Michele Beccari 856608 - BioInformatica - Progetto spaced seeds - Giugno 2023

1 Progetto Spaced Seeds

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, cioe $\exists\ r'$ sottostringa di r tale che $hamming(r, r') = 1$
- **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 =AAAAATCGG e r=ATAGG, chiaramente r differisce in posizione 2 da r' = ATCGG a causa dello SNP C che sostituisce A.

1.1 Metodo proposto

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

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

Ad esempio il k-mer AA*GG indica che al posti di * possiamo mettere qualunque simbolo.
Uno spaced seed può matchare diversi k-meri di R.
Ad esempio AA*GG matcha AAAGG oppure AAGGG.

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

Abbiamo diversi k-mer 1-approsimati 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 match con i k-mers memorizzati per R.

1.2 Implementazione

Utilizzando ntHash2 è possibile indicizzare la stringa R con k-mer 1-approssimati, distinguendo l'hashing per * in posizione $1, 2, 3...k$
L'idea è quella di:

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", utilizzando
// 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 gli 0 rappresentano gli asterischi.
Potenzialmente è possibile generare anche più di un hash per ogni sequenza per gestire eventuali conflitti.