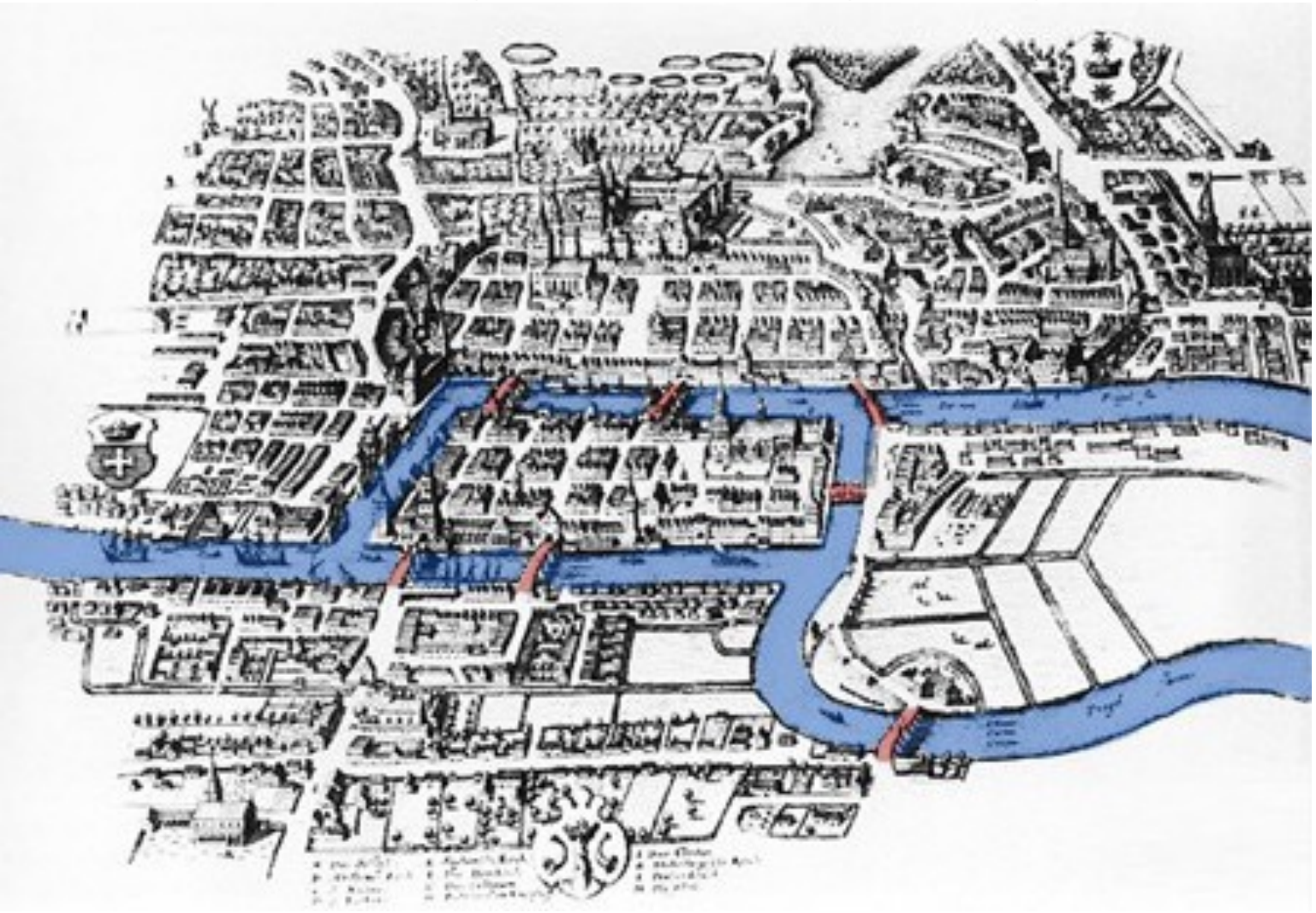
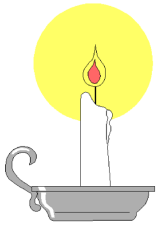
Senior Researcher
Rolf Sommer Kaas, PhD

De Novo assembly

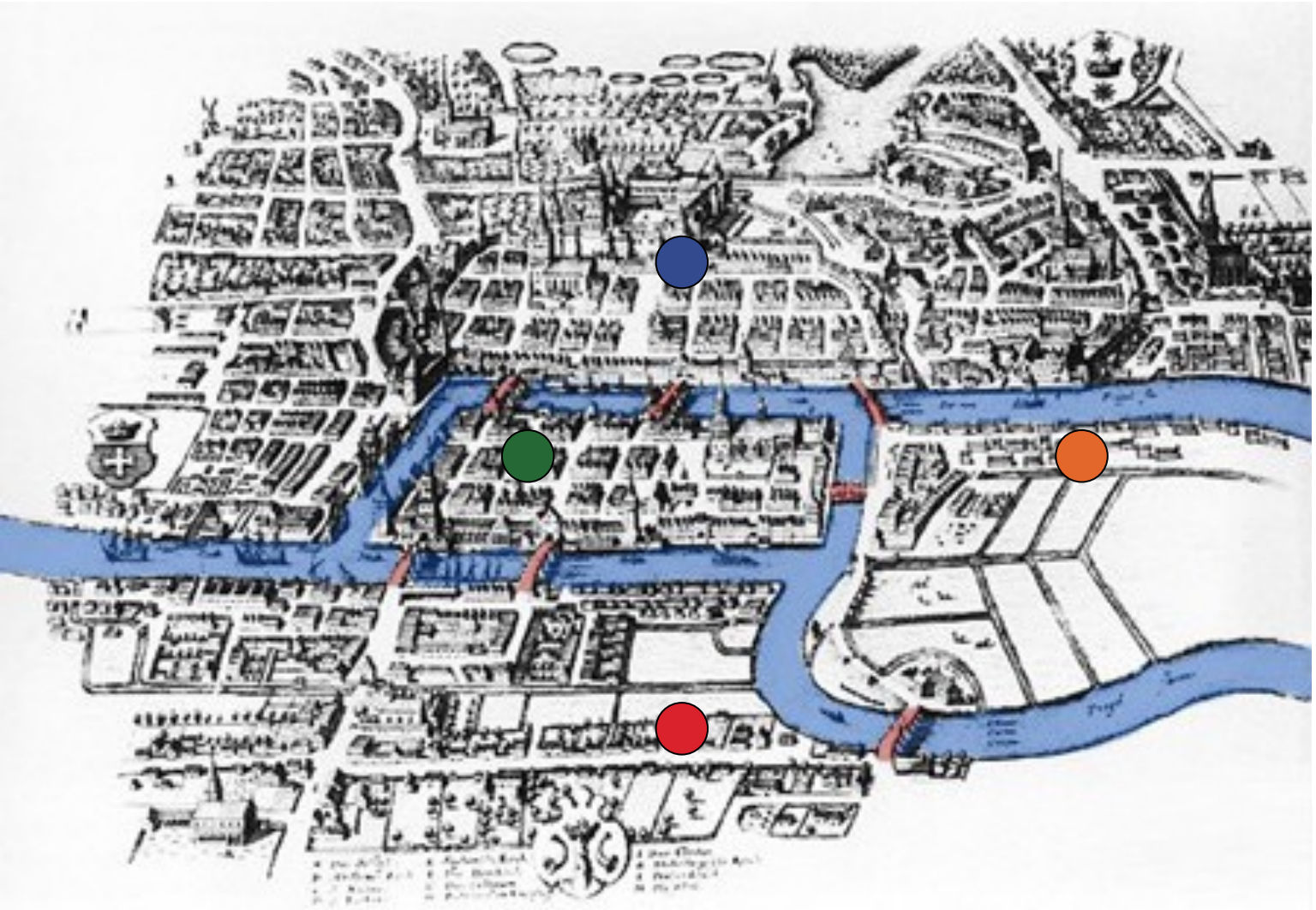

KONIGSBERG.
1. Pregel Fluß
2. Brandeburgische Thor.
3. Vestung Friedrichs Burg.
4. Haberbragische Kirch.
5. Kneiphofisches Wittwen Stifft.
6. Königl. Lacent und Pack-Hauß.
7. Neu Rosegartische Kirch.
8. Vorstättisch Hospital.
9. Die Kauffmans Börse.
10. Lange gaßen Thurn.
11. Thurn bey dem altstättischen gemein garten.
12. Pollnische Kirch.
13. Tragheimische Kirch.
14. Pfeiffen Thurn.
15. Kneiphofische Rath-Hauß.
16. Alt Städtische Kirche.
17. Königl. Schloß
18. Der blaue Thurn.
19. Altstättische Rath-Hauß.
20. Schnidt thor Thurn.
21. Kneiphöfische Thum Kirch und Collegium.
22. Pollnischner gaßen Thurn.
23. Zucht-Haus.
24. Pulver Thurn.
25. Der Mönch Hof.
26. Löbenichtsches Rath-Hauß.
27. Reformirte Kirch.
28. Löbenichtsche Kirche.
29. König. Hospital.
30. Die Catholische Kirch.
31. Sackheimsche Kirche.
32. Rosegartische Kirche.
33. Littauische Kirche.
34. Königl. Wayßenhauß.
35. Das Sackheimische Thor.
36. Friedlandsche Thor.
Stadt Wiesen
№ 57.

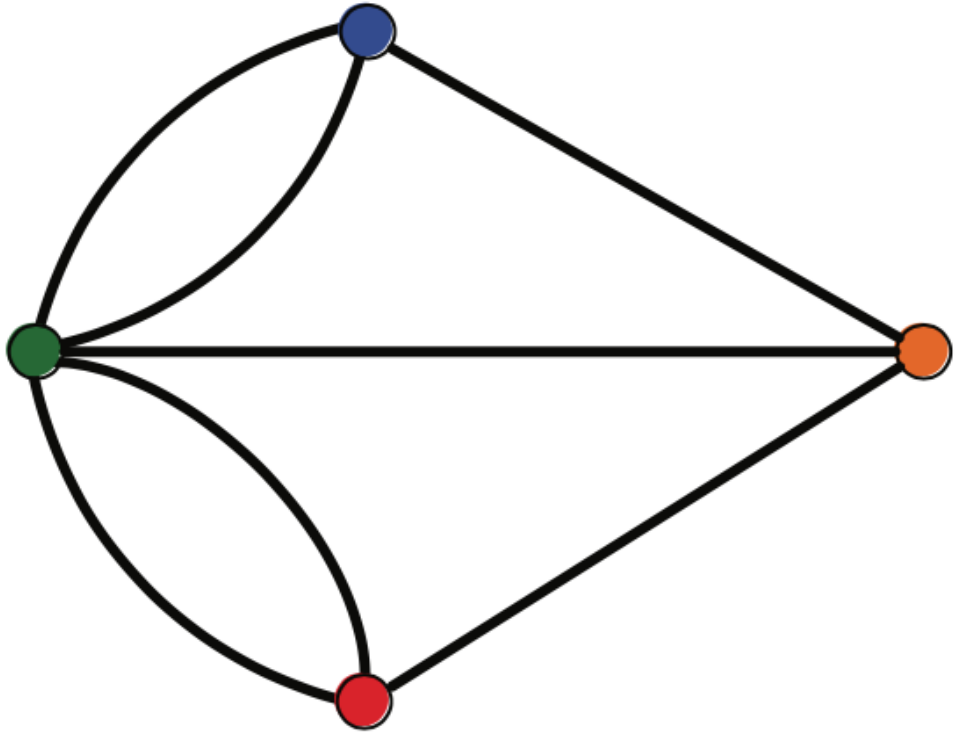
Bridges of Königsberg problem





Bridges of Königsberg problem





De Bruijn Graphs

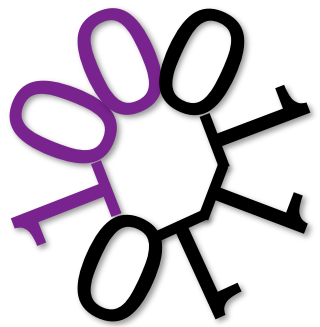
That's de Braun!

Nicolaas de Bruijn

```
000, 001,
010, 011,
100, 101,
110, 111
```

The Superstring problem

Find a shortest circular 'superstring' that contains all possible 'substrings' of length k (k-mers) over a given alphabet A containing n symbols.



00011101

De Bruijn Graphs

Example

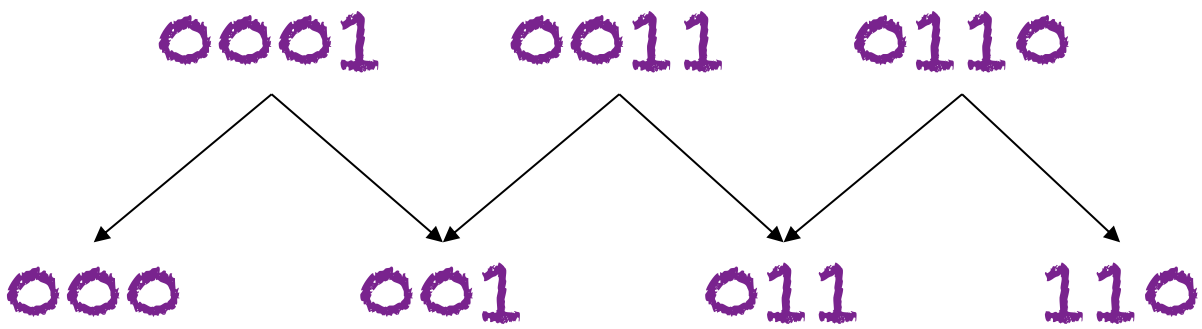
K = 4

n = 2

A = { 0, 1 }

Possible 4-mers: $2^4 = 16$

0000, 0001,
0011, 0110,
1100, 1001,
0010, 0101,
1011, 0111,
1111, 1110,
1101, 1010,
0100, 1000



...and so on.

Construct a graph B for which every possible (k – 1)-mer is assigned to a node.

De Bruijn Graphs

Example

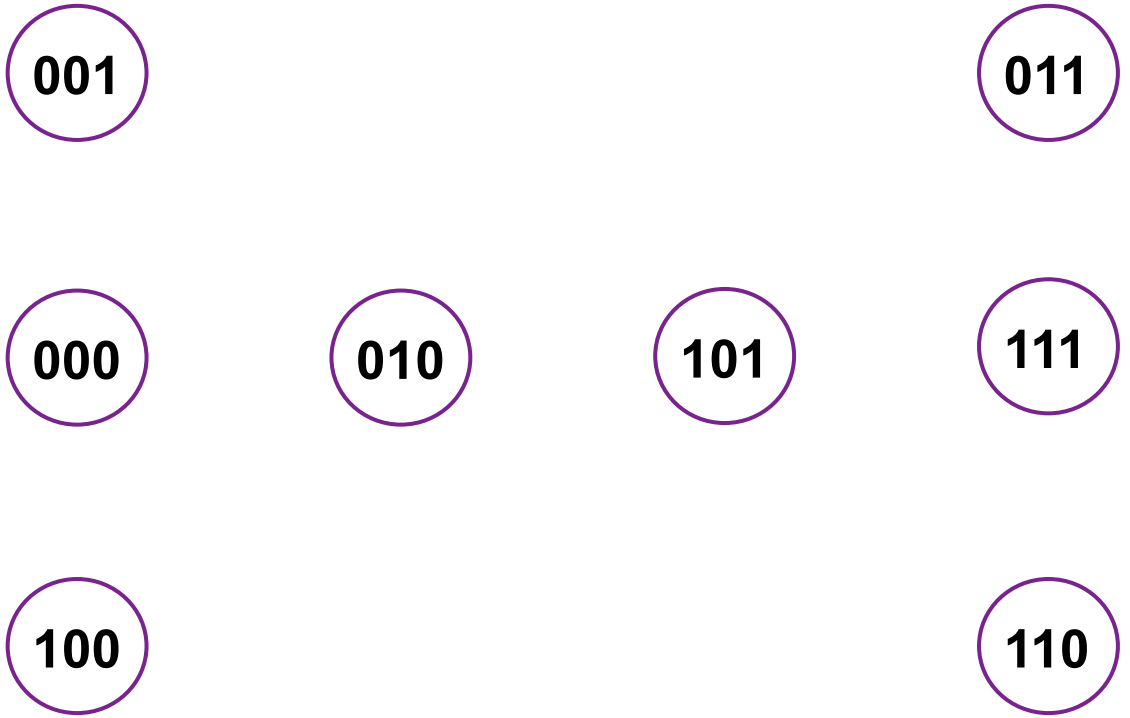
K = 4

n = 2

A = { 0, 1 }

Possible 4-mers: $2^4 = 16$

0000, 0001,
0011, 0110,
1100, 1001,
0010, 0101,
1011, 0111,
1111, 1110,
1101, 1010,
0100, 1000



Construct a graph B for which every possible (k – 1)-mer is assigned to a node. Connect one (k – 1)-mer by a directed edge to a second (k – 1)-mer if there is some k-mer whose prefix is the former and whose suffix is the latter.

De Bruijn Graphs

Example

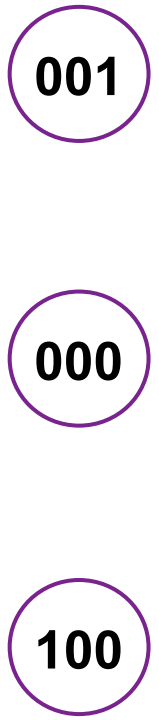
K = 4

n = 2

A = { 0, 1 }

Possible 4-mers: $2^4 = 16$

0000, 0001,
0011, 0110,
1100, 1001,
0010, 0101,
1011, 0111,
1111, 1110,
1101, 1010,
0100, 1000



100

Construct a graph B for which every possible (k – 1)-mer is assigned to a node. Connect one (k – 1)-mer by a directed edge to a second (k – 1)-mer if there is some k-mer whose prefix is the former and whose suffix is the latter.

De Bruijn Graphs

Example

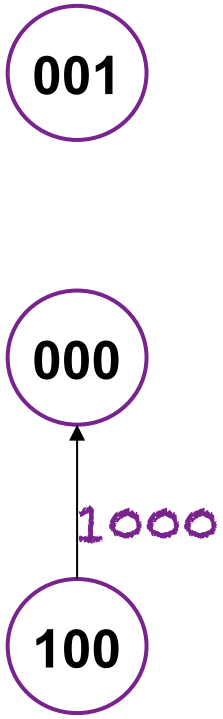
K = 4

n = 2

A = { 0, 1 }

Possible 4-mers: $2^4 = 16$

0000, 0001,
0011, 0110,
1100, 1001,
0010, 0101,
1011, 0111,
1111, 1110,
1101, 1010,
0100, 1000



1000

Construct a graph B for which every possible (k – 1)-mer is assigned to a node. Connect one (k – 1)-mer by a directed edge to a second (k – 1)-mer if there is some k-mer whose prefix is the former and whose suffix is the latter.

De Bruijn Graphs

Example

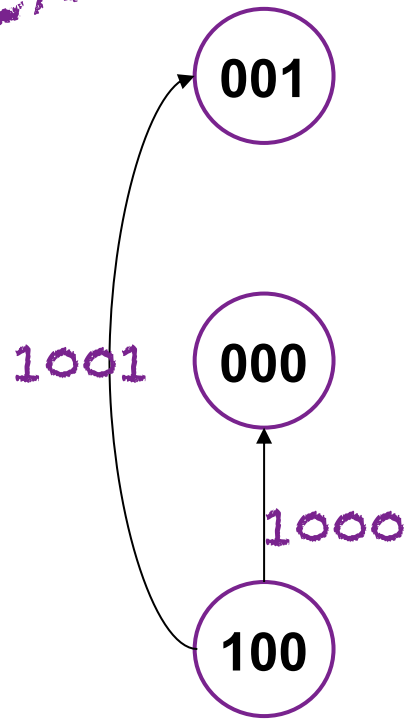
K = 4

n = 2

A = { 0, 1 }

Possible 4-mers: $2^4 = 16$

0000, 0001,
0011, 0110,
1100, 1001,
0010, 0101,
1011, 0111,
1111, 1110,
1101, 1010,
0100, 1000



1001

Construct a graph B for which every possible (k – 1)-mer is assigned to a node. Connect one (k – 1)-mer by a directed edge to a second (k – 1)-mer if there is some k-mer whose prefix is the former and whose suffix is the latter.

De Bruijn Graphs

Example

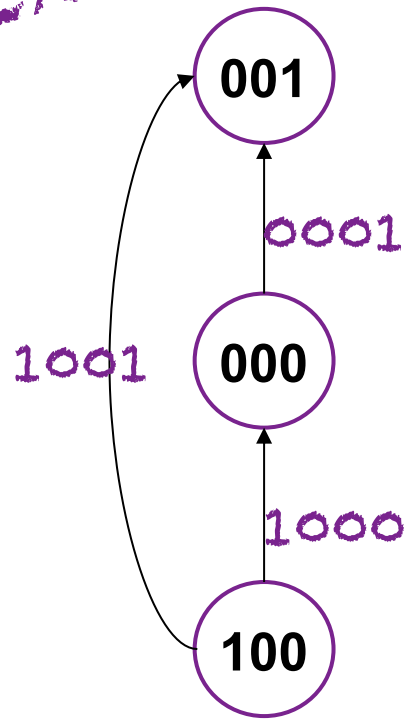
K = 4

n = 2

A = { 0, 1 }

Possible 4-mers: $2^4 = 16$

0000, 0001,
0011, 0110,
1100, 1001,
0010, 0101,
1011, 0111,
1111, 1110,
1101, 1010,
0100, 1000



0001

Construct a graph B for which every possible (k – 1)-mer is assigned to a node. Connect one (k – 1)-mer by a directed edge to a second (k – 1)-mer if there is some k-mer whose prefix is the former and whose suffix is the latter.

De Bruijn Graphs

Example

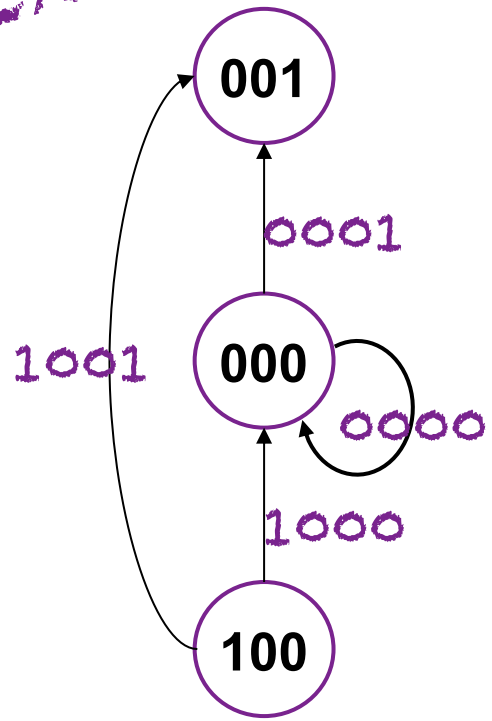
K = 4

n = 2

A = { 0, 1 }

Possible 4-mers: $2^4 = 16$

0000, 0001,
0011, 0110,
1100, 1001,
0010, 0101,
1011, 0111,
1111, 1110,
1101, 1010,
0100, 1000



0000

Construct a graph B for which every possible (k – 1)-mer is assigned to a node. Connect one (k – 1)-mer by a directed edge to a second (k – 1)-mer if there is some k-mer whose prefix is the former and whose suffix is the latter.

De Bruijn Graphs

0000110010111101

Example

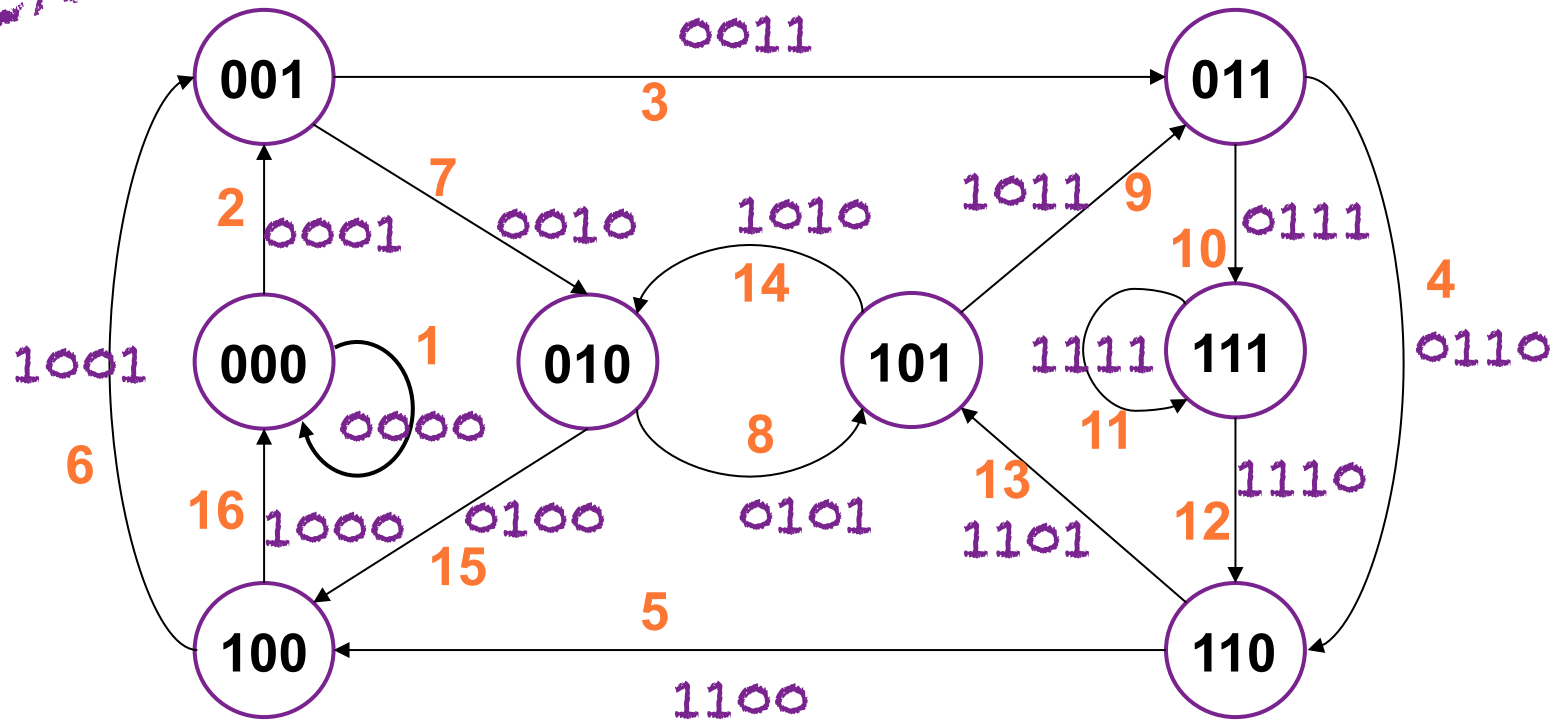
K = 4

n = 2

A = { 0, 1 }

Possible 4-mers: $2^4 = 16$

0000, 0001,
0011, 0110,
1100, 1001,
0010, 0101,
1011, 0111,
1111, 1110,
1101, 1010,
0100, 1000



Construct a graph B for which every possible (k – 1)-mer is assigned to a node. Connect one (k – 1)-mer by a directed edge to a second (k – 1)-mer if there is some k-mer whose prefix is the former and whose suffix is the latter.

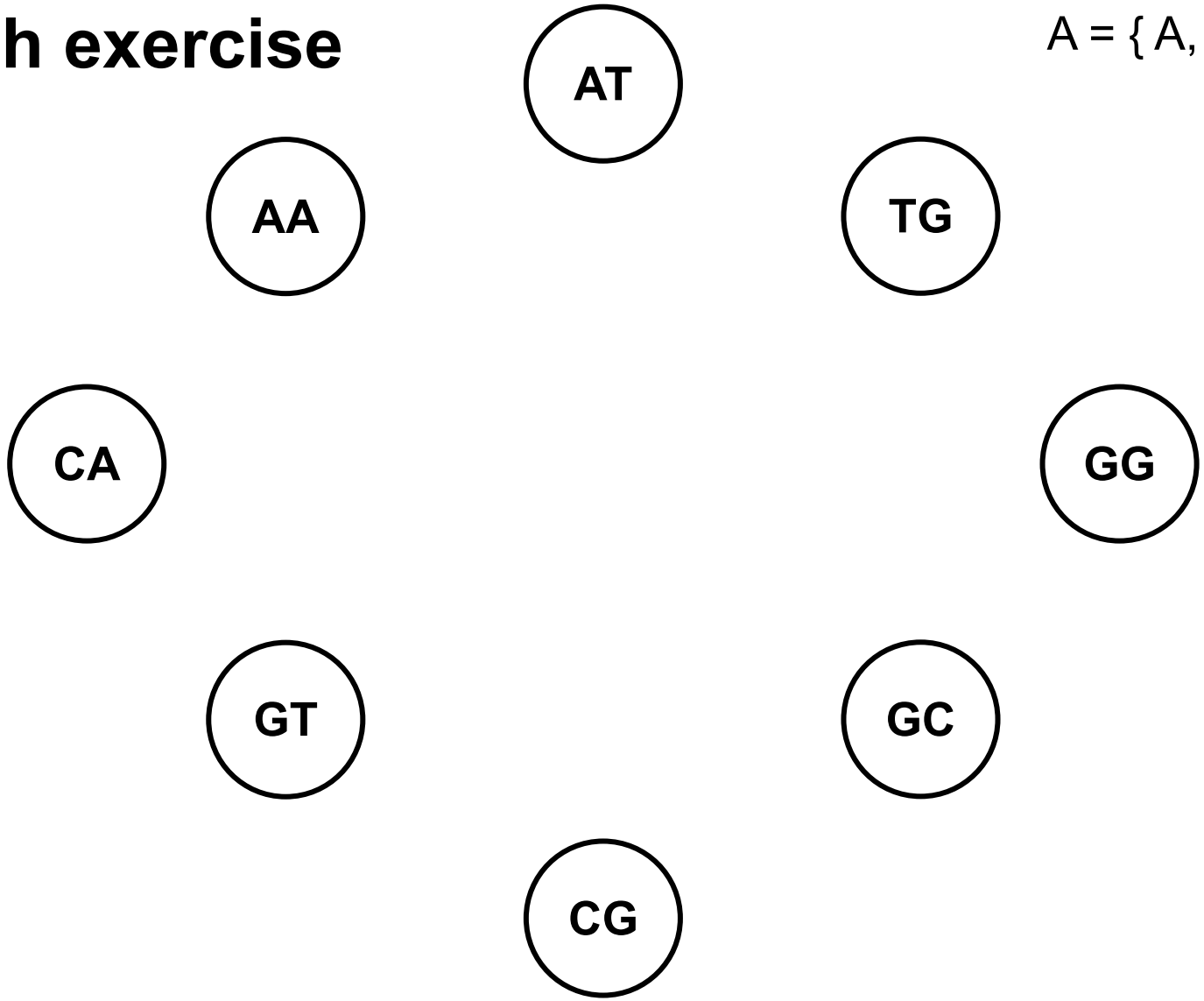
De Bruijn Graph exercise

K = 3

n = 4

A = { A, T, G, C }

ATG	GCA
TGG	CAA
TGC	AAT
GCG	GTG
CGT	GGC



De Bruijn Graph – Size of k

The choice of k affects the construction of the de Bruijn graph.

Smaller values of k collapse more repeats together, making the graph more tangled. Larger values of k may fail to detect overlaps between reads, particularly in low coverage regions, making the graph more fragmented.

(Bankevich et al., “SPAdes.”)

k = tangled K = fragmented

Larger k results in longer contigs in high coverage areas, but will tend to break contigs in areas with low coverage.

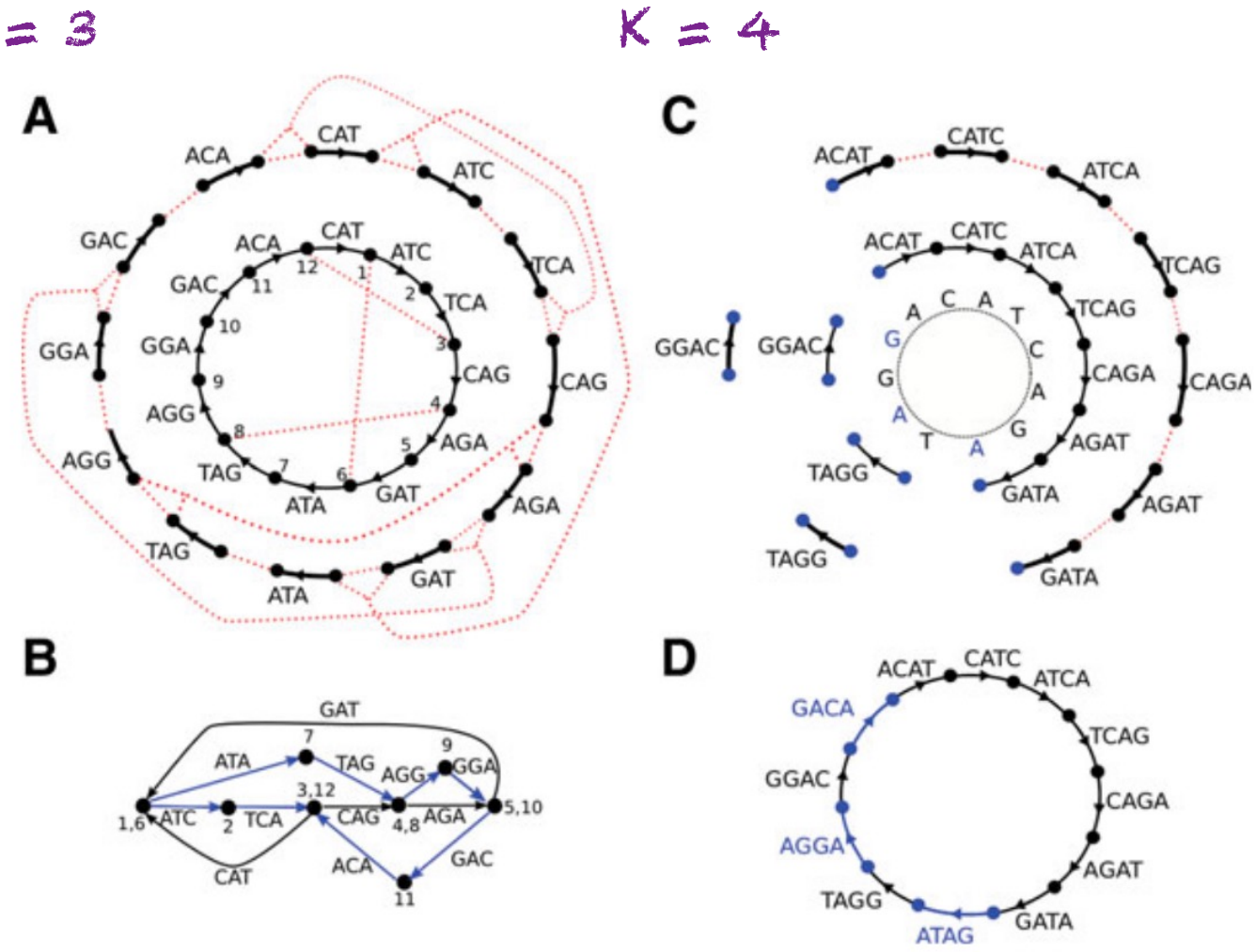
De Bruijn Graph – Large k-mers vs. small k-mers

Sequence:
CATCAGATAGGA

Reads (input data):
ACAT CATC ATCA
TCAG CAGA AGAT
GATA TAGG GGAC

4-mers not in data:
GATA TAGG GGAC

All 3-mers are found in data

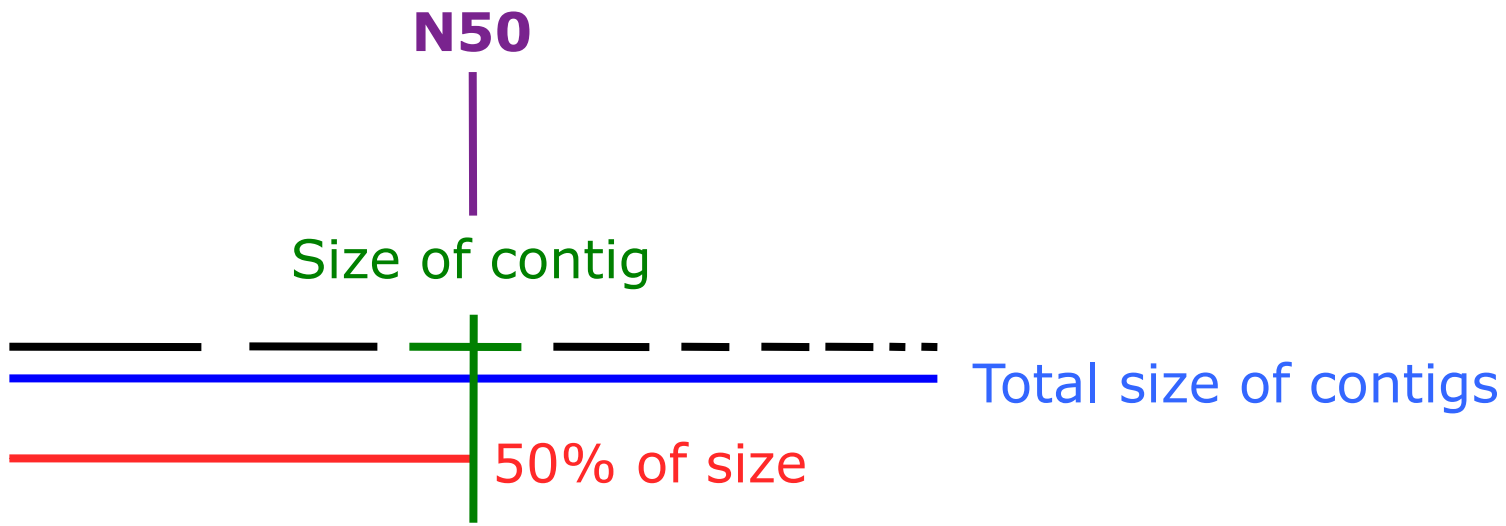


De Bruijn Graph – The assumptions we made

- We can generate all k-mers present in the genome
- All k-mers are error free
- Each k-mer appears at most once
- Single circular chromosome

De novo assembly quality

- Number of contigs
- Size of largest contig
- Assembly size
- N50



Small N50 exercise

ACT

AGGGTCCA

ACTTCGACCAATGC

ACGTT

What is the N50 for this collection of strings?