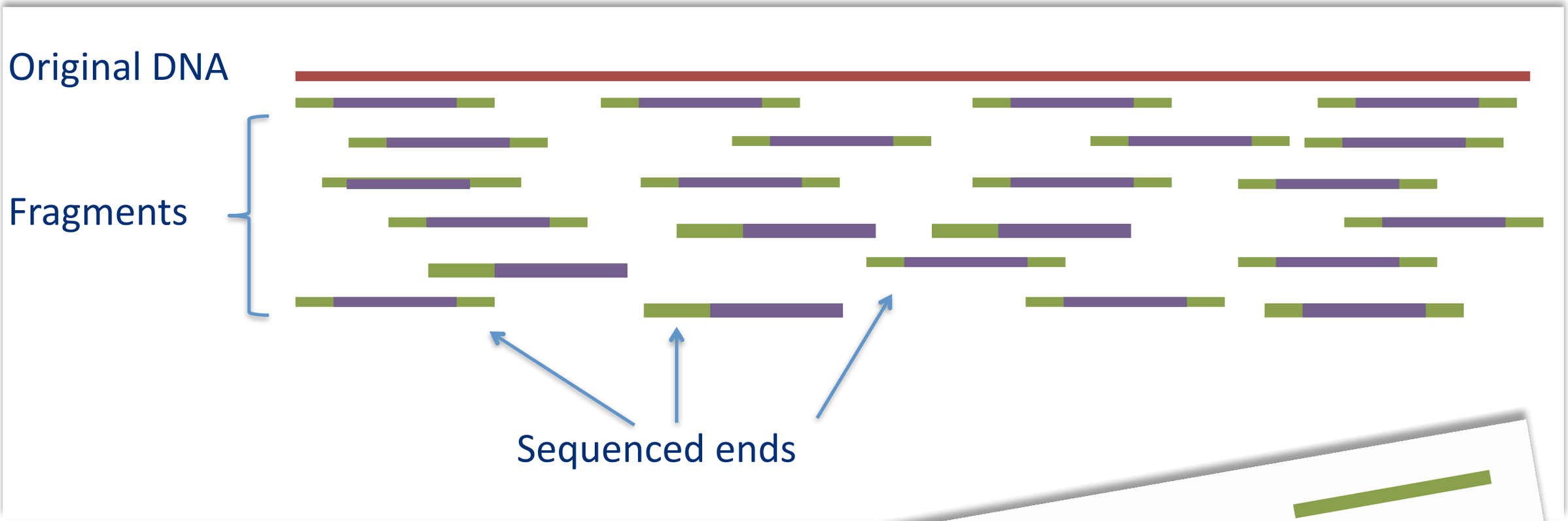
3 - A simple genome assembly

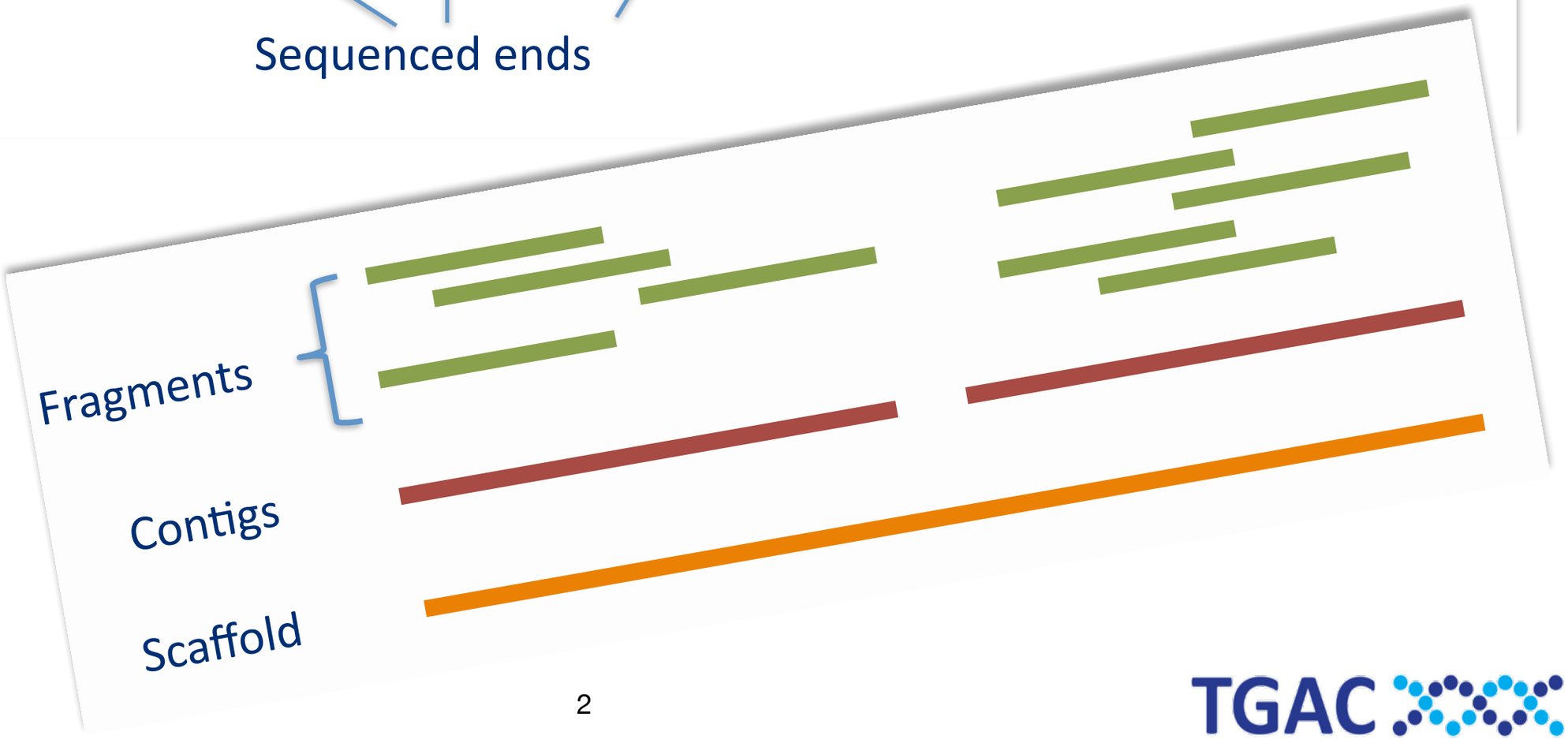
Wednesday morning

Bernardo J. Clavijo
Richard Smith-Unna
Gonzalo Garcia / Jon Wright

TGAC

The genome assembly problem (WGS)





TGAC

A correct assembly has:

The right *motifs*,

the correct number of times,

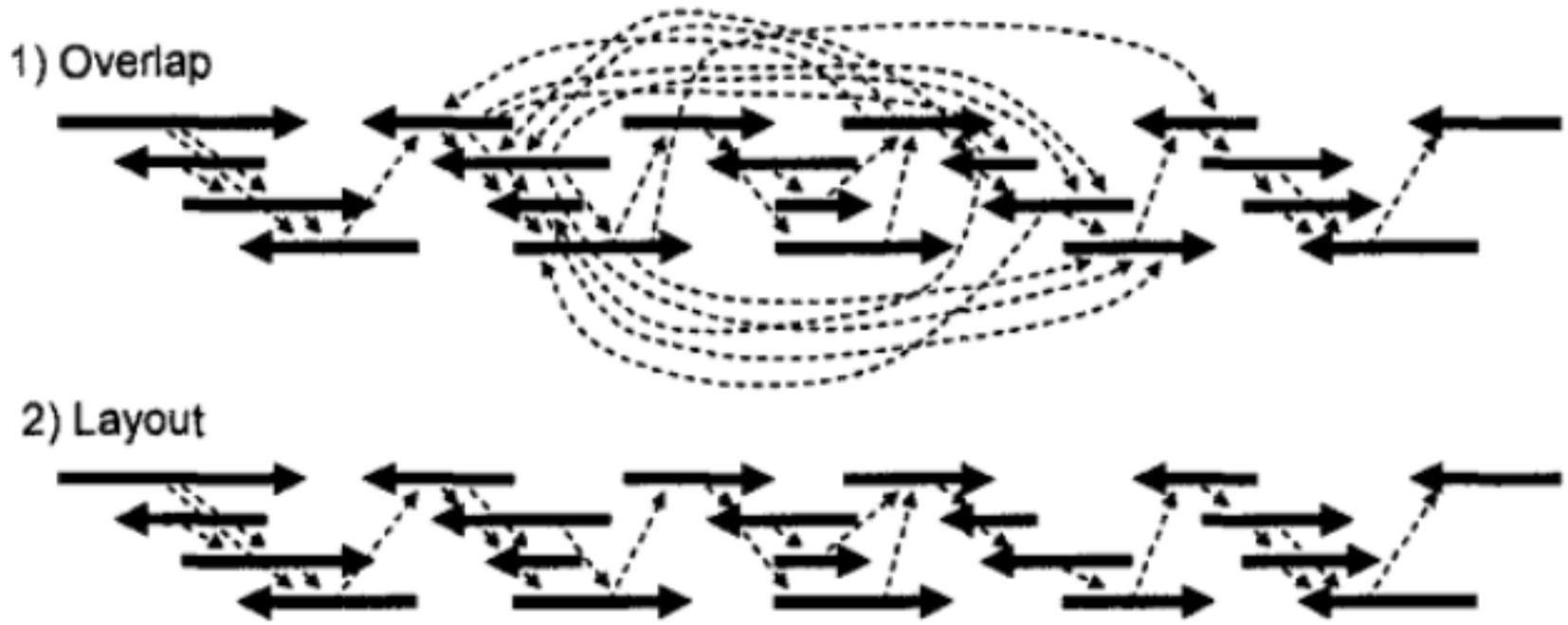
in correct order and position.

None of which is assessed by length stats.

TGAC

Overlap Layout Consensus

Overlap - Layout - Consensus



3) Consensus

```
            CCTATG-TAGTCAGTCG
                ATGCTAGTCAG
                  GCTAGTCGGTCGATCTACC
                         CAGTCGATCTGCCGGT
                       GTCAGTC-ATCTAC-GGTTAGCATTGC
Consensus   CCTATGCTAGTCAGTCGATCTACCGGTTAGCATTGC
```


TGAC

Overlap Layout Consensus: Key points

- Finding overlaps and defining them is key.
- The layout can be quite difficult.
- The method tracks every read.
- The consensus is constructed from the reads.

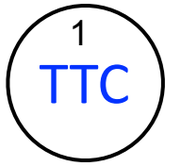
De Bruijn Graphs

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

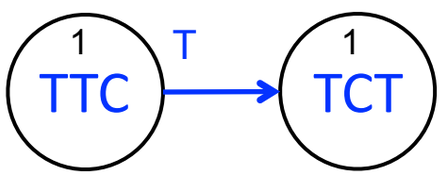
Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



Assembling a De Bruijn Graph

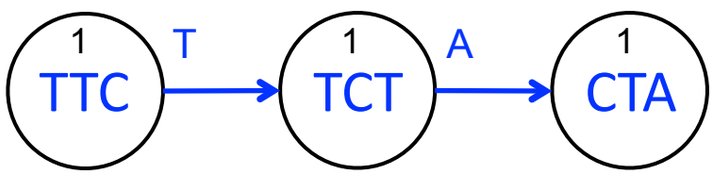
```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



TGAC

Assembling a De Bruijn Graph

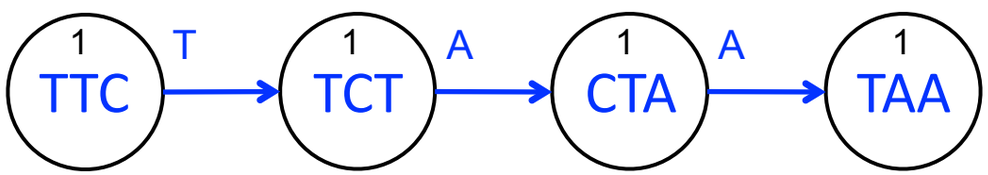
```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



TGAC

Assembling a De Bruijn Graph

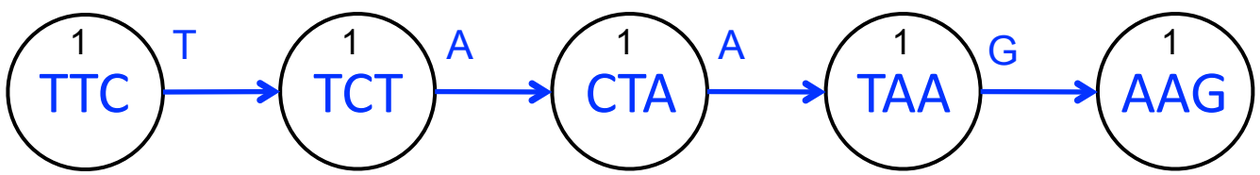
```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



TGAC

Assembling a De Bruijn Graph

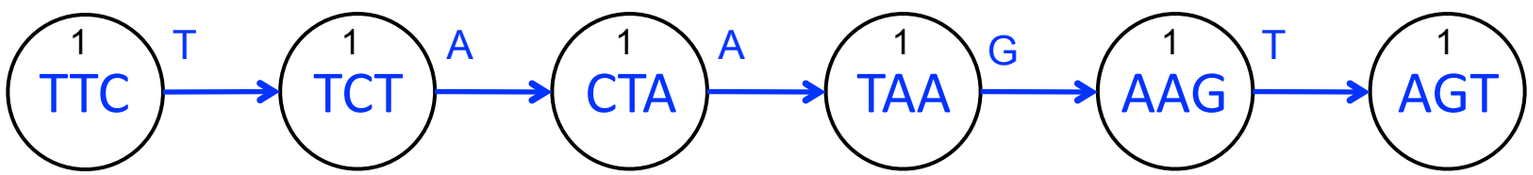
```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



TGAC

Assembling a De Bruijn Graph

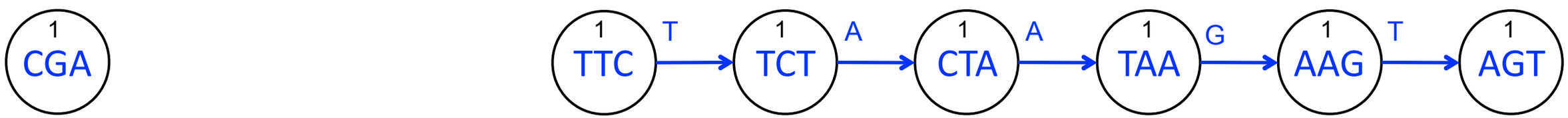
```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



TGAC

Assembling a De Bruijn Graph

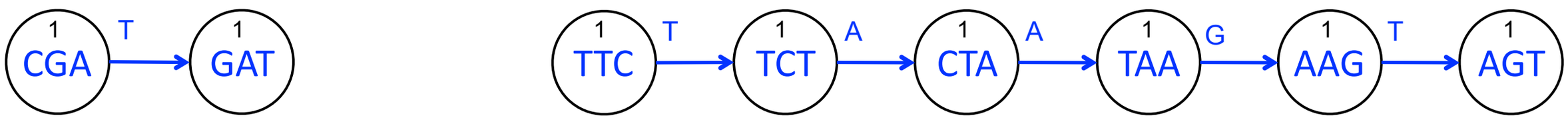
```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



TGAC

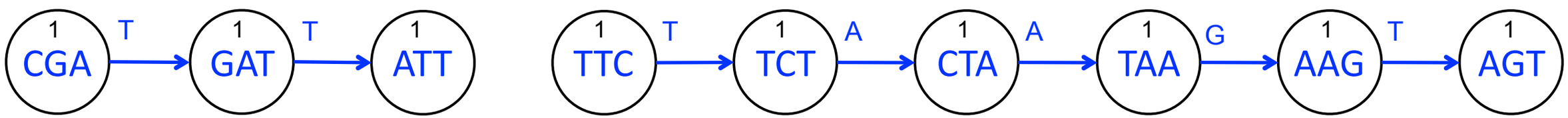
Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



Assembling a De Bruijn Graph

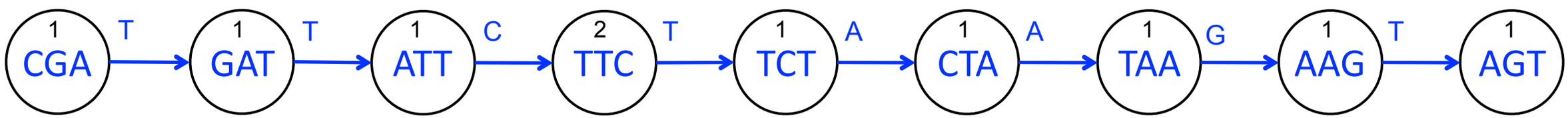
```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



TGAC

Assembling a De Bruijn Graph

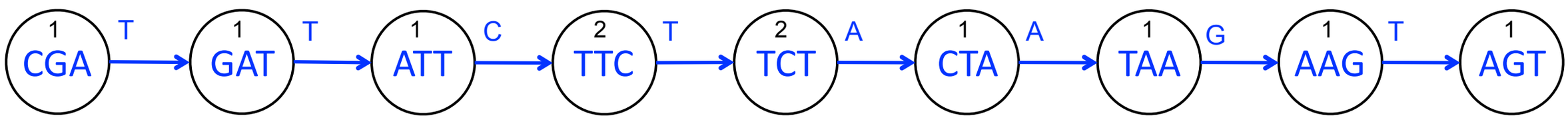
```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



TGAC

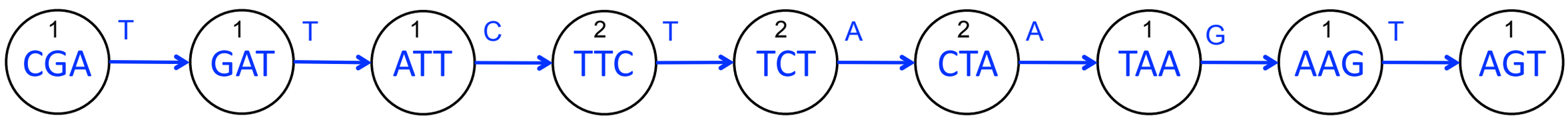
Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



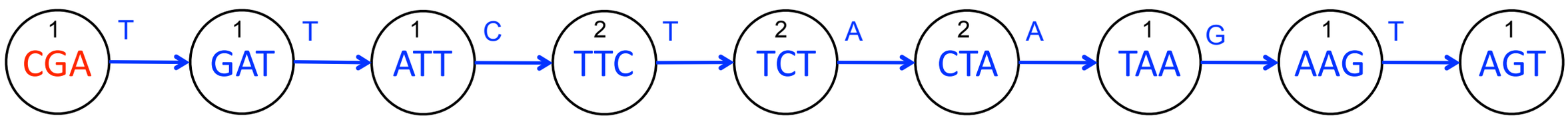
Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

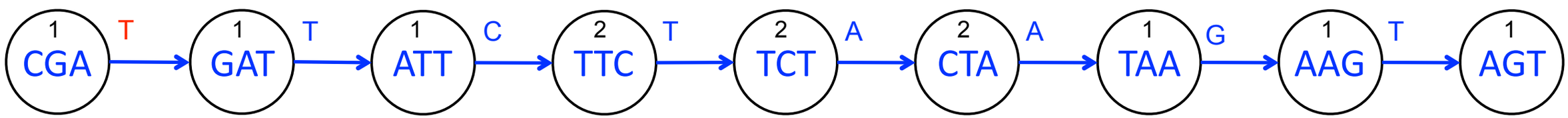


CGA

TGAC

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

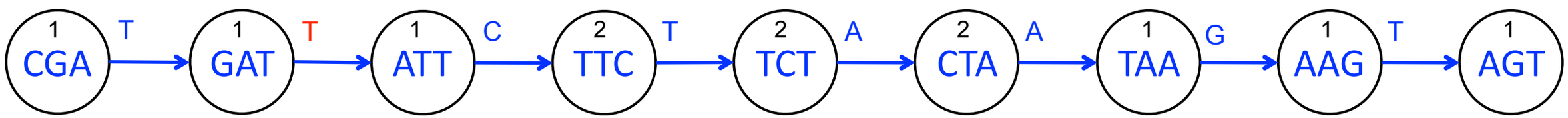


CGAT

TGAC

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

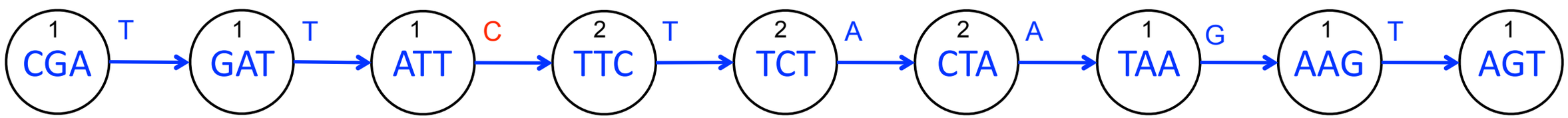


CGATT

TGAC

Assembling a De Bruijn Graph

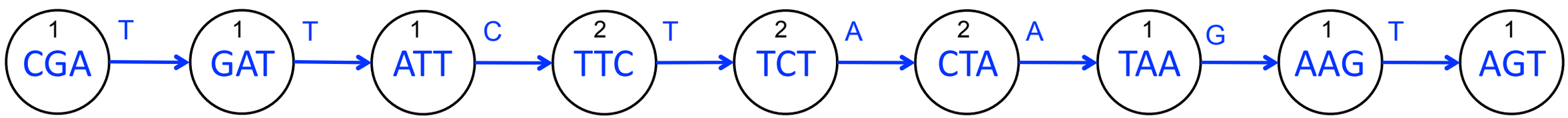
```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```



CGATTC

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

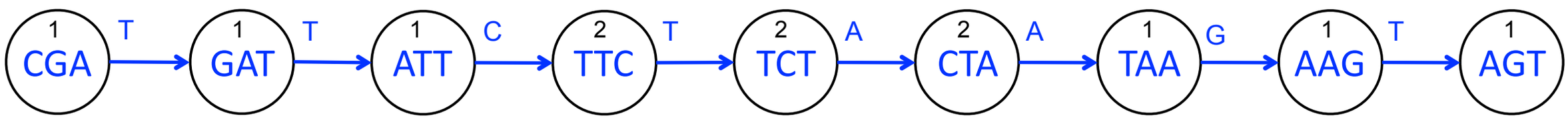


CGATTCTAAGT

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

```
>seq3
CGATTGTAAGT
```

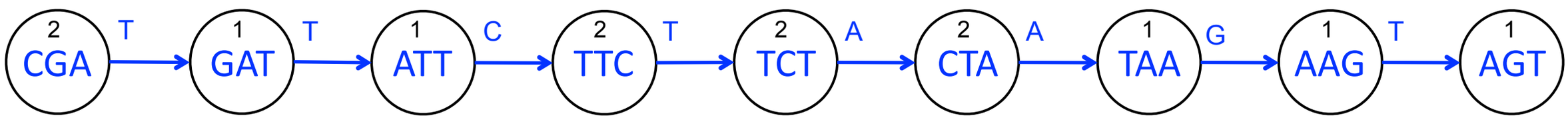


CGATTCTAAGT

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

```
>seq3
CGATTGTAAGT
```

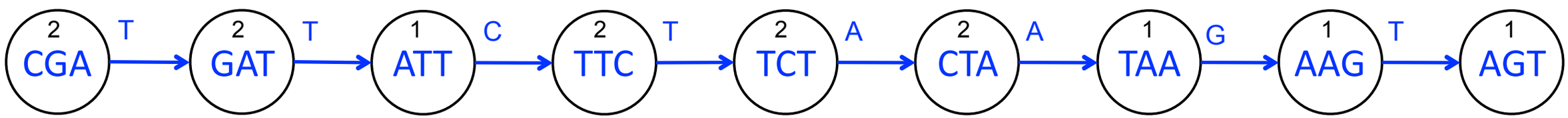


TGAC

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

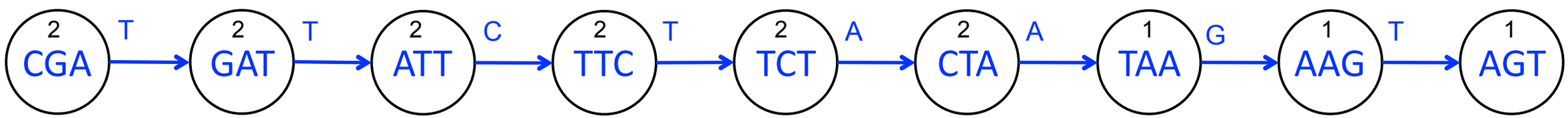
```
>seq3
CGATTGTAAGT
```



Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

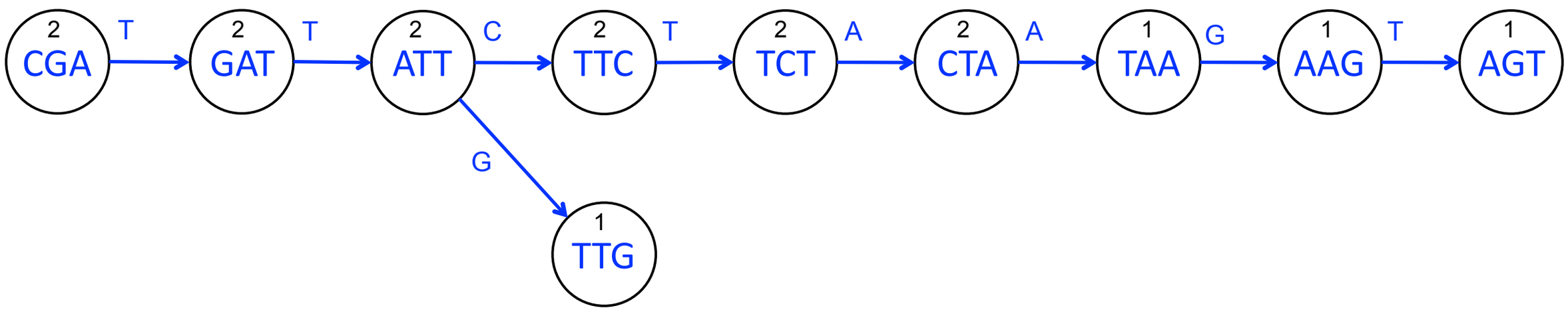
```
>seq3
CGATTGTAAGT
```



Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

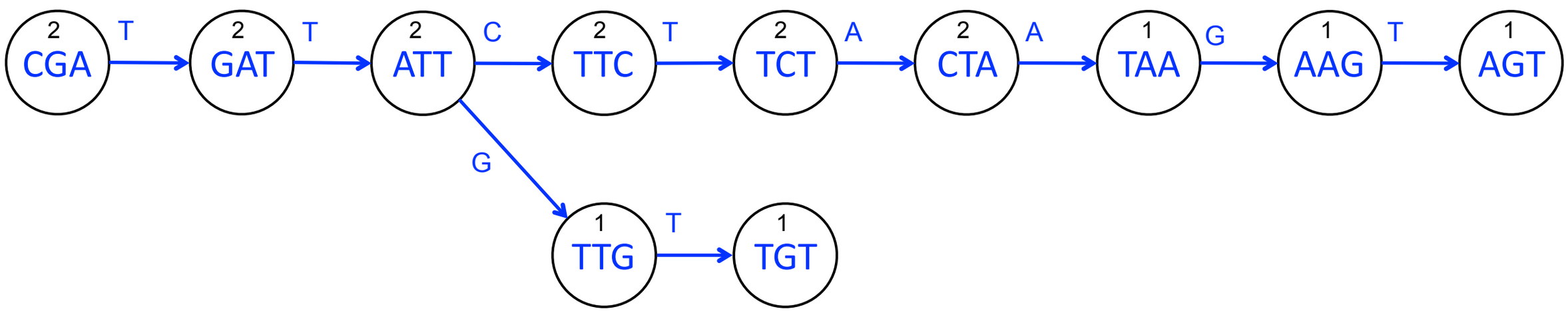
```
>seq3
CGATTGTAAGT
```



Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

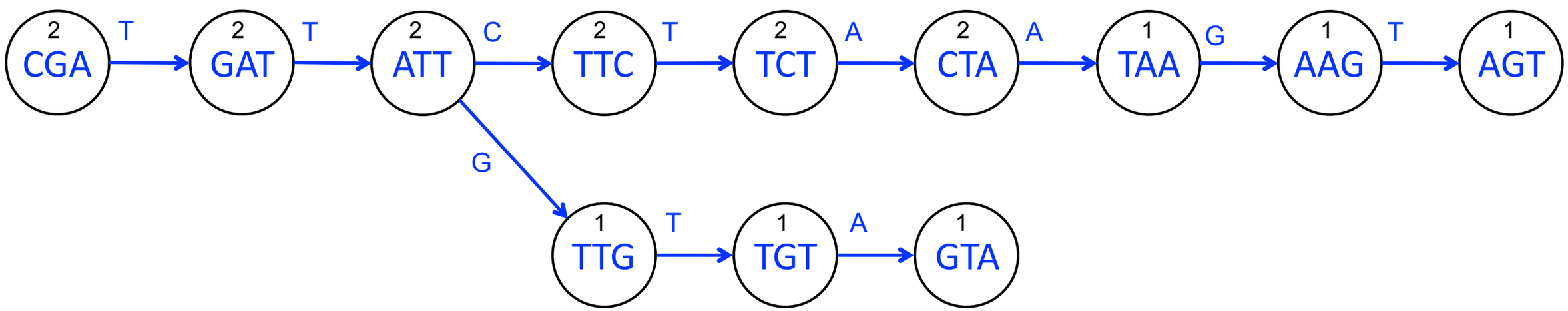
```
>seq3
CGATTGTAAGT
```



Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

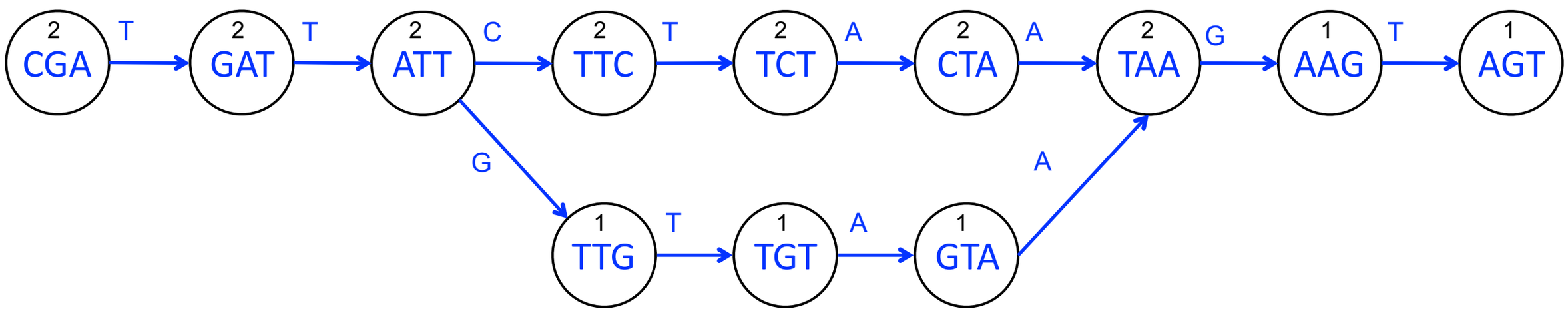
```
>seq3
CGATTGTAAGT
```



Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

```
>seq3
CGATTGTAAGT
```

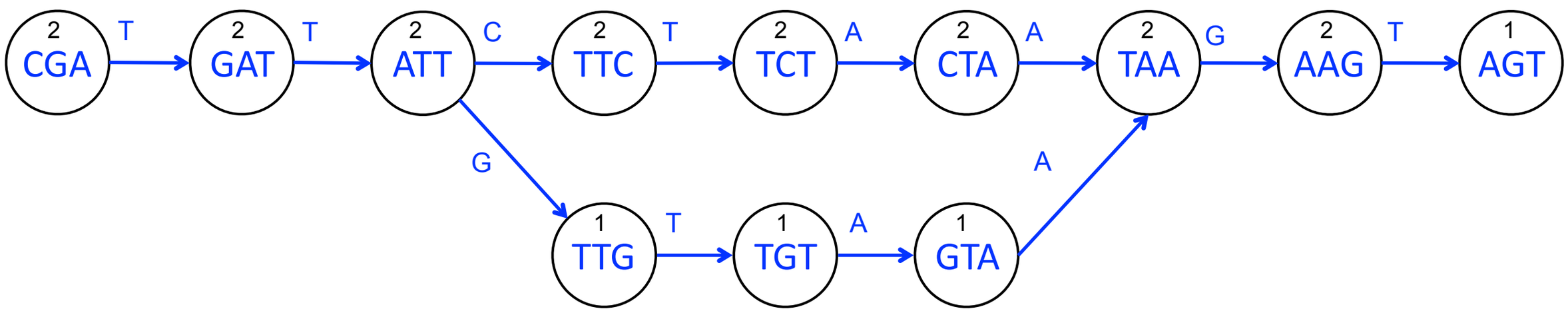


TGAC

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

```
>seq3
CGATTGTAAGT
```

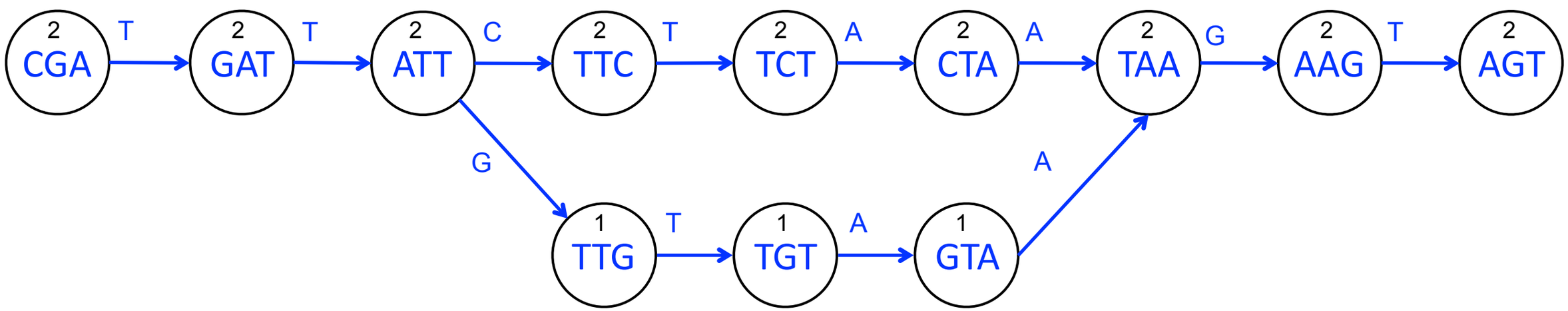


TGAC

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

```
>seq3
CGATTGTAAGT
```

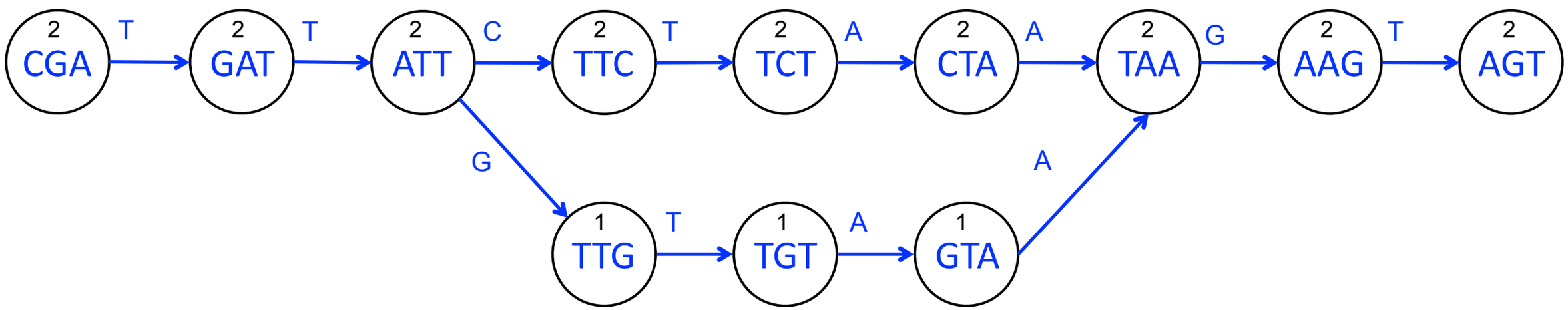


TGAC

Assembling a De Bruijn Graph

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

```
>seq3
CGATTGTAAGT
```



CGATTCTAAGT
CGATTGTAAGT

Graphs get complicated

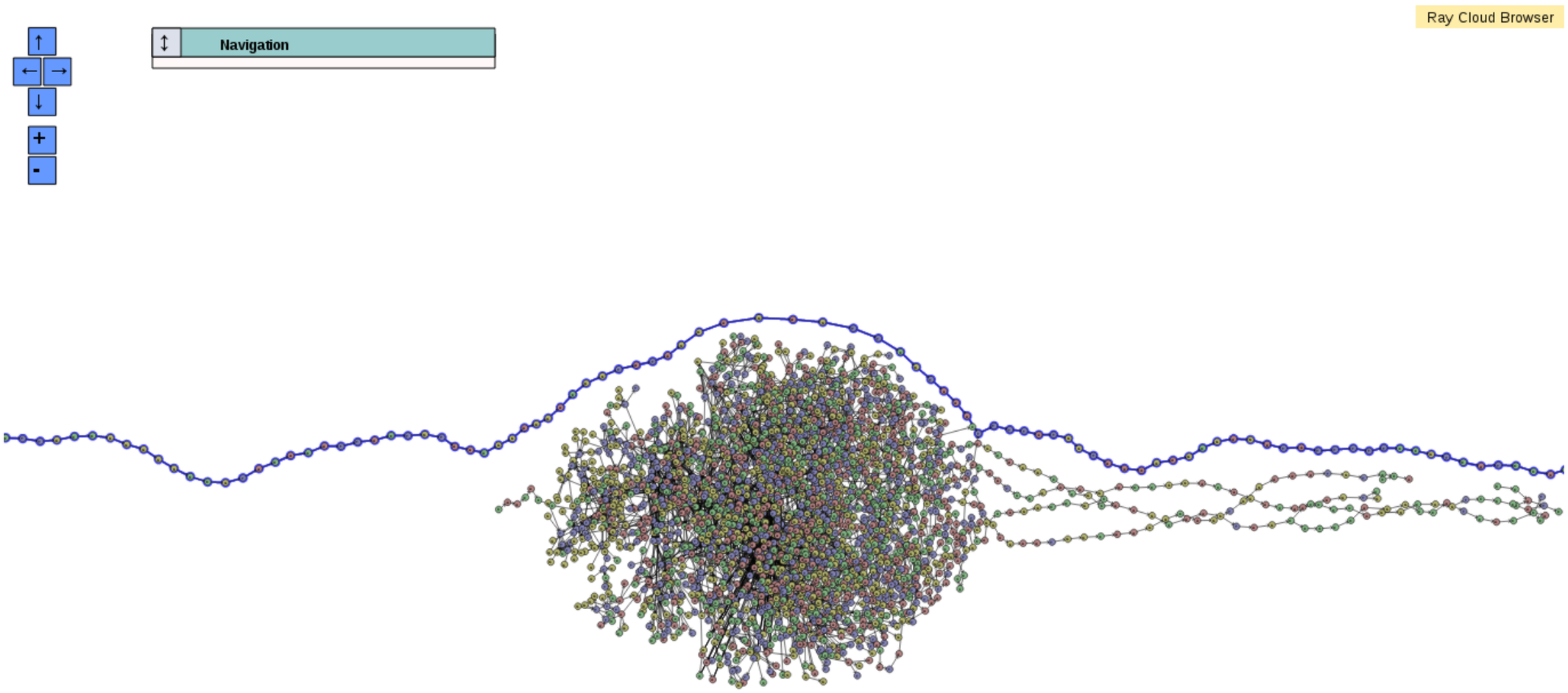
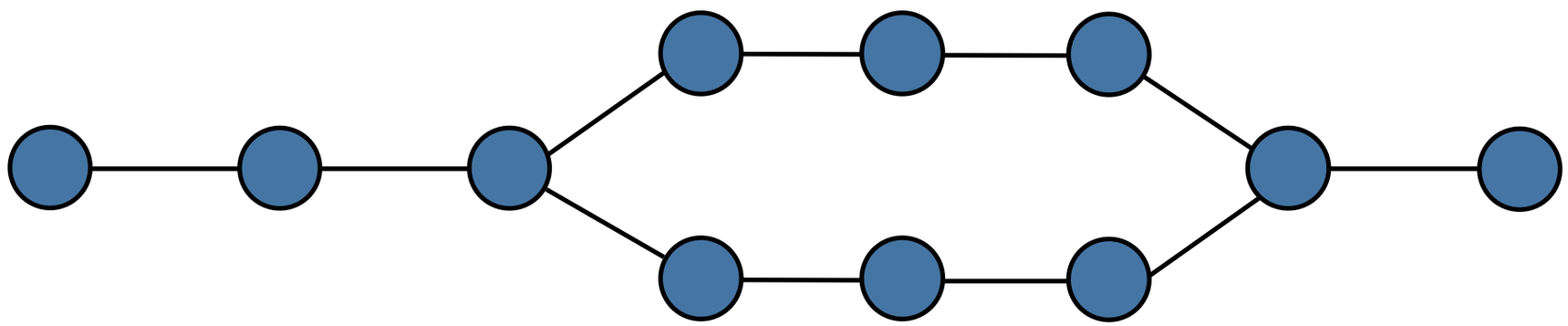


Image from dskernel.blogspot.co.uk

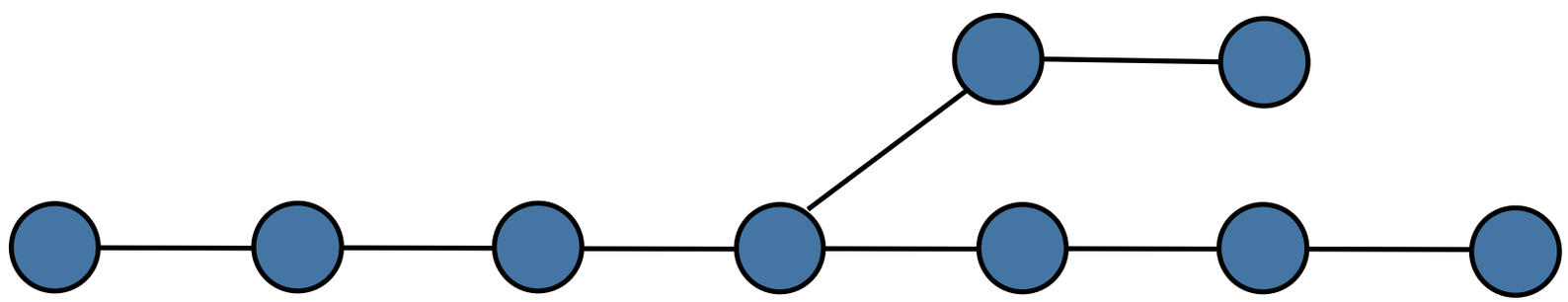
TGAC

Common structures

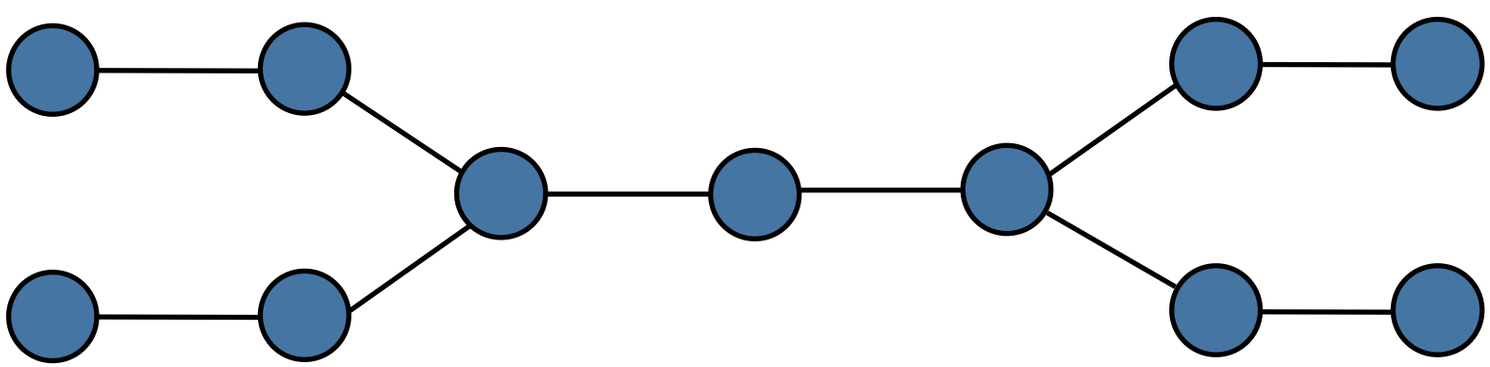
Polymorphism



Errant base call

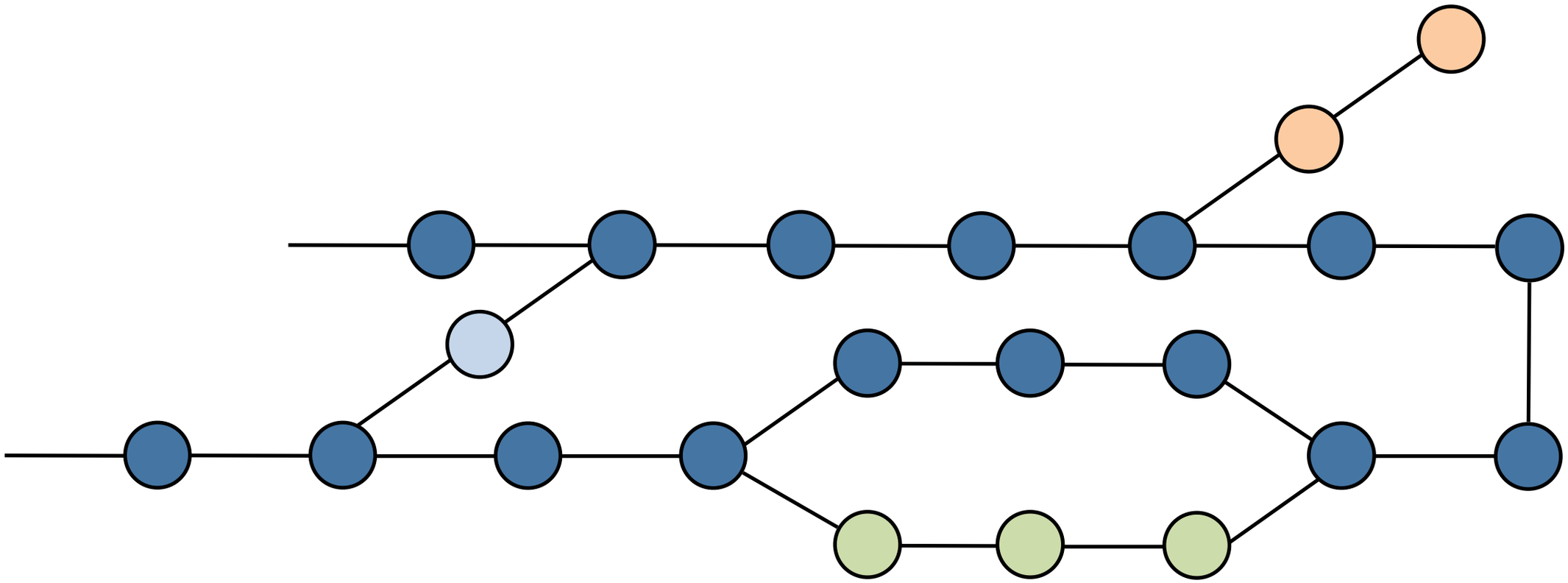


Repeating element

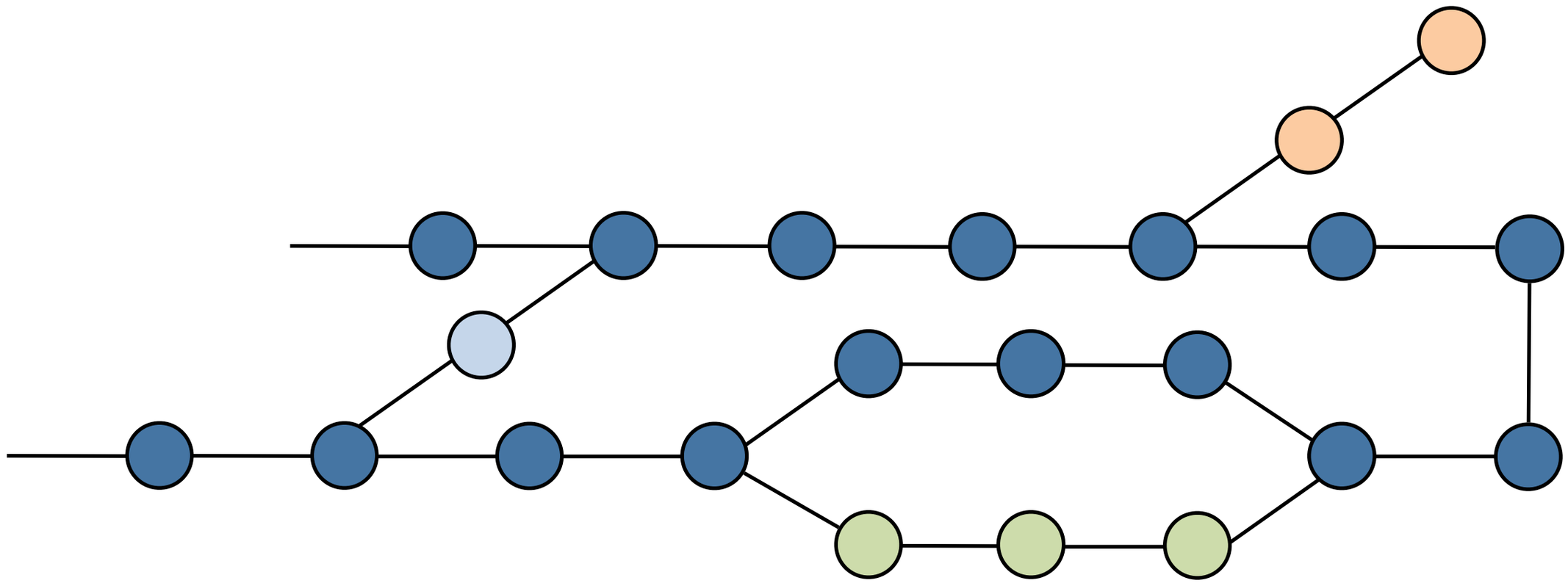


TGAC

Cleaning graphs

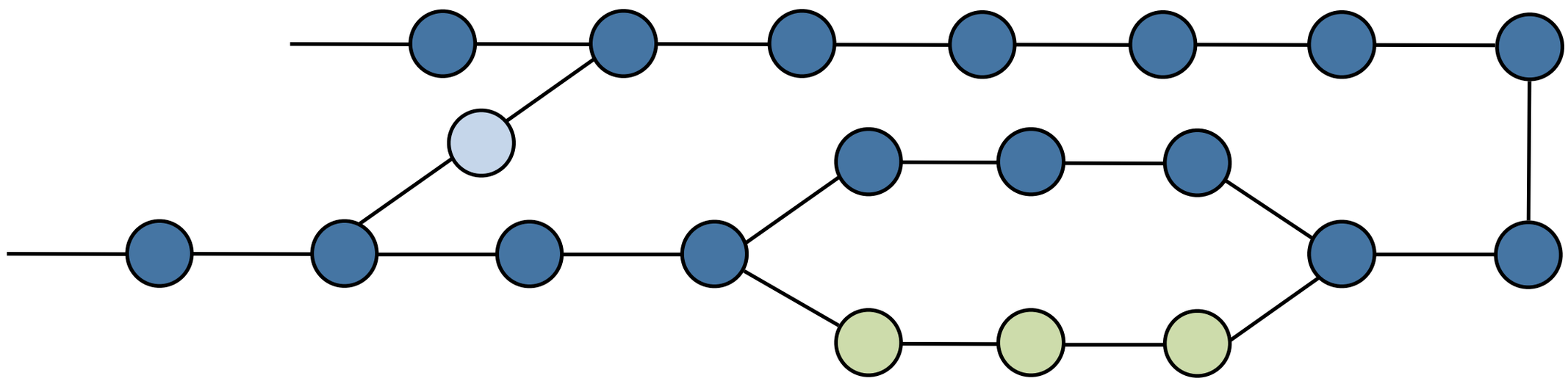


Cleaning graphs



Clip tips

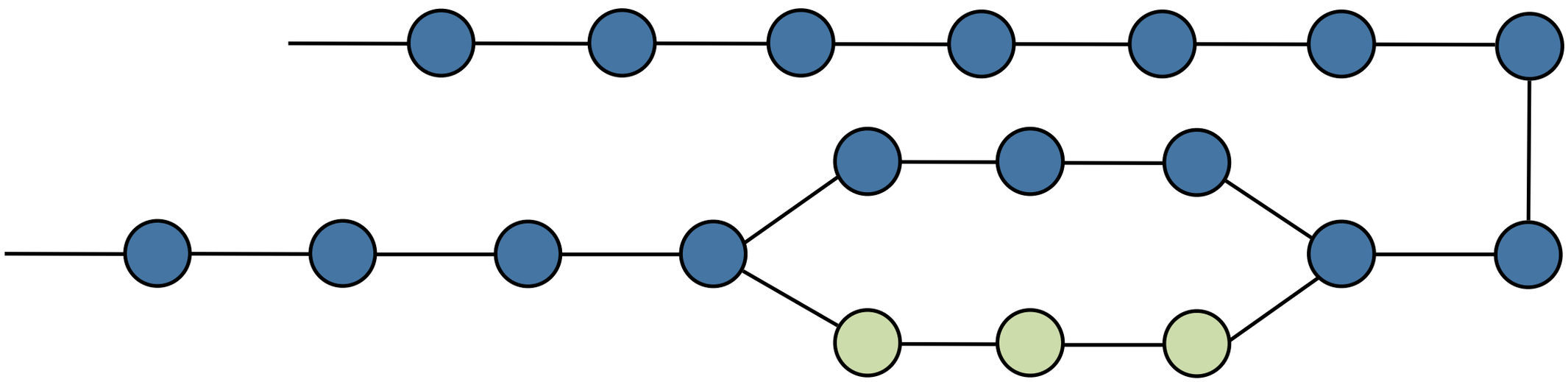
Cleaning graphs



Clip tips

Remove low coverage nodes

Cleaning graphs

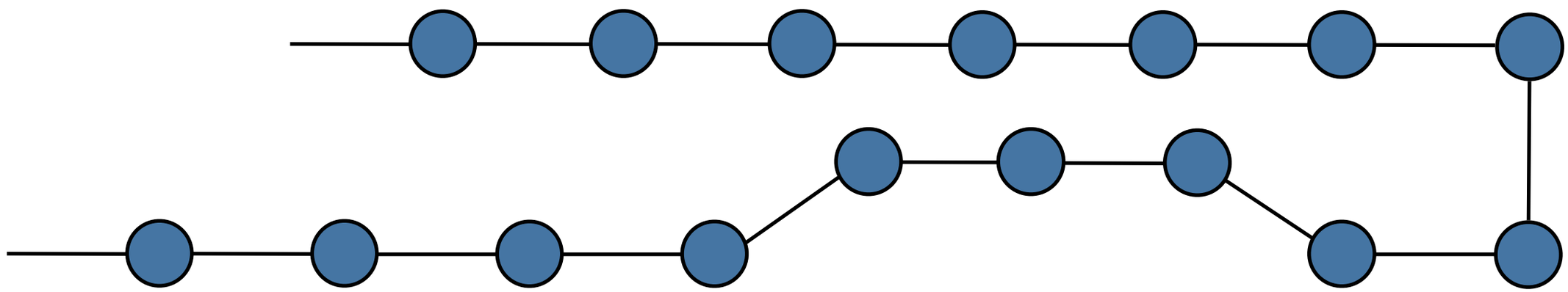


Clip tips

Remove low coverage nodes

Remove bubbles

Cleaning graphs

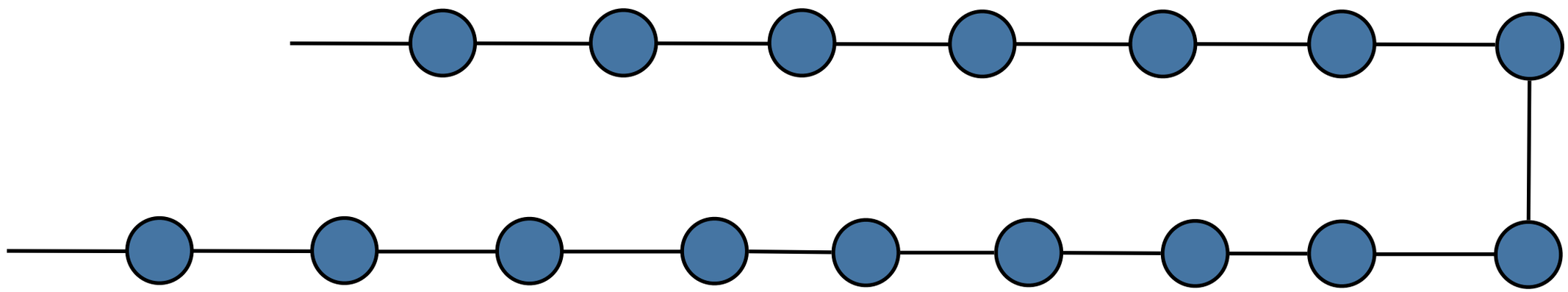


Clip tips

Remove low coverage nodes

Remove bubbles

Cleaning graphs



Clip tips

Remove low coverage nodes

Remove bubbles

OLC vs. De bruijn

	OLC	De bruijn
Representation of the problem	(Reads) Overlap Graph	De bruijn (kmer) graph
Steps to add a read	Insert read, compare to every read already included and insert overlaps	Insert new kmers or update count for those already present
Strengths	• Tracks reads • Intuitive representation • Consensus	• Computational speed • Ability to handle big datasets
Optimal depth	Just enough to cover genome and give accurate consensus	The higher the better (to grow SNR)
Typical sequencing technologies processed	Sanger, 454, Pacbio	Illumina, Ion Torrent

TGAC

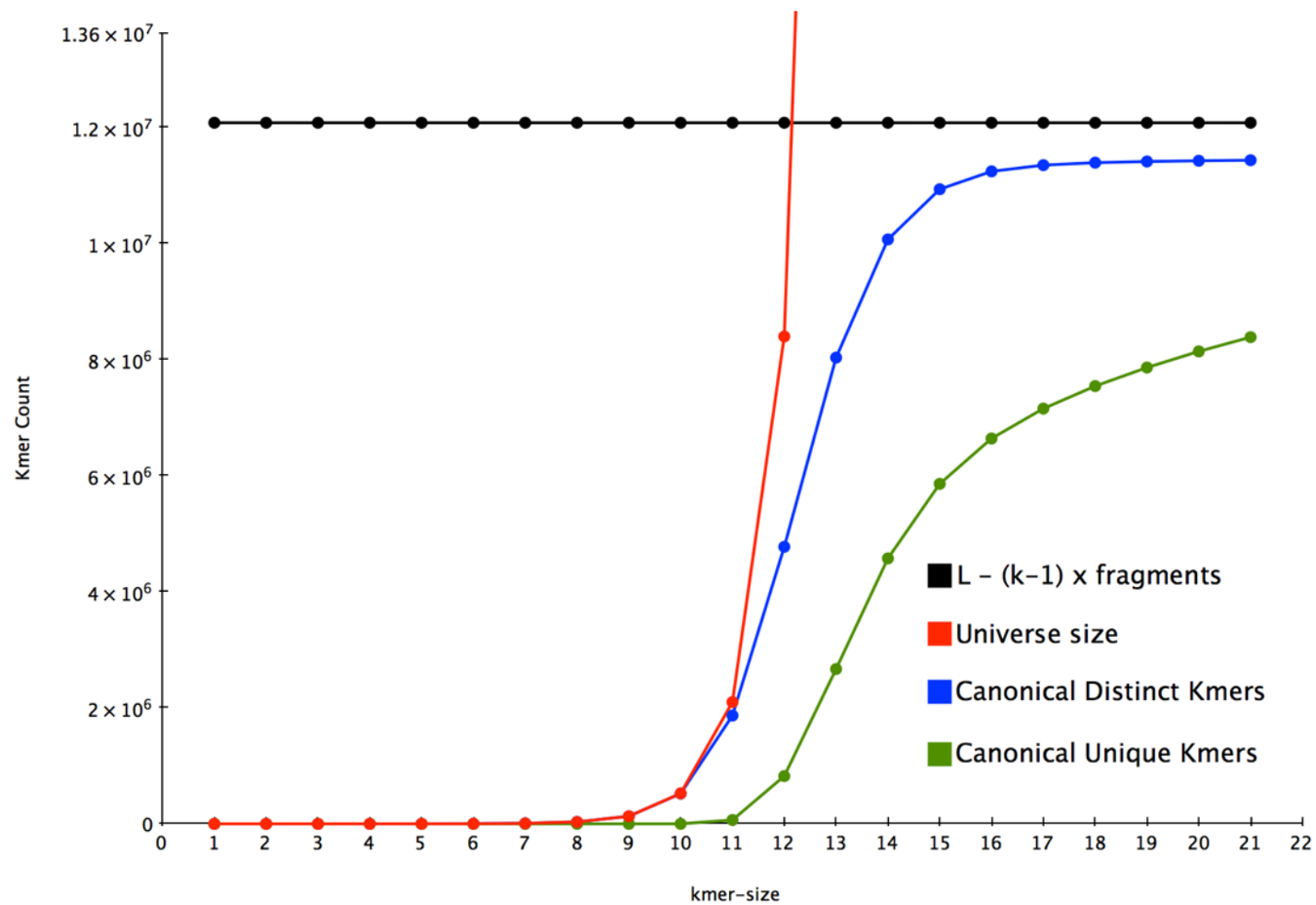
The size of the universe

	K is odd	K is even
Non-canonical representation	4^k	4^k
Canonical representation	$\frac{4^k}{2}$	$\frac{4^k + 4^{k/2}}{2}$

TGAC

The K tradeoff

- **Longer kmers** are more **unique** in the target, **disentangling** the graph.
- **Smaller kmers** will **overlap** more often, favouring **contiguity**.
- Every read produces **L-k+1** kmers.
  - Higher k -> less coverage.
- Every single error affects **k** kmers.
  - Higher k -> more errors.



- A typical choice for 100bp reads is k=71.

Resolving repeats using reads

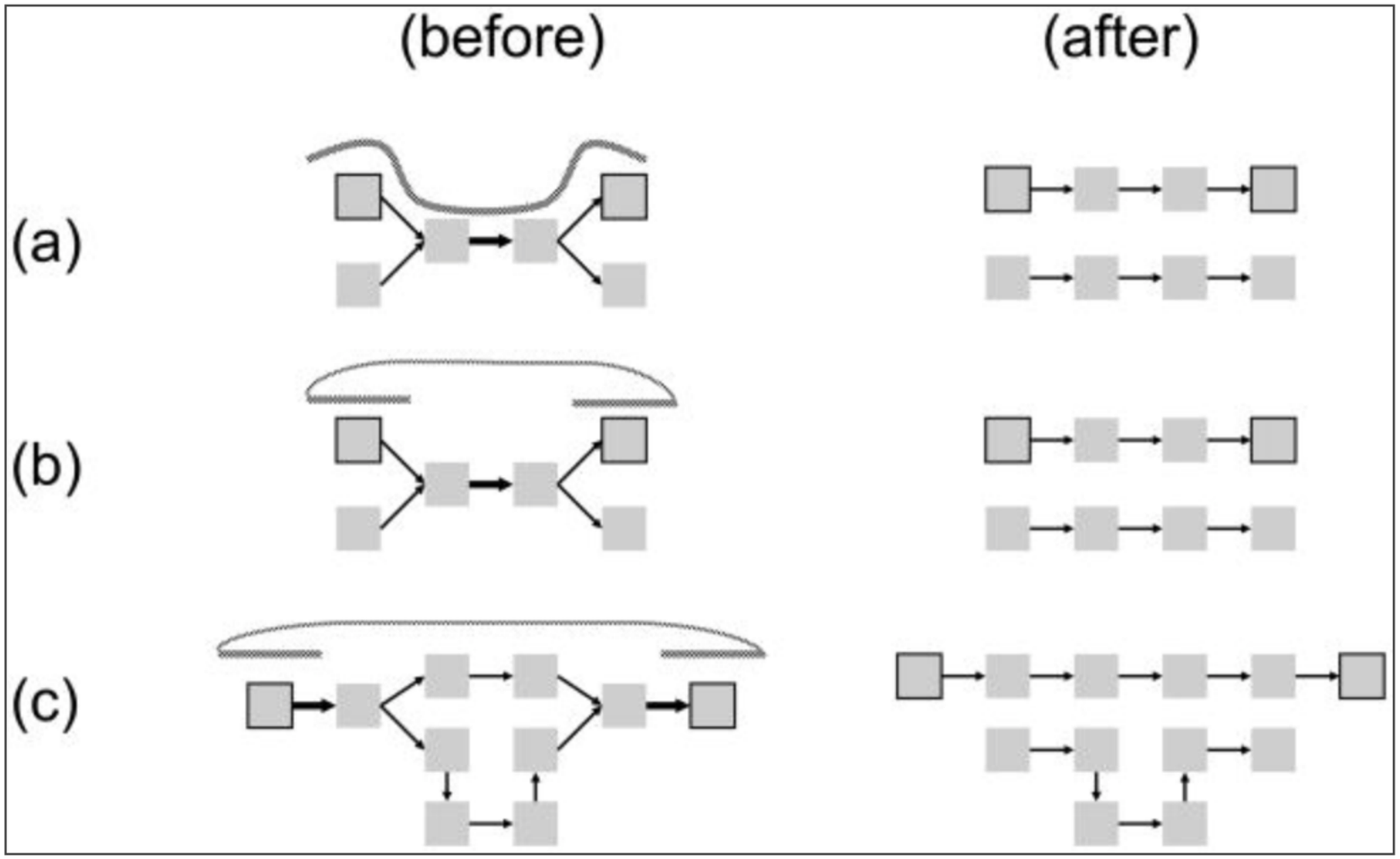


Figure 4

Three methods to resolve graph complexity. (a) Read threading joins paths across collapsed repeats that are shorter than the read lengths. (b) Mate threading joins paths across collapsed repeats that are shorter than the paired-end distances. (c) Path following chooses one path if its length fits the paired-end constraint. Reads and mates are shown as patterned lines. Not all tangles can be resolved by reads and mates. The non-branching paths are illustrative; they could be simplified to single edges or nodes.

Assembly Algorithms for Next-Generation Sequencing Data
Genomics. 2010 June;95(6):315-327.

TGAC

A correct assembly has:

The right *motifs*,

the correct number of times,

in correct order and position.

None of which is assessed by length stats.

TGAC

Example datasets

- Dataset A:
  - Target size ~5Mbp.
  - Illumina 2x300bp.
  - Pacbio long reads.
- Dataset B:
  - Target Size ~12Mbp.
  - Illumina 2x100bp.
  - Nextera LMP (2x300bp).

```
dataset_A
dataset_A/pacbio_long.fastq
dataset_A/pe_1.fastq
dataset_A/pe_2.fastq
dataset_B
dataset_B/nextera_lmp_1.fastq
dataset_B/nextera_lmp_2.fastq
dataset_B/pe_1.fastq
dataset_B/pe_2.fastq
```


TGAC

Questions?