Searching massive amounts of sequencing data using K-mers and graphs

Finlay Maguire

February 15, 2023

FCS, Dalhousie

Learning Objectives

- Describe the scale of available sequencing data

Learning Objectives

- Describe the scale of available sequencing data
- Understand that assembly loses information

Learning Objectives

- Describe the scale of available sequencing data
- Understand that assembly loses information
- Know why decomposing into k-mers is useful

Learning Objectives

- Describe the scale of available sequencing data
- Understand that assembly loses information
- Know why decomposing into k-mers is useful
- Describe how a de Bruijn graph can be an efficient representation of a k-mer set

Learning Objectives

- Describe the scale of available sequencing data
- Understand that assembly loses information
- Know why decomposing into k-mers is useful
- Describe how a de Bruijn graph can be an efficient representation of a k-mer set
- Understand how a de Bruijn graph can be coloured to represent multiple datasets

Learning Objectives

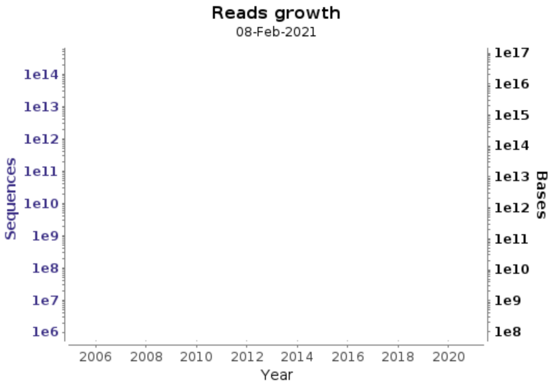
- Describe the scale of available sequencing data
- Understand that assembly loses information
- Know why decomposing into k-mers is useful
- Describe how a de Bruijn graph can be an efficient representation of a k-mer set
- Understand how a de Bruijn graph can be coloured to represent multiple datasets
- Outline the two main strategies for k-mer indexing: colour aggregative and k-mer aggregative

Learning Objectives

- Describe the scale of available sequencing data
- Understand that assembly loses information
- Know why decomposing into k-mers is useful
- Describe how a de Bruijn graph can be an efficient representation of a k-mer set
- Understand how a de Bruijn graph can be coloured to represent multiple datasets
- Outline the two main strategies for k-mer indexing: colour aggregative and k-mer aggregative
- Describe the core algorithm used by BlastFrost (colour aggregative) and BIGSI (k-mer aggregative)

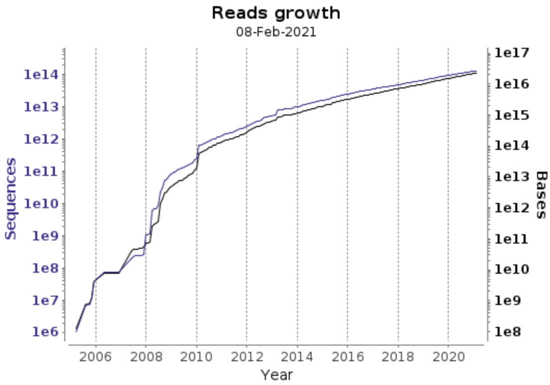
Massive datasets?

Sequencing Data Explosion



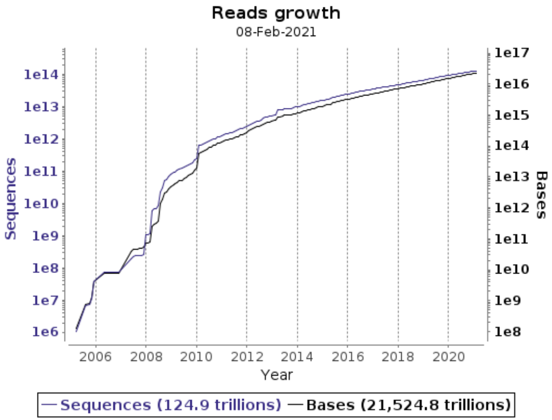
European Nucleotide Archive: Read Data

Sequencing Data Explosion



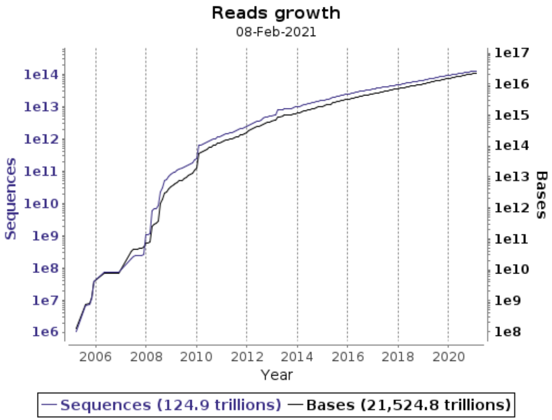
European Nucleotide Archive: Read Data

Sequencing Data Explosion



European Nucleotide Archive: Read Data

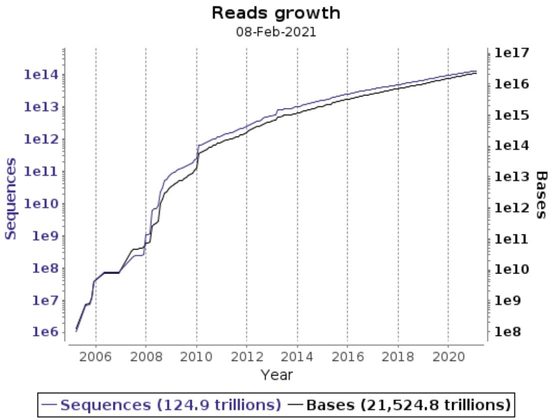
Sequencing Data Explosion



European Nucleotide Archive: Read Data

- Uncompressed at 2-bits per base:

Sequencing Data Explosion



European Nucleotide Archive: Read Data

- Uncompressed at 2-bits per base:
- 5,381.2 TB (without any metadata or accession information)

Searching all this data: surveillance of colistin resistance

Countries reporting plasmid-mediated colistin resistance encoded by *mcr-1*

Data source: Al-Tawfiq, J. A., Laxminarayan, R. & Mendelson, M. How should we respond to the emergence of plasmid-mediated colistin resistance in humans and animals? Int. J. Infect. Dis. (2016). doi:10.1016/j.ijid.2016.11.415

CDDEP THE CENTER FOR Disease Dynamics, Economics & Policy WASHINGTON DC • NEW DELHI

Searching all this data: surveillance of colistin resistance

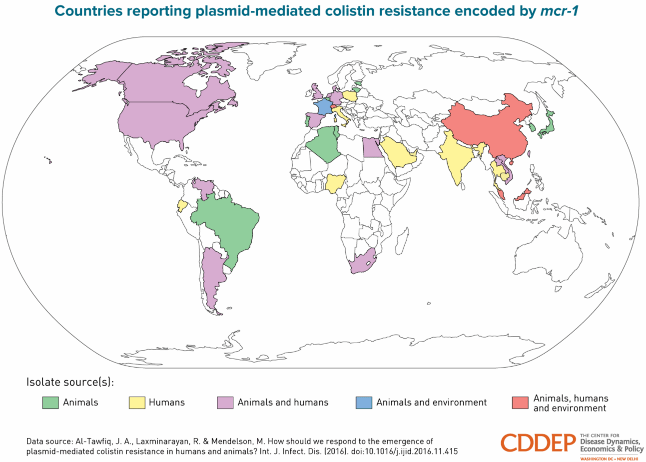
Countries reporting plasmid-mediated colistin resistance encoded by *mcr-1*

Animals | Humans | Animals and humans | Animals and environment | Animals, humans and environment

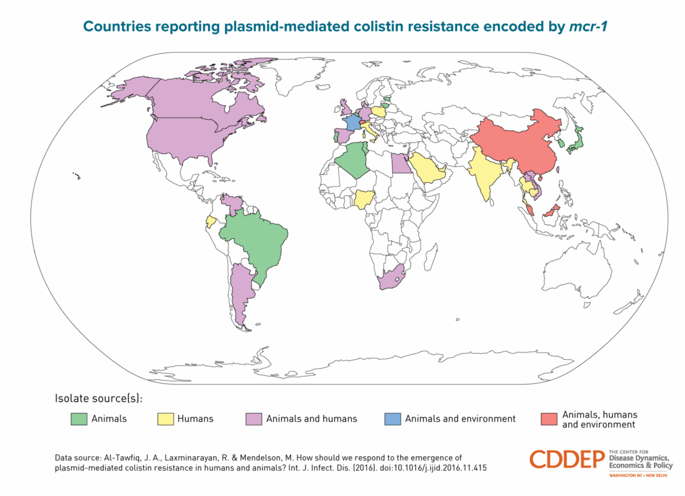
Data source: Al-Tawfiq, J. A., Laxminarayan, R. & Mendelson, M. How should we respond to the emergence of plasmid-mediated colistin resistance in humans and animals? Int. J. Infect. Dis. (2016). doi:10.1016/j.ijid.2016.11.415

CDDEP THE CENTER FOR Disease Dynamics, Economics & Policy WASHINGTON DC • NEW DELHI

Searching all this data: surveillance of colistin resistance



Searching all this data: surveillance of colistin resistance



- Which genome and metagenome read sets from all over the world contain MCR-1?

Formal Problem

- $\mathcal{D}$ is a collection of n sets of reads

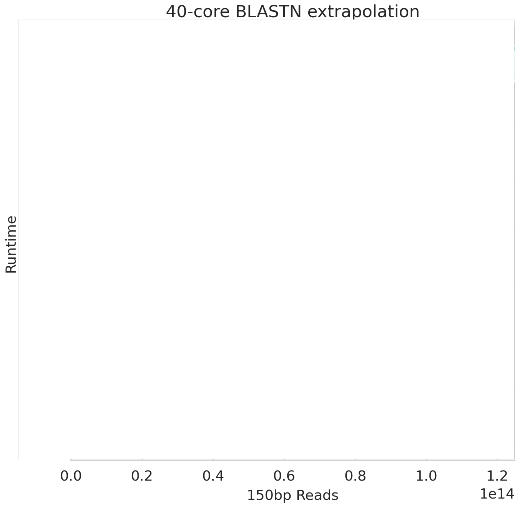
Formal Problem

- $\mathcal{D}$ is a collection of n sets of reads
- S is a query sequence of arbitrary length (including $> len(read)$)

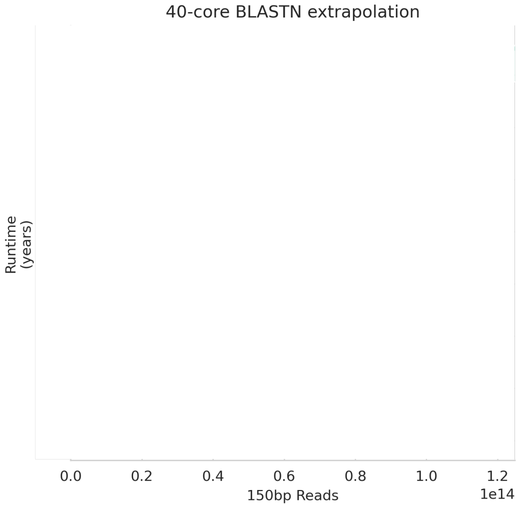
Formal Problem

- $\mathcal{D}$ is a collection of n sets of reads
- S is a query sequence of arbitrary length (including $> len(read)$)
- Identify which sets of reads in $\mathcal{D}$ contain S

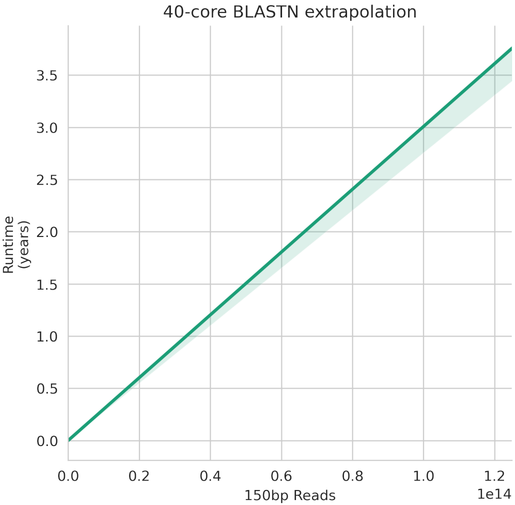
Just use BLAST?



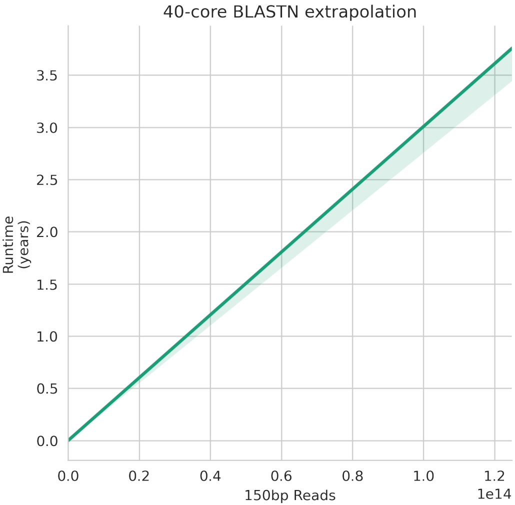
Just use BLAST?



Just use BLAST?

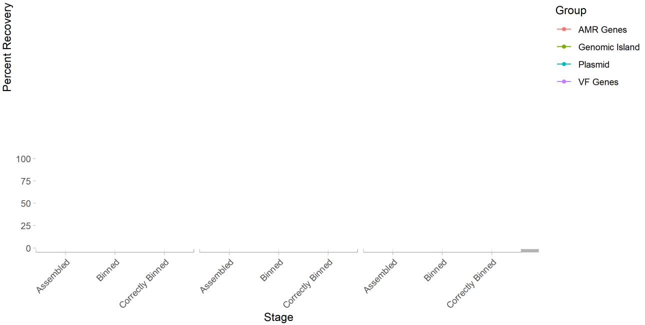


Just use BLAST?

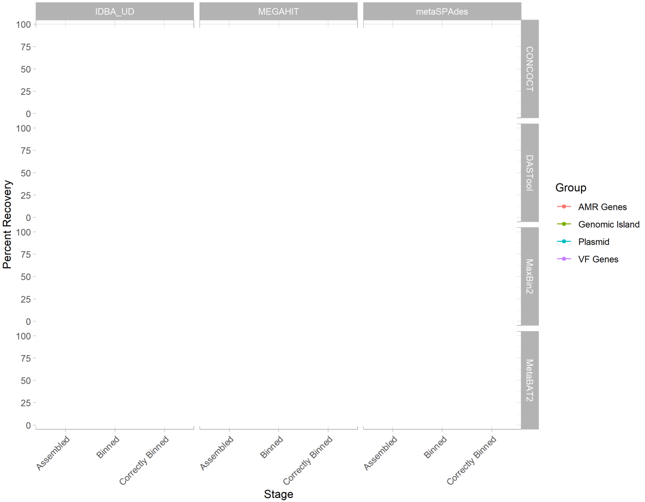


- By the end, there will be $\sim 3x$ more data than at the start.

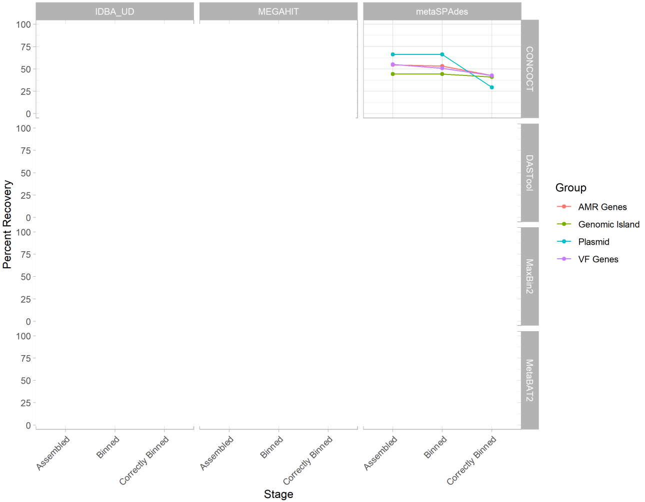
What about only using assembled data?



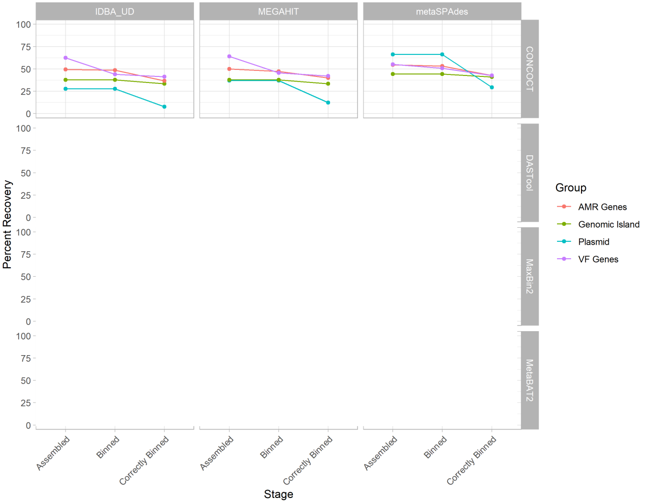
What about only using assembled data?



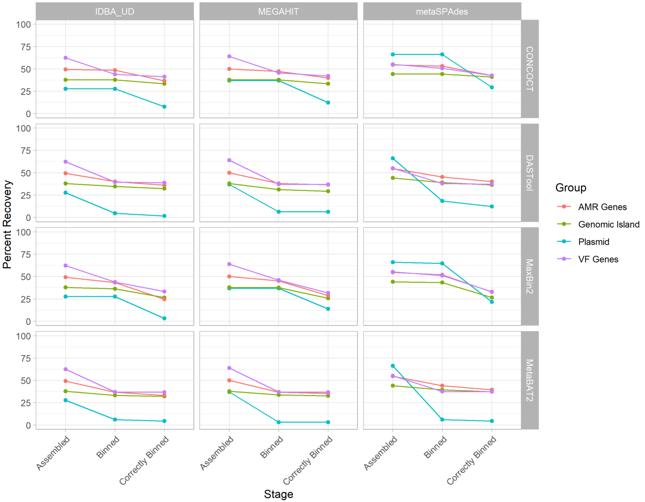
What about only using assembled data?



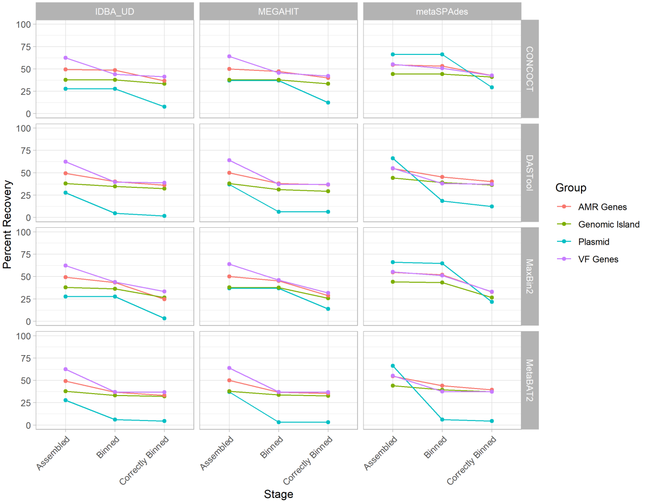
What about only using assembled data?



What about only using assembled data?



What about only using assembled data?



- No matter the method, assembly causes loss of information.

Let’s complicate but actually simplify this problem

k-mers!

Sequence Sets

k-mers!

Sequence Sets

AGCTCA

GGCTCA

k-mers!

Sequence Sets

AGCTCA

GGCTCA

Decompose to K-mers

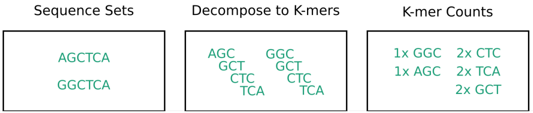
k-mers!

Sequence Sets
AGCTCA
GGCTCA
Decompose to K-mers
AGC
GCT
CTC
TCA
GGC
GCT
CTC
TCA

k-mers!

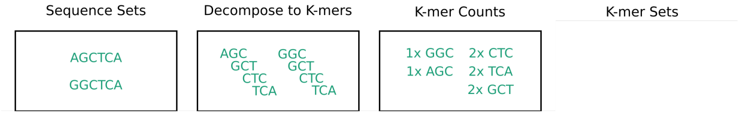

Sequence Sets
Decompose to K-mers
K-mer Counts
AGCTCA
GGCTCA
AGC
GCT
CTC
TCA
GGC
GCT
CTC
TCA

k-mers!

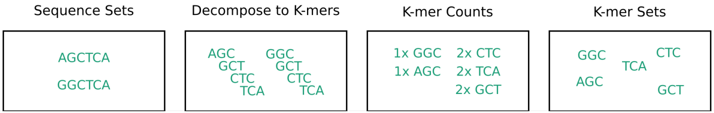


Sequence Sets
AGCTCA
GGCTCA
Decompose to K-mers
AGC GCT CTC TCA
GGC GCT CTC TCA
K-mer Counts
1x GGC
1x AGC
2x CTC
2x TCA
2x GCT

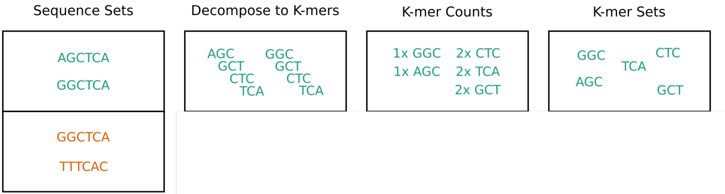
k-mers!


Sequence Sets
AGCTCA
GGCTCA
Decompose to K-mers
AGC GCT CTC TCA
GGC GCT CTC TCA
K-mer Counts
1x GGC
1x AGC
2x CTC
2x TCA
2x GCT
K-mer Sets

k-mers!

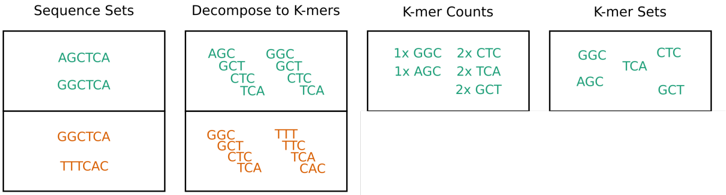

Sequence Sets
AGCTCA
GGCTCA
Decompose to K-mers
AGC GGC
GCT GCT
CTC CTC
TCA TCA
K-mer Counts
1x GGC 2x CTC
1x AGC 2x TCA
2x GCT
K-mer Sets
GGC CTC
TCA
AGC GCT

k-mers!



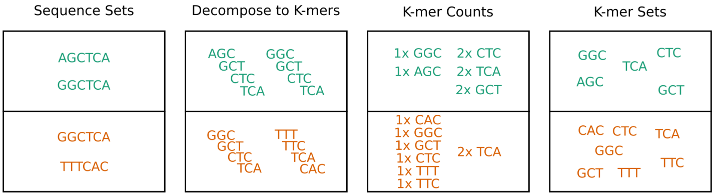
Sequence Sets
AGCTCA
GGCTCA
GGCTCA
TTTCAC
Decompose to K-mers
AGC
GCT
CTC
TCA
GGC
GCT
CTC
TCA
K-mer Counts
1x GGC
1x AGC
2x CTC
2x TCA
2x GCT
K-mer Sets
GGC
AGC
TCA
CTC
GCT

k-mers!



Sequence Sets
Decompose to K-mers
K-mer Counts
K-mer Sets
AGCTCA
GGCTCA
AGC GCT CTC TCA
GGC GCT CTC TCA
1x GGC 2x CTC
1x AGC 2x TCA
2x GCT
GGC CTC TCA AGC GCT
GGCTCA
TTTCAC
GGC GCT CTC TCA
TTT TTC TCA CAC

k-mers!


Sequence Sets
Decompose to K-mers
K-mer Counts
K-mer Sets
AGCTCA
GGCTCA
AGC
GCT
CTC
TCA
GGC
GCT
CTC
TCA
1x GGC
1x AGC
2x CTC
2x TCA
2x GCT
GGC
TCA
CTC
AGC
GCT
GGCTCA
TTTCAC
GGC
GCT
CTC
TCA
TTT
TTC
TCA
CAC
1x CAC
1x GGC
1x GCT
1x CTC
1x TTT
1x TTC
2x TCA
CAC
CTC
TCA
GGC
TTC
GCT
TTT

Formal Problem: querying the set of sets of k-mers

- $\mathcal{D}$ is a collection of n sets of ~~reads~~ k-mers

Formal Problem: querying the set of sets of k-mers

- $\mathcal{D}$ is a collection of n sets of ~~reads~~ k-mers
- S is a query sequence of arbitrary length (including $>$~~read-length~~ k)

Formal Problem: querying the set of sets of k-mers

- $\mathcal{D}$ is a collection of n sets of ~~reads~~ k-mers
- S is a query sequence of arbitrary length (including $>$~~read-length~~ k)
- Identify which sets of ~~reads~~ k-mers in $\mathcal{D}$ contain S

Formal Problem: querying the set of sets of k-mers

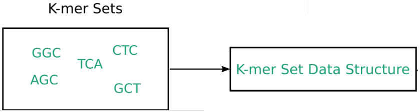
- $\mathcal{D}$ is a collection of n sets of ~~reads~~ k-mers
- S is a query sequence of arbitrary length (including $>$~~read-length~~ k)
- Identify which sets of ~~reads~~ k-mers in $\mathcal{D}$ contain S
- Bonus: also applicable to anything you can decompose into k-mers e.g., assembled sequences and long-reads

Algorithms to query a set of k-mer sets

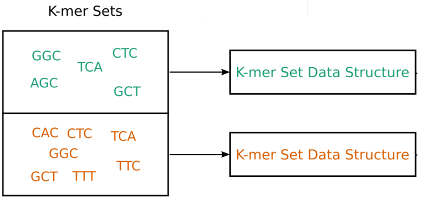
Components of a solution

K-mer Sets

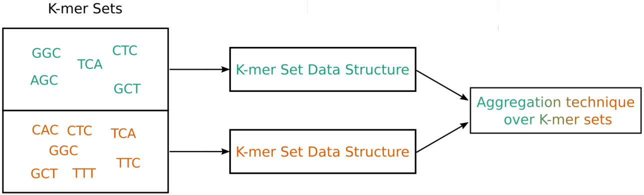
Components of a solution



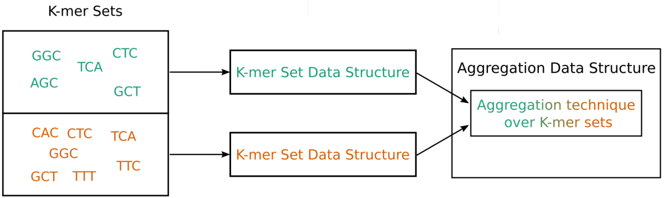
Components of a solution



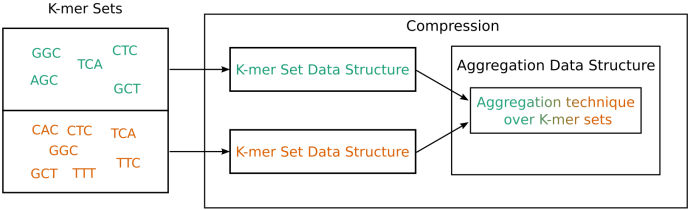
Components of a solution



Components of a solution



Components of a solution



Indexing a single set of k-mers

K-mer Set Data Structure: de Bruijn graphs

sequence ATGGAAGTCGCGGAATC

homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs

sequence ATGGAAGTCGCGGAATC

7mers

homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs

sequence	**ATGGAAGTCGCGGAATC**
7mers	ATGGAAG

homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs

sequence	ATGGAAGTCGCGGAATC
7mers	ATGGAAG TGGAAGT

homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs

sequence ATGGAAGTCGCGGAATC

7mers

```
ATGGAAG
 TGGAAGT
  GGAAGTC
   GAAGTCG
    AAGTCGC
     AGTCGCG
      GTCGCGG
       TCGCGGA
        CGCGGAA
         GCGGAAT
          CGGAATC
```

homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs

sequence **ATGGAAGTCGCGGAATC**

7mers

```
ATGGAAG
 TGGAAGT
  GGAAGTC
   GAAGTCG
    AAGTCGC
     AGTCGCG
      GTCGCGG
       TCGCGGA
        CGCGGAA
         GCGGAAT
          CGGAATC
```

de Bruijn graph

homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs

sequence **ATGGAAGTCGCGGAATC**

7mers

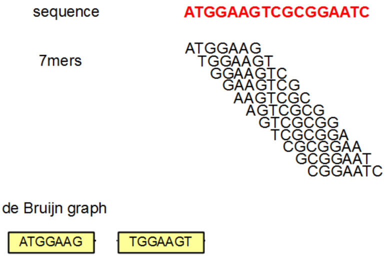
```
ATGGAAG
 TGGAAGT
  GGAAGTC
   GAAGTCG
    AAGTCGC
     AGTCGCG
      GTCGCGG
       TCGCGGA
        CGCGGAA
         GCGGAAT
          CGGAATC
```

de Bruijn graph

$\boxed{\text{ATGGAAG}}$

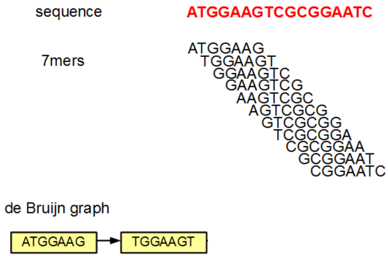
homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs



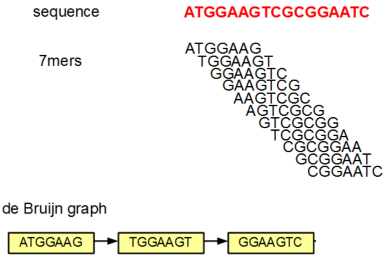
homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs



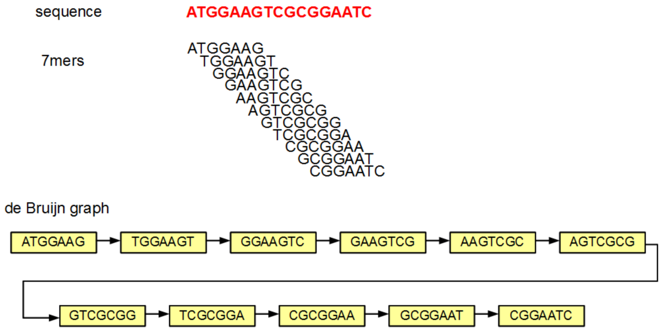
homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs



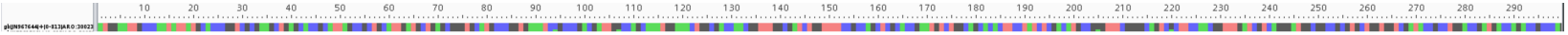
homolog.us/Tutorials/book4/p2.1.html

K-mer Set Data Structure: de Bruijn graphs



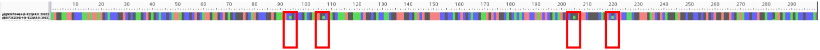
homolog.us/Tutorials/book4/p2.1.html

de Bruijn graph collapses diversity: NDM

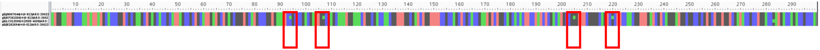


de Bruijn graph collapses diversity: NDM

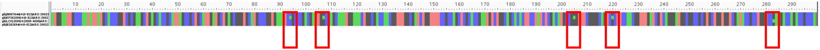
de Bruijn graph collapses diversity: NDM



de Bruijn graph collapses diversity: NDM

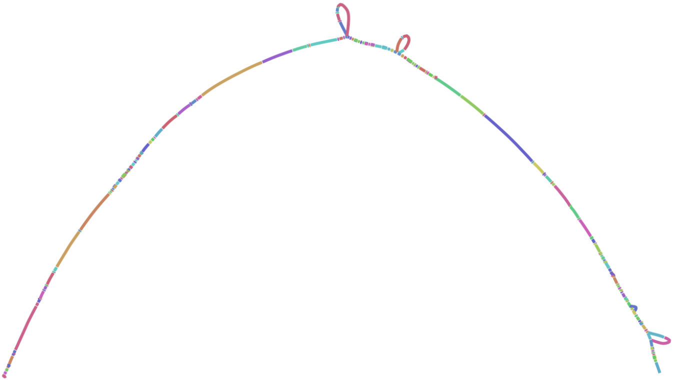


de Bruijn graph collapses diversity: NDM

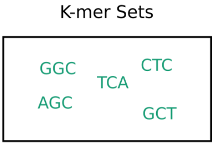


de Bruijn graph collapses diversity: NDM

de Bruijn graph collapses diversity: NDM

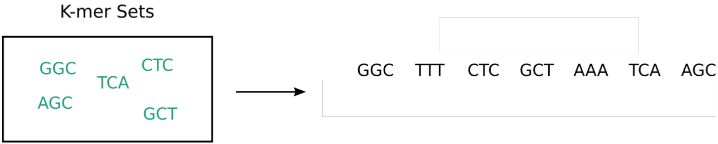


K-mer Set Data Structure: Bit-Vector



K-mer Set Data Structure: Bit-Vector

K-mer Set Data Structure: Bit-Vector



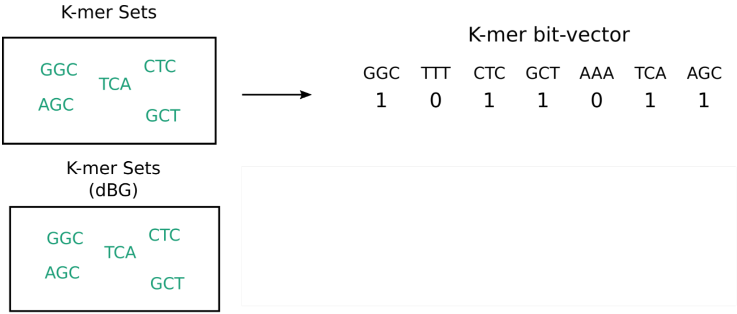
K-mer Set Data Structure: Bit-Vector

K-mer Set Data Structure: Bit-Vector

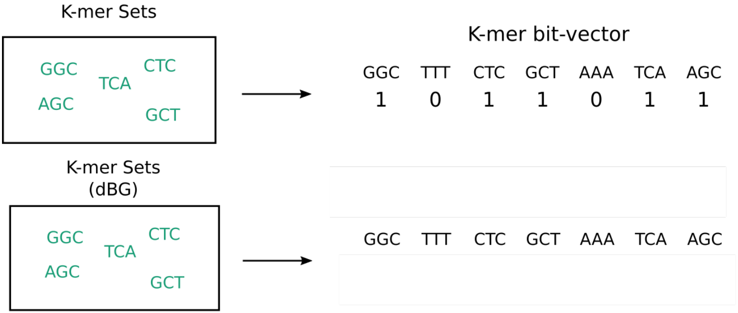
K-mer Set Data Structure: Bit-Vector

K-mer Set Data Structure: Bit-Vector

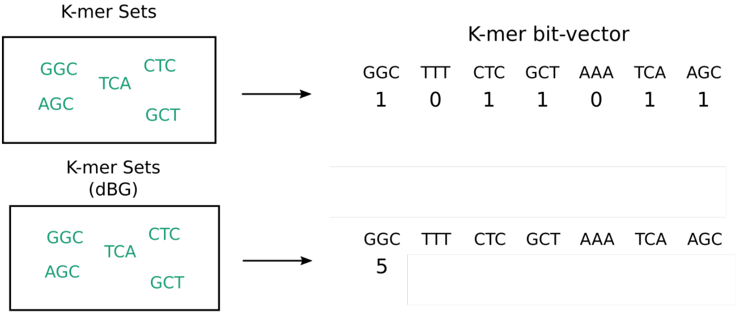
K-mer Set Data Structure: Bit-Vector



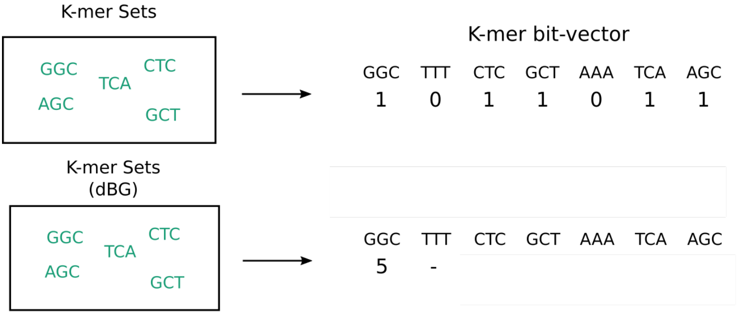
K-mer Set Data Structure: Bit-Vector



K-mer Set Data Structure: Bit-Vector

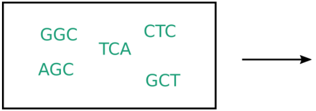


K-mer Set Data Structure: Bit-Vector



K-mer Set Data Structure: Bit-Vector

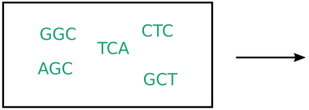
K-mer Sets



K-mer bit-vector

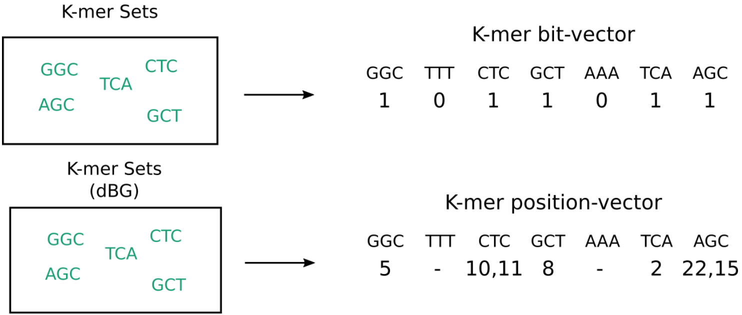
GGC	TTT	CTC	GCT	AAA	TCA	AGC
1	0	1	1	0	1	1

K-mer Sets
(dBG)



GGC	TTT	CTC	GCT	AAA	TCA	AGC
5	-	10,11	8	-	2	22,15

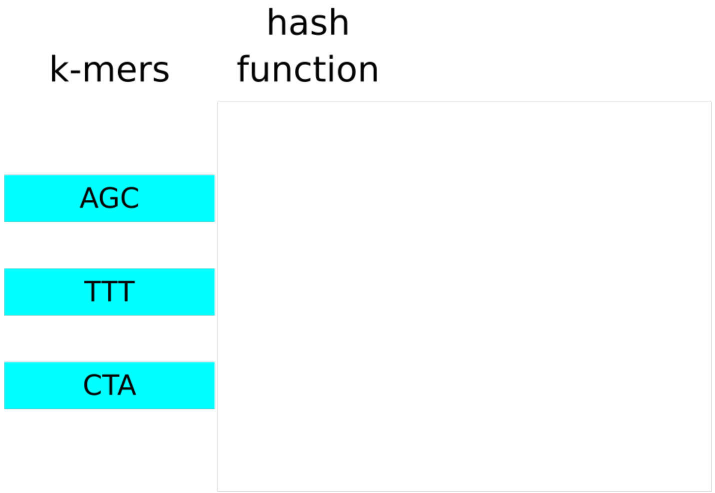
K-mer Set Data Structure: Bit-Vector



K-mer Set Data Structure: making bit-vectors more efficient

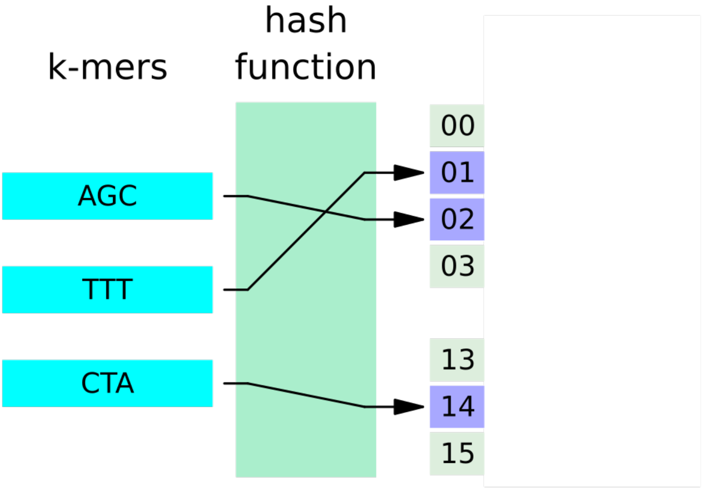
Adapted from *wikimedia.org/wiki/File:Hash_table_3_1_1_0_1_0_0_SP.svg*

K-mer Set Data Structure: making bit-vectors more efficient



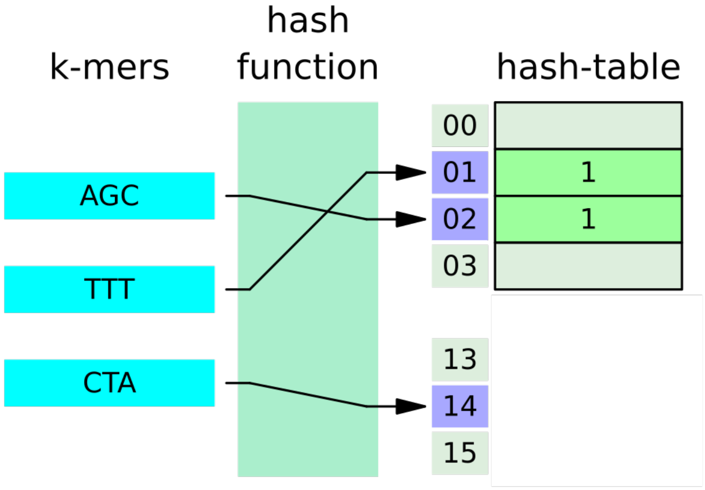
Adapted from `wikimedia.org/wiki/File:Hash_table_3_1_1_0_1_0_0_SP.svg`

K-mer Set Data Structure: making bit-vectors more efficient



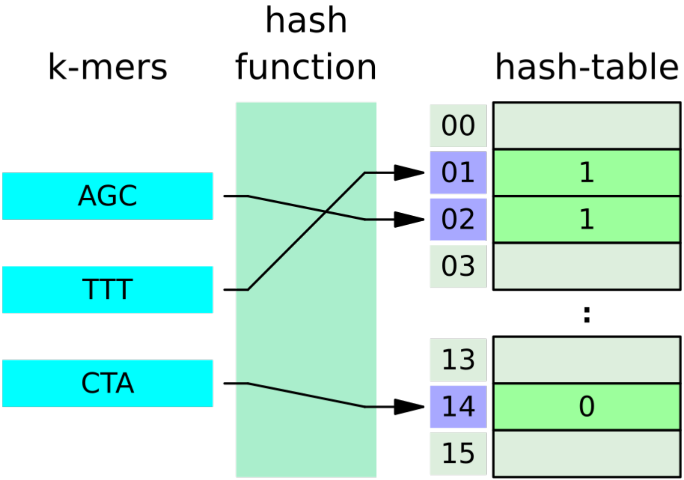
Adapted from `wikimedia.org/wiki/File:Hash_table_3_1_1_0_1_0_0_SP.svg`

K-mer Set Data Structure: making bit-vectors more efficient



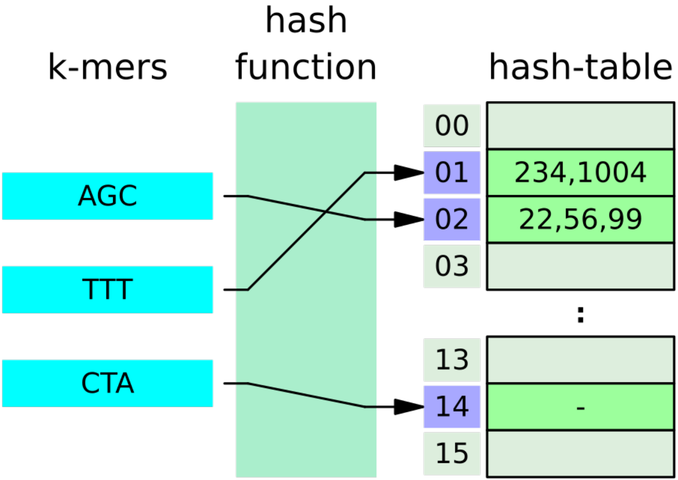
Adapted from `wikimedia.org/wiki/File:Hash_table_3_1_1_0_1_0_0_SP.svg`

K-mer Set Data Structure: making bit-vectors more efficient



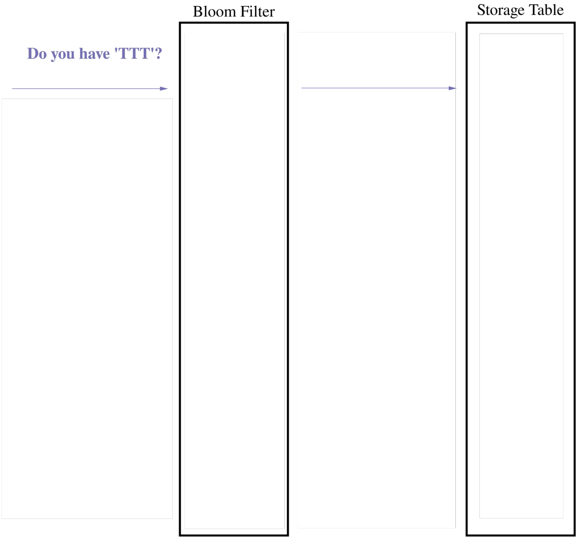
Adapted from `wikimedia.org/wiki/File:Hash_table_3_1_1_0_1_0_0_SP.svg`

K-mer Set Data Structure: making bit-vectors more efficient



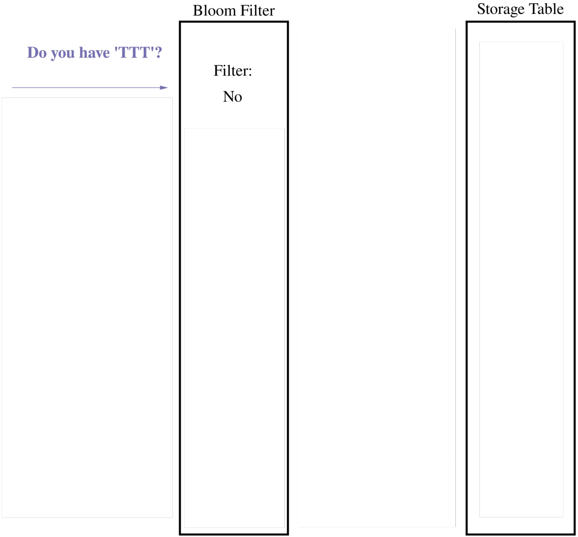
Adapted from `wikimedia.org/wiki/File:Hash_table_3_1_1_0_1_0_0_SP.svg`

K-mer Set Data Structure: bloom filters



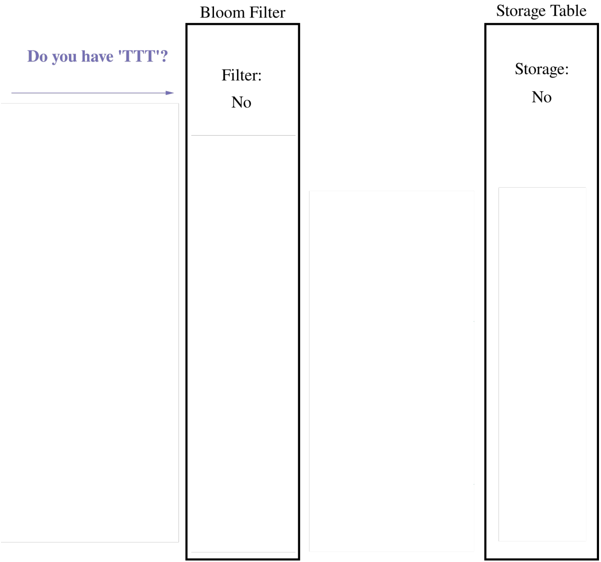
Adapted from *wikimedia.org/wiki/File:Bloom_filter_speed.svg*

K-mer Set Data Structure: bloom filters



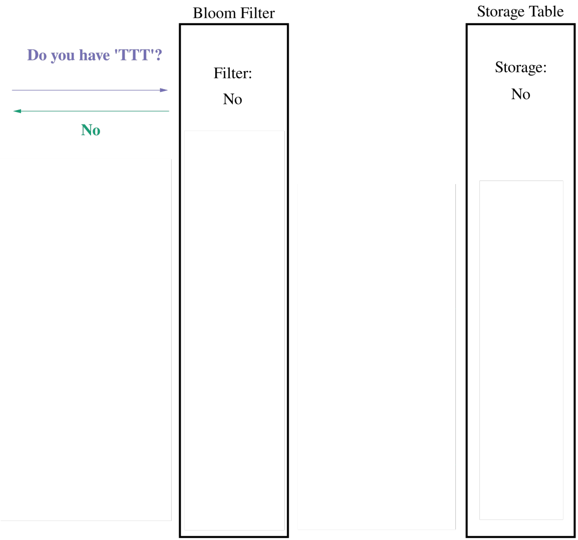
Adapted from *wikimedia.org/wiki/File:Bloom_filter_speed.svg*

K-mer Set Data Structure: bloom filters



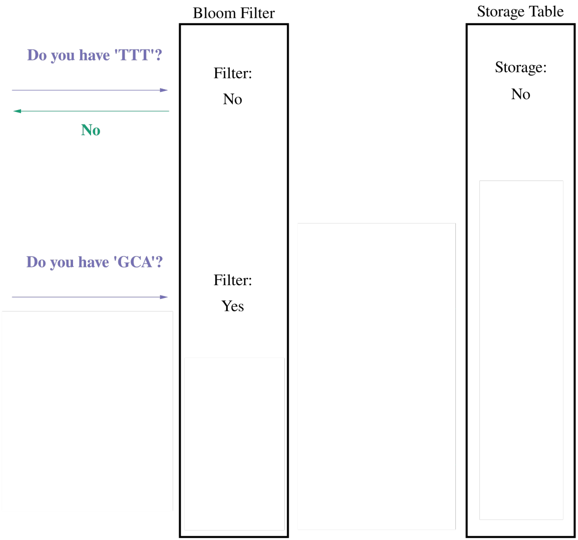
Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

K-mer Set Data Structure: bloom filters



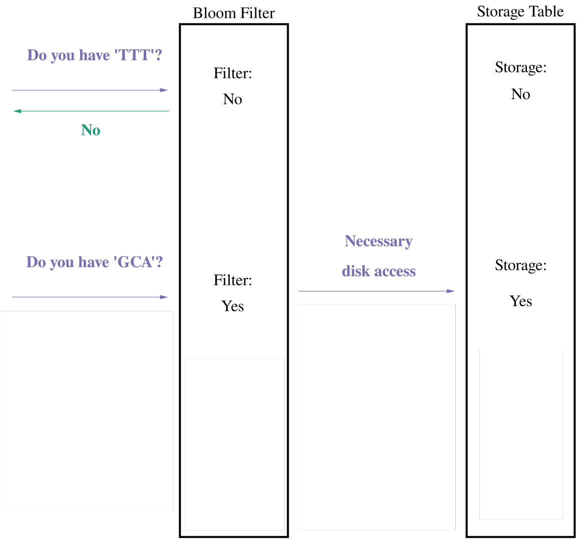
Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

K-mer Set Data Structure: bloom filters



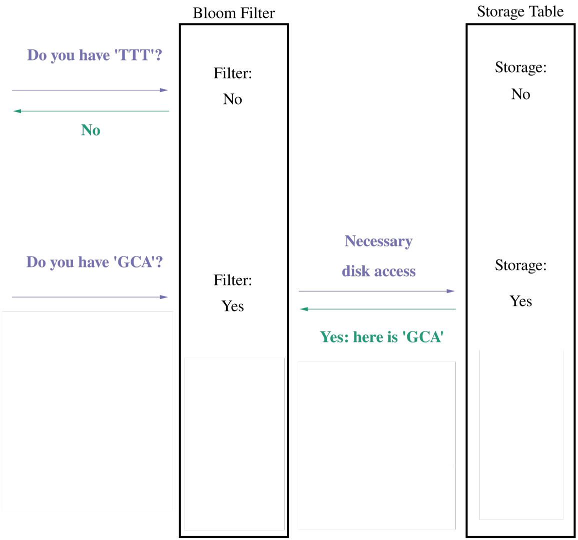
Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

K-mer Set Data Structure: bloom filters



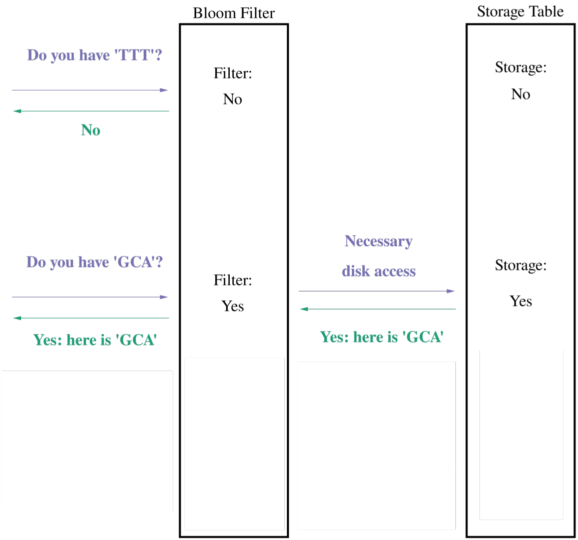
Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

K-mer Set Data Structure: bloom filters



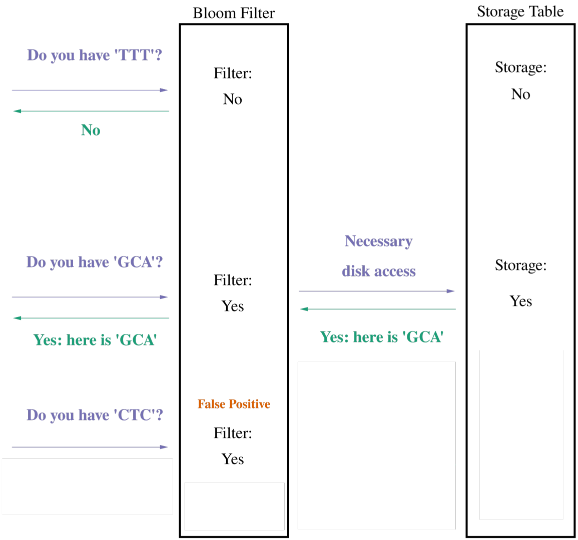
Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

K-mer Set Data Structure: bloom filters



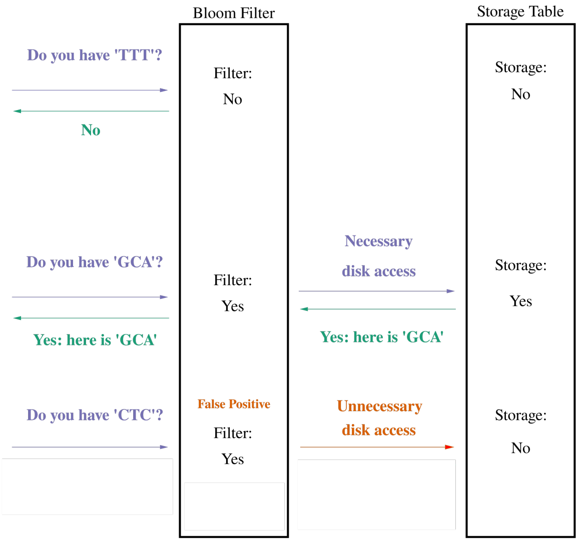
Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

K-mer Set Data Structure: bloom filters



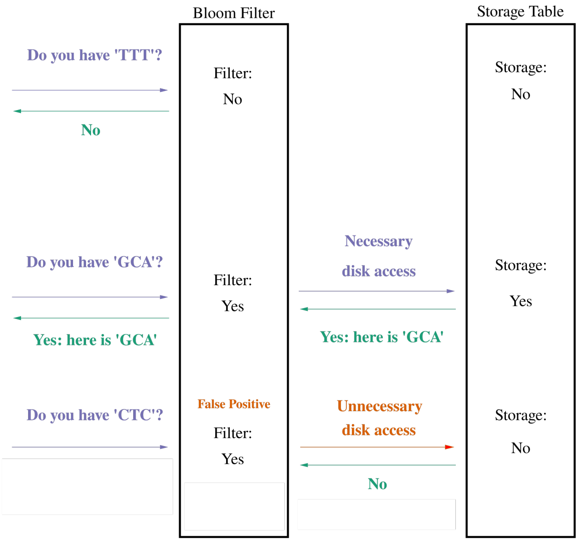
Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

K-mer Set Data Structure: bloom filters



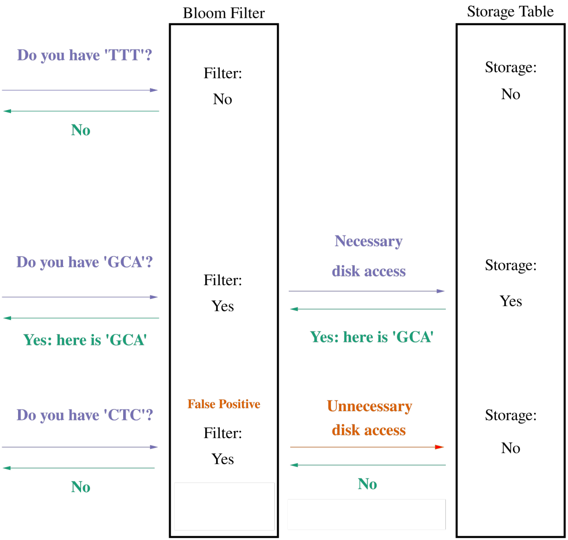
Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

K-mer Set Data Structure: bloom filters



Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

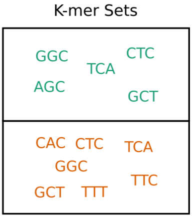
K-mer Set Data Structure: bloom filters



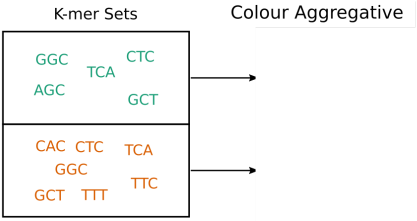
Adapted from `wikimedia.org/wiki/File:Bloom_filter_speed.svg`

How do we index across sets of k-mers?

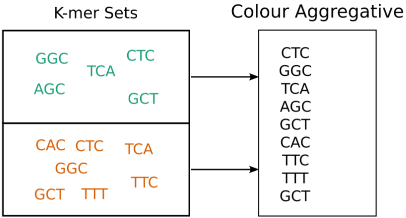
Two possible approaches: colour or k-mer aggregative



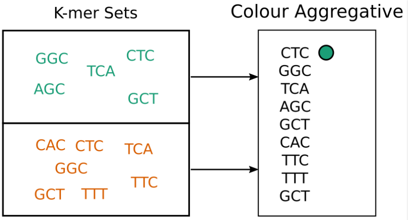
Two possible approaches: colour or k-mer aggregative



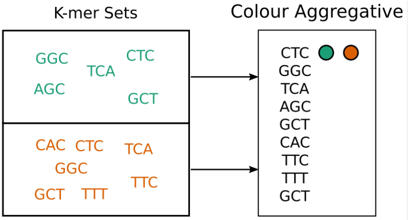
Two possible approaches: colour or k-mer aggregative



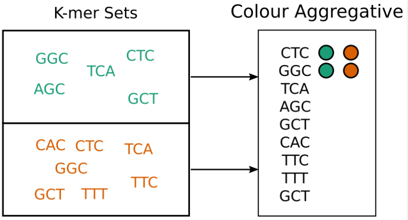
Two possible approaches: colour or k-mer aggregative



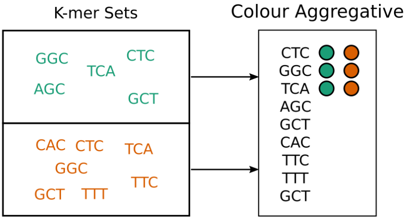
Two possible approaches: colour or k-mer aggregative



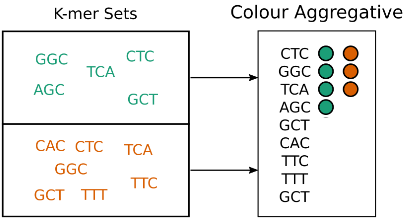
Two possible approaches: colour or k-mer aggregative



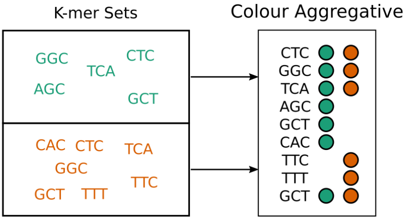
Two possible approaches: colour or k-mer aggregative



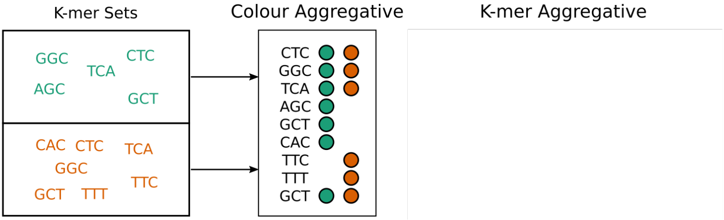
Two possible approaches: colour or k-mer aggregative



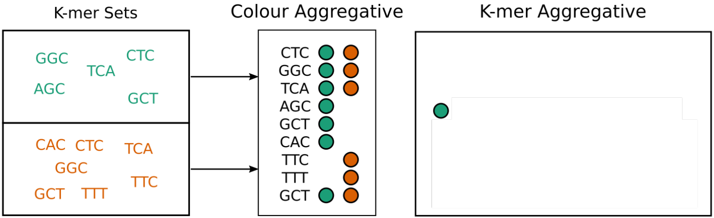
Two possible approaches: colour or k-mer aggregative



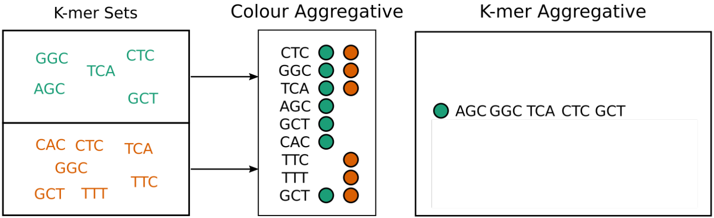
Two possible approaches: colour or k-mer aggregative



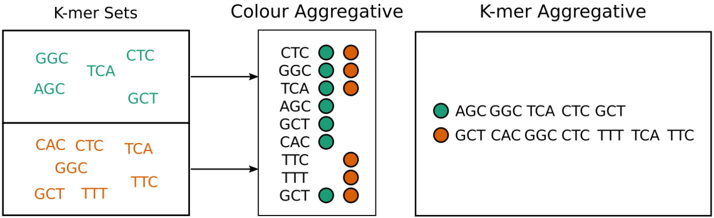
Two possible approaches: colour or k-mer aggregative



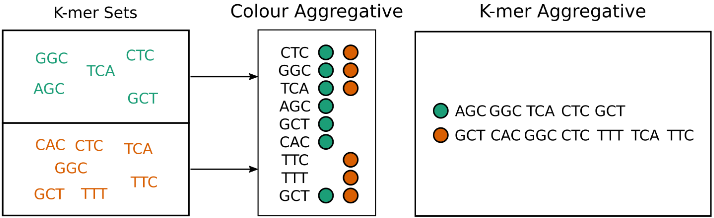
Two possible approaches: colour or k-mer aggregative



Two possible approaches: colour or k-mer aggregative

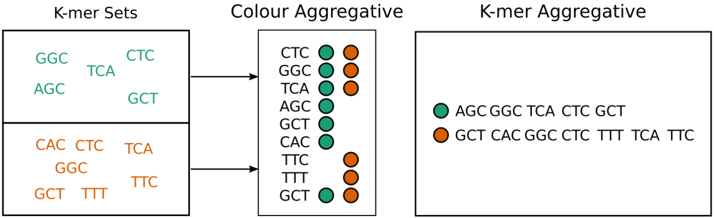


Two possible approaches: colour or k-mer aggregative



- Colour aggregative: k-mer -> sample(s)

Two possible approaches: colour or k-mer aggregative

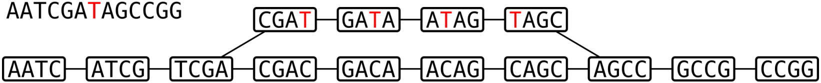


- Colour aggregative: k-mer -> sample(s)
- K-mer aggregative: sample -> k-mer(s)

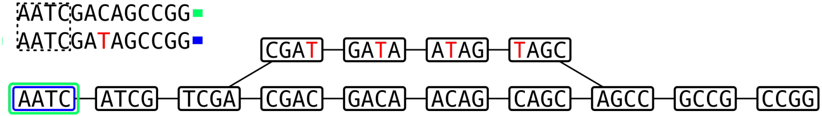
Colour aggregative methods

Coloured de Bruijn graph

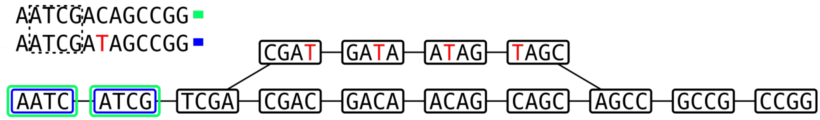
AATCGACAGCCGG
AATCGATAGCCGG



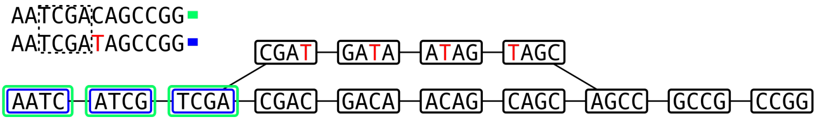
Coloured de Bruijn graph



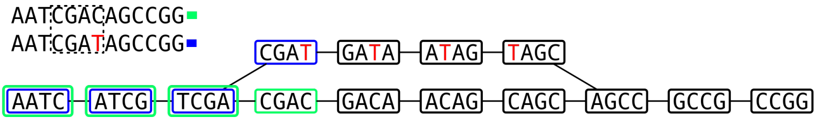
Coloured de Bruijn graph



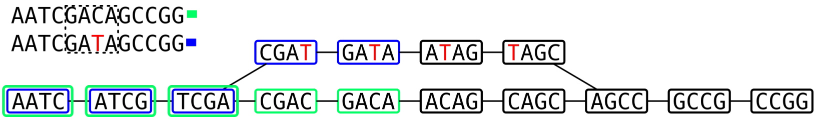
Coloured de Bruijn graph



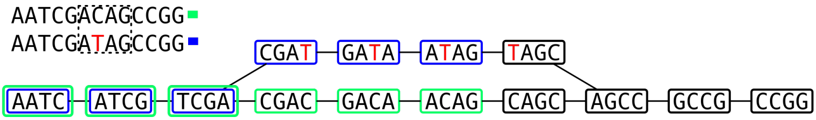
Coloured de Bruijn graph



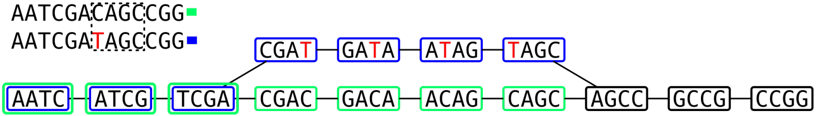
Coloured de Bruijn graph



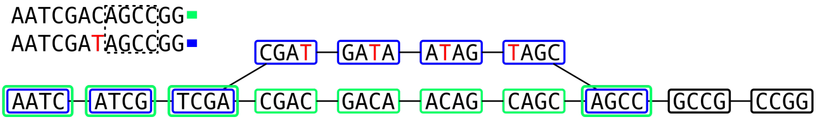
Coloured de Bruijn graph



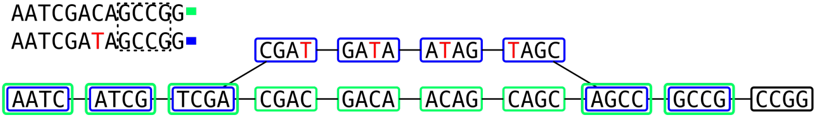
Coloured de Bruijn graph



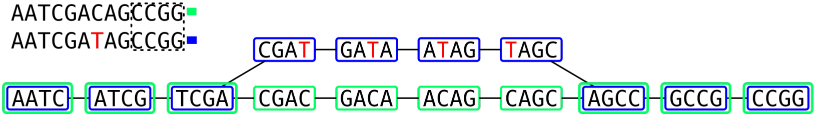
Coloured de Bruijn graph



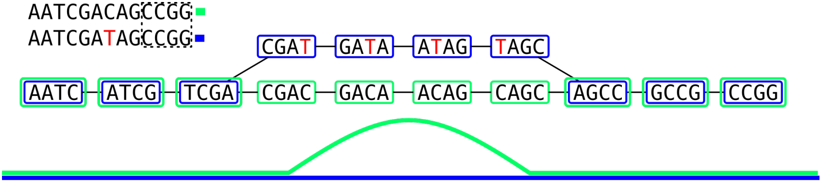
Coloured de Bruijn graph



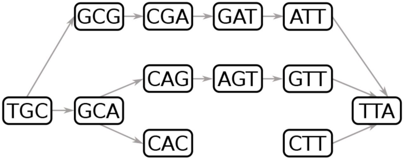
Coloured de Bruijn graph



Coloured de Bruijn graph

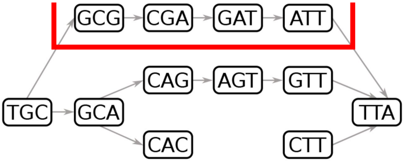


Succinct/Compacted coloured de Bruijn graphs



[Holley and Melsted, 2019]

Succinct/Compacted coloured de Bruijn graphs

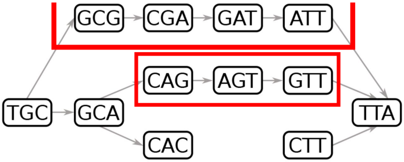


(a)

(b)

[Holley and Melsted, 2019]

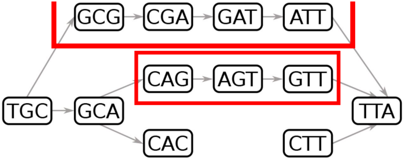
Succinct/Compacted coloured de Bruijn graphs



(a)

(b)

[Holley and Melsted, 2019]

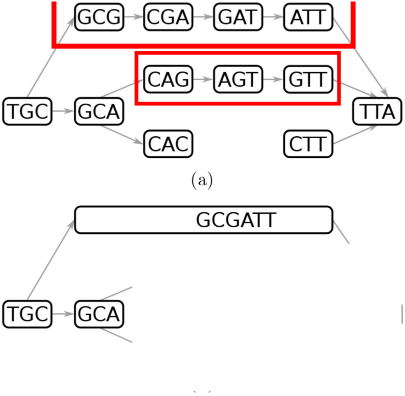


(a)

(b)

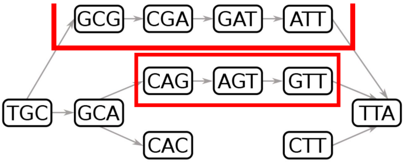
[Holley and Melsted, 2019]

Succinct/Compacted coloured de Bruijn graphs

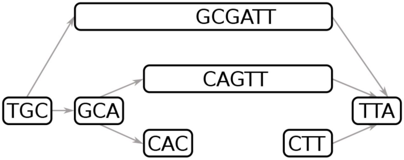


[Holley and Melsted, 2019]

Succinct/Compacted coloured de Bruijn graphs



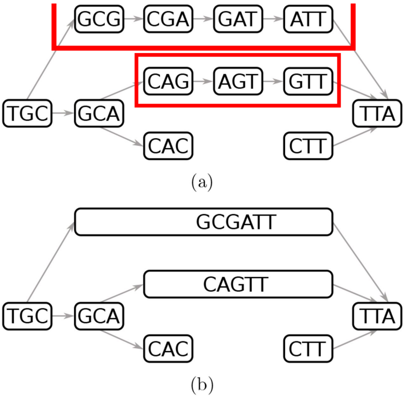
(a)



(b)

[Holley and Melsted, 2019]

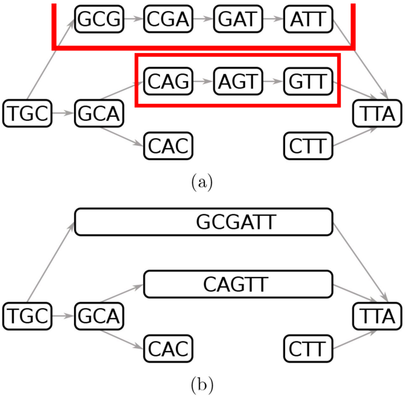
Succinct/Compacted coloured de Bruijn graphs



[Holley and Melsted, 2019]

- Compact maximal non-branching paths into untigs

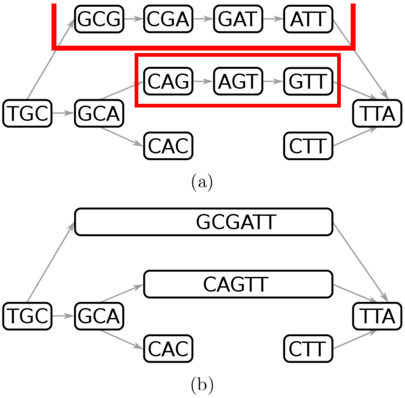
Succinct/Compacted coloured de Bruijn graphs



[Holley and Melsted, 2019]

- Compact maximal non-branching paths into untigs
- Use probabilistic data structures e.g. bloomfilters, minhash sketches, minimisers

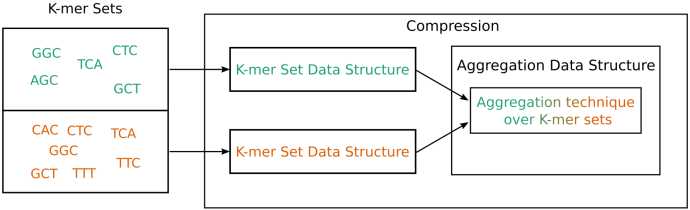
Succinct/Compacted coloured de Bruijn graphs



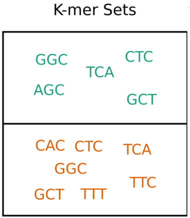
[Holley and Melsted, 2019]

- Compact maximal non-branching paths into untigs
- Use probabilistic data structures e.g. bloomfilters, minhash sketches, minimisers
- AKA make things more approximate but smaller!

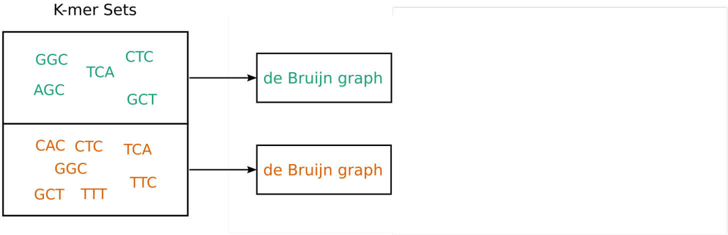
BlastFrost


K-mer Sets
GGC
TCA
CTC
AGC
GCT
CAC
CTC
TCA
GGC
TTC
GCT
TTT
Compression
K-mer Set Data Structure
K-mer Set Data Structure
Aggregation Data Structure
Aggregation technique
over K-mer sets

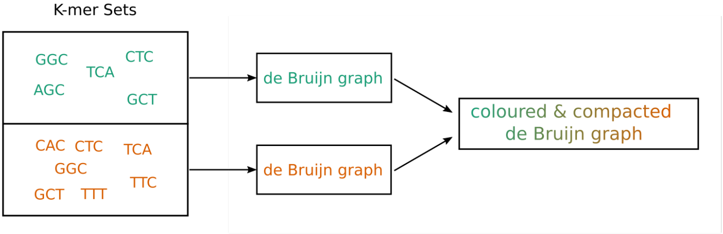
BlastFrost



BlastFrost

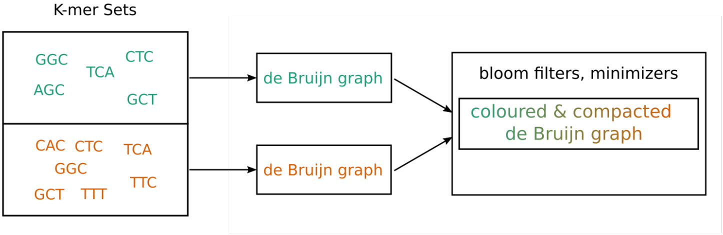

K-mer Sets
GGC
TCA
CTC
AGC
GCT
CAC
CTC
TCA
GGC
TTC
GCT
TTT
de Bruijn graph
de Bruijn graph

BlastFrost

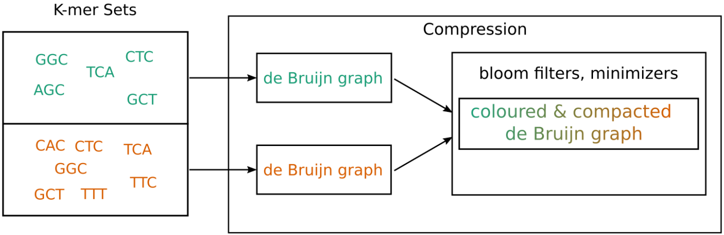


K-mer Sets
GGC
TCA
CTC
AGC
GCT
CAC
CTC
TCA
GGC
TTC
GCT
TTT
de Bruijn graph
de Bruijn graph
coloured & compacted
de Bruijn graph

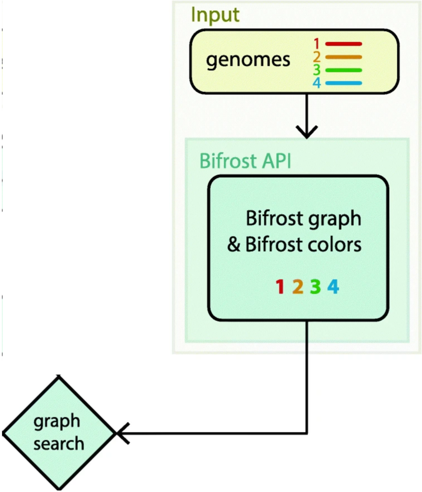
BlastFrost


K-mer Sets
GGC
TCA
CTC
AGC
GCT
CAC
CTC
TCA
GGC
TTC
GCT
TTT
de Bruijn graph
de Bruijn graph
bloom filters, minimizers
coloured & compacted
de Bruijn graph

BlastFrost

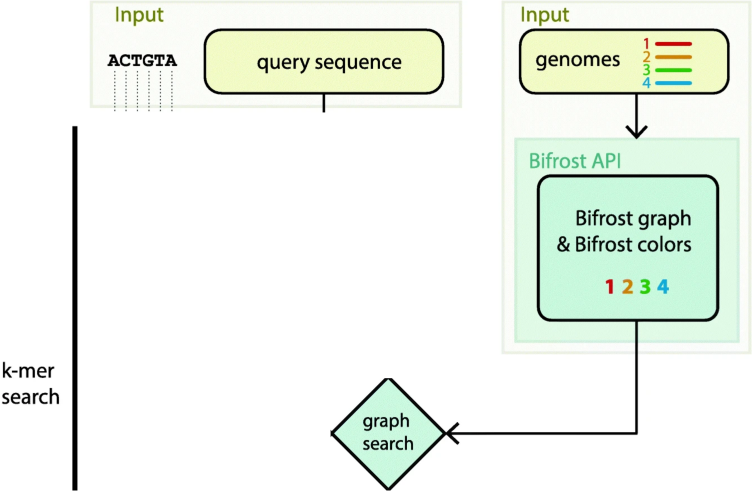

K-mer Sets
GGC
TCA
CTC
AGC
GCT
CAC
CTC
TCA
GGC
TTC
GCT
TTT
Compression
de Bruijn graph
de Bruijn graph
bloom filters, minimizers
coloured & compacted
de Bruijn graph

BlastFrost: Similar but for bigger sequences!



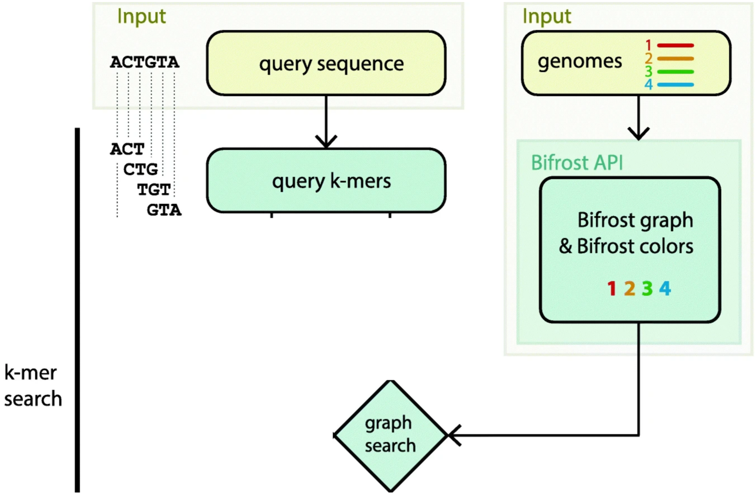
[Luhmann et al., 2020]

BlastFrost: Similar but for bigger sequences!



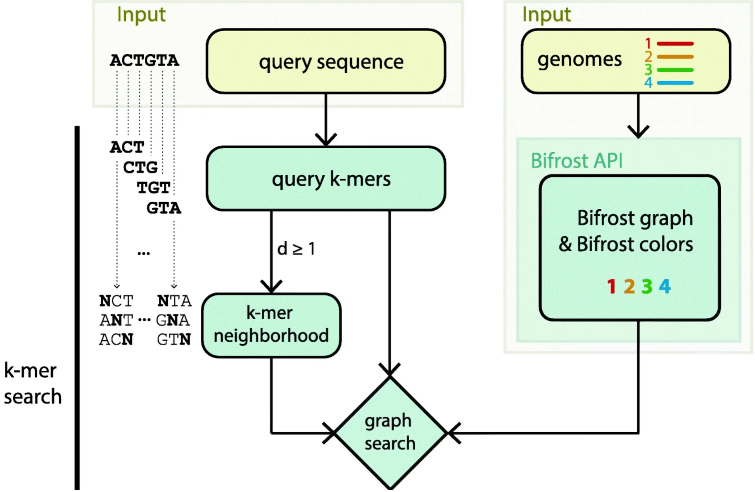
[Luhmann et al., 2020]

BlastFrost: Similar but for bigger sequences!



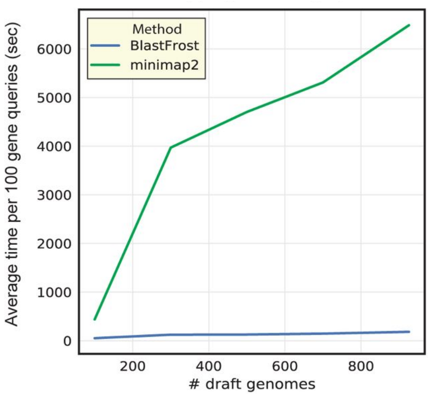
[Luhmann et al., 2020]

BlastFrost: Similar but for bigger sequences!



[Luhmann et al., 2020]

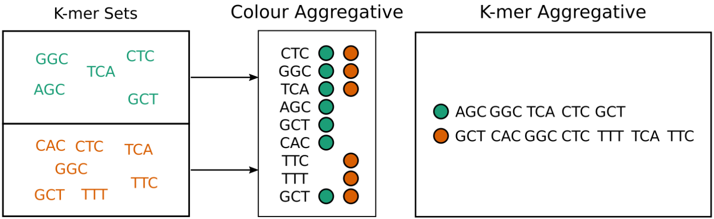
BlastFrost scaling



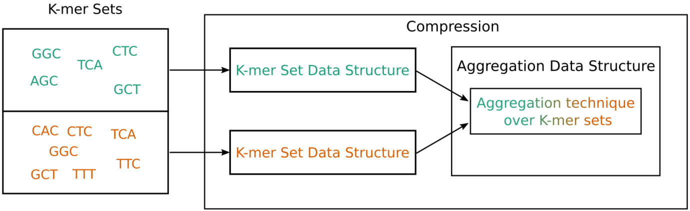
[Luhmann et al., 2020]

K-mer aggregative methods

Index based on sample -> k-mer(s)

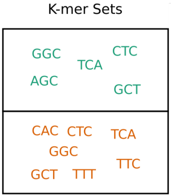


BIGSI



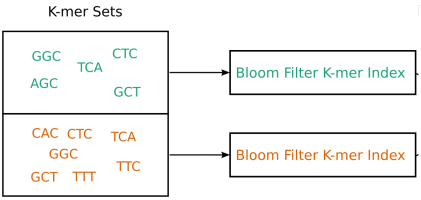
[Bradley et al., 2019]

BIGSI



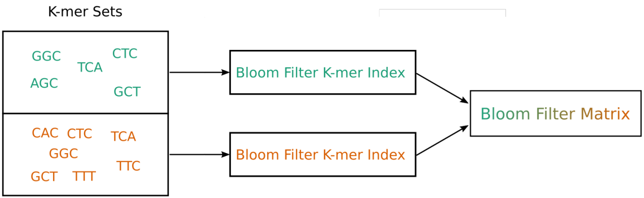
[Bradley et al., 2019]

BIGSI



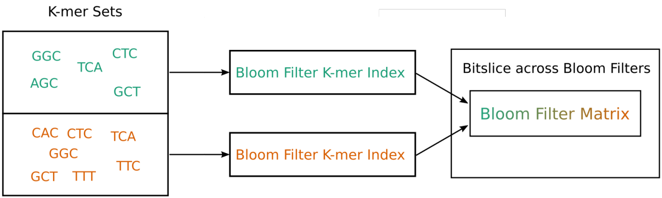
[Bradley et al., 2019]

BIGSI



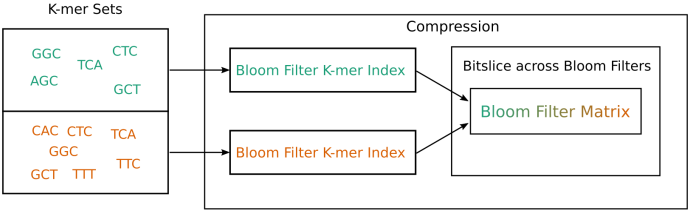
[Bradley et al., 2019]

BIGSI



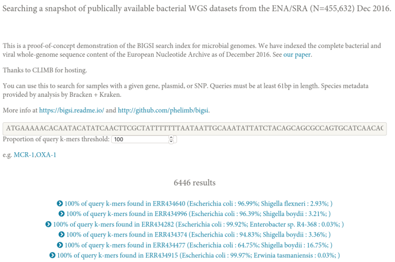
[Bradley et al., 2019]

BIGSI



[Bradley et al., 2019]

BIGSI indexing of ENA



Searching a snapshot of publically available bacterial WGS datasets from the ENA/SRA (N=455,632) Dec 2016.

This is a proof-of-concept demonstration of the BIGSI search index for microbial genomes. We have indexed the complete bacterial and viral whole-genome sequence content of the European Nucleotide Archive as of December 2016. See our paper.

Thanks to CLIMB for hosting.

You can use this to search for samples with a given gene, plasmid, or SNP. Queries must be at least 61bp in length. Species metadata provided by analysis by Bracken + Kraken.

More info at https://bigsi.readme.io/ and http://github.com/phelimb/bigsi.

ATGAAAAACACAATACATATCAACTTCGCTATTTTTTTAATAATTGCAAATATTATCTACAGCAGCGCCAGTGCATCAACAC

Proportion of query k-mers threshold: 100

e.g. MCR-1,OXA-1

6446 results

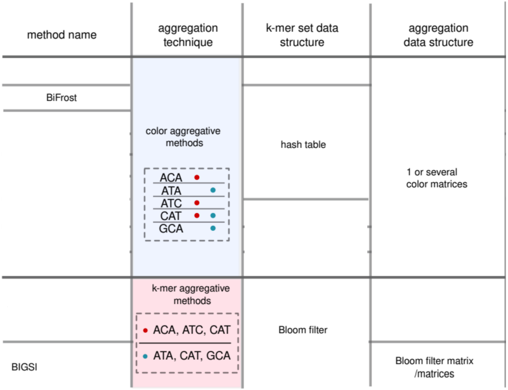
- 100% of query k-mers found in ERR434640 (Escherichia coli : 96.99%; Shigella flexneri : 2.93%;)
- 100% of query k-mers found in ERR434996 (Escherichia coli : 96.39%; Shigella boydii : 3.21%;)
- 100% of query k-mers found in ERR434282 (Escherichia coli : 99.92%; Enterobacter sp. R4-368 : 0.03%;)
- 100% of query k-mers found in ERR434374 (Escherichia coli : 94.83%; Shigella boydii : 3.36%;)
- 100% of query k-mers found in ERR434477 (Escherichia coli : 64.75%; Shigella boydii : 16.75%;)
- 100% of query k-mers found in ERR434915 (Escherichia coli : 99.97%; Erwinia tasmaniensis : 0.03%;)

[Bradley et al., 2019]

- Indexing all bacterial, viral and parasitic reads in ENA (500,000 sets, 170TB of data)
- 1.5TB index that be queried near instantaneously

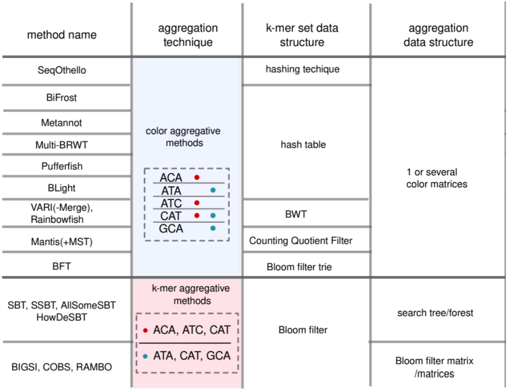
Which method?

Many Options



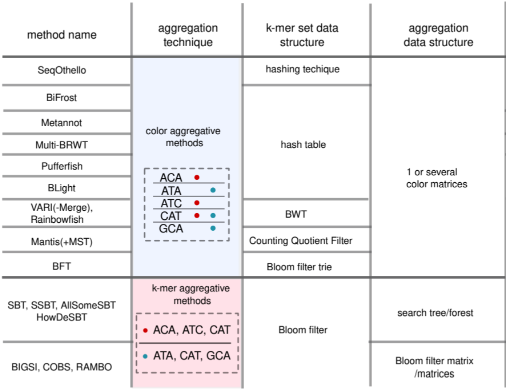
[Marchet et al., 2021]

Many Options



[Marchet et al., 2021]

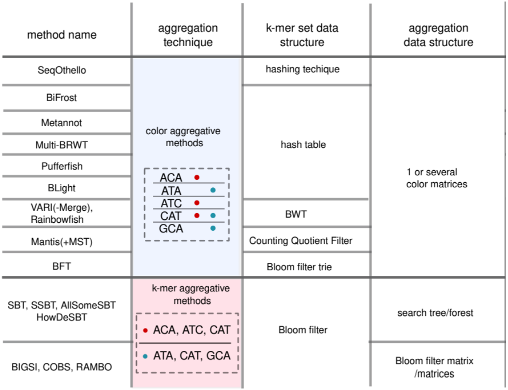
Many Options



[Marchet et al., 2021]

- It depends: complexity, sequence length, query length

Many Options



[Marchet et al., 2021]

- It depends: complexity, sequence length, query length
- What features you need e.g., inserting new sets, space vs time trade-offs

Summary

Learning Objectives

- Vast amount of sequence data and it is growing rapidly

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths
- Fixed-size of k-mers makes them more tractable

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths
- Fixed-size of k-mers makes them more tractable
- dBG encoded as vector gracefully handles k-mer redundancy

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths
- Fixed-size of k-mers makes them more tractable
- dBG encoded as vector gracefully handles k-mer redundancy
- coloured dBG (cdBG) represent multiple samples

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths
- Fixed-size of k-mers makes them more tractable
- dBG encoded as vector gracefully handles k-mer redundancy
- coloured dBG (cdBG) represent multiple samples
- When indexing across samples:

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths
- Fixed-size of k-mers makes them more tractable
- dBG encoded as vector gracefully handles k-mer redundancy
- coloured dBG (cdBG) represent multiple samples
- When indexing across samples:
  - Map from union of all k-mers to samples they contain (colour aggregative) e.g., BiFrost

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths
- Fixed-size of k-mers makes them more tractable
- dBG encoded as vector gracefully handles k-mer redundancy
- coloured dBG (cdBG) represent multiple samples
- When indexing across samples:
    - Map from union of all k-mers to samples they contain (colour aggregative) e.g., BiFrost
    - BiFrost uses compacted cdBG to index across samples

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths
- Fixed-size of k-mers makes them more tractable
- dBG encoded as vector gracefully handles k-mer redundancy
- coloured dBG (cdBG) represent multiple samples
- When indexing across samples:
    - Map from union of all k-mers to samples they contain (colour aggregative) e.g., BiFrost
    - BiFrost uses compacted cdBG to index across samples
    - Map from each sample to the k-mers it contains (k-mer aggregative) e.g., BIGSI

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths
- Fixed-size of k-mers makes them more tractable
- dBG encoded as vector gracefully handles k-mer redundancy
- coloured dBG (cdBG) represent multiple samples
- When indexing across samples:
    - Map from union of all k-mers to samples they contain (colour aggregative) e.g., BiFrost
    - BiFrost uses compacted cdBG to index across samples
    - Map from each sample to the k-mers it contains (k-mer aggregative) e.g., BIGSI
    - BIGSI creates a big matrix of bloom filters where each column is a sample

Learning Objectives

- Vast amount of sequence data and it is growing rapidly
- Assembly just a path through graph NOT all possible paths
- Fixed-size of k-mers makes them more tractable
- dBG encoded as vector gracefully handles k-mer redundancy
- coloured dBG (cdBG) represent multiple samples
- When indexing across samples:
    - Map from union of all k-mers to samples they contain (colour aggregative) e.g., BiFrost
    - BiFrost uses compacted cdBG to index across samples
    - Map from each sample to the k-mers it contains (k-mer aggregative) e.g., BIGSI
    - BIGSI creates a big matrix of bloom filters where each column is a sample
- Active field and choosing best method is very data and task specific

Questions?

Bradley, P., Den Bakker, H. C., Rocha, E. P., McVean, G., and Iqbal, Z. (2019).
Ultrafast search of all deposited bacterial and viral genomic data.
Nature biotechnology, 37(2):152–159.

Holley, G. and Melsted, P. (2019).
Bifrost–highly parallel construction and indexing of colored and compacted de bruijn graphs.
BioRxiv, page 695338.

Luhmann, N., Holley, G., and Achtman, M. (2020).
Blastfrost: Fast querying of 100,000 s of bacterial genomes in bifrost graphs.
BioRxiv.

Marchet, C., Boucher, C., Puglisi, S. J., Medvedev, P., Salson, M., and Chikhi, R. (2021).
Data structures based on k-mers for querying large collections of sequencing data sets.
Genome Research, 31(1):1–12.