# Linear-time Computation of Minimal Absent Words Using Suffix Array

Carl Barton[1], Alice Heliou[2],
Laurent Mouchard[3] and Solon P. Pissis[1]

[1]Department of Informatics
King's College London, UK

[2]Laboratoire d'Informatique, LIX
École Polytechnique, France

[3]University of Rouen, LITIS EA 4108, TIBS
Rouen, France

LSD& LAW, 5$^{th}$ February 2015

# Table of Contents

$$\begin{array}{l} \\ y = \end{array} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \mathtt{A} & \mathtt{A} & \mathtt{C} & \mathtt{A} & \mathtt{C} & \mathtt{A} & \mathtt{C} & \mathtt{C} \end{array}$$

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

$$\begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ A & A & C & A & C & A & C & C \end{array}$$
$$y = AACACACC$$

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

## Minimal Absent Word (MAW)

A **word** is **absent** of $y$ if it doesn't occur in $y$.
An **absent word** is **minimal** if all its proper factors occur in $y$.
The number of minimal absent words is **upper bounded** by $\mathcal{O}(\sigma n)$
Crochemore et al. 1998 and Mignosi et al. 2002

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

AAA, AACACC, AACC, CAA, CACACA, CCA, CCC

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

## Minimal Absent Word (MAW)

A **word** is **absent** of $y$ if it doesn't occur in $y$.
An **absent word** is **minimal** if all its proper factors occur in $y$.
The number of minimal absent words is **upper bounded** by $\mathcal{O}(\sigma n)$
Crochemore et al. 1998 and Mignosi et al. 2002

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

AAA, AACACC, AACC, CAA, CACACA, CCA, CCC

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

## Minimal Absent Word (MAW)

A **word** is **absent** of $y$ if it doesn't occur in $y$.
An **absent word** is **minimal** if all its proper factors occur in $y$.
The number of minimal absent words is **upper bounded** by $\mathcal{O}(\sigma n)$
Crochemore et al. 1998 and Mignosi et al. 2002

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

AAA, AACACC, AACC, CAA, CACACA, CCA, CCC

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

## Minimal Absent Word (MAW)

A **word** is **absent** of $y$ if it doesn't occur in $y$.
An **absent word** is **minimal** if all its proper factors occur in $y$.
The number of minimal absent words is **upper bounded** by $\mathcal{O}(\sigma n)$
Crochemore et al. 1998 and Mignosi et al. 2002

$$y = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}$$

AAA, AACACC, AACC, CAA, CACACA, CCA, CCC

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

## Minimal Absent Word (MAW)

A **word** is **absent** of $y$ if it doesn't occur in $y$.
An **absent word** is **minimal** if all its proper factors occur in $y$.
The number of minimal absent words is **upper bounded** by $\mathcal{O}(\sigma n)$
Crochemore et al. 1998 and Mignosi et al. 2002

$$y = \overset{0}{A}\,\overset{1}{A}\,\overset{2}{C}\,\overset{3}{A}\,\overset{4}{C}\,\overset{5}{A}\,\overset{6}{C}\,\overset{7}{C}$$

AAA, AACACC, AACC, CAA, CACACA, CCA, CCC

## Proper Factor

A proper factor is a factor different from the empty word and from the word itself.

## Minimal Absent Word (MAW)

A **word** is **absent** of $y$ if it doesn't occur in $y$.
An **absent word** is **minimal** if all its proper factors occur in $y$.
The number of minimal absent words is **upper bounded** by $\mathcal{O}(\sigma n)$
Crochemore et al. 1998 and Mignosi et al. 2002

# Context

## Biology

- Nullomers: Really a matter of natural selection? [Acquisti et al.], 2007
- Minimal Absent Words in Four Human Genome Assemblies, [Garcia et al.], 2011

## Computer Science

- Data Compression Using Antidictionaries, [Crochemore et al.], 2000
- DCA using Suffix Arrays, [Fiala and Holub], 2008

# State of the Art

| References | Time<br>for fixed size alphabet | Space | Structure |
|---|---|---|---|
| Crochemore et al. 1998<br>Automata and forbidden words | $\mathcal{O}(n)$ | $\mathcal{O}(n)$ | suffix automata |
| Pinho et al. 2009<br>On finding minimal absent words | $\mathcal{O}(n^2)$ | $\mathcal{O}(n)$ | suffix array |
| Belazzougui et al. 2013<br>Versatile Succinct Representations of the Bidirectional Burrows-Wheeler Transform. | $\mathcal{O}(n+output)$ | $\mathcal{O}(n)$ | compact bidirectional BWT |

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$\overset{0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8}{T = \mathrm{A\,A\,C\,A\,C\,A\,C\,C\,\#}}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 2 | C | A | C | A | C | C | # | A | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 4 | C | A | C | C | # | A | A | C | A |
| 5 | A | C | C | # | A | A | C | A | C |
| 6 | C | C | # | A | A | C | A | C | A |
| 7 | C | # | A | A | C | A | C | A | C |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$\overset{0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8}{T = \mathrm{A\,A\,C\,A\,C\,A\,C\,C\,\#}}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 2 | C | A | C | A | C | C | # | A | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 4 | C | A | C | C | # | A | A | C | A |
| 5 | A | C | C | # | A | A | C | A | C |
| 6 | C | C | # | A | A | C | A | C | A |
| 7 | C | # | A | A | C | A | C | A | C |
| **8** | # | **A** | **A** | **C** | **A** | **C** | **A** | **C** | **C** |

Rotations of $T$

$\Rightarrow$

| *pos* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | # | A | A | C | A | C | A | C | C |

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$\overset{0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8}{T = \mathrm{A\,A\,C\,A\,C\,A\,C\,C\,\#}}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **0** | **A** | **A** | **C** | **A** | **C** | **A** | **C** | **C** | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 2 | C | A | C | A | C | C | # | A | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 4 | C | A | C | C | # | A | A | C | A |
| 5 | A | C | C | # | A | A | C | A | C |
| 6 | C | C | # | A | A | C | A | C | A |
| 7 | C | # | A | A | C | A | C | A | C |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

$\Rightarrow$

| *pos* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | # | A | A | C | A | C | A | C | C |
| 0 | A | A | C | A | C | A | C | C | # |

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$\overset{0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8}{T = \mathrm{A\,A\,C\,A\,C\,A\,C\,C\,\#}}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| **1** | **A** | **C** | **A** | **C** | **A** | **C** | **C** | # | **A** |
| 2 | C | A | C | A | C | C | # | A | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 4 | C | A | C | C | # | A | A | C | A |
| 5 | A | C | C | # | A | A | C | A | C |
| 6 | C | C | # | A | A | C | A | C | A |
| 7 | C | # | A | A | C | A | C | A | C |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

$\Rightarrow$

| *pos* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | # | A | A | C | A | C | A | C | C |
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$\overset{0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8}{T = A\,A\,C\,A\,C\,A\,C\,C\,\#}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 2 | C | A | C | A | C | C | # | A | A |
| **3** | **A** | **C** | **A** | **C** | **C** | # | **A** | **A** | **C** |
| 4 | C | A | C | C | # | A | A | C | A |
| 5 | A | C | C | # | A | A | C | A | C |
| 6 | C | C | # | A | A | C | A | C | A |
| 7 | C | # | A | A | C | A | C | A | C |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

$\Rightarrow$

| *pos* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | # | A | A | C | A | C | A | C | C |
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 3 | A | C | A | C | C | # | A | A | C |

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$T = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}\overset{8}{\#}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 2 | C | A | C | A | C | C | # | A | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 4 | C | A | C | C | # | A | A | C | A |
| **5** | **A** | **C** | **C** | # | **A** | **A** | **C** | **A** | **C** |
| 6 | C | C | # | A | A | C | A | C | A |
| 7 | C | # | A | A | C | A | C | A | C |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

$\Rightarrow$

| *pos* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | # | A | A | C | A | C | A | C | C |
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 5 | A | C | C | # | A | A | C | A | C |

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$T = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}\overset{8}{\#}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 2 | C | A | C | A | C | C | # | A | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 4 | C | A | C | C | # | A | A | C | A |
| 5 | A | C | C | # | A | A | C | A | C |
| 6 | C | C | # | A | A | C | A | C | A |
| **7** | **C** | **#** | **A** | **A** | **C** | **A** | **C** | **A** | **C** |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

$\Rightarrow$

| *pos* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | # | A | A | C | A | C | A | C | C |
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 5 | A | C | C | # | A | A | C | A | C |
| 7 | C | # | A | A | C | A | C | A | C |

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$T = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}\overset{8}{\#}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| **2** | **C** | **A** | **C** | **A** | **C** | **C** | # | **A** | **A** |
| 3 | A | C | A | C | C | # | A | A | C |
| 4 | C | A | C | C | # | A | A | C | A |
| 5 | A | C | C | # | A | A | C | A | C |
| 6 | C | C | # | A | A | C | A | C | A |
| 7 | C | # | A | A | C | A | C | A | C |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

$\Rightarrow$

| *pos* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | # | A | A | C | A | C | A | C | C |
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 5 | A | C | C | # | A | A | C | A | C |
| 7 | C | # | A | A | C | A | C | A | C |
| 2 | C | A | C | A | C | C | # | A | A |

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$T = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}\overset{8}{\#}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 2 | C | A | C | A | C | C | # | A | A |
| 3 | A | C | A | C | C | # | A | A | C |
| **4** | **C** | **A** | **C** | **C** | # | **A** | **A** | **C** | **A** |
| 5 | A | C | C | # | A | A | C | A | C |
| 6 | C | C | # | A | A | C | A | C | A |
| 7 | C | # | A | A | C | A | C | A | C |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

$\Rightarrow$

| *pos* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | # | A | A | C | A | C | A | C | C |
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 5 | A | C | C | # | A | A | C | A | C |
| 7 | C | # | A | A | C | A | C | A | C |
| 2 | C | A | C | A | C | C | # | A | A |
| 4 | C | A | C | C | # | A | A | C | A |

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$\overset{0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8}{T = A\,A\,C\,A\,C\,A\,C\,C\,\#}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 2 | C | A | C | A | C | C | # | A | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 4 | C | A | C | C | # | A | A | C | A |
| 5 | A | C | C | # | A | A | C | A | C |
| **6** | **C** | **C** | **#** | **A** | **A** | **C** | **A** | **C** | **A** |
| 7 | C | # | A | A | C | A | C | A | C |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

$\Rightarrow$

| *pos* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | # | A | A | C | A | C | A | C | C |
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 5 | A | C | C | # | A | A | C | A | C |
| 7 | C | # | A | A | C | A | C | A | C |
| 2 | C | A | C | A | C | C | # | A | A |
| 4 | C | A | C | C | # | A | A | C | A |
| 6 | C | C | # | A | A | C | A | C | A |

Ordered rotations of $T$

## Suffix Array Manber& Myers 1990 and Burrows-Wheeler Transform 1994

**Suffix Array**: Index allowing fast localisation of patterns.
**BWT**: Reversible permutation used in compression and indexing.
**LCP**: Longest Common Prefix between two rows of the suffix array.

$$\overset{0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8}{T = \mathrm{A\,A\,C\,A\,C\,A\,C\,C\,\#}}$$

$$BWT(T) = \mathrm{C\#ACCCAAA}, \text{ and } SA(T) = [8,0,1,3,5,7,2,4,6]$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A | A | C | A | C | A | C | C | # |
| 1 | A | C | A | C | A | C | C | # | A |
| 2 | C | A | C | A | C | C | # | A | A |
| 3 | A | C | A | C | C | # | A | A | C |
| 4 | C | A | C | C | # | A | A | C | A |
| 5 | A | C | C | # | A | A | C | A | C |
| 6 | C | C | # | A | A | C | A | C | A |
| 7 | C | # | A | A | C | A | C | A | C |
| 8 | # | A | A | C | A | C | A | C | C |

Rotations of $T$

$\Rightarrow$

| *LCP* | *SA* | | | | | | | | | *BWT* |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 8 | # | A | A | C | A | C | A | C | C |
| 0 | 0 | A | A | C | A | C | A | C | C | # |
| 1 | 1 | A | C | A | C | A | C | C | # | A |
| 4 | 3 | A | C | A | C | C | # | A | A | C |
| 2 | 5 | A | C | C | # | A | A | C | A | C |
| 0 | 7 | C | # | A | A | C | A | C | A | C |
| 1 | 2 | C | A | C | A | C | C | # | A | A |
| 3 | 4 | C | A | C | C | # | A | A | C | A |
| 1 | 6 | C | C | # | A | A | C | A | C | A |

Ordered rotations of $T$

# Table of Contents

## Problem

**Input**: A word $y$ of length $n$ on $\Sigma$ a fixed size alphabet ($\sigma = \mathcal{O}(1)$)
**Output**: Set of minimal absent words of $y$

## Our contribution

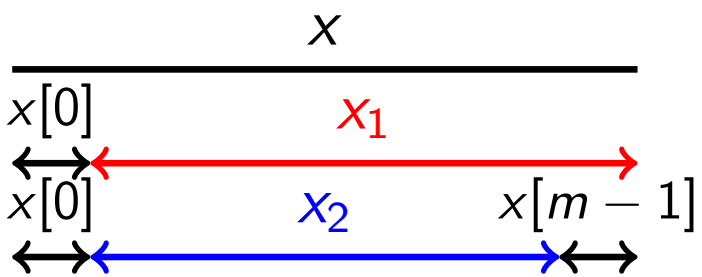
Algorithm linear in time and space based on the suffix array structure

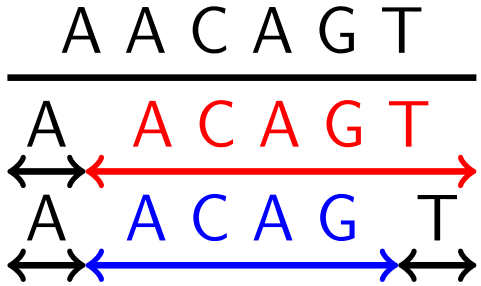
## Remark

$x = x[0\,..\,m-1]$ is a MAW of $y$ **if and only if** $x[1\,..\,m-1]$ and $x[0\,..\,m-2]$ are factors of $y$ and $x$ is not a factor of $y$

## Remark

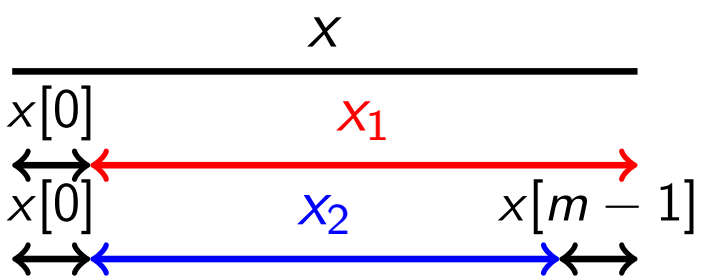
$x = x[0\,..\,m-1]$ is a MAW of $y$ **if and only if** $x[1\,..\,m-1]$ and $x[0\,..\,m-2]$ are factors of $y$ and $x$ is not a factor of $y$

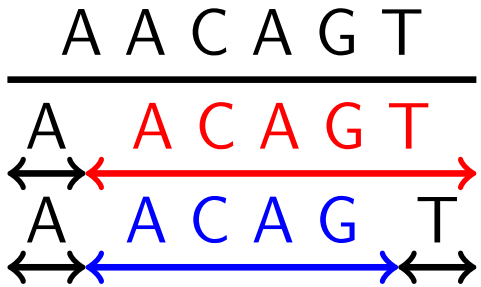


## Remark

$x = x[0\,..\,m-1]$ is a MAW of $y$ **if and only if** $x[1\,..\,m-1]$ and $x[0\,..\,m-2]$ are factors of $y$ and $x$ is not a factor of $y$

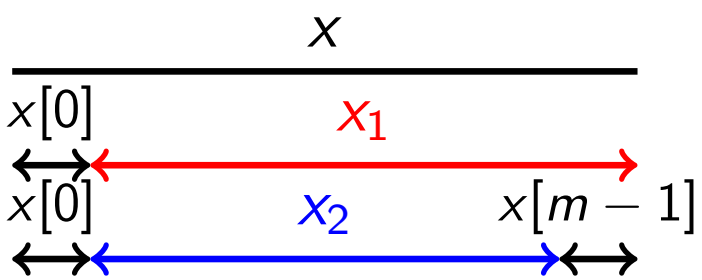


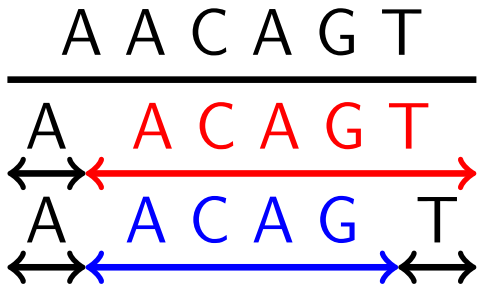




B($x_1$)={ y[j-1] : $j$ starting position of an occurrence of $x_1$ in $y$}
B($x_2$)={ y[j-1] : $j$ starting position of an occurrence of $x_2$ in $y$}

## Remark

$x = x[0\,..\,m-1]$ is a MAW of $y$ **if and only if** $x[1\,..\,m-1]$ and $x[0\,..\,m-2]$ are factors of $y$ and $x$ is not a factor of $y$

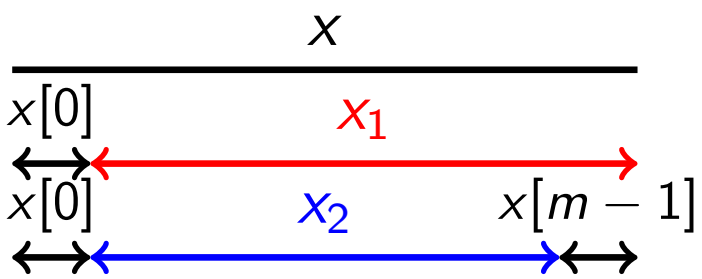


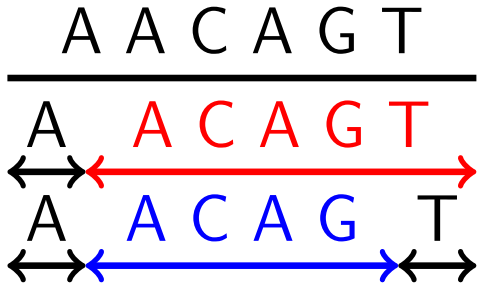




B($x_1$)={ y[j-1] : $j$ starting position of an occurrence of $x_1$ in $y$}
B($x_2$)={ y[j-1] : $j$ starting position of an occurrence of $x_2$ in $y$}

## Lemma 1

$x$ is a MAW of $y$ if and only if $x[0] \in \mathrm{B}(x_2)$ and $x[0] \notin \mathrm{B}(x_1)$ with $x_1 = x[1\,..\,m-1]$ and $x_2 = x[1\,..\,m-2]$.



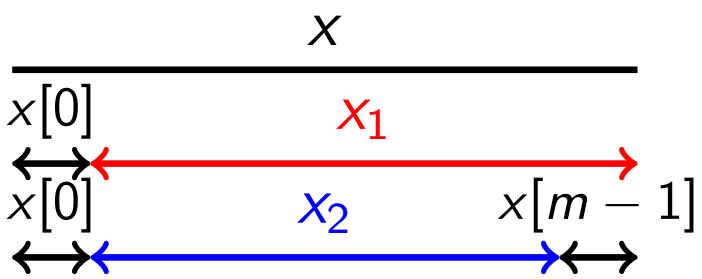

## Lemma 1

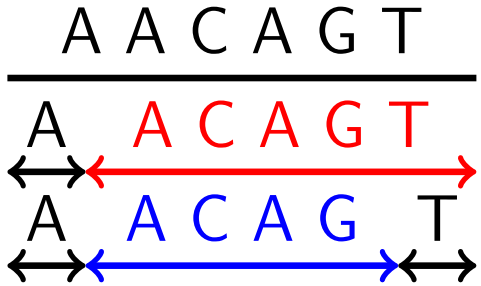
$x$ is a MAW of $y$ if and only if: $x[0] \in \mathrm{B}(x_2)$ and $x[0] \notin \mathrm{B}(x_1)$ with $x_1 = x[1\,..\,m-1]$ and $x_2 = x[1\,..\,m-2]$.

$\Rightarrow$ Let $x$ be a MAW of $y$

$x[0]x_2$ is a factor of $y$, consequently $x[0] \in \mathrm{B}(x_2)$.

$x$ is not a factor of $y$ so $x[0] \notin \mathrm{B}(x_1)$.




## Lemma 1

$x$ is a MAW of $y$ if and only if: $x[0] \in \mathrm{B}(x_2)$ and $x[0] \notin \mathrm{B}(x_1)$ with $x_1 = x[1\,..\,m-1]$ and $x_2 = x[1\,..\,m-2]$.

$\Rightarrow$ Let $x$ be a MAW of $y$
$x[0]x_2$ is a factor of $y$, consequently $x[0] \in \mathrm{B}(x_2)$.
$x$ is not a factor of $y$ so $x[0] \notin \mathrm{B}(x_1)$.

$\Leftarrow$ Let $x_1$ be a factor of $y$
$x_2$ its longest prefix and $p \in \mathrm{B}(x_2) \backslash \mathrm{B}(x_1)$.
Then $px_2$ is a factor of $y$ and $px_1$ is not so $px_1$ is a MAW of $y$. □

## Lemma 2

If $x$ is a minimal absent word of $y$ then:
There is $i \in [0;n\text{-}1]$ such that $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i] + \mathrm{LCP}[i]] = x_1$ or $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i] + \mathrm{LCP}[i+1]] = x_1$, with $x_1 = x[1\,..\,m-1]$

## Lemma 2

If $x$ is a minimal absent word of $y$ then:
There is $i \in [0;n\text{-}1]$ such that $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i]] = x_1$ or $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i+1]] = x_1$, with $x_1 = x[1\,..\,m-1]$

Let $j$ be the starting position of an occurrence of $x[0\,..\,m\text{-}2]$ in $y$ and $k$ the starting position of an occurrence of $x_1$

## Lemma 2

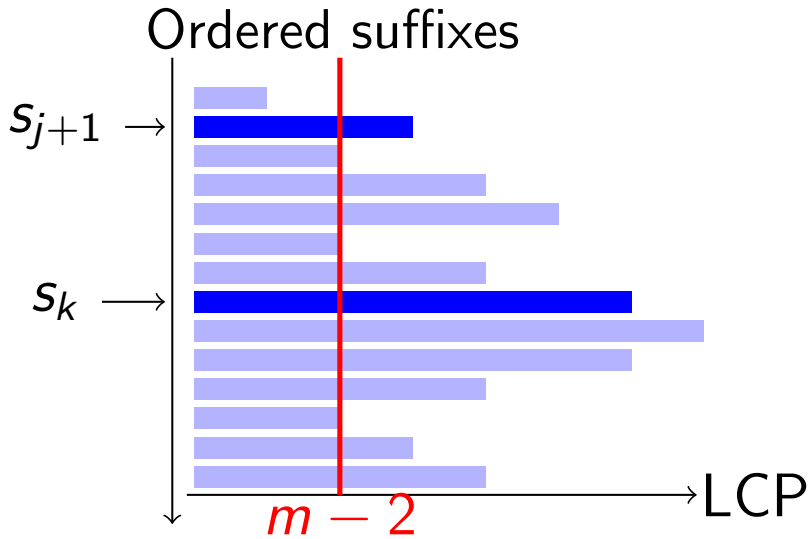
If $x$ is a minimal absent word of $y$ then:
There is $i \in [0;n\text{-}1]$ such that $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i]] = x_1$ or $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i+1]] = x_1$, with $x_1 = x[1 \,..\, m-1]$

Let $j$ be the starting position of an occurrence of $x[0 \,..\, m\text{-}2]$ in $y$ and $k$ the starting position of an occurrence of $x_1$
$y[j+1 \,..\, n\text{-}1]$ and $y[k \,..\, n\text{-}1]$ share $x_2 = x[1 \,..\, m\text{-}2]$ as a prefix.

## Lemma 2

If $x$ is a minimal absent word of $y$ then:
There is $i \in [0;n\text{-}1]$ such that $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i]] = x_1$ or $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i+1]] = x_1$, with $x_1 = x[1 \,..\, m-1]$

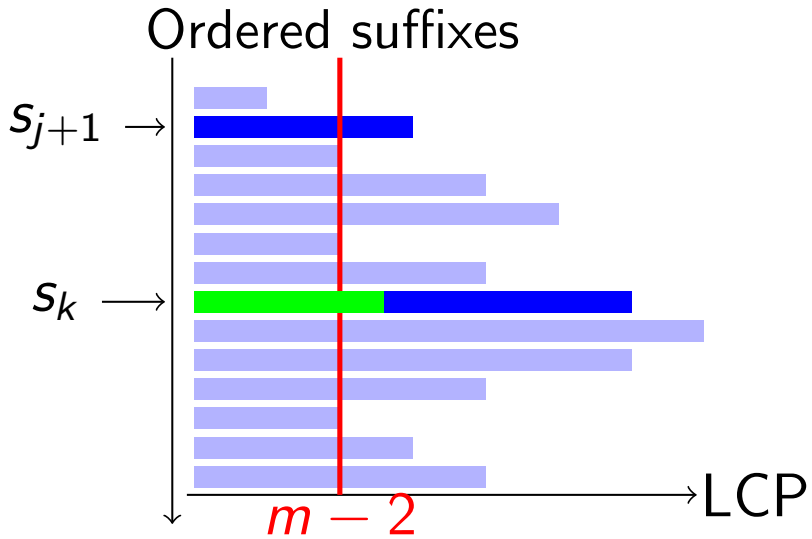
Let $j$ be the starting position of an occurrence of $x[0 \,..\, m\text{-}2]$ in $y$ and $k$ the starting position of an occurrence of $x_1$
$y[j+1 \,..\, n\text{-}1]$ and $y[k \,..\, n\text{-}1]$ share $x_2 = x[1 \,..\, m\text{-}2]$ as a prefix.



## Lemma 2

If $x$ is a minimal absent word of $y$ then:
There is $i \in [0;n\text{-}1]$ such that $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i]] = x_1$ or $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i+1]] = x_1$, with $x_1 = x[1 \,..\, m-1]$

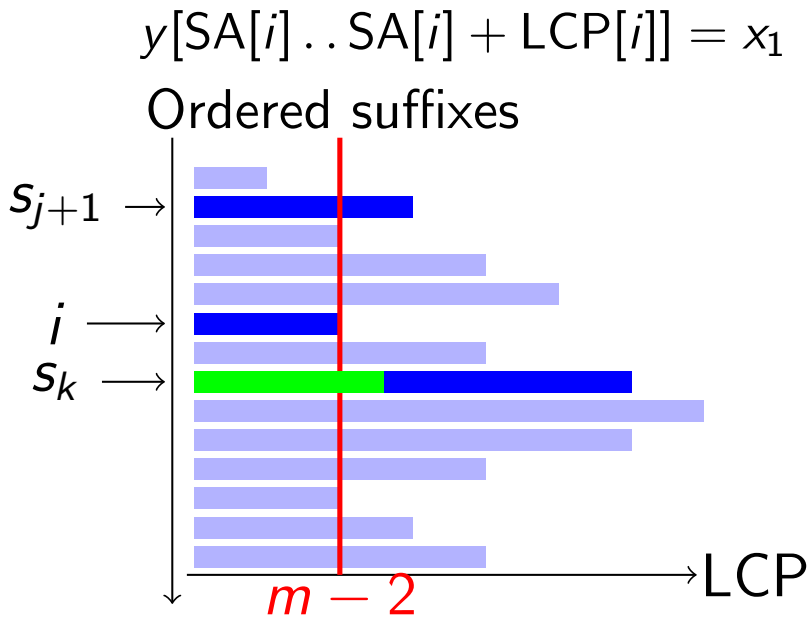
Let $j$ be the starting position of an occurrence of $x[0 \,..\, m\text{-}2]$ in $y$ and $k$ the starting position of an occurrence of $x_1$
$y[j+1 \,..\, n\text{-}1]$ and $y[k \,..\, n\text{-}1]$ share $x_2 = x[1 \,..\, m\text{-}2]$ as a prefix.



## Lemma 2

If $x$ is a minimal absent word of $y$ then:
There is $i \in [0;n\text{-}1]$ such that $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i]] = x_1$ or $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i+1]] = x_1$, with $x_1 = x[1 \,..\, m-1]$

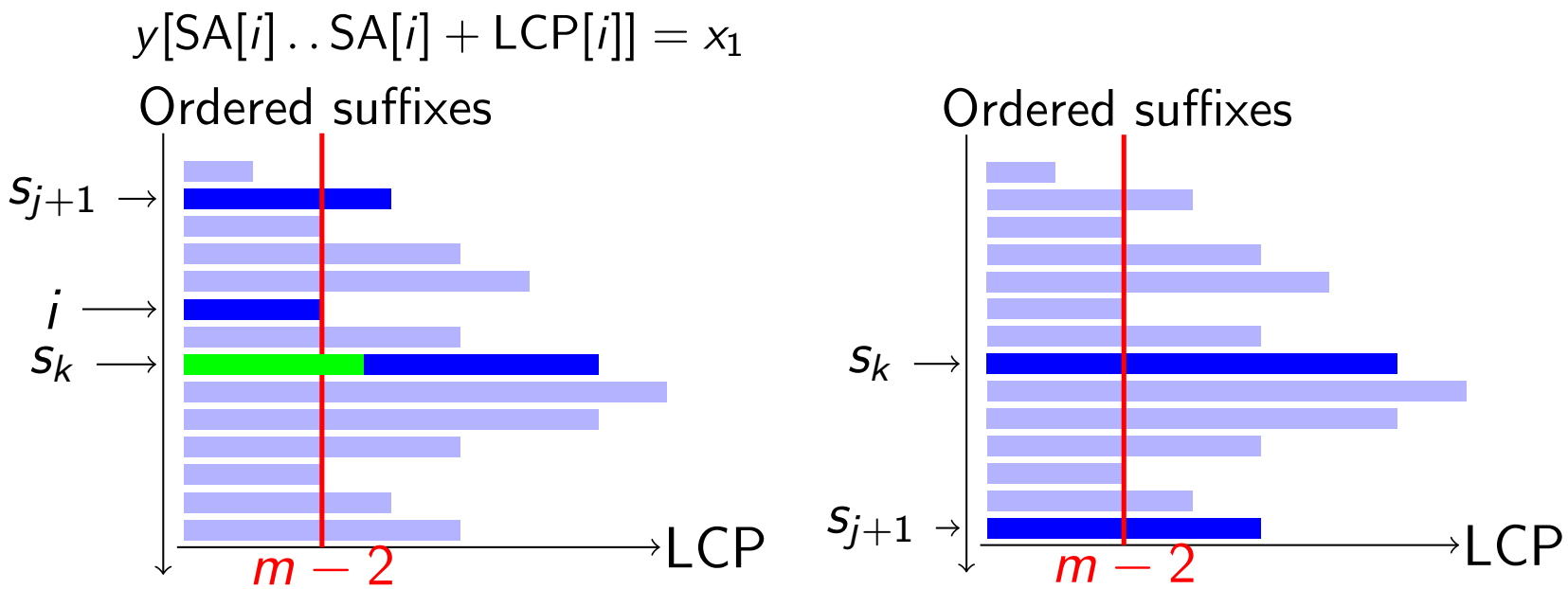
Let $j$ be the starting position of an occurrence of $x[0 \,..\, m\text{-}2]$ in $y$ and $k$ the starting position of an occurrence of $x_1$
$y[j+1 \,..\, n\text{-}1]$ and $y[k \,..\, n\text{-}1]$ share $x_2 = x[1 \,..\, m\text{-}2]$ as a prefix.



## Lemma 2

If $x$ is a minimal absent word of $y$ then:
There is $i \in [0;n\text{-}1]$ such that $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i]] = x_1$ or $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i+1]] = x_1$, with $x_1 = x[1 \,..\, m-1]$

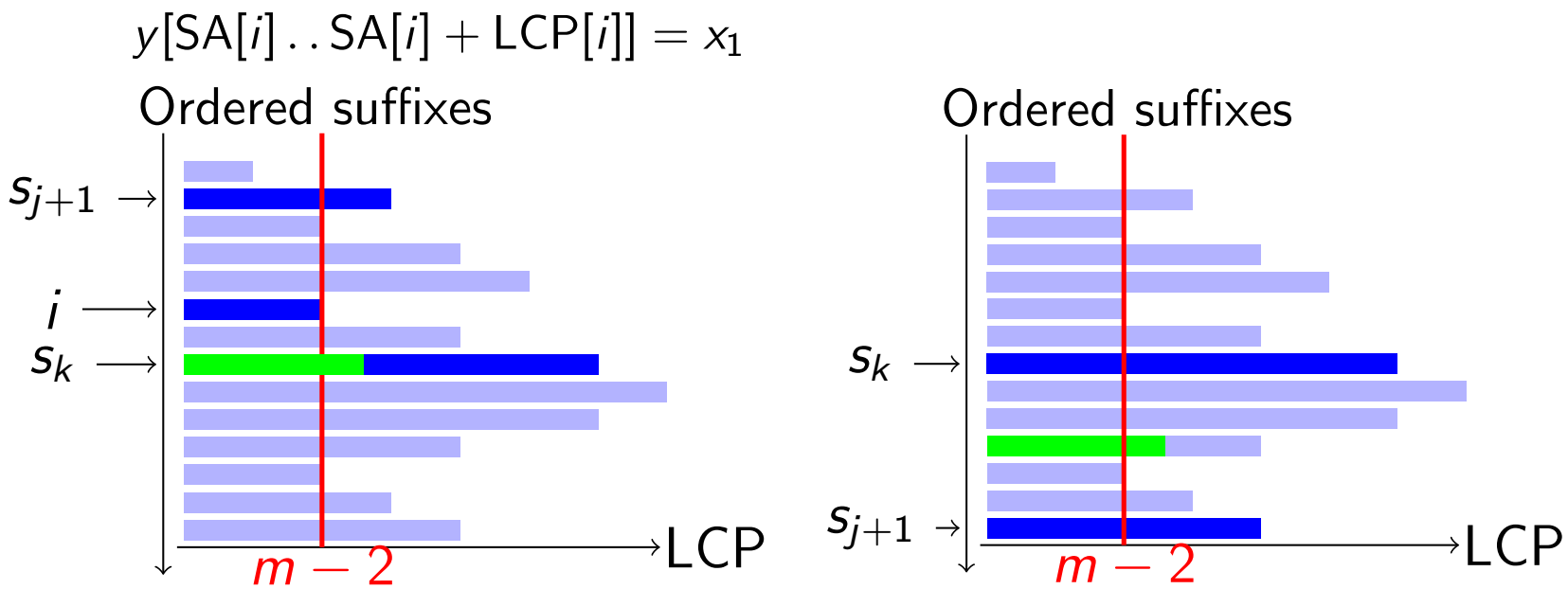
Let $j$ be the starting position of an occurrence of $x[0 \,..\, m\text{-}2]$ in $y$ and $k$ the starting position of an occurrence of $x_1$
$y[j+1 \,..\, n\text{-}1]$ and $y[k \,..\, n\text{-}1]$ share $x_2 = x[1 \,..\, m\text{-}2]$ as a prefix.



## Lemma 2

If $x$ is a minimal absent word of $y$ then:
There is $i \in [0;n\text{-}1]$ such that $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i]] = x_1$ or $y[\mathrm{SA}[i] \,..\, \mathrm{SA}[i] + \mathrm{LCP}[i+1]] = x_1$, with $x_1 = x[1 \,..\, m-1]$

Let $j$ be the starting position of an occurrence of $x[0 \,..\, m\text{-}2]$ in $y$ and $k$ the starting position of an occurrence of $x_1$
$y[j+1 \,..\, n\text{-}1]$ and $y[k \,..\, n\text{-}1]$ share $x_2 = x[1 \,..\, m\text{-}2]$ as a prefix.

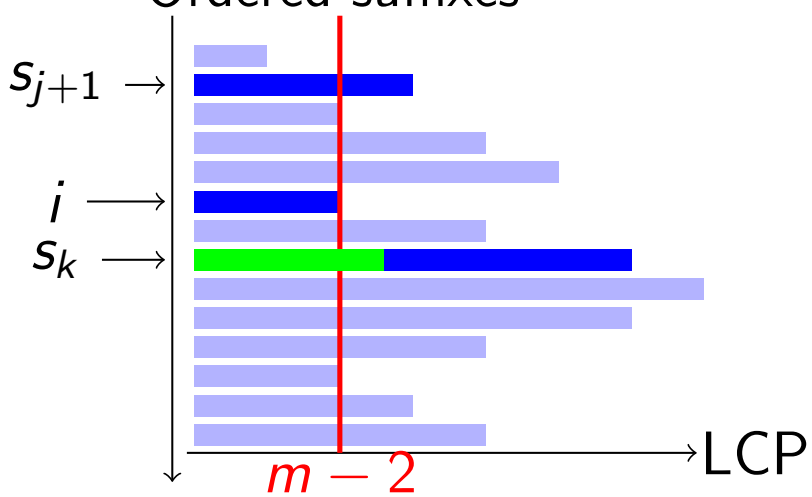


## Lemma 2

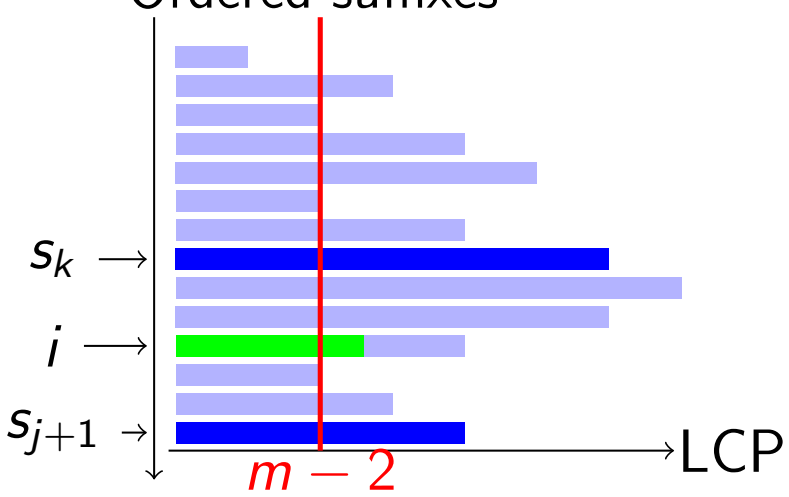
If $x$ is a minimal absent word of $y$ then:
There is $i \in [0;n\text{-}1]$ such that $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i]] = x_1$ or $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i+1]] = x_1$, with $x_1 = x[1\,..\,m-1]$

Let $j$ be the starting position of an occurrence of $x[0\,..\,m\text{-}2]$ in $y$ and $k$ the starting position of an occurrence of $x_1$
$y[j+1\,..\,n\text{-}1]$ and $y[k\,..\,n\text{-}1]$ share $x_2 = x[1\,..\,m\text{-}2]$ as a prefix.

$y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i]] = x_1$




$y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i+1]] = x_1$



Lemma 2 $\Rightarrow$ we need to consider only $2n$ factors $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i]]$ and $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i+1]]$

Lemma 2 $\Rightarrow$ we need to consider only $2n$ factors $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i]]$ and $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i+1]]$

Lemma 1 $\Rightarrow$ for each one we have to compute the set of letters that precede them, and the set that precede their longest prefix

Lemma 2 $\Rightarrow$ we need to consider only $2n$ factors $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i]]$ and $y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i+1]]$

Lemma 1 $\Rightarrow$ for each one we have to compute the set of letters that precede them, and the set that precede their longest prefix

We use:

- the Suffix Array;
- the LCP table;
- two tables (of size 2n) to store those sets of letters;
- an additional table: Interval.
  Interval[$\ell$] contains the set of letters that occur before an occurrence of the ongoing $\ell$-prefix.

# Pre-computation

Lemma 2 $\Rightarrow$ we need to look at only $2n$ factors of $y$

$y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i]]$ and

$y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i+1]]$

| $i$ | LCP | SA | Suffixes |
|---|---|---|---|
| 0 | 0 | 0 | A A C A C A C C |
| 1 | 1 | 1 | A C A C A C C |
| 2 | 4 | 3 | A C A C C |
| 3 | 2 | 5 | A C C |
| 4 | 0 | 7 | C |
| 5 | 1 | 2 | C A C A C C |
| 6 | 3 | 4 | C A C C |
| 7 | 1 | 6 | C C |

Table: LCP table, Suffix Array and suffixes for y=A A C A C A C C

# Pre-computation

Lemma 2 $\Rightarrow$ we need to look at only $2n$ factors of $y$

$y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i]]$ and

$y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i+1]]$

| $i$ | LCP | SA | Suffixes |
|---|---|---|---|
| 0 | 0 | 0 | A A C A C A C C |
| | | | A A C A C A C C |
| 1 | 1 | 1 | A C A C A C C |
| | | | A C A C A C C |
| 2 | 4 | 3 | A C A C C |
| | | | A C A C C |
| 3 | 2 | 5 | A C C |
| | | | A C C |
| 4 | 0 | 7 | C |
| | | | C |
| 5 | 1 | 2 | C A C A C C |
| | | | C A C A C C |
| 6 | 3 | 4 | C A C C |
| | | | C A C C |
| 7 | 1 | 6 | C C |

Table: LCP table, Suffix Array and suffixes for y=$\overset{0}{\mathrm{A}}\,\overset{1}{\mathrm{A}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{A}}\,\overset{4}{\mathrm{C}}\,\overset{5}{\mathrm{A}}\,\overset{6}{\mathrm{C}}\,\overset{7}{\mathrm{C}}$

# Pre-computation

Lemma 2 $\Rightarrow$ we need to look at only $2n$ factors of $y$

$y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i]]$ and

$y[\mathrm{SA}[i]\,..\,\mathrm{SA}[i]+\mathrm{LCP}[i+1]]$

| $i$ | LCP | SA | Suffixes |
|---|---|---|---|
| 0 | 0 | 0 | A A C A C A C C |
| | | | A A C A C A C C |
| 1 | 1 | 1 | A C A C A C C |
| | | | A C A C A C C |
| 2 | 4 | 3 | A C A C C |
| | | | A C A C C |
| 3 | 2 | 5 | A C C |
| | | | A C C |
| 4 | 0 | 7 | C |
| | | | C |
| 5 | 1 | 2 | C A C A C C |
| | | | C A C A C C |
| 6 | 3 | 4 | C A C C |
| | | | C A C C |
| 7 | 1 | 6 | C C |

Table: LCP table, Suffix Array and suffixes for y=A A C A C A C C (positions 0 1 2 3 4 5 6 7)

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | | | | | |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | | | | | |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | | | | | |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | | | | | |
| | | | A | | | | | |
| 4 | 0 | 7 | C | | | | | |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | | | | | |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | | | | | |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | | | | | |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | | |
| 1 | | |
| 2 | | |
| 3 | | |
| 4 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |
| 11 | | |
| 12 | | |
| 13 | | |
| 14 | | |

Table: Top-down pass for the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | ∅ | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | | | | | |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | | | | | |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | | | | | |
| | | | A | | | | | |
| 4 | 0 | 7 | C | | | | | |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | | | | | |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | | | | | |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | | | | | |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | | |
| 3 | | |
| 4 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |
| 11 | | |
| 12 | | |
| 13 | | |
| 14 | | |

Table: Top-down pass for the word $y = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | ∅ | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | A | A | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | | | | | |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | | | | | |
| | | | A | | | | | |
| 4 | 0 | 7 | C | | | | | |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | | | | | |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | | | | | |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | | | | | |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |
| 11 | | |
| 12 | | |
| 13 | | |
| 14 | | |

Table: Top-down pass for the word $y = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | ∅ | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | A | A | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | AC | ∅ | ∅ | AC |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | | | | | |
| | | | A | | | | | |
| 4 | 0 | 7 | C | | | | | |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | | | | | |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | | | | | |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | | | | | |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |
| 11 | | |
| 12 | | |
| 13 | | |
| 14 | | |

Table: Top-down pass for the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | ∅ | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | A | A | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | AC | ∅ | ∅ | AC |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | AC | AC | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | | | | | |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | | | | | |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | | | | | |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | | | | | |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | | |
| 9 | | |
| 10 | | |
| 11 | | |
| 12 | | |
| 13 | | |
| 14 | | |

Table: Top-down pass for the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | ∅ | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | A | A | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | AC | ∅ | ∅ | AC |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | AC | AC | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | | | | | |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | | | | | |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | | | | | |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | C | AC |
| 9 | C | C |
| 10 | | |
| 11 | | |
| 12 | | |
| 13 | | |
| 14 | | |

Table: Top-down pass for the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | ∅ | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | A | A | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | AC | ∅ | ∅ | AC |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | AC | AC | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | AC | AC | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | | | | | |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | | | | | |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | C | AC |
| 9 | C | C |
| 10 | A | AC |
| 11 | A | A |
| 12 | | |
| 13 | | |
| 14 | | |

Table: Top-down pass for the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | ∅ | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | A | A | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | AC | ∅ | ∅ | AC |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | AC | AC | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | AC | AC | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | AC | AC | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | | | | | |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | C | AC |
| 9 | C | C |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | | |

Table: Top-down pass for the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | ∅ | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | A | A | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | AC | ∅ | ∅ | AC |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | AC | AC | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | AC | AC | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | AC | AC | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | AC | AC | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | C | AC |
| 9 | C | C |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Top-down pass for the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The top-down pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | ∅ | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | A | A | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | AC | ∅ | ∅ | AC |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | AC | AC | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | AC | AC | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | AC | AC | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | AC | AC | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | C | AC |
| 9 | C | C |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Top-down pass for the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step. Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | | | | | |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | | | | | |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | | | | | |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | | | | | |
| | | | A | | | | | |
| 4 | 0 | 7 | C | | | | | |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | | | | | |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | | | | | |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | | | | | |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | C | AC |
| 9 | C | C |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y = \overset{0}{\mathrm{A}}\,\overset{1}{\mathrm{A}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{A}}\,\overset{4}{\mathrm{C}}\,\overset{5}{\mathrm{A}}\,\overset{6}{\mathrm{C}}\,\overset{7}{\mathrm{C}}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | | | | | |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | | | | | |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | | | | | |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | | | | | |
| | | | A | | | | | |
| 4 | 0 | 7 | C | | | | | |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | | | | | |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | | | | | |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | A | A | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | C | AC |
| 9 | C | C |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | | | | | |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | | | | | |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | | | | | |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | | | | | |
| | | | A | | | | | |
| 4 | 0 | 7 | C | | | | | |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | | | | | |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | A | A | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | A | A | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | C | AC |
| 9 | C | C |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | | | | | |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | | | | | |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | | | | | |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | | | | | |
| | | | A | | | | | |
| 4 | 0 | 7 | C | | | | | |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | A | A | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | A | A | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | A | A | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | C | AC |
| 9 | C | C |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | | | | | |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | | | | | |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | | | | | |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | | | | | |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | A | A | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | A | A | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | A | A | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | AC | AC |
| 9 | C | AC |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | | | | | |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | | | | | |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | | | | | |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | ∅ | C | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | A | A | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | A | A | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | A | A | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | AC | AC |
| 9 | C | AC |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | | | | | |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | | | | | |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | ∅ | C | ∅ | C |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | ∅ | C | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | A | A | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | A | A | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | A | A | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | A | A |
| 3 | A | A |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | AC | AC |
| 9 | C | AC |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y = \overset{0}{\mathrm{A}}\overset{1}{\mathrm{A}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{A}}\overset{4}{\mathrm{C}}\overset{5}{\mathrm{A}}\overset{6}{\mathrm{C}}\overset{7}{\mathrm{C}}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | | | | | |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | AC | AC | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | ∅ | C | ∅ | C |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | ∅ | C | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | A | A | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | A | A | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | A | A | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | ∅ | ∅ |
| 1 | ∅ | ∅ |
| 2 | AC | AC |
| 3 | A | AC |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | AC | AC |
| 9 | C | AC |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y =$ $\overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | AC | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | AC | AC | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | ∅ | C | ∅ | C |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | ∅ | C | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | A | A | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | A | A | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | A | A | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | AC | AC |
| 1 | ∅ | AC |
| 2 | AC | AC |
| 3 | A | AC |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | AC | AC |
| 9 | C | AC |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

# The bottom-up pass

| $i$ | LCP | SA | Factor | Intval[0] | Intval[1] | Intval[2] | Intval[3] | Intval[4] |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | A | AC | ∅ | ∅ | ∅ | ∅ |
| | | | A A | | | | | |
| 1 | 1 | 1 | A C | AC | AC | ∅ | ∅ | ∅ |
| | | | A C A C A | | | | | |
| 2 | 4 | 3 | A C A C C | AC | ∅ | C | ∅ | C |
| | | | A C A | | | | | |
| 3 | 2 | 5 | A C C | AC | ∅ | C | ∅ | ∅ |
| | | | A | | | | | |
| 4 | 0 | 7 | C | AC | ∅ | ∅ | ∅ | ∅ |
| | | | C | | | | | |
| 5 | 1 | 2 | C A | A | A | ∅ | ∅ | ∅ |
| | | | C A C A | | | | | |
| 6 | 3 | 4 | C A C C | A | A | ∅ | A | ∅ |
| | | | C A | | | | | |
| 7 | 1 | 6 | C C | A | A | ∅ | ∅ | ∅ |

| $j$ | $B(x_1)$ | $B(x_2)$ |
|---|---|---|
| 0 | AC | AC |
| 1 | ∅ | AC |
| 2 | AC | AC |
| 3 | A | AC |
| 4 | C | AC |
| 5 | AC | AC |
| 6 | C | AC |
| 7 | AC | AC |
| 8 | AC | AC |
| 9 | C | AC |
| 10 | A | AC |
| 11 | A | A |
| 12 | A | A |
| 13 | A | AC |
| 14 | A | AC |

Table: Bottom-up of the word $y = \overset{0}{A}\overset{1}{A}\overset{2}{C}\overset{3}{A}\overset{4}{C}\overset{5}{A}\overset{6}{C}\overset{7}{C}$ step by step.
Factors $x_2$ are in orange and factors $x_1$ are in orange and red.

| $j$ | $B(x_1)$ | $B(x_2)$ | Factors | Minimal Absent Words | Output |
|---|---|---|---|---|---|
| 0 | AC | AC | A | | |
| 1 | $\varnothing$ | AC | A A | AAA, CAA | $<$A$,(0,1)>$, $<$C$,(0,1)>$ |
| 2 | AC | AC | A C | | |
| 3 | A | AC | A C A C A | CACACA | $<$C$,(1,5)>$ |
| 4 | C | AC | A C A C C | AACACC | $<$A$,(3,7)>$ |
| 5 | AC | AC | A C A | | |
| 6 | C | AC | A C C | AACC | $<$A$,(5,7)>$ |
| 7 | AC | AC | A | | |
| 8 | AC | AC | C | | |
| 9 | C | AC | C | We do not consider this row as it corresponds to the end of $y$ | |
| 10 | A | AC | C A | CCA | $<$C$,(2,3)>$ |
| 11 | A | A | C A C A | | |
| 12 | A | A | C A C C | | |
| 13 | A | AC | C A | CCA | This is a duplicate so we ignore it |
| 14 | A | AC | C C | CCC | $<$C$,(6,7)>$ |

Table: $y = \overset{0}{\mathrm{A}}\,\overset{1}{\mathrm{A}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{A}}\,\overset{4}{\mathrm{C}}\,\overset{5}{\mathrm{A}}\,\overset{6}{\mathrm{C}}\,\overset{7}{\mathrm{C}}$;

We find 7 Minimal Absent Words

{AAA, AACACC, AACC, CAA, CACACA, CCA, CCC}

# Table of Contents

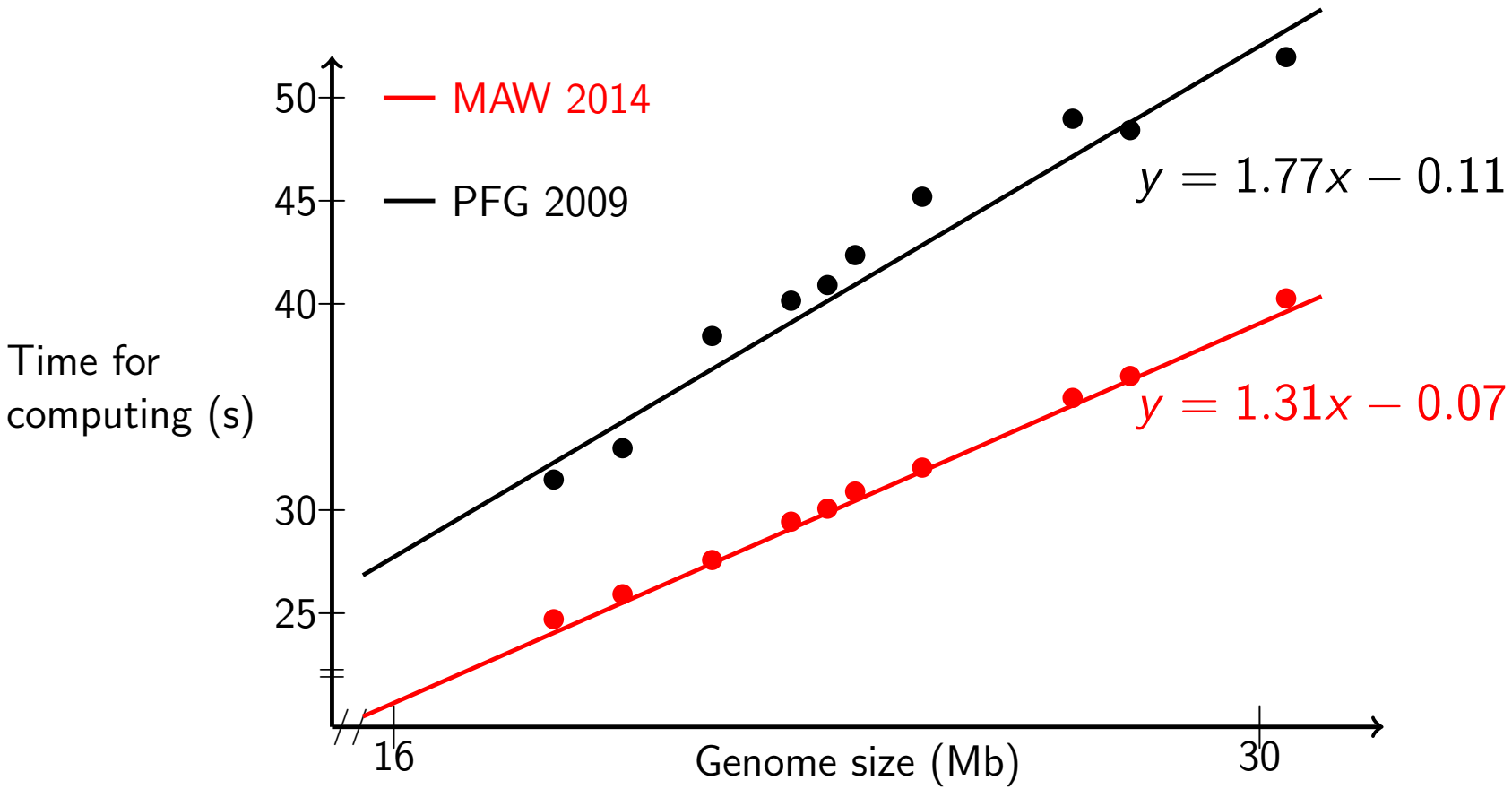
# Correctness

| Species | Genome size (bp) | $M_{11}$ | $M_{14}$ | $M_{17}$ | $M_{24}$ |
|---|---|---|---|---|---|
| Ba | 5,227,293 | 1,113,398 | 1,001,357 | 32,432 | 46 |
| Bs | 4,214,630 | 951,273 | 1,703,309 | 86,372 | 226 |
| Ec | 4,639,675 | 1,072,074 | 1,125,653 | 36,395 | 247 |
| Hi | 1,830,023 | 722,860 | 294,353 | 12,158 | 91 |
| Hp | 1,667,825 | 564,308 | 336,122 | 19,276 | 75 |
| Lc | 3,079,196 | 1,126,363 | 502,861 | 13,083 | 246 |
| Ll | 2,365,589 | 764,006 | 507,490 | 25,667 | 183 |
| Mg | 1,664,957 | 246,342 | 66,324 | 2,737 | 28 |
| Sa | 2,814,816 | 755,483 | 704,147 | 32,054 | 138 |
| Sp | 2,209,198 | 904,815 | 327,713 | 10,390 | 234 |
| Xc | 5,148,708 | 804,034 | 1,746,214 | 179,346 | 633 |

Table: Number of minimal absent words of lengths 11, 14, 17, and 24 in the genomes of thirteen bacteria.

| *Arabidopsis thaliana* chr. | Size (bp) | MAW (s) | PFG (s) |
|---|---|---|---|
| 1 | 30,427,671 | 40.20 | 51.90 |
| 2 | 19,698,289 | 25.86 | 32.94 |
| 3 | 23,459,830 | 30.84 | 42.30 |
| 4 | 18,585,056 | 24.65 | 31.42 |
| 5 | 26,975,502 | 35.38 | 48.91 |
| *Drosophila melanogaster* chr. | Size (bp) | MAW (s) | PFG (s) |
| 2L | 23,011,544 | 30.01 | 40.85 |
| 2R | 21,146,708 | 27.52 | 38.38 |
| 3L | 24,543,557 | 32.00 | 45.13 |
| 3R | 27,905,053 | 36.44 | 48.36 |
| X | 22,422,827 | 29.38 | 40.09 |

Table: Elapsed-time comparison of MAW and PFG for computing all minimal absent words in the genome of *Arabidopsis thaliana* and *Drosophila melanogaster*




Elapsed-time comparison of MAW and PFG for computing all minimal absent words in the genome of *Arabidopsis thaliana* and *Drosophila melanogaster*

## Comparison using synthetic data

| Sequence | Size (bp) | MAW (s) | PFG (s) |
|---|---|---|---|
| $S_1$ | 248,956,422 | 435.63 | 746.93 |
| $S_2$ | 248,956,422 | 438.52 | 733.69 |
| $S_3$ | 248,956,422 | 444.62 | 726.34 |
| $S_4$ | 248,956,422 | 444.06 | 743.29 |
| $S_5$ | 248,956,422 | 449.25 | 741.01 |

Table: Elapsed-time comparison of MAW and PFG for computing all minimal absent words in synthetic data

We created five instances from chr 1 of Hs: $S_1$, $S_2$, $S_3$, $S_4$ and $S_5$. Choosing randomly $10\%, 20\%, 30\%, 40\%, 50\%$ of the positions and randomly replacing them to one of {A,C,G,T}

## Our contribution

- A $\mathcal{O}(n)$-time and $\mathcal{O}(n)$-space algorithm for computing all minimal absent words based on the suffix array;
- An available implementation that outperforms existing tools: `http://github.com/solonas13/maw`
- Carl Barton, Alice Heliou, Laurent Mouchard and Solon P. Pissis Linear-time Computation of Minimal Absent Words Using Suffix Array, BMC Bioinformatics 2014

## Our contribution

- A $\mathcal{O}(n)$-time and $\mathcal{O}(n)$-space algorithm for computing all minimal absent words based on the suffix array;
- An available implementation that outperforms existing tools: `http://github.com/solonas13/maw`
- Carl Barton, Alice Heliou, Laurent Mouchard and Solon P. Pissis Linear-time Computation of Minimal Absent Words Using Suffix Array, BMC Bioinformatics 2014

## Perspectives

- Implementation for symmetric multiprocessing systems;
- Implement a fast space-efficient solution based on the construction of compressed full-text indexes (Belazzougi et al. ESA 2013).

# Garcia and Pinho, 2011, Minimal Absent Words in Four Human Genome Assemblies

**Table 1.** Four human genome assemblies.

| | GRCh37 | | HuRef | | NA12878 | | YH | |
|---|---|---|---|---|---|---|---|---|
| | noRC | withRC | noRC | withRC | noRC | withRC | noRC | withRC |
| **Sequencing** | capillary-based | | ABI3730xl | | Illumina | | Illumina | |
| **Assembly** | PHRAP & GigAssembler | | Celera | | ALLPATHS-LG | | SOAPdenovo | |
| **Fragment type** | chromosomes | | chromosomes | | scaffolds | | scaffolds | |
| Genome size (bp) | 2,861,327,131 | 5,722,654,262 | 2,782,339,374 | 5,564,678,748 | 2,613,381,835 | 5,226,763,670 | 2,218,539,040 | 4,437,078,080 |
| Number of MAWs | 4,217,129,944 | 8,317,669,642 | 4,155,779,040 | 8,235,214,304 | 3,962,196,417 | 7,861,209,250 | 3,546,060,591 | 7,059,225,195 |
| Longest MAW (bp) | 67,633 | 119,821 | 9,385 | 31,117 | 9,769 | 34,342 | 1,281 | 1,657 |

GRCh37 is the reference human genome assembly build 37.1, HuRef is the genome of Craig Venter, NA12878 is the human genome assembly from cell line GM12878, and YH is the genome of a Han Chinese individual. Genome size is the number of A,C,G and T base pairs (bp). The number of minimal absent words (MAWs) indicates the total number of minimal absent words in the assembly. The noRC columns display results without considering the reversed complement and the withRC columns display results considering the reversed complement.

doi:10.1371/journal.pone.0029344.t001