Algoritmi per la trasformata di Burrows–Wheeler posizionale con compressione run-length

Davide Cozzi
matr. 829827

Slide 1

Buongiorno, sono Davide Cozzi e oggi presento la mia tesi dal titolo: "Algoritmi per la trasformata di Burrows–Wheeler posizionale con compressione run-length"

Slide 2

Negli ultimi anni si è assistito a un cambio di paradigma nel campo della bioinformatica, ovvero il passaggio dallo studio della sequenza lineare di un singolo genoma a quello di un insieme di genomi, provenienti da un gran numero di individui, al fine di poter considerare anche le varianti geniche. Questo nuovo concetto è stato introdotto da Tettelin, nel 2005, con il termine di pangenoma. Grazie ai risultati ottenuti in pangenomica, ci sono stati miglioramenti sia nel campo della biologia che in quello della medicina personalizzata, grazie al fatto che, con il pangenoma, si migliora la precisione della rappresentazione di multipli genomi e delle loro differenze. Il genoma umano di riferimento (GRCh38.p14), è composto da circa 3.1 miliardi di basi, con più di 88 milioni varianti tra i genomi sequenziati, secondo i risultati ottenuti nel 1000 Genome Project. Considerando che, grazie al miglioramento delle tecnologie di sequenziamento, la quantità dei dati di sequenziamento sia destinata ad aumentare esponenzialmente nei prossimi anni, risulta necessaria la costruzione di algoritmi e strutture dati efficienti per gestire una tale mole di dati. A questo scopo, uno degli approcci più usati per rappresentare il pangenoma è attraverso un pannello di aplotipi, ovvero, da un punto di vista computazionale, una matrice di M righe, corrispondenti agli individui, e N colonne, corrispondenti ai siti con le varianti. Si specifica che, con il termine aplotipo, si intende l'insieme di alleli, ovvero di varianti che, a meno di mutazioni, un organismo eredita da ogni genitore. In questo contesto trova spazio uno dei problemi fondamentali della bioinformatica, ovvero quello del pattern matching. Inizialmente tale problema era relativo alla ricerca di una stringa (pattern) all'interno di un testo di grandi dimensioni, cioè il genoma di riferimento. Ora, con l'introduzione del pangenoma, il problema deve essere risolto sulle nuove strutture di rappresentazione del pangenoma.

Slide 3

Lo scopo di questa tesi è progettare strutture dati e algoritmi efficienti per risolvere il problema del pattern matching, inteso come ricerca dei set-maximal exact match (SMEM) tra un aplotipo esterno e un pannello di aplotipi, in una delle strutture dati più utilizzata per la rappresentazione del pangenoma: la trasformata di Burrows–Wheeler Posizionale (PBWT).
Questo progetto, svolto in collaborazione con il laboratorio BIAS e con diversi ricercatori internazionali (University of Florida, Dalhousie University e Tokyo Medical and Dental University), permetterà la gestione e lo studio (ad esempio nei GWAS) dei sempre più grandi dati provenienti dalle tecnologie di sequenziamento. Inoltre, con tale progetto, si è confermata l'ovvia correlazione tra la BWT e la PBWT, estendendo tale correlazione anche alle rispettive varianti run-length.

Slide 4

In questa breve presentazione è impossibile entrare nei dettagli di tutti i concetti teorici alla base di questo progetto. Tra di essi si hanno:

- bitvector e bitvector sparsi, strutture succinte alla base del lavoro
- intvector compressi, strutture compresse che hanno permesso di lavorare con valori interi
- straight-line program (SLP) e longest common extension (LCE) query, una grammatica context-free compressa che permette random access e LCE query in tempo logaritmico
- trasformata di Burrows–Wheeler (BWT), (inverse) suffix array ((I)SA), (permuted) longest common prefix ((P)PLCP), funzione φ, FM-index, LF-mapping, maximal exact matches (MEM), e tutte le altre teorie allo stato dell'arte su questa trasformata
- trasformata di Burrows–Wheeler run-length encoded (RLBWT), r-index, Toheold lemma, matching statistics (MS) e tutti i più recenti studi sull'uso del run-length encoded
- trasformata di Burrows–Wheeler posizionale (PBWT) e set-maximal exact match (SMEM), che invece per ovvie ragioni vedremo un po' più nel dettaglio

Slide 5

In merito alla RLBWT bisogna citare due recenti lavori, che sfruttano tale trasformata per calcolare le matching statistics e da qui calcolare MEM. Il primo è MONI (di Rossi et al.), che sfrutta il concetto di threshold (minimo lcp in una run) e il random access al pannello per il calcolo delle matching statistics.
Il secondo è PHONI (di Boucher, Rossi et al.), che sfrutta invece le LCE query per

fare il calcolo in una singola passata sul pattern, ottimizzando ancor di più la memoria necessaria.
Tali lavori sono da citare in quanto, in questo progetto, si sono create le varianti ispirate ad entrambi i lavori per la PBWT. A tal fine, come vedremo, tutti i concetti teorici della RLBWT sono stati ripensati in ottica posizionale.

Slide 6

La Trasformata di Burrows–Wheeler Posizionale (PBWT), presentata da Durbin nel 2014, viene costruita a partire da un pannello di aplotipi, rappresentato, riferendosi al solo caso biallelico, tramite una matrice binaria. La motivazione essenziale della PBWT è considerare match, e quindi anche SMEM, dove anche le posizioni di inizio e fine sono rispettate. Tale vincolo, da cui deriva il termine "posizionale", non è soddisfacibile dalla BWT ed è dovuto al fatto che ogni colonna (o indice della query) rappresenta un preciso sito per una specifica variante genica. Il funzionamento della PBWT prevede la costruzione di due insiemi di array, tramite l'ordinamento dei prefissi inversi a ogni colonna del pannello, detti insieme dei prefix array (che tiene traccia degli indici degli ordinamenti) e insieme dei divergence array (che tiene traccia della colonna d'inizio del prefisso inverso più lungo tra una riga e la precedente nel riordinamento ad una certa colonna). Il pannello, permutato tramite l'insieme dei prefix array, è detto matrice PBWT.
Qui, ad esempio, vediamo la costruzione della trasformata alla colonna 6, basata sul riordinamento fino alla quinta, e la conseguente produzione dei due array. Si noti che il divergence può anche essere sostituito dal Reverse Longest Common prefix, che memorizza la lunghezza del prefisso comune.

Slide 7

La PBWT permette di calcolare gli SMEM, di cui un esempio è qui disponibile, tra un aplotipo esterno e il pannello in tempo Avg. $\mathcal{O}(N + c)$ (dove c è il numero complessivo di SMEM), mentre una soluzione semplice impiegherebbe $\mathcal{O}(N^2M)$ tramite il famoso algoritmo 5 di Durbin che si basa sul mantenere ed eventualmente estendere un intervallo sui prefix array che contiene gli indici delle righe che hanno uno SMEM fino a quella colonna. Se l'intervallo non è più estendibile si reporta lo SMEM e si sfrutta il divergence array per computare il nuovo intervallo. Il tradeoff di questo algoritmo è la richiesta in termini di spazio (13NM bytes), dovuto ad ulteriori array necessari in memoria per il "mapping", ovvero il forward step, tra una colonna e la successiva nella matrice PBWT. Superare questo limite è l'obiettivo principale di questo progetto di tesi.

Slide 8