



Recherche de motifs : Median String

Cours 4

The Median String Problem

- 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.

The Median String Problem

- 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.

$$d_H(\text{A}\textcolor{red}{A}\textcolor{red}{A}\textcolor{red}{A}\textcolor{red}{A}\textcolor{red}{A}\textcolor{red}{A}, \text{A}\textcolor{blue}{C}\textcolor{blue}{A}\textcolor{blue}{A}\textcolor{blue}{A}\textcolor{blue}{C}) = 2$$

A~~A~~A~~A~~A~~A~~A~~A~~
A~~C~~A~~A~~A~~C~~

The Median String Problem

- 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 ($1 \leq i \leq n - k$) 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.
 - distance (v, ADN) = $\min d_H(v, x)$,
où s est l'ensemble des positions de départ s_1, s_2, \dots, s_t

Exemple: The Median String Problem

- Etant donné $v = \text{"acgtacgt"}$ et x ci-dessous

```
          acgtacgt
cctgatatagacgctatctggctatccacgtacAtaggtcctctgtgcgaatctatgcgttttccaaccat
          acgtacgt
agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc
acgtacgt
aaaAgtCcggtgcaccctctttcttctcgtggctctggccaacgagggctgatgtataagacgaaaatttt
                                     acgtacgt
agcctccgatgtaagtcatagctgtaactattacctgccacccctattacatcttacgtacgtataca
                                     acgtacgt
ctgttatacaacgcgctcatggcgggggtatgcgtttttggtcgtcgtacgctcgatcgttaacgtaGgtc
```

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

Example: The Median String Problem

- Etant donné $v = \text{"acgtacgt"}$ et x ci-dessous

$d_H(v, x) = 1$



```

cctgatagacgctatctggctatccacgtacgtaggctcctctgtgcgaatctatgcggtttccaaccat
      acgtacgt
agtactggtgtacatttgataacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc
    acgtacgt
aaaAgtCcggtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatttt
                                     acgtacgt
agcctccgatgtaagtcatagctgtaactattacctgccaccctattacatcttacgtacgttataca
                                     acgtacgt
ctgttatacaacgcgctcatggcggggtatgcggttttggtcgtcgtacgctcgatcgttaacgtaGgtc

```

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

Example: The Median String Problem

- Etant donné $v = \text{"acgtacgt"}$ et x ci-dessous

$d_H(v, x) = 1$ → acgtacgt
 $d_H(v, x) = 0$ → acgtacgt
 cctgatatagacgctatctggctatccacgtacgtaggtcctctgtgcgaatctatgcggtttccaacat
 agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc
acgtacgt
 aaaAgtCcggtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatttt
acgtacgt
 agcctccgatgtaagtcatagctgtaactattacctgccaccctattacatcttacgtacgtataca
acgtacgt
 ctggtatacaacgcgctcatggcggggtatgcggttttggtcgtcgtacgctcgatcggttaacgtaGgtc

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

Example: The Median String Problem

- Etant donné $v = \text{"acgtacgt"}$ et x ci-dessous



$$\text{TotalDistance}(v, \text{DNA}) = 1 + 0 + 2 + 0 + 1 = 4$$

The Median String Problem

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

The Median String Problem

1. MedianStringSearch (DNA, t, n, k)
2. $bestWord \leftarrow AAA...A$
3. $bestDistance \leftarrow \infty$
4. for each k -mer v from $AAA...A$ to $TTT...T$ if
 $TotalDistance(v, DNA) < bestDistance$
5. $bestDistance \leftarrow TotalDistance(v, DNA)$
6. $bestWord \leftarrow v$
7. return $bestWord$

The Median String Problem

1. MedianStringSearch (DNA, t, n, l)
2. $bestWord \leftarrow AAA...A$
3. $bestDistance \leftarrow \infty$
4. for each l -mer v from $AAA...A$ to $TTT...T$ if
 $TotalDistance(v, DNA) < bestDistance$
5. $bestDistance \leftarrow TotalDistance(v, DNA)$
6. $bestWord \leftarrow v$
7. return $bestWord$

Pour éviter les comparaisons inutiles nous pouvons éliminer les motifs peu complexes

Median String Problem x Brute Force Method

- La méthode “Brute force” nécessite 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 $\sim O(k^4)$