Recherche de motifs : Median String

Cours 4

- Etant donnée un ensemble de t séquences d'ADN, trouver un motif de taille k qui apparaît dans toutes les séquences t avec le nombre minimum de mutations.
- Différence du Brute Force Method :
 - Plutôt que de varier les positions de départ et d'essayer de trouver une séquence consensus représentant un motif,
 - Nous allons au contraire chercher parmi tous les motifs possibles le motif le plus fréquent.

Hamming distance

• Étant donné deux séquences v et x, d_H(v, x) est le nombre de paires de nucléotides qui ne correspondent pas lorsque v et x sont alignés.

$$d_H(AAAAAA, ACAAAC) = 2$$

AAAAA

ACAAAC

Distance total

- Pour chaque séquence d'ADN s, calculer tous $d_H(v, x)$, où x est un **motif** que commence à la position de départ s_i (1 <= i <n k) et v est une séquence consensus de taille k
- Trouver le $d_H(v, x)$ minimum parmi tous les motifs de la séquence s
- TotalDistance (v,ADN) est la somme des distances Hamming minimales pour chaque séquence d'ADN.
 - o distance $(v,ADN) = min d_H(v,x)$,
 - où s est l'ensemble des positions de départ s₁, s₂, ..., s_t

Etant donné v = "acgtacgt" et x ci-dessous

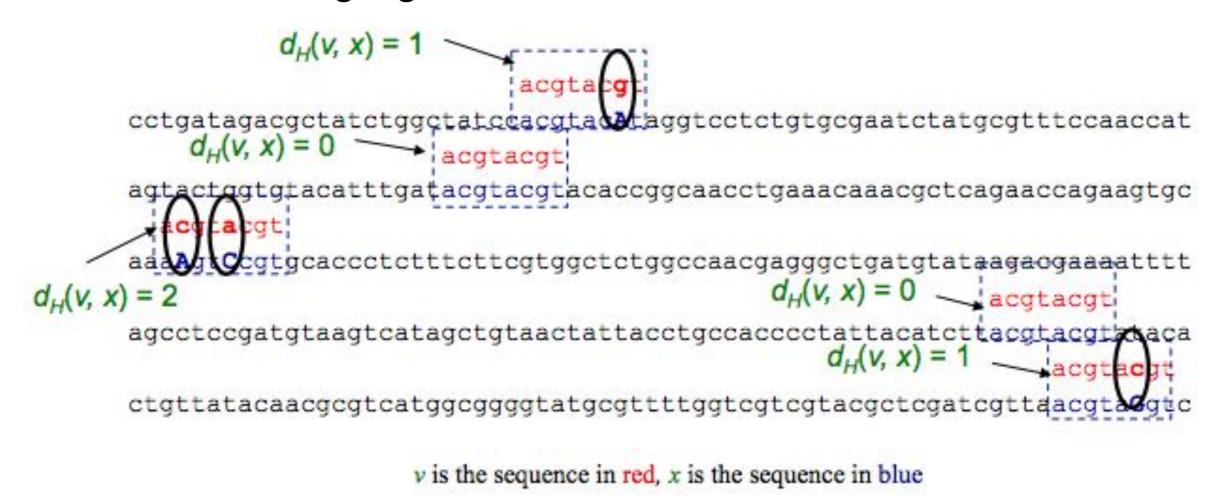
Etant donné v = "acgtacgt" et x ci-dessous

Etant donné v = "acgtacgt" et x ci-dessous

```
d_H(v,\,x)=1 acgtacgt acceptate aggreetet green acceptate aggreetet green acceptate aggreetet green acceptate aggreetet green acceptate acceptat
```

v is the sequence in red, x is the sequence in blue

Etant donné v = "acgtacgt" et x ci-dessous



TotalDistance(v, DNA) =
$$1+0+2+0+1=4$$

Definition formelle

- but: Étant donné un ensemble de séquences d'ADN, trouvez le median string.
- Entrée: Une matrice d'ADN de t x n, et k, la longueur du motif à trouver.
- Sortie: median string v qui minimise TotalDistance (v, ADN) sur toutes les séquences de cette longueur.

- 1. MedianStringSearch (DNA, t, n, k)
- 2. bestWord ← AAA...A
- 3. bestDistance ← ∞
- for each k-mer v from AAA...A to TTT...T if TotalDistance(v,DNA) < bestDistance
- bestDistance←TotalDistance(v,DNA)
- bestWord ← v
- 7. return bestWord

- MedianStringSearch (DNA, t, n, l)
- 2. bestWord ← AAA...A
- 3. bestDistance ← ∞
- 4. for each I-mer v from AAA...A to TTT...T if TotalDistance(v,DNA) < bestDistance
- bestDistance←TotalDistance(v, DNA)
- bestWord ← v
- 7. return bestWord

Pour éviter les comparaisons inutile nous pouvons éliminer les motifs peu complexe

Median String Problem x Brute Force Method

• La méthode "Brute force" nécessités de O(kn)^t

- La méthode "Median String Problem" doit examiner toutes les combinaisons de O(4^k) pour v.
- Ce nombre est typiquement plus petit, mais si k est grand, l'utilisation du algorithme "Median String Problem" sera toujours irréalisable.

A retenir

- L'algorithme median String peut trouver des motifs invariables de taille k dans les séquences régulatrices.
- Il explore tout l'espace de recherche en variant toutes les letters d'un motif de taille k.
- Complexité ~ O (k⁴)