Maximal Exponent Repeats

MAXIME CROCHEMORE

King's College London **Université Paris-Est**

&

joint work with

CHALITA TOOPSUWAN & GOLNAZ BADKOBEH

Stringologists' Mascot?

Stringocephalus [Wikipedia, 2011]

- ⋆ extinct genus; between 360 to 408 million years old
- ⋆ usually found as fossils in Devonian marine rocks
- ⋆ found in western North America, northern Europe (especially Poland), Asia and the Canning Basin of Western Australia

A beautiful mind!

[based on fake etymology]

Repeats

- **String = text = word = sequence of symbols**

- **Repetition = periodic string = power of exponent** ≥ 2

$$\text{length} = 17$$

`a b a a b a b a a b a b a a b a b`

$$\text{period} = 5$$

$$\text{exponent} = \frac{\text{length}}{\text{period}} = \frac{17}{5} = 3.4$$

Repeats

- **String = text = word = sequence of symbols**

- **Repetition = periodic string = power of exponent** ≥ 2

  `abaab abaab abaab ab` $= ($`abaab`$)^{17/5}$

Repeats

- **String = text = word = sequence of symbols**

- **Repetition = periodic string = power of exponent ≥ 2**

$$\texttt{abaab abaab abaab ab} = (\texttt{abaab})^{17/5}$$

$$\texttt{alfalfa} = (\texttt{alf})^{7/3} \qquad \texttt{entente} = (\texttt{ent})^{7/3}$$

Repeats

- **String = text = word = sequence of symbols**

- **Repetition = periodic string = power of exponent** ≥ 2

  $\texttt{abaab abaab abaab ab} = (\texttt{abaab})^{17/5}$

  $\texttt{alfalfa} = (\texttt{alf})^{7/3}$ $\qquad$ $\texttt{entente} = (\texttt{ent})^{7/3}$

- **Repeat:** $1 <$ **exponent** ≤ 2

  length $= 15$

  a b a a b c c c c c a b a a b

  period $= 10$ $\qquad$ border

$$\text{exponent} = \frac{\text{length}}{\text{period}} = 1 + \frac{\text{border}}{\text{period}} = \frac{15}{10} = 1.5$$

Repeats

⋆ **String = text = word = sequence of symbols**

⋆ **Repetition = periodic string = power of exponent** ≥ 2

$$\texttt{abaab abaab abaab ab} = (\texttt{abaab})^{17/5}$$

$$\texttt{alfalfa} = (\texttt{alf})^{7/3} \qquad \texttt{entente} = (\texttt{ent})^{7/3}$$

⋆ **Repeat:** $1 <$ **exponent** ≤ 2

$$\texttt{abaab ccccc abaab} = (\texttt{abaabccccc})^{15/10}$$

Repeats

- String = text = word = sequence of symbols

- **Repetition** = periodic string = power of exponent ≥ 2

$$\texttt{abaab abaab abaab ab} = (\texttt{abaab})^{17/5}$$

$$\texttt{alfalfa} = (\texttt{alf})^{7/3} \qquad \texttt{entente} = (\texttt{ent})^{7/3}$$

- **Repeat:** $1 <$ exponent ≤ 2

$$\texttt{abaab ccccc abaab} = (\texttt{abaabccccc})^{15/10}$$

$$\texttt{restore} = (\texttt{resto})^{7/5} \qquad \texttt{all in all} = (\texttt{all in })^{10/7}$$

Motivation

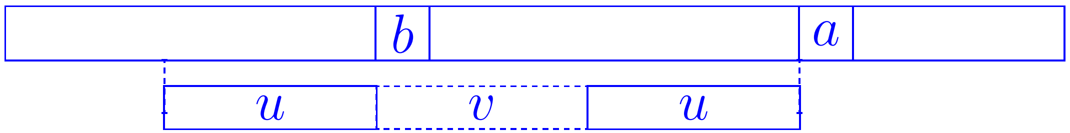
- **Combinatorics on words**
  Avoidability of repetitions, Interaction between periods, Counting repetitions
- **Pattern matching algorithms**
  String Matching, Time-space optimal String Matching: local and global periods, Indexing
- **Text Compression**
  Generalised run-length encoding
  Dictionary-based compression
- **Analysis of biological molecular sequences**
  Intensive study of satellites, Simple Sequence Repeats, or Tandem Repeats in DNA sequences
  Molecular structure prediction
- **Analysis of music**
  Rhythm detection, Chorus location

Maximal Exponent

- ⋆ **String y of length n drawn from a fixed alphabet maximal exponent of all factors of y?**
- ⋆ **RUN: maximal periodicity in y (exponent ≥ 2)**
- ⋆ **Linear number of runs, linear-time computation on fixed alphabet [Kolpakov, Kucherov, 1999]**
- ⋆ **Question 1: Compute the maximal exponent of all repeats in an overlap-free string**

Maximal-Exponent Repeats

- **MER: maximal exponent repeat occurring in y**
  a MER occurrence is maximal



  `abacada.........axayaza`
  $\Omega(n^2)$ **maximal repeats but** $\lfloor \frac{n}{2} \rfloor$ **MER occ. of exponent** $\frac{3}{2}$

- **Question 2: locate all MER occurrences in an overlap-free string**

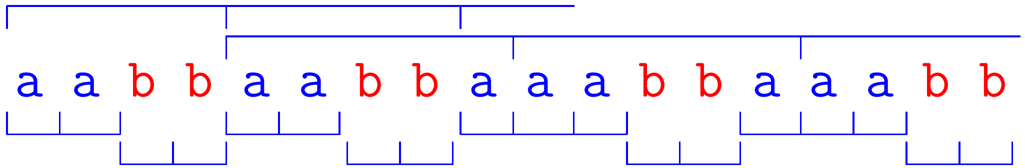
Theorem 1 ([Badkobeh, C., Toopsuwan, 2012]) ***All the occurrences of maximal-exponent repeats in an overlap-free string over a fixed alphabet can be listed in linear time.***

All powers

- ⋆ **Finding all occurrences of powers efficiently**
  - **Problem: too many occurrences**
  - **Solution: select some, encode them in a compact form**
- ⋆ **All primitively-rooted right-maximal integer exponent:** $O(n \log n)$ **time [C., 1981]**
- ⋆ **All primitively-rooted right-maximal:** $O(n \log n)$ **time [Apostolico, Preparata, 1983]**
- ⋆ **All primitively-rooted maximal:** $O(n \log n)$ **time [Main, Lorentz, 1985]**
- ⋆ **All leftmost maximal:** $O(n \log a)$ **time [Main, 1989] extension of [C., 1983]**
- ⋆ **All runs in Fibonacci strings:** $O(n)$ **time [Iliopoulos, Moore, Smyth, 1997]**

Computing runs

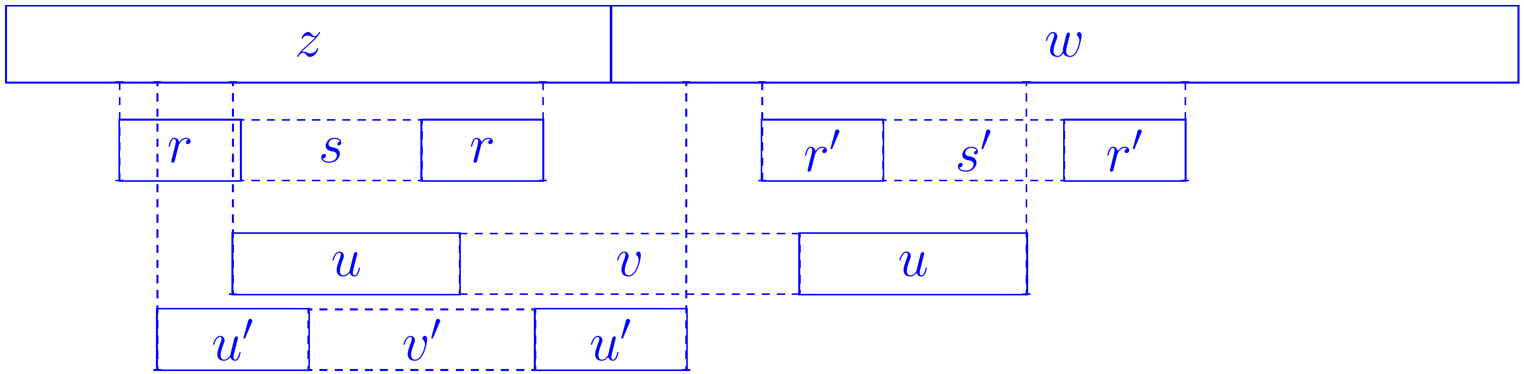
* Run: maximal periodicity
  Runs encode all powers



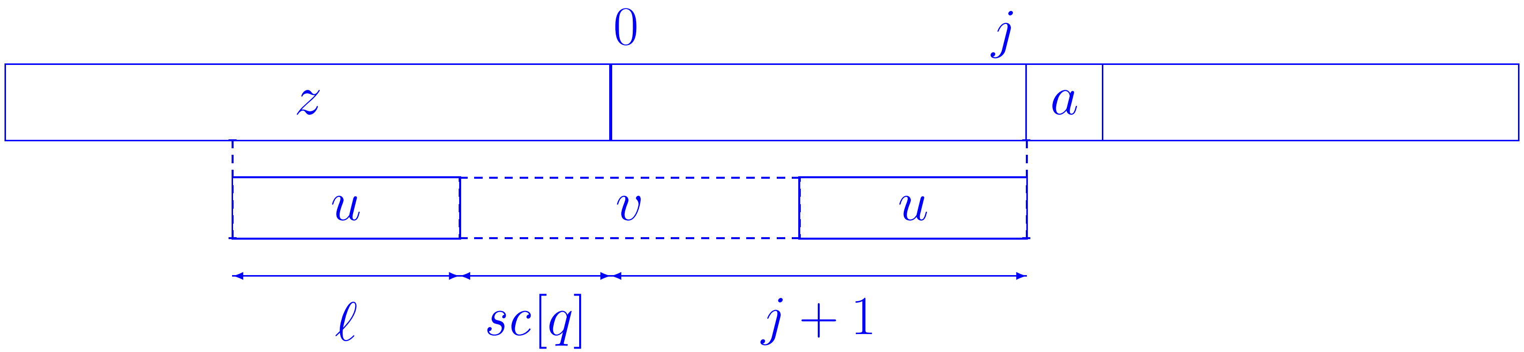
* Computation in $O(n \log a)$ time
  [Kolpakov, Kucherov, 1999]
* .. based on
  - modified Main's algorithm
  - f-factorization (kind of Ziv-Lempel factorization)
  - linear upper bound on the number of runs
* Explicit best known bound:
  $\text{runs}(n) \leq \mathbf{1.029}n$ [C., Ilie, Tinta, 2008]
  $\text{runs}(n) \geq \mathbf{0.944}n$ [Simpson, 2009]

Computing the maximal exponent

- **Naive algorithm**
  border/period/exponent in linear time (Morris-Pratt algo)
  yields $O(n^3)$ time solution, reducible to $O(n^2)$
- **Divide and conquer**

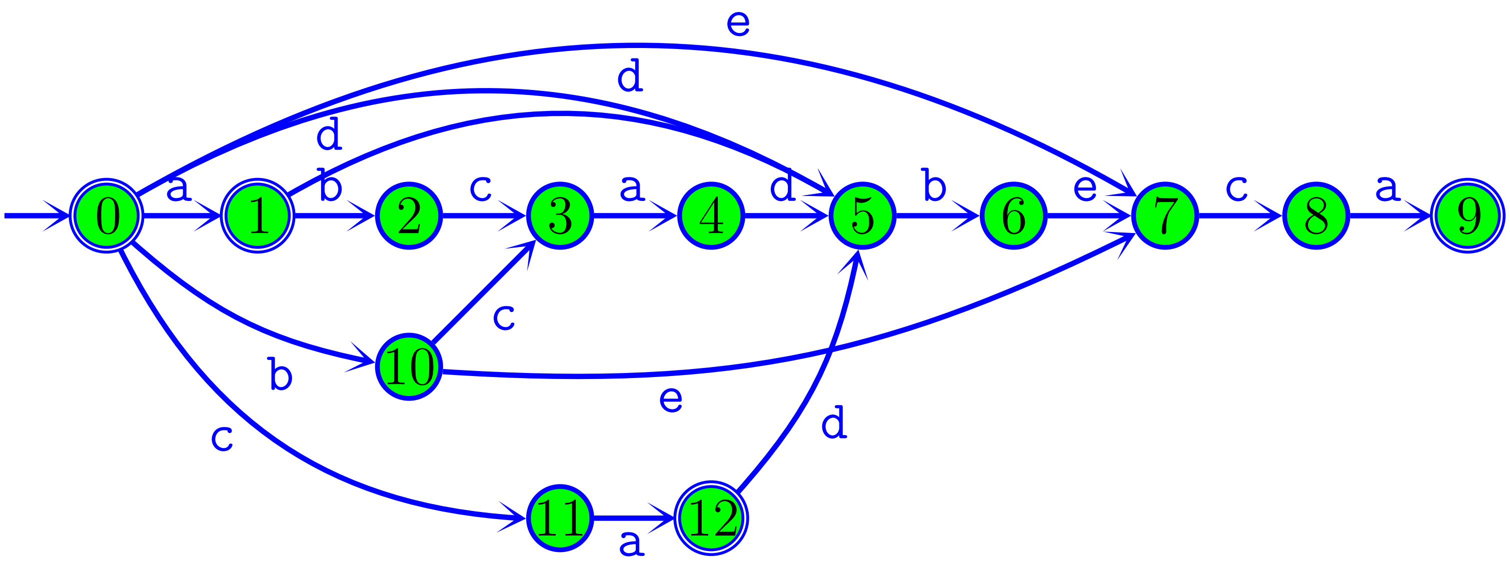


Repeat in a product



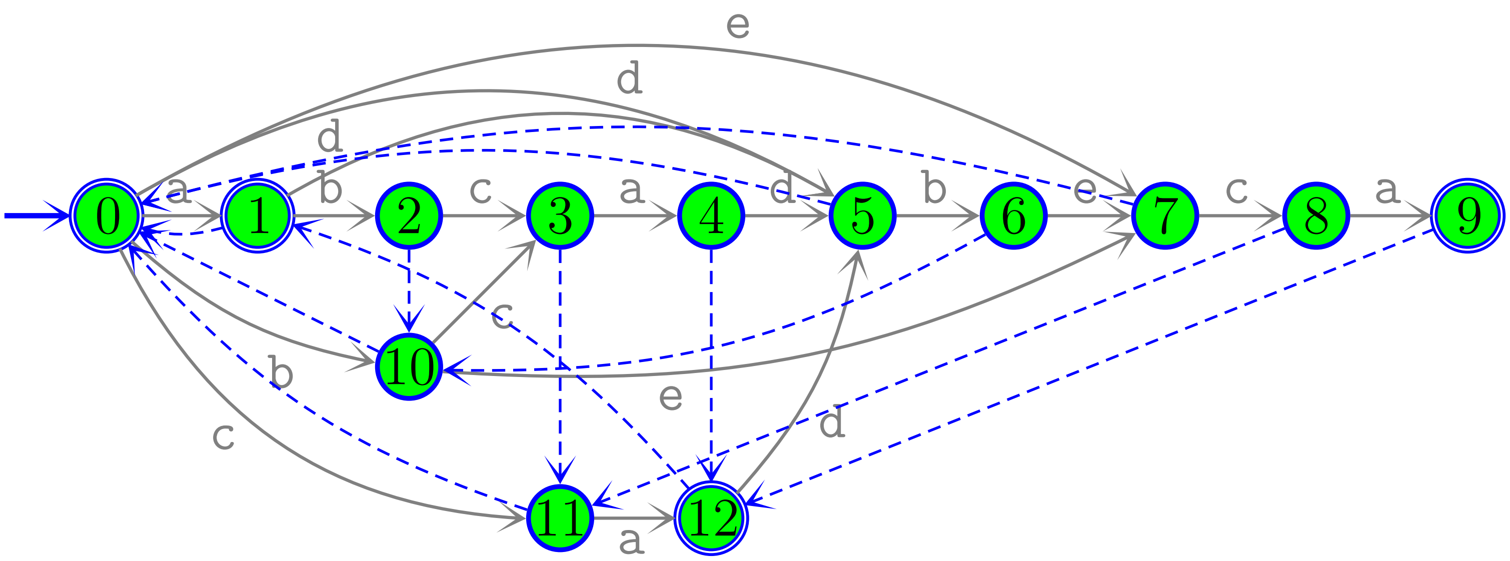
- **With the Suffix Automaton of z:**
  - u **longest factor of** z **ending at** j; $\ell = |u|$
  - **state** $q = \delta(\textbf{initial}, u)$
    $sc[q]$ **locates the last occurrence of** u **in** z
  - **exponent** $= \frac{\ell+sc[q]+j+1}{sc[q]+j+1}$
- **failure links on states to locate suffixes of** u

Suffix Automaton



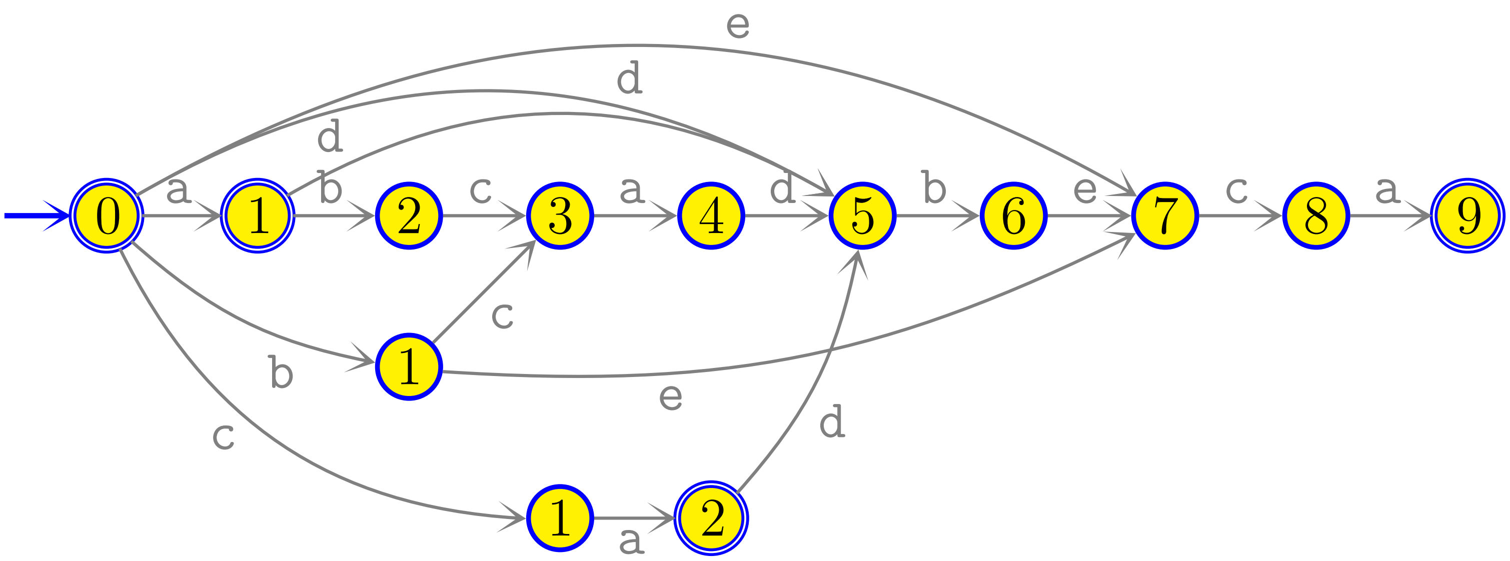
⋆ **Used to locate rightmost occurrences of border u in z**

Suffix Automaton



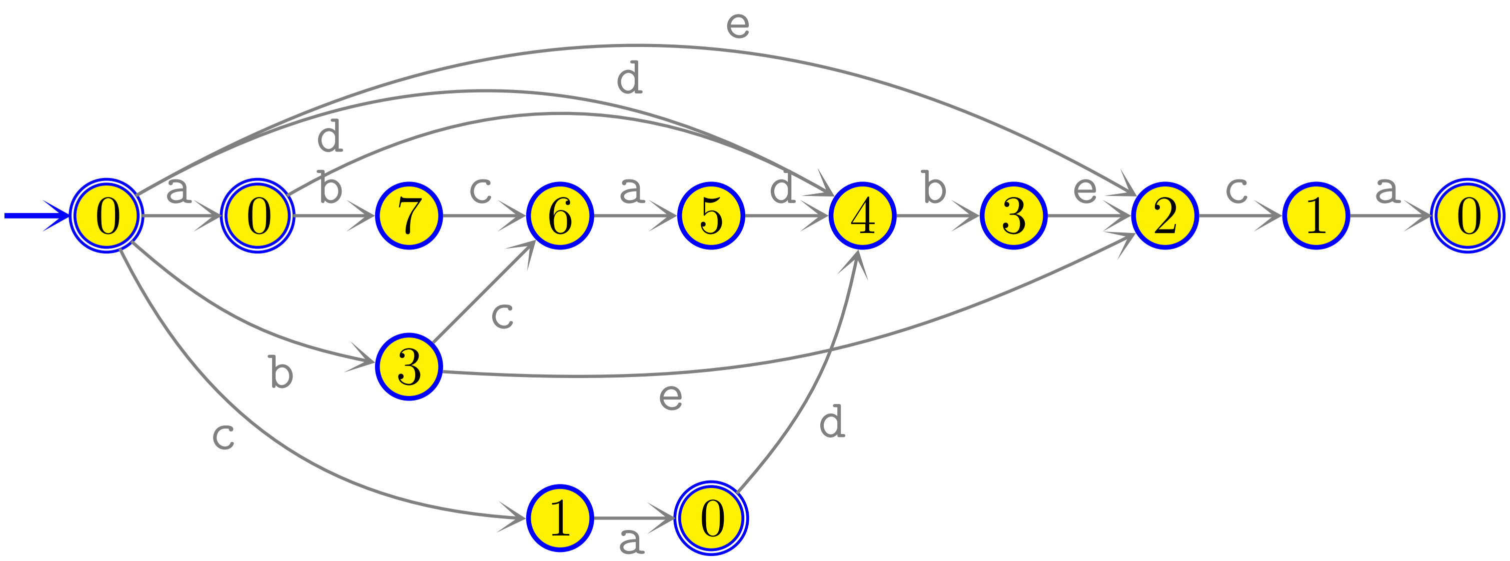
- ⋆ **Used to locate rightmost occurrences of border u in z**
- ⋆ **Equipped with: Failure links**

Suffix Automaton



- ⋆ **Used to locate rightmost occurrences of border u in z**
- ⋆ **Equipped with: Failure links**
  $L[q]$ = **maximal length of words reaching state** q

Suffix Automaton



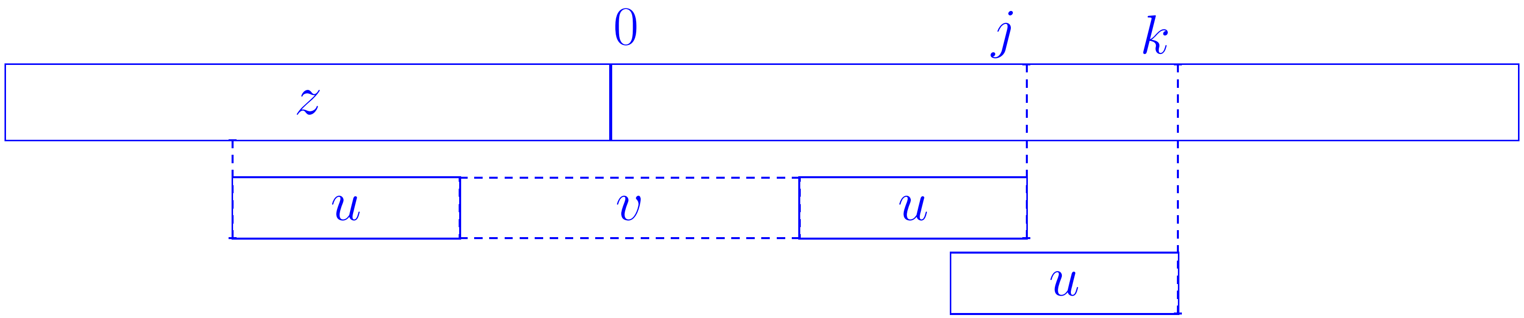
- ⋆ **Used to locate rightmost occurrences of border u in z**
- ⋆ **Equipped with:**
  Failure links
  $L[q]$ = **maximal length of words reaching state** q
  $sc[q]$ = **shortest length to a terminal state**

Core algorithm

MaxExp(z, w, e)

$\mathcal{S} \leftarrow$ ***Suffix Automaton of*** z
mark $\text{initial}(\mathcal{S})$
$(q, \ell) \leftarrow (F[\text{last}(\mathcal{S})], L[F[\text{last}(\mathcal{S})]])$
for $j \leftarrow 0$ **to** $\min\{\lfloor |z|/(e-1) - 1 \rfloor, |w| - 1\}$ **do**
  while $\text{goto}(q, w[j]) = \text{NIL}$ **and** $q \neq \text{initial}(\mathcal{S})$ **do**
    $(q, \ell) \leftarrow (F[q], L[F[q]])$
  if $\text{goto}(q, w[j]) \neq \text{NIL}$ **then**
    $(q, \ell) \leftarrow (\text{goto}(q, w[j]), \ell + 1)$
    $(q', \ell') \leftarrow (q, \ell)$
    while q' **unmarked do**
      $e \leftarrow \max\{e, (\ell' + sc[q'] + j + 1)/(sc[q'] + j + 1)\}$
      if $\ell' = L[q']$ **then**
        mark q'
      $(q', \ell') \leftarrow (F[q'], L[F[q']])$
return e

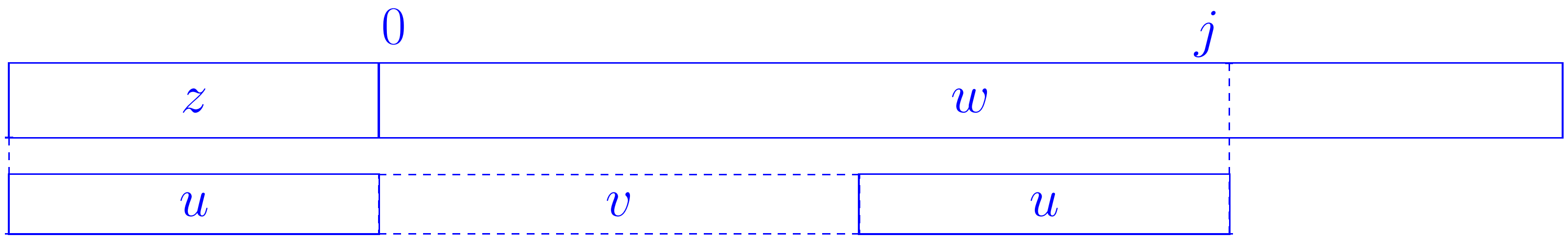
Marking states



- **At** j**, state** q **is marked if** u **longest with** $q = \delta(\textbf{initial}, u)$
- **then no more exponent computation if** q **met later**
- **it happens when a failure link is used**
  then no more than $2|z|$ **extra exponent computations**

Runtime

- Linear number of exponent computations due to marking



- j needs not be larger than $|z|/(e-1)-1$ (e current exponent)
- for long enough y, $e \geq \mathbf{RT}(a)$ then

$$j \leq \frac{|z|}{\mathbf{RT}(a)-1} - 1$$

- $O(|z|)$ time to compute exponents $\geq \mathbf{RT}(a)$ in zw independent of $|w|$

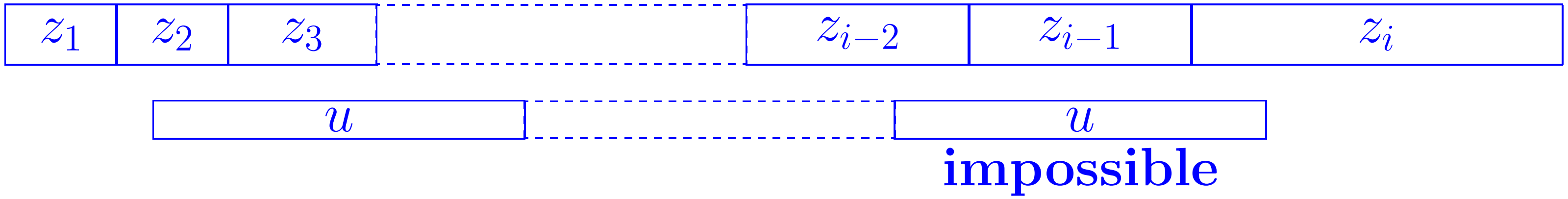
Maximal exponent

- Balanced divide and conquer: $O(n \log n)$
- Use of the f-factorisation: $O(n)$
- Phrase = longest factor occurring before (no overlap)
- Example of $y =$ abaababababaaababb

a b a a b a a a b a a a b a a b

- Computation with the suffix tree of y: $O(n \log a)$ time [Storer, Szymanski, 1982]
- .. however possible in linear time with the suffix array of y [C., Tischler, 2010], extension of [C., Ilie, 2008]

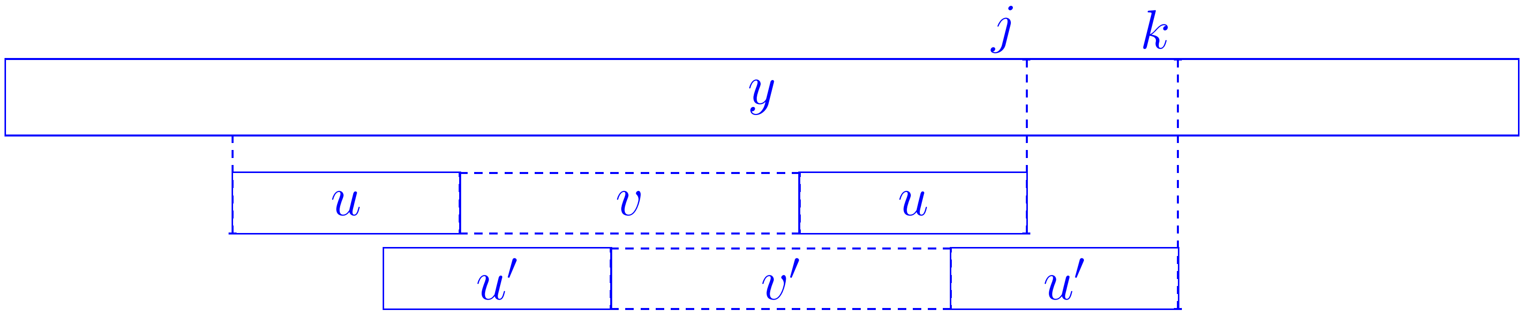
Maximal exponent



- **No phrase in the second occurrence of** u
- **for each** i
  - $\text{MaxExp}(z_{i-1}, z_i)$
  - $\text{MaxExp}(\widetilde{z_i}, \widetilde{z_{i-1}})$
  - $\text{MaxExp}(\widetilde{z_{i-1}} z_i, \widetilde{z_1 \cdots z_{i-2}})$
- **Running time:** $O(\Sigma z_i) = O(n)$

Theorem 2 (**[Badkobeh, C., Toopsuwan]**) ***The maximal exponent of repeats can be computed in linear time on a fixed alphabet.***

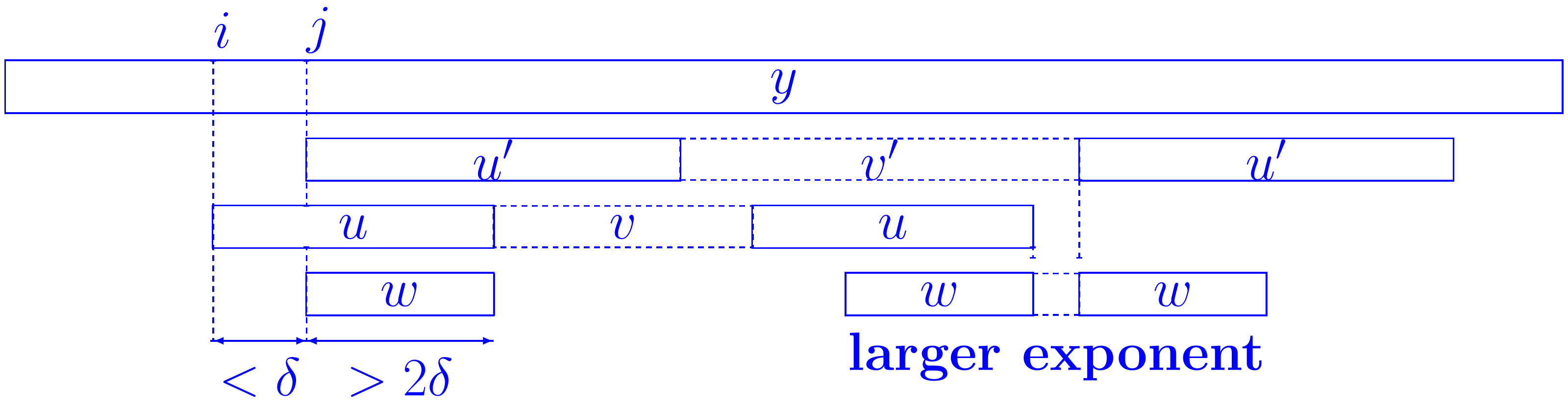
Counting MER occurrences (1)



- Impossible when uvu and $u'v'u'$ have the same border length: then $k - j > |u|$
- no more than $n/(b+1)$ MER occurrences of border length b
- total number of occurrences:

$$\leq \sum_{b=1}^{N} \frac{n}{b+1}$$

Counting MER occurrences (2)



- δ**-MER: MER whose border length satisfies** $3\delta \leq b < 4\delta$**.**
- **two** δ**-MER occurrences at** i **and** j**, then** $j - i \geq \delta$
- **total number of occurrences:**

$$\leq \sum_{\delta \in \Delta} \frac{n}{\delta} = n \left(3 + \frac{3}{2} + 1 + \frac{3}{4} + \left(\frac{3}{4}\right)^2 + \ldots \right) < 8.5\, n.$$

Counting MER occurrences (3)

- **Combining**
- **for border lengths up to** 11

$$\leq \sum_{b=1}^{11} \frac{n}{b+1} = 2.103211\, n$$

- **for border length from** 12**,** $\Gamma = \{4, 4(4/3), 4(4/3)^2, \ldots\}$,

$$\leq \sum_{\delta \in \Gamma} \frac{n}{\delta} = \frac{1}{4}\left(1 + \frac{3}{4} + \left(\frac{3}{4}\right)^2 + \ldots\right) n = n$$

Theorem 3 ***There are less than*** $3.11\, n$ ***occurrences of MERs in a string of length*** n***.***

- **Consequence: linear computation of all MER occurrences with upgraded algorithm**

Conclusion and questions

* **Linear computation of MER occurrences and of runs on a fixed alphabet**
* **MER computation in the comparison model?**
  Note: optimal $O(n \log n)$ time algorithm for runs
  [C., Rytter, Tyczyński, 2012]
* **Exact bound on the maximal number of MER occurrences?**
  less than n? tested up to length 20 on alphabet sizes 2, 3 and 4

Conclusion and questions

- **Linear computation of MER occurrences and of runs on a fixed alphabet**
- **MER computation in the comparison model?**
  Note: optimal $O(n \log n)$ time algorithm for runs
  [**C., Rytter, Tyczyński, 2012**]
- **Exact bound on the maximal number of MER occurrences?**
  less than n? tested up to length 20 on alphabet sizes 2, 3 and 4
- **2 is a threshold exponent:**
  above, at most a linear number of runs
  below, possible quadratic number of maximal occurrences of repeats, but at most a linear number of MER occurrences
- **Any other threshold?**
  Note: no more than $\frac{1}{\epsilon} n \ln n$ maximal repetitions of exponent more than $1 + \epsilon$ [**Kolpakov, Kucherov, Ochem, 2010**]