Сборка генома
Часть I

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

Введение

ДНК

ДНК = строка в алфавите {A, C, G, T}

Нуклеотиды, основания, базы (base pair, bp)

Бактерии ~ 3 Mbp (10^6)

Человек ~ 3 Gbp (10^9)

Процесс чтения ДНК — секвенирование.

Первые технологии

Конец 1970-х: Уолтер Гилберт и Фредерик Сэнгер независимо разрабатывают методы секвенирования

1980: Нобелевская премия

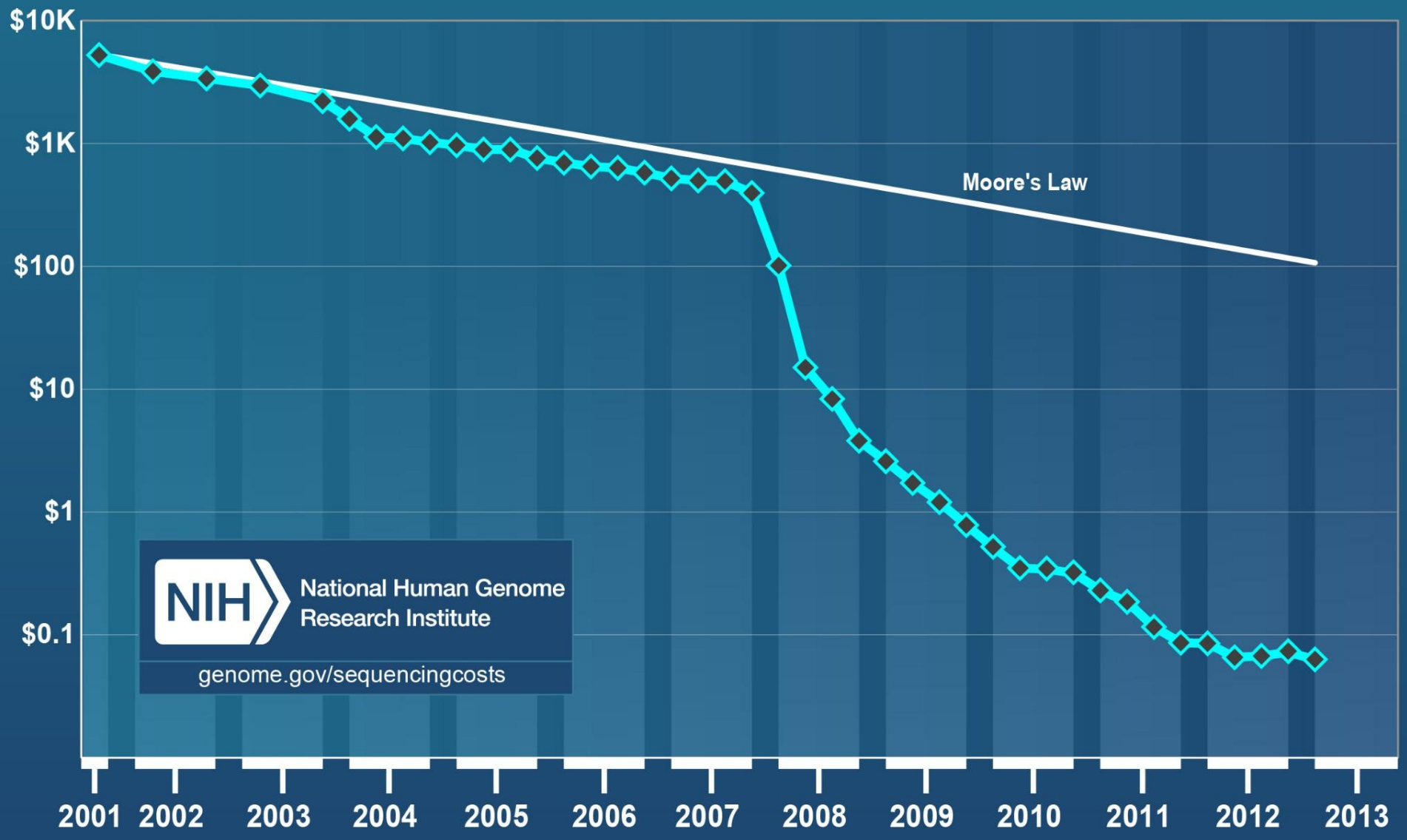
Позволяет прочитать небольшие фрагменты

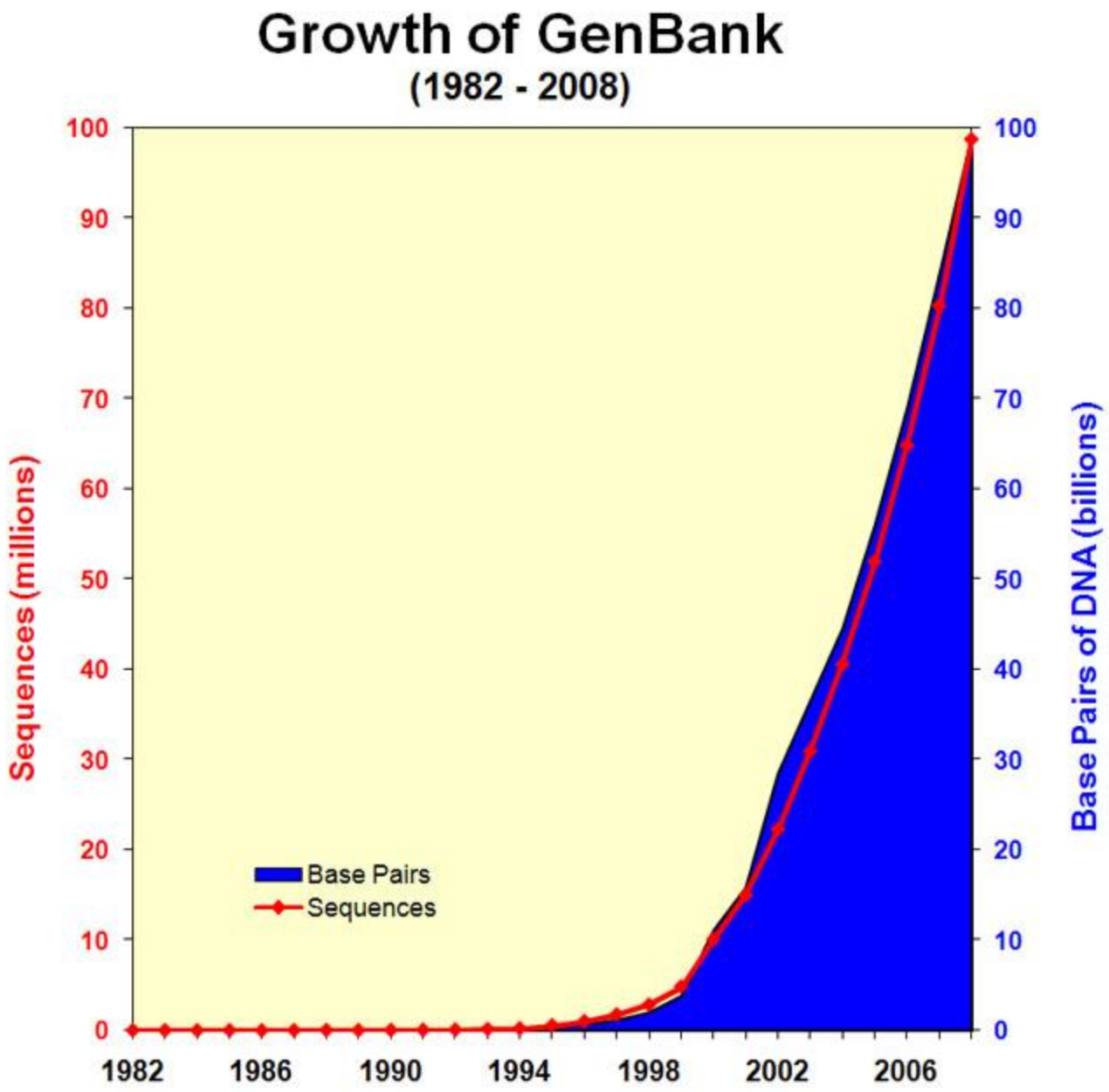
NGS революция

Начало 2000-х: первых NGS технологий

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

Cost per Raw Megabase of DNA Sequence





http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/

Технологии секвенирования

- Sanger

- Illumina
- 454
- Solid
- IonTorrent

- PacBio
- Nanopore

Method	Sanger	Illumina	454	SOLiD	Ion Torrent	PacBio
Read length	400 to 900 bp	50 to 250 bp	700 bp	50+35 or 50+50 bp	200 bp	2900 bp average
Accuracy	99.9%	98%	99.9%	99.9%	98%	87% (read length mode), 99% (accuracy mode)
Reads per run	N/A	up to 3 billion	1 million	1.2 to 1.4 billion	up to 5 million	35–75 thousand
Time per run	20 minutes to 3 hours	1 to 10 days, depending upon specified read length	24 hours	1 to 2 weeks	2 hours	30 minutes to 2 hours
Cost per 1 million bases (in US$)	$2400	$0.05 to $0.15	$10	$0.13	$1	$2

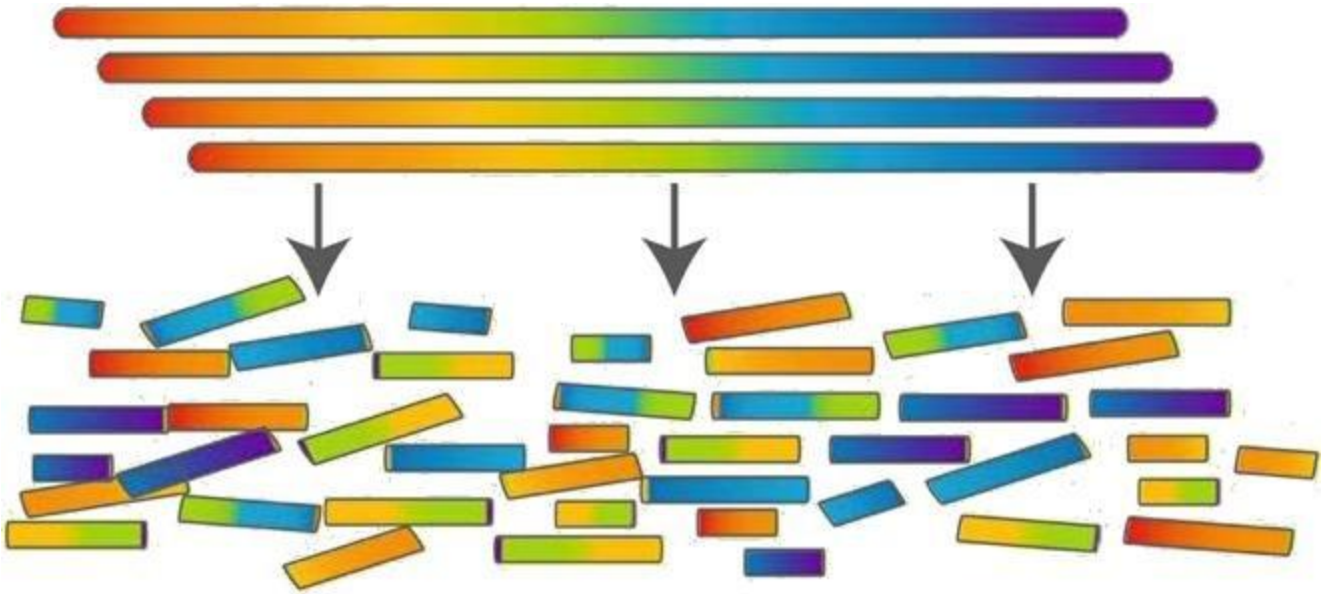
http://en.wikipedia.org/wiki/Next-generation_sequencing

Виды анализа

1. *De novo* сборка
2. Основанные на прикладывании

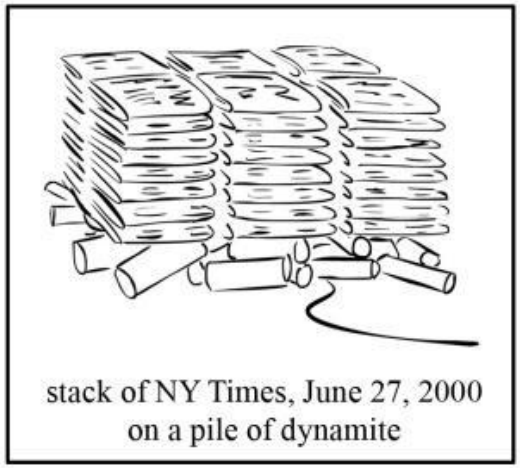
Сборка

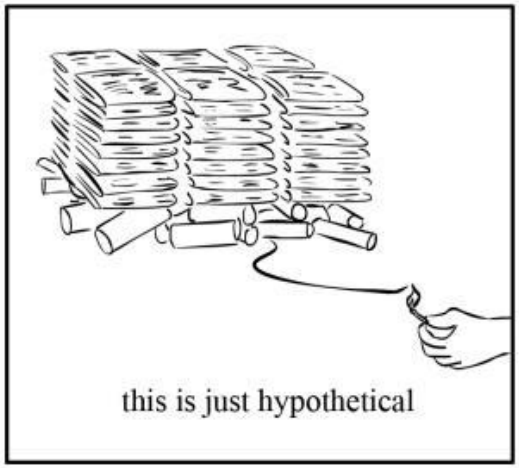
Whole genome shotgun sequencing

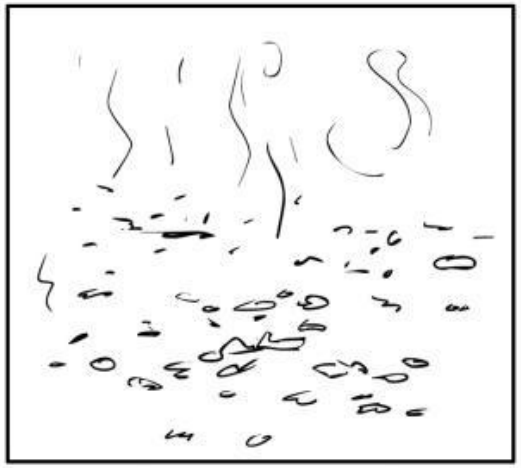


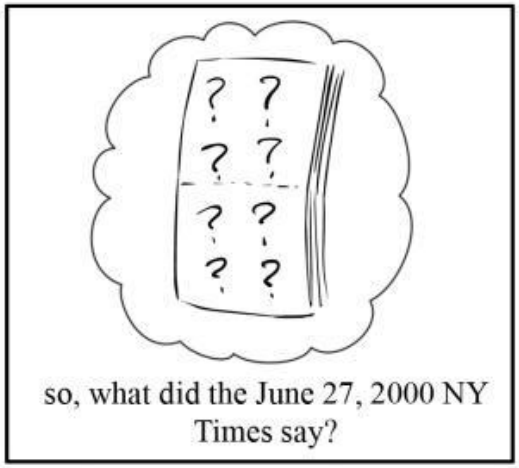
Сборка (assembly) -- восстановление участков изначальной последовательности

Задача сборки









SSP

Дано: множество строк $\mathbf{S_i}$

Найти: кратчайшую строку **S**, содержащую все $\mathbf{S_i}$

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

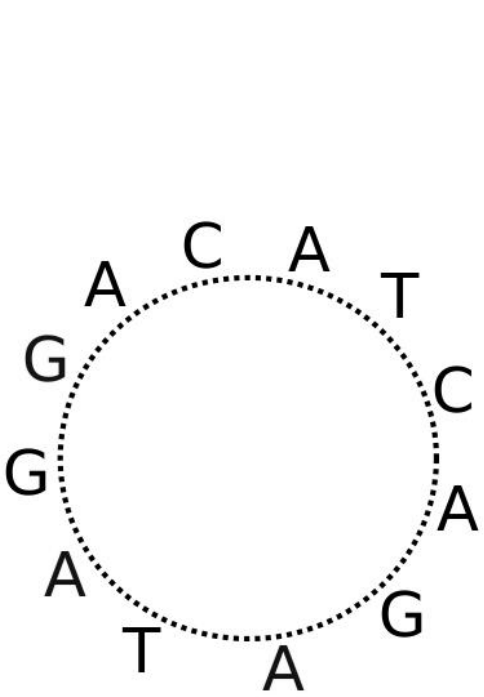
- Velvet
- IDBA
- SOAP-denovo
- Ray
- ABySS
- Allpaths
- EULER
- Minia
- SPAdes

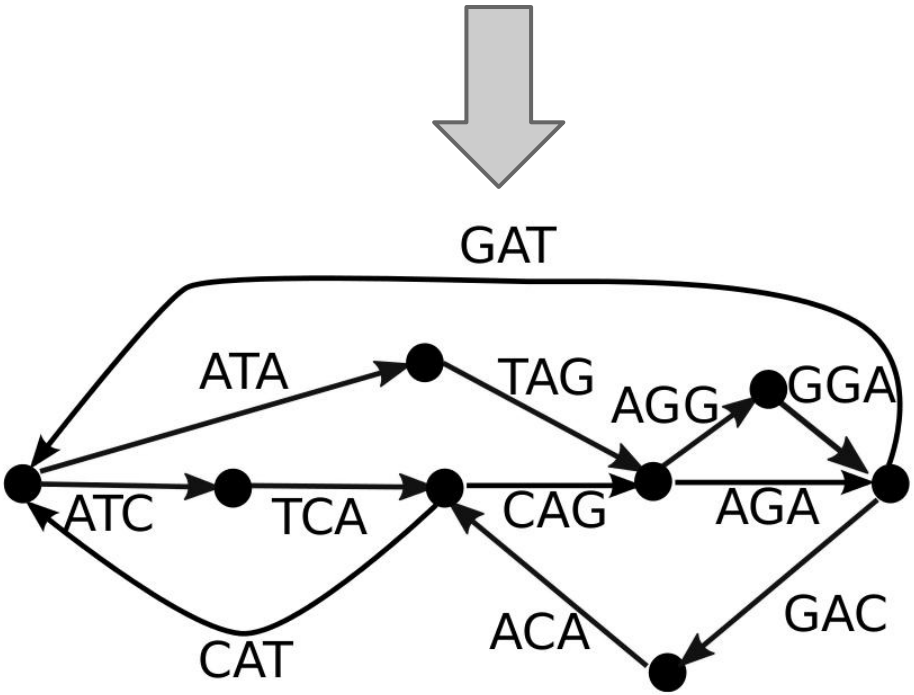
Графы де Брюйна

- 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*

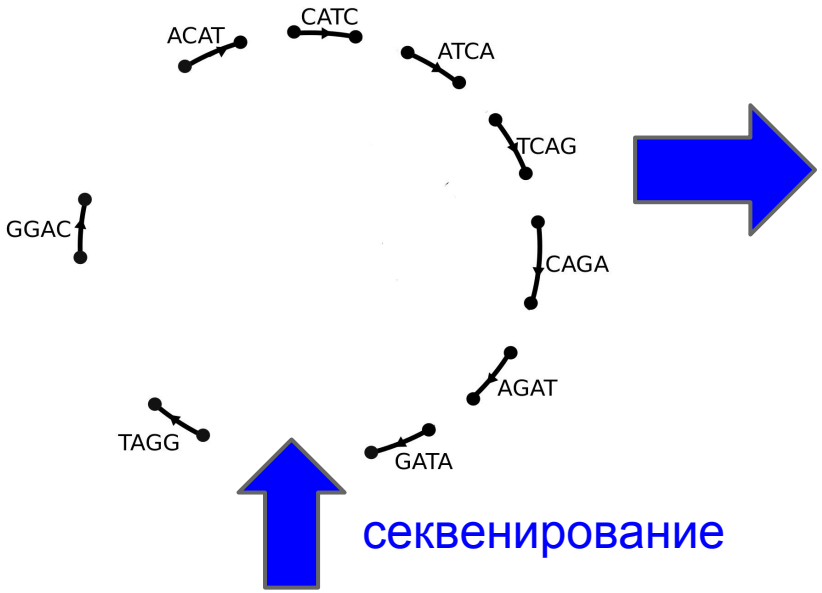
Графы де Брюйна

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



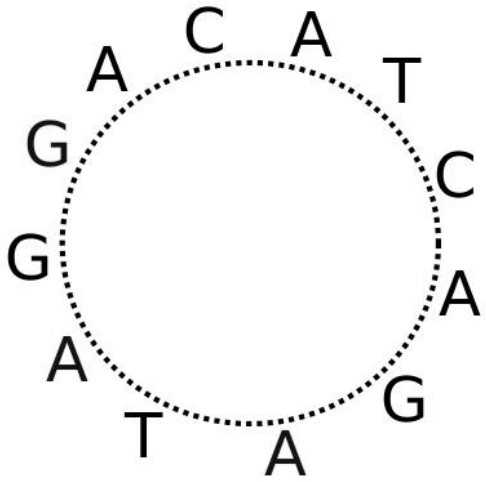


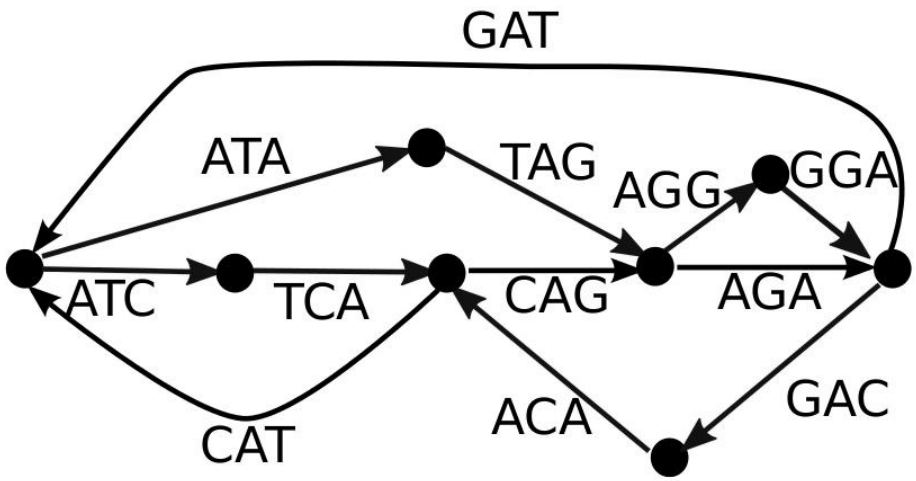
Графы де Брюйна



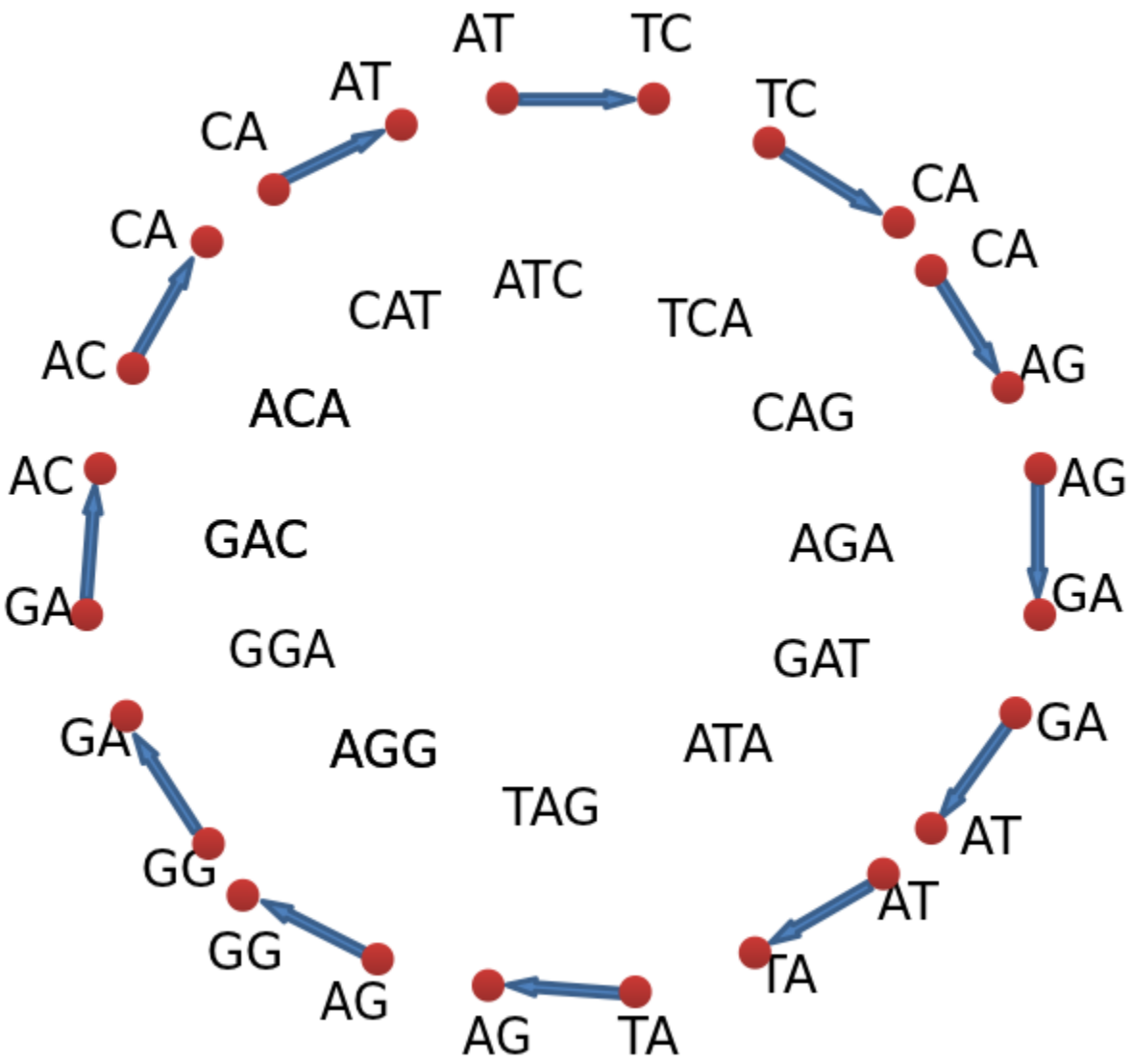
секвенирование

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

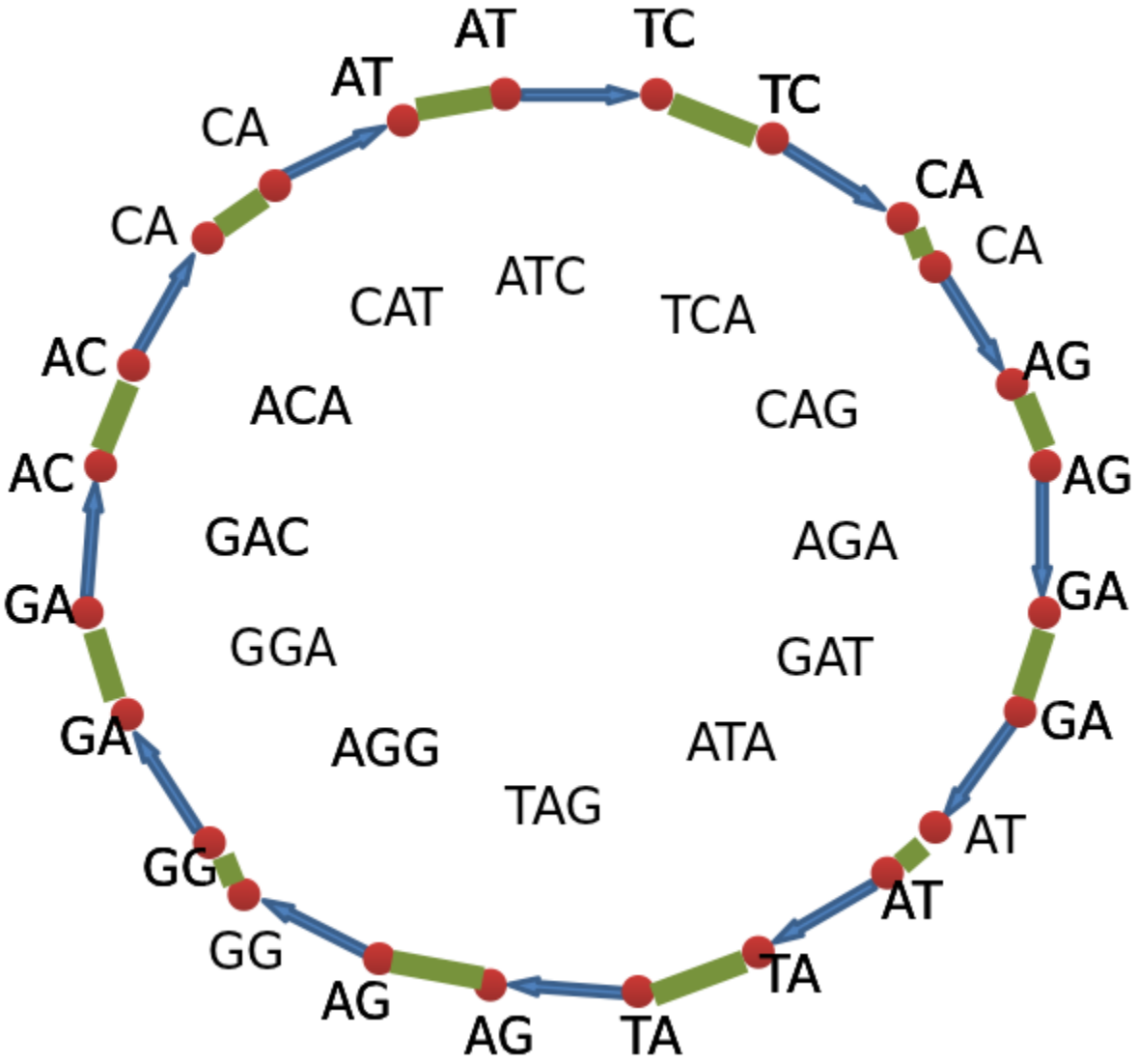




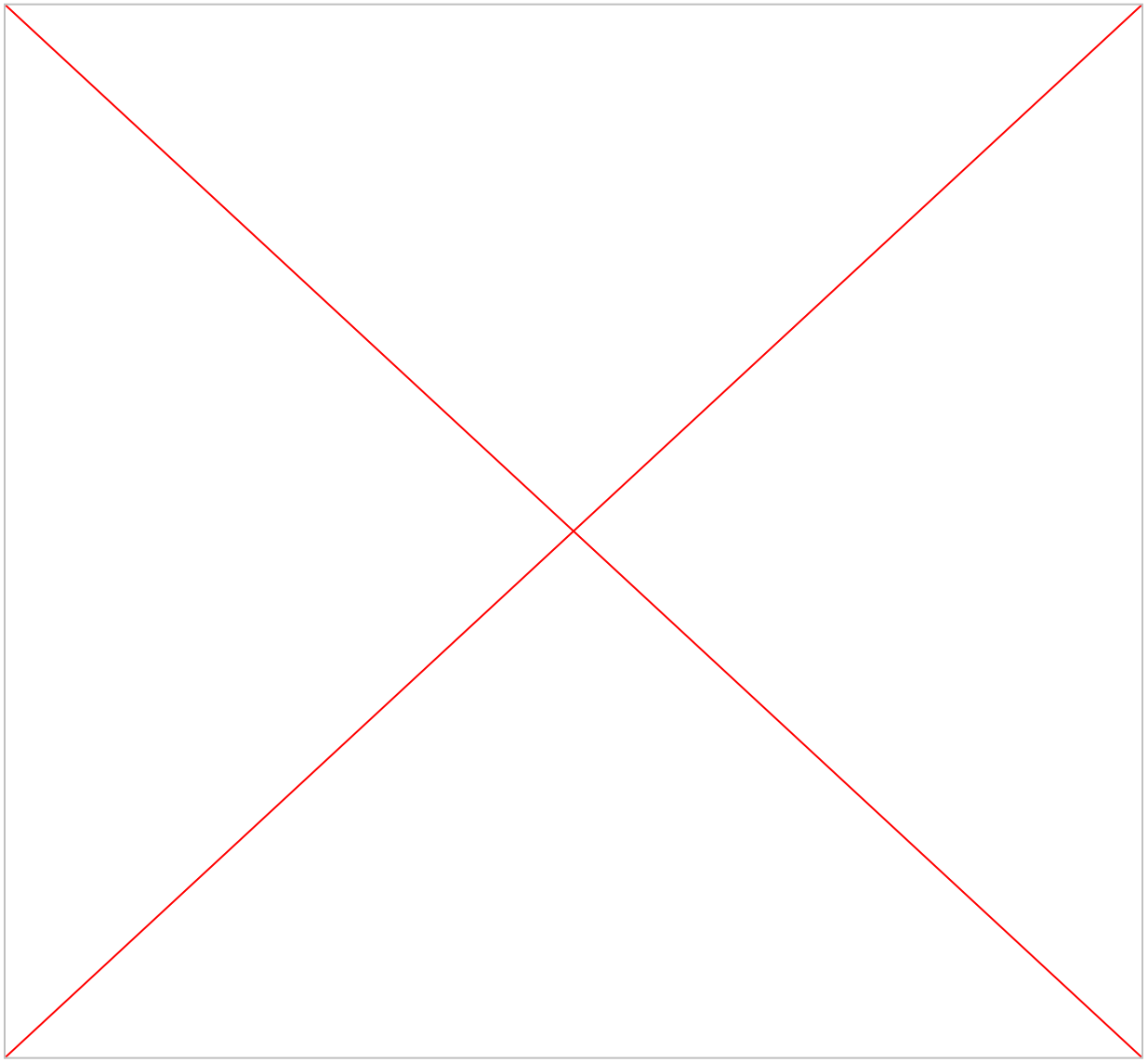
Графы де Брюйна



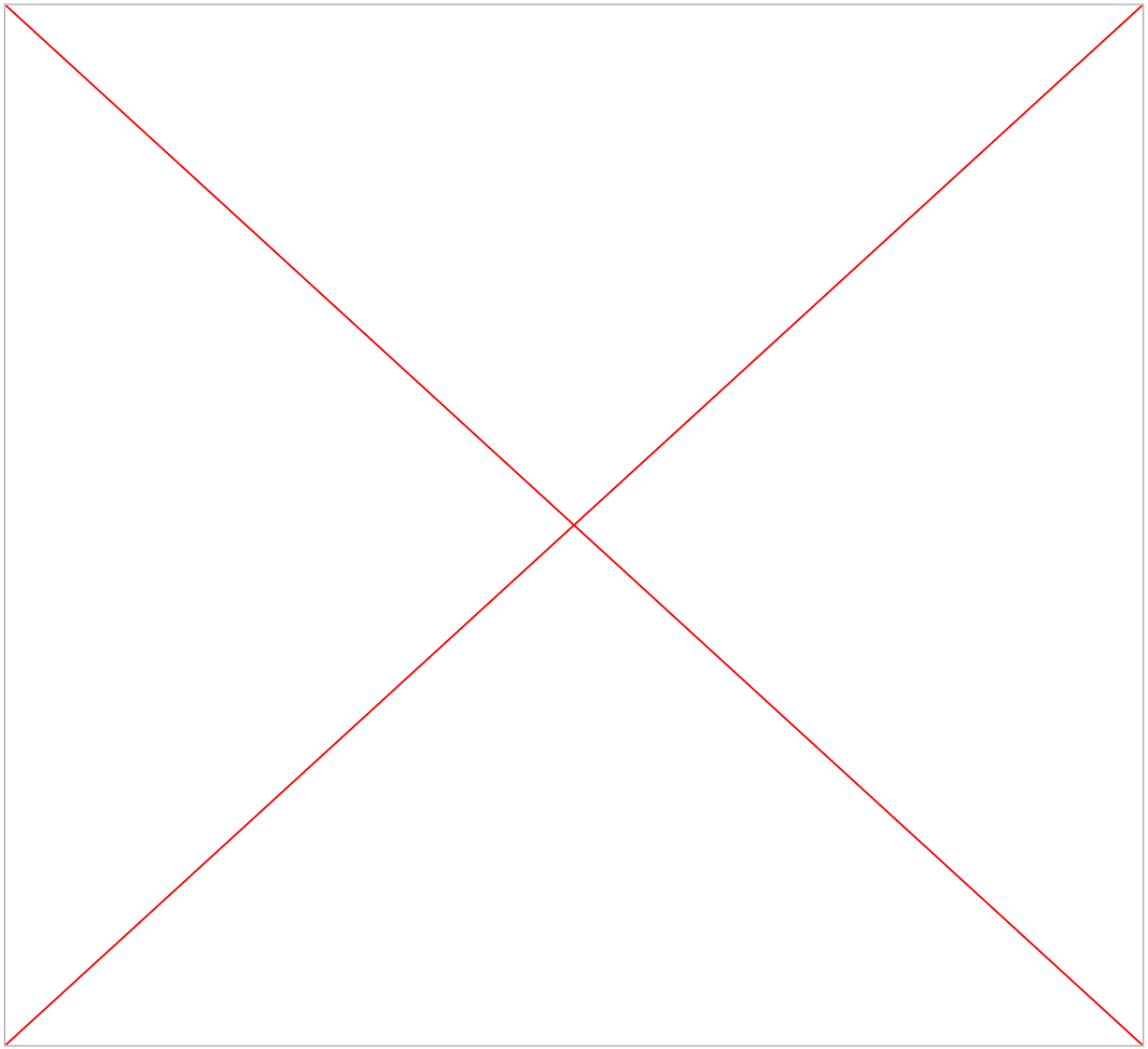
Графы де Брюйна



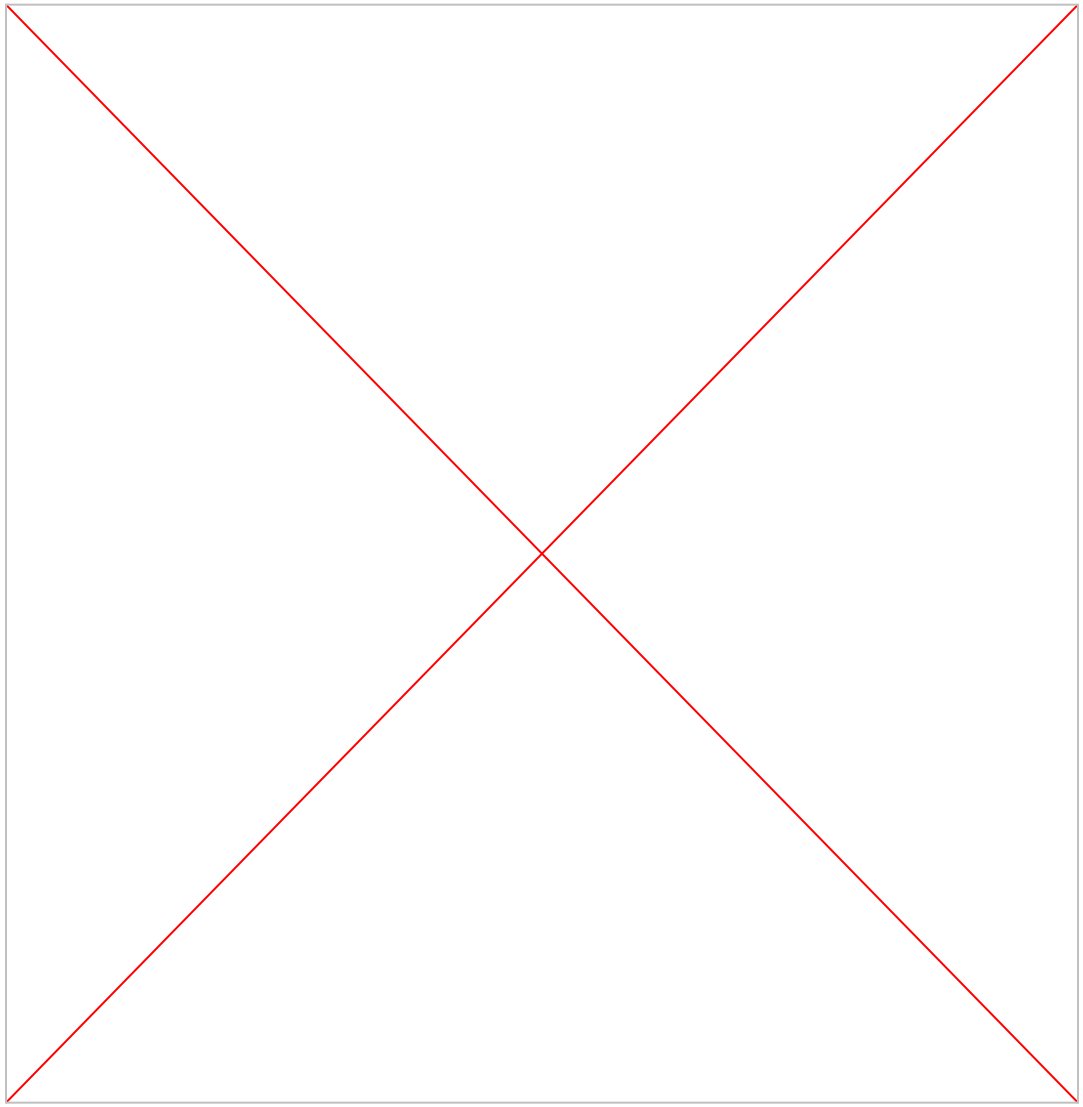
Графы де Брюйна



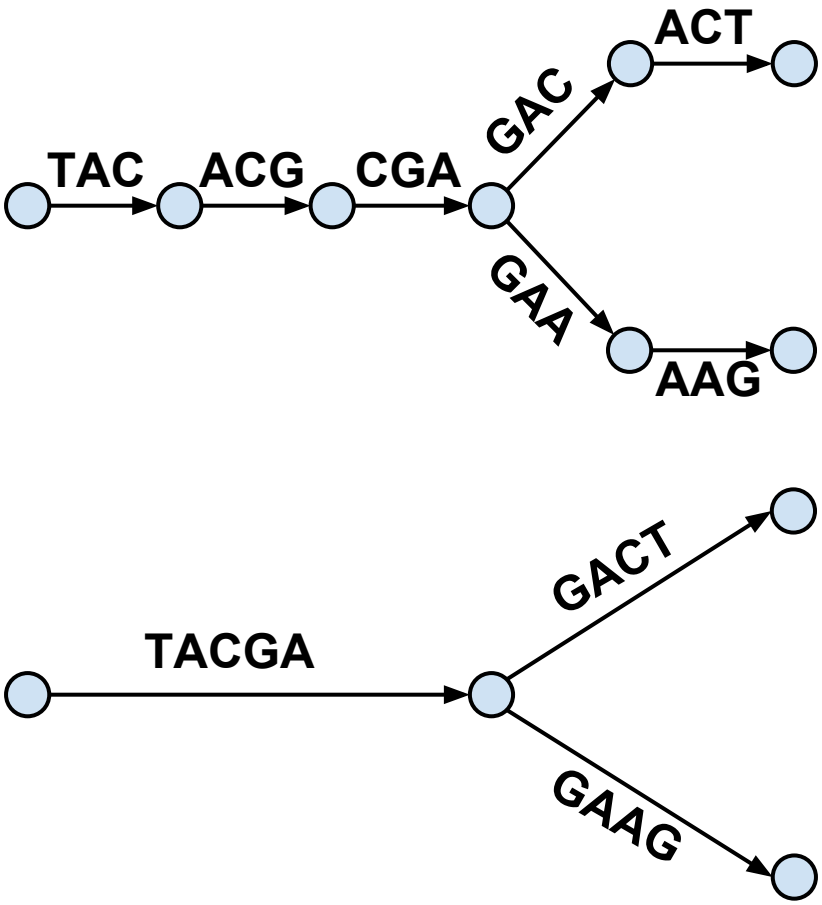
Графы де Брюйна



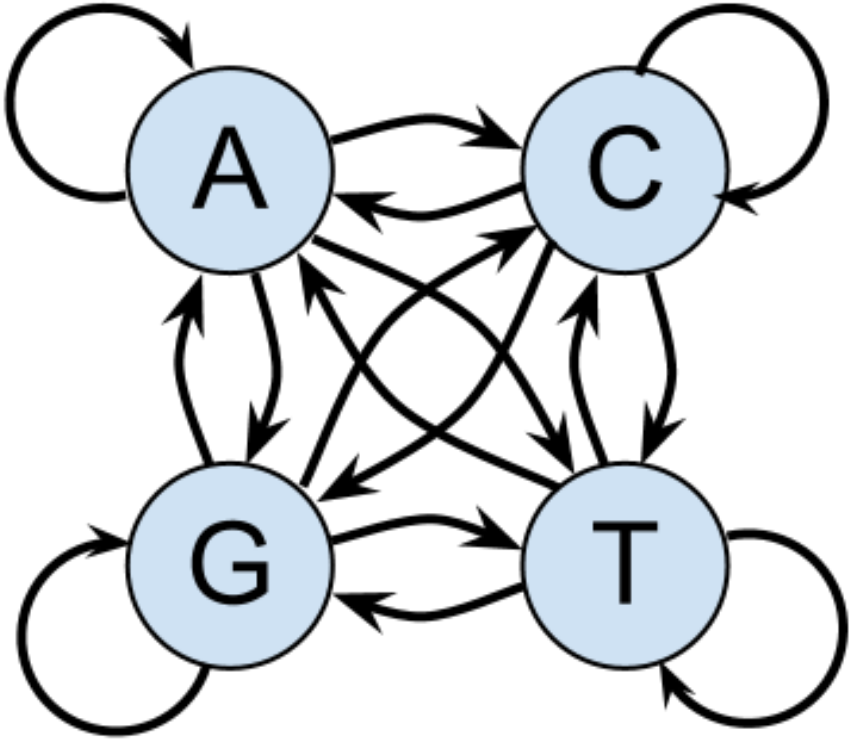
Графы де Брюйна



Сжатый граф



K имеет значение!

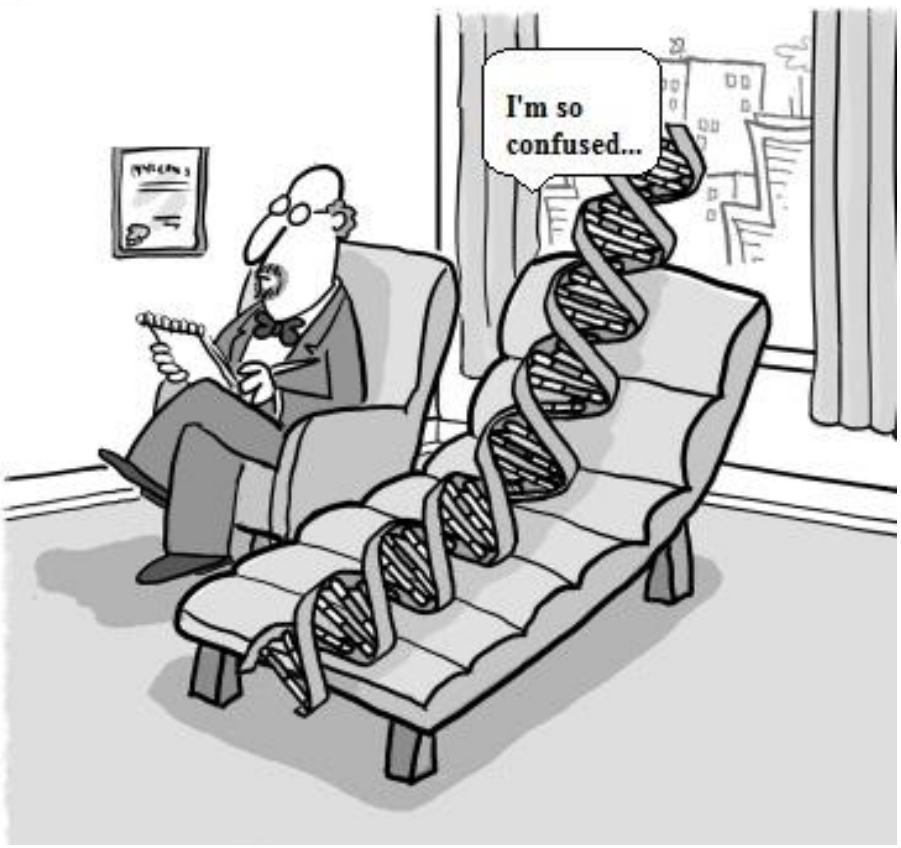


Проблема повторов

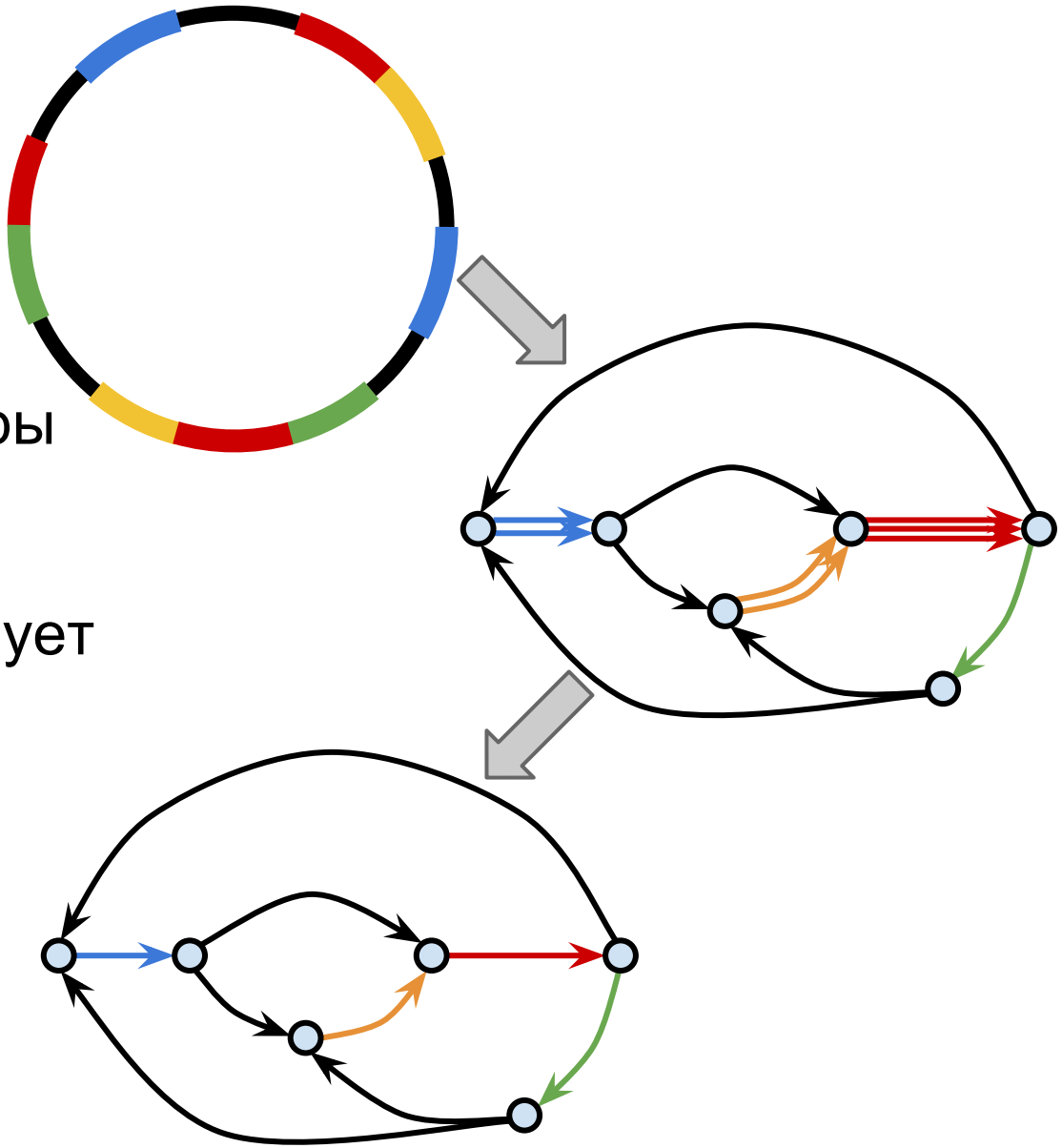
ALU

длина: 300

кратность: 1000000

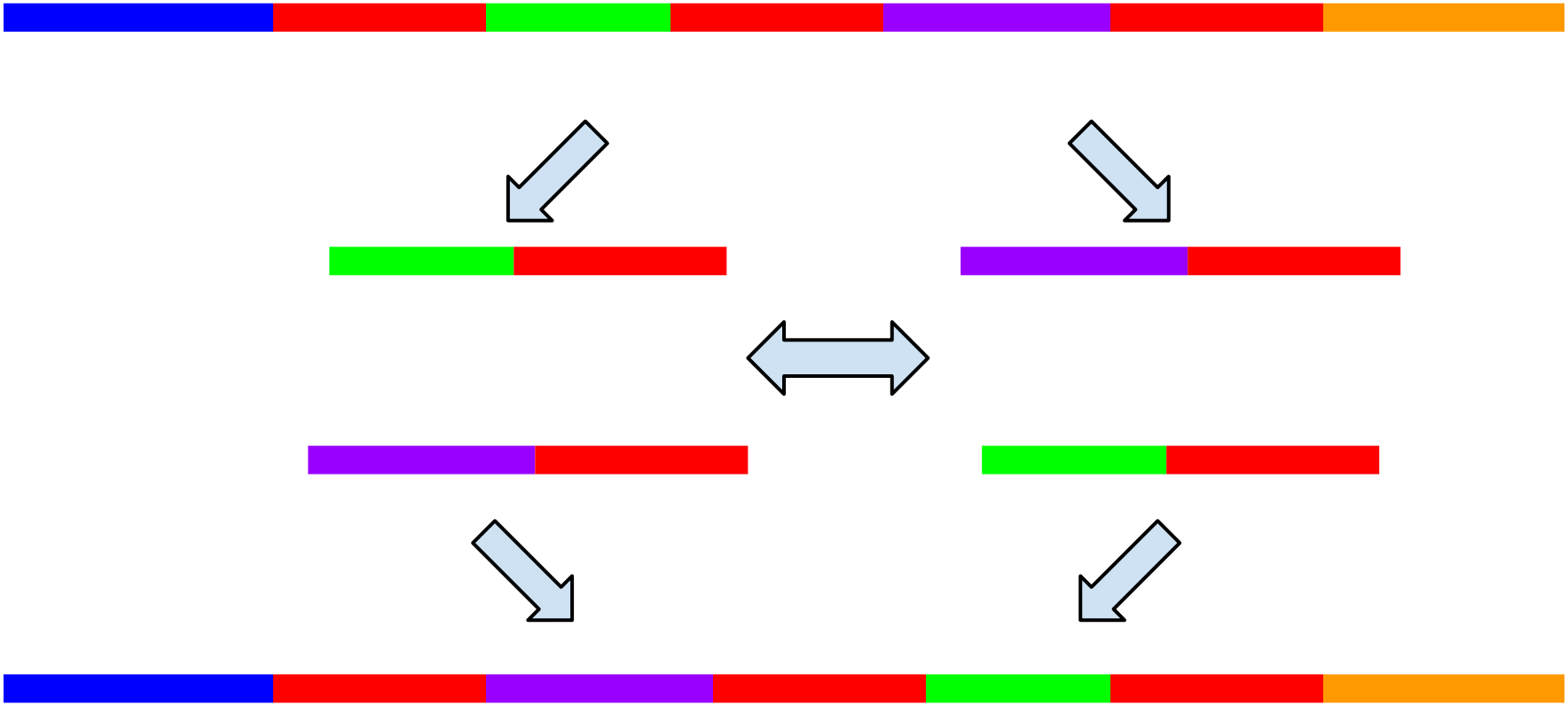


Заметки про граф де Брюйна

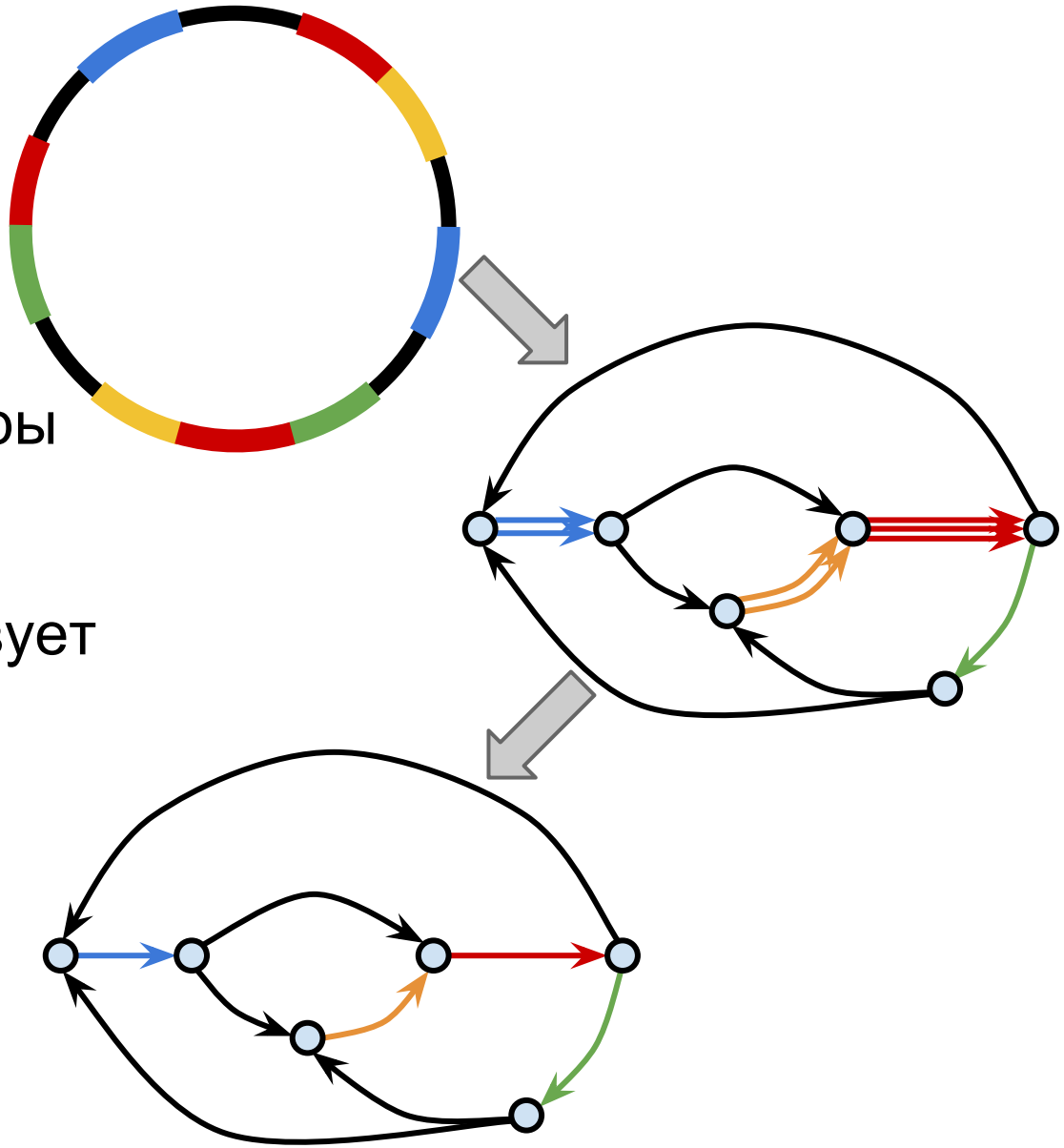


1. Склеивает повторы (длиннее **k**)
2. Геном соответствует циклу в графе

Проблема повторов



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 проблемы

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

Разрывы в покрытии

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

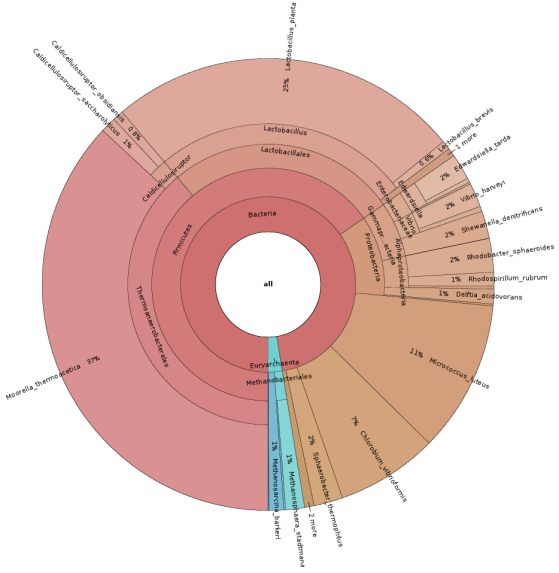
Чтобы снизить вероятность разрыва, приходится использовать *k* значительно меньше длины рида

Разрывы в покрытии

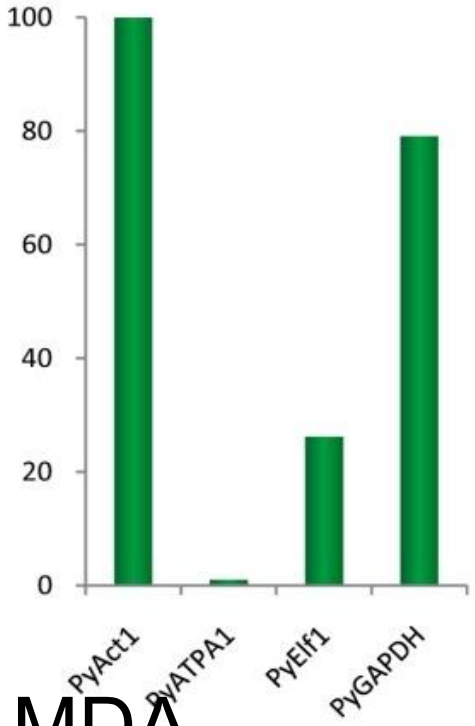
- Длина генома: L
- Количество ридов: n
- Эффективная длина рида: l-k
- Покрытие генома: C = n (l-k) / L
- Вопрос: Какое покрытие необходимо?
  - Модель Лэндера-Уотермана: В предположении о равномерном распределении ридов и покрытии C = 10, на каждый миллион нуклеотидов генома приходится 1 разрыв покрытия

Неравномерное покрытие

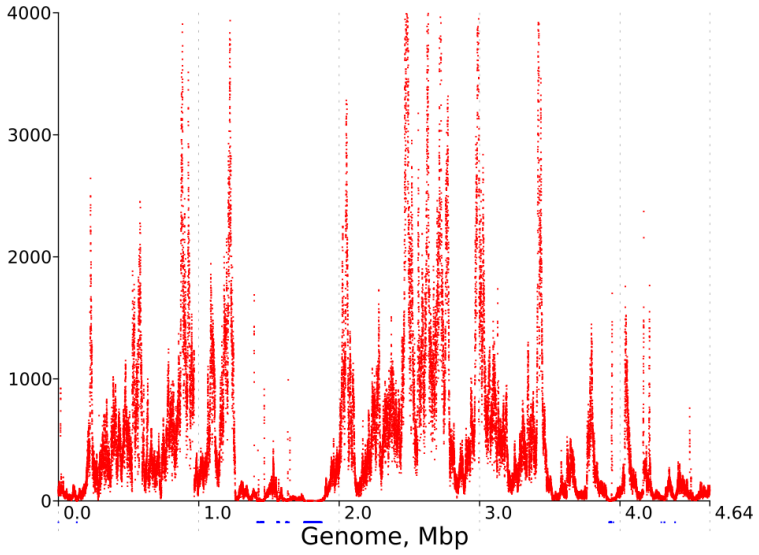
1. Метагеномные данные



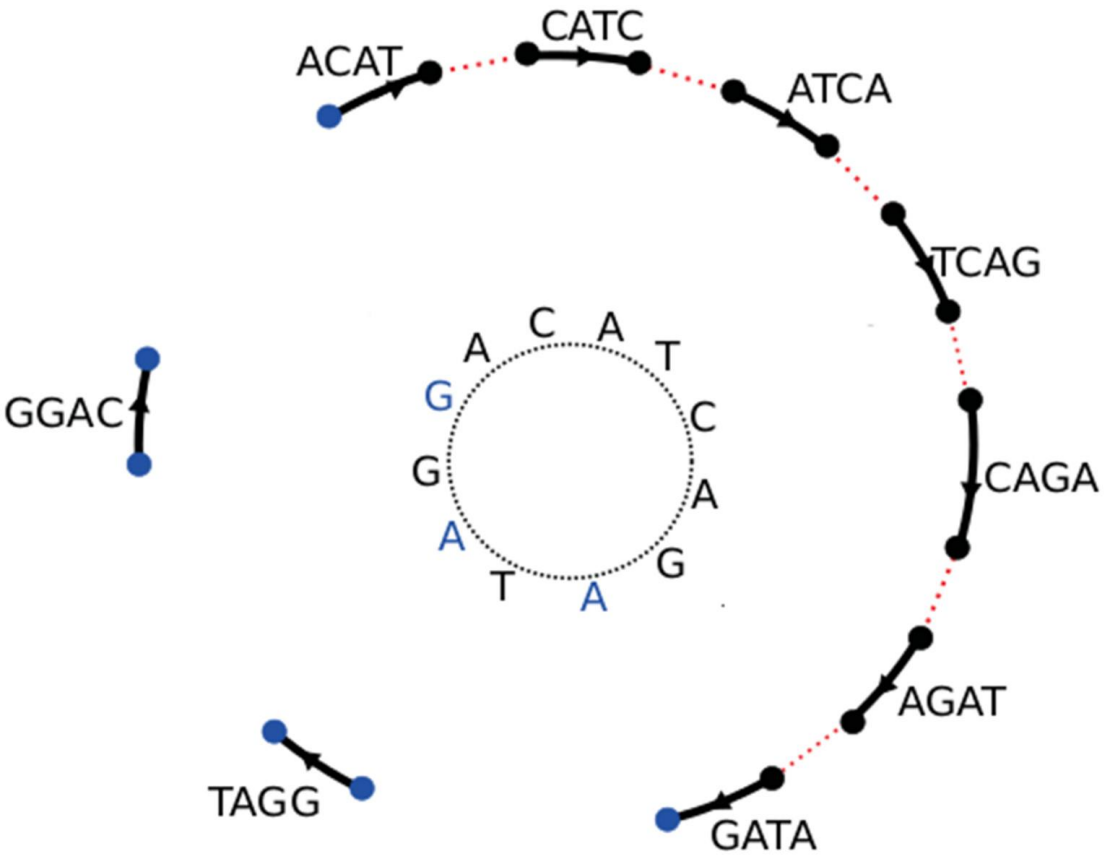
2. RNA-seq

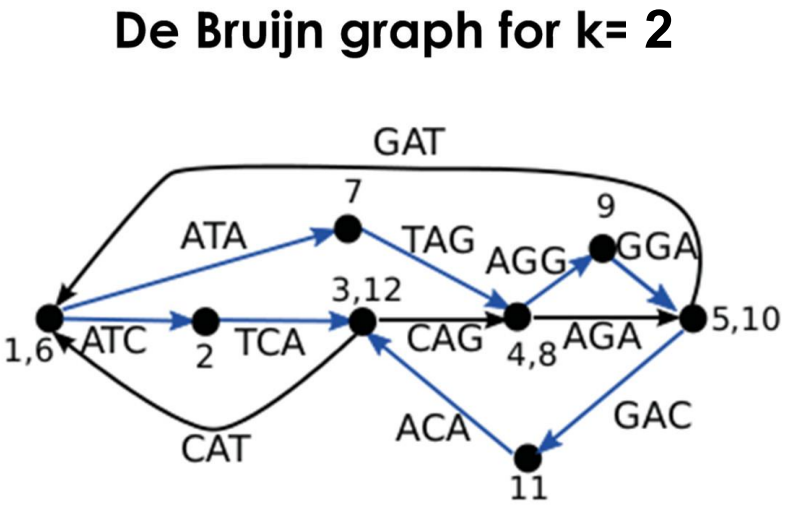


3. Single-cell MDA

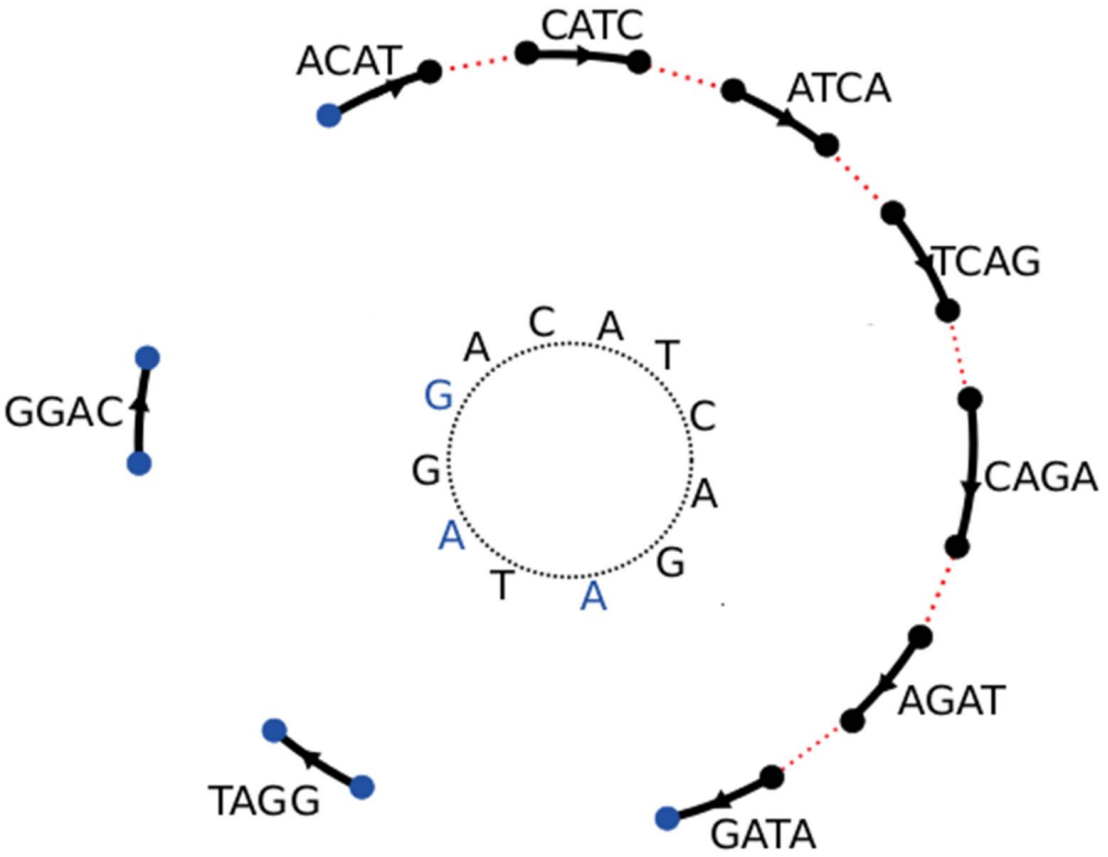


Борьба с разрывами

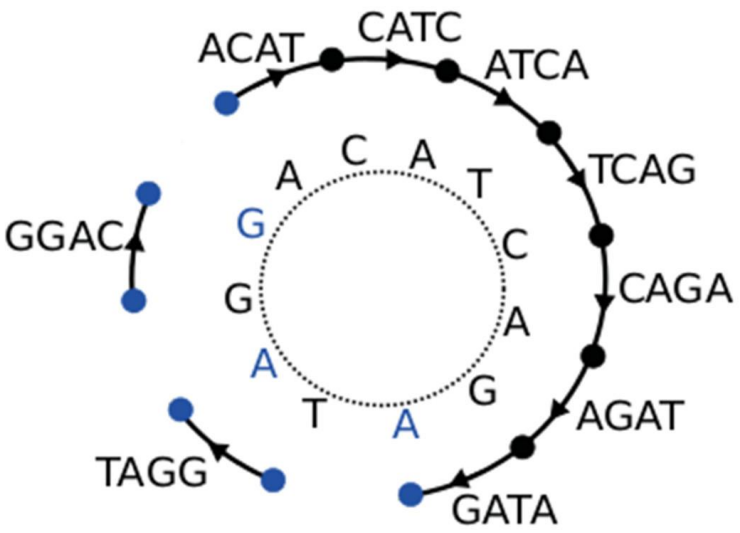


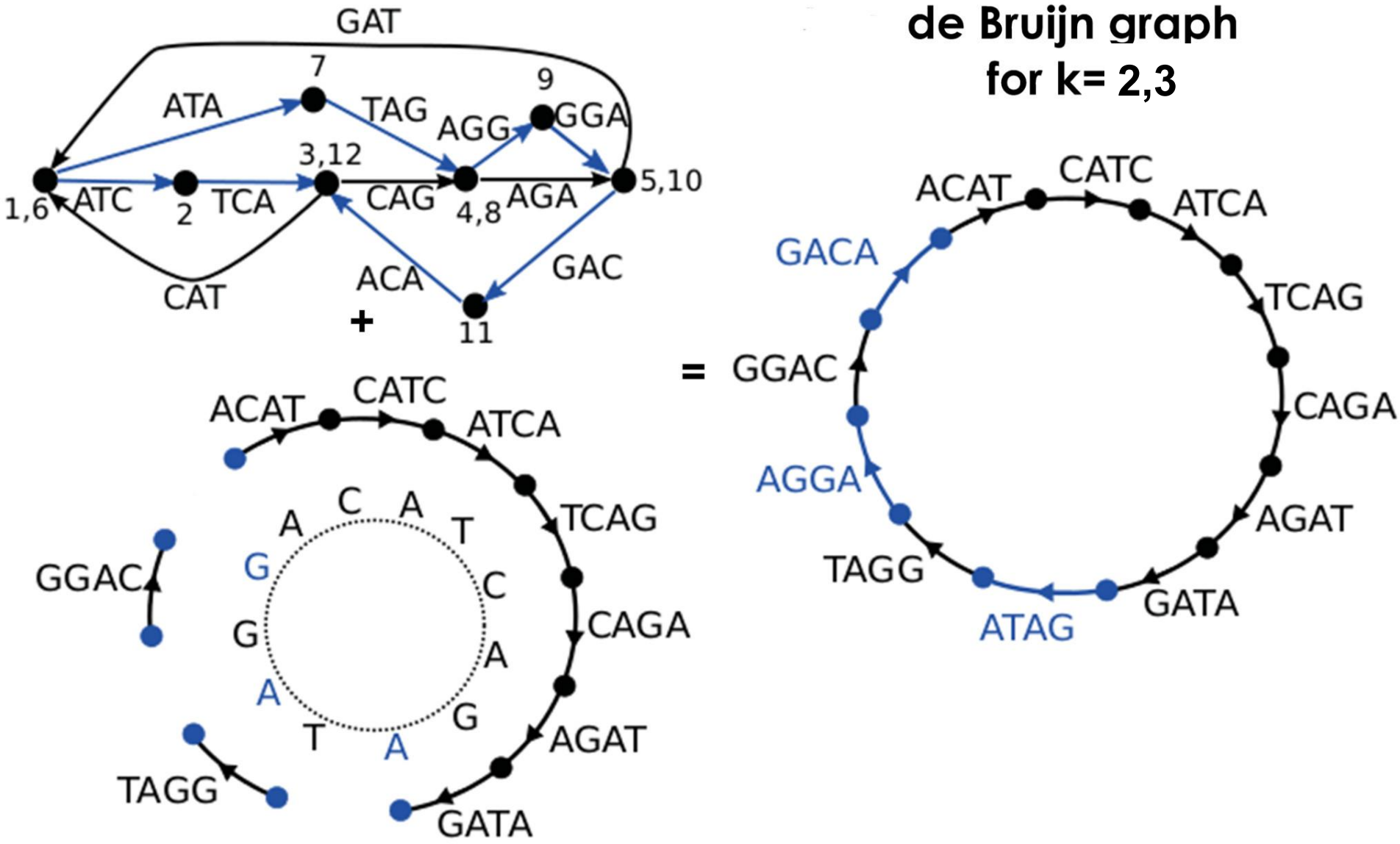


Борьба с разрывами



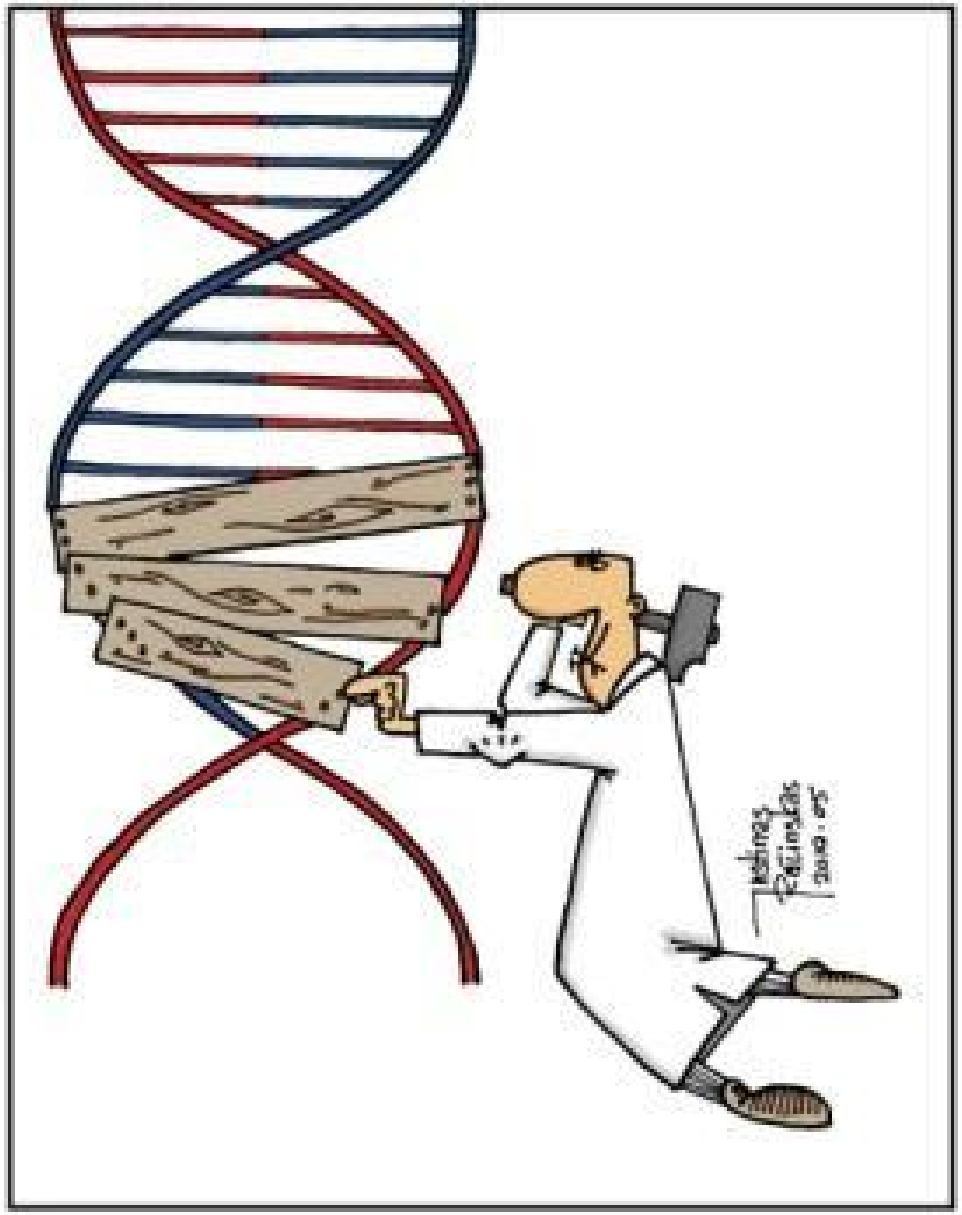
De Bruijn graph for k= 3




de Bruijn graph
for k= 2,3
GAT
7
9
ATA
TAG
AGG
GGA
3,12
5,10
1,6
ATC
2
TCA
CAG
4,8
AGA
CAT
ACA
GAC
11
+
=
ACAT
CATC
ATCA
TCAG
CAGA
AGAT
GATA
TAGG
ATAG
AGGA
GACA
GGAC
A
C
A
T
G
C
G
A
A
T
A
G

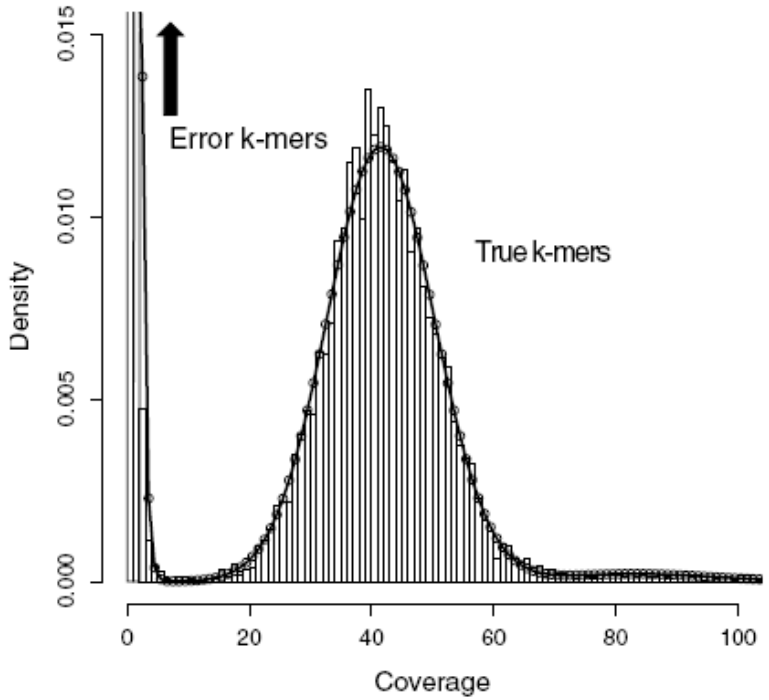
Ошибки секвенирования

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

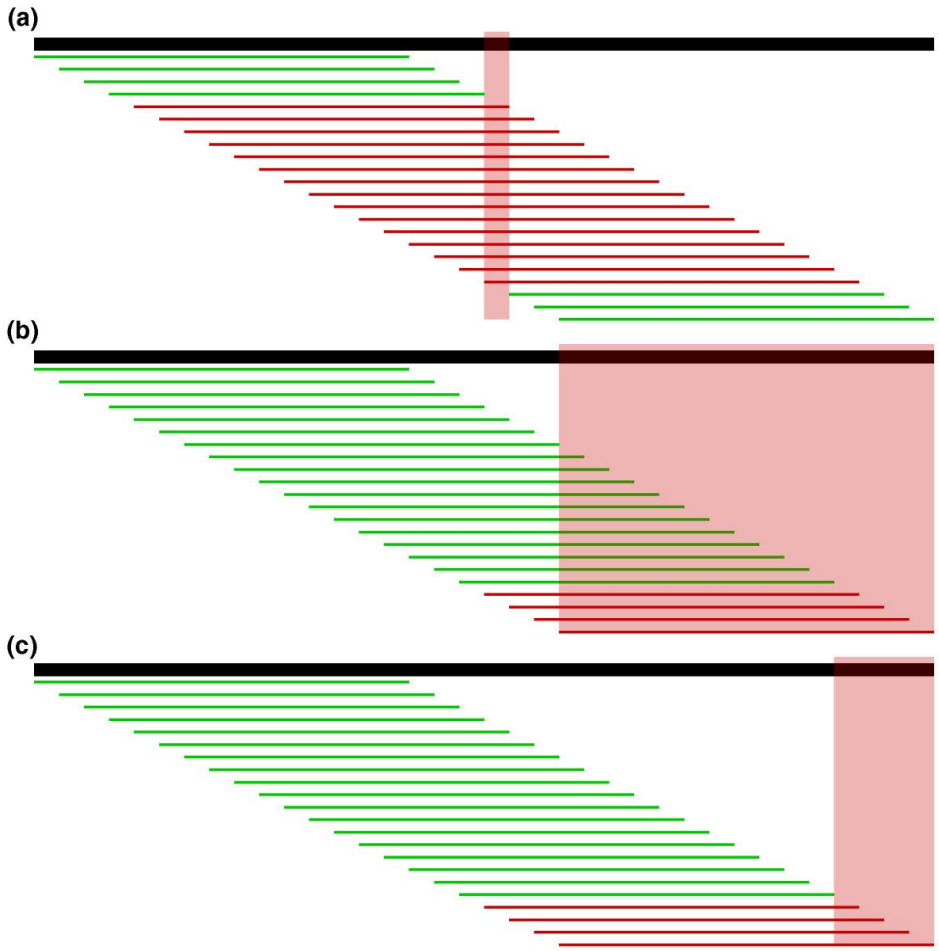


Quake. Надежные k-меры

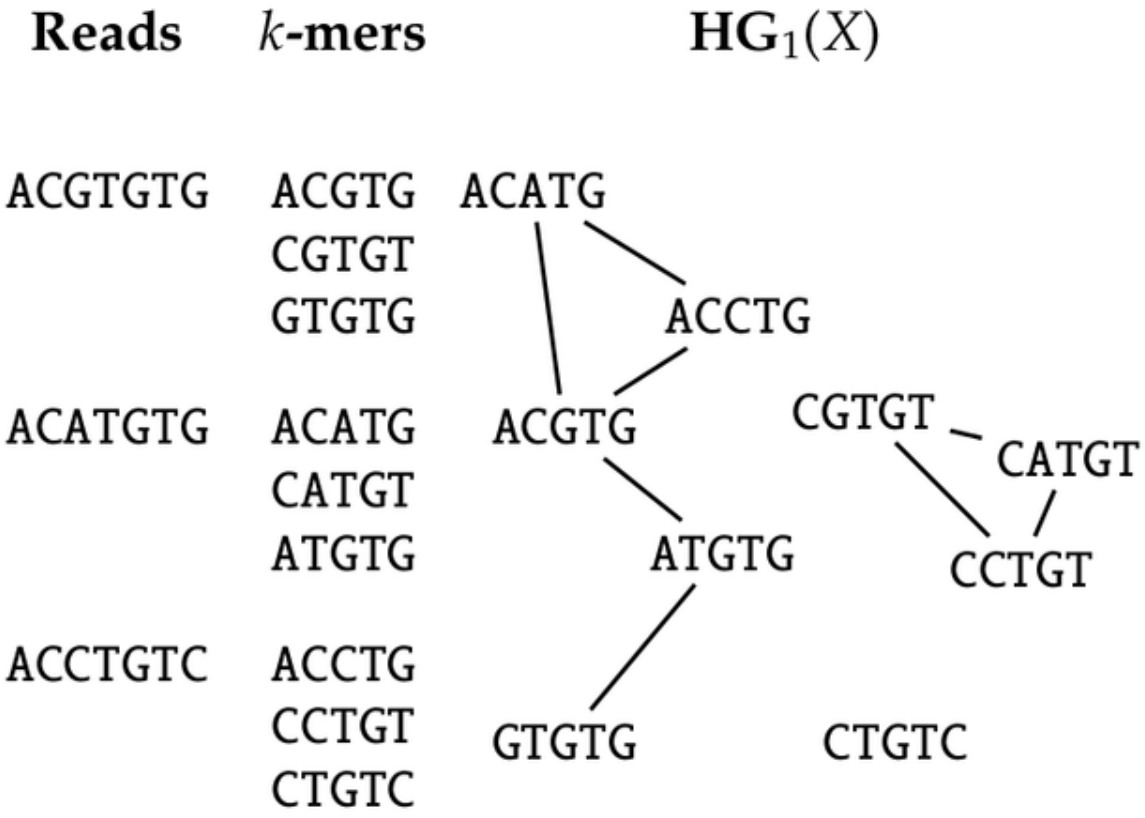
- "Хорошо" покрытые k-меры объявляются надёжными.
- Отсечка определяется исходя из распределения покрытия.



Quake. Коррекция ридов



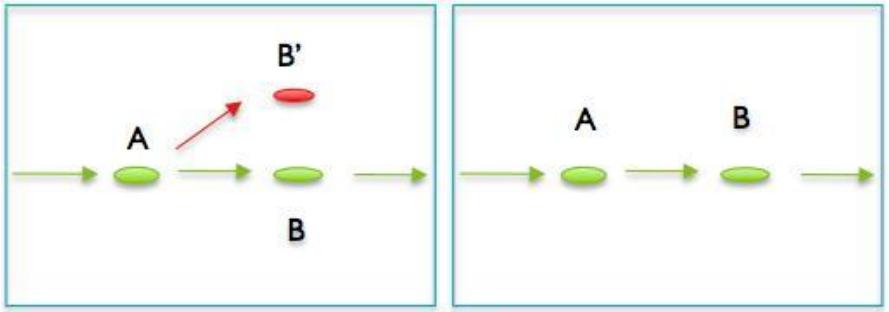
Hammer


Reads
k-mers
HG1(X)
ACGTGTG
ACGTG
CGTGT
GTGTG
ACATGTG
ACATG
CATGT
ATGTG
ACCTGTC
ACCTG
CCTGT
CTGTC
ACATG
ACCTG
ACGTG
CGTGT
CATGT
ATGTG
CCTGT
GTGTG
CTGTC

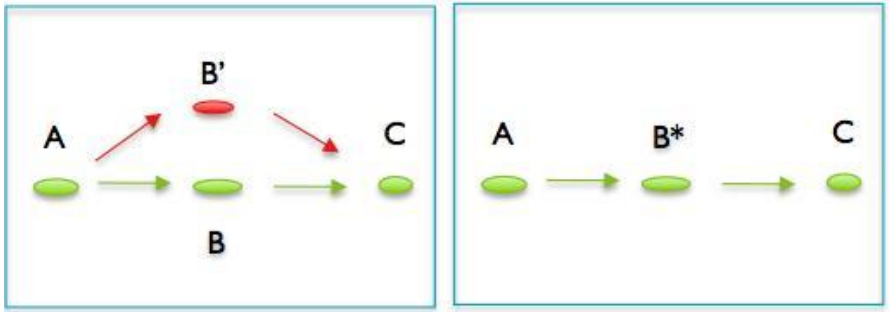
Ошибки в графе

Неисправленные ошибки превращаются в "лишние" ребра в графе

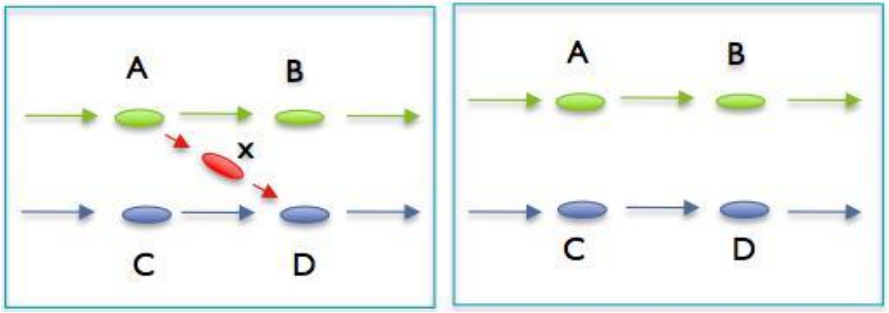
tip



bulge



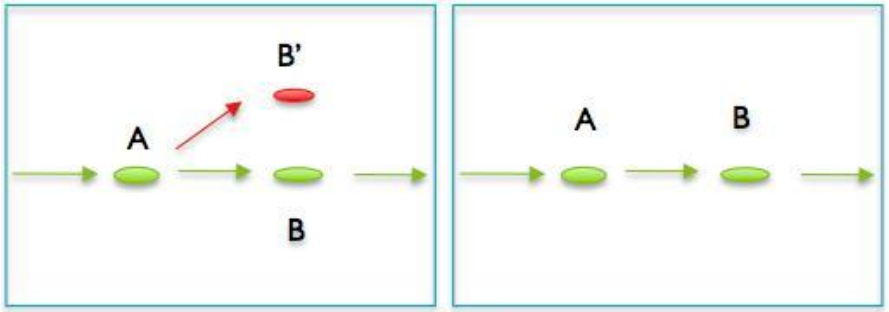
chimeric connection



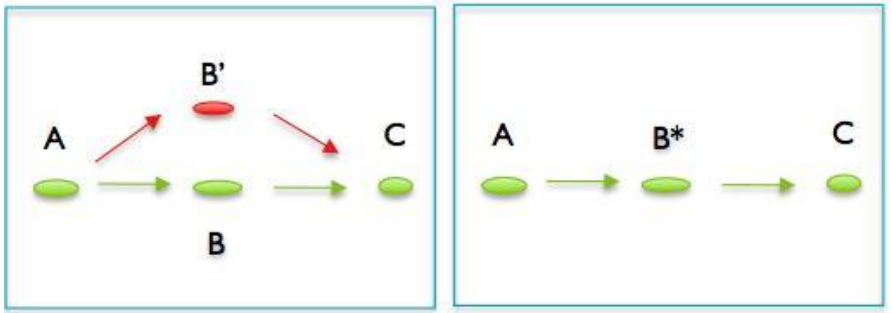
Ошибки в графе

Неисправленные ошибки превращаются в "лишние" ребра в графе

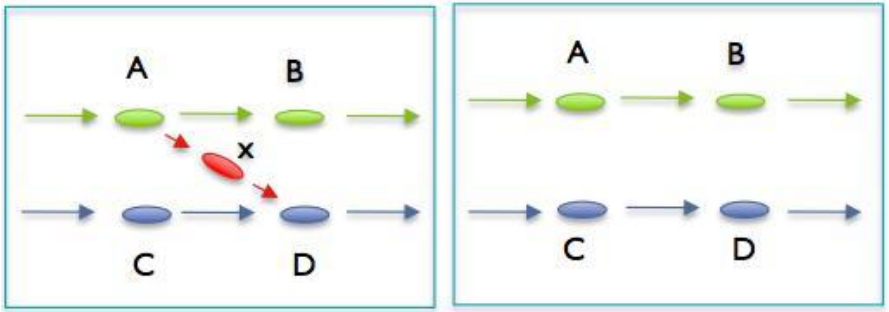
кончики

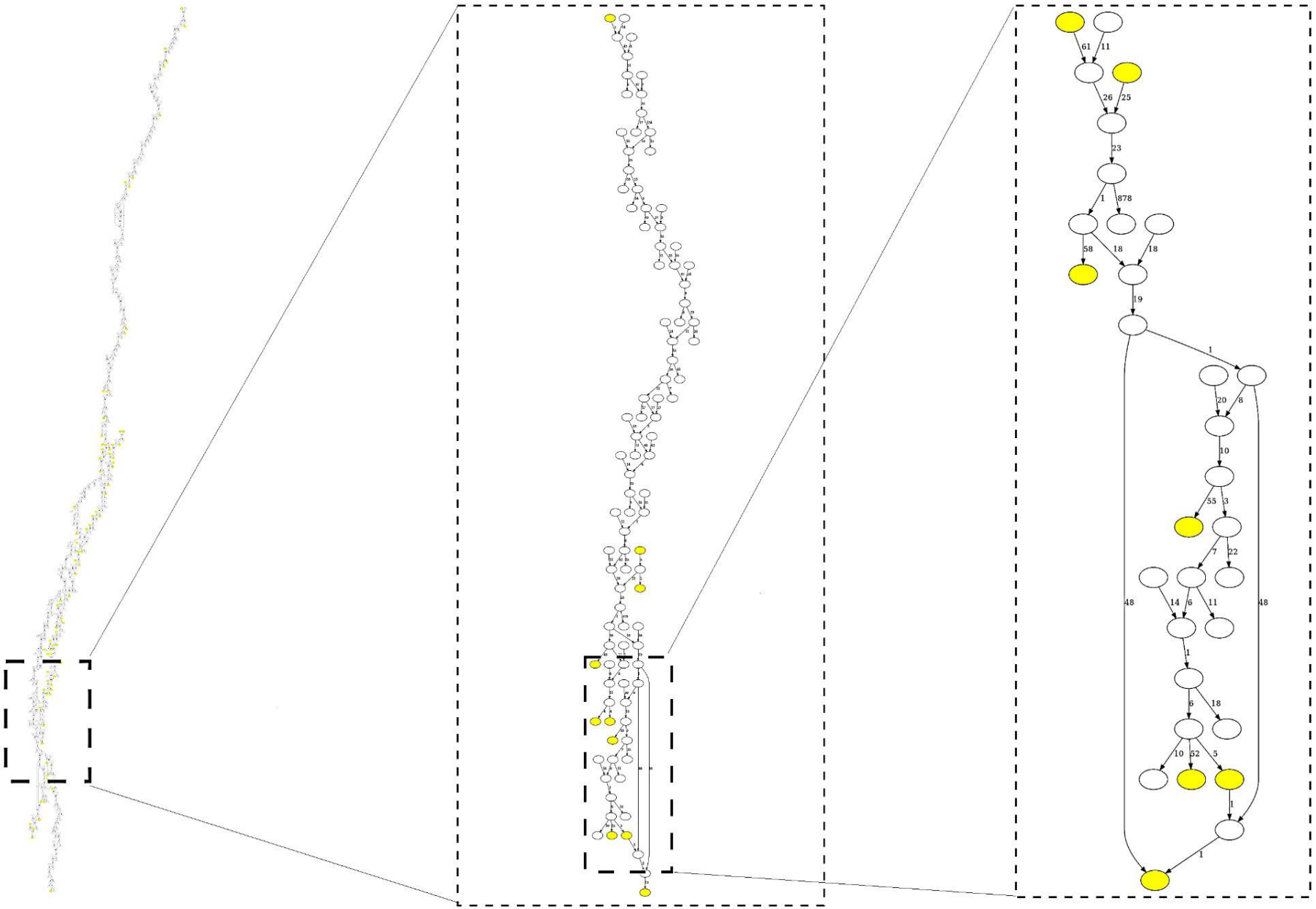


пузыри



химерные соединения




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

Техника

Представление графа

- Память
- Время

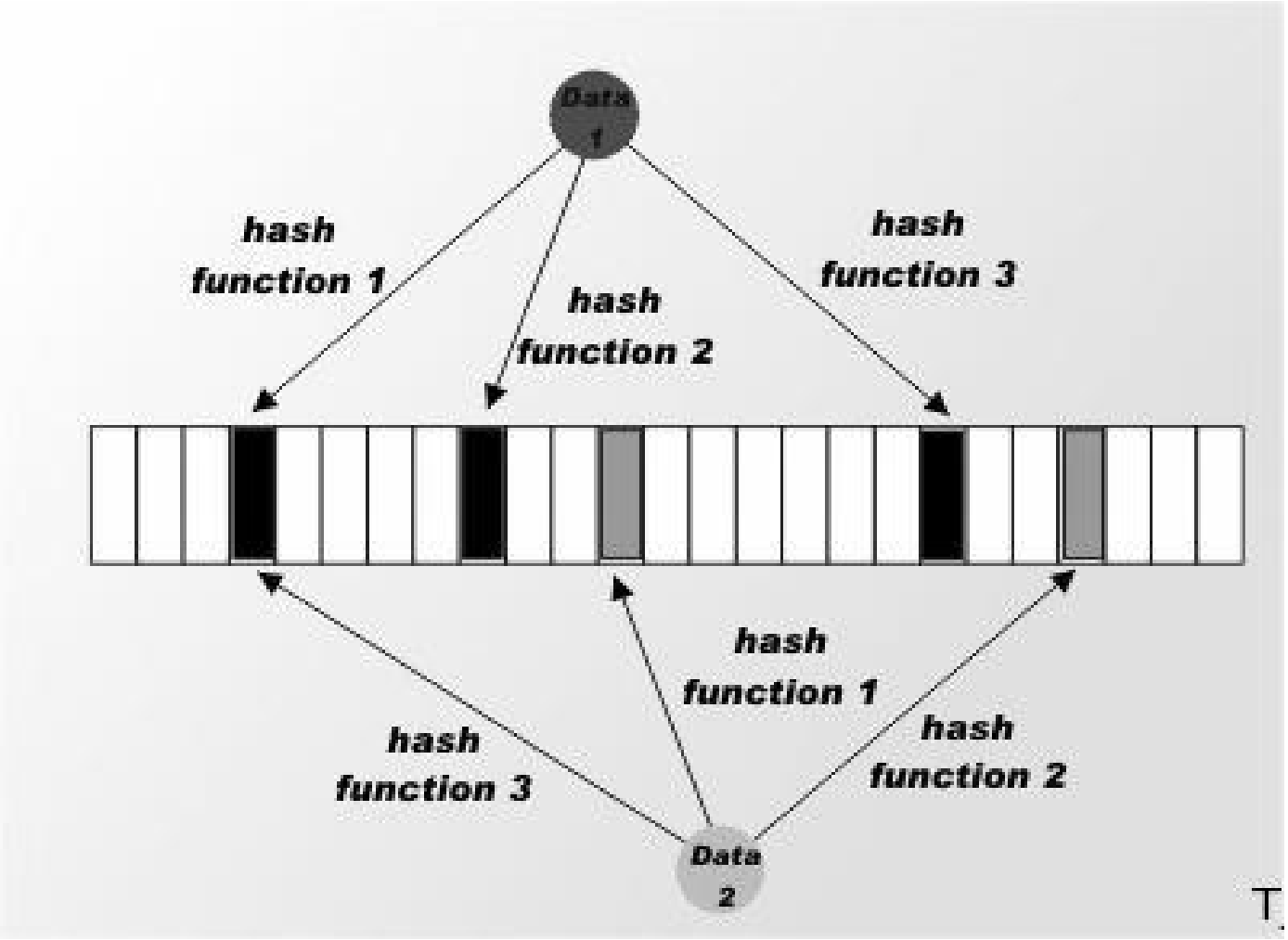
Представление графа

Требования:

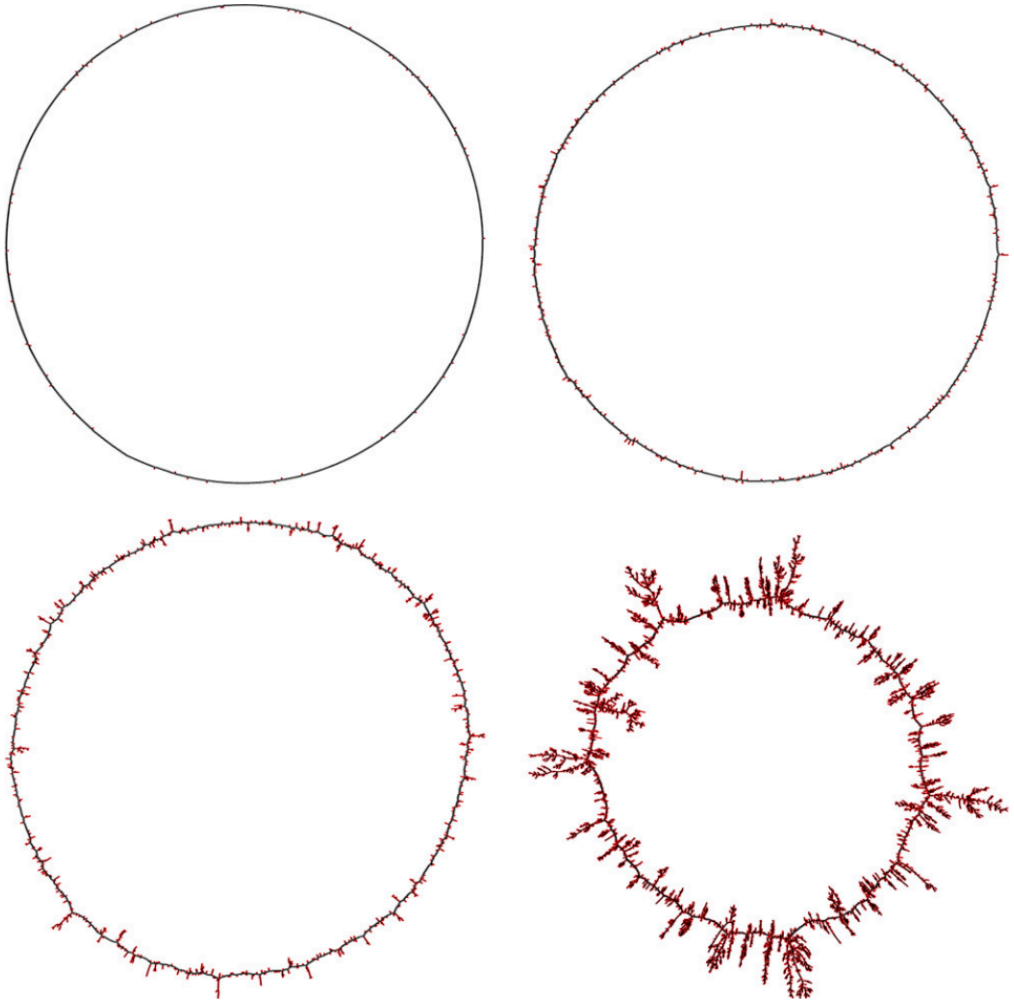
- Возможность перебрать все *k*-меры
- Возможность найти соседей *k*-мера

Пример: Множество всех *(k+1)*-меров

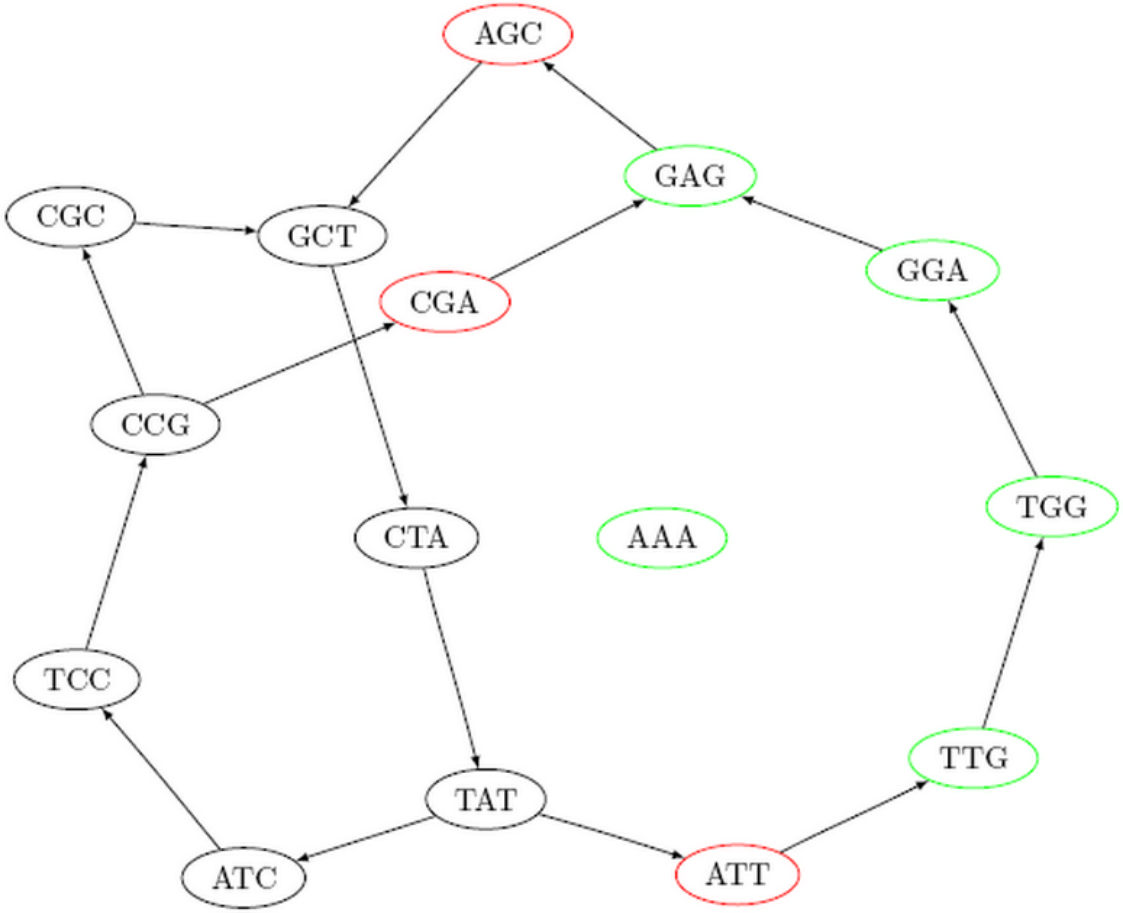
Фильтр Блума



Вероятностный граф де Брюйна



Точное представление



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. http://bioinf.spbau.ru/en/spades

Вопросы
???