Rapport développeur

# Indexation

L’indexation d’une séquence suit le processus suivant :

Le tableau SA[i] est obtenu avec la fonction *tks.simple\_kark\_sort* du package [tools\_karkkainen\_sanders](https://github.com/michaelmior/pysuffix/blob/master/tools_karkkainen_sanders.py). Cette fonction crée, à partir d’une séquence à laquelle on a ajouté un « $ », l’ensemble des séquences différentes existantes en faisant varier de positions le caractère $. Elle trie ensuite ces séquences par ordre lexicographique. Une petite fonction permet ensuite de récupérer la dernière lettre de chaque séquence ordonnée sous forme de liste (BWT).

# Mapping

## Fonction *search\_querry*

Cette fonction recherche les correspondances des k-mers d’un read dans une séquence référence, grâce au FM-index (les premières lettres des séquences ordonnées, donc le nombre de chaque nucléotides trié alphabétiquement) et à la liste BWT.

Elle retourne un dictionnaire avec les informations suivantes :

{('TCTGA', 0): [[301, 500], 0, '+'], ('CTGAT', 1): [501, 1, '+']}

**Clés**

**Valeurs**

**k-mer**

**Position du k-mer dans le read**

**Position(s) correspondante(s) dans la séquence référence**

**Sens de lecture du read contenant le k-mer**

## Fonction *comparison*

Cette fonction aligne deux séquences à partir d’une position d’ancrage de la première séquence dans la seconde et retourne le nombre de substitutions entre les deux séquences.

### Fonction *seed\_and\_extend*

Cette fonction retourne un fichier au format VCF, avec la structure suivante :

|  |  |  |  |
| --- | --- | --- | --- |
| Position de la substitution | Nucleotide original | Nucléotide variant | Nombre d’observations de la mutation |
| X | X | X | X |

La fonction se structure comme ci-après :

**Pour chaque read**

NB : Lors de la comparaison des k-mers avec la séquence référence, on vérifie bien que les positions comparées ne l’ont pas déjà été, afin de gagner du temps d’exécution.

## Main

On vérifie dans un premier temps que le parser est complet, auquel cas on écrit dans un fichier la structure VCF à l’aide de la fonction *seed\_and\_extend*. Si le parser est incomplet, un message d’erreur est retourné.