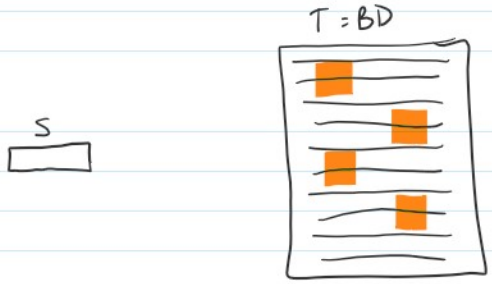
BLAST

12 septembre 2023 09:13

Blast Local Alignment Search Tool



But: trouver les régions de T similaires à S

Alignement local: parfait pour cette tâche mais
$O(|S||T|)$ trop lent
e.g. $|T| \sim 100\,000\,000\,000$

Heuristique: algo rapide mais sans garantie de précision

⓪ Prétraitement sur la BD T: indexation des k-mers
k-mer: séquence de longueur k

k=3
ACA: 1
CAT: 2
...

```
H = dict()        // clé: k-mer, val: liste des positions
pour i = 1,.., |T| - k + 1
    X = T[i..i+k-1]
    H[x].append(i)
```

O(|T|) itér.
O(1)

AAA: 10, 12, 108
AAC: 1, 4
:

① Recherche de S

- Lister tous les k-mers de S

k=3 S = ACATACCA

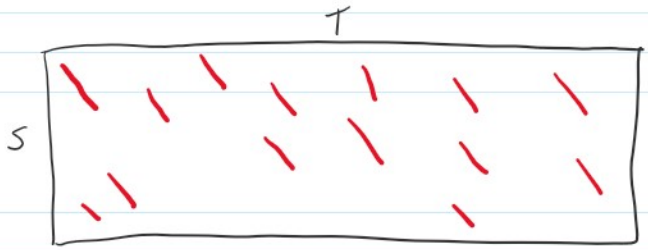
ACA TAC
CAT ACC
ATA CCA

- Lister tous les k-mers à distance Hamming $\leq d$ des k-mers listés (Hamming = #pos différents)

d=1 ACA CCA GCA TCA AGA ATA ...
CAT AAT ...

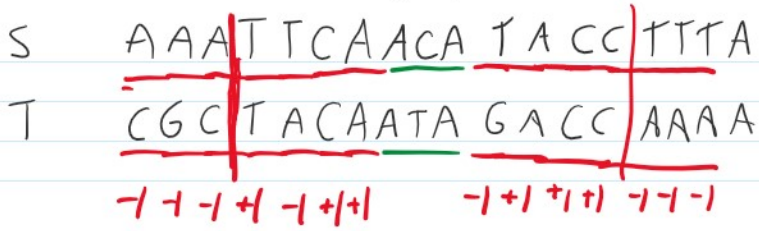
Pour tous les k-mers listés, trouver les

Pour tous les k-mers listés, trouver les occurrences dans T. (utiliser H)

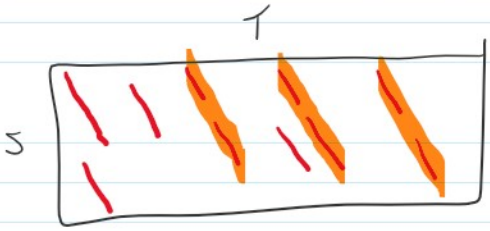


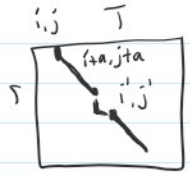
Les occurrences trouvées sont appelées matches

(2) ∀ match, on étend le match à gauche et à droite maximalement, i.e. tant que l'extension améliore le score, jusqu'à un seuil.



(3) Joindre les matchs étendus qui sont proches





Match $(S[i, i+a], T[j, j+a])$

$(S[i', i'+b], T[j', j'+b])$

Joindre si $i' > i+a$ $j' > j+a$

et $i' - (i+a) = j' - (j+a) \leq p$

p = paramètre

(4) Retourner les matchs étendus/joints

(5) Exécuter SW sur les matchs trouvés

⑤ Exécuter SW sur les matchs trouvés

⑥ Trier les matchs par score

E-Value : pertinence d'un score S

ε-val(s) = nb espéré d'align. locaux de score $\geq s$
si S et T sont des chaînes aléatoire.

Cette espérance tend vers

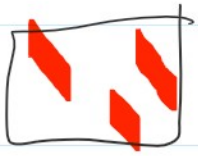
$$\varepsilon\text{-val}(s) \sim \frac{K \cdot |S| \cdot |T|}{e^{\lambda \cdot s}}$$

K, λ dépendent de la matrice de scores

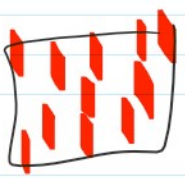
↓ bas : rare
↑ haut = pas rare

ε-val(s) $\leq 10^{-5}$: bon match
$\leq 10^{-100}$: très bon match
≥ 1 : mauvais

Paramètre k : k grand ⇒ − de matches
− de précisim
+ rapide



k petit ⇒ + de matches
+ de precision
− rapide



En pratique : nucléotide : k = 11 $|\Sigma| = 4$

En pratique : nucléotide : $k = 11$ $\quad |\Sigma| = 4$

acide aminés : $k = 5$ $\quad |\Sigma| = 20$