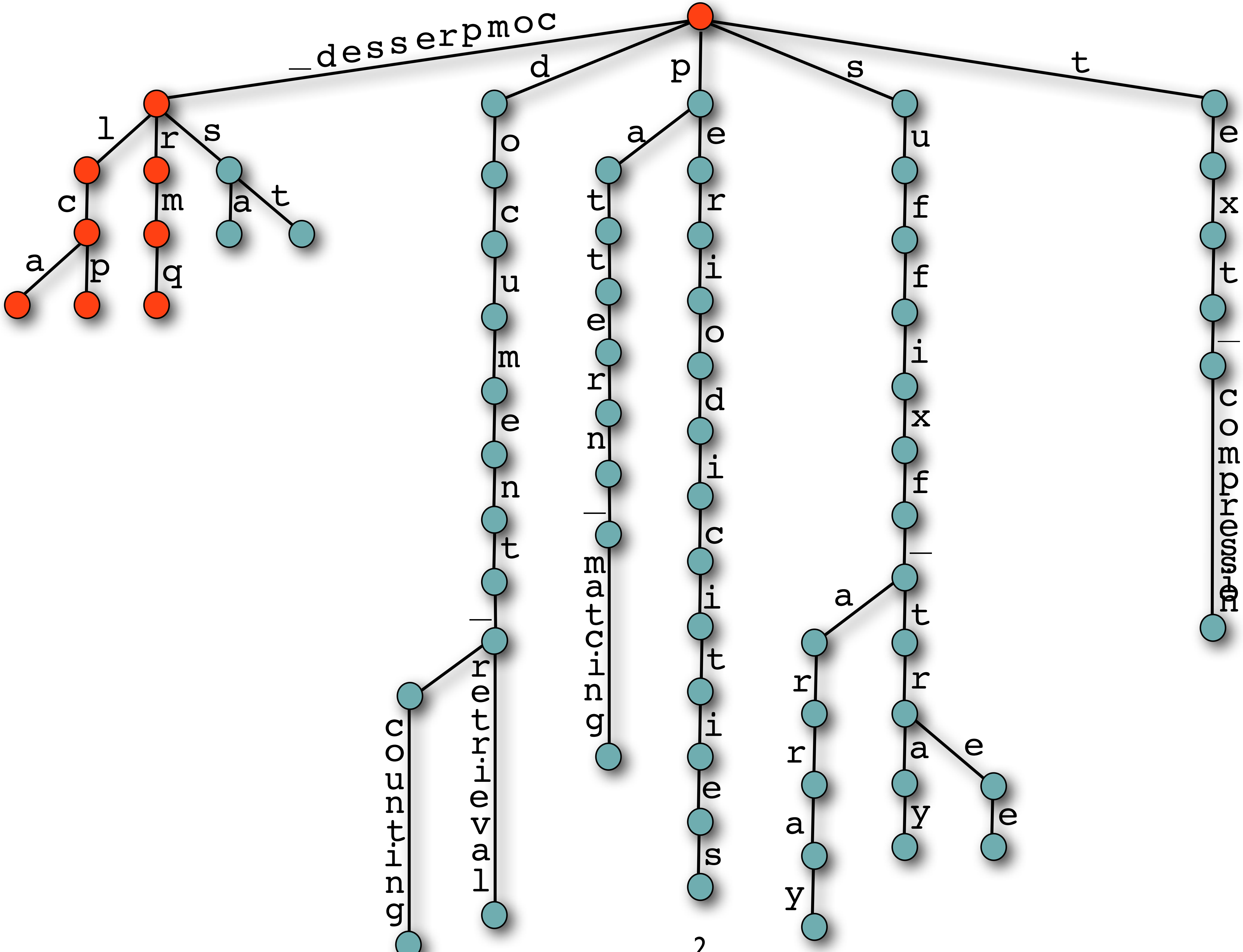
Lecture 11: Compressed LCPs/RMQs/PSVs/NSVs

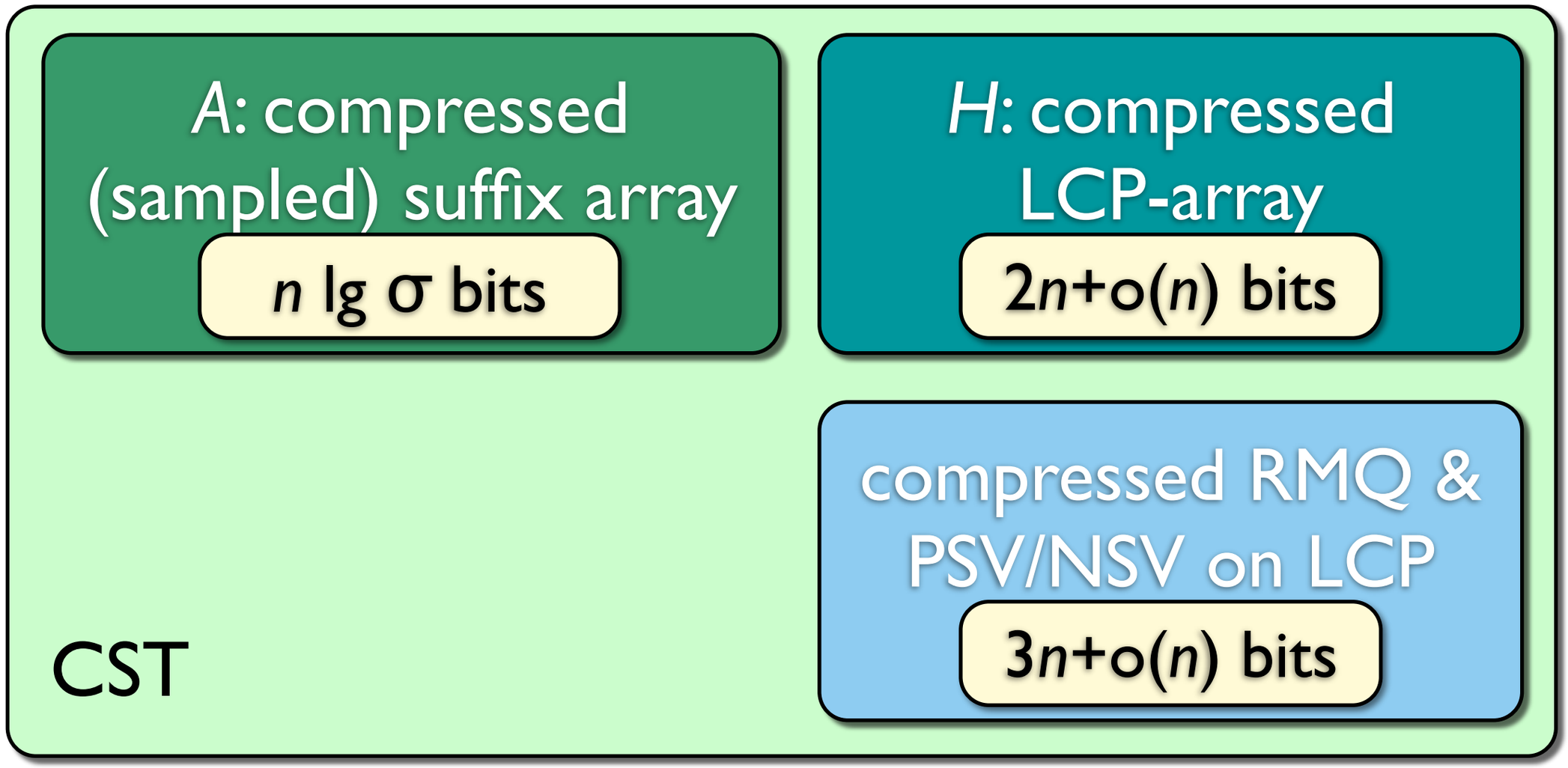
Johannes Fischer

Topics



_desserpmoc
d
p
s
t
l
r
s
c
m
a
t
a
p
q
o
c
u
m
e
n
t
_
counting
retrieval
a
t
t
e
r
n
_
matching
e
r
i
o
d
i
c
i
t
i
e
s
u
f
f
i
x
f
_
a
r
r
a
y
t
r
a
y
e
e
e
x
t
_
compression

3 Components of CST



ST Operations

Operation	Description	Time
ROOT()	return root	O(1)
COUNT(*v*)	count leaves below *v*	O(1)
ISANCESTOR(*v*,*w*)	true if *v* is an ancestor of *w*	O(1)
ISLEAF (*v*)	true if *v* is a leaf	O(1)
LEAFLABEL(*v*)	suffix number represented by leaf *v*	$O(t_{SA})$
SDEPTH(*v*)	string depth of *v*	$O(t_{SA}+t_{LCP})$
PARENT(*v*)	parent node of *v*	$O(t_{LCP}+t_{PNSV})$
FIRSTCHILD(*v*)	first (alphabetically smallest) child of *v*	$O(t_{RMQ})$
NEXTSIBLING(*v*)	next sibling of *v*	$O(t_{LCP}+t_{PNSV}+t_{RMQ})$
EDGELABEL(*v*,*i*)	*i*'th letter on the edge leading to *v*	$O(t_{SA}+t_{LCP}+t_{PNSV}+t_{RMQ})$
LCA(*v*,*w*)	lowest common ancestor of *v* and *w*	$O(t_{RMQ}+t_{PNSV})$

Part I:

Compressing LCP

Compressed LCP-Array

- *H'*=LCP-values in **text order**: values do not decrease dramatically

H' = 2,3,2,1,0,1,0,-1

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

H = -1,0,1,2,1,3,0,2

→**Lemma**: $H'[i] \geq H'[i-1]-1$ for all $i>1$

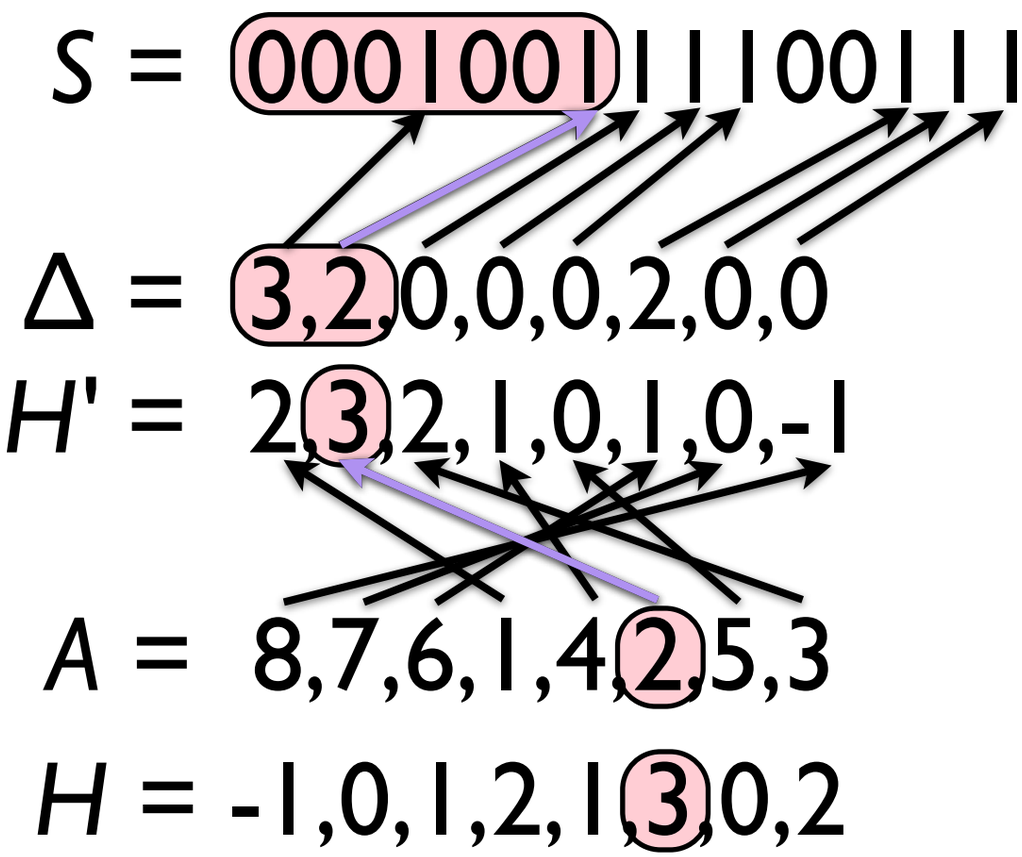
Compressed LCP-Array

- Lemma: $H'[i] \geq H'[i-1]-1$ for all $i>1$
  - $\Delta[i]:=H'[i]-H'[i-1]+1 \geq 0$ for all i (say $H'[0]=0$)
- **Idea**: encode $\Delta[1], \Delta[2],..., \Delta[n]$ in **unary**

$$S = 0^{\Delta[1]}1\,0^{\Delta[2]}1\,...0^{\Delta[n]}1$$

- Size of S is $\boldsymbol{2n-1}$ bits:
  - #1's is n
  - $\#0\text{'s} = \Sigma\Delta[i] = \Sigma(H'[i]-H'[i-1]+1) = n-\overbrace{H'[0]}^{=0}+\overbrace{H'[n]}^{=-1}$

Compressed LCP-Array



- $H[i]=H'[A[i]]=\Sigma_{j\le A[i]}\Delta[j]-A[i]$

  $\Sigma_{j\le A[i]}\Delta[j]=\text{rank}_0(S, \text{select}_1(S, A[i])) =\text{select}_1(S, A[i])-A[i]$

Compressed LCP-Array

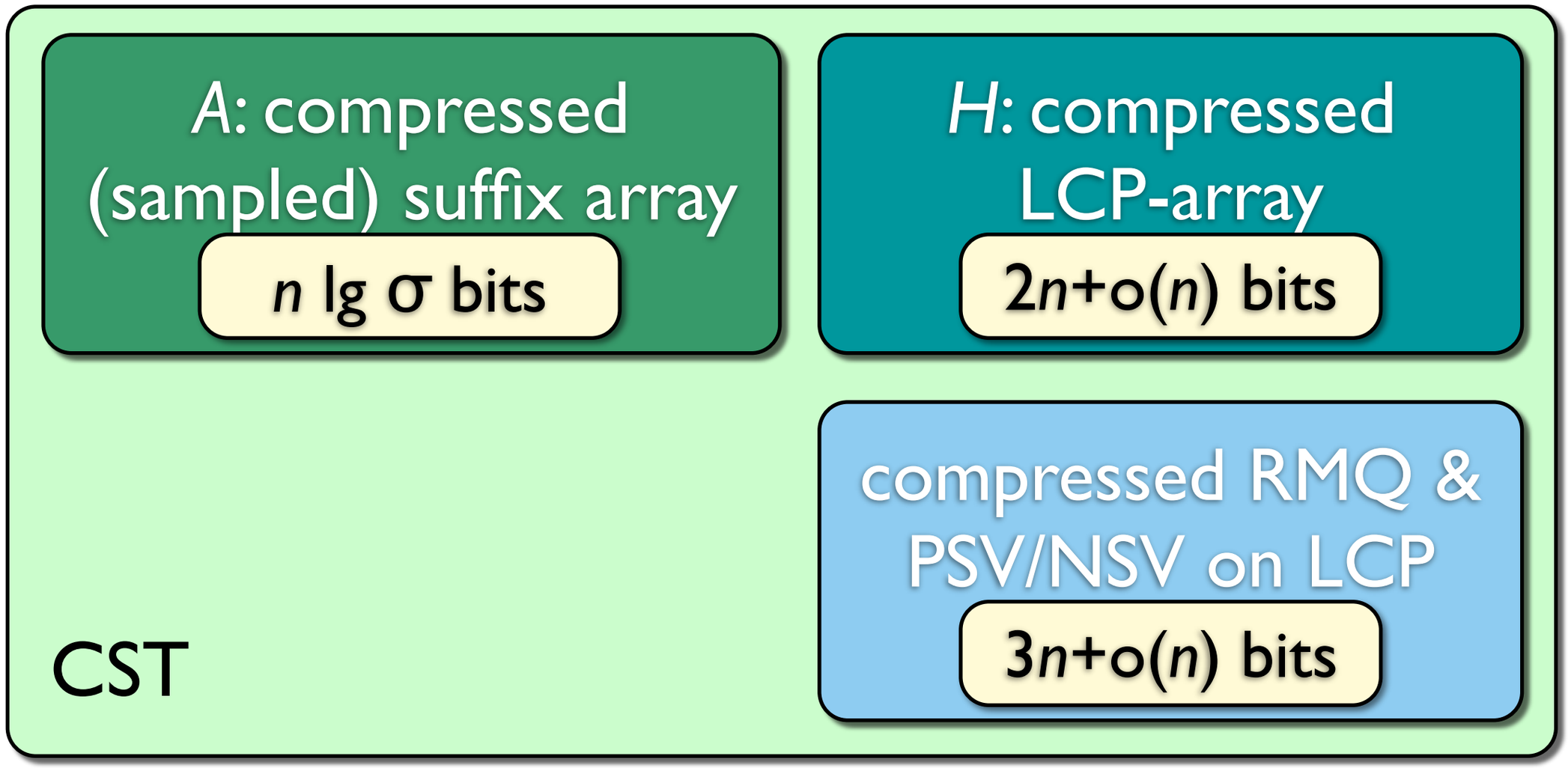
- **Summary**:
  - Replace LCP array H by bit-vector S
  - + $O(1)$ select on S

  $\Longrightarrow 2n+o(n)$ bits
- Get $H[i]$ by $\text{select}_1(S, A[i])-2A[i]$ in time t_{SA}
- in **practice**: $\approx 3n$ bits, $\approx 10\mu s$/query + t_{SA}

Part II:

Compressing RMQ/ PSV/NSV

3 Components of CST



ST Operations

Operation	Description	Time
ROOT()	return root	O(1)
COUNT(*v*)	count leaves below *v*	O(1)
ISANCESTOR(*v*,*w*)	true if *v* is an ancestor of *w*	O(1)
ISLEAF (*v*)	true if *v* is a leaf	O(1)
LEAFLABEL(*v*)	suffix number represented by leaf *v*	O(lg *n*)
SDEPTH(*v*)	string depth of *v*	O(lg *n*)
PARENT(*v*)	parent node of *v*	O(lg *n*)
FIRSTCHILD(*v*)	first (alphabetically smallest) child of *v*	O(1)
NEXTSIBLING(*v*)	next sibling of *v*	O(lg *n*)
EDGELABEL(*v*,*i*)	*i*'th letter on the edge leading to *v*	O(lg *n*)
LCA(*v*,*w*)	lowest common ancestor of *v* and *w*	O(1)

$t_{LCP}=O(t_{SA})=O(\lg n)$, $t_{RMQ}=t_{PNSV}=O(1)$