Alignement de lectures issues de séquenceurs « nouvelle génération »
ou
"Mapping de reads NGS"

Plan

1. Algorithmes de mapping ;
2. Outils ;
3. Expériences et résultats.

Algorithme naïf

Pour chaque read, je parcours le génome de référence à la recherche de correspondances.

- Si on cherche avec un nombre maximal donné de mismatches ;
- Si on cherche avec mismatches et gaps : Smith-Waterman.

Smith-Waterman est couteux (même si GPU & SSE2), donc limiter les appels.

Dans la suite : reads de longueur 40.

Mapping par hashing

Indexation

✺ Requête SQL :

```
SELECT prenom FROM annuaire WHERE nom = 'Dupond';
```

✺ Indexation des *3-mers* du génome G=ATCTGCTAGCTA :

ATC	0	GCT	4, 8
TCT	1	CTA	5, 9
CTG	2	TAG	6
TGC	3	AGC	7

✺ Table de hachage (*hash table*) : ensemble de paires (clef, valeur).

- La **clef** c'est un entier, encodage (hash) du k-mer.
- La **valeur** c'est un pointeur vers les occurrences du k-mer dans la référence.

Indexation — le retour

Théorème des chaussettes (ou des pigeons)

4 tiroirs dans la commode et 3 chaussettes à ranger.

Exemple : on cherche des alignements avec au plus 3 mismatches :

- découper les reads en 4 tronçons de 10 nucléotides ;
- indexer la référence (des 10-mers) ;
- lire les reads (découpés en 4 tronçons) et chercher dans la table de hachage ;
- si un tronçon s'aligne (ou plus), lancer un Smith-Waterman sur la région.

Idée : *Seed and extend*.

Indexation — graines et chaussettes

Limite : 4^{10} 10-mers, soit environ 1 million de 10-mers. On les retrouve quasi-tous dans un génome de taille moyenne. Donc beaucoup (trop) de lancement de Smith-Waterman.

Deux approches :

- q-**gram filter** : plein de petites seeds matchent.

- **Seeds (graines)** : 1111111111 (poids et taille).
Graine espacée : 1010110010011010 1.
Filtrage plus efficace (choix de la graine ou de la famille de graines) : moins de hits potentiels sur lesquels lancer un SW.

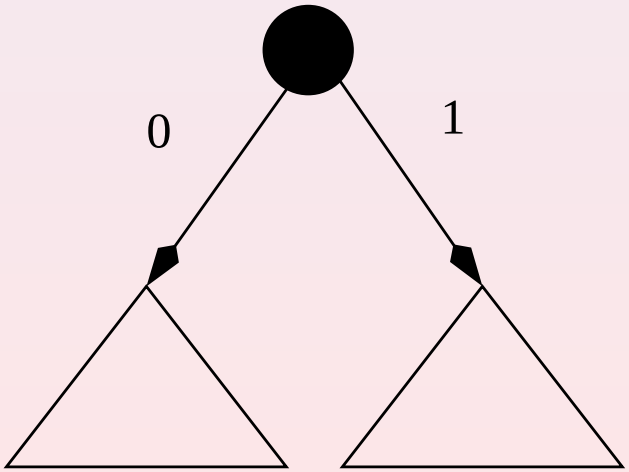
Mapping par Burrows-Wheeler

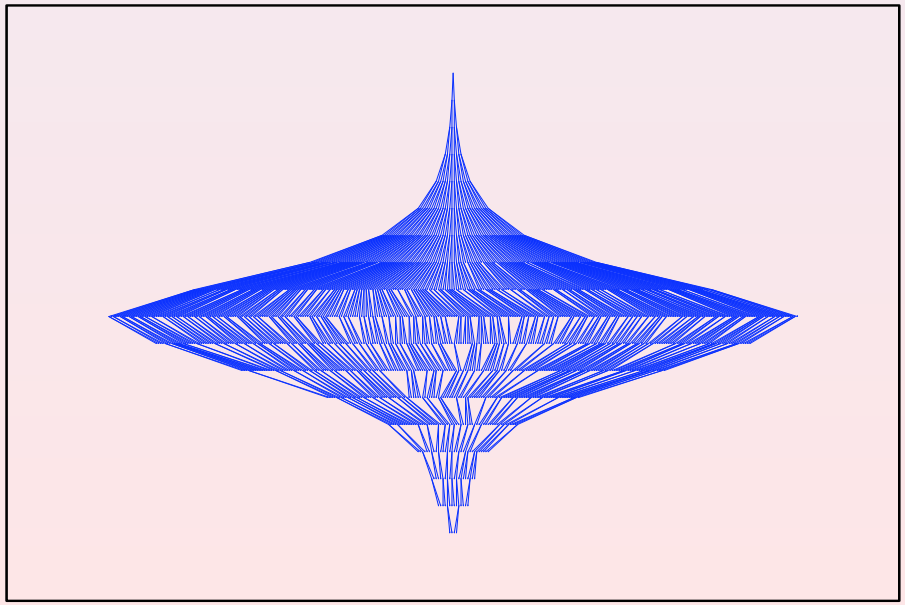
Tries [de la Briandais'59 et Fredkin'60]

Définition récursive : X un ensemble fini de chaînes infinies sur l'alphabet

$$\text{trie}(X) = \begin{cases} \emptyset & \text{si } |X| = 0, \\ \bullet & \text{si } |X| = 1, \\ \langle \bullet, \text{trie}(X\backslash 0), \text{trie}(X\backslash 1) \rangle & \text{sinon,} \end{cases}$$

où $X\backslash\alpha$ est l'ensemble des textes commençant par la lettre α auxquels on retire cette lettre initiale.





Arbres des suffixes

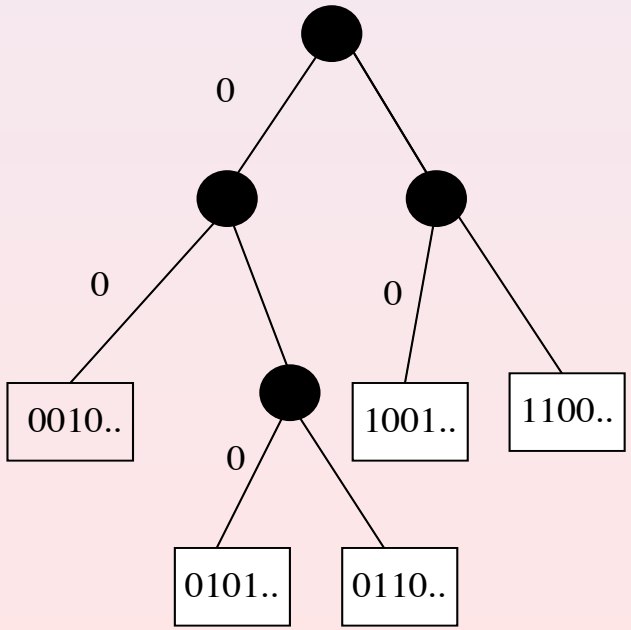
Arbre des suffixes et tries

- Même définition récursive ;
- Ensemble de base différent.

Soit T un mot infini (texte) et n un entier. X est l'ensemble des n premiers suffixes de T.

$T = 011001010\ldots \qquad n = 5$

$$X = \{011001010\ldots, 11001010\ldots, 1001010\ldots, 001010\ldots, 01010\ldots\}$$



Arbre des suffixes, suffix arrays, prefix trie, FM-index

Construction d'un arbre des suffixes du génome de référence.
On tronque l'arbre à la profondeur 40 (*sliding window*) : regarder les 40-gram (ou 40-mer).

⇛ Taille de la structure de donnée ;
⇛ Gérer les erreurs et gaps.

Structures de données assez proches

- Suffix array : tableau des suffixes triés par ordre alphabétique.
- Prefix trie : suffix trie, mais sens de lecture inversé.
- FM-index : index Ferragina-Manzini.

Transformée de Burrows-Wheeler

➢ utilisée dans `bzip2`.

`ATCAG$` → → `GC$TAA`.

Et on peut reconstruire la séquence initiale (permutation reversible).
On ne gagne rien en taille (même longueur).

Description des outils

Outils	**Format**	**Algorithme**	**Input**	**Threads**	**Gaps**
bwa	SAM	Burrows-Wheeler	NT & color	oui	oui
Bowtie	SAM	Burrows-Wheeler	NT & color	oui	non
Novoalign	SAM	indexe la réf	NT & color	-	-
MOM	perso	hash sur réf/reads	NT	oui	non
ProbeMatch	perso	hash sur réf	NT	non	oui
SOAP2	perso	Burrows-Wheeler	NT	oui	non
BFAST	SAM	hash sur réf	NT & color	oui	oui
SHRiMP	SAM	hash sur reads	NT & color	oui	oui
maq	SAM	hash sur reads	NT	non	non
SSAHA2	SAM	hash sur réf	NT	non	oui

Expériences et résultats

2.3GHz (64 bits) CPU. 16 Go memoire.

10 M de reads tirés aléatoirement (presque)-uniformément dans

- A concaténation du génome humain ;
- B concaténation de 900 génomes bactériens.

Est-ce qu'on retrouve la position originelle parmi les hits ?

Résultats 0 mismatches

				Unique reads	
Software	Indexing time	Mapping time	Nb mapped reads	Nb	Orig pos retrieved
bwa	1h 28mn	48mn	9999998 100%	8739090	8330833 95.32%
Novoalign	23mn	10h 50mn	9999320 99.99%	8875324	8874639 99.99%
Bowtie	3h 32mn	21mn	9999950 99.99%	8874680	8874631 99.99%
SOAP2	1h 34mn	56mn	9999958	8877067	8877066
BFAST	14mn+ 10× 1d 10h	13h 39mn	9296090 92.9%	8851822	8831146 88.3%
maq	6 mn	8 h 46 mn	9999958	9999958	7007096
SSAHA2	29 mn	1d 12 h	9964125	8886204	8876357

Résultats 3 mismatches

				Unique reads	
Software	Indexing time	Mapping time	Nb mapped reads	Nb	Orig pos retrieved
bwa	1h 28mn	3h 16mn	5781876	4790181	4566774
Novoalign	23mn	4d 8h	9999949	8695303	8471634
Bowtie	3h 32mn	3h 31mn	9999950	8495019	8494971
SOAP2			665457	202437	
BFAST	14mn + 10 $\times$ 1d 10h	10h 30mn	9921349	8759952	8392474
maq	6 mn	1 d 3h	9999957	9999957	7273335
SSAHA2	29 mn	5 d 22 h	9999787	8286416	5200503

bwa et bowtie

bwa

```
bwa index ../Data/refJIP.fasta
bwa aln -o 0 -n 3 refJIP.fasta
../Data/readsSangerTHC.fastq > THCvsJIP.sai
bwa samse refJIP.fasta THCvsJIP.sai
../Data/readsSangerTHC.fastq > THCvsJIP.sam
```

bowtie

```
bowtie-build -f ../Data/Refs/refJIP.fasta refJIP
bowtie --sam -v 3 -k 200 -f refJIP
../Data/Reads/readsTHCTrimmed.fasta THCvsJIP.sam
```

samtools

Création d'un fichier BAM

```
samtools faidx ../Data/Refs/refJIP.fasta
samtools view -bt ../Data/Refs/refJIP.fasta.fai
refJIP.sam > refJIP.bam
samtools sort refJIP.bam refJIP.sort
samtools index refJIP.sort.bam
```

Statistiques

```
samtools idxstats refJIP.sort.bam
```

Visualisation

```
samtools tview refJIP.sort.bam
```

Format SAM/BAM

```
AZOTE:3:1:3:926#0 0 AM398681 2749095 37 22M * 0 0
ATAATGAACAATTAGAAATGAC BCCCCCCCCBCACCABBCCBACB
```

- Format d'alignement le plus usité.
- Format très flexible.
- Format textuel (vous pouvez le lire).

Commence par un en-tête (@). Chaque ligne renseigne l'alignement d'un read (selon les réglages de l'outil d'alignement, plusieurs lignes pour le même read).

Le FLAG peut être 0 (read aligné en sens direct), 4 (read pas aligné), 16 (read aligné en sens invcomp).
MAPQ : qualité du mapping (échelle Phred). CIGAR : changements entre le read et l'alignement.

BAM : version comprimée du fichier SAM. Prend moins de place. Illisible à l'œil humain.