CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Feb 14

- Continue FM-Index and application to pattern matching
- Friday: FM-Index in practice

Notes:

- Office hours today: 1-3pm
- Lab 3 + DBG challenge (challenge.py) due tonight!
- Fill in partner form for Lab 4

FM-Index and application to read mapping

FM-Index: data structure for pattern matching

- Set of auxiliary data structures computed from the BWT of a string S
- The FM-Index consists of 3 parts:

FM-Index: data structure for pattern matching

- Set of auxiliary data structures computed from the BWT of a string *S*
- The FM-Index consists of 3 parts:

(a) The BWT of S, i.e. the L column of $\pi^{\text{sorted}}(\mathrm{S})$

(b) $M[c]$, the first index of c in F (note that F is actually not part of the FM-Index)

(c) $\mathrm{occ}(c, i)$, the number of times c occurs in $L[1 \cdots i]$, inclusive

FM-Index: data structure for pattern matching

- Set of auxiliary data structures computed from the BWT of a string *S*
- The FM-Index consists of 3 parts:

(a) The BWT of S, i.e. the L column of $\pi^{\text{sorted}}(\mathrm{S})$

(b) $M[c]$, the first index of c in F (note that F is actually not part of the FM-Index)

(c) $\mathrm{occ}(c, i)$, the number of times c occurs in $L[1 \cdots i]$, inclusive

- The suffix array *A* is not technically part of the FM-Index, but we will need it for the last step of finding out where pattern *P* occurs in the original string *S*
- A[i] is the index of F[i] in the original string

Example: S = abaaba$, P = aba

1234567

i	F		L	A	occ($)	occ(a)	occ(b)
1	$\$_1$	abaab	a_1				
2	a_1	$abaa	b_1				
3	a_2	aba$a	b_2				
4	a_3	ba$ab	a_2				
5	a_4	baaba	$\$_1$				
6	b_1	a$aba	a_3				
7	b_2	aaba$	a_4				

Step 1: compute the BWT of S

Example: S = abaaba\$, P = aba
1234567

i	F		L	A	occ(\$)	occ(a)	occ(b)
1	$\$_1$	abaab	a_1	7			
2	a_1	\$abaa	b_1	6			
3	a_2	aba\$a	b_2	3			
4	a_3	ba\$ab	a_2	4			
5	a_4	baaba	$\$_1$	1			
6	b_1	a\$aba	a_3	5			
7	b_2	aaba\$	a_4	2			

Step 2: compute the suffix array, where A[i] = index of F[i] in the original sequence

Example: S = abaaba$, P = aba
1234567

i	F		L	A	occ($)	occ(a)	occ(b)
1	$\$_1$	abaab	a_1	7	0		
2	a_1	$abaa	b_1	6	0		
3	a_2	aba$a	b_2	3	0		
4	a_3	ba$ab	a_2	4	0		
5	a_4	baaba	$\$_1$	1	1		
6	b_1	a$aba	a_3	5	1		
7	b_2	aaba$	a_4	2	1		

Step 3: compute the occurrence table for each character c (# times c in L[1…i])

Example: S = abaaba$, P = aba

1234567

i	F	L	A	occ($)	occ(a)	occ(b)
1	$\$_1$ abaab a_1		7	0	1	
2	a_1 $abaa b_1		6	0	1	
3	a_2 aba$a b_2		3	0	1	
4	a_3 ba$ab a_2		4	0	2	
5	a_4 baaba $\$_1$		1	1	2	
6	b_1 a$aba a_3		5	1	3	
7	b_2 aaba$ a_4		2	1	4	

Step 3: compute the occurrence table for each character c (# times c in L[1…i])

Example: S = abaaba$, P = aba
1234567

i	F		L	A	occ($)	occ(a)	occ(b)
1	$\$_1$	abaab	a_1	7	0	1	0
2	a_1	$abaa	b_1	6	0	1	1
3	a_2	aba$a	b_2	3	0	1	2
4	a_3	ba$ab	a_2	4	0	2	2
5	a_4	baaba	$\$_1$	1	1	2	2
6	b_1	a$aba	a_3	5	1	3	2
7	b_2	aaba$	a_4	2	1	4	2

Step 3: compute the occurrence table for each character c (# times c in L[1…i])

Example: S = abaaba$, P = aba
1234567

	i	F		L	A	occ($)	occ(a)	occ(b)
	1	$\$_1$	abaab	a_1	7	0	1	0
sp(a) →	2	a_1	$abaa	b_1	6	0	1	1
	3	a_2	aba$a	b_2	3	0	1	2
	4	a_3	ba$ab	a_2	4	0	2	2
ep(a) →	5	a_4	baaba	$\$_1$	1	1	2	2
	6	b_1	a$aba	a_3	5	1	3	2
	7	b_2	aaba$	a_4	2	1	4	2

Step 4: for pattern P, start with its last char and compute the start and end points

Example: S = abaaba\$, P = aba

1234567

i	F		L	A	occ(\$)	occ(a)	occ(b)
1	$\$_1$	abaab	a_1	7	0	1	0
2	a_1	\$abaa	b_1	6	0	1	1
3	a_2	aba\$a	b_2	3	0	1	2
4	a_3	ba\$ab	a_2	4	0	2	2
5	a_4	baaba	$\$_1$	1	1	2	2
6	b_1	a\$aba	a_3	5	1	3	2
7	b_2	aaba\$	a_4	2	1	4	2

0 -> 2 means we must have seen b_1 and b_2 in the L column

Step 5: for each new character, find the correct number of occurrences in L

Example: S = abaaba\$, P = a$\boxed{\text{ba}}$

1234567

i	F		L	A	occ(\$)	occ(a)	occ(b)
1	$\$_1$	abaab	a_1	7	0	1	0
2	a_1	\$abaa	b_1	6	0	1	1
3	a_2	aba\$a	b_2	3	0	1	2
4	a_3	ba\$ab	a_2	4	0	2	2
5	a_4	baaba	$\$_1$	1	1	2	2
6	b_1	a\$aba	a_3	5	1	3	2
7	b_2	aaba\$	a_4	2	1	4	2

Find where b_1 and b_2 are in the F column, and repeat the process

Step 5: for each new character, find the correct number of occurrences in L

Example: S = abaaba\$, P = aba

1234567

	i	F		L	A	occ(\$)	occ(a)	occ(b)
	1	$\$_1$	abaab	a_1	7	0	1	0
	2	a_1	\$abaa	b_1	6	0	1	1
	3	a_2	aba\$a	b_2	3	0	1	2
	4	a_3	ba\$ab	a_2	4	0	2	2
	5	a_4	baaba	$\$_1$	1	1	2	2
sp(*ba*) →	6	b_1	a\$aba	a_3	5	1	3	2
ep(*ba*) →	7	b_2	aaba\$	a_4	2	1	4	2

Step 5: for each new character, find the correct number of occurrences in L

Example: S = abaaba\$, P = a$\boxed{\text{ba}}$

1234567

i	F L	A	occ(\$)	occ(a)	occ(b)
1	$\$_1$ abaab a_1	7	0	1	0
2	a_1 \$abaa b_1	6	0	1	1
3	a_2 aba\$a b_2	3	0	1	2
4	a_3 ba\$ab a_2	4	0	2	2
5	a_4 baaba $\$_1$	1	1	2	2
6	b_1 a\$aba a_3	5	1	3	2
7	b_2 aaba\$ a_4	2	1	4	2

2 -> 4 means we must have seen a_3 and a_4 in the L column

Step 5: for each new character, find the correct number of occurrences in L

Example: S = abaaba$, P = aba

1234567

i	F L	A	occ($)	occ(a)	occ(b)
1	$\$_1$ abaab a_1	7	0	1	0
2	a_1 \$abaa b_1	6	0	1	1
3	a_2 aba\$a b_2	3	0	1	2
4	a_3 ba\$ab a_2	4	0	2	2
5	a_4 baaba $\$_1$	1	1	2	2
6	b_1 a\$aba a_3	5	1	3	2
7	b_2 aaba\$ a_4	2	1	4	2

Find where a_3 and a_4 are in the F column, done since P ended

Step 5: for each new character, find the correct number of occurrences in L

Example: S = abaaba$, P = $\boxed{\text{aba}}$

1234567

i	F		L	A	occ($)	occ(a)	occ(b)
1	$\$_1$	abaab	a_1	7	0	1	0
2	a_1	$abaa	b_1	6	0	1	1
3	a_2	aba$a	b_2	3	0	1	2
sp(*aba*) → 4	a_3	ba$ab	a_2	4	0	2	2
ep(*aba*) → 5	a_4	baaba	$\$_1$	1	1	2	2
6	b_1	a$aba	a_3	5	1	3	2
7	b_2	aaba$	a_4	2	1	4	2

Note that start and end points are inclusive

Step 6: when we reach the end of *P*, we should have the start/end points in F

Example: S = aba aba$, P = aba

1234567

	i	F		L	A	occ($)	occ(a)	occ(b)
	1	$\$_1$	abaab	a_1	7	0	1	0
	2	a_1	$abaa	b_1	6	0	1	1
	3	a_2	aba$a	b_2	3	0	1	2
sp(*aba*) →	4	a_3	ba$ab	a_2	4	0	2	2
ep(*aba*) →	5	a_4	baaba	$\$_1$	1	1	2	2
	6	b_1	a$aba	a_3	5	1	3	2
	7	b_2	aaba$	a_4	2	1	4	2

Use A (suffix array) to find the original locations of *P* in *S*

Step 7: we are not truly done until we find the locations in the original string!

Handout 8 example: S = `barbara$`, P = `ba`

12345678

i	F L	A	occ($)	occ(a)	occ(b)	occ(r)
1	$\$_1$ barbar a_1					
2	a_1 \$barba r_1					
3	a_2 ra\$bar b_1					
4	a_3 rbara\$ b_2					
5	b_1 ara\$ba r_2					
6	b_2 arbara $\$_1$					
7	r_1 a\$barb a_2					
8	r_2 bara\$b a_3					

Work with a partner!

1) Fill in a column for A as well

2) Try to come up with a formula for **sp** and **ep** in terms of **M** and **occ**

Handout 8 example: S = `barbara$`, P = `ba`

12345678

i	F L	A	occ($)	occ(a)	occ(b)	occ(r)
1	$\$_1$ barbar a_1	8	0	1	0	0
2	a_1 $barba r_1	7	0	1	0	1
3	a_2 ra$bar b_1	5	0	1	1	1
4	a_3 rbara$ b_2	2	0	1	2	1
5	b_1 ara$ba r_2	4	0	1	2	2
6	b_2 arbara $\$_1$	1	1	1	2	2
7	r_1 a$barb a_2	6	1	2	2	2
8	r_2 bara$b a_3	3	1	3	2	2

Handout 8 example: S = barbara\$, P = ba

12345678

i	F		L	A	occ(\$)	occ(a)	occ(b)	occ(r)
1	$\$_1$	barbar	a_1	8	0	1	0	0
2	a_1	\$barba	r_1	7	0	1	0	1
3	a_2	ra\$bar	b_1	5	0	1	1	1
4	a_3	rbara\$	b_2	2	0	1	2	1
5	b_1	ara\$ba	r_2	4	0	1	2	2
6	b_2	arbara	$\$_1$	1	1	1	2	2
7	r_1	a\$barb	a_2	6	1	2	2	2
8	r_2	bara\$b	a_3	3	1	3	2	2

c	M[c]
\$	1
a	2
b	5
r	7

M[*c*] is the first index of character *c* in F (Store instead of F)

Handout 8 example: S = barbara\$, P = ba

12345678

i	F		L	A	occ(\$)	occ(a)	occ(b)	occ(r)
1	$\$_1$	barbar	a_1	8	0	1	0	0
2	a_1	\$barba	r_1	7	0	1	0	1
3	a_2	ra\$bar	b_1	5	0	1	1	1
4	a_3	rbara\$	b_2	2	0	1	2	1
5	b_1	ara\$ba	r_2	4	0	1	2	2
6	b_2	arbara	$\$_1$	1	1	1	2	2
7	r_1	a\$barb	a_2	6	1	2	2	2
8	r_2	bara\$b	a_3	3	1	3	2	2

c	M[c]
\$	1
a	2
b	5
r	7

M[*c*] is the first index of character *c* in F (Store instead of F)

sp(*a*) = 2

ep(*a*) = 4

Handout 8 example: S = barbara\$, P = ba

12345678

i	F L	A	occ(\$)	occ(a)	occ(b)	occ(r)
1	$\$_1$ barbar a_1	8	0	1	0	0
2	a_1 \$barba r_1	7	0	1	0	1
3	a_2 ra\$bar b_1	5	0	1	1	1
4	a_3 rbara\$ b_2	2	0	1	2	1
5	b_1 ara\$ba r_2	4	0	1	2	2
6	b_2 arbara $\$_1$	1	1	1	2	2
7	r_1 a\$barb a_2	6	1	2	2	2
8	r_2 bara\$b a_3	3	1	3	2	2

c	M[c]
\$	1
a	2
b	5
r	7

M[*c*] is the first index of character *c* in F (Store instead of F)

sp(*ba*) = M[*b*] + # *b*'s we saw right before the first *a*

ep(*ba*) = M[*b*] + # *b*'s we saw up until the last *a*

Handout 8 example: S = barbara$, P = ba

12345678

i	F		L	A	occ($)	occ(a)	occ(b)	occ(r)
1	$\$_1$	barbar	a_1	8	0	1	0	0
2	a_1	$barba	r_1	7	0	1	0	1
3	a_2	ra$bar	b_1	5	0	1	1	1
4	a_3	rbara$	b_2	2	0	1	2	1
5	b_1	ara$ba	r_2	4	0	1	2	2
6	b_2	arbara	$\$_1$	1	1	1	2	2
7	r_1	a$barb	a_2	6	1	2	2	2
8	r_2	bara$b	a_3	3	1	3	2	2

c	M[c]
$	1
a	2
b	5
r	7

M[*c*] is the first index of character *c* in F (Store instead of F)

sp(*ba*) = 5 + 0

ep(*ba*) = 5 + 2 – 1 (subtract 1 since we are being *inclusive*)

Handout 8 example: S = barbara\$, P = ba

12345678

i	F		L	A	occ(\$)	occ(a)	occ(b)	occ(r)
1	$\$_1$	barbar	a_1	8	0	1	0	0
2	a_1	\$barba	r_1	7	0	1	0	1
3	a_2	ra\$bar	b_1	5	0	1	1	1
4	a_3	rbara\$	b_2	2	0	1	2	1
5	b_1	ara\$ba	r_2	4	0	1	2	2
6	b_2	arbara	$\$_1$	1	1	1	2	2
7	r_1	a\$barb	a_2	6	1	2	2	2
8	r_2	bara\$b	a_3	3	1	3	2	2

c	M[c]
\$	1
a	2
b	5
r	7

M[*c*] is the first index of character *c* in F (Store instead of F)

sp(*ba*) = 5

ep(*ba*) = 6

Handout 8 example: S = `barbara$`, P = `ba`

`12345678`

i	F		L	A	occ($)	occ(a)	occ(b)	occ(r)
1	$\$_1$	barbar	a_1	8	0	1	0	0
2	a_1	$barba	r_1	7	0	1	0	1
3	a_2	ra$bar	b_1	5	0	1	1	1
4	a_3	rbara$	b_2	2	0	1	2	1
5	b_1	ara$ba	r_2	4	0	1	2	2
6	b_2	arbara	$\$_1$	1	1	1	2	2
7	r_1	a$barb	a_2	6	1	2	2	2
8	r_2	bara$b	a_3	3	1	3	2	2

c	M[c]
$	1
a	2
b	5
r	7

sp(*ba*) = 5

ep(*ba*) = 6

Use A to find locations in original string

Recursive Formulas

base case

$$sp[a] = M[a] = 2$$

$$ep[a] = M[\text{next char}] - 1$$

$$= 5 - 1 \quad 4$$

recursion

$$sp(c\sigma)$$

$$= M[c] + occ(c, \overbrace{sp(\sigma)}^{\text{recursion}} - 1)$$

$$ep(c\sigma)$$

$$= M[c] + occ(c, ep(\sigma)) - 1$$

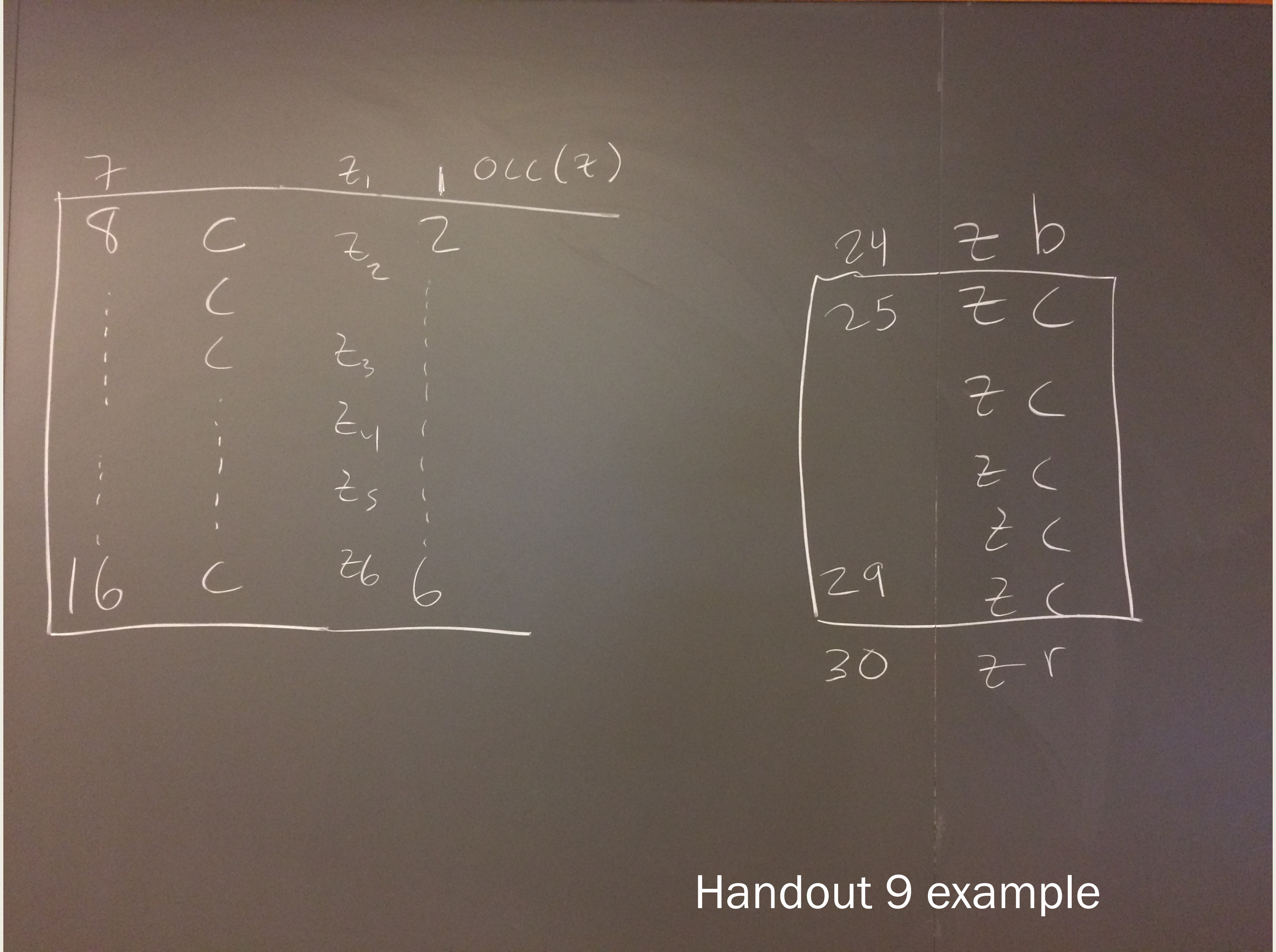
(inclusive)

c	M[c]
$	1
a	2
b	6
c	8
d	17
f	19
r	21
t	23
z	24

$$sp(c) = M[c] = 8$$
$$ep(c) = M[d] - 1 = 16$$

$$sp(zc) = M[z] + occ(z, 8-1)$$
$$= 24 + 1 = 25$$

Handout 9 example



Handout 9 example