Analisi algoritmo invMap

1. DEFINIZIONE DI INVERSIONE

Nel paper è riportata la definizione di variazione strutturale (SV), che viene poi utilizzata durante tutta la trattazione dell'algoritmo: "Le variazioni strutturali (SVs) sono definite come variazioni nel genoma aventi una lunghezza maggiore di 50bp". La definizione di inversione adottata finora, invece, è una variazione tale per cui la stringa target risulta nell'ordine opposto e complementata rispetto al genoma di riferimento.

2. DEFINIZONE DI STRINGA SPECIFICA

La stringa specifica è stata definite come una più corta sottostringa che occorre nel target T, non occorre nel genoma di riferimento (reference) R, ma tutte le sue sottostringhe occorrono in R.

3. ALGORITMO INVMAP

Il funzionamento dell'algoritmo invMap presentato nel paper si articola in diverse fasi:

- Indicizza il genoma di riferimento attaverso una tabella di hash per permettere di trovare in tempo lineare la posizione di un dato k-mer in input all'interno del genoma di riferimento;

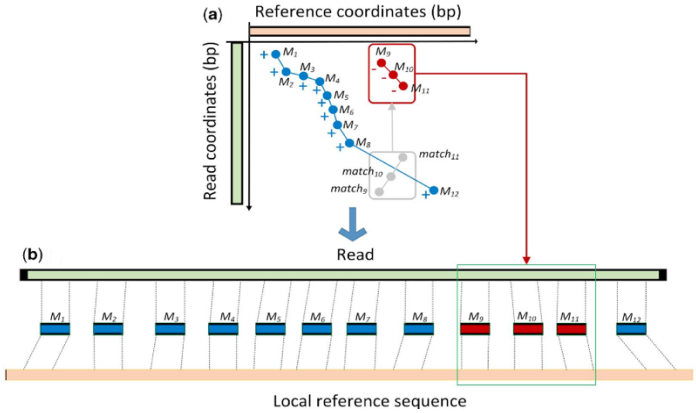
- Trova la catena di allineamento principale, definita come il sottoinsieme di anchor (k-mer che esegue match con il genoma di riferimento) tra loro connesse avente punteggio di allineamento massimo. Tale catena viene determinata utilizzando la programmazione dinamica. Si parte calcolando lo score per ogni anchor, e si prosegue selezionando il percorso tra anchor che massimizza lo score totale, derivando la catena principale di allineamento;

- Una volta trovata, viene localizzata la regione del genoma di riferimento corrispondente alla catena principale della read, e vengono prese in considerazione tutte le anchor appartenenti a tale regione. Dunque, si sa che le anchor che identificano una potenziale inversione non saranno allineate alle restanti della catena principale. Si procede quindi a rendere lineari le anchor che non lo sono, tenendo traccia anche della direzione in cui si allineano al genoma di riferimento ("+" o "-"). Se sono contraddistinte dal -, allora saranno allineate nella direzione invertita rispetto al genoma di riferimento, e quindi non lineari rispetto alla catena principale. A questo punto, invMap calcola il punteggio di score di tutte le anchor che non sono incluse nella catena principale (ma fanno comunque parte della regione allineata, essendo quella presa in considerazione). Si trova quindi una

nuova catena, più corta della principale, avente punteggio di allineamento massimo tra tutte le anchor non appartenenti alla catena principale. Se si trova che tutte le anchor hanno direzione "-", allora si ha una potenziale inversione. In particolare, per essere etichettata come tale, si deve avere che:

- La catena trovata è composta da almeno tre anchor;
- Il segmento di genoma di riferimento individuato da queste anchor è lungo almeno 50bp (per rispettare la definizione di variazione strutturale).

- A questo punto, l'algoritmo procede a generare il punteggio di allineamento dell'intera read fornita in input. Per le anchor interne, si usa la programmazione dinamica attraverso l'applicazione dell'algoritmo di allineamento globale. Per



quelle ai bordi, utilizza l'algoritmo di allineamento con gap.

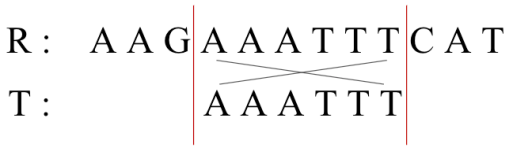
4. ESEMPI DI INVERSIONI INDIVIDUATE DA STRINGHE SPECIFICHE E NON

Basandosi sulla definizione di stringa invertita e complementata utilizzata finora, qualsiasi variazione avvenuta in T rispetto alla porzione corrispondente di R dovrebbe portare alla generazione di almeno una stringa specifica per T, il che significa che dovunque l'inversione porti a mismatch tra R e T, allora i breakpoint della porzione invertita (e complementata)

R : A G A A A A T T G C T C
T : A C A A T T T C T

saranno individuati da una stringa specifica. Nell'esempio in figura, le stringhe specifiche AC e TTC (ma anche TCT) sono esattamente posizionate in corrispondenza dei breakpoint.

Ci sono però anche inversioni che non sono individuate dalle stringhe specifiche, proprio perché non portano alla generazione delle stesse, come tutti quei casi in cui l'inversione e la complementazione della stringa target T rispetto al genoma di riferimento R porta all'ottenimento di una stringa T identica alla porzione corrispondente di R.



Proprio perché non comportano alcuna modifica rispetto al genoma di riferimento, c'è da chiedersi se questi casi siano effettivamente rilevanti al fine della trattazione delle variazioni strutturali.

5. INTEGRAZIONE DI INVMAP CON LE STRINGHE SPECIFICHE

Si potrebbe implementare un approccio molto simile a quanto avviene con invMap dell'individuazione di inversioni mediante stringhe specifiche, nei casi in cui effettivamente queste ultime portino all'individuazione dei breakpoint del segmento invertito. Si potrebbe infatti confrontare lo score di allineamento della porzione interna ai breakpoint con R con quello della stessa stringa ma "linearizzata", ovvero invertita e complementata. Se il punteggio di questa ultima risultasse significativamente maggiore, allora sarebbe molto probabile una inversione.