

Spaced seeds

Michele Beccari 856608

Corso di Bioinformatica

2023

Lo scopo del progetto è individuare uno SNP o un singolo errore su una read r rispetto al reference R .

INPUT: due sequenze r e R . r differisce di uno SNP o errore da una sottostringa di R , cioè $\exists r'$ sottostringa di R tale che $\text{hamming}(r, r') = 1$

OUTPUT: la posizione (se esiste) in r di uno SNP oppure di un errore, ovvero la posizione in cui r differisce per una sottostringa r' da R da cui r può essere derivata.

Esempio

Dato $R = \text{AAAAATCGG}$ e $r = \text{ATAGG}$, chiaramente r differisce in posizione 2 da $r' = \text{ATCGG}$ a causa dello SNP C che sostituisce A.

Una possibile soluzione è utilizzare gli **spaced seeds**

Spaced seed

Uno spaced seed consiste in un k-mer dove alcune posizioni sono indicate come * che sta per “do not care”.

Definiamo un k-mer 1-approssimato se solo una posizione del k-mer è *.

Esempio

AA*GG può matchare con AAAGG e anche con AAGGG.

Abbiamo diversi k -mer 1-approssimati per le varie posizioni da 1 a k .
Ipotizzando che per ogni posizione i di r ci sia uno SNP si può capire se saltando una determinata posizione i di r si ottiene un k -mer che matcha con uno dei k -mers di R .

Utilizzando ntHash2 è possibile indicizzare la stringa R con k -mer 1-approssimati, distinguendo l'hashing per $*$ in posizione $1, 2, 3 \dots k$. Per trovare lo SNP gli step saranno quindi:

- 1 Ottenere tutti k -mer 1-approssimati della stringa r .
- 2 Per ogni k -mer 1-approssimato nella stringa R eseguire un confronto con i k -mer 1-approssimati trovati al punto 1.
- 3 Quando eventualmente si trova una match, verificare il carattere corrispondente alla posizione dell' $"*"$ nella stringa r per trovare lo SNP.

Per generare gli hash dei k-mer 1-approssimati la libreria offre degli oggetti appositi:

```
// Oggetto che genera un hash
// per i k-meri di lunghezza 5 della stringa "ATAGG",
// i seed forniti.
// Gli hash sono degli interi senza segno a 64 bit
nthash::SeedNtHash("ATAGG", seeds, 1, 5)
```

Per generare gli hash con un "*" in una determinata posizione è necessario fornire all'oggetto della libreria uno o più *seed*, ovvero una stringa composta da 0 e 1 dove 0 rappresentano gli asterischi. Potenzialmente è possibile generare anche più di un hash per ogni sequenza per gestire eventuali conflitti.