

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

Davide Cozzi

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

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

26 Ottobre 2022



- 1 Introduzione e scopo della tesi
- 2 Run-length encoded PBWT
- 3 Risultati sperimentali
- 4 Conclusioni e sviluppi futuri

# Il pangenoma

## Un cambio di paradigma

- singolo genoma  $\implies$  insieme di genomi
- studio delle varianti geniche tra genomi di diversi individui

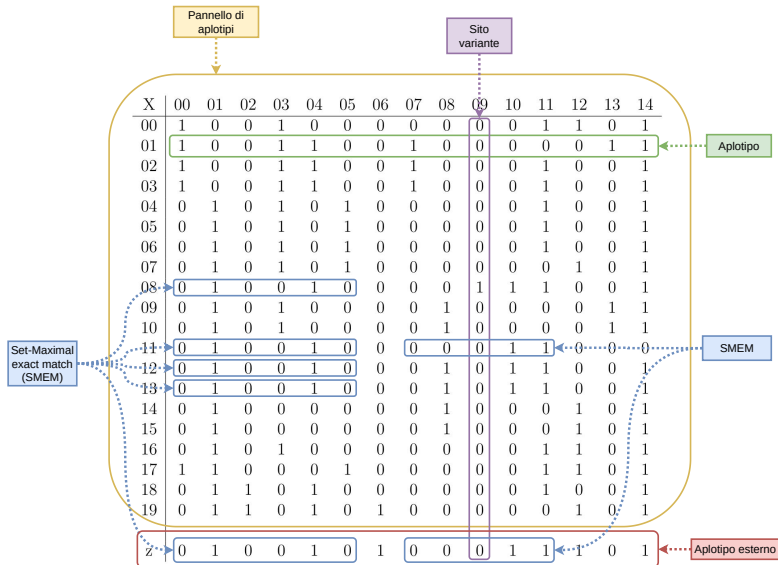
## Rappresentazioni del pangenoma

- grafo del pangenoma
- **pannello di aplotipi**

## Aplotipo

Insieme di alleli, ovvero di varianti che un organismo eredita da ogni genitore, a meno di mutazioni.

# Il pangenoma




# Trasformata di Burrows–Wheeler posizionale - *Durbin, Bioinformatics, 2014*

Prefix array			00	01	02	03	04	05	06	07	08	09	10	11	12	13	14
i	$d_5$	$a_5$															
00	05	14	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
01	00	15	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
02	01	17	1	1	0	0	0	1	0	0	0	0	0	1	1	0	1
03	04	00	1	0	0	1	0	0	0	0	0	0	0	1	1	0	1
04	02	04	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
05	00	05	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
06	00	06	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
07	00	07	0	1	0	1	0	1	0	0	0	0	0	0	1	0	1
08	00	09	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
09	00	10	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
10	00	16	0	1	0	1	0	0	0	0	0	0	0	1	1	0	1
11	05	08	0	1	0	0	1	0	0	0	0	1	1	1	0	0	1
12	00	11	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0
13	00	12	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
14	00	13	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
15	03	18	0	1	1	0	1	0	0	0	0	0	0	1	0	0	1
16	00	19	0	1	1	0	1	0	1	0	0	0	0	0	1	0	1
17	04	01	1	0	0	1	1	0	0	1	0	0	0	0	0	1	1
18	00	02	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1
19	00	03	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1

Trasformata di Burrows–Wheeler posizionale - *Durbin, Bioinformatics, 2014*

Divergence array																		
i	$d_5$	$a_5$	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	
00	05	14	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	
01	00	15	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	
02	01	17	1	1	0	0	0	1	0	0	0	0	0	1	1	0	1	
03	04	00	1	0	0	1	0	0	0	0	0	0	0	1	1	0	1	
04	02	04	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1	
05	00	05	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1	
06	00	06	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1	
07	00	07	0	1	0	1	0	1	0	0	0	0	0	0	1	0	1	
08	00	09	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1	
09	00	10	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1	
10	00	16	0	1	0	1	0	0	0	0	0	0	0	1	1	0	1	
11	05	08	0	1	0	0	0	1	0	0	0	1	1	1	0	0	1	
12	00	11	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	
13	00	12	0	1	0	0	0	1	0	0	1	0	1	1	0	0	1	
14	00	13	0	1	0	0	0	1	0	0	1	0	1	1	0	0	1	
15	03	18	0	1	1	0	0	1	0	0	0	0	0	1	0	0	1	
16	00	19	0	1	1	0	0	1	0	1	0	0	0	0	1	0	1	
17	04	01	1	0	0	1	0	1	0	0	1	0	0	0	0	1	1	
18	00	02	1	0	0	1	0	1	0	0	1	0	0	1	0	0	1	
19	00	03	1	0	0	1	0	1	0	0	1	0	0	1	0	0	1	

Trasformata di Burrows–Wheeler posizionale - *Durbin, Bioinformatics, 2014*


i	$d_5$	$a_5$	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14
00	05	14	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
01	00	15	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
02	01	17	1	1	0	0	0	1	0	0	0	0	0	1	1	0	1
03	04	00	1	0	0	1	0	0	0	0	0	0	0	1	1	0	1
04	02	04	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
05	00	05	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
06	00	06	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
07	00	07	0	1	0	1	0	1	0	0	0	0	0	0	1	0	1
08	00	09	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
09	00	10	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
10	00	16	0	1	0	1	0	0	0	0	0	0	0	1	1	0	1
11	05	08	0	1	0	0	1	0	0	0	0	1	1	1	0	0	1
12	00	11	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0
13	00	12	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
14	00	13	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
15	03	18	0	1	1	0	1	0	0	0	0	0	0	1	0	0	1
16	00	19	0	1	1	0	1	0	1	0	0	0	0	0	1	0	1
17	04	01	1	0	0	1	1	0	0	1	0	0	0	0	0	1	1
18	00	02	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1
19	00	03	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1

# Scopo della tesi

Lo scopo di questa tesi è quello di creare una variante run-length encoded della PBWT (RLPBWT) che permetta il calcolo degli SMEM con aplotipo esterno, in modo efficiente dal punto di vista della memoria richiesta.

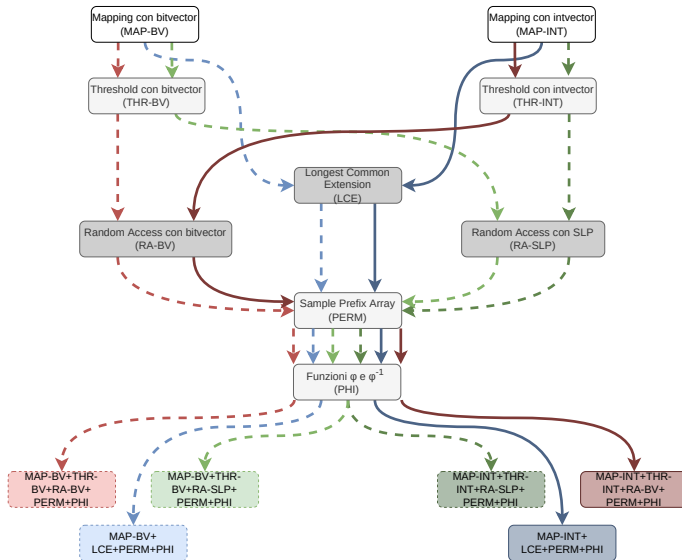
Calcolo degli SMEM naïve ( $N$  siti e  $M$  aplotipi):  $\mathcal{O}(N^2M)$

Calcolo degli SMEM per Durbin ( $c$  numero SMEM):

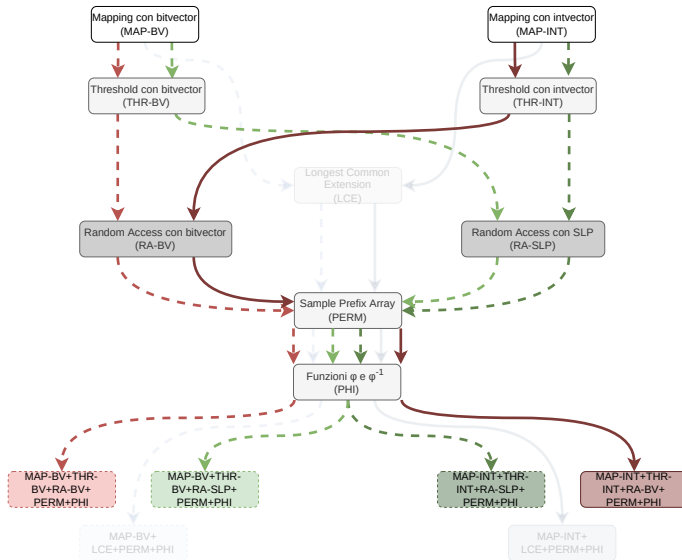
- tempo:  $\mathcal{O}(NM) + \text{Avg. } \mathcal{O}(N + c)$
- spazio:  $\mathcal{O}(NM) \Rightarrow \underline{13NM \text{ byte}}$



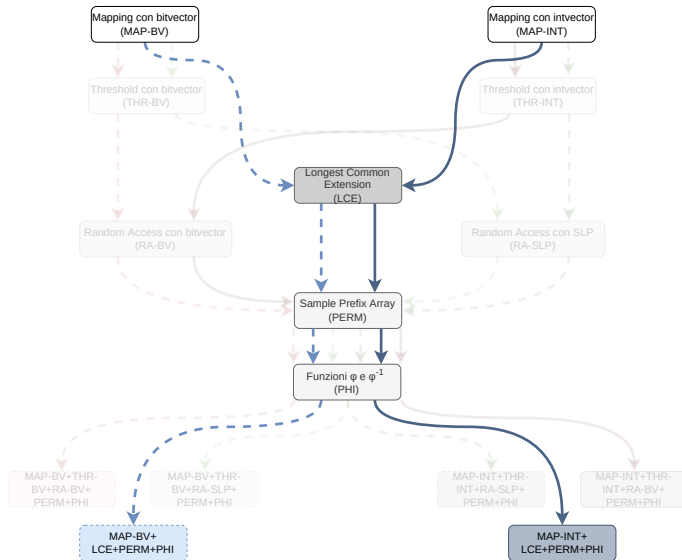
# Sottostrutture e strutture dati composte per la RLPBWT



# Sottostrutture e strutture dati composte per la RLPBWT

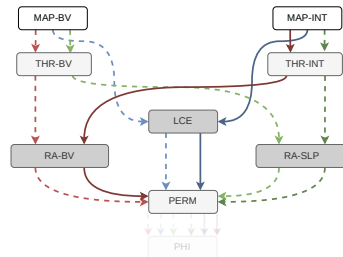


# Sottostrutture e strutture dati composte per la RLPBWT



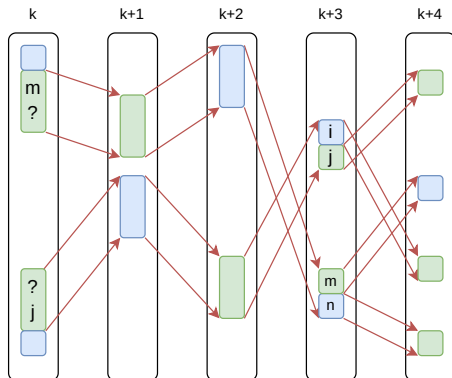
# Matching statistics e calcolo degli SMEM

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

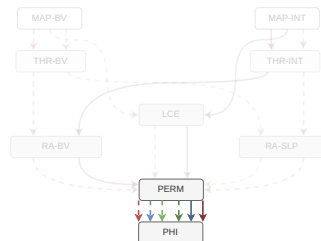


MS \ k	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14
row	19	19	16	15	13	13	19	19	19	19	11	11	17	17	17
len	1	2	3	4	5	6	4	5	6	7	4	5	2	3	4

# Struttura per le funzioni $\varphi$ e $\varphi^{-1}$



...



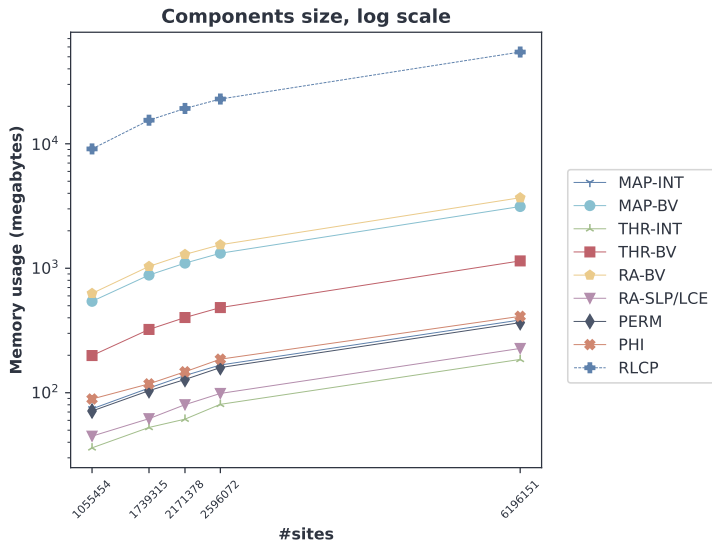
# Dati in input

## Pannelli di varianti reali

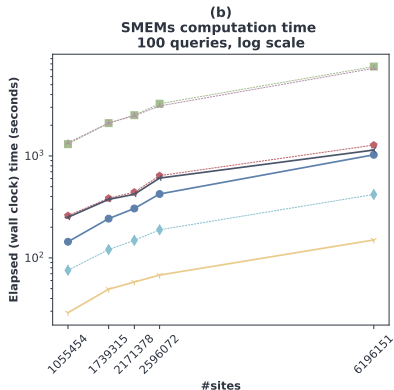
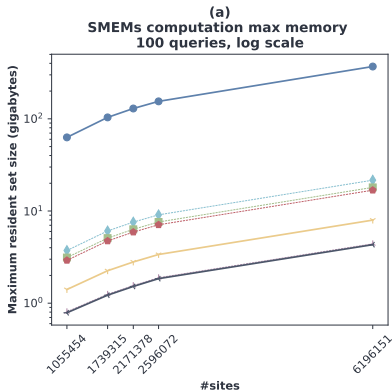
- dati reali relativi alla phase 3 del **1000 Genome Project**
- numero costante di aplotipi: 5008
- estrazione di 100 query  $\implies$  pannelli con 4908 aplotipi
- numero variabile di siti

Chr	#Siti	Media run
chr22	1.055.454	14
chr20	1.739.315	11
chr18	2.171.378	11
chr16	2.596.072	12
chr1	6.196.151	11

# Costo in memoria delle componenti



# Performance del calcolo degli SMEM con 100 query





# Considerazioni e sviluppi futuri

## Considerazioni finali

- le strutture dati e gli algoritmi proposti hanno confermato le 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

- gestione di pannelli di query
- RLPBWT con dati mancanti
- SMEM interni con RLPBWT
- RLPBWT multiallelica

## Ulteriori dettagli

**Bonizzoni, Boucher, Cozzi, Gagie, Kashgouli, Köppl e Rossi:**

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

17<sup>th</sup> *Workshop on Compression, Text and Algorithms (WCTA)* - Concepción (Cile)  
11 Novembre 2022



Grazie per l'attenzione



- [1] Richard Durbin.  
Efficient haplotype matching and storage using the positional BurrowsWheeler transform (PBWT).  
*Bioinformatics*, 30(9):1266–1272, 01 2014.
- [2] Jasmijn A Baaijens, Paola Bonizzoni, Christina Boucher, Gianluca Della Vedova, Yuri Pirola, Raffaella Rizzi, and Jouni Sirén.  
Computational graph pangenomics: a tutorial on data structures and their applications.  
*Natural Computing*, pages 1–28, 2022.
- [3] Travis Gagie, Giovanni Manzini, Gonzalo Navarro, Hiroshi Sakamoto, Louisa Seelbach Benkner, Yoshimasa Takabatake, et al.  
Practical random access to SLP-compressed texts.  
In *International Symposium on String Processing and Information Retrieval*, pages 221–231. Springer, 2020.
- [4] 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.
- [5] 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.
- [6] 1000 Genomes Project Consortium et al.  
A global reference for human genetic variation.  
*Nature*, 526(7571):68, 2015.

## PBWT - *Durbin, Bioinformatics, 2014*

Dato pannello di  $M$  aplotipi, lunghi  $N$  siti (biallelici:  $\Sigma = \{0, 1\}$ ), si definisce PBWT del pannello una collezione di  $N + 1$  coppie di array  $(a_k, d_k)$ ,  $0 \leq k \leq N$ , dove:

- $a_k$  è il **prefix array** della colonna  $k$
  - $d_k$  è il **divergence array** della colonna  $k$
- 
- la PBWT è basata sul riordinamento co-lessicografico a ogni colonna
  - il pannello, riordinato in ogni colonna  $k$  con  $a_k$ , è detto matrice PBWT
  - aplotipi simili, riordinati consecutivamente alla colonna  $k$ , è molto probabile presentino il medesimo carattere in colonna  $k + 1$

## Run-length encoding

Il run-length encoding consiste nel memorizzare le *run*, ovvero sequenze massimali di caratteri uguali, come coppie:

(carattere, lunghezza della run)

000000  $\implies$  (0,6)

# Definizione MS e calcolo SMEM

## Matching statistics per la PBWT

Dato un pannello  $X = \{x_0, \dots, x_{M-1}\}$ ,  $x_i = N$ , e un aplotipo esterno  $z$ , tale che  $|z| = N$ , si definisce **matching statistics** di  $z$  su  $X$  un array  $MS$  di coppie  $(row, len)$ ,  $|MS| = N$ , tale che:

- $x_{MS[i].row}[i - MS[i].len + 1, i] = z[i - MS[i].len + 1, i]$ , ovvero si ha che l'aplotipo esterno ha un match, lungo  $MS[i].len$  e terminante in colonna  $i$ , con la riga  $MS[i].row$ -esima del pannello
- $z[i - MS[i].len, i]$  non è un suffisso, terminante in colonna  $i$ , di un qualsiasi sottoinsieme di righe di  $X$

## SMEM da MS

Dato un array di matching statistics  $MS$  si ha che  $z[i - l + 1, i]$  presenta uno SMEM di lunghezza  $l$  con la riga  $MS[i].row$ -esima del pannello  $X$  sse:  
 $MS[i].len = l \wedge (i = N - 1 \vee MS[i].len \geq MS[i + 1].len)$

# Mapping e threshold

## Colonna matrice PBWT

$$y^5 = 00101111000000000000$$

## intvector compressi

$$p_5 = [0, 2, 3, 4, 8]$$

$$uv_5 = [0, 2, 1, 3, 5], \quad c[5] = 15, \quad start_5 = \top$$

$$t_5 = [0, 3, 3, 4, 11]$$

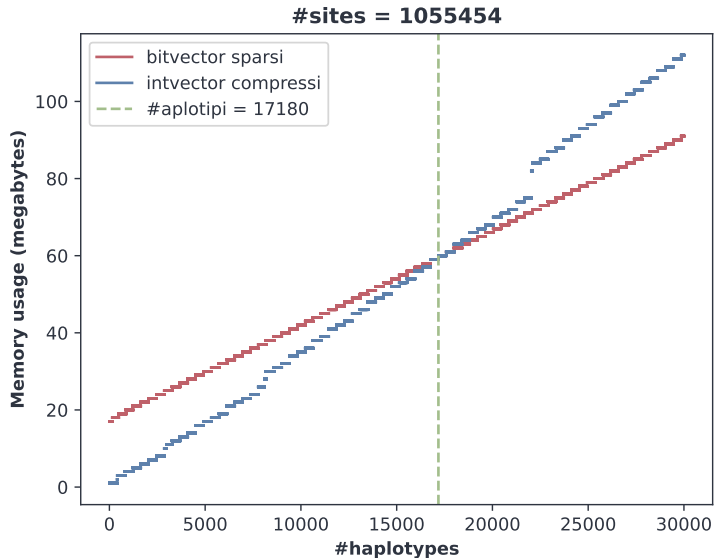
## bitvector sparsi

$$h_5 = 01110001000000000001$$

$$u_5 = 0110000000000001, \quad v_5 = 10001, \quad c[5] = 15, \quad start_5 = \top$$

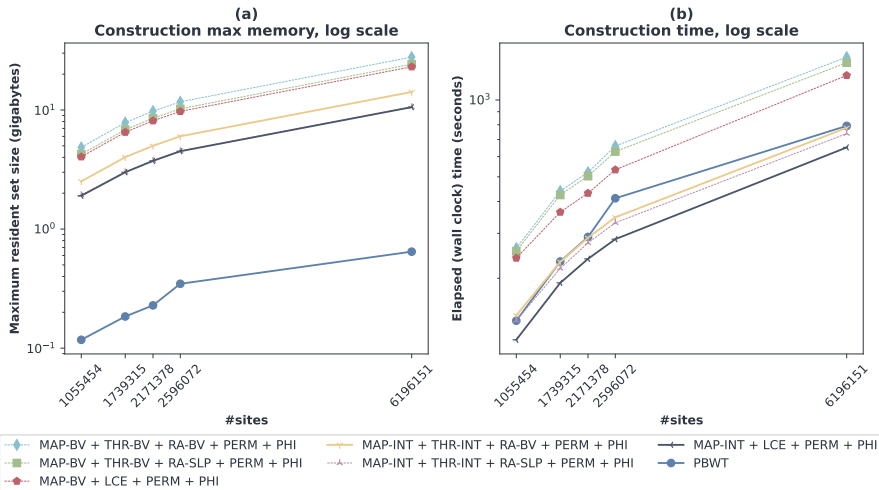
$$t_5 = 10111000000100000000$$

# Bitvector vs Intvector, spazio

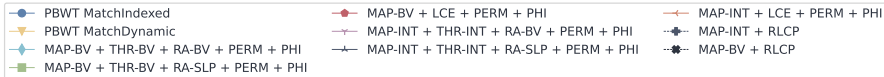
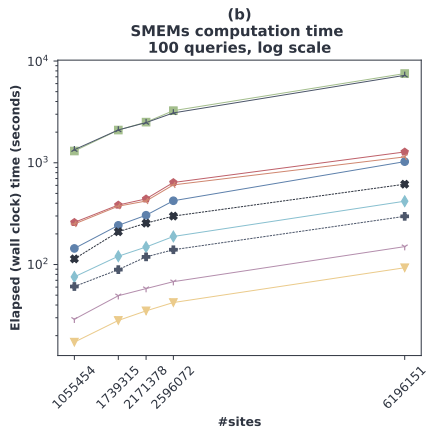
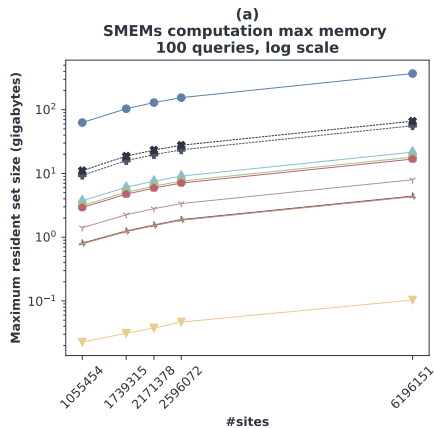




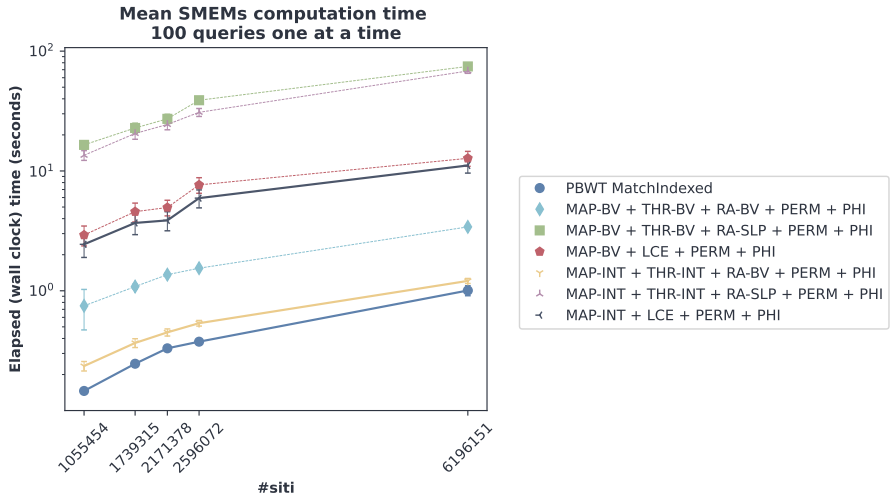
# Performance costruzione strutture dati



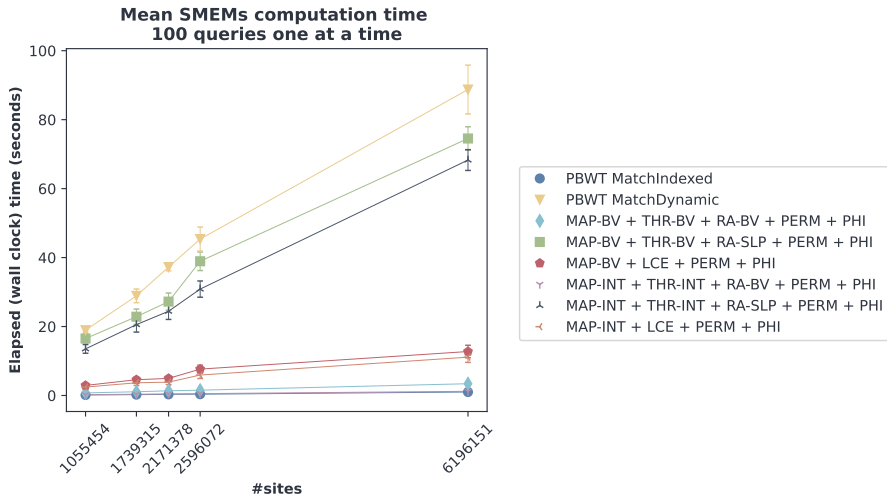
# Performance calcolo degli SMEM con 100 query



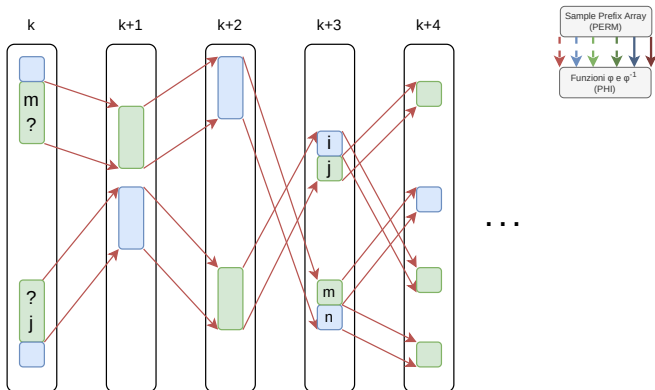
# Performance calcolo degli SMEM per singole query



# Performance calcolo degli SMEM per singole query



# Struttura per le funzioni $\varphi$ e $\varphi^{-1}$



$$\Phi_j = [0, 0, 0, 1, 0, \dots], \quad \Phi_m^{-1} = [0, 0, 0, 1, 0, \dots], \quad \Phi_{supp} = [i, \dots], \quad \Phi_{supp}^{-1} = [n, \dots]$$

$$\Phi_{supp}^j[\text{rank}_j^\varphi(0)] = \Phi_{supp}^j[0] = i, \quad \Phi_{supp}^{-1}{}^m[\text{rank}_m^{\varphi^{-1}}(0)] = \Phi_{supp}^{-1}{}^m[0] = n$$