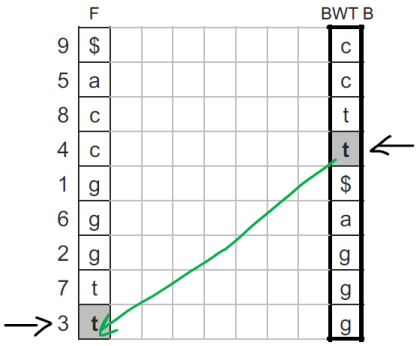
Lezione 19 14/12/2023

BWT

B[i] è l'ultimo simbolo della rotazione r_i che è la i-esima nell'ordinamento lessicografico e che sarà la q-rotazione per un certo valore di q.
B[i] è il primo simbolo della (q-1)-rotazione.
Dove si trova la (q-1)-rotazione nell'ordinamento?

La BWT è un vettore di n elementi, dove la posizione i è occupata da un simbolo che rappresenta l'ultimo simbolo dell'-i esima rotazione nell'ordinamento lessicografico. Questa rotazione è una q-rotazione, cioè che inizia dall'indice q.

La posizione lessicografica, e il suo inizio nel testo, non sono uguali e se lo sono è un caso.



B[4] coincide con F[9]

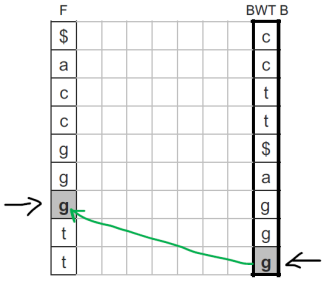
Il simbolo salvato nella posizione i della bwt è il primo simbolo della (q-1)rotazione.

Troviamo quindi la 3 rotazione con gli indici, ma se non avessimo gli indici?

Il ranking viene mantenuto, dall'alto nel BWT è la seconda "t", e anche nel vettore F p la seconda "f".

Proprietà 2

L'r-esimo simbolo σ in B e l'r-esimo simbolo σ in F sono lo stesso simbolo del testo T
→ Proprietà P2: Last-First Mapping



Abbiamo quindi che B[9] coincide con F[7] perchè B[9] è la terza g in B, e F[7] è la terza g in F.

è una coincidenza forte, non è solo lo stesso simbolo "a" o "b", è proprio lo stesso.

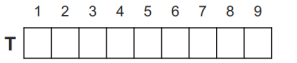
Last-First (LF) function

j = LF(i)

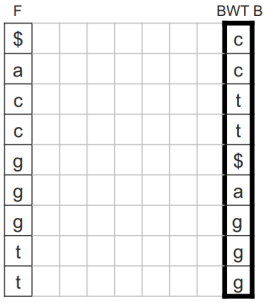
tale che B[i] e F[j] sono lo stesso simbolo del testo

In questo caso avrebbe come input 9 e come output 7. L'idea è che la funzione farà questo mapping in tempo costante, per poterlo usare nella ricerca di un pattern in tempo costante.

Ricostruzione di T da B



Si determina F riordinando i simboli di B

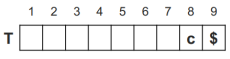


Può capitare in esame.

T ha la stessa lunghezza di B.

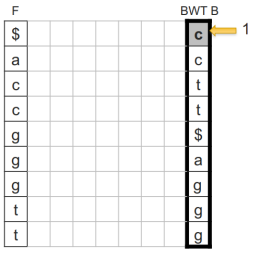
Si determina F riordinando i simboli di B.

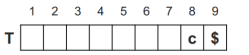
Si parte dalla prima posizione di F che contiene sempre un $.



Proprietà P1:

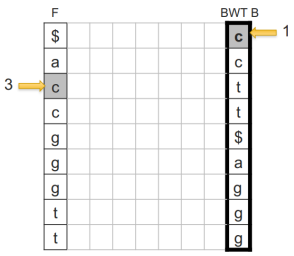
B[1] = **c**
precede in T
F[1] = $

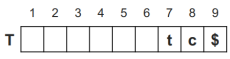




Proprietà P2:

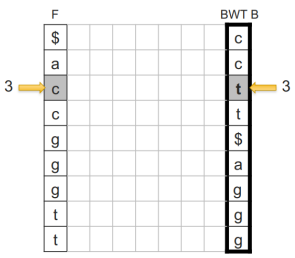
B[1] = **c** è il primo simbolo c in B e corrisponde al primo simbolo c in F, cioè F[3]

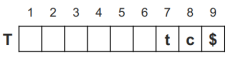




Proprietà P1:

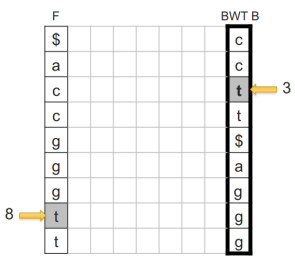
B[3] = **t**
precede in T
F[3] = c

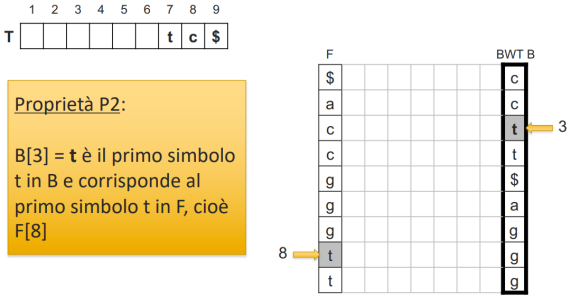




Proprietà P2:

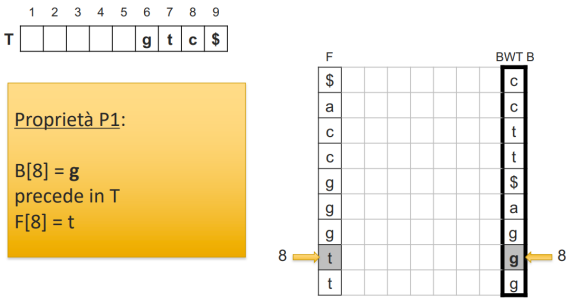
B[3] = **t** è il primo simbolo t in B e corrisponde al primo simbolo t in F, cioè F[8]





Proprietà P2:

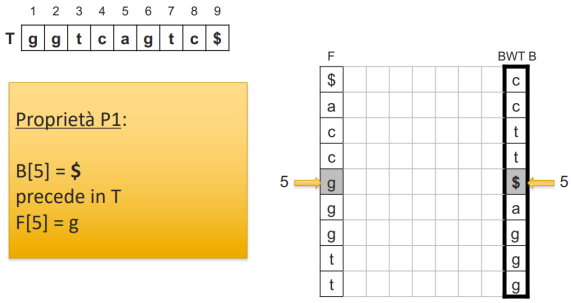
B[3] = **t** è il primo simbolo t in B e corrisponde al primo simbolo t in F, cioè F[8]



Proprietà P1:

B[8] = **g** precede in T
F[8] = t

Quando arrivo al dollaro mi fermo.



Proprietà P1:

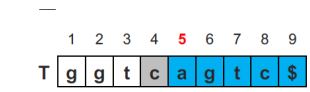
B[5] = **$** precede in T
F[5] = g

BWT e Suffix Array

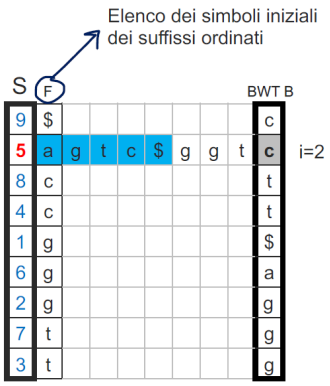
Pur essendo due strutture diverse, possono essere uniti. Recap:

> Il simbolo B[i] della BWT è l'ultimo simbolo della i-esima rotazione r_i.
>
> Supponiamo che r_i sia la rotazione che inizia in posizione q del testo, cioè è la q-rotazione.
>
> L'ultimo simbolo di r_i è il simbolo T[q-1].
>
> Quindi B[i], che è l'ultimo simbolo di r_i, sarà il simbolo T[q-1].
>
> La rotazione r_i inizia con il q-suffisso e quindi B[i] è il simbolo che precede il q-suffisso.
>
> Sicuramente, il q-suffisso è l'i-esimo nell'ordinamento lessicografico dei suffissi di T.
>
> In conclusione, B[i] è il simbolo che precede l'i-esimo suffisso nell'ordinamento lessicografico dei simboli di T

Esempio di questa frase in blu:



ESEMPIO:
per i=2, B[i] = **c** precede il secondo suffisso che è il **5**-suffisso.



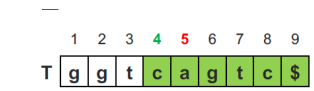
La colonna S è quella del suffix array. Quindi gli indici nell'ordinamento lessicografico delle rotazioni coincidono con il suffix array. Questo funziona perché abbiamo la terminazione con il $.

Quindi F è l'elenco dei simboli iniziali dei suffissi ordinati (nel suffix array).

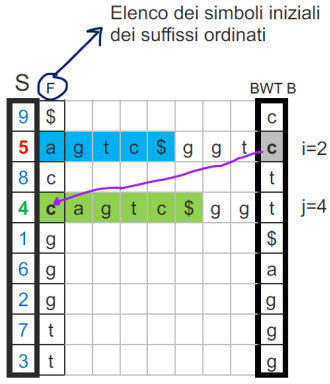
1. B[i] è il simbolo che in T precede l'i-esimo suffisso
2. S[i] è l'indice dell'i-esimo suffisso.

⇒ B[i] il simbolo di T in posizione S[i]-1 e quindi è il simbolo iniziale del suffisso di indice S[i]-1.

Esempio:



ESEMPIO:
per i=2, B[i] = **c** precede il secondo suffisso che è il **5**-suffisso.
B[i] = **c** è il simbolo iniziale del **4**-suffisso



Quindi B[i] è il simbolo iniziale del q-suffisso di q precedente.

Questa freccia è la funzione last-first mapping.

Last-First Mapping: il suffisso che inizia con l'r-esimo simbolo σ della BWT è l'r-esimo suffisso che inizia con σ nell'ordinamento lessicografico dei suffissi di T

La funzione **j = LF(i)** fornisce la posizione j (nel Suffix Array) del suffisso che inizia con B[i]

Quindi in input della funzione 2, restituisce 4, che è la posizione del suffisso che inizia con il carattere in B[2]

Calcolo di BWT da SA

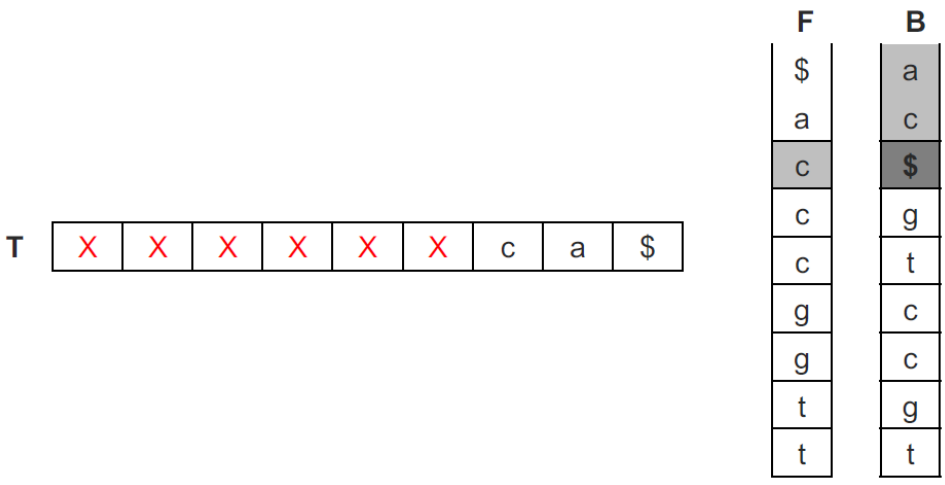
```
Procedura Costruisci-BWT-da-SA(S, T)
   n ← |S|
   for i from 1 to n do
     B[i] ← T[S[i]-1]

   return B
```

Esercizi

Esercizio 1

> Dire se la stringa ac$gtccgt può essere la BWT di un testo T.

Bisogna provare a ricostruire il testo.



Quindi non può essere una BWT.

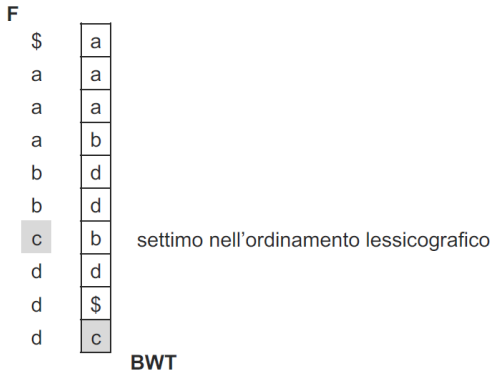
Esercizio 2

Sapendo che la BWT di un testo T è B = aaabdddbd$c, specificare senza ricostruire T e motivando la risposta:

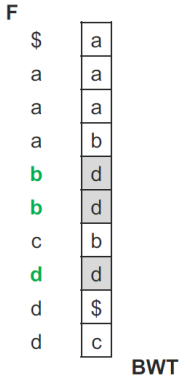
① quanti sono i suffissi che iniziano con il simbolo c e in che posizioni stanno nel Suffix Array

② quali simboli sono preceduti nel testo da un simbolo d

Numero di suffissi che iniziano con c → 1



Simboli che seguono un simbolo d → **b**, **b** e **d**

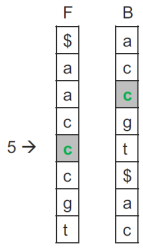


Quindi l'unico suffisso che inizia con la c è il settimo nell'ordinamento lessicografico, perché la c mappata è la settima nel vettore F.

Esercizio 3

Data la Burrows-Wheeler Transform B = accgt$ac di un testo T, si richiede di specificare l'array F e usarlo per individuare la posizione nel Suffix Array del suffisso che inizia con il terzo simbolo della BWT.

Posizione nel SA del suffisso che inizia il terzo simbolo di B?



Il suffisso che inizia con B[3]=**c** è il quinto in ordine lessicografico

Esercizio 4

Determinare il Suffix Array S del testo T = dacdbbac$. Sulla base di S individuare l'indice del quarto suffisso in ordine lessicografico ed evidenziarlo su T. Determinare inoltre la BWT di T ricavandola da S.

Suffix Array di T?

1	d	a	c	d	b	b	a	c	$
2	a	c	d	b	b	a	c	$	
3	c	d	b	b	a	c	$		
4	d	b	b	a	c	$			
5	b	b	a	c	$				
6	b	a	c	$					
7	a	c	$						
8	c	$							
9	$								

9	$								
7	a	c	$						
2	a	c	d	b	b	a	c	$	
6	b	a	c	$					
5	b	b	a	c	$				
8	c	$							
3	c	d	b	b	a	c	$		
1	d	a	c	d	b	b	a	c	$
4	d	b	b	a	c	$			

S

Indice del quarto suffisso di T?

S
9
7
2
6
5
8
3
1
4

indice del quarto suffisso nell'ordinamento lessicografico → 6

BWT di T?

S	B
9	c
7	
2	
6	
5	
8	
3	
1	
4	

simbolo che precede S[1]-suffisso → T[8]

T = dacdbbac$

S	B
9	c
7	b
2	
6	
5	
8	
3	
1	
4	

simbolo che precede S[2]-suffisso → T[6]

T = dacdbbac$

S	B
9	c
7	b
2	d
6	b
5	d
8	a
3	a
1	$
4	c

simbolo che precede S[9]-suffisso → T[3]

T = dacdbbac$

Esercizio 5

Determinare la BWT B del testo T = agctgga$.

BWT di T?

a	g	c	t	g	g	a	$
g	c	t	g	g	a	$	a
c	t	g	g	a	$	a	g
t	g	g	a	$	a	g	c
g	g	a	$	a	g	c	t
g	a	$	a	g	c	t	g
a	$	a	g	c	t	g	g
$	a	g	c	t	g	g	a

$	a	g	c	t	g	g	a
a	$	a	g	c	t	g	g
a	g	c	t	g	g	a	$
c	t	g	g	a	$	a	g
g	a	$	a	g	c	t	g
g	c	t	g	g	a	$	a
g	g	a	$	a	g	c	t
t	g	g	a	$	a	g	c

B

Esercizio 6

Il Suffix Array di un generico testo T ($-terminato) è:

S = 8, 6, 4, 3, 2, 1, 7, 5

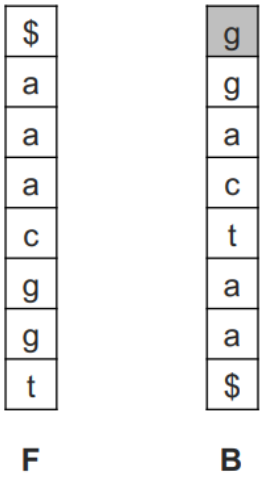
Elencare i simboli della BWT (dal primo all'ultimo) esprimendoli tramite la loro posizione nel testo.

B = T[7], T[5], T[3], T[2], T[1], T[8], T[6], T[4]

Esercizio 7

La BWT di un testo T è B = ggactaa$. Senza ricostruire T, specificare il primo e l'ultimo carattere di T.

T = …..g$



B[1] è l'utlimo simbolo di T prima del simbolo $

F[i], tale che B[i] = $, è il primo simbolo di T