



UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di Laurea Magistrale in Informatica

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

**Relatore:** *Prof.ssa Raffaella Rizzi*

**Correlatore:**

**Tesi di Laurea Magistrale di:**

*Davide Cozzi*

*Matricola 829827*

**Anno Accademico 2021-2022**

# Abstract

# Indice

<b>1</b>	<b>Introduzione</b>	<b>4</b>
<b>2</b>	<b>Preliminari</b>	<b>5</b>
2.1	Motivazioni Biologiche . . . . .	6
2.2	Bitvector sparsi . . . . .	6
2.3	Straight-Line Program . . . . .	6
2.3.1	Random access . . . . .	6
2.3.2	Longest Common Extension . . . . .	6
2.4	Trasformata di Burrows-Wheeler . . . . .	6
2.4.1	Trasformata di Burrows-Wheeler run-length . . . . .	6
2.4.2	Matching Statistics . . . . .	6
2.4.3	R-index . . . . .	6
2.4.4	MONI . . . . .	6
2.4.5	PHONI . . . . .	6
2.5	Trasformata di Burrows-Wheeler posizionale . . . . .	6
2.5.1	Implementazione originale . . . . .	6
2.5.2	Varianti della PBWT . . . . .	6
<b>3</b>	<b>Metodo</b>	<b>7</b>
3.1	Introduzione agli strumenti usati . . . . .	8
3.1.1	SDSL . . . . .	8
3.1.2	BigRepair . . . . .	8
3.1.3	ShapedSlp . . . . .	8
3.2	Introduzione alle varianti della RLPBWT . . . . .	8
3.2.1	Perché un'implementazione run-length . . . . .	8
3.3	Mapping nella RLPBWT . . . . .	8
3.4	RLPBWT naive . . . . .	8
3.4.1	Algoritmo per match massimali . . . . .	8
3.5	RLPBWT con bitvectors . . . . .	8
3.5.1	Algoritmo per match massimali . . . . .	8
3.6	RLPBWT con pannello . . . . .	8

3.6.1	Algoritmo con matching statistics . . . . .	8
3.7	RLPBWT con SLP . . . . .	8
3.7.1	Algoritmo con matching statistics . . . . .	8
3.8	Funzione Phi . . . . .	8
3.8.1	Costruzione della struttura di supporto . . . . .	8
3.8.2	Estensione dei match . . . . .	8
<b>4</b>	<b>Risultati</b>	<b>9</b>
4.1	Ambiente di benchmark . . . . .	9
4.1.1	Descrizione input . . . . .	9
4.2	Analisi temporale . . . . .	9
4.3	Analisi spaziale . . . . .	9
<b>5</b>	<b>Conclusioni</b>	<b>10</b>
5.1	Sviluppi futuri . . . . .	10
	<b>Bibliografia e sitografia</b>	<b>10</b>



# Capitolo 1

## Introduzione

### 1.1 Motivazioni Biologiche

### 1.2 Bitvector sparsi

### 1.3 Straight-Line Program

#### 1.3.1 Random access

#### 1.3.2 Longest Common Extension

### 1.4 Trasformata di Burrows-Wheeler

#### 1.4.1 Trasformata di Burrows-Wheeler run-length

#### 1.4.2 Matching Statistics

#### 1.4.3 R-index

#### 1.4.4 MONI

#### 1.4.5 PHONI

### 1.5 Trasformata di Burrows-Wheeler posizionale

#### 1.5.1 Implementazione originale

Gli algoritmi di Durbin

Limiti spaziali

#### 1.5.2 Varianti della PBWT <sup>5</sup>

PBWT multi-allelica

PBWT con struttura LEAP

PBWT dinamica

PBWT bidirezionale

Recenti sviluppi



# Capitolo 2

## Metodo

### 2.1 Introduzione agli strumenti usati

#### 2.1.1 SDSL

#### 2.1.2 BigRepair

#### 2.1.3 ShapedSlp

Ricostruzione del panel

### 2.2 Introduzione alle varianti della RLPBWT

#### 2.2.1 Perché un'implementazione run-length

### 2.3 Mapping nella RLPBWT

### 2.4 RLPBWT naive

#### 2.4.1 Algoritmo per match massimali

### 2.5 RLPBWT con bitvectors

#### 2.5.1 Algoritmo per match massimali

### 2.6 RLPBWT con pannello

#### 2.6.1 Algoritmo con matching statistics

### 2.7 RLPBWT con SLP<sub>7</sub>

#### 2.7.1 Algoritmo con matching statistics

### 2.8 Funzione Phi

#### 2.8.1 Costruzione della struttura di supporto

#### 2.8.2 Estensione dei match



# Capitolo 3

## Risultati

### 3.1 Ambiente di benchmark

#### 3.1.1 Descrizione input

### 3.2 Analisi temporale

### 3.3 Analisi spaziale

# Capitolo 4

## Conclusioni

### 4.1 Sviluppi futuri