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 $R$
- **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 $R$ =AAAAAGGGG e $r$=AACGG, chiaramente $r$ differisce in posizione 3 da $r'$ = AAGGG a causa dello SNP C che sostituisce G.

## 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.

## 1.2 Implementazione

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

Ipotizzando che per ogni posizione $i$ di $r$ ci sia uno SNP o un errore si può utilizzare il risultato precedente per capire se saltando una determinata posiziione $i$ di $r$ si ottiene un match con i k-mers memorizzati per R.