Mapping-friendly Sequence Reductions to process compressed genomic data

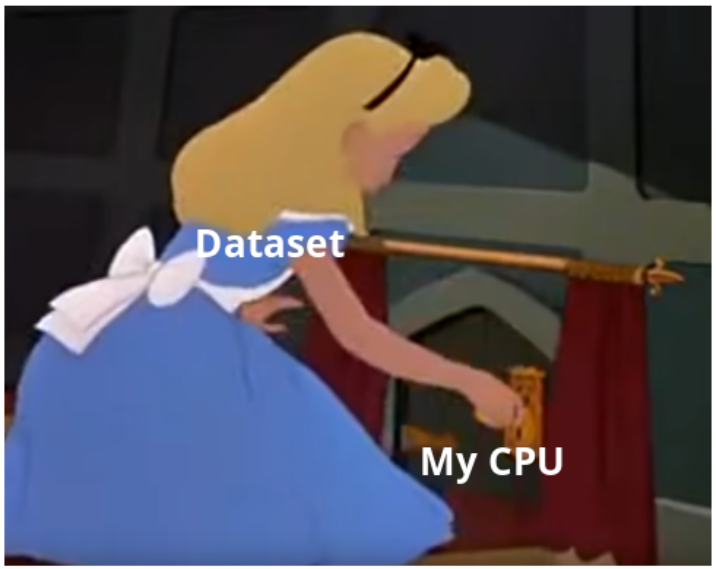
Roland Faure[1,2], Baptiste Hilaire[2], Dominique Lavenier[2]

[1]Université libre de Bruxelles (ULB) - Belgium

[2]Université de Rennes, IRISA - France

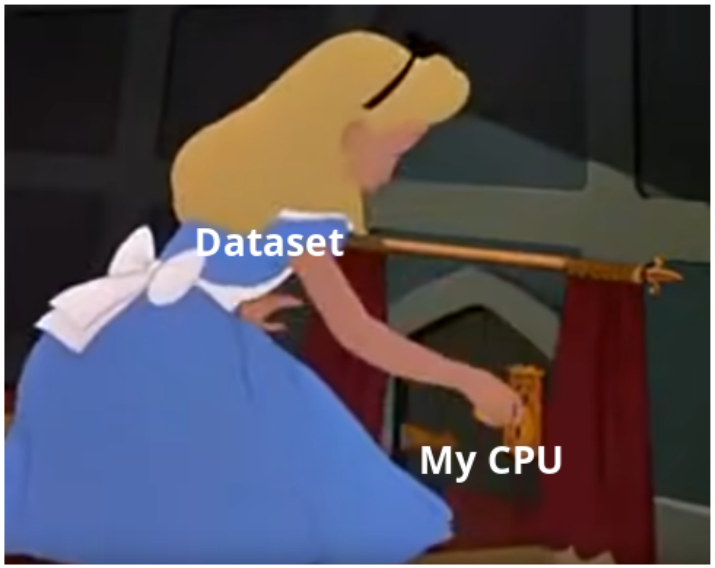
SeqBIM Lille 2023

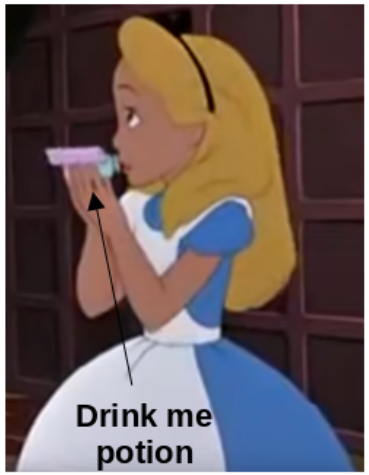
Big data

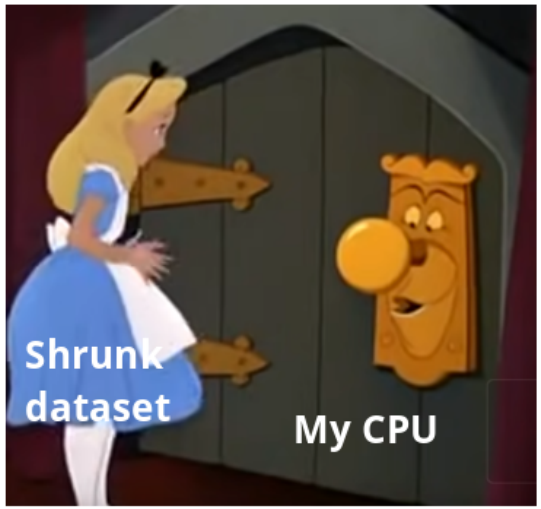


Credits: Alice in Wonderland, Lewis, Disney

Big data

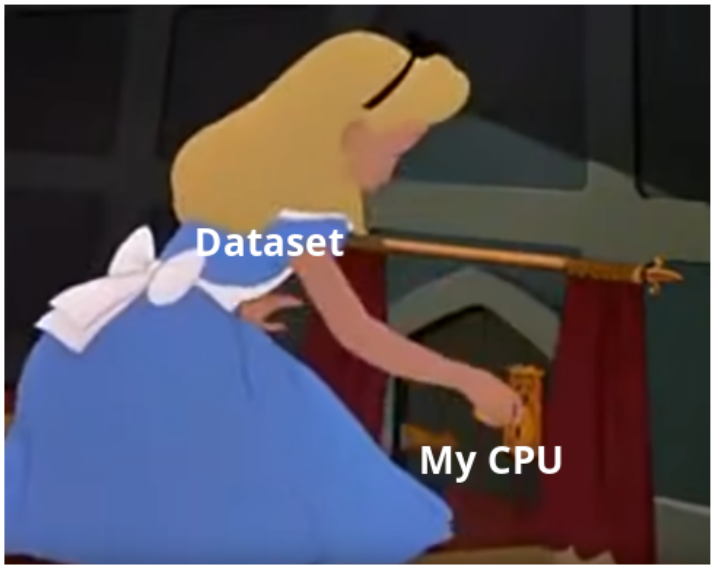


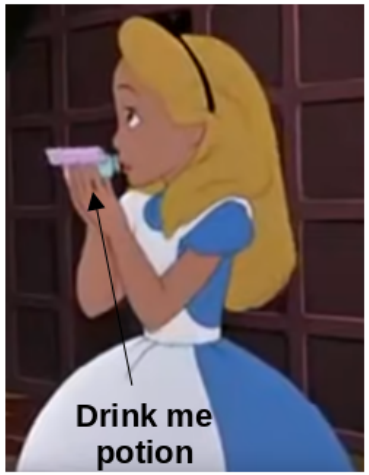


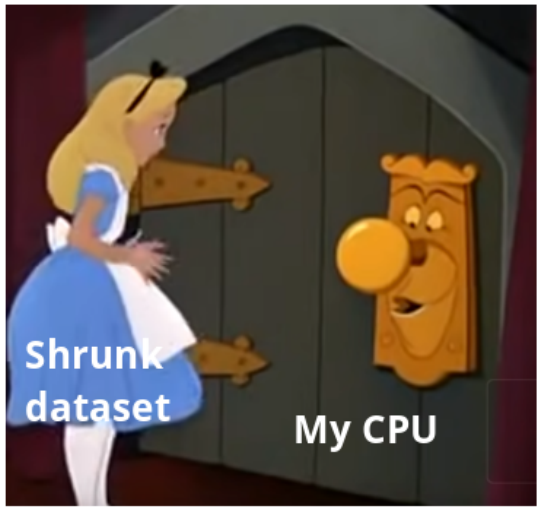


Credits: Alice in Wonderland, Lewis, Disney

Big data







Credits: Alice in Wonderland, Lewis, Disney

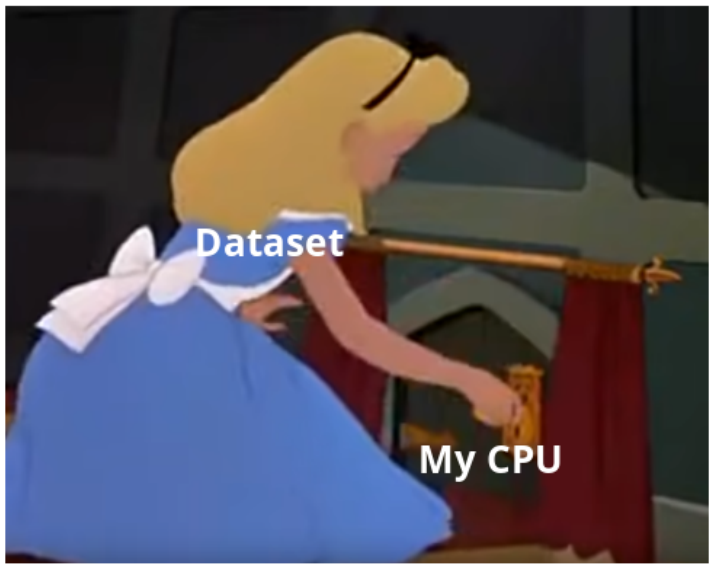
Shrinking sequence information to process shrunk datasets

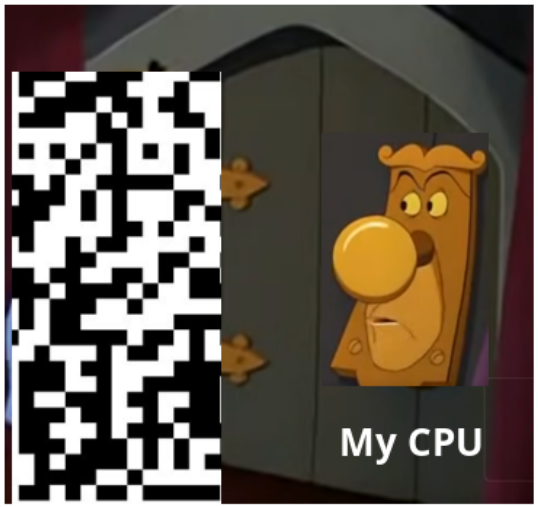
Shrinking sequences

- gzip, xz...

Shrinking sequences

- gzip, xz...

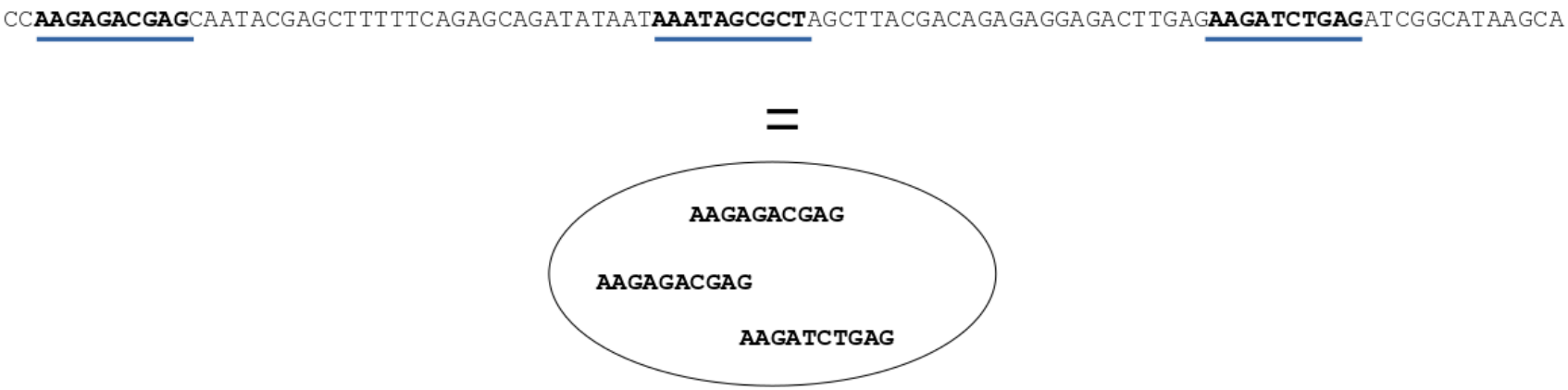




Credits: Alice in Wonderland, Lewis, Disney

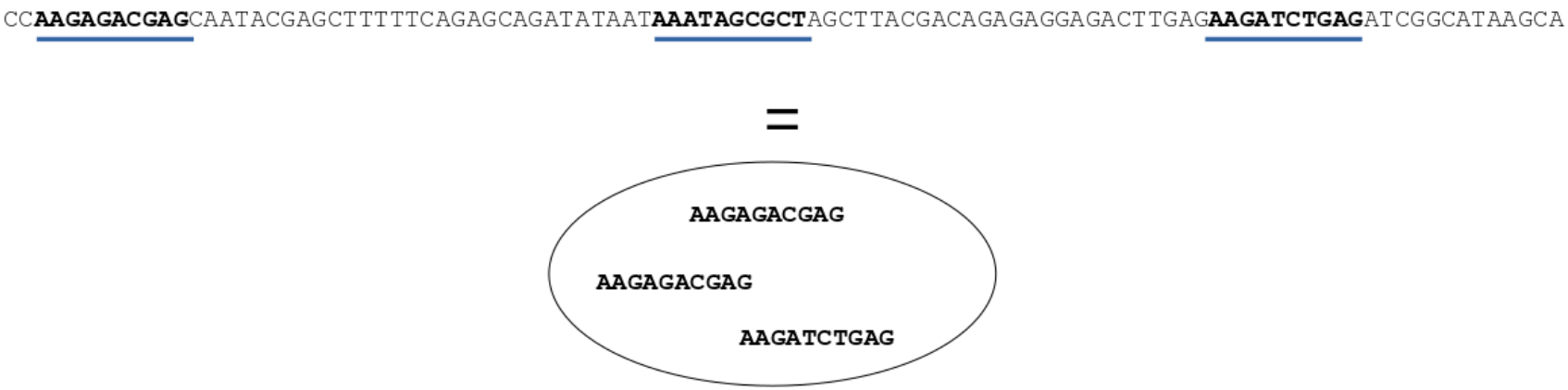
Shrinking sequence information to process shrunk datasets

State-of-the-art: shrinking sequences with k-mers



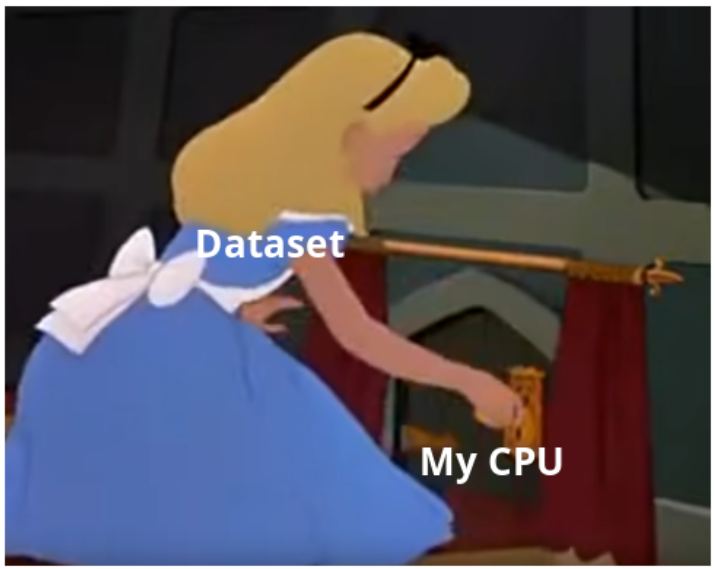
- subsets of k-mers are compressed representation of sequences

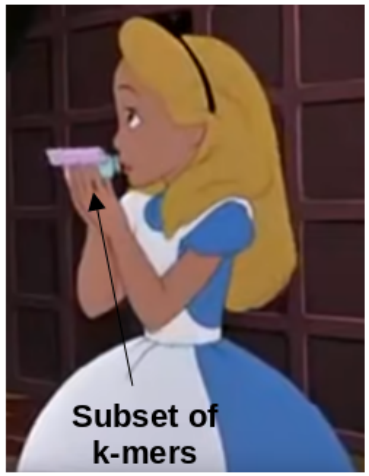
State-of-the-art: shrinking sequences with k-mers



- subsets of k-mers are compressed representation of sequences
- minimap2 (alignment), mDBG (assembly), mash (sequence similarity), PebbleScout (sequence querying)...

Going beyond shrinking sequences with k-mers

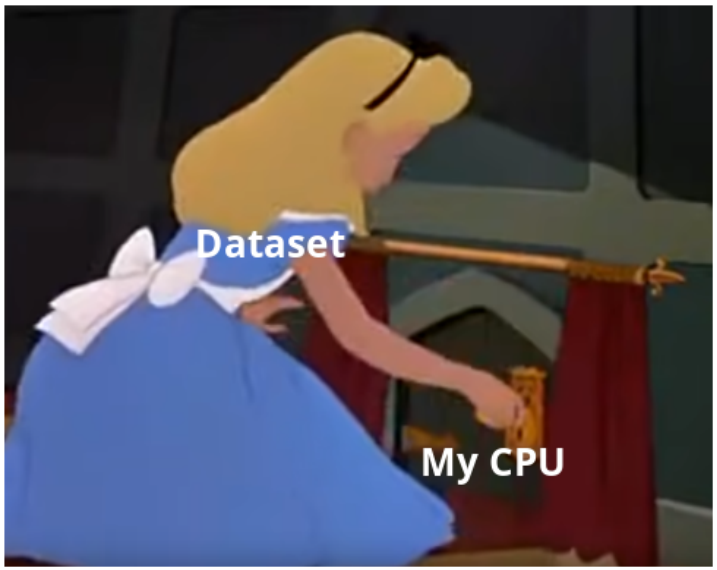




Credits: Alice in Wonderland, Lewis, Disney

- No structural information
- Big chunks of data missing
- Awkward tasks: SVs, gene detection, assembly...

MSR compression to shrink sequences



Credits: Alice in Wonderland, Lewis, Disney

- Different properties compared to subset of k-mers
- Different problems

Mapping-friendly Sequence Reduction (MSR)

- Introduced in iScience 2022 [1] to improve mapping accuracy

[1] Blassel, Luc Medvedev, Paul Chikhi, Rayan. (2022). Mapping-friendly sequence reductions: Going beyond homopolymer compression. iScience.

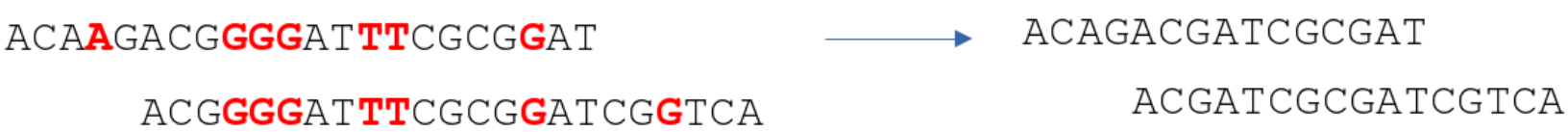
Mapping-friendly Sequence Reduction (MSR)

- Introduced in iScience 2022 [1] to improve mapping accuracy

[1] Blassel, Luc Medvedev, Paul Chikhi, Rayan. (2022). Mapping-friendly sequence reductions: Going beyond homopolymer compression. iScience.

Mapping-friendly Sequence Reduction (MSR)

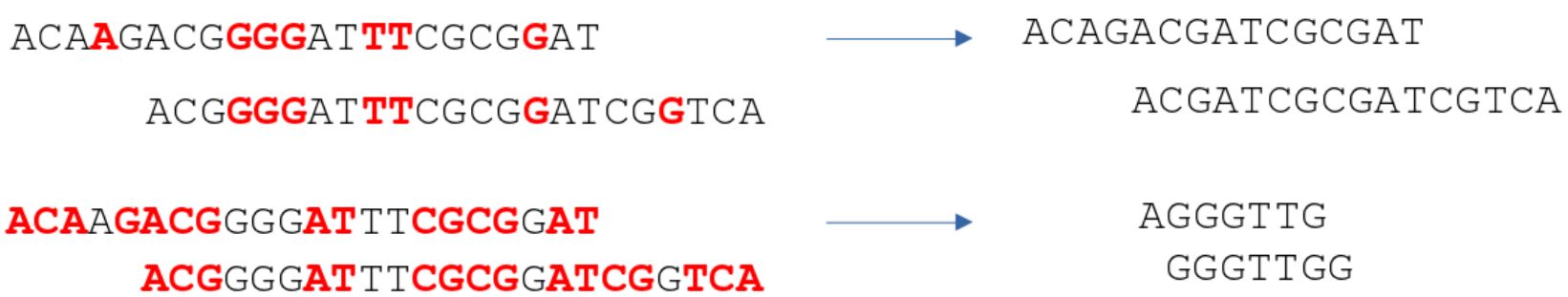
- Introduced in iScience 2022 [1] to improve mapping accuracy



[1] Blassel, Luc Medvedev, Paul Chikhi, Rayan. (2022). Mapping-friendly sequence reductions: Going beyond homopolymer compression. iScience.

Mapping-friendly Sequence Reduction (MSR)

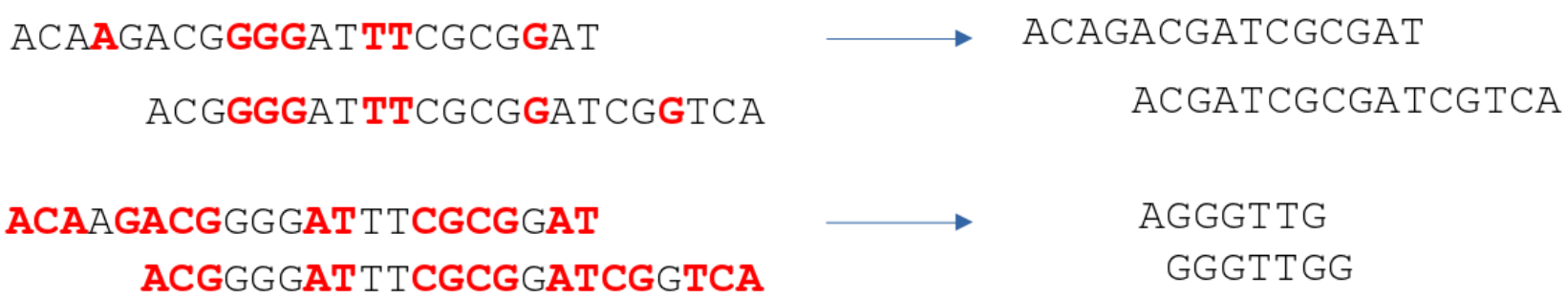
- Introduced in iScience 2022 [1] to improve mapping accuracy



[1] Blassel, Luc Medvedev, Paul Chikhi, Rayan. (2022). Mapping-friendly sequence reductions: Going beyond homopolymer compression. iScience.

Mapping-friendly Sequence Reduction (MSR)

- Introduced in iScience 2022 [1] to improve mapping accuracy



- Mapping-friendly
- Computed in a streaming fashion
- Reverse-complement insensitive

[1] Blassel, Luc Medvedev, Paul Chikhi, Rayan. (2022). Mapping-friendly sequence reductions: Going beyond homopolymer compression. iScience.

Computing a MSR

$$f : \{A, C, G, T\}^3 \longrightarrow \{A, C, G, T, \_\}$$

AGT $\rightarrow G$	CAG $\rightarrow A$
ACT $\rightarrow C$	CCG $\rightarrow C$
ATT $\rightarrow T$	CGG $\rightarrow G$
AAT $\rightarrow A$	CTG $\rightarrow T$

else $\rightarrow$ _

sequence **CAG**TATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence **A**

Computing a MSR

$$f : \{A, C, G, T\}^3 \longrightarrow \{A, C, G, T, \_\}$$

AGT $\rightarrow G$	CAG $\rightarrow A$
ACT $\rightarrow C$	CCG $\rightarrow C$
ATT $\rightarrow T$	CGG $\rightarrow G$
AAT $\rightarrow A$	CTG $\rightarrow T$

else $\rightarrow$ _

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence **AG**

Computing a MSR

$$f : \{A, C, G, T\}^3 \longrightarrow \{A, C, G, T, \_\}$$

AGT $\rightarrow G$	CAG $\rightarrow A$
ACT $\rightarrow C$	CCG $\rightarrow C$
ATT $\rightarrow T$	CGG $\rightarrow G$
AAT $\rightarrow A$	CTG $\rightarrow T$

else $\rightarrow$ _

sequence CA**GTA**TGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence **AG_**

Computing a MSR

$$f : \{A, C, G, T\}^3 \longrightarrow \{A, C, G, T, \_\}$$

AGT $\rightarrow G$	CAG $\rightarrow A$
ACT $\rightarrow C$	CCG $\rightarrow C$
ATT $\rightarrow T$	CGG $\rightarrow G$
AAT $\rightarrow A$	CTG $\rightarrow T$

else $\rightarrow$ _

sequence CAG**TAT**GGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence **AG___**

Computing a MSR

$$f : \{A, C, G, T\}^3 \longrightarrow \{A, C, G, T, \_\}$$

AGT $\rightarrow G$	CAG $\rightarrow A$
ACT $\rightarrow C$	CCG $\rightarrow C$
ATT $\rightarrow T$	CGG $\rightarrow G$
AAT $\rightarrow A$	CTG $\rightarrow T$

else $\rightarrow$ _

sequence CAGT**ATG**GATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence **AG____**

Computing a MSR

$$f : \{A, C, G, T\}^3 \longrightarrow \{A, C, G, T, \_\}$$

AGT $\rightarrow G$	CAG $\rightarrow A$
ACT $\rightarrow C$	CCG $\rightarrow C$
ATT $\rightarrow T$	CGG $\rightarrow G$
AAT $\rightarrow A$	CTG $\rightarrow T$

else $\rightarrow$ _

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCA**GAG**

Reduced sequence **AG**_____**A**_____**G**_____**C**_____**A**___

Computing a MSR

$$f : \{A, C, G, T\}^3 \longrightarrow \{A, C, G, T, \_\}$$

$$\begin{array}{ll} \mathrm{AGT} \rightarrow G & \mathrm{CAG} \rightarrow A \\ \mathrm{ACT} \rightarrow C & \mathrm{CCG} \rightarrow C \\ \mathrm{ATT} \rightarrow T & \mathrm{CGG} \rightarrow G \\ \mathrm{AAT} \rightarrow A & \mathrm{CTG} \rightarrow T \end{array}$$

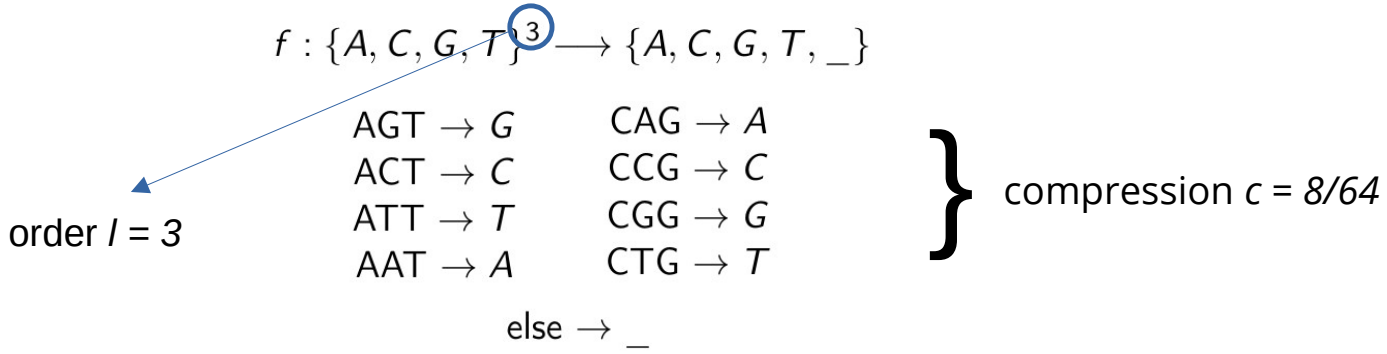
$$\mathrm{else} \rightarrow \_$$

} compression c = *8/64*

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCA**GAG**

Reduced sequence **AG**_____**A**_____**G**_____**C**_____**A**___

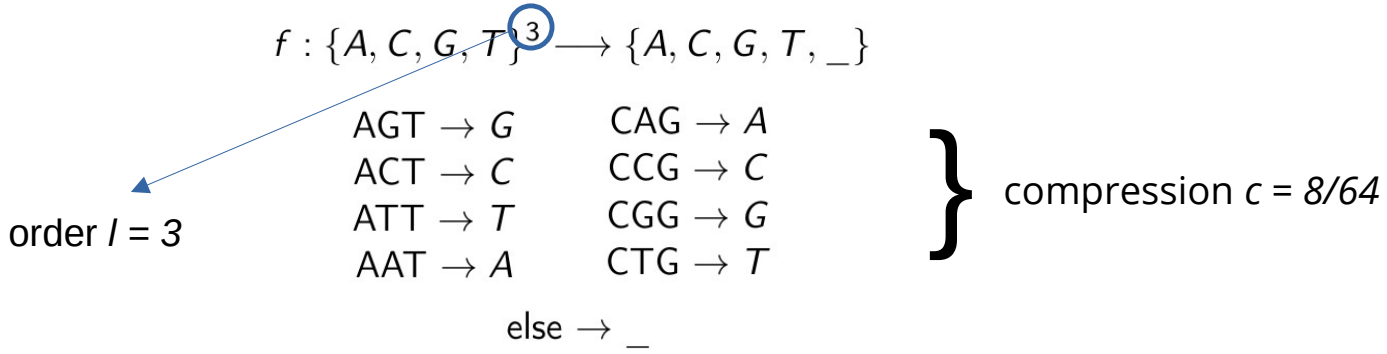
Computing a MSR



sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCA**GAG**

Reduced sequence **AG**_____**A**_____**G**_____**C**_____**A**___

Computing a MSR



sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCA**GAG**

Reduced sequence **AG**_____**A**_____**G**_____**C**_____**A**___

- Let's use this to shrink the sequences of our datasets

Demo: an experimental HiFi assembler

Demo: an experimental HiFi assembler

- Step 1: Compress the reads c=0.05 l=201

Demo: an experimental HiFi assembler

- Step 1: Compress the reads c=0.05 l=201
- Step 2: Assemble the compressed reads BCALM2[2] k=41

[2]Rayan Chikhi, Antoine Limasset and Paul Medvedev, Compacting de Bruijn graphs from sequencing data quickly and in low memory, Proceedings of ISMB 2016, Bioinformatics, 32 (12): i201-i208.

Demo: an experimental HiFi assembler

- Step 1: Compress the reads c=0.05 l=201
- Step 2: Assemble the compressed reads BCALM2[2] k=41
- Step 3: Inflate the contigs

Credits: Alice in Wonderland, Lewis, Disney

[2]Rayan Chikhi, Antoine Limasset and Paul Medvedev, Compacting de Bruijn graphs from sequencing data quickly and in low memory, Proceedings of ISMB 2016, Bioinformatics, 32 (12): i201-i208.

Going back to uncompressed space

- MSR is lossy $\rightarrow$ not strictly reversible

Going back to uncompressed space

- MSR is lossy $\rightarrow$ not strictly reversible
- Keep a record while compressing

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence **AG**_____**A**_____**G**_____**C**_____**A**__

Record {AGAG→CAGTATGGATACAGATGGAGATATCATCGAGT,
GAGC→AGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACT,
AGCA→CAGATGGAGATATCATCGAGTAGGGGCACTGTACCAG}

Why use MSR ?

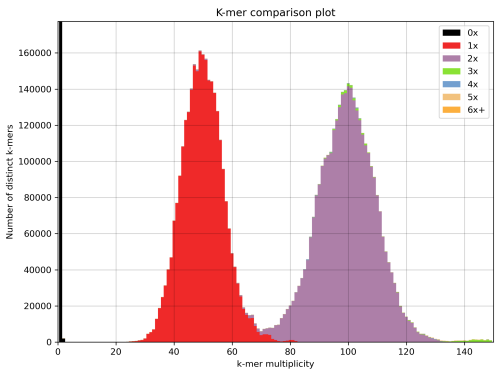
- Dataset: mix of *E. coli* strains 12009 and HS, simulated HiFi
- hifiasm: perfect assembly, 22351s
- mDBG: not-so-good assembly, 30s

Why use MSR ?

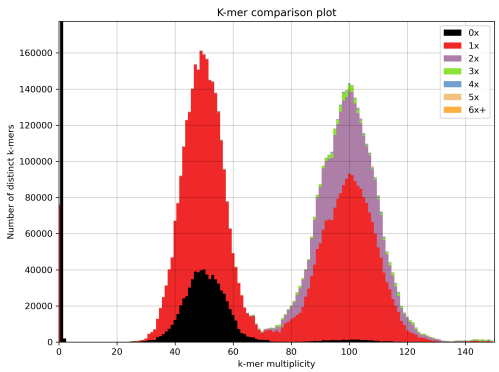
- Dataset: mix of *E. coli* strains 12009 and HS, simulated HiFi
- hifiasm: perfect assembly, 22351s
- mDBG: not-so-good assembly, 30s
- MSR assembler: 90s

Why use MSR ?

- Dataset: mix of *E. coli* strains 12009 and HS, simulated HiFi
- hifiasm: perfect assembly, 22351s
- mDBG: not-so-good assembly, 30s
- MSR assembler: 90s



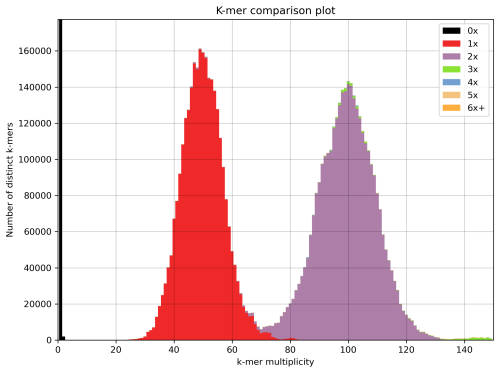
KAT of hifiasm



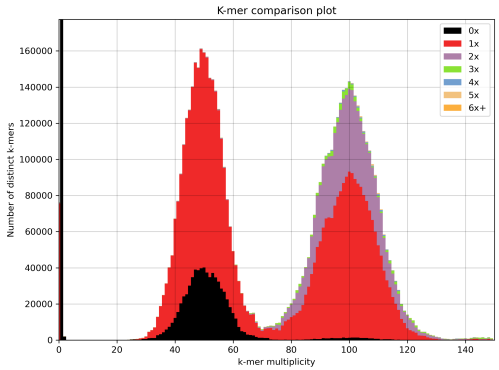
KAT of mDBG

Why use MSR ?

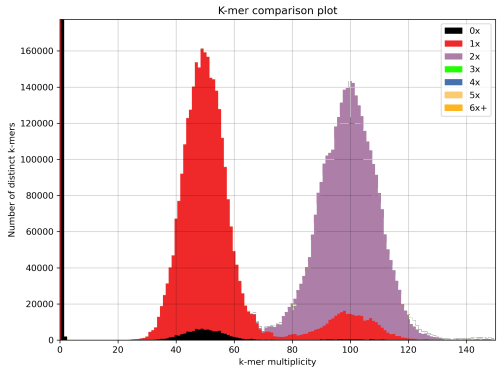
- Dataset: mix of *E. coli* strains 12009 and HS, simulated HiFi
- hifiasm: perfect assembly, 22351s
- mDBG: not-so-good assembly, 30s
- MSR assembler: 90s



KAT of hifiasm



KAT of mDBG



KAT of MSR

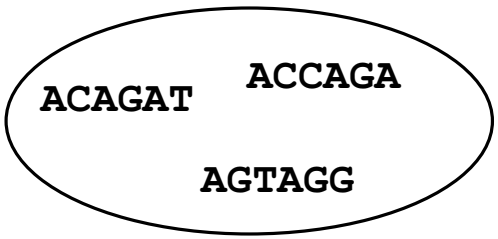
Difference between minimizers and MSRs

CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

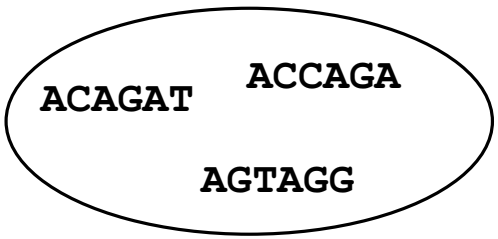
CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

Difference between minimizers and MSRs

CAGTATGGAT**ACAGAT**GGAGATATCATCG**AGTAGG**GGCACTGT**ACCAGA**G



CAGTATGGAT**ACAGAT**GGAGATAT**G**ATCG**AGTAGG**GGCACTGT**ACCAGA**G



Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence

sequence CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence

Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence **CAGTATGGATACAGA**TGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _

sequence **CAGTATGGATACAGA**TGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _

Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A

sequence CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A

Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence CA**GTATGGATACAGATG**GAGATATCATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A_

sequence CA**GTATGGATACAGATG**GAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A_

Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence CAGTATGGA**TACAGATGGAGATAT**CATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A_____G_T_

sequence CAGTATGGA**TACAGATGGAGATATG**ATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A______G_T_

Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence CAGTATGGAT**ACAGATGGAGATATC**ATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A_____G_T__

sequence CAGTATGGAT**ACAGATGGAGATATG**ATCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A_____G_T__

Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence CAGTATGGATAC**AGATGGAGATATCA**TCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A_____G_T___G

sequence CAGTATGGATAC**AGATGGAGATATGA**TCGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A_____G_T_____

Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence CAGTATGGATAC**AGATGGAGATATCAT**CGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A_____G_T____G_

sequence CAGTATGGATACA**GATGGAGATATGAT**CGAGTAGGGGCACTGTACCAGAG

Reduced sequence _A_____G_T_____T

Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGG**GGCACTGTACCAGAG**

Reduced sequence _A_____G_T___G_____A_C_A_T__G___AC__T

sequence CAGTATGGATACAGATGGAGATAT**G**ATCGAGTAGG**GGCACTGTACCAGAG**

Reduced sequence _A_____G_T____TA_____G__T__G___AC__T

Difference between minimizers and MSRs

- Example $c = 0.2$, $l = 15$

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGG**GGCACTGTACCAGAG**

Reduced sequence _A_____G_T___G_____A_C_A_T__G___AC__T

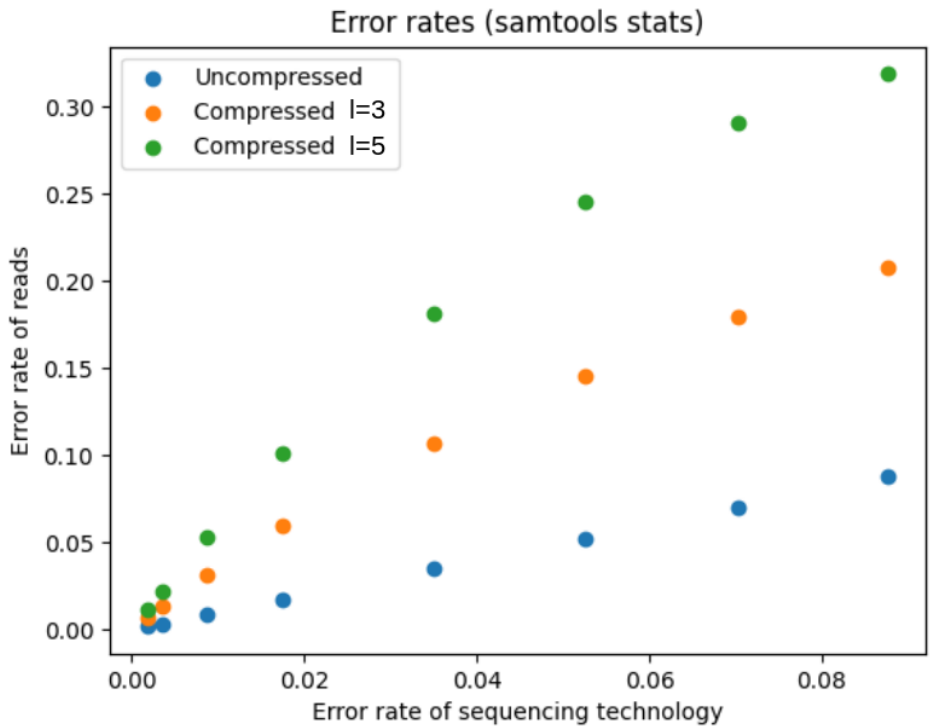
sequence CAGTATGGATACAGATGGAGATAT**G**ATCGAGTAGG**GGCACTGTACCAGAG**

Reduced sequence _A_____G_T____TA_____G__T__G___AC__T

- With high l, all bases are used to generate the compression

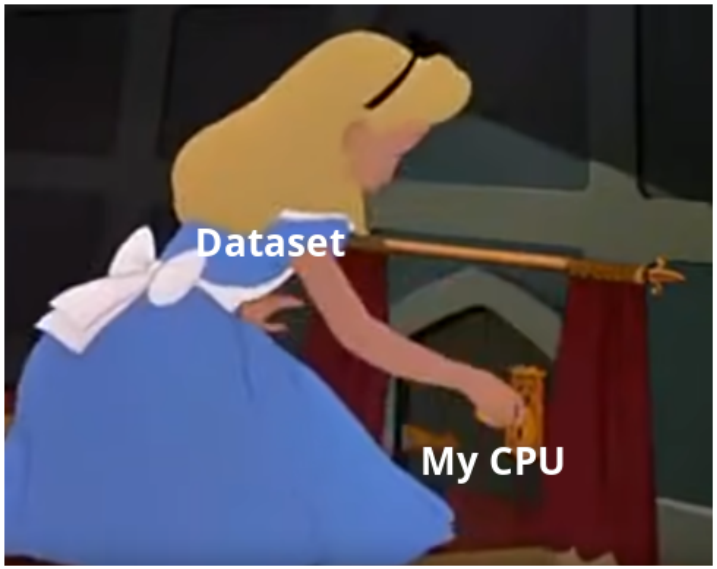
Error rates

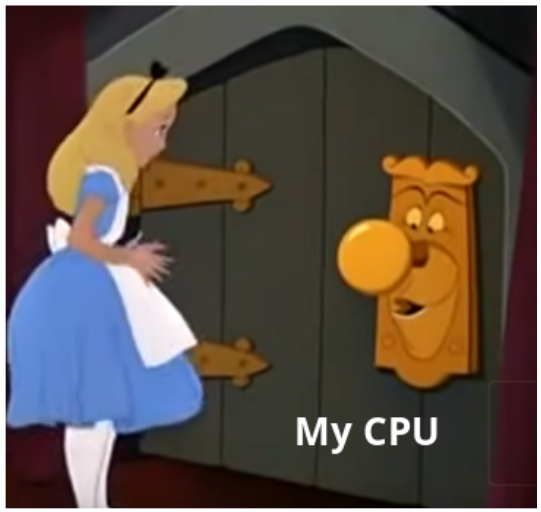
- Sequencing errors get amplified too



- Rule of thumb for low l: $new\_error = l * old\_error$

Summary: MSR as a compression technique





Credits: Alice in Wonderland, Lewis, Disney

- Retains contiguity
- No big chunks of data missing
- MSRs preserve and amplify differences
- Compatible with existing software

Perspectives

- SNP/SV calling
- pangenome graph building
- data indexing
- taxonomic assignment

Perspectives

- SNP/SV calling
- pangenome graph building
- data indexing
- taxonomic assignment

- Inflating the data: open question
- Using different MSRs
- Apply to more noisy data
- ...

Acknowledgments

Parameters

- How I chose my parameters (c=0.05 l=201, Bcalm k=31)

$$\frac{31}{0.05} + 201 = 821$$