

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 , 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.

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-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 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.