Suffix Arrays: the suffix tree is hiding

Ben Langmead

Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).
Suffix array

Both encode trees

Further: the suffix tree can be recovered from the suffix array
Pre-compute and record LCEs for each adjacent pair of suffixes
Pre-compute and record LCEs for each adjacent pair of suffixes
Pre-compute and record LCEs for each adjacent pair of suffixes
Suffix array

Pre-compute and record LCEs for each adjacent pair of suffixes
Suffix array

Pre-compute and record LCEs for each adjacent pair of suffixes
**Suffix array**

Su
ffi
x array

14 0 7 3 12 5 10 1 8 4 13 6 11 2 9

LCE: 0 4 1 1 2

Pre-compute and record LCEs for each adjacent pair of suffixes
Precompute and record LCEs for each adjacent pair of suffixes
Suffix array

Pre-compute and record LCEs for each adjacent pair of suffixes
Suffix array

Pre-compute and record LCEs for each adjacent pair of suffixes

LCE: 0 4 1 1 2 3 0 3 0 0 1 2 0 2
Suffix array

Pre-compute and record LCEs for each adjacent pair of suffixes

Add "bookends" equal to -1
Certain **intervals** of the SA are $\ell$-**intervals**

An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
Suffix array

An interval \( SA[i, j] \) is an \( \ell \)-interval if:

1. LCEs to either side are \(< \ell\)
2. At least one LCE in the interval is \(= \ell\)
3. All other LCEs in the interval are \(> \ell\)
An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
### Suffix array

<table>
<thead>
<tr>
<th>$</th>
<th>abracadabra$</th>
<th>abrad$</th>
<th>acadabra$</th>
<th>ad$</th>
<th>adabra$</th>
<th>brad$</th>
<th>cadabra$</th>
<th>d$</th>
<th>dabrad$</th>
<th>dad$</th>
<th>radabr$</th>
<th>radad$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA:</td>
<td>14</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td>12</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>13</td>
<td>6</td>
</tr>
</tbody>
</table>

#### LCE:

-1 0 4 1 1 2 3 0 3 0 0 1 2 0 2 -1

An interval SA[$i, j$] is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
See any 2-intervals?

An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
An interval SA\([i, j]\) is an \(\ell\)-interval if:

1. LCEs to either side are < \(\ell\)
2. At least one LCE in the interval is = \(\ell\)
3. All other LCEs in the interval are > \(\ell\)
Suffix array

<table>
<thead>
<tr>
<th>LCE:</th>
<th>-1</th>
<th>0</th>
<th>4</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>0</th>
<th>3</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>0</th>
<th>2</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA:</td>
<td>14</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td>12</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>13</td>
<td>6</td>
<td>11</td>
<td>2</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

An interval SA[i, j] is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
Is there a 0-interval?

An interval $SA[i, j]$ is an $\ell$-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$
An interval \( SA[i, j] \) is an \( \ell \)-interval if:

1. LCEs to either side are \( < \ell \)
2. At least one LCE in the interval is \( = \ell \)
3. All other LCEs in the interval are \( > \ell \)
Suffix array

$ \quad abracadabra$ ad$d$ ad$ ad$ bradad$ dad$ dad$ dad$ radad$

SA: 14 0 7 3 12 5 10 1 8 4 13 6 11 2 9

LCE: -1 0 4 1 1 2 3 0 3 0 0 1 2 0 2 -1

$\ell$-intervals correspond to internal nodes of the suffix tree
Suffix array

Which node is this?
Suffix array

ad$  adabradad$  adad$

12 5 10

1 2 3 0

2-ival

ad$  adabradad$  adad$
Suffix array

Which node is this?
1-intervals are at (label) depth of 1, 2-intervals at depth of 2, etc
Why is this not a 1-interval?

1. LCEs to either side are not both <1

2. It's not an internal node!
What is the "meaning" of the LCEs that are \( \ell \)?

Correspond to "turnovers" from child edge to child edge.
Suffix array

What is the "meaning" of the LCEs that are $= \ell$?

Correspond to "turnovers" from child edge to child edge
What is the "meaning" of the LCEs that are $= \ell$?

Correspond to "turnovers" from child edge to child edge.
What is the "meaning" of the LCEs that are $= \ell$?

Correspond to "turnovers" from child edge to child edge
Suffix array

What is the "meaning" of the LCEs that are $= \ell$?

Correspond to "turnovers" from child edge to child edge
**Suffix array**

What is the "meaning" of the LCEs that are $\ell$?

```
-1 0 4 1 1 2 3 0 3 0 0 1 2 0 2 -1
```

Correspond to "turnovers" from child edge to child edge.
Suffix array

ℓ-intervals correspond to internal nodes

LCEs = ℓ in an ℓ-interval correspond to child "turnovers"

...so quickly finding = ℓ LCEs allows us to quickly find child ℓ-intervals. We can traverse the tree!
Suffix array

-Intervals correspond to internal nodes

LCEs = ℓ in an ℓ-interval correspond to child "turnovers"

...so quickly finding = ℓ LCEs allows us to quickly find child ℓ-intervals. We can traverse the tree!
Suffix array

\( \ell \)-intervals correspond to internal nodes

LCEs = \( \ell \) in an \( \ell \)-interval correspond to child "turnovers"

...so quickly finding = \( \ell \) LCEs allows us to quickly find child \( \ell \)-intervals. We can traverse the tree!

1-interval
Suffix array

ℓ-intervals correspond to internal nodes

LCEs = ℓ in an ℓ-interval correspond to child "turnovers"

...so quickly finding = ℓ LCEs allows us to quickly find child ℓ-intervals. We can traverse the tree!
Suffix array

$\ell$-intervals correspond to internal nodes

LCEs $= \ell$ in an $\ell$-interval correspond to child "turnovers"

...so quickly finding $= \ell$ LCEs allows us to quickly find child $\ell$-intervals. We can traverse the tree!
Suffix array

1-interval  3-interval  1-interval  2-interval

-1  0  4  1  1  2  3  0  3  0  0  1  2  0  2  -1
Suffix array

Recurse
Suffix array

4-interval  2-interval

0 4 1 1 2 3 0

0-interval 2-interval

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$

$ a b r a d a b r a d a d a d$
Suffix array

Recurse

[Diagram of a suffix array with nodes labeled from 0 to 14 and edges connecting them based on the suffixes of the string 'cadabradad$'.]
Suffix array

3-interval

1 2 3 0

14

$ a b r a c a d a b r a d a d$

13

d $ a c a d a b r a d a d$

dad$

4

$ a b r a c a d a b r a d a d$

12

d $ a c a d a b r a d a d$

dad$

3 11

$ a b r a c a d a b r a d a d$

9

$ a b r a c a d a b r a d a d$

d$ dad$

2

$ a b r a c a d a b r a d a d$

10

$ a b r a c a d a b r a d a d$

d$ dad$

18

$ a b r a c a d a b r a d a d$


Suffix array

How to accomplish fast jumping between \( \ell \) LCEs?

Pre-compute

Rank minimum queries

Super cartesian trees
