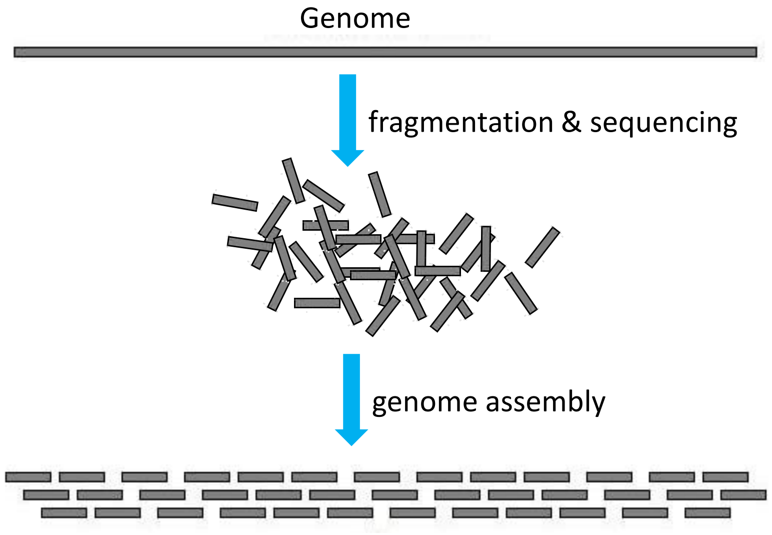
Genome Assembly

by Ahmet Sacan



Slides adapted from:
https://www.coursera.org/learn/dna-sequencing/lecture/lscgN/lecture-overlap-graphs

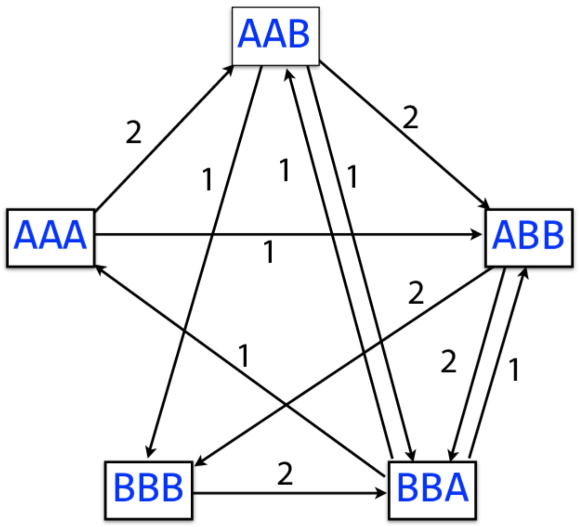
Genome Assembly

- Given Reads:
  - AAA, AAB, ABB, BBA, BBB
- What is the genome sequence?

- One solution:
  - AAAAABABBBBABBB

Two classes of methods

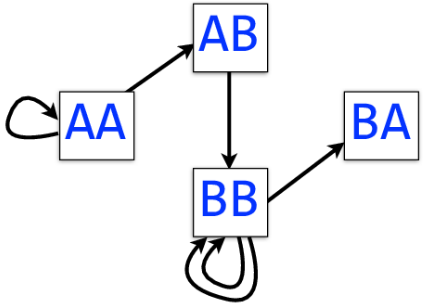
Overlap graph



Overlap-Layout-Consensus
(OLC) assembly

One node for each read

De Bruijn graph

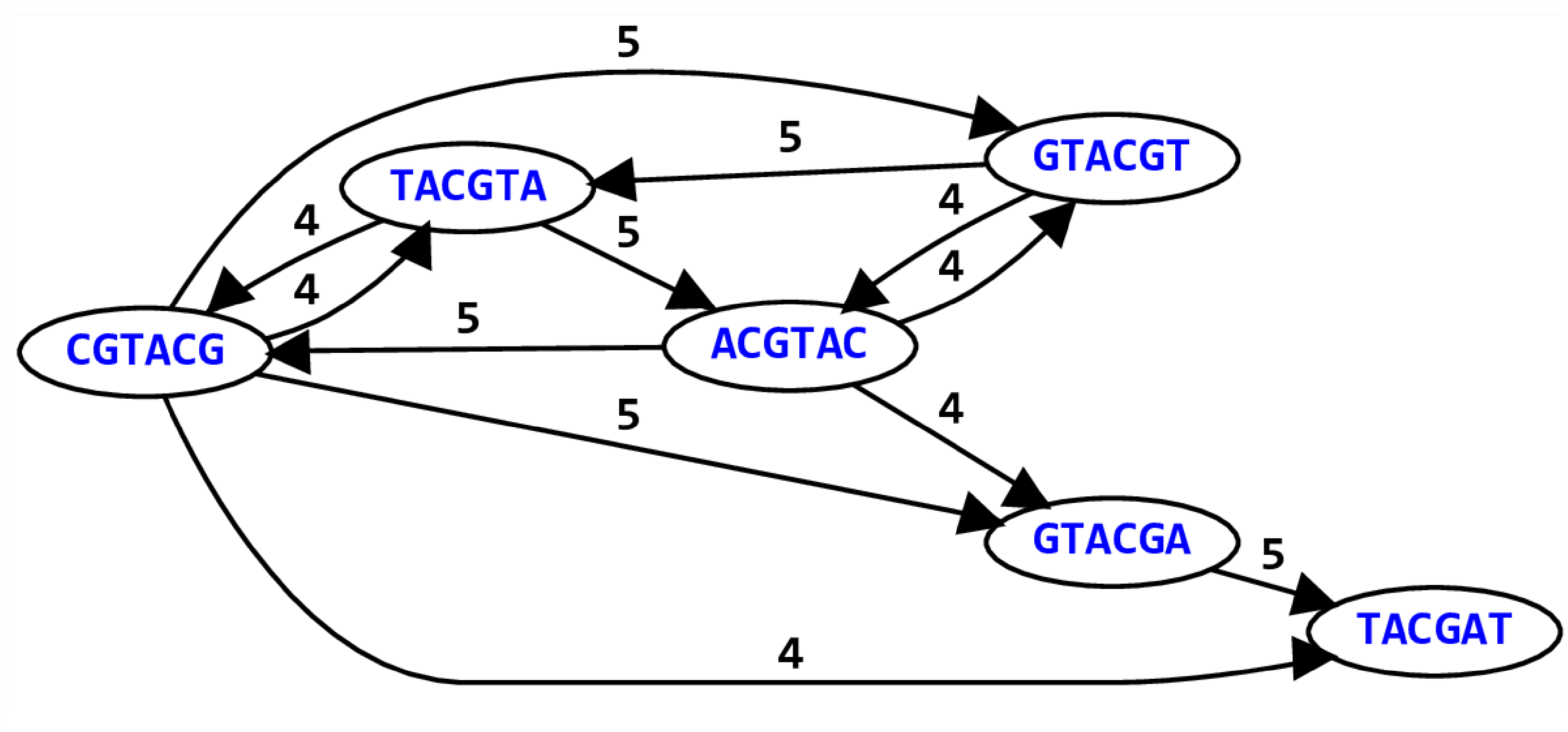


De Bruijn Graph based
(DBG) assembly

One edge for each read

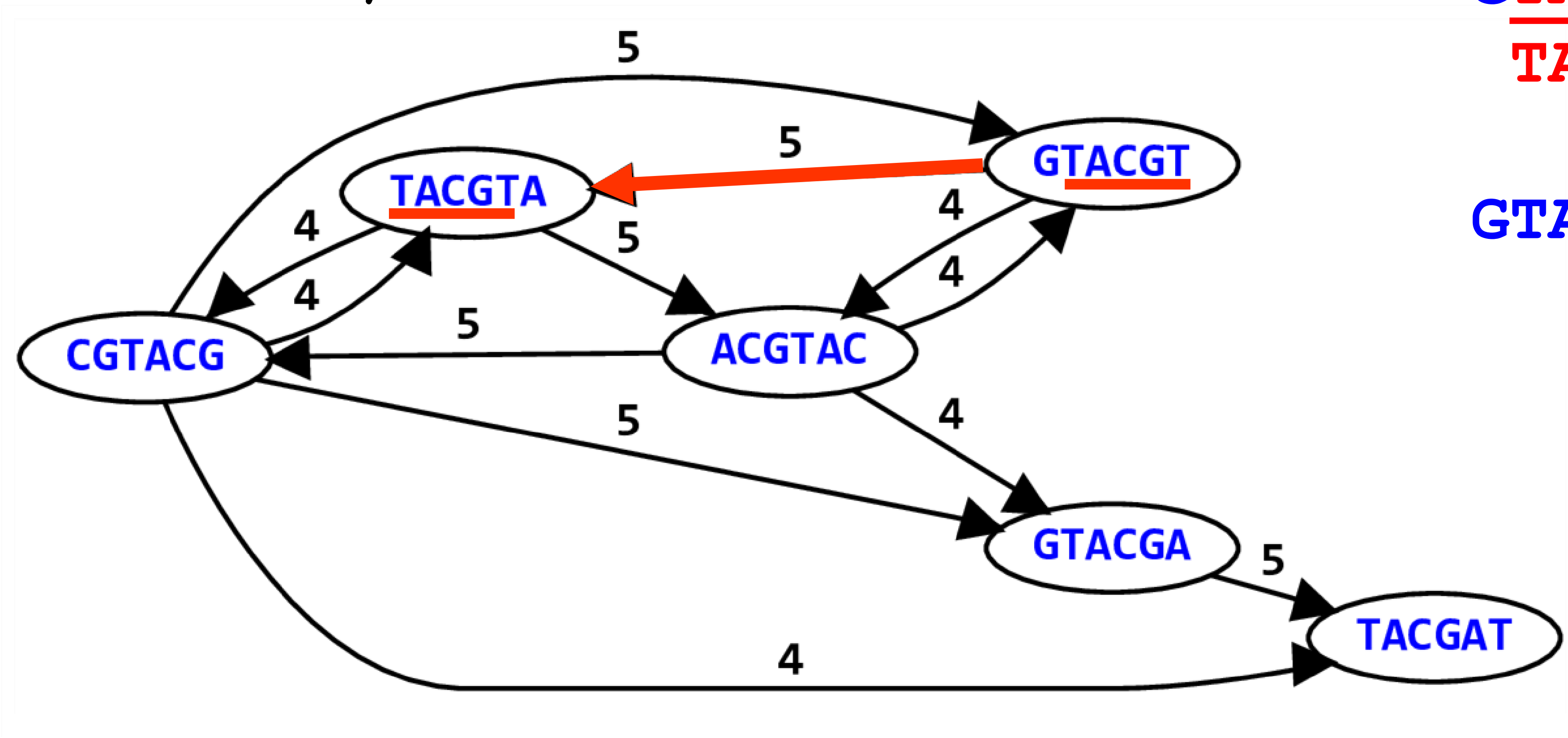
Overlap graph

- Reads: CGTACG, TACGTA, GTACGT, ACGTAC, GTACGA, TACGAT



Overlap graph

- Reads: CGTACG, TACGTA, GTACGT, ACGTAC, GTACGA, TACGAT

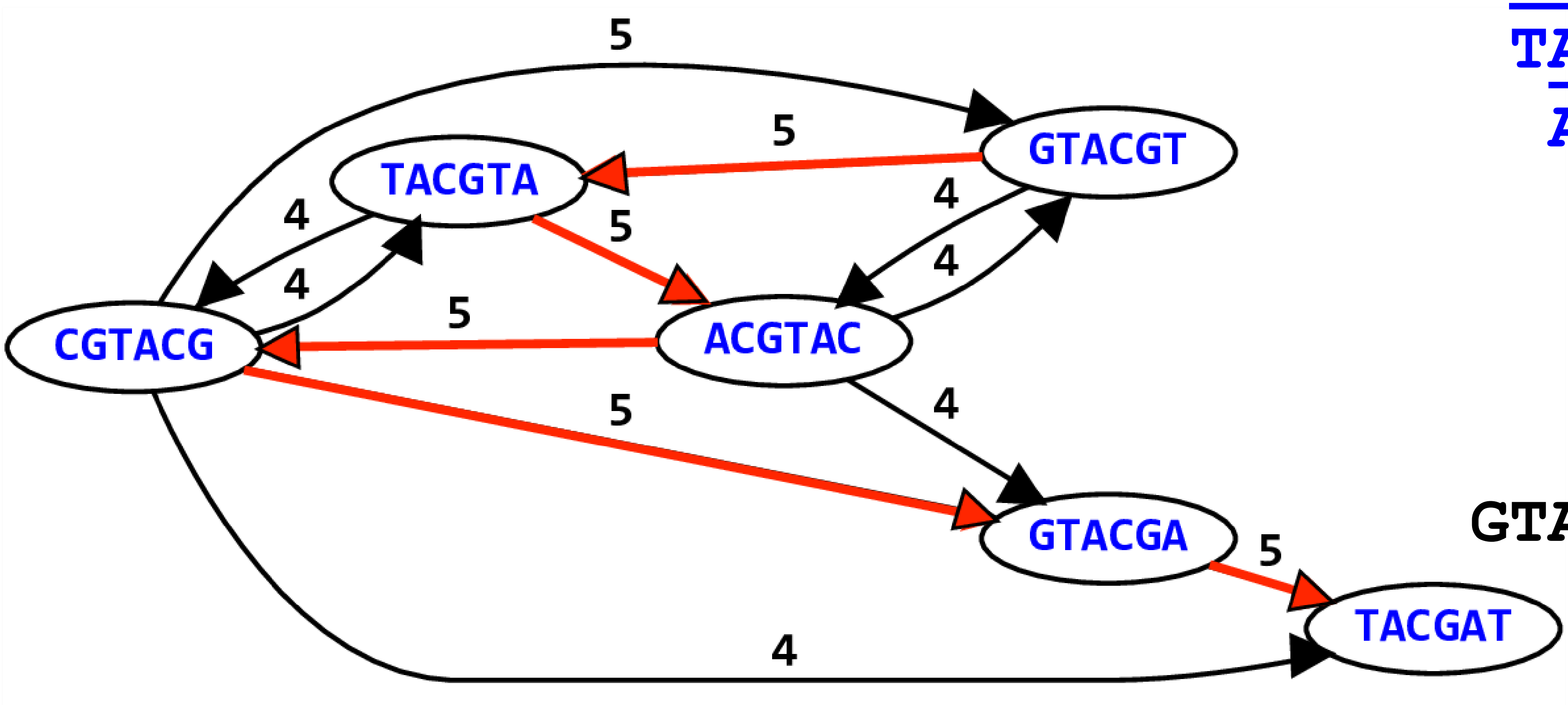


GTACGT
 TACGTA

GTACGTA

Reconstruction of Genome from Overlap Graph

- Find a walk that visits every node once. (Hamiltonian Path)

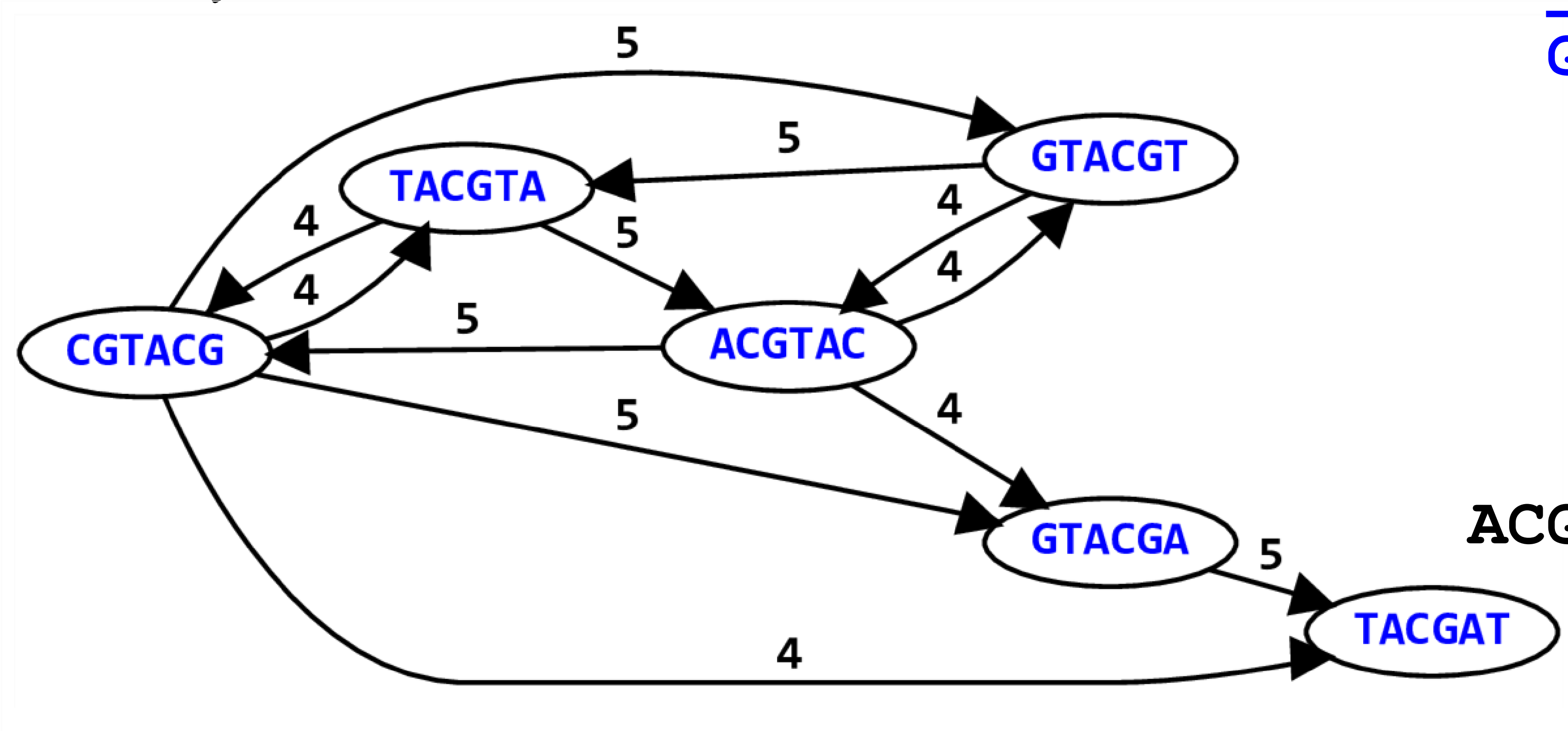


GTACGT
TACGTA
ACGTAC
CGTACG
GTACGA
TACGAT

GTACGTACGAT

Reconstruction of Genome from Overlap Graph

- Find a walk that visits every node once. (Hamiltonian Path)



ACGTAC
GTACGT
TACGTA
CGTACG
GTACGA
TACGAT

ACGTACGTACGAT

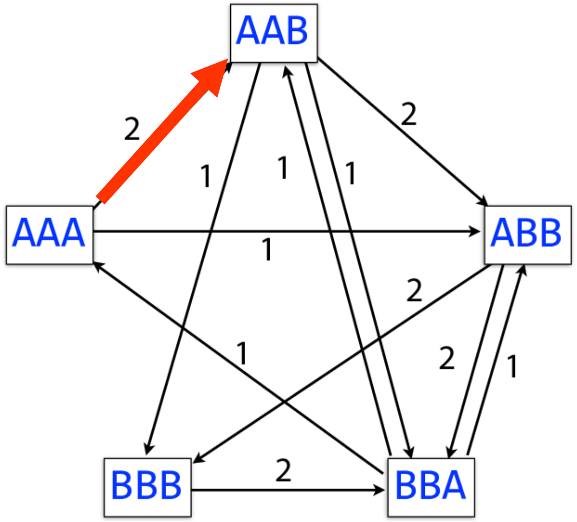
Shortest Common Superstring

- SCS: Given a set of substrings, find the shortest superstring that contains these substrings

- e.g., given reads:
  - AAA, AAB, ABB, BBA, BBB
- What is the shortest genome sequence?

Greedy Solution to SCS

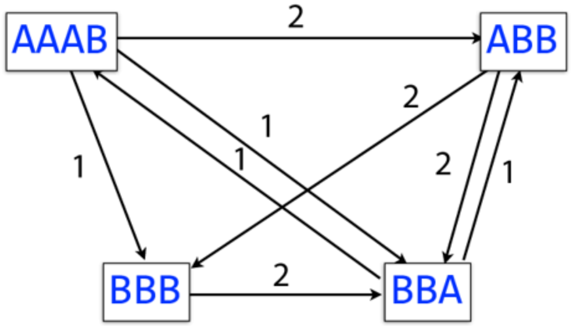
Greedy shortest common superstring



AAA AAB ABB BBA BBB

Greedy Solution to SCS

Greedy shortest common superstring

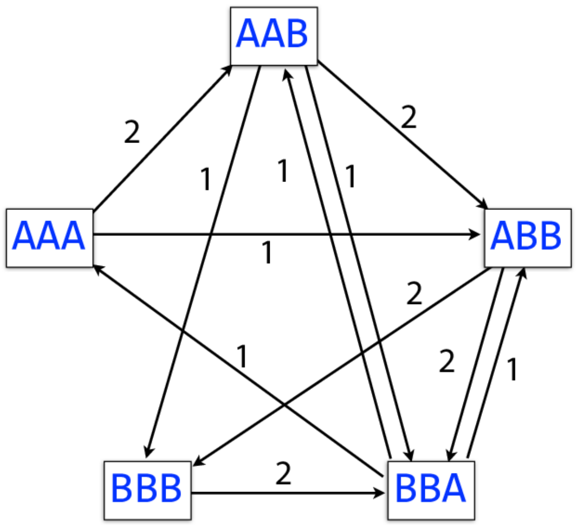


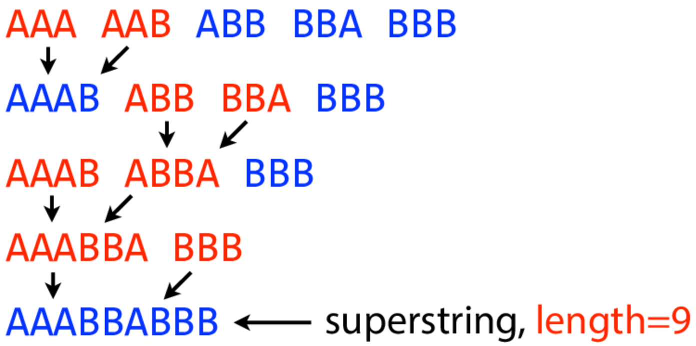
AAA AAB ABB BBA BBB

AAAB ABB BBA BBB

Greedy Solution to SCS

Greedy shortest common superstring





AAA AAB ABB BBA BBB

AAAB ABB BBA BBB

AAAB ABBA BBB

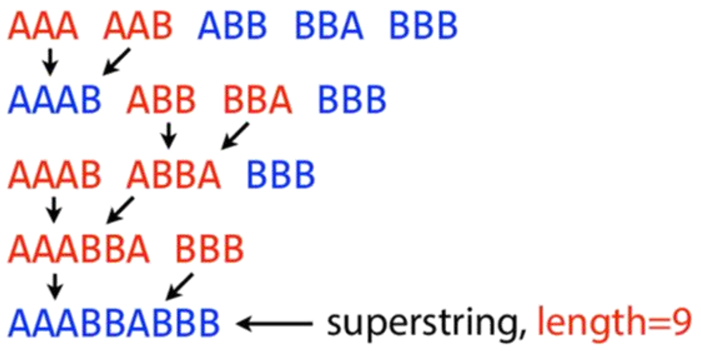
AAABBA BBB

AAABBABBB ← superstring, length=9

Greedy Solution to SCS

Greedy shortest common superstring

AAABBABBB ← superstring, length=9



AAA AAB ABB BBA BBB

AAAB ABB BBA BBB

AAAB ABBA BBB

AAABBA BBB

AAABBABBB ← superstring, length=9

Shorter Superstring: AAABBBA ← superstring, length=7

Problems with Overlap Graph

- No known efficient solution to SCS or Hamiltonian Path
- Heuristic approaches do not guarantee best solution
- It over-collapses the repeats in the genome, resulting in fewer copies than present in the genome.

De Bruijn Graph

WILL   ILLY   LLYN   LYNI   YNIL   NILL

```
WILL
 ILLY
  LLYN
   LYNI
    YNIL
     NILL
      ILLY
WILLYNILLY
```

De Bruijn Graph

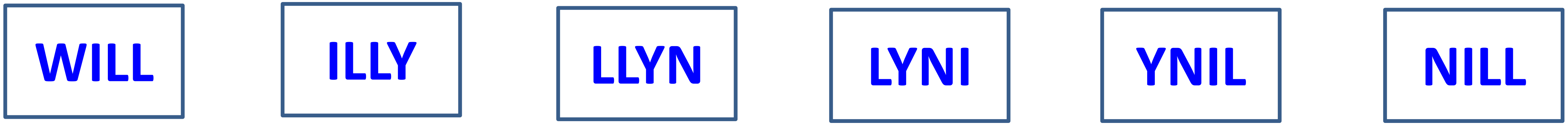
W

ILLY

N

```
W
 ILLY
     N
      ILLY
WILLYNILLY
```

De Bruijn Graph



WILL

ILLY

LLYN

LYNI

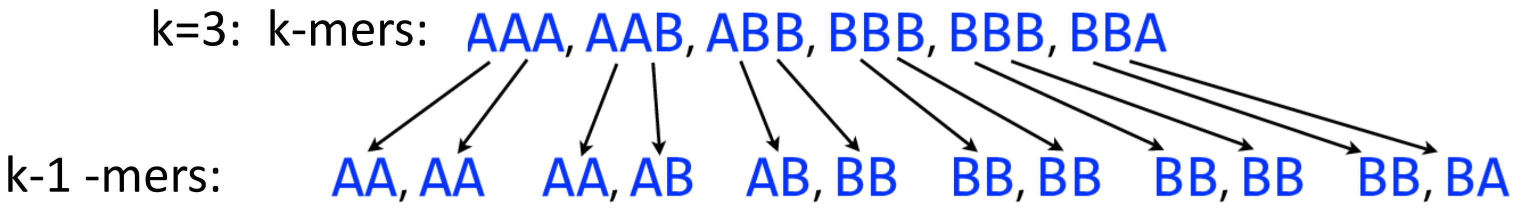
YNIL

NILL

```
WILL
 ILLY
  LLYN
   LYNI
    YNIL
     NILL
      ILLY
WILLYNILLY
```

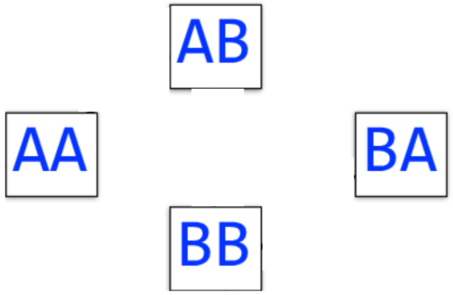
De Bruijn Graph

genome: AAABBBBA



k=3: k-mers: AAA, AAB, ABB, BBB, BBB, BBA

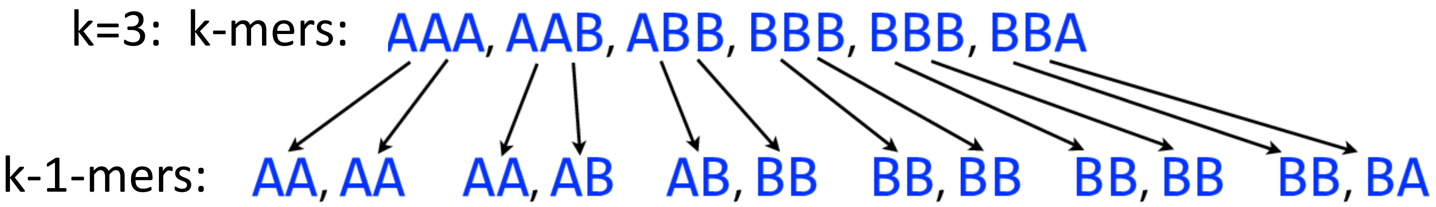
k-1 -mers: AA, AA AA, AB AB, BB BB, BB BB, BB BB, BA



One node per distinct k-1-mer

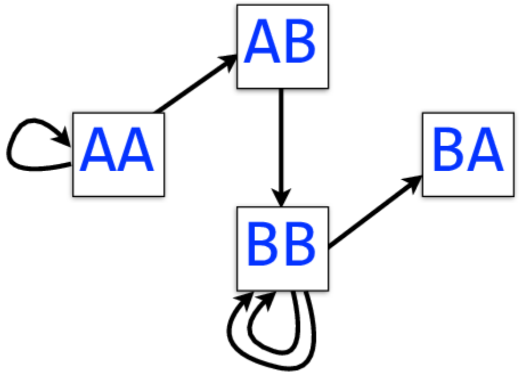
De Bruijn Graph

genome: AAABBBBA



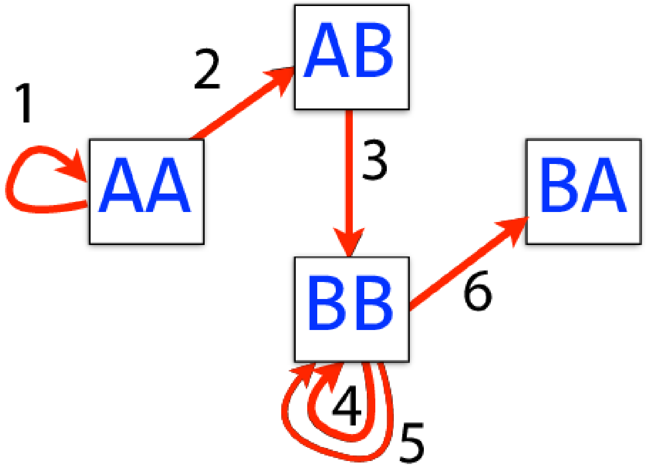
k=3: k-mers: AAA, AAB, ABB, BBB, BBB, BBA

k-1-mers: AA, AA AA, AB AB, BB BB, BB BB, BB BB, BA



One node per distinct k-1-mer
One edge per k-mer

Genome Reconstruction from De Brujin Graph

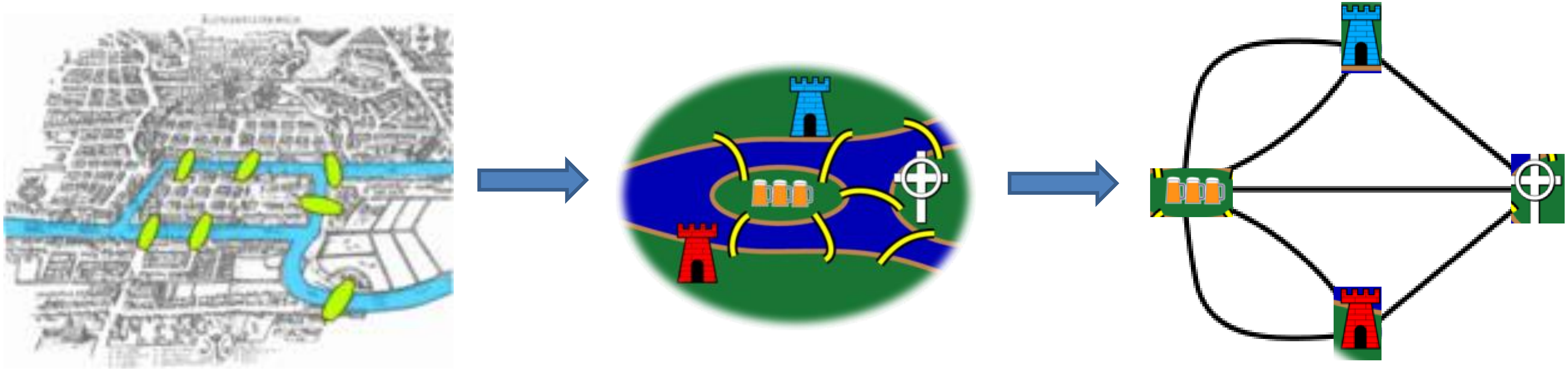


AAABBBBA

Walk crossing each edge exactly once (Eulerian Path) gives a reconstruction of the genome.

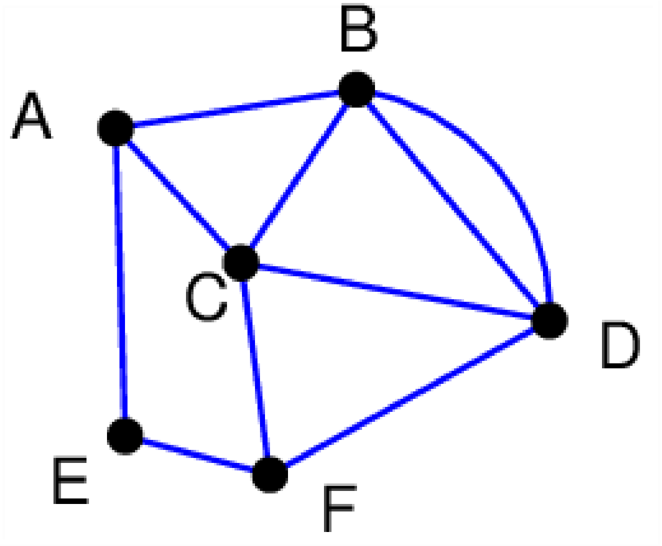
Seven Bridges of Königsberg

- Find a walk that crosses each bridge exactly once.



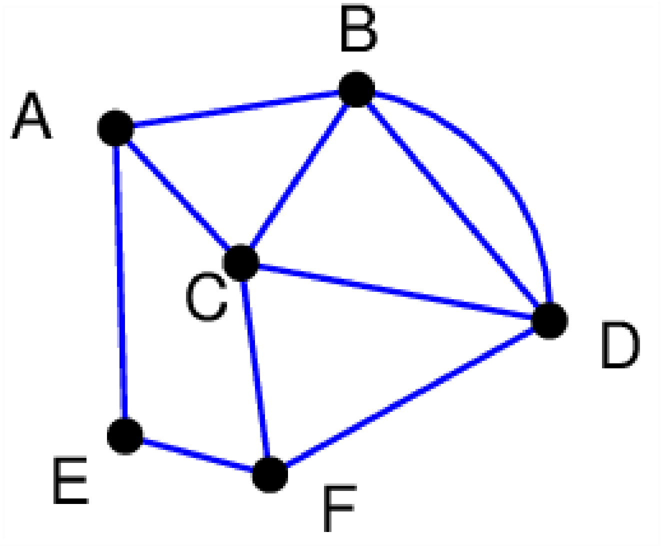
Euler Path/Cycle

- Is there an Euler Path/Cycle?



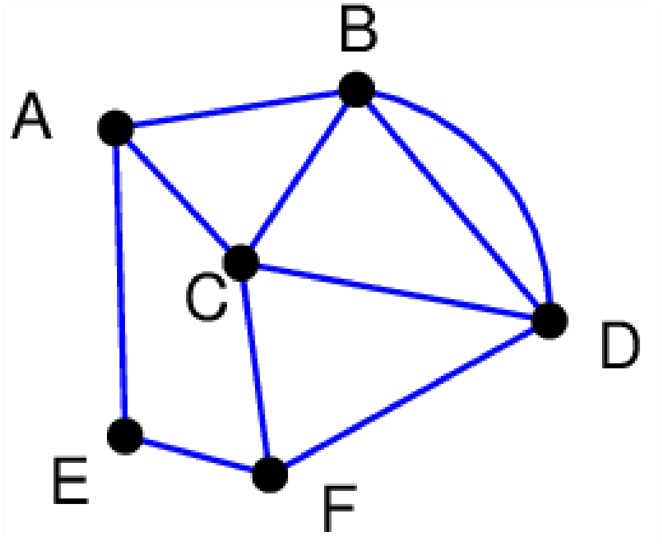
Euler Path/Cycle

- Is there an Euler Path/Cycle?



Euler Path/Cycle

- Find a Euler Path in the following graph



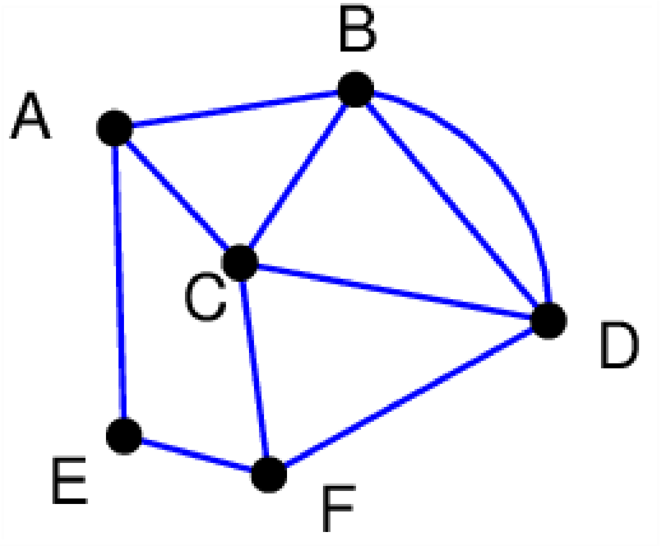
Fleury's Algorithm

- Refuse if graph doesn't have 0 or 2 odd nodes.

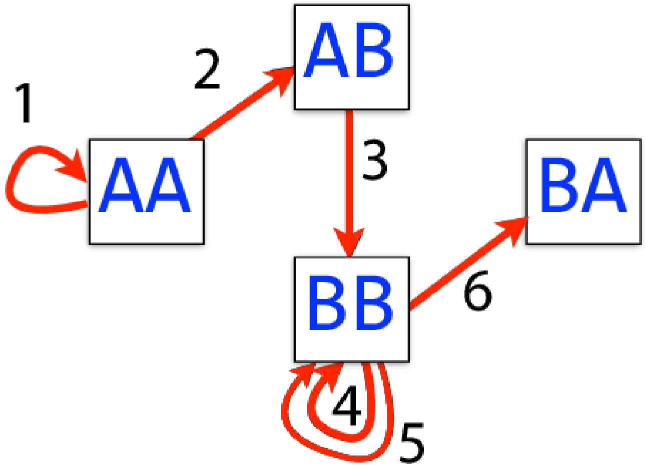
- Start:
  - If 2 odd nodes: start from one of the odd nodes.
  - If no odd node: start from any node
- Keep walking.
- If you have a choice between a "bridge" and a "non-bridge" edge, always choose the non-bridge edge.
  - A "bridge" edge is one whose removal would disconnect the remaining graph

Fleury's Algorithm

- Find a Euler Path in the following graph



Genome Reconstruction from De Brujin Graph



AAABBBBA

Walk crossing each edge exactly once (Eulerian Path) gives a reconstruction of the genome.

DeBruijn + Euler Path
Genome Reconstruction Example (k=5)

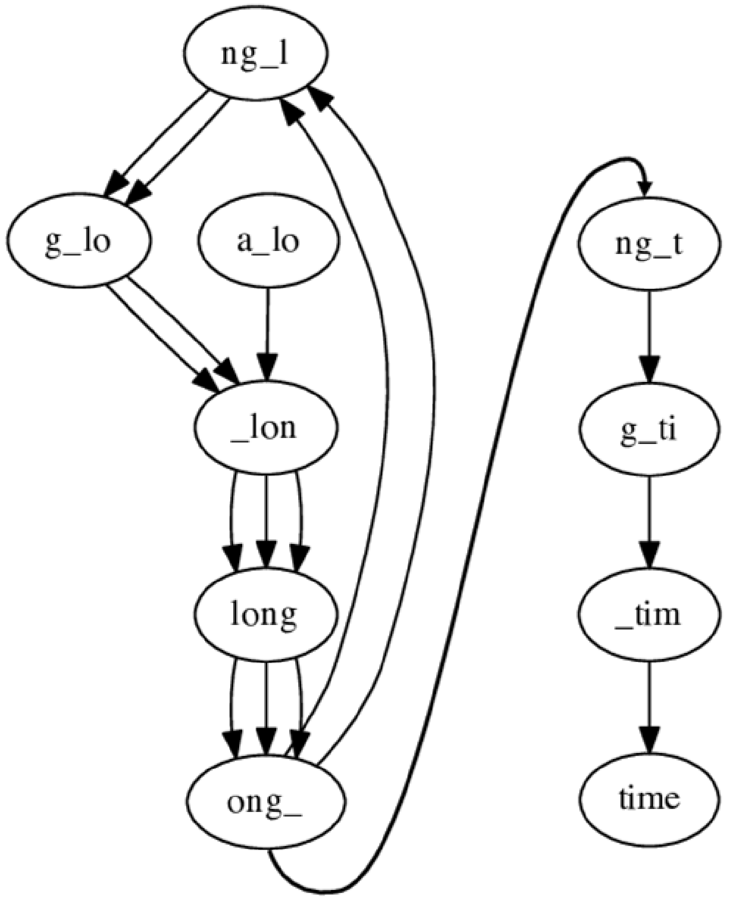
```
a_long_long_long_time
```

```
a_lon    ng_lo
 _long    g_lon
  long_    _long
   ong_l    long_
    ng_lo    ong_t
     g_lon    ng_ti
      _long    g_tim
       long_    _time
        ong_l
```

DeBruijn + Euler Path
Genome Reconstruction Example (k=5)

```
a_lon
 _long
  long_
   ong_l
    ng_lo
     g_lon
      _long
       long_
        ong_l
```

```
ng_lo
 g_lon
  _long
   long_
    ong_t
     ng_ti
      g_tim
       _time
```



a_long_long_long_time

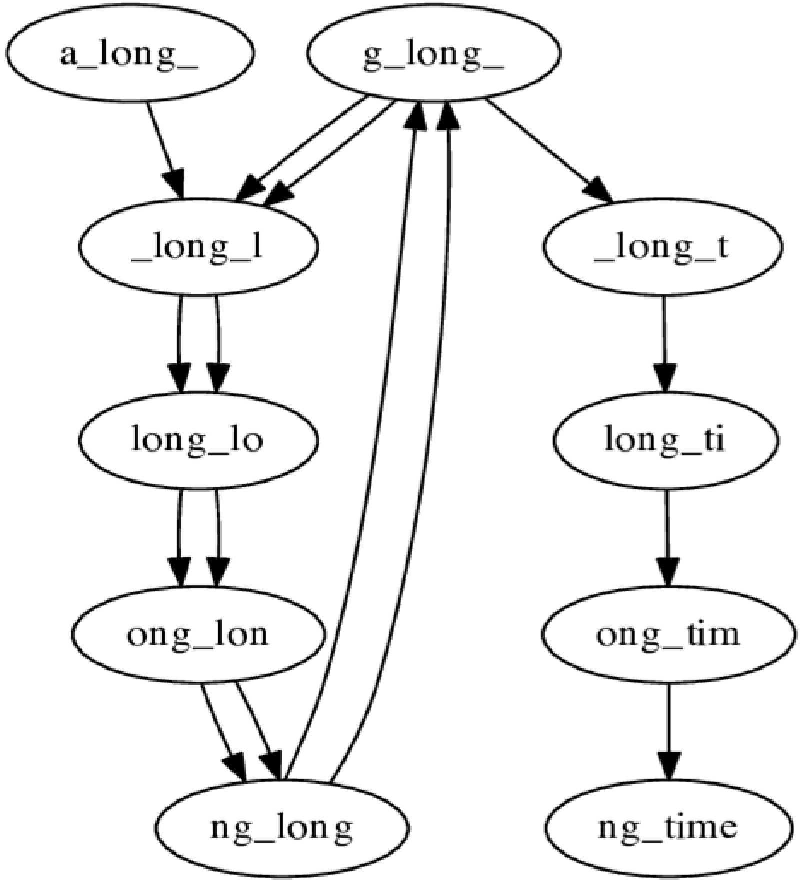
Problem: Reads are not perfect

- Reads are:
  - longer than k
  - non-uniform
  - incomplete

Genome: a_long_long_long_time

Reads: a_long_long_long, ng_long_l, g_long_time

```
k=8:  k-mers:  a_long_l            ng_long_   g_long_t
                _long_lo            g_long_l   _long_ti
                 long_lon                       long_tim
                  ong_long                       ong_time
                   ng_long_
                    g_long_l
                     _long_lo
                      long_lon
                       ong_long
```

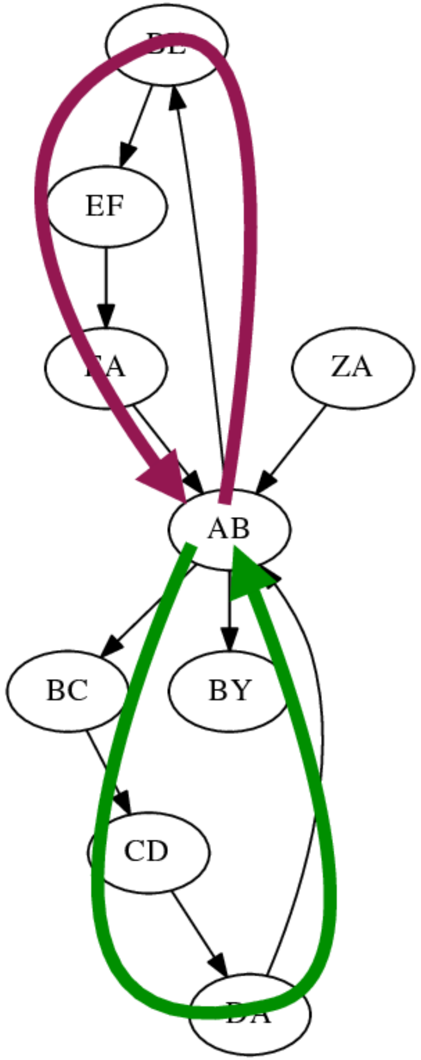


Biggest Problem: Repeats

Right: graph for ZABCDABEFABY, $k = 3$

ZA → AB → BE → EF → FA → AB → BC → CD → DA → AB → BY

ZA → AB → BC → CD → DA → AB → BE → EF → FA → AB → BY

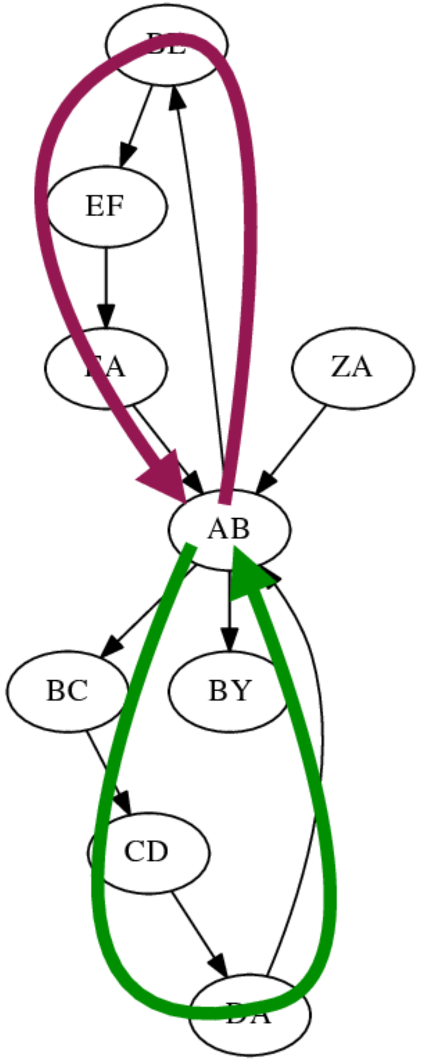


Biggest Problem: Repeats

Right: graph for **ZABCDABEFABY**, $k = 3$

ZA → AB → BE → EF → FA → AB → BC → CD → DA → AB → BY

ZA → AB → BC → CD → DA → AB → BE → EF → FA → AB → BY



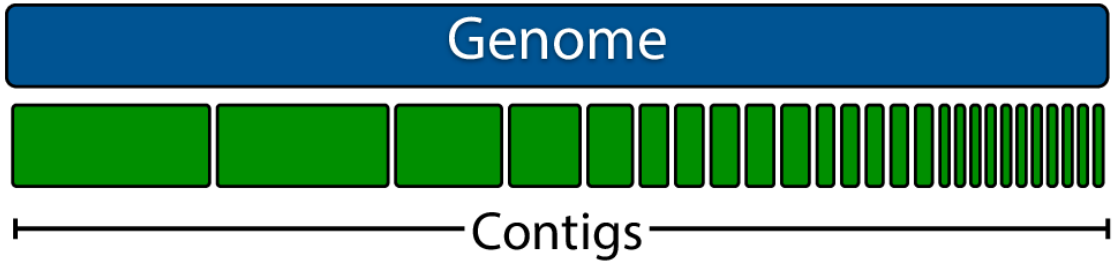
Contigs:

ZA → AB ⟹ ZAB

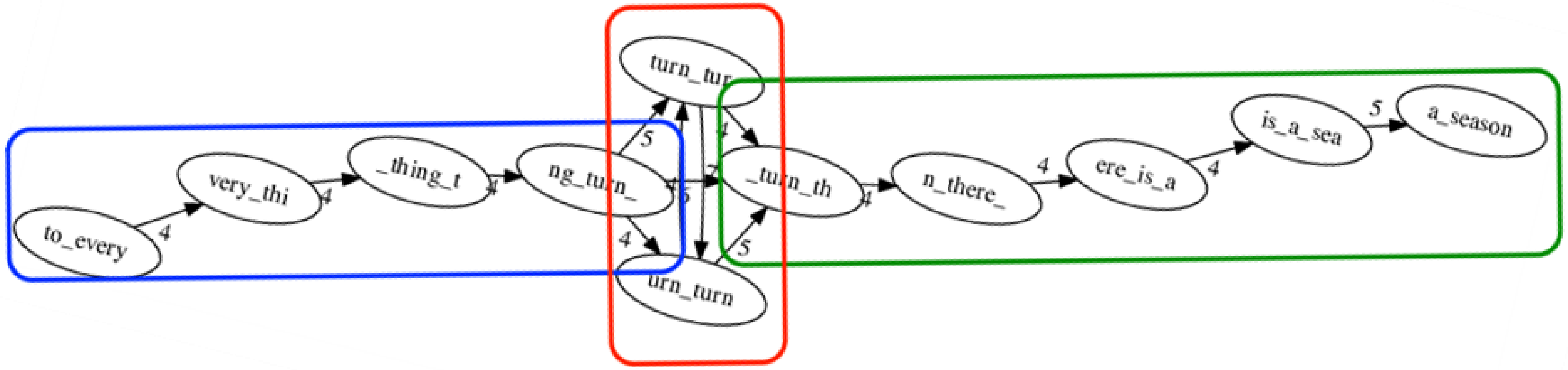
AB → BY ⟹ ABY

AB → BE → EF → FA → AB ⟹ ABEFA

AB → BC → CD → DA → AB ⟹ ABCDA

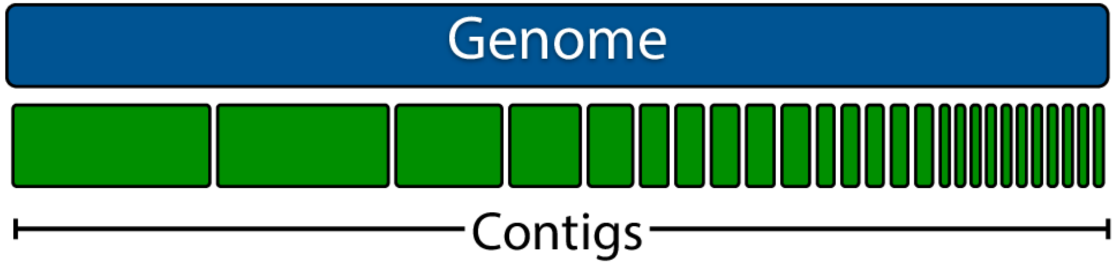


More repeats

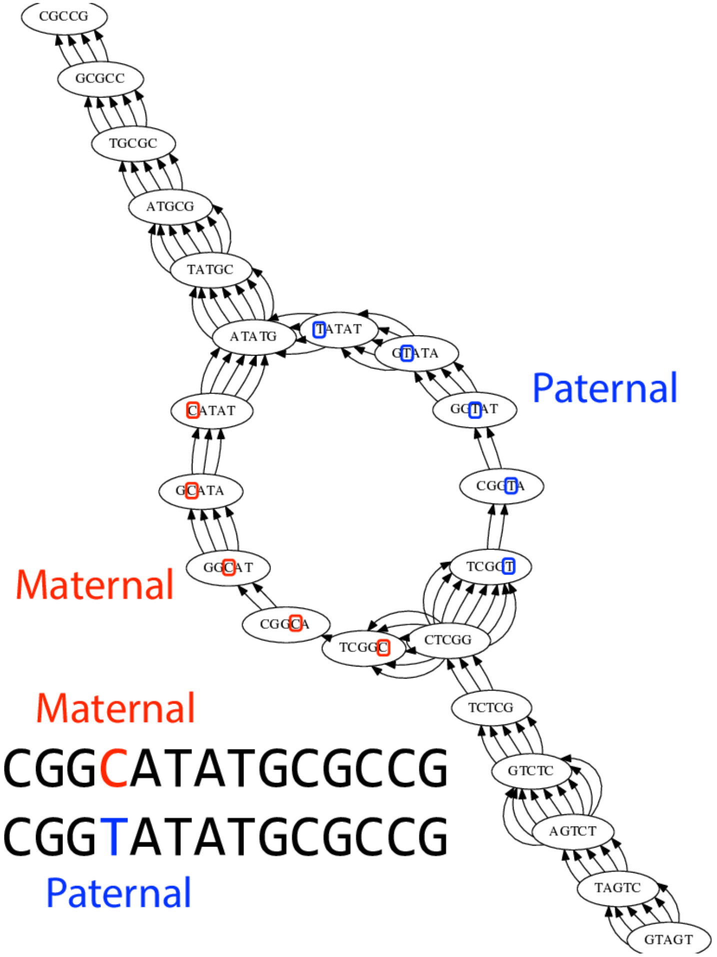


to_every_thing_turn_ _turn_there_is_a_season

_turn (repeated)



More Problems: Polyploidy



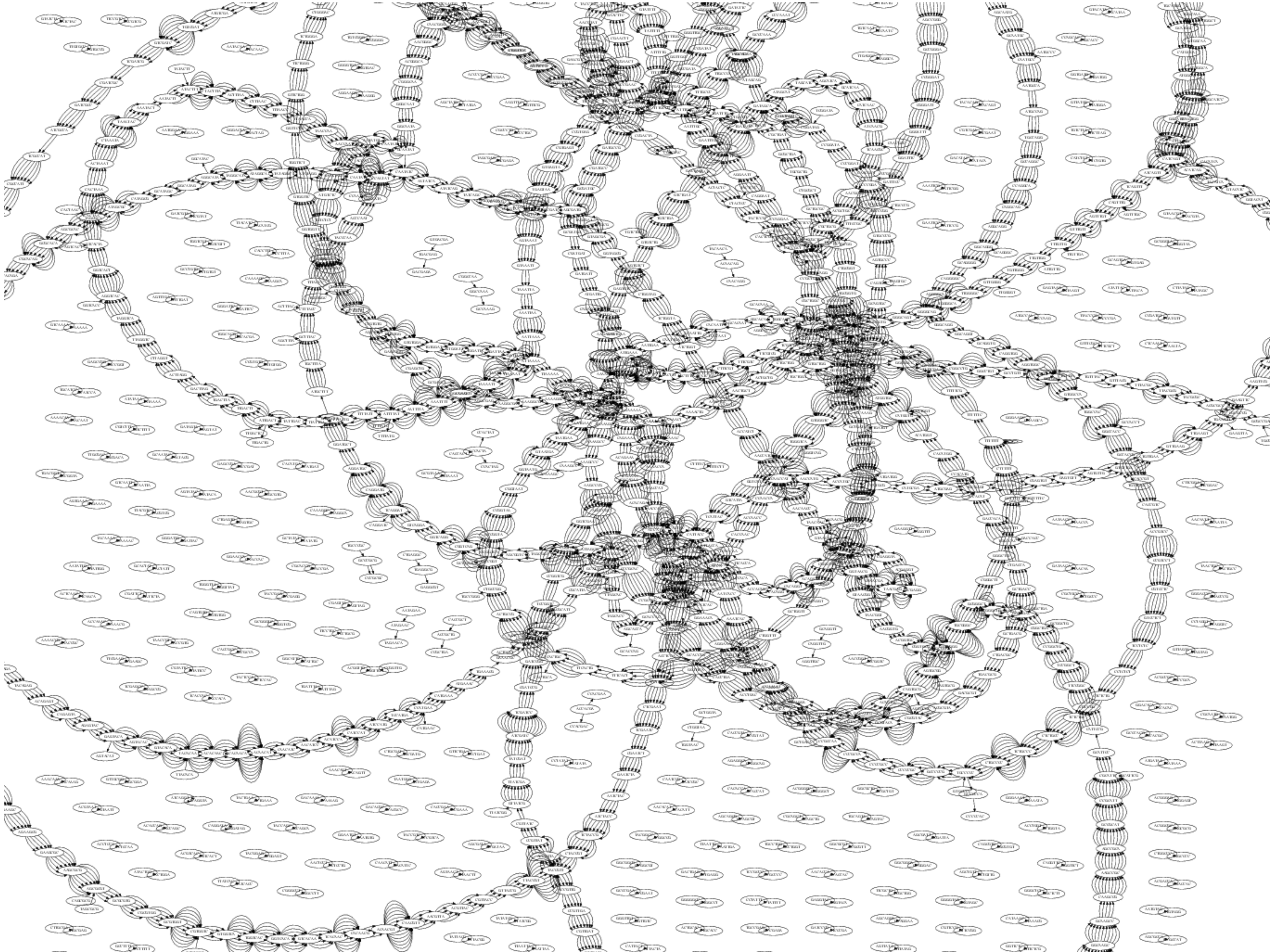
Maternal

GTAGTCTCGGCATATGCGCCG

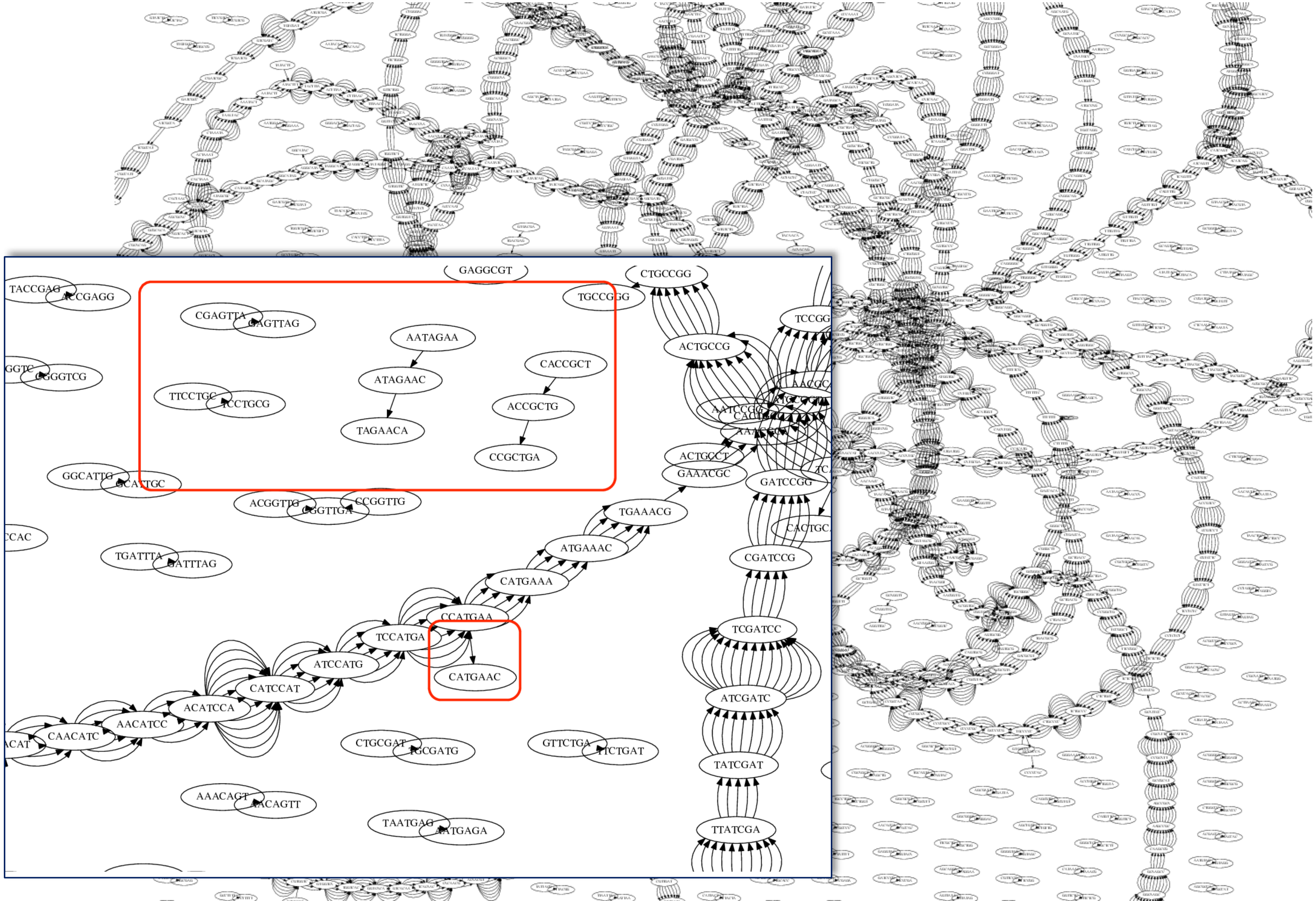
GTAGTCTCGGTATATGCGCCG

Paternal

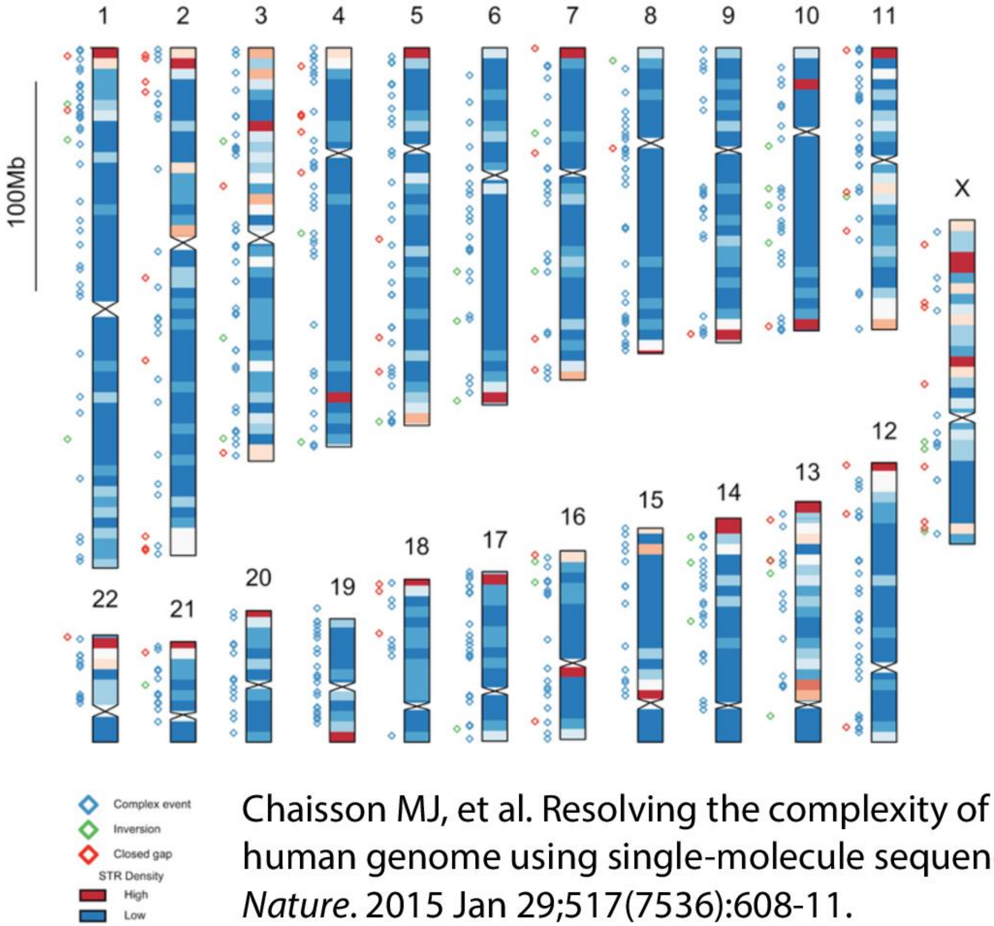
More Problems: Sequencing Errors



More Problems: Sequencing Errors



Reference Genomes are incomplete



Chaisson MJ, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*. 2015 Jan 29;517(7536):608-11.