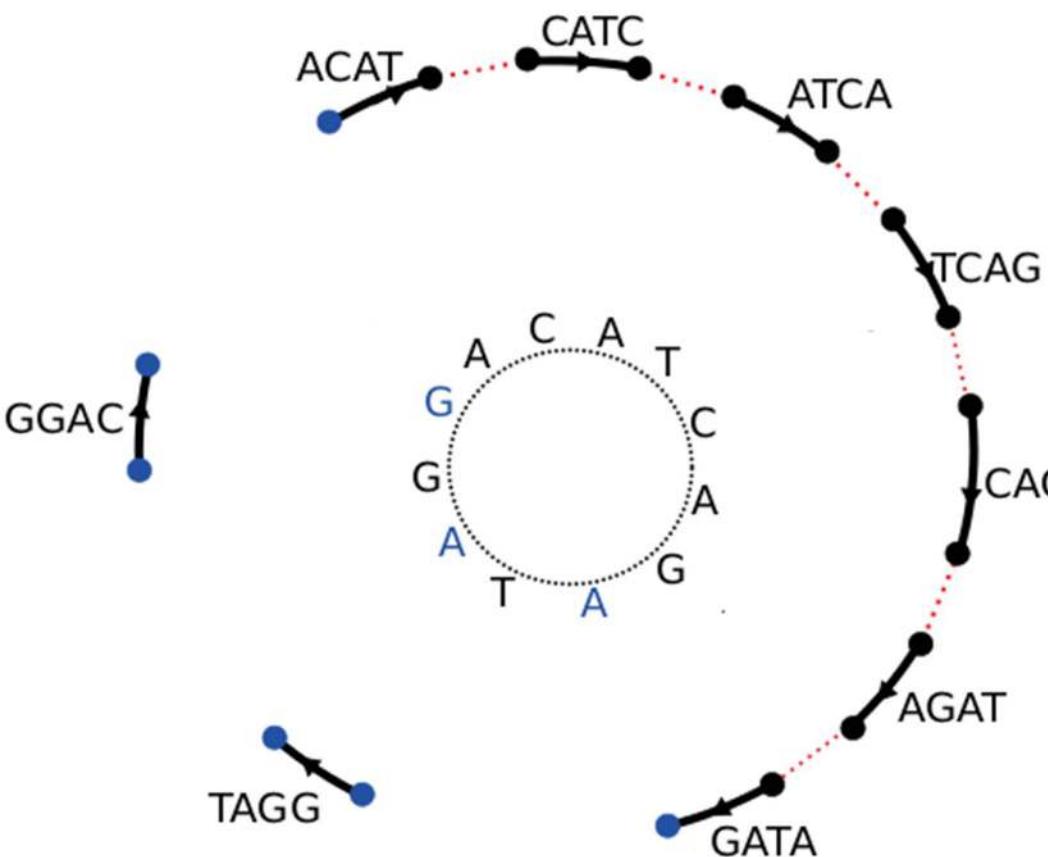
Борьба с разрывами



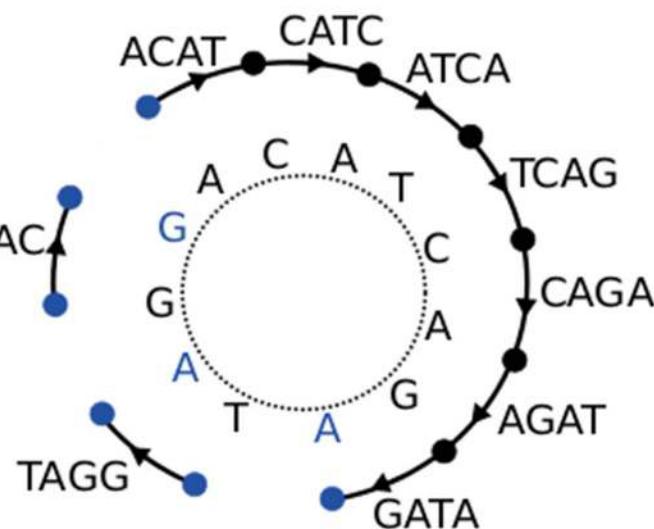
De Bruijn graph for $k=2$



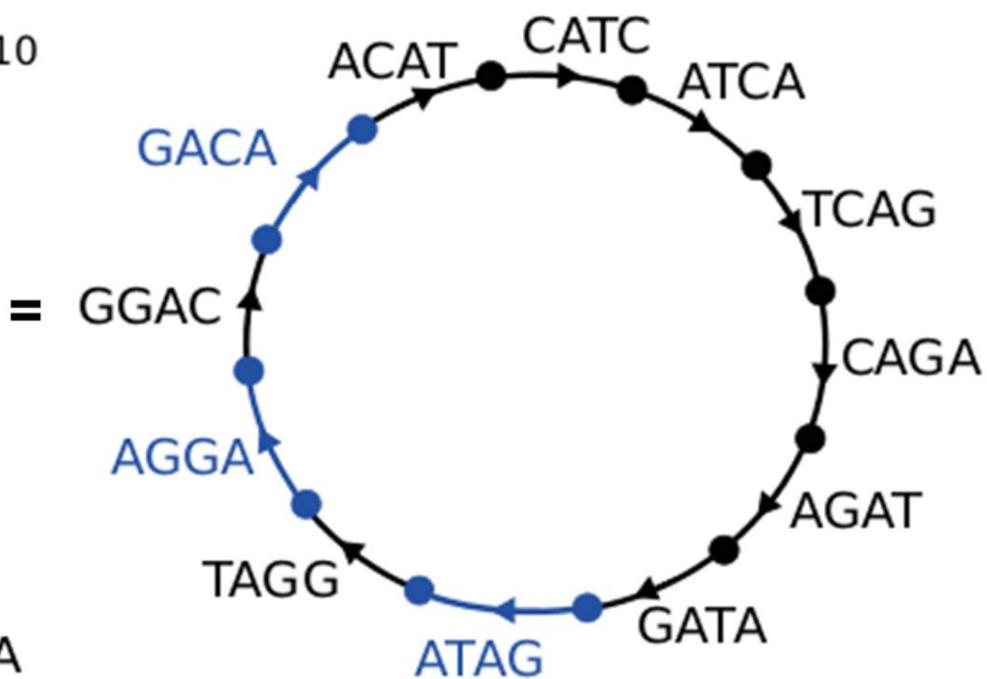
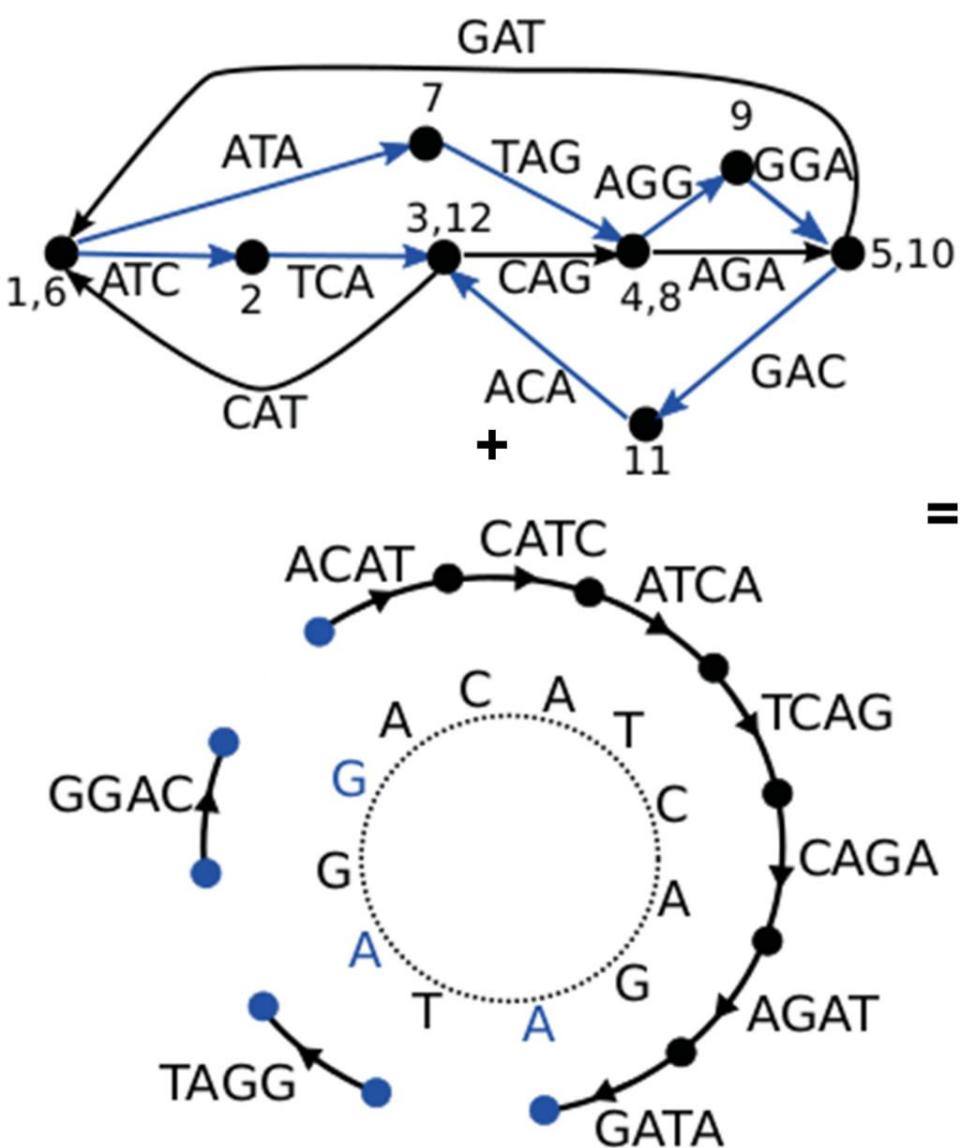
Борьба с разрывами



De Bruijn graph for $k= 3$

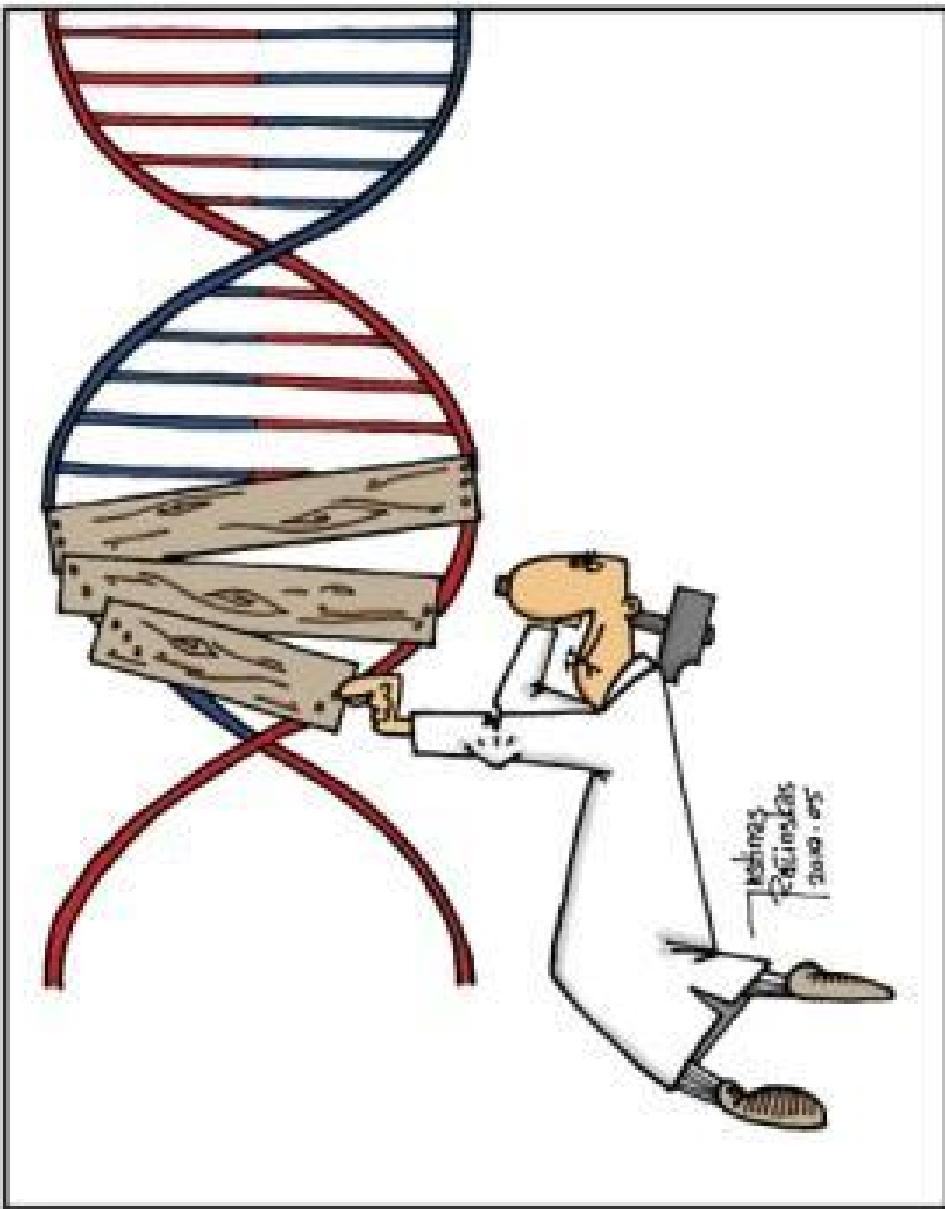


de Bruijn graph for $k=2,3$



Ошибки секвенирования

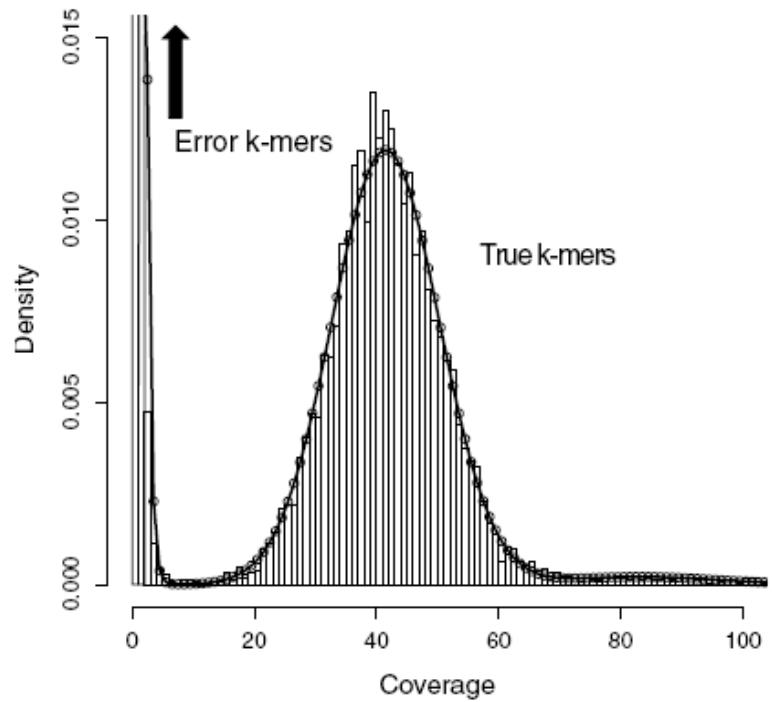
- Тип и частота зависят от технологий
- Секвенаторы предоставляют информацию о качестве каждого нуклеотида в риде
- Предобработка ридов: Quake, BayesHammer



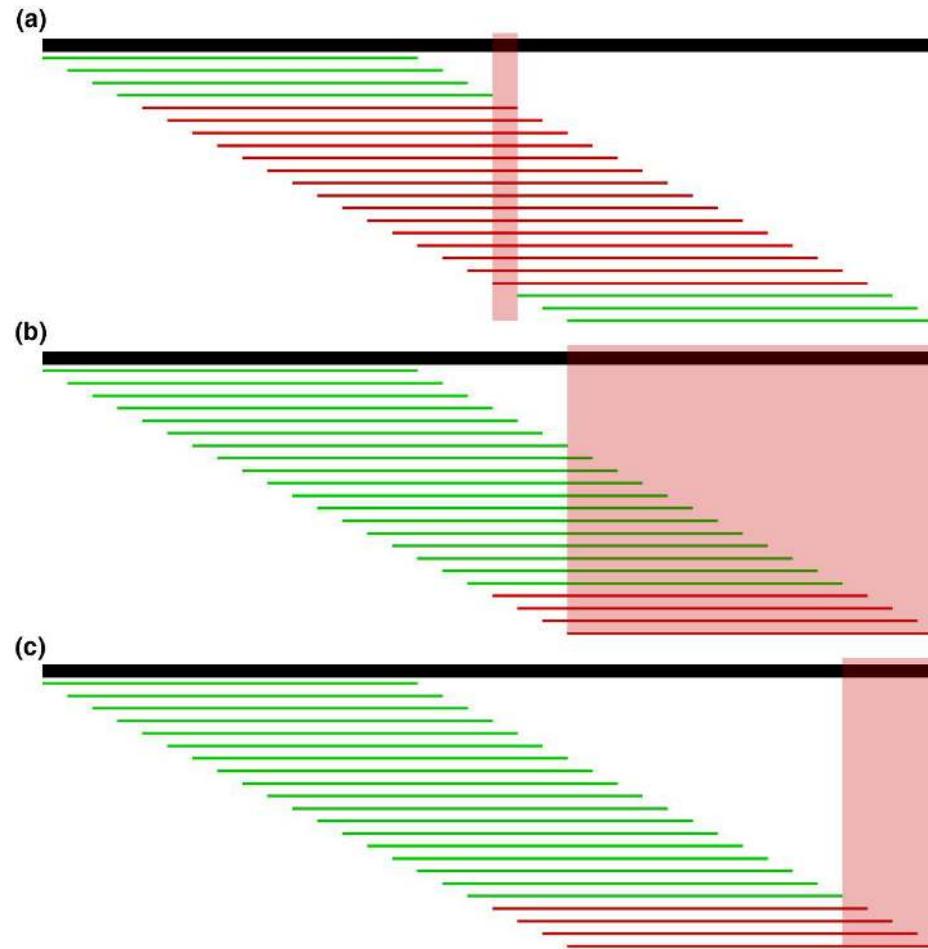
Thomas
Pfeiffer
2010 · 05

Quake. Надежные k-меры

- "Хорошо" покрытые k-меры объявляются надёжными
- Отсечка определяется исходя из распределения покрытия



Quake. Коррекция ридов



Hammer

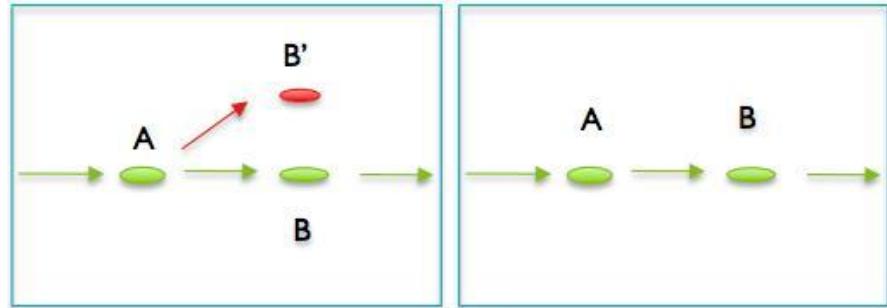


Reads	k -mers	$\text{HG}_1(X)$
ACGTGTG	ACGTG CGTGT GTGTG	ACATG ACCTG CGTGT
ACATGTG	ACATG CATGT ATGTG	ACGTG ATGTG CCTGT
ACCTGTC	ACCTG CCTGT CTGTC	GTGTG CTGTC

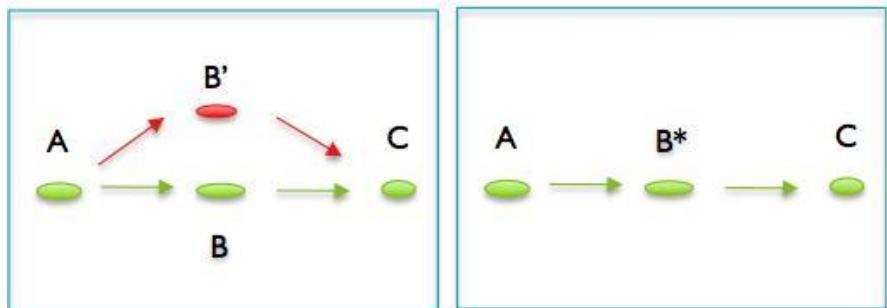
Ошибки в графе

Неисправленные ошибки превращаются в "лишние" ребра в графе

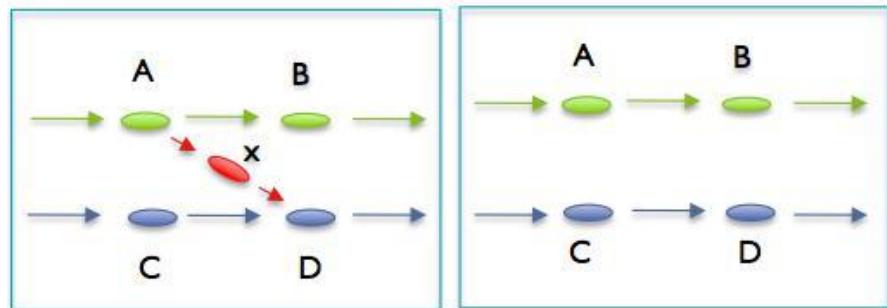
tip

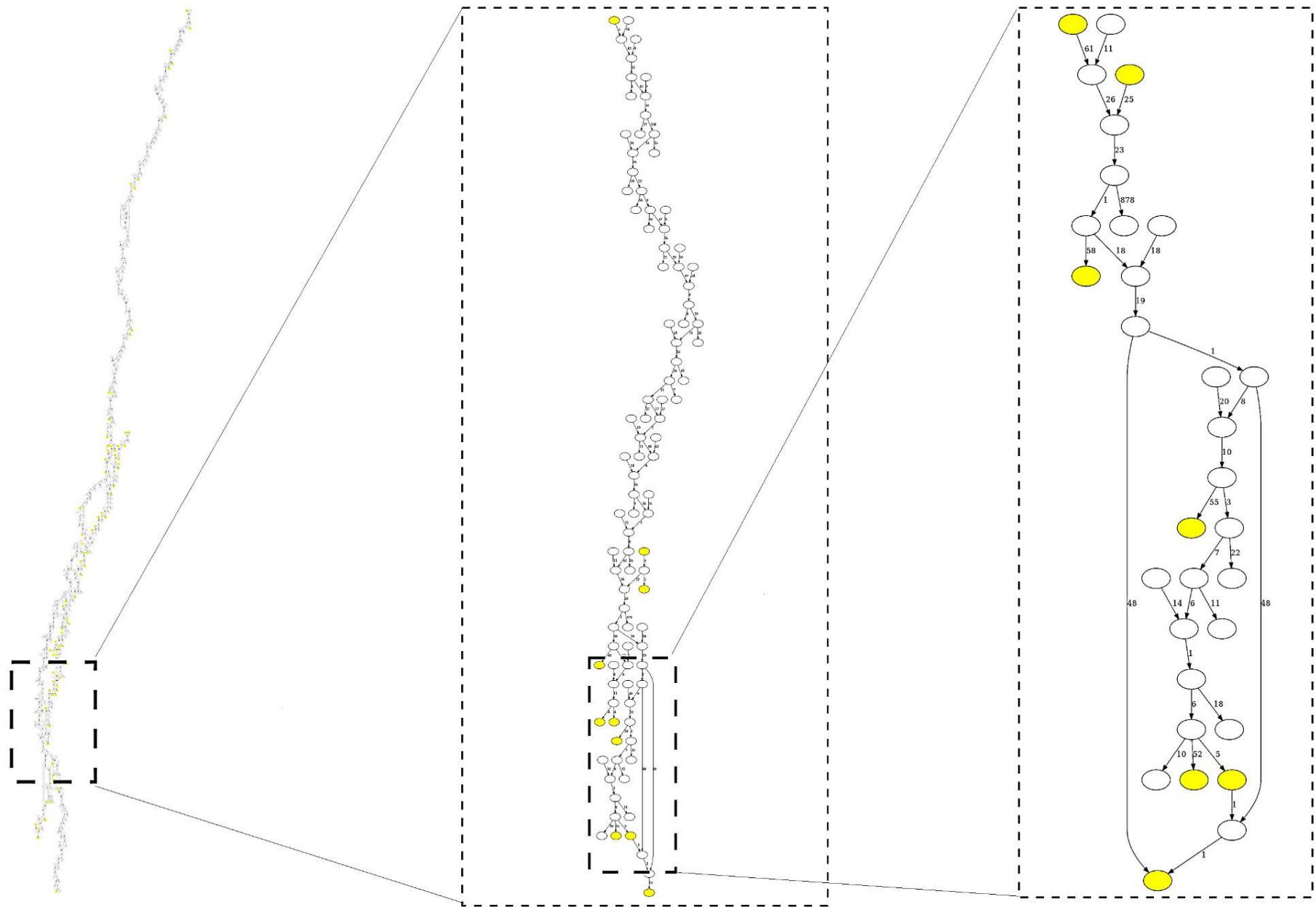


bulge



chimeric connection





Техника



Представление графа

- Память
- Время

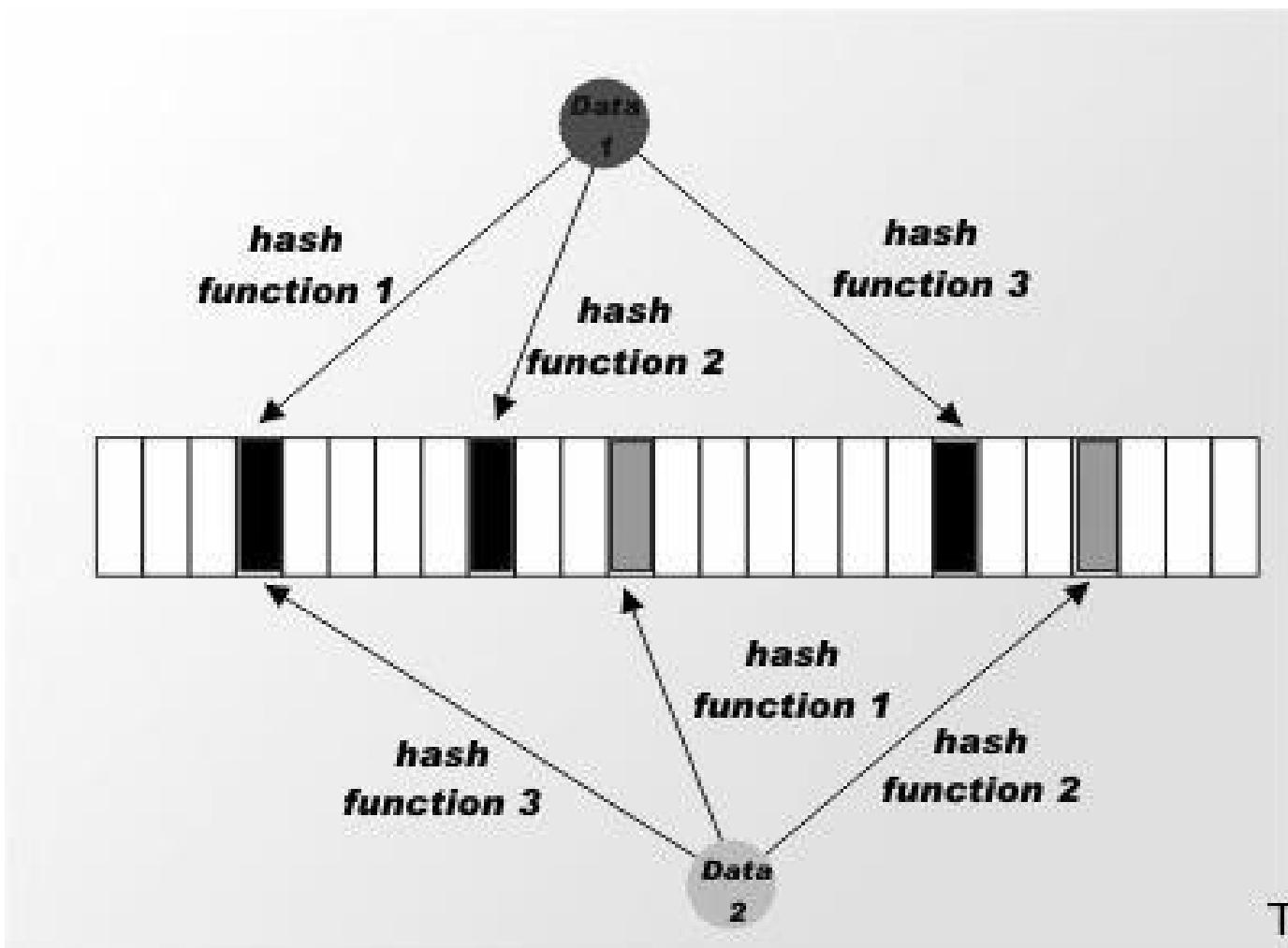
Представление графа

Требования:

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

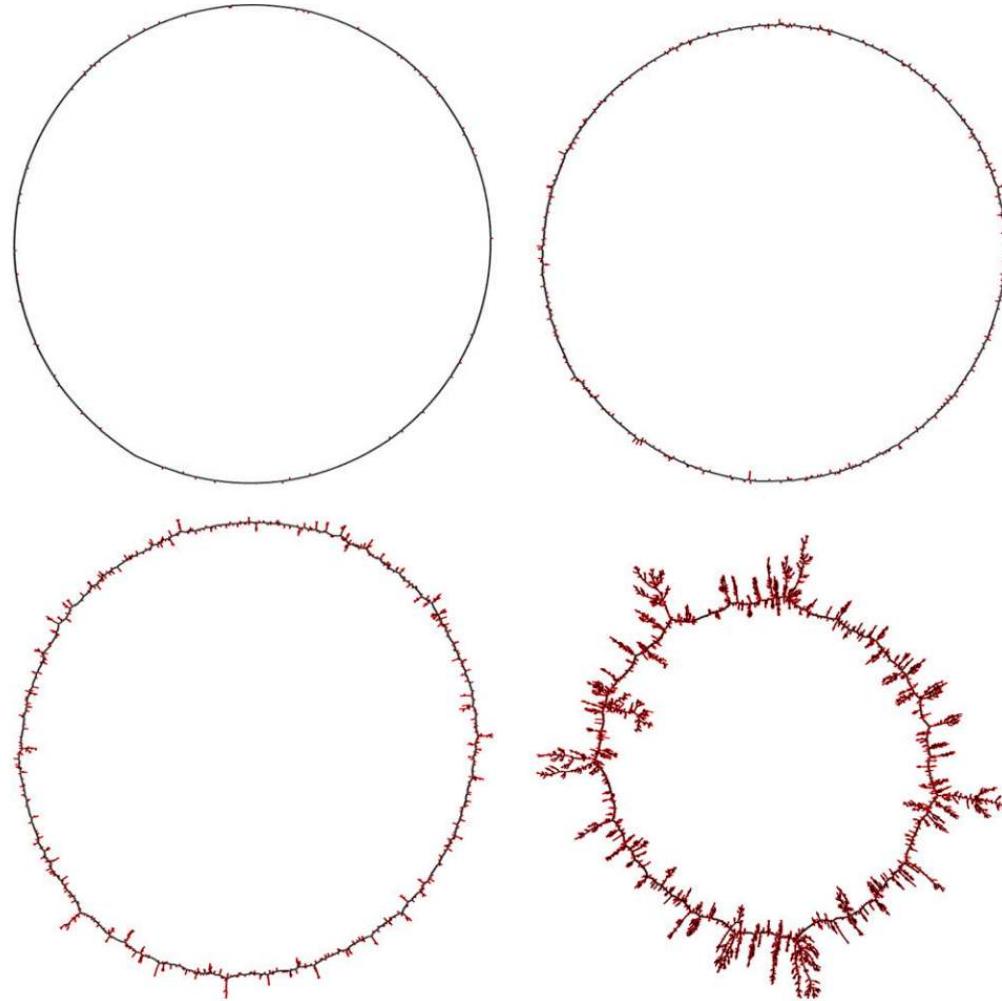
Пример: Множество всех $(k+1)$ -меров

Фильтр Блума

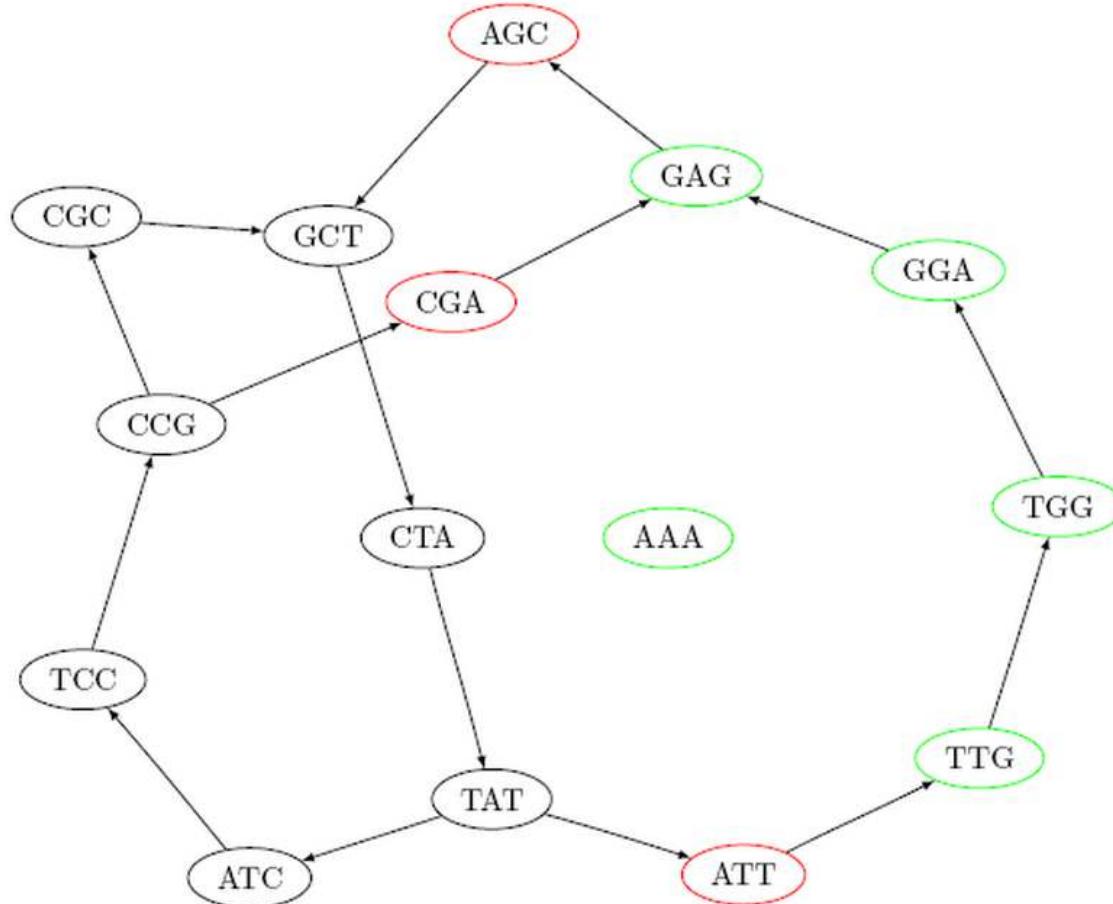


T.

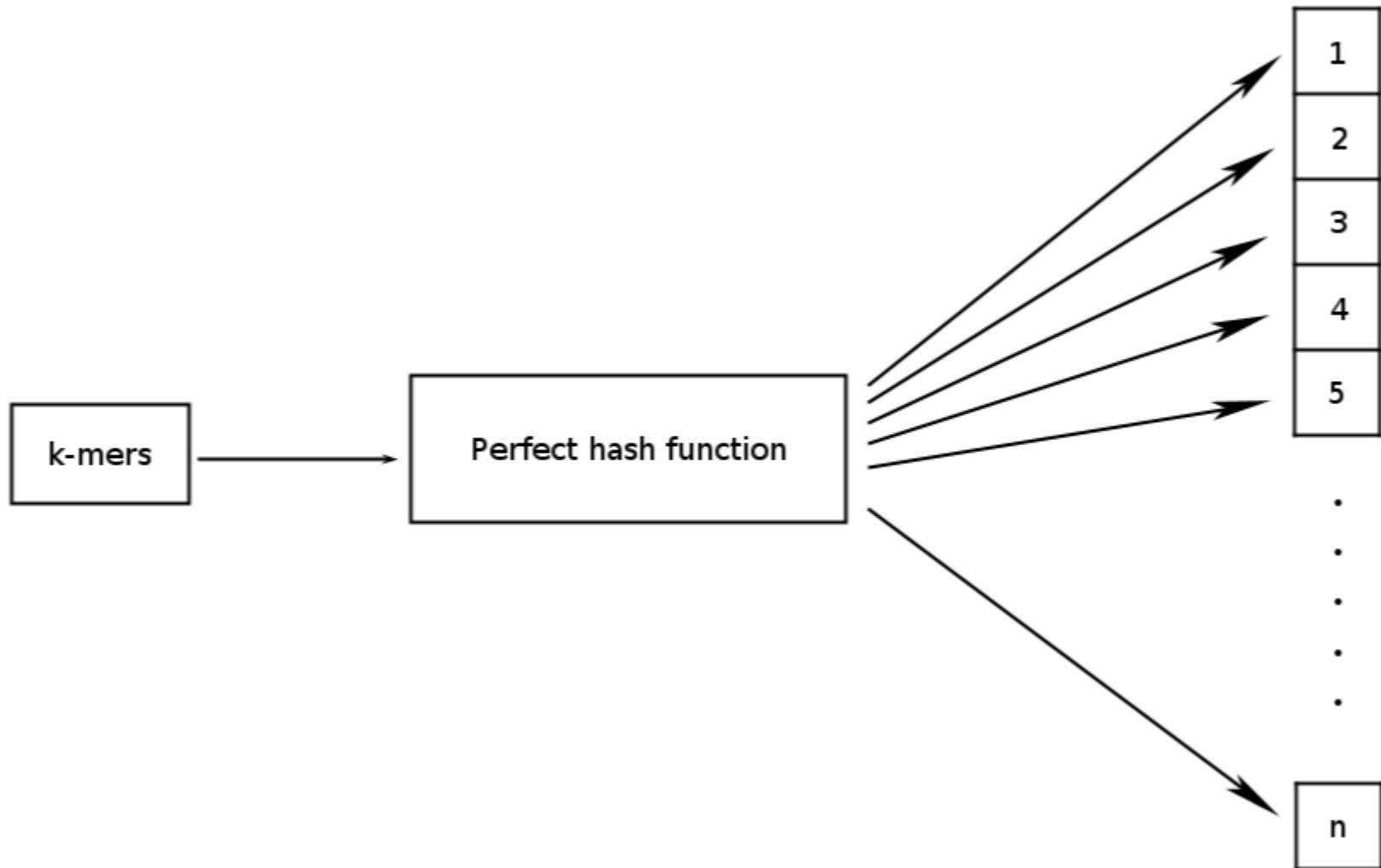
Вероятностный граф де Брюйна



Точное представление



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)

Ссылки

1. "Genome Reconstruction: A Puzzle with a Billion Pieces", P.Compeau, P. Pevzner
2. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing", A. Bankevich et al.
3. "Quake: quality-aware detection and correction of sequencing errors", D. Kelley et al.
4. "BayesHammer: Bayesian clustering for error correction in single-cell sequencing", S.Nikolenko et al.
5. "Scaling metagenome sequence assembly with probabilistic de Bruijn graphs", Jason Pell et al.
6. "Space-efficient and exact de Bruijn graph representation based on a Bloom filter", Rayan Chikhi, Guillaume Rizk
7. "External Perfect Hashing for Very Large Key Sets", Fabiano C. Botelho, Nivio Ziviani
8. "*De novo* assembly and genotyping of variants using colored de Bruijn graphs", Z.Iqubal et al.
9. <http://bioinf.spbau.ru/en/spades>

Вопросы
???