

SUBSTRNG-SEARCH \Rightarrow SUFFIX ARRAY

given $T = [1, n]$

search $P = [1, p]$

\Rightarrow return the occurrences of P as a
suffixes of T

$T = \text{mississippi}$ \Rightarrow build and sort the S.A.s over containing
 $\begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \end{matrix}$ all the prefixes

$\begin{array}{l} T_1 \text{ mississippi\$} \\ T_2 \text{ ississippi\$M} \\ T_3 \text{ ssissippi\$M} \\ T_4 \text{ sissippi\$MIS} \\ T_5 \text{ isissippi\$MISS} \\ T_6 \text{ si ppi\$MISSI} \\ T_7 \text{ i ppi\$MISSIS} \\ T_8 \text{ ppi\$MISSISSI} \\ T_9 \text{ pi\$MISSISSIP} \\ T_{10} \text{ i\$MISSISSIPP} \\ T_{11} \text{ \$mississippi} \\ T_{12} \text{ mississippi} \end{array}$

$\begin{array}{l} 12 \$ \\ 11 i \$ \\ 8 i pp\$ \\ 5 issipp\$ \\ 2 ississipp\$ \\ 1 mississippi\$ \\ 10 pi\$ \\ 9 pp\$ \\ 7 si pp\$ \\ 4 si sippi \\ 6 ssipp\$ \\ 3 ssissippi\$ \end{array}$
 \Rightarrow

$T[1, n]$ sorted
alphabetically

PSEUDO:

SA-build (char* T , int n , char* SA)

for ($i=0$; $i < n$; $i++$)

$SA[i] = T + i;$

QSORT (SA, n , sizeof(char*), suffix-cmp)}

suffix-cmp (char* P , char* Q) {

return strcmp (* P , * Q) }

$O(n^2 \log n)$

L2 77

T =

WIMBOW

COPY < distance, length, char ->
of the copy

T = | A | B | A B A B C |
2 1 $\underbrace{\hspace{1cm}}$
<0, 0, A> <0, 0, B> <2, 4, C>

T = | A | A C | A A C A B |

<0, 0, A> <1, 1, C> <3, 4, B>

L7 SS: points: copy, not copy
<d, l> or <0, characters
 \downarrow
no extra character

T = A | A | C | A A C A | B

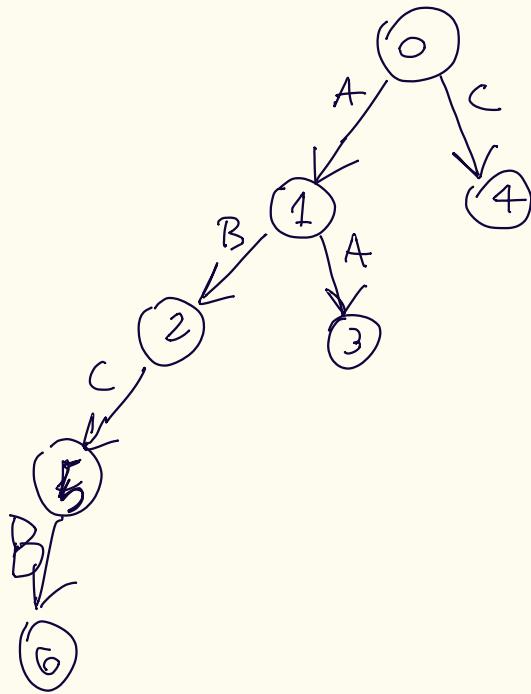
<0, A> <1, 1> <0, C> <3, 4> <0, B>

L7 78

Build tree while processing;

$$T = \begin{array}{c|c} A & A \mid B \\ \hline A & A \mid C \\ \hline A & B \mid C \\ \hline A & B \mid C \mid B \end{array}$$

$\langle 0, A \rangle \langle 1, B \rangle \langle 1, A \rangle \langle 0, C \rangle \langle 2, C \rangle \langle 5, D \rangle >$



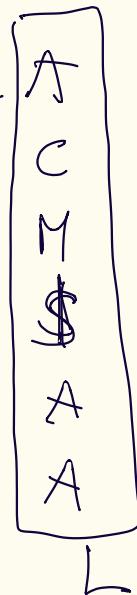
BWT + MTF + RLφ + HUFFMAN

S = A M A C A \$

+ A M A C A \$
 + M A C A \$ A
 + A C A \$ A M
 + C A \$ A M A
 + A \$ A M A C
 • \$ A M A C A

sort

\$ A M A C
 A \$ A M A C
 A C A \$ A M
 A M A C A \$
 C A \$ A M A C
 M A C A \$ A



L = A C M \$ A A
 0 1 2 3

BW(L) = < L without \$, POS of \$ > = < \underbrace{A C M A A}_1, 3 >

2) MTF TO $\stackrel{1}{\square}$

When output more char to front

list of symbols sorted

$L = \{ \begin{matrix} 0 & 1 & 2 \\ A & C & M \end{matrix} \}$

C A M

M C A

A M C

3) RLφ TO MTF of $\stackrel{1}{\square}$

1st ADD 4 to # that are not 0

0 1 2 2 0 $\xrightarrow{\text{ADD 1 to non zero bits}}$ 0 2 3 3 0

ENCODE THE RUNS OF ZEROES WHEELER'S WORK \rightarrow , M & run etc.

\Rightarrow

0 2 3 3 0

JUST ON ZERO

000
0

- take left
- add 1 ($l+1$)

$$l=1 \rightarrow \text{sm}(l+1) = \text{sm}(2) \leq 10$$

\Rightarrow drop 1

\Rightarrow OUTPUT 0

- binary ($lRl+1$)

$$\text{sm}(4) = 100$$

- remove most significant bit

\Rightarrow OUTPUT: 00

0 2 3 3 0

$$\Rightarrow l=2 \quad \text{sm}(2) = 10$$

\Rightarrow OUTPUT 0

0 2 3 3 0

if I've

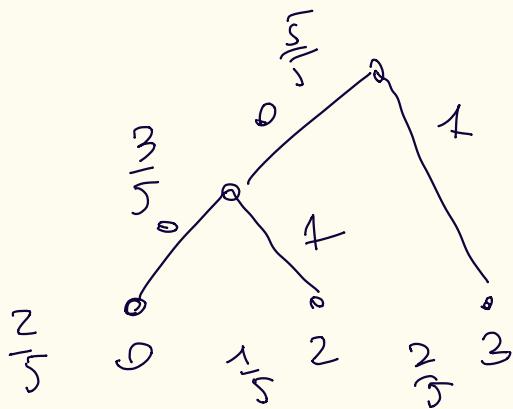
2 3 0 0

\Rightarrow 231

$$l=2$$

$$\text{sm}(l+1) = \text{sm}(3) = 11$$

\Rightarrow OUT 1



0 \Rightarrow 00

2 \Rightarrow 01

3 \Rightarrow 1

0 2 3 3 0 \Rightarrow 00011100

DECODE : 023301 $k=3$ $S = \{ A, C, M \}$

Bwt + NTF + RLEφ

1

→ RLEφ 0 2 3 3 0
 $\Rightarrow \begin{matrix} 0 \\ 2 \\ 3 \\ 3 \\ 0 \end{matrix}$

-1 #s $\Rightarrow \overline{01220}$

→ NTF

0 1 2 2 0
 $\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ TA & C & M & A & A \end{matrix}$

$S = \{ \begin{matrix} 1 \\ 2 \\ A \\ C \\ M \end{matrix} \}$
 $C = \{ \begin{matrix} A \\ M \\ C \\ T \end{matrix} \}$
 $M = \{ \begin{matrix} A \\ C \\ M \end{matrix} \}$

$I = \langle A, C, M, A, A, 3 \rangle$

$L = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ A & C & M & $ & A & A & A \end{bmatrix}$

$C = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \$ & A & C & M & A & C & M \end{bmatrix}$

$L \rightarrow C \rightarrow LF$

$LF = [\begin{matrix} 1 & 4 & 5 & 0 & 2 & 3 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{matrix}]$

values by LF
 $k = 0; 1; 4; 2; 5$

$S =$

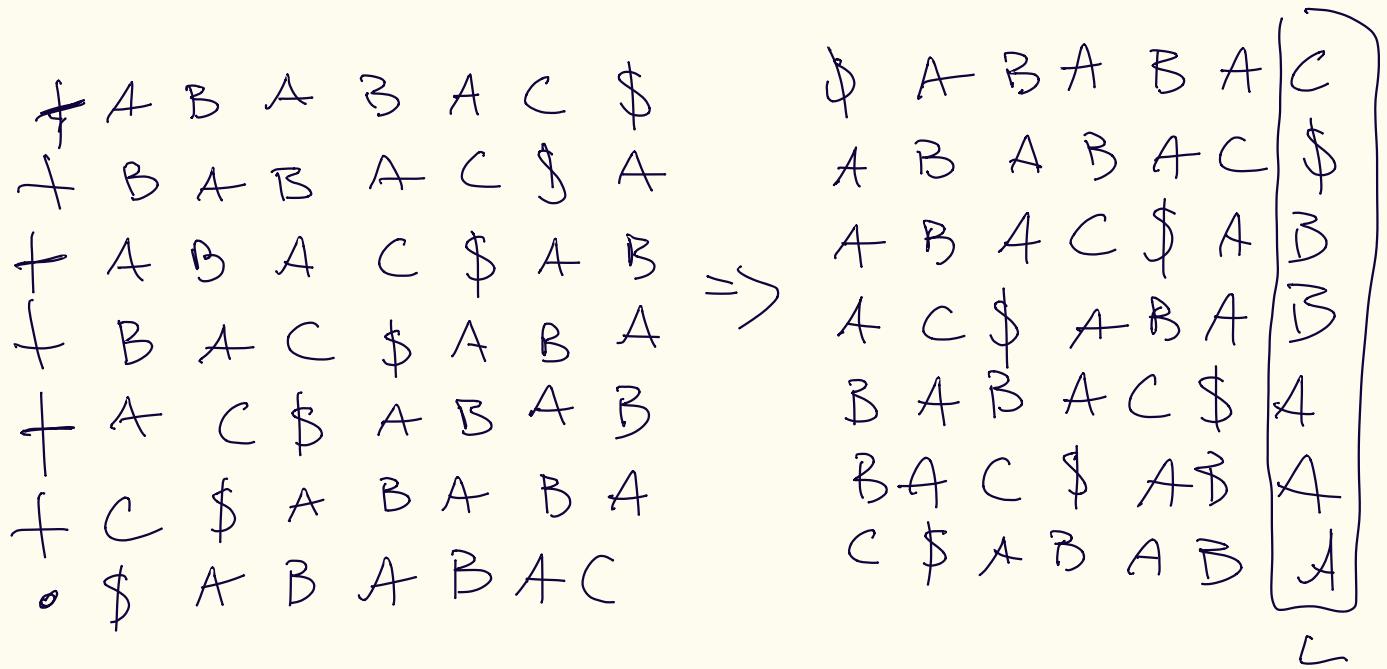
A M A C A \downarrow 1

$K \rightarrow L \rightarrow S$

update
 K with LF

BWT + MTF + runed + HUFFMAN

S = A B A B A C



$$L = \text{BWT}(S) = \langle \text{MTF}(\text{Huffman}(S)), 1 \rangle$$

— 0 —

MTF 1

~~\$~~ B B A A A

↓ ↓ ↓ ↓ ↓

2 2 0 2 0 0

$\{ \overset{1}{A}, \overset{2}{B}, \overset{3}{C} \}$

C A B

B C A

A B C

RLE \emptyset

3 3 0 3 0 0
3 3 0 3 1

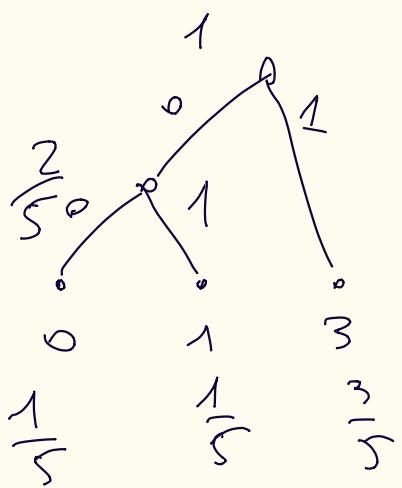
$$\ell(\#z_1) = 2$$

$$\text{sum}(z) = \#0$$

$$e(3) = 3$$

$$\text{sum}(3) = \#1$$

Huffman 3 3 0 3 1 => 11 00 100



0 ~> 00

1 ~> 01

3 ~> 1

BWT + MTF + RLE + HUFFMAN

S = C A B A B C A

+ C A B + B C A \$
 + A B A B C A \$ C
 + B A B C A \$ C A
 + A B C A \$ C A B
 # B C A \$ C A B A \Rightarrow
 + C A \$ C A B A B
 + A \$ C A B A B C
 + \$ C A B A B C A

\$ C A B + B C C
 A \$ C A B A B
 A B A B C A \$ C
 A B C A \$ C A B
 B A B C A \$ C A
 B C A \$ C A B A
 C A \$ C A B A B
 C A B A B C A

BW(L) = {A C C B A A B, +}

MTF(L)

A C C B A A B
 + Z O Z Z Z O 1
 0

L 0 1 2
 L A B C
 L1 C A B
 L2 B C A

L3 A B C
 L4 B A C

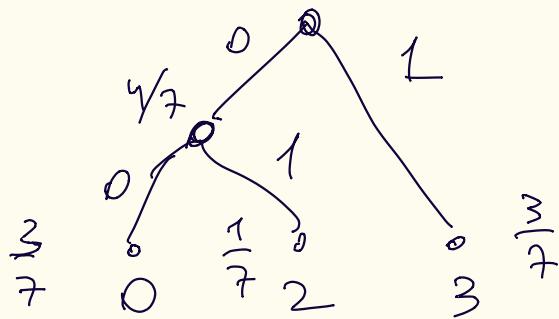
RLE ϕ

0 3 0 3 3 0 2
 0 3 0 3 3 0 2

$l(L_1) = 2$ $l_m(L_2) = 10$

idley
idley

HUFFMAN



$0 \rightarrow Q_0$
 $2 \rightarrow Q_1$
 $3 \rightarrow 1$

$0 \ 3 \ 0 \ 3 \ 3 \ 0 \ 2$
 $\Rightarrow 00100 + 10001$

Decode 0303302 $R=7$ $S = \{A, B, E\}$

BWT
 $nif + nif$.
 \Rightarrow reuse
 that
 nif

$\neg R \in \emptyset$ $0303302 \Rightarrow 0$ still zero

0303302 ($1+6+0$)
 $\Rightarrow 0202201$

MTF

$02 \ 02201$
 $\downarrow \downarrow \downarrow \downarrow \downarrow$
 $A \ C \ C \ B \ A \ A \ B$

$\hat{L} : LACCBABAAB$

$\begin{cases} 0 \\ 1 \\ ? \end{cases}$
 $S = \{A, B, C\}$
 $C \ A \ B$
 $B \ C \ A$
 $A \ B \ C$

$L = - A \ C \ C \ B \ A \ A \ B \$$

$\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix}$

 $C = \left[\begin{matrix} 1 & 3 & 4 & 6 & 5 & 7 & 8 \\ 0 & 2 & 4 & 4 & 6 & 7 & 8 \\ A & B & C & C & B & A & C \end{matrix} \right]$

$L^T = [\begin{matrix} 1 & 5 & 7 & 4 & 2 & 3 & 5 & 0 \end{matrix}]$

$L \rightarrow C \rightarrow LF$

$K \rightarrow LF \rightarrow S$

$K = 0; 1; 6; 5; 3; 4; 2; 7$

$S = CABABCAC$



BWT + MTF RLEφ + Huffman

S = amata

+ A M A T A \$
+ M A T A \$ A
+ A T A \$ A M
+ T A \$ A M A \Rightarrow
+ A \$ A M A T
+ \$ A M A T A

\$ A M A T A
A \$ A M A T
A M A T A \$
A T A \$ A M
M A T A \$ A
T A \$ A M A

Bwt(L) = < A T ^M A A, 2 >

MTF(L)

A T M A A
↓ ↓ ↓ ↓ ↓
0 2 2 2 0

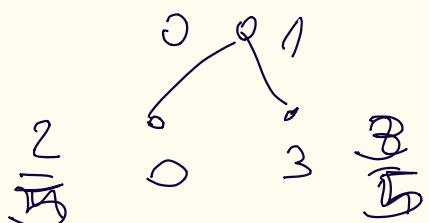
L = 0 1 2
A N T
T A M
N T A
A M T

RLEφ: (0 2 2 0) \rightarrow (0 3 3 0)

bm(1) = 10

$\Rightarrow \boxed{0 3 3 0}$ idm (\Rightarrow bm(2))

Huffman



0 => 0
3 => 1

0 3 3 0 \Rightarrow 0 4 1 1 0

BWT + MET + RLEq + HUFFMATE

S = amata

A M A T A \$
 M A T A \$ A
 A T A \$ A M
 T A \$ A M A
 A \$ A M A T
 \$ A M A T A



\$	A	M	A	T	A
A	\$	A	M	A	T
A	M	A	T	A	S
A	T	A	\$	A	M
M	A	T	A	\$	A
T	A	\$	A	M	A



BW(L) = < \underbrace{ATMAA}_{L}, 2 >

MTF(L)

A T M A A
 ↓ ↓ ↓ ↓ ↓
 0 2 2 2 0

L : {0 1 2 }
 A M T
 T A M
 M T A
 A M T

RL ≠ φ

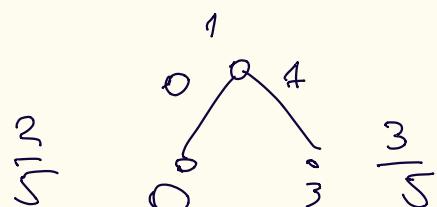
0 2 2 2 0
 => 0 3 3 3 0
 l=1 l=1
 => 0 3 3 3 0

sum (l+1) = 10
 of num

sum (l+1) = 10

Huffman encoding

0 => 0
 3 => 1



0 3 3 3 0
 0 1 1 1 0



BWT + MTF + RLEφ + Huffman

$S = C B A B A A \bar{A}$

L

$\begin{array}{r} + C B A B A B A A \$ \\ + B A B A B A A \$ C \\ + \underline{A B A} B A A \$ C B \\ + B A B A A \$ C B A \\ + \underline{A B A} A \$ C B A B \\ + \underline{B A A} \$ C B A B A \\ + \underline{A A} \$ C B A B A B \\ + \underline{A \$} C B A B A B A \\ + \$ C B A B A B A A \end{array} \Rightarrow$

\$	C	B	A	B	A	B	A	A
A	\$	C	B	A	B	A	B	A
AA	\$	C	B	A	B	A	B	B
AB	AA	\$	C	B	A	B	A	B
AB	AB	A	A	\$	C	B	B	B
BAA	\$	C	B	A	B	A	\$	A
BAB	AA	\$	C	B	A	B	A	A
BABA	B	A	B	A	A	\$	C	C
cBABA	B	A	B	A	A	A	\$	

$BW(L) = \langle \overbrace{A A B B B A A C}, 8 \rangle$

$MTF(\vec{z}):$

$\begin{array}{ccccccccc} A & A & B & B & B & A & A & C \\ \downarrow & \downarrow \\ 0 & 0 & 1 & 0 & 0 & * & 0 & 2 \end{array}$

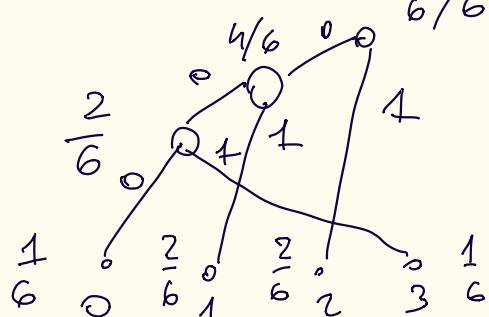
$I = \left\{ \begin{array}{ccc} 0 & 1 & 2 \\ A & B & C \\ B & A & C \\ A & B & C \\ C & A & B \end{array} \right\}$

$RLE\phi(00100102)$

$\Rightarrow \underbrace{00}_e 2 \underbrace{00}_e 2 \underbrace{0}_l 3$

 $\boxed{1 \ 2 \ 1 \ 2 \ 0 \ 3}$

 $6/6$



$bin(2+1) = 11$

$bin(2+1) = 11$

$bin(1+1) = 10$

$0 \Rightarrow 000$
 $1 \Rightarrow 01$
 $2 \Rightarrow 1$
 $3 \Rightarrow 001$

121203

$= 01101100001$

DECODE (BWT + RLE + Huffman...) the compressed sequence:

[0 3 3 3 1] with $R=3 \rightarrow$ position of the $\$$
assuming an alphabet $\{A, B, C\}$

1) INVERT RLE $\$$

$\begin{array}{c} 0 \\ \downarrow \\ 0 \end{array} \quad \begin{array}{c} 3 \\ \downarrow \\ 3 \end{array} \quad \begin{array}{c} 3 \\ \downarrow \\ 3 \end{array} \quad \begin{array}{c} 1 \\ \downarrow \\ \text{run} \end{array} \rightarrow \boxed{0 \ 3 \ 3 \ 3 \ 0 \ 0}$

always $0 \Rightarrow$ reserved
 4 run

$\Rightarrow \boxed{0 \ 2 \ 2 \ 2 \ 0 \ 0}$

$\begin{array}{c} 1 \\ \downarrow \\ 2 \\ \downarrow \\ 2 \end{array} \rightarrow \begin{array}{c} 2 \\ \downarrow \\ \text{zeros} \end{array}$

$\ell = 2 \quad \sin(\frac{2}{3}+1) = \frac{5}{3} = 1.67 \approx 2$

$\Rightarrow \text{ok}$

2) INVERT HTF

$\begin{array}{cccccc} 0 & 2 & 2 & 2 & 0 & 0 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ A & C & B & A & A & A = \Sigma \end{array}$

$R=3 \quad \text{BW}(L) \langle A \ C \ B, A \ A \ A, 3 \rangle$

$L = \{ \begin{array}{c} 0 \ 1 \ 2 \\ A \ B \ C \\ C \ A \ B \\ B \ C \ A \\ A \ B \ C \end{array} \}$

$L = A \ C \ B \$ A \ A \ A$

3) INVERT BWT

$C = \begin{bmatrix} 0 & 1 & 5 & 6 \\ \$ & A & B & C \end{bmatrix}$

C-ARRAY

& symbol, count in L are # of occurrences < symbol considered
ex: \\$ has no symbol's occurrences < than itself

$|LF| = |L|$
 $LF = [1 \ 6 \ 5 \ 0 \ 2 \ 3 \ 4]$
 INDEX
 0 1 2 3 4 5 6

FOR $i = 0$ to $|L|-1$
 $LF[i] = C[L[i]];$
 $C[L[i]]++;$

$L = A \ C \ B \ $ \ A \ A$
 $C = [1 \ 0 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9]$
 INDEX
 0 1 2 3 4 5 6 7 8 9

$LF = [1 \ 6 \ 5 \ 0 \ 2 \ 3 \ 4]$
 INDEX
 0 1 2 3 4 5 6

$k=0$
 FOR $i = |S|-1$ DOWN TO 0

$S[i] = L[k]$
 $k = LF[k]$

RECONSTRUCT S
 UPDATE BY POS GIVEN BY LF
 $k = 2$
 $i = 6; 5; 4; 3; 2; 1$
 $S = A B A A C A$
 $L = A \ C \ B \ $ \ A \ A \ A$
 INDEX
 0 1 2 3 4 5 6

A B A A C A

Decode, skipping Huffman shot: BWT + MTF + RLE

0 3 3 3 0 ; R=2

$S = \{A, M, T\}$

→ RLE 0 3 3 3 0 → 0 2 2 2 0

→ 0 2 2 2 0 $\text{wt bin}(\ell_{1,1}) = 10$

→ MTF

0 2 2 2 0
 ↓ ↓ ↓ ↓ ↓
 A T M A A

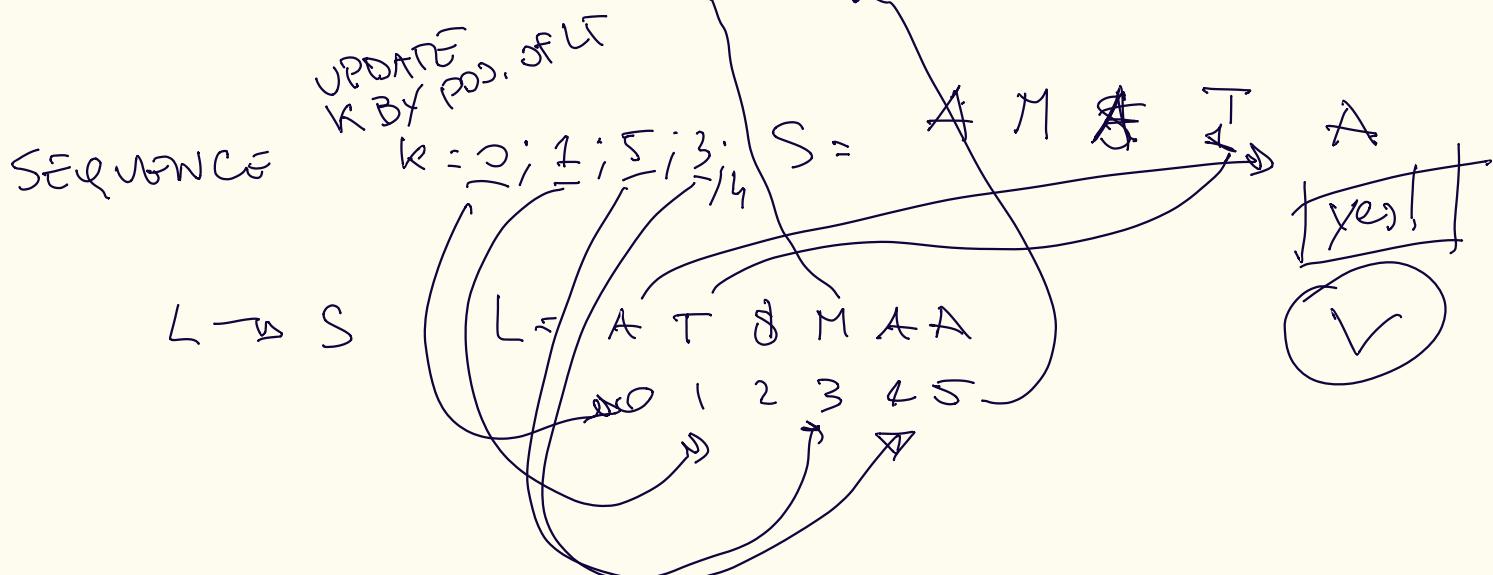
L = {0, 1, 2}
 T A M
 M T A
 A M T

} = <ATM AA, 2>

L = A T \$ M A A
 ↓ ↓ ↓ ↑ ↑ ↑
 C = [0 1 2 3 4 5 6]
 ↓ ↓ ↓ ↓ ↓ ↓
 A M T

L → C → DLF

L_F = [1 5 0 4 2 3]
 0 1 2 3 4 5

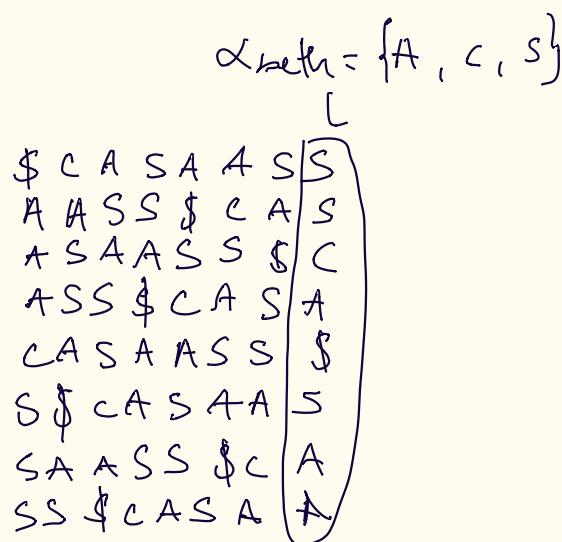


BWT + MTF + RLE + Huffman costs

S = CASAASS

+ C A S A A S S \$
+ A S A A S S \$ C
+ S A A S S \$ C A
+ A A S S \$ C A S
+ A S S \$ C A S A
+ S S \$ C A S A A
+ S \$ C A S A A S
+ \$ C # S A A S S

\Rightarrow



L = SSCA \$ S A A

$BW(L) = 2SS \underbrace{C A S A A}_{L}, 4>$

MTF

$\hat{L} = \begin{matrix} S & S & C & A & S & A & A \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 0 & 2 & 2 & 1 & 0 \end{matrix}$

RLE

+ 1 + 0 + 0

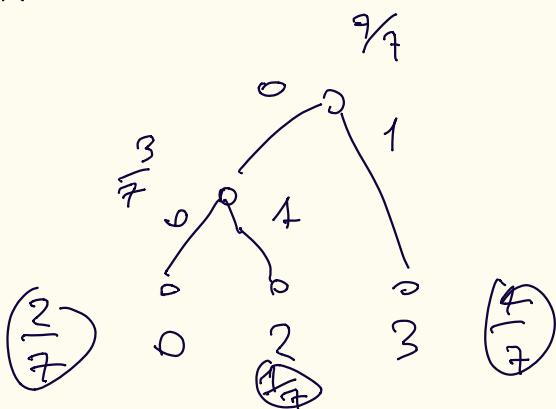
$\Rightarrow \begin{matrix} 3 & 0 & 3 & 3 & 3 & 2 & 0 \\ \downarrow & \downarrow & & & & \downarrow & \downarrow \\ P & & & & & P & P \end{matrix}$

$\Rightarrow 3033320$

$\begin{cases} 0 & 1 & 2 \\ A & C & S \end{cases}$
S A C
C S A
A C S
S A C
A S C

1st run
2nd run
 $b \ln(\frac{L_{\text{out}}}{\text{out}} + 1) \leq \beta = b \ln(2) = 10$
 $\Rightarrow \text{out} = \text{for } \alpha \text{ & } \beta$

Huffman



0 $\Rightarrow 00$
2 $\Rightarrow 01$
3 $\Rightarrow 1$

$3033320 \Rightarrow 1001110100$

$s = CASAASS$

$\Sigma = \{ A, C, S \}$

BWT

↗ CA S A A S S \$
 ↘ A S A A S S \$ C →
 ↗ \$ C A A S S \$ C A
 ↗ A A S S \$ C A S
 ↗ A S S \$ C A S A
 ↗ S S \$ C A S A A
 ↗ S \$ C A S A A S
 ↗ \$ C A S A A S S

↗
 ↗ \$ C A S A A S S
 ↗ A A S S \$ C A S
 ↗ A S A A S S \$ C
 ↗ A S S \$ C A S A S
 ↗ C A S A A S S \$
 ↗ S \$ C A S A A S
 ↗ S A A S S \$ C A
 ↗ S S \$ C A S A A S

$L = S S C A \$ S A A$ $Bw(L) = \langle S S C A S A A, 4 \rangle$

MIF(C)

S S C A S A A
 ↓ ↓ ↓ ↓ ↓ ↓
 2 0 1 2 2 1 0
 ↓
 3 0 2 3 3 2 0
 ↑
 1 s_m(n) = x_0

+1

0 1 2
 S - A C S
 S C A
 C S A
 A C S
 S A C
 A S C

3 0 2 3 3 2 0

Mutantech --

=> double

length

3 0 2 3 3 2 0 =>

3 0 2 3 3 2 0

MIF
 2 0 1 2 2 1 0
 ↓
 S

2 0 1 2 2 1 0
 ↓
 A C S
 S

LZ77

$$S = |A|L|A\ B|A\ R|A\ L\ A\ L|A\ B\ A\ R\ D\ A$$

$\langle 0, 0, A \rangle \langle 0, 0, L \rangle \langle 2, 1, B \rangle \langle 2, 1, R \rangle$

$\langle 6, 3, L \rangle \langle 8, 4, D \rangle \langle 3, 1, \underbrace{?}_{\text{EOF}} \rangle$

DISTANCE,
copy,
char

Decode \rightarrow

A L A B A R - A L A L A B A R D A /

LZ77 with window size ≤ 4 1st param ≤ 4

$$S = |A|L|A\ B|A\ R|A\ L\ A\ B|A\ R|D|A|$$

$\langle 0, 0, A \rangle \langle 0, 0, L \rangle \langle 2, 1, B \rangle \langle 2, 1, R \rangle$

$\langle 2, 1, L \rangle \langle 2, 3, B \rangle \langle 2, 1, R \rangle \langle 0, 0, D \rangle$

$\langle 3, 1, EOF \rangle$

L7 SS \Rightarrow pairs

copy, not copy
 $\langle d, l \rangle$ or $\langle 0, \text{character} \rangle$
 ↓
 no extra character

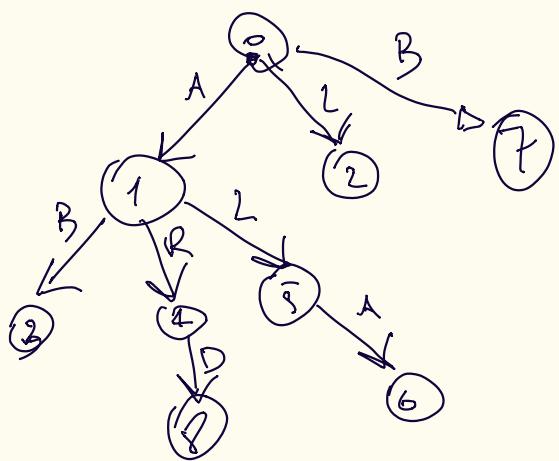
$$S = |A|L|A|B|A|R|A\ L\ A|L\ A\ B\ A\ R|D|A|$$

$\langle 0, A \rangle \langle 0, L \rangle \langle 2, 1 \rangle \langle 0, B \rangle \langle 2, 1 \rangle \langle 0, R \rangle$

$\langle 6, 3 \rangle \langle 8, 5 \rangle \langle 0, D \rangle \langle 3, 1 \rangle$

↙ 7 73

o Construct Dictionary



↳ source node, char applied

$$\langle 0, A \rangle \langle 0, L \rangle \langle 1, B \rangle$$

$\langle 1, R \rangle$ $\langle 1, L \rangle$ $\langle 5, A \rangle$

$\langle \text{O}, \text{B} \rangle \langle \text{A}, \text{D} \rangle \langle \text{O}, \text{A} \rangle$

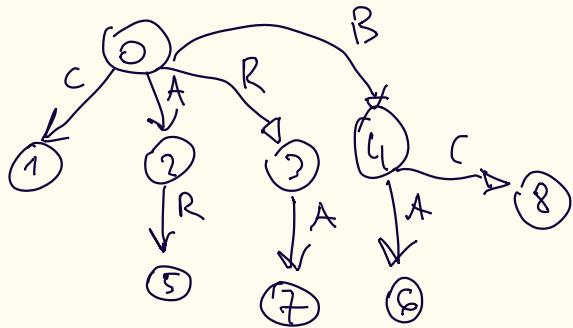
$\angle 1, \text{ e.o.t}$

$$S = C | A | R | B | A \ R \ B \ A \ R \ A | B \ C$$

L277

$\langle O_1, O_1, C \rangle \langle O_1, O_1, A \rangle \langle O_1, O_1, R \rangle \langle O_1, O_1, B \rangle \langle S_1, S_1, A \rangle$
 $\langle 4, 1, C \rangle$

(L278) $S = C | A | R | B | A \ R | B \ A | R \ A | B \ C$

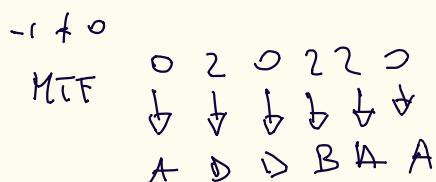
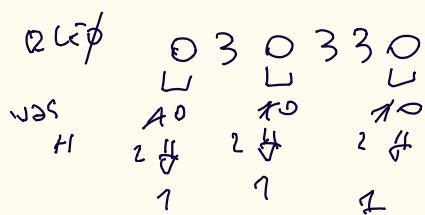


$\langle O_1, C \rangle \langle O_1, A \rangle$
 $\langle O_1, R \rangle \langle O_1, B \rangle$
 $\langle 2, R \rangle \langle 4, A \rangle$
 $\langle 3, A \rangle \langle 4, C \rangle$

DECODE BZIP:

$$\Sigma = \{A, B, C\}$$

is in Σ^* pos

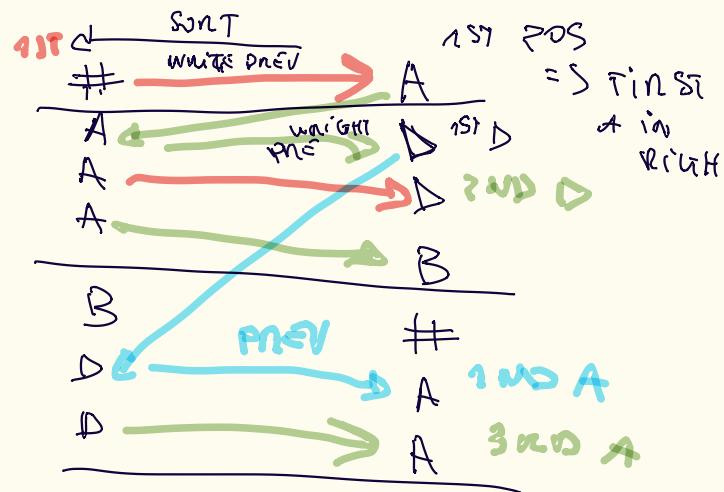


$$L = A D D B \# A A$$

BWT

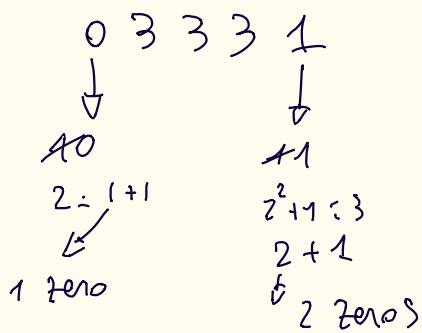
B A D A D A #

$$\begin{array}{c} 0 \quad 1 \quad 2 \\ \downarrow \quad A \quad B \quad D \\ D \quad A \quad B \\ B \quad D \quad A \\ A \quad B \quad D \end{array}$$



DECODER

RLE0



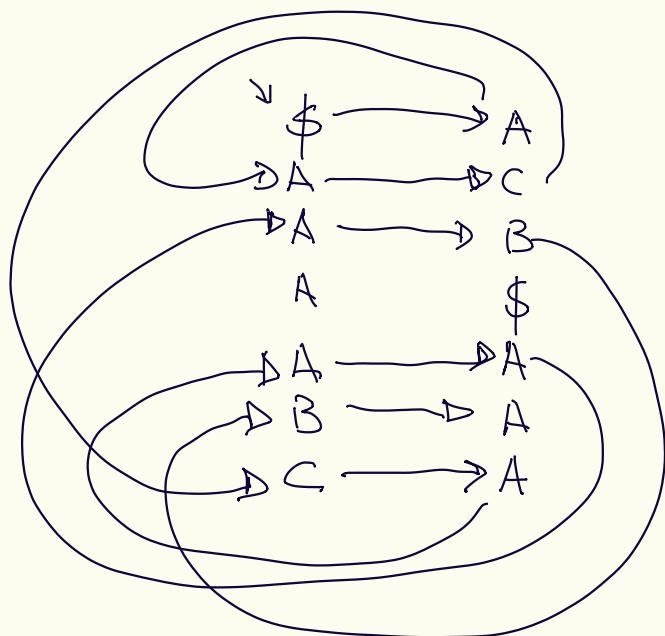
$$\Sigma = (A, B, C)$$

$R \approx 3$

$\Rightarrow 0 2 2 2 0 0$
MTP A C B A A A

$\Rightarrow L: A \subset B \subset \$ \subset A \subset A$

$\begin{matrix} & 1 & 2 \\ A, & B, & C \\ C & \uparrow & B \\ B & \subset & A \\ A & \subset & C \end{matrix}$



ABAAACAA\$