Indexer des fragments d'ADN

En collaboration avec:
N. Philippe, T. Commes, É. Rivals
M. Léonard, T. Lecroq

Journées au vert

21 juin 2011

Introduction

Contexte

- Séquenceurs haut-débit
- Des giga-octets de données par jour

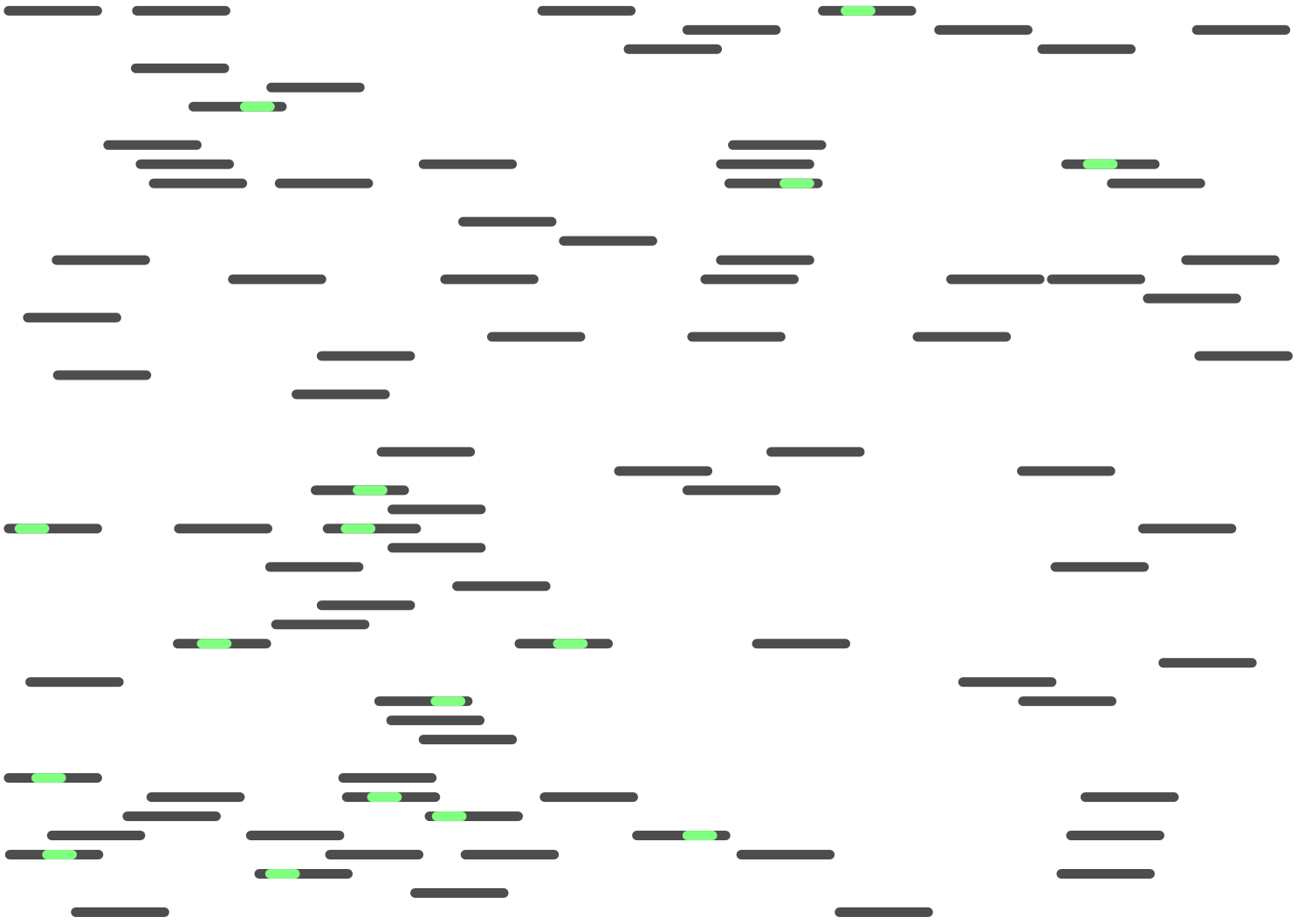
Questions

- Comment tirer parti de tous les fragments produits ?
- Comment les organiser de façon intelligente ?
- Comment ne pas saturer nos machines (portables, ordinateurs de bureau, serveurs, . . .) ?

Des points communs. . .

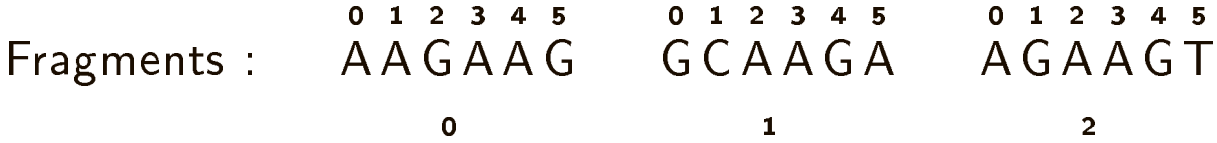
Des points communs. . .

Des points communs. . .



Des points communs. . .

Un annuaire de k-mers



Un annuaire de k-mers

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
	0					

	0	1	2	3	4	5
	G	C	A	A	G	A
	1					

	0	1	2	3	4	5
	A	G	A	A	G	T
	2					

Annuaire (ou index. . .)

		F	P
0	AAG	0	0
1	AAG	0	3
2	AAG	1	2
3	AAG	2	2
4	AGT	2	3
5	AGA	0	1
6	AGA	1	3
7	AGA	2	0
8	CAA	1	1
9	GAA	0	2
10	GAA	2	1
11	GCA	1	0

Un annuaire de k-mers

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
	0					

	0	1	2	3	4	5
	G	C	A	A	G	A
	1					

	0	1	2	3	4	5
	A	G	A	A	G	T
	2					

Annuaire (ou index. . .)

		F	P
0	AAG	0	0
1	AAG	0	3
2	AAG	1	2
3	AAG	2	2
4	AGT	2	3
5	AGA	0	1
6	AGA	1	3
7	AGA	2	0
8	CAA	1	1
9	GAA	0	2
10	GAA	2	1
11	GCA	1	0

Problème

Comment rendre notre annuaire plus « puissant » et plus « efficace » ?

Un annuaire de k-mers

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
			0			

	0	1	2	3	4	5
	G	C	A	A	G	A
			1			

	0	1	2	3	4	5
	A	G	A	A	G	T
			2			

Annuaire (ou index. . .)

		F	P
0	AAG	0	0
1	AAG	0	3
2	AAG	1	2
3	AAG	2	2
4	AGT	2	3
5	AGA	0	1
6	AGA	1	3
7	AGA	2	0
8	CAA	1	1
9	GAA	0	2
10	GAA	2	1
11	GCA	1	0

Problème

Comment rendre notre annuaire plus « puissant » et plus « efficace » ?

Plus puissant

En ayant l'adresse d'un k-mer je veux retrouver l'adresse de ses homonymes

Un annuaire de k-mers

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
	0					

	0	1	2	3	4	5
	G	C	A	A	G	A
	1					

	0	1	2	3	4	5
	A	G	A	A	G	T
	2					

Annuaire (ou index. . .)

		F	P
0	AAG	0	0
1	AAG	0	3
2	AAG	1	2
3	AAG	2	2
4	AGT	2	3
5	AGA	0	1
6	AGA	1	3
7	AGA	2	0
8	CAA	1	1
9	GAA	0	2
10	GAA	2	1
11	GCA	1	0

Problème

Comment rendre notre annuaire plus « puissant » et plus « efficace » ?

Plus puissant

En ayant l'adresse d'un k-mer je veux retrouver l'adresse de ses homonymes

Plus efficace

Éliminer les informations redondantes

Retrouver l'adresse des homonymes

	0	1	2	3	4	5
Fragments :	A	A	G	A	A	G
	0					

	0	1	2	3	4	5
	G	C	A	A	G	A
	1					

	0	1	2	3	4	5
	A	G	A	A	G	T
	2					

Annuaire (ou index. . .)

		F P
0	AAG	0 0
1	AAG	0 3
2	AAG	1 2
3	AAG	2 2
4	AGT	2 3
5	AGA	0 1
6	AGA	1 3
7	AGA	2 0
8	CAA	1 1
9	GAA	0 2
10	GAA	2 1
11	GCA	1 0

Retrouver l'adresse des homonymes

	0	1	2	3	4	5		0	1	2	3	4	5		0	1	2	3	4	5
Fragments :	A	A	G	A	A	G		G	C	A	A	G	A		A	G	A	A	G	T
			0							1							2			

Annuaire (ou index. . .)

		F	P	
0	AAG	0	0	4
1	AAG	0	3	
2	AAG	1	2	
3	AAG	2	2	
4	AGT	2	3	1
5	AGA	0	1	3
6	AGA	1	3	
7	AGA	2	0	
8	CAA	1	1	1
9	GAA	0	2	2
10	GAA	2	1	
11	GCA	1	0	1

Retrouver l'adresse des homonymes

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
	0					

	0	1	2	3	4	5
	G	C	A	A	G	A
	1					

	0	1	2	3	4	5
	A	G	A	A	G	T
	2					

Annuaire (ou index. . .)

		F P		
0	AAG	0 0	0	4
1	AAG	0 3		
2	AAG	1 2		
3	AAG	2 2		
4	AGT	2 3	1	1
5	AGA	0 1	2	3
6	AGA	1 3		
7	AGA	2 0		
8	CAA	1 1	3	1
9	GAA	0 2	4	2
10	GAA	2 1		
11	GCA	1 0	5	1

Retrouver l'adresse des homonymes

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
0						

	0	1	2	3	4	5
	G	C	A	A	G	A
1						

	0	1	2	3	4	5
	A	G	A	A	G	T
2						

Annuaire (ou index...)

		F	P		
0	AAG	0	0	0	4
1	AAG	0	3		
2	AAG	1	2		
3	AAG	2	2		
4	AGT	2	3	1	1
5	AGA	0	1	2	3
6	AGA	1	3		
7	AGA	2	0		
8	CAA	1	1	3	1
9	GAA	0	2	4	2
10	GAA	2	1		
11	GCA	1	0	5	1

Annuaire inverse

F	P
0	0
0	1
0	2
0	3
1	0
1	1
1	2
1	3
2	0
2	1
2	2
2	3

Retrouver l'adresse des homonymes

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
0						

	0	1	2	3	4	5
	G	C	A	A	G	A
1						

	0	1	2	3	4	5
	A	G	A	A	G	T
2						

Annuaire (ou index. . .)

		F	P		
0	AAG	0	0	0	4
1	AAG	0	3		
2	AAG	1	2		
3	AAG	2	2		
4	AGT	2	3	1	1
5	AGA	0	1	2	3
6	AGA	1	3		
7	AGA	2	0		
8	CAA	1	1	3	1
9	GAA	0	2	4	2
10	GAA	2	1		
11	GCA	1	0	5	1

Annuaire inverse

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Retrouver l'adresse des homonymes

Fragments :

	0	1	2	3	4	5
0	A	A	G	A	A	G
1	G	C	A	A	G	A
2	A	G	A	A	G	T

Annuaire (ou index...)

		F	P		
0	AAG	0	0		
1	AAG	0	3	0	4
2	AAG	1	2		
3	AAG	2	2		
4	AGT	2	3	1	1
5	AGA	0	1		
6	AGA	1	3	2	3
7	AGA	2	0		
8	CAA	1	1	3	1
9	GAA	0	2	4	2
10	GAA	2	1		
11	GCA	1	0	5	1

Annuaire inverse

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Retrouver l'adresse des homonymes

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
0						

	0	1	2	3	4	5
	G	C	A	A	G	A
1						

	0	1	2	3	4	5
	A	G	A	A	G	T
2						

Annuaire (ou index. . .)

		F	P		
0	AAG	0	0	0	4
1	AAG	0	3		
2	AAG	1	2		
3	AAG	2	2		
4	AGT	2	3	1	1
5	AGA	0	1	2	3
6	AGA	1	3		
7	AGA	2	0		
8	CAA	1	1	3	1
9	GAA	0	2	4	2
10	GAA	2	1		
11	GCA	1	0	5	1

Annuaire inverse

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Retrouver l'adresse des homonymes

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
0						

	0	1	2	3	4	5
	G	C	A	A	G	A
1						

	0	1	2	3	4	5
	A	G	A	A	G	T
2						

Annuaire (ou index...)

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

Annuaire inverse

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Des requêtes intéressantes. . .

Fragments :

	0	1	2	3	4	5
0	A	A	G	A	A	G
1	G	C	A	A	G	A
2	A	G	A	A	G	T

Annuaires (ou index. . .)

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

0	4
1	1
2	3
3	1
4	2
5	1

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Des requêtes intéressantes...

Fragments :

	0	1	2	3	4	5
0	A	A	G	A	A	G
1	G	C	A	A	G	A
2	A	G	A	A	G	T

Annuaires (ou index...)

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

0	4
1	1
2	3
3	1
4	2
5	1

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Requêtes

1. Combien de fois voit-on le même k-mer qu'en 0 2 ?

Des requêtes intéressantes...

Fragments :

	0	1	2	3	4	5
0	A	A	G	A	A	G
1	G	C	A	A	G	A
2	A	G	A	A	G	T

Annuaires (ou index...)

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

0	4
1	1
2	3
3	1
4	2
5	1

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Requêtes

1. Combien de fois voit-on le même k-mer qu'en 0 2 ?

Des requêtes intéressantes...

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
0						

	0	1	2	3	4	5
	G	C	A	A	G	A
1						

	0	1	2	3	4	5
	A	G	A	A	G	T
2						

Annuaires (ou index...)

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

0	4	
1	1	
2	3	
3	1	
4	2	
5	1	

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Requêtes

1. Combien de fois voit-on le même k-mer qu'en 0 2 ?

Des requêtes intéressantes...

Fragments :

	0 1 2 3 4 5		0 1 2 3 4 5		0 1 2 3 4 5
	A A G A A G		G C A A G A		A G A A G T
	0		1		2

Annuaires (ou index...)

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

0	4
1	1
2	3
3	1
4	2
5	1

F	P
0 0	0
0 1	2
0 2	4
0 3	0
1 0	5
1 1	3
1 2	0
1 3	2
2 0	2
2 1	4
2 2	0
2 3	1

Requêtes

1. Combien de fois voit-on le même k-mer qu'en 0 2 ? $O(1)$

Des requêtes intéressantes...

Fragments :

	0	1	2	3	4	5
0	A	A	G	A	A	G
1	G	C	A	A	G	A
2	A	G	A	A	G	T

Annuaires (ou index...)

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

0	4
1	1
2	3
3	1
4	2
5	1

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Requêtes

1. Combien de fois voit-on le même k-mer qu'en 0 2 ? $O(1)$
2. À quels endroits trouve-t-on ce même k-mer ?

Des requêtes intéressantes...

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
0						

	0	1	2	3	4	5
	G	C	A	A	G	A
1						

	0	1	2	3	4	5
	A	G	A	A	G	T
2						

Annuaires (ou index...)

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

0	4
1	1
2	3
3	1
4	2
5	1

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Requêtes

1. Combien de fois voit-on le même k-mer qu'en 0 2 ? $O(1)$
2. À quels endroits trouve-t-on ce même k-mer ?

Des requêtes intéressantes...

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
0						

	0	1	2	3	4	5
	G	C	A	A	G	A
1						

	0	1	2	3	4	5
	A	G	A	A	G	T
2						

Annuaires (ou index...)

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

0	4	
1	1	
2	3	
3	1	
4	2	
5	1	

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Requêtes

1. Combien de fois voit-on le même k-mer qu'en 0 2 ? $O(1)$
2. À quels endroits trouve-t-on ce même k-mer ? $O(\text{occ})$

Des requêtes intéressantes...

Fragments :

	0	1	2	3	4	5
0	A	A	G	A	A	G
1	G	C	A	A	G	A
2	A	G	A	A	G	T

Annuaires (ou index...)

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

0	4	
1	1	
2	3	
3	1	
4	2	
5	1	

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Requêtes

1. Combien de fois voit-on le même k-mer qu'en 0 2 ? $O(1)$
2. À quels endroits trouve-t-on ce même k-mer ? $O(\text{occ})$

Variantes

1. Combien de fragments partagent un même k-mer ?
2. Quels fragments contiennent ce k-mer ?

Des requêtes intéressantes...

Fragments :

	0	1	2	3	4	5
	A	A	G	A	A	G
0						

	0	1	2	3	4	5
	G	C	A	A	G	A
1						

	0	1	2	3	4	5
	A	G	A	A	G	T
2						

Annuaires (ou index...)

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

0	4	
1	1	
2	3	
3	1	
4	2	
5	1	

F	P	
0	0	0
0	1	2
0	2	4
0	3	0
1	0	5
1	1	3
1	2	0
1	3	2
2	0	2
2	1	4
2	2	0
2	3	1

Requêtes

1. Combien de fois voit-on le même k-mer qu'en 0 2 ? $O(1)$
2. À quels endroits trouve-t-on ce même k-mer ? $O(\text{occ})$

Variantes

1. Combien de fragments partagent un même k-mer ? $O(\text{occ})$
2. Quels fragments contiennent ce k-mer ? $O(\text{occ})$

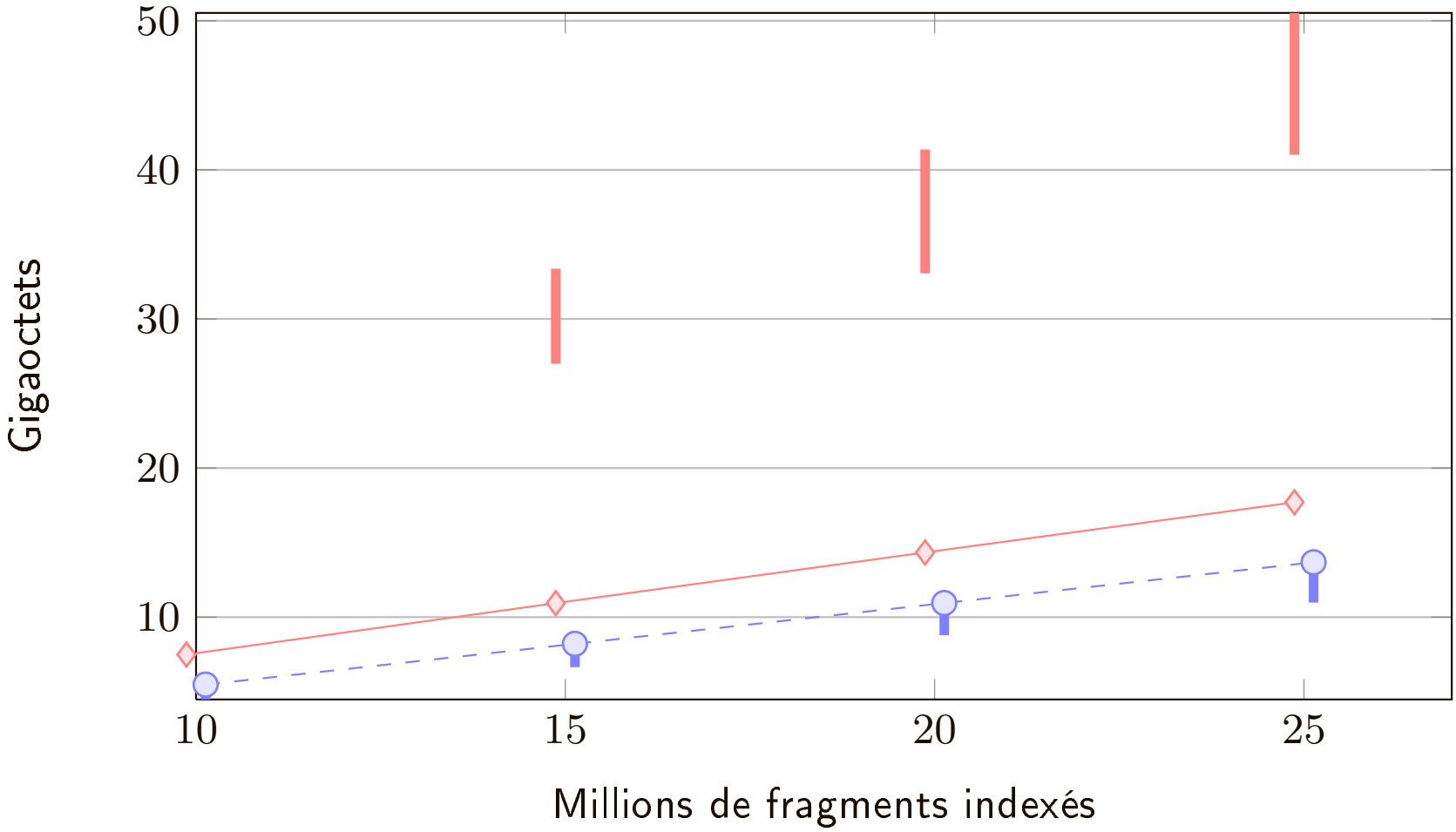
Quelques expériences

Ce qui nous intéresse

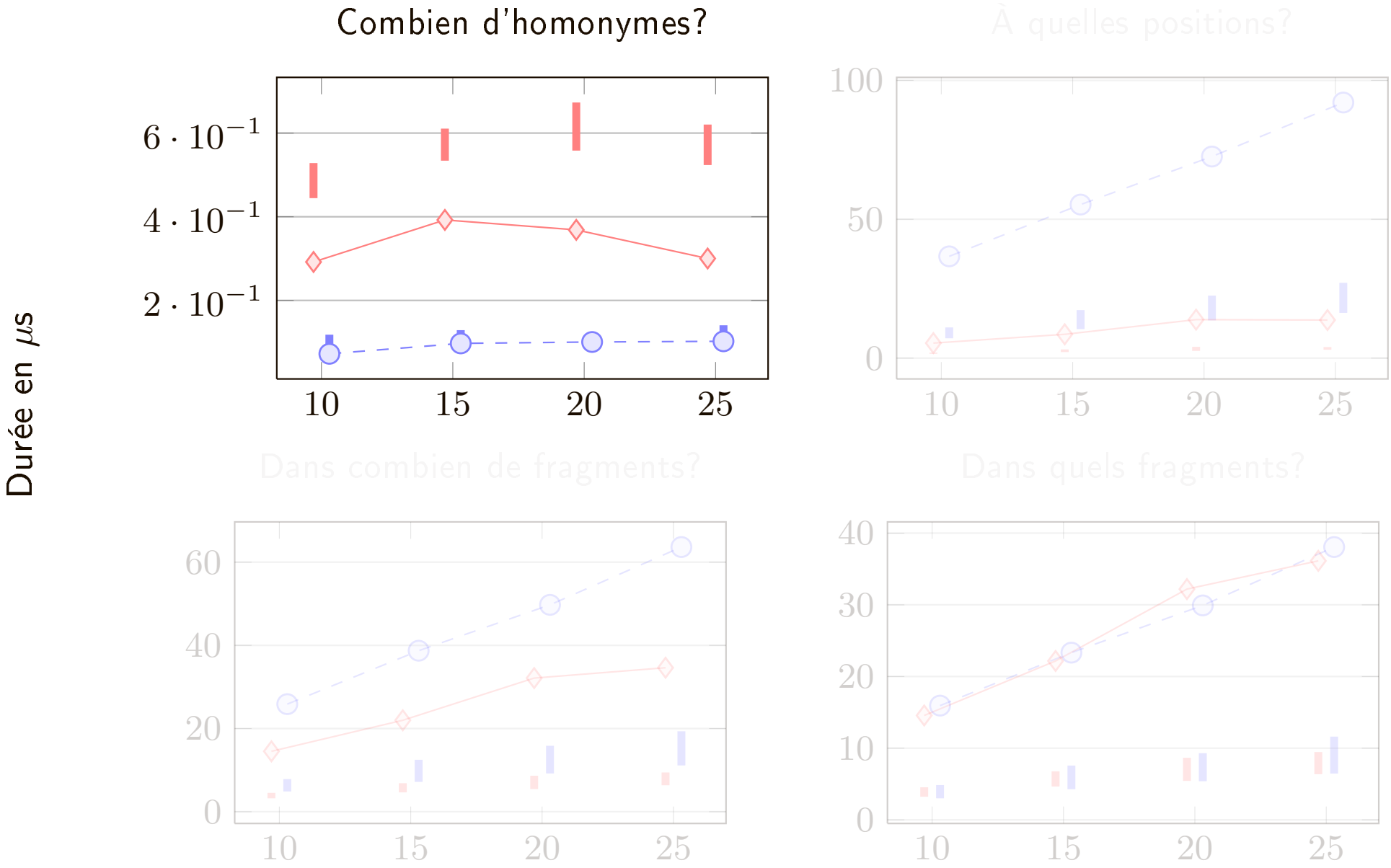
- Combien ça pèse ?
- Combien de temps ça prend de faire une requête ?
- Est-ce mieux qu'une solution alternative (table de hâchage) ?

Quel espace?

Indexation de fragments de longueur 75, pour k allant de 10 à 30

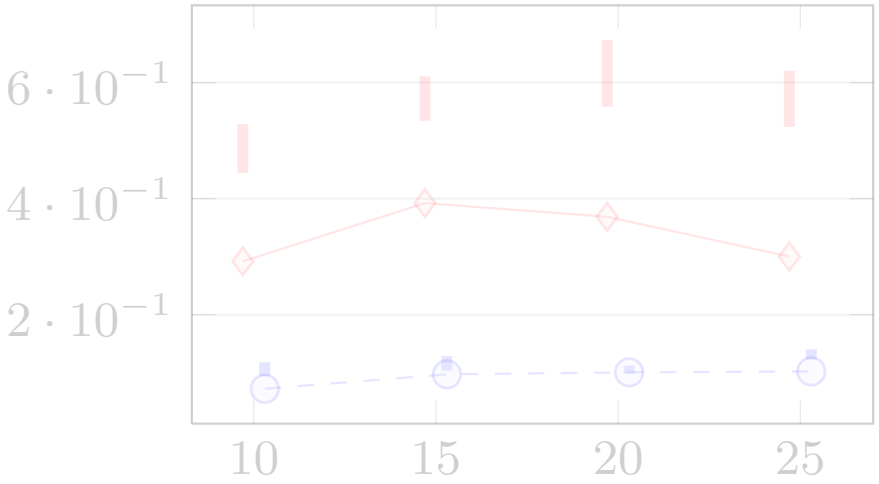


Quelle rapidité?

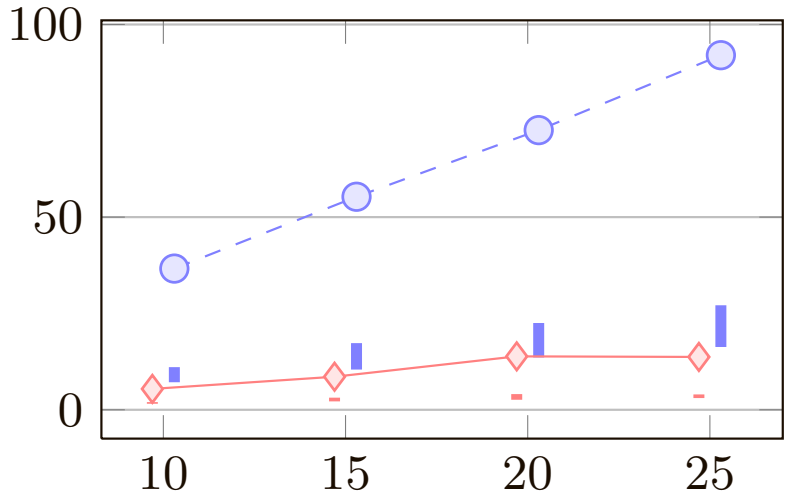


Quelle rapidité?

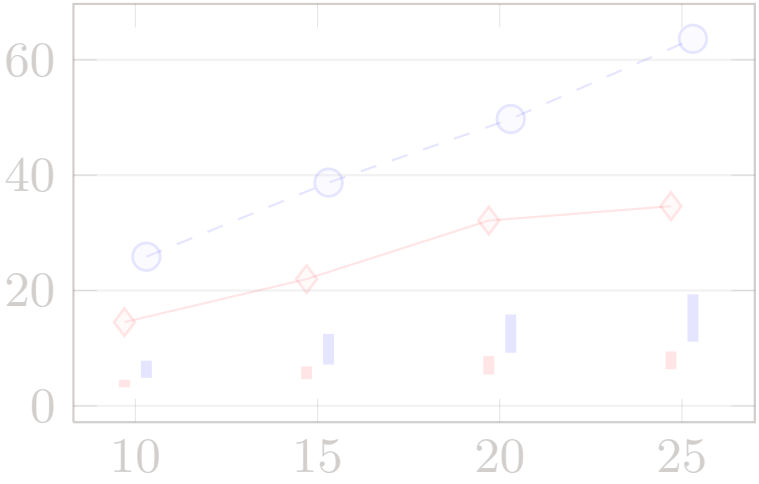
Combien d'homonymes?



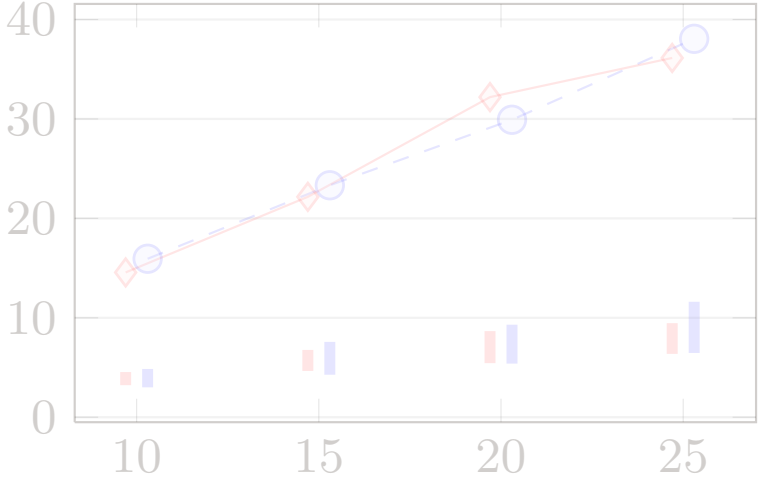
À quelles positions?



Dans combien de fragments?



Dans quels fragments?

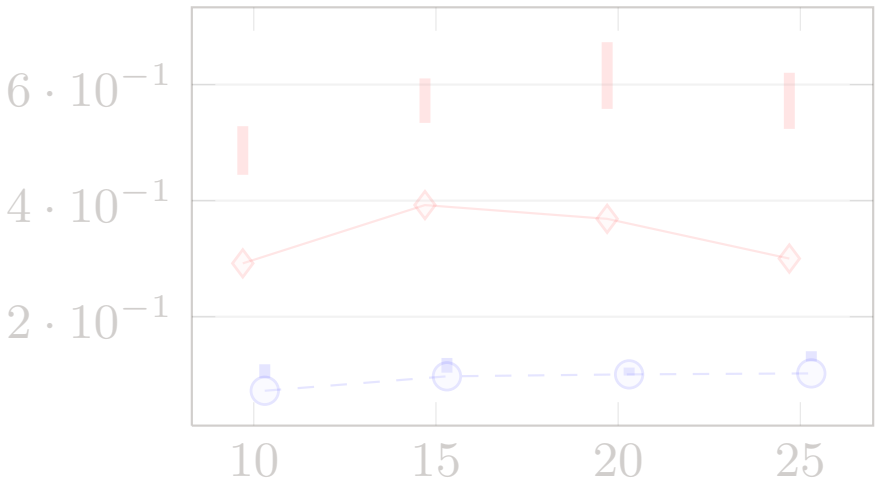


Durée en μs

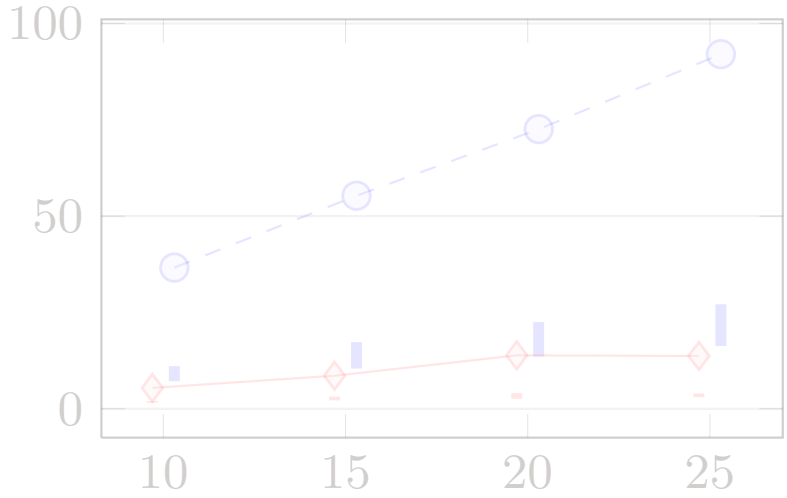
Millions de fragments indexés

Quelle rapidité?

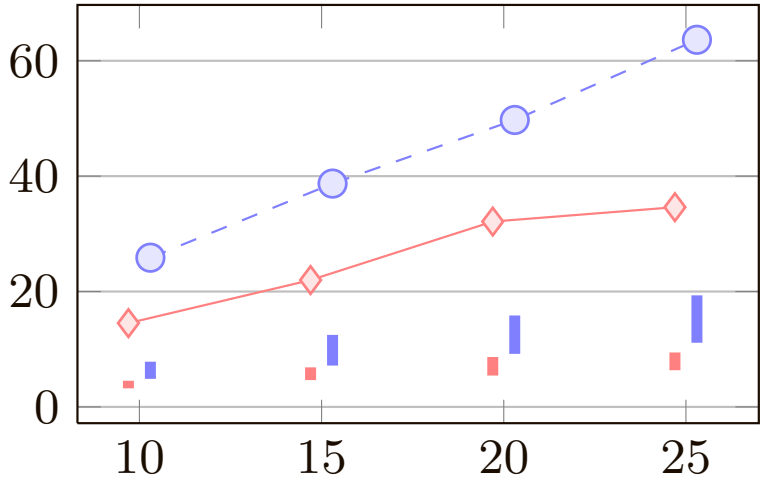
Combien d'homonymes?



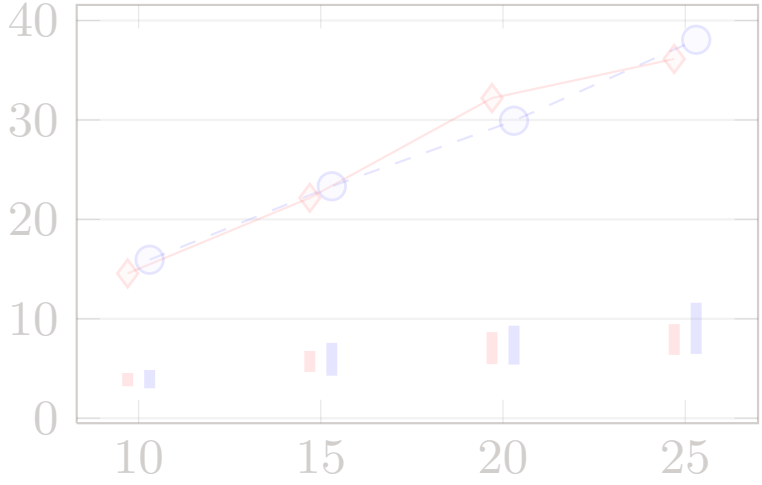
À quelles positions?



Dans combien de fragments?



Dans quels fragments?



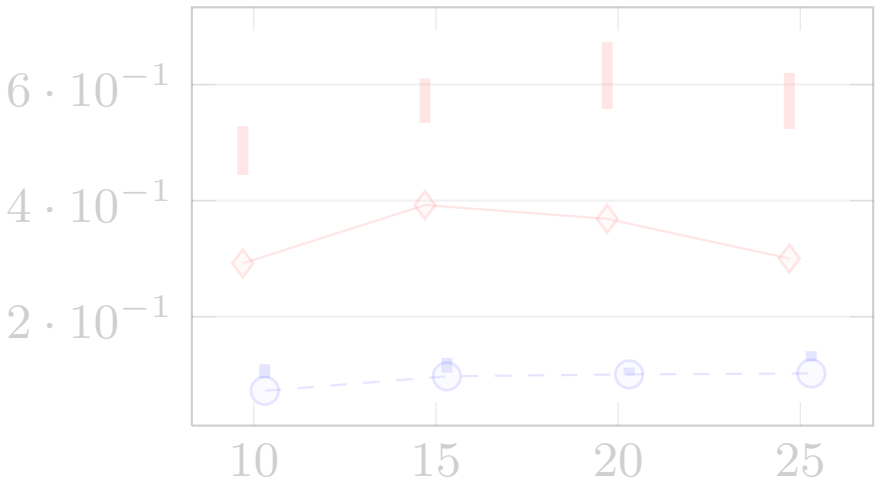
Durée en μs

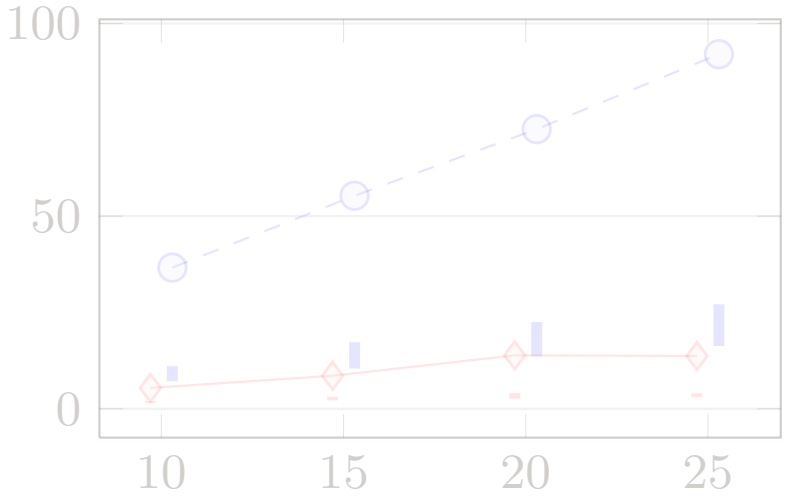
Millions de fragments indexés

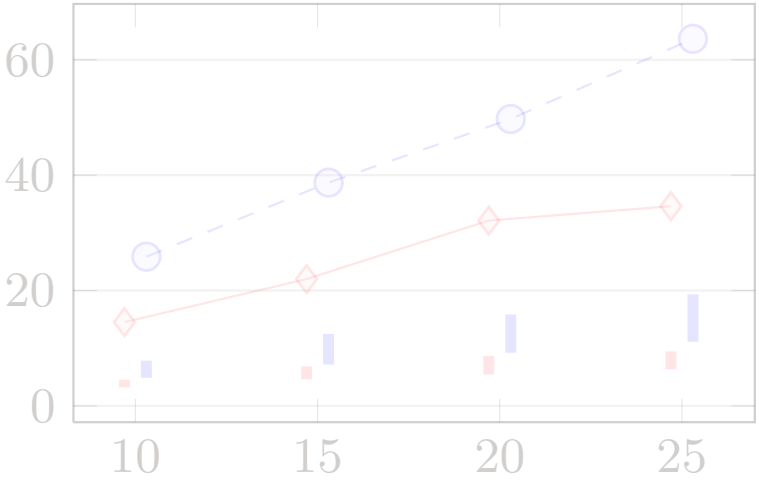
Quelle rapidité?

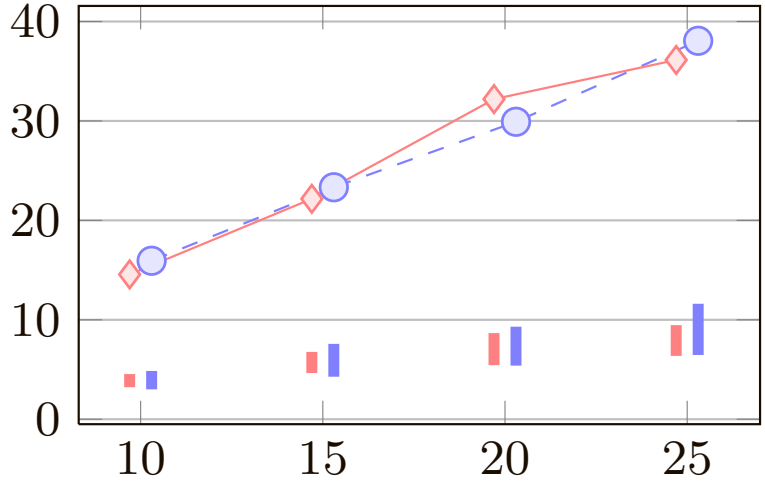
Durée en μs

Millions de fragments indexés

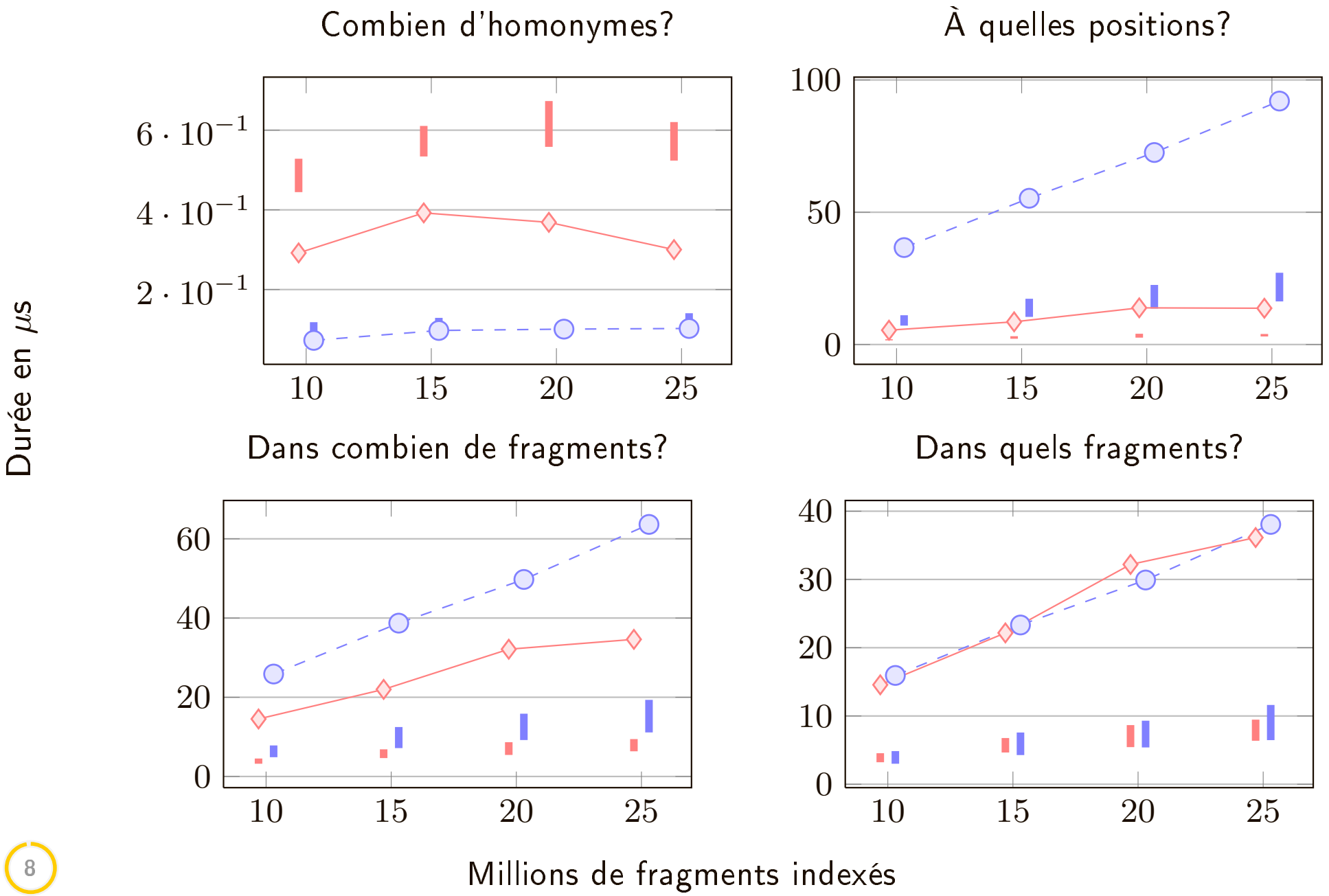




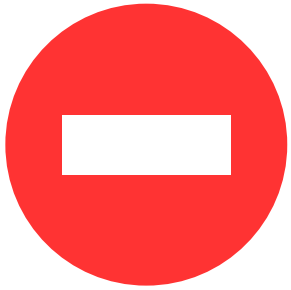




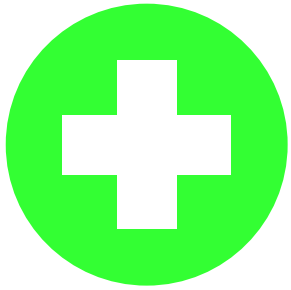
Quelle rapidité?



Conclusions



- Mauvaises performances avec un grand nombre d'occurrences (k petit)
- Espace peu dépendant de k



- Prend moins de place que les tables de hâchage
- Espace peu dépendant de k
- Requêtes à peu près aussi rapides qu'une table de hâchage
- Peut gérer des reads de longueur variable