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

Outline

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**.

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 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

BV e SLP

BV

0	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0	**1**	0	**1**	0	**1**	0	**1**	0	0	**1**	0

$$\texttt{rank}(6) = 3 \qquad \texttt{select}(5) = 9$$

BV e SLP

BV

0	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0	**1**	0	**1**	0	**1**	0	**1**	0	0	**1**	0

$$\texttt{rank}(6) = 3 \qquad \texttt{select}(5) = 9$$

SLP

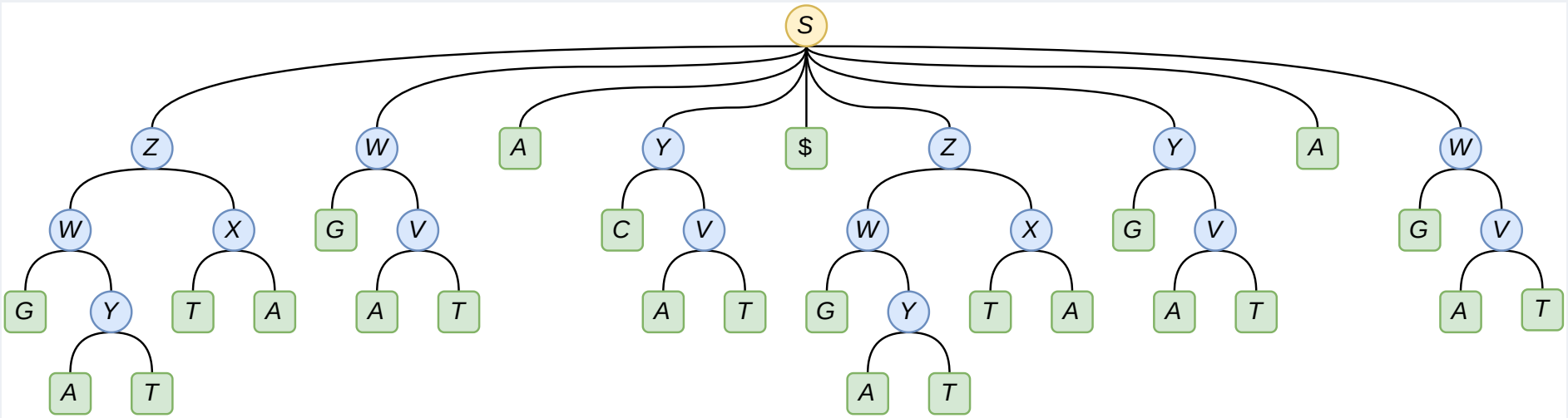
$$s = \text{GATTAGATACAT\$GATTACATAGAT}$$

$$\text{S} \rightarrow \text{ZWAY\,\$ZYAW}$$

- $\text{Z} \rightarrow \text{WX}$
- $\text{Y} \rightarrow \text{CV}$
- $\text{X} \rightarrow \text{TA}$
- $\text{W} \rightarrow \text{GV}$
- $\text{V} \rightarrow \text{AT}$

BV e SLP

SLP



RLBWT

Esempio[1]

Thresholds							
A	T	SA	SA sample	BWT	Run heads	LCP	$\mathcal{M}$
		15	15	A	A	0	$ATTAGATTACATTA
*		14	14	T	T	0	A$ATTAGATTACATT
		9		T		1	ACATTA$ATTAGATT
		4		T		1	AGATTACATTA$ATT
		11	11	C	C	1	ATTA$ATTAGATTAC
		6	6	G	G	4	ATTACATTA$ATTAG
		1	1	$	$	4	ATTAGATTACATTA$
	*	10	10	A	A	0	CATTA$ATTAGATTA
		5		A		0	GATTACATTA$ATTA
*		13	13	T	T	0	TA$ATTAGATTACAT
		8		T		2	TACATTA$ATTAGAT
		3		T		2	TAGATTACATTA$AT
		12	12	A	A	1	TTA$ATTAGATTACA
		7		A		3	TTACATTA$ATTAGA
		2		A		3	TTAGATTACATTA$A

MS e MEM

MS

Dato un testo T, con $|T| = n$, e un pattern P, con $|P| = m$, si definisce **matching statistics** di P su T un array MS di coppie (pos, len), lungo quanto il pattern, tale che:

- $T[MS[i].pos, MS[i].pos + MS[i].len - 1] = P[i, i + MS[i].len - 1]$, quindi si ha un match tra P e T lungo $MS[i].len$ a partire da $MS[i].pos$ in T e da i in P
- $P[i, i + MS[i].len]$ non occorre in T, quindi il match non è ulteriormente estendibile

MS e MEM

MS

Dato un testo T, con $|T| = n$, e un pattern P, con $|P| = m$, si definisce **matching statistics** di P su T un array MS di coppie (pos, len), lungo quanto il pattern, tale che:

- $T[MS[i].pos, MS[i].pos + MS[i].len - 1] = P[i, i + MS[i].len - 1]$, quindi si ha un match tra P e T lungo $MS[i].len$ a partire da $MS[i].pos$ in T e da i in P
- $P[i, i + MS[i].len]$ non occorre in T, quindi il match non è ulteriormente estendibile

MEM

Dato un testo T, con $|T| = n$, e un pattern P, con $|P| = m$, si definisce una sottostringa $P[i, i + l - 1]$, di lunghezza l, **MEM** di P in T se:

- $P[i, i + l - 1]$ è una sottostringa di T
- $P[i - 1, i + l - 1]$ non è una sottostringa di T (non si può estendere a sinistra) e $P[i, i + l]$ non è una sottostringa di T (non si può estendere a destra)

Un MEM si può calcolare dalle MS:

$$MS[i].len = l \wedge MS[i - 1].len \leq MS[i].len$$

MONI e PHONI

MONI

Rossi et al., nel 2021, sfruttarono tutte le conoscenze relative alla **RLBWT**, all'**r-index** e alle **matching statistics** per ideare **MONI: *A Pangenomics Index for Finding MEMs*** [2]. In questa soluzione si ha quindi la costruzione, in due *sweep*, tramite l'uso delle *threshold* (**algoritmo di Bannai**), dell'array delle *matching statistics*.

MONI e PHONI

MONI

Rossi et al., nel 2021, sfruttarono tutte le conoscenze relative alla **RLBWT**, all'**r-index** e alle **matching statistics** per ideare **MONI: *A Pangenomics Index for Finding MEMs*** [2]. In questa soluzione si ha quindi la costruzione, in due *sweep*, tramite l'uso delle *threshold* (**algoritmo di Bannai**), dell'array delle *matching statistics*.

LCE

Dato un testo T, tale che $|T| = n$, il risultato della **LCE query** tra due posizioni i e j, tali che $0 \leq i, j < n$, corrisponde al più lungo prefisso comune tra le sotto-stringhe che hanno come indice di partenza i e j, avendo quindi il più lungo prefisso comune tra $T[i, n-1]$ e $T[j, n-1]$.

MONI e PHONI

MONI

Rossi et al., nel 2021, sfruttarono tutte le conoscenze relative alla **RLBWT**, all'**r-index** e alle **matching statistics** per ideare **MONI: *A Pangenomics Index for Finding MEMs*** [2]. In questa soluzione si ha quindi la costruzione, in due *sweep*, tramite l'uso delle *threshold* (**algoritmo di Bannai**), dell'array delle *matching statistics*.

LCE

Dato un testo T, tale che $|T| = n$, il risultato della **LCE query** tra due posizioni i e j, tali che $0 \leq i, j < n$, corrisponde al più lungo prefisso comune tra le sotto-stringhe che hanno come indice di partenza i e j, avendo quindi il più lungo prefisso comune tra $T[i, n-1]$ e $T[j, n-1]$.

PHONI

Nel 2021, Boucher, Gagie, Rossi et al. proposero un ulteriore miglioramento di quanto fatto in *MONI*, con **PHONI: *Streamed Matching Statistics with Multi-Genome References***[3], usando le *LCE query* al posto delle *threshold*.

PBWT

Prefix array

Dato un aplotipo i, appartenente al pannello X, e un indice di colonna k, si definisce il **prefix array** a_k come una permutazione degli indici $0, \ldots, M-1$ tale che $a_k[i] = j$ sse x_j è l'i-esimo aplotipo di X nell'ordinamento inverso dei prefissi ottenuto alla colonna k. Quindi $a_k[i] = m$, con $m < M$, altro non è che l'indice della sequenza x_m del pannello X da cui deriva il prefisso i-esimo nell'ordine inverso in colonna k.

PBWT

Prefix array

Dato un aplotipo i, appartenente al pannello X, e un indice di colonna k, si definisce il **prefix array** a_k come una permutazione degli indici $0, \ldots, M-1$ tale che $a_k[i] = j$ sse x_j è l'i-esimo aplotipo di X nell'ordinamento inverso dei prefissi ottenuto alla colonna k. Quindi $a_k[i] = m$, con $m < M$, altro non è che l'indice della sequenza x_m del pannello X da cui deriva il prefisso i-esimo nell'ordine inverso in colonna k.

Divergence array

Si definisce **divergence array** l'array d_k tale che $d_k[i]$ è l'indice colonna iniziale del match massimale a sinistra terminante in k tra l'i-esimo aplotipo e il suo precedente nell'ordinamento ottenuto alla colonna k-esima. Formalmente, dato $i > 0$, si definisce $d_k[i]$ come il più piccolo j tale che $y_i^k[j, k) = y_{i-1}^k[j, k)$. Ne segue che
$y_i^k[k-1] \neq y_{i-1}^k[k-1] \implies d_k[i] = k$ (per definizione $d_k[0] = k$).

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

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

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

Set-maximal exact match

SMEM

Dato un pannello X, con M aplotipi/righe e N siti/colonne, e un aplotipo query z, tale che $|z| = N$, si definisce un **Set-Maximal Exact Match (*SMEM*)**, iniziante in colonna e_k e terminante il colonna k, tra la query z e le righe del pannello indicizzate dai valori compresi nell'intervallo $[f_k, g_k)$ in a_k sse:

$$z[e_k, k) = y_i^k[e_k, k) \wedge z[e_k - 1] \neq y_i^k[e_k - 1], \forall\, i \text{ t.c. } f_k \leq i < g_k$$

Si noti che $g_k = M$ sse y_{M-1}^k appartiene alle righe per le quali si ha tale *SMEM*.

Il calcolo viene effettuato tramite il cosiddetto **algoritmo 5 di Durbin**[4] in tempo Avg.$\mathcal{O}(N + c)$[5], avendo N aplotipi e c SMEM, con una memoria richiesta di 13NM byte.

Outline

Outline

Outline

Outline

Bibliografia I

[1] Paola Bonizzoni, Christina Boucher, Davide Cozzi, Travis Gagie, Sana Kashgouli, Dominik Köppl, and Massimiliano Rossi.
Compressed data structures for population-scale positional Burrows–Wheeler transforms.
bioRxiv, 09 2022.

[2] Massimiliano Rossi, Marco Oliva, Ben Langmead, Travis Gagie, and Christina Boucher.
MONI: A pangenomic index for finding maximal exact matches.
Journal of Computational Biology, 02 2022.

[3] Christina Boucher, Travis Gagie, I Tomohiro, Dominik Köppl, Ben Langmead, Giovanni Manzini, Gonzalo Navarro, Alejandro Pacheco, and Massimiliano Rossi.
PHONI: Streamed matching statistics with multi-genome references.
In *2021 Data Compression Conference (DCC)*, pages 193–202. IEEE, 2021.

[4] Richard Durbin.
Efficient haplotype matching and storage using the positional BurrowsWheeler transform (PBWT).
Bioinformatics, 30(9):1266–1272, 01 2014.

[5] Ahsan Sanaullah, Degui Zhi, and Shaojie Zhang.
d-PBWT: dynamic positional BurrowsWheeler transform.
Bioinformatics, 37(16):2390–2397, 02 2021.

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