

# Algoritmi per la trasformata di Burrows-Wheeler Posizionale con compressione run-length

Davide Cozzi  
matr. 829827

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.

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)*. Il progetto di tesi, svolto in collaborazione con Prof. Gagie (Dalhousie University) e Prof. Boucher (University of Florida), ha quindi permesso lo sviluppo di una variante **run-length encoded** della **PBWT**, detta **RLPBWT**, che risolve questo problema riducendo la memoria impiegata rispetto alla soluzione attualmente allo stato dell'arte.

## Stato dell'arte

Gli algoritmi più efficienti per risolvere il problema del pattern matching tra un pattern e un testo sono attualmente basati sulla *Trasformata di Burrows-Wheeler (BWT)*, a sua volta strettamente legata ad altre due strutture dati: il Suffix Array (SA) e il Longest Common Prefix (LCP).

Invece, allo scopo di risolvere il problema del calcolo degli SMEM tra un pannello di aplotipi e un aplotipo esterno, Durbin, nel 2014, propose una struttura dati ispirata alla BWT, detta *Trasformata di Burrows-Wheeler Posizionale (PBWT)*. Tale trasformata viene costruita a partire da un pannello di aplotipi, rappresentato, riferendosi al solo caso bi-allelico, tramite una matrice binaria. La PBWT permette di calcolare gli SMEM 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  $\mathcal{O}(N^2M)$ . 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 tradeoff di questo algoritmo è la richiesta in termini di spazio ( $13NM$  bytes). Superare questo limite è l'obiettivo principale di questo progetto di tesi. 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* e *insieme dei divergence array*. Il pannello, permutato tramite l'insieme dei prefix array, è detto matrice PBWT.

Un altro obiettivo raggiunto negli ultimi anni è stato quello di studiare la costruzione di un'unica BWT partendo da multipli genomi. A tal fine si è sfruttata una caratteristica della BWT, ovvero la tendenza a produrre una sequenza formata da run, ovvero caratteri uguali in posizioni consecutive. La causa di tale fenomeno si ritrova nelle ripetizioni di determinate sottostringhe nel testo, ripetizioni che sono frequenti se si studiano diverse sequenze genomiche concatenate. In letteratura è stata quindi proposta la *trasformata di Burrows-Wheeler Run-Length encoded (RLBWT)* dove ogni run viene memorizzata in modo efficiente come coppia (carattere, lunghezza della run). Ad esempio, la stringa `aaaaaa` sarebbe memorizzata come `(a,6)`. Recentemente, per questa struttura dati compressa, è stato proposto un nuovo tipo di indicizzazione: il cosiddetto *r-index*. Tale indice riduce lo spazio di memoria richiesto perché non è lineare sulla lunghezza del testo ma sul numero di run della sua BWT. L'*r-index* include la RLBWT e un suffix array sample, ovvero i valori del SA all'inizio e alla fine di ogni run. Tali articoli hanno portato alla produzione di due tool, *MONI* e *PHONI*, alle cui tecniche è fortemente ispirata questa tesi. Queste due soluzioni hanno entrambe l'obiettivo di calcolare i maximal exact match (MEM) tra un testo e un pattern, ovvero sottostringhe del pattern che sono anche sottostringhe del testo, avendo che tale match non può essere esteso, in entrambe le direzioni, senza introdurre un mismatch.

Il calcolo dei MEM è direttamente correlato alla costruzione dell'array delle *Matching Statistics (MS)*. Tale array, lungo quanto il pattern e formato da coppie (posizione *pos*, lunghezza *len*), annota, per ogni posizione del pattern, la lunghezza *len* di un match, anche non massimale, con una sottostringa nel testo avente indice iniziale *pos*. Estruendo

da tali match i MEM si possono poi ottenere tutte le altre occorrenze del medesimo MEM nel testo. In MONI, per il calcolo delle MS, è stato usato anche il concetto di *threshold*, definito come il minimo valore dell'array LCP tra due run consecutive dello stesso carattere. Un ulteriore miglioramento si è registrato in PHONI con l'uso delle *LCE query* per il calcolo dell'array MS. Infatti, tale struttura, dati due indici del testo  $i$  e  $j$ , restituisce il più lungo prefisso comune tra i suffissi  $i$ -esimo e  $j$ -esimo del testo.

## Contributo

Questo lavoro di tesi è stato incentrato sullo sviluppo di approcci run-length encoded per la PBWT, ispirandosi ai risultati già ottenuti per la RLBWT. Essendoci diverse soluzioni possibili, si è deciso di suddividere il lavoro in varie strutture dati (componenti) che fornissero strutture dati composte atte al calcolo degli SMEM.

Alcuni di questi approcci si basano sull'uso dei bitvector sparsi, una struttura dati succinta che permette di memorizzare e interrogare vettori binari efficientemente. Infatti, tali strutture permettono due operazioni, dette rank e select, in tempo costante. La prima permette di conteggiare il numero di simboli  $\sigma = 1$  fino ad una certa posizione mentre la seconda di ottenere l'indice sul bitvector dell' $i$ -esimo simbolo  $\sigma = 1$ . Quindi, le strutture necessarie all'indicizzazione delle run e al cosiddetto mapping, ovvero il "seguire" una riga dalla sua posizione permutata in una certa colonna alla posizione permutata nella colonna successiva, sono state implementate, anche, tramite bitvector sparsi (componente MAP-BV), con un numero di simboli  $\sigma = 1$  proporzionale al numero run. È stata anche proposta, d'altro canto, una struttura per il mapping basata sull'uso di vettori di interi compressi, con un numero di valori proporzionale a quello delle run (componente MAP-INT).

Al fine di avvicinarsi alle idee proposte in MONI e PHONI è quindi servito uno studio teorico preliminare per ridefinire: matching statistics (e conseguente calcolo degli SMEM), threshold e LCE query. In particolare, grazie agli ultimi due concetti, si è potuto evitare di memorizzare il divergence array. Le matching statistics sono definite tramite un array, lungo quanto il pattern e formato da coppie (riga *row*, lunghezza *len*), che, per ogni colonna, tiene traccia del suffisso comune che non sia estendibile a sinistra, terminate in quella colonna e lungo *len*, tra la riga *row* del pannello e l'aplotipo query. Una volta calcolato tale array è possibile estrarre uno SMEM, terminate in colonna  $k$  tra la query e la riga *row*, ed estendere, tramite una struttura dati a supporto, tale risultato a tutte le altre righe del pannello che presentano il medesimo SMEM. Il primo metodo di calcolo di tale array è stato basato sull'utilizzo delle threshold. Ragionando sulla matrice PBWT, una threshold è definita come l'indice del massimo valore del divergence array all'interno di una run (comprendendo anche la testa, qualora esistente, della run successiva, essendo il suo valore del divergence array calcolato sfruttando anche la coda della run corrente). L'insieme delle threshold per una certa colonna è stato memorizzato come bitvector sparso (componente THR-BV) oppure, come per il mapping, tramite vettore di interi compresso (componente THR-INT). Inoltre, si è provveduto a tenere in memoria i sample di prefix array a inizio e fine di ogni run (componente PERM) e una struttura (componente PHI) in grado di estendere uno SMEM a tutte le righe del

pannello che presentano il medesimo SMEM con il pattern.

Grazie all'uso delle *threshold* si è potuto sviluppare un algoritmo efficiente, dal punto di vista della memoria richiesta, per il calcolo delle matching statistics in due “sweep” sul pattern, calcolandone prima le posizioni e poi, tramite random access al pannello, anche le lunghezze. Tale soluzione richiedeva in memoria l'intero pannello, per il quale si è scelto di usare uno Straight-Line Program (SLP), ovvero una grammatica context-free che genera una e una sola parola, molto compatta in memoria, sulla quale è possibile effettuare random access (in tempo, da un punto di vista prettamente teorico,  $\mathcal{O}(\log s)$  dove  $s$  è la lunghezza della parola prodotta tramite l'SLP). Ad esempio, un pannello  $4.908 \times 6.196.151$  richiede, in forma non compressa, circa 28GB di memoria, mentre l'SLP richiede solamente in 0,2GB (garantendo random access). Per riferimento, comprimere il pannello con una tecnica standard (GZip) richiede 0,5GB. Oltre all'uso dell'SLP (componente RA-SLP), si è anche proposto l'uso di un pannello di bitvector per la memorizzazione del pannello (componente RA-BV), avendo maggior spazio in memoria (3,6GB) ma minor tempo d'interrogazione, garantendo random access in tempo costante.

Inoltre, è possibile risparmiare ulteriore spazio eliminando l'uso delle *threshold* e usando, per il calcolo delle matching statistics, le LCE query (componente LCE), ridefinite, in questo caso, come il suffisso comune più lungo tra due righe del pannello, eventualmente fissando il termine a una colonna precisata. Il loro calcolo è permesso in modo efficiente dall'uso dell'SLP. Calcolando la lunghezza di tale suffisso comune, è possibile computare le lunghezze delle matching statistics contemporaneamente al calcolo delle righe. Quindi, si è ottenuto il calcolo completo di tale array in un singolo “sweep” sul pattern e, avendo che il calcolo degli SMEM viene fatto contemporaneamente al calcolo delle lunghezze, si è ridotto lo spazio necessario in quanto sono sufficienti la coppia (riga *row*, lunghezza *len*) corrente e quella precedente dell'array delle matching statistics per calcolare gli SMEM.

## Risultati sperimentali

L'obiettivo della fase sperimentale è stato quello di confrontarsi con l'algoritmo di Durbin per il calcolo degli SMEM sulla PBWT al fine (1) di verificare se le soluzioni per la RLPBWT riducevano la memoria necessaria al calcolo degli SMEM e (2) di quantificare l'aumento dei tempi di calcolo. Infatti, per quanto riguarda (1), la caratterizzazione asintotica teorica non è sufficientemente stringente da garantire una riduzione in pratica, in quanto fortemente dipendente dalle caratteristiche dei dati impiegati. Per quanto riguarda (2), un aumento dei tempi è atteso in quanto la RLPBWT, a differenza dell'algoritmo di Durbin, utilizza strutture dati succinte che comportano tempi di calcolo superiori per le operazioni di accesso e/o interrogazione. I test sono stati effettuati su 5 pannelli della phase 3 del 1000 Genome Project, uno dei più importanti progetti di catalogazione delle variazioni geniche umane. Tali pannelli sono relativi ai cromosomi 22, 20, 18, 16 e 1 (scelto in quanto è il più grande). Tutti i pannelli presentano 5008 sample. Per questa sintesi, si è scelto di riportare alcuni risultati relativi al pannello del cromosoma 1, di dimensioni  $4.908 \times 6.196.151$ , che è stato interrogato con 100 query (estratte dal pannello originale), al fine di ridurre la fluttuazione statistica dei risultati. I

risultati qui riportati, relativi alla PBWT, sono stati ottenuti usando l'implementazione ufficiale della stessa. Per la RLPBWT si riportano, per onor di sintesi, solo i risultati dell'implementazione con le LCE query e mapping tramite vettori di interi compressi, essendo la soluzione con minor richiesta di memoria.

In merito al tempo di calcolo si è ottenuto il risultato atteso: con la PBWT si ottiene il calcolo degli SMEM in 1026s mentre con la RLPBWT in 1142s. La differenza, per quanto attesa, è comunque molto piccola. Si è passato, quindi, ad analizzare i risultati in termini di memoria. Secondo le stime di Durbin, gli array necessari al funzionamento del suo algoritmo richiederebbero  $\sim 13NM$  bytes, implicando, per il pannello in analisi, circa 368GB di memoria richiesti. Sperimentalmente si sono registrati picchi di memoria prossimi ai 369GB di memoria, confermando le stime. Invece, per la variante della RLPBWT presa in esame, il picco registrato è stato di appena 4GB, ottenendo una riduzione di quasi il 99% di memoria. Tale riduzione renderebbe possibile l'esecuzione del calcolo anche su un comune portatile, assumendo di aver a disposizione l'SLP.

## Conclusioni e sviluppi futuri

In conclusione, le strutture dati composte per la RLPBWT, basate sul calcolo dell'array delle matching statistics, sono state tutte in grado, al variare della soluzione, di ridurre sensibilmente la quantità di memoria necessaria al problema del calcolo degli SMEM tra un pannello di aplotipi e un aplotipo esterno.

Seppure questa tesi abbia raggiunto l'obiettivo prefissato, il lavoro può essere ulteriormente esteso in diverse direzioni. Tra di esse, si ha un'ulteriore ottimizzazione della struttura, sia in termini di gestione del mapping che, eventualmente, di gestione di più query contemporaneamente. In merito a quest'ultimo sviluppo, si segnala che Durbin ha proposto anche un algoritmo per il calcolo degli SMEM tra un pannello di aplotipi e un pannello di query. Questo algoritmo è basato sulla creazione di un unico pannello, su cui costruire la PBWT, e sul calcolo dei match interni al pannello stesso. Tale implementazione, nel setup di test descritto nella sezione precedente (che si noti comportare il perfetto input per tale soluzione), risulta essere molto performante sia in termini di tempo (93s) che di memoria utilizzata (0,1GB) durante l'esecuzione. Si segnala che, rispetto ai tempi di calcolo, tale algoritmo diventa inefficiente, rispetto alle altre soluzioni proposte, al diminuire del numero di query.

Inoltre, sono possibili diverse generalizzazioni rispetto alle caratteristiche del pannello, come ad esempio pannelli multi-allelici o con dati mancanti. Tali caratteristiche sono, infatti, comuni in dati reali. In particolare, la gestione di dati mancanti, molto rilevante a causa della non perfezione delle tecnologie di sequenziamento, risulta essere un problema aperto in bioinformatica. Sarà quindi richiesto lo sviluppo di nuove metodologie, basate, eventualmente, su algoritmi parametrici o algoritmi approssimati, per la risoluzione del problema del calcolo degli SMEM su pannelli di questo tipo.

Le potenzialità di tale struttura sono quindi molteplici e, grazie al ridotto consumo di memoria, si hanno le giuste premesse perché venga utilizzata per gestire e interrogare grosse moli di dati reali, incrementando le capacità di studio, previsione e inferenza che si possono avere grazie allo studio del pangenoma.