Mansoura University
Faculty of Computers and Information
Department of Computer Science
First Semester: 2020-2021

[MED121] Bioinformatics: Sequence Assembly Algorithms

Grade: Third Year (Medical Informatics Program)

Sara El-Metwally, Ph.D.

Faculty of Computers and Information,

Mansoura University,

Egypt.

AGENDA

- Sequence Assembly
- Sequence Assembly Challenges
- Genome Assembly Terminology
- Comparative Vs. De Novo Assembly
- Overlap-layout-consensus Approach
- De Bruijn Graph
- Assembly Evaluation metrics

SEQUENCING TECHNOLOGIES

ps://www.biocompare.com/23967-Ne

https://ngisweden.scilifelab.se/technologies/pacific-biosciences/pacbio-sequel/

https://www.technologyreview.com/2016/02/24/8993/with-patent-suit-illumina-looks-to-tame-emerging-british-rival-oxford-nanopore/

SEQUENCING TECHNOLOGIES

Throughput

Read length

sequencing errors

Sequencing cost

Bioinformatics tools available for data analysis and processing

NOTES

Reads

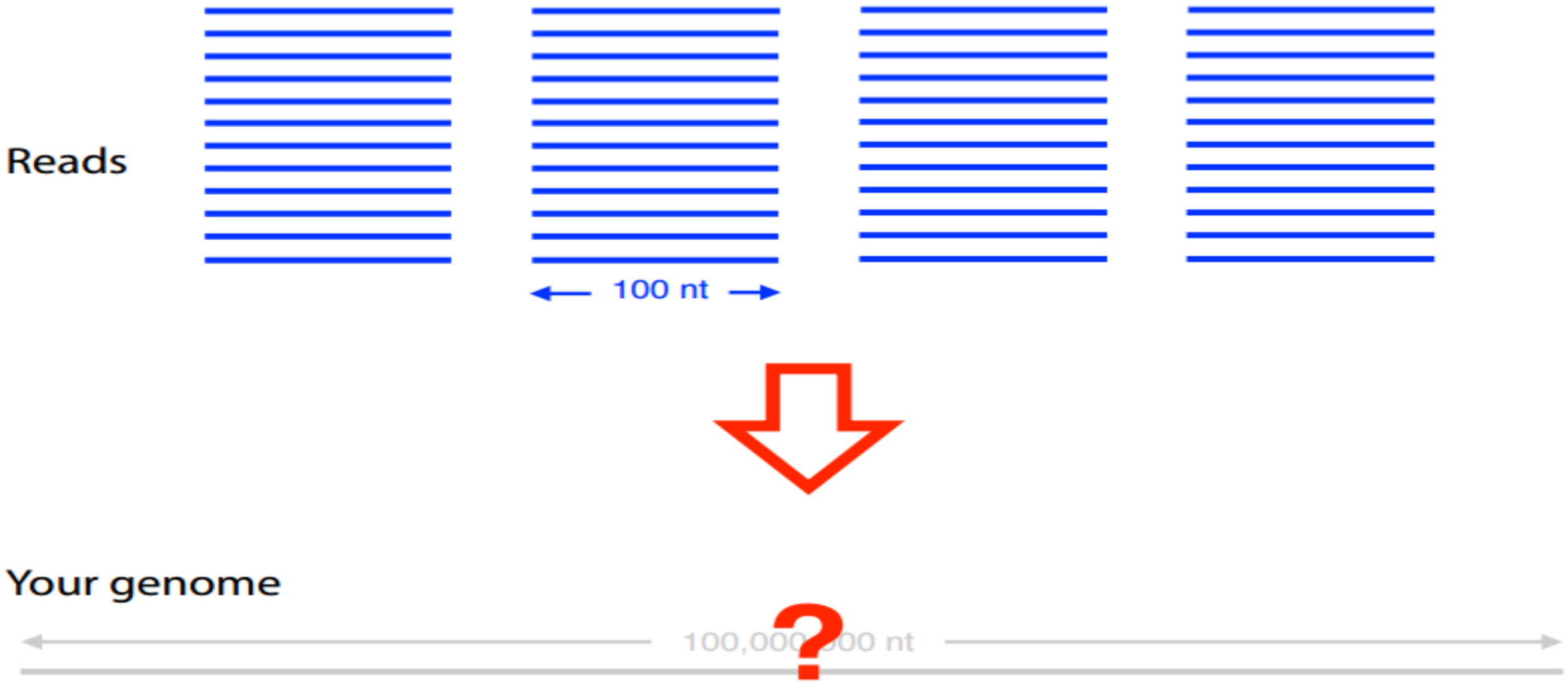
GTATGCACGCGATAG	TATGTCGCAGTATCT	CACCCTATGTCGCAG	GAGACGCTGGAGCCG
TAGCATTGCGAGACG	GGTATGCACGCGATA	TGGAGCCGGAGCACC	CGCTGGAGCCGGAGC
TGTCTTTGATTCCTG	CGCGATAGCATTGCG	GCATTGCGAGACGCT	CCTATGTCGCAGTAT
GACGCTGGAGCCGGA	GCACCCTATGTCGCA	GTATCTGTCTTTGAT	CCTCATCCTATTATT
TATCGCACCTACGTT	CAATATTCGATCATG	GATCACAGGTCTATC	ACCCTATTAACCACT
CACGGGAGCTCTCCA	TGCATTTGGTATTTT	CGTCTGGGGGGTATG	CACGCGATAGCATTG
GTATGCACGCGATAG	ACCTACGTTCAATAT	TATTTATCGCACCTA	CCACTCACGGGAGCT
GCGAGACGCTGGAGC	CTATCACCCTATTAA	CTGTCTTTGATTCCT	ACTCACGGGAGCTCT
CCTACGTTCAATATT	GCACCTACGTTCAAT	GTCTGGGGGGTATGC	AGCCGGAGCACCCTA
GACGCTGGAGCCGGA	GCACCCTATGTCGCA	GTATCTGTCTTTGAT	CCTCATCCTATTATT
TATCGCACCTACGTT	CAATATTCGATCATG	GATCACAGGTCTATC	ACCCTATTAACCACT
CACGGGAGCTCTCCA	TGCATTTGGTATTTT	CGTCTGGGGGGTATG	CACGCGATAGCATTG

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

NOTES



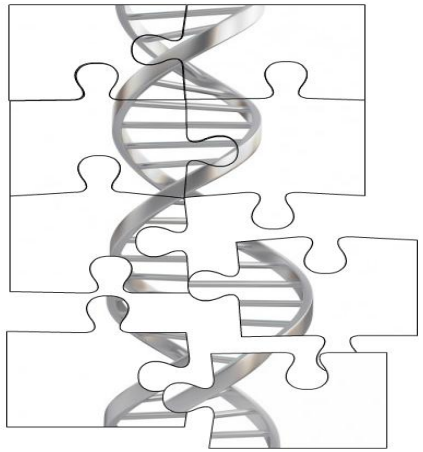
Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

What is Next?

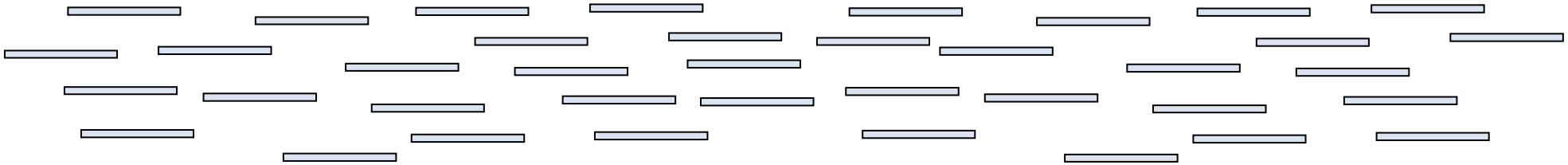
Sequence Analysis

Alignment

Assembly



DNA SEQUENCE ASSEMBLER
(BACKGROUND)



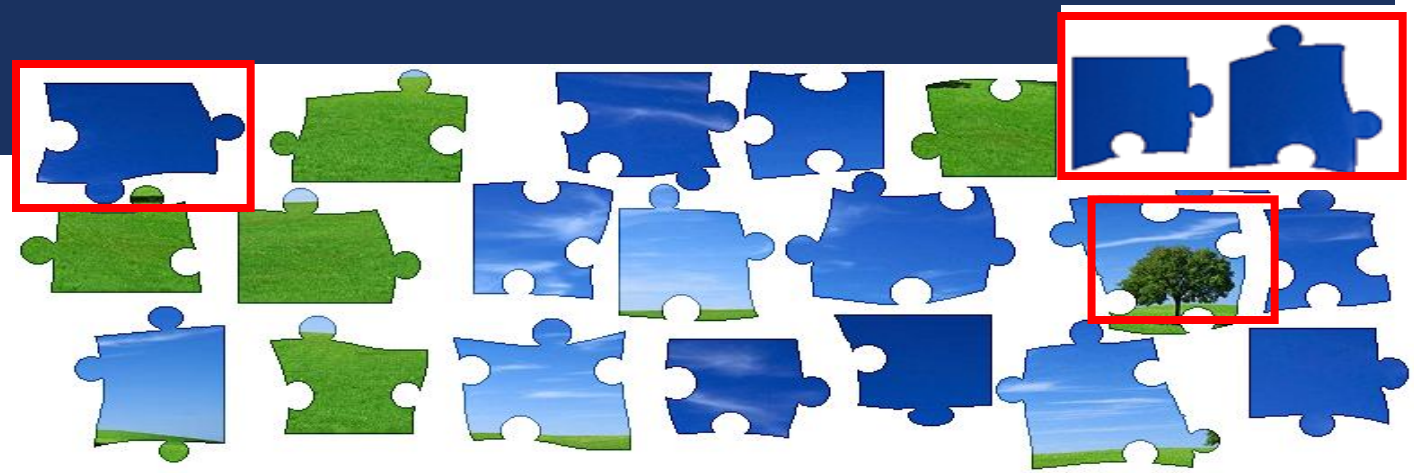
DNA SEQUENCE ASSEMBLER
(BACKGROUND)

- Assembler is a computer program that stitches the sequencing reads together into longer sequences to reconstruct the original genome.

- Sequence assembly is the initial step towards downstream data analysis of the sequencing data.

ASSEMBLY = JIGSAW PUZZLE

(ASSEMBLY CHALLENGES)



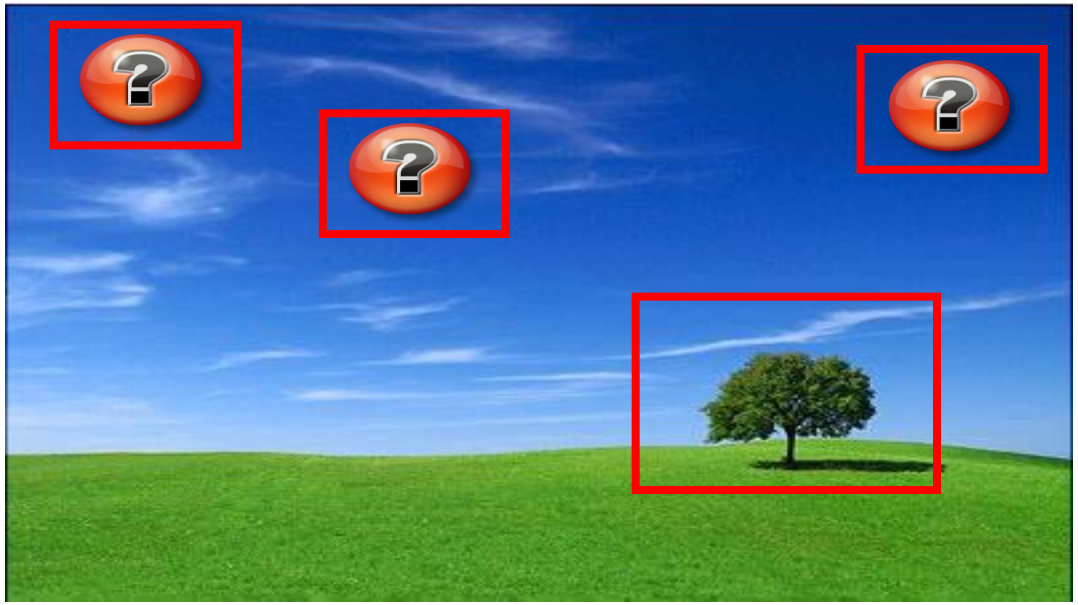
Assembling the sequencing reads like solving a jigsaw puzzle…

Some pieces are similar and repeated

Some pieces have unique features

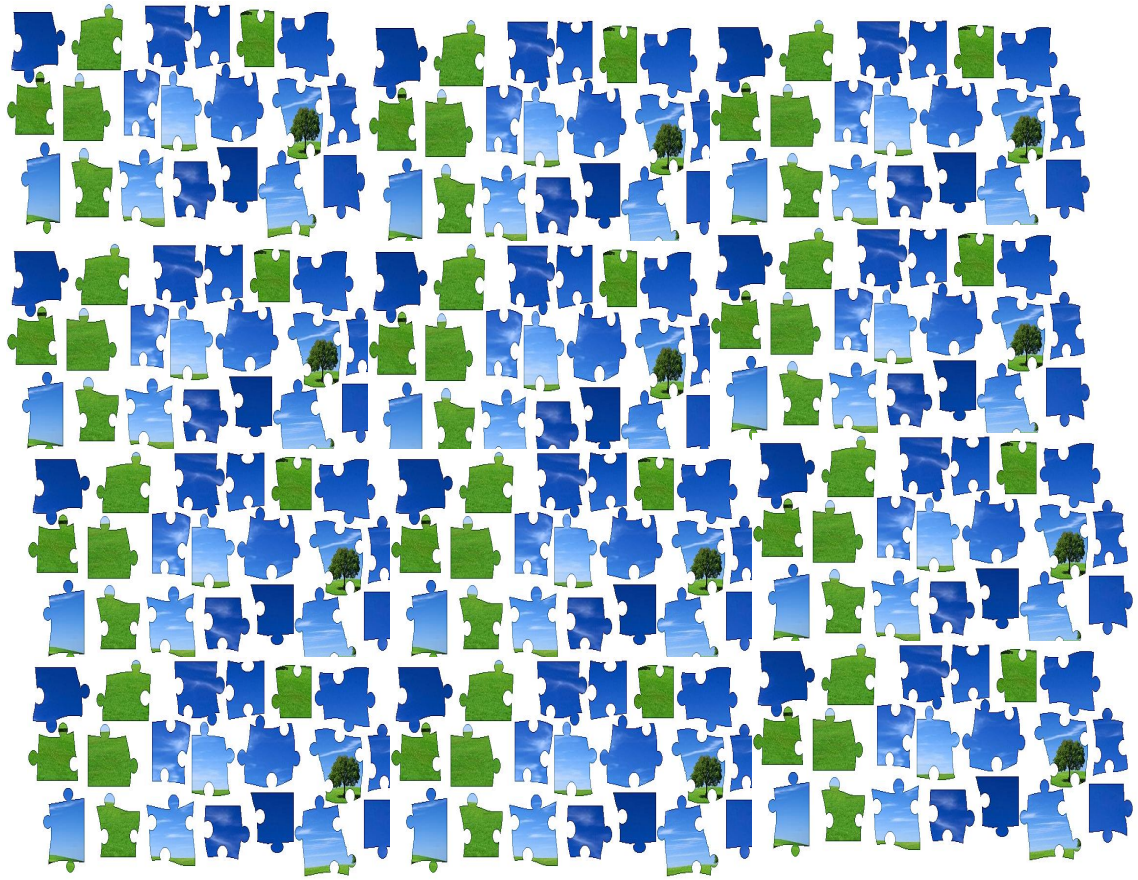
Some pieces are missing

Some pieces have errors



ASSEMBLY = JIGSAW PUZZLE

(ASSEMBLY CHALLENGES)



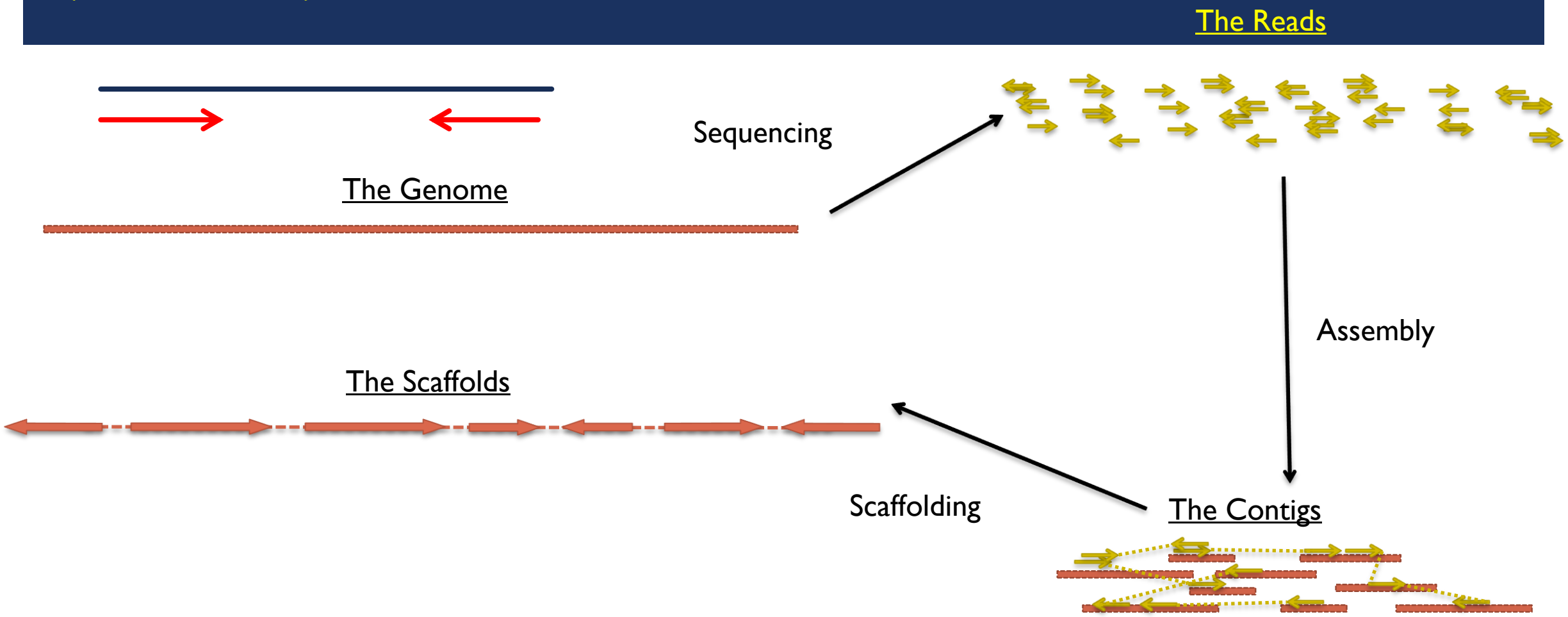
When the number of pieces is increased and the size of each piece is decreased ,the process of solving the puzzle becomes more complicated .

Try to solve the puzzle without a reference picture ??

- **Human genome puzzle could have 3 billion pieces, each with 100 copies.**

GENOME ASSEMBLY TERMINOLOGY

(BACKGROUND)



COMPARATIVE VS. DE NOVO

- **Comparative Assembly** : reference based assembly or mapping to genome of a closely related species .

- **De Novo Assembly** : assembly in the strict sense . No or little information about the genome , transcriptome or proteins.

EXAMPLE

read

ACTGAGTACTGCAT

ACTGAGTACTGAGT

CATGAGTACACTGT

ACTACTGA

EXAMPLE

Ovelap

ACTGAGTACTGCAT

Ovelap Length = 7 chars

ACTGAGTACTGAGT

CATGAGTACACTGT

ACTACTGA

Layout

ACTGAGTACTGAGTACTGCAT

EXAMPLE

ACTGAGTACTGCAT

ACTGAGTACTGAGT

CATGAGTACACTGT

ACTACTGA

ACTGAGTACTGAGTACTGCAT

EXAMPLE

ACTGAGTACTGCAT

ACTGAGTACTGAGT

CATGAGTACACTGT

ACTACTGA

ACTACTGAGTACTGAGTACTGCAT

EXAMPLE

ACTGAGTACTGCAT

ACTGAGTACTGAGT

CATGAGTACACTGT

ACTACTGA

ACTACTGAGTACTGAGTACTGCAT

EXAMPLE

ACTGAGTACTGCAT

ACTGAGTACTGAGT

CATGAGTACACTGT

ACTACTGA

Consensus

ACTACTGAGTACTGAGTACTGCATGAGTACACTGT

EXAMPLE

Target Genome ATTTGCGCAGAGACCTAAGGCATTAGCTTGGCCCTAAAG

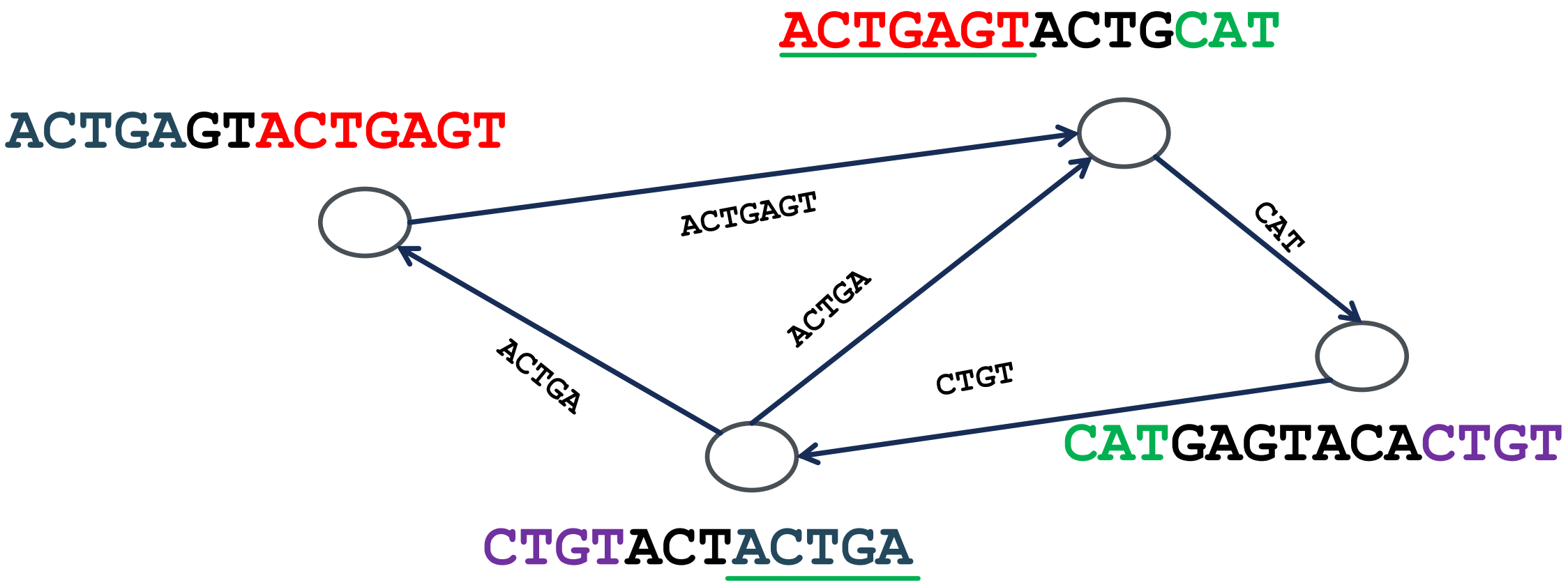
Reads ATTTGC AGAGACCTAAG TTAGCTTGGC AAG
TGCGCAGA TGGCCCTAA

Overlapping ATTTGC AGAGACCTAAG TTAGCTTGGC AAG
TGCGCAGA TGGCCCTAA

Contigs ATTTGCGCAGAGACCTAAG TTAGCTTGGCCCTAAAG

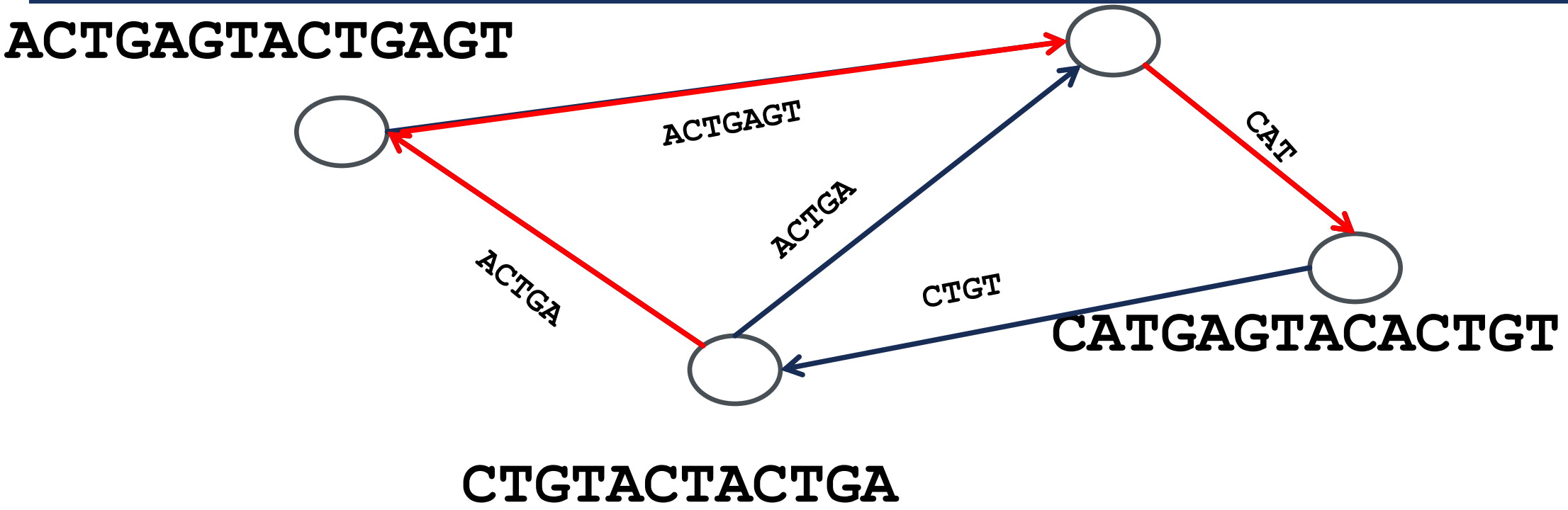
Image credit:
https://towardsdatascience.com/genome-assembly-the-holy-grail-of-genome-analysis-fae8fc9ef09c

PROBLEM FORMULATION



PROBLEM FORMULATION

ACTGAGTACTGCAT



CTGTACTACTGAGTACTGAGTACTGCATGAGTACACTGT

OVERLAP LAYOUT CONSENSUS

- **Nodes** = reads
- **Edges** = connection between overlapping reads
- Based on all **pairwise comparisons** .
- **Consensus** : combine the overlapping reads in the graph.
- **Layout** : find Hamiltonian path in the graph (order of visiting nodes in the graph) the nodes.
- **Programs using OLC:** Arachne , Celera , newbler ,Edena ,PCAP

GREEDY GRAPH

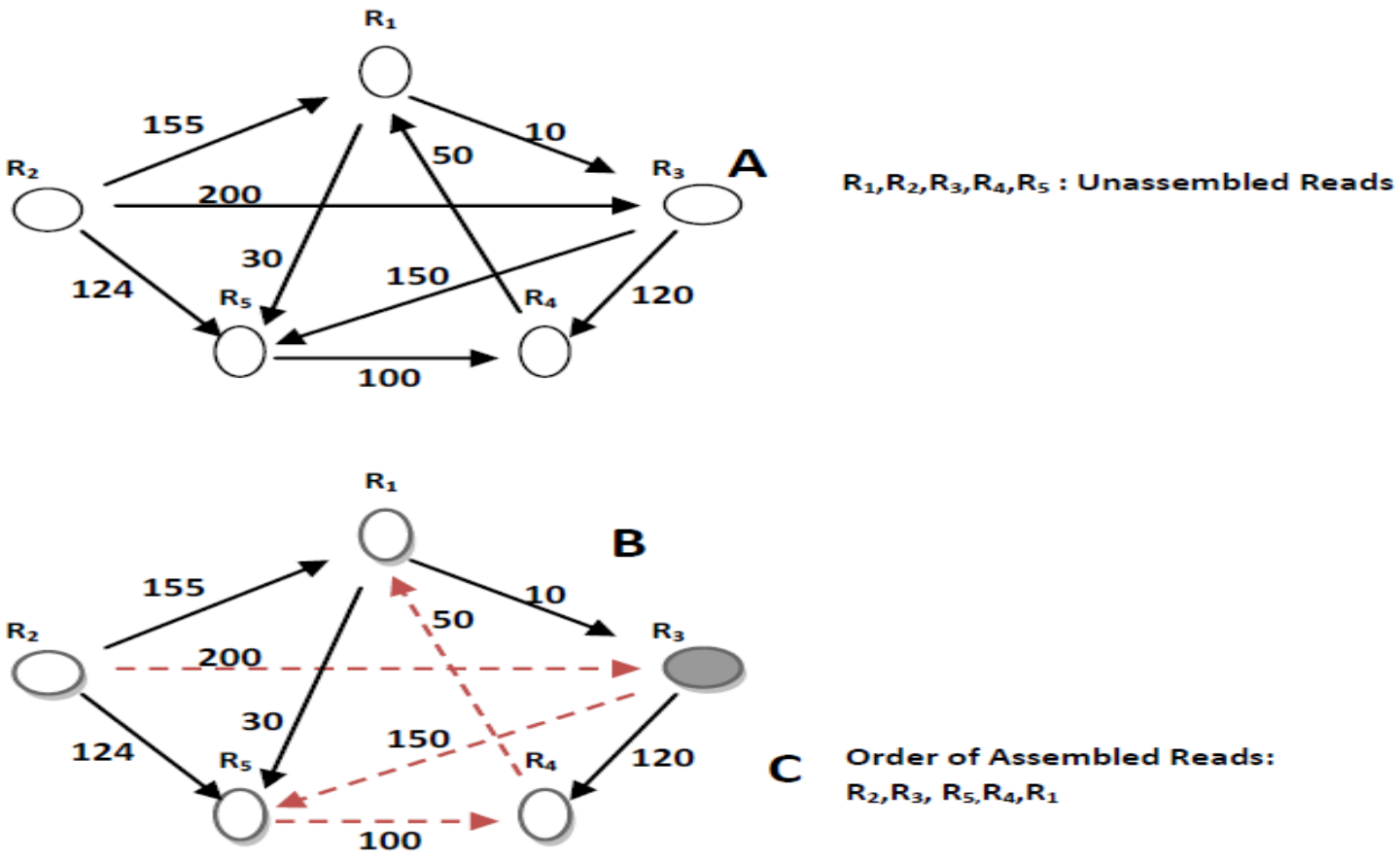
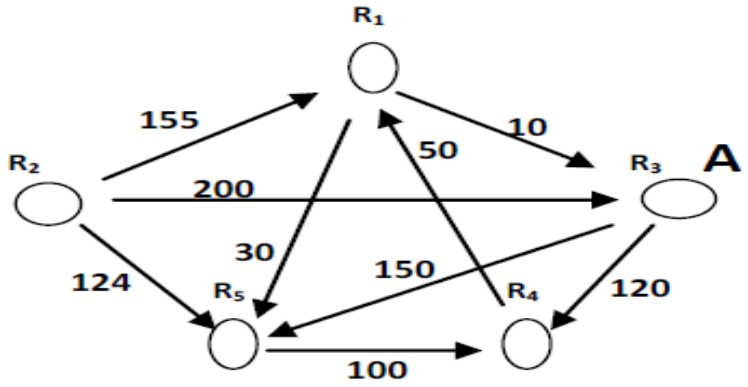
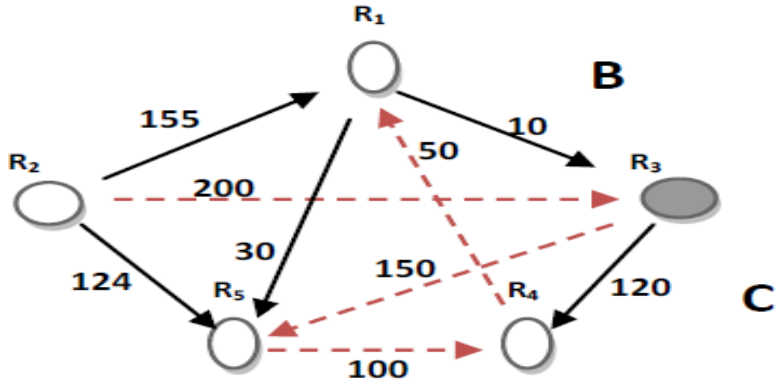


Image credit: El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput Biol.* 2013;9(12):e1003345. doi:10.1371/journal.pcbi.1003345

GREEDY GRAPH



R_1, R_2, R_3, R_4, R_5 : Unassembled Reads



Order of Assembled Reads:
R_2, R_3, R_5, R_4, R_1

Image credit: El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput Biol.* 2013;9(12):e1003345. doi:10.1371/journal.pcbi.1003345

DE BRUIJN GRAPH

k=4

$R_1 = \boxed{GACTGTA}$

GACT
ACTG
CTGT
TGTA

$R_2 = ACTGTAC$

ACTG
CTGT
TGTA
GTAC

$R_3 = GACTGCA$

GACT
ACTG
CTGC
TGCA

k=4

L=7

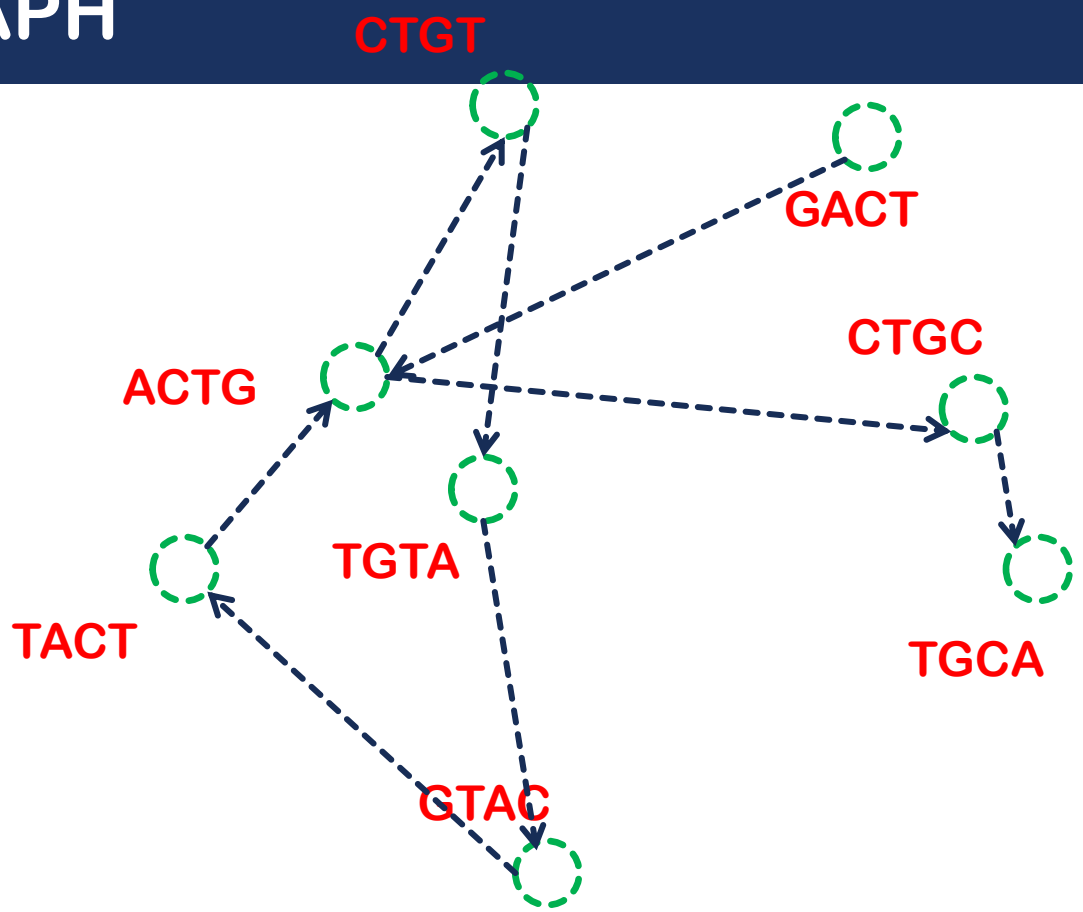
L-k+1=4 kmers

DE BRUIJN GRAPH

$R_1 = \mathrm{GACTGTA}$

$R_2 = \mathrm{ACTGTAC}$

$R_3 = \mathrm{GACTGCA}$

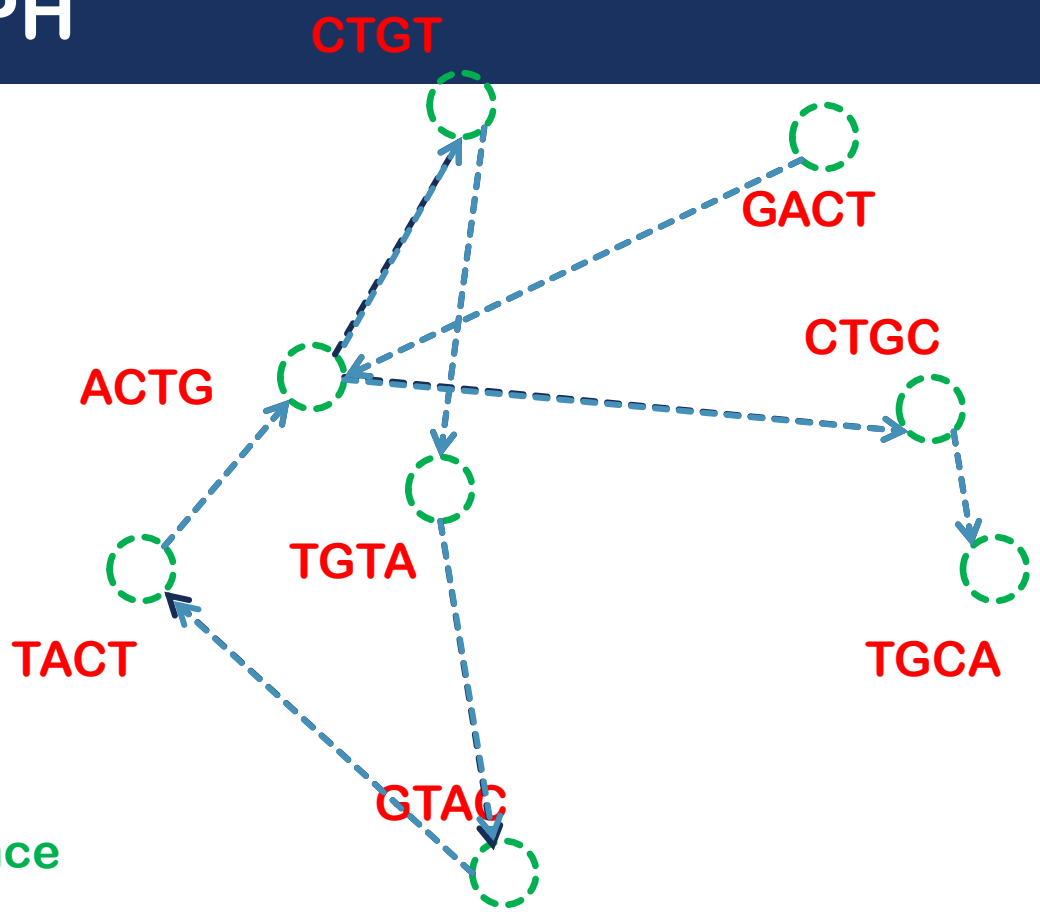


DE BRUIJN GRAPH

$R_1 = \text{GACTGTA}$

$R_2 = \text{ACTGTAC}$

$R_3 = \text{GACTGCA}$



Fixed overlap k-1

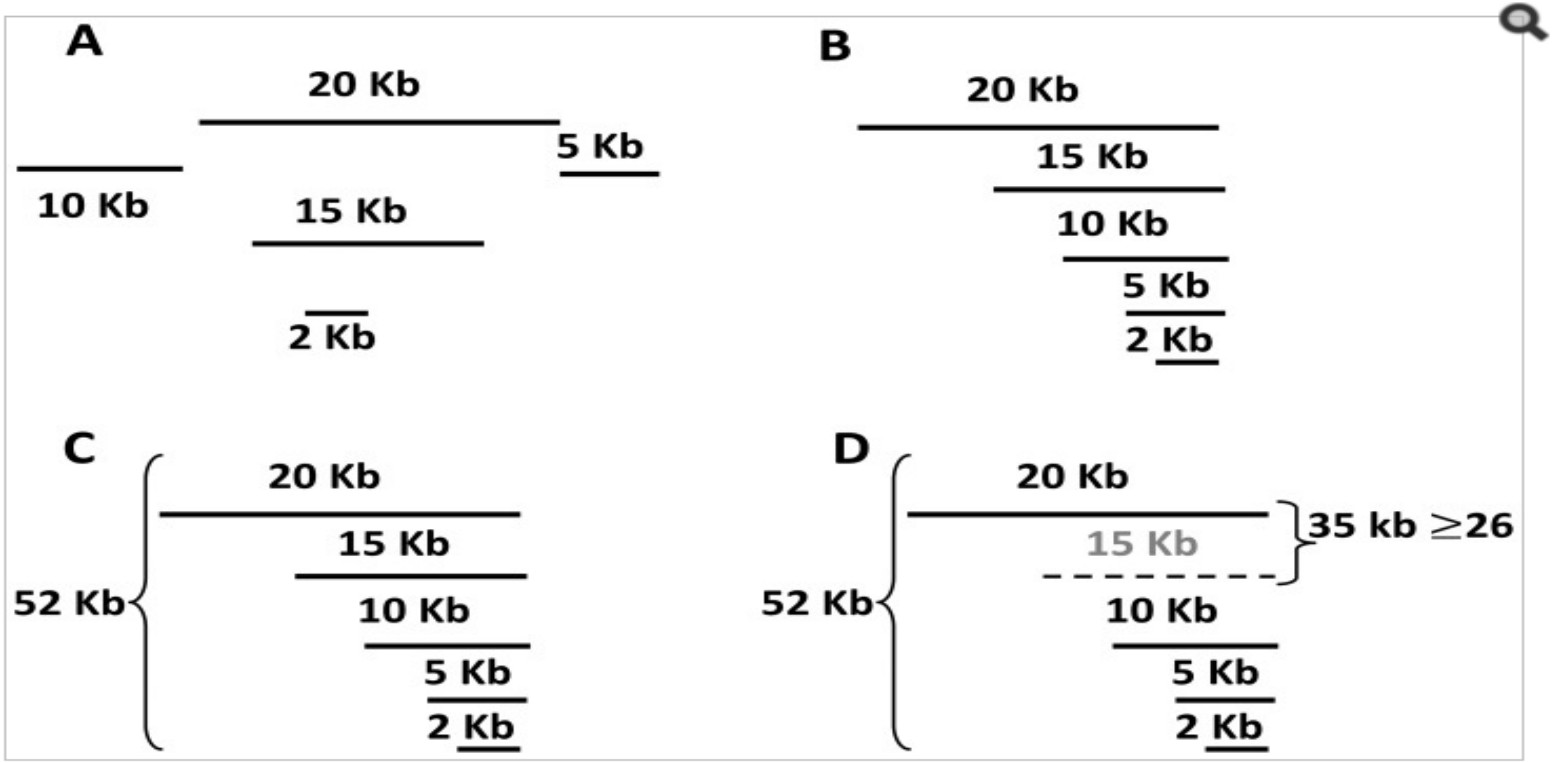
Repeated kmers used only once

GACTGTACTGCA

DE BRUIJN GRAPH

- **Nodes** = kmers
- **Edges** = overlapped kmers by k-1 chars.
- **Consensus** : combine the overlapping kmers in the graph.
- **Layout** : find Eulerian path in the graph (order of visiting Edges in the graph).
- **Programs: Velvet, ABySS, ALLPATHS-LG, SOAPdenovo**

N_{50} SCORE



N_{50} calculation method.

(A) Set of contigs with their length. (B) Contigs are sorted in descending order. (C) Lengths of all contigs are added (20+15+10+5+2=52 kb) and divided by 2 (52/2=26 kb). (D) Lengths are added again until the sum exceeds 26 kb, and hence exceeds 50% of the total length of all contigs: 20+15=35 kb≥26; then, N_{50} is the last added contig, which is 15 kb.

Image credit: El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput Biol.* 2013;9(12):e1003345. doi:10.1371/journal.pcbi.1003345

PLOS COMPUTATIONAL BIOLOGY

Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges

Sara El-Metwally, Taher Hamza, Magdi Zakaria, Mohamed Helmy

Published: December 12, 2013 • https://doi.org/10.1371/journal.pcbi.1003345

Article | Authors | Metrics | Comments | Media Coverage

Abstract

Decoding DNA symbols using next-generation sequencers was a major breakthrough in genomic research. Despite the many advantages of next-generation sequencers, e.g., the high-throughput sequencing rate and relatively low cost of sequencing, the assembly of the reads produced by these sequencers still remains a major challenge. In this review, we address the

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003345

OXFORD ACADEMIC

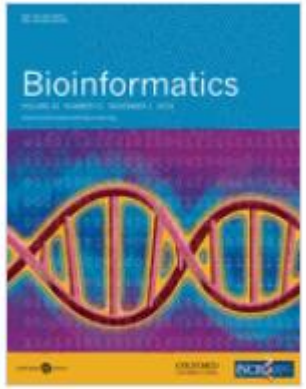
Bioinformatics

LightAssembler: fast and memory-efficient assembly algorithm for high-throughput sequencing reads FREE

Sara El-Metwally, Magdi Zakaria, Taher Hamza Author Notes

Bioinformatics, Volume 32, Issue 21, 1 November 2016, Pages 3215–3223,
https://doi.org/10.1093/bioinformatics/btw470
Published: 13 July 2016 **Article history** ▾

PDF Split View Cite Permissions Share ▾



Volume 32, Issue 21
1 November 2016

Motivation: The deluge of current sequenced data has exceeded Moore's Law, more than doubling every 2 years since the next-generation sequencing (NGS) technologies were

https://academic.oup.com/bioinformatics/article/32/21/3215/2415317

Purchase PDF

PODCAST

Review Article

A Roadmap to Sequence Assembly Evaluation Tools

(E-pub Ahead of Print)

Author(s): Sara El-Metwally*, Eslam Hamouda, Mayada Tarek

Journal Name: Current Bioinformatics

DOI : 10.2174/1574893615999201111140419

Journal Home

https://www.youtube.com/watch?v=boWiht0CTiw&list=UUFzSVHgGkjFW2Q8WV8-gzUw&index=7

https://www.eurekaselect.com/187845/article

Thank you!