Evolution of genome assembly methods: solutions and delusions

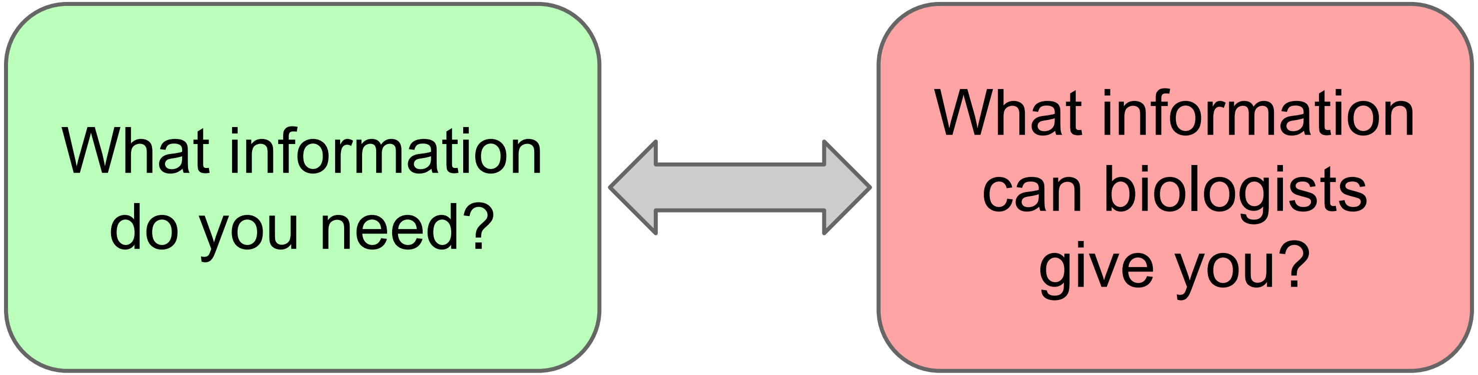
Anton Bankevich

SPbAU 2014

Assembly problem

- In 1944 the role of DNA as a primary information carrier was discovered.
- Before starting to understand information you should get it

How can one reconstruct 3Gb long sequence?

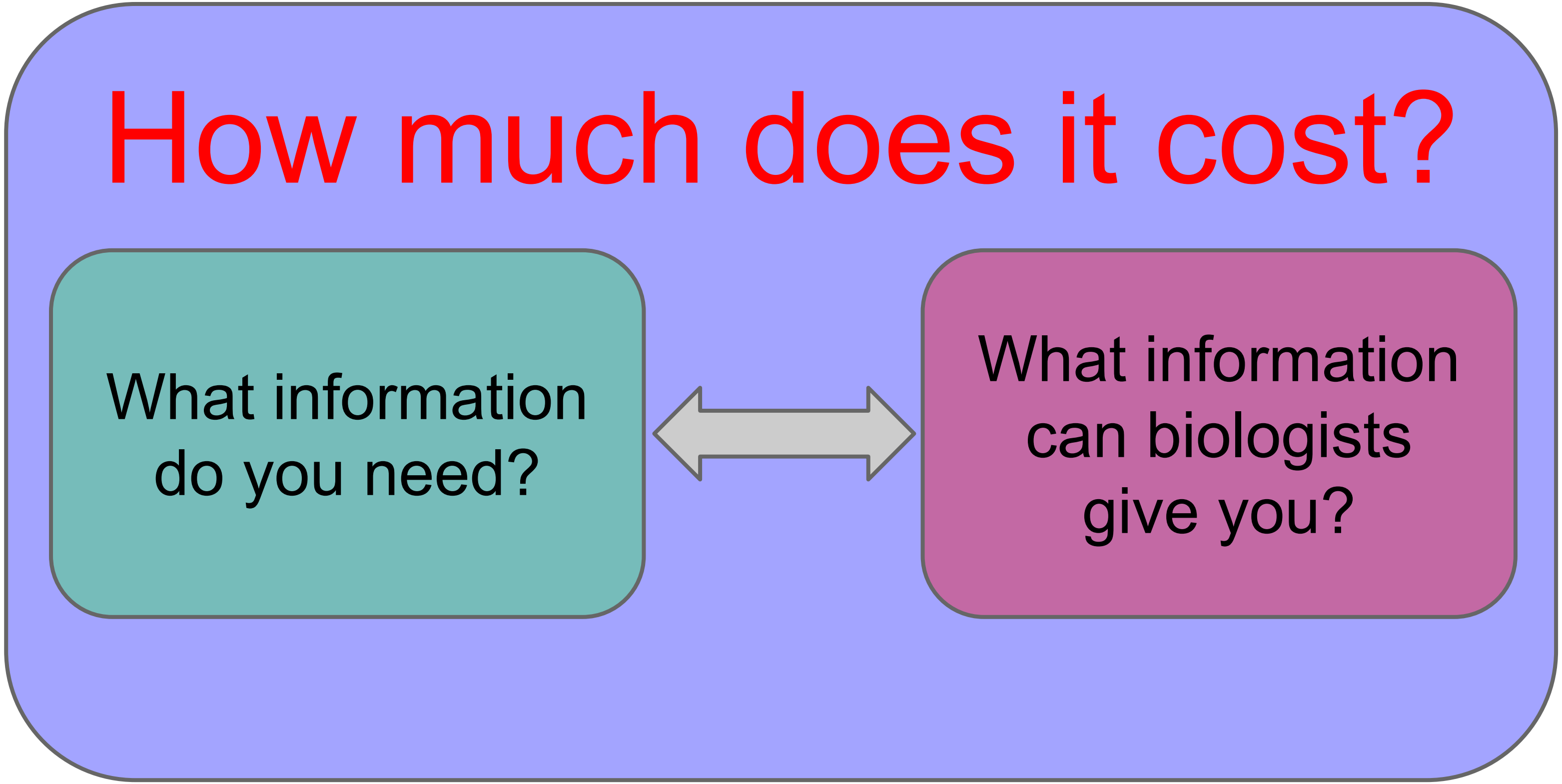


What information do you need?

What information can biologists give you?

Interaction between these two questions brings evolution into genome assembly problem.
Along with one more small question.

How can one reconstruct 3Gb long sequence?



How much does it cost?

What information do you need?

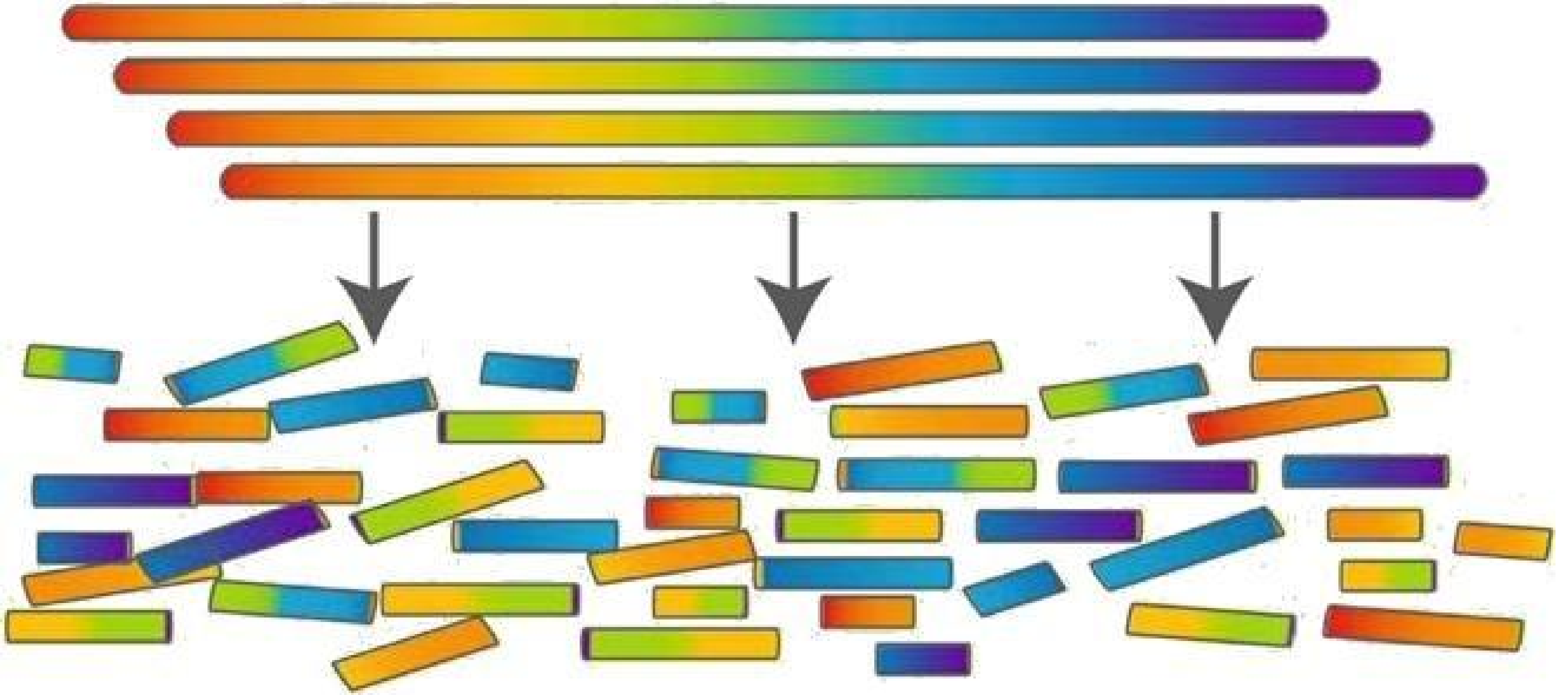
What information can biologists give you?

Read concept

Sequences of small pieces DNA can be read.

Reconstruction of the whole genome turns into solution of a puzzle.

Read concept



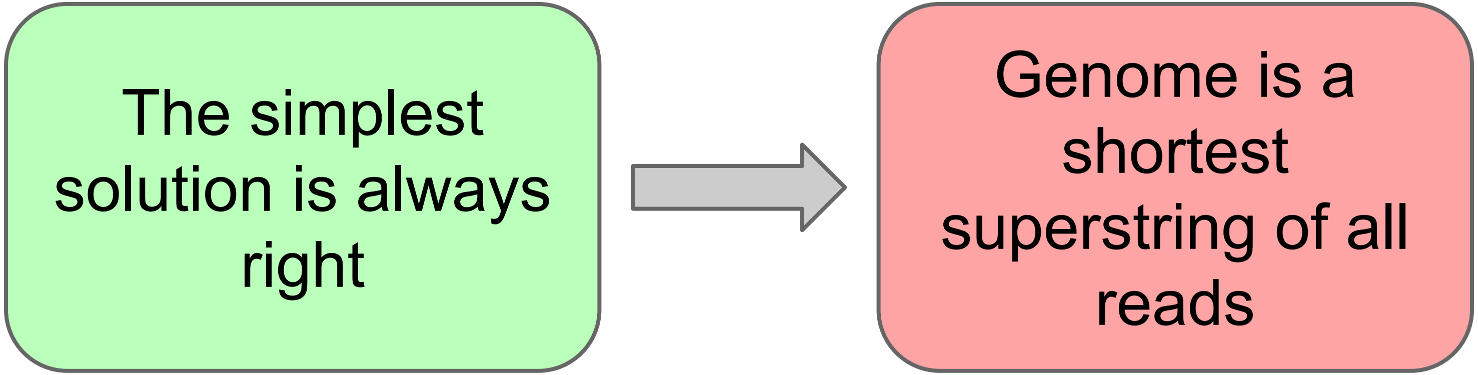
Assembly problem

Given: a set of subsequences of genome sequence

Goal: reconstruct whole genome

Question: is it even possible?

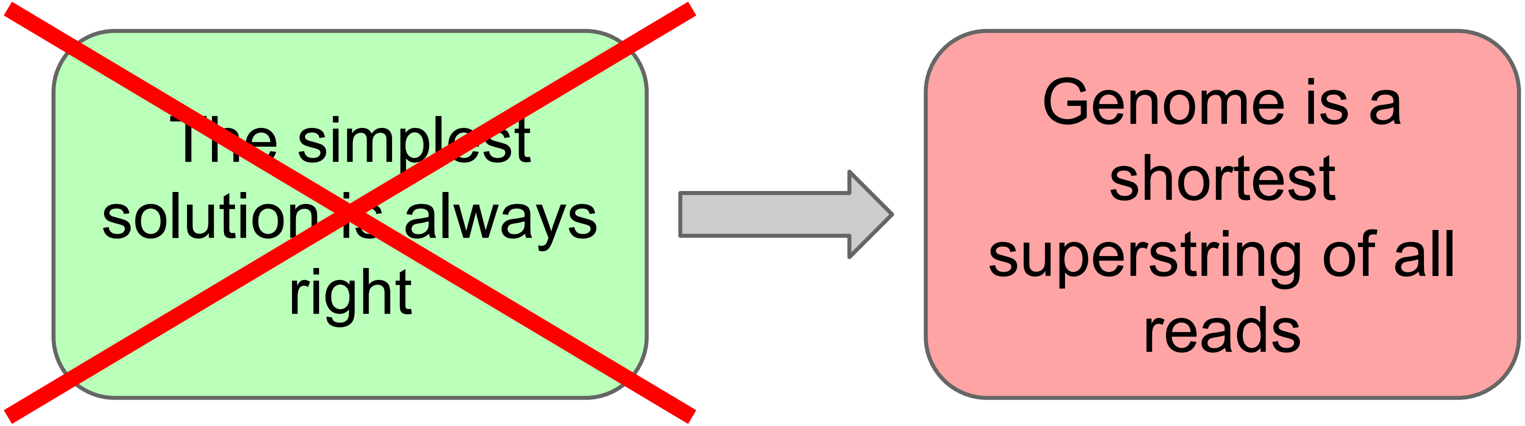
Shortest superstring delusion



The simplest solution is always right

Genome is a shortest superstring of all reads

Shortest superstring delusion



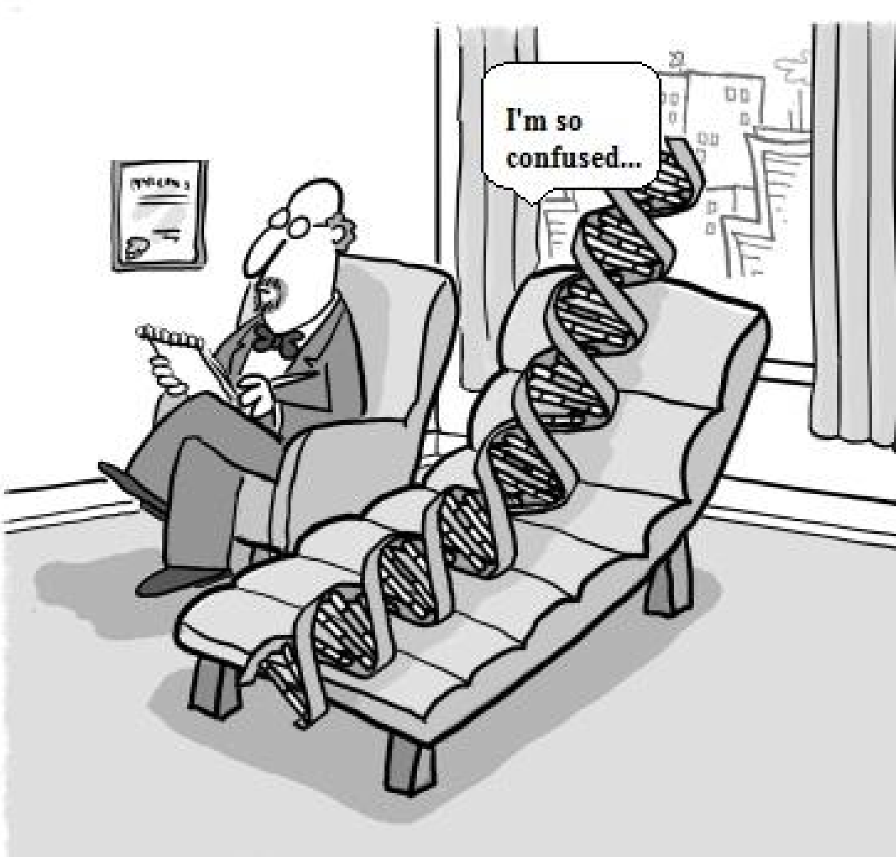
~~The simplest solution is always right~~

Genome is a shortest superstring of all reads

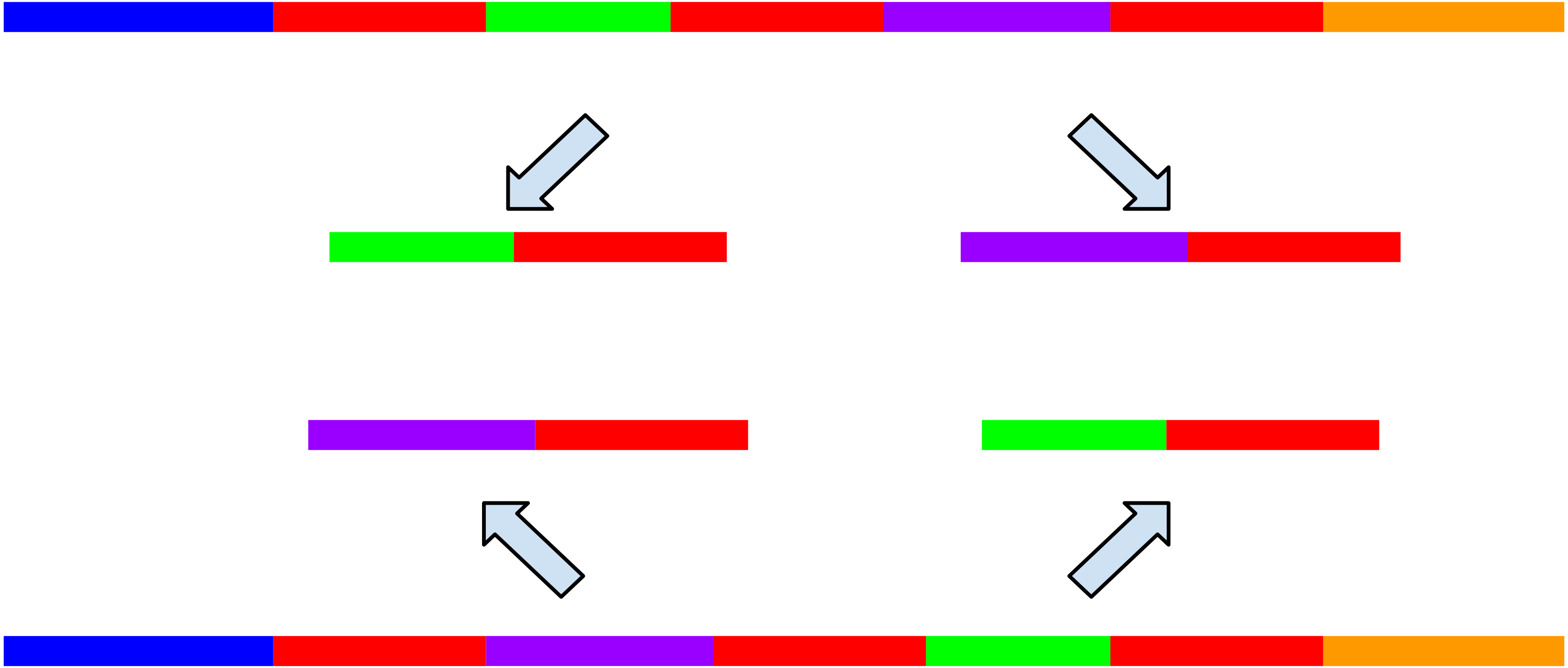
Repeat structure

Alu: длина 300, кратность 1 000 000

L1: длина 6000, кратность 516 000



Repeat structure



Assembly problem

Formal statement:

Assembly problem

Formal statement:

Has never been formulated...

Assembly problem

Wishes to assembly result:

- Answer is a set of sequences (contigs)
- We want longer contigs
- We want contigs to be subsequences of genome
- We do not want contigs to intersect

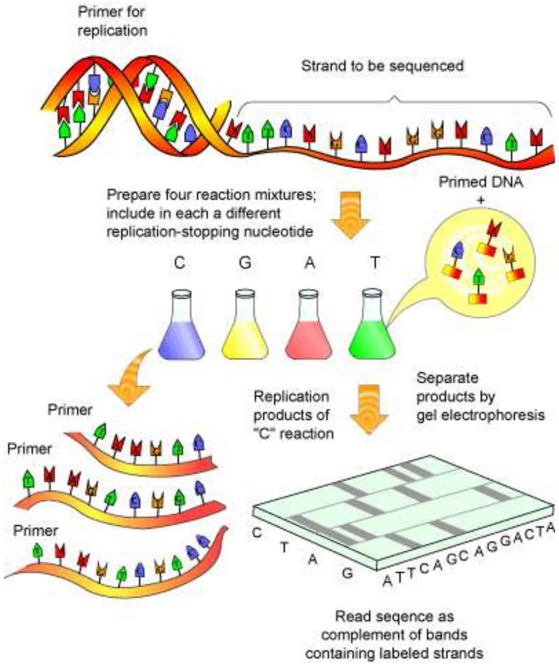
Assembly problem

Wishes to assembly result:

- Answer is a set of sequences (contigs)
- We want longer contigs
- We want contigs to be subsequences of genome
- We do not want contigs to intersect

So what’s the input?

Sanger


Primer for replication
Strand to be sequenced
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
Primed DNA
+
C
G
A
T
Primer
Primer
Primer
Replication products of "C" reaction
Separate products by gel electrophoresis
C
T
A
G
ATTCAGCAGGACTA
Read seqence as complement of bands containing labeled strands

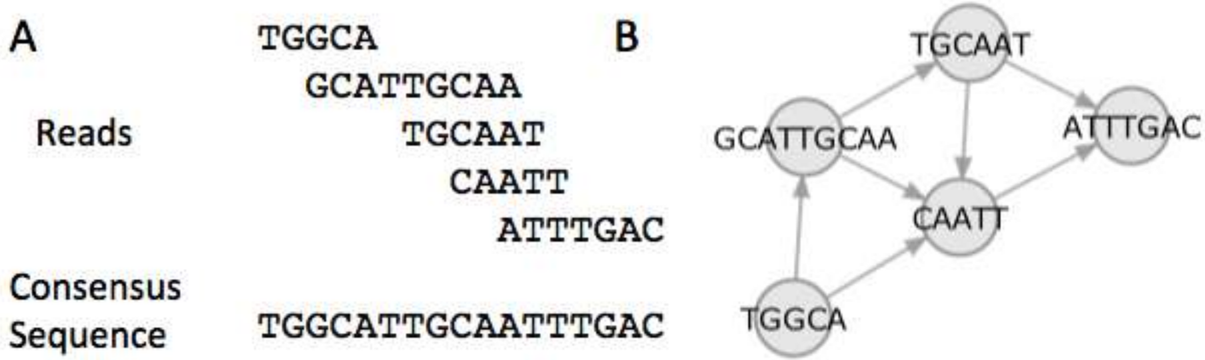
Length	up to 900 bp
Accuracy *	99.9 %
Cost per 10^6 bp	2400 $

Overlap graph: theory

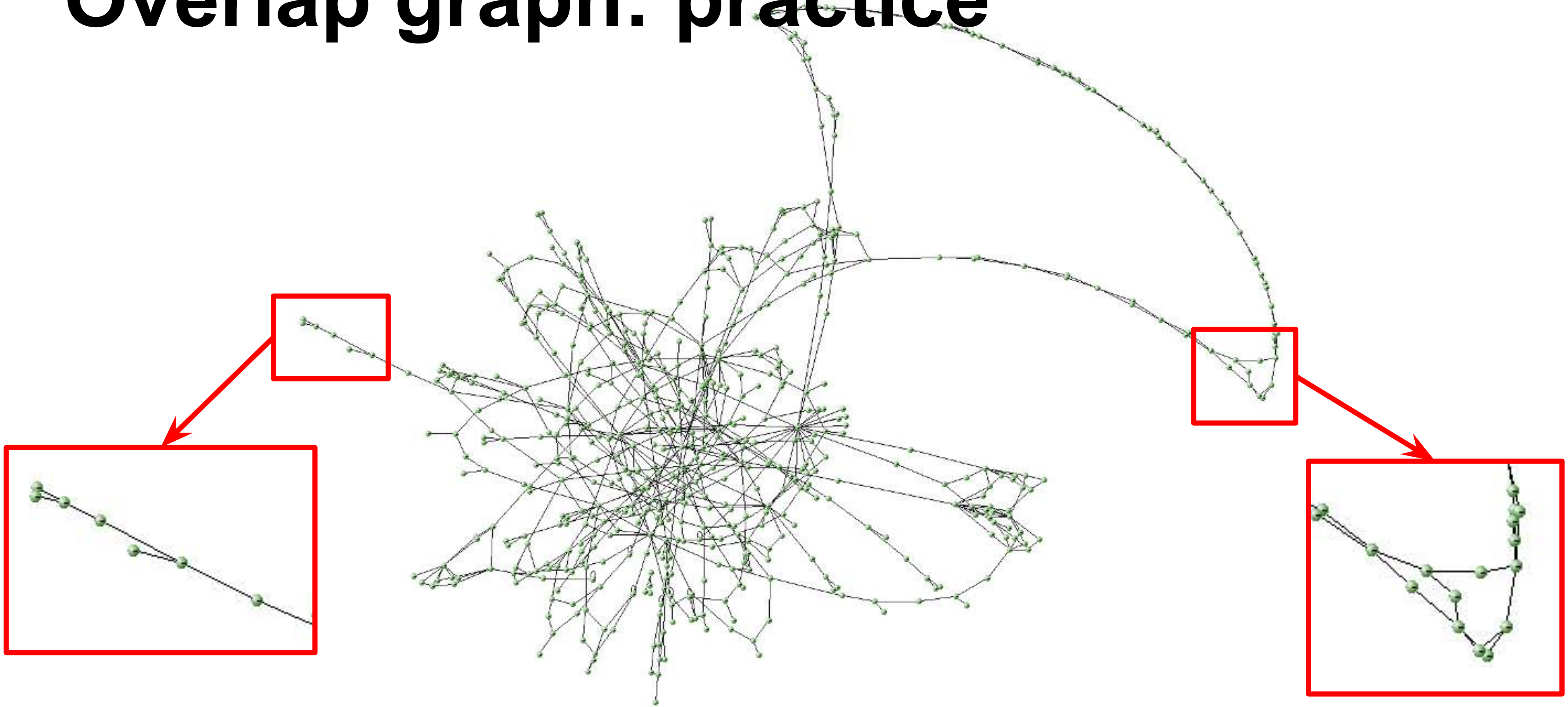
R = {r_1, …, r_n} is a set of reads

V = *R*, *e = (r_i, r_j)* in *E*, if read *r_i* overlaps read *r_j*

w(e) - weight of edge *e* is a size of overlap



Overlap graph: practice



overlap graph constructed from long reads for bacterial genome (output of Newler)

Delusions about overlap graph

- Finding Hamiltonian path of maximal weight **does not** give us superstring of *R* for genome with repeats

- Overlap graph constructed from reads contains erroneous edges due to sequencing errors

- $|V| = |R|$, max $|E| = R$^2

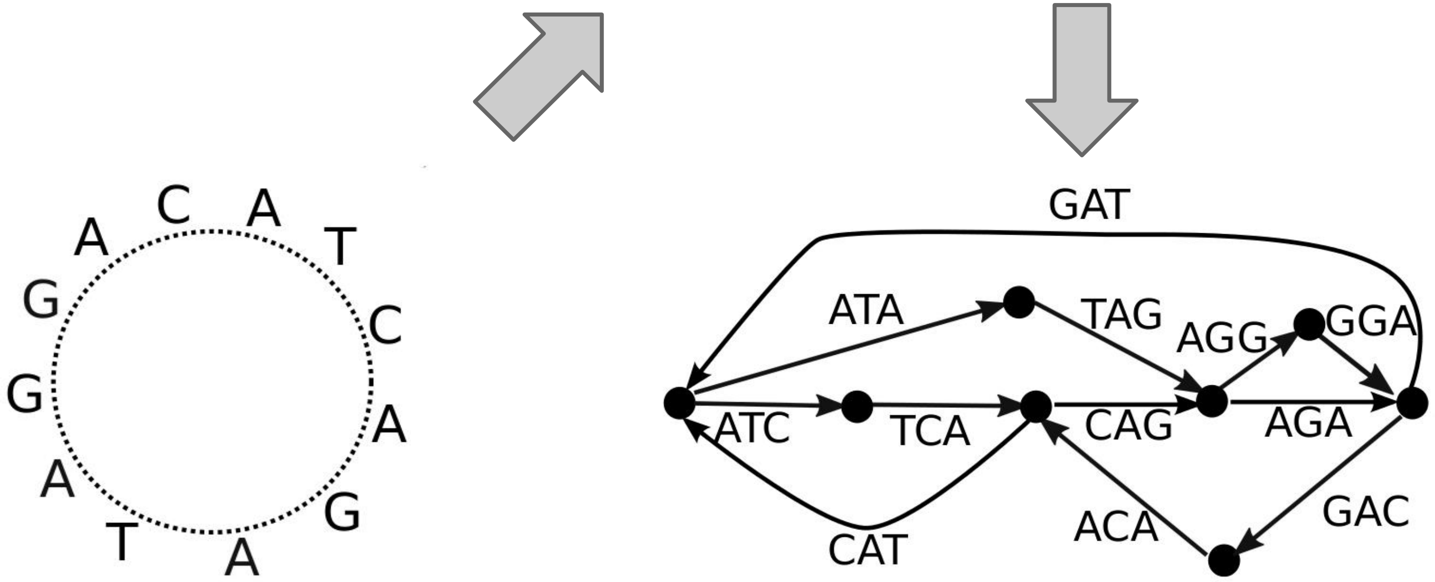
Next Generation Sequencing

Starting from 2005 several new sequencing technologies hit the market: 454, Solexa (Illumina).

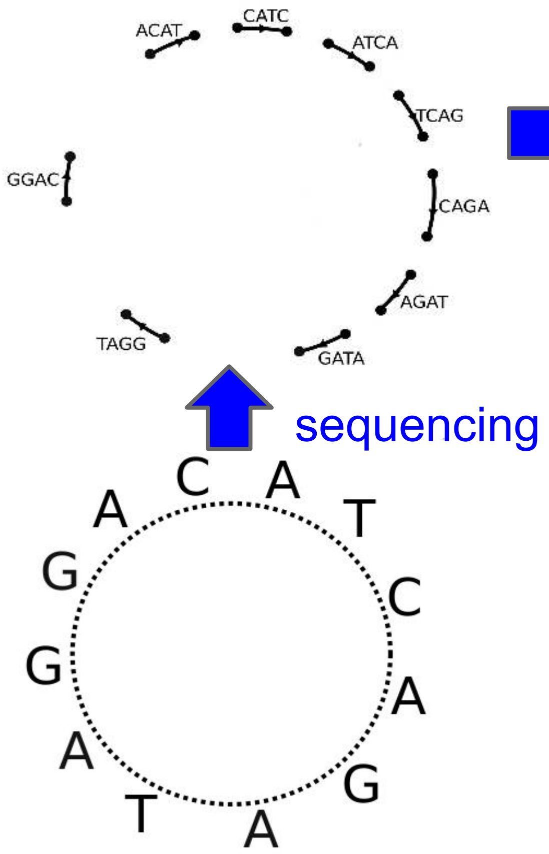
New technologies produced millions of short reads making overlap graph approach inapplicable.

de Bruijn graph

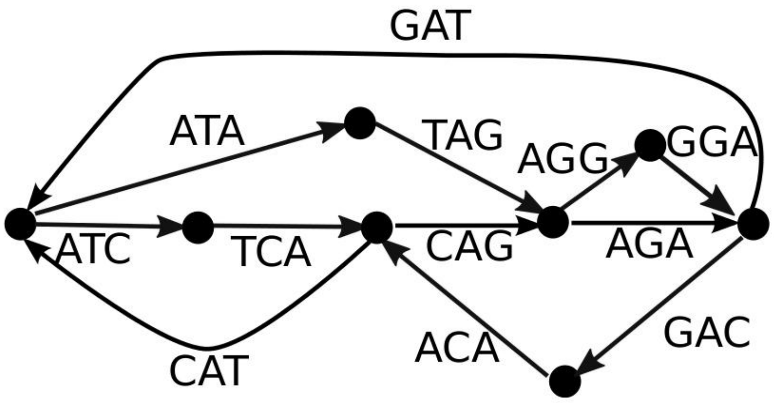
Vertices: k-mers from genome
Edges: (k+1)-mers from genome
k=2: 3-mer ACG gives AC -> CG



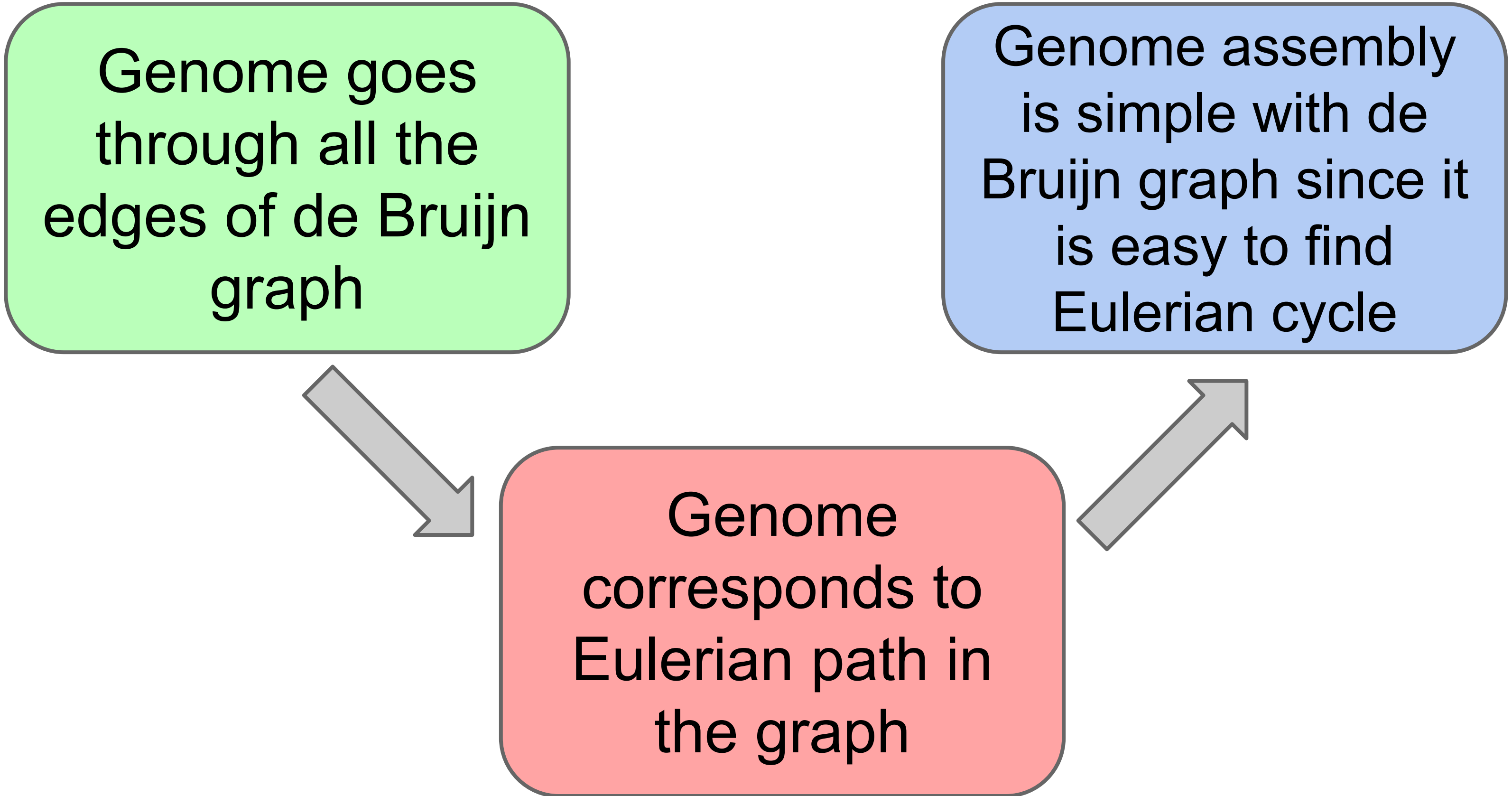
de Bruijn graph



Vertices: k-mers from reads
Edges: (k+1)-mers from reads
k=2: 3-mer ACG gives AC -> CG



Delusions about de Bruijn graph

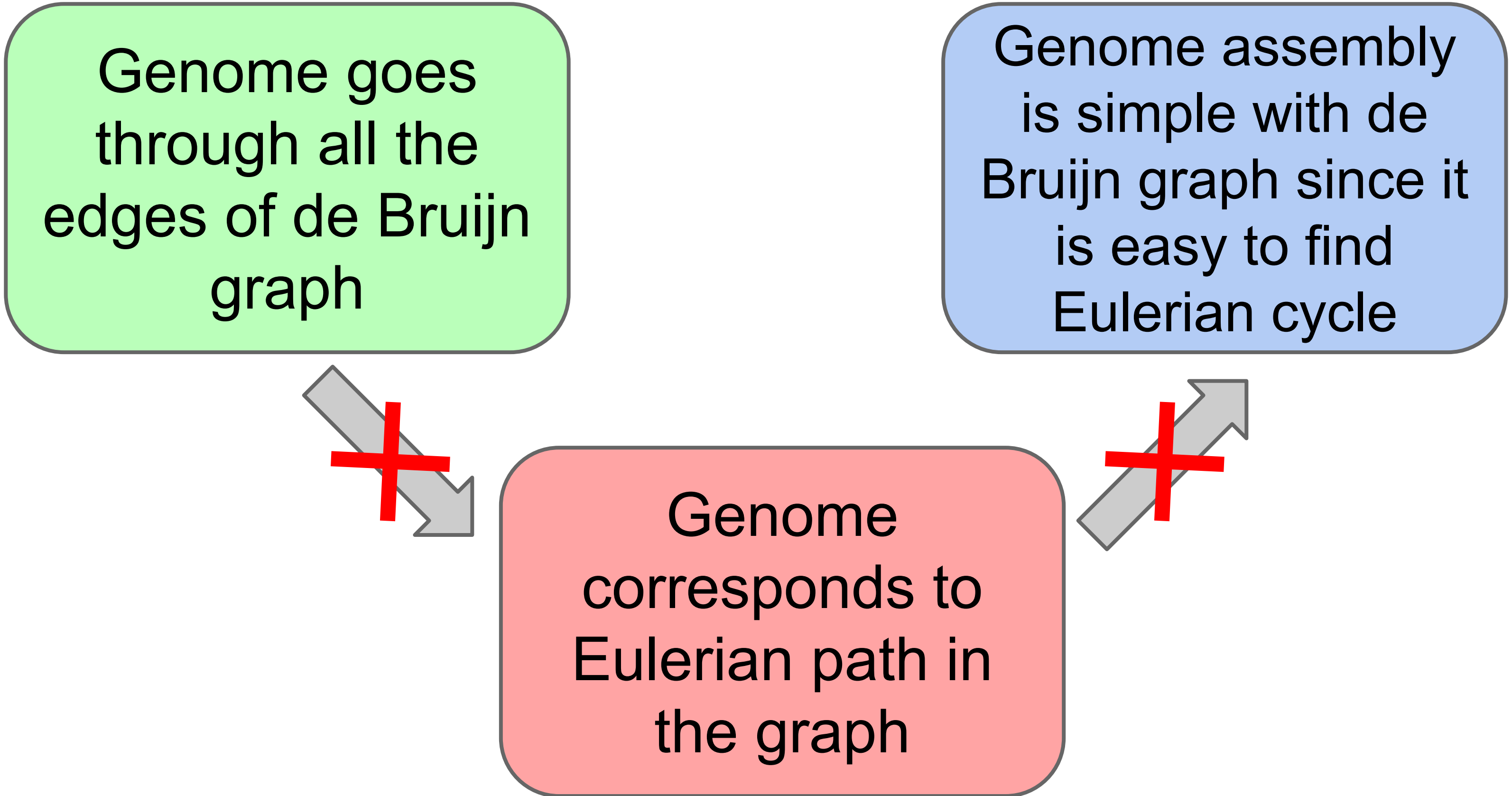


Genome goes through all the edges of de Bruijn graph

Genome corresponds to Eulerian path in the graph

Genome assembly is simple with de Bruijn graph since it is easy to find Eulerian cycle

Delusions about de Bruijn graph

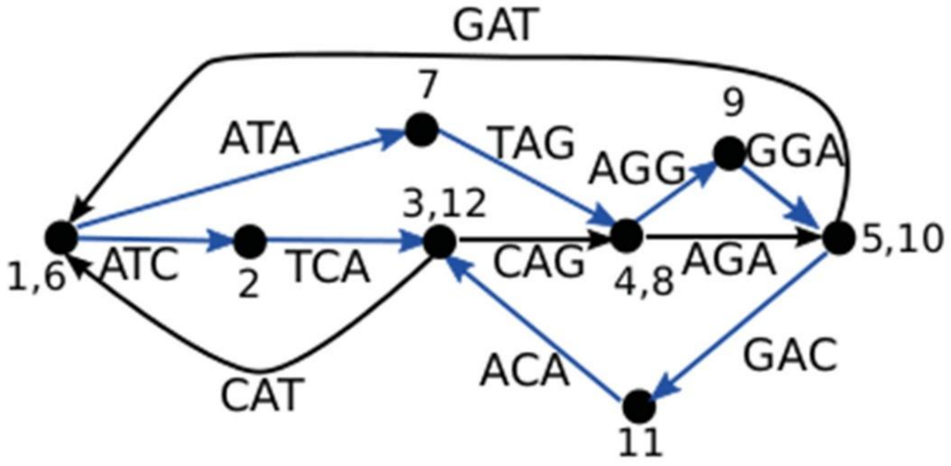


Genome goes through all the edges of de Bruijn graph

Genome corresponds to Eulerian path in the graph

Genome assembly is simple with de Bruijn graph since it is easy to find Eulerian cycle

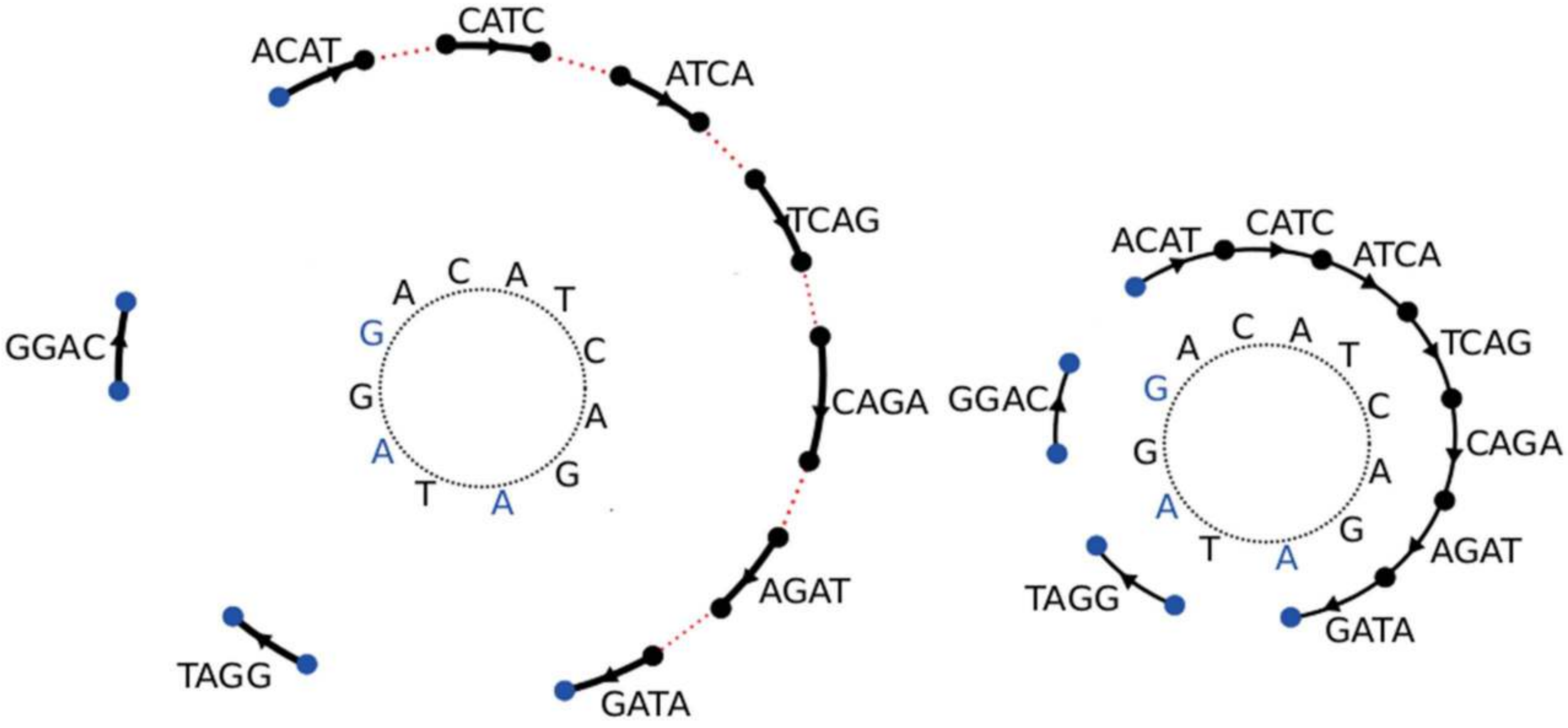
Assembly using de Bruijn graph



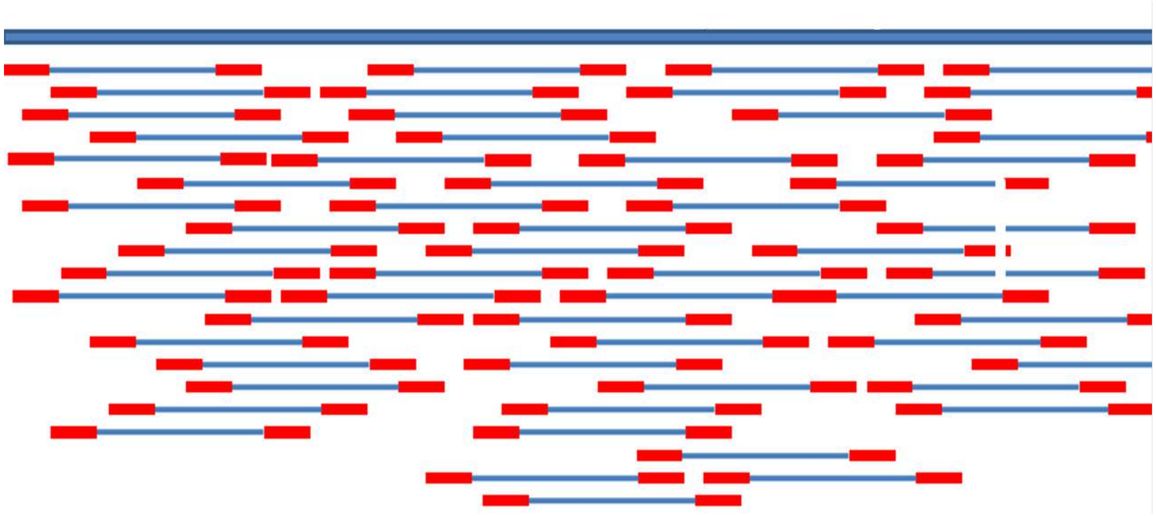
Output contigs are formed by non branching paths in de Bruijn graph (highlighted in blue)

Only repeats of length k can be resolved this way

Choice of k


ACAT
CATC
ATCA
TCAG
CAGA
AGAT
GATA
GGAC
TAGG
A C A T C A G A T A G G
ACAT
CATC
ATCA
TCAG
CAGA
AGAT
GATA
GGAC
TAGG
A C A T C A G A T A G G

Paired reads



Paired reads allow us to construct long genomic paths in de Bruijn graph instead of just using non branching paths.

De Bruijn graph challenges

- De Bruijn graph of human genome contains 3 billion vertices.
- De Bruijn graph of reads from human genome contains at least 10 times more vertices.
- Even if we use 10 bytes/kmer we need 300 Gb memory
- If we consider all additional information and overheads we hit the roof.

De Bruijn graph as a data structure

Operations of de Bruijn graph data structure

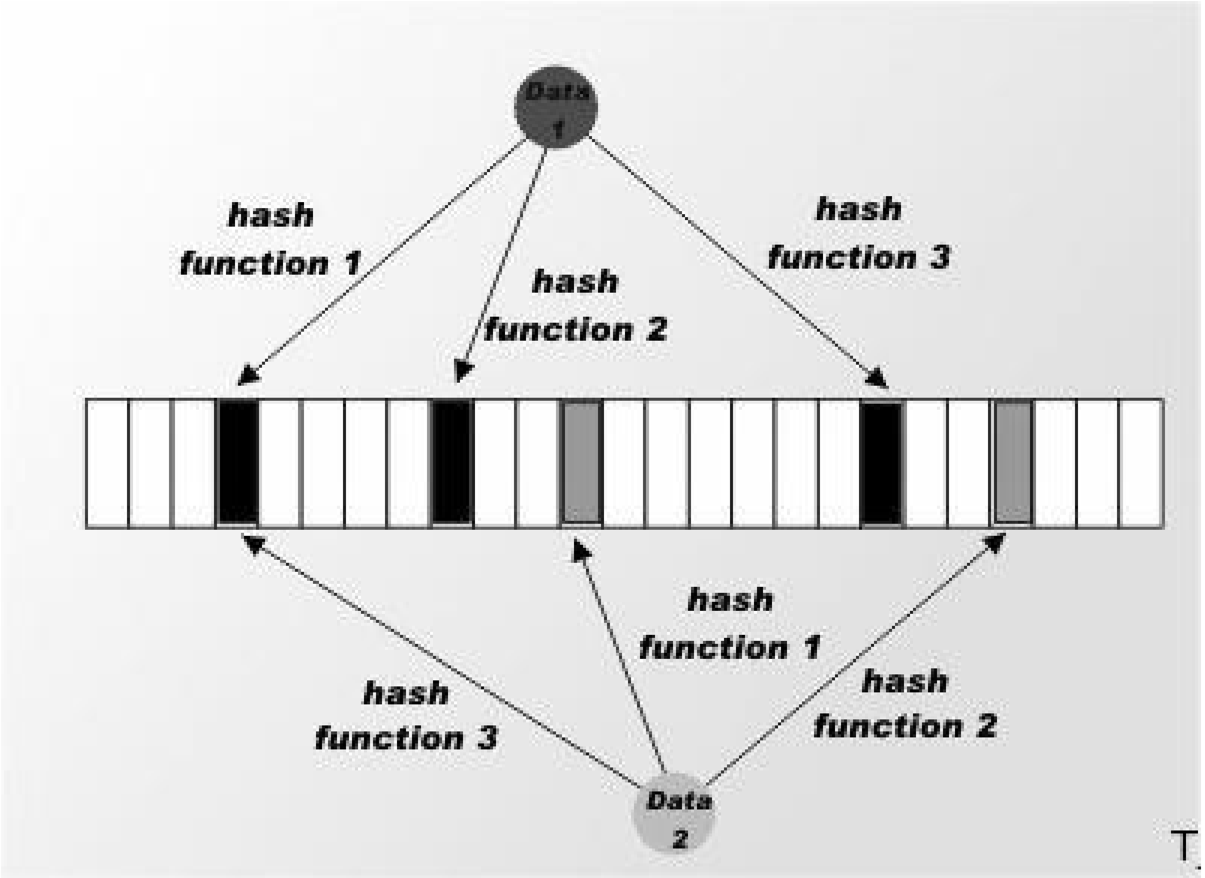
- Find all adjacent k-mers for a given k-mer
- Iterate over all k-mers

How to do it using as little memory as possible.

Distributed storage (Abyss)

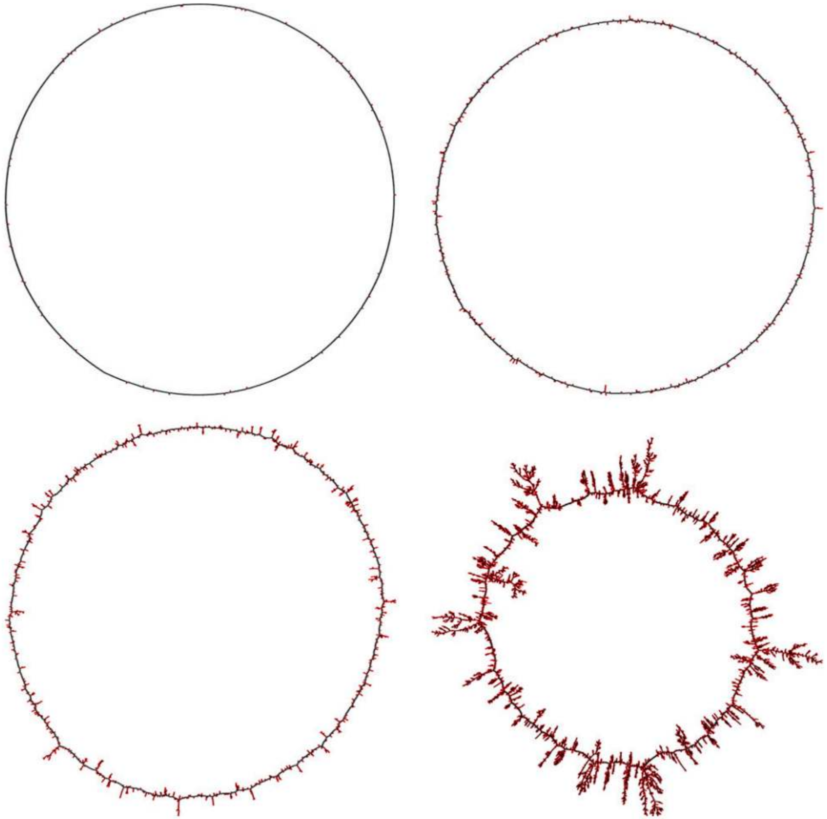
- Each k-mer is stored in one of the nodes based on some hash.
- Hash is chosen wisely in such a way that adjacent vertices have high probability to be stored on the same node

Bloom filters (Minia)

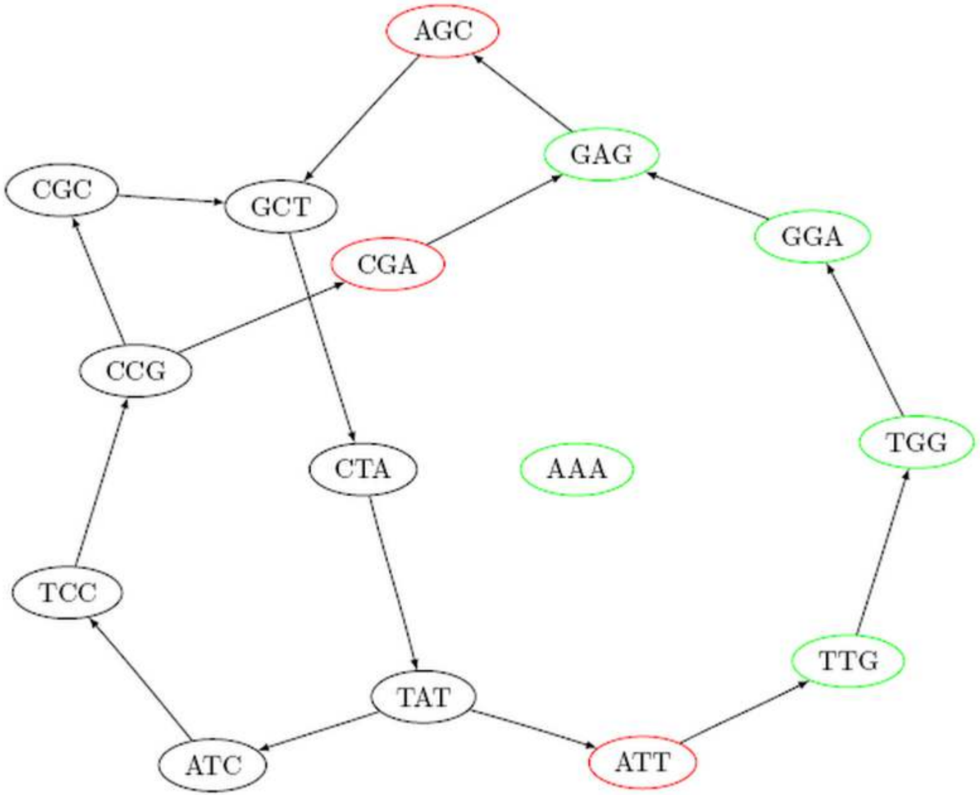


T

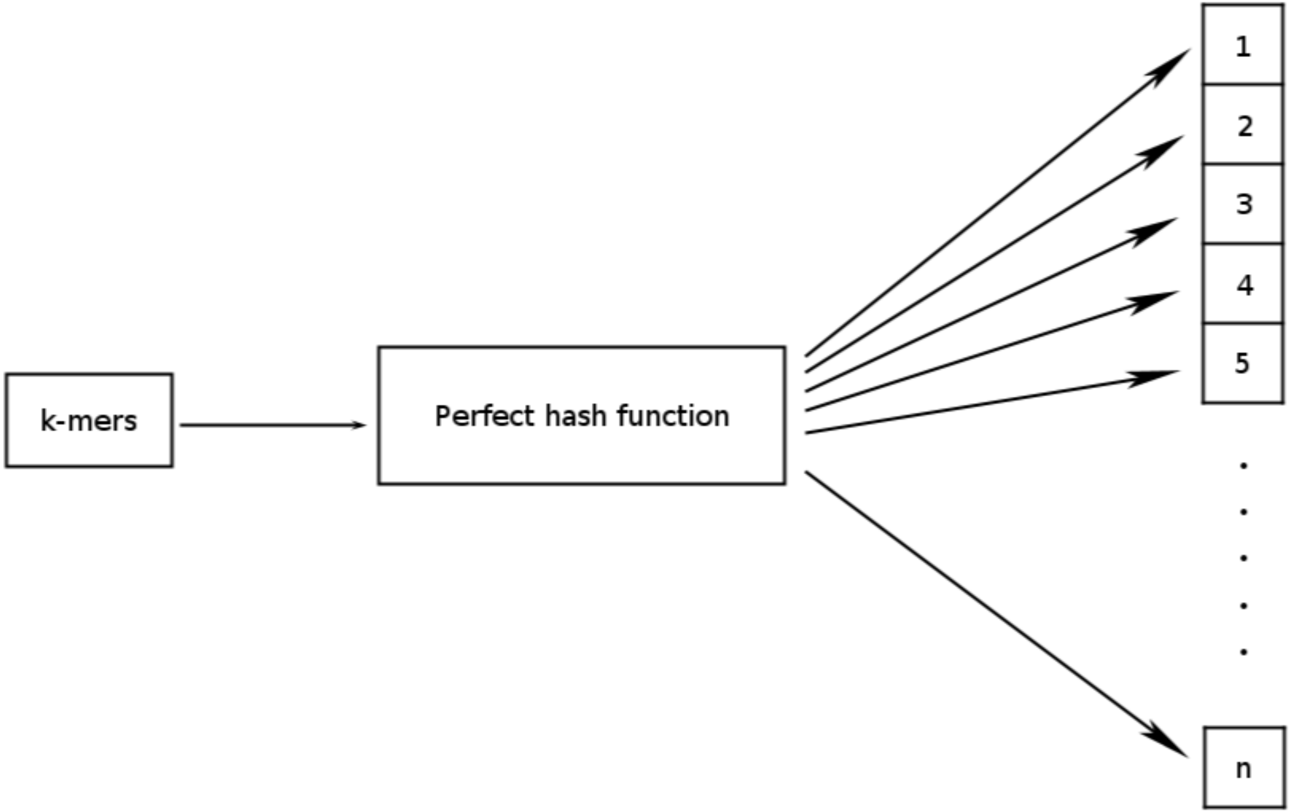
Probabilistic de Bruijn graph



Exact representation using bloom filters



Perfect hashing (Miraculous, SPAdes)



Perfect hashing

- Requires only 2.7 bits/kmer
- Any additional information can be stored without overhead
- Allows to modify de Bruijn graph
- Does not allow to check if given k-mer is in the graph

Sparse de Bruijn graphs (SparseAssembler)

Idea: A lot of information in de Bruijn graph is not useful.

Solution: remove all highly covered and all low-covered k-mers/vertices and consider reads as edges.

Result: simultaneous graph cleaning and repeat resolution.

Results of NGS era

- Assembly technology development result: Dozens of excellent assemblers appeared and used to assemble various species
- Sequencing technology development result: NGS technologies came close to the quality of Sanger. Today Illumina produces accurate and cheap reads with length up to 300bp.

SGA

SGA assembler uses FM-index to construct overlap graph using very low amount of memory.

The problem of storing de Bruijn graph in FM-index remains open.

Comparison of approaches

Overlap graph

- Can use the whole read length for repeat resolution
- Is tolerant for moderate error rate in reads

De Bruijn graph

- Can be used with very short reads (obsolete?)
- Can be used with read libraries of very high coverage
- Can be used with very high error rate

Third generation and new challenges

NGS proved unable to assemble large and even complex small genomes efficiently.

Third generation is coming with more challenges:

- Pacbio technology provides very long but inaccurate reads
- Bionano technology creates maps of 100Kb long genome fragments
- TruSeq technology from Illumina provides 10Kb long, accurate but “virtual” reads.

Thank you

Questions?

Reference assisted vs de novo

Alignment problem

MicroChips

Genome map concept

Technologies