

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

Davide Cozzi

Relatore: *Prof. Raffaella Rizzi* **Correlatore:** *Dr. Yuri Pirola*

*Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo)
Università degli Studi di Milano Bicocca*

25 Ottobre 2022

Outline

- 1 Introduzione
- 2 Preliminari
- 3 Metodo
- 4 Risultati sperimentali
- 5 Conclusioni e sviluppi futuri

Outline

- 1 Introduzione
- 2 Preliminari
- 3 Metodo
- 4 Risultati sperimentali
- 5 Conclusioni e sviluppi futuri

Un punto di vista per il pangenoma

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**.

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.

Un **aplotipo** è l'insieme di alleli, ovvero di varianti che, a meno di mutazioni, un organismo eredita da ogni genitore.

Outline

- 1 Introduzione
- 2 Preliminari**
- 3 Metodo
- 4 Risultati sperimentali
- 5 Conclusioni e sviluppi futuri

MONI e PHONI

MONI

Rossi et al., nel 2021, sfruttarono le conoscenze relative alla RLBWT e all'r-index per ideare **MONI**. In questa soluzione si ha la costruzione, in due sweep, tramite l'uso delle threshold (*algoritmo di Bannai*), dell'array delle matching statistics, da cui si computano i maximal exact match.

Rossi et al.: *MONI: A pangenomic index for finding maximal exact matches, 2021*

PHONI

Nel 2021, Boucher, Gagie, Rossi et al. proposero un ulteriore miglioramento di quanto fatto in MONI, con **PHONI**, usando le LCE query al posto delle threshold, ottenendo un algoritmo “online”.

Boucher et al.: *PHONI: Streamed matching statistics with multi-genome references, 2021*

PBWT

| X | 00 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 14 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 15 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 00 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 09 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 10 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 16 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 08 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| 11 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 12 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 13 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 18 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 19 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 01 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 02 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 03 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 17 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 04 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 05 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 06 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 07 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |

$a_6 = [14, 15, 0, 9, 10, 16, 8, 11, 12, 13, 18, 19, 1, 2, 3, 17, 4, 5, 6, 7]$

$d_6 = [6, 0, 4, 2, 0, 0, 5, 0, 0, 0, 3, 0, 4, 0, 0, 6, 4, 0, 0, 0]$

Durbin: *Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT)*, 2014

Set-maximal exact match

Calcolo SMEM via **algoritmo 5 di Durbin** in tempo $\mathcal{O}(NM) + \text{Avg.}\mathcal{O}(N + c)$,
richiedendo $13NM$ byte

| X | 00 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 00 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 01 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 02 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 03 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 04 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 05 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 06 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 07 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 08 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| 09 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 10 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 11 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 12 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 13 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 14 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 15 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 16 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 17 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 18 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 19 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| z | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |

Outline

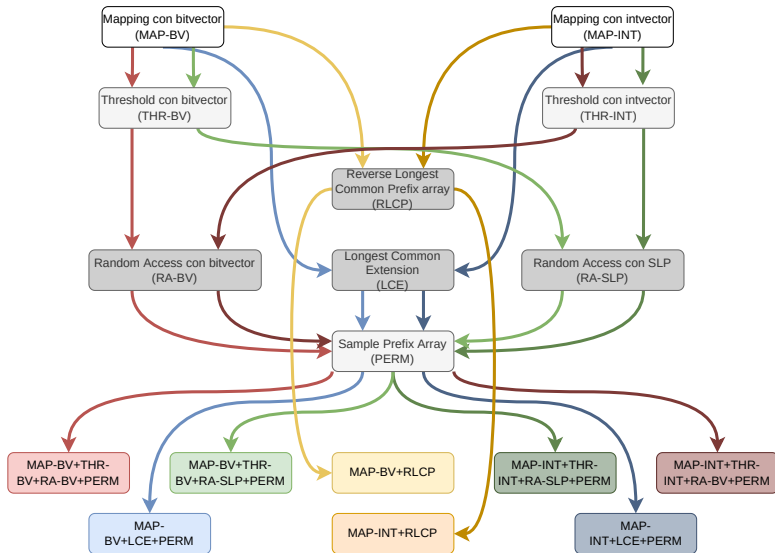
- 1 Introduzione
- 2 Preliminari
- 3 Metodo**
- 4 Risultati sperimentali
- 5 Conclusioni e sviluppi futuri

Le componenti

Componenti derivate dallo studio della RLBWT in ottica PBWT

- componenti per il mapping tra una colonna e la successiva nella PBWT, tramite bitvector sparsi (MAP-BV) o intvector compressi (MAP-INT)
- componenti per le threshold (THR-BV/THR-INT)
- componente per i prefix array sample (PERM)
- componenti per il random access, tramite pannello di bitvector (RA-BV) tramite SLP (RA-SLP)
- componente per le LCE query con SLP (LCE)
- componente per il calcolo delle funzioni φ e φ^{-1} (PHI)
- componente per il reverse longest common prefix (RLCP)

Componenti e strutture dati, una panoramica



Matching statistics

| X | 00 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 00 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 01 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 02 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 03 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 04 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 05 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 06 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 07 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 08 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| 09 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 10 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 11 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 12 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 13 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 14 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 15 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 16 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 17 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 18 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 19 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| z | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |

| k | 00 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| row | 19 | 19 | 16 | 15 | 13 | 13 | 19 | 19 | 19 | 19 | 11 | 11 | 17 | 17 | 17 |
| len | 1 | 2 | 3 | 4 | 5 | 6 | 4 | 5 | 6 | 7 | 4 | 5 | 2 | 3 | 4 |

Outline

- 1 Introduzione
- 2 Preliminari
- 3 Metodo
- 4 Risultati sperimentali**
- 5 Conclusioni e sviluppi futuri

Sperimentazione e dati

Implementazione e sperimentazione

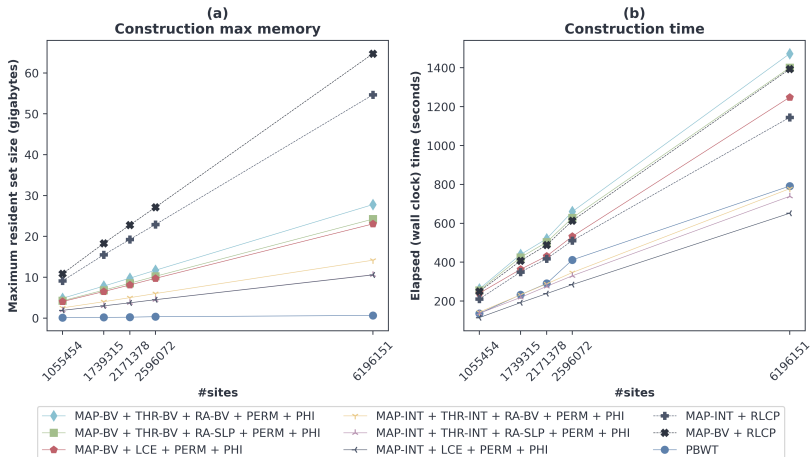
La sperimentazione, orchestrata tramite `snakemake`, è stata effettuata su una macchina con processore Intel Xeon E5-2640 V4 (2,40GHz), 756GB di RAM, 768GB di swap e sistema operativo Ubuntu 20.04.4 LTS.

Si sono confrontate l'implementazione in C++ della RLPBWT e l'implementazione in C ufficiale della PBWT.

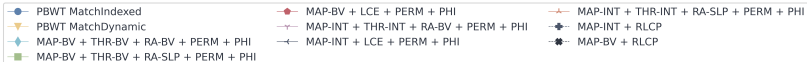
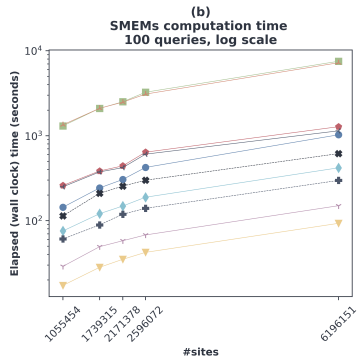
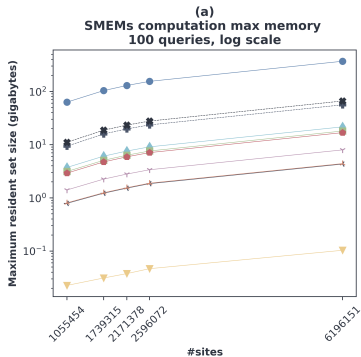
Pannelli del 1000 Genome Project con 4908 sample, avendone estratti 100 come query.

| Chr | #Siti | #Run totale | Max run | Media run |
|-------|-----------|-------------|---------|-----------|
| chr22 | 1.055.454 | 14.772.105 | 2.450 | 14 |
| chr20 | 1.739.315 | 19.966.504 | 2.176 | 11 |
| chr18 | 2.171.378 | 24.288.263 | 2.365 | 11 |
| chr16 | 2.596.072 | 31.187.856 | 2.330 | 12 |
| chr1 | 6.196.151 | 69.671.952 | 2.721 | 11 |

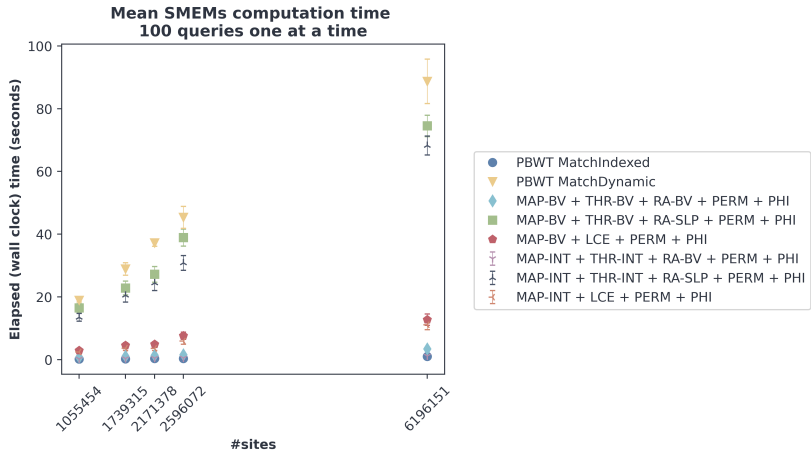
Performance costruzione strutture dati



Performance calcolo degli SMEM con 100 query



Performance calcolo degli SMEM per singole query



Outline

- 1 Introduzione
- 2 Preliminari
- 3 Metodo
- 4 Risultati sperimentali
- 5 Conclusioni e sviluppi futuri**

Considerazioni e sviluppi

Alcune considerazioni

- le strutture dati e gli algoritmi proposti hanno confermato la potenzialità dell'uso di strutture run-length encoded in pangenomica
- l'obiettivo della tesi, ovvero lo sviluppo di un algoritmo, efficiente in spazio, per il calcolo degli SMEM di un aplotipo esterno contro un pannello, è stato raggiunto con risultati molto interessanti

Sviluppi futuri

- SMEM interni con RLPBWT
- RLPBWT multiallelica
- RLPBWT con dati mancanti
- calcolo K-SMEM con RLPBWT

Considerazioni e sviluppi

Alcune considerazioni

- le strutture dati e gli algoritmi proposti hanno confermato la potenzialità dell'uso di strutture run-length encoded in pangenomica
- l'obiettivo della tesi, ovvero lo sviluppo di un algoritmo, efficiente in spazio, per il calcolo degli SMEM di un aploipo esterno contro un pannello, è stato raggiunto con risultati molto interessanti

Sviluppi futuri

- SMEM interni con RLPBWT
- RLPBWT multiallelica
- RLPBWT con dati mancanti
- calcolo K-SMEM con RLPBWT

Bonizzoni, Boucher, Cozzi, Gagic, Kashgouli, Köppl e Rossi:

Compressed data structures for population-scale positional Burrows–Wheeler transforms, 2022

Grazie per l'attenzione

Davide Cozzi

Relatore: *Prof. Raffaella Rizzi* **Correlatore:** *Dr. Yuri Pirola*

*Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo)
Università degli Studi di Milano Bicocca*

25 Ottobre 2022