De Novo Assembly

London School of Hygiene and Tropical Medicine

Matthew.Higgins@lshtm.ac.uk

Outline

This practical has 3 core objectives:

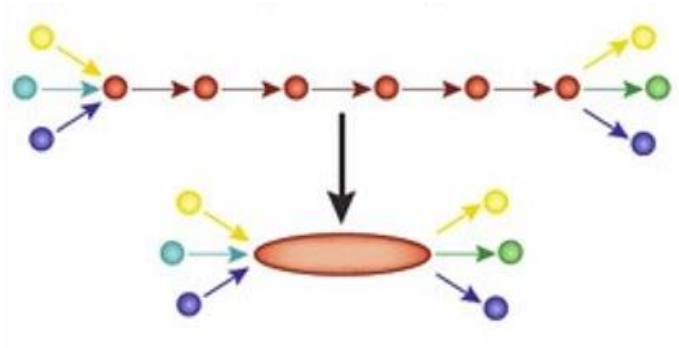
- Introduce assembly.
- Highlight applications of assembly.
- Introduce assembly algorithms & principles.

What is De Novo Assembly?

De Novo Assembly is process of reconstructing a complete genome from short fragments of DNA, without using a reference genome for guidance.

(1)

Find overlaps between fragments.



(2)

Assemble fragments into contigs.

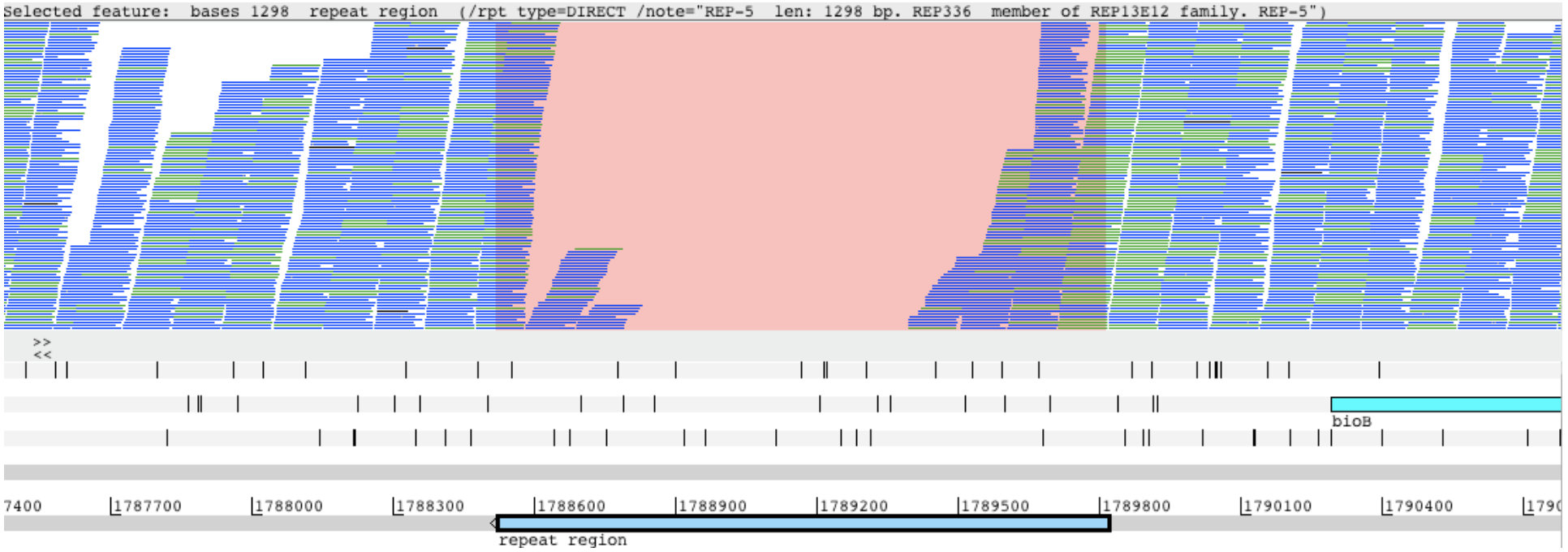
(3)

Assemble contigs into scaffolds.

Why Assemble?

- No reference genomes for target organism.
- Highly variable / unstable regions in target organism.
- To investigate large structural variants, including:
    - Insertions.
    - Deletions
    - Inversions
- To investigate novel transcripts in a transcriptome.

Why Assemble?

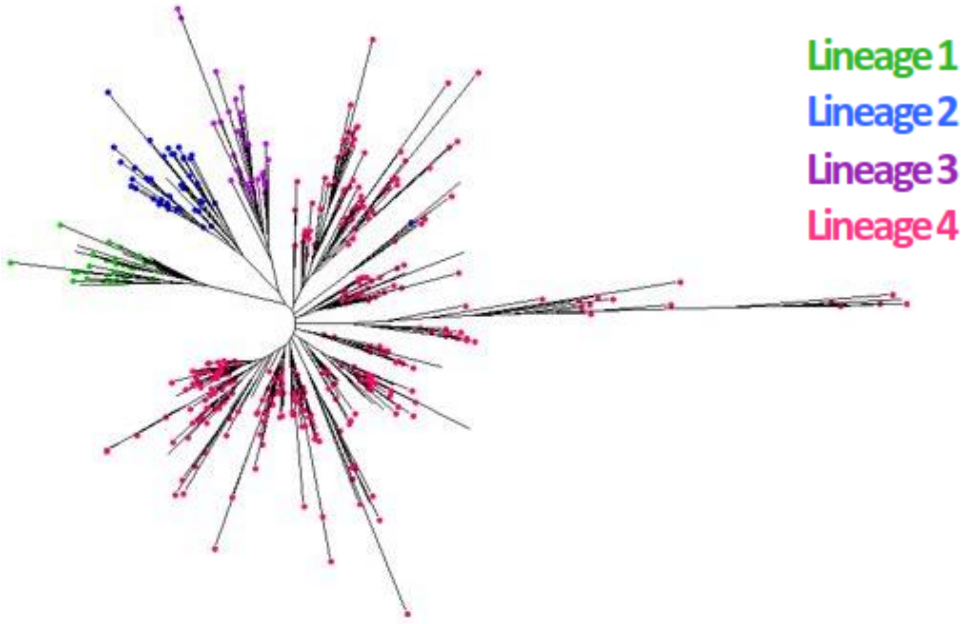


Overcome limitations of traditional mapping approaches which may fail to cover **repeat or highly variable regions,** specifically for short read data.

Assembly Applications

PE/PPE genes in *Mycobacterium tuberculosis (TB)* are typically excluded in downstream analysis due to:

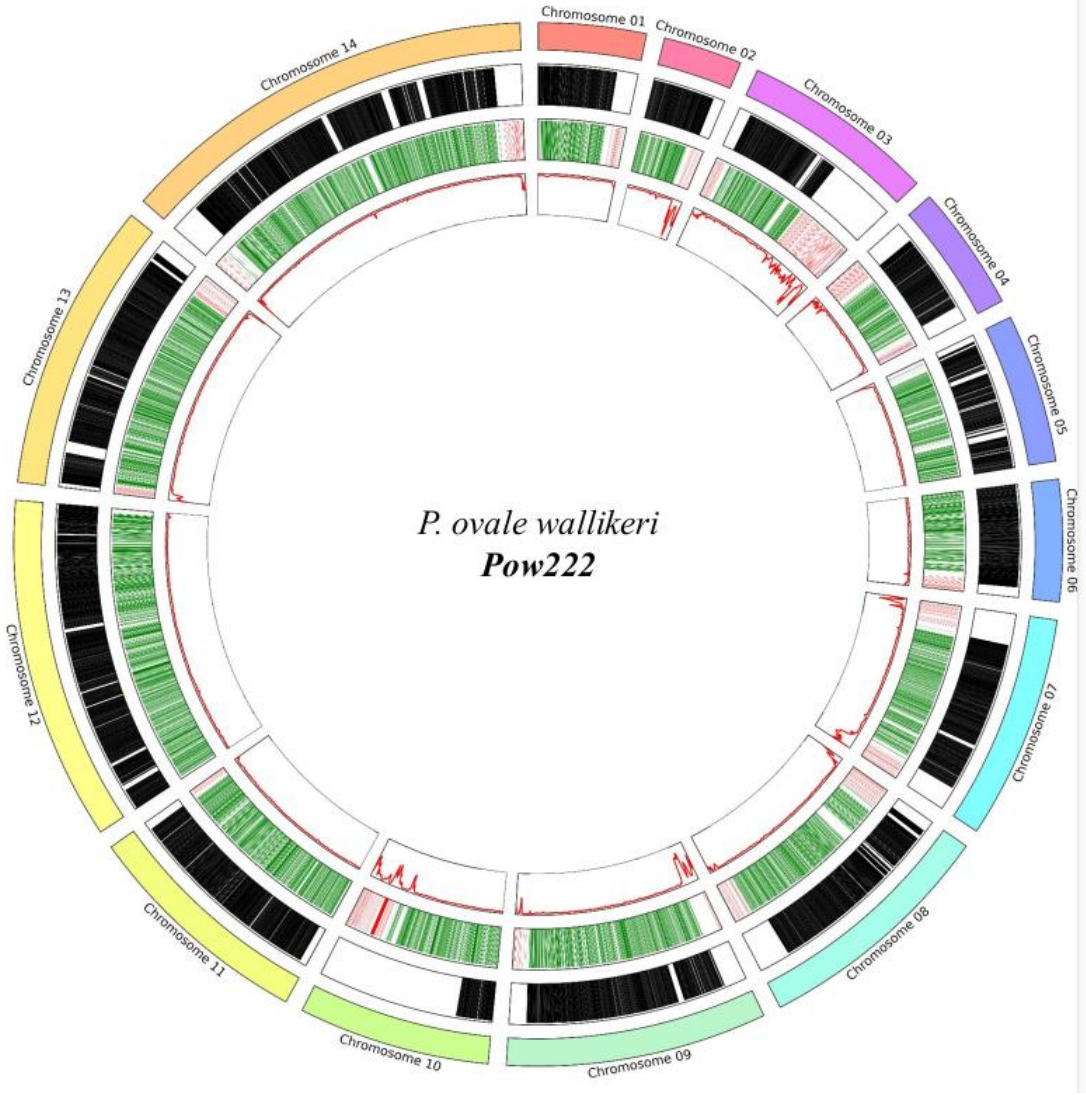
- High GC content
- Repetitive nature



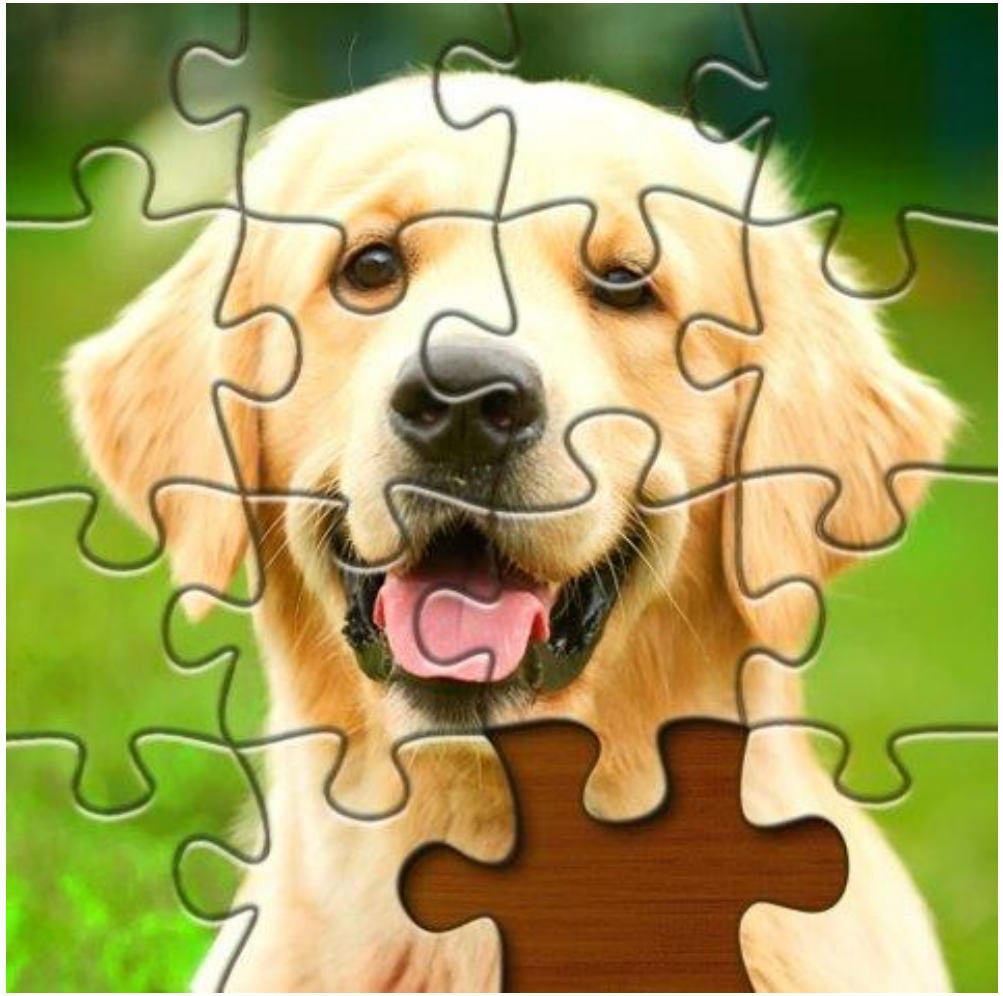
An assembly approach allows us to reconstruct these gene enabling accurate downstream analysis across 500 clinical samples.

Assembly Applications

Creation of new ***P. ovale wallikeri*** reference genome, leading to the recovery of Chromosome 10 which was previously missing in old reference.



Genomic Jigsaw Puzzle



Complete genome = *A*ssembled jigsaw puzzle.

Read = individual jigsaw piece

De novo Assembly can be thought of assembling the puzzle **without** knowing what the final picture should be.

Compared to **Mapping** where we know the final picture.

Assembly Challenges

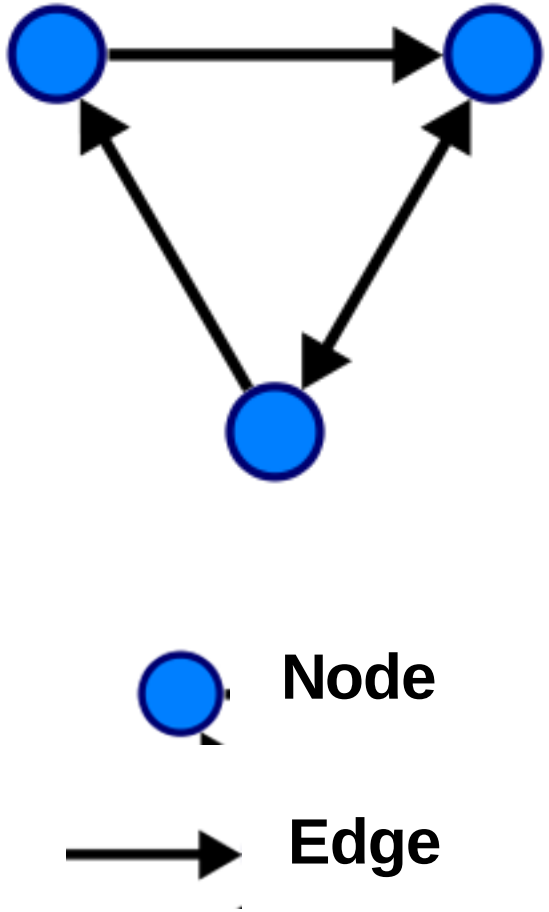
Challenges:

- Individual reads may fit together in more than one way and so need to optimise assembly to minimise introduction of errors.
- Require high levels of coverage for accurate and complete assembly.
- Requires stringent quality control to prevent error introduction.
- Computationally intensive.

Assembly Algorithms

To understand assembly algorithms you should understand basics of **Graph theory.**

A **Graph i**s a mathematical structure that is used to represent objects and the connections between them.

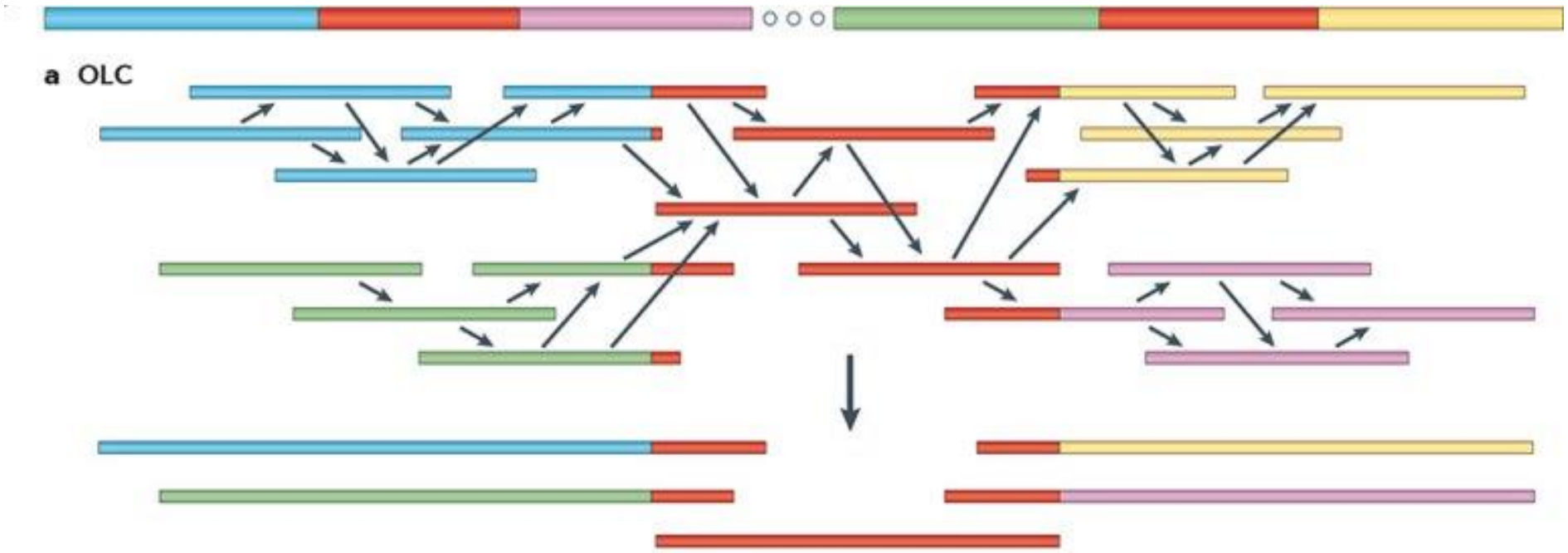


Overlap Layout Consensus (OLC) Algorithm

Pairwise alignment of all reads. Then a graph is built using**:**

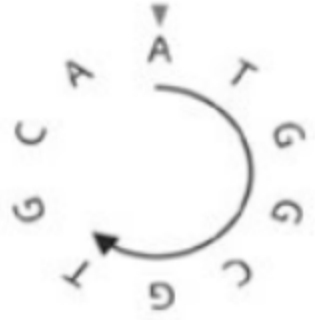
Reads = Nodes

Overlap = Edges

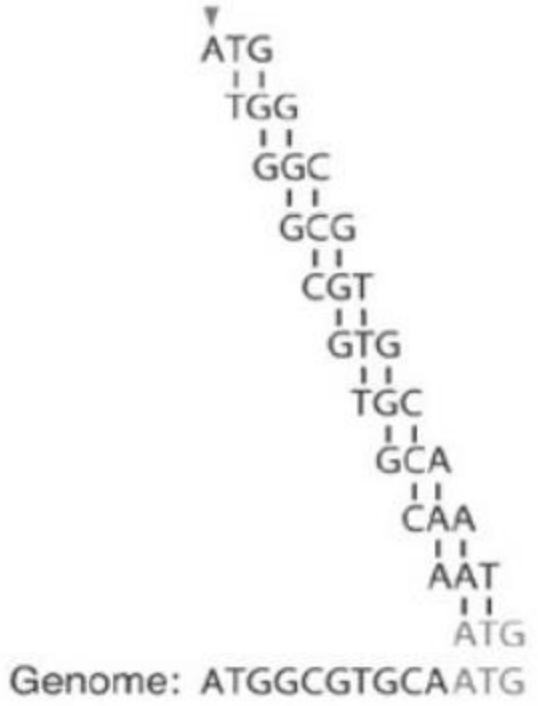


From the construction of the graph we can begin to bundle **reads** into **contigs** as shown.

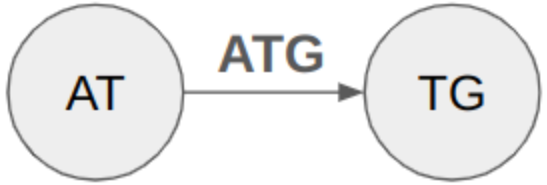
De Bruijn Graph Algorithm



(1)
Example DNA Sequence (e.g. read)



(2)
Split into kmer's via sliding window

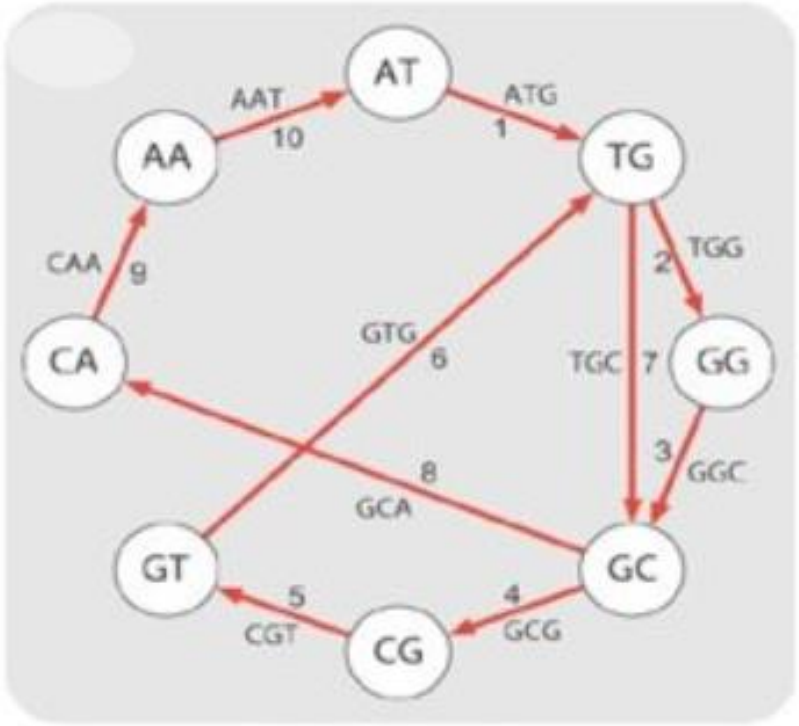


(3)
Split kmer into node & edges

De Bruijn Graph Algorithm

(4) Constructed graph using all kmers.

(5) Find the path which uses each edge once, this will stich kmers together and create our **assembly**.

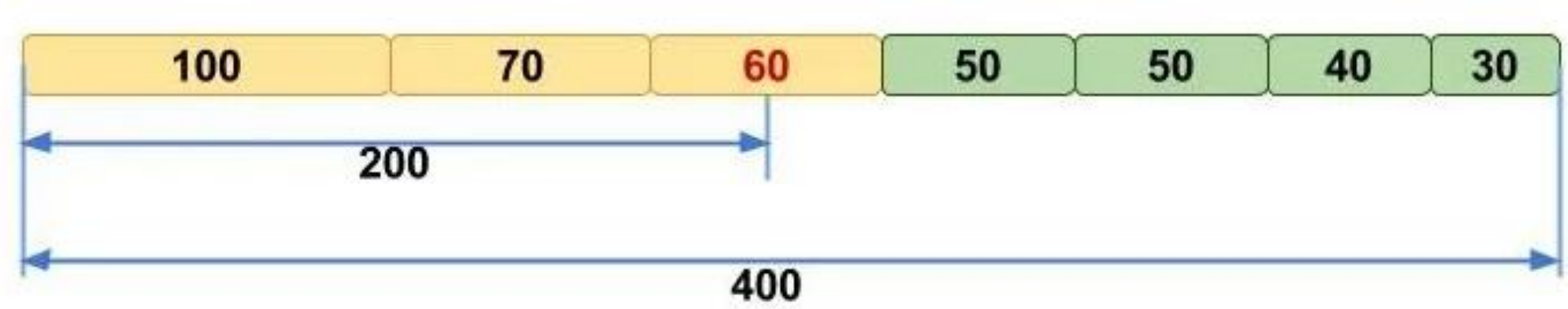


Eulerian cycle
Visit each edge once
(easier to solve)

Evaluating Assemblies

In the absence of a high-quality reference genome, assemblies are often evaluated on the basis of:

- **N50 Metric** = 50% of assembly is contained within contigs of this size or bigger..



Evaluating Assemblies

In the absence of a high-quality reference genome, assemblies are often evaluated on the basis of:

- **BUSCO Score** = Identifies presence and completeness of single copy orthologs expected.

- The proportion of reads that can be assembled.

Practical Overview

Perform de novo assembly of *Mycobacterium tuberculosis* and look to identify and validate structural variants.

1. Spades. *De Novo* assembly
2. ABACAS. Contig Ordering.
3. Artemis. Visualisation.
4. BLAST. Structural Variant Validation.