

Méthodes sans alignement

École (JC)2BIM

Mikaël Salson

`mikael.salson@univ-lille.fr`



Les méthodes sans-alignement

Quantification d'ARN

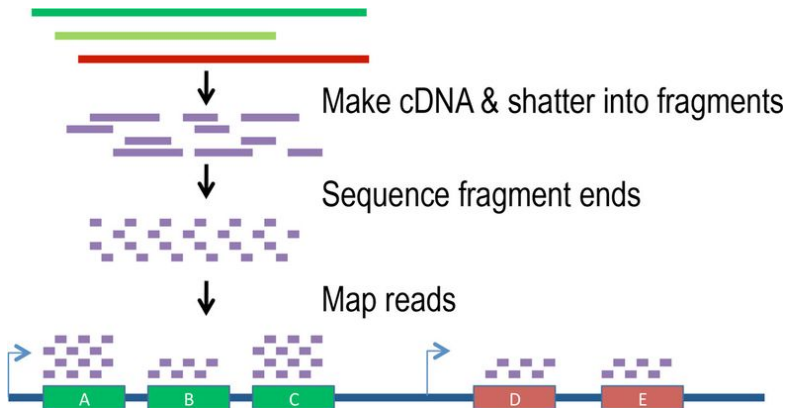
Kallisto, Salmon, RapMap

Données de 3^e génération

Sans parler métagénomique, phylogénie, analyses RNA-Seq

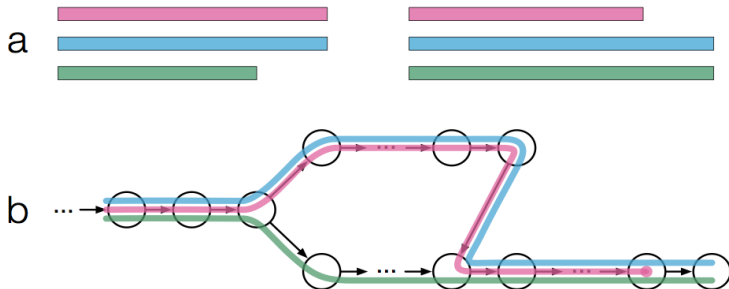
MHAP, Minimap2

La quantification de gènes au temps de l'alignement



Kallisto : approche par « pseudo-alignement »

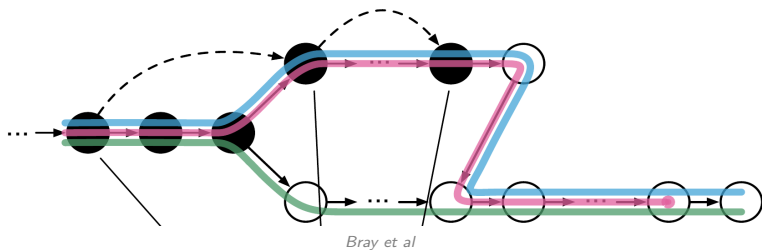
Bray et al, 2016 (sur arXiv en 2015)



Bray et al

Kallisto : approche par « pseudo-alignement »

Bray et al, 2016 (sur arXiv en 2015)



Salmon : approche par « alignement léger »

Patro *et al*, 2015

Utilisent les Super Maximal Exact Matches

Salmon : approche par « alignement léger »

Patro *et al*, 2015

Utilisent les Super Maximal Exact Matches

read ...ATAGCTCGTA...

transcriptome ...CGCTAGACTGCTCTAGTAC...

Salmon : approche par « alignement léger »

Patro et al, 2015

Utilisent les Super Maximal Exact Matches

read

...ATAGCTCGTA...

Maximal exact match

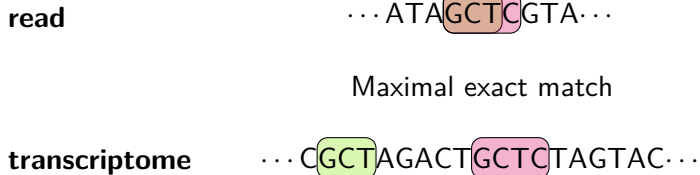
transcriptome

...CGCTAGACTGCTCTAGTAC...

Salmon : approche par « alignement léger »

Patro *et al*, 2015

Utilisent les Super Maximal Exact Matches



Salmon : approche par « alignement léger »

Patro et al, 2015

Utilisent les Super Maximal Exact Matches

read

...ATAGCTCGTA...

Super maximal exact match

transcriptome

...CGCTAGACTGCTCTAGTAC...

Salmon : approche par « alignement léger »

Patro et al, 2015

Utilisent les Super Maximal Exact Matches

read

...ATAGCTCGTA...

Super maximal exact match

transcriptome

...CGCTAGACTGCTCTAGTAC...

Salmon : approche par « alignement léger »

Patro et al, 2015

Utilisent les Super Maximal Exact Matches

read

...ATAGCTCGTA...

Super maximal exact match

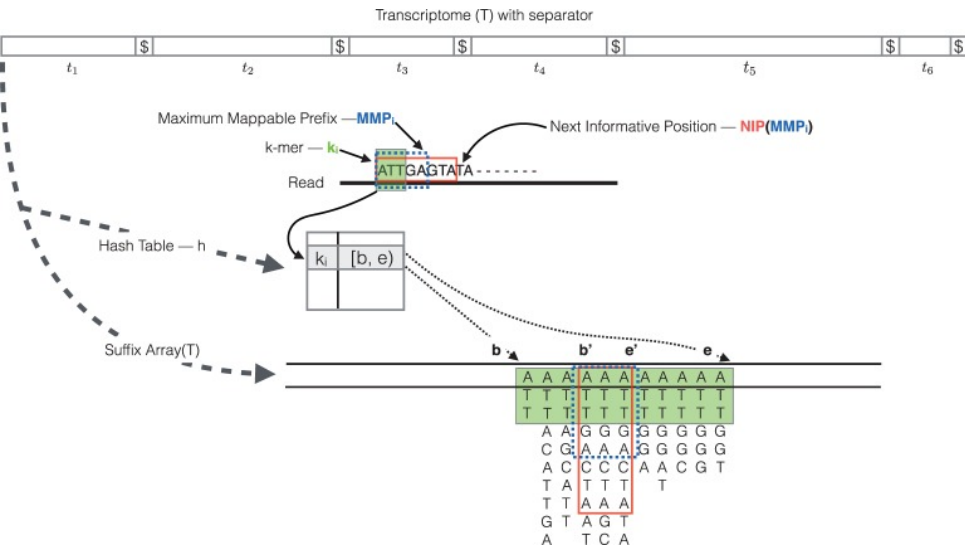
transcriptome

...CGCTAGACTGCTCTAGTAC...

Conservent les chaînes de SMEMs cohérents entre eux (colinéaires)

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017)  sur bioRxiv en 2016 



RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche A T T G A G T A T A C G A ...

Table des suffixes

i	A T T A C A A C G C ...
$i + 1$	A T T A C A T T T G ...
$i + 2$	A T T A G C A T T A ...
$i + 3$	A T T G A C T A A T ...
$i + 4$	A T T G A C T A G C ...
$i + 5$	A T T G A C T A T A ...
$i + 6$	A T T G G A C G T A ...
$i + 7$	A T T G G A T A C A ...
$i + 8$	A T T G G C A A T C ...
$i + 9$	A T T G G G A A C T ...
$i + 10$	A T T G G T C A G A ...

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche **ATT** G A G T A T A C G A ...
k-mer

Table des suffixes

<i>i</i>	A T T A C A A C G C ...
<i>i</i> + 1	A T T A C A T T T G ...
<i>i</i> + 2	A T T A G C A T T A ...
<i>i</i> + 3	A T T G A C T A A T ...
<i>i</i> + 4	A T T G A C T A G C ...
<i>i</i> + 5	A T T G A C T A T A ...
<i>i</i> + 6	A T T G G A C G T A ...
<i>i</i> + 7	A T T G G A T A C A ...
<i>i</i> + 8	A T T G G C A A T C ...
<i>i</i> + 9	A T T G G G A A C T ...
<i>i</i> + 10	A T T G G T C A G A ...

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

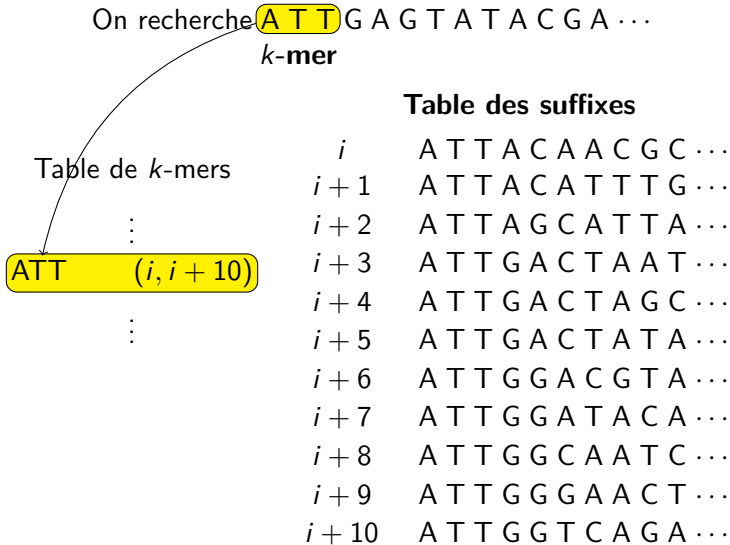
On recherche **ATT** G A G T A T A C G A ...
k-mer

Table des suffixes

Table de <i>k</i> -mers ATT (<i>i</i> , <i>i</i> + 10)	<i>i</i>	A T T A C A A C G C ...
	<i>i</i> + 1	A T T A C A T T T G ...
	<i>i</i> + 2	A T T A G C A T T A ...
	<i>i</i> + 3	A T T G A C T A A T ...
	<i>i</i> + 4	A T T G A C T A G C ...
	<i>i</i> + 5	A T T G A C T A T A ...
	<i>i</i> + 6	A T T G G A C G T A ...
	<i>i</i> + 7	A T T G G A T A C A ...
	<i>i</i> + 8	A T T G G C A A T C ...
	<i>i</i> + 9	A T T G G G A A C T ...
	<i>i</i> + 10	A T T G G T C A G A ...

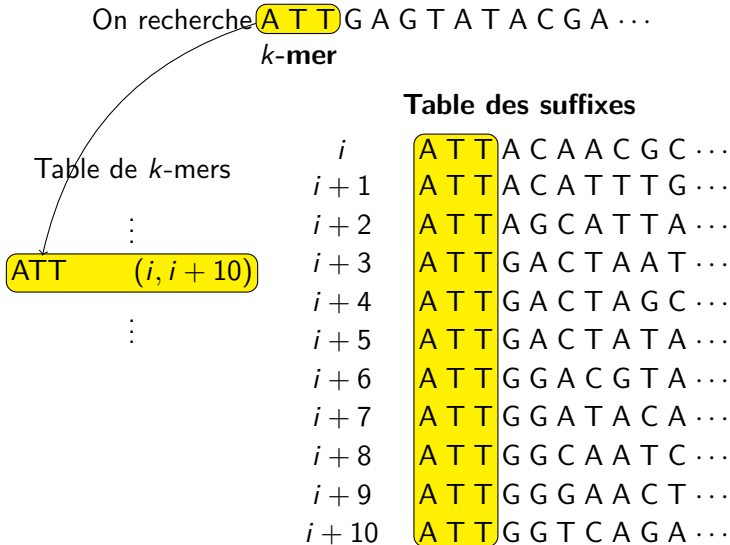
RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 



RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 



RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche **ATTG** A G T A T A C G A ...

		Table des suffixes	
Table de k -mers ATT ($i, i + 10$)	i		A T T A C A A C G C ...
	$i + 1$		A T T A C A T T T G ...
	$i + 2$		A T T A G C A T T A ...
	$i + 3$		A T T G A C T A A T ...
	$i + 4$		A T T G A C T A G C ...
	$i + 5$		A T T G A C T A T A ...
	$i + 6$		A T T G G A C G T A ...
	$i + 7$		A T T G G A T A C A ...
	$i + 8$		A T T G G C A A T C ...
	$i + 9$		A T T G G G A A C T ...
	$i + 10$		A T T G G T C A G A ...

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche **A T T G A** G T A T A C G A ...

		Table des suffixes	
Table de k -mers ATT ($i, i + 10$)	i	A T T A C A A C G C ...	
	$i + 1$	A T T A C A T T T G ...	
	$i + 2$	A T T A G C A T T A ...	
	$i + 3$	A T T G A C T A A T ...	
	$i + 4$	A T T G A C T A G C ...	
	$i + 5$	A T T G A C T A T A ...	
	$i + 6$	A T T G G A C G T A ...	
	$i + 7$	A T T G G A T A C A ...	
	$i + 8$	A T T G G C A A T C ...	
	$i + 9$	A T T G G G A A C T ...	
	$i + 10$	A T T G G T C A G A ...	

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche **A T T G A G** T A T A C G A ...

		Table des suffixes	
Table de k -mers ATT ($i, i + 10$)	i	A	T T A C A A C G C ...
	$i + 1$	A	T T A C A T T T G ...
	$i + 2$	A	T T A G C A T T A ...
	$i + 3$	A	T T G A C T A A T ...
	$i + 4$	A	T T G A C T A G C ...
	$i + 5$	A	T T G A C T A T A ...
	$i + 6$	A	T T G G A C G T A ...
	$i + 7$	A	T T G G A T A C A ...
	$i + 8$	A	T T G G C A A T C ...
	$i + 9$	A	T T G G G A A C T ...
	$i + 10$	A	T T G G T C A G A ...

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche **A T T G A** G T A T A C G A ...

MEM

Table des suffixes

Table de k -mers ATT $(i, i + 10)$	i	A T T A C A A C G C ...
	$i + 1$	A T T A C A T T T G ...
	$i + 2$	A T T A G C A T T A ...
	$i + 3$	A T T G A C T A A T ...
	$i + 4$	A T T G A C T A G C ...
	$i + 5$	A T T G A C T A T A ...
	$i + 6$	A T T G G A C G T A ...
	$i + 7$	A T T G G A T A C A ...
	$i + 8$	A T T G G C A A T C ...
	$i + 9$	A T T G G G A A C T ...
	$i + 10$	A T T G G T C A G A ...

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche **ATTGA**GTATACGA...

MEM

Table des suffixes

Table de k -mers ATT $(i, i + 10)$	i	ATTACAACGC...
	$i + 1$	ATTACATTTG...
	$i + 2$	ATTAGCATTAA...
	$i + 3$	ATTGA CTA AT...
	$i + 4$	ATTGA CTA GC...
	$i + 5$	ATTGA CTA TA...
	$i + 6$	ATTGGACGTA...
	$i + 7$	ATTGGATACA...
	$i + 8$	ATTGGCAATC...
	$i + 9$	ATTGGGAAC T...
	$i + 10$	ATTGGTCAGA...

Positions non informatives

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche **A T T G A** **G T A** T A C G A ...

MEM

Table des suffixes

Table de k -mers ATT $(i, i + 10)$	i	A T T A C A A C G C ...
	$i + 1$	A T T A C A T T T G ...
	$i + 2$	A T T A G C A T T A ...
	$i + 3$	A T T G A C T A A T ...
	$i + 4$	A T T G A C T A G C ...
	$i + 5$	A T T G A C T A T A ...
	$i + 6$	A T T G G A C G T A ...
	$i + 7$	A T T G G A T A C A ...
	$i + 8$	A T T G G C A A T C ...
	$i + 9$	A T T G G G A A C T ...
	$i + 10$	A T T G G T C A G A ...

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche **A T T G A** **G T A T** A C G A ...

		Table des suffixes	
Table de k -mers ATT ($i, i + 10$)	i	A T T A C A A C G C ...	
	$i + 1$	A T T A C A T T T G ...	
	$i + 2$	A T T A G C A T T A ...	
	$i + 3$	A T T G A C T A A T ...	
	$i + 4$	A T T G A C T A G C ...	
	$i + 5$	A T T G A C T A T A ...	
	$i + 6$	A T T G G A C G T A ...	
	$i + 7$	A T T G G A T A C A ...	
	$i + 8$	A T T G G C A A T C ...	
	$i + 9$	A T T G G G A A C T ...	
	$i + 10$	A T T G G T C A G A ...	

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017) , sur bioRxiv en 2016 

On recherche **A T T G A** **G T A T** A C G A ...

		Table des suffixes	
Table de k -mers ATT $(i, i + 10)$	i		A T T A C A A C G C ...
	$i + 1$	6	A T T A C A T T T G ...
	$i + 2$	4	A T T A G C A T T A ...
	$i + 3$	3	A T T G A C T A A T ...
	$i + 4$	8	A T T G A C T A G C ...
	$i + 5$	8	A T T G A C T A T A ...
	$i + 6$	4	A T T G G A C G T A ...
	$i + 7$	6	A T T G G A T A C A ...
	$i + 8$	5	A T T G G C A A T C ...
	$i + 9$	5	A T T G G G A A C T ...
	$i + 10$	5	A T T G G T C A G A ...

LCP

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017)  sur bioRxiv en 2016 

On recherche **A T T G A** **G T A T** A C G A ...

		Table des suffixes	
Table de k -mers		i	A T T A C A A C G C ...
ATT	$(i, i + 10)$	$i + 1$	6 A T T A C A T T T G ...
		$i + 2$	4 A T T A G C A T T A ...
		$i + 3$	3 A T T G A C T A A T ...
		$i + 4$	8 A T T G A C T A G C ...
		$i + 5$	8 A T T G A C T A T A ...
		$i + 6$	4 A T T G G A C G T A ...
		$i + 7$	6 A T T G G A T A C A ...
		$i + 8$	5 A T T G G C A A T C ...
		$i + 9$	5 A T T G G G A A C T ...
		$i + 10$	5 A T T G G T C A G A ...
		LCP	

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017)  sur bioRxiv en 2016 

On recherche **A T T G A** **G T A T** A C G A ...

		Table des suffixes	
Table de k -mers ATT $(i, i + 10)$	i	6	A T T A C A A C G C ...
	$i + 1$	4	A T T A C A T T T G ...
	$\vdots$		
	$i + 2$	3	A T T A G C A T T A ...
	$i + 3$	8	A T T G A C T A A T ...
	$i + 4$	8	A T T G A C T A G C ...
	$i + 5$	4	A T T G A C T A T A ...
	$i + 6$	6	A T T G G A C G T A ...
	$i + 7$	5	A T T G G A T A C A ...
	$i + 8$	5	A T T G G C A A T C ...
	$i + 9$	5	A T T G G G A A C T ...
	$i + 10$		A T T G G T C A G A ...
LCP			

RapMap : approche par « quasi-mapping »

Srivastava *et al*, 2016  utilisé dans Salmon (Patro *et al*, 2017)  sur bioRxiv en 2016 

On recherche **A T T G A G T A T A C G** A ...
k-mer

		Table des suffixes	
Table de <i>k</i> -mers ATT (<i>i</i> , <i>i</i> + 10)	<i>i</i>	6	A T T A C A A C G C ...
	<i>i</i> + 1	4	A T T A C A T T T G ...
	⋮		
	<i>i</i> + 2	3	A T T A G C A T T A ...
	<i>i</i> + 3	8	A T T G A C T A A T ...
	<i>i</i> + 4	8	A T T G A C T A G C ...
	⋮		
	<i>i</i> + 5	4	A T T G A C T A T A ...
	<i>i</i> + 6	6	A T T G G A C G T A ...
	<i>i</i> + 7	5	A T T G G A T A C A ...
	<i>i</i> + 8	5	A T T G G C A A T C ...
	<i>i</i> + 9	5	A T T G G G A A C T ...
	<i>i</i> + 10		A T T G G T C A G A ...
LCP			

Chevauchement de reads de 3^e génération

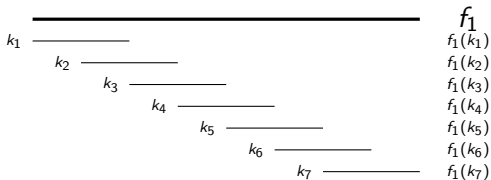
Les reads sont longs et bruités

Plusieurs (dizaines) de kb, 10–30 % d'erreur

MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

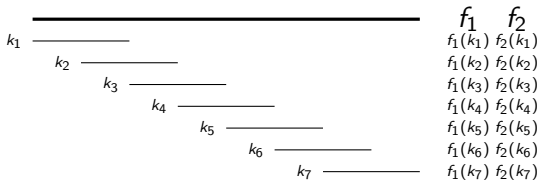
read



MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

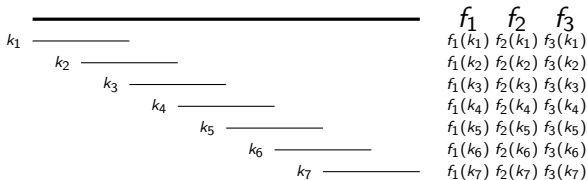
read



MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

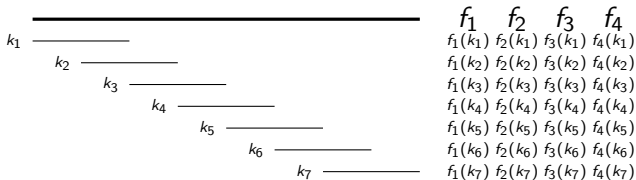
read



MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

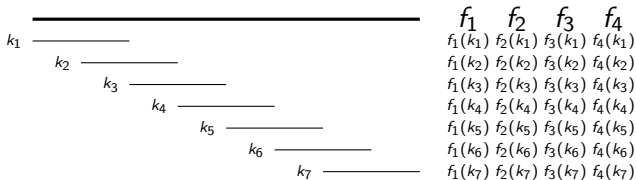
read



MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

read



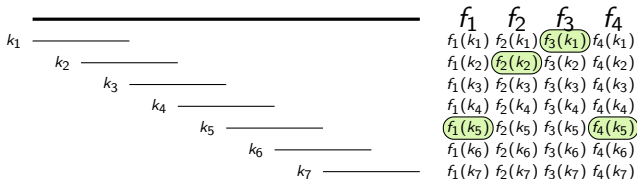
On ne prend qu'une valeur par fonction

Généralement le minimum

MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

read



minimizers

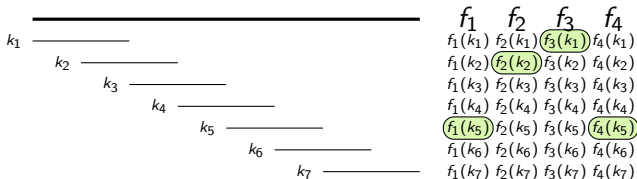
On ne prend qu'une valeur par fonction

Généralement le minimum

MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

read



minimizers

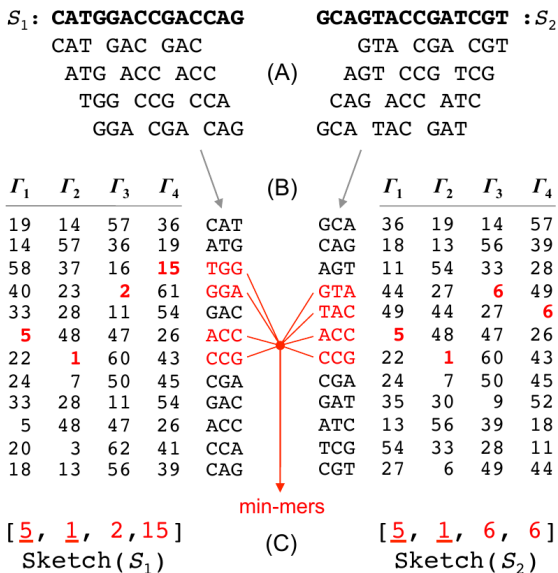
On ne prend qu'une valeur par fonction

Généralement le minimum

Le read est résumé à $[f_1(k_5), f_2(k_2), f_3(k_1), f_4(k_5)]$

MHAP : approche par « alignement probabiliste »

Berlin et al, 2015



MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

S_1 : **CATGGACCGACCAG**

CAT GAC GAC

ATG ACC ACC

TGG CCG CCA

GGA CGA CAG

GCAGTACCGATCGT : S_2

GTA CGA CGT

AGT CCG TCG

CAG ACC ATC

GCA TAC GAT

(A)

Γ_1	Γ_2	Γ_3	Γ_4	
19	14	57	36	CAT
14	57	36	19	ATG
58	37	16	15	TGG
33	28	11	54	GAC
5	48	47	26	ACC
22	1	60	43	CCG
24	7	50	45	CGA
33	28	11	54	GAC
5	48	47	26	ACC
20	3	62	41	CCA
18	13	56	39	CAG

(B)

Γ_1	Γ_2	Γ_3	Γ_4	
36	19	14	57	GCA
18	13	56	39	CAG
11	54	33	28	AGT
49	44	27	6	TAC
5	48	47	26	ACC
22	1	60	43	CCG
24	7	50	45	CGA
35	30	9	52	GAT
13	56	39	18	ATC
54	33	28	11	TCG
27	6	49	44	CGT

Faut-il comparer toutes les paires de reads ?

min-mers

[5, 1, 2, 15]
Sketch(S_1)

(C)

[5, 1, 6, 6]
Sketch(S_2)

MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

Chaque f_i a sa table de hash T_i

On y enregistre la valeur minimale (clé) pour chaque read (valeur)

MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

Chaque f_i a sa table de hash T_i

On y enregistre la valeur minimale (clé) pour chaque read (valeur)

$f_1(.)$

4 r_4 pos. 2907 r_{12} pos. 6295 r_{569} pos. 42810

7 r_9 pos. 271

13 r_6 pos. 5817 r_7 pos. 211

MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

Chaque f_i a sa table de hash T_i

On y enregistre la valeur minimale (clé) pour chaque read (valeur)

$f_1(.)$

4 r_4 pos. 2907 r_{12} pos. 6295 r_{569} pos. 42810

7 r_9 pos. 271

13 r_6 pos. 5817 r_7 pos. 211

Pour les paires de reads ayant suffisamment de minimizers conservés on va s'intéresser à l'ordre des k-mers

On peut choisir des k-mers plus petits

MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

Chaque f_i a sa table de hash T_i

On y enregistre la valeur minimale (clé) pour chaque read (valeur)

$f_1(.)$

4 r_4 pos. 2907 r_{12} pos. 6295 r_{569} pos. 42810

7 r_9 pos. 271

13 r_6 pos. 5817 r_7 pos. 211

Pour les paires de reads ayant suffisamment de minimizers conservés on va s'intéresser à l'ordre des k-mers

On peut choisir des k-mers plus petits



MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

Chaque f_i a sa table de hash T_i

On y enregistre la valeur minimale (clé) pour chaque read (valeur)

$f_1(.)$

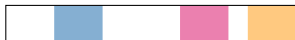
4 r_4 pos. 2907 r_{12} pos. 6295 r_{569} pos. 42810

7 r_9 pos. 271

13 r_6 pos. 5817 r_7 pos. 211

Pour les paires de reads ayant suffisamment de minimizers conservés on va s'intéresser à l'ordre des k-mers

On peut choisir des k-mers plus petits



MHAP : approche par « alignement probabiliste »

Berlin et al, 2015

Chaque f_i a sa table de hash T_i

On y enregistre la valeur minimale (clé) pour chaque read (valeur)

$f_1(.)$

4 r_4 pos. 2907 r_{12} pos. 6295 r_{569} pos. 42810

7 r_9 pos. 271

13 r_6 pos. 5817 r_7 pos. 211

Pour les paires de reads ayant suffisamment de minimizers conservés on va s'intéresser à l'ordre des k-mers

On peut choisir des k-mers plus petits

En pratique : $k_1 = 16$; $k_2 = 12$; 512 à 1 256 fonctions de hash

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Problème de MHAP : le nombre de minimizers est fixe

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash

Une fenêtre glissante sur w k-mers

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash

Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash

Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

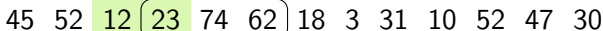
Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30



Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs

45 52 12 23 74 62 18 3 31 10 52 47 30

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs 45 52 12 23 74 62 18 3 31 10 52 47 30

Il y a toujours (au moins) un minimizer parmi w positions consécutives

Minimap2 : minimizers et fenêtre glissante (*winnowing*)

Li, 2018 qui s'appuie sur Minimap Li, 2016

Une seule fonction de hash
Une fenêtre glissante sur w k-mers

Hashs 45 52 12 23 74 62 18 3 31 10 52 47 30

Il y a toujours (au moins) un minimizer parmi w positions consécutives

En pratique: $k = 15$, $w = 5$

Combien bien choisir ses minimizers ?

Marçais et al, 2017 

Ordre alphabétique ?

Combien bien choisir ses minimizers ?

Marçais et al, 2017 

~~Ordre alphabétique ?~~

AAA... est le plus petit

Combien bien choisir ses minimizers ?

Marçais et al, 2017

~~Ordre alphabétique ?~~

AAA... est le plus petit

Ordre alphabétique bidouillé

KMC2 : Retrait des k-mers commençant par
AAA, ACA ou contenant AA (sauf au début)

Combien bien choisir ses minimizers ?

Marçais et al, 2017 

~~Ordre alphabétique ?~~

AAA... est le plus petit

Ordre alphabétique bidouillé

KMC2 : Retrait des k-mers commençant par
AAA, ACA ou contenant AA (sauf au début)

Ordre aléatoire

Kraken : XOR avec valeur aléatoire ;
Minimap : fonction inversible

Combien bien choisir ses minimizers ?

Marçais et al, 2017 

~~Ordre alphabétique ?~~

AAA... est le plus petit

Ordre alphabétique bidouillé

KMC2 : Retrait des k-mers commençant par
AAA, ACA ou contenant AA (sauf au début)

Ordre aléatoire

Kraken : XOR avec valeur aléatoire ;
Minimap : fonction inversible

Universal hitting sets

DOCKS 

Combien bien choisir ses minimizers ?

Marçais et al, 2017 

~~Ordre alphabétique ?~~

AAA... est le plus petit

Ordre alphabétique bidouillé

KMC2 : Retrait des k-mers commençant par
AAA, ACA ou contenant AA (sauf au début)

Ordre aléatoire

Kraken : XOR avec valeur aléatoire ;
Minimap : fonction inversible

Universal hitting sets

DOCKS 

Comment bien choisir ses minimizers ?

Marçais et al, 2017 

$$k = 7, w = 11$$

	Ordering	distance mean $\pm$ stdev	low sep. %
human	lexico.	5.1(34)	33
	random	6.0(32)	19
	Minimap	5.8(33)	22
	KMC2	5.9(33)	19
	UMD Ovl	6.1(31)	17
	Kraken	6.2(30)	13
	DOCKS	6.7(26)	6.2

Comment bien choisir ses minimizers ?

Marçais et al, 2017 [a](#)

$$k = 7, w = 11$$

	Ordering	# bins	avg size (mega bases)	max ratio
human	lexico.	16285	0.175	96.4
	random	11388	0.251	38.3
	Minimap	12280	0.233	69.5
	KMC2	12287	0.233	74.5
	UMD Ovl	11015	0.259	26.1
	Kraken	10389	0.275	26.3
	DOCKS	4046	0.706	23.5

© Marçais et al, 2017