# Query-time false positives k-mer filtration and MultiFilters

Lucas Robidou and Pierre Peterlongo

Inria, Rennes

February 11, 2021

(naío-

1/36

æ

・ロト ・ 日 ・ ・ ヨ ・ ・ ヨ ・ ・

Lucas Robidou and Pierre Peterlongo Query-time filtration and MultiFilters



2 How to identify false positive results ?

#### 3 Multifilters



Э.

シペペ 2/36

イロト イポト イヨト イヨト

About Bloom filters Some characteristics

#### **Bloom Filters**

*Ínría* ≣ २९९ 3/36

Lucas Robidou and Pierre Peterlongo Query-time filtration and MultiFilters

About Bloom filters Some characteristics

## About Bloom filters

A Bloom filter is a data structure designed to tell rapidly and memory-efficiently whether an element W is present in a set S.

- If W is in S, the the query return "true".
- If W is not in S, the query may still return "true" with a probability of ε.



Figure: A Bloom filter (each element have 3 hash functions)

イロト イヨト イヨト イヨト

source: https://www.abhishek-tiwari.com/bloom-filters-iselement-x-in-set-s/

About Bloom filters Some characteristics

## Some characteristics:

- The possible answers are only booleans ("yes" or "no")
- If n represents the number of elements stored, a Bloom filter requires at least nln(e)/(ln(2))<sup>2</sup> bits. (the memory grows in O(n))

Innin\_

イロト イヨト イヨト イヨト

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

・ロト ・同ト ・ヨト ・ヨト

#### How to identify false positive results ?

Ínría

6/36

æ

Lucas Robidou and Pierre Peterlongo Query-time filtration and MultiFilters

Context

Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

・ロト ・同ト ・ヨト ・ヨト



Compute similarity of an index with a queried sequence:

- extract every k-mer
- query each k-mer
- compute the similarity

(nría\_

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イヨト イヨト イヨト

### Exemple of a query

#### indexed sequence:



Figure: Execution of a query on a perfect filter

Innía

Exemple of a query

9/36

### Exemple of a query



Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イポト イヨト イヨト

### Exemple of a query

indexed sequence:



Figure: Execution of a query on a perfect filter

Ínnía-

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イポト イヨト イヨト

### Exemple of a query

indexed sequence:





11/36

Figure: Execution of a query on a perfect filter

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イポト イヨト イヨト

### Exemple of a query

indexed sequence:



Figure: Execution of a query on a perfect filter

Innia

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イポト イヨト イヨト

### Exemple of a query

indexed sequence:



Figure: Execution of a query on a perfect filter

Innia

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イポト イヨト イヨト

### Exemple of a query

indexed sequence:



Figure: Execution of a query on a perfect filter

Innia

Context Exemple of a query What does a single positive result means? Main idea : query-time filtration Preliminary results

イロト イヨト イヨト イヨト

#### Exemple of a query

indexed sequence:



Perfect filter



Figure: Execution of a query on a perfect filter

Innin\_

Context Exemple of a query What does a single positive result means? Main idea : query-time filtration Preliminary results

イロト イヨト イヨト イヨト

### Exemple of a query

indexed sequence:



**Bloom** filter



Figure: Execution of a query on a Bloom filter

Innin\_

16/36

Lucas Robidou and Pierre Peterlongo Query-time filtration and MultiFilters

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イヨト イヨト イヨト

# Time for some definitions

- Positive k-mer: k-mer that is found in a Bloom filter (might be a FP)
- Neighbours of a k-mer: k-mers located on the "left" and on the "right" of a given k-mer
- Z-positive stretch: stretch of Z positives neighbours
- Isolated k-mer: k-mer part of a 1-positive stretch (no positive neighbours)

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イヨト イヨト イヨト

Exemple of a query with only one isolated kmer



Figure: Execution of a query on a Bloom filter

Low probability to have one isolated shared kmer between two similar sequences.

Ínnía -

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イポト イヨト イヨト



Compute similarity of an index with a queried sequence:

- choose Z
- extract every k-mer
- query each k-mer
- if a kmer is not part of a >Z positive stretch, consider it as a negative
- compute the similarity



Innin -

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

## Preliminary results

Using a Bloom filter:

- $\epsilon = 4\%$
- *k* = 32

Indexing whole gnenomes of:

- 2 E. Coli
- 1 Listeria phage
- 1 Penicillium chrysogenum

Querying another E. Coli with Z = 6 We were able to :

• decrease the false positive rate from 4% to  $\approx$  0.005% (FPR =  $\frac{FP}{FP+TN})$ 

• however: we had a false negative rate of  $\approx 3.65\%$ ( $FNR = \frac{FN}{FN+TP}$ )

Context Exemple of a query What does a single positive result means ? Main idea : query-time filtration Preliminary results

イロト イポト イヨト イヨト

21/36



Figure: *FPR* anf *FNR* of (Bloom filter + query-time filtration) vs "paive" Bloom filter Context Bloom Filters How to identify false positive results ? Multifilters Multifi



Figure: FPR anf FNR of Bloom filter + query-time filtration

イロト イポト イヨト イヨ

Main idea of Multifilters Main problem Applying query-time filtration Preliminary results

(日) (四) (日) (日) (日)

### Multifilters

*Ínría* ≣ ೨९৫ 23/36

Main idea of Multifilters Main problem Applying query-time filtration Preliminary results

イロト イヨト イヨト イヨト

## Main idea of Multifilters

Remember : "The possible answers are only booleans ("yes" or "no")"

 $->\!\!Storing$  kmer abundance: one bloom filter per abundance (or range of abundances)

Ínnía\_

Main idea of Multifilters Main problem Applying query-time filtration Preliminary results

indexed sequence:



**Bloom** filter



イロト イヨト イヨト イヨト



Figure: Main idea of Multifilters

(naío-

シュペ 25/36

Main problem

## Usage of Multifilters

#### indexed sequence:



Main problem

## Usage of Multifilters

#### indexed sequence:



Main problem

28/36

## Usage of Multifilters

#### indexed sequence:



Main problem

## Usage of Multifilters

#### indexed sequence:



Main idea of Multifilters Main problem Applying query-time filtration Preliminary results

イロト イヨト イヨト イヨト

#### Usage of Multifilters



Ínnía\_

シュペ 30/36

Lucas Robidou and Pierre Peterlongo Query-time filtration and MultiFilters

Main idea of Multifilters Main problem Applying query-time filtration Preliminary results

# Main problem Multifilters

More queries : more false positives !

One possible solution is to lower  $\epsilon$  for each filter, but each "sub filter" ends up taking as much space as the original Bloom filter...

If only we had a solution to handle so many false positives !

indexed sequence:



イロト イポト イヨト イヨト

Main idea of Multifilters Main problem **Applying query-time filtration** Preliminary results

# Applying query-time filtration

Compute similarity of an index with a queried sequence:

- extract every k-mer
- query each k-mer on each sub filter
- if a kmer is not part of a >Z positive stretch **on a sub filter**, consider it as a negative
- else, select the filter for which the number of positives neighbours is maximized
- compute the similarity



イロト イポト イヨト イヨト

Main idea of Multifilters Main problem Applying query-time filtration Preliminary results

## Exemple of an execution

This way, we are able to get rid of most false positives:



Figure: Execution of a query on a Multifilter with query time filtration

イロト イヨト イヨト イヨト

Main idea of Multifilters Main problem Applying query-time filtration Preliminary results

## Exemple of an execution

This way, we are able to get rid of most false positives:



Figure: Execution of a query on a Multifilter with query time filtration

イロト イタト イヨト イヨト

Main idea of Multifilters Main problem Applying query-time filtration Preliminary results

イロト イポト イヨト イヨト

35/36



Figure: *FPR* anf *FNR* of (Bloom filter + query-time filtration) vs (Multifilters + query-time filtration)

Main idea of Multifilters Main problem Applying query-time filtration Preliminary results

イロト イヨト イヨト イヨト

## Take home message

By using one Bloom filter per abundance, we can:

- index kmers along with their abundance
- without needing more memory
- with about 800 times less false positives

The price to pay is about 4% of false negatives rate, which is 4% more than Bloom filters.