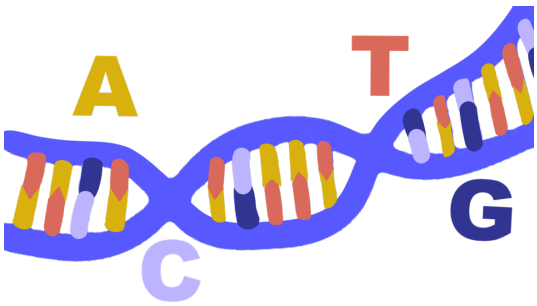
Recherche textuelle

i Nigel Horspool est britannique de naissance, mais est maintenant citoyen canadien. Il a inventé l' algorithme Boyer-Moore-Horspool , un algorithme de recherche de chaînes rapide adapté de l' algorithme de recherche de chaînes Boyer-Moore. L'algorithme fut publié en 1980 et reste depuis l'algorithme de recherche le plus efficace.
Horspool est aussi co-inventeur de la compression dynamique de Markov et a été rédacteur associé puis rédacteur en chef de la revue Software : Practice and Experience de 2007 à 2017.
Il prend sa retrait en tant que professeur d' informatique de l' Université de Victoria en 2016.

•Situation



L'information génétique présente dans nos cellules est portée par les molécules d'ADN. Les molécules d'ADN sont, entre autres, composées de bases azotées ayant pour noms : Adénine (représenté par un A), Thymine (représenté par un T), Guanine (représenté par un G) et Cytosine (représenté par un C).

Il est donc souvent nécessaire de rechercher un motif dans une chaîne d'ADN.

Dans ce cours, nous allons prendre les éléments suivants :

- **Chaîne d'ADN :**

  ```
  CAAGCGCAGAAGACGCGGCAGACCTTCGTTATAGG CGATGATTTCGAACCTACTAGTGGGTCTCTTAAG
  GGCCGAGCGGTTCCGAGAGATAGTGAAAGATGGCT GGGCTGTGAAGGGAAGGAGTCGTGAAAGCGCGAG
  ACACGAGTGTGCGCAAGCGCAGCGCCTTAGTATGC TCCAGTGTAGAAGCTCCGGCGTCCCGTCTAACCG
  TACGCTGTCCCCGGTACATGGAGCTAATAGGCTTT ACTGCCCAATATGACCCCGCGCCGCGACAAAACA
  ATAACAGTTTGCTGTATGTTCCATGGTGGCCAATC CGTCTCTTTTCGACAGCACGGCCAATTCTCCTAG
  GAAGCCAGCTCAATTTCAACGAAGTCGGCTGTTGA ACAGCGAGGTATGGCGTCGGTGGCTCTATTAGTG
  GTGAGCGAATTGAAATTCGGTGGCCTTACTTGTAC CAGAGCGATCCCTTCCCACCATTCTTATGCGTCG
  TCTGTTACCTGGCTTGGCAT
  ```

- **Motif recherché :**

  ```
  CAGA
  ```

•Algorithme naif

i L'algorithme naïf consiste simplement à comparer un à un, de gauche à droite, les caractères du texte apparaissant dans la fenêtre avec ceux du motif.
On compare le motif avec le texte caractère par caractère, puis si le caractère considéré correspond au premier caractère du mot, nous comparerons les caractères suivants à ceux du mot.
En cas de non-correspondance, on avance simplement la fenêtre d'un caractère vers la droite, si la recherche s'avère fructueuse (c'est à dire que le nombre de correspondances est égal à la longueur du motif) on renvoie True.

•Exemple d'exécution

Étudions les étapes de la recherche naive de note motif (`CAGA`) dans le brin d'ADN `CAAGCGCAGAAGACGCGGCAGACCTTCGTTA...`

Etape 1	Index : 0 Correspondance : Oui Nb de correspondance : 1	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 2	Index : 0 Correspondance : Oui Nb de correspondance : 2	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 3	Index : 0 Correspondance : Non Nb de correspondance : 0	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 4	Index : 1 Correspondance : Non Nb de correspondance : 0	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... →CAGA
Etape 5	Index : 2 Correspondance : Non Nb de correspondance : 0	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... →CAGA
Etape 6	Index : 3 Correspondance : Non Nb de correspondance : 0	CAAGGCAGAAGACGCGGCAGACCTTCGTTA... →CAGA
Etape 7	Index : 4 Correspondance : Oui Nb de correspondance : 1	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... →CAGA
Etape 8	Index : 4 Correspondance : Non Nb de correspondance : 0	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 9	Index : 5 Correspondance : Non Nb de correspondance : 0	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... →CAGA
Etape 10	Index : 6 Correspondance : Oui Nb de correspondance : 1	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... →CAGA
Etape 11	Index : 6 Correspondance : Oui Nb de correspondance : 2	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 12	Index : 6 Correspondance : Oui Nb de correspondance : 3	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 13	Index : 6 Correspondance : Oui Nb de correspondance : 3	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA

Il faut donc 13 étapes pour retrouver le motif.

•Algorithme Boyer-Moore-Horspool

i L'agorithme de Boyer-Moore-Horspool étudié ici nécessite un prétraitement du motif recherché : on doit déterminer le pour chaque lettre du motif le décallage nécessaire.

•• Règle de calcul du décalage

Pour chaque lettre du motif

Pour chaque lettre du motif la règle est la suivante : le décalage de la lettre est obtenue par la longueur du motif - indice de la lettre - 1

$$d = longueur(motif) - indice - 1$$

⚠ Les indices commence à 0.

⚠ Les exceptions :

- Pour les caractères apparaissant en double, le décalage est calculé en fonction du dernier indice du caractère dans le motif
- Pour le dernier caractère du motif, on lui attribut un décalage égal à la longueur du motif.

Pour les autres caractères

Tous les caractères n'appartenant pas motif ont un un décalage égal à la longueur du motif.

Exemple

Calculs du décalage des différentes lettres du motif `CAGA`

Lettre	Index retenu	Calcul	Décalage
A	3 (1 ou 3)		4
C	0	4-0-1	3
G	2	4-2-1	1
*	4		1

⚠ * symbolise tous les autres caractères.

•• Recherche

Définition

`Algorithme de recherche`

L'algorithme de recherche de Boyer-Moore-Horspool repose sur 2 idées :

1. Effectuer les comparaisons de la gauche vers la droite.
2. Utiliser la table de décalage pour determiner la nouvelle position de la fenêtre de recherche dès qu'une correspondance n'est pas valide. Pour cela on calcule la différences entre le décalage correspond à la lettre moins le nombre de correspondance déjà trouvées.

   ⚠ Si ce nombre est négatif, on décale de la longueur du mot - le nombre de correspondances.

•Exemple d'exécution

Étudions les étapes de la recherche naive de note motif (`CAGA`) dans le brin d'ADN `CAAGCGCAGAAGACGCGGCAGACCTTCGTTA...`

Etape 1	Index : 0 Correspondance : Non Nb de correspondances : 0 Décalage : G(1) - Cor(0) = 1	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 2	Index : 1 Correspondance : Non Nb de correspondances : 0 Décalage : C(3) - Cor(0) = 3	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... →CAGA
Etape 3	Index : 4 Correspondance : Oui Nb de correspondances : 1 Décalage :	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... →CAGA
Etape 4	Index : 4 Correspondance : Non Nb de correspondances : 1 Décalage : C(3) - Cor(1) -> 2	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 5	Index : 2 Correspondance : Oui Nb de correspondances : 1 Décalage : C(3) - Cor(1) -> 2	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 6	Index : 6 Correspondance : Oui Nb de correspondances : 2 Décalage :	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 7	Index : 6 Correspondance : Oui Nb de correspondances : 3 Décalage :	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA
Etape 8	Index : 6 Correspondance : Oui Nb de correspondances : 4 Décalage :	CAAGCGCAGAAGACGCGGCAGACCTTCGTTA... CAGA

•Sources

- Grafikart
- Mon lycée numérique
- Pixees