Indexing a large set of reads

Nicolas Philippe[1,2] Mikaël Salson[3] Thierry Lecroq[4]
Martine Léonard[4] Thérèse Commes[2] Éric Rivals[1]

[1] LIRMM, CNRS and Université de Montpellier 2
[2] IGH, CNRS, Montpellier
[3] LIFL, CNRS and Université de Lille I – INRIA Lille-Nord Europe
[4] LITIS, Université de Rouen

10 January 2011

Introduction

Context

Next generation sequencers produce gigabytes of reads in a single run

Introduction

Context

Next generation sequencers produce gigabytes of reads in a single run

Problem

How to search efficiently any relevant information?

Introduction

Context

Next generation sequencers produce gigabytes of reads in a single run

Problem

How to search efficiently any relevant information?

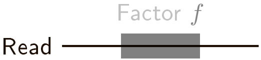
Introduction

Context

Next generation sequencers produce gigabytes of reads in a single run

Problem

How to search efficiently any relevant information?



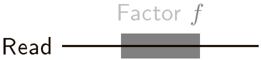
Introduction

Context

Next generation sequencers produce gigabytes of reads in a single run

Problem

How to search efficiently any relevant information?



Interesting questions

- How many reads share this factor f?
- Which reads share this factor? At which positions?

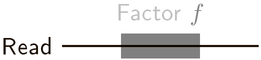
Introduction

Context

Next generation sequencers produce gigabytes of reads in a single run

Problem

How to search efficiently any relevant information?



Interesting questions

- How many reads share this factor f?
- Which reads share this factor? At which positions?

Why is it interesting?

- Genome assembly
- Read mapping
- ...

Fixed-length factors

Question

Should we consider every factor?

Fixed-length factors

Question

Should we consider every factor?

Remarks

- Factors of length 2 are quite uninformative
- At a certain point, increasing factor lengths does not help in identifying unique genome location ([Philippe *et al.*, 2009])

Fixed-length factors

Question

Should we consider every factor?

Remarks

- Factors of length 2 are quite uninformative
- At a certain point, increasing factor lengths does not help in identifying unique genome location ([Philippe *et al.*, 2009])

Conclusion

We only consider k-length factors (k-factors or k-mers), k being fixed

Queries

Queries for k-factors of a given read

Given a read, and a k-factor in that read, we would like to know:

- Q1 the number of times this k-factor appears in the whole set of reads
- Q2 the reads and the positions in the reads in which it occurs
- Q3 the number of distinct reads in which it occurs
- ...

An immediate solution

Remark

We need to search patterns in a text

An immediate solution

Remark

We need to search patterns in a text

Classical solution

Use a text index

An immediate solution

Remark

We need to search patterns in a text

Classical solution

Use a text index

- Suffix tree
- Suffix array
- Compressed text index (FM-index, LZ-index, . . .)

An immediate solution

Remark

We need to search patterns in a text

Classical solution

Use a text index

- Suffix tree
- Suffix array
- Compressed text index (FM-index, LZ-index, . . .)

Ok, let's try a suffix array!

Using a Suffix Array for querying reads

Reads: $r_0 = \overset{0}{\mathrm{A}}\,\overset{1}{\mathrm{T}}\,\overset{2}{\mathrm{A}}\,\overset{3}{\mathrm{A}}\,\overset{4}{\mathrm{C}}\,\overset{5}{\mathrm{G}}$ $r_1 = \overset{6}{\mathrm{A}}\,\overset{7}{\mathrm{T}}\,\overset{8}{\mathrm{A}}\,\overset{9}{\mathrm{G}}\,\overset{10}{\mathrm{T}}\,\overset{11}{\mathrm{C}}$ $r_2 = \overset{12}{\mathrm{G}}\,\overset{13}{\mathrm{A}}\,\overset{14}{\mathrm{T}}\,\overset{15}{\mathrm{A}}\,\overset{16}{\mathrm{A}}\,\overset{17}{\mathrm{C}}$

Using a Suffix Array for querying reads

Reads:

	0	1	2	3	4	5		6	7	8	9	10	11		12	13	14	15	16	17
$r_0 =$	A	T	A	A	C	G	$r_1 =$	A	T	A	G	T	C	$r_2 =$	G	A	T	A	A	C

$$R = r_0 \cdot r_1 \cdot r_2 \cdot \$$$

Using a Suffix Array for querying reads

Reads: $r_0 = \begin{array}{cccccc}0&1&2&3&4&5\\ \mathrm{A}&\mathrm{T}&\mathrm{A}&\mathrm{A}&\mathrm{C}&\mathrm{G}\end{array}$ $r_1 = \begin{array}{cccccc}6&7&8&9&10&11\\ \mathrm{A}&\mathrm{T}&\mathrm{A}&\mathrm{G}&\mathrm{T}&\mathrm{C}\end{array}$ $r_2 = \begin{array}{cccccc}12&13&14&15&16&17\\ \mathrm{G}&\mathrm{A}&\mathrm{T}&\mathrm{A}&\mathrm{A}&\mathrm{C}\end{array}$

$$R = r_0 \cdot r_1 \cdot r_2 \cdot \$ = \begin{array}{ccccccccccccccccccc}0&1&2&3&4&5&6&7&8&9&10&11&12&13&14&15&16&17&18\\ \mathrm{A}&\mathrm{T}&\mathrm{A}&\mathrm{A}&\mathrm{C}&\mathrm{G}&\mathrm{A}&\mathrm{T}&\mathrm{A}&\mathrm{G}&\mathrm{T}&\mathrm{C}&\mathrm{G}&\mathrm{A}&\mathrm{T}&\mathrm{A}&\mathrm{A}&\mathrm{C}&\$\\ &&r_0&&&&&&r_1&&&&&&r_2&&&&\end{array}$$

Using a Suffix Array for querying reads

Reads: $r_0 = \begin{array}{cccccc} 0&1&2&3&4&5\\ \text{A}&\text{T}&\text{A}&\text{A}&\text{C}&\text{G}\end{array}$ $r_1 = \begin{array}{cccccc} 6&7&8&9&10&11\\ \text{A}&\text{T}&\text{A}&\text{G}&\text{T}&\text{C}\end{array}$ $r_2 = \begin{array}{cccccc} 12&13&14&15&16&17\\ \text{G}&\text{A}&\text{T}&\text{A}&\text{A}&\text{C}\end{array}$

$$R = r_0 \cdot r_1 \cdot r_2 \cdot \$ = \begin{array}{ccccccccccccccccccc} 0&1&2&3&4&5&6&7&8&9&10&11&12&13&14&15&16&17&18\\ \text{A}&\text{T}&\text{A}&\text{A}&\text{C}&\text{G}&\text{A}&\text{T}&\text{A}&\text{G}&\text{T}&\text{C}&\text{G}&\text{A}&\text{T}&\text{A}&\text{A}&\text{C}&\$ \\ &&&r_0&&&&&&r_1&&&&&&r_2&&& \end{array}$$

Let's build the suffix array (sort suffixes in lexicographic ascending order)

SA	Suffixes
18	$
15	AAC$
2	AACGATAGTCGATAAC$
16	AC$
3	ACGATAGTCGATAAC$
8	AGTCGATAAC$
13	ATAAC$
0	ATAACGATAGTCGATAAC$
17	C$
11	CGATAAC$
⋮	⋮

Using a Suffix Array for querying reads

Reads: $r_0 = $ ATAACG $r_1 = $ ATAGTC $r_2 = $ GATAAC

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C

$$R = r_0 \cdot r_1 \cdot r_2 \cdot \$ = \underbrace{ATAACG}_{r_0}\underbrace{ATAGTC}_{r_1}\underbrace{GATAAC}_{r_2}\$$$

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$

Let's build the suffix array (sort suffixes in lexicographic ascending order)

SA	Suffixes
18	$
15	AAC$
2	AACGATAGTCGATAAC$
16	AC$
3	ACGATAGTCGATAAC$
8	AGTCGATAAC$
13	ATAAC$
0	ATAACGATAGTCGATAAC$
17	C$
11	CGATAAC$
⋮	⋮

Remark

Only the k first letters of each suffix are interesting ($k = 3$)

Using a Suffix Array for querying reads

Reads:

	0	1	2	3	4	5
$r_0 =$	A	T	A	A	C	G

	6	7	8	9	10	11
$r_1 =$	A	T	A	G	T	C

	12	13	14	15	16	17
$r_2 =$	G	A	T	A	A	C

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R = r_0 \cdot r_1 \cdot r_2 \cdot \$ =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
			r_0						r_1						r_2				

Let's build the suffix array (sort suffixes in lexicographic ascending order)

SA	Suffixes
18	$
15	AAC$
2	AACGATAGTCGATAAC$
16	AC$
3	ACGATAGTCGATAAC$
8	AGTCGATAAC$
13	ATAAC$
0	ATAACGATAGTCGATAAC$
17	C$
11	CGATAAC$
⋮	⋮

Remark

Only the k first letters of each suffix are interesting ($k = 3$)

Using a Suffix Array for querying reads

Reads: $r_0 = \begin{array}{cccccc} 0&1&2&3&4&5\\ A&T&A&A&C&G \end{array}$ $r_1 = \begin{array}{cccccc} 6&7&8&9&10&11\\ A&T&A&G&T&C \end{array}$ $r_2 = \begin{array}{cccccc} 12&13&14&15&16&17\\ G&A&T&A&A&C \end{array}$

$$R = r_0 \cdot r_1 \cdot r_2 \cdot \$ = \begin{array}{ccccccccccccccccccc} 0&1&2&3&4&5&6&7&8&9&10&11&12&13&14&15&16&17&18\\ A&T&A&A&C&G&A&T&A&G&T&C&G&A&T&A&A&C&\$\\ \multicolumn{6}{c}{r_0}&\multicolumn{6}{c}{r_1}&\multicolumn{6}{c}{r_2}& \end{array}$$

Let's build the suffix array (sort suffixes in lexicographic ascending order)

SA	Suffixes
18	$
15	AAC$
2	AACGATAGTCGATAAC$
16	AC$
3	ACGATAGTCGATAAC$
8	AGTCGATAAC$
13	ATAAC$
0	ATAACGATAGTCGATAAC$
17	C$
11	CGATAAC$
⋮	⋮

Remark

Only the k first letters of each suffix are interesting ($k = 3$)

Remark

Factors overlapping two reads are undesirable

Using a Suffix Array for querying reads

Reads:

	0	1	2	3	4	5
$r_0 =$	A	T	A	A	C	G

	6	7	8	9	10	11
$r_1 =$	A	T	A	G	T	C

	12	13	14	15	16	17
$r_2 =$	G	A	T	A	A	C

$R = r_0 \cdot r_1 \cdot r_2 \cdot \$ =$

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$

r_0 r_1 r_2

Let's build the suffix array (sort suffixes in lexicographic ascending order)

SA	Suffixes
18	~~$~~
15	AAC$
2	AACGATAGTCGATAAC$
16	~~AC$~~
3	ACGATAGTCGATAAC$
8	AGTCGATAAC$
13	ATAAC$
0	ATAACGATAGTCGATAAC$
17	~~C$~~
11	~~CGA~~TAAC$
⋮	⋮

Remark

Only the k first letters of each suffix are interesting ($k = 3$)

Remark

Factors overlapping two reads are undesirable

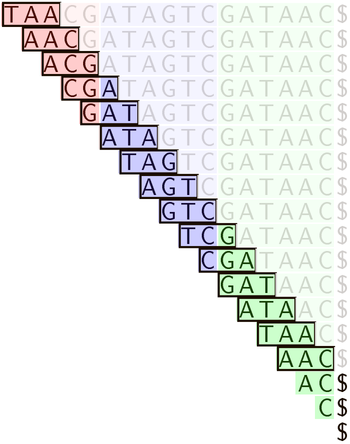
Discarding useless positions

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$

```
 TAACGATAGTCGATAAC$
  AACGATAGTCGATAAC$
   ACGATAGTCGATAAC$
    CGATAGTCGATAAC$
     GATAGTCGATAAC$
      ATAGTCGATAAC$
       TAGTCGATAAC$
        AGTCGATAAC$
         GTCGATAAC$
          TCGATAAC$
           CGATAAC$
            GATAAC$
             ATAAC$
              TAAC$
               AAC$
                AC$
                 C$
                  $
```

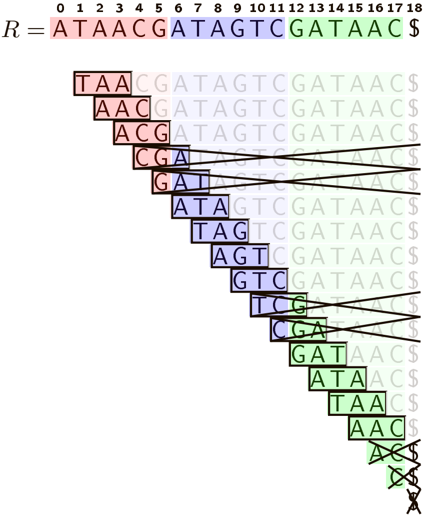
Discarding useless positions

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$

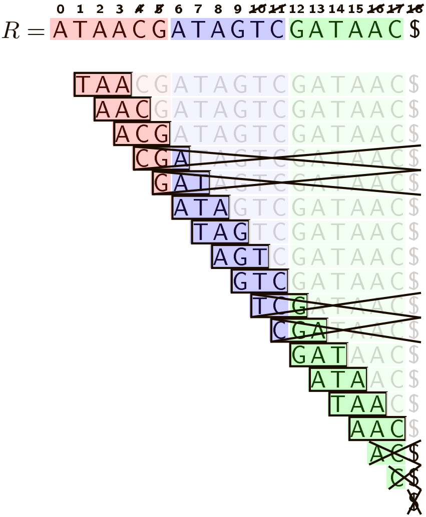


```
TAACGATAGTCGATAAC$
 AACGATAGTCGATAAC$
  ACGATAGTCGATAAC$
   CGATAGTCGATAAC$
    GATAGTCGATAAC$
     ATAGTCGATAAC$
      TAGTCGATAAC$
       AGTCGATAAC$
        GTCGATAAC$
         TCGATAAC$
          CGATAAC$
           GATAAC$
            ATAAC$
             TAAC$
              AAC$
               AC$
                C$
                 $
```

Discarding useless positions



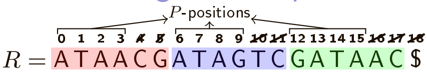
Discarding useless positions



0 1 2 3 ~~4~~ ~~5~~ 6 7 8 9 ~~10~~ ~~11~~ 12 13 14 15 ~~16~~ ~~17~~ ~~18~~

$R =$ ATAACG ATAGTC GATAAC $

Discarding useless positions



P-positions

0 1 2 3 ~~4~~ ~~5~~ 6 7 8 9 ~~10~~ ~~11~~ 12 13 14 15 ~~16~~ ~~17~~ ~~18~~

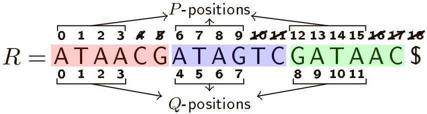
$R =$ ATAACG ATAGTC GATAAC \$

P-positions

- Set of positions where a k-factor belonging to a single read starts.
- This set is not a permutation

A k-factor starting at a P-position is called a P-k-factor

Discarding useless positions



P-positions

0 1 2 3 ~~4~~ ~~5~~ 6 7 8 9 ~~10~~ ~~11~~ 12 13 14 15 ~~16~~ ~~17~~ ~~18~~

$R =$ ATAACG ATAGTC GATAAC \$

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

P-positions

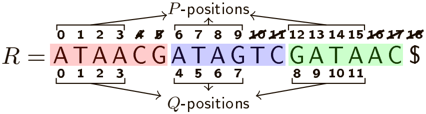
- Set of positions where a k-factor belonging to a single read starts.
- This set is not a permutation

A k-factor starting at a P-position is called a P-k-factor

Q-positions

Renumbered P-positions so that the set of Q-positions is a permutation

Generalized k-Factor Array (GkFA)



P-positions

0 1 2 3 ~~4~~ ~~5~~ 6 7 8 9 ~~10~~ ~~11~~ 12 13 14 15 ~~16~~ ~~17~~ ~~18~~

$R =$ ATAACG ATAGTC GATAAC \$

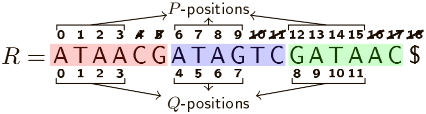
0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

Generalized k-factor array

Index suffixes starting at P-positions. Positions are renumbered to Q-positions.

Generalized k-Factor Array (GkFA)



$R =$ ATAACG ATAGTC GATAAC $

P-positions: 0 1 2 3 ~~4~~ ~~5~~ 6 7 8 9 ~~10~~ ~~11~~ 12 13 14 15 ~~16~~ ~~17~~ ~~18~~

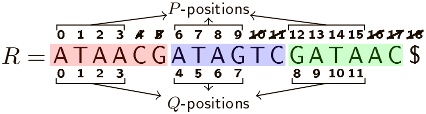
Q-positions: 0 1 2 3 4 5 6 7 8 9 10 11

Generalized k-factor array

Index suffixes starting at P-positions. Positions are renumbered to Q-positions.

SA	Suffixes
18	$
15	AAC$
2	AACGATAGTCGATAAC$
16	AC$
3	ACGATAGTCGATAAC$
8	AGTCGATAAC$
13	ATAAC$
0	ATAACGATAGTCGATAAC$
6	ATAGTCGATAAC$
17	C$
11	CGATAAC$
4	CGATAGTCGATAAC$
12	GATAAC$
5	GATAGTCGATAAC$
9	GTCGATAAC$
14	TAAC$
1	TAACGATAGTCGATAAC$
7	TAGTCGATAAC$
10	TCGATAAC$

Generalized k-Factor Array (GkFA)

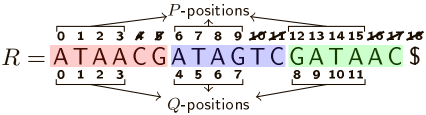


SA	Suffixes
~~18~~	~~$~~
15	AAC$
2	AACGATAGTCGATAAC$
~~16~~	~~AC$~~
3	ACGATAGTCGATAAC$
8	AGTCGATAAC$
13	ATAAC$
0	ATAACGATAGTCGATAAC$
6	ATAGTCGATAAC$
~~17~~	~~C$~~
~~11~~	~~CGATAAC$~~
~~4~~	~~CGATAGTCGATAAC$~~
12	GATAAC$
~~5~~	~~GATAGTCGATAAC$~~
9	GTCGATAAC$
14	TAAC$
1	TAACGATAGTCGATAAC$
7	TAGTCGATAAC$
~~10~~	~~TCGATAAC$~~

Generalized k-factor array

Index suffixes starting at P-positions. Positions are renumbered to Q-positions.

Generalized k-Factor Array (GkFA)

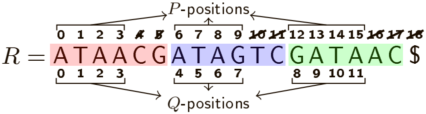


SA	Suffixes	GkFA (P-positions)
~~18~~	~~$~~	
15	AAC$	15
2	AACGATAGTCGATAAC$	2
~~16~~	~~AC$~~	
3	ACGATAGTCGATAAC$	3
8	AGTCGATAAC$	8
13	ATAAC$	13
0	ATAACGATAGTCGATAAC$	0
6	ATAGTCGATAAC$	6
~~17~~	~~C$~~	
~~11~~	~~CGATAAC$~~	
~~4~~	~~CGATAGTCGATAAC$~~	
12	GATAAC$	12
~~5~~	~~GATAGTCGATAAC$~~	
9	GTCGATAAC$	9
14	TAAC$	14
1	TAACGATAGTCGATAAC$	1
7	TAGTCGATAAC$	7
~~10~~	~~TCGATAAC$~~	

Generalized k-factor array

Index suffixes starting at P-positions. Positions are renumbered to Q-positions.

Generalized k-Factor Array (GkFA)

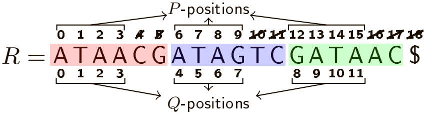


SA	Suffixes	GkFA (P-positions)	GkFA (Q-positions)
~~18~~	~~$~~		
15	AAC$	15	11
2	AACGATAGTCGATAAC$	2	2
~~16~~	~~AC$~~		
3	ACGATAGTCGATAAC$	3	3
8	AGTCGATAAC$	8	6
13	ATAAC$	13	9
0	ATAACGATAGTCGATAAC$	0	0
6	ATAGTCGATAAC$	6	4
~~17~~	~~C$~~		
~~11~~	~~CGATAAC$~~		
~~4~~	~~CGATAGTCGATAAC$~~		
12	GATAAC$	12	8
~~5~~	~~GATAGTCGATAAC$~~		
9	GTCGATAAC$	9	7
14	TAAC$	14	10
1	TAACGATAGTCGATAAC$	1	1
7	TAGTCGATAAC$	7	5
~~10~~	~~TCGATAAC$~~		

Generalized k-factor array

Index suffixes starting at P-positions. Positions are renumbered to Q-positions.

Generalized k-Factor Array (GkFA)



SA	Suffixes	GkFA (Q-positions)
~~18~~	~~$~~	
15	AAC$	11
2	AACGATAGTCGATAAC$	2
~~16~~	~~AC$~~	
3	ACGATAGTCGATAAC$	3
8	AGTCGATAAC$	6
13	ATAAC$	9
0	ATAACGATAGTCGATAAC$	0
6	ATAGTCGATAAC$	4
~~17~~	~~C$~~	
~~11~~	~~CGATAAC$~~	
~~4~~	~~CGATAGTCGATAAC$~~	
12	GATAAC$	8
~~5~~	~~GATAGTCGATAAC$~~	
9	GTCGATAAC$	7
14	TAAC$	10
1	TAACGATAGTCGATAAC$	1
7	TAGTCGATAAC$	5
~~10~~	~~TCGATAAC$~~	

Generalized k-factor array

Index suffixes starting at P-positions. Positions are renumbered to Q-positions.

Generalized k Count Factor Array

$$
\begin{array}{ccccccccccccccccccccc}
R = & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{G} & \mathsf{T} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \$ \\
 & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & & & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} & & & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} & & &
\end{array}
$$

Q-positions

Generalized k Count Factor Array

$$R = \begin{array}{ccccccccccccccccccc} \mathsf{A}&\mathsf{T}&\mathsf{A}&\mathsf{A}&\mathsf{C}&\mathsf{G}&\mathsf{A}&\mathsf{T}&\mathsf{A}&\mathsf{G}&\mathsf{T}&\mathsf{C}&\mathsf{G}&\mathsf{A}&\mathsf{T}&\mathsf{A}&\mathsf{A}&\mathsf{C}&\$ \\ \mathbf{0}&\mathbf{1}&\mathbf{2}&\mathbf{3}&&&\mathbf{4}&\mathbf{5}&\mathbf{6}&\mathbf{7}&&&\mathbf{8}&\mathbf{9}&\mathbf{10}&\mathbf{11}&&& \end{array}$$

Q-positions

GkCFA

Count the number of occurrences of a P-k-factor

Generalized k Count Factor Array

$$R = \begin{array}{ccccccccccccccccccc} \mathsf{A}&\mathsf{T}&\mathsf{A}&\mathsf{A}&\mathsf{C}&\mathsf{G}&\mathsf{A}&\mathsf{T}&\mathsf{A}&\mathsf{G}&\mathsf{T}&\mathsf{C}&\mathsf{G}&\mathsf{A}&\mathsf{T}&\mathsf{A}&\mathsf{A}&\mathsf{C}&\$ \\ \mathbf{0}&\mathbf{1}&\mathbf{2}&\mathbf{3}&&&\mathbf{4}&\mathbf{5}&\mathbf{6}&\mathbf{7}&&&\mathbf{8}&\mathbf{9}&\mathbf{10}&\mathbf{11}&&& \end{array}$$

Q-positions

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$$R = \begin{array}{ccccccccccccccccccc} \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{G} & \mathsf{T} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \$ \\ 0 & 1 & 2 & 3 & & & 4 & 5 & 6 & 7 & & & 8 & 9 & 10 & 11 & & & \end{array}$$

Q-positions

GkFA	k-factor
11	AAC
2	AAC
3	ACG
6	AGT
9	ATA
0	ATA
4	ATA
8	GAT
7	GTC
10	TAA
1	TAA
5	TAG

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R =$ A T A A C G A T A G T C G A T A A C \$

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

	GkFA	k-factor
0	11	AAC
	2	AAC
	3	ACG
	6	AGT
	9	ATA
	0	ATA
	4	ATA
	8	GAT
	7	GTC
	10	TAA
	1	TAA
	5	TAG

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R =$ A T A A C G A T A G T C G A T A A C \$

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
	3	ACG	
	6	AGT	
	9	ATA	
	0	ATA	
	4	ATA	
	8	GAT	
	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$$R = \underset{0}{A}\,\underset{1}{T}\,\underset{2}{A}\,\underset{3}{A}\,C\,G\,\underset{4}{A}\,\underset{5}{T}\,\underset{6}{A}\,\underset{7}{G}\,T\,C\,\underset{8}{G}\,\underset{9}{A}\,\underset{10}{T}\,\underset{11}{A}\,A\,C\,\$$$

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	
	6	AGT	
	9	ATA	
	0	ATA	
	4	ATA	
	8	GAT	
	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	\$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
	6	AGT	
	9	ATA	
	0	ATA	
	4	ATA	
	8	GAT	
	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$$R = \begin{array}{ccccccccccccccccccc} A & T & A & A & C & G & A & T & A & G & T & C & G & A & T & A & A & C & \$ \\ 0 & 1 & 2 & 3 & & & 4 & 5 & 6 & 7 & & & 8 & 9 & 10 & 11 & & & \end{array}$$

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	
	9	ATA	
	0	ATA	
	4	ATA	
	8	GAT	
	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7				8	9	10	11		

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
	9	ATA	
	0	ATA	
	4	ATA	
	8	GAT	
	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R =$ A T A A C G A T A G T C G A T A A C \$

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	
	0	ATA	
	4	ATA	
	8	GAT	
	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$$R = \underset{0}{A}\underset{1}{T}\underset{2}{A}\underset{3}{A}CG\underset{4}{A}\underset{5}{T}\underset{6}{A}\underset{7}{G}TC\underset{8}{G}\underset{9}{A}\underset{10}{T}\underset{11}{A}AC\$$$

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
	8	GAT	
	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R = \mathtt{A\,T\,A\,A\,C\,G\,A\,T\,A\,G\,T\,C\,G\,A\,T\,A\,A\,C\,\$}$

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
4	8	GAT	
	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
4	8	GAT	1
	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

Generalized k Count Factor Array

$$\begin{array}{cccccccccccccccccccc} R = & A & T & A & A & C & G & A & T & A & G & T & C & G & A & T & A & A & C & \$ \\ & 0 & 1 & 2 & 3 & & & 4 & 5 & 6 & 7 & & & 8 & 9 & 10 & 11 & & & \end{array}$$

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
4	8	GAT	1
5	7	GTC	
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$$R = \underset{0}{\mathrm{A}}\,\underset{1}{\mathrm{T}}\,\underset{2}{\mathrm{A}}\,\underset{3}{\mathrm{A}}\,\mathrm{C}\,\mathrm{G}\,\underset{4}{\mathrm{A}}\,\underset{5}{\mathrm{T}}\,\underset{6}{\mathrm{A}}\,\underset{7}{\mathrm{G}}\,\mathrm{T}\,\mathrm{C}\,\underset{8}{\mathrm{G}}\,\underset{9}{\mathrm{A}}\,\underset{10}{\mathrm{T}}\,\underset{11}{\mathrm{A}}\,\mathrm{A}\,\mathrm{C}\,\$$$

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
4	8	GAT	1
5	7	GTC	1
	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R = \mathtt{ATAACGATAGTCGATAAC\$}$

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
4	8	GAT	1
5	7	GTC	1
6	10	TAA	
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
4	8	GAT	1
5	7	GTC	1
6	10	TAA	2
	1	TAA	
	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$$R = \begin{array}{ccccccccccccccccccc} A&T&A&A&C&G&A&T&A&G&T&C&G&A&T&A&A&C&\$\\ 0&1&2&3&&&4&5&6&7&&&8&9&10&11&&& \end{array}$$

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
	9	ATA	
3	0	ATA	3
	4	ATA	
4	8	GAT	1
5	7	GTC	1
6	10	TAA	2
	1	TAA	
7	5	TAG	

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$R = \text{A T A A C G A T A G T C G A T A A C \$}$

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

	GkFA	k-factor	
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
4	8	GAT	1
5	7	GTC	1
6	10	TAA	2
	1	TAA	
7	5	TAG	1

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

Generalized k Count Factor Array

$$R = \underset{0}{\mathrm{A}}\underset{1}{\mathrm{T}}\underset{2}{\mathrm{A}}\underset{3}{\mathrm{A}}\mathrm{C}\mathrm{G}\underset{4}{\mathrm{A}}\underset{5}{\mathrm{T}}\underset{6}{\mathrm{A}}\underset{7}{\mathrm{G}}\mathrm{T}\mathrm{C}\underset{8}{\mathrm{G}}\underset{9}{\mathrm{A}}\underset{10}{\mathrm{T}}\underset{11}{\mathrm{A}}\mathrm{A}\mathrm{C}\$$$

Q-positions

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

	GkFA	k-factor	GkCFA
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
4	8	GAT	1
5	7	GTC	1
6	10	TAA	2
	1	TAA	
7	5	TAG	1

GkCFA

Generalized k Count Factor Array

$$R = \underset{0}{\mathsf{A}}\,\underset{1}{\mathsf{T}}\,\underset{2}{\mathsf{A}}\,\underset{3}{\mathsf{A}}\,\mathsf{C}\,\mathsf{G}\,\underset{4}{\mathsf{A}}\,\underset{5}{\mathsf{T}}\,\underset{6}{\mathsf{A}}\,\underset{7}{\mathsf{G}}\,\mathsf{T}\,\mathsf{C}\,\mathsf{G}\,\underset{8}{\mathsf{A}}\,\underset{9}{\mathsf{T}}\,\underset{10}{\mathsf{A}}\,\underset{11}{\mathsf{A}}\,\mathsf{C}\,\$$$

Q-positions

GkCFA

Count the number of occurrences of a P-k-factor

Purpose

Compute the read coverage of a given region *inside* a read

IDs	GkFA	k-factor	GkCFA
0	11	AAC	2
	2	AAC	
1	3	ACG	1
2	6	AGT	1
3	9	ATA	3
	0	ATA	
	4	ATA	
4	8	GAT	1
5	7	GTC	1
6	10	TAA	2
	1	TAA	
7	5	TAG	1

Generalized k Inverse Factor Array (GkIFA)

$$R = \underset{0}{\mathsf{A}}\,\underset{1}{\mathsf{T}}\,\underset{2}{\mathsf{A}}\,\underset{3}{\mathsf{A}}\,\mathsf{C}\,\mathsf{G}\,\underset{4}{\mathsf{A}}\,\underset{5}{\mathsf{T}}\,\underset{6}{\mathsf{A}}\,\underset{7}{\mathsf{G}}\,\mathsf{T}\,\mathsf{C}\,\underset{8}{\mathsf{G}}\,\underset{9}{\mathsf{A}}\,\underset{10}{\mathsf{T}}\,\underset{11}{\mathsf{A}}\,\mathsf{A}\,\mathsf{C}\,\$$$

Q-positions

Generalized k Inverse Factor Array (GkIFA)

$$R = \underset{0}{\mathrm{A}}\,\underset{1}{\mathrm{T}}\,\underset{2}{\mathrm{A}}\,\underset{3}{\mathrm{A}}\,\mathrm{C}\,\mathrm{G}\,\underset{4}{\mathrm{A}}\,\underset{5}{\mathrm{T}}\,\underset{6}{\mathrm{A}}\,\underset{7}{\mathrm{G}}\,\mathrm{T}\,\mathrm{C}\,\underset{8}{\mathrm{G}}\,\underset{9}{\mathrm{A}}\,\underset{10}{\mathrm{T}}\,\underset{11}{\mathrm{A}}\,\mathrm{A}\,\mathrm{C}\,\$$$

Q-positions

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Generalized k Inverse Factor Array (GkIFA)

```
R = A T A A C G A T A G T C G A T A A C $
    0 1 2 3     4 5 6 7     8 9 10 11
             Q-positions
```

GkIFA

- “Inverse” of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

Generalized k Inverse Factor Array (GkIFA)

$$R = \begin{array}{ccccccccccccccccccc} \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{G} & \mathsf{T} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \$ \\ 0 & 1 & 2 & 3 & & & 4 & 5 & 6 & 7 & & & 8 & 9 & 10 & 11 & & & \end{array}$$

Q-positions

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	
	2	AAC	1	
1	3	ACG	2	
2	6	AGT	3	
3	9	ATA	4	
	0	ATA	5	
	4	ATA	6	
4	8	GAT	7	
5	7	GTC	8	
6	10	TAA	9	
	1	TAA	10	
7	5	TAG	11	

Generalized k Inverse Factor Array (GkIFA)

$R =$ ATAACGATAGTCGATAAC$

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

GkIFA

- “Inverse” of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	
	2	AAC	1	
1	3	ACG	2	0
2	6	AGT	3	
3	9	ATA	4	
	0	ATA	5	
	4	ATA	6	
4	8	GAT	7	
5	7	GTC	8	
6	10	TAA	9	
	1	TAA	10	
7	5	TAG	11	0

Generalized k Inverse Factor Array (GkIFA)

$R =$ A T A A C G A T A G T C G A T A A C \$

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	
	2	AAC	1	
1	3	ACG	2	0
2	6	AGT	3	1
3	9	ATA	4	
	0	ATA	5	
	4	ATA	6	
4	8	GAT	7	
5	7	GTC	8	
6	10	TAA	9	
	1	TAA	10	
7	5	TAG	11	0

Generalized k Inverse Factor Array (GkIFA)

$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	
	2	AAC	1	
1	3	ACG	2	0
2	6	AGT	3	1
3	9	ATA	4	
	0	ATA	5	
	4	ATA	6	2
4	8	GAT	7	
5	7	GTC	8	
6	10	TAA	9	
	1	TAA	10	
7	5	TAG	11	0

Generalized k Inverse Factor Array (GkIFA)

$$R = \underset{0}{\mathrm{A}}\,\underset{1}{\mathrm{T}}\,\underset{2}{\mathrm{A}}\,\underset{3}{\mathrm{A}}\,\mathrm{C}\,\mathrm{G}\,\underset{4}{\mathrm{A}}\,\underset{5}{\mathrm{T}}\,\underset{6}{\mathrm{A}}\,\underset{7}{\mathrm{G}}\,\mathrm{T}\,\mathrm{C}\,\underset{8}{\mathrm{G}}\,\underset{9}{\mathrm{A}}\,\underset{10}{\mathrm{T}}\,\underset{11}{\mathrm{A}}\,\mathrm{A}\,\mathrm{C}\,\$$$

Q-positions

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	3
	2	AAC	1	
1	3	ACG	2	0
2	6	AGT	3	1
3	9	ATA	4	3
	0	ATA	5	
	4	ATA	6	2
4	8	GAT	7	
5	7	GTC	8	
6	10	TAA	9	3
	1	TAA	10	
7	5	TAG	11	0

Generalized k Inverse Factor Array (GkIFA)

$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	\$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	3
	2	AAC	1	
1	3	ACG	2	0
2	6	AGT	3	1
3	9	ATA	4	3
	0	ATA	5	
	4	ATA	6	2
4	8	GAT	7	
5	7	GTC	8	4
6	10	TAA	9	3
	1	TAA	10	
7	5	TAG	11	0

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

Generalized k Inverse Factor Array (GkIFA)

$$R = \begin{array}{ccccccccccccccccccc} A & T & A & A & C & G & A & T & A & G & T & C & G & A & T & A & A & C & \$ \\ 0 & 1 & 2 & 3 & & & 4 & 5 & 6 & 7 & & & 8 & 9 & 10 & 11 & & & \end{array}$$

Q-positions

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	3
	2	AAC	1	
1	3	ACG	2	0
2	6	AGT	3	1
3	9	ATA	4	3
	0	ATA	5	
	4	ATA	6	2
4	8	GAT	7	5
5	7	GTC	8	4
6	10	TAA	9	3
	1	TAA	10	
7	5	TAG	11	0

Generalized k Inverse Factor Array (GkIFA)

$$R = \begin{array}{ccccccccccccccccccc} \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{G} & \mathsf{T} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \$ \\ \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & & & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} & & & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} & & & \end{array}$$

Q-positions

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	3
	2	AAC	1	6
1	3	ACG	2	0
2	6	AGT	3	1
3	9	ATA	4	3
	0	ATA	5	
	4	ATA	6	2
4	8	GAT	7	5
5	7	GTC	8	4
6	10	TAA	9	3
	1	TAA	10	6
7	5	TAG	11	0

Generalized k Inverse Factor Array (GkIFA)

$R =$ A T A A C G A T A G T C G A T A A C $

0 1 2 3 4 5 6 7 8 9 10 11

Q-positions

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	3
	2	AAC	1	6
1	3	ACG	2	0
2	6	AGT	3	1
3	9	ATA	4	3
	0	ATA	5	7
	4	ATA	6	2
4	8	GAT	7	5
5	7	GTC	8	4
6	10	TAA	9	3
	1	TAA	10	6
7	5	TAG	11	0

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

Generalized k Inverse Factor Array (GkIFA)

$$R = \underset{0}{A}\underset{1}{T}\underset{2}{A}\underset{3}{A}CG\underset{4}{A}\underset{5}{T}\underset{6}{A}\underset{7}{G}TC\underset{8}{G}\underset{9}{A}\underset{10}{T}\underset{11}{A}AC\$$$

Q-positions

GkIFA

- "Inverse" of GkFA
- Given a Q-position in R, stores the id associated to the corresponding k-factor

Purpose

Quickly find the id associated to a k-factor coming from a read

	GkFA	k-factor	i	GkIFA
0	11	AAC	0	3
	2	AAC	1	6
1	3	ACG	2	0
2	6	AGT	3	1
3	9	ATA	4	3
	0	ATA	5	7
	4	ATA	6	2
4	8	GAT	7	5
5	7	GTC	8	4
6	10	TAA	9	3
	1	TAA	10	6
7	5	TAG	11	0

Using Gk arrays (I)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	\$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Using Gk arrays (I)

$$
\begin{array}{rccccccccccccccccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 \\
R = & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \mathrm{T} & \mathrm{C} & \mathrm{G} & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{A} & \mathrm{C} & \$ \\
 & 0 & 1 & 2 & 3 & & & 4 & 5 & 6 & 7 & & & 8 & 9 & 10 & 11 & & &
\end{array}
$$

Q-positions

Searching a k-factor of a given read

Read r_2 at position 0 → GAT

Using Gk arrays (I)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
Q-positions	0	1	2	3			4	5	6	7			8	9	10	11			

Searching a k-factor of a given read

Read r_2 at position 0 → GAT

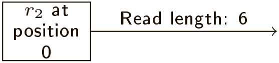
r_2 at position 0

Using Gk arrays (I)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
Q-positions	0	1	2	3			4	5	6	7			8	9	10	11			

Searching a k-factor of a given read

Read r_2 at position 0 → GAT



r_2 at position 0

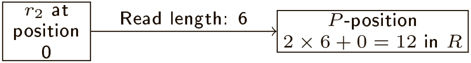
Read length: 6

Using Gk arrays (I)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
Q-positions	0	1	2	3			4	5	6	7			8	9	10	11			

Searching a k-factor of a given read

Read r_2 at position 0 → GAT



r_2 at position 0

Read length: 6

P-position $2 \times 6 + 0 = 12$ in R

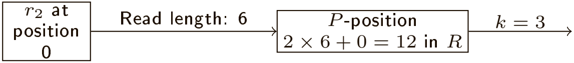
Using Gk arrays (I)

$$
\begin{array}{rccccccccccccccccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 \\
R = & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \mathrm{T} & \mathrm{C} & \mathrm{G} & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{A} & \mathrm{C} & \$ \\
 & 0 & 1 & 2 & 3 & & & 4 & 5 & 6 & 7 & & & 8 & 9 & 10 & 11 & & &
\end{array}
$$

Q-positions

Searching a k-factor of a given read

Read r_2 at position 0 → GAT



r_2 at position 0

Read length: 6

P-position $2 \times 6 + 0 = 12$ in R

$k = 3$

Using Gk arrays (I)

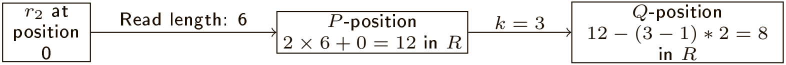
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_2 at position 0 → GAT

Position 0 in read r_2 corresponds to Q-position 8



r_2 at position 0

Read length: 6

P-position $2 \times 6 + 0 = 12$ in R

$k = 3$

Q-position $12 - (3-1) * 2 = 8$ in R

Using Gk arrays (I)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	\$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_2 at position 0 → GAT

Position 0 in read r_2 corresponds to Q-position 8

Using Gk arrays (I)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
Q-positions	0	1	2	3			4	5	6	7			8	9	10	11			

Searching a k-factor of a given read

Read r_2 at position 0 → GAT

Position 0 in read r_2 corresponds to Q-position 8

i	GkIFA
0	3
1	6
2	0
3	1
4	3
5	7
6	2
7	5
8	4
9	3
10	6
11	0

b	GkCFA	j	GkFA
0	2	0	11
		1	2
1	1	2	3
2	1	3	6
3	3	4	9
		5	0
		6	4
4	1	7	8
5	1	8	7
6	2	9	10
		10	1
7	1	11	5

Using Gk arrays (I)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_2 at position 0 → GAT

Position 0 in read r_2 corresponds to Q-position 8

i	GkIFA
0	3
1	6
2	0
3	1
4	3
5	7
6	2
7	5
8	4
9	3
10	6
11	0

b	GkCFA	j	GkFA
0	2	0	11
		1	2
1	1	2	3
2	1	3	6
3	3	4	9
		5	0
		6	4
4	1	7	8
5	1	8	7
6	2	9	10
		10	1
7	1	11	5

Using Gk arrays (I)

$$\begin{array}{r*{19}{c}} & 0&1&2&3&4&5&6&7&8&9&10&11&12&13&14&15&16&17&18 \\ R = & \mathsf{A}&\mathsf{T}&\mathsf{A}&\mathsf{A}&\mathsf{C}&\mathsf{G}&\mathsf{A}&\mathsf{T}&\mathsf{A}&\mathsf{G}&\mathsf{T}&\mathsf{C}&\mathsf{G}&\mathsf{A}&\mathsf{T}&\mathsf{A}&\mathsf{A}&\mathsf{C}&\$ \\ & 0&1&2&3&&&4&5&6&7&&&8&9&10&11&&& \end{array}$$

Q-positions

Searching a k-factor of a given read

Read r_2 at position 0 → GAT

Position 0 in read r_2 corresponds to Q-position 8

i	GkIFA
0	3
1	6
2	0
3	1
4	3
5	7
6	2
7	5
8	4
9	3
10	6
11	0

b	GkCFA	j	GkFA
0	2	0	11
		1	2
1	1	2	3
2	1	3	6
3	3	4	9
		5	0
		6	4
4	1	7	8
5	1	8	7
6	2	9	10
		10	1
7	1	11	5

Using Gk arrays (I)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
Q-positions	0	1	2	3			4	5	6	7			8	9	10	11			

Searching a k-factor of a given read

Read r_2 at position 0 → GAT

Position 0 in read r_2 corresponds to Q-position 8

i	GkIFA
0	3
1	6
2	0
3	1
4	3
5	7
6	2
7	5
8	4
9	3
10	6
11	0

b	GkCFA	j	GkFA
0	2	0	11
		1	2
1	1	2	3
2	1	3	6
3	3	4	9
		5	0
		6	4
4	1	7	8
5	1	8	7
6	2	9	10
		10	1
7	1	11	5

Information

There is only 1 P-k-factor GAT in R.

Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Using Gk arrays (II)

$$
\begin{array}{rccccccccccccccccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 \\
R = & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \mathrm{T} & \mathrm{C} & \mathrm{G} & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{A} & \mathrm{C} & \$ \\
 & 0 & 1 & 2 & 3 & & & 4 & 5 & 6 & 7 & & & 8 & 9 & 10 & 11 & & &
\end{array}
$$

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Using Gk arrays (II)

$$
\begin{array}{r@{\;}ccccccccccccccccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 \\
R = & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{G} & \mathsf{T} & \mathsf{C} & \mathsf{G} & \mathsf{A} & \mathsf{T} & \mathsf{A} & \mathsf{A} & \mathsf{C} & \$ \\
 & 0 & 1 & 2 & 3 & & & 4 & 5 & 6 & 7 & & & 8 & 9 & 10 & 11 & & &
\end{array}
$$

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

r_1 at position 0

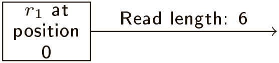
Using Gk arrays (II)

$$R = \begin{array}{ccccccccccccccccccc} 0&1&2&3&4&5&6&7&8&9&10&11&12&13&14&15&16&17&18\\ \mathrm{A}&\mathrm{T}&\mathrm{A}&\mathrm{A}&\mathrm{C}&\mathrm{G}&\mathrm{A}&\mathrm{T}&\mathrm{A}&\mathrm{G}&\mathrm{T}&\mathrm{C}&\mathrm{G}&\mathrm{A}&\mathrm{T}&\mathrm{A}&\mathrm{A}&\mathrm{C}&\$\\ 0&1&2&3&&&4&5&6&7&&&8&9&10&11&&& \end{array}$$

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA



r_1 at position 0

Read length: 6

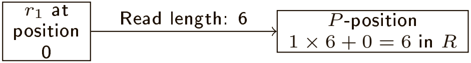
Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA



r_1 at position 0

Read length: 6

P-position

$1 \times 6 + 0 = 6$ in R

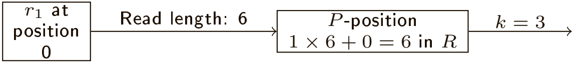
Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA



Using Gk arrays (II)

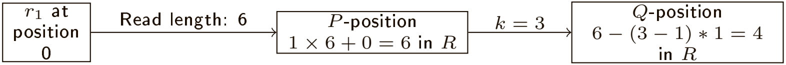
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Position 0 in read r_1 corresponds to Q-position 4



r_1 at position 0

Read length: 6

P-position $1 \times 6 + 0 = 6$ in R

$k = 3$

Q-position $6 - (3 - 1) * 1 = 4$ in R

Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Position 0 in read r_1 corresponds to Q-position 4

Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
Q-positions	0	1	2	3			4	5	6	7			8	9	10	11			

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Position 0 in read r_1 corresponds to Q-position 4

i	GkIFA
0	3
1	6
2	0
3	1
4	3
5	7
6	2
7	5
8	4
9	3
10	6
11	0

b	GkCFA	j	GkFA
0	2	0	11
		1	2
1	1	2	3
2	1	3	6
3	3	4	9
		5	0
		6	4
4	1	7	8
5	1	8	7
6	2	9	10
		10	1
7	1	11	5

Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Position 0 in read r_1 corresponds to Q-position 4

i	GkIFA
0	3
1	6
2	0
3	1
4	3
5	7
6	2
7	5
8	4
9	3
10	6
11	0

b	GkCFA	j	GkFA
0	2	0	11
		1	2
1	1	2	3
2	1	3	6
3	3	4	9
		5	0
		6	4
4	1	7	8
5	1	8	7
6	2	9	10
		10	1
7	1	11	5

Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
Q-positions	0	1	2	3			4	5	6	7			8	9	10	11			

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Position 0 in read r_1 corresponds to Q-position 4

i	GkIFA	b	GkCFA	j	GkFA
0	3	0	2	0	11
1	6			1	2
2	0	1	1	2	3
3	1	2	1	3	6
4	3	3	3	4	9
5	7			5	0
6	2			6	4
7	5	4	1	7	8
8	4	5	1	8	7
9	3	6	2	9	10
10	6			10	1
11	0	7	1	11	5

Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Position 0 in read r_1 corresponds to Q-position 4

i	GkIFA	b	GkCFA	j	GkFA
0	3	0	2	0	11
1	6			1	2
2	0	1	1	2	3
3	1	2	1	3	6
4	3	3	3	4	9
5	7			5	0
6	2			6	4
7	5	4	1	7	8
8	4	5	1	8	7
9	3	6	2	9	10
10	6			10	1
11	0	7	1	11	5

Information

There are 3 P-k-factors ATA in R.

But. . . where are they?

Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
Q-positions	0	1	2	3			4	5	6	7			8	9	10	11			

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Position 0 in read r_1 corresponds to Q-position 4

i	GkIFA
0	3
1	6
2	0
3	1
4	3
5	7
6	2
7	5
8	4
9	3
10	6
11	0

b	GkCFA	j	GkFA
0	2	0	11
		1	2
1	1	2	3
2	1	3	6
3	3	4	9
		5	0
		6	4
4	1	7	8
5	1	8	7
6	2	9	10
		10	1
7	1	11	5

Information

There are 3 P-k-factors ATA in R.

But. . . where are they?

Using Gk arrays (II)

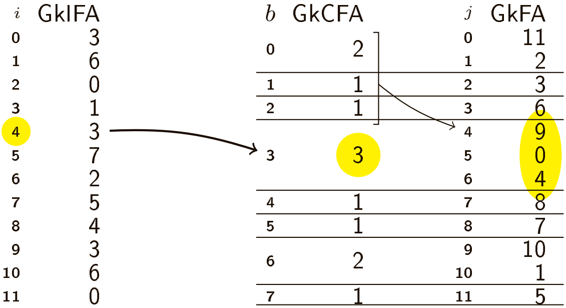
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
	0	1	2	3			4	5	6	7			8	9	10	11			

Q-positions

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Position 0 in read r_1 corresponds to Q-position 4



i	GkIFA
0	3
1	6
2	0
3	1
4	3
5	7
6	2
7	5
8	4
9	3
10	6
11	0

b	GkCFA	j	GkFA
0	2	0	11
		1	2
1	1	2	3
2	1	3	6
3	3	4	9
		5	0
		6	4
4	1	7	8
5	1	8	7
6	2	9	10
		10	1
7	1	11	5

Information

There are 3 P-k-factors ATA in R.
But. . . where are they?

Q-positions of ATA

9 r_2, position 1

0 r_0, position 0

4 r_1, position 0

Using Gk arrays (II)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$R =$	A	T	A	A	C	G	A	T	A	G	T	C	G	A	T	A	A	C	$
Q-positions	0	1	2	3			4	5	6	7			8	9	10	11			

Searching a k-factor of a given read

Read r_1 at position 0 → ATA

Position 0 in read r_1 corresponds to Q-position 4

i	GkIFA
0	3
1	6
2	0
3	1
4	3
5	7
6	2
7	5
8	4
9	3
10	6
11	0

b	GkCFA	GkCFPS	j	GkFA
0	2	2	0	11
			1	2
1	1	3	2	3
2	1	4	3	6
3	3	7	4	9
			5	0
			6	4
4	1	8	7	8
5	1	9	8	7
6	2	11	9	10
			10	1
7	1	12	11	5

Information

There are 3 P-k-factors ATA in R.
But. . . where are they?

Q-positions of ATA

9 r_2, position 1

0 r_0, position 0

4 r_1, position 0

Multiplicity of P-k-factors

Problem

What if a P-k-factor occurs many times in the same read?

Multiplicity of P-k-factors

Problem

What if a P-k-factor occurs many times in the same read?

$\Rightarrow$ The number of reads in which occurs a P-k-factor is not its total number of occurrences in R

Multiplicity of P-k-factors

Problem

What if a P-k-factor occurs many times in the same read?

$\Rightarrow$ The number of reads in which occurs a P-k-factor is not its total number of occurrences in R

(Modified) Example

GkCFA	i	GkFA
	$\vdots$	$\vdots$
3	4	11
	5	0
	6	4
	7	9
	$\vdots$	$\vdots$

Multiplicity of P-k-factors

Problem

What if a P-k-factor occurs many times in the same read?

$\Rightarrow$ The number of reads in which occurs a P-k-factor is not its total number of occurrences in R

(Modified) Example

GkCFA	i	GkFA	
	$\vdots$	$\vdots$	
3	4	11	$\longrightarrow$ read 2
	5	0	$\longrightarrow$ read 0
	6	4	$\longrightarrow$ read 1
	7	9	$\longrightarrow$ read 2
	$\vdots$	$\vdots$	

Multiplicity of P-k-factors

Problem

What if a P-k-factor occurs many times in the same read?

$\Rightarrow$ The number of reads in which occurs a P-k-factor is not its total number of occurrences in R

(Modified) Example

GkCFA	i	GkFA	
	⋮	⋮	
3	4	11	$\longrightarrow$ read 2
	5	0	$\longrightarrow$ read 0
	6	4	$\longrightarrow$ read 1
	7	9	$\longrightarrow$ read 2
	⋮	⋮	

Solutions for counting reads

- Use a mask to known which read have already been counted

Multiplicity of P-k-factors

Problem

What if a P-k-factor occurs many times in the same read?

$\Rightarrow$ The number of reads in which occurs a P-k-factor is not its total number of occurrences in R

(Modified) Example

GkCFA	i	GkFA	
	$\vdots$	$\vdots$	
3	4	11	$\longrightarrow$ read 2
	5	0	$\longrightarrow$ read 0
	6	4	$\longrightarrow$ read 1
	7	9	$\longrightarrow$ read 2
	$\vdots$	$\vdots$	

Solutions for counting reads

- Use a mask to known which read have already been counted
- Sort the entries (when querying or at construction).

Multiplicity of P-k-factors

Problem

What if a P-k-factor occurs many times in the same read?

$\Rightarrow$ The number of reads in which occurs a P-k-factor is not its total number of occurrences in R

(Modified) Example

GkCFA	i	GkFA	
	⋮	⋮	
3	4	0	$\longrightarrow$ read 0
	5	4	$\longrightarrow$ read 1
	6	9	$\longrightarrow$ read 2
	7	11	$\longrightarrow$ read 2
	⋮	⋮	

Solutions for counting reads

- Use a mask to known which read have already been counted
- Sort the entries (when querying or at construction).

Complexities

Space complexities

GkFA, GkIFA	Number of entries: Number of reads × (Read length $-k+1$)
GkCFA	Number of entries: Number of distinct P-k-factors

Complexities

Space complexities

GkFA, GkIFA	Number of entries: Number of reads × (Read length $-k+1$)
GkCFA	Number of entries: Number of distinct P-k-factors

Time complexities

Q1	(counting P-k-factors)	$O(1)$
Q2	(retrieving positions in reads)	$O(occ)$
Q3	(counting reads)	$O(occ)$

where occ is the number of occurrences of the P-k-factor in the reads.

Complexities with the classical solution

SA-based solution

Build the suffix array of R, the inverse suffix array and the LCP array.

Complexities with the classical solution

SA-based solution

Build the suffix array of R, the inverse suffix array and the LCP array.

Space complexities

Three arrays containing (number of reads × length of the reads) elements each

Complexities with the classical solution

SA-based solution

Build the suffix array of R, the inverse suffix array and the LCP array.

Space complexities

Three arrays containing (number of reads × length of the reads) elements each

Time complexities

Q1	(counting P-k-factors)	$O(occ_R)$
Q2	(retrieving positions in reads)	$O(occ_R)$
Q3	(counting reads)	$O(occ_R+$number of reads$)$

where occ_R is the number of occurrences of the k-factors in R.

Complexities with the classical solution

SA-based solution

Build the suffix array of R, the inverse suffix array and the LCP array.

Space complexities

Three arrays containing (number of reads × length of the reads) elements each

Time complexities

Q1	(counting P-k-factors)	$O(occ_R)$
Q2	(retrieving positions in reads)	$O(occ_R)$
Q3	(counting reads)	$O(occ_R+$number of reads$)$

where occ_R is the number of occurrences of the k-factors in R.

Improvements over a SA-based solution

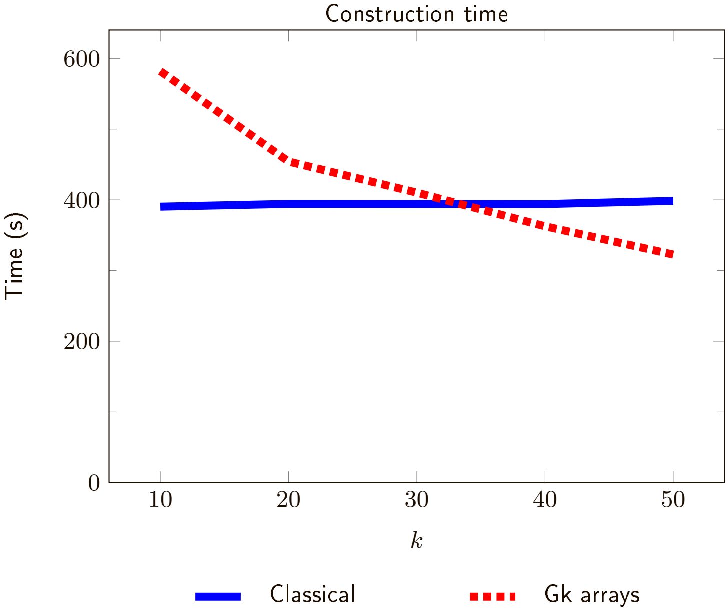
Space	At least $(3 \times (k-1)\times$ number of reads$)$ elements
Time	No dependency on the number of reads, no dependency on the number of occurrences in R

Time and space construction in practice

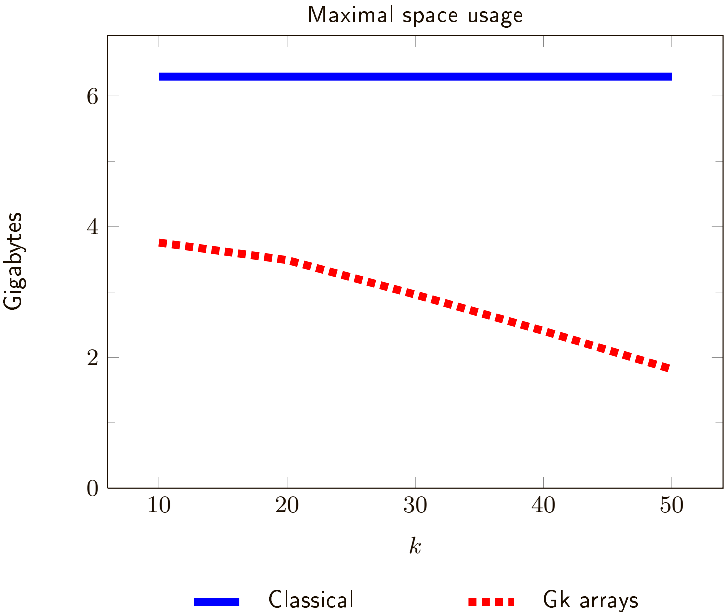
Data

- Fruit fly sequences from a Genome Analyzer II
- 7,000,000 reads
- read length: 75

Time and space construction in practice

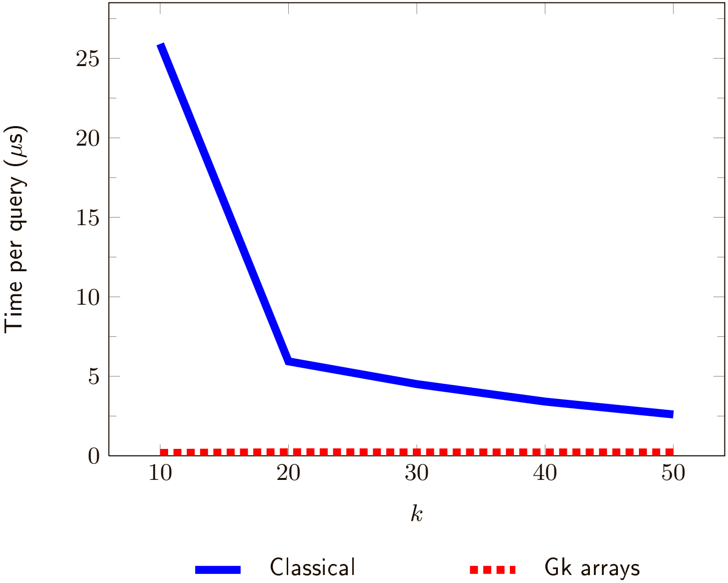


Time and space construction in practice



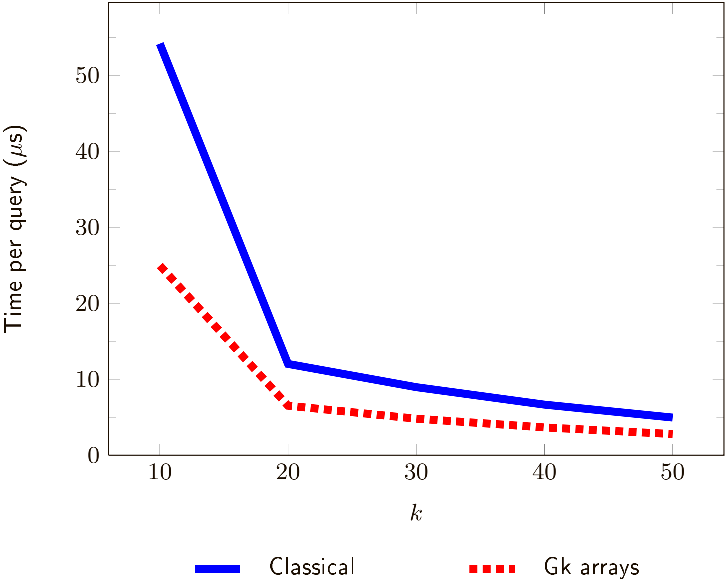
Query time

What is the number of occurrences of f in the reads?

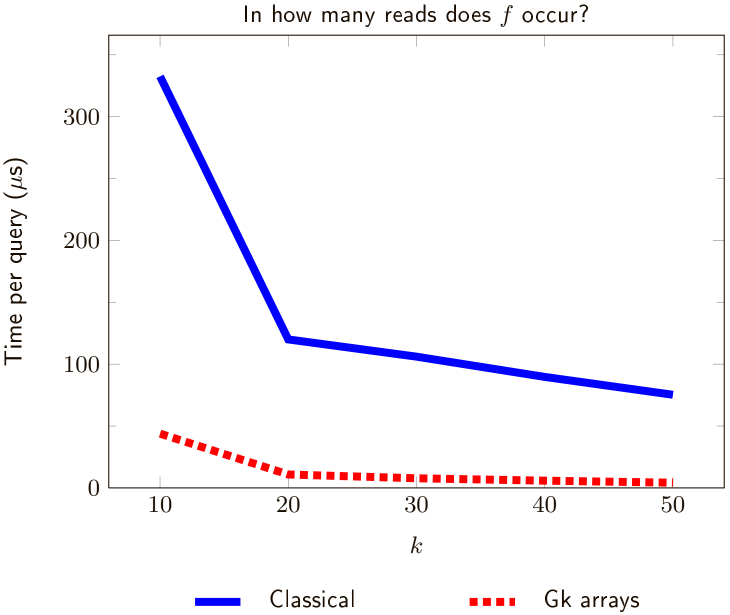


Query time

What are the occurrence positions of f in the reads?



Query time


In how many reads does f occur?
Time per query (μs)
300
200
100
0
10
20
30
40
50
k
Classical
Gk arrays

Conclusions and Perspectives

Efficiency

Gk arrays allow to query more reads in less time than a suffix array-based method

Conclusions and Perspectives

Efficiency

Gk arrays allow to query more reads in less time than a suffix array-based method

Variable-length reads

We can deal with variable-length reads by adding a bit vector for identifying the end of reads in R

Conclusions and Perspectives

Efficiency

Gk arrays allow to query more reads in less time than a suffix array-based method

Variable-length reads

We can deal with variable-length reads by adding a bit vector for identifying the end of reads in R

Compressing Gk arrays

Can we adapt compression techniques to Gk arrays?

$\rightarrow$ new space/time tradeoff

Conclusions and Perspectives

Efficiency

Gk arrays allow to query more reads in less time than a suffix array-based method

Variable-length reads

We can deal with variable-length reads by adding a bit vector for identifying the end of reads in R

Compressing Gk arrays

Can we adapt compression techniques to Gk arrays?

→ new space/time tradeoff

Updating Gk arrays

Can we efficiently update Gk arrays?

→ read correction