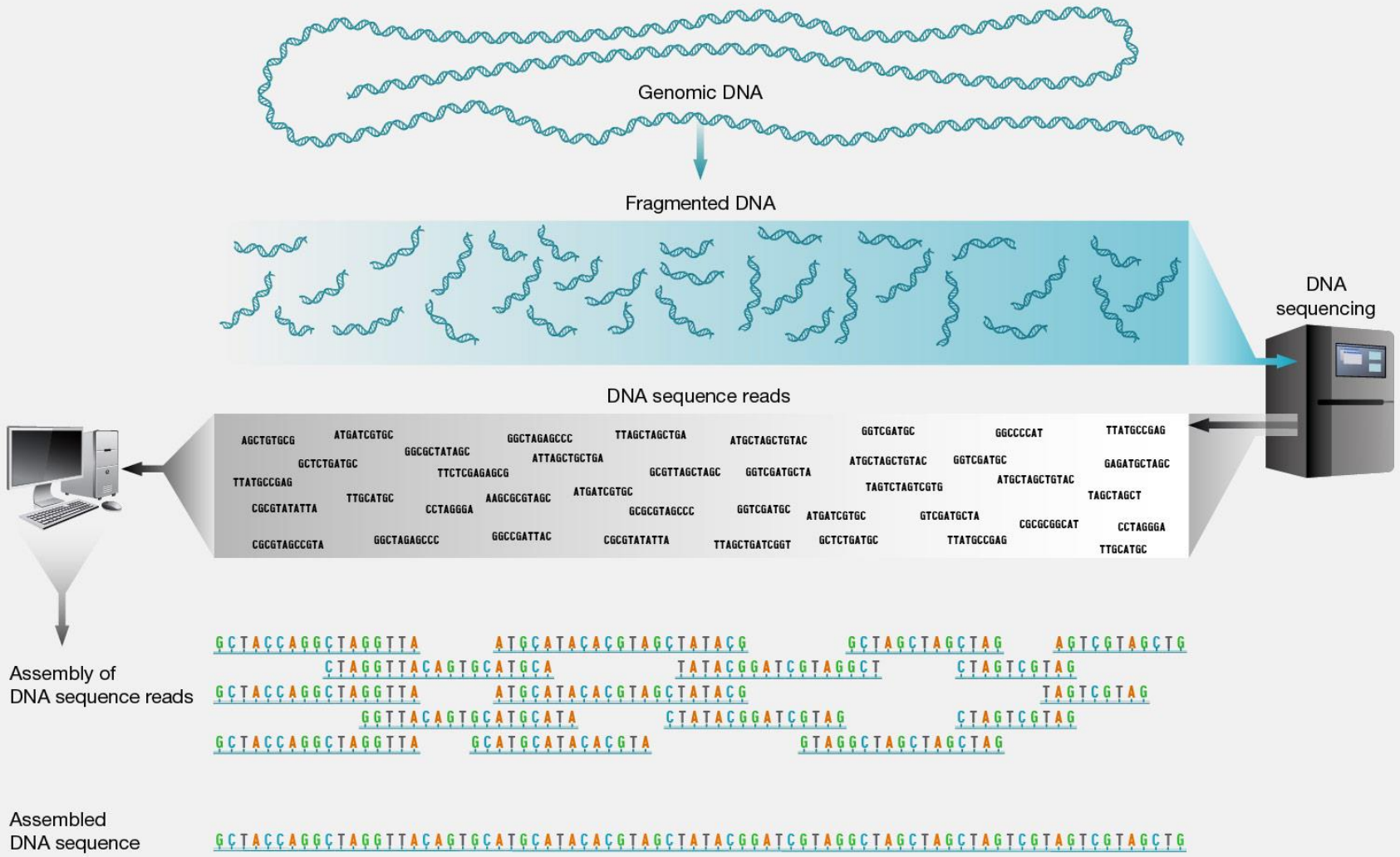
DNA assembly using De Bruijn graphs

University of Bologna

May 6, 2024


Genomic DNA
Fragmented DNA
DNA sequencing
DNA sequence reads
Assembly of
DNA sequence reads
Assembled
DNA sequence

DNA assembly methods

- Assemble fragments by **aligning** them with respect to a known reference sequence.

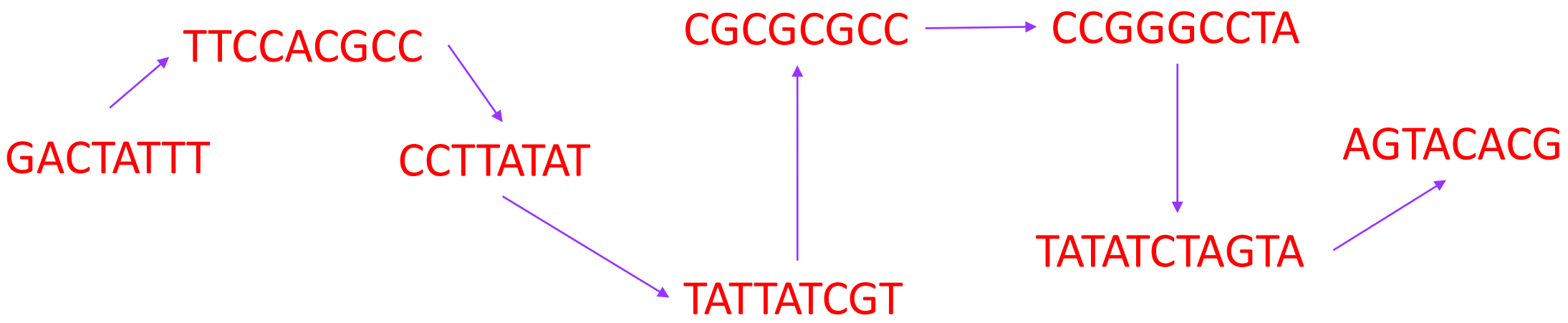
GACTATTTCCACGCCTTATATTATCGTACGCGCGCCGGGCCTATATCTAGTACACG

GACTATTT CCTTATAT CCGGGCCTA AGTACACG

TTCCACGCC TATTATCGT CGCGCGCC TATATCTAGTA

DNA assembly methods

- Assemble fragments based on **mutual overlaps** in sequence (*de novo* assembly)



GACTATTTCCACGCCTTATATTATCGTACGCGCGCCGGGCCTATATCTAGTACACG

De Bruijn graph

Sequence:

De Bruijn graph

Sequence: **AATCGACCGA**

5-mer:

De Bruijn graph

Sequence: **AATCGACCGA**

5-mer: **AATCG**

De Bruijn graph

Sequence: **AATCGACCGA**

5-mer:

AATCG
ATCGA
TCGAC
CGACC
GACCG
ACCGA

De Bruijn graph

Sequence: **AATCGACCGA**

5-mer:

AATCG → **AATC**, **ATCG**

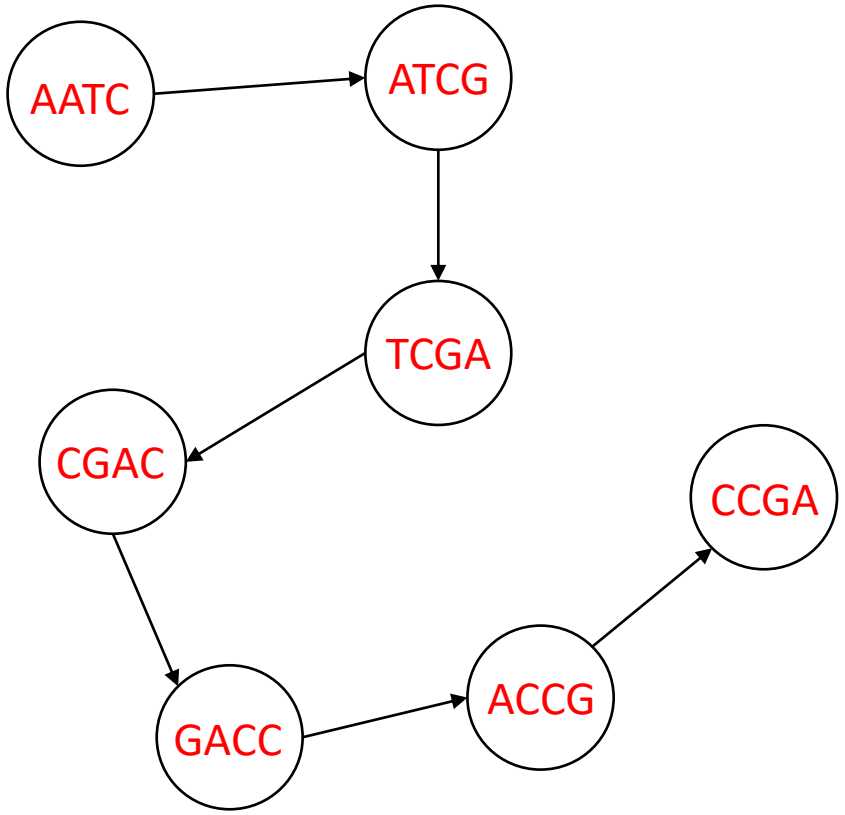
ATCGA → **ATCG**, **TCGA**

TCGAC → **TCGA**, **CGAC**

CGACC → **CGAC**, **GACC**

GACCG → **GACC**, **ACCG**

ACCGA → **ACCG**, **CCGA**



De Bruijn graph

Sequence: **AATCGACCGA**

5-mer:

AATCG → **AATC**, **ATCG**

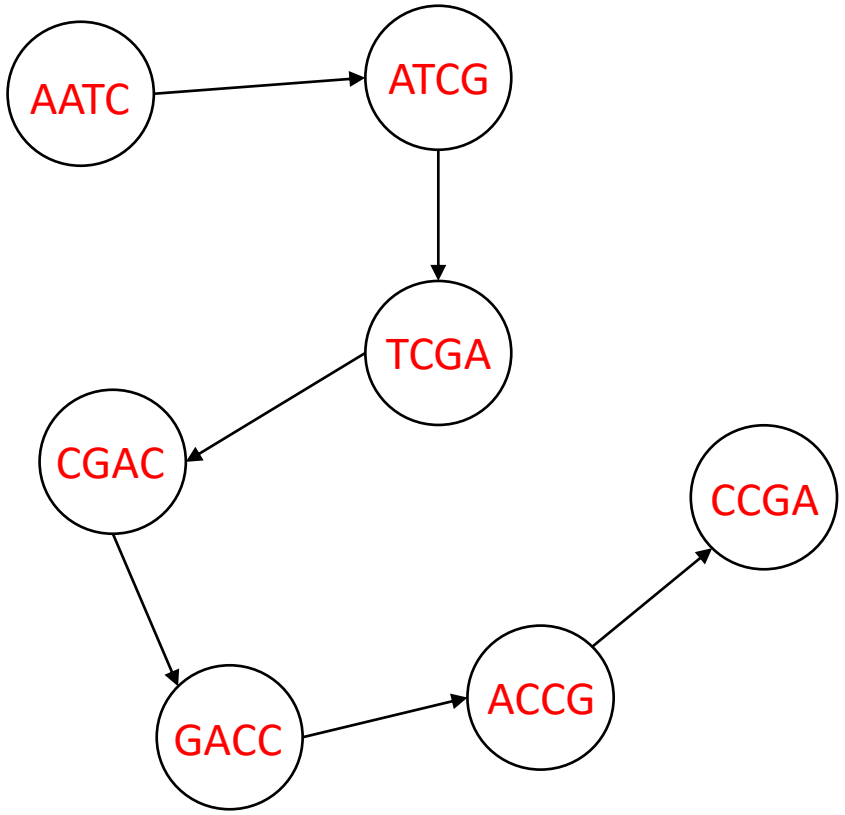
ATCGA → **ATCG**, **TCGA**

TCGAC → **TCGA**, **CGAC**

CGACC → **CGAC**, **GACC**

GACCG → **GACC**, **ACCG**

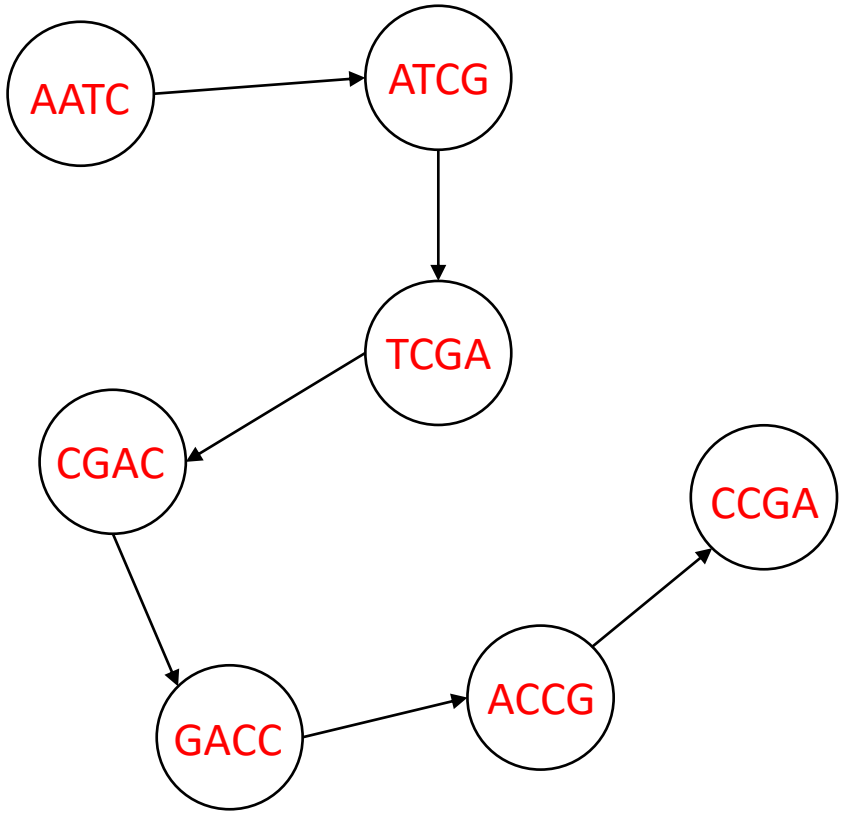
ACCGA → **ACCG**, **CCGA**



Sequence reconstruction

I have to find an **eulerian path** that is a path that visits **each edge** exactly **once**.

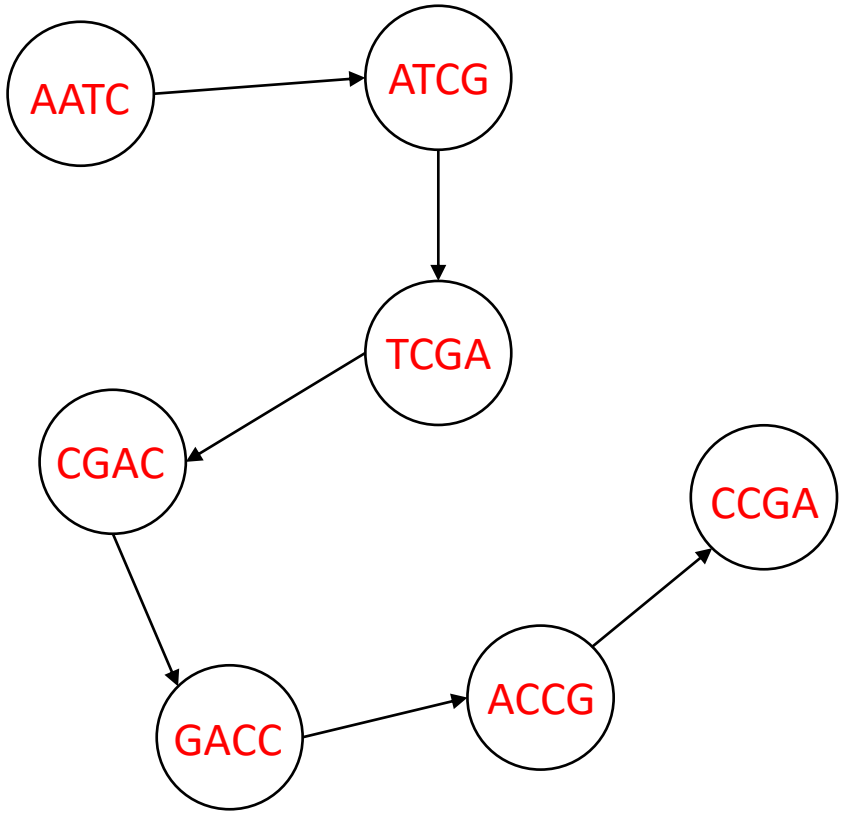
I have an **eulerian path** within a directed and connected graph, **if and only if** at most 2 nodes are **semi-balanced** and all other nodes are **balanced**.



Sequence reconstruction

A node is **balanced** if indegree equals outdegree.

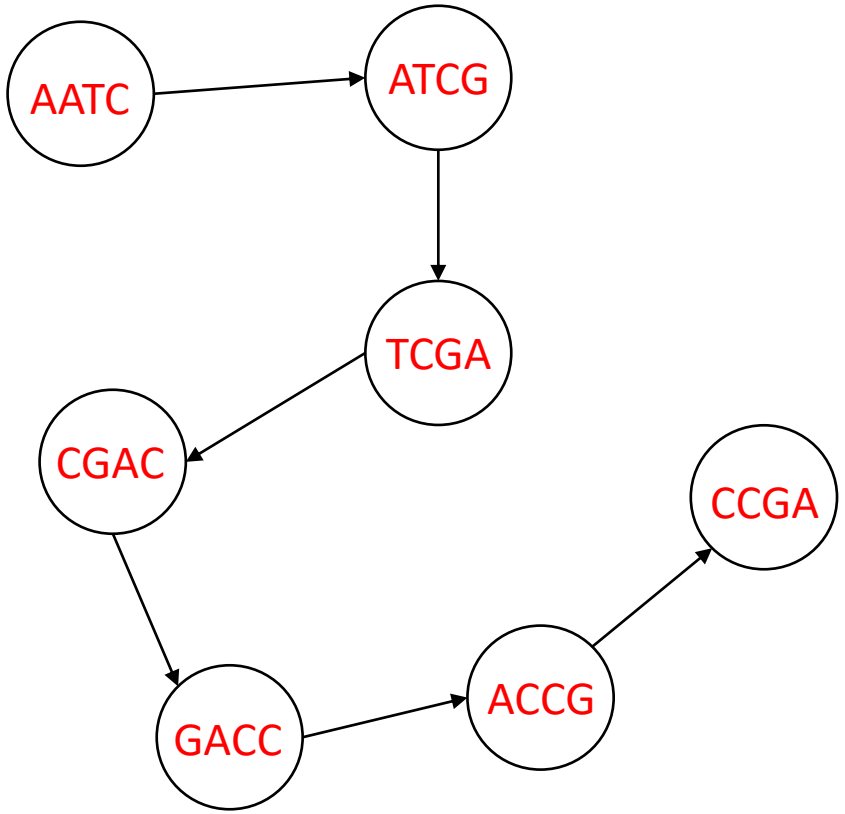
A node is **semi-balanced** if indegree differs from outdegree by 1.



Sequence reconstruction

How to assemble the (k-1)-mers to get the original sequence?

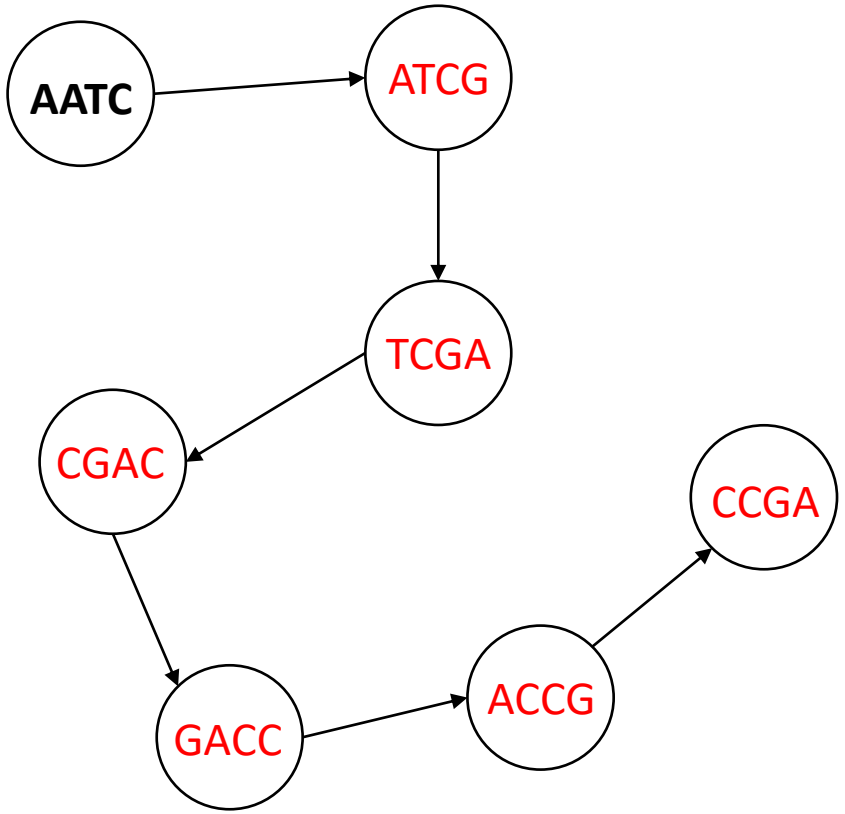
Original sequence: **AATCGACCGA**



Sequence reconstruction

How to assemble the (k-1)-mers to get the original sequence?

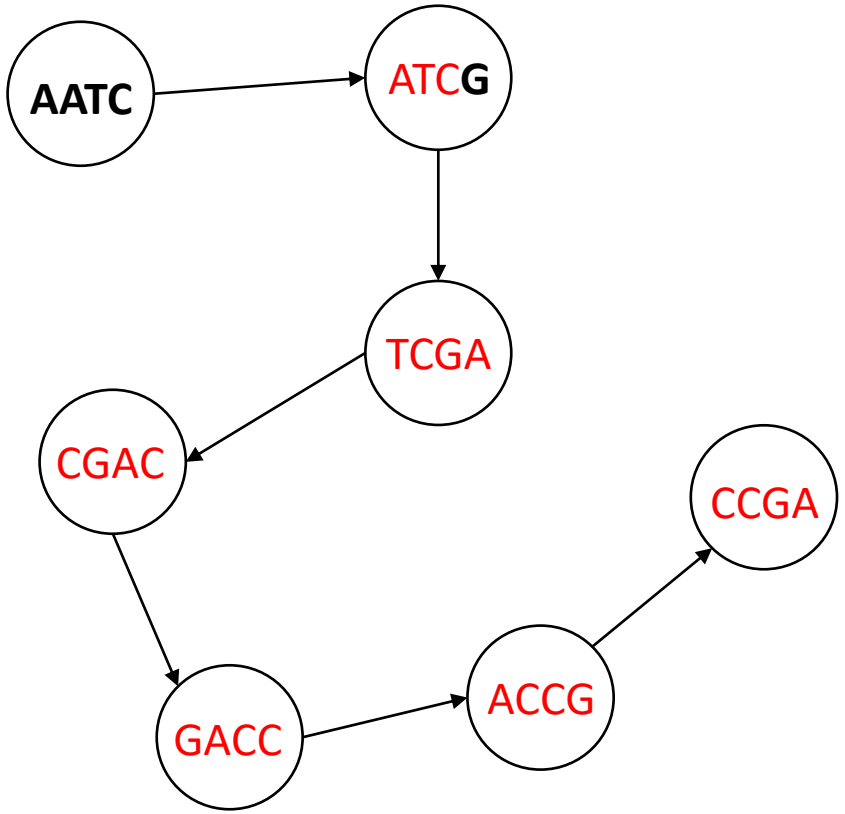
Original sequence: **AATCGACCGA**



Sequence reconstruction

How to assemble the (k-1)-mers to get the original sequence?

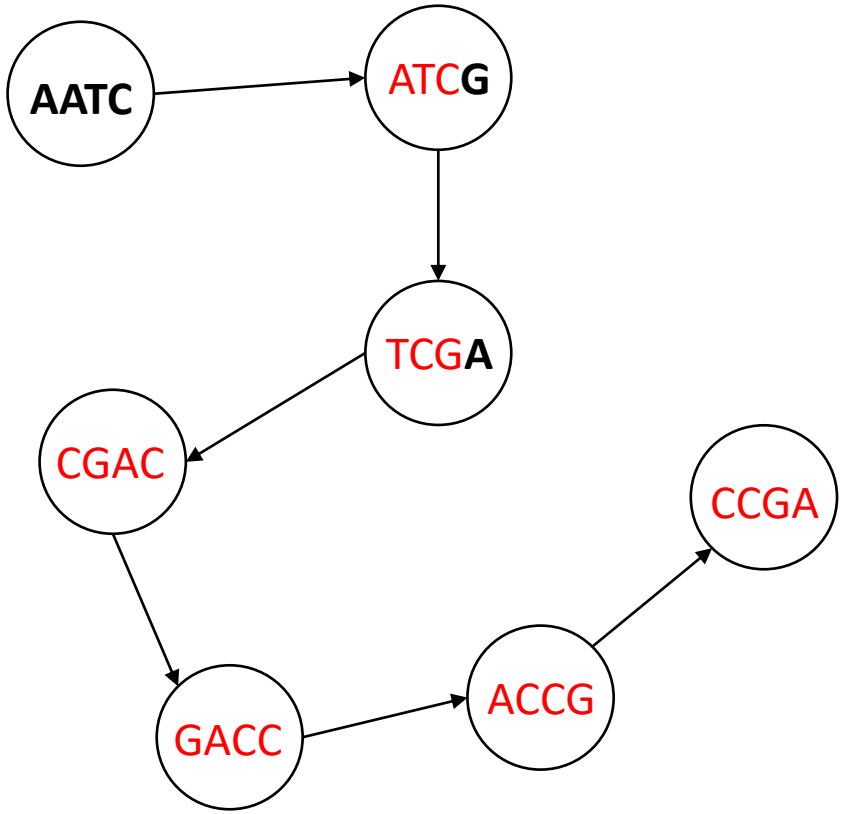
Original sequence: **AATCGACCGA**



Sequence reconstruction

How to assemble the (k-1)-mers to get the original sequence?

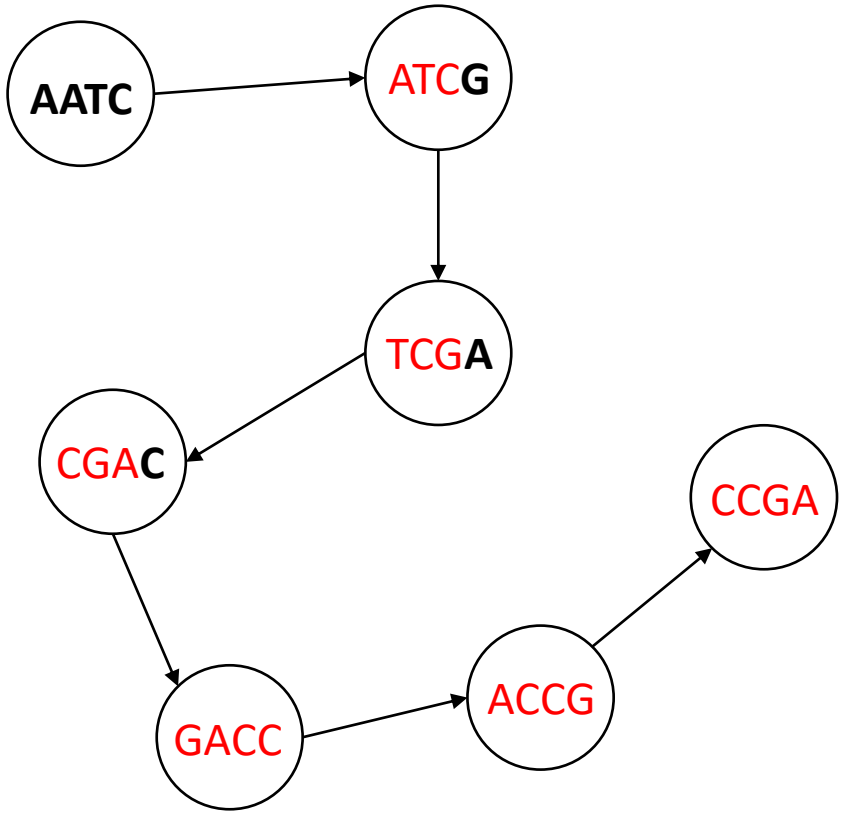
Original sequence: **AATCGACCGA**



Sequence reconstruction

How to assemble the (k-1)-mers to get the original sequence?

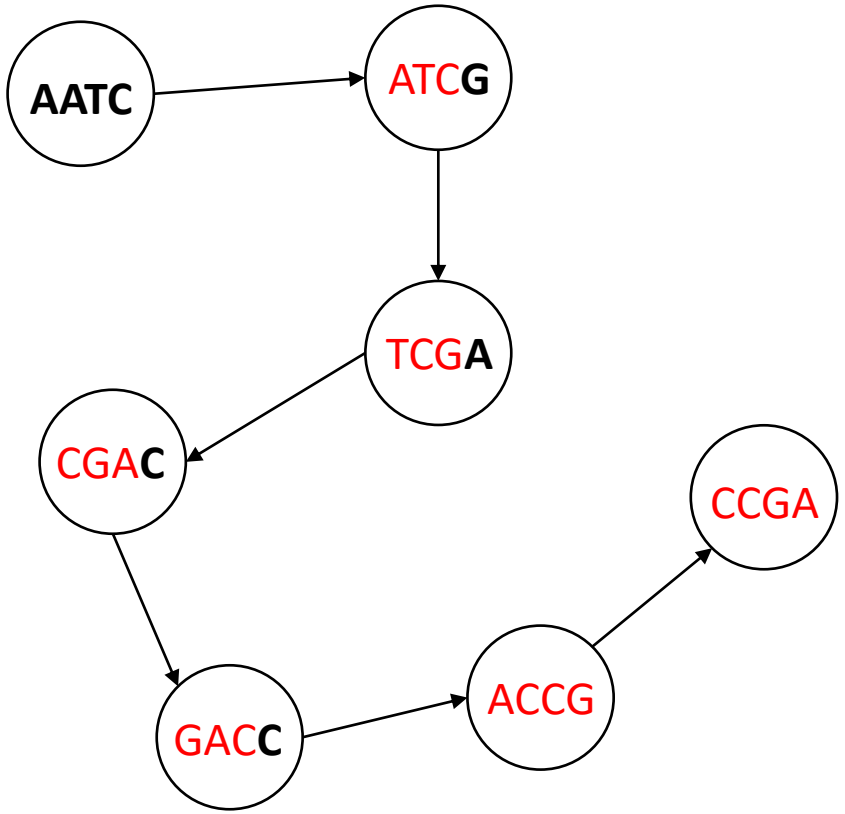
Original sequence: **AATCGACCGA**



Sequence reconstruction

How to assemble the (k-1)-mers to get the original sequence?

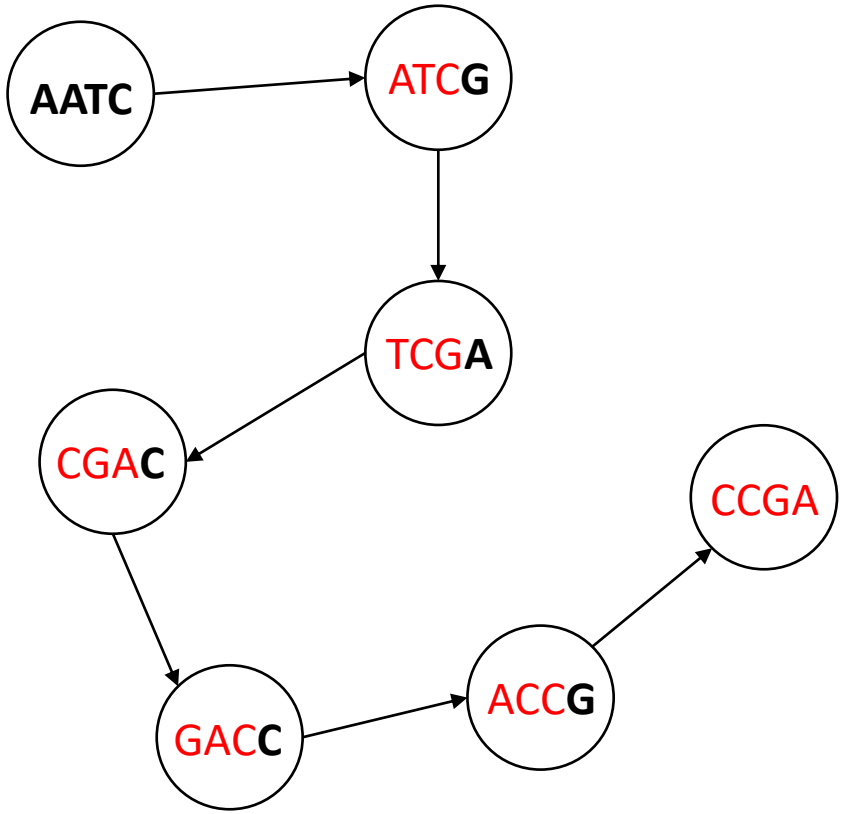
Original sequence: **AATCGACCGA**



Sequence reconstruction

How to assemble the (k-1)-mers to get the original sequence?

Original sequence: **AATCGACCGA**



Sequence reconstruction

How to assemble the (k-1)-mers to get the original sequence?

Original sequence: **AATCGACCGA**

Reconstructed sequence: **AATCGACCGA**